Artificial Intelligence 1
Winter Semester 2023/24

– Lecture Notes –

Prof. Dr. Michael Kohlhase

Professur für Wissensrepräsentation und -verarbeitung
Informatik, FAU Erlangen-Nürnberg
Michael.Kohlhase@FAU.de

2023-09-20

0.1 Preface

0.1.1 Course Concept

Objective: The course aims at giving students a solid (and often somewhat theoretically oriented) foundation of the basic concepts and practices of artificial intelligence. The course will predominantly cover symbolic AI – also sometimes called "good old-fashioned AI (GofAI)" – in the first semester and offers the very foundations of statistical approaches in the second. Indeed, a full account sub symbolic, machine learning based AI deserves its own specialization courses and needs much more mathematical prerequisites than we can assume in this course.

Context: The course "Artificial Intelligence" (AI 1 & 2) at FAU Erlangen is a two-semester course in the "Wahlpflichtbereich" (specialization phase) in semesters 5/6 of the Bachelor program "Computer Science" at FAU Erlangen. It is also available as a (somewhat remedial) course in the "Vertiefungsmodul Künstliche Intelligenz" in the Computer Science Master's program.

Prerequisites: AI-1 & 2 builds on the mandatory courses in the FAU Bachelor's program, in particular the course "Grundlagen der Logik in der Informatik" [Glo], which already covers a lot of the materials usually presented in the "knowledge and reasoning" part of an introductory AI course. The AI 1& 2 course also minimizes overlap with the course.

The course is relatively elementary, we expect that any student who attended the mandatory CS courses at FAU Erlangen can follow it.

Open to external students:

Other Bachelor programs are increasingly co-opting the course as specialization option. There is no inherent restriction to computer science students in this course. Students with other study biographies – e.g. students from other Bachelor programs our external Master's students should be able to pick up the prerequisites when needed.

0.1.2 Course Contents

Goal: To give students a solid foundation of the basic concepts and practices of the field of Artificial Intelligence. The course will be based on Russell/Norvig's book "*Artificial Intelligence; A modern Approach*" [RN09]

Artificial Intelligence I (the first semester): introduces AI as an area of study, discusses "rational agents" as a unifying conceptual paradigm for AI and covers problem solving, search, constraint propagation, logic, knowledge representation, and planning.

Artificial Intelligence II (the second semester): is more oriented towards exposing students to the basics of statistically based AI: We start out with reasoning under uncertainty, setting the foundation with Bayesian Networks and extending this to rational decision theory. Building on this we cover the basics of machine learning.

0.1.3 This Document

Format: The document mixes the slides presented in class with comments of the instructor to give students a more complete background reference.

Caveat: This document is made available for the students of this course only. It is still very much a draft and will develop over the course of the current course and in coming academic years. **Licensing:** **Knowledge Representation Experiment:** This document is also an experiment in knowledge representation. Under the hood, it uses the sTeX package [Koh08; sTeX], a TeX/LaTeX extension for semantic markup, which allows to export the contents into active documents that adapt to the reader and can be instrumented with services based on the explicitly represented meaning of the documents.

0.1.4 Acknowledgments

Materials: Most of the materials in this course is based on Russel/Norvik's book "Artificial Intelligence — A Modern Approach" (AIMA [RN95]). Even the slides are based on a LaTeX-based slide set, but heavily edited. The section on search algorithms is based on materials obtained from Bernhard Beckert (then Uni Koblenz), which is in turn based on AIMA. Some extensions have been inspired by an AI course by Jörg Hoffmann and Wolfgang Wahlster at Saarland University in 2016. Finally Dennis Müller suggested and supplied some extensions on AGI. Florian Rabe, Max Rapp and Katja Berčič have carefully re-read the text and pointed out problems.

All course materials have bee restructured and semantically annotated in the STeX format, so that we can base additional semantic services on them.

AI Students: The following students have submitted corrections and suggestions to this and earlier versions of the notes: Rares Ambrus, Ioan Sucan, Yashodan Nevatia, Dennis Müller, Simon Rainer, Demian Vöhringer, Lorenz Gorse, Philipp Reger, Benedikt Lorch, Maximilian Lösch, Luca Reeb, Marius Frinken, Peter Eichinger, Oskar Herrmann, Daniel Höfer, Stephan Mattejat, Matthias Sonntag, Jan Urfei, Tanja Würsching, Adrian Kretschmer, Tobias Schmidt, Maxim Onciul, Armin Roth, Liam Corona, Tobias Völk, Lena Voigt, Yinan Shao, Michael Girstl, Matthias Vietz, Anatoliy Cherepantsev, Stefan Musevski, Matthias Lobenhofer, Philipp Kaludercic, Diwarkara Reddy, Martin Helmke, Stefan Müller, Dominik Mehlich, Paul Martini, Vishwang Dave, Arthur Miehlich, Christian Schabesberger, Vishaal Saravanan, Simon Heilig, Michelle Fribrance, Wenwen Wang, Xinyuan Tu, Lobna Eldeeb.

0.1.5 Recorded Syllabus

In this subsection, we record the progress of the course in the academic year 2023/24 in the form of a "recorded syllabus", i.e. a syllabus that is created after the fact rather than before. For the topics planned for this course, see subsection 0.1.2.

Syllabus – Winter 2023/24: The recorded syllabus for this semester is in the course page in the ALeA system at `https://courses.voll-ki.fau.de/course-home/ai-1`. The table of contents in the AI-1 notes at `https://courses.voll-ki.fau.de` indicates the material covered to date in yellow.

The recorded syllabus of AI-2 can be found at `https://courses.voll-ki.fau.de/course-home/ai-2`

Contents

Chapter 1

Preliminaries

In this chapter, we want to get all the organizational matters out of the way, so that we can get into the discussion of artificial intelligence content unencumbered. We will talk about the necessary administrative details, go into how students can get most out of the course, talk about where the various resources provided with the course can be found, and finally introduce the ALeA system, an experimental – using AI methods – learning support system for the AI course.

1.1 Administrative Ground Rules

We will now go through the ground rules for the course. This is a kind of a social contract between the instructor and the students. Both have to keep their side of the deal to make learning as efficient and painless as possible.

Prerequisites for AI-1

- ▷ **Content Prerequisites:** The mandatory courses in CS@FAU; Sem 1-4, in particular:
  - ▷ Course "Algorithmen und Datenstrukturen". (Algorithms & Data Structures)
  - ▷ Course "Grundlagen der Logik in der Informatik" (GLOIN). (Logic in CS)
  - ▷ Course "Berechenbarkeit und Formale Sprachen". (Theoretical CS)
- ▷ **Skillset Prerequisite:** Coping with mathematical formulation of the structures
  - ▷ Mathematics is the language of science (in particular computer science)
  - ▷ It allows us to be very precise about what we mean. (good for you)
- ▷ **Intuition:** (take them with a kilo of salt)
  - ▷ This is what I assume you know! (I have to assume something)
  - ▷ In most cases, the dependency on these is partial and "in spirit".
  - ▷ If you have not taken these (or do not remember), read up on them as needed!
- ▷ **Real Prerequisites:** Motivation, interest, curiosity, hard work.(AI-1 is non-trivial)
- ▷ You can do this course if you want! (and I hope you are successful)

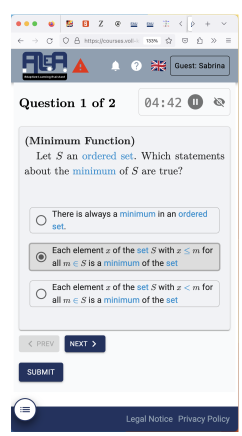
Now we come to a topic that is always interesting to the students: the grading scheme.

Assessment, Grades

- **Overall (Module) Grade:**
  - Grade via the exam (Klausur) $\leadsto$ 100% of the grade.
  - Up to 10% bonus on-top for an exam with $\geq 50\%$ points.($\leq 50\%$ $\leadsto$ no bonus)
  - Bonus points $\widehat{=}$ percentage sum of the best 10 tuesday quizzes divided by 100.
- **Exam:** 90 minutes exam conducted in presence on paper (~ April 1. 2024)
- **Retake Exam:** 90 min exam six months later (~ October 1. 2024)
- ⚠ You have to register for exams in campo in the first month of classes.
- **Note:** You can de-register from an exam on campo up to three working days before.
- **Tuesday Quizzes:** Every tuesday we start the lecture with a 10 min online quiz – the tuesday quiz – about the material from the previous week. (starts in week 2)

Tuesday Quizzes

- **Tuesday Quizzes:** Every tuesday we start the lecture with a 10 min online quiz – the tuesday quiz – about the material from the previous week. (starts in week 2)
- **Motivations:** We do this to
  - keep you prepared and working continuously. (primary)
  - update the ALeA learner model (fringe benefit)
- The tuesday quiz will be given in the ALeA system
  - `https://courses.voll-ki.fau.de/ai-1/quiz`
  - You have to be logged into ALeA!
  - You can take the quiz on your laptop or phone, ...
  - ... in the lecture or at home ...
  - ... via WLAN or 4G Network. (do not overload)
  - Quizzes will only be available 16:15-16:25!



Tomorrow: Pretest

- ⚠ Tomorrow we will try out the tuesday quiz infrastructure with a pretest!
  - **Presence**: bring your laptop or cellphone.
  - **Online**: you can and should take the pretest as well.
  - Have a recent firefox or chrome (chrome: $\geq$ March 2023)
  - Make sure that you are logged into ALeA (via FAU IDM; see below)
- **Definition 1.1.1.** A pretest is an assessment for evaluating the preparedness of learners for further studies.
- **Concretely:** This pretest
  - establishes a baseline for the competency expectations in AI-1 and
  - tests the ALeA quiz infrastructure for the tuesday quizzes.
- Participation in this test is optional; it will not influence your grades in any way.
- The test covers the prerequisites of AI-1 and some of the material that may have been covered in other courses.
- The test will be also used to refine the ALeA learner model, which may make learning experience in ALeA better. (see below)

Due to the current AI hype, the course Artificial Intelligence is very popular and thus many degree programs at FAU have adopted it for their curricula. Sometimes the course setup that fits for the CS program does not fit the other's very well, therefore there are some special conditions. I want to state here.

⚠ Special Admin Conditions ⚠

- Some degree programs do not "import" the course Artificial Intelligence, and thus you may not be able to register for the exam via `https://campus.fau.de`.
  - Just send me an e-mail and come to the exam, we will issue a "Schein".
  - Tell your program coordinator about AI-1/2 so that they remedy this situation
- In "Wirtschafts-Informatik" you can only take AI-1 and AI-2 together in the "Wahlpflicht-bereich".
  - ECTS credits need to be divisible by five $\leadsto 7.5 + 7.5 = 15$.

I can only warn of what I am aware, so if your degree program lets you jump through extra hoops, please tell me and then I can mention them here.

1.2 Getting Most out of AI-1

1.2.1 I

n this subsection we will discuss a couple of measures that students may want to consider to get most out of the AI-1 course.

None of them – homeworks, tutorials, study groups, and attendance – are mandatory, but most of them are very clearly correlated with success (i.e. passing the exam and getting a good grade).

AI-1 Homework Assignments

- ▷ **Homework Assignments:** Small individual problem/programming/proof task
  - ▷ but take time to solve (at least read them directly ⇝ questions)
- ▷ ⚠ Homeworks give no bonus points, but without trying you are unlikely to pass the exam.
- ▷ **Homework/Tutorial Discipline:**
  - ▷ Start early! (many assignments need more than one evening's work)
  - ▷ Don't start by sitting at a blank screen (talking & study group help)
  - ▷ Humans will be trying to understand the text/code/math when grading it.
  - ▷ Go to the tutorials, discuss with your TA! (they are there for you!)
- ▷ ⚠ We will not be able to grade all homework assignments!
- ▷ **Graded Assignments:** To keep things running smoothly
  - ▷ Homeworks will be posted on StudOn.
  - ▷ Sign up for AI-1 under `https://www.studon.fau.de/crs4622069.html`.
  - ▷ Homeworks are handed in electronically there. (plain text, program files, PDF)
  - ▷ Do not sign up for the "AI-2 Übungen" on StudOn (we do not use them)
- ▷ **Ungraded Assignments:** Are peer-feedbacked in ALeA (see below)

It is very well-established experience that without doing the homework assignments (or something similar) on your own, you will not master the concepts, you will not even be able to ask sensible questions, and take very little home from the course. Just sitting in the course and nodding is not enough! If you have questions please make sure you discuss them with the instructor, the teaching assistants, or your fellow students. There are three sensible venues for such discussions: online in the lecture, in the tutorials, which we discuss now, or in the course forum – see below. Finally, it is always a very good idea to form study groups with your friends.

Tutorials for Artificial Intelligence 1

- ▷ **Approach:** Weekly tutorials and homework assignments (first one in week two)
- ▷ **Goal 1:** Reinforce what was taught in class. (you need practice)

- **Goal 2:** Allow you to ask any question you have in a protected environment.
- **Instructor/Lead TA:** Florian Rabe (KWARC Postdoc)
  - Room: 11.137 @ Händler building, `florian.rabe@fau.de`
- **Tutorials:** one each taught by Florian Rabe (lead); Joshua Chacko, Mahd Mantash, Ahmed Aboelela, Ilia Dudnik, and Jovial Silatsa Tchatchum
- **Life-saving Advice:** Go to your tutorial, and prepare for it by having looked at the slides and the homework assignments!
- **Caveat:** We cannot grade all submissions with 5 TAs and ~ 1000 students.
- **Also:** Group submission has not worked well in the past! (too many freeloaders)

Collaboration

- **Definition 1.2.1.** Collaboration (or cooperation) is the process of groups of agents working or acting together for common, mutual, or some underlying benefit, as opposed to working in competition for selfish benefit. In a collaboration, every agent contributes to the common goal.
- In learning situations, the benefit is "better learning outcomes".
- **Observation:** In collaborative learning, the overall result can be significantly better than in competitive learning.
- **Good Practice:** Form study groups. (long- or short-term)
  - ⚠ those learners who work most, learn most
  - ⚠ freeloaders – indicivuals who only watch – learn very little!
- It is OK to collaborate on homework assignments in AI-1! (no bonus points)
- Choose your study group well (We will (eventually) help via ALeA)

What I am going to go into next is – or should be – obvious, but there is an important point I want to make.

Do I need to attend the lectures

- Attendance is not mandatory for the AI-1 lecture
- There are two ways of learning AI-1: (both are OK, your mileage may vary)
  - Approach B: Read a Book
  - Approach I: come to the lectures, be involved, interrupt me whenever you have a question.

The only advantage of I over B is that books do not answer questions (yet! $\leftsquigarrow$ we are working on this in AI research)

- Approach S: come to the lectures and sleep does not work!
- **I really mean it:** If you come to class, be involved, ask questions, challenge me with comments, tell me about errors, . . .
  - I would much rather have a lively discussion than get through all the slides
  - You learn more, I have more fun (Approach B serves as a backup)
  - You may have to change your habits, overcome shyness, . . . (please do!)
- This is what I get paid for, and I am more expensive than most books (get your money's worth)

1.3 Learning Resources for AI-1

But what if you are not in a lecture or tutorial and want to find out more about the AI-1 topics?

Textbook, Handouts and Information, Forums, Videos

- **Textbook:** *Russel/Norvig: Artificial Intelligence, A modern Approach* [RN09].
  - basically "broad but somewhat shallow"
  - great to get intuitions on the basics of AI

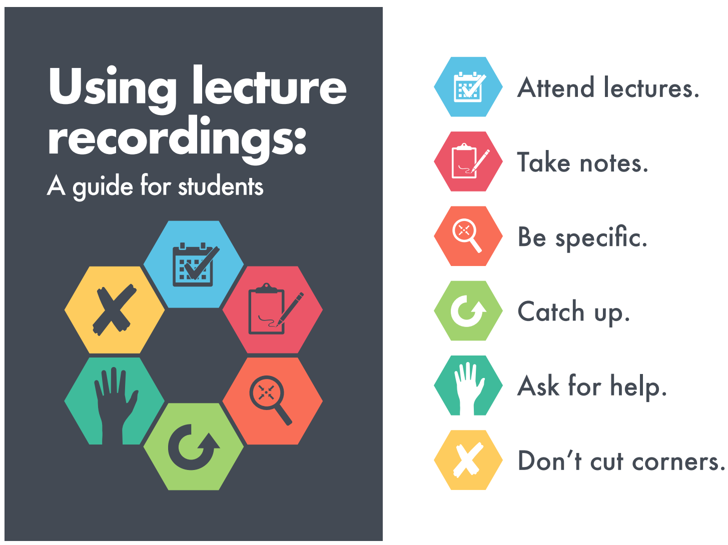
  Make sure that you read the edition ≥ 3 $\leftsquigarrow$ vastly improved over ≤ 2.
- **Course notes:** will be posted at `http://kwarc.info/teaching/AI/notes.pdf`
  - more detailed than [RN09] in some areas
  - I mostly prepare them as we go along (semantically preloaded $\rightsquigarrow$ research resource)
  - please e-mail me any errors/shortcomings you notice. (improve for the group)
- **StudOn Forum:** `https://www.studon.fau.de/crs4622069.html` for
  - announcements, homeworks (my view on the forum)
  - questions, discussion among your fellow students (your forum too, use it!)
- **Course Videos:** AI-1 will be streamed/recorded at `https://fau.tv/course/id/3180`
  - **Organized**: Video course nuggets are available at `https://fau.tv/course/id/1690` (short; organized by topic)
  - **Backup**: The lectures from WS 2016/17 to SS 2018 have been recorded (in English and German), see `https://www.fau.tv/search/term.html?q=Kohlhase`
- **Do not let the videos mislead you:** Coming to class is highly correlated with passing the course!

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 10 2023-09-20 SOME RIGHTS RESERVED

FAU has issued a very insightful guide on using lecture recordings. It is a good idea to heed these recommendations, even if they seem annoying at first.

Practical recommendations on Lecture Resources

▷ Excellent Guide: [Nor+18a] (german Version at [Nor+18b])



FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 11 2023-09-20 SOME RIGHTS RESERVED

1.4 AI-Supported Learning

In this section we introduce the ALeA (Adaptive Learning Assistant) system, a learning support system we have developed using symbolic AI methods – the stuff we learn about in AI-1 – and which we will use to support students in the course. As such ALeA does double duty in this course it supports learning activities and serves as a showcase, what methods can to in an important application.

ALeA: Adaptive Learning Assistant

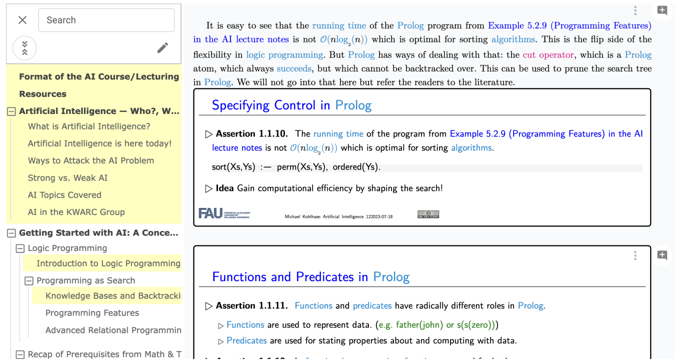
▷ **Idea:** Use AI methods to help teach/learn AI (AI4AI)

▷ **Concretely:** Provide HTML versions of the AI-1 slides/notes and embed learning support services into them. (for pre/postparation of lectures)

▷ **Definition 1.4.1.** Call a document active, iff it is interactive and adapts to specific information needs of the readers. (course notes on steroids)

▷ **Intuition:** ALeA servies active course materials. (PDF mostly inactive)

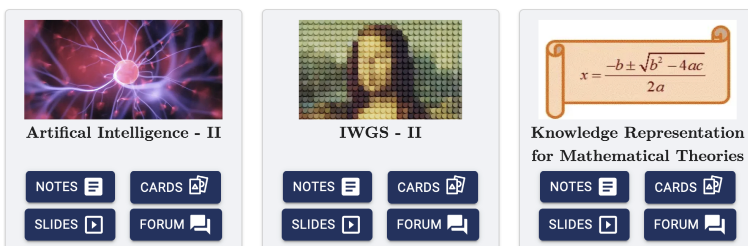
▷ **Example 1.4.2 (Course Notes).** $\widehat{=}$ Slides + Comments



⇝ yellow parts in table of contents (left) already covered in lecture.

VoLL-KI Portal at `https://courses.voll-ki.fau.de`

▷ **Portal for ALeA Courses:** `https://courses.voll-ki.fau.de`



▷ **AI-1 in ALeA:** `https://courses.voll-ki.fau.de/course-home/ai-1`

- ▷ All details for the course.
- ▷ recorded syllabus (keep track of material covered in course)
- ▷ syllabus of the last semester (for over/preview)

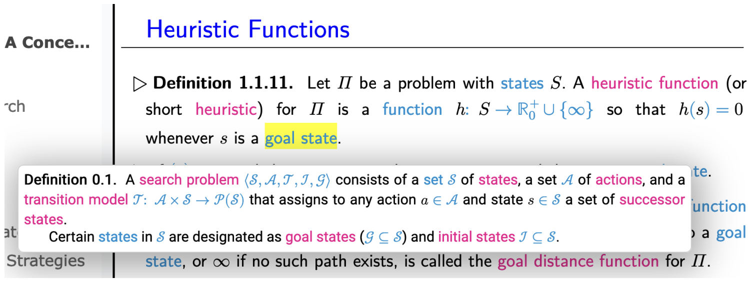
▷ **ALeA Status:** The ALEA system is deployed at FAU for over 1000 students taking six courses

- ▷ (some) students use the system actively (our logs tell us)
- ▷ reviews are mostly positive/enthusiastic (error reports pour in)

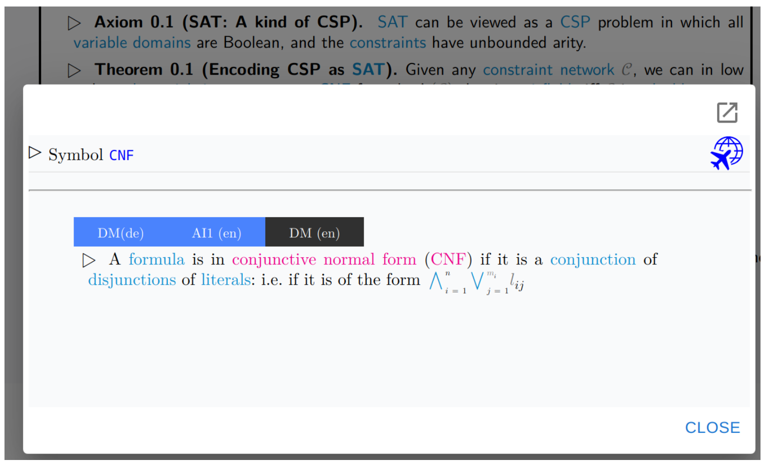
Learning Support Services in ALEA

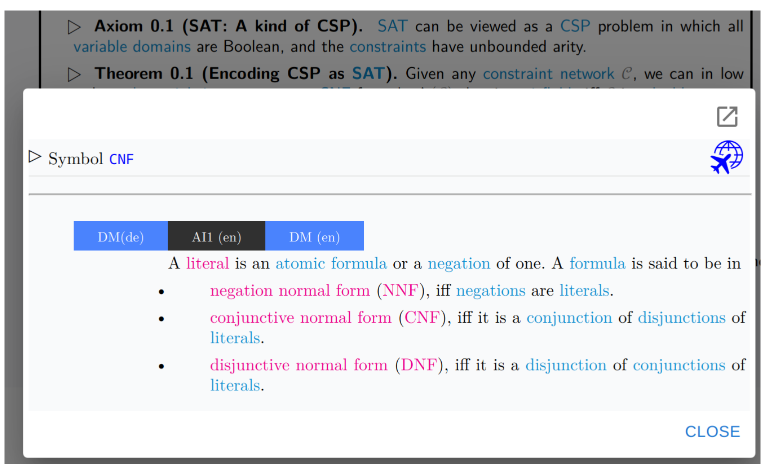
▷ **Idea:** Embed learning support services into active course materials.

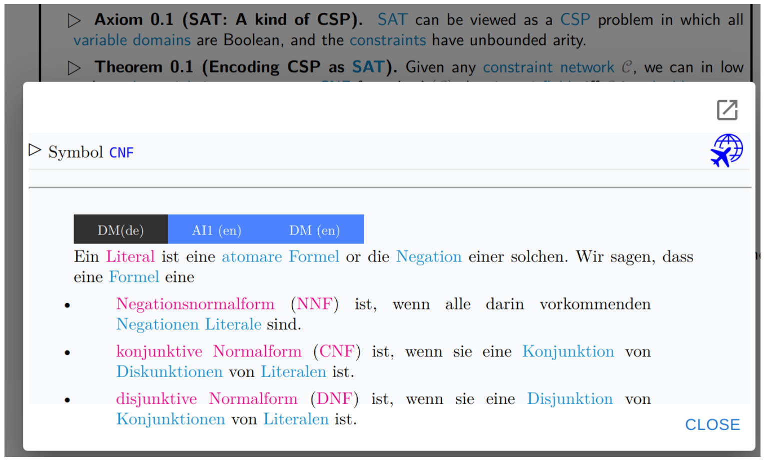
▷ **Example 1.4.3 (Definition on Hover).** Hovering on a (cyan) term reference reminds us of the definition. (even works recursively)



▷ **Example 1.4.4 (More Definitions on Click).** Clicking on a (cyan) term reference shows us more definitions from other contexts.







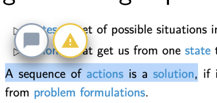
▷ **Example 1.4.5 (Guided Tour).** A guided tour for a concept c assembles definitions/etc. into a self-cont

c = countable ⇝

▷ . . . your idea here . . . (the sky is the limit)

Localized Interactions with the Community

▷ Selecting text brings up localized – ancored on selection – interactions:



▷ post a (public) comment or take (private) note

▷ report an error to the course authors/instructors

▷ Localized comments induce a thread in the ALeA forum (like the StudOn Forum, but targeted towards specific learning objects)

▷ Answering questions gives karma $\widehat{=}$ a public measure of helpfulness

▷ Notes can be anonymous ($\leadsto$ generate no karma)

Let us briefly look into how the learning support services introduced above might work, focusing on where the necessar information might come from.

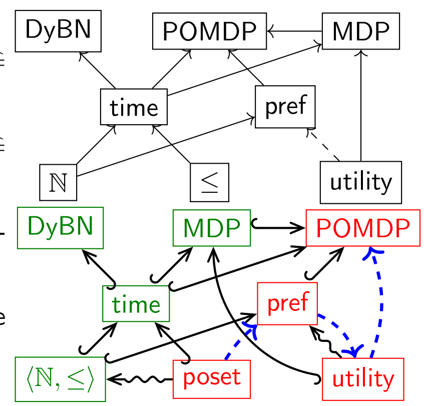
ALeA $\widehat{=}$ Data-Driven & AI-enabled Learning Assistance

▷ **Ingredient 1:** Domain model $\widehat{=}$ knowledge/theory graph

▷ **Ingredient 2:** Learner model $\widehat{=}$ adding competency estimations

▷ **Ingredient 3:** A collection of ready-formulated learning objects

▷ **Ingredient 4:** Educational dialogue planner $\leadsto$ guided tours



A theory graph provides (modular representation of the domain)

▷ symbols with URIs for all concepts, objects, and relations

▷ definitions, notations, and verbalizations for all symbols

▷ "object-oriented inheritance" and views between theories.

The learner model is a function from learner IDs $\times$ symbol URIs to competency values

▷ competency comes in six cognitive dimensions: `remember`, `understand`, `analyze`, `evaluate`, `apply`, and `create`.

▷ ALeA logs all learner interactions (keeps data learner-private)

▷ each interaction updates the learner model function.

Learning objects are the text fragments learners see and interact with; they are structured by

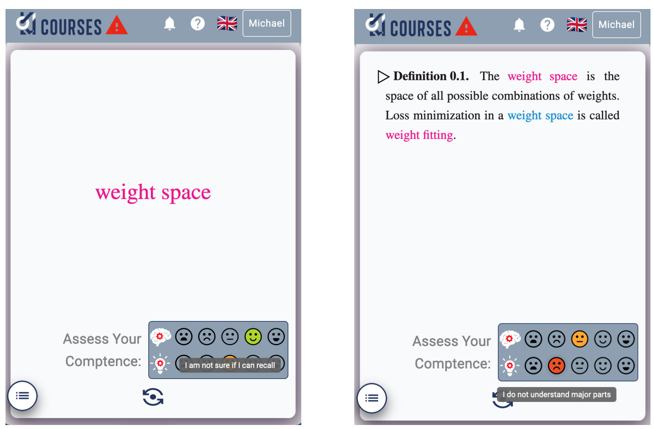
- ▷ didactic relations, e.g. tasks have prerequisites and learning objectives
- ▷ rhetoric relations, e.g. introduction, elaboration, and transition

The dialogue planner assembles learning objects into active course materials using

- ▷ the domain model and didactic relations to determine the order of LOs
- ▷ the learner model to determine what to show
- ▷ the rhetoric relations to make the dialogue coherent

New Feature: Drilling with Flashcards

- ▷ Flashcards challenge you with a task (term/problem) on the front...



...and the definition/answer is on the back.

- ▷ Self-assessment updates the learner model (before/after)
- ▷ **Bonus:** Flashcards can be generated from existing semantic markup (educational equivalent to free beer)

Learner Data and Privacy in ALeA

- ▷ **Observation:** Most learning support services in ALeA use the learner model; they
  - ▷ need the learner model data to adapt to the invidivual learner!
  - ▷ collect learner interaction data (to update the learner model)
- ▷ **Consequence:** You need to be logged in (via your FAU IDM credentials) for useful learning support services!

▷ **Problem:** Learner model data is highly sensitive personal data!

▷ **ALeA Promise:** The ALEA team does the utmost to keep your personal data safe. (SSO via FAU IDM/eduGAIN, ALEA trust zone)

▷ **ALeA Privacy Axioms:**

1. ALEA only collects learner models data about logged in users.
2. Personally identifiable learner model data is only accessible to its subject (delegation possible)
3. Learners can always query the learner model about its data.
4. All learner model data can be purged without negative consequences (except usability deterioration)
5. Logging into ALEA is completely optional.

▷ **Observation:** Authentication for bonus quizzes are somewhat less optional, but you can always purge the learner model later.

Chapter 2

Artificial Intelligence – Who?, What?, When?, Where?, and Why?

We start the course by giving an overview of (the problems, methods, and issues of) Artificial Intelligence, and what has been achieved so far.

Naturally, this will dwell mostly on philosophical aspects – we will try to understand what the important issues might be and what questions we should even be asking. What the most important avenues of attacks may be and where AI research is being carried out.

In particular the discussion will be very non-technical – we have very little basis to discuss technicalities yet. But stay with me, this will drastically change very soon. **A Video Nugget** covering the introduction of this chapter can be found at `https://fau.tv/clip/id/21467`.

Plot for this chapter

▷ Motivation, overview, and finding out what you already know

- ▷ What is Artificial Intelligence?
- ▷ What has AI already achieved?
- ▷ A (very) quick walk through the AI-1 topics.
- ▷ How can you get involved with AI at KWARC?

2.1 What is Artificial Intelligence?

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21701`.
The first question we have to ask ourselves is "What is Artificial Intelligence?", i.e. how can we define it. And already that poses a problem since the natural definition *like human intelligence, but artificially realized* presupposes a definition of Intelligence, which is equally problematic; even Psychologists and Philosophers – the subjects nominally "in charge" of human intelligence – have problems defining it, as witnessed by the plethora of theories e.g. found at [WHI].

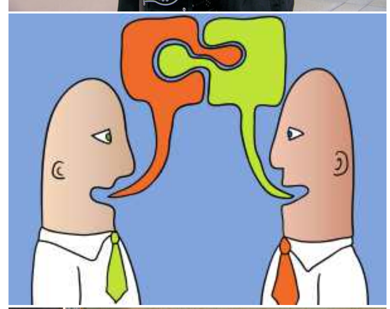
What is Artificial Intelligence? Definition

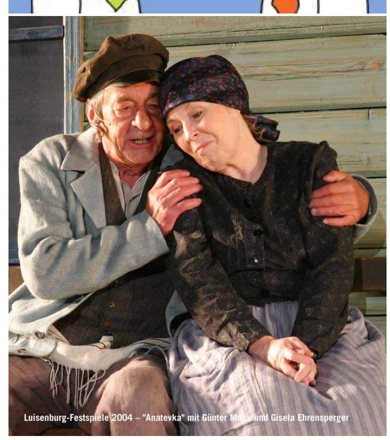
- ▷ **Definition 2.1.1 (According to Wikipedia).** Artificial Intelligence (AI) is intelligence exhibited by machines
- ▷ **Definition 2.1.2 (also).** Artificial Intelligence (AI) is a sub-field of computer science that is concerned with the automation of intelligent behavior.
- ▷ **BUT:** it is already difficult to define "Intelligence" precisely
- ▷ **Definition 2.1.3 (Elaine Rich).** Artificial Intelligence (AI) studies how we can make the computer do things that humans can still do better at the moment.

Maybe we can get around the problems of defining "what Artificial intelligence is", by just describing the necessary components of AI (and how they interact). Let's have a try to see whether that is more informative.

What is Artificial Intelligence? Components

- **Elaine Rich:** AI studies how we can make the computer do things that humans can still do better at the moment.
- This needs a combination of
  - the ability to learn
  - inference
  - perception
  - language understanding
  - emotion



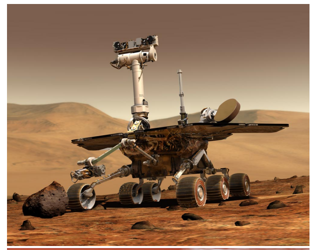


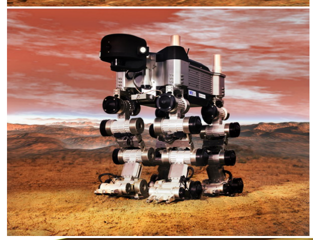
2.2 Artificial Intelligence is here today!

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21697`.

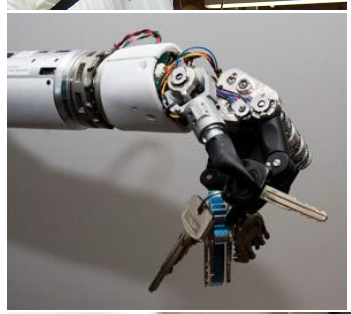
The components of Artificial Intelligence are quite daunting, and none of them are fully understood, much less achieved artificially. But for some tasks we can get by with much less. And indeed that is what the field of Artificial Intelligence does in practice – but keeps the lofty ideal around. This practice of "trying to achieve AI in selected and restricted domains" (cf. the discussion starting with slide 29) has borne rich fruits: systems that meet or exceed human capabilities in such areas. Such systems are in common use in many domains of application.

Artificial Intelligence is here today!



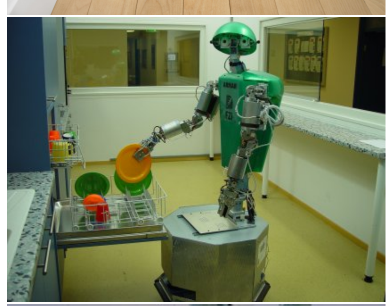


- ▷ in outer space
  - ▷ in outer space systems need autonomous control:
  - ▷ remote control impossible due to time lag

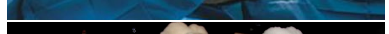


- ▷ in artificial limbs
  - ▷ the user controls the prosthesis via existing nerves, can e.g. grip a sheet of paper.

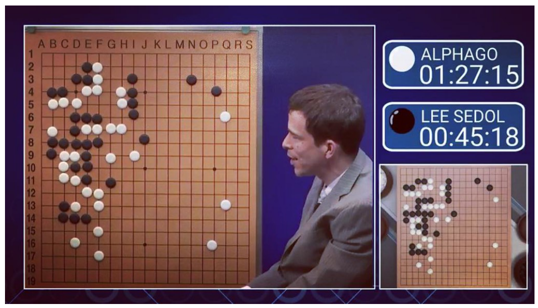
- ▷ in household appliances
  - ▷ The iRobot Roomba vacuums, mops, and sweeps in corners, . . . , parks, charges, and discharges.
  - ▷ general robotic household help is on the horizon.



- ▷ in hospitals
  - ▷ in the USA 90% of the prostate operations are carried out by RoboDoc
  - ▷ Paro is a cuddly robot that eases solitude in nursing homes.



And here's what you all have been waiting for . . .



CC-BY-SA: Buster Benson@
https://www.flickr.com/photos/erikbenson/25717574115

- ▷ AlphaGo is a program by Google DeepMind to play the board game go.
- ▷ In March 2016, it beat Lee Sedol in a five-game match, the first time a go program has beaten a 9 dan professional without handicaps. In December 2017 AlphaZero, a successor of AlphaGo "learned" the games go, chess, and shogi in 24 hours, achieving a superhuman level of play in these three games by defeating world-champion programs. By September 2019, AlphaStar, a variant of AlphaGo, attained "grandmaster level" in Starcraft II, a real time strategy game with partially observable state. AlphaStar now among the top 0.2% of human players.

We will conclude this section with a note of caution.

The AI Conundrum

- ▷ **Observation:** Reserving the term "Artificial Intelligence" has been quite a land grab!
- ▷ **But:** researchers at the Dartmouth Conference (1956) really thought they would solve/reach AI in two/three decades.
- ▷ **Consequence:** AI still asks the big questions.
- ▷ **Another Consequence:** AI as a field is an incubator for many innovative technologies.
- ▷ **AI Conundrum:** Once AI solves a subfield it is called "computer science". (becomes a separate subfield of CS)
- ▷ **Example 2.2.1.** Functional/Logic Programming, automated theorem proving, Planning, machine learning, Knowledge Representation, . . .
- ▷ **Still Consequence:** AI research was alternatingly flooded with money and cut off brutally.

2.3 Ways to Attack the AI Problem

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21717`.
There are currently three main avenues of attack to the problem of building artificially intelligent systems. The (historically) first is based on the symbolic representation of knowledge about the world and uses inference-based methods to derive new knowledge on which to base action decisions. The second uses statistical methods to deal with uncertainty about the world state and learning methods to derive new (uncertain) world assumptions to act on.

Three Main Approaches to Artificial Intelligence

- ▷ **Definition 2.3.1.** Symbolic AI is based on the assumption that many aspects of intelligence can be achieved by the manipulation of symbols, combining them into structures (expressions) and manipulating them (using processes) to produce new expressions.
- ▷ **Definition 2.3.2.** Statistical AI remedies the two shortcomings of symbolic AI approaches: that all concepts represented by symbols are crisply defined, and that all aspects of the world are knowable/representable in principle. Statistical AI adopts sophisticated mathematical models of uncertainty and uses them to create more accurate world models and reason about them.
- ▷ **Definition 2.3.3.** Subsymbolic AI attacks the assumption of symbolic and statistical AI that intelligence can be achieved by reasoning about the state of the world. Instead it posits that intelligence must be embodied i.e. situated in the world, equipped with a "body" that can interact with it via sensors and actuators. The main method for realizing intelligent behavior is by learning from the world, i.e. machine learning.

As a consequence, the field of Artificial Intelligence (AI) is an engineering field at the intersection of computer science (logic, programming, applied statistics), cognitive science (psychology, neuroscience), philosophy (can machines think, what does that mean?), linguistics (natural language understanding), and mechatronics (robot hardware, sensors).
Subsymbolic AI and in particular machine learning is currently hyped to such an extent, that many people take it to be synonymous with "Artificial Intelligence". It is one of the goals of this course to show students that this is a very impoverished view.

Two ways of reaching Artificial Intelligence?

- ▷ We can classify the AI approaches by their coverage and the analysis depth (they are complementary)

Deep	symbolic AI-1	not there yet cooperation?
Shallow	no-one wants this	statistical/sub symbolic AI-2
Analysis ↑ vs. Coverage →	Narrow	Wide

- ▷ **This semester** we will cover foundational aspects of symbolic AI (deep/narrow processing)
- ▷ **next semester** concentrate on statistical/subsymbolic AI. (shallow/wide-coverage)

We combine the topics in this way in this course, not only because this reproduces the historical development but also as the methods of statistical and subsymbolic AI share a common basis.
It is important to notice that all approaches to AI have their application domains and strong points. We will now see that exactly the two areas, where symbolic AI and statistical/subsymbolic AI have their respective fortes correspond to natural application areas.

Environmental Niches for both Approaches to AI

- ▷ **Observation:** There are two kinds of applications/tasks in AI
  - ▷ Consumer tasks: consumer grade applications have tasks that must be fully generic and wide coverage. (e.g. machine translation like Google Translate)
  - ▷ Producer tasks: producer grade applications must be high-precision, but can be domain-specific (e.g. multilingual documentation, machinery-control, program verification, medical technology)

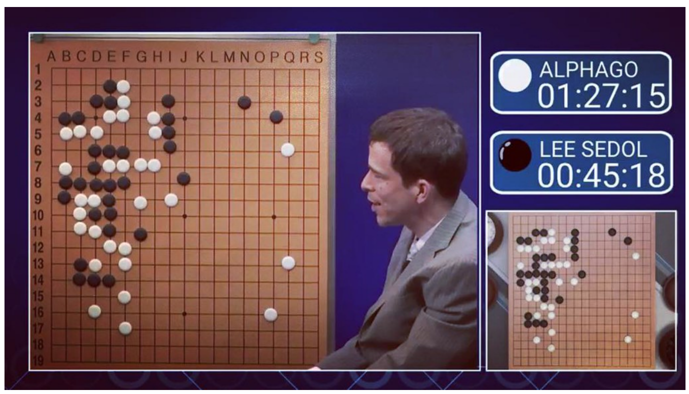
Precision			
100%	Producer Tasks		
50%		Consumer Tasks	
	$10^{3\pm1}$ Concepts	$10^{6\pm1}$ Concepts	**Coverage**

- ▷ **General Rule:** Subsymbolic AI is well suited for consumer tasks, while symbolic AI is better suited for producer tasks.
- ▷ A domain of producer tasks I am interested in: mathematical/technical documents.

An example of a producer task – indeed this is where the name comes from – is the case of a machine tool manufacturer T, which produces digitally programmed machine tools worth multiple million Euro and sells them into dozens of countries. Thus T must also comprehensive machine operation manuals, a non-trivial undertaking, since no two machines are identical and they must be translated into many languages, leading to hundreds of documents. As those manual share a lot of semantic content, their management should be supported by AI techniques. It is critical that these methods maintain a high precision, operation errors can easily lead to very costly machine damage and loss of production. On the other hand, the domain of these manuals is quite restricted. A machine tool has a couple of hundred components only that can be described by a comple of thousand attribute only.

Indeed companies like T employ high-precision AI techniques like the ones we will cover in this course successfully; they are just not so much in the public eye as the consumer tasks.

To get this out of the way . . .



- AlphaGo = search + neural networks (symbolic + subsymbolic AI)
  - we do search this semester and cover neural networks in AI-2.
  - I will explain AlphaGo a bit in chapter 7.

2.4 Strong vs. Weak AI

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21724`.
To get this out of the way before we begin: We now come to a distinction that is often muddled in popular discussions about "Artificial Intelligence", but should be cristal clear to students of the course AI-1 – after all, you are upcoming "AI-specialists".

Strong AI vs. Narrow AI

- **Definition 2.4.1.** With the term narrow AI (also weak AI, instrumental AI, applied AI) we refer to the use of software to study or accomplish *specific* problem solving or reasoning tasks (e.g. playing chess/go, controlling elevators, composing music, . . .)
- **Definition 2.4.2.** With the term strong AI (also full AI, AGI) we denote the quest for software performing at the full range of human cognitive abilities.
- **Definition 2.4.3.** Problems requiring strong AI to solve are called AI hard.

▷ **In short:** We can characterize the difference intuitively:

- ▷ narrow AI: What (most) computer scientists think AI is / should be.
- ▷ strong AI: What Hollywood authors think AI is / should be.

▷ **Needless to say** we are only going to cover narrow AI in this course!

One can usually defuse public worries about "is AI going to take control over the world" by just explaining the difference between strong AI and weak AI clearly.
I would like to add a few words on AGI, that – if you adopt them; they are not universally accepted – will strengthen the arguments differentiating between strong and weak AI.

A few words on AGI...

▷ The conceptual and mathematical framework (agents, environments etc.) is the same for strong AI and weak AI.

▷ AGI research focuses mostly on abstract aspects of machine learning (reinforcement learning, neural nets) and decision/game theory ("which goals should an AGI pursue?").

▷ Academic respectability of AGI fluctuates massively, recently increased (again). (correlates somewhat with AI winters and golden years)

▷ Public attention increasing due to talk of "existential risks of AI" (e.g. Hawking, Musk, Bostrom, Yudkowsky, Obama, ...)

▷ **Kohlhase's View:** Weak AI is here, strong AI is very far off. (not in my lifetime)
But even if that is true, weak AI will affect all of us deeply in everyday life.

▷ **Example 2.4.4.** You should not train to be an accountant or truck driver! (bots will replace you)

I want to conclude this section with an overview over the recent protagonists – both personal and institutional – of AGI.

AGI Research and Researchers

▷ "Famous" research(ers) / organizations

- ▷ MIRI (Machine Intelligence Research Institute), Eliezer Yudkowsky (Formerly known as "Singularity Institute")
- ▷ Future of Humanity Institute Oxford (Nick Bostrom),
- ▷ Google (Ray Kurzweil),
- ▷ AGIRI / OpenCog (Ben Goertzel),
- ▷ `petrl.org` (People for the Ethical Treatment of Reinforcement Learners). (Obviously somewhat tongue-in-cheek)

▷ ⚠ Be highly skeptical about any claims with respect to AGI! (Kohlhase's View)

2.5 AI Topics Covered

A Video Nugget covering this section can be found at https://fau.tv/clip/id/21719.

We will now preview the topics covered by the course "Artificial Intelligence" in the next two semesters.

Topics of AI-1 (Winter Semester)

- Getting Started
  - What is Artificial Intelligence? (situating ourselves)
  - Logic programming in Prolog (An influential paradigm)
  - Intelligent Agents (a unifying framework)
- Problem Solving
  - Problem Solving and search (Black Box World States and Actions)
  - Adversarial Search (Game playing) (A nice application of Search)
  - constraint satisfaction problems (Factored World States)
- Knowledge and Reasoning
  - Formal Logic as the mathematics of Meaning
  - Propositional logic and satisfiability (Atomic Propositions)
  - First-order logic and theorem proving (Quantification)
  - Logic programming (Logic + Search$\leadsto$ Programming)
  - Description logics and semantic web
- Planning
  - Planning Frameworks
  - Planning Algorithms
  - Planning and Acting in the real world

Topics of AI-2 (Summer Semester)

- Uncertain Knowledge and Reasoning
  - Uncertainty
  - Probabilistic reasoning
  - Making Decisions in Episodic Environments
  - Problem Solving in Sequential Environments
- Foundations of machine learning

- ▹ Learning from Observations
- ▹ Knowledge in Learning
- ▹ Statistical Learning Methods

▷ Communication (If there is time)

- ▹ Natural Language Processing
- ▹ Natural Language for Communication

AI1SysProj: A Systems/Project Supplement to AI-1

▷ The AI-1 course concentrates on concepts, theory, and algorithms of symbolic AI.

▷ **Problem:** Engineering/Systems Aspects of AI are very important as well.

▷ **Partial Solution:** Getting your hands dirty in the homeworks and the Kalah Challenge

▷ **Full Solution:** AI1SysProj: AI-1 Systems Project (10 ECTS, 30-50places)

- ▹ For each Topic of AI-1, where will be a mini-project in AI1SysProj
- ▹ e.g. for game-play there will be Chinese Checkers (more difficult than Kalah)
- ▹ e.g. for CSP we will schedule TechFak courses or exams (from real data)
- ▹ solve challenges by implementing the AI-1 algorithms or use SoA systems

▷ **Question:** Should I take AI1SysProj in my first semester? (i.e. now)

▷ **Answer:** It depends . . . (on your situation)

- ▹ most master's programs require a 10-ECTS "Master's Project"(Master AI: two)
- ▹ there will be a great pressure on project places (so reserve one early)
- ▹ BUT 10 ECTS $\widehat{=}$ 250-300 hours involvement by definition (1/3 of your time/ECTS)

▷ **BTW:** There will also be an AI2SysProj next semester! (another chance)

2.6 AI in the KWARC Group

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21725`.

Now allow me to beat my own drum. In my research group at FAU, we do research on a particular kind of Artificial Intelligence: logic, language, and information. This may not be the most fashionable or well-hyped area in AI, but it is challenging, well-respected, and – most importantly – fun.

The KWARC Research Group

- **Observation:** The ability to represent knowledge about the world and to draw logical inferences is one of the central components of intelligent behavior.
- **Thus:** reasoning components of some form are at the heart of many AI systems.
- **KWARC Angle:** Scaling up (web-coverage) without dumbing down (too much)
  - Content markup instead of full formalization (too tedious)
  - User support and quality control instead of "The Truth" (elusive anyway)
  - use Mathematics as a test tube (Mathematics $\widehat{=}$ Anything Formal)
  - care more about applications than about philosophy (we cannot help getting this right anyway as logicians)
- The KWARC group was established at Jacobs Univ. in 2004, moved to FAU Erlangen in 2016
- see `http://kwarc.info` for projects, publications, and links

Research in the KWARC group ranges over a variety of topics, which range from foundations of mathematics to relatively applied web information systems. I will try to organize them into three pillars here.

Overview: KWARC Research and Projects

<table>
<tr><td colspan="3">Applications: eMath 3.0, Active Documents, Active Learning, Semantic Spreadsheets/CAD/CAM, Change Mangagement, Global Digital Math Library, Math Search Systems, SMGloM: Semantic Multilingual Math Glossary, Serious Games, ...</td></tr>
<tr><td>Foundations of Math:
▷ MathML, OpenMath
▷ advanced Type Theories
▷ MMT: Meta Meta Theory
▷ Logic Morphisms/Atlas
▷ Theorem Prover/CAS Interoperability
▷ Mathematical Models/Simulation</td><td>KM & Interaction:
▷ Semantic Interpretation (aka. Framing)
▷ math-literate interaction
▷ MathHub: math archives & active docs
▷ Active documents: embedded semantic services
▷ Model-based Education</td><td>Semantization:
▷ LaTeXML: LaTeX → XML
▷ sTeX: Semantic LaTeX
▷ invasive editors
▷ Context-Aware IDEs
▷ Mathematical Corpora
▷ Linguistics of Math
▷ ML for Math Semantics Extraction</td></tr>
<tr><td colspan="3">Foundations: Computational Logic, Web Technologies, OMDoc/MMT</td></tr>
</table>

For all of these areas, we are looking for bright and motivated students to work with us. This can take various forms, theses, internships, and paid student assistantships.

Research Topics in the KWARC Group

- We are always looking for bright, motivated KWARCies.
- We have topics in for all levels! (Enthusiast, Bachelor, Master, Ph.D.)

- List of current topics: `https://gl.kwarc.info/kwarc/thesis-projects/`
  - Automated Reasoning: Maths Representation in the Large
  - Logics development, $(\text{Meta})^n$-Frameworks
  - Math Corpus Linguistics: Semantics Extraction
  - Serious Games, Cognitive Engineering, Math Information Retrieval, Legal Reasoning, ...
- We always try to find a topic at the intersection of your and our interests.
- We also often have positions!. (HiWi, Ph.D.: $\frac{1}{2}$, PostDoc: full)

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 37 2023-09-20

Sciences like physics or geology, and engineering need high-powered equipment to perform measurements or experiments. computer science and in particular the KWARC group needs high powered human brains to build systems and conduct thought experiments.

The KWARC group may not always have as much funding as other AI research groups, but we are very dedicated to give the best possible research guidance to the students we supervise.

So if this appeals to you, please come by and talk to us.

Part I

Getting Started with AI: A Conceptual Framework

This part of the course note sets the stage for the technical parts of the course by establishing a common framework (Rational Agents) that gives context and ties together the various methods discussed in the course.

After having seen what AI can do and where AI is being employed today (see chapter 2), we will now

1. introduce a programming language to use in the course,
2. prepare a conceptual framework in which we can think about "intelligence" (natural and artificial), and
3. recap some methods and results from theoretical computer science that we will need throughout the course.

ad 1. Prolog:
For the programming language w e choose Prolog, historically one of the most influential "AI programming languages". While the other AI programming language: `Lisp` which gave rise to the functional programming programming paradigm has been superseded by typed languages like SML, Haskell, `Scala`, and F#, Prolog is still the prime example of the declarative programming paradigm. So using Prolog in this course gives students the opportunity to explore this paradigm. At the same time, Prolog is well-suited for trying out algorithms in symbolic AI the topic of this semester since it internalizes the more complex primitives of the algorithms presented here.

ad 2. Rational Agents: The conceptual framework centers around rational agents which combine aspects of purely cognitive architectures (an original concern for the field of AI) with the more recent realization that intelligence must interact with the world (embodied AI) to grow and learn. The cognitive architectures aspect allows us to place and relate the various algorithms and methods we will see in this course. Unfortunately, the "situated AI" aspect will not be covered in this course due to the lack of time and hardware.

ad 3. Topics of Theoretical Computer Science: When we evaluate the methods and algorithms introduced in AI-1, we will need to judge their suitability as agent functions. The main theoretical tool for that is complexity theory; we will give a short motivation and overview of the main methods and results as far as they are relevant for AI-1 in section 4.1.

In the second half of the semester we will transition from search-based methods for problem solving to inference-based ones, i.e. where the problem formulation is described as expressions of a formal language which are transformed until an expression is reached from which the solution can be read off. Phrase structure grammars are the method of choice for describing such languages; we will introduce/recap them in section 4.2.

Enough philosophy about "Intelligence" (Artificial or Natural)

- ▷ So far we had a nice philosophical chat, about "intelligence" et al.
- ▷ As of today, we look at technical stuff!
- ▷ Before we go into the algorithms and data structures proper, we will
  1. introduce a programming language for AI-1
  2. prepare a conceptual framework in which we can think about "intelligence" (natural and artificial), and
  3. recap some methods and results from theoretical computer science.

Chapter 3

Logic Programming

We will now learn a new programming paradigm: logic programming, which is one of the most influential paradigms in AI. We are going to study Prolog (the oldest and most widely used) as a concrete example of ideas behind logic programming and use it for our homeworks in this course.
As Prolog is a representative of a programming paradigm that is new to most students, programming will feel weird and tedious at first. But subtracting the unusual syntax and program organization logic programming really only amounts to recursive programming just as in functional programming (the other declarative programming paradigm). So the usual advice applies, keep staring at it and practice on easy examples until the pain goes away.

3.1 Introduction to Logic Programming and ProLog

Logic programming is a programming paradigm that differs from functional and imperative programming in the basic procedural intuition. Instead of transforming the state of the memory by issuing instructions (as in imperative programming), or computing the value of a function on some arguments, logic programming interprets the program as a body of knowledge about the respective situation, which can be queried for consequences.

This is actually a very natural conception of program; after all we usually run (imperative or functional) programs if we want some question answered. **Video Nuggets** covering this section can be found at `https://fau.tv/clip/id/21752` and `https://fau.tv/clip/id/21753`.

.

Logic Programming

- ▷ **Idea:** Use logic as a programming language!
- ▷ We state what we know about a problem (the program) and then ask for results (what the program would compute).
- ▷ **Example 3.1.1.**

Program	Leibniz is human Sokrates is human Sokrates is a greek Every human is fallible	$x + 0 = x$ If $x + y = z$ then $x + s(y) = s(z)$ 3 is prime
Query	Are there fallible greeks?	is there a z with $s(s(0)) + s(0) = z$
Answer	Yes, Sokrates!	yes $s(s(s(0)))$

- **How to achieve this?** Restrict a logic calculus sufficiently that it can be used as computational procedure.
- **Remark:** This idea leads a totally new programming paradigm: logic programming.
- **Slogan:** Computation = Logic + Control (Robert Kowalski 1973; [Kow97])
- We will use the programming language Prolog as an example.

We now formally define the language of Prolog, starting off the atomic building blocks.

Prolog Programs: Terms and Literals

- **Definition 3.1.2.** Prolog programs express knowledge about the world via
  - constants denoted by lower case strings,
  - variables denoted by upper-case strings or starting with _, and
  - functions and predicates (lower-case strings) applied to terms.
- **Definition 3.1.3.** A Prolog term is
  - a Prolog variable, or constant, or
  - a Prolog function applied to terms.

  A Prolog literal is a constant or a predicate applied to terms.
- **Example 3.1.4.** The following are
  - Prolog terms: john, X, _, father(john), ...
  - Prolog literals: loves(john,mary), loves(john,_), loves(john,wife_of(john)),...

Now we build up Prolog programs from those building blocks.

Prolog Programs: Facts and Rules

- **Definition 3.1.5.** A Prolog program is a sequence of clauses, i.e.
  - facts of the form l., where l is a literal, (a literal and a dot)
  - rules of the form h:−b_1,...,b_n, where h is called the head literal (or simply head) and the b_i are together called the body of the rule.

  A rule h: b_1,...,b_n, should be read as *h (is true) if b_1 and ... and b_n are*.
- **Example 3.1.6.** The following is a Prolog program:

```
human(leibniz).
human(sokrates).
greek(sokrates).
fallible(X):−human(X).
```

The first three lines are Prolog facts and the last a rule.

▷ **Definition 3.1.7.** The knowledge base given by a Prolog program is the set of facts that can be derived from it under the if/and reading above.

Definition 3.1.7 introduces a very important distinction: that between a Prolog program and the knowledge base it induces. Whereas the former is a finite, syntactic object (essentially a string), the latter may be an infinite set of facts, which represents the totality of knowledge about the world or the aspects described by the program.

As knowledge bases can be infinite, we cannot pre compute them. Instead, logic programming languages compute fragments of the knowledge base by need; i.e. whenever a user wants to check membership; we call this approach querying: the user enters a query expression and the system answers yes or no. This answer is computed in a depth first search process.

Querying the Knowledge Base: Size Matters

▷ **Idea:** We want to see whether a fact is in the knowledge base.

▷ **Definition 3.1.8.** A query is a list of Prolog terms called goal literal (also subgoal or simply goals). We write a query as ?$-A_1$, …, A_n. where A_i are goals.

▷ **Problem:** Knowledge bases can be big and even infinite. (cannot pre compute)

▷ **Example 3.1.9.** The knowledge base induced by the Prolog program

```
nat(zero).
nat(s(X)) :− nat(X).
```

contains the facts nat(zero), nat(s(zero)), nat(s(s(zero))), …

Querying the Knowledge Base: Backchaining

▷ **Definition 3.1.10.** Given a query Q: ?− A_1, …, A_n. and rule R: h:− b_1,…,b_n, backchaining computes a new query by

1. finding terms for all variables in h to make h and A_1 equal and
2. replacing A_1 in Q with the body literals of R, where all variables are suitably replaced.

▷ Backchaining motivates the names goal/subgoal:

- ▹ the literals in the query are "goals" that have to be satisfied,
- ▹ backchaining does that by replacing them by new "goals".

▷ **Definition 3.1.11.** The Prolog interpreter keeps backchaining from the top to the bottom of the program until the query

- ▹ succeeds, i.e. contains no more goals, or (answer: **true**)
- ▹ fails, i.e. backchaining becomes impossible. (anser: false)

▷ **Example 3.1.12 (Backchaining).** We continue Example 3.1.9

```
?− nat(s(s(zero))).
?− nat(s(zero)).
?− nat(zero).
true
```


Note that backchaining replaces the current query with the body of the rule suitably instantiated. For rules with a long body this extends the list of current goals, but for facts (rules without a body), backchaining shortens the list of current goals. Once there are no goals left, the Prolog interpreter finishes and signals success by issuing the string **true**.
If no rules match the current goal, then the interpreter terminates and signals failure with the string false,

Querying the Knowledge Base: Failure

▷ If no instance of a query can be derived from the knowledge base, then the Prolog interpreter reports failure.

▷ **Example 3.1.13.** We vary Example 3.1.12 using 0 instead of zero.

```
?− nat(s(s(0))).
?− nat(s(0)).
?− nat(0).
FAIL
false
```


We can extend querying from simple yes/no answers to programs that return values by simply using variables in queries. In this case, the Prolog interpreter returns a substitution.

Querying the Knowledge base: Answer Substitutions

▷ **Definition 3.1.14.** If a query contains variables, then Prolog will return an answer substitution, i.e the values for all the query variables accumulated during repeated backchaining.

▷ **Example 3.1.15.** We talk about (Bavarian) cars for a change, and use a query with a variables

```
has_wheels(mybmw,4).
has_motor(mybmw).
car(X):−has_wheels(X,4),has_motor(X).
?− car(Y) % query
?− has_wheels(Y,4),has_motor(Y). % substitution X = Y
?− has_motor(mybmw). % substitution Y = mybmw
Y = mybmw % answer substitution
true
```


In Example 3.1.15 the first backchaining step binds the variable X to the query variable Y, which gives us the two subgoals has_wheels(Y,4),has_motor(Y). which again have the query variable Y.

The next backchaining step binds this to mybmw, and the third backchaining step exhausts the subgoals. So the query succeeds with the (overall) answer substitution Y = mybmw. With this setup, we can already do the "fallible Greeks" example from the introduction.

PROLOG: Are there Fallible Greeks?

▷ **Program:**

```
human(leibniz).
human(sokrates).
greek(sokrates).
fallible(X):—human(X).
```

▷ **Example 3.1.16 (Query).** ?—fallible(X),greek(X).

▷ **Answer substitution:** [sokrates/X]

3.2 Programming as Search

In this section, we want to really use Prolog as a programming language, so let use first get our tools set up.

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/21754` and `https://fau.tv/clip/id/21827`.

3.2.1 Knowledge Bases and Backtracking

We will now discuss how to use a Prolog interpreter to get to know the language. The SWI Prolog interpreter can be downloaded from `http://www.swi-prolog.org/`. To start the Prolog interpreter with pl or prolog or swipl from the shell. The SWI manual is available at `http://www.swi-prolog.org/pldoc/`

We will introduce working with the interpreter using unary natural numbers as examples: we first add the fact[1] to the knowledge base

```
unat(zero).
```

which asserts that the predicate unat[2] is **true** on the term zero. Generally, we can add a fact to the knowledge base either by writing it into a file (e.g. example.pl) and then "consulting it" by writing one of the following three commands into the interpreter:

```
[example]
consult('example.pl').
consult('example').
```

or by directly typing

```
assert(unat(zero)).
```

into the Prolog interpreter. Next tell Prolog about the following rule

```
assert(unat(suc(X)) :— unat(X)).
```

which gives the Prolog runtime an initial (infinite) knowledge base, which can be queried by

[1]for "unary natural numbers"; we cannot use the predicate nat and the constructor function s here, since their meaning is predefined in Prolog

[2]for "unary natural numbers".

```
?- unat(suc(suc(zero))).
```

Even though we can use any text editor to program Prolog, but running Prolog in a modern editor with language support is incredibly nicer than at the command line, because you can see the whole history of what you have done. Its better for debugging too. We will use `emacs` as an example in the following.

If you've never used `emacs` before, it still might be nicer, since its pretty easy to get used to the little bit of `emacs` that you need. (Just type "emacs \&" at the `UNIX` command line to run it; if you are on a remote terminal without visual capabilities, you can use "emacs −nw".).

If you don't already have a file in your home directory called ".emacs" (note the dot at the front), create one and put the following lines in it. Otherwise add the following to your existing .emacs file:

```
(autoload 'run-prolog "prolog" "Start a Prolog sub-process." t)
    (autoload 'prolog-mode "prolog" "Major mode for editing Prolog programs." t)
    (setq prolog-program-name "swipl"); or whatever the prolog executable name is
    (add-to-list 'auto-mode-alist '("\\pl$" . prolog-mode))
```

The file prolog.el, which provides prolog−mode should already be installed on your machine, otherwise download it at `http://turing.ubishops.ca/home/bruda/emacs-prolog/`

Now, once you're in `emacs`, you will need to figure out what your "meta" key is. Usually its the alt key. (Type "control" key together with "h" to get help on using `emacs`). So you'll need a "meta−X" command, then type "run−prolog". In other words, type the meta key, type "x", then there will be a little buffer at the bottom of your `emacs` window with "M−x", where you type run−prolog[3]. This will start up the SWI Prolog interpreter, . . . et voilà!

The best thing is you can have two buffers "within" your `emacs` window, one where you're editing your program and one where you're running Prolog. This makes debugging easier.

Depth-First Search with Backtracking

- So far, all the examples led to direct success or to failure. (simpl. KB)
- **Definition 3.2.1 (Prolog Search Procedure).** The Prolog interpreter employes top-down, left-right depth first search, concretely, Prolog search:
  - works on the subgoals in left right order.
  - matches first query with the head literals of the clauses in the program in top-down order.
  - if there are no matches, fail and backtrack to the (chronologically) last backtrack point.
  - otherwise backchain on the first match, keep the other matches in mind for backtracking via backtrack points.
- We can force backtracking to get more solutions by typing ;.

Note: With the Prolog search procedure detailed above, computation can easily go into infinite loops, even though the knowledge base could provide the correct answer. Consider for instance the simple program

```
p(X):- p(X).
p(X):- q(X).
q(X).
```

[3]Type "control" key together with "h" then press "m" to get an exhaustive mode help.

If we query this with ?− p(john), then DFS will go into an infinite loop because Prolog expands by default the first predicate. However, we can conclude that p(john) is true if we start expanding the second predicate.

In fact this is a necessary feature and not a bug for a programming language: we need to be able to write non-terminating programs, since the language would not be Turing complete otherwise. The argument can be sketched as follows: we have seen that for Turing machines the halting problem is undecidable. So if all Prolog programs were terminating, then Prolog would be weaker than Turing machines and thus not Turing complete.

We will now fortify our intuition about the Prolog search procedure by an example that extends the setup from Example 3.1.15 by a new choice of a vehicle that could be a car (if it had a motor).

Backtracking by Example

▷ **Example 3.2.2.** We extend Example 3.1.15:

```
has_wheels(mytricycle,3).
has_wheels(myrollerblade,3).
has_wheels(mybmw,4).
has_motor(mybmw).
car(X):-has_wheels(X,3),has_motor(X). % cars sometimes have three wheels
car(X):-has_wheels(X,4),has_motor(X). % and sometimes four.
?- car(Y).
?- has_wheels(Y,3),has_motor(Y). % backtrack point 1
Y = mytricycle % backtrack point 2
?- has_motor(mytricycle).
FAIL % fails, backtrack to 2
Y = myrollerblade % backtrack point 2
?- has_motor(myrollerblade).
FAIL % fails, backtrack to 1
?- has_wheels(Y,4),has_motor(Y).
Y = mybmw
?- has_motor(mybmw).
Y=mybmw
true
```


In general, a Prolog rule of the form $A{:-}B,C$ reads as *A, if B and C*. If we want to express *A if B or C*, we have to express this two separate rules $A{:-}B$ and $A{:-}C$ and leave the choice which one to use to the search procedure.

In Example 3.2.2 we indeed have two clauses for the predicate car/1; one each for the cases of cars with three and four wheels. As the three-wheel case comes first in the program, it is explored first in the search process.

Recall that at every point, where the Prolog interpreter has the choice between two clauses for a predicate, chooses the first and leaves a backtrack point. In Example 3.2.2 this happens first for the predicate car/1, where we explore the case of three-wheeled cars. The Prolog interpreter immediately has to choose again – between the tricycle and the rollerblade, which both have three wheels. Again, it chooses the first and leaves a backtrack point. But as tricycles do not have motors, the subgoal has_motor(mytricycle) fails and the interpreter backtracks to the chronologically nearest backtrack point (the second one) and tries to fulfill has_motor(myrollerblade). This fails again, and the next backtrack point is point 1 – note the stack-like organization of backtrack points which is in keeping with the depth-first search strategy – which chooses the case of four-wheeled cars. This ultimately succeeds as before with y=mybmw.

3.2.2 Programming Features

We now turn to a more classical programming task: computing with numbers. Here we turn to our initial example: adding unary natural numbers. If we can do that, then we have to consider

Prolog a programming language.

Can We Use This For Programming?

- ▷ **Question:** What about functions? E.g. the addition function?
- ▷ **Question:** We cannot define functions, in Prolog!
- ▷ **Idea (back to math):** use a three-place predicate.
- ▷ **Example 3.2.3.** add(X,Y,Z) stands for X+Y=Z
- ▷ Now we can directly write the recursive equations $X + 0 = X$ (base case) and $X + s(Y) = s(X + Y)$ into the knowledge base.

```
add(X,zero,X).
add(X,s(Y),s(Z)) :− add(X,Y,Z).
```

- ▷ Similarly with multiplication and exponentiation.

```
mult(X,zero,zero).
mult(X,s(Y),Z) :− mult(X,Y,W), add(X,W,Z).

expt(X,zero,s(zero)).
expt(X,s(Y),Z) :− expt(X,Y,W), mult(X,W,Z).
```


Note: Viewed through the right glasses logic programming is very similar to functional programming; the only difference is that we are using $n+1$ ary relations rather than n ary function. To see how this works let us consider the addition function/relation example above: instead of a binary function $+$ we program a ternary relation add, where relation add(X,Y,Z) means $X + Y = Z$. We start with the same defining equations for addition, rewriting them to relational style.

The first equation is straight-forward via our correspondence and we get the Prolog fact add(X,zero,X). For the equation $X + s(Y) = s(X + Y)$ we have to work harder, the straight-forward relational translation add(X,s(Y),s(X+Y)) is impossible, since we have only partially replaced the function $+$ with the relation add. Here we take refuge in a very simple trick that we can always do in logic (and mathematics of course): we introduce a new name Z for the offending expression $X + Y$ (using a variable) so that we get the fact add(X,s(Y),s(Z)). Of course this is not universally true (remember that this fact would say that "$X + s(Y) = s(Z)$ for all X, Y, and Z"), so we have to extend it to a Prolog rule add(X,s(Y),s(Z)):−add(X,Y,Z). which relativizes to mean "$X + s(Y) = s(Z)$ for all X, Y, and Z with $X + Y = Z$".

Indeed the rule implements addition as a recursive predicate, we can see that the recursion relation is terminating, since the left hand sides have one more constructor for the successor function. The examples for multiplication and exponentiation can be developed analogously, but we have to use the naming trick twice.

We now apply the same principle of recursive programming with predicates to other examples to reinforce our intuitions about the principles.

More Examples from elementary Arithmetic

- ▷ **Example 3.2.4.** We can also use the add relation for subtraction without changing the implementation. We just use variables in the "input positions" and ground terms in the other two. (possibly very inefficient "generate and test approach")

```
?−add(s(zero),X,s(s(s(zero)))).
X = s(s(zero))
true
```

▷ **Example 3.2.5.**

Computing the n^{th} Fibonacci number (0, 1, 1, 2, 3, 5, 8, 13,...; add the last two to get the next), using the addition predicate above.

```
fib(zero,zero).
fib(s(zero),s(zero)).
fib(s(s(X)),Y):−fib(s(X),Z),fib(X,W),add(Z,W,Y).
```

▷ **Example 3.2.6.** Using Prolog's internal arithmetic: a goal of the form ?− D **is** e. — where e is a ground arithmetic expression binds D to the result of evaluating e.

```
fib(0,0).
fib(1,1).
fib(X,Y):− D is X − 1, E is X − 2,fib(D,Z),fib(E,W), Y is Z + W.
```


Note: Note that the **is** relation does not allow "generate and test" inversion as it insists on the right hand being ground. In our example above, this is not a problem, if we call the fib with the first ("input") argument a ground term. Indeed, if match the last rule with a goal ?− g,Y., where g is a ground term, then g−1 and g−2 are ground and thus D and E are bound to the (ground) result terms. This makes the input arguments in the two recursive calls ground, and we get ground results for Z and W, which allows the last goal to succeed with a ground result for Y. Note as well that re-ordering the bodys literal of the rule so that the recursive calls are called before the computation literals will lead to failure.

We will now add the primitive data structure of lists to Prolog; they are constructed by prepending an element (the head) to an existing list (which becomes the rest list or "tail" of the constructed one).

Adding Lists to Prolog

▷ Lists are represented by terms of the form [a,b,c,...]

▷ First/rest representation [F|R], where R is a rest list.

▷ predicates for member, append and reverse of lists in default Prolog representation.

```
member(X,[X|_]).
member(X,[_|R]):−member(X,R).

append([],L,L).
append([X|R],L,[X|S]):−append(R,L,S).

reverse([],[]).
reverse([X|R],L):−reverse(R,S),append(S,[X],L).
```


Just as in functional programming languages, we can define list operations by recursion, only that we program with relations instead of with functions.

Logic programming is the third large programming paradigm (together with functional programming and imperative programming).

Relational Programming Techniques

▷ **Example 3.2.7.** Parameters have no unique direction "in" or "out"

```
?− rev(L,[1,2,3]).
?− rev([1,2,3],L1).
?− rev([1|X],[2|Y]).
```

▷ **Example 3.2.8.** Symbolic programming by structural induction

```
rev([],[]).
rev([X|Xs],Ys) :− ...
```

▷ **Example 3.2.9.**

Generate and test:

```
sort(Xs,Ys) :− perm(Xs,Ys), ordered(Ys).
```


From a programming practice point of view it is probably best understood as "relational programming" in analogy to functional programming, with which it shares a focus on recursion.
The major difference to functional programming is that "relational programming" does not have a fixed input/output distinction, which makes the control flow in functional programs very direct and predictable. Thanks to the underlying search procedure, we can sometime make use of the flexibility afforded by logic programming.

If the problem solution involves search (and depth first search is sufficient), we can just get by with specifying the problem and letting the Prolog interpreter do the rest. In Example 3.2.9 we just specify that list Xs can be sorted into Ys, iff Ys is a permutation of Xs and Ys is ordered. Given a concrete (input) list Xs, the Prolog interpreter will generate all permutations of Ys of Xs via the predicate perm/2 and then test them whether they are ordered.

This is a paradigmatic example of logic programming. We can (sometimes) directly use the specification of a problem as a program. This makes the argument for the correctness of the program immediate, but may make the program execution non optimal.

3.2.3 Advanced Relational Programming

It is easy to see that the running time of the Prolog program from Example 3.2.9 is not $\mathcal{O}(n\log_2(n))$ which is optimal for sorting algorithms. This is the flip side of the flexibility in logic programming. But Prolog has ways of dealing with that: the cut operator, which is a Prolog atom, which always succeeds, but which cannot be backtracked over. This can be used to prune the search tree in Prolog. We will not go into that here but refer the readers to the literature.

Specifying Control in Prolog

▷ *Remark 3.2.10.* The running time of the program from Example 3.2.9 is not $\mathcal{O}(n\log_2(n))$ which is optimal for sorting algorithms.

```
sort(Xs,Ys) :− perm(Xs,Ys), ordered(Ys).
```

▷ **Idea:** Gain computational efficiency by shaping the search!

Functions and Predicates in Prolog

▷ *Remark 3.2.11.* Functions and predicates have radically different roles in Prolog.

- ▷ Functions are used to represent data. (e.g. father(john) or s(s(zero)))
- ▷ Predicates are used for stating properties about and computing with data.

▷ *Remark 3.2.12.* In functional programming, functions are used for both. (even more confusing than in Prolog if you think about it)

▷ **Example 3.2.13.** Consider again the reverse program for lists below:
An input datum is e.g. [1,2,3], then the output datum is [3,2,1].

```
reverse([],[]).
reverse([X|R],L):−reverse(R,S),append(S,[X],L).
```

We "define" the computational behavior of the predicate rev, but the list constructors [. . .] are just used to construct lists from arguments.

▷ **Example 3.2.14 (Trees and Leaf Counting).** We represent (unlabelled) trees via the function t from tree lists to trees. For instance, a balanced binary tree of depth 2 is t([t([t([]),t([])]),t([t([]),t([])])]). We count leaves by

```
leafcount(t([]),1).
leafcount(t([X|R]),Y) :− leafcount(X,Z), leafcount(t(R,W)), Y is Z + W.
```


For more information on Prolog

RTFM ($\widehat{=}$ "read the fine manuals")

▷ **RTFM Resources:** There are also lots of good tutorials on the web,

- ▷ I personally like [Fis; LPN],
- ▷ [Fla94] has a very thorough logic-based introduction,
- ▷ consult also the SWI Prolog Manual [SWI],

Chapter 4

Recap of Prerequisites from Math & Theoretical Computer Science

In this chapter we will briefly recap some of the prerequisites from theoretical computer science that are needed for understanding Artificial Intelligence 1.

4.1 Recap: Complexity Analysis in AI?

We now come to an important topic which is not really part of Artificial Intelligence but which adds an important layer of understanding to this enterprise: We (still) live in the era of Moore's law (the computing power available on a single CPU doubles roughly every two years) leading to an exponential increase. A similar rule holds for main memory and disk storage capacities. And the production of computer (using CPUs and memory) is (still) very rapidly growing as well; giving mankind as a whole, institutions, and individual exponentially grow of computational resources.

In public discussion, this development is often cited as the reson why (strong) AI is inevitable. But the argument is fallacious if all the algorithms we have are of very high complexity (i.e. at least exponential in either time or space). So, to judge the state of play in Artificial Intelligence, we have to know the complexity of our algorithms.

In this section, we will give a very brief recap of some aspects of elementary complexity theory and make a case of why this is a generally important for computer scientists.

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21839` and `https://fau.tv/clip/id/21840`.

In order to get a feeling what we mean by "fast algorithm", we to some preliminary computations.

Performance and Scaling

- ▷ Suppose we have three algorithms to choose from. (which one to select)
- ▷ Systematic analysis reveals performance characteristics.
- ▷ **Example 4.1.1.** For a problem of size n we have

	performance		
size	linear	quadratic	exponential
n	$100n\mu s$	$7n^2\mu s$	$2^n\mu s$
1	$100\mu s$	$7\mu s$	$2\mu s$
5	$.5ms$	$175\mu s$	$32\mu s$
10	$1ms$	$.7ms$	$1ms$
45	$4.5ms$	$14ms$	$1.1Y$
100	...	...	...
1 000	...	...	...
10 000	...	...	...
1 000 000	...	...	...

What?! One year?

- $2^{10} = 1\,024$ ($1024\mu s \simeq 1ms$)
- $2^{45} = 35\,184\,372\,088\,832$ ($3.5{\times}10^{13}\mu s \simeq 3.5{\times}10^7 s \simeq 1.1Y$)
- **Example 4.1.2.** we denote all times that are longer than the age of the universe with —

	performance		
size	linear	quadratic	exponential
n	$100n\mu s$	$7n^2\mu s$	$2^n\mu s$
1	$100\mu s$	$7\mu s$	$2\mu s$
5	$.5ms$	$175\mu s$	$32\mu s$
10	$1ms$	$.7ms$	$1ms$
45	$4.5ms$	$14ms$	$1.1Y$
< 100	$100ms$	$7s$	$10^{16}Y$
1 000	$1s$	$12min$	—
10 000	$10s$	$20h$	—
1 000 000	$1.6min$	$2.5mon$	—

So it does make a difference for larger problems what algorithm we choose. Considerations like the one we have shown above are very important when judging an algorithm. These evaluations go by the name of "complexity theory".
Let us now recapitulate some notions of elementary complexity theory: we are interested in the worst case growth of the resources (time and space) required by an algorithm in terms of the sizes of its arguments. Mathematically we look at the functions from input size to resource size and classify them into "big-O" classes, abstracting from constant factors (which depend on the machine thealgorithm runs on and which we cannot control) and initial (algorithm startup) factors.

Recap: Time/Space Complexity of Algorithms

- We are mostly interested in worst-case complexity in AI-1.
- **Definition:** Let $S \subseteq \mathbb{N} \to \mathbb{N}$ be a set of natural number functions, then we say that analgorithm α that terminates in time $t(n)$ for all inputs of size n has running

time $T(\alpha){:=}t$.

We say that α has time complexity in S (written $T(\alpha){\in}S$ or colloquially $T(\alpha){=}S$), iff $t{\in}S$. We say α has space complexity in S, iff α uses only memory of size $s(n)$ on inputs of size n and $s{\in}S$.

▷ Time/space complexity depends on size measures. (no canonical one)

▷ **Definition:** The following sets are often used for S in $T(\alpha)$:

Landau set	class name	rank	Landau set	class name	rank
$\mathcal{O}(1)$	constant	1	$\mathcal{O}(n^2)$	quadratic	4
$\mathcal{O}(\ln(n))$	logarithmic	2	$\mathcal{O}(n^k)$	polynomial	5
$\mathcal{O}(n)$	linear	3	$\mathcal{O}(k^n)$	exponential	6

where $\mathcal{O}(g) = \{f | \exists k{>}0. f {\leq_a} k \cdot g\}$ and $f{\leq_a}g$ (f is asymptotically bounded by g), iff there is an $n_0{\in}\mathbb{N}$, such that $f(n){\leq}g(n)$ for all $n{>}n_0$.

For $k'{>}2$ and $k{>}1$ we have

$$\mathcal{O}(1){\subset}\mathcal{O}(\log n){\subset}\mathcal{O}(n){\subset}\mathcal{O}(n^2){\subset}\mathcal{O}(n^{k'}){\subset}\mathcal{O}(k^n)$$

▷ **For AI-1:** I expect that given analgorithm, you can determine its complexity class. (next)

OK, that was the theory, ... but how do we use that in practice.

Determining the Time/Space Complexity of Algorithms

▷ Given a function γ that maps variables v to sets $\Gamma(v)$, we compute $T_\Gamma(\alpha)$ and $C_\Gamma(\alpha)$ of an imperative algorithm α by induction on the structure of α:

- ▹ constant: can be accessed in constant time If $\alpha = \delta$ for a data constant δ, then $T_\Gamma(\alpha){\in}\mathcal{O}(1)$.
- ▹ variable: need the complexity of the value
  If $\alpha = v$ with $v{\in}\mathbf{dom}(\Gamma)$, then $T_\Gamma(\alpha){\in}\mathcal{O}(\Gamma(v))$.
- ▹ application: compose the complexities of the function and the argument
  If $\alpha = \varphi(\psi)$ with $T_\Gamma(\varphi){\in}\mathcal{O}(f)$ and $T_{\Gamma\cup C_\Gamma(\varphi)}(\psi){\in}\mathcal{O}(g)$, then $T_\Gamma(\alpha){\in}\mathcal{O}(f \circ g)$ and $C_\Gamma(\alpha) = C_{\Gamma\cup C_\Gamma(\varphi)}(\psi)$.
- ▹ assignment: has to compupte the value $\rightsquigarrow$ has its complexity
  If α is $v{:=}\varphi$ with $T_\Gamma(\varphi){\in}S$, then $T_\Gamma(\alpha){\in}S$ and $C_\Gamma(\alpha) = \Gamma \cup (v{,}S)$.
- ▹ composition: has the maximal complexity of the components
  If α is $\varphi\,;\psi$, with $T_\Gamma(\varphi){\in}P$ and $T_{\Gamma\cup C_\Gamma(\psi)}(\psi){\in}Q$, then $T_\Gamma(\alpha){\in}\max\,\{P, Q\}$ and $C_\Gamma(\alpha) = C_{\Gamma\cup C_\Gamma(\psi)}(\psi)$.
- ▹ branching: has the maximal complexity of the condition and branches
  If α is **if**γ**then**φ**else**ψ**end**, with $T_\Gamma(\gamma){\in}C$, $T_{\Gamma\cup C_\Gamma(\gamma)}(\varphi){\in}P$, $T_{\Gamma\cup C_\Gamma(\gamma)}(\varphi){\in}Q$, and then $T_\Gamma(\alpha){\in}\max\,\{C, P, Q\}$ and $C_\Gamma(\alpha) = \Gamma \cup C_\Gamma(\gamma) \cup C_{\Gamma\cup C_\Gamma(\gamma)}(\varphi) \cup C_{\Gamma\cup C_\Gamma(\gamma)}(\psi)$.

- ▷ looping: multiplies complexities
  If α is **while**γ**do**φ**end**, with $T_\Gamma(\gamma)\in\mathcal{O}(f)$, $T_{\Gamma\cup C_\Gamma(\gamma)}(\varphi)\in\mathcal{O}(g)$, then $T_\Gamma(\alpha)\in\mathcal{O}(f(n)\cdot g(n))$ and $C_\Gamma(\alpha)=C_{\Gamma\cup C_\Gamma(\gamma)}(\varphi)$.

The time complexity $T(\alpha)$ is just $T_\emptyset(\alpha)$, where $\emptyset$ is the empty function.

▷ Recursion is much more difficult to analyze ⇝ recurrence relations and Master's theorem.

Please excuse the chemistry pictures, public imagery for CS is really just quite boring, this is what people think of when they say "scientist". So, imagine that instead of a chemist in a lab, it's me sitting in front of a computer.

Why Complexity Analysis? (General)

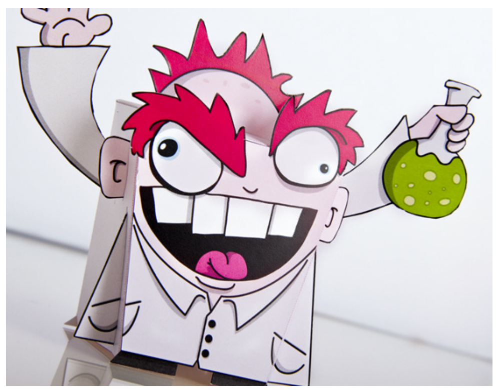
▷ **Example 4.1.3.** Once upon a time I was trying to invent an efficient algorithm.

- ▷ My first algorithm attempt didn't work, so I had to try harder.

- ▷ But my 2nd attempt didn't work either, which got me a bit agitated.

- ▷ The 3rd attempt didn't work either...

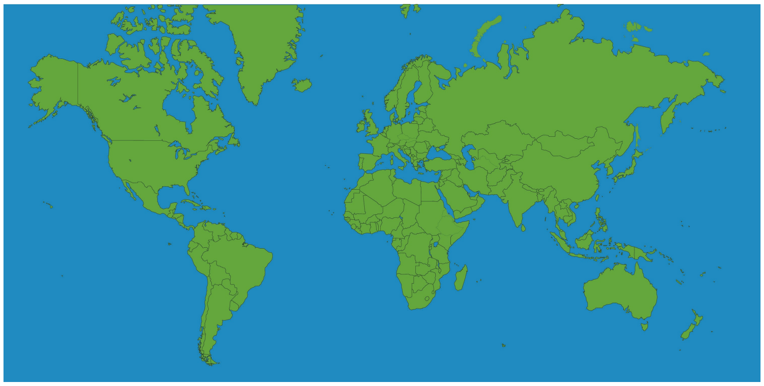
▷ And neither the 4th. But then:



▷ Ta-da ... when, for once, I turned around and looked in the other direction–CAN one actually solve this efficiently? – NP hardness was there to rescue me.

Why Complexity Analysis? (General)

▷ **Example 4.1.4.** Trying to find a sea route east to India (from Spain) (does not exist)



- ▷ **Observation:** Complexity theory saves you from spending lots of time trying to invent algorithms that do not exist.

It's like, you're trying to find a route to India (from Spain), and you presume it's somewhere to the east, and then you hit a coast, but no; try again, but no; try again, but no; ... if you don't have a map, that's the best you can do. But NP hardness gives you the map: you can check that there actually is no way through here.

But what is this notion of NP completness alluded to above? We observe that we can analyze the complexity of problems by the complexity classcomplexity of the algorithms that solve them. This gives us a notion of what to expect from solutions to a given problem class, and thus whether efficient (i.e. polynomial time) algorithms can exist at all.

Reminder (?): NP and PSPACE (details ⇝ e.g. [GJ79])

- ▷ **Turing Machine:** Works on a tape consisting of cells, across which its Read/Write head moves. The machine has internal states. There is a transition function that specifies – given the current cell content and internal state – what the subsequent internal state will be, how what the R/W head does (write a symbol and/or move). Some internal states are accepting.
- ▷ Decision problems are in **NP** if there is a non deterministic Turing machine that halts with an answer after time polynomial in the size of its input. Accepts if *at least one* of the possible runs accepts.
- ▷ Decision problems are in **NPSPACE**, if there is a non deterministic Turing machine that runs in space polynomial in the size of its input.
- ▷ **NP vs. PSPACE:** Non-deterministic polynomial space can be simulated in deterministic polynomial space. Thus **PSPACE** = **NPSPACE**, and hence (trivially) **NP** ⊆ **PSPACE**.

  It is commonly believed that **NP**⊉**PSPACE**. (similar to **P** ⊆ **NP**)

The Utility of Complexity Knowledge (NP-Hardness)

- ▷ **Assume:** In 3 years from now, you have finished your studies and are working in your first industry job. Your boss Mr. X gives you a problem and says *Solve It!*. By which he means, *write a program that solves it efficiently*.
- ▷ **Question:** Assume further that, after trying in vain for 4 weeks, you got the next meeting with Mr. X. How could knowing about NP hardness help?
- ▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

4.2 Recap: Formal Languages and Grammars

One of the main ways of designing rational agents in this course will be to define formal languages that represent the state of the agent environment and let the agent use various inference techniques to predict effects of its observations and actions to obtain a world model. In this section we recap the basics of formal languages and grammars that form the basis of a compositional theory for them.

The Mathematics of Strings

- ▷ **Definition 4.2.1.** An alphabet A is a finite set; we call each element $a \in A$ a character, and an n tuple $s \in A^n$ a string (of length n over A).
- ▷ **Definition 4.2.2.** Note that $A^0 = \{\langle\rangle\}$, where $\langle\rangle$ is the (unique) 0-tuple. With the definition above we consider $\langle\rangle$ as the string of length 0 and call it the empty string and denote it with ϵ.
- ▷ **Note:** Sets $\neq$ strings, e.g. $\{1,2,3\} = \{3,2,1\}$, but $\langle 1,2,3\rangle \neq \langle 3,2,1\rangle$.
- ▷ **Notation:** We will often write a string $\langle c_1, \ldots, c_n\rangle$ as "$\mathtt{c_1 \ldots c_n}$", for instance "abc" for $\langle \mathtt{a}, \mathtt{b}, \mathtt{c}\rangle$
- ▷ **Example 4.2.3.** Take $A = \{\mathtt{h}, \mathtt{1}, /\}$ as an alphabet. Each of the members $\mathtt{h}$, $\mathtt{1}$, and $/$ is a character. The vector $\langle /, /, \mathtt{1}, \mathtt{h}, \mathtt{1}\rangle$ is a string of length 5 over A.
- ▷ **Definition 4.2.4 (String Length).** Given a string s we denote its length with $|s|$.
- ▷ **Definition 4.2.5.** The concatenation $\mathrm{conc}(s,t)$ of two strings $s = \langle s_1, \ldots, s_n\rangle \in A^n$ and $t = \langle t_1, \ldots, t_m\rangle \in A^m$ is defined as $\langle s_1, \ldots, s_n, t_1, \ldots, t_m\rangle \in A^{n+m}$.

  We will often write $\mathrm{conc}(s,t)$ as $s+t$ or simply st
- ▷ **Example 4.2.6.** $\mathrm{conc}(\text{"text"}, \text{"book"}) = \text{"text"} + \text{"book"} = \text{"textbook"}$

We have multiple notations for concatenation, since it is such a basic operation, which is used so often that we will need very short notations for it, trusting that the reader can disambiguate based on the context.

Now that we have defined the concept of a string as a sequence of characters, we can go on to give ourselves a way to distinguish between good strings (e.g. programs in a given programming language) and bad strings (e.g. such with syntax errors). The way to do this by the concept of a formal language, which we are about to define.

Formal Languages

- **Definition 4.2.7.** Let A be an alphabet, then we define the sets $A^+ := \bigcup_{i \in \mathbb{N}^+} A^i$ of nonempty string and $A^* := A^+ \cup \{\epsilon\}$ of strings.
- **Example 4.2.8.** If $A = \{\mathsf{a}, \mathsf{b}, \mathsf{c}\}$, then $A^* = \{\epsilon, \mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{aa}, \mathsf{ab}, \mathsf{ac}, \mathsf{ba}, \ldots, \mathsf{aaa}, \ldots\}$.
- **Definition 4.2.9.** A set $L \subseteq A^*$ is called a formal language over A.
- **Definition 4.2.10.** We use $\mathsf{c}^{[n]}$ for the string that consists of the character c repeated n times.
- **Example 4.2.11.** $\#^{[5]} = \langle \#, \#, \#, \#, \# \rangle$
- **Example 4.2.12.** The set $M := \{\mathsf{ba}^{[n]} | n \in \mathbb{N}\}$ of strings that start with character b followed by an arbitrary numbers of a's is a formal language over $A = \{\mathsf{a}, \mathsf{b}\}$.
- **Definition 4.2.13 (Operations on Languages).** Let L, L_1, and L_2 be formal languages over the same alphabet, then we define language level operations: $L_1 L_2 := \{s_1 s_2 | s_1 \in L_1 \wedge s_2 \in L_2\}$, $L^+ := \{s^+ | s \in L\}$, and $L^* := \{s^* | s \in L\}$.

There is a common misconception that a formal language is something that is difficult to understand as a concept. This is not true, the only thing a formal language does is separate the "good" from the bad strings. Thus we simply model a formal language as a set of stings: the "good" strings are members, and the "bad" ones are not.

Of course this definition only shifts complexity to the way we construct specific formal languages (where it actually belongs), and we have learned two (simple) ways of constructing them: by repetition of characters, and by concatenation of existing languages.

As mentioned above, the purpose of a formal language is to distinguish "good" from "bad" strings. It is maximally general, but not helpful, since it does not support computation and inference. In practice we will be interested in formal languages that have some structure, so that we can represent formal languages in a finite manner (recall that a formal language is a subset of A^*, which may be infinite and even undecidable – even though the alphabet A is finite).

To remedy this, we will now introduce phrase structure grammars (or just grammars), the standard tool for describing structured formal languages.

Phrase Structure Grammars (Theory)

- **Recap:** A formal language is an arbitrary set of symbol sequences.
- **Problem:** This may be infinite and even undecidable even if A is finite.
- **Idea:** Find a way of representing formal languages with structure finitely.
- **Definition 4.2.14.** A phrase structure grammar (or just grammar) is a tuple $\langle N, \Sigma, P, S \rangle$ where
  - N is a finite set of nonterminal symbols,
  - Σ is a finite set of terminal symbols, members of $\Sigma \cup N$ are called symbols.

- P is a finite set of production rules: pairs $p{:=}h{\to}b$ (also written as $h{\Rightarrow}b$), where $h{\in}(\Sigma\cup N)^*N(\Sigma\cup N)^*$ and $b{\in}(\Sigma\cup N)^*$. The string h is called the head of p and b the body.
- $s{\in}N$ is a distinguished symbol called the start symbol (also sentence symbol).

- **Intuition:** Production rules map strings with at least one nonterminal to arbitrary other strings.
- **Notation:** If we have n rules $h{\to}b_i$ sharing a head, we often write $h{\to}b_1\mid\ldots\mid b_n$ instead.

We fortify our intuition about these – admittedly very abstract – constructions by an example and introduce some more vocabulary.

Phrase Structure Grammars (cont.)

- **Example 4.2.15.** A simple phrase structure grammar G:

$$\begin{array}{rcl} S & \to & NP;Vi \\ NP & \to & Article;N \\ Article & \to & \textbf{the}\mid\textbf{a}\mid\textbf{an} \\ N & \to & \textbf{dog}\mid\textbf{teacher}\mid\ldots \\ Vi & \to & \textbf{sleeps}\mid\textbf{smells}\mid\ldots \end{array}$$

  Here S, is the start symbol, NP, VP, $Article$, N, and Vi are nonterminals.

- **Definition 4.2.16.** The subset of lexical rules, i.e. those whose body consists of a single terminal is called its lexicon and the set of body symbols the alphabet. The nonterminals in their heads are called lexical categories.
- **Definition 4.2.17.** The non-lexicon production rules are called structural, and the nonterminals in the heads are called phrasal categories.

Now we look at just how a grammar helps in analyzing formal languages. The basic idea is that a grammar accepts a word, iff the start symbol can be rewritten into it using only the rules of the grammar.

Phrase Structure Grammars (Theory)

- **Idea:** Each symbol sequence in a formal language can be analyzed/generated by the grammar.
- **Definition 4.2.18.** Given a phrase structure grammar $G{:=}\langle N,\Sigma,P,S\rangle$, we say G derives $t{\in}(\Sigma\cup N)^*$ from $s{\in}(\Sigma\cup N)^*$ in one step, iff there is a production rule $p{\in}P$ with $p=h{\to}b$ and there are $u,v{\in}(\Sigma\cup N)^*$, such that $s=suhv$ and $t=ubv$. We write $s{\to}_G^p t$ (or $s{\to}_G t$ if p is clear from the context) and use ${\to}_G^*$ for the transitive reflexive closure of ${\to}_G$. We call $s{\to}_G^* t$ a G derivation of t from s.
- **Definition 4.2.19.** Given a phrase structure grammar $G{:=}\langle N,\Sigma,P,S\rangle$, we say

that $s \in (N \cup \Sigma)^*$ is a sentential form of G, iff $S \rightarrow_G^* s$. A sentential form that does not contain nontermials is called a sentence of G, we also say that G accepts s.

▷ **Definition 4.2.20.** The language $\mathbf{L}(G)$ of G is the set of its sentences.

Definition 4.2.21. We call two grammars equivalent, iff they have the same languages.

▷ **Definition 4.2.22.** Parsing, syntax analysis, or syntactic analysis is the process of analyzing a string of symbols, either in a formal or a natural language by means of a grammar.

Phrase Structure Grammars (Example)

▷ **Example 4.2.23.** In the grammar G from Example 30.9.2:

1. $\mathit{Article}; \mathbf{teacher}; \mathit{Vi}$ is a sentential form,

$$\begin{array}{lll} S & \rightarrow_G & \mathit{NP}; \mathit{Vi} \\ & \rightarrow_G & \mathit{Article}; N; \mathit{Vi} \\ & \rightarrow_G & \mathit{Article}; \mathbf{teacher}; \mathit{Vi} \end{array}$$

2. *The teacher sleeps* is a sentence.

$$\begin{array}{lll} S & \rightarrow_G^* & \mathit{Article}; \mathbf{teacher}; \mathit{Vi} \\ & \rightarrow_G & \mathbf{the}; \mathbf{teacher}; \mathit{Vi} \\ & \rightarrow_G & \mathbf{the}; \mathbf{teacher}; \mathbf{sleeps} \end{array}$$

$$\begin{array}{rcl} S & \rightarrow & \mathit{NP}; \mathit{Vi} \\ \mathit{NP} & \rightarrow & \mathit{Article}; N \\ \mathit{Article} & \rightarrow & \mathbf{the} \mid \mathbf{a} \mid \mathbf{an} \mid \ldots \\ N & \rightarrow & \mathbf{dog} \mid \mathbf{teacher} \mid \ldots \\ \mathit{Vi} & \rightarrow & \mathbf{sleeps} \mid \mathbf{smells} \mid \ldots \end{array}$$

Note that this process indeed defines a formal language given a grammar, but does not provide an efficient algorithm for parsing, even for the simpler kinds of grammars we introduce below.

Grammar Types (Chomsky Hierarchy [Cho65])

▷ **Observation:** The shape of the grammar determines the "size" of its language.

▷ **Definition 4.2.24.** We call a grammar and the formal language it accepts:

1. context-sensitive, if the bodies of production rules have no less symbols than the heads,
2. context-free, if the heads have exactly one symbol,
3. regular, if additionally, bodies consist of a nonterminal, optionally followed by a terminal symbol.

By extension, a formal language L is called context-sensitive/context-free/regular, iff it is the language of a respective grammar. Context-free grammars are sometimes CFLs and context-free languages CFGs.

▷ **Example 4.2.25 (Languages and their Grammars).**

▷ Context-sensitive: The language $\{a^{[n]}b^{[n]}c^{[n]}\}$ is accepted by

$$\begin{array}{rcl} S & \rightarrow & \mathbf{a};\mathbf{b};\mathbf{c} \mid A \\ A & \rightarrow & \mathbf{a};A;B;\mathbf{c} \mid \mathbf{a};\mathbf{b};\mathbf{c} \\ \mathbf{c};B & \rightarrow & B;\mathbf{c} \\ \mathbf{b};B & \rightarrow & \mathbf{b};\mathbf{b} \end{array}$$

▷ Context-free: The language $\{a^{[n]}b^{[n]}\}$ is accepted by $S \rightarrow \mathbf{a};S;\mathbf{b} \mid \epsilon$.

▷ Regular: The language $\{a^{[n]}\}$ is accepted by $S \rightarrow S;\mathbf{a}$

▷ **Observation:** Natural languages are probably context-sensitive but parsable in real time! (like languages low in the hierarchy)

Useful Extensions of Phrase Structure Grammars

▷ **Definition 4.2.26.** The Bachus Naur form or Backus normal form (BNF) is a metasyntax notation for context-free grammars.

It extends the body of a production rule by mutiple (admissible) constructors:

▷ alternative: $s_1 \mid \ldots \mid s_n$,

▷ repetition: s^* (arbitrary many s) and s^+ (at least one s),

▷ optional: $[s]$ (zero or one times), and

▷ grouping: $(s_1 ; \ldots ; s_n)$, useful e.g. for repetition.

▷ **Observation:** All of these can be eliminated, .e.g (⇝ many more rules)

▷ replace $X \rightarrow Z;(s^*);W$ with the production rules $X \rightarrow Z;Y;W$, $Y \rightarrow \epsilon$, and $Y \rightarrow Y;s$.

▷ replace $X \rightarrow Z;(s^+);W$ with the production rules $X \rightarrow Z;Y;W$, $Y \rightarrow s$, and $Y \rightarrow Y;s$.

An Grammar Notation for AI-1

▷ **Problem:** In grammars, notations for nonterminal symbols should be

▷ short and mnemonic (for the use in the body)

▷ close to the official name of the syntactic category (for the use in the head)

▷ In AI-1 we will only use context-free grammars (simpler, but problem still applies)

▷ **in AI-1:** I will try to give "grammar overviews" that combine those, e.g. the grammar of first-order logic.

variables	X	$\in$	$\mathcal{V}_1$	
functions	f^k	$\in$	Σ_k^f	
predicates	p^k	$\in$	Σ_k^p	
terms	t	$::=$	X	variable
		$\mid$	f^0	constant
		$\mid$	$f^k(t_1,\ldots,t_k)$	application
formulae	$\mathbf{A}$	$::=$	$p^k(t_1,\ldots,t_k)$	atomic
		$\mid$	$\neg\mathbf{A}$	negation
		$\mid$	$\mathbf{A}_1 \wedge \mathbf{A}_2$	conjunction
		$\mid$	$\forall X.\mathbf{A}$	quantifier

We will generally get by with context-free grammars, which have highly efficient into parsing algorithms, for the formal language we use in this course, but we will not cover the algorithms in AI-1.

4.3 Mathematical Language Recap

Mathematical Structures

- **Observation:** Mathematicians often cast object classes as mathematical structures.
- We have just seen this: repeated here for convenience.
- **Definition 4.3.1.** A phrase structure grammar (or just grammar) is a tuple $\langle N, \Sigma, P, S\rangle$ where
  - N is a finite set of nonterminal symbols,
  - Σ is a finite set of terminal symbols, members of $\Sigma \cup N$ are called symbols.
  - P is a finite set of production rules: pairs $p{:=}h{\rightarrow}b$ (also written as $h{\Rightarrow}b$), where $h{\in}(\Sigma \cup N)^*N(\Sigma \cup N)^*$ and $b{\in}(\Sigma \cup N)^*$. The string h is called the head of p and b the body.
  - $s{\in}N$ is a distinguished symbol called the start symbol (also sentence symbol).
- **Observation:** Even though we call production rules "pairs" above, they are also mathematical structures $\langle h, b\rangle$ with a funny notation $h{\rightarrow}b$.

Mathematical Structures in Programming

- Most programming languages have some way of creating "named structures". Referencing components is usually done via "dot notation"
- **Example 4.3.2 (Structs in C).**

```
// Create strutures grule grammar
```

```
struct grule {
  char[][] head;
  char[][] body;
}

struct grammar {
  char[][] nterminals;
  char[][] termininals;
  grule[] grules;
  char[] start;
}

int main() {
  struct grule r1;
  r1.head = "foo";
  r1.body = "bar";
}
```


In AI-1 we use a mixture between Math and Programming Styles

- ▷ In AI-1 we use mathematical notation, ...
- ▷ I will try to always give "structure overviews", that combine notations with "type" information and accessor names, e.g.

$$\text{grammar} = \left\langle \begin{array}{lll} N & \text{set} & \text{nonterminal symbols}, \\ \Sigma & \text{set} & \text{terminal symbols}, \\ P & \{h{\rightarrow}b | \ldots\} & \text{production rules}, \\ S & N & \text{start symbol} \end{array} \right\rangle$$

$$\text{production rule} \quad h{\rightarrow}b = \left\langle \begin{array}{lll} h & (\Sigma \cup N)^*, N, (\Sigma \cup N)^* & \text{head}, \\ b & (\Sigma \cup N)^* & \text{body} \end{array} \right\rangle$$

Chapter 5

Rational Agents: a Unifying Framework for Artificial Intelligence

In this chapter, we introduce a framework that gives a comprehensive conceptual model for the multitude of methods and algorithms we cover in this course. The framework of rational agents accommodates two traditions of AI.

Initially, the focus of AI research was on symbolic methods concentrating on the mental processes of problem solving, starting from Newell/Simon's "physical symbol hypothesis":

> *A physical symbol system has the necessary and sufficient means for general intelligent action.*
> [NS76]

Here a symbol is a representation an idea, object, or relationship that is physically manifested in (the brain of) an intelligent agent (human or artificial).

Later – in the 1980s – the proponents of embodied AI posited that most features of cognition, whether human or otherwise, are shaped – or at least critically influenced – by aspects of the entire body of the organism. The aspects of the body include the motor system, the perceptual system, bodily interactions with the environment (situatedness) and the assumptions about the world that are built into the structure of the organism. They argue that symbols are not always necessary since

> *The world is its own best model. It is always exactly up to date. It always has every detail there is to be known. The trick is to sense it appropriately and often enough.* [Bro90]

The framework of rational agents initially introduced by Russell and Wefald in [RW91] – accommodates both, it situates agents with percepts and actions in an environment, but does not preclude physical symbol systems – i.e. systems that manipulate symbols as agent functions. Russell and Norvig make it the central metaphor of their book "Artificial Intelligence – A modern approach" [RN03], which we follow in this course.

5.1 Introduction: Rationality in Artificial Intelligence

We now introduce the notion of rational agents as entities in the world that act optimally (given the available information). We situate rational agents in the scientific landscape by looking at variations of the concept that lead to slightly different fields of study.

What is AI? Going into Details

▷ **Recap:** AI studies how we can make the computer do things that humans can still do better at the moment. (humans are proud to be rational)

▷ **What is AI?:** Four possible answers/facets: Systems that

think like humans	think rationally
act like humans	act rationally

expressed by four different definitions/quotes:

	Humanly	**Rational**
Thinking	"*The exciting new effort to make computers think . . . machines with human-like minds*" [Hau85]	"*The formalization of mental faculties in terms of computational models*" [CM85]
Acting	"*The art of creating machines that perform actions requiring intelligence when performed by people*" [Kur90]	"*The branch of CS concerned with the automation of appropriate behavior in complex situations*" [LS93]

▷ **Idea:** Rationality is performance-oriented rather than based on imitation.

So, what does modern AI do?

▷ **Acting Humanly:** Turing test, not much pursued outside Loebner prize

- ▷ $\widehat{=}$ building pigeons that can fly so much like real pigeons that they can fool pigeons
- ▷ Not reproducible, not amenable to mathematical analysis

▷ **Thinking Humanly:** $\rightsquigarrow$ Cognitive Science.

- ▷ How do humans think? How does the (human) brain work?
- ▷ Neural networks are a (extremely simple so far) approximation

▷ **Thinking Rationally:** Logics, Formalization of knowledge and inference

- ▷ You know the basics, we do some more, fairly widespread in modern AI

▷ **Acting Rationally:** How to make good action choices?

- ▷ Contains logics (one possible way to make intelligent decisions)
- ▷ We are interested in making good choices in practice (e.g. in AlphaGo)

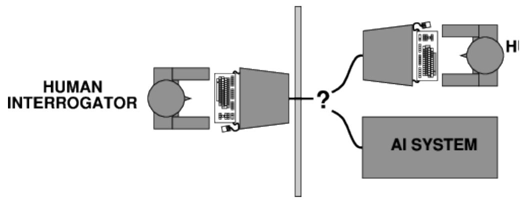
We now discuss all of the four facets in a bit more detail, as they all either contribute directly to our discussion of AI methods or characterize neighboring disciplines.

Acting humanly: The Turing test

▷ Introduced by Alan Turing (1950) "Computing machinery and intelligence" [Tur50]:

▷ "Can machines think?" ⟶ "Can machines behave intelligently?"

▷ **Definition 5.1.1.** The Turing test is an operational test for intelligent behavior based on an imitation game over teletext (arbitrary topic)



▷ It was predicted that by 2000, a machine might have a 30% chance of fooling a lay person for 5 minutes.

▷ **Note:** In [Tur50], Alan Turing

- ▷ anticipated all major arguments against AI in following 50 years and
- ▷ suggested major components of AI: knowledge, reasoning, language understanding, learning

▷ **Problem:** Turing test is not reproducible, constructive, or amenable to mathematical analysis!

Thinking humanly: Cognitive Science

▷ **1960s:** "cognitive revolution": information processing psychology replaced prevailing orthodoxy of behaviorism.

▷ Requires scientific theories of internal activities of the brain

▷ What level of abstraction? "Knowledge" or "circuits"?

▷ **How to validate?:** Requires

1. Predicting and testing behavior of human subjects or (top-down)
2. Direct identification from neurological data. (bottom-up)

▷ **Definition 5.1.2.** Cognitive Science is the interdisciplinary, scientific study of the mind and its processes. It examines the nature, the tasks, and the functions of cognition.

▷ **Definition 5.1.3.** Cognitive Neuroscience studies the biological processes and aspects that underlie cognition, with a specific focus on the neural connections in the brain which are involved in mental processes.

▷ Both approaches/disciplines are now distinct from AI.

▷ Both share with AI the following characteristic: *the available theories do not explain (or engender) anything resembling human-level general intelligence*

▷ Hence, all three fields share one principal direction!

Thinking rationally: Laws of Thought

- ▷ Normative (or prescriptive) rather than descriptive
- ▷ Aristotle: what are correct arguments/thought processes?
- ▷ Several Greek schools developed various forms of logic: *notation* and *rules of derivation* for thoughts; may or may not have proceeded to the idea of mechanization.
- ▷ Direct line through mathematics and philosophy to modern AI
- ▷ Problems
  1. Not all intelligent behavior is mediated by logical deliberation
  2. What is the purpose of thinking? What thoughts *should* I have out of all the thoughts (logical or otherwise) that I *could* have?

Acting Rationally

- ▷ **Idea:** Rational behavior $\widehat{=}$ doing the right thing!
- ▷ **Definition 5.1.4.** Rational behavior consists of always doing what is expected to maximize goal achievement given the available information.
- ▷ Rational behavior does not necessarily involve thinking e.g., blinking reflex — but thinking should be in the service of rational action.
- ▷ **Aristotle:** *Every art and every inquiry, and similarly every action and pursuit, is thought to aim at some good.* (Nicomachean Ethics)

The Rational Agents

- ▷ **Definition 5.1.5.** An agent is an entity that perceives and acts.
- ▷ **Central Idea:** This course is about designing agent that exhibit rational behavior, i.e. for any given class of environments and tasks, we seek the agent (or class of agents) with the best performance.
- ▷ **Caveat:** *Computational limitations make perfect rationality unachievable*
  $\leadsto$ design best program for given machine resources.

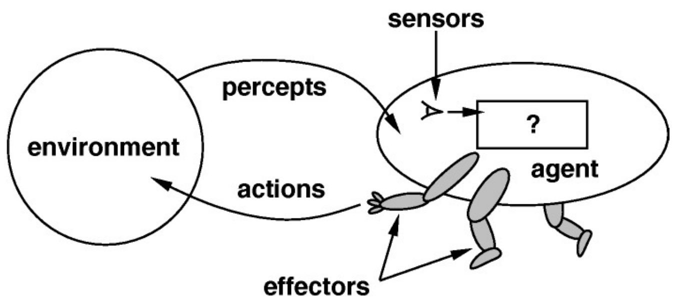
5.2 Agents and Environments as a Framework for AI

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21843`.

Agents and Environments

▷ **Definition 5.2.1.** An agent is anything that

- ▷ perceives its environment via sensors (a means of sensing the environment)
- ▷ acts on it with actuators (means of changing the environment).



▷ **Example 5.2.2.** Agents include humans, robots, softbots, thermostats, etc.

Modeling Agents Mathematically and Computationally

▷ **Definition 5.2.3.** A percept is the perceptual input of an agent at a specific instant.

▷ **Definition 5.2.4.** Any recognizable, coherent employment of the actuators of an agent is called an action.

▷ **Definition 5.2.5.** The agent function f_a of an agent a maps from percept histories to actions:

$$f_a\colon \mathcal{P}^* \to \mathcal{A}$$

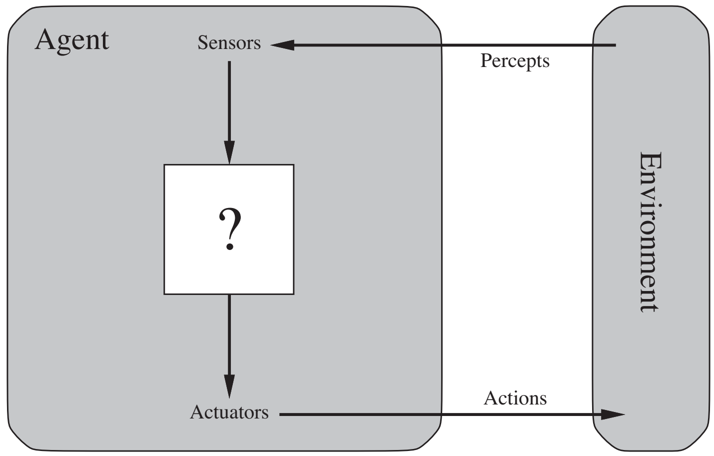
▷ We assume that agents can always perceive their own actions. (but not necessarily their consequences)

▷ **Problem:** agent functions can become very big (theoretical tool only)

▷ **Definition 5.2.6.** An agent function can be implemented by an agent program that runs on a physical agent architecture.

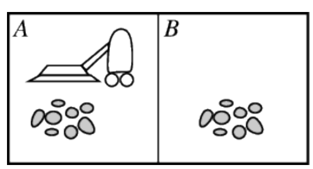
Agent Schema: Visualizing the Internal Agent Structure

▷ **Agent Schema:** We will use the following kind of schema to visualize the internal structure of an agent:



Different agents differ on the contents of the white box in the center.

Example: Vacuum-Cleaner World and Agent



- ▷ percepts: location and contents, e.g., $[A, Dirty]$
- ▷ actions: $Left$, $Right$, $Suck$, $NoOp$

Percept sequence	Action
$[A, Clean]$	$Right$
$[A, Dirty]$	$Suck$
$[B, Clean]$	$Left$
$[B, Dirty]$	$Suck$
$[A, Clean]$, $[A, Clean]$	$Right$
$[A, Clean]$, $[A, Dirty]$	$Suck$
$[A, Clean]$, $[B, Clean]$	$Left$
$[A, Clean]$, $[B, Dirty]$	$Suck$
$[A, Dirty]$, $[A, Clean]$	$Right$
$[A, Dirty]$, $[A, Dirty]$	$Suck$
⋮	⋮
$[A, Clean]$, $[A, Clean]$, $[A, Clean]$	$Right$
$[A, Clean]$, $[A, Clean]$, $[A, Dirty]$	$Suck$
⋮	⋮

- ▷ **Science Question:** What is the *right* agent function?
- ▷ **AI Question:** Is there an agent architecture and an agent program that implements it.

Example: Vacuum-Cleaner World and Agent

- ▷ **Example 5.2.7 (Agent Program).**

procedure Reflex−Vacuum−Agent $[location, status]$ **returns** an action
  if $status$ = Dirty **then return** Suck
  else if $location$ = A **then return** Right
  else if $location$ = B **then return** Left

Table-Driven Agents

- ▷ **Idea:** We can just implement the agent function as a table and look up actions.
- ▷ We can directly implement this:

function Table—Driven—Agent($percept$) **returns** an action
  persistent $table$ /* a table of actions indexed by percept sequences */
  var $percepts$ /* a sequence, initially empty */
  append $percept$ **to** the **end** of $percepts$
  $action$:= lookup($percepts$, $table$)
  return $action$

- ▷ **Problem:** Why is this not a good idea?
  - ▷ The table is much too large: even with n binary percepts whose order of occurrence does not matter, we have 2^n rows in the table.
  - ▷ Who is supposed to write this table anyways, even if it "only" has a million entries?

5.3 Good Behavior $\rightsquigarrow$ Rationality

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21844`.

Rationality

- ▷ **Idea:** Try to design agents that are successful! (aka. "do the right thing")
- ▷ **Definition 5.3.1.** A performance measure is a function that evaluates a sequence of environments.
- ▷ **Example 5.3.2.** A performance measure for the vacuum cleaner world could
  - ▷ award one point per square cleaned up in time T?
  - ▷ award one point per clean square per time step, minus one per move?
  - ▷ penalize for $> k$ dirty squares?
- ▷ **Definition 5.3.3.** An agent is called rational, if it chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date.
- ▷ **Question:** Why is rationality a good quality to aim for?

Consequences of Rationality: Exploration, Learning, Autonomy

- **Note:** a rational agent need not be perfect
  - only needs to maximize expected value (rational $\neq$ omniscient)
    - need not predict e.g. very unlikely but catastrophic events in the future
  - percepts may not supply all relevant information (rational $\neq$ clairvoyant)
    - if we cannot perceive things we do not need to react to them.
    - but we may need to try to find out about hidden dangers (exploration)
  - action outcomes may not be as expected (rational $\neq$ successful)
    - but we may need to take action to ensure that they do (more often) (learning)
- **Note:** rational $\rightsquigarrow$ exploration, learning, autonomy
- **Definition 5.3.4.** An agent is called autonomous, if it does not rely on the prior knowledge about the environment of the designer.
- Autonomy avoids fixed behaviors that can become unsuccessful in a changing environment. (anything else would be irrational)
- The agent has to learning agentlearn all relevant traits, invariants, properties of the environment and actions.

PEAS: Describing the Task Environment

- **Observation:** To design a rational agent, we must specify the task environment in terms of performance measure, environment, actuators, and sensors, together called the PEAS components.
- **Example 5.3.5.** When designing an automated taxi:
  - **Performance measure:** safety, destination, profits, legality, comfort, ...
  - **Environment:** US streets/freeways, traffic, pedestrians, weather, ...
  - **Actuators:** steering, accelerator, brake, horn, speaker/display, ...
  - **Sensors:** video, accelerometers, gauges, engine sensors, keyboard, GPS, ...
- **Example 5.3.6 (Internet Shopping Agent).**

  The task environment:
  - Performance measure: price, quality, appropriateness, efficiency
  - Environment: current and future WWW sites, vendors, shippers
  - Actuators: display to user, follow URL, fill in form
  - Sensors: HTML pages (text, graphics, scripts)

Examples of Agents: PEAS descriptions

Agent Type	Performance measure	Environment	Actuators	Sensors
Chess/Go player	win/loose/draw	game board	moves	board position
Medical diagnosis system	accuracy of diagnosis	patient, staff	display questions, diagnoses	keyboard entry of symptoms
Part-picking robot	percentage of parts in correct bins	conveyor belt with parts, bins	jointed arm and hand	camera, joint angle sensors
Refinery controller	purity, yield, safety	refinery, operators	valves, pumps, heaters, displays	temperature, pressure, chemical sensors
Interactive English tutor	student's score on test	set of students, testing accuracy	display exercises, suggestions, corrections	keyboard entry

Agents

- ▷ Which are agents?

  (A) James Bond.

  (B) Your dog.

  (C) Vacuum cleaner.

  (D) Thermometer.

- ▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

5.4 Classifying Environments

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21869`.

It is important to understand that the type of the environment has a very profound effect on the agent design. Depending on the type, different types of agents are needed to be successful. So before we discuss common types of agents in section 5.5, we will classify types of environments.

Environment types

- ▷ **Observation 5.4.1.** *Agent design is largely determined by the type of environment it is intended for.*
- ▷ **Problem:** There is a vast number of possible kinds of environments in AI.
- ▷ **Solution:** Classify along a few "dimensions". (independent characteristics)
- ▷ **Definition 5.4.2.** For an agent a we classify the environment e of a by its type, which is one of the following. We call e

1. fully observable, iff the a's sensors give it access to the complete state of the environment at any point in time, else partially observable.
2. deterministic, iff the next state of the environment is completely determined by the current state and a's action, else stochastic.
3. episodic, iff a's experience is divided into atomic episodes, where it perceives and then performs a single action. Crucially the next episode does not depend on previous ones. Non-episodic environments are called sequential.
4. dynamic, iff the environment can change without an action performed by a, else static. If the environment does not change but a's performance measure does, we call e semidynamic.
5. discrete, iff the sets of e's state and a's actions are countable, else continuous.
6. single agent, iff only a acts on e; else multi agent (when must we count parts of e as agents?)

Some examples will help us understand the classification of environments better.

Environment Types (Examples)

▷ **Example 5.4.3.** Some environments classified:

	Solitaire	Backgammon	Internet shopping	Taxi
fully observable	No	Yes	No	No
deterministic	Yes	No	Partly	No
episodic	No	No	No	No
static	Yes	Semi	Semi	No
discrete	Yes	Yes	Yes	No
single agent	Yes	No	Yes (except auctions)	No

▷ **Observation 5.4.4.** *The real world is (of course) a partially observable, stochastic, sequential, dynamic, continuous, and multi agent environment.* (worst case for AI)

In the AI-1 course we will work our way from the simpler environment types to the more general ones. Each environment type wil need its own agent types specialized to surviving and doing well in them.

5.5 Types of Agents

We will now discuss the main types of agents we will encounter in this course, get an impression of the variety, and what they can and cannot do. We will start from simple reflex agents, add state, and utility, and finally add learning. **A Video Nugget** covering this section can be found at `https://fau.tv/clip/id/21926`.

Agent types

▷ **Observation:** So fare we have described (and analyzed) agents only by their behavior (cf. agent function $f\colon \mathcal{P}^* \to \mathcal{A}$).

▷ **Problem:** This does not help us to build agents. (the goal of AI)

▷ To build an agent, we need to fix an agent architecture and come up with an agent program that runs on it.

▷ **Preview:** Four basic types of agent architectures in order of increasing generality:

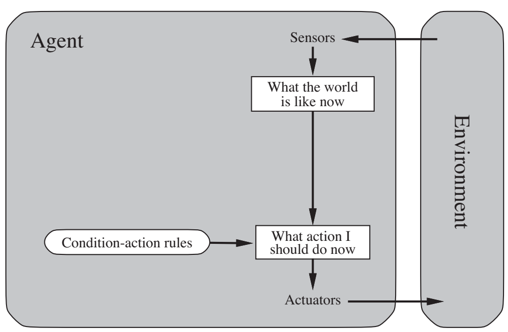
1. simple reflex agents
2. model-based agents
3. goal-based agents
4. utility-based agents

All these can be turned into learning agents.

Simple reflex agents

▷ **Definition 5.5.1.** A simple reflex agent is an agent a that only bases its actions on the last percept: so the agent function simplifies to $f_a\colon \mathcal{P} \to \mathcal{A}$.

▷ **Agent Schema:**



▷ **Example 5.5.2 (Agent Program).**

```
procedure Reflex—Vacuum—Agent [location,status] returns an action
  if status = Dirty then . . .
```


Simple reflex agents (continued)

▷ **General Agent Program:**

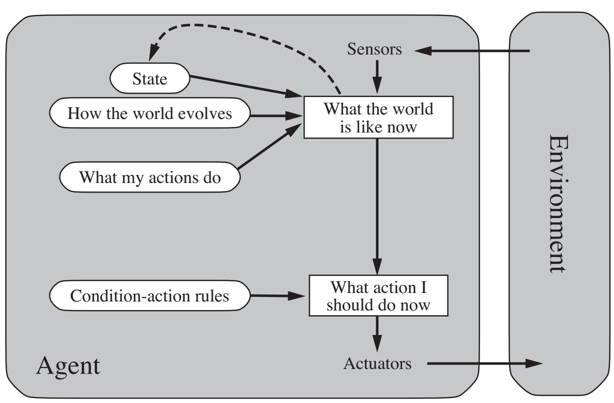
```
function Simple—Reflex—Agent (percept) returns an action
  persistent: rules /* a set of condition—action rules*/
```

```
state := Interpret−Input(percept)
rule := Rule−Match(state,rules)
action := Rule−action[rule]
return action
```

- **Problem:** Simple reflex agents can only react to the perceived state of the environment, not to changes.
- **Example 5.5.3.** Automobile tail lights signal braking by brightening. A simple reflex agent would have to compare subsequent percepts to realize.
- **Problem:** Partially observable environments get simple reflex agents into trouble.
- **Example 5.5.4.** Vacuum cleaner robot with defective location sensor $\leadsto$ infinite loops.

Model-based Reflex Agents: Idea

- **Idea:** Keep track of the state of the world we cannot see in an internal model.
- **Agent Schema:**



Model-based Reflex Agents: Definition

- **Definition 5.5.5.** A model-based agent (also called reflex agent with state) is an agent whose function depends on
  - a world model: a set $\mathcal{S}$ of possible states.
  - a sensor model S that given a state s and percepts determines a new state s'.
  - (optionally) a transition model T, that predicts a new state s'' from a state s' and an action a .
  - An action function f that maps (new) states to actions.

The agent function is iteratively computed via $e \mapsto f(S(s,e))$.

▷ **Note:** As different percept sequences lead to different states, so the agent function $f_a \colon \mathcal{P}^* \to \mathcal{A}$ no longer depends only on the last percept.

▷ **Example 5.5.6 (Tail Lights Again).** Model-based agents can do the 98 if the states include a concept of tail light brightness.

Model-Based Agents (continued)

▷ **Observation 5.5.7.** *The agent program for a model-based agent is of the following form:*

```
function Model—Based—Agent (percept) returns an action
  var state /* a description of the current state of the world */
  persistent rules /* a set of condition—action rules */
  var action /* the most recent action, initially none */

  state := Update—State(state,action,percept)
  rule := Rule—Match(state,rules)
  action := Rule—action(rule)
  return action
```

▷ **Problem:** Having a world model does not always determine what to do (rationally).

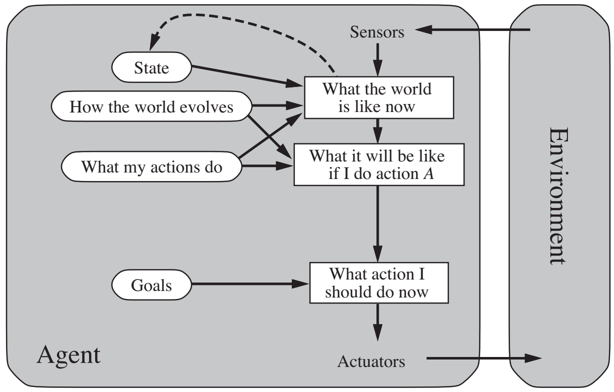
▷ **Example 5.5.8.** Coming to an intersection, where the agent has to decide between going left and right.

Goal-based Agents

▷ **Problem:** A world model does not always determine what to do (rationally).

▷ **Observation:** Having a goal in mind does! (determines future actions)

▷ **Agent Schema:**



Goal-based agents (continued)

▷ **Definition 5.5.9.** A goal-based agent is a model-based agent with transition model T that deliberates actions based on goals and a world model: It employs

▷ a set $\mathcal{G}$ of goals and a goal function f that given a (new) state s' selects an action a to best reach $\mathcal{G}$.

The action function is then $s \mapsto f(T(s), \mathcal{G})$.

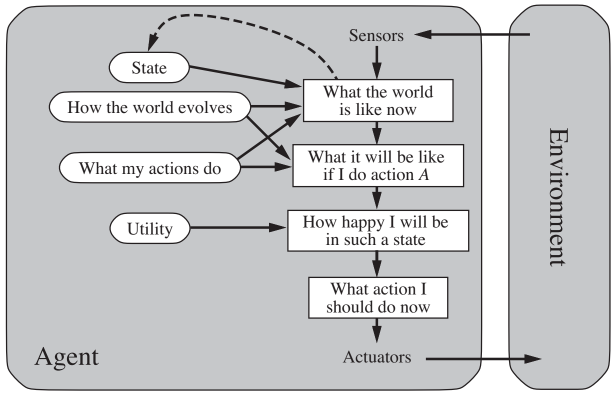
▷ **Observation:** A goal-based agent is more flexible in the knowledge it can utilize.

▷ **Example 5.5.10.** A goal-based agent can easily be changed to go to a new destination, a model-based agent's rules make it go to exactly one destination.

Utility-based Agents

▷ **Definition 5.5.11.** A utility-based agent uses a world model along with a utility function that models its preferences among the states of that world. It chooses the action that leads to the best expected utility.

▷ **Agent Schema:**



Utility-based vs. Goal-based Agents

▷ **Question:** What is the difference between goal-based and utility-based agents?

▷ **Utility-based Agents are a Generalization:** We can always force goal-directedness by a utility function that only rewards goal states.

▷ **Goal-based Agents can do less:** A utility function allows rational decisions where mere goals are inadequate:

- ▷ conflicting goals (utility gives tradeoff to make rational decisions)
- ▷ goals obtainable by uncertain actions (utility $\times$ likelihood helps)

Learning Agents

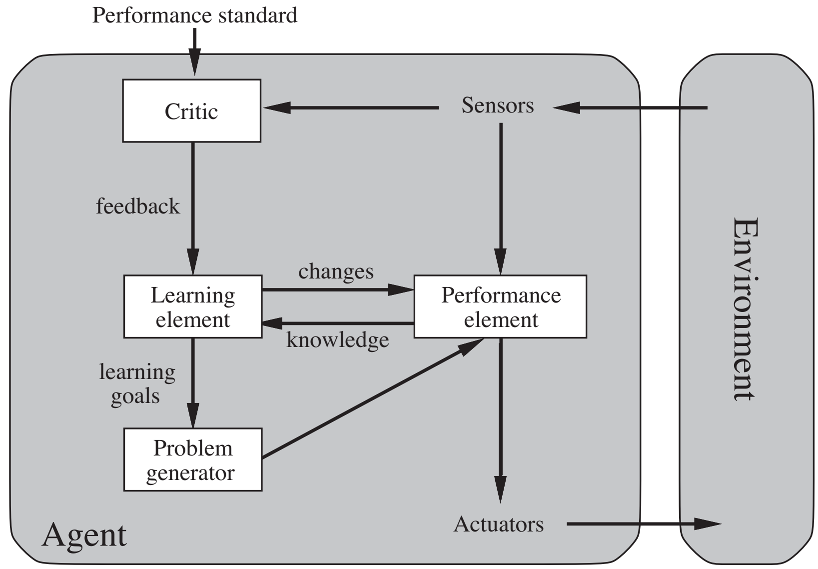
▷ **Definition 5.5.12.** A learning agent is an agent that augments the performance element – which determines actions from percept sequences with

- ▷ a learning element which makes improvements to the agent's components,
- ▷ a critic which gives feedback to the learning element based on an external performance standard,
- ▷ a problem generator which suggests actions that lead to new and informative experiences.

▷ The performance element is what we took for the whole agent above.

Learning Agents

▷ **Agent Schema:**



Learning Agents: Example

- ▷ **Example 5.5.13 (Learning Taxi Agent).** It has the components
  - ▷ Performance element: the knowledge and procedures for selecting driving actions. (this controls the actual driving)
  - ▷ critic: observes the world and informs the learning element (e.g. when passengers complain brutal braking)
  - ▷ Learning element modifies the braking rules in the performance element (e.g. earlier, softer)
  - ▷ Problem generator might experiment with braking on different road surfaces
- ▷ The learning element can make changes to any "knowledge components" of the diagram, e.g. in the
  - ▷ model from the percept sequence (how the world evolves)
  - ▷ success likelihoods by observing action outcomes (what my actions do)
- ▷ **Observation:** here, the passenger complaints serve as part of the "external performance standard" since they correlate to the overall outcome – e.g. in form of tips or blacklists.

Domain-Specific vs. General Agents

Domain-Specific Agent	vs.	General Agent
 Duell Kasparow gegen Deep Blue (1997): *Demütigende Niederlage*	vs.	
Solver specific to a particular problem ("domain").	vs.	Solver based on *description* in a general problem-description language (e.g., the rules of any board game).
More efficient.	vs.	Much less design/maintenance work.

▷ What kind of agent are you?

Michael Kohlhase: Artificial Intelligence 1 109 2023-09-20

5.6 Representing the Environment in Agents

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21925`.

We now come to a very important topic, which has a great influence on agent design: how does the agent represent the environment. After all, in all agent designs above (except the simple reflex agent) maintain a notion of world state and how the world state evolves given percepts and actions. The form of this model determines the algorithms.

Representing the Environment in Agents

▷ We have seen various components of agents that answer questions like

- ▷ *What is the world like now?*
- ▷ *What action should I do now?*
- ▷ *What do my actions do?*

▷ **Next natural question:** How do these work? (see the rest of the course)

▷ **Important Distinction:** How the agent implement the wold model.

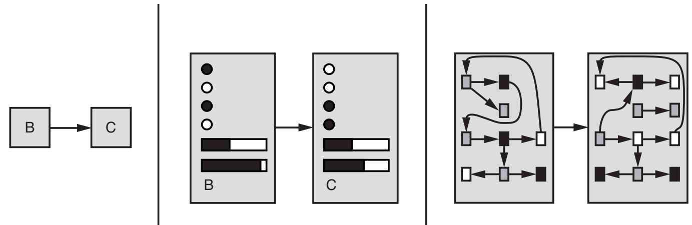
▷ **Definition 5.6.1.** We call a state representation

- ▷ atomic, iff it has no internal structure (black box)
- ▷ factored, iff each state is characterized by attributes and their values.
- ▷ structured, iff the state includes representations of objects and their relationships.

Michael Kohlhase: Artificial Intelligence 1 110 2023-09-20

Atomic/Factored/Structured State Representations

▷ **Schematically:** we can visualize the three kinds by



(a) Atomic (b) Factored (b) Structured

▷ **Example 5.6.2.** Consider the problem of finding a driving route from one end of a country to the other via some sequence of cities.

- ▷ In an atomic representation the state is represented by the name of a city.
- ▷ In a factored representation we may have attributes "gps-location", "gas",... (allows information sharing between states and uncertainty)
- ▷ But how to represent a situation, where a large truck blocking the road, since it is trying to back into a driveway, but a loose cow is blocking its path. (attribute "TruckAheadBackingIntoDairyFarmDrivewayBlockedByLooseCow" is unlikely)
- ▷ In a structured representation, we can have objects for trucks, cows, etc. and their relationships.

Summary

- ▷ Agents interact with environments through actuators and sensors.
- ▷ The agent function describes what the agent does in all circumstances.
- ▷ The performance measure evaluates the environment sequence.
- ▷ A perfectly rational agent maximizes expected performance.
- ▷ Agent programs implement (some) agent functions.
- ▷ PEAS descriptions define task environments.
- ▷ Environments are categorized along several dimensions: fully observable? deterministic? episodic? static? discrete? single agent?
- ▷ Several basic agent architectures exist: reflex, model-based, goal-based, utility-based

Part II

General Problem Solving

This part introduces search-based methods for general problem solving using atomic and factored representations of states.

Concretely, we discuss the basic techniques of search-based symbolic AI. First in the shape of classical and heuristic search and adversarial search paradigms. Then in constraint propagation, where we see the first instances of inference-based methods.

Chapter 6

Problem Solving and Search

In this chapter, we will look at a class of algorithms called search algorithms. These are algorithms that help in quite general situations, where there is a precisely described problem, that needs to be solved. Hence the name "General Problem Solving" for the area.

6.1 Problem Solving

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21927`.

Before we come to the search algorithms themselves, we need to get a grip on the types of problems themselves and how we can represent them, and on what the variuous types entail for the problem solving process.
The first step is to classify the problem solving process by the amount of knowledge we have available. It makes a difference, whether we know all the factors involved in the problem before we actually are in the situation. In this case, we can solve the problem in the abstract, i.e. make a plan before we actually enter the situation (i.e. offline), and then when the problem arises, only execute the plan. If we do not have complete knowledge, then we can only make partial plans, and have to be in the situation to obtain new knowledge (e.g. by observing the effects of our actions or the actions of others). As this is much more difficult we will restrict ourselves to offline problem solving.

Problem Solving: Introduction

- ▷ **Recap:** Agents perceive the environment and compute an action.
- ▷ **In other words:** Agents continually solve "the problem of what to do next".
- ▷ **AI Goal:** Find algorithms that help solving problems in general.
- ▷ **Idea:** If we can describe/represent problems in a standardized way, we may have a chance to find general algorithms.
- ▷ **Concretely:** We will use the following two concepts to describe problems
  - ▷ States: A set of possible situations in our problem domain (≘ environments)
  - ▷ Actions: that get us from one state to another. (≘ agents)

  A sequence of actions is a solution, if it brings us from an initial state to a goal state. Problem solving computes solutions from problem formulations.

▷ **Definition 6.1.1.** In offline problem solving an agent computing an action sequence based complete knowledge of the environment.

▷ *Remark 6.1.2.* Offline problem solving only works in fully observable, deterministic, static, and episodic environments.

▷ **Definition 6.1.3.** In online problem solving an agent computes one action at a time based on incoming perceptions.

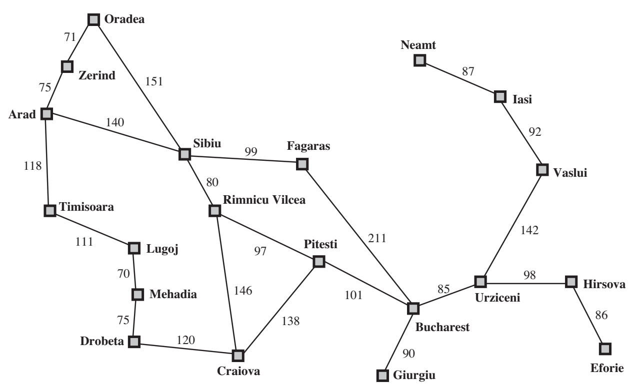
▷ **This Semester:** We largely restrict ourselves to offline problem solving. (easier)

Michael Kohlhase: Artificial Intelligence 1 113 2023-09-20

We will use the following problem as a running example. It is simple enough to fit on one slide and complex enough to show the relevant features of the problem solving algorithms we want to talk about.

Example: Traveling in Romania

▷ **Scenario:** An agent is on holiday in Romania; currently in Arad; flight home leaves tomorrow from Bucharest; how to get there? We have a map:



▷ **Formulate the Problem:**

- ▹ States: various cities.
- ▹ Actions: drive between cities.

▷ **Solution:** Appropriate sequence of cities, e.g.: Arad, Sibiu, Fagaras, Bucharest

Michael Kohlhase: Artificial Intelligence 1 114 2023-09-20

Given this example to fortify our intuitions, we can now turn to the formal definition of problem formulation and their solutions.

Problem Formulation

▷ **Definition 6.1.4.** A problem formulation models a situation using states and actions at an appropriate level of abstraction. (do not model things like "put on my left sock", etc.)

- ▹ it describes the initial state (we are in Arad)

- it also limits the objectives by specifying goal states. (excludes, e.g. to stay another couple of weeks.)

A solution is a sequence of actions that leads from the initial state to a goal state.

Problem solving computes solutions from problem formulations.

- Finding the right level of abstraction and the required (not more!) information is often the key to success.

The Math of Problem Formulation: Search Problems

- **Definition 6.1.5.** A search problem $\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ consists of a set $\mathcal{S}$ of states, a set $\mathcal{A}$ of actions, and a transition model $\mathcal{T}\colon\mathcal{A}\times\mathcal{S}\to\mathcal{P}(\mathcal{S})$ that assigns to any action $a\in\mathcal{A}$ and state $s\in\mathcal{S}$ a set of successor states.

  Certain states in $\mathcal{S}$ are designated as goal states ($\mathcal{G}\subseteq\mathcal{S}$) and initial states $\mathcal{I}\subseteq\mathcal{S}$.
- **Definition 6.1.6.** We say that an action $a\in\mathcal{A}$ is applicable in a state $s\in\mathcal{S}$, iff $\mathcal{T}(a,s)\neq\emptyset$. We call $\mathcal{T}_a\colon\mathcal{S}\to\mathcal{P}(\mathcal{S})$ with $\mathcal{T}_a(s){:=}\mathcal{T}(a,s)$ the result relation for a and $\mathcal{T}_\mathcal{A}{:=}\bigcup_{a\in\mathcal{A}}\mathcal{T}_a$ the result relation of Π. The graph $\langle\mathcal{S},\mathcal{T}_\mathcal{A}\rangle$ is called the state space induced by Π.
- **Definition 6.1.7.** A solution for a search problem $\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ consists of a sequence $a_1,\ldots,a_n$ of actions such that for all $1\leq i<n$
  - a_i is applicable to state $s_{(i-1)}$, where $s_0\in\mathcal{I}$,
  - $s_i\in\mathcal{T}_{a_i}(s_{(i-1)})$, and $s_n\in\mathcal{G}$.
- **Idea:** A solution bring us from $\mathcal{I}$ to a goal state.
- **Definition 6.1.8.** Often we add a cost function $c\colon\mathcal{A}\to\mathbb{R}_0^+$ that associates a step cost $c(a)$ to an action $a\in\mathcal{A}$. The cost of a solution is the sum of the step costs of its actions.

Observation:

The formulation of problems from Definition 6.1.5 uses an atomic (black-box) state representation. It has enough functionality to construct the state space but nothing else. We will come back to this in slide 119.

Remark 6.1.9. Note that search problems formalize problem formulations by making many of the implicit constraints explicit.

Structure Overview: Search Problem

- The structure overview for search problems:

$$\text{search problem} = \left\langle \begin{array}{lll} \mathcal{S} & \text{set} & \text{states}, \\ \mathcal{A} & \text{set} & \text{actions}, \\ \mathcal{T} & \mathcal{A}\times\mathcal{S}\to\mathcal{P}(\mathcal{S}) & \text{transition model}, \\ \mathcal{I} & \mathcal{S} & \text{initial state}, \\ \mathcal{G} & \mathcal{S} & \text{goal state} \end{array} \right\rangle$$

We will now specialize Definition 6.1.5 to deterministic, fully observable environments, i.e. environments where actions only have one – assured – outcome state.

Search Problems in deterministic, fully observable Environments

- This semester, we will restrict ourselves to search problems, where(extend in AI II)
  - $|\mathcal{T}(a,s)|\leq 1$ for the transition models and ($\leftsquigarrow$ deterministic environment)
  - $\mathcal{I} = \{s_0\}$ ($\leftsquigarrow$ fully observable environment)
- **Definition 6.1.10.** Then $\mathcal{T}_a$ induces partial function $S_a\colon \mathcal{S}\rightharpoonup\mathcal{S}$ whose natural domain is the set of states where a is applicable: $S(s){:=}s'$ if $\mathcal{T}_a = \{s'\}$ and undefined at s otherwise.

  We call S_a the successor function for a and $S_a(s)$ the successor state of s. $S_{\mathcal{A}}{:=}\bigcup_{a\in\mathcal{A}} S_a$ the successor relation of $\mathcal{P}$.
- **Definition 6.1.11.** The predicate that tests for goal states is called a goal test.

Blackbox/Declarative Problem Descriptions

- **Observation:** $\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ from Definition 6.1.5 is essentially a blackbox description; it (think programming API)
  - provides the functionality needed to construct a state space, but
  - gives the algorithm no information about the problem.
- **Definition 6.1.12.** A declarative description (also called whitebox description) describes the problem itself $\leadsto$ problem description language
- **Example 6.1.13 (Planning Problems as Declarative Descriptions).**

  The STRIPS language describes planning problems in terms of
  - a set P of propositional variables (propositions)
  - a set $I \subseteq P$ of propositions true in the initial state.
  - a set $G \subseteq P$, where state $s \subseteq P$ is a goal if $G \subseteq s$
  - a set A of actions, each $a\in A$ with preconditions pre_a, add list add_a, and delete list del_a: a is applicable, if $\mathrm{pre}_a \subseteq s$, the result state is then $s \cup \mathrm{add}_a \backslash \mathrm{del}_a$,
  - a function c that maps all actions a to their cost $c(a)$.

▷ **Observation 6.1.14.** *Declarative descriptions are strictly more powerful than blackbox descriptions: they induce blackbox descriptions, but also allow to analyze/simplify the problem.*

▷ We will come back to this later ⇝ planning.

6.2 Problem Types

Note that the definition of a search problem is very general, it applies to many many real-world problems. So we will try to characterize these by difficulty. **A Video Nugget** covering this section can be found at `https://fau.tv/clip/id/21928`.

Problem types

▷ **Definition 6.2.1.** A search problem is called a single state problem, iff it is

- ▷ fully observable (at least the initial state)
- ▷ deterministic (i.e. the successor of each state is determined)
- ▷ static (states do not change other than by our own actions)
- ▷ discrete (a countable number of states)

▷ **Definition 6.2.2.** A search problem is called a multi state problem

- ▷ states partially observable (e.g. multiple initial states)
- ▷ deterministic, static, discrete

▷ **Definition 6.2.3.** A search problem is called a contingency problem, iff

- ▷ the environment is non deterministic (solution can branch, depending on contingencies)
- ▷ the state space is unknown (like a baby, agent has to learn about states and actions)

We will explain these problem types with another example. The problem $\mathcal{P}$ is very simple: We have a vacuum cleaner and two rooms. The vacuum cleaner is in one room at a time. The floor can be dirty or clean.

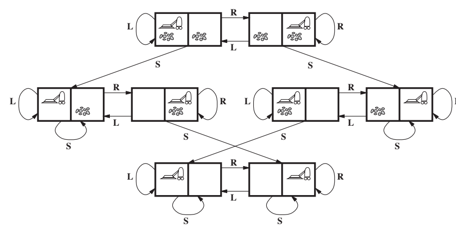
The possible states are determined by the position of the vacuum cleaner and the information, whether each room is dirty or not. Obviously, there are eight states: $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ for simplicity.

The goal is to have both rooms clean, the vacuum cleaner can be anywhere. So the set $\mathcal{G}$ of goal states is $\{7, 8\}$. In the single-state version of the problem, $[right, suck]$ shortest solution, but $[suck, right, suck]$ is also one. In the multiple-state version we have

$$[right\{2, 4, 6, 8\}, suck\{4, 8\}, left\{3, 7\}, suck\{7\}]$$

Example: vacuum-cleaner world

- **Single-state Problem:**
  - Start in 5
  - **Solution**: $[right, suck]$



- **Multiple-state Problem:**
  - Start in $\{1, 2, 3, 4, 5, 6, 7, 8\}$
  - **Solution**: $[right, suck, left, suck]$

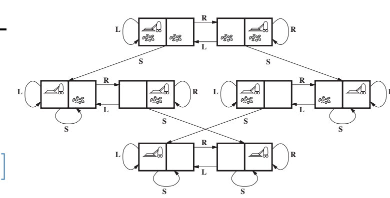
$$\begin{array}{ll} right & \to \{2, 4, 6, 8\} \\ suck & \to \{4, 8\} \\ left & \to \{3, 7\} \\ suck & \to \{7\} \end{array}$$

Michael Kohlhase: Artificial Intelligence 1 121 2023-09-20

Example: Vacuum-Cleaner World (continued)

- **Contingency Problem:**
  - Murphy's Law: $suck$ can dirty a clean carpet
  - Local sensing: $dirty/notdirty$ at location only
  - Start in: $\{1, 3\}$
  - **Solution**: $[suck, right, suck]$

$$\begin{array}{ll} suck & \to \{5, 7\} \\ right & \to \{6, 8\} \\ suck & \to \{6, 8\} \end{array}$$



- **better:** $[suck, right, \textbf{if } dirt \textbf{ then } suck]$ (decide whether in 6 or 8 using local sensing)

Michael Kohlhase: Artificial Intelligence 1 122 2023-09-20

In the contingency version of $\mathcal{P}$ a solution is the following:

$$[suck\{5, 7\}, right \to \{6, 8\}, suck \to \{\mathbf{6}, 8\}, suck\{\mathbf{5}, 7\}]$$

etc. Of course, local sensing can help: narrow $\{6, 8\}$ to $\{6\}$ or $\{8\}$, if we are in the first, then suck.

Single-state problem formulation

▷ Defined by the following four items

1. Initial state: (e.g. $\mathcal{S}Arad$)
2. Successor function S: (e.g. $S(\mathcal{S}Arad) = \{\langle goZer, \text{Zerind}\rangle, \langle goSib, \text{Sibiu}\rangle, \ldots\}$)
3. Goal test: (e.g. $x = \mathcal{S}Bucharest$ (explicit test))
   $noDirt(x)$ (implicit test)
4. Path cost (optional):(e.g. sum of distances, number of operators executed, etc.)

▷ Solution: A sequence of actions leading from the initial state to a goal state.

Michael Kohlhase: Artificial Intelligence 1 123 2023-09-20

"Path cost": There may be more than one solution and we might want to have the "best" one in a certain sense.

Selecting a state space

▷ **Abstraction:** Real world is absurdly complex!
State space must be abstracted for problem solving.

▷ **(Abstract) state:** Set of real states.

▷ **(Abstract) operator:** Complex combination of real actions.

▷ **Example:** $Arad \rightarrow Zerind$ represents complex set of possible routes.

▷ **(Abstract) solution:** Set of real paths that are solutions in the real world.

Michael Kohlhase: Artificial Intelligence 1 124 2023-09-20

"State": e.g., we don't care about tourist attractions found in the cities along the way. But this is problem dependent. In a different problem it may well be appropriate to include such information in the notion of state.

"Realizability": one could also say that the abstraction must be sound wrt. reality.

Example: The 8-puzzle

7	2	4
5		6
8	3	1

Start State

	1	2
3	4	5
6	7	8

Goal State

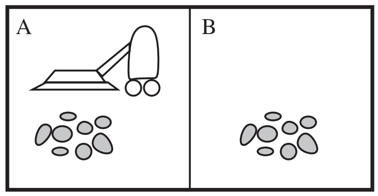
States	integer locations of tiles
Actions	$left, right, up, down$
Goal test	= goal state?
Path cost	1 per move

Michael Kohlhase: Artificial Intelligence 1 125 2023-09-20

How many states are there? N factorial, so it is not obvious that the problem is in **NP**. One needs to show, for example, that polynomial length solutions do always exist. Can be done by

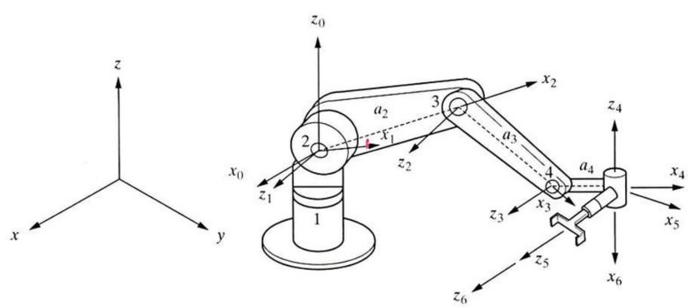
combinatorial arguments on state space graph (really ?).
Some rule-books give a different goal state for the 8-puzzle: starting with 1, 2, 3 in the top row and having the hold in the lower right corner. This is completely irrelevant for the example and its significance to AI-1.

Example: Vacuum-cleaner



States	integer dirt and robot locations
Actions	$left, right, suck, noOp$
Goal test	$notdirty?$
Path cost	1 per operation (0 for $noOp$)

Example: Robotic assembly



States	real-valued coordinates of robot joint angles and parts of the object to be assembled
Actions	continuous motions of robot joints
Goal test	assembly complete?
Path cost	time to execute

General Problems

▷ **Question:** Which are "Problems"?

(A) You didn't understand any of the lecture.

(B) Your bus today will probably be late.

(C) Your vacuum cleaner wants to clean your apartment.

(D) You want to win a chess game.

▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

6.3 Search

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/21956`.

Tree Search Algorithms

▷ **Note:** The state space of a search problem $\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ is a graph $\langle\mathcal{S},\mathcal{T}_{\mathcal{A}}\rangle$.

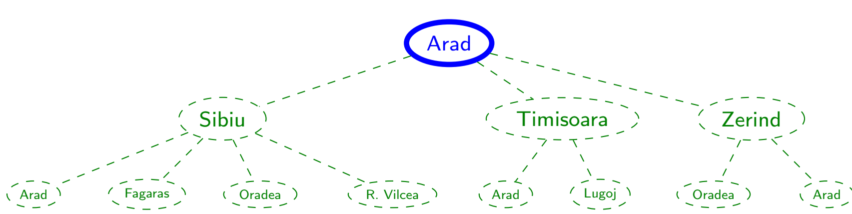
▷ As graphs are difficult to compute with, we often compute a corresponding tree and work on that. (standard trick in graph algorithms)

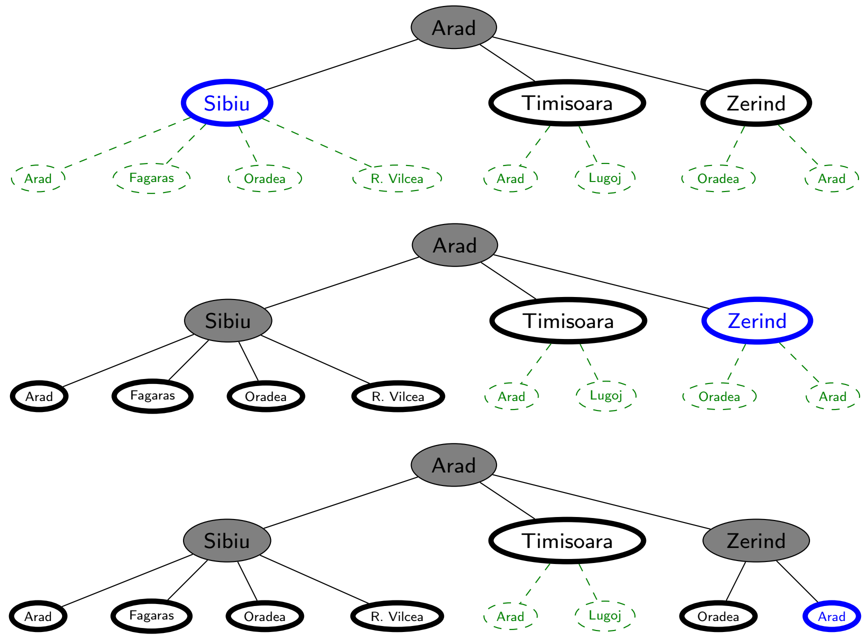
▷ **Definition 6.3.1.** Given a search problem $\mathcal{P}{:=}\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$, the tree search algorithm consists of the simulated exploration of state space $\langle\mathcal{S},\mathcal{T}_{\mathcal{A}}\rangle$ in a search tree formed by successively expanding already explored states. (offline algorithm)

```
procedure Tree—Search (problem, strategy) : <a solution or failure>
  <initialize the search tree using the initial state of problem>
  loop
     if <there are no candidates for expansion> <return failure> end if
    <choose a leaf node for expansion according to strategy>
     if <the node contains a goal state> return <the corresponding solution>
     else <expand the node and add the resulting nodes to the search tree>
     end if
  end loop
end procedure
```

We expand a node n by generating all successors of n and inserting them as children of n in the search tree.

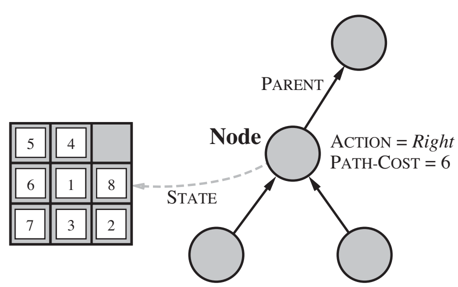
Tree Search: Example





Implementation: States vs. nodes

- ▷ **Recap:** A state is a (representation of) a physical configuration.
- ▷ **Remark:** The nodes of a search tree are implemented as a data structure that includes accessors for parent, children, depth, path cost, etc.



- ▷ **Observation:** Paths in the search tree correspond to paths in the state space.
- ▷ **Definition 6.3.2.** We define the path cost of a node n in a search tree T to be the sum of the step costs on the path from n to the root of T.

Implementation of Search Algorithms

```
procedure Tree_Search (problem,strategy)
  fringe := insert(make_node(initial_state(problem)))
```

```
    loop
      if fringe <is empty> fail end if
        node := first(fringe,strategy)
        if NodeTest(State(node)) return State(node)
        else fringe := insert_all(expand(node,problem),strategy)
        end if
    end loop
end procedure
```

▷ **Definition 6.3.3.** The fringe is a list nodes not yet considered in tree search.

▷

It is ordered by the strategy. (see below)

- State gives the state that is represented by *node*
- Expand = creates new nodes by applying possible actions to *node*
- Make-Queue creates a queue with the given elements.
- fringe holds the queue of nodes not yet considered.
- Remove-First returns first element of queue and as a side effect removes it from fringe.
- State gives the state that is represented by *node*.
- Expand applies all operators of the problem to the current node and yields a set of new nodes.
- Insert inserts an element into the current *fringe* queue. This can change the behavior of the search.
- Insert-All Perform Insert on set of elements.

Search strategies

▷ **Definition 6.3.4.** A strategy is a function that picks a node from the fringe of a search tree. (equivalently, orders the fringe and picks the first.)

▷ **Definition 6.3.5 (Important Properties of Strategies).**

completeness	does it always find a solution if one exists?
time complexity	number of nodes generated/expanded
space complexity	maximum number of nodes in memory
optimality	does it always find a least cost solution?

▷ **Time and space complexity measured in terms of:**

b	maximum branching factor of the search tree
d	minimal graph depth of a solution in the search tree
m	maximum graph depth of the search tree (may be ∞)

Complexity means here always *worst-case* complexity!

Note that there can be infinite branches, see the search tree for Romania.

6.4 Uninformed Search Strategies

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/21994` and `https://fau.tv/clip/id/21995`.

Uninformed search strategies

- **Definition 6.4.1.** We speak of an uninformed search algorithm, if it only uses the information available in the problem definition.
- **Next:** Frequently used search algorithms
  - Breadth first search
  - Uniform cost search
  - Depth first search
  - Depth limited search
  - Iterative deepening search

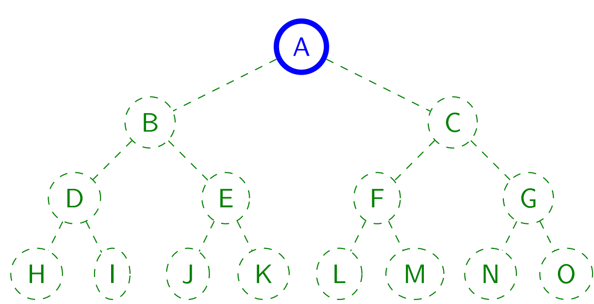
The opposite of uninformed search is informed or heuristic search that uses a heuristicheuristic function that adds external guidance to the search process. In the Romania example, one could add the heuristic to prefer cities that lie in the general direction of the goal (here SE).

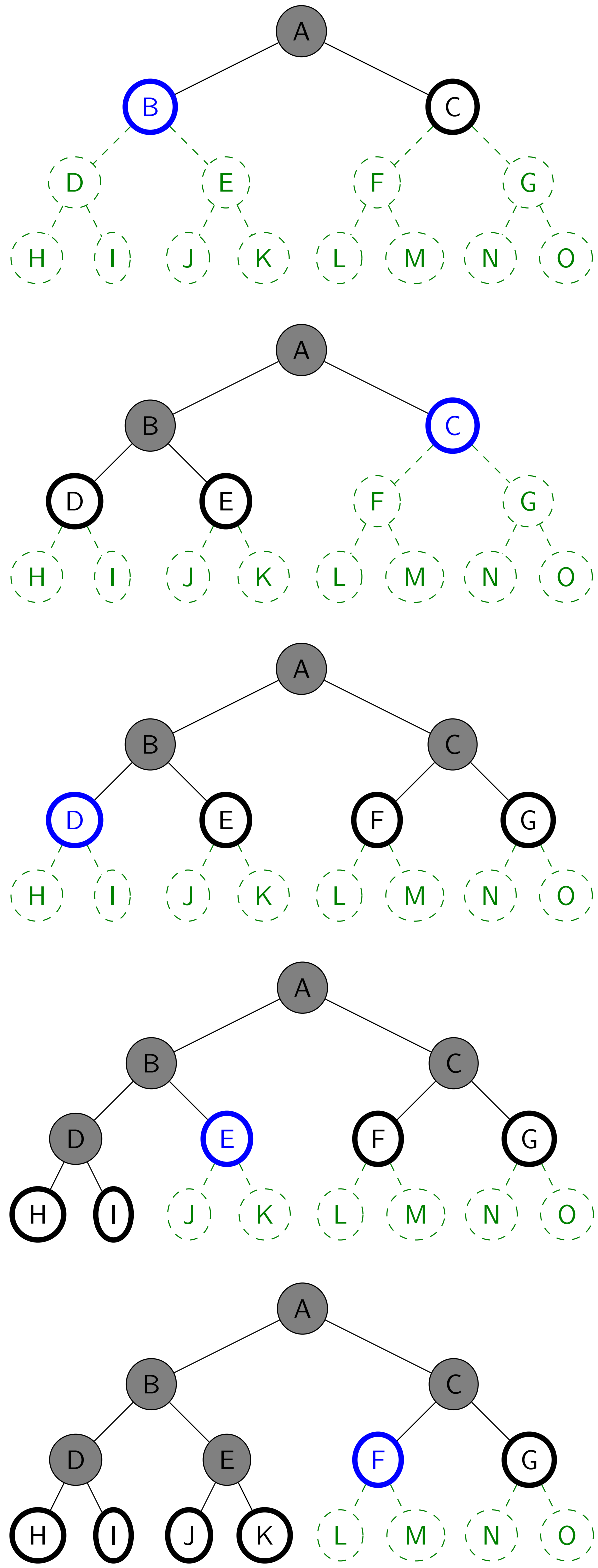
Even though heuristic search is usually much more efficient, uninformed search is important nonetheless, because many problems do not allow to extract good heuristics.

6.4.1 Breadth-First Search Strategies

Breadth-First Search

- **Idea:** Expand the shallowest unexpanded node.
- **Definition 6.4.2.** The breadth first search (BFS) strategy treats the fringe as a FIFO queue, i.e. successors go in at the end of the fringe.
- **Example 6.4.3 (Synthetic).**

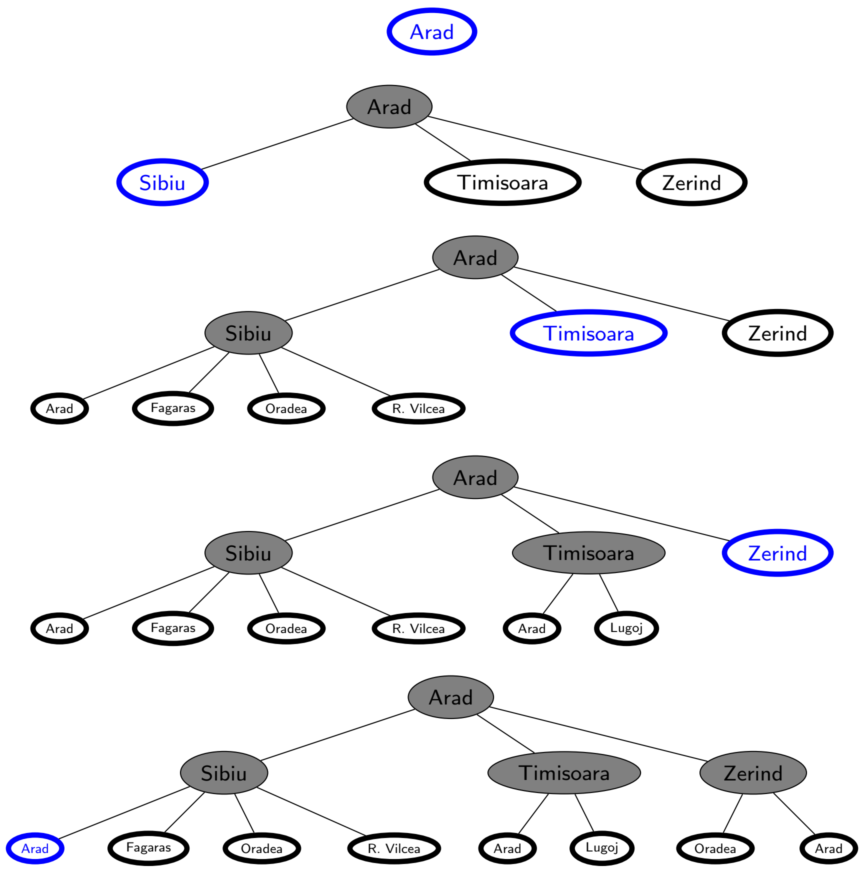




We will now apply the breadth first search strategy to our running example: Traveling in Romania.

Note that we leave out the green dashed nodes that allow us a preview over what the search tree will look like (if expanded). This gives a much cleaner picture we assume that the readers already have grasped the mechanism sufficiently.

Breadth-First Search: Romania



Breadth-first search: Properties

▷

Completeness	Yes (if b is finite)
Time complexity	$1 + b + b^2 + b^3 + \ldots + b^d$, so $\mathcal{O}(b^d)$, i.e. exponential in d
Space complexity	$\mathcal{O}(b^d)$ (fringe may be whole level)
Optimality	Yes (if cost = 1 per step), not optimal in general

▷ **Disadvantage:** Space is the big problem (can easily generate nodes at 500MB/sec $\widehat{=}$ 1.8TB/h)

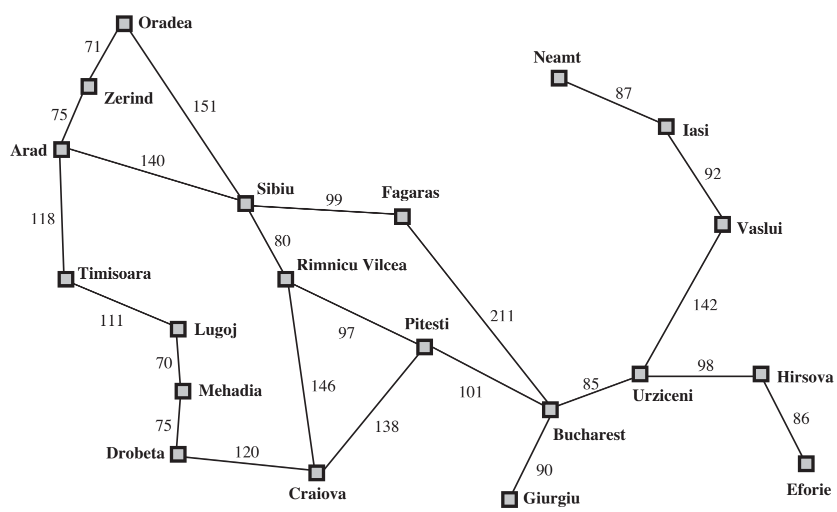
▷ **Optimal?:** No! If cost varies for different steps, there might be better solutions below the level of the first one.

▷ An alternative is to generate *all* solutions and then pick an optimal one. This works only, if m is finite.

The next idea is to let cost drive the search. For this, we will need a non-trivial cost function: we will take the distance between cities, since this is very natural. Alternatives would be the driving time, train ticket cost, or the number of tourist attractions along the way.

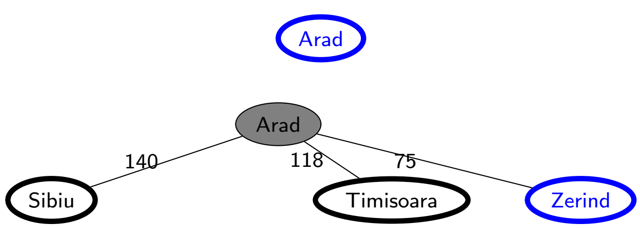
Of course we need to update our problem formulation with the necessary information.

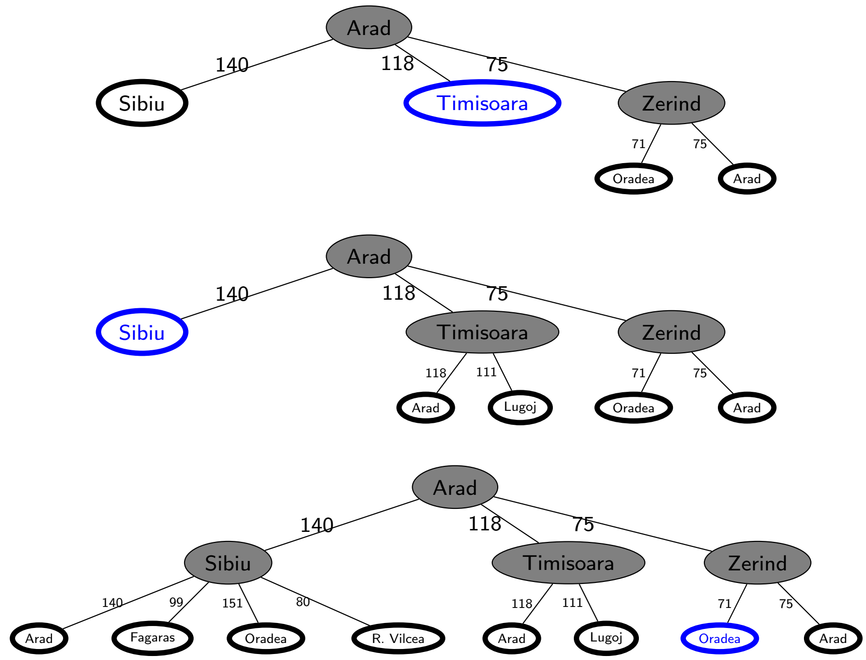
Romania with Step Costs as Distances



Uniform-cost search

- **Idea:** Expand least cost unexpanded node.
- **Definition 6.4.4.** Uniform-cost search (UCS) is the strategy where the fringe is ordered by increasing path cost.
- **Note:**
  Equivalent to breadth first search if all step costs are equal.
- **Synthetic Example:**





Michael Kohlhase: Artificial Intelligence 1 139 2023-09-20

Note that we must sum the distances to each leaf. That is, we go back to the first level after the third step.

Uniform-cost search: Properties

Completeness	Yes (if step costs $\geq \epsilon > 0$)
Time complexity	number of nodes with path cost less than that of optimal solution
Space complexity	number of nodes with path cost less than that of optimal solution
Optimality	Yes

Michael Kohlhase: Artificial Intelligence 1 140 2023-09-20

If step cost is negative, the same situation as in breadth first search can occur: later solutions may be cheaper than the current one.

If step cost is 0, one can run into infinite branches. UCS then degenerates into depth first search, the next kind of search algorithm we will encounter. Even if we have infinite branches, where the sum of step costs converges, we can get into trouble, since the search is forced down these infinite paths before a solution can be found.

Worst case is often worse than BFS, because large trees with small steps tend to be searched first. If step costs are uniform, it degenerates to BFS.

6.4.2 Depth-First Search Strategies

Depth-first Search

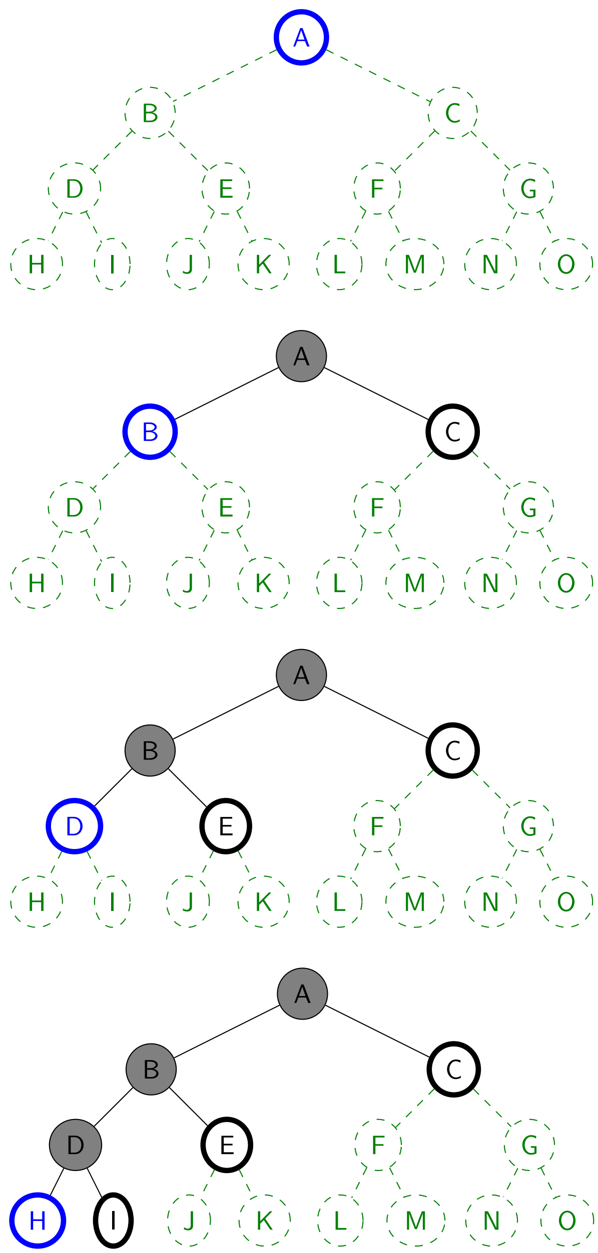
▷ **Idea:** Expand deepest unexpanded node.

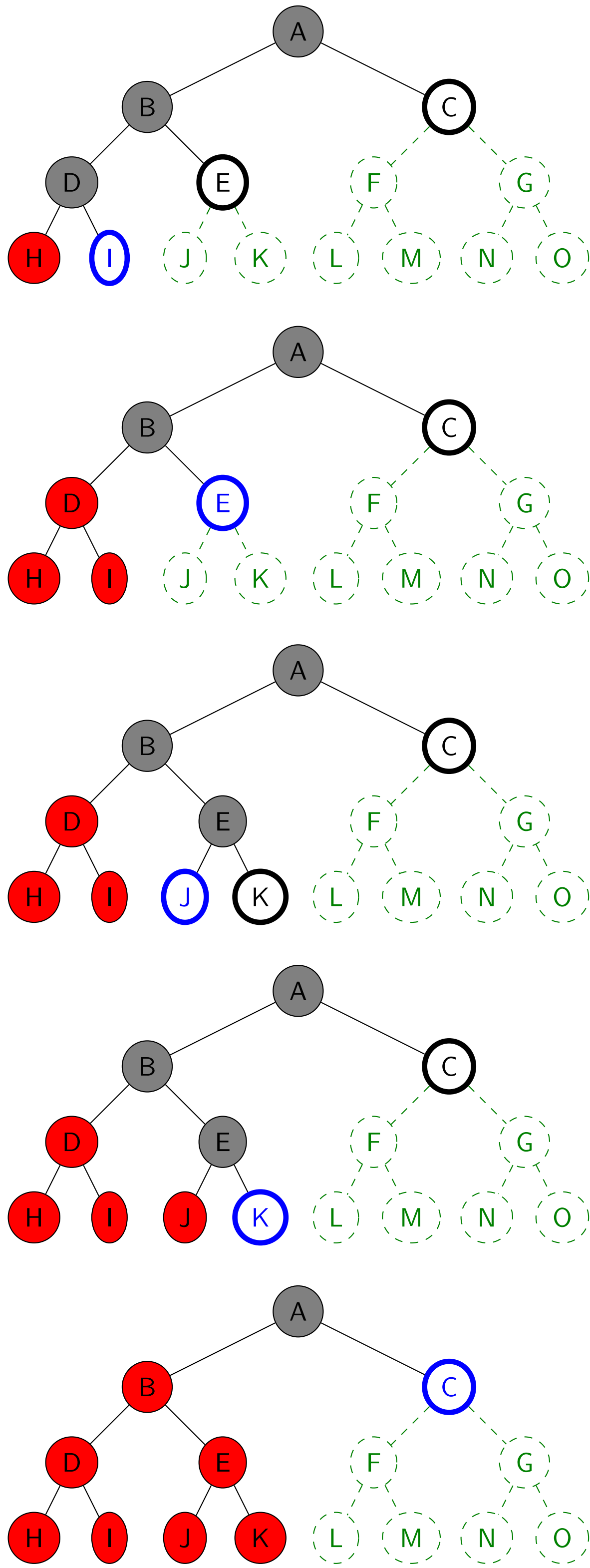
▷ **Definition 6.4.5.** Depth-first search (DFS) is the strategy where the fringe is organized as a (LIFO) stack i.e. successor go in at front of the fringe.

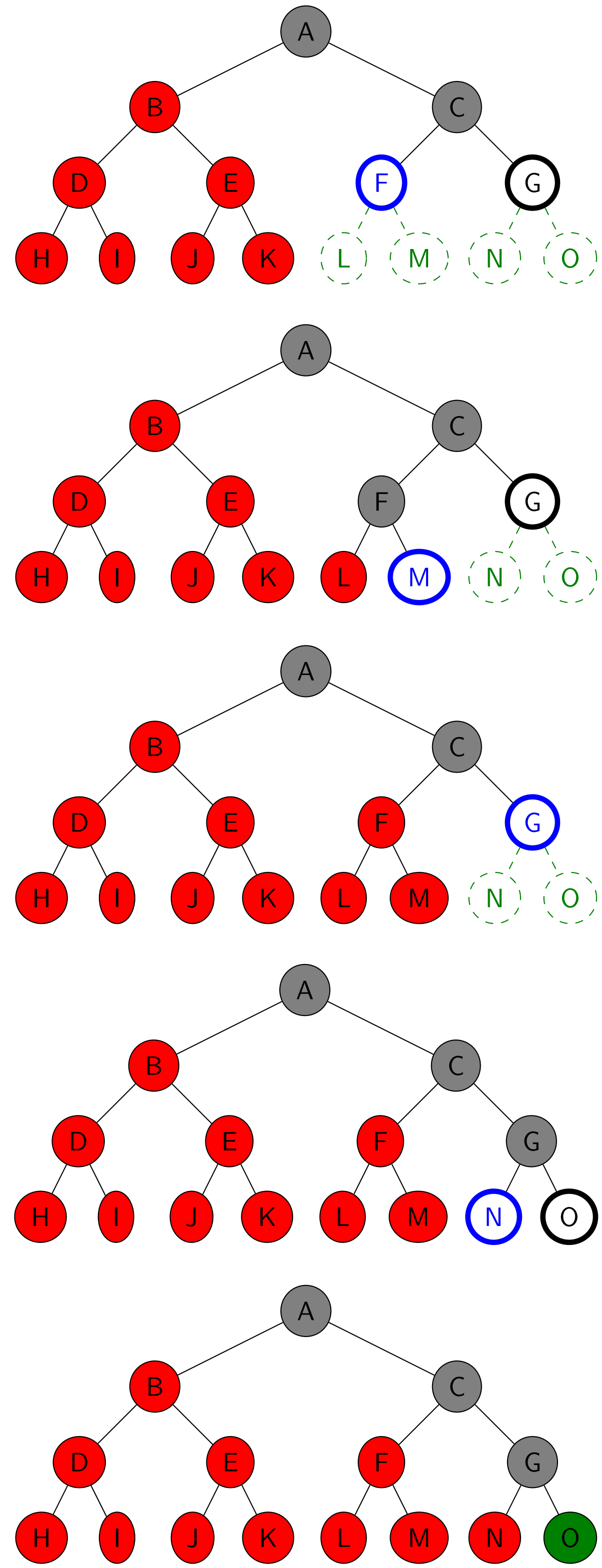
▷ **Note:** Depth first search can perform infinite cyclic excursions
Need a finite, non cyclic state space (or repeated state checking)

Depth-First Search

▷ **Example 6.4.6 (Synthetic).**

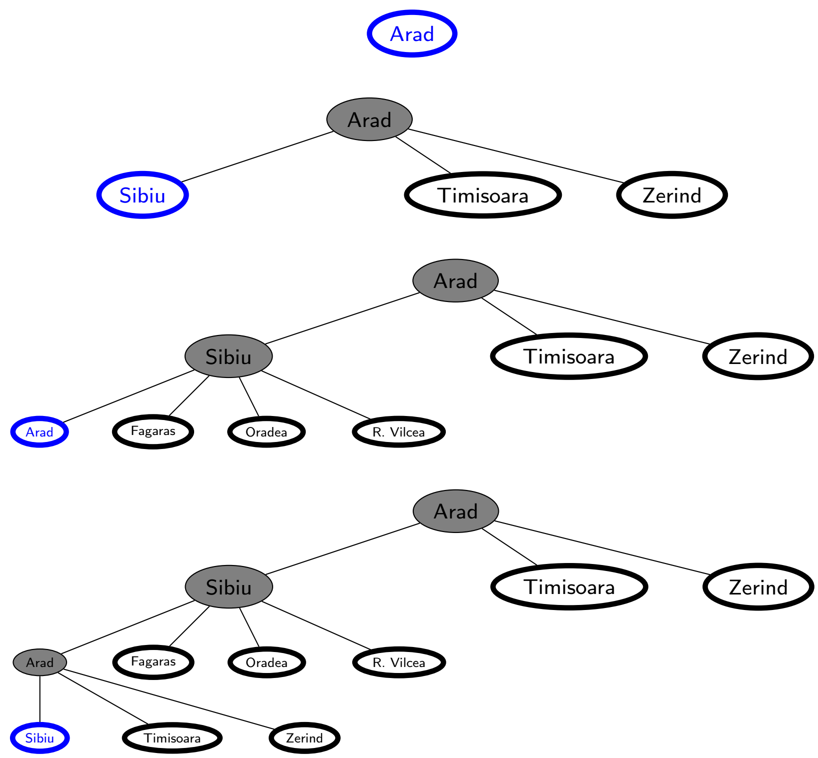



A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O

Depth-First Search: Romania

▷ **Example 6.4.7 (Romania).**



Depth-first search: Properties

▷

Completeness	Yes: if state space finite No: if search tree contains infinite paths or loops
Time complexity	$\mathcal{O}(b^m)$ (we need to explore until max depth m in any case!)
Space complexity	$\mathcal{O}(bm)$ (i.e. linear space) (need at most store m levels and at each level at most b nodes)
Optimality	No (there can be many better solutions in the unexplored part of the search tree)

▷ **Disadvantage:** Time terrible if m much larger than d.

▷ **Advantage:** Time may be much less than breadth first search if solutions are dense.

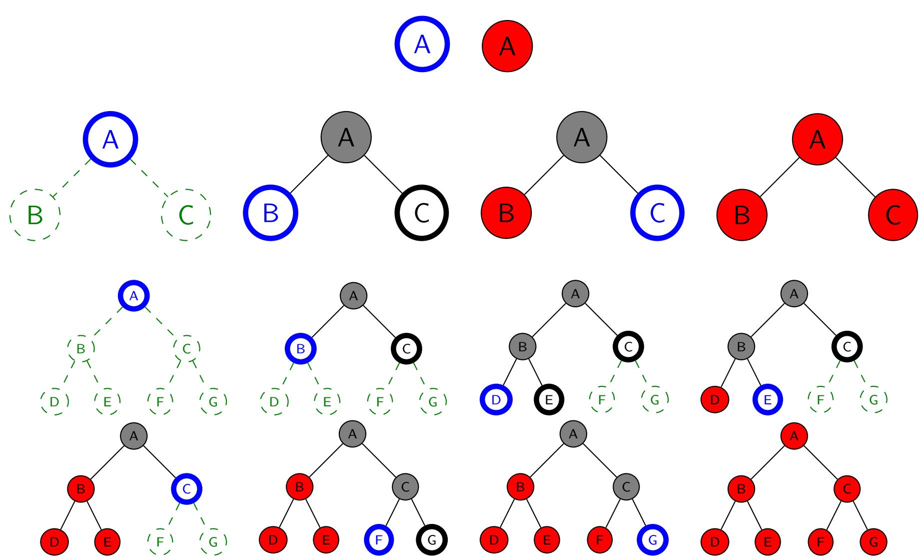
Iterative deepening search

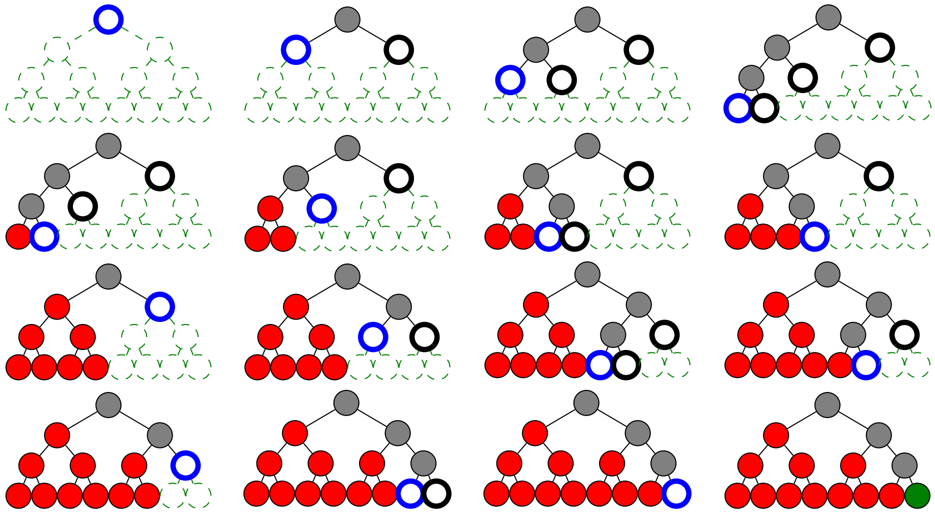
▷ **Definition 6.4.8.** Depth limited search is depth first search with a depth limit.

▷ **Definition 6.4.9.** Iterative deepening search (IDS) is depth limited search with ever increasing depth limits.

▷ **procedure** Tree_Search (problem)
  <initialize the search tree using the initial state of problem>
  for depth = 0 **to** ∞
    result := Depth_Limited_search(problem,depth)
    if depth $\neq$ cutoff **return** result **end if**
  end for
end procedure

Ilustration: Iterative Deepening Search at various Limit Depths





Iterative deepening search: Properties

▷

Completeness	Yes
Time complexity	$(d+1)\cdot b^0 + d\cdot b^1 + (d-1)\cdot b^2 + \ldots + b^d \in \mathcal{O}(b^{d+1})$
Space complexity	$\mathcal{O}(b\cdot d)$
Optimality	Yes (if step cost $= 1$)

▷ **Consequence:** IDS used in practice for search spaces of large, infinite, or unknown depth.

Note:

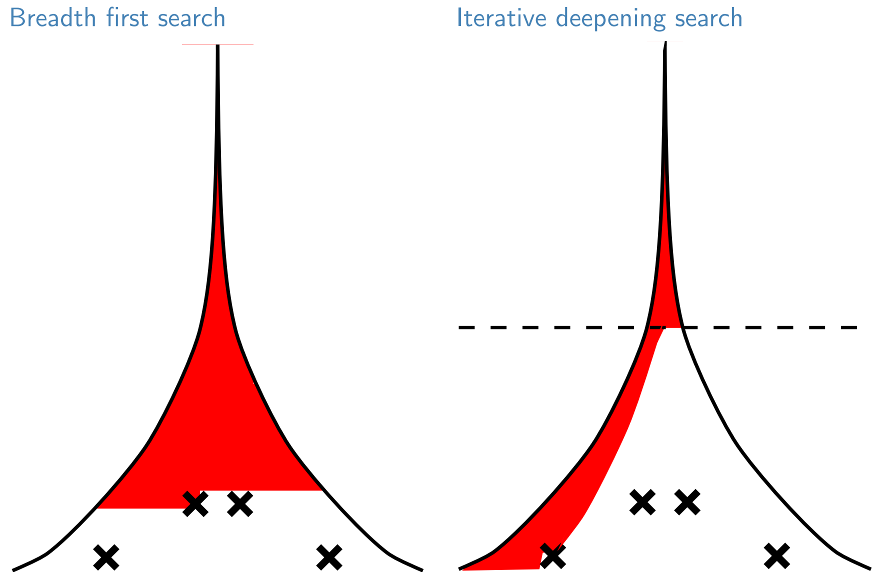
To find a solution (at depth d) we have to search the whole tree up to d. Of course since we do not save the search state, we have to re-compute the upper part of the tree for the next level. This seems like a great waste of resources at first, however, IDS tries to be complete without the space penalties.

However, the space complexity is as good as DFS, since we are using DFS along the way. Like in BFS, the whole tree on level d (of optimal solution) is explored, so optimality is inherited from there. Like BFS, one can modify this to incorporate uniform cost search behavior.

As a consequence, variants of IDS are the method of choice if we do not have additional information.

Comparison BFS (optimal) and IDS (not)

▷ **Example 6.4.10.** IDS may fail to be be optimal at step sizes > 1.



Michael Kohlhase: Artificial Intelligence 1 148 2023-09-20

6.4.3 Further Topics

Tree Search vs. Graph Search

- ▷ We have only covered tree search algorithms.
- ▷ States duplicated in nodes are a huge problem for efficiency.
- ▷ **Definition 6.4.11.** A graph search algorithm is a variant of a tree search algorithm that prunes nodes whose state has already been considered (duplicate pruning), essentially using a DAG data structure.
- ▷ **Observation 6.4.12.** *Tree search is memory intensive it has to store the fringe so keeping a list of "explored states" does not lose much.*
- ▷ Graph versions of all the tree search algorithms considered here exist, but are more difficult to understand (and to prove properties about).
- ▷ The (time complexity) properties are largely stable under duplicate pruning. (no gain in the worst case)
- ▷ **Definition 6.4.13.** We speak of a search algorithm, when we do not want to distinguish whether it is a tree or graph search algorithm. (difference considered in implementation detail)

Michael Kohlhase: Artificial Intelligence 1 149 2023-09-20

Uninformed Search Summary

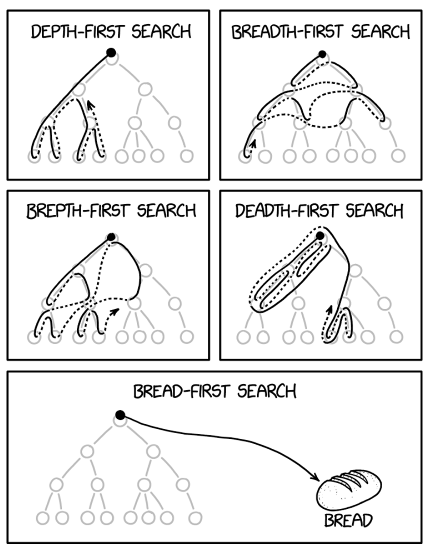
- ▷ **Tree/Graph Search Algorithms:** Systematically explore the state tree/graph induced by a search problem in search of a goal state. Search strategies only differ by the treatment of the fringe.

▷ **Search Strategies and their Properties:** We have discussed

Criterion	Breadth first	Uniform cost	Depth first	Iterative deepening
Completeness	Yes[1]	Yes[2]	No	Yes
Time complexity	b^d	$\approx b^d$	b^m	b^{d+1}
Space complexity	b^d	$\approx b^d$	bm	bd
Optimality	Yes*	Yes	No	Yes*
Conditions	[1] b finite	[2] $0<\epsilon\leq$cost		

Search Strategies; the XKCD Take

▷ **More Search Strategies?:** (from https://xkcd.com/2407/)



6.5 Informed Search Strategies

Summary: Uninformed Search/Informed Search

▷ Problem formulation usually requires abstracting away real-world details to define a state space that can feasibly be explored.

▷ Variety of uninformed search strategies.

▷ Iterative deepening search uses only linear space and not much more time than other uninformed algorithms.

- ▷ **Next Step:** Introduce additional knowledge about the problem(heuristic search)
  - ▷ Best-first-, A^*-strategies (guide the search by heuristics)
  - ▷ Iterative improvement algorithms.
- ▷ **Definition 6.5.1.** A search algorithm is called informed, iff it uses some form of external information – that is not part of the search problem – to guide the search.

6.5.1 Greedy Search

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/22015`.

Best-first search

- ▷ **Idea:** Order the fringe by estimated "desirability" (Expand most desirable unexpanded node)
- ▷ **Definition 6.5.2.** An evaluation function assigns a desirability value to each node of the search tree.
- ▷ **Note:** A evaluation function is not part of the search problem, but must be added externally.
- ▷ **Definition 6.5.3.** In best first search, the fringe is a queue sorted in decreasing order of desirability.
- ▷ **Special cases:** Greedy search, A^* search

This is like UCS, but with evaluation function related to problem at hand replacing the path cost function.

If the heuristics is arbitrary, we expect incompleteness!

Depends on how we measure "desirability".

Concrete examples follow.

Greedy search

- ▷ **Idea:** Expand the node that *appears* to be closest to the goal.
- ▷ **Definition 6.5.4.** A heuristic is an evaluation function h on states that estimates the cost from n to the nearest goal state.
- ▷ **Note:** All nodes for the same state must have the same h-value!
- ▷ **Definition 6.5.5.** Given a heuristic h, greedy search is the strategy where the fringe is organized as a queue sorted by decreasing h value.
- ▷ **Example 6.5.6.** Straight-line distance from/to Bucharest.

▷ **Note:** Unlike uniform cost search the node evaluation function has nothing to do with the nodes explored so far

internal search control $\rightsquigarrow$ external search control
partial solution cost $\rightsquigarrow$ goal cost estimation

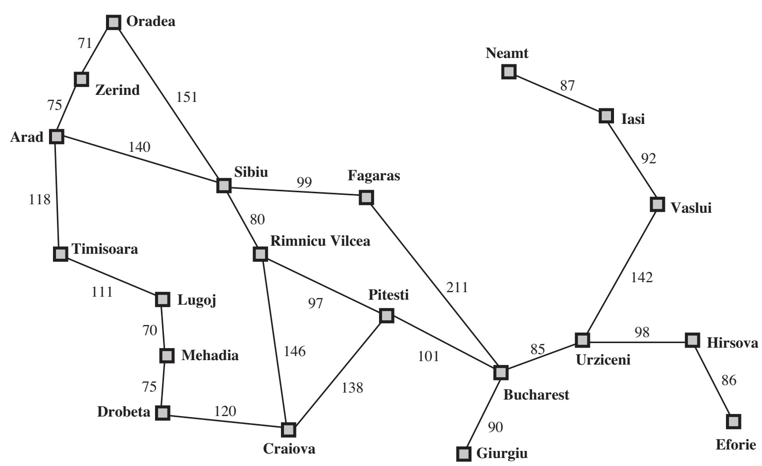
In greedy search we replace the *objective* cost to *construct* the current solution with a heuristic or *subjective* measure from which we think it gives a good idea how far we are from a *solution*. Two things have shifted:

- we went from internal (determined only by features inherent in the search space) to an external/heuristic cost
- instead of measuring the cost to build the current partial solution, we estimate how far we are from the desired goal

Romania with Straight-Line Distances

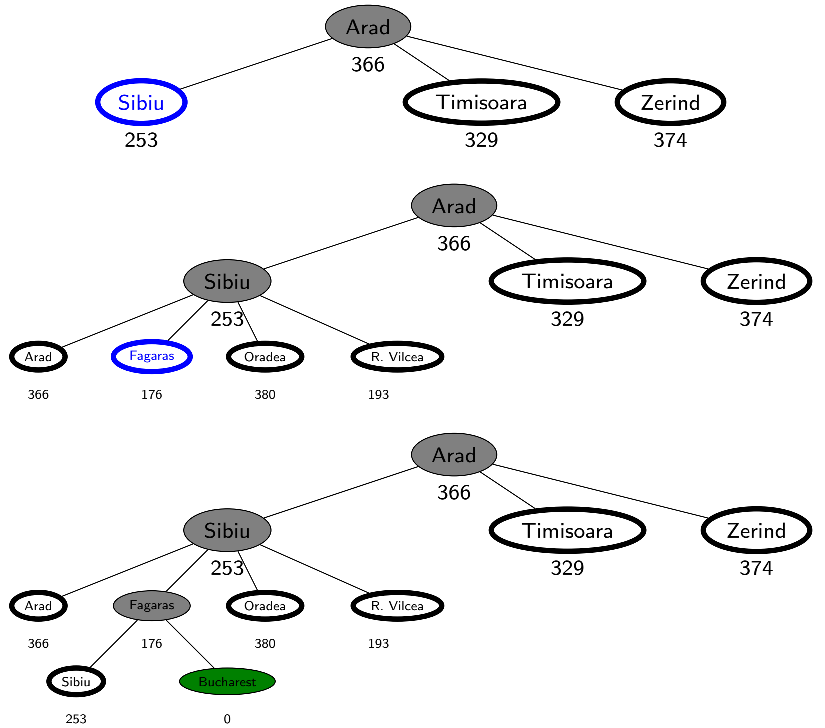
▷ **Example 6.5.7 (Informed Travel).** $h_{\mathsf{SLD}}(n)$ = straight-line distance to Bucharest

Arad	366	Mehadia	241	Bucharest	0	Neamt	234
Craiova	160	Oradea	380	Drobeta	242	Pitesti	100
Eforie	161	Rimnicu Vilcea	193	Fragaras	176	Sibiu	253
Giurgiu	77	Timisoara	329	Hirsova	151	Urziceni	80
Iasi	226	Vaslui	199	Lugoj	244	Zerind	374



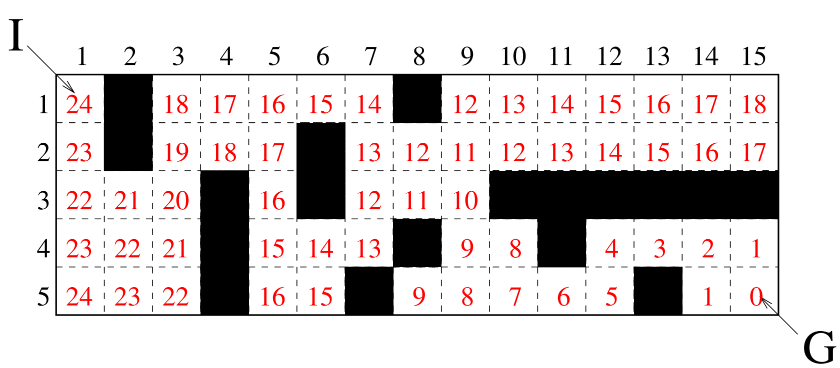
Greedy Search: Romania

Arad
366

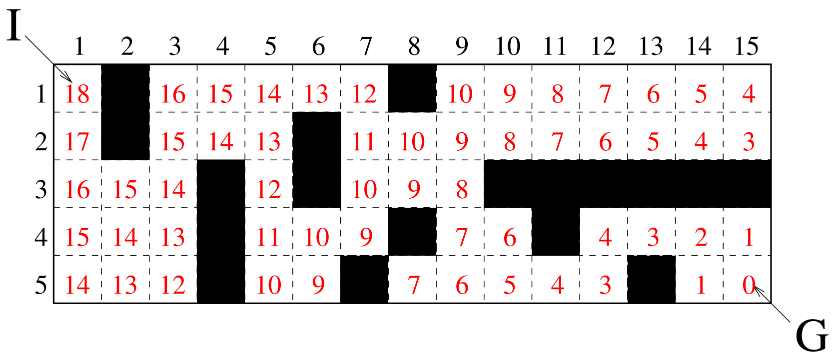


Heuristic Functions in Path Planning

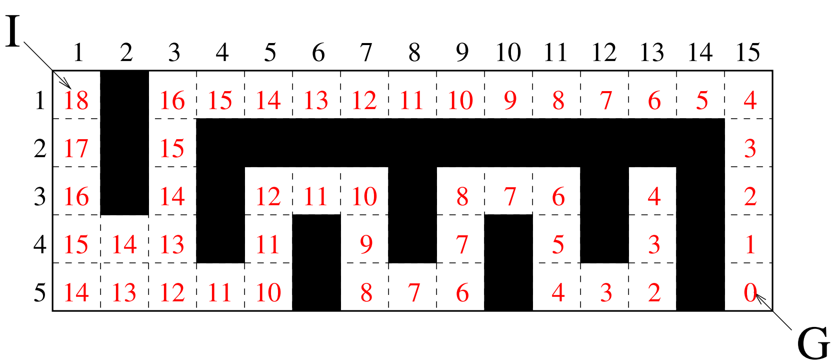
▷ **Example 6.5.8 (The maze solved).** We indicate h^* by giving the goal distance



▷ **Example 6.5.9 (Maze Heuristic: the good case).** We use the Manhattan distance to the goal as a heuristic



▷ **Example 6.5.10 (Maze Heuristic: the bad case).** We use the Manhattan distance to the goal as a heuristic again

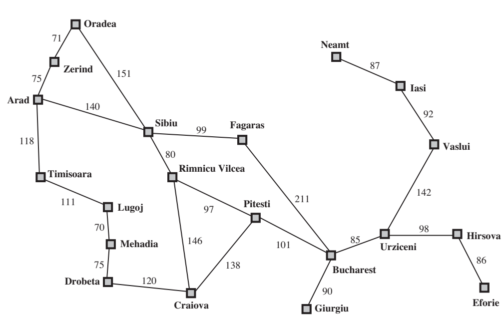


Greedy search: Properties

▷

Completeness	No: Can get stuck in loops Complete in finite space with repeated state checking
Time complexity	$\mathcal{O}(b^m)$
Space complexity	$\mathcal{O}(b^m)$
Optimality	No

▷ **Example 6.5.11.** Greedy search can get stuck going from Iasi to Oradea: Iasi → Neamt → Iasi → Neamt → ⋯



- **Worst-case Time:** Same as depth first search.
- **Worst-case Space:** Same as breadth first search.
- **But:** A good heuristic can give dramatic improvements.

Remark 6.5.12. Greedy Search is similar to UCS. Unlike the latter, the node evaluation function has nothing to do with the nodes explored so far. This can prevent nodes from being enumerated systematically as they are in UCS and BFS.

For completeness, we need repeated state checking as the example shows. This enforces complete enumeration of the state space (provided that it is finite), and thus gives us completeness.

Note that nothing prevents from *all* nodes being searched in worst case; e.g. if the heuristic function gives us the same (low) estimate on all nodes except where the heuristic mis-estimates the distance to be high. So in the worst case, greedy search is even worse than BFS, where d (depth of first solution) replaces m.

The search procedure cannot be optimal, since actual cost of solution is not considered.

For both, completeness and optimality, therefore, it is necessary to take the actual cost of partial solutions, i.e. the path cost, into account. This way, paths that are known to be expensive are avoided.

6.5.2 Heuristics and their Properties

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/22019`.

Heuristic Functions

- **Definition 6.5.13.** Let Π be a problem with states S. A heuristic function (or short heuristic) for Π is a function $h\colon S{\to}\mathbb{R}^+_0 \cup \{\infty\}$ so that $h(s) = 0$ whenever s is a goal state.
- $h(s)$ is intended as an estimate between state s and the nearest goal state.
- **Definition 6.5.14.** Let Π be a problem with states S, then the function $h^*\colon S{\to}\mathbb{R}^+_0 \cup \{\infty\}$, where $h^*(s)$ is the cost of a cheapest path from s to a goal state, or ∞ if no such path exists, is called the goal distance function for Π.
- **Notes:**
  - $h(s) = 0$ on goal states: If your estimator returns "I think it's still a long way" on a goal state, then its "intelligence" is, um ...
  - Return value ∞: To indicate dead ends, from which the goal can't be reached anymore.
  - The distance estimate depends only on the state s, not on the node (i.e., the path we took to reach s).

Where does the word "Heuristic" come from?

- Ancient Greek word $\epsilon\upsilon\rho\iota\sigma\kappa\epsilon\iota\nu$ ($\widehat{=}$ "I find") (aka. $\epsilon\upsilon\rho\epsilon\kappa\alpha$!)

- Popularized in modern science by George Polya: "How to solve it" [Pól73]
- same word often used for "rule of thumb" or "imprecise solution method".

Heuristic Functions: The Eternal Trade-Off

- "Distance Estimate"? (h is an arbitrary function in principle)
  - In practice, we want it to be *accurate* (aka: *informative*), i.e., close to the actual goal distance.
  - We also want it to be fast, i.e., a small overhead for computing h.
  - These two wishes are in contradiction!
- **Example 6.5.15 (Extreme cases).**
  - $h=0$: no overhead at all, completely un-informative.
  - $h=h^*$: perfectly accurate, overhead $\widehat{=}$ solving the problem in the first place.
- **Observation 6.5.16.** *We need to trade off the accuracy of h against the overhead for computing it.*

Properties of Heuristic Functions

- **Definition 6.5.17.** Let Π be a search problem with states S and actions A. We say that a heuristic h for Π is admissible if $h(s) \leq h^*(s)$ for all $s \in S$.

  We say that h is consistent if $h(s) - h(s') \leq c(a)$ for all $s \in S$ and $a \in A$.
- **In other words . . . :**
  - h is admissible if it is a lower bound on goal distance.
  - h is consistent if, when applying an action a, the heuristic value cannot decrease by more than the cost of a.

Properties of Heuristic Functions, ctd.

- Let Π be a problem, and let h be a heuristic for Π. If h is consistent, then h is admissible.
- *Proof:* we prove $h(s) \leq h^*(s)$ for all $s \in S$ by induction over the length of the cheapest path to a goal state.
  1. base case
     1.1. $h(s)=0$ by definition of heuristic, so $h(s) \leq h^*(s)$ as desired.
  2. step case

2.1. We assume that $h(s')\leq h^*(s)$ for all states s' with a cheapest goal path of length n.
2.2. Let s be a state whose cheapest goal path has length $n+1$ and the first transition is $o = (s,s')$.
2.3. By consistency, we have $h(s) - h(s')\leq c(o)$ and thus $h(s)\leq h(s') + c(o)$.
2.4. By construction, $h^*(s)$ has a cheapest goal path of length n and thus, by induction hypothesis $h(s')\leq h^*(s')$.
2.5. By construction, $h^*(s) = h^*(s') + c(o)$.
2.6. Together this gives us $h(s)\leq h^*(s)$ as desired.

▷ Consistency is a sufficient condition for admissibility (easier to check)

Properties of Heuristic Functions: Examples

▷ **Example 6.5.18.** Straight line distance is admissible and consistent by the triangle inequality.
If you drive 100km, then the straight line distance to Rome can't decrease by more than 100km.

▷ **Observation:** In practice, admissible heuristics are typically consistent.

▷ **Example 6.5.19 (An admissible, but inconsistent heuristic).** When traveling to Rome, let $h(Munich) = 300$ and $h(Innsbruck) = 100$.

▷ **Inadmissible heuristics:** typically arise as approximations of admissible heuristics that are too costly to compute. (see later)

6.5.3 A-Star Search

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/22020`.

A^* Search: Evaluation Function

▷ **Idea:** Avoid expanding paths that are already expensive(make use of actual cost)
The simplest way to combine heuristic and path cost is to simply add them.

▷ **Definition 6.5.20.** The evaluation function for A^* search is given by $f(n) = g(n) + h(n)$, where $g(n)$ is the path cost for n and $h(n)$ is the estimated cost to goal from n.

▷ Thus $f(n)$ is the estimated total cost of the path through n to a goal.

▷ **Definition 6.5.21.** Best first search with evaluation function $g + h$ is called A^* search.

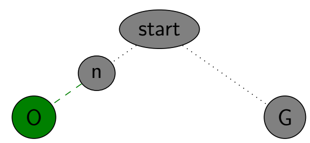
This works, provided that h does not overestimate the true cost to achieve the goal. In other words, h must be *optimistic* wrt. the real cost h^*. If we are too pessimistic, then non-optimal solutions have a chance.

A^* Search: Optimality

▷ **Theorem 6.5.22.** *A^* search with admissible heuristic is optimal.*

▷ *Proof:* We show that sub-optimal nodes are never selected by A^*

1. Suppose a suboptimal goal G has been generated then we are in the following situation:

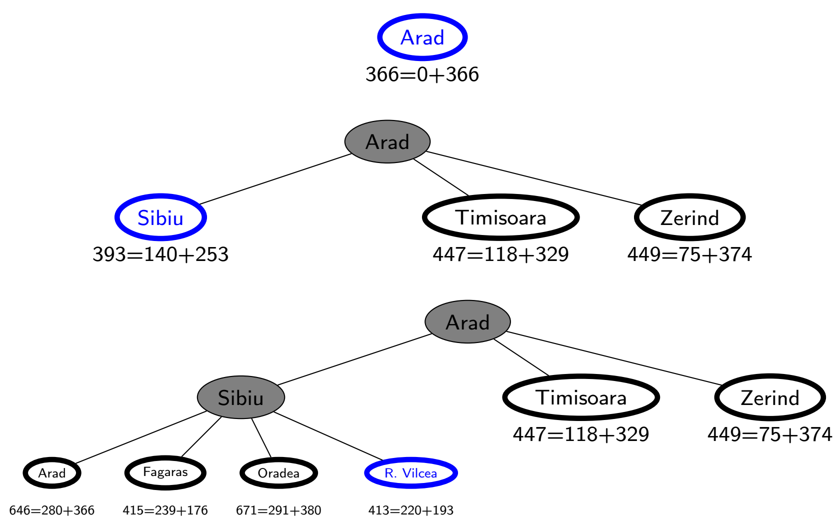


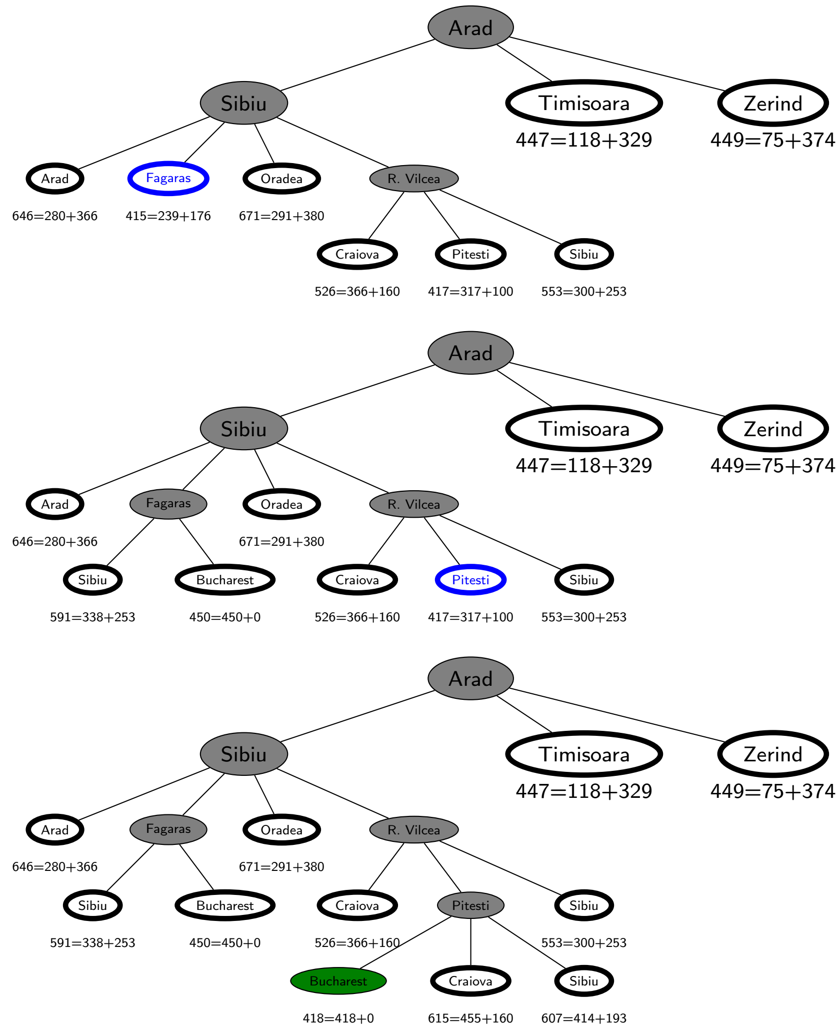
2. Let n be an unexpanded node on a path to an optimality goal O, then

$f(G) = g(G)$	since $h(G) = 0$
$g(G) > g(O)$	since G suboptimal
$g(O) = g(n) + h^*(n)$	n on optimal path
$g(n) + h^*(n) \geq g(n) + h(n)$	since h is admissible
$g(n) + h(n) = f(n)$	

3. Thus, $f(G) > f(n)$ and A^* never selects G for expansion.

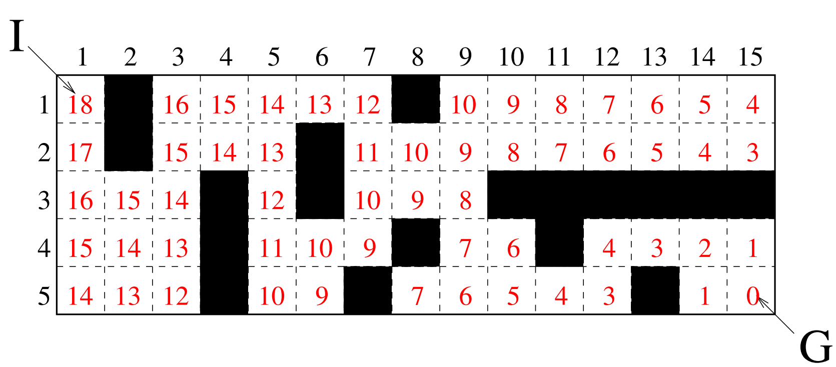
A^* Search Example





Additional Observations (Not Limited to Path Planning)

▷ **Example 6.5.23 (Greedy best-first search, "good case").**



We will find a solution with little search.

Additional Observations (Not Limited to Path Planning)

▷ **Example 6.5.24 (A^* ($g + h$), "good case").**

I

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	18		22	22	22	22	22		24	24	24	24	24	24	24
2	18		20	20	20		22	22	22	22	22	22	22	22	22
3	18	18	18		20		22	22	22						
4	18	18	18		20	20	20		22	22		24	24	24	24
5	18	18	18		20	20		24	22	22	22	22		24	24

G

- ▷ In A^* with a consistent heuristic, $g + h$ always increases monotonically (h cannot decrease more than g increases)
- ▷ We need more search, in the "right upper half". This is typical: Greedy best first search tends to be faster than A^*.

Additional Observations (Not Limited to Path Planning)

▷ **Example 6.5.25 (Greedy best-first search, "bad case").**

I

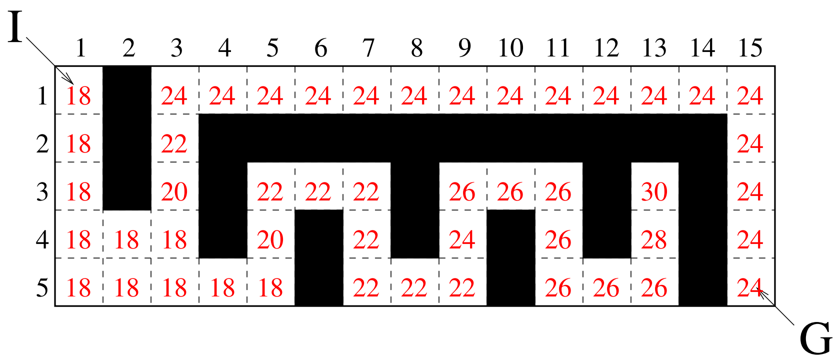
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	18		16	15	14	13	12	11	10	9	8	7	6	5	4
2	17		15												3
3	16		14		12	11	10		8	7	6		4		2
4	15	14	13		11		9		7		5		3		1
5	14	13	12	11	10		8	7	6		4	3	2		0

G

Search will be mis-guided into the "dead-end street".

Additional Observations (Not Limited to Path Planning)

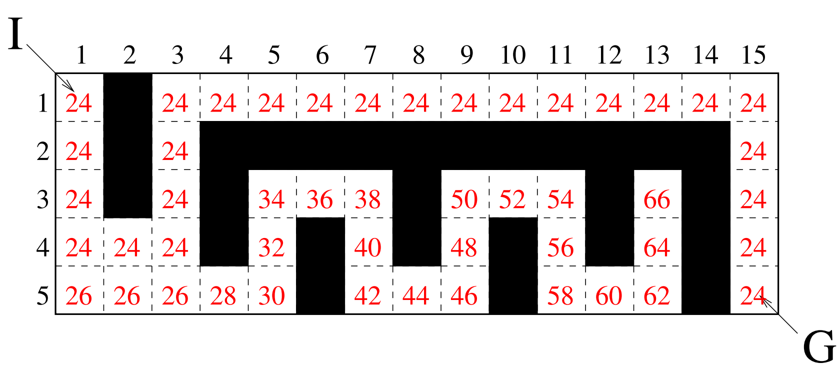
▷ **Example 6.5.26 (A^* ($g+h$), "bad case").**



We will search less of the "dead-end street". Sometimes $g+h$ gives better search guidance than h. ($\leadsto$ A^* is faster there)

Additional Observations (Not Limited to Path Planning)

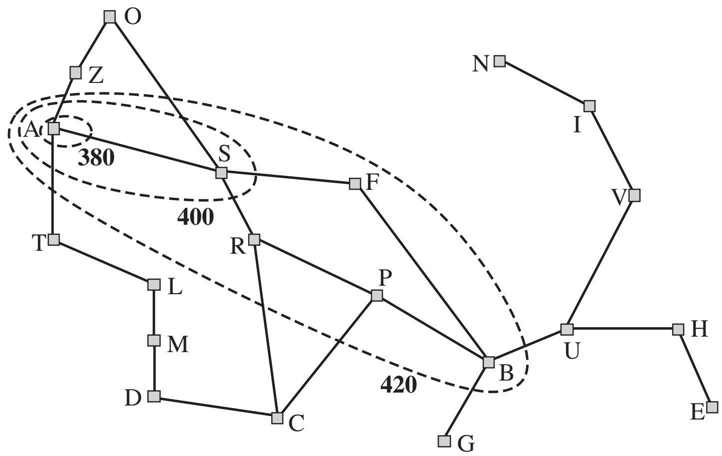
▷ **Example 6.5.27 (A^* ($g+h$) using h^*).**



In A^*, node values always increase monotonically (with any heuristic). If the heuristic is perfect, they remain constant on optimal paths.

A^* search: f-contours

▷ A^* gradually adds "f-contours" of nodes



A^* search: Properties

▷ Properties or A^*

Completeness	Yes (unless there are infinitely many nodes n with $f(n) \leq f(0)$)
Time complexity	Exponential in [relative error in h × length of solution]
Space complexity	Same as time (variant of BFS)
Optimality	Yes

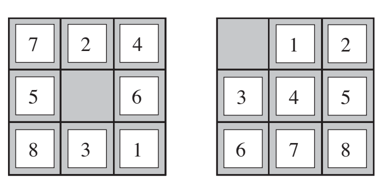
▷ A^* expands all (some/no) nodes with $f(n) < h^*(n)$

▷ The run-time depends on how good we approximated the real cost h^* with h.

6.5.4 Finding Good Heuristics

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/22021`.
Since the availability of admissible heuristics is so important for informed search (particularly for A^*), let us see how such heuristics can be obtained in practice. We will look at an example, and then derive a general procedure from that.

Admissible heuristics: Example 8-puzzle



Start State Goal State

▷ **Example 6.5.28.** Let $h_1(n)$ be the number of misplaced tiles in node n ($h_1(S) = 6$)

▷ **Example 6.5.29.** Let $h_2(n)$ be the total Manhattan distance from desired location of each tile. ($h_2(S) = 2+0+3+1+0+1+3+4 = 14$)

▷ **Observation 6.5.30 (Typical search costs).** *(IDS $\widehat{=}$ iterative deepening search)*

nodes explored	IDS	$A^*(h_1)$	$A^*(h_2)$
$d = 14$	3,473,941	539	113
$d = 24$	too many	39,135	1,641

Dominance

▷ **Definition 6.5.31.** Let h_1 and h_2 be two admissible heuristics we say that h_2 dominates h_1 if $h_2(n) \geq h_1(n)$ for all n.

▷ **Theorem 6.5.32.** *If h_2 dominates h_1, then h_2 is better for search than h_1.*

Relaxed problems

▷ **Observation:** Finding good admissible heuristics is an art!

▷ **Idea:** Admissible heuristics can be derived from the *exact* solution cost of a relaxed version of the problem.

▷ **Example 6.5.33.** If the rules of the 8-puzzle are relaxed so that a tile can move *anywhere*, then we get heuristic h_1.

▷ **Example 6.5.34.** If the rules are relaxed so that a tile can move to *any adjacent square*, then we get heuristic h_2. (Manhattan distance)

▷ **Definition 6.5.35.** Let $\Pi := \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{I}, \mathcal{G} \rangle$ be a search problem, then we call a search problem $\mathcal{P}^r := \langle \mathcal{S}, \mathcal{A}^r, \mathcal{T}^r, \mathcal{I}^r, \mathcal{G}^r \rangle$ a relaxed problem (wrt. Π; or simply relaxation of Π), iff $\mathcal{A} \subseteq \mathcal{A}^r$, $\mathcal{T} \subseteq \mathcal{T}^r$, $\mathcal{I} \subseteq \mathcal{I}^r$, and $\mathcal{G} \subseteq \mathcal{G}^r$.

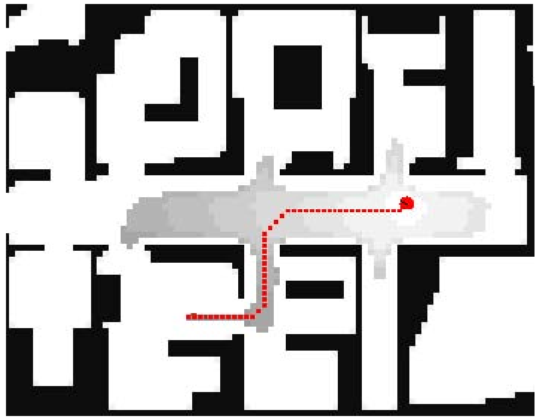
▷ **Lemma 6.5.36.** *If Pr relaxes Π, then every solution for Π is one for Pr.*

▷ **Key point:** The optimal solution cost of a relaxed problem is not greater than the optimal solution cost of the real problem.

Relaxation means to remove some of the constraints or requirements of the original problem, so that a solution becomes easy to find. Then the cost of this easy solution can be used as an optimistic approximation of the problem.

Empirical Performance: A^* in Path Planning

▷ **Example 6.5.37 (Live Demo vs. Breadth-First Search).**



See http://qiao.github.io/PathFinding.js/visual/

▷ **Difference to Breadth-first Search?:** That would explore all grid cells in a *circle* around the initial state!

6.6 Local Search

Video Nuggets covering this section can be found at https://fau.tv/clip/id/22050 and https://fau.tv/clip/id/22051.

Systematic Search vs. Local Search

▷ **Definition 6.6.1.** We call a search algorithm systematic, if it considers all states at some point.

▷ **Example 6.6.2.**

All tree search algorithms (except pure depth first search) are systematic. (given reasonable assumptions e.g. about costs.)

▷ **Observation 6.6.3.** *Systematic search algorithms are complete.*

▷ **Observation 6.6.4.** *In systematic search algorithms there is no limit of the number of nodes that are kept in memory at any time.*

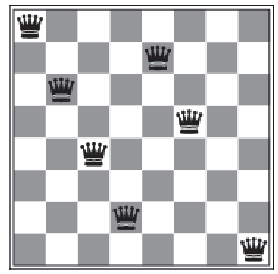
▷ **Alternative:** Keep only one (or a few) nodes at a time

- ▷ ⇝ no systematic exploration of all options, ⇝ incomplete.

Local Search Problems

▷ **Idea:** Sometimes the path to the solution is irrelevant.

▷ **Example 6.6.5 (8 Queens Problem).** Place 8 queens on a chess board, so that no two queens threaten each other.



▷ This problem has various solutions (the one of the right isn't one of them)

▷ **Definition 6.6.6.** A local search algorithm is a search algorithm that operates on a single state, the current state (rather than multiple paths). (advantage: constant space)

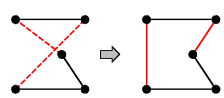
▷ Typically local search algorithms only move to successor of the current state, and do not retain search paths.

▷ Applications include: integrated circuit design, factory-floor layout, job-shop scheduling, portfolio management, fleet deployment,...

Local Search: Iterative improvement algorithms

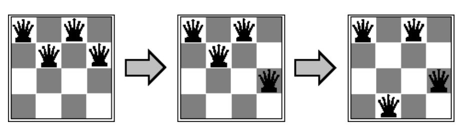
▷ **Definition 6.6.7 (Traveling Salesman Problem).** Find shortest trip through set of cities such that each city is visited exactly once.

▷ **Idea:** Start with any complete tour, perform pairwise exchanges



▷ **Definition 6.6.8 (n-queens problem).** Put n queens on $n \times n$ board such that no two queens in the same row, columns, or diagonal.

▷ **Idea:** Move a queen to reduce number of conflicts



Hill-climbing (gradient ascent/descent)

▷ **Idea:** Start anywhere and go in the direction of the steepest ascent.

▷ **Definition 6.6.9.** Hill climbing (also gradient ascent)is a local search algorithm that iteratively selects the best successor:

```
procedure Hill−Climbing (problem) /* a state that is a local minimum */
  local current, neighbor /* nodes */
  current := Make−Node(Initial−State[problem])
```

```
  loop
    neighbor := <a highest−valued successor of current>
    if Value[neighbor] < Value[current] return [current] end if
    current := neighbor
  end loop
end procedure
```

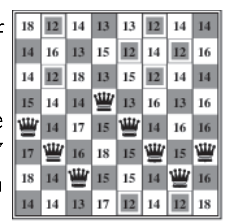
- ▷ **Intuition:** Like best first search without memory.
- ▷ Works, if solutions are dense and local maxima can be escaped.

In order to understand the procedure on a more intuitive level, let us consider the following scenario: We are in a dark landscape (or we are blind), and we want to find the highest hill. The search procedure above tells us to start our search anywhere, and for every step first feel around, and then take a step into the direction with the steepest ascent. If we reach a place, where the next step would take us down, we are finished.

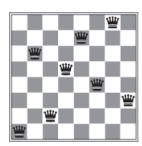
Of course, this will only get us into local maxima, and has no guarantee of getting us into global ones (remember, we are blind). The solution to this problem is to re-start the search at random (we do not have any information) places, and hope that one of the random jumps will get us to a slope that leads to a global maximum.

Example Hill-Climbing with 8 Queens

- ▷ **Idea:** Consider $h \mathrel{\widehat{=}}$ number of queens that threaten each other.
- ▷ **Example 6.6.10.** An 8-queens state with heuristic cost estimate $h = 17$ showing h-values for moving a queen within its column

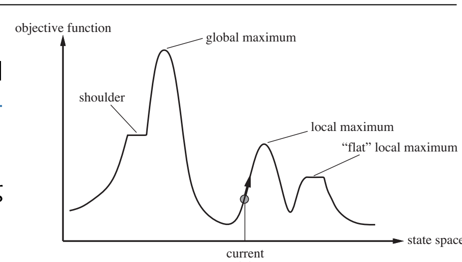


- ▷ **Problem:** The state space has local minima. e.g. the board on the right has $h = 1$ but every successor has $h > 1$.



Hill-climbing

- ▷ **Problem:** Depending on initial state, can get stuck on local maxima/minima and plateaux.
- ▷ "Hill-climbing search is like climbing Everest in thick fog with amnesia".



- ▷ **Idea:** Escape local maxima by allowing some "bad" or random moves.
- ▷ **Example 6.6.11.** local search, simulated annealing...

▷ **Properties:** All are incomplete, nonoptimal.

▷ Sometimes performs well in practice (if (optimal) solutions are dense)

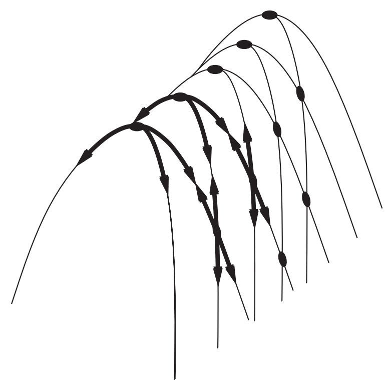
Recent work on hill climbing algorithms tries to combine complete search with randomization to escape certain odd phenomena occurring in statistical distribution of solutions.

Simulated annealing (Idea)

▷ **Definition 6.6.12.** Ridges are ascending successions of local maxima.

▷ **Problem:** They are extremely difficult to bv navigate for local search algorithms.

▷ **Idea:** Escape local maxima by allowing some "bad" moves, but gradually decrease their size and frequency.



▷ Annealing is the process of heating steel and let it cool gradually to give it time to grow an optimal cristal structure.

▷ Simulated annealing is like shaking a ping pong ball occasionally on a bumpy surface to free it. (so it does not get stuck)

▷ Devised by Metropolis et al for physical process modelling [Met+53]

▷ Widely used in VLSI layout, airline scheduling, etc.

Simulated annealing (Implementation)

▷ **Definition 6.6.13.** The following algorithm is called simulated annealing:

```
procedure Simulated−Annealing (problem,schedule) /∗ a solution state ∗/
  local node, next /∗ nodes ∗/
   local T /∗ a ‘‘temperature’’ controlling prob.~of downward steps ∗/
   current := Make−Node(Initial−State[problem])
   for t :=1 to ∞
    T := schedule[t]
      if T = 0 return current end if
      next := <a randomly selected successor of current>
     Δ(E) := Value[next]−Value[current]
      if Δ(E) > 0 current := next
      else
       current := next <only with probability> e^(Δ(E)/T)
     end if
  end for
end procedure
```

A schedule is a mapping from time to "temperature".

Properties of simulated annealing

- At fixed "temperature" T, state occupation probability reaches Boltzman distribution

$$p(x) = \alpha e^{\frac{E(x)}{kT}}$$

T decreased slowly enough $\leadsto$ always reach best state x^* because

$$\frac{e^{\frac{E(x^*)}{kT}}}{e^{\frac{E(x)}{kT}}} = e^{\frac{E(x^*)-E(x)}{kT}} \gg 1$$

for small T.

- **Question:** Is this necessarily an interesting guarantee?

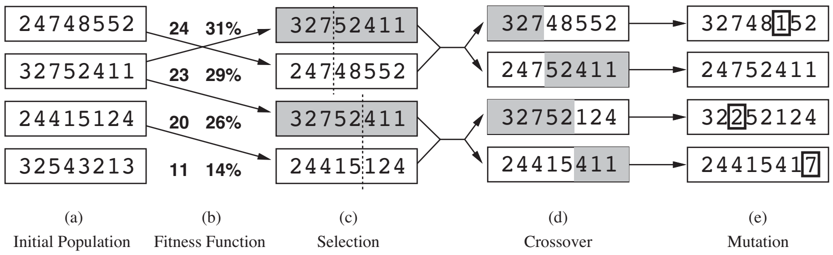
Local beam search

- **Definition 6.6.14.** Local beam search is a search algorithm that keep k states instead of 1 and chooses the top k of all their successors.
- **Observation:** Local beam search is Not the same as k searches run in parallel! (Searches that find good states recruit other searches to join them)
- **Problem:** Quite often, all k searches end up on same local hill!
- **Idea:** Choose k successors randomly, biased towards good ones. (Observe the close analogy to natural selection!)

Genetic algorithms (very briefly)

- **Definition 6.6.15.** A genetic algorithm is a variant of local beam search that generates successors by
  - randomly modifying states (mutation)
  - mixing pairs of states (sexual reproduction or crossover)

  to optimize a fitness function. (survival of the fittest)
- **Example 6.6.16.** Generating successors for 8 Queens

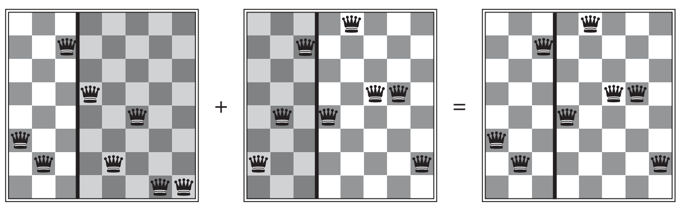


Genetic algorithms (continued)

▷ **Problem:** Genetic algorithms require states encoded as strings.

▷ Crossover only helps iff substrings are meaningful components.

▷ **Example 6.6.17 (Evolving 8 Queens).** First crossover



▷ **Note:** Genetic algorithms $\neq$ evolution: e.g., real genes also encode replication machinery!

Chapter 7

Adversarial Search for Game Playing

A Video Nugget covering this chapter can be found at https://fau.tv/clip/id/22079.

7.1 Introduction

Video Nuggets covering this section can be found at https://fau.tv/clip/id/22060 and https://fau.tv/clip/id/22061.

The Problem (cf. chapter 6)

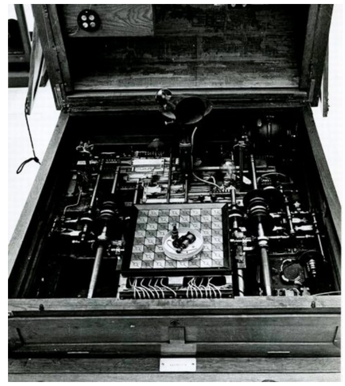
"Adversarial search" = Game playing against an opponent.

Why Game Playing?

- ▷ **What do you think?**
  - ▷ Playing a game well clearly requires a form of "intelligence".
  - ▷ Games capture a pure form of competition between opponents.
  - ▷ Games are abstract and precisely defined, thus very easy to formalize.

- ▷ Game playing is one of the oldest sub-areas of AI (ca. 1950).
- ▷ The dream of a machine that plays chess is, indeed, *much* older than AI!

"Schachtürke" (1769)



"El Ajedrecista" (1912)

"Game" Playing? *Which* Games?

- ▷ . . . sorry, we're not gonna do soccer here.
- ▷ **Restrictions:**
  - ▷ Game states discrete, number of game states finite.
  - ▷ Finite number of possible moves.
  - ▷ The game state is fully observable.
  - ▷ The outcome of each move is deterministic.
  - ▷ Two players: Max and Min.
  - ▷ Turn-taking: It's each player's turn alternatingly. Max begins.
  - ▷ Terminal game states have a utiliy u. Max tries to maximize u, Min tries to minimize u.
  - ▷ In that sense, the utility for Min is the exact opposite of the utility for Max ("zero sum").
  - ▷ There are no infinite runs of the game (no matter what moves are chosen, a terminal state is reached after a finite number of steps).

An Example Game

▷ Game states: Positions of figures.

▷ Moves: Given by rules.

▷ Players: White (Max), Black (Min).

▷ Terminal states: Checkmate.

▷ Utility of terminal states, e.g.:

- ▷ $+100$ if Black is checkmated.
- ▷ 0 if stalemate.
- ▷ -100 if White is checkmated.

"Game" Playing? Which Games *Not*?

▷ Soccer (sorry guys; not even RoboCup)

▷ Important types of games that we don't tackle here:

- ▷ Chance. (E.g., Backgammon)
- ▷ More than two players. (E.g., Halma)
- ▷ Hidden information. (E.g., most card games)
- ▷ Simultaneous moves. (E.g., Diplomacy)
- ▷ Not zero-sum, i.e., outcomes may be beneficial (or detrimental) for both players. (cf. Game theory: Auctions, elections, economy, politics, ...)

▷ Many of these more general game types can be handled by similar/extended algorithms.

(A Brief Note On) Formalization

▷ **Definition 7.1.1 (Game State Space).** A game state space is a 6 tuple $\Theta:=\langle S,A,T,I,S^T,u\rangle$ where:

- ▷ states S, actions A, deterministic transition relation T, initial state I are as in classical search problems, except:
  - ▷ S is the disjoint union of S^{Max}, S^{Min}, and S^T.
  - ▷ A is the disjoint union of A^{Max} and A^{Min}.
  - ▷ For $a\in A^{\mathrm{Max}}$, if $s\xrightarrow{a}s'$ then $s\in S^{\mathrm{Max}}$ and $s'\in(S^{\mathrm{Min}}\cup S^T)$.
  - ▷ For $a\in A^{\mathrm{Min}}$, if $s\xrightarrow{a}s'$ then $s\in S^{\mathrm{Min}}$ and $s'\in(S^{\mathrm{Max}}\cup S^T)$.
- ▷ S^T is the set of terminal states.
- ▷ $u\colon S^T\to\mathbb{R}$ is the utility function.

- **Definition 7.1.2 (Commonly used terminology).** position $\widehat{=}$ state, end state $\widehat{=}$ terminal state, move $\widehat{=}$ action.
- A round of the game – one move Max, one move Min – is often referred to as a "move", and individual actions as "half-moves". We *don't* do that here.

Why Games are Hard to Solve: I

- What is a "solution" here?
- **Definition 7.1.3.** Let Θ be a game state space, and let $X \in \{\text{Max}, \text{Min}\}$. A strategy for X is a function $\sigma^X : S^X \to A^X$ so that a is applicable to s whenever $\sigma^X(s) = a$.
- We don't know how the opponent will react, and need to prepare for all possibilities.
- **Definition 7.1.4.** A strategy is called optimal if it yields the best possible utility for X assuming perfect opponent play (not formalized here).
- In (almost) all games, computing a strategy is infeasible. Instead, compute the next move "on demand", given the current game state.

Why Games are hard to solve II

- **Example 7.1.5.** Number of reachable states in chess: 10^{40}.
- **Example 7.1.6.** Number of reachable states in go: 10^{100}.
- **It's even worse:** Our algorithms here look at search trees (game trees), no duplicate checking.
  - Chess: $35^{100} \simeq 10^{154}$.
  - Go: $200^{300} \simeq 10^{690}$.

How To Describe a Game State Space?

- Like for classical search problems, there are three possible ways to describe a game: blackbox/API description, declarative description, explicit game state space.
- **Question:** Which ones do humans use?
  - Explicit $\approx$ Hand over a book with all 10^{40} moves in Chess.
  - Blackbox $\approx$ Give possible Chess moves on demand but don't say how they are generated.

▷ **Answer:** Declarative!
With "game description language" $\widehat{=}$ natural language.

Specialized vs. General Game Playing

▷ And which game descriptions do computers use?

- ▷ Explicit: Only in illustrations.
- ▷ Blackbox/API: Assumed description in (This Chapter)
  - ▷ Method of choice for all those game players out there in the market (Chess computers, video game opponents, you name it).
  - ▷ Programs designed for, and specialized to, a particular game.
  - ▷ Human knowledge is key: evaluation functions (see later), opening databases (Chess!!), end game databases.
- ▷ Declarative: General Game Playing, active area of research in AI.
  - ▷ Generic Game Description Language (GDL), based on logic.
  - ▷ Solvers are given only "the rules of the game", no other knowledge/input whatsoever (cf. chapter 6).
  - ▷ Regular academic competitions since 2005.

Our Agenda for This Chapter

▷ **Minimax Search**: How to compute an optimal strategy?

- ▷ Minimax is the canonical (and easiest to understand) algorithm for *solving* games, i.e., computing an optimal strategy.

▷ **Evaluation Functions**: But what if we don't have the time/memory to solve the entire game?

- ▷ Given limited time, the best we can do is look ahead as far as we can. Evaluation functions tell us how to evaluate the leaf states at the cut off.

▷ **Alpha-Beta Search**: How to prune unnecessary parts of the tree?

- ▷ Often, we can detect early on that a particular action choice cannot be part of the optimal strategy. We can then stop considering this part of the game tree.

▷ State of the art: What is the state of affairs, for prominent games, of computer game playing vs. human experts?

- ▷ Just FYI (not part of the technical content of this course).

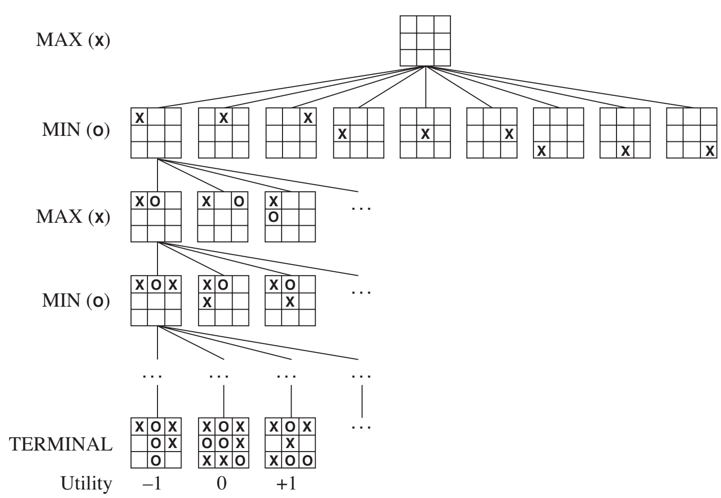
7.2 Minimax Search

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22061`.

"Minimax"?

- ▷ We want to compute an optimal strategy for player "Max".
  - ▷ In other words: *We are Max, and our opponent is Min.*
- ▷ **Recall:**
  - ▷ We compute the strategy offline, before the game begins. During the game, whenever it's our turn, we just lookup the corresponding action.
  - ▷ Max attempts to maximize the utility $u(s)$ of the terminal state that will be reached during play.
  - ▷ Min attempts to minimize $u(s)$.
- ▷ So what?
  - ▷ The computation alternates between minimization and maximization ⇝ hence "minimax".

Example Tic-Tac-Toe



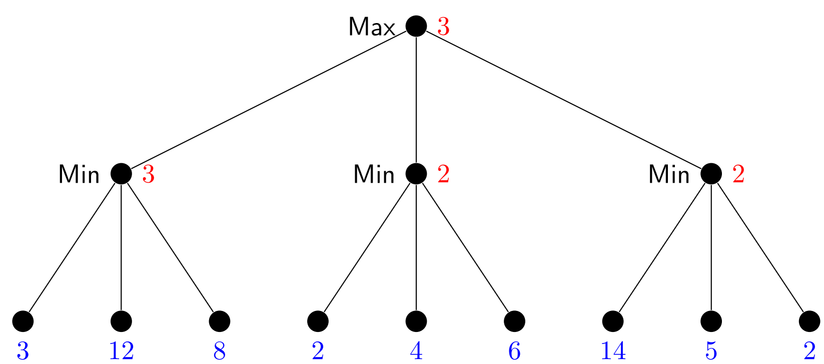
- ▷ Game tree, current player marked on the left.
- ▷ Last row: terminal positions with their utility.

Minimax: Outline

▷ **We max, we min, we max, we min ...**

1. Depth first search in game tree, with Max in the root.
2. Apply utility function to terminal positions.
3. Bottom-up for each inner node n in the tree, compute the utility $u(n)$ of n as follows:
   - ▷ If it's Max's turn: Set $u(n)$ to the maximum of the utilities of n's successor nodes.
   - ▷ If it's Min's turn: Set $u(n)$ to the minimum of the utilities of n's successor nodes.
4. Selecting a move for Max at the root: Choose one move that leads to a successor node with maximal utility.

Minimax: Example



▷ **Blue numbers:** Utility function u applied to terminal positions.

▷ **Red numbers:** Utilities of inner nodes, as computed by the minimax algorithm.

The Minimax Algorithm: Pseudo-Code

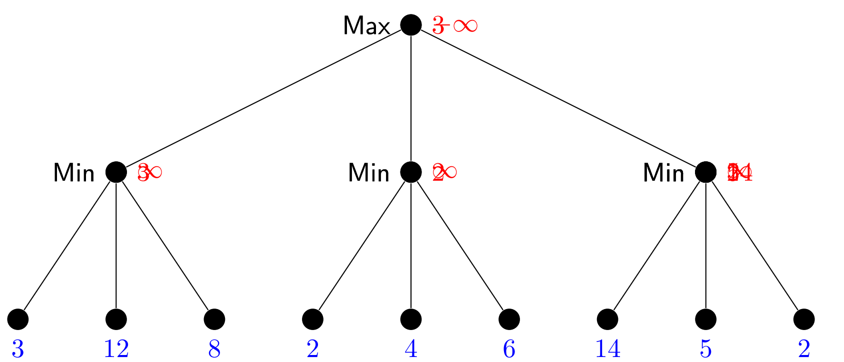
▷ **Definition 7.2.1.** The minimax algorithm (often just called minimax) is given by the following function whose input is a state $s \in S^{\text{Max}}$, in which Max is to move.

function Minimax—Decision(s) **returns** an action
  v := Max—**Value**(s)
  return an action yielding value v **in** the previous **function** call

function Max—**Value**(s) **returns** a utility value
  if Terminal—Test(s) **then return** $u(s)$
  $v := -\infty$
  for each $a \in$ Actions(s) **do**
    v := max(v,Min—**Value**(ChildState(s,a)))
  return v

```
function Min—Value(s) returns a utility value
  if Terminal—Test(s) then return u(s)
  v := +∞
  for each a ∈ Actions(s) do
    v := min(v,Max—Value(ChildState(s,a)))
  return v
```


Minimax: Example, Now in Detail



- ▷ So which action for Max is returned?
- ▷ Leftmost branch.
- ▷ Note: The maximal possible pay-off is higher for the rightmost branch, but assuming perfect play of Min, it's better to go left. (Going right would be "relying on your opponent to do something stupid".)

Minimax, Pro and Contra

- ▷ **Minimax advantages:**
  - ▷ Minimax is the simplest possible (reasonable) search algorithm for games. (If any of you sat down, prior to this lecture, to implement a Tic-Tac-Toe player, chances are you either looked this up on Wikipedia, or invented it in the process.)
  - ▷ Returns an optimal action, assuming perfect opponent play.
    - ▷ No matter how the opponent plays, the utility of the terminal state reached will be at least the value computed for the root.
    - ▷ If the opponent plays perfectly, exactly that value will be reached.
  - ▷ There's no need to re-run minimax for every game state: Run it once, offline before the game starts. During the actual game, just follow the branches taken in the tree. Whenever it's your turn, choose an action maximizing the value of the successor states.

▷ **Minimax disadvantages:** It's completely infeasible in practice.

  ▷ When the search tree is too large, we need to limit the search depth and apply an evaluation function to the cut off states.

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 208 2023-09-20

7.3 Evaluation Functions

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22064`.

Evaluation Functions for Minimax

▷ **Problem:** Game tree too big so search through in minimax.

▷ **Solution:** We impose a search depth limit (also called horizon) d, and apply an evaluation function to the non-terminal cut off states, i.e. states s with $\mathrm{dp}(s){>}d$.

▷ **Definition 7.3.1.** An evaluation function f maps game states to numbers:

  ▷ $f(s)$ is an estimate of the actual value of s (as would be computed by unlimited-depth Minimax for s).

  ▷ If cut off state is terminal: Just use u instead of f.

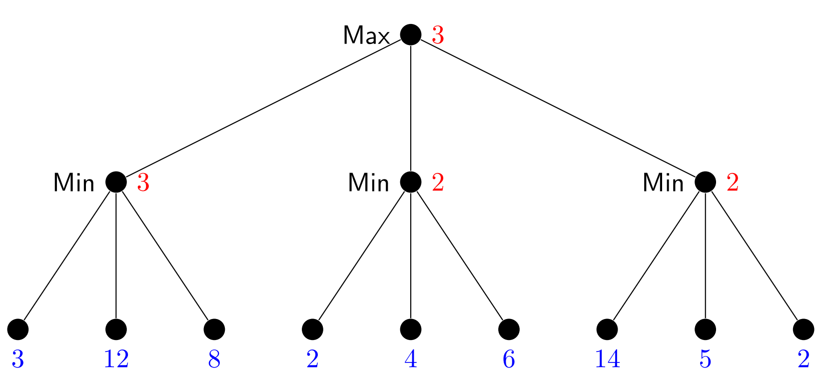
▷ Analogy to heuristic functions (cf. section 6.5): We want f to be both (a) accurate and (b) fast.

▷ Another analogy: (a) and (b) are in contradiction $\leadsto$ need to trade-off accuracy against overhead.

  ▷ In typical game playing algorithms today, f is inaccurate but very fast. (Usually no good methods known for computing accurate f)

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 209 2023-09-20

Our Example, Revisited: Minimax With Depth Limit $d = 2$



▷ Blue numbers: evaluation function f, applied to the *cut-off states at* $d = 2$.

▷ Red numbers: utilities of inner nodes, as computed by minimax using d, f.

Example Chess

- ▷ Evaluation function in Chess:
  - ▷ **Material**: Pawn 1, Knight 3, Bishop 3, Rook 5, Queen 9.
  - ▷ 3 points advantage $\rightsquigarrow$ safe win.
  - ▷ **Mobility**: How many fields do you control?
  - ▷ King safety, Pawn structure, ...
- ▷ Note how simple this is! (probably is not how Kasparov evaluates his positions)

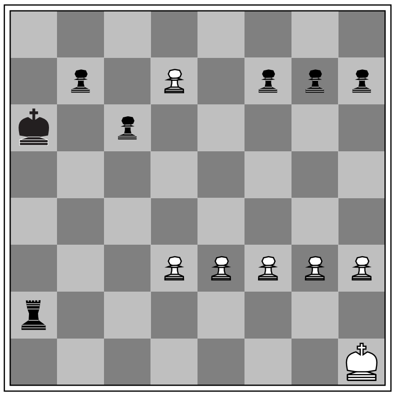
Linear Evaluation Functions

- ▷ **Paragraph:** How to come up with evaluation functions?
- ▷ **Definition 7.3.2.** A common approach is to use a weighted linear function for f, i.e. given a set of features $f_i\colon S{\rightarrow}\mathbb{R}$ and a corresponding sequence of weights $w_i{\in}\mathbb{R}$, f is of the form $f(s){:=}w_1 \cdot f_1(s) + w_2 \cdot f_2(s) + \cdots + w_n \cdot f_n(s)$
- ▷ **Problem:** How to obtain these weighted linear functions?
  - ▷ Weights w_i can be learned automatically.
  - ▷ The features f_i, however, have to be designed by human experts.
- ▷ **Note:**
  - ▷ Very fast, very simplistic.
  - ▷ Can be computed incrementally: In transition $s \xrightarrow{a} s'$, adapt $f(s)$ to $f(s')$ by considering only those features whose values have changed.

This assumes that the features (their contribution towards the actual value of the state) are independent. That's usually not the case (e.g. the value of a Rook depends on the Pawn structure).

The Horizon Problem

- ▷ **Problem:** Critical aspects of the game can be cut-off by the horizon.



Black to move

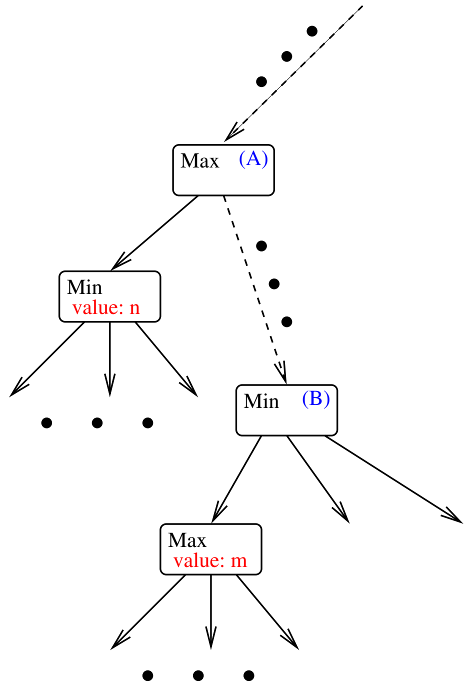
- ▷ Who's gonna win here?
  - ▷ White wins (Pawn cannot be prevented from becoming a queen.)
  - ▷ Black has a $+4$ advantage in material, so if we cut-off here then our evaluation function will say "100, black wins".
  - ▷ The loss for black is "beyond our horizon" unless we search extremely deeply: Black can hold off the end by repeatedly giving check to White's king.

So, How Deeply to Search?

- ▷ **Goal:** In given time, search as deeply as possible.
- ▷ **Problem:** Very difficult to predict search running time. (need an anytime algorithm)
- ▷ **Solution:** iterative deepening search.
  - ▷ Search with depth limit $d = 1, 2, 3, \ldots$
  - ▷ Time's up: Return result of deepest completed search.
- ▷ **Definition 7.3.3 (Better Solution).** The quiescent search algorithm uses a dynamically adapted search depth d: It searches more deeply in unquiet positions, where value of evaluation function changes a lot in neighboring states.
- ▷ **Example 7.3.4.** In quiescent search for chess:
  - ▷ piece exchange situations ("you take mine, I take yours") are very unquiet
  - ▷ $\rightsquigarrow$ Keep searching until the end of the piece exchange is reached.

7.4 Alpha-Beta Search

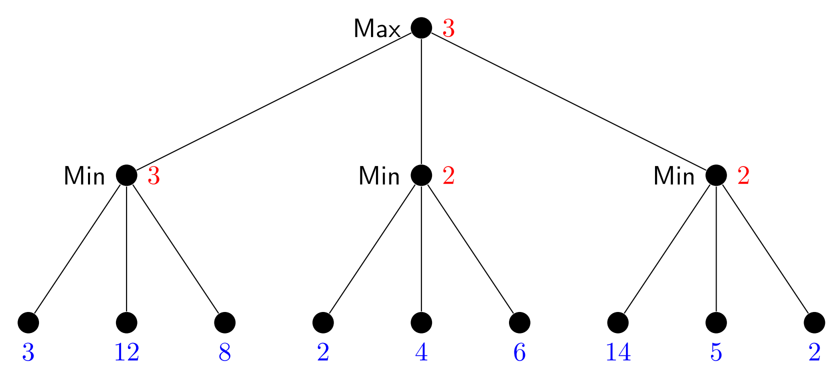
When We Already Know We Can Do Better Than This



- Say $n > m$.
- By choosing to go to the left in search node (A), Max already can get utility at least n in this part of the game.
- So, if "later on" (further down in the same sub-tree), in search node (B) we already know that Min can force Max to get value $m < n$.
- Then Max will play differently in (A) so we will never actually get to (B).

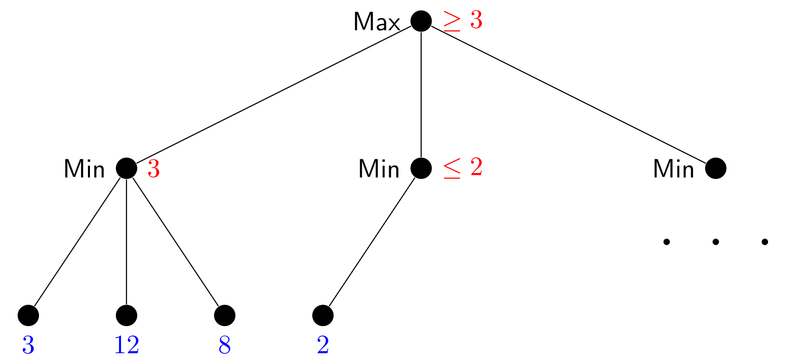
Alpha Pruning: Basic Idea

- **Question:** Can we save some work here?



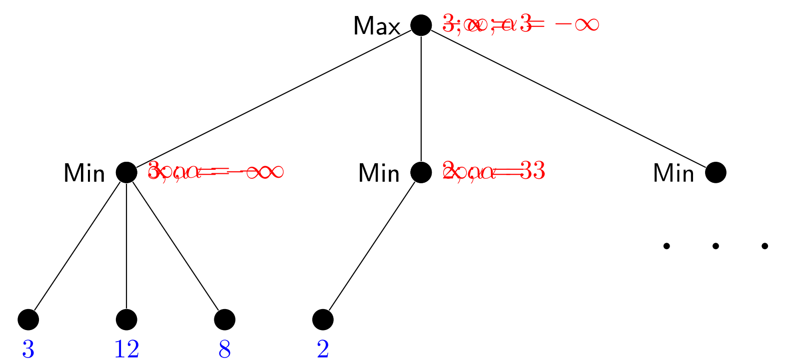
Alpha Pruning: Basic Idea (Continued)

- **Answer:** Yes! We already know at this point that the middle action won't be taken by Max.



Alpha Pruning

- What is α? For each search node n, the highest Max-node utility that search has encountered on its path from the root to n.



- **How to use α?:** In a Min node n, if one of the successors already has utility $\leq \alpha$, then stop considering n. (Pruning out its remaining successors.)

Alpha-Beta Pruning

- **Recall:**
  - **What is α:** For each search node n, the highest Max-node utility that search has encountered on its path from the root to n.
  - **How to use α:** In a Min node n, if one of the successors already has utility $\leq \alpha$, then stop considering n. (Pruning out its remaining successors.)
- **Idea:** We can use a dual method for Min:
  - **What is β:** For each search node n, the lowest Min-node utility that search has encountered on its path from the root to n.

- **How to use β:** In a Max node n, if one of the successors already has utility $\geq \beta$, then stop considering n. (Pruning out its remaining successors.)

- . . . and of course we can use both together!

Alpha-Beta Search: Pseudo-Code

- **Definition 7.4.1.** The alphabeta search algorithm is given by the following pseudo-code

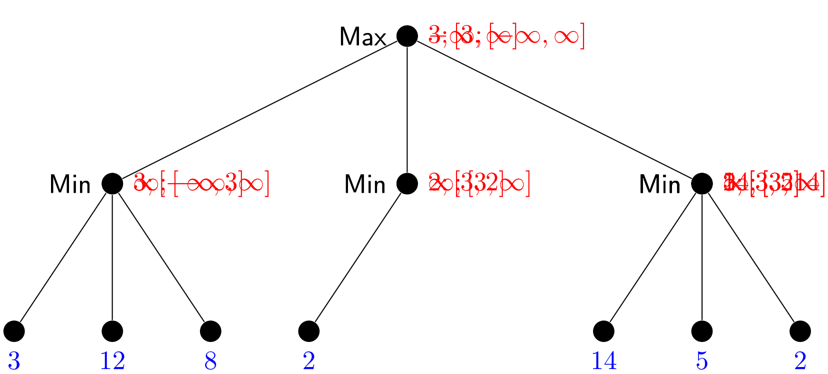
```
function Alpha−Beta−Search (s) returns an action
  v := Max−Value(s, −∞, +∞)
  return an action yielding value v in the previous function call

function Max−Value(s, α, β) returns a utility value
  if Terminal−Test(s) then return u(s)
  v:= −∞
  for each a ∈ Actions(s) do
    v := max(v,Min−Value(ChildState(s,a), α, β))
      α := max(α, v)
      if v ≥ β then return v /∗ Here: v ≥ β ⇔ α ≥ β ∗/
  return v

function Min−Value(s, α, β) returns a utility value
  if Terminal−Test(s) then return u(s)
  v := +∞
  for each a ∈ Actions(s) do
      v := min(v,Max−Value(ChildState(s,a), α, β))
      β := min(β, v)
      if v ≤ α then return v /∗ Here: v ≤ α ⇔ α ≥ β ∗/
  return v
```

$\widehat{=}$ Minimax (slide 206) + α/β book-keeping and pruning.

Note: Note that α only gets assigned a value in Max nodes, and β only gets assigned a value in Min nodes.

Alpha-Beta Search: Example

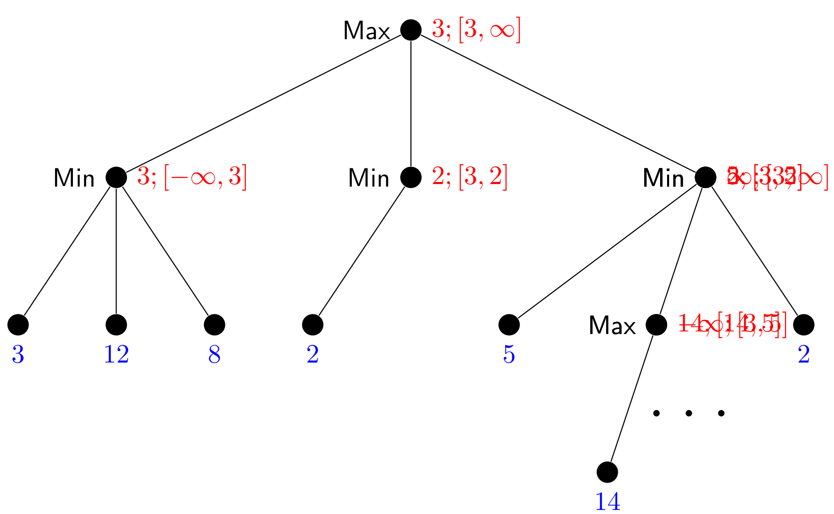
- **Notation:** $v; [\alpha, \beta]$



▷ **Note:** We could have saved work by choosing the opposite order for the successors of the rightmost Min node. Choosing the best moves (for each of Max and Min) first yields more pruning!

Alpha-Beta Search: Modified Example

▷ **Showing off some actual β pruning:**



How Much Pruning Do We Get?

▷ Choosing the best moves first yields most pruning in alphabeta search.

- ▷ The maximizing moves for Max, the minimizing moves for Min.

▷ **Assuming game tree with branching factor b and depth limit d:**

- ▷ Minimax would have to search b^d nodes.
- ▷ Best case: If we always choose the best moves first, then the search tree is reduced to $b^{\frac{d}{2}}$ nodes!
- ▷ Practice: It is often possible to get very close to the best case by simple move-ordering methods.

▷ **Example Chess**:

- ▷ Move ordering: Try captures first, then threats, then forward moves, then backward moves.
- ▷ From 35^d to $35^{\frac{d}{2}}$. E.g., if we have the time to search a billion (10^9) nodes, then Minimax looks ahead $d = 6$ moves, i.e., 3 rounds (white-black) of the game. Alpha-beta search looks ahead 6 rounds.

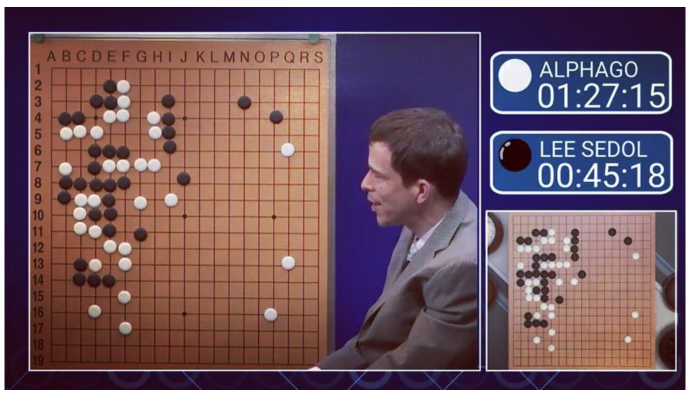
7.5 Monte-Carlo Tree Search (MCTS)

Video Nuggets covering this section can be found at https://fau.tv/clip/id/22259 and https://fau.tv/clip/id/22262.

We will now come to the most visible game-play program in recent times: The AlphaGo system for the game of Go. This has been out of reach of the state of the art (and thus for alphabeta search) until 2016. This challenge was cracked by a different technique, which we will discuss in this section.

And now ...

▷ **AlphaGo = Monte Carlo tree search (AI-1) + neural networks (AI-2)**



CC-BY-SA: Buster Benson@ https://www.flickr.com/photos/erikbenson/25717574115

Monte-Carlo Tree Search: Basic Ideas

▷ **Observation:** We do not always have good evaluation functions.

▷ **Definition 7.5.1.** For Monte Carlo sampling we evaluate actions through sampling.

▷ When deciding which action to take on game state s:

while time not up **do**
  select action a applicable **to** s
  run a random sample from a **until** terminal state t
return an a **for** s with maximal average $u(t)$

▷ **Definition 7.5.2.** For the Monte Carlo tree search algorithm (MCTS) we maintain a search tree T, the MCTS tree.

while time not up **do**
  apply actions within T **to** select a leaf state s'
  select action a' applicable **to** s', run random sample from a'
  add s' **to** T, update averages etc.
return an a **for** s with maximal average $u(t)$

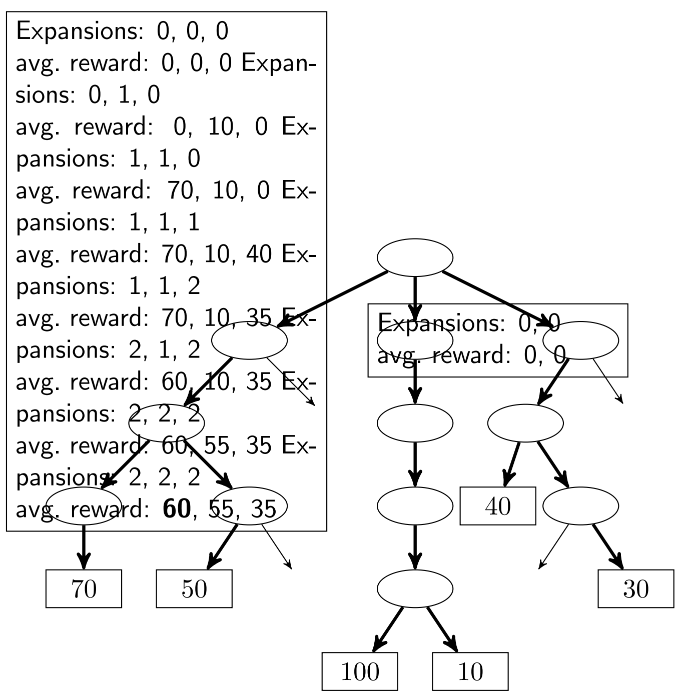
When executing a, keep the part of T below a.

- ▷ Compared to alphabeta search: no exhaustive enumeration.
  - ▷ **Pro**: running time & memory.
  - ▷ **Contra**: need good guidance how to "select" and "sample".

This looks only at a fraction of the search tree, so it is crucial to have good guidance *where to go*, i.e. which part of the search tree to look at.

Monte-Carlo Sampling: Illustration of Sampling

- ▷ **Idea:** Sample the search tree keeping track of the average utilities.
- ▷ **Example 7.5.3 (Single-player, for simplicity).** (with adversary, distinguish max/min nodes)

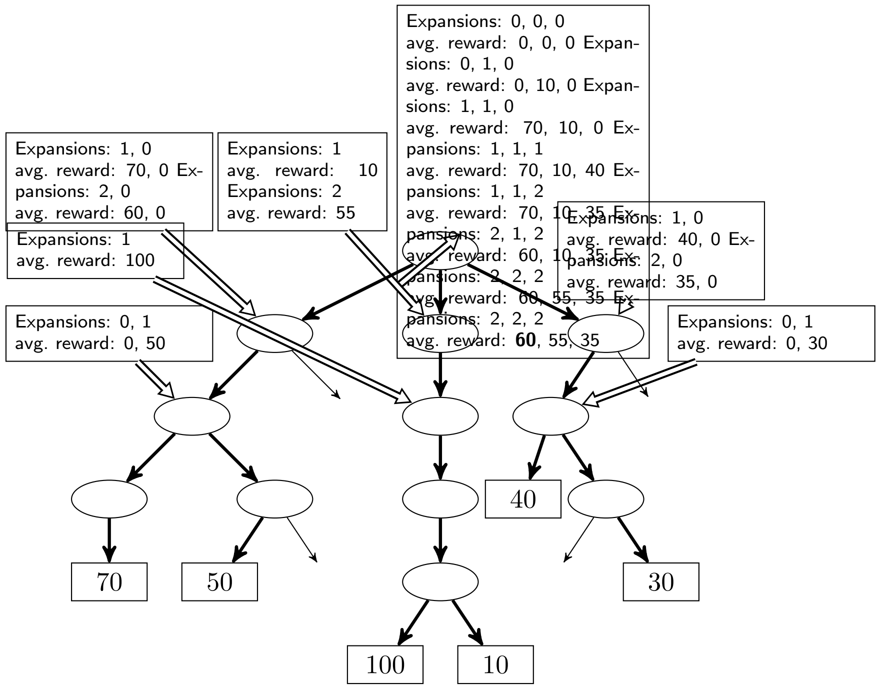


The sampling goes middle, left, right, right, left, middle. Then it stops and selects the highest-average action, 60, left. After first sample, when values in initial state are being updated, we have the following "expansions" and "avg. reward fields": small number of expansions favored for exploration: visit parts of the tree rarely visited before, what is out there? avg. reward: high values favored for exploitation: focus on promising parts of the search tree.

Monte-Carlo Tree Search: Building the Tree

▷ **Idea:** we can save work by building the tree as we go along.

▷ **Example 7.5.4 (Redoing the previous example).**



This is the exact same search as on previous slide, but incrementally building the search tree, by always keeping the first state of the sample. The first three iterations middle, left, right, go to show the tree extension; do point out here that, like the root node, the nodes added to the tree have expansions and avg reward counters for every applicable action. Then in next iteration right, after 30 leaf node was found, an important thing is that the averages get updated *along the entire path*, i.e., not only in the root as we did before, but also in the nodes along the way. After all six iterations have been done, as before we select the action left, value 60; but we keep the part of the tree below that action, "saving relevant work already done before".

How to Guide the Search in MCTS?

▷ **How to "sample"?:** What exactly is "random"?

▷ **Classical formulation:** balance exploitation vs. exploration.

- ▷ Exploitation: Prefer moves that have high average already (interesting regions of state space)
- ▷ Exploration: Prefer moves that have not been tried a lot yet (don't overlook other, possibly better, options)

▷ UCT: "Upper Confidence bounds applied to Trees" [KS06].

- ▷ Inspired by Multi-Armed Bandit (as in: Casino) problems.
- ▷ Basically a formula defining the balance. Very popular (buzzword).

- Recent critics (e.g. [FD14]): Exploitation in search is very different from the Casino, as the "accumulated rewards" are fictitious (we're only thinking about the game, not actually playing and winning/losing all the time).

AlphaGo: Overview

- **Definition 7.5.5 (Neural Networks in AlphaGo).**
  - Policy networks: Given a state s, output a probability distribution over the actions applicable in s.
  - Value networks: Given a state s, output a number estimating the game value of s.
- **Combination with MCTS:**
  - Policy networks bias the action choices within the MCTS tree (and hence the leaf state selection), and bias the random samples.
  - Value networks are an additional source of state values in the MCTS tree, along with the random samples.
- And now in a little more detail

Neural Networks in AlphaGo

- **Neural network training pipeline and architecture:**

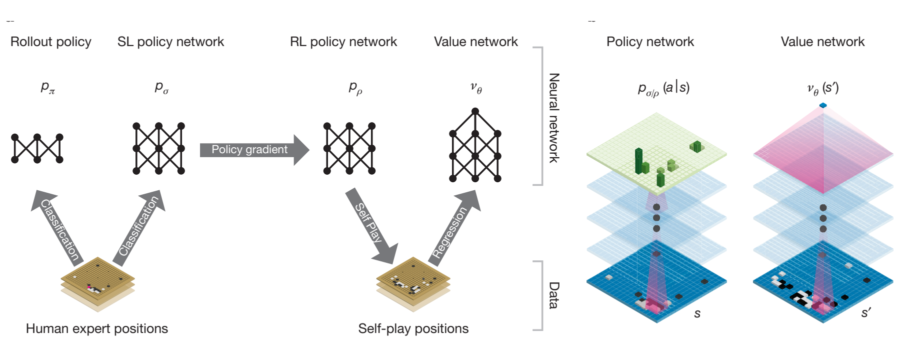


Illustration taken from [Sil+16] .

- *Rollout policy p_π:* Simple but fast, ≈ prior work on Go.
- *SL policy network p_σ:* Supervised learning, human-expert data ("learn to choose an expert action").
- *RL policy network p_ρ:* Reinforcement learning, self-play ("learn to win").
- *Value network v_θ:* Use self-play games with p_ρ as training data for game-position evaluation v_θ ("predict which player will win in this state").

Comments on the Figure:

a A fast rollout policy p_π and supervised learning (SL) policy network p_σ are trained to predict human expert moves in a data set of positions. A reinforcement learning (RL) policy network p_ρ is initialized to the SL policy network, and is then improved by policy gradient learning to maximize the outcome (that is, winning more games) against previous versions of the policy network. A new data set is generated by playing games of self-play with the RL policy network. Finally, a value network v_θ is trained by regression to predict the expected outcome (that is, whether the current player wins) in positions from the self-play data set.

b Schematic representation of the neural network architecture used in AlphaGo. The policy network takes a representation of the board position s as its input, passes it through many convolutional layers with parameters σ (SL policy network) or ρ (RL policy network), and outputs a probability distribution $p_\sigma(a|s)$ or $p_\rho(a|s)$ over legal moves a, represented by a probability map over the board. The value network similarly uses many convolutional layers with parameters θ, but outputs a scalar value $v_\theta(s')$ that predicts the expected outcome in position s'.

Neural Networks + MCTS in AlphaGo

▷ **Monte Carlo tree search in AlphaGo:**

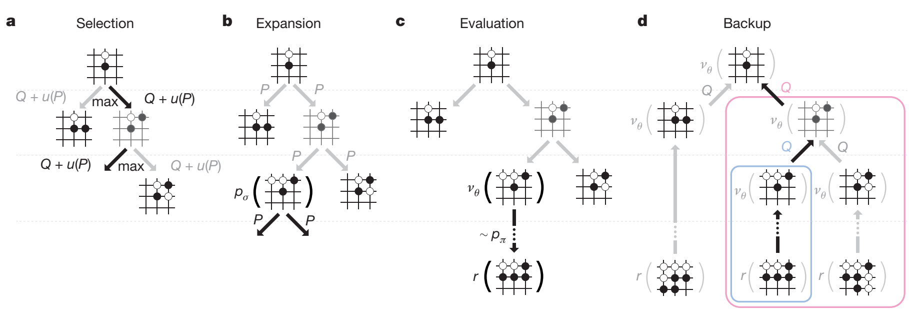


Illustration taken from [Sil+16]

- ▷ *Rollout policy* p_π: Action choice in random samples.
- ▷ *SL policy network* p_σ: Action choice bias within the UCTS tree (stored as "P", gets smaller to "$u(P)$" with number of visits); along with quality Q.
- ▷ *RL policy network* p_ρ: Not used here (used only to learn v_θ).
- ▷ *Value network* v_θ: Used to evaluate leaf states s, in linear sum with the value returned by a random sample on s.

Comments on the Figure:

a Each simulation traverses the tree by selecting the edge with maximum action value Q, plus a bonus $u(P)$ that depends on a stored prior probability P for that edge.

b The leaf node may be expanded; the new node is processed once by the policy network p_σ and the output probabilities are stored as prior probabilities P for each action.

c At the end of a simulation, the leaf node is evaluated in two ways:

- using the value network v_θ,
- and by running a rollout to the end of the game

with the fast rollout policy $p\ \pi$, then computing the winner with function r.

d Action values Q are updated to track the mean value of all evaluations $r(\cdot)$ and $v_\theta(\cdot)$ in the subtree below that action.

AlphaGo, Conclusion?: This is definitely a great achievement!

- "Search + neural networks" looks like a great formula for general problem solving.
- expect to see lots of research on this in the coming decade(s).
- The AlphaGo design is quite intricate (architecture, learning workflow, training data design, neural network architectures, . . .).
- How much of this is reusable in/generalizes to other problems?
- Still lots of human expertise in here. Not as much, like in Chess, about the game itself. But rather, in the design of the neural networks + learning architecture.

7.6 State of the Art

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22250`.

State of the Art

▷ **Some well-known board games:**

- ▷ **Chess**: Up next.
- ▷ **Othello (Reversi)**: In 1997, "Logistello" beat the human world champion. Best computer players now are clearly better than best human players.
- ▷ **Checkers (Dame)**: Since 1994, "Chinook" is the offical world champion. In 2007, it was shown to be *unbeatable*: Checkers is *solved*. (We know the exact value of, and optimal strategy for, the initial state.)
- ▷ **Go**: In 2016, AlphaGo beat the Grandmaster Lee Sedol, cracking the "holy grail" of board games. In 2017, "AlphaZero" – a variant of AlphaGo with zero prior knowledge beat all reigning champion systems in all board games (including AlphaGo) 100/0 after 24h of self-play.
- ▷ **Intuition:** Board Games are considered a "solved problem" from the AI perspective.

Computer Chess: "Deep Blue" beat Garry Kasparov in 1997

Duell Kasparow gegen Deep Blue (1997): *Demütigende Niederlage*

- ▷ 6 games, final score 3.5 : 2.5.
- ▷ Specialized Chess hardware, 30 nodes with 16 processors each.
- ▷ Alphabeta search plus human knowledge. (more details in a moment)
- ▷ Nowadays, standard PC hardware plays at world champion level.

Computer Chess: Famous Quotes

▷ The chess machine is an ideal one to start with, since (Claude Shannon (1949))

1. the problem is sharply defined both in allowed operations (the moves) and in the ultimate goal (checkmate),
2. it is neither so simple as to be trivial nor too difficult for satisfactory solution,
3. chess is generally considered to require "thinking" for skilful play, [...]
4. the discrete structure of chess fits well into the digital nature of modern computers.

▷ Chess is the drosophila of Artificial Intelligence. (Alexander Kronrod (1965))

Computer Chess: Another Famous Quote

▷ In 1965, the Russian mathematician Alexander Kronrod said, "Chess is the Drosophila of artificial intelligence."

However, computer chess has developed much as genetics might have if the geneticists had concentrated their efforts starting in 1910 on breeding racing Drosophilae. We would have some science, but mainly we would have very fast fruit flies. (John McCarthy (1997))

7.7 Conclusion

Summary

- ▷ Games (2-player turn-taking zero-sum discrete and finite games) can be understood as a simple extension of classical search problems.
- ▷ Each player tries to reach a terminal state with the best possible utility (maximal vs. minimal).

▷ Minimax searches the game depth-first, max'ing and min'ing at the respective turns of each player. It yields perfect play, but takes time $\mathcal{O}(b^d)$ where b is the branching factor and d the search depth.

▷ Except in trivial games (Tic-Tac-Toe), Minimax needs a depth limit and apply an evaluation function to estimate the value of the cut-off states.

▷ Alpha-beta search remembers the best values achieved for each player elsewhere in the tree already, and prunes out sub-trees that won't be reached in the game.

▷ Monte Carlo tree search (MCTS) samples game branches, and averages the findings. AlphaGo controls this using neural networks: evaluation function ("value network"), and action filter ("policy network").

Suggested Reading:

- *Chapter 5: Adversarial Search*, Sections 5.1 – 5.4 [RN09].
  - Section 5.1 corresponds to my "Introduction", Section 5.2 corresponds to my "Minimax Search", Section 5.3 corresponds to my "Alpha-Beta Search". I have tried to add some additional clarifying illustrations. RN gives many complementary explanations, nice as additional background reading.
  - Section 5.4 corresponds to my "Evaluation Functions", but discusses additional aspects relating to narrowing the search and look-up from opening/termination databases. Nice as additional background reading.
  - I suppose a discussion of MCTS and AlphaGo will be added to the next edition ...

Chapter 8

Constraint Satisfaction Problems

In the last chapters we have studied methods for "general problem", i.e. such that are applicable to all problems that are expressible in terms of states and "actions". It is crucial to realize that these states were atomic, which makes the algorithms employed (search algorithms) relatively simple and generic, but does not let them exploit the any knowledge we might have about the *internal structure of states*.

In this chapter, we will look into algorithms that do just that by progressing to factored states representations. We will see that this allows for algorithms that are many orders of magnitude more efficient than search algorithms.

To give an intuition for factored states representations we, we present some motivational examples in section 8.1 and go into detail of the Waltz algorithm, which gave rise to the main ideas of constraint satisfaction algorithms in section 8.2. section 8.3 and section 8.4 define constraint satisfaction problems formally and use that to develop a class of backtracking/search based algorithms. The main contribution of the factored states representations is that we can formulate advanced search heuristics that guide search based on the structure of the states.

8.1 Constraint Satisfaction Problems: Motivation

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22251`.

A (Constraint Satisfaction) Problem

▷ **Example 8.1.1 (Tournament Schedule).** Who's going to play against who, when and where?

Constraint Satisfaction Problems (CSPs)

▷ Standard search problem: state is a "black box" any old data structure that supports goal test, eval, successor state, ...

▷ **Definition 8.1.2.** A constraint satisfaction problem (CSP) is a search problem, where the states are given by a finite set $V:=\{X_1,\ldots,X_n\}$ of variables and domains $\{D_v|v\in V\}$ and the goal state are specified by a set of constraints specifying allowable combinations of values for subsets of variables.

▷ **Definition 8.1.3.** A constraint network is satisfiable, iff it has a solution a total, consistent variable assignment.

▷ **Definition 8.1.4.** The process of finding solutions to CSPs is called constraint solving.

▷ *Remark 8.1.5.* We are using factored representation for world states now.

▷ Simple example of a *formal representation language*

▷ Allows useful *general-purpose* algorithms with more power than standard tree search algorithm.

Another Constraint Satisfaction Problem

▷ **Example 8.1.6 (SuDoKu).** Fill the cells with row/column/block-unique digits

2	5			3		9		1
	1				4			
4		7				2		8
		5	2					
				9	8	1		
	4				3			
			3	6			7	2
	7							3
9		3				6		4

$\rightsquigarrow$

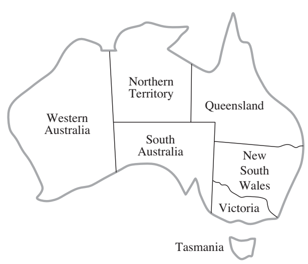
2	5	8	7	3	6	9	4	1
6	1	9	8	2	4	3	5	7
4	3	7	9	1	5	2	6	8
3	9	5	2	7	1	4	8	6
7	6	2	4	9	8	1	3	5
8	4	1	6	5	3	7	2	9
1	8	4	3	6	9	5	7	2
5	7	6	1	4	2	8	9	3
9	2	3	5	8	7	6	1	4

- ▷ Variables: The 81 cells.
- ▷ Domains: Numbers $1, \ldots, 9$.
- ▷ Constraints: Each number only once in each row, column, block.

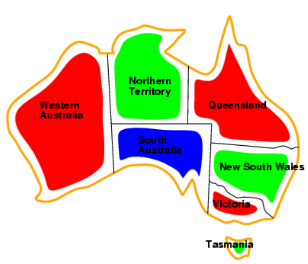
CSP Example: Map-Coloring

▷ **Definition 8.1.7.** Given a map M, the map coloring problem is to assign colors to regions in a map so that no adjoining regions have the same color.

▷ **Example 8.1.8 (Map coloring in Australia).**



- ▷ Variables: WA, NT, Q, NSW, V, SA, T
- ▷ Domains: $D_i = \{\text{red}, \text{green}, \text{blue}\}$
- ▷ Constraints: adjacent regions must have different colors e.g., $\text{WA} \neq \text{NT}$ (if the language allows this), or $\langle \text{WA}, \text{NT} \rangle \in \{\langle \text{red}, \text{green} \rangle, \langle \text{red}, \text{blue} \rangle, \langle \text{green}, \text{red} \rangle, \ldots\}$



- ▷ **Intuition**: solutions map variables to domain values satisfying all constraints,
- ▷ e.g., $\{\text{WA} = \text{red}, \text{NT} = \text{green}, \ldots\}$

Bundesliga Constraints

▷ Variables: $v_{\text{Avs.}B}$ where A and B are teams, with domains $\{1, \ldots, 34\}$: For each match, the index of the weekend where it is scheduled.

▷ (Some) constraints:

▷ If $\{A, B\} \cap \{C, D\} \neq \emptyset$: $v_{A\text{vs.}B} \neq v_{C\text{vs.}D}$ (each team only one match per day).

▷ If $\{A, B\} = \{C, D\}$: $v_{A\text{vs.}B} \leq 17 < v_{C\text{vs.}D}$ or $v_{C\text{vs.}D} \leq 17 < v_{A\text{vs.}B}$ (each pairing exactly once in each half-season).

▷ If $A = C$: $v_{A\text{vs.}B} + 1 \neq v_{C\text{vs.}D}$ (each team alternates between home matches and away matches).

▷ Leading teams of last season meet near the end of each half-season.

▷ ...

How to Solve the Bundesliga Constraints?

▷ 306 nested for-loops (for each of the 306 matches), each ranging from 1 to 306. Within the innermost loop, test whether the current values are (a) a permutation and, if so, (b) a legal Bundesliga schedule.

  ▷ **Estimated running time**: End of this universe, and the next couple billion ones after it ...

▷ Directly enumerate all permutations of the numbers $1, \ldots, 306$, test for each whether it's a legal Bundesliga schedule.

  ▷ **Estimated running time**: Maybe only the time span of a few thousand universes.

▷ View this as variables/constraints and use backtracking (this chapter)

  ▷ **Executed running time**: About 1 minute.

▷ **How do they actually do it?:** Modern computers and CSP methods: fractions of a second. 19th (20th/21st?) century: Combinatorics and manual work.

▷ **Try it yourself:** with an off-the shelf CSP solver, e.g. Minion [Min]

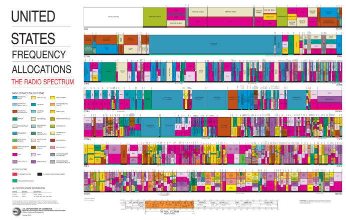
More Constraint Satisfaction Problems

Traveling Tournament Problem

Scheduling

Timetabling

Radio Frequency Assignment



1. U.S. Major League Baseball, 30 teams, each 162 games. There's one crucial additional difficulty, in comparison to Bundesliga. Which one? Travel is a major issue here!! Hence "Traveling Tournament Problem" in reference to the TSP.

2. This particular scheduling problem is called "car sequencing", how to most efficiently get cars through the available machines when making the final customer configuration (non-standard/flexible/custom extras).

3. Another common form of scheduling ...

4. The problem of assigning radio frequencies so that all can operate together without noticeable interference. Variabledomains are available frequencies, constraints take form of $|x - y| > \delta_{xy}$, where delta depends on the position of x and y as well as the physical environment.

Our Agenda for This Topic

- ▷ Our treatment of the topic "Constraint Satisfaction Problems" consists of Chapters 7 and 8. in [RN03]
- ▷ **This Chapter**: Basic definitions and concepts; naïve backtracking search.
  - ▷ Sets up the framework. Backtracking underlies many successful algorithms for solving constraint satisfaction problems (and, naturally, we start with the simplest version thereof).
- ▷ **Next Chapter**: Inference and decomposition methods.
  - ▷ Inference reduces the search space of backtracking. Decomposition methods break the problem into smaller pieces. Both are crucial for efficiency in practice.

Our Agenda for This Chapter

- How are constraint networks, and assignments, consistency, solutions: How are constraint satisfaction problems defined? What is a solution?
  - Get ourselves on firm ground.
- **Naïve Backtracking**: How does backtracking work? What are its main weaknesses?
  - Serves to understand the basic workings of this wide-spread algorithm, and to motivate its enhancements.
- **Variable- and Value Ordering**: How should we guide backtracking search?
  - Simple methods for making backtracking aware of the structure of the problem, and thereby reduce search.

8.2 The Waltz Algorithm

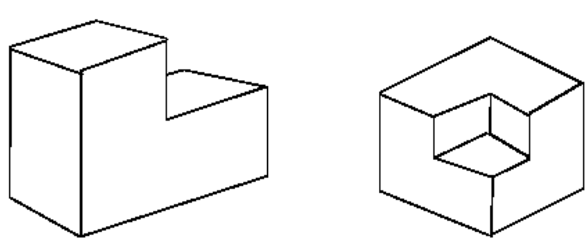
We will now have a detailed look at the problem (and innovative solution) that started the field of constraint satisfaction problems.

Background:

Adolfo Guzman worked on an algorithm to count the number of simple objects (like children's blocks) in a line drawing. David Huffman formalized the problem and limited it to objects in general position, such that the vertices are always adjacent to three faces and each vertex is formed from three planes at right angles (trihedral). Furthermore, the drawings could only have three kinds of lines: object boundary, concave, and convex. Huffman enumerated all possible configurations of lines around a vertex. This problem was too narrow for real-world situations, so Waltz generalized it to include cracks, shadows, non-trihedral vertices and light. This resulted in over 50 different line labels and thousands of different junctions. [ILD]

The Waltz Algorithm

- **Remark:** One of the earliest examples of applied CSPs.
- **Motivation:** Interpret line drawings of polyhedra.



- **Problem:** Are intersections convex or concave? (interpret $\widehat{=}$ label as such)

▷ **Idea:** Adjacent intersections impose constraints on each other. Use CSP to find a unique set of labelings.

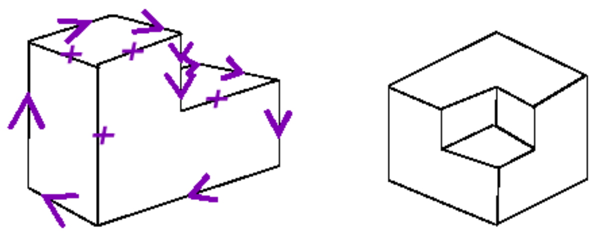
Waltz Algorithm on Simple Scenes

▷ **Assumptions:** All objects

- ▷ have no shadows or cracks,
- ▷ have only three-faced vertices,
- ▷ are in "general position", i.e. no junctions change with small movements of the eye.

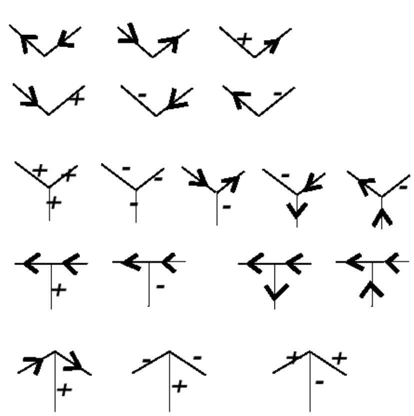
▷ **Observation 8.2.1.** *Then each line on the images is one of the following:*

- ▷ *a boundary line (edge of an object) (<) with right hand of arrow denoting "solid" and left hand denoting "space"*
- ▷ *an interior convex edge* (label with "+")
- ▷ *an interior concave edge* (label with "-")



18 Legal Kinds of Junctions

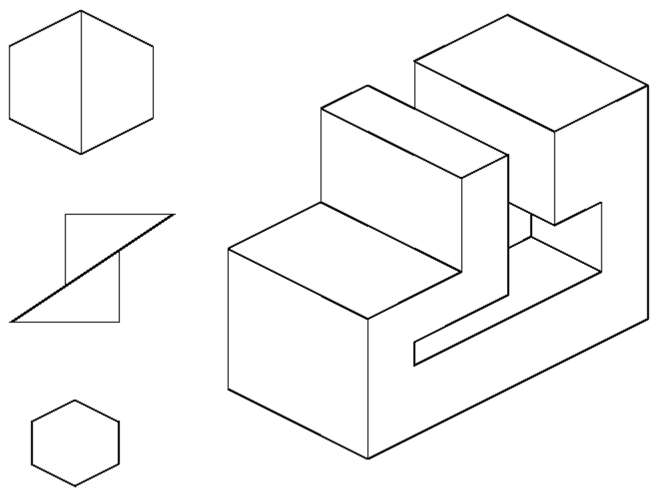
▷ **Observation 8.2.2.** *There are only 18 "legal" kinds of junctions:*



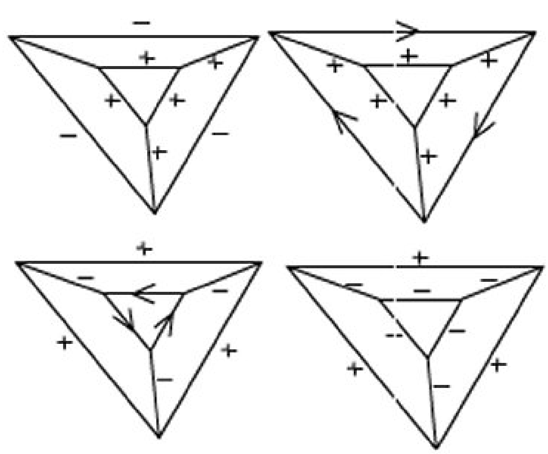
- **Idea:** given a representation of a diagram
  - label each junction in one of these manners (lots of possible ways)
  - junctions must be labeled, so that lines are labeled consistently
- **Fun Fact:** CSP always works perfectly! (early success story for CSP [Wal75])

Waltz's Examples

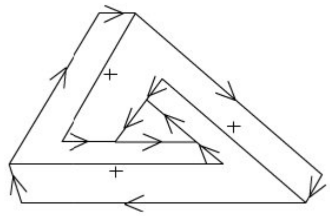
- In his dissertation 1972 [Wal75] David Waltz used the following examples



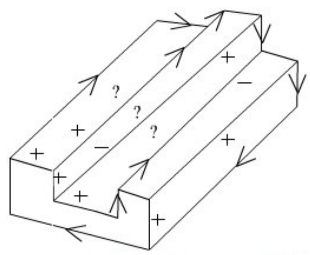
Waltz Algorithm (More Examples): Ambiguous Figures



Waltz Algorithm (More Examples): Impossible Figures



Consistent labelling for impossible figure



No consistent labelling possible

8.3 CSP: Towards a Formal Definition

We will now work our way towards a definition of CSPs that is formal enough so that we can define the concept of a solution. This gives use the necessary grounding to talk about algorithms later. **Video Nuggets** covering this section can be found at `https://fau.tv/clip/id/22277` and `https://fau.tv/clip/id/22279`.

Types of CSPs

- **Definition 8.3.1.** We call a CSP discrete, iff all of the variables have countable domains; we have two kinds:
  - finite domains (size $d \rightsquigarrow \mathcal{O}(d^n)$ solutions)
    - e.g., Boolean CSPs (solvability $\widehat{=}$ Boolean satisfiability $\rightsquigarrow$ NP complete)
  - infinite domains (e.g. integers, strings, etc.)
    - e.g., job scheduling, variables are start/end days for each job
    - need a "constraint language", e.g., $StartJob_1 + 5 \leq StartJob_3$
    - linear constraints decidable, nonlinear ones undecidable
- **Definition 8.3.2.** We call a CSP continuous, iff one domain is uncountable.
- **Example 8.3.3.** Start/end times for Hubble Telescope observations form a continuous CSP.

▷ **Theorem 8.3.4.** *Linear constraints solvable in poly time by linear programming methods.*

▷ **Theorem 8.3.5.** *There cannot be optimal algorithms for nonlinear constraint systems.*

Types of Constraints

▷ We classify the constraints by the number of variables they involve.

▷ **Definition 8.3.6.** Unary constraints involve a single variable, e.g., SA $\neq green$.

▷ **Definition 8.3.7.** Binary constraints involve pairs of variables, e.g., SA $\neq$ WA.

▷ **Definition 8.3.8.** Higher-order constraints involve $n = 3$ or more variables, e.g., cryptarithmetic column constraints.

The number n of variables is called the order of the constraint.

▷ **Definition 8.3.9.** Preferences (soft constraint) (e.g., red is better than green) are often representable by a cost for each variable assignment $\rightsquigarrow$ constrained optimization problems.

Non-Binary Constraints, e.g. "Send More Money"

▷ **Example 8.3.10 (Send More Money).** A student writes home:

$$\begin{array}{cccccc} & & S & E & N & D \\ + & & M & O & R & E \\ \hline & M & O & N & E & Y \end{array}$$

Puzzle: letters stand for digits, addition should work out (parents send MONEY€)

- ▷ Variables: S, E, N, D, M, O, R, Y, each with domain $\{0, \ldots, 9\}$.
- ▷ Constraints:
  1. all variables should have different values: $S \neq E$, $S \neq N$, ...
  2. first digits are non-zero: $S \neq 0$, $M \neq 0$.
  3. the addition scheme should work out: i.e. $1000 \cdot S + 100 \cdot E + 10 \cdot N + D + 1000 \cdot M + 100 \cdot O + 10 \cdot R + E = 10000 \cdot M + 1000 \cdot 0 + 100 \cdot N + 10 \cdot E + Y$.

BTW: The solution is $S \mapsto 9, E \mapsto 5, N \mapsto 6, D \mapsto 7, M \mapsto 1, O \mapsto 0, R \mapsto 8, Y \mapsto 2 \rightsquigarrow$ parents send 10652€

▷ **Definition 8.3.11.** Problems like the one in Example 8.3.10 are called crypto arithmetic puzzles.

Encoding Higher-Order Constraints as Binary ones

- **Problem:** The last constraint is of order 8. ($n = 8$ variables involved)
- **Observation 8.3.12.** *We can write the addition scheme constraint column wise using auxiliary variables, i.e. variables that do not "occur" in the original problem.*

$$\begin{aligned} D + E &= Y + 10 \cdot X_1 \\ X_1 + N + R &= E + 10 \cdot X_2 \\ X_2 + E + O &= N + 10 \cdot X_3 \\ X_3 + S + M &= O + 10 \cdot M \end{aligned}$$

$$\begin{array}{ccccc} & S & E & N & D \\ + & M & O & R & E \\ \hline M & O & N & E & Y \end{array}$$

These constraints are of order ≤ 5.

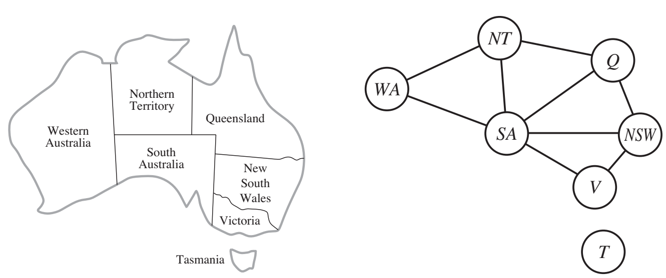
- **General Recipe:** For $n \geq 3$, encode $C(v_1, \ldots, v_{n-1}, v_n)$ as

$$C(p_1(x), \ldots, p_{n-1}(x), v_n) \wedge v_1 = p_1(x) \wedge \ldots \wedge v_{n-1} = p_{n-1}(x)$$

- **Problem:** The problem structure gets hidden. (search algorithms can get confused)

Constraint Graph

- **Definition 8.3.13.** A binary CSP is a CSP where each constraint is binary.
- **Observation 8.3.14.** *A binary CSP forms a graph called the constraint graph whose nodes are variables, and whose edges represent the constraints.*
- **Example 8.3.15.** Australia as a binary CSP



- **Intuition:** General-purpose CSP algorithms use the graph structure to speed up search. (E.g., Tasmania is an independent subproblem!)

Real-world CSPs

- **Example 8.3.16 (Assignment problems).** e.g., who teaches what class
- **Example 8.3.17 (Timetabling problems).** e.g., which class is offered when and where?
- **Example 8.3.18 (Hardware configuration).**
- **Example 8.3.19 (Spreadsheets).**
- **Example 8.3.20 (Transportation scheduling).**
- **Example 8.3.21 (Factory scheduling).**
- **Example 8.3.22 (Floorplanning).**
- **Note:** many real-world problems involve real-valued variables $\rightsquigarrow$ continuous CSPs.

Constraint Satisfaction Problems (Formal Definition)

- **Definition 8.3.23.** A constraint network is a triple $\langle V, D, C\rangle$, where
  - V is a finite set of variables,
  - $D{:=}\{D_v | v{\in}V\}$ the set of their domains, and
  - $C{:=}\{C_{uv} \subseteq D_u{\times}D_v | u,v{\in}V \text{ and } u \neq v\}$ is a set of constraints with $C_{uv} = C_{vu}^{-1}$.

  We call the undirected graph $\langle V, \{(u,v){\in}V^2 | C_{uv} \neq D_u \times D_v\}\rangle$, the constraint graph of γ.
- We will talk of CSPs and mean constraint networks.
- **Remarks:** The mathematical formulation gives us a lot of leverage:
  - $C_{uv} \subseteq D_u{\times}D_v \mathrel{\widehat{=}}$ possible assignments to variables u and v
  - Relations are the most general formalization, generally we use symbolic formulations, e.g. "$u = v$" for the relation $C_{uv} = \{(a,b) | a = b\}$ or "$u \neq v$".
  - We can express unary constraint C_u by restricting the domain of v: $D_v{:=}C_v$.

Example: SuDoKu as a Constraint Network

- **Example 8.3.24 (Formalize SuDoKu).** We use the added formality to encode SuDoKu as a constraint network, not just as a CSP as Example 8.1.6.

2	5			3		9		1
	1				4			
4		7				2		8
		5	2					
				9	8	1		
	4				3			
			3	6			7	2
	7							3
9		3				6		4

- ▷ Variables: $V = \{v_{ij} | 1 \leq i,j \leq 9\}$: v_{ij} =cell row i column j.
- ▷ Domains For all $v \in V$: $D_v = D = \{1,\ldots,9\}$.
- ▷ Unary constraint: $C_{v_{ij}} = \{d\}$ if cell i,j is pre-filled with d.
- ▷ (Binary) constraint: $C_{v_{ij}v_{i'j'}} \mathrel{\widehat{=}}$ "$v_{ij} \neq v_{i'j'}$", i.e. $C_{v_{ij}v_{i'j'}} = \{(d,d') \in D \times D | d \neq d'\}$, for: $i = i'$ (same row), or $j = j'$ (same column), or $(\lceil\frac{i}{3}\rceil, \lceil\frac{j}{3}\rceil) = (\lceil\frac{i'}{3}\rceil, \lceil\frac{j'}{3}\rceil)$ (same block).

Note that the ideas are still the same as Example 8.1.6, but in constraint networks we have a language to formulate things precisely.

Constraint Networks (Solutions)

- ▷ Let $\gamma := \langle V, D, C\rangle$ be a constraint network.
- ▷ **Definition 8.3.25.** We call a partial function $a\colon V \rightharpoonup \bigcup_{u \in V} D_u$ a variable assignment if $a(v) \in D_v$ for all $v \in \mathbf{dom}(V)$.
- ▷ **Definition 8.3.26.** Let $\mathcal{C} := \langle V, D, C\rangle$ be a constraint network and $a\colon V \rightharpoonup \bigcup_{v \in V} D_v$ a variable assignment. We say that a satisfies (otherwise violates) a constraint C_{uv}, iff $(a(u),a(v)) \in C_{uv}$. a is called consistent in $\mathcal{C}$, iff it satisfies all constraints in $\mathcal{C}$. A value $v \in D_u$ is legal for a variable u in $\mathcal{C}$, iff $\{(u,v)\}$ is a consistent assignment in $\mathcal{C}$. A variable with illegal value under a is called conflicted.
- ▷ **Example 8.3.27.** The empty assignment ϵ is (trivially) consistent in any constraint network.
- ▷ **Definition 8.3.28.** Let f and g be variable assignments, then we say that f extends (or is an extension of) g, iff $\mathbf{dom}(g) \subset \mathbf{dom}(f)$ and $f|_{\mathbf{dom}(g)} = g$.
- ▷ **Definition 8.3.29.** We call a consistent (total) assignment a solution for γ and γ itself solvable or satisfiable.

How it all fits together

- ▷ **Lemma 8.3.30.** *Higher-order constraints can be transformed into equi-satisfiable* *binary constraints using auxiliary variables.*

▷ **Corollary 8.3.31.** *Any CSP can be represented by a constraint network.*

▷ **In other words** The notion of a constraint network is a refinement of that of a CSP.

▷ So we will stick to constraint networks in this course.

▷ **Observation 8.3.32.** *We can view a constraint network as a search problem, if we take the states as the variable assignments, the actions as assignment extensions, and the goal states as consistent assignments.*

▷ **Idea:** We will explore that idea for algorithms that solve constraint networks.

Michael Kohlhase: Artificial Intelligence 1 261 2023-09-20

8.4 CSP as Search

We now follow up on Observation 8.3.32 to use search algorithms for solving constraint networks.

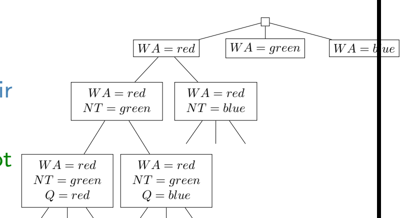
The key point of this section is that the factored states representations realized by constraint networks allow the formulation of very powerful heuristics. **A Video Nugget** covering this section can be found at `https://fau.tv/clip/id/22319`.

Standard search formulation (incremental)

▷ **Idea:** Every constraint network induces a single state problem.

▷ State are defined by the values assigned so far

- ▹ States are variable assignments
- ▹ Initial state: the empty assignment, $\emptyset$
- ▹ Actions: extend current assignment a by a pair (x,v) that does not conflicted with a.
- ▹ $\leadsto$ fail if no consistent assignments exist (not fixable!)
- ▹ Goal test: the current assignment is total.



▷ **Remark:** This is the same for all CSPs! ☺

▷ **Observation:** Every solution appears at depth n with n variables.

▷ **Idea:** Use depth first search!

▷ Path is irrelevant, so can also use complete-state formulation

▷ Branching factor $b = (n - \ell)d$ at depth ℓ, hence $n!d^n$ leaves!!!! ☹

Michael Kohlhase: Artificial Intelligence 1 262 2023-09-20

Backtracking Search

▷ Assignments for different variables are independent!

▷ e.g. first WA = red then NT = green vs. first NT = green then WA = red

▷ $\leadsto$ we only need to consider assignments to a single variable at each node

▷ $\leadsto$ $b = d$ and there are d^n leaves.

▷ **Definition 8.4.1.** Depth first search for CSPs with single-variable assignment extensions actions is called backtracking search.

▷ Backtracking search is the basic uninformed algorithm for CSPs.

▷ Can solve the n-queens problem for $\approxeq n = 25$.

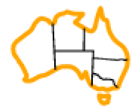
Backtracking Search (Implementation)

▷ **Definition 8.4.2.** The generic backtracking search algorithm

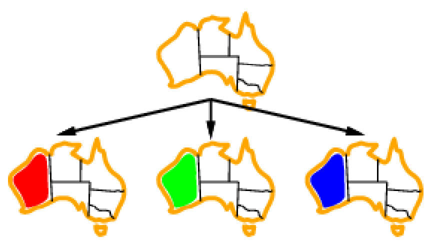
```
procedure Backtracking−Search(csp ) returns solution/failure
    return Recursive−Backtracking (∅, csp)

procedure Recursive−Backtracking (assignment) returns soln/failure
  if assignment is complete then return assignment
  var := Select−Unassigned−Variable(Variables[csp], assignment, csp)
  foreach value in Order−Domain−Values(var, assignment, csp) do
      if value is consistent with assignment given Constraints[csp] then
          add {var = value} to assignment
          result := Recursive−Backtracking(assignment,csp)
          if result ≠ failure then return result
      remove {var= value} from assignment
  return failure
```


Backtracking in Australia

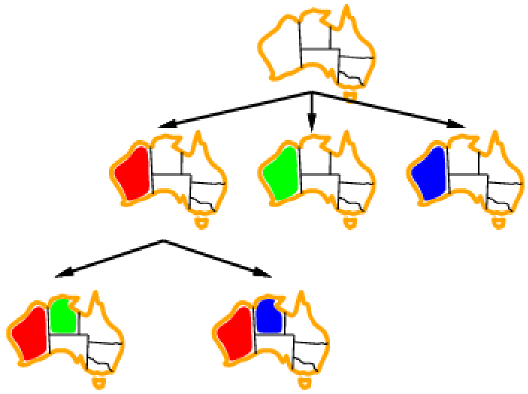
▷ **Example 8.4.3.** We apply backtracking search for a map coloring problem:
Step 1:



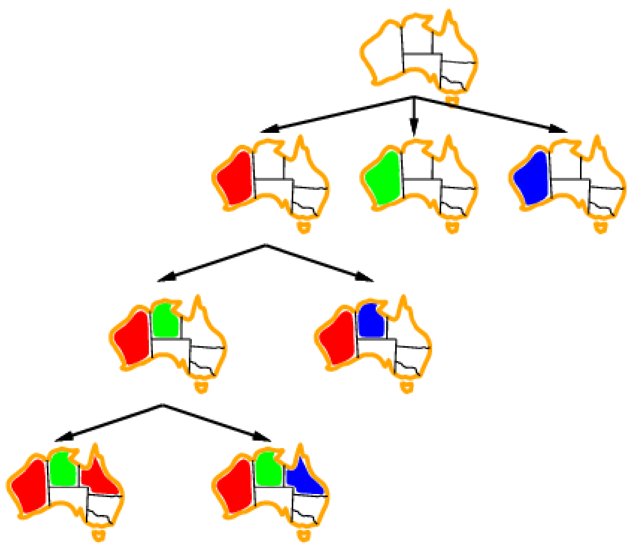
Step 2:



Step 3:



Step 4:



Improving backtracking efficiency

▷ General-purpose methods can give huge gains in speed for backtracking search.

▷ Answering the following questions well helps find powerful heuristics:

1. Which variable should be assigned next? (i.e. a variable ordering heuristic)
2. In what order should its values be tried? (i.e. a value ordering heuristic)
3. Can we detect inevitable failure early? (for pruning strategies)

4. Can we take advantage of problem structure? ($\rightsquigarrow$ inference)

▷ **Observation:** Questions 1/2 correspond to the missing subroutines Select—Unassigned—Variable and Order—Domain—Values from Definition 8.4.2.

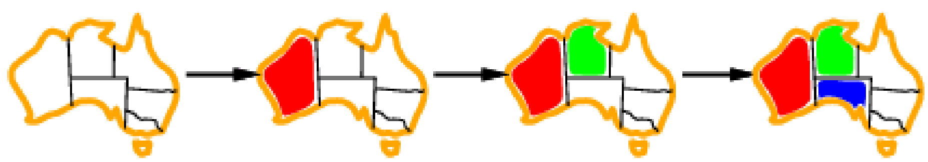
Heuristic: Minimum Remaining Values (Which Variable)

▷ **Definition 8.4.4.** The minimum remaining values (MRV) heuristic for backtracking search always chooses the variable with the fewest legal values, i.e. a variable that minimizes $\#(\{d{\in}D_v | a \cup \{v{\mapsto}d\}$ is consistent$\})$.

▷ **Intuition:** By choosing a most constrained variable v first, we reduce the outdegreebranching factor (number of sub trees generated for v) and thus reduce the size of our search tree.

▷ **Extreme case:** If $\#(\{d{\in}D_v | a \cup \{v{\mapsto}d\}$ is consistent$\}) = 1$, then the value assignment to v is forced by our previous choices.

▷ **Example 8.4.5.** In step 3 of Example 8.4.3, there is only one remaining value for SA!



Degree Heuristic (Variable Order Tie Breaker)

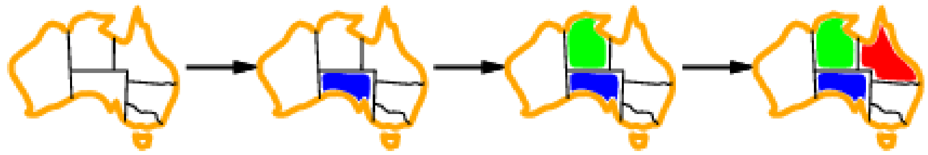
▷ **Problem:** Need a tie-breaker among MRV variables! (there was no preference in step 1,2)

▷ **Definition 8.4.6.** The degree heuristic in backtracking search always chooses a most constraining variable, i.e. always pick a v with $\#(\{v{\in}(V \backslash \mathbf{dom}(a)) | C_{uv}{\in}C\})$ maximal.

▷ By choosing a most constraining variable first, we detect inconsistencies earlier on and thus reduce the size of our search tree.

▷ **Commonly used strategy combination:** From the set of most constrained variable, pick a most constraining variable.

▷ **Example 8.4.7.**



Degree heuristic: SA = 5, T = 0, all others 2 or 3.

Where in Example 8.4.7 does the most constraining variable play a role in the choice? SA (only possible choice), NT (all choices possible except WA, V, T). Where in the illustration does most constrained variable play a role in the choice? NT (all choices possible except T), Q (only Q and WA possible).

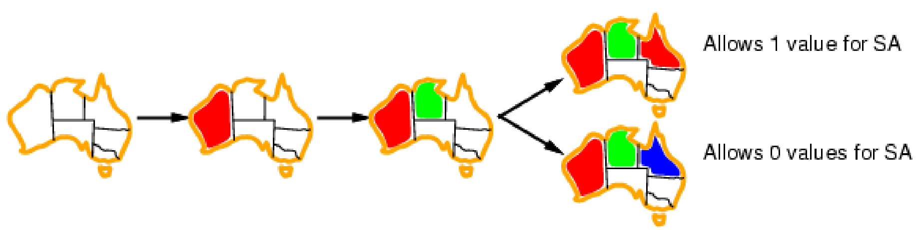
Least Constraining Value Heuristic (Value Ordering)

▷ **Definition 8.4.8.** Given a variable, the least constraining value heuristic chooses the least constraining value: the one that rules out the fewest values in the remaining variables, i.e. for a given variable v pick a value $d\in D_v$ that minimizes

$$\#(\{e\in(D_u\backslash\mathbf{dom}(a))|C_{uv}\in C \text{ and } (e,d)\notin C_{uv}\})$$

▷ By choosing the least constraining value first, we increase the chances to not rule out the solutions below the current node.

▷ **Example 8.4.9.**



▷ Combining these heuristics makes 1000 queens feasible.

8.5 Conclusion & Preview

Summary & Preview

▷ Summary of "CSP as Search":

- ▹ Constraint networks γ consist of variables, associated with finite domains, and constraints which are binary relations specifying permissible value pairs.
- ▹ A variable assignment a maps some variables to values. a is consistent if it complies with all constraints. A consistent total assignment is a solution.

- ▷ The constraint satisfaction problem (CSP) consists in finding a solution for a constraint network. This has numerous applications including, e.g., scheduling and timetabling.
- ▷ Backtracking search assigns variable one by one, pruning inconsistent variable assignments.
- ▷ Variable orderings in backtracking can dramatically reduce the size of the search tree. Value orderings have this potential (only) in solvable sub trees.

▷ **Up next:** Inference and decomposition, for improved efficiency.

Suggested Reading:

- *Chapter 6: Constraint Satisfaction Problems*, Sections 6.1 and 6.3, in [RN09].
  - Compared to our treatment of the topic "Constraint Satisfaction Problems" (chapter 8 and chapter 9), RN covers much more material, but less formally and in much less detail (in particular, my slides contain many additional in-depth examples). Nice background/additional reading, can't replace the lecture.
  - Section 6.1: Similar to my "Introduction" and "Constraint Networks", less/different examples, much less detail, more discussion of extensions/variations.
  - Section 6.3: Similar to my "Naïve Backtracking" and "Variable- and Value Ordering", with less examples and details; contains part of what I cover in chapter 9 (RN does inference first, then backtracking). Additional discussion of *backjumping*.

Chapter 9

Constraint Propagation

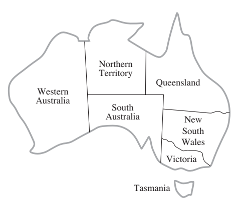
In this chapter we discuss another idea that is central to symbolic AI as a whole. The first component is that with the factored states representations, we need to use a representation language for (sets of) states. The second component is that instead of state-level search, we can graduate to representation-level search (inference), which can be much more efficient that state level search as the respective representation language actions correspond to groups of state-level actions.

9.1 Introduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22321`.

Illustration: Inference

▷ **Example 9.1.1.** A constraint network γ:



▷ **Question:** An additional constraint we can add without losing any solutions?

▷ **Example 9.1.2.** C_{WAQ} := "=". If WA and Q are assigned different colors, then NT must be assigned the 3rd color, leaving no color for SA.

▷ **Intuition:** Adding constraints without losing solutions = obtaining an equivalent network with a "tighter description" and hence with a smaller number of consistent variable assignments.

Illustration: Decomposition

▷ **Example 9.1.3.** constraint network γ:

- ▷ We can separate this into two independent constraint networks.
- ▷ Tasmania is not adjacent to any other state. Thus we can color Australia first, and assign an arbitrary color to Tasmania afterwards.
- ▷ Decomposition methods exploit the structure of the constraint network. They identify separate parts (sub-networks) whose inter-dependencies are "simple" and can be handled efficiently.
- ▷ **Example 9.1.4 (Extreme case).** No inter-dependencies at all, as in our example here.

Our Agenda for This Chapter

- ▷ Inference: How does inference work in principle? What are relevant practical aspects?
  - ▷ Fundamental concepts underlying inference, basic facts about its use.
- ▷ Forward checking: What is the simplest instance of inference?
  - ▷ Gets us started on this subject.
- ▷ Arc consistency: How to make inferences between variables whose value is not fixed yet?
  - ▷ Details a state of the art inference method.
- ▷ **Decomposition**: constraint graphs, and two simple cases
  - ▷ How to capture dependencies in a constraint network? What are "simple cases"?
  - ▷ Basic results on this subject.
- ▷ Cutset conditioning: What if we're not in a simple case?
  - ▷ Outlines the most easily understandable technique for decomposition in the general case.

9.2 Inference

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22326`.

Inference: Basic Facts

▷ **Definition 9.2.1.** Inference in constraint networks consists in deducing additional constraints, that follow from the already known constraints, i.e. that are legalsatisfied in all solutions.

▷ **Example 9.2.2.** It's what you do all the time when playing SuDoKu:

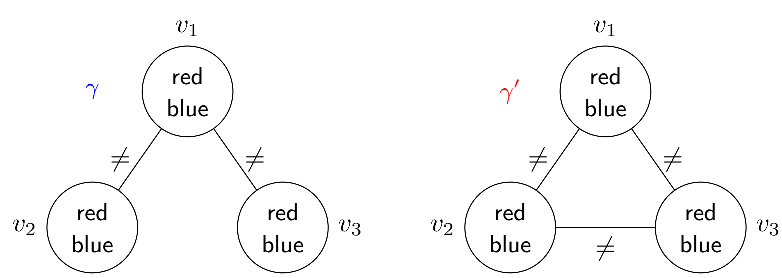
	5	8	7		6	9	4	1
		9	8		4	3	5	7
4		7	9		5	2	6	8
3	9	5	2	7	1	4	8	6
7	6	2	4	9	8	1	3	5
8	4	1	6	5	3	7	2	9
1	8	4	3	6	9	5	7	2
5	7	6	1	4	2	8	9	3
9	2	3	5	8	7	6	1	4

▷ **Formally:** Replace γ by an equivalent and strictly tighter constraint network γ'.

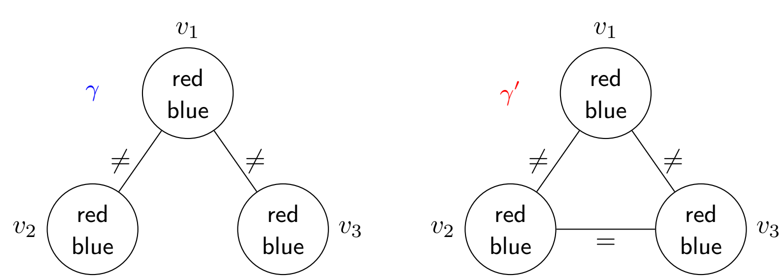
Equivalent Constraint Networks

▷ **Definition 9.2.3.** We say that two constraint networks $\gamma{:=}\langle V, D, C\rangle$ and $\gamma'{:=}\langle V, D', C'\rangle$ sharing the same set of variables are equivalent, (write $\gamma'\equiv\gamma$), if they have the same solutions.

▷ **Example 9.2.4.**



Are these constraint networks equivalent? No.



Are these constraint networks equivalent? Yes.

Tightness

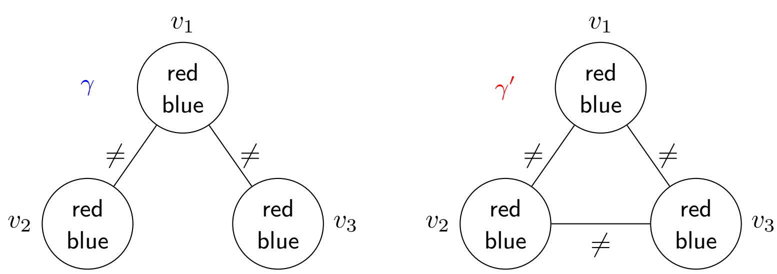
▷ **Definition 9.2.5 (Tightness).** Let $\gamma{:=}\langle V, D, C\rangle$ and $\gamma' = \langle V_{\gamma'}, D_{\gamma'}, C_{\gamma'}\rangle$ be constraint networks sharing the same set of variables, then γ' is tighter than γ, (write $\gamma' \sqsubseteq \gamma$), if:

(i) For all $v{\in}V$: $D_v \subseteq D_v$.

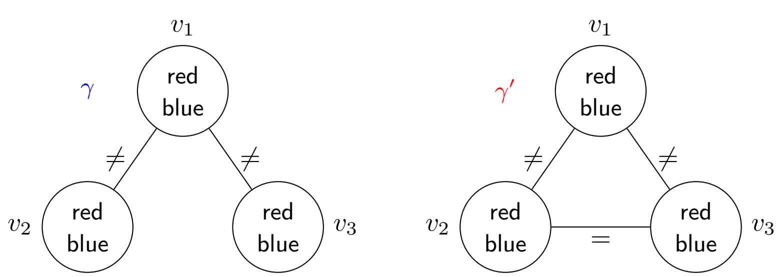
(ii) For all $u \neq v \in V$ and $C_{uv}{\in}C_{\gamma'}$: either $C_{uv}{\notin}C$ or $C_{uv} \subseteq C_{uv}$.

γ' is strictly tighter than γ, (written $\gamma' \sqsubset \gamma$), if at least one of these inclusions is proper.

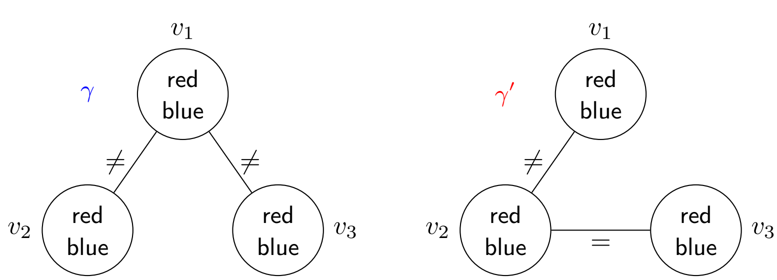
▷ **Example 9.2.6.**



Here, we do have $\gamma' \sqsubseteq$¶!.



Here, we do have $\gamma' \sqsubseteq$¶!.



Here, we do not have $\gamma' \sqsubseteq$¶!.

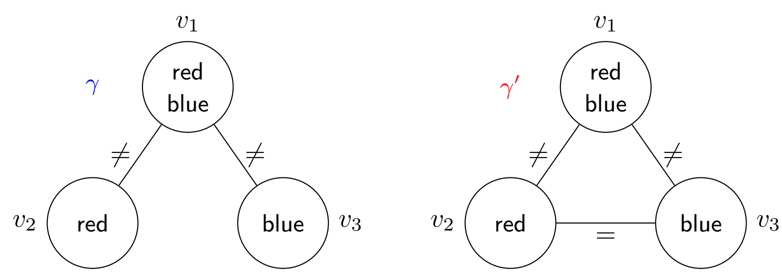
▷ **Intuition:** Strict tightness $\widehat{=}$ γ' *has the same constraints as* ¶!, *plus some.*

Equivalence + Tightness = Inference

▷ **Theorem 9.2.7.** *Let γ and γ' be constraint networks such that $\gamma' \equiv \gamma$ and $\gamma' \sqsubseteq \gamma$. Then γ' has the same solutions as, but fewer consistent assignments than, γ.*

▷ $\leadsto$ γ' is a better encoding of the underlying problem.

▷ **Example 9.2.8.**



a cannot be extended to a solution (neither in γ nor in γ' because they're equivalent). a is consistent with γ, but not with γ'.

How to Use Inference in CSP Solvers?

▷ **Simple:** Inference as a pre process:

- ▷ **When**: Just once before search starts.
- ▷ **Effect**: Little running time overhead, little pruning power. Not considered here.

▷ **More Advanced:** Inference during search:

- ▷ **When**: At every recursive call of backtracking.
- ▷ **Effect**: Strong pruning power, may have large running time overhead.

▷ **Search vs. Inference:** The more complex the inference, the *smaller* the number of search nodes, but the *larger* the running time needed at each node.

▷ **Idea:** Encode variable assignments as unary constraints (i.e., for $a(v) = d$, set the unary constraint $D_v = \{d\}$), so that inference reasons about *the network restricted to the commitments already made.*

Backtracking With Inference

▷ **Definition 9.2.9.** The general algorithm for backtracking with inference is

function BacktrackingWithInference(γ,a) **returns** a solution, or ''inconsistent''
  if a is inconsistent **then return** ''inconsistent''
  if a is a total assignment **then return** a
  γ' := a copy of γ /* $\gamma' = (V_{\gamma'}, D_{\gamma'}, C_{\gamma'})$ */

γ' := Inference(γ')
if exists v with $D_v = \emptyset$ **then return** ‘‘inconsistent’’
select some variable v **for** which a is not defined
for each $d \in$ copy of D_v **in** some order **do**
  $a' := a \cup \{v = d\}$; $D_v := \{d\}$ /* makes a explicit as a constraint */
  a'' := BacktrackingWithInference(γ',a')
  if $a'' \neq$ “inconsistent” **then return** a''
return ‘‘inconsistent’’

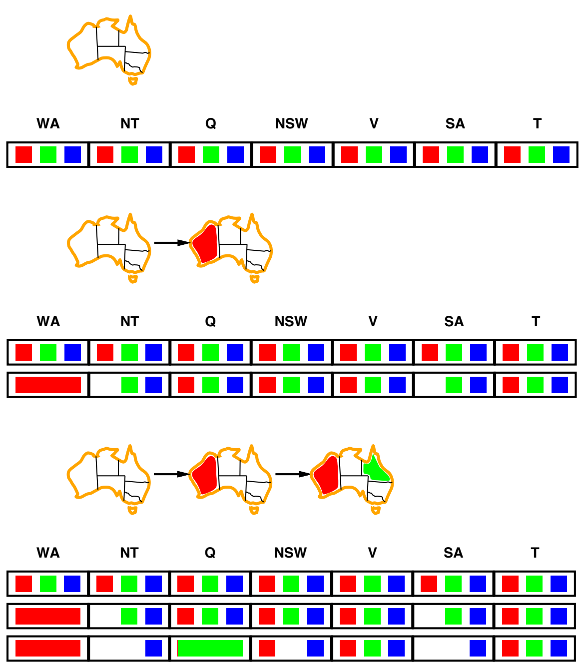
- Exactly the same as Definition 8.4.2, only line 5 new!
- Inference(): Any procedure delivering a (tighter) equivalent network.
- Inference() typically prunes domains; indicate unsolvability by $D_v = \emptyset$.
- When backtracking out of a search branch, retract the inferred constraints: these were dependent on a, the search commitments so far.

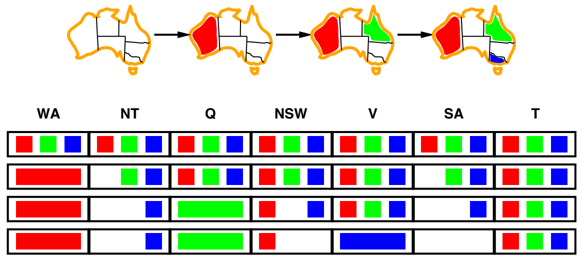
9.3 Forward Checking

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22326`.

Forward Checking

- **Definition 9.3.1.** Forward checking propagates information about illegal values: Whenever a variable u is assigned by a, delete all values inconsistent with $a(u)$ from every D_v for all variables v connected with u by a constraint.
- **Example 9.3.2.** Forward checking in Australia





▷ **Definition 9.3.3 (Inference, Version 1).** Forward checking implemented

function ForwardChecking(γ,a) **returns** modified γ
  for each v where $a(v)=d'$ is defined **do**
    for each u where $a(u)$ is undefined and $C_{uv}\in C$ **do**
      $D_u := \{d\in D_u|(d,d')\in C_{uv}\}$
  return γ

Note: It's a bit strange that we start with d' here; this is to make link to arc consistency – coming up next – as obvious as possible (same notations u, and d vs. v and d').

Forward Checking: Discussion

▷ **Definition 9.3.4.** An inference procedure is called sound, iff for any input γ the output γ' have the same solutions.

▷ **Lemma 9.3.5.** *Forward checking is sound*

Proof sketch: Recall here that the assignment a is represented as unary constraints inside γ.

▷ **Corollary 9.3.6.** *γ and γ' are equivalent.*

▷ Incremental computation: Instead of the first for loop in 0Definition 9.3.3, use only the inner one every time a new assignment $a(v)=d'$ is added.

▷ **Practical Properties:**

- ▷ Cheap but useful inference method.
- ▷ Rarely a good idea to not use forward checking (or a stronger inference method subsuming it).

▷ **Up next:** A stronger inference method (subsuming forward checking).

▷ **Definition 9.3.7.** Let p and q be inference procedures, then p subsumes q, if $p(\gamma)\sqsubseteq q(\gamma)$ for any input γ.

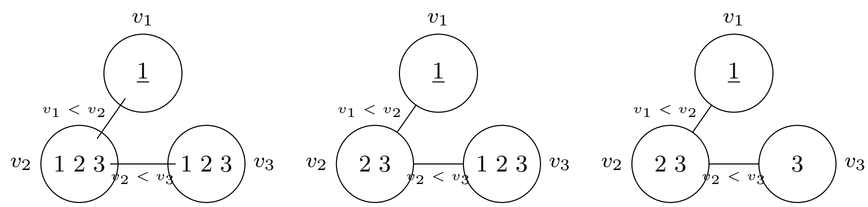
9.4 Arc Consistency

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/22350` and `https://fau.tv/clip/id/22351`.

When Forward Checking is Not Good Enough I

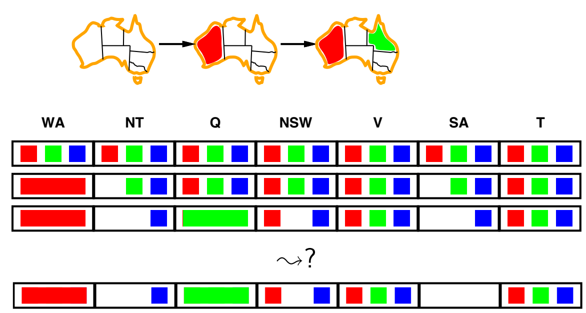
▷ **Problem:** Forward checking makes inferences only from assigned to unassigned variables.

▷ **Example 9.4.1.**



When Forward Checking is Not Good Enough II

▷ **Example 9.4.2.**



Arc Consistency: Definition

▷ **Definition 9.4.3 (arc consistency).** let $\gamma:=\langle V, D, C\rangle$ be a constraint network.

1. A variable $u\in V$ is arc consistent relative to another variable $v\in V$ if either $C_{uv}\notin C$, or for every value $d\in D_u$ there exists a value $d'\in D_v$ such that $(d,d')\in C_{uv}$.
2. The constraint network γ is arc consistent if every variable $u\in V$ is arc consistent relative to every other variable $v\in V$.

▷ **Intuition:** Arc consistency $\widehat{=}$ for every domain value and constraint, at least one value on the other side of the constraint "works".

▷ **Note** the asymmetry between u and v: arc consistency is directed.

▷ **Example 9.4.4 (Arc Consistency (previous slide)).**

- ▷ **Question**: On top, middle, is v_3 arc consistent relative to v_2?

- **Answer**: No. For values 1 and 2, D_{v_2} does not have a value that works.
- **Note**: Enforcing arc consistency for one variable may lead to further reductions on another variable!
- **Question**: And on the right?
- **Anser**: Yes. (But v_2 is not arc consistent relative to v_3)
- **Note**: SA is not arc consistent relative to NT in Example 9.4.2, 3rd row.

Enforcing Arc Consistency: General Remarks

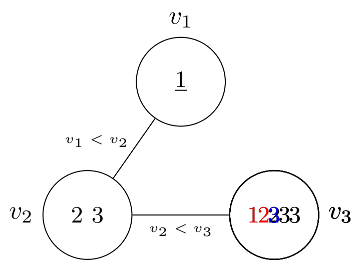
- **Inference, version 2:** "Enforcing Arc Consistency" = removing domain values until γ is arc consistent. (Up next)
- **Note:** Assuming such an inference method AC(γ)
- **Lemma 9.4.5.** *AC(γ) is sound: guarantees to deliver an equivalent network.*
- *Proof sketch:* If, for $d \in D_u$, there does not exist a value $d' \in D_v$ such that $(d,d') \in C_{uv}$, then $u = d$ cannot be part of any solution.
- **Observation 9.4.6.** *AC(γ) subsumes forward checking: AC(γ)$\sqsubseteq$ForwardChecking(γ).*
- *Proof:* Recall from slide 276 that $\gamma' \sqsubseteq \gamma$ means γ' is tighter than γ.
  1. Forward checking removes d from D_u only if there is a constraint C_{uv} such that $D_v = \{d'\}$ (i.e. when v was assigned the value d'), and $(d,d') \notin C_{uv}$.
  2. Clearly, enforcing arc consistency of u relative to v removes d from D_u as well.

Enforcing Arc Consistency for *One* Pair of Variables

- **Definition 9.4.7 (Revise).** An algorithm enforcing arc consistency of u relative to v

```
function Revise(γ,u,v) returns modified γ
  for each d∈D_u do
    if there is no d'∈D_v with (d,d')∈C_uv then D_u := D_u\{d}
  return γ
```

- **Lemma 9.4.8.** *If d is maximal domain size in γ and the test "$(d,d') \in C_{uv}$?" has running time $\mathcal{O}(1)$, then the running time of Revise(γ,u,v) is $\mathcal{O}(d^2)$.*
- **Example 9.4.9.** Revise(γ, v_3, v_2)



AC-1: Enforcing Arc Consistency (Version 1)

▷ **Idea:** Apply Revise pairwise up to a fixed point.

▷ **Definition 9.4.10.2** AC-1 enforces arc consistency in constraint networks:

```
function AC−1(γ) returns modified γ
  repeat
    changesMade := False
    for each constraint C_u0 v do
      Revise(γ,u,v) /∗ if D_u reduces, set changesMade := True ∗/
      Revise(γ,v,u) /∗ if D_v reduces, set changesMade := True ∗/
  until changesMade = False
  return γ
```

▷ **Observation:** Obviously, this does indeed enforce arc consistency for γ.

▷ **Lemma 9.4.11.** *If γ has n variables, m constraints, and maximal domain size d, then the running time of AC1(γ) is $\mathcal{O}(md^2nd)$.*

▷ *Proof sketch:* $\mathcal{O}(md^2)$ for each inner loop, fixed point reached at the latest once all nd variable values have been removed.

▷ **Problem:** There are redundant computations.

▷ **Question:** Do you see what these redundant computations are?

▷ **Redundant computations:** u and v are revised even if theirdomains haven't changed since the last time.

▷ Better algorithm avoiding this: AC 3 (coming up)

AC-3: Enforcing Arc Consistency (Version 3)

▷ **Idea:** Remember the potentially inconsistent variable pairs.

▷ **Definition 9.4.12.** AC-3 optimizes AC-1 for enforcing arc consistency.

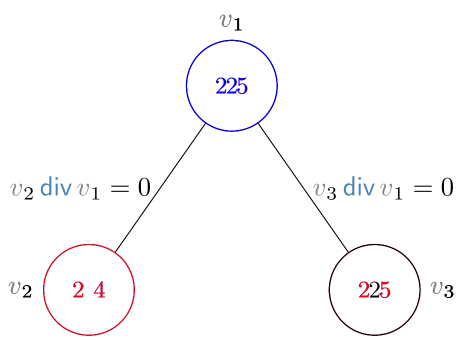
```
function AC−3(γ) returns modified γ
  M := ∅
  for each constraint C_uv∈C do
```

$M := M \cup \{(u,v), (v,u)\}$
while $M \neq \emptyset$ **do**
remove any element (u,v) from M
Revise(γ, u, v)
if D_u has changed **in** the call **to** Revise **then**
for each constraint $C_{wu} \in C$ where $w \neq v$ **do**
$M := M \cup \{(w,u)\}$
return γ

- ▷ **Question:** AC-3(γ) enforces arc consistency because?
- ▷ **Answer:** At any time during the while-loop, if $(u,v) \notin M$ then u is arc consistent relative to v.
- ▷ **Question:** Why only "where $w \neq v$"?
- ▷ **Answer:** If $w = v$ is the reason why D_u changed, then w is still arc consistent relative to u: the values just removed from D_u did not match any values from D_w anyway.

AC-3: Example

- ▷ **Example 9.4.13.** $y \operatorname{div} x = 0$: y modulo x is 0, i.e., y is divisible by x



M
(v_2,v_1)
(v_1,v_2)
(v_3,v_1)
(v_1,v_3)

M
(v_2,v_1)
(v_1,v_2)
(v_3,v_1)
(v_1,v_3)

M
(v_2,v_1)
(v_1,v_2)
(v_3,v_1)

M
(v_2,v_1)
(v_1,v_2)

M
(v_2,v_1)

M
(v_2,v_1)
(v_3,v_1)

M
(v_2,v_1)
(v_3,v_1)

M
(v_2,v_1)

M

AC-3: Runtime

▷ **Theorem 9.4.14 (Runtime of AC-3).** *Let $\gamma{:=}\langle V, D, C\rangle$ be a constraint network with m constraints, and maximal domain size d. Then AC-3(γ) runs in time $\mathcal{O}(md^3)$.*

▷ *Proof:* by counting how often Revise is called.

1. Each call to Revise(γ, u, v) takes time $\mathcal{O}(d^2)$ so it suffices to prove that at most $\mathcal{O}(md)$ of these calls are made.
2. The number of calls to Revise(γ, u, v) is the number of iterations of the while-loop, which is at most the number of insertions into M.
3. Consider any constraint C_{uv}.
4. Two variable pairs corresponding to C_{uv} are inserted in the for-loop. In the while loop, if a pair corresponding to C_{uv} is inserted into M, then
5. beforehand the domain of either u or v was reduced, which happens at most $2d$ times.
6. Thus we have $\mathcal{O}(d)$ insertions per constraint, and $\mathcal{O}(md)$ insertions overall, as desired.

9.5 Decomposition: Constraint Graphs, and Three Simple Cases

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22353`.

Reminder: The Big Picture

▷ Say γ is a constraint network with n variables and maximal domain size d.

- ▷ d^n total assignments must be tested in the worst case to solve γ.

▷ **Inference:** One method to try to avoid/ameliorate this explosion in practice.

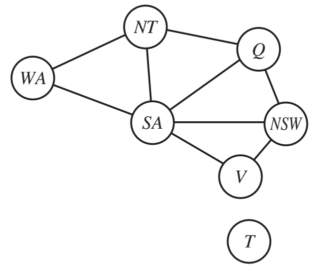
- ▷ Often, from an assignment to some variables, we can easily make inferences regarding other variables.

▷ **Decomposition:** Another method to avoid/ameliorate this explosion in practice.

- ▷ Often, we can exploit the *structure* of a network to *decompose* it into smaller parts that are easier to solve.
- ▷ **Question**: What is "structure", and how to "decompose"?

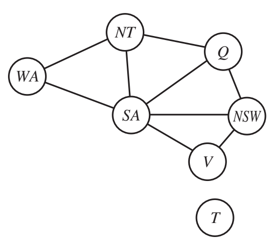
Problem structure

- ▷ Tasmania and mainland are "independent subproblems"
- ▷ **Definition 9.5.1.** Independent subproblems are identified as connected components of constraint graphs.
- ▷ Suppose each subproblem has c variables out of n total
- ▷ Worst-case solution cost is $n \operatorname{div} c \cdot d^c$ (linear in n)
- ▷ E.g., $n = 80$, $d = 2$, $c = 20$
  - ▷ $2^{80} \mathrel{\widehat{=}}$ 4 billion years at 10 million nodes/sec
  - ▷ $42^{20} \mathrel{\widehat{=}}$ 0.4 seconds at 10 million nodes/sec



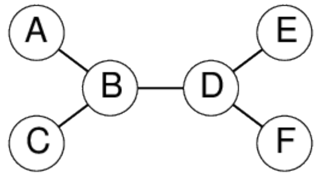
"Decomposition" 1.0: Disconnected Constraint Graphs

- ▷ **Theorem 9.5.2 (Disconnected Constraint Graphs).** *Let $\gamma{:=}\langle V, D, C\rangle$ be a constraint network. Let a_i be a solution to each connected component γ_i of the constraint graph of γ. Then $a{:=}\bigcup_i a_i$ is a solution to γ.*
- ▷ *Proof:*
  1. a satisfies all C_{uv} where u and v are inside the same connected component.
  2. The latter is the case for all C_{uv}.
  3. If two parts of γ are not connected, then they are independent.
- ▷ **Example 9.5.3.** Color Tasmania separately in Australia



- ▷ **Example 9.5.4 (Doing the Numbers).**
  - ▷ γ with $n = 40$ variables, each domain size $k = 2$. Four separate connected components each of size 10.
  - ▷ Reduction of worst-case when using decomposition:
    - ▷ No decomposition: 2^{40}. With: $4 \cdot 2^{10}$. Gain: $2^{28} \approx 280.000.000$.

Tree-structured CSPs



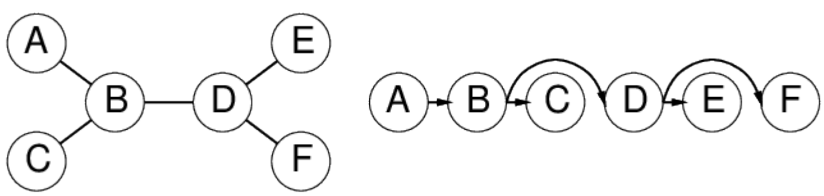
▷ **Theorem 9.5.5.** *If the constraint graph has no cycles, the CSP can be solved in $\mathcal{O}(nd^2)$ time.*

▷ Compare to general CSPs, where worst case time is $\mathcal{O}(d^n)$.

▷ This property also applies to logical and probabilistic reasoning: an important example of the relation between syntactic restrictions and the complexity of reasoning.

Algorithm for tree-structured CSPs

1. Choose a variable as root, order variables from root to leaves such that every node's parent precedes it in the ordering



2. For j from n down to 2, apply

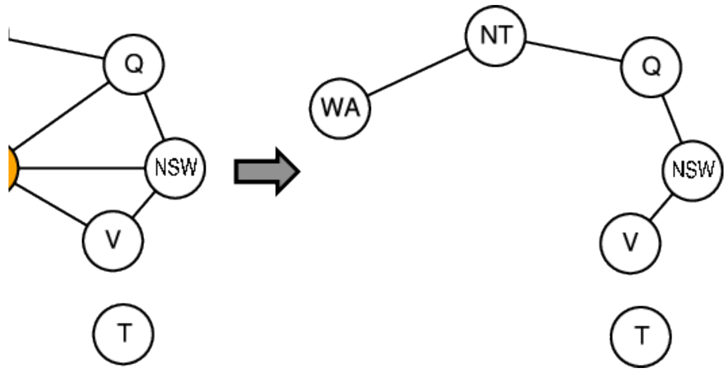
   RemoveInconsistent(Parent(X_j,X_j)

3. For j from 1 to n, assign X_j consistently with $Parent(X_j)$

Nearly tree-structured CSPs

▷ **Definition 9.5.6.** Conditioning: instantiate a variable, prune its neighbors'domains.

▷ **Example 9.5.7.**



▷ **Definition 9.5.8.** Cutset conditioning: instantiate (in all ways) a set of variables such that the remaining constraint graph is a tree.

▷ Cutset size $c \rightsquigarrow$ running time $\mathcal{O}(d^c(n-c)d^2)$, very fast for small c.

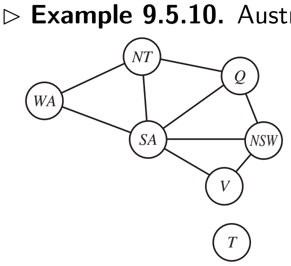
"Decomposition" 2.0: Acyclic Constraint Graphs

▷ **Theorem 9.5.9 (Acyclic Constraint Graphs).** *Let $\gamma{:=}\langle V, D, C\rangle$ be a constraint network with n variables and maximal domain size k, whose constraint graph is acyclic. Then we can find a solution for γ, or prove γ to be unsatisfiable, in time $\mathcal{O}(nk^2)$.*

▷ *Proof sketch:* See the algorithm on the next slide

▷ Constraint networks with acyclic constraint graphs can be solved in (low order) PTIMEpolynomial time.

▷ **Example 9.5.10.** Australia is not acyclic. (But see next section)



▷ **Example 9.5.11 (Doing the Numbers).**

- ▷ γ with $n = 40$ variables, each domain size $k = 2$. Acyclic constraint graph.
- ▷ Reduction of worst-case when using decomposition:
  - ▷ No decomposition: 2^{40}. With decomposition: $40 \cdot 2^2$. Gain: 2^{32}.

Acyclic Constraint Graphs: How To

▷ **Definition 9.5.12.**

Algorithm AcyclicCG(γ):

1. Obtain a directed tree from γ's constraint graph, picking an arbitrary variable v as the root, and directing arcs outwards.[a]
2. Order the variables topologically, i.e., such that each vertex is ordered before its children; denote that order by $v_1, \ldots, v_n$.
3. **for** $i := n, n-1, \ldots, 2$ **do**:
   (a) Revise($\gamma, v_{parent(i)}, v_i$).
   (b) **if** $D_{v_{parent(i)}} = \emptyset$ **then return** "inconsistent"

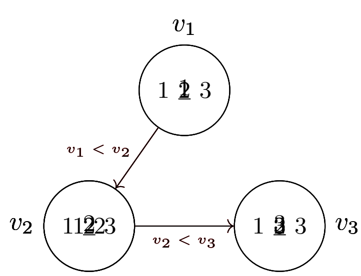
   Now, every variable is arc consistent relative to its children.
4. Run BacktrackingWithInference with forward checking, using the variable order $v_1, \ldots, v_n$.

▷ **Lemma 9.5.13.** *This algorithm will find a solution without ever having to backtrack!*

[a]We assume here that γ's constraint graph is connected. If it is not, do this and the following for each component separately.

AcyclicCG(γ): Example

▷ **Example 9.5.14 (AcyclicCG() execution).**



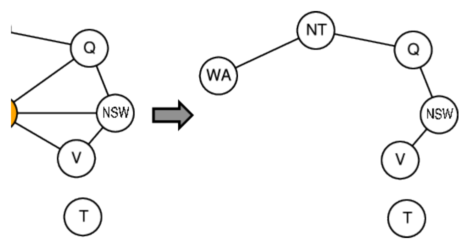
Input network γ. Step 1: Directed tree for root v_1.
Step 2: Order v_1, v_2, v_3.
Step 3: After Revise(γ, v_2, v_3).
Step 3: After Revise(γ, v_1, v_2).
Step 4: After $a(v_1) := 1$ and forward checking.
Step 4: After $a(v_2) := 2$ and forward checking.
Step 4: After $a(v_3) := 3$ (and forward checking).

9.6 Cutset Conditioning

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22354`.

"Almost" Acyclic Constraint Graphs

▷ **Example 9.6.1 (Coloring Australia).**



▷ **Cutset Conditioning: Idea:**

1. Recursive call of backtracking on a s.t. the sub-graph of the constraint graph induced by $\{v{\in}V\,|\,a(v)$ is undefined$\}$ is acyclic.
   - ▷ Then we can solve the remaining sub-problem with AcyclicCG().
2. Choose the variable order so that removing the first d variables renders the constraint graph acyclic.
   - ▷ Then with (1) we won't have to search deeper than d ...!

"Decomposition" 3.0: Cutset Conditioning

▷ **Definition 9.6.2 (Cutset).** Let $\gamma{:=}\langle V, D, C\rangle$ be a constraint network, and $V_0 \subseteq V$. Then V_0 is a cutset for γ if the subgraph of γ's constraint graph induced by $V \backslash V_0$ is acyclic. V_0 is called optimal if its size is minimal among all cutsets for γ.

▷ **Definition 9.6.3.** The cutset conditioning algorithm, computes an optimal cutset, from γ and an existing cutset V_0.

function CutsetConditioning(γ,V_0,a) **returns** a solution, or "inconsistent"
  γ' := a copy of γ; γ' := ForwardChecking(γ',a)
  if ex. v with $D_v = \emptyset$ **then return** "inconsistent"
  if ex. $v{\in}V_0$ s.t. $a(v)$ is undefined **then** select such v
  else a' := AcyclicCG(γ'); **if** $a' \neq$ "inconsistent" **then return** $a \cup a'$ **else return** "inconsistent"
  for each $d{\in}$copy of D_v **in** some order **do**
    a' := $a \cup \{v = d\}$; D_v := $\{d\}$;
    a'' := CutsetConditioning(γ',V_0,a')
  if $a'' \neq$ "inconsistent" **then return** a'' **else return** "inconsistent"

▷ Forward checking is required so that "$a \cup$ AcyclicCG(γ')" is consistent in γ.

▷ **Observation 9.6.4.** *Running time is exponential only in $\#(V_0)$, not in $\#(V)$!*

▷ *Remark 9.6.5.* Finding optimal cutsets is **NP** hard, but approximations exist.

9.7 Constraint Propagation with Local Search

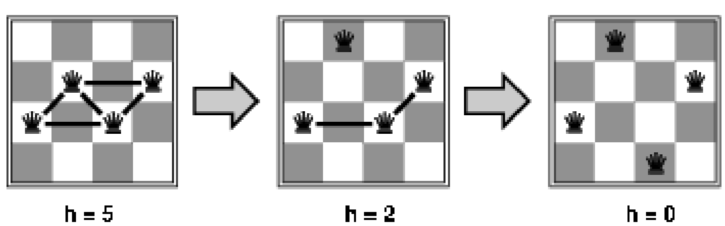
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22355`.

Iterative algorithms for CSPs

- ▷ Local search algorithms like hill climbing and simulated annealing typically work with "complete" states, i.e., all variables are assigned
- ▷ To apply to CSPs: allow states with unsatisfied constraints, actions reassign variable values.
- ▷ **Variable selection:** randomly select any conflicted variable.
- ▷ **Value selection:** by min conflicts heuristic: choose value that violates the fewest constraints i.e., hill climb with $h(n)$:=total number of violated constraints

Example: 4-Queens

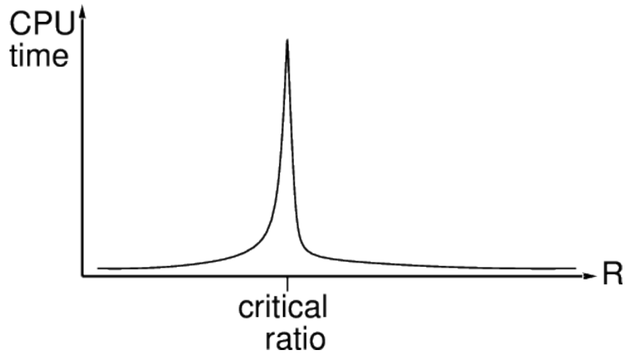
- ▷ States: 4 queens in 4 columns ($4^4 = 256$ states)
- ▷ Actions: move queen in column
- ▷ Goal state: no attacks
- ▷ Heuristic: $h(n) \widehat{=}$ number of attacks



Performance of min-conflicts

- ▷ Given random initial state, can solve n-queens in almost constant time for arbitrary n with high probability (e.g., n = 10,000,000)
- ▷ The same appears to be true for any randomly-generated CSP *except* in a narrow range of the ratio

$$R = \frac{\text{number of constraints}}{\text{number of variables}}$$



9.8 Conclusion & Summary

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22356`.

Conclusion & Summary

- ▷ γ and γ' are equivalent if they have the same solutions. γ' is tighter than γ if it is more constrained.
- ▷ Inference tightens γ without losing equivalence, during backtracking. This reduces the amount of search needed; that benefit must be traded off against the running time overhead for making the inferences.
- ▷ Forward checking removes values conflicting with an assignment already made.
- ▷ Arc consistency removes values that do not comply with any value still available at the other end of a constraint. This subsumes forward checking.
- ▷ The constraint graph captures the dependencies between variables. Separate connected components can be solved independently. Networks with acyclic constraint graphs can be solved in low order polynomial time.
- ▷ A cutset is a subset of variables removing which renders the constraint graph acyclic. Cutset decomposition backtracks only on such a cutset, and solves a sub-problem with acyclic constraint graph at each search leaf.

Topics We Didn't Cover Here

- ▷ **Path consistency**, **k-consistence**: Generalizes arc consistency to size k subsets of variables. Path consistency $\widehat{=}$ 3-consistency.
- ▷ **Tree decomposition**: Instead of instantiating variables until the leaf nodes are trees, distribute the variables and constraints over sub CSPs whose connections form a tree.
- ▷ **Backjumping**: Like backtracking, but with ability to back up *across several levels*

(to a previous assignment identified to be responsible for failure).

▷ **No-Good Learning**: Inferring additional constraints based on information gathered during backtracking.

▷ **Local search**: In space of total (but not necessarily consistent) assignments. (E.g., 8 Queens in chapter 6)

▷ **Tractable CSP**: Classes of CSPs that can be solved in **P**.

▷ **Global Constraints**: Constraints over many/all variables, with associated specialized inference methods.

▷ **Constraint Optimization Problems (COP)**: Utility function over solutions, need an optimal one.

Suggested Reading:

- *Chapter 6: Constraint Satisfaction Problems* in [RN09], in particular Sections 6.2, 6.3.2, and 6.5.
  - Compared to our treatment of the topic "Constraint Satisfaction Problems" (chapter 8 and chapter 9), RN covers much more material, but less formally and in much less detail (in particular, my slides contain many additional in-depth examples). Nice background/additional reading, can't replace the lecture.
  - Section 6.3.2: Somewhat comparable to my "Inference" (except that equivalence and tightness are not made explicit in RN) together with "Forward Checking".
  - Section 6.2: Similar to my "Arc Consistency", less/different examples, much less detail, additional discussion of path consistency and global constraints.
  - Section 6.5: Similar to my "Decomposition: Constraint Graphs, and Two Simple Cases" and "Cutset Conditioning", less/different examples, much less detail, additional discussion of tree decomposition.

Part III

Knowledge and Inference

A Video Nugget covering this part can be found at `https://fau.tv/clip/id/22466`.

This part of the course introduces representation languages and inference methods for structured state representations for agents: In contrast to the atomic and factored state representations from Part II, we look at state representations where the relations between objects are not determined by the problem statement, but can be determined by inference-based methods, where the knowledge about the environment is represented in a formal langauge and new knowledge is derived by transforming expressions of this language.

We look at propositional logic – a rather weak representation langauge – and first-order logic – a much stronger one – and study the respective inference procedures. In the end we show that computation in Prolog is just an inference problem as well.

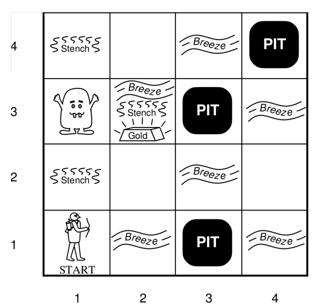
Chapter 10

Propositional Logic & Reasoning, Part I: Principles

10.1 Introduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/22455`.

The Wumpus World

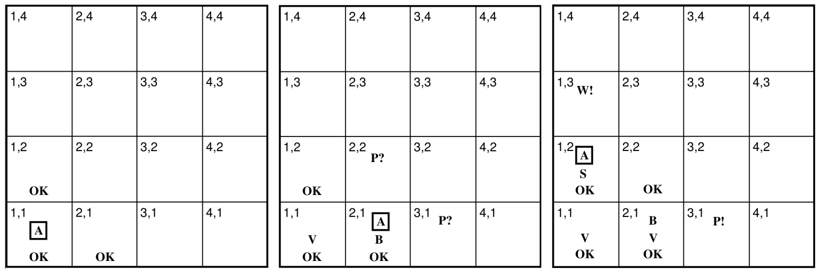


Definition 10.1.1. The Wumpus world is a simple game where an agent explores a cave with 16 cells that can contain pits, gold, and the Wumpus with the goal of getting back out alive with the gold.

- ▷ **Definition 10.1.2 (Actions).** The agent can perform the following actions: goForward, turnRight (by 90°), turnLeft (by 90°), shoot arrow in direction you're facing (you got exactly one arrow), grab an object in current cell, leave cave if you're in cell $[1,1]$.
- ▷ **Definition 10.1.3 (Initial and Terminal States).** Initially, the agent is in cell $[1,1]$ facing east. If the agent falls down a pit or meets live Wumpus it dies.
- ▷ **Definition 10.1.4 (Percepts).** The agent can experience the following percepts: stench, breeze, glitter, bump, scream, none.
  - ▷ Cell adjacent (i.e. north, south, west, east) to Wumpus: stench (else: none).
  - ▷ Cell adjacent to pit: breeze (else: none).
  - ▷ Cell that contains gold: glitter (else: none).
  - ▷ You walk into a wall: bump (else: none).
  - ▷ Wumpus shot by arrow: scream (else: none).

Reasoning in the Wumpus World

▷ **Example 10.1.5 (Reasoning in the Wumpus World).** **A**: agent, **V**: visited, **OK**: safe, **P**: pit, **W**: Wumpus, **B**: breeze, **S**: stench, **G**: gold.



(1) Initial state (2) One step to right (3) Back, and up to [1,2]

▷ *The Wumpus is in [1,3]*! How do we know?

▷ No stench in [2,1], so the stench in [1,2] can only come from [1,3].

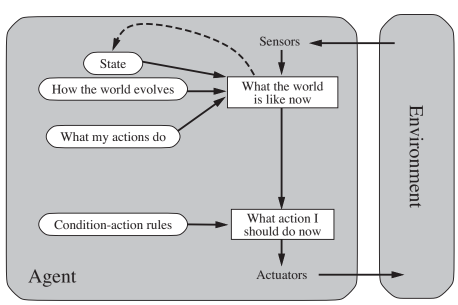
▷ *There's a pit in [3,1]*! How do we know?

▷ No breeze in [1,2], so the breeze in [2,1] can only come from [3,1].

Agents that Think Rationally

▷ **Idea:** Think Before You Act!
"Thinking" = Inference about knowledge represented using logic.

▷ **Definition 10.1.6.** A logic-based agent is a model-based agent that represents the world state as a logical formula and uses inference to think about the state of the environment and its own actions.



function KB−AGENT ($percept$) **returns** an action
  persistent: KB, a knowledge base
    t, a counter, initially 0, indicating time
  TELL(KB, MAKE−PERCEPT−SENTENCE($percept$,t))
  $action$:= ASK(KB, MAKE−ACTION−QUERY(t))
  TELL(KB, MAKE−ACTION−SENTENCE($action$,t))
  t := t+1
  return $action$

Logic: Basic Concepts (Representing Knowledge)

- **Definition 10.1.7.** Syntax: What are legal statements (formulae) $\mathbf{A}$ in the logic?
- **Example 10.1.8.** "W" and "$W \Rightarrow S$". ($W \widehat{=}$ *Wumpus is here*, $S \widehat{=}$ *it stinks*)
- **Definition 10.1.9.** Semantics: Which formulas $\mathbf{A}$ are true under which assignment φ, written $\varphi \models \mathbf{A}$?
- **Example 10.1.10.** If $\varphi := \{W \mapsto \mathsf{T}, S \mapsto \mathsf{F}\}$, then $\varphi \models W$ but $\varphi \not\models W \Rightarrow S$.
- **Intuition:** Knowledge about the state of the world is described by formulae, interpretations evaluate them in the current world (they should turn out true!)

Logic: Basic Concepts (Reasoning about Knowledge)

- **Definition 10.1.11.** Entailment: Which $\mathbf{B}$ are entailed by $\mathbf{A}$, written $\mathbf{A} \models \mathbf{B}$, meaning that, for all φ with $\varphi \models \mathbf{A}$, we have $\varphi \models \mathbf{B}$? E.g., $P \wedge (P \Rightarrow Q) \models Q$.
- **Intuition:** Entailment $\widehat{=}$ ideal outcome of reasoning, everything that we can possibly conclude. e.g. determine Wumpus position as soon as we have enough information
- **Definition 10.1.12.** Deduction: Which statements $\mathbf{B}$ can be derived from $\mathbf{A}$ using a set $\mathcal{C}$ of inference rules (a calculus), written $\mathbf{A} \vdash_{\mathcal{C}} \mathbf{B}$?
- **Example 10.1.13.** If $\mathcal{C}$ contains $\dfrac{\mathbf{A} \quad \mathbf{A} \Rightarrow \mathbf{B}}{\mathbf{B}}$ then $P, P \Rightarrow Q \vdash_{\mathcal{C}} Q$
- **Intuition:** Deduction $\widehat{=}$ process in an actual computer trying to reason about entailment. E.g. a mechanical process attempting to determine Wumpus position.
- **Definition 10.1.14.** Soundness: whenever $\mathbf{A} \vdash_{\mathcal{C}} \mathbf{B}$, we also have $\mathbf{A} \models \mathbf{B}$.
- **Definition 10.1.15.** Completeness: whenever $\mathbf{A} \models \mathbf{B}$, we also have $\mathbf{A} \vdash_{\mathcal{C}} \mathbf{B}$.

General Problem Solving using Logic

- **Idea:** Any problem that can be formulated as reasoning about logic. $\rightsquigarrow$ use off-the-shelf reasoning tool.
- *Very* successful using propositional logic and modern SAT solvers! (Propositional satisfiability testing; chapter 13)

Propositional Logic and Its Applications

- Propositional logic = canonical form of knowledge + reasoning.
  - Syntax: Atomic propositions that can be either true or false, connected by "and, or, not".
  - Semantics: Assign value to every proposition, evaluate connectives.
- **Applications:** Despite its simplicity, widely applied!
  - **Product configuration** (e.g., Mercedes). Check consistency of customized combinations of components.
  - **Hardware verification** (e.g., Intel, AMD, IBM, Infineon). Check whether a circuit has a desired property p.
  - **Software verification**: Similar.
  - **CSP applications**: propositional logic can be (successfully!) used to formulate and solve constraint satisfaction problems. (see chapter 8)
- chapter 9 gives an example for verification.

Our Agenda for This Topic

- **This section**: Basic definitions and concepts; tableaux, resolution.
  - Sets up the framework. Resolution is the quintessential reasoning procedure underlying most successful SAT solvers.
- **chapter 13**: The Davis Putnam procedure and clause learning; practical problem structure.
  - State-of-the-art algorithms for reasoning about propositional logic, and an important observation about how they behave.

Our Agenda for This Chapter

- Propositional logic: What's the syntax and semantics? How can we capture deduction?
  - We study this logic formally.
- **Tableaux, Resolution**: How can we make deduction mechanizable? What are its properties?
  - Formally introduces the most basic machine-oriented reasoning methods.
- **Killing a Wumpus**: How can we use all this to figure out where the Wumpus is?

▷ Coming back to our introductory example.

10.2 Propositional Logic (Syntax/Semantics)

Video Nuggets covering this section can be found at https://fau.tv/clip/id/22457 and https://fau.tv/clip/id/22458.

Propositional Logic (Syntax)

▷ **Definition 10.2.1 (Syntax).** The formulae of propositional logic (write PL^0) are made up from

- ▷ propositional variables: $\mathcal{V}_0 := \{P, Q, R, P^1, P^2, \ldots\}$ (countably infinite)
- ▷ constants/constructors called connectives: $\Sigma_0 := \{T, F, \neg, \vee, \wedge, \Rightarrow, \Leftrightarrow, \ldots\}$

We define the set $wff_0(\mathcal{V}_0)$ of well-formed propositional formula (wffs) as

- ▷ propositional variables,
- ▷ the logical constants T and F,
- ▷ negations $\neg\mathbf{A}$,
- ▷ conjunctions $\mathbf{A} \wedge \mathbf{B}$($\mathbf{A}$ and $\mathbf{B}$ are called conjuncts),
- ▷ disjunctions $\mathbf{A} \vee \mathbf{B}$ ($\mathbf{A}$ and $\mathbf{B}$ are called disjuncts),
- ▷ implications $\mathbf{A} \Rightarrow \mathbf{B}$, and
- ▷ equivalences (or biimplication). $\mathbf{A} \Leftrightarrow \mathbf{B}$,

where $\mathbf{A}, \mathbf{B} \in wff_0(\mathcal{V}_0)$ themselves.

▷ **Example 10.2.2.** $P \wedge Q, P \vee Q, (\neg P \vee Q) \Leftrightarrow (P \Rightarrow Q) \in wff_0(\mathcal{V}_0)$

▷ **Definition 10.2.3.** Propositional formulae without connectives are called atomic (or an atom) and complex otherwise.

Propositional Logic Grammar Overview

▷ **Grammar for Propositional Logic:**

propositional variables	X	::=	$\mathcal{V}_0 = \{P, Q, R, \ldots, \ldots\}$	variables
propositional formulae	$\mathbf{A}$	::=	X	variable
		|	$\neg\mathbf{A}$	negation
		|	$\mathbf{A}_1 \wedge \mathbf{A}_2$	conjunction
		|	$\mathbf{A}_1 \vee \mathbf{A}_2$	disjunction
		|	$\mathbf{A}_1 \Rightarrow \mathbf{A}_2$	implication
		|	$\mathbf{A}_1 \Leftrightarrow \mathbf{A}_2$	equivalence

Alternative Notations for Connectives

Here	Elsewhere		
$\neg\mathbf{A}$	$\sim\mathbf{A}$	$\overline{\mathbf{A}}$	
$\mathbf{A}\wedge\mathbf{B}$	$\mathbf{A}\,\&\,\mathbf{B}$	$\mathbf{A}\bullet\mathbf{B}$	$\mathbf{A},\mathbf{B}$
$\mathbf{A}\vee\mathbf{B}$	$\mathbf{A}+\mathbf{B}$	$\mathbf{A}\,\vert\,\mathbf{B}$	$\mathbf{A};\mathbf{B}$
$\mathbf{A}\Rightarrow\mathbf{B}$	$\mathbf{A}\rightarrow\mathbf{B}$	$\mathbf{A}\supset\mathbf{B}$	
$\mathbf{A}\Leftrightarrow\mathbf{B}$	$\mathbf{A}\leftrightarrow\mathbf{B}$	$\mathbf{A}\equiv\mathbf{B}$	
F	$\bot$	0	
T	$\top$	1	

Semantics of PL^0 (Models)

- **Definition 10.2.4.** A model $\mathcal{M}:=\langle\mathcal{D}_o,\mathcal{I}\rangle$ for propositional logic consists of
  - the universe $\mathcal{D}_o=\{\mathsf{T},\mathsf{F}\}$
  - the interpretation $\mathcal{I}$ that assigns values to essential connectives.
  - $\mathcal{I}(\neg)\colon\mathcal{D}_o\to\mathcal{D}_o;\mathsf{T}\mapsto\mathsf{F},\mathsf{F}\mapsto\mathsf{T}$
  - $\mathcal{I}(\wedge)\colon\mathcal{D}_o\times\mathcal{D}_o\to\mathcal{D}_o;\langle\alpha,\beta\rangle\mapsto\mathsf{T}$, iff $\alpha=\beta=\mathsf{T}$

  We call a constructor a logical constant, iff its value is fixed by the interpretation
- Treat the other connectives as abbreviations, e.g. $\mathbf{A}\vee\mathbf{B}\widehat{=}\neg(\neg\mathbf{A}\wedge\neg\mathbf{B})$ and $\mathbf{A}\Rightarrow\mathbf{B}\widehat{=}\neg\mathbf{A}\vee\mathbf{B}$, and $T\widehat{=}P\vee\neg P$ (only need to treat $\neg,\wedge$ directly)

Semantics of PL^0 (Evaluation)

- **Problem:** The interpretation function only assigns meaning to connectives.
- **Definition 10.2.5.** A variable assignment $\varphi\colon\mathcal{V}_0\to\mathcal{D}_o$ assigns values to propositional variables.
- **Definition 10.2.6.** The value function $\mathcal{I}_\varphi\colon\mathit{wff}_0(\mathcal{V}_0)\to\mathcal{D}_o$ assigns values to PL^0 formulae. It is recursively defined,
  - $\mathcal{I}_\varphi(P)=\varphi(P)$ (base case)
  - $\mathcal{I}_\varphi(\neg\mathbf{A})=\mathcal{I}(\neg)(\mathcal{I}_\varphi(\mathbf{A}))$.
  - $\mathcal{I}_\varphi(\mathbf{A}\wedge\mathbf{B})=\mathcal{I}(\wedge)(\mathcal{I}_\varphi(\mathbf{A}),\mathcal{I}_\varphi(\mathbf{B}))$.

▷ Note that $\mathcal{I}_\varphi(\mathbf{A} \vee \mathbf{B}) = \mathcal{I}_\varphi(\neg(\neg\mathbf{A} \wedge \neg\mathbf{B}))$ is only determined by $\mathcal{I}_\varphi(\mathbf{A})$ and $\mathcal{I}_\varphi(\mathbf{B})$, so we think of the defined connectives as logical constants as well.

▷ **Definition 10.2.7.** Two formulae $\mathbf{A}$ and $\mathbf{B}$ are called equivalent, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathcal{I}_\varphi(\mathbf{B})$ for all variable assignments φ.

Computing Semantics

▷ **Example 10.2.8.** Let $\varphi := [\mathsf{T}/P_1], [\mathsf{F}/P_2], [\mathsf{T}/P_3], [\mathsf{F}/P_4], \ldots$ then

$$\begin{aligned}
& \mathcal{I}_\varphi(P_1 \vee P_2 \vee \neg(\neg P_1 \wedge P_2) \vee P_3 \wedge P_4) \\
= \; & \mathcal{I}(\vee)(\mathcal{I}_\varphi(P_1 \vee P_2), \mathcal{I}_\varphi(\neg(\neg P_1 \wedge P_2) \vee P_3 \wedge P_4)) \\
= \; & \mathcal{I}(\vee)(\mathcal{I}(\vee)(\mathcal{I}_\varphi(P_1), \mathcal{I}_\varphi(P_2)), \mathcal{I}(\vee)(\mathcal{I}_\varphi(\neg(\neg P_1 \wedge P_2)), \mathcal{I}_\varphi(P_3 \wedge P_4))) \\
= \; & \mathcal{I}(\vee)(\mathcal{I}(\vee)(\varphi(P_1), \varphi(P_2)), \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathcal{I}_\varphi(\neg P_1 \wedge P_2)), \mathcal{I}(\wedge)(\mathcal{I}_\varphi(P_3), \mathcal{I}_\varphi(P_4)))) \\
= \; & \mathcal{I}(\vee)(\mathcal{I}(\vee)(\mathsf{T}, \mathsf{F}), \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathcal{I}(\wedge)(\mathcal{I}_\varphi(\neg P_1), \mathcal{I}_\varphi(P_2))), \mathcal{I}(\wedge)(\varphi(P_3), \varphi(P_4)))) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathcal{I}(\wedge)(\mathcal{I}(\neg)(\mathcal{I}_\varphi(P_1)), \varphi(P_2))), \mathcal{I}(\wedge)(\mathsf{T}, \mathsf{F}))) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathcal{I}(\wedge)(\mathcal{I}(\neg)(\varphi(P_1)), \mathsf{F})), \mathsf{F})) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathcal{I}(\wedge)(\mathcal{I}(\neg)(\mathsf{T}), \mathsf{F})), \mathsf{F})) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathcal{I}(\wedge)(\mathsf{F}, \mathsf{F})), \mathsf{F})) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathcal{I}(\vee)(\mathcal{I}(\neg)(\mathsf{F}), \mathsf{F})) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathcal{I}(\vee)(\mathsf{T}, \mathsf{F})) \\
= \; & \mathcal{I}(\vee)(\mathsf{T}, \mathsf{T}) \\
= \; & \mathsf{T}
\end{aligned}$$

▷ What a mess!

Now we will also review some propositional identities that will be useful later on. Some of them we have already seen, and some are new. All of them can be proven by simple truth table arguments.

Propositional Identities

▷ We have the following identities in propositional logic:

Name	for $\wedge$	for $\vee$
Idenpotence	$\varphi \wedge \varphi = \varphi$	$\varphi \vee \varphi = \varphi$
Identity	$\varphi \wedge T = \varphi$	$\varphi \vee F = \varphi$
Absorption I	$\varphi \wedge F = F$	$\varphi \vee T = T$
Commutativity	$\varphi \wedge \psi = \psi \wedge \varphi$	$\varphi \vee \psi = \psi \vee \varphi$
Associativity	$\varphi \wedge (\psi \wedge \theta) = (\varphi \wedge \psi) \wedge \theta$	$\varphi \vee (\psi \vee \theta) = (\varphi \vee \psi) \vee \theta$
Distributivity	$\varphi \wedge (\psi \vee \theta) = \varphi \wedge \psi \vee \varphi \wedge \theta$	$\varphi \vee \psi \wedge \theta = (\varphi \vee \psi) \wedge (\varphi \vee \theta)$
Absorption II	$\varphi \wedge (\varphi \vee \theta) = \varphi$	$\varphi \vee \varphi \wedge \theta = \varphi$
De Morgan	$\neg(\varphi \wedge \psi) = \neg\varphi \vee \neg\psi$	$\neg(\varphi \vee \psi) = \neg\varphi \wedge \neg\psi$
Double negation	$\neg\neg\varphi = \varphi$	
Definitions	$\varphi \Rightarrow \psi = \neg\varphi \vee \psi$	$\varphi \Leftrightarrow \psi = (\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$

We will now use the distribution of values of a Boolean expression under all (variable) assignments to characterize them semantically. The intuition here is that we want to understand theorems, examples, counterexamples, and inconsistencies in mathematics and everyday reasoning[1].

The idea is to use the formal language of Boolean expressions as a model for mathematical language. Of course, we cannot express all of mathematics as Boolean expressions, but we can at least study the interplay of mathematical statements (which can be true or false) with the copula "and", "or" and "not".

Semantic Properties of Propositional Formulae

- **Definition 10.2.9.** Let $\mathcal{M}:=\langle\mathcal{U},\mathcal{I}\rangle$ be our model, then we call $\mathbf{A}$
  - true under φ (φ satisfies $\mathbf{A}$) in $\mathcal{M}$, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{T}$ (write $\mathcal{M}\models^\varphi\mathbf{A}$)
  - false under φ (φ falsifies $\mathbf{A}$) in $\mathcal{M}$, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{F}$ (write $\mathcal{M}\not\models^\varphi\mathbf{A}$)
  - satisfiable in $\mathcal{M}$, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{T}$ for some assignment φ
  - valid in $\mathcal{M}$, iff $\mathcal{M}\models^\varphi\mathbf{A}$ for all assignments φ
  - falsifiable in $\mathcal{M}$, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{F}$ for some assignments φ
  - unsatisfiable in $\mathcal{M}$, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{F}$ for all assignments φ
- **Example 10.2.10.** $x \vee x$ is satisfiable and falsifiable.
- **Example 10.2.11.** $x \vee \neg x$ is valid and $x \wedge \neg x$ is unsatisfiable.
- **Alternative Notation:** Write $[\![\mathbf{A}]\!]_\varphi$ for $\mathcal{I}_\varphi(\mathbf{A})$, if $\mathcal{M} = \langle\mathcal{U},\mathcal{I}\rangle$. (and $[\![\mathbf{A}]\!]$, if $\mathbf{A}$ is ground, and $[\![\mathbf{A}]\!]$, if $\mathcal{M}$ is clear)
- **Definition 10.2.12 (Entailment).** (aka. logical consequence)
  We say that $\mathbf{A}$ entails $\mathbf{B}$ ($\mathbf{A}\models\mathbf{B}$), iff $\mathcal{I}_\varphi(\mathbf{B}) = \mathsf{T}$ for all φ with $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{T}$ (i.e. all assignments that make $\mathbf{A}$ true also make $\mathbf{B}$ true)

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 323 2023-09-20

Let us now see how these semantic properties model mathematical practice.

In mathematics we are interested in assertions that are true in all circumstances. In our model of mathematics, we use variable assignments to stand for circumstances. So we are interested in Boolean expressions which are true under all variable assignments; we call them valid. We often give examples (or show situations) which make a conjectured assertion false; we call such examples counterexamples, and such assertions "falsifiable". We also often give examples for certain assertions to show that they can indeed be made true (which is not the same as being valid yet); such assertions we call "satisfiable". Finally, if an assertion cannot be made true in any circumstances we call it "unsatisfiable"; such assertions naturally arise in mathematical practice in the form of refutation proofs, where we show that an assertion (usually the negation of the theorem we want to prove) leads to an obviously unsatisfiable conclusion, showing that the negation of the theorem is unsatisfiable, and thus the theorem valid.

A better mouse-trap: Truth Tables

[1]Here (and elsewhere) we will use mathematics (and the language of mathematics) as a test tube for understanding reasoning, since mathematics has a long history of studying its own reasoning processes and assumptions.

▷ Truth tables visualize truth functions:

$\neg$	
$\top$	F
$\bot$	T

$\wedge$	$\top$	$\bot$
$\top$	T	F
$\bot$	F	F

$\vee$	$\top$	$\bot$
$\top$	T	T
$\bot$	T	F

▷ If we are interested in values for all assignments (e.g $z \wedge x \vee \neg(z \wedge y)$)

assignments			intermediate results			full
x	y	z	$e_1 := z \wedge y$	$e_2 := \neg e_1$	$e_3 := z \wedge x$	$e_3 \vee e_2$
F	F	F	F	T	F	T
F	F	T	F	T	F	T
F	T	F	F	T	F	T
F	T	T	T	F	F	F
T	F	F	F	T	F	T
T	F	T	F	T	T	T
T	T	F	F	T	F	T
T	T	T	T	F	T	T

Hair Color in Propositional Logic

▷ There are three persons, Stefan, Nicole, and Jochen.

1. Their hair colors are black, red, or green.
2. Their study subjects are AI, Physics, or Chinese at least one studies AI.
   (a) Persons with red or green hair do not study AI.
   (b) Neither the Physics nor the Chinese students have black hair.
   (c) Of the two male persons, one studies Physics, and the other studies Chinese.

▷ **Question:** Who studies AI?
(A) Stefan (B) Nicole (C) Jochen (D) Nobody

▷ **Answer:** You can solve this using PL^0, if we accept $bla(S)$, etc. as propositional variables.

We first express what we know: For every $x \in \{S, N, J\}$ (Stefan, Nicole, Jochen) we have

1. $bla(x) \vee red(x) \vee gre(x)$; (note: three formulae)
2. $ai(x) \vee phy(x) \vee chi(x)$ and $ai(S) \vee ai(N) \vee ai(J)$
   (a) $ai(x) \Rightarrow \neg red(x) \wedge \neg gre(x)$.
   (b) $phy(x) \Rightarrow \neg bla(x)$ and $chi(x) \Rightarrow \neg bla(x)$.
   (c) $phy(S) \wedge chi(J) \vee phy(J) \wedge chi(S)$.

Now, we obtain new knowledge via entailment steps:

3. 1. together with 2.1 entails that $ai(x) \Rightarrow bla(x)$ for every $x \in \{S, N, J\}$,
4. thus $\neg bla(S) \wedge \neg bla(J)$ by 3. and 2.2 and
5. so $\neg ai(S) \wedge \neg ai(J)$ by 3. and 4.
6. With 2.3 the latter entails $ai(N)$.

10.3 Predicate Logic Without Quantifiers

In the hair-color example we have seen that we are able to model complex situations in PL^0.

The trick of using variables with fancy names like $bla(N)$ is a bit dubious, and we can already imagine that it will be difficult to support programmatically unless we make names like $bla(N)$ into first class citizens i.e. expressions of the logic language themselves.

Individuals and their Properties/Relations

- **Observation:** We want to talk about individuals like Stefan, Nicole, and Jochen and their properties, e.g. being blond, or studying AI and relationships, e.g. that *Stefan loves Nicole*.
- **Idea:** Re-use PL^0, but replace propositional variables with something more expressive! (instead of fancy variable name trick)
- **Definition 10.3.1.** A first-order signature consists of pairwise disjoint, countable sets for each $k \in \mathbb{N}$
  - function constants: $\Sigma_k^f = \{f, g, h, \ldots\}$ – denoting functions on individuals
  - predicate constants: $\Sigma_k^p = \{p, q, r, \ldots\}$ – denoting relationships among individuals.

  We set $\Sigma^f := \bigcup_{k\in\mathbb{N}} \Sigma_k^f$, $\Sigma^p := \bigcup_{k\in\mathbb{N}} \Sigma_k^p$, and $\Sigma_1 := \Sigma^f \cup \Sigma^p$.
- **Definition 10.3.2.**
  The formulae of PLnq are given by the following grammar

functions	f^k	$\in$	Σ_k^f	
predicates	p^k	$\in$	Σ_k^p	
terms	t	::=	X	variable
		|	f^0	constant
		|	$f^k(t_1, \ldots, t_k)$	application
formulae	$\mathbf{A}$	::=	$p^k(t_1, \ldots, t_k)$	atomic
		|	$\neg\mathbf{A}$	negation
		|	$\mathbf{A}_1 \wedge \mathbf{A}_2$	conjunction

PLNQ Semantics

- **Definition 10.3.3.** Universes $\mathcal{D}_0 = \{\mathsf{T}, \mathsf{F}\}$ of truth values and $\mathcal{D}_\iota \neq \emptyset$ of individuals.
- **Definition 10.3.4.** Interpretation $\mathcal{I}$ assigns values to constants, e.g.
  - $\mathcal{I}(\neg) \colon \mathcal{D}_0 \to \mathcal{D}_0; \mathsf{T} \mapsto \mathsf{F}; \mathsf{F} \mapsto \mathsf{T}$ and $\mathcal{I}(\wedge) = \ldots$ (as in PL^0)
  - $\mathcal{I} \colon \Sigma_0^f \to \mathcal{D}_\iota$ (interpret individual constants as individuals)
  - $\mathcal{I} \colon \Sigma_k^f \to {\mathcal{D}_\iota}^k \to \mathcal{D}_\iota$ (interpret function constants as functions)mo
  - $\mathcal{I} \colon \Sigma_k^p \to \mathcal{P}({\mathcal{D}_\iota}^k)$ (interpret predicates as arbitrary relations)
- **Definition 10.3.5.** The value function $\mathcal{I}$ assigns values to formulae (recursively)

- $\mathcal{I}(f(\mathbf{A}^1, \ldots, \mathbf{A}^k)) := \mathcal{I}(f)(\mathcal{I}(\mathbf{A}^1), \ldots, \mathcal{I}(\mathbf{A}^k))$
- $\mathcal{I}(p(\mathbf{A}^1, \ldots, \mathbf{A}^k)) := \mathsf{T}$, iff $\langle \mathcal{I}(\mathbf{A}^1), \ldots, \mathcal{I}(\mathbf{A}^k) \rangle \in \mathcal{I}(p)$
- $\mathcal{I}(\neg \mathbf{A}) = \mathcal{I}(\neg)(\mathcal{I}(\mathbf{A}))$ and $\mathcal{I}(\mathbf{A} \wedge \mathbf{B}) = \mathcal{I}(\wedge)(\mathcal{I}(\mathbf{A}), \mathcal{I}(\mathbf{G}))$ (just as in PL^0)

▷ **Definition 10.3.6.** Model: $\mathcal{M} = \langle \mathcal{D}_\iota, \mathcal{I} \rangle$ varies in $\mathcal{D}_\iota$ and $\mathcal{I}$.

▷ **Theorem 10.3.7.** *$\mathrm{PL}^{\mathrm{nq}}$ is isomorphic to PL^0* (*interpret atoms as prop. variables*)

A Model for $\mathrm{PL}^{\mathrm{nq}}$

▷ **Example 10.3.8.** Let $L := \{a, b, c, d, e, P, Q, R, S\}$, we set the domain $\mathcal{D} := \{\clubsuit, \spadesuit, \heartsuit, \diamondsuit\}$, and the interpretation function $\mathcal{I}$ by setting

- $a \mapsto \clubsuit$, $b \mapsto \spadesuit$, $c \mapsto \heartsuit$, $d \mapsto \diamondsuit$, and $e \mapsto \diamondsuit$ for individual constants,
- $P \mapsto \{\clubsuit, \spadesuit\}$ and $Q \mapsto \{\spadesuit, \diamondsuit\}$, for unary predicate constants.
- $R \mapsto \{\langle \heartsuit, \diamondsuit \rangle, \langle \diamondsuit, \heartsuit \rangle\}$, and $S \mapsto \{\langle \diamondsuit, \spadesuit \rangle, \langle \spadesuit, \clubsuit \rangle\}$ for binary predicate constants.

▷ **Example 10.3.9 (Computing Meaning in this Model).**

- $\mathcal{I}(R(a,b) \wedge P(c)) = \mathsf{T}$, iff
- $\mathcal{I}(R(a,b)) = \mathsf{T}$ and $\mathcal{I}(P(c)) = \mathsf{T}$, iff
- $\langle \mathcal{I}(a), \mathcal{I}(b) \rangle \in \mathcal{I}(R)$ and $\mathcal{I}(c) \in \mathcal{I}(P)$, iff
- $\langle \clubsuit, \spadesuit \rangle \in \{\langle \heartsuit, \diamondsuit \rangle, \langle \diamondsuit, \heartsuit \rangle\}$ and $\heartsuit \in \{\clubsuit, \spadesuit\}$

So, $\mathcal{I}(R(a,b) \wedge P(c)) = \mathsf{F}$.

$\mathrm{PL}^{\mathrm{nq}}$ and PL^0 are Isomorphic

▷ **Observation:** For every choice of Σ of signature, the set $\mathcal{A}_\Sigma$ of atomic $\mathrm{PL}^{\mathrm{nq}}$ formulae is countable, so there is a $\mathcal{V}_\Sigma \subseteq \mathcal{V}_0$ and a bijection $\theta_\Sigma \colon \mathcal{A}_\Sigma \to \mathcal{V}_\Sigma$.

θ_Σ can be extended to formulae as $\mathrm{PL}^{\mathrm{nq}}$ and PL^0 share connectives.

▷ **Lemma 10.3.10.** *For every model $\mathcal{M} = \langle \mathcal{D}_\iota, \mathcal{I} \rangle$, there is a variable assignment $\varphi_\mathcal{M}$, such that $\mathcal{I}_{\varphi_\mathcal{M}}(\mathbf{A}) = \mathcal{I}(\mathbf{A})$.*

▷ *Proof sketch:* We just define $\varphi_\mathcal{M}(X) := \mathcal{I}(\theta_\Sigma^{-1}(X))$

▷ **Lemma 10.3.11.** *For every variable assignment $\psi \colon \mathcal{V}_\Sigma \to \{\mathsf{T}, \mathsf{F}\}$ there is a model $\mathcal{M}^\psi = \langle \mathcal{D}^\psi, \mathcal{I}^\psi \rangle$, such that $\mathcal{I}_\psi(\mathbf{A}) = \mathcal{I}^\psi(\mathbf{A})$.*

▷ *Proof sketch:* see next slide

▷ **Corollary 10.3.12.** *$\mathrm{PL}^{\mathrm{nq}}$ is isomorphic to PL^0, i.e. the following diagram commutes:*

$$\begin{tikzcd}
\langle \mathcal{D}^\psi, \mathcal{I}^\psi \rangle & \mathcal{V}_\Sigma \to \{\mathsf{T},\mathsf{F}\} \arrow[l, "\psi \mapsto \mathcal{M}^\psi"'] \\
PL^{nq}(\Sigma_1) \arrow[u, "\mathcal{I}^\psi()"] \arrow[r, "\theta_\Sigma"] & PL^0(\mathcal{A}_\Sigma) \arrow[u, "\mathcal{I}_{\varphi_\mathcal{M}}()"']
\end{tikzcd}$$

▷ **Note:** This constellation with a language isomorphism and a corresponding model isomorphism (in converse direction) is typical for a logic isomorphism.

Valuation and Satisfiability

▷ **Lemma 10.3.13.** *For every variable assignment $\psi\colon \mathcal{V}_\Sigma \to \{\mathsf{T},\mathsf{F}\}$ there is a model $\mathcal{M}^\psi = \langle \mathcal{D}^\psi, \mathcal{I}^\psi \rangle$, such that $\mathcal{I}_\psi(\mathbf{A}) = \mathcal{I}^\psi(\mathbf{A})$.*

▷ *Proof:* We construct $\mathcal{M}^\psi = \langle \mathcal{D}^\psi, \mathcal{I}^\psi \rangle$ and show that it works as desired.

1. Let $\mathcal{D}^\psi$ be the set of $\mathrm{PL}^{\mathrm{nq}}$ terms over Σ, and
   - ▹ $\mathcal{I}^\psi(f)\colon {\mathcal{D}_\iota}^k \to {\mathcal{D}^\psi}^k; \langle \mathbf{A}_1, \ldots, \mathbf{A}_k \rangle \mapsto f(\mathbf{A}_1, \ldots, \mathbf{A}_k)$ for $f \in \Sigma^f k$
   - ▹ $\mathcal{I}^\psi(p) := \{\langle \mathbf{A}_1, \ldots, \mathbf{A}_k \rangle \mid \psi(\theta_\psi^{-1} p(\mathbf{A}_1, \ldots, \mathbf{A}_k)) = \mathsf{T}\}$ for $p \in \Sigma^p$.
2. We show $\mathcal{I}^\psi(\mathbf{A}) = \mathbf{A}$ for terms $\mathbf{A}$ by induction on $\mathbf{A}$: If $\mathbf{A} = f(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ then $\mathcal{I}^\psi(\mathbf{A}) = \mathcal{I}^\psi(f)(\mathcal{I}(\mathbf{A}_1), \ldots, \mathcal{I}(\mathbf{A}_n)) = \mathcal{I}^\psi(f)(\mathbf{A}_1, \ldots, \mathbf{A}_k) = \mathbf{A}$.
3. For a $\mathrm{PL}^{\mathrm{nq}}$ formula $\mathbf{A}$ we show that $\mathcal{I}^\psi(\mathbf{A}) = \mathcal{I}_\psi(\mathbf{A})$ by induction on $\mathbf{A}$.
   - 3.1. If $\mathbf{A} = p(\mathbf{A}_1, \ldots, \mathbf{A}_k)$, then $\mathcal{I}^\psi(\mathbf{A}) = \mathcal{I}^\psi(p)(\mathcal{I}(\mathbf{A}_1), \ldots, \mathcal{I}(\mathbf{A}_n)) = \mathsf{T}$, iff $\langle \mathbf{A}_1, \ldots, \mathbf{A}_k \rangle \in \mathcal{I}^\psi(p)$, iff $\psi(\theta_\psi^{-1}\mathbf{A}) = \mathsf{T}$, so $\mathcal{I}^\psi(\mathbf{A}) = \mathcal{I}_\psi(\mathbf{A})$ as desired.
   - 3.2. If $\mathbf{A} = \neg\mathbf{B}$, then $\mathcal{I}^\psi(\mathbf{A}) = \mathsf{T}$, iff $\mathcal{I}^\psi(\mathbf{B}) = \mathsf{F}$, iff $\mathcal{I}^\psi(\mathbf{B}) = \mathcal{I}_\psi(\mathbf{B})$, iff $\mathcal{I}^\psi(\mathbf{A}) = \mathcal{I}_\psi(\mathbf{A})$.
   - 3.3. If $\mathbf{A} = \mathbf{B} \wedge \mathbf{C}$ then we argue similarly
4. Hence $\mathcal{I}^\psi(\mathbf{A}) = \mathcal{I}_\psi(\mathbf{A})$ for all $\mathrm{PL}^{\mathrm{nq}}$ formulae and we have concluded the proof.

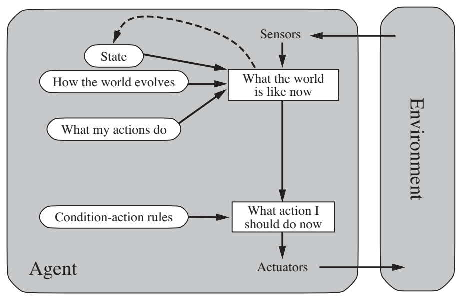
10.4 Inference in Propositional Logics

We have now defined syntax (the language agents can use to represent knowledge) and its semantics (how expressions of this language relate to the world the agent's environment). Theoretically, an agent could use the entailment relation to derive new knowledge percepts and the existing state representation – in the MAKE−PERCEPT−SENTENCE and MAKE−ACTION−SENTENCE subroutines below. But as we have seen in above, this is very tedious. A much better way would be to have a set of rules that directly act on the state representations.

Agents that Think Rationally

▷ **Idea:** Think Before You Act!
"Thinking" = Inference about knowledge represented using logic.

▷ **Definition 10.4.1.** A logic-based agent is a model-based agent that represents the world state as a logical formula and uses inference to think about the state of the environment and its own actions.



function KB−AGENT ($percept$) **returns** an action
  persistent: KB, a knowledge base
        t, a counter, initially 0, indicating time
  TELL(KB, MAKE−PERCEPT−SENTENCE($percept$,t))
  $action$:= ASK(KB, MAKE−ACTION−QUERY(t))
  TELL(KB, MAKE−ACTION−SENTENCE($action$,t))
  $t := t+1$
  return $action$

A Simple Formal System: Prop. Logic with Hilbert-Calculus

- **Formulae:** built from propositional variables: $P, Q, R...$ and implication: $\Rightarrow$
- **Semantics:** $\mathcal{I}_\varphi(P) = \varphi(P)$ and $\mathcal{I}_\varphi(\mathbf{A} \Rightarrow \mathbf{B}) = \mathsf{T}$, iff $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{F}$ or $\mathcal{I}_\varphi(\mathbf{B}) = \mathsf{T}$.
- **Definition 10.4.2.** The Hilbert calculus $\mathcal{H}^0$ consists of the inference rules:

$$\frac{}{P \Rightarrow Q \Rightarrow P}\,\mathsf{K} \qquad \frac{}{(P \Rightarrow Q \Rightarrow R) \Rightarrow (P \Rightarrow Q) \Rightarrow P \Rightarrow R}\,\mathsf{S}$$

$$\frac{\mathbf{A} \Rightarrow \mathbf{B} \quad \mathbf{A}}{\mathbf{B}}\,\mathsf{MP} \qquad \frac{\mathbf{A}}{[\mathbf{B}/X](\mathbf{A})}\,\mathsf{Subst}$$

- **Example 10.4.3.** A $\mathcal{H}^0$ theorem $\mathbf{C} \Rightarrow \mathbf{C}$ and its proof

Proof: We show that $\emptyset \vdash_{\mathcal{H}^0} \mathbf{C} \Rightarrow \mathbf{C}$

1. $(\mathbf{C} \Rightarrow (\mathbf{C} \Rightarrow \mathbf{C}) \Rightarrow \mathbf{C}) \Rightarrow (\mathbf{C} \Rightarrow \mathbf{C} \Rightarrow \mathbf{C}) \Rightarrow \mathbf{C} \Rightarrow \mathbf{C}$ (S with $[\mathbf{C}/P], [\mathbf{C} \Rightarrow \mathbf{C}/Q], [\mathbf{C}/R]$)
2. $\mathbf{C} \Rightarrow (\mathbf{C} \Rightarrow \mathbf{C}) \Rightarrow \mathbf{C}$ (K with $[\mathbf{C}/P], [\mathbf{C} \Rightarrow \mathbf{C}/Q]$)
3. $(\mathbf{C} \Rightarrow \mathbf{C} \Rightarrow \mathbf{C}) \Rightarrow \mathbf{C} \Rightarrow \mathbf{C}$ (MP on P.1 and P.2)
4. $\mathbf{C} \Rightarrow \mathbf{C} \Rightarrow \mathbf{C}$ (K with $[\mathbf{C}/P], [\mathbf{C}/Q]$)
5. $\mathbf{C} \Rightarrow \mathbf{C}$ (MP on P.3 and P.4)

This is indeed a very simple formal system, but it has all the required parts:

- A formal language: expressions built up from variables and implications.
- A semantics: given by the obvious interpretation function
- A calculus: given by the two axioms and the two inference rules.

The calculus gives us a set of rules with which we can derive new formulae from old ones. The axioms are very simple rules, they allow us to derive these two formulae in any situation. The proper inference rules are slightly more complicated: we read the formulae above the horizontal line as assumptions and the (single) formula below as the conclusion. An inference rule allows us to derive the conclusion, if we have already derived the assumptions.
Now, we can use these inference rules to perform a proof – a sequence of formulae that can be derived from each other. The representation of the proof in the slide is slightly compactified to fit onto the slide: We will make it more explicit here. We first start out by deriving the formula

$$(P \Rightarrow Q \Rightarrow R) \Rightarrow (P \Rightarrow Q) \Rightarrow P \Rightarrow R \tag{10.1}$$

which we can always do, since we have an axiom for this formula, then we apply the rule Subst, where $\mathbf{A}$ is this result, $\mathbf{B}$ is $\mathbf{C}$, and X is the variable P to obtain

$$(\mathbf{C} \Rightarrow Q \Rightarrow R) \Rightarrow (\mathbf{C} \Rightarrow Q) \Rightarrow \mathbf{C} \Rightarrow R \tag{10.2}$$

Next we apply the rule Subst to this where $\mathbf{B}$ is $\mathbf{C} \Rightarrow \mathbf{C}$ and X is the variable Q this time to obtain

$$(\mathbf{C} \Rightarrow (\mathbf{C} \Rightarrow \mathbf{C}) \Rightarrow R) \Rightarrow (\mathbf{C} \Rightarrow \mathbf{C} \Rightarrow \mathbf{C}) \Rightarrow \mathbf{C} \Rightarrow R \tag{10.3}$$

And again, we apply the inference rulerule Subst this time, $\mathbf{B}$ is $\mathbf{C}$ and X is the variable R yielding the first formula in our proof on the slide. To conserve space, we have combined these three steps into one in the slide. The next steps are done in exactly the same way.
In general formulae can be used to represent facts about the world as propositions; they have a semantics that is a mapping of formulae into the real world (propositions are mapped to truth values.) We have seen two relations on formulae: the entailment relation and the deduction relation. The first one is defined purely in terms of the semantics, the second one is given by a calculus, i.e. purely syntactically. Is there any relation between these relations?

Soundness and Completeness

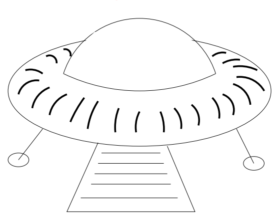
▷ **Definition 10.4.4.** Let $\mathcal{L}:=\langle\mathcal{L},\mathcal{K},\models\rangle$ be a logical system, then we call a calculus $\mathcal{C}$ for $\mathcal{L}$, iff

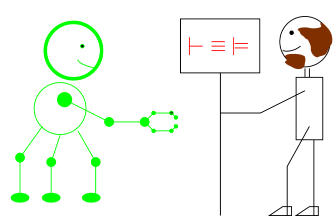
- ▷ sound (or correct), iff $\mathcal{H}\models\mathbf{A}$, whenever $\mathcal{H}\vdash_{\mathcal{C}}\mathbf{A}$, and
- ▷ complete, iff $\mathcal{H}\vdash_{\mathcal{C}}\mathbf{A}$, whenever $\mathcal{H}\models\mathbf{A}$.

▷ **Goal:** Find calculi C, such that $\vdash_{C}\mathbf{A}$ iff $\models\mathbf{A}$ (provability and validity coincide)

- ▷ To TRUTH through PROOF (CALCULEMUS [Leibniz ∼1680])

▷





Ideally, both relations would be the same, then the calculus would allow us to infer all facts that can be represented in the given formal language and that are true in the real world, and only those. In other words, our representation and inference is faithful to the world.

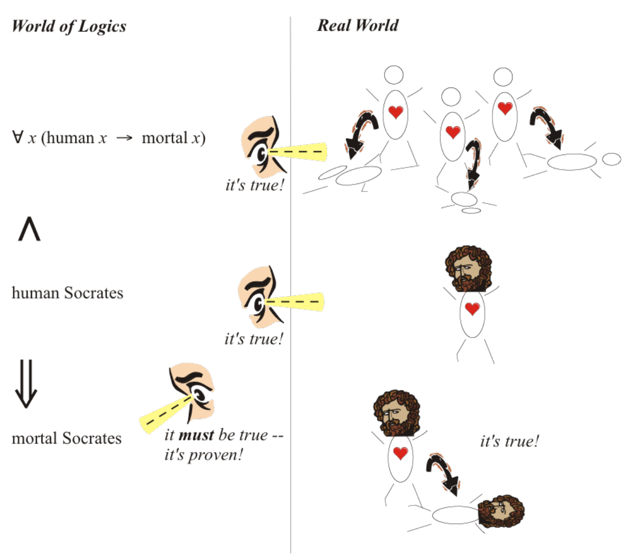
A consequence of this is that we can rely on purely syntactical means to make predictions about the world. Computers rely on formal representations of the world; if we want to solve a problem on our computer, we first represent it in the computer (as data structures, which can be seen as a formal language) and do syntactic manipulations on these structures (a form of calculus). Now, if the provability relation induced by the calculus and the validity relation coincide (this will be quite difficult to establish in general), then the solutions of the program will be correct, and we will find all possible ones.

Of course, the logics we have studied so far are very simple, and not able to express interesting facts about the world, but we will study them as a simple example of the fundamental problem of computer science: How do the formal representations correlate with the real world. Within the world of logics, one can derive new propositions (the *conclusions*, here: *Socrates is mortal*) from given ones (the *premises*, here: *Every human is mortal* and *Sokrates is human*). Such derivations are *proofs*.

In particular, logics can describe the internal structure of real-life facts; e.g. individual things, actions, properties. A famous example, which is in fact as old as it appears, is illustrated in the slide below.

The miracle of logics

▷ Purely formal derivations are true in the real world!



If a logic is correct, the conclusions one can prove are true (= hold in the real world) whenever the premises are true. This is a miraculous fact (think about it!)

10.5 Propositional Natural Deduction Calculus

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/22520` and `https://fau.tv/clip/id/22525`.

We will now introduce the "natural deduction" calculus for propositional logic. The calculus was created in order to model the natural mode of reasoning e.g. in everyday mathematical practice. In particular, it was intended as a counter-approach to the well-known Hilbert style calculi, which were mainly used as theoretical devices for studying reasoning in principle, not for modeling particular reasoning styles. We will introduce natural deduction in two styles/notation, both were invented by Gerhard Gentzen in the 1930's and are very much related. The Natural Deduction style (ND) uses "local hypotheses" in proofs for hypothetical reasoning, while the "sequent style" is a rationalized version and extension of the ND calculus that makes certain meta-proofs simpler to push through by making the context of local hypotheses explicit in the notation. The sequent notation also constitutes a more adequate data struture for implementations, and user interfaces.

Rather than using a minimal set of inference rules, we introduce a natural deduction calculus that provides two/three inference rules for every logical constant, one "introduction rule" (an inference rule that derives a formula with that symbol at the head) and one "elimination rule" (an inference rule that acts on a formula with this head and derives a set of subformulae).

Calculi: Natural Deduction ($\mathcal{ND}_0$; Gentzen [Gen34])

- **Idea:** $\mathcal{ND}_0$ tries to mimic human argumentation for theorem proving.
- **Definition 10.5.1.** The propositional natural deduction calculus $\mathcal{ND}_0$ has inference rules for the introduction and elimination of connectives:

Introduction	Elimination	Axiom
$\dfrac{\mathbf{A}\quad\mathbf{B}}{\mathbf{A}\land\mathbf{B}}\land I$	$\dfrac{\mathbf{A}\land\mathbf{B}}{\mathbf{A}}\land E_l \qquad \dfrac{\mathbf{A}\land\mathbf{B}}{\mathbf{B}}\land E_r$	$\dfrac{}{\mathbf{A}\lor\neg\mathbf{A}}\mathrm{TND}$
$\dfrac{\begin{array}{c}[\mathbf{A}]^1\\ \hline\hline \\ \mathbf{B}\end{array}}{\mathbf{A}\Rightarrow\mathbf{B}}\Rightarrow I^1$	$\dfrac{\mathbf{A}\Rightarrow\mathbf{B}\quad\mathbf{A}}{\mathbf{B}}\Rightarrow E$	

  $\Rightarrow I$ proves $\mathbf{A}\Rightarrow\mathbf{B}$ by exhibiting a $\mathcal{ND}_0$ derivation $\mathcal{D}$ (depicted by the double horizontal lines) of $\mathbf{B}$ from the local hypothesis $\mathbf{A}$; $\Rightarrow I$ then discharges (get rid of $\mathbf{A}$, which can only be used in $\mathcal{D}$) the hypothesis and concludes $\mathbf{A}\Rightarrow\mathbf{B}$. This mode of reasoning is called hypothetical reasoning.

- **Definition 10.5.2.**

  Given a set $\mathcal{H}\subseteq wff_0(\mathcal{V}_0)$ of assumptions and a conclusion $\mathbf{C}$, we write $\mathcal{H}\vdash_{\mathcal{ND}_0}\mathbf{C}$, iff there is a $\mathcal{ND}_0$ derivation tree whose leaves are in $\mathcal{H}$.

- **Note:** TND is used only in classical logic (otherwise constructive/intuitionistic)

The most characteristic rule in the natural deduction calculus is the $\Rightarrow I$ rule and the hypothetical reasoning it introduce. $\Rightarrow I$ corresponds to the mathematical way of proving an implication $\mathbf{A}\Rightarrow\mathbf{B}$: We assume that $\mathbf{A}$ is true and show $\mathbf{B}$ from this local hypothesis. When we can do this we discharge the assumption and conclude $\mathbf{A}\Rightarrow\mathbf{B}$.

Note that the local hypothesis is discharged by the rule $\Rightarrow I$, i.e. it cannot be used in any other part of the proof. As the $\Rightarrow I$ rules may be nested, we decorate both the rule and the corresponding assumption with a marker (here the number 1).

Let us now consider an example of hypothetical reasoning in action.

Natural Deduction: Examples

▷ **Example 10.5.3 (Inference with Local Hypotheses).**

$$\dfrac{\dfrac{\dfrac{[\mathbf{A}\wedge\mathbf{B}]^1}{\mathbf{B}}\wedge E_r \quad \dfrac{[\mathbf{A}\wedge\mathbf{B}]^1}{\mathbf{A}}\wedge E_l}{\mathbf{B}\wedge\mathbf{A}}\wedge I}{\mathbf{A}\wedge\mathbf{B}\Rightarrow\mathbf{B}\wedge\mathbf{A}}\Rightarrow I^1 \qquad \begin{array}{c}[A]^1\\ [B]^2\\ \dfrac{\dfrac{A}{B\Rightarrow A}\Rightarrow I^2}{A\Rightarrow B\Rightarrow A}\Rightarrow I^1\end{array}$$

Here we see hypothetical reasoning with local hypotheses at work. In the left example, we assume the formula $\mathbf{A}\wedge\mathbf{B}$ and can use it in the proof until it is discharged by the rule $\wedge E_l$ on the bottom – therefore we decorate the hypothesis and the rule by corresponding numbers (here the label "1"). Note the assumption $\mathbf{A}\wedge\mathbf{B}$ is *local to the proof fragment* delineated by the corresponding local hypothesishypothesis and the discharging rule, i.e. even if this proof is only a fragment of a larger proof, then we cannot use its local hypothesishypothesis anywhere else.

Note also that we can use as many copies of the local hypothesis as we need; they are all discharged at the same time.
In the right example we see that local hypotheses can be nested as long as they are kept local. In particular, we may not use the hypothesis $\mathbf{B}$ after the $\Rightarrow I^2$, e.g. to continue with a $\Rightarrow E$.
One of the nice things about the natural deduction calculus is that the deduction theorem is almost trivial to prove. In a sense, the triviality of the deduction theorem is the central idea of the calculus and the feature that makes it so natural.

A Deduction Theorem for $\mathcal{ND}_0$

▷ **Theorem 10.5.4.** $\mathcal{H},\mathbf{A}\vdash_{\mathcal{ND}_0}\mathbf{B}$, *iff* $\mathcal{H}\vdash_{\mathcal{ND}_0}\mathbf{A}\Rightarrow\mathbf{B}$.

▷ *Proof:* We show the two directions separately

1. If $\mathcal{H},\mathbf{A}\vdash_{\mathcal{ND}_0}\mathbf{B}$, then $\mathcal{H}\vdash_{\mathcal{ND}_0}\mathbf{A}\Rightarrow\mathbf{B}$ by $\Rightarrow I$, and
2. If $\mathcal{H}\vdash_{\mathcal{ND}_0}\mathbf{A}\Rightarrow\mathbf{B}$, then $\mathcal{H},\mathbf{A}\vdash_{\mathcal{ND}_0}\mathbf{A}\Rightarrow\mathbf{B}$ by weakening and $\mathcal{H},\mathbf{A}\vdash_{\mathcal{ND}_0}\mathbf{B}$ by $\Rightarrow E$.

Another characteristic of the natural deduction calculus is that it has inference rules (introduction and elimination rules) for all connectives. So we extend the set of rules from Definition 10.5.1 for disjunction, negation and falsity.

More Rules for Natural Deduction

▷ **Note:** $\mathcal{ND}_0$ does not try to be minimal, but comfortable to work in!x

▷ **Definition 10.5.5.** $\mathcal{ND}_0$ has the following additional inference rules for the remaining connectives.

$$\frac{\mathbf{A}}{\mathbf{A}\vee\mathbf{B}}\vee I_l \qquad \frac{\mathbf{B}}{\mathbf{A}\vee\mathbf{B}}\vee I_r \qquad \frac{\mathbf{A}\vee\mathbf{B} \quad \begin{array}{c}[\mathbf{A}]^1\\ \vdots\\ \mathbf{C}\end{array} \quad \begin{array}{c}[\mathbf{B}]^1\\ \vdots\\ \mathbf{C}\end{array}}{\mathbf{C}}\vee E^1$$

$$\frac{\begin{array}{c}[\mathbf{A}]^1\\ \vdots\\ \mathbf{C}\end{array} \quad \begin{array}{c}[\mathbf{A}]^1\\ \vdots\\ \neg\mathbf{C}\end{array}}{\neg\mathbf{A}}\neg I^1 \qquad \frac{\neg\neg\mathbf{A}}{\mathbf{A}}\neg E$$

$$\frac{\neg\mathbf{A} \quad \mathbf{A}}{F}FI \qquad \frac{F}{\mathbf{A}}FE$$

▷ **Again:** $\neg E$ is used only in classical logic (otherwise constructive/intuitionistic)

Natural Deduction in Sequent Calculus Formulation

▷ **Idea:** Represent hypotheses explicitly. (lift calculus to judgments)

▷ **Definition 10.5.6.** A judgment is a meta statement about the provability of propositions.

▷ **Definition 10.5.7.** A sequent is a judgment of the form $\mathcal{H}\vdash\mathbf{A}$ about the provability of the formula $\mathbf{A}$ from the set $\mathcal{H}$ of hypotheses. We write $\vdash\mathbf{A}$ for $\emptyset\vdash\mathbf{A}$.

▷ **Idea:** Reformulate $\mathcal{ND}_0$ inference rules so that they act on sequents.

▷ **Example 10.5.8.** We give the sequent style version of Example 10.5.3:

$$\cfrac{\cfrac{\cfrac{\cfrac{}{\mathbf{A}\wedge\mathbf{B}\vdash\mathbf{A}\wedge\mathbf{B}}\mathrm{Ax}}{\mathbf{A}\wedge\mathbf{B}\vdash\mathbf{B}}\wedge E_r \quad \cfrac{\cfrac{}{\mathbf{A}\wedge\mathbf{B}\vdash\mathbf{A}\wedge\mathbf{B}}\mathrm{Ax}}{\mathbf{A}\wedge\mathbf{B}\vdash\mathbf{A}}\wedge E_l}{\mathbf{A}\wedge\mathbf{B}\vdash\mathbf{B}\wedge\mathbf{A}}\wedge I}{\vdash\mathbf{A}\wedge\mathbf{B}\Rightarrow\mathbf{B}\wedge\mathbf{A}}\Rightarrow I \qquad \cfrac{\cfrac{\cfrac{}{\mathbf{A},\mathbf{B}\vdash\mathbf{A}}\mathrm{Ax}}{\mathbf{A}\vdash\mathbf{B}\Rightarrow\mathbf{A}}\Rightarrow I}{\vdash\mathbf{A}\Rightarrow\mathbf{B}\Rightarrow\mathbf{A}}\Rightarrow I$$

▷ **Note:** Even though the antecedent of a sequent is written like a sequence, it is actually a set. In particular, we can permute and duplicate members at will.

Sequent-Style Rules for Natural Deduction

▷ **Definition 10.5.9.** The following inference rules make up the propositional sequent style natural deduction calculus $\mathcal{ND}^0_\vdash$:

$$\frac{}{\Gamma, \mathbf{A} \vdash \mathbf{A}}\mathsf{Ax} \qquad \frac{\Gamma \vdash \mathbf{B}}{\Gamma, \mathbf{A} \vdash \mathbf{B}}\mathsf{weaken} \qquad \frac{}{\Gamma \vdash \mathbf{A} \vee \neg \mathbf{A}}\mathsf{TND}$$

$$\frac{\Gamma \vdash \mathbf{A} \quad \Gamma \vdash \mathbf{B}}{\Gamma \vdash \mathbf{A} \wedge \mathbf{B}}\wedge I \qquad \frac{\Gamma \vdash \mathbf{A} \wedge \mathbf{B}}{\Gamma \vdash \mathbf{A}}\wedge E_l \qquad \frac{\Gamma \vdash \mathbf{A} \wedge \mathbf{B}}{\Gamma \vdash \mathbf{B}}\wedge E_r$$

$$\frac{\Gamma \vdash \mathbf{A}}{\Gamma \vdash \mathbf{A} \vee \mathbf{B}}\vee I_l \qquad \frac{\Gamma \vdash \mathbf{B}}{\Gamma \vdash \mathbf{A} \vee \mathbf{B}}\vee I_r \qquad \frac{\Gamma \vdash \mathbf{A} \vee \mathbf{B} \quad \Gamma, \mathbf{A} \vdash \mathbf{C} \quad \Gamma, \mathbf{B} \vdash \mathbf{C}}{\Gamma \vdash \mathbf{C}}\vee E$$

$$\frac{\Gamma, \mathbf{A} \vdash \mathbf{B}}{\Gamma \vdash \mathbf{A} \Rightarrow \mathbf{B}}\Rightarrow I \qquad \frac{\Gamma \vdash \mathbf{A} \Rightarrow \mathbf{B} \quad \Gamma \vdash \mathbf{A}}{\Gamma \vdash \mathbf{B}}\Rightarrow E$$

$$\frac{\Gamma, \mathbf{A} \vdash F}{\Gamma \vdash \neg \mathbf{A}}\neg I \qquad \frac{\Gamma \vdash \neg\neg \mathbf{A}}{\Gamma \vdash \mathbf{A}}\neg E$$

$$\frac{\Gamma \vdash \neg \mathbf{A} \quad \Gamma \vdash \mathbf{A}}{\Gamma \vdash F}FI \qquad \frac{\Gamma \vdash F}{\Gamma \vdash \mathbf{A}}FE$$

Michael Kohlhase: Artificial Intelligence 1 340 2023-09-20

Linearized Notation for (Sequent-Style) ND Proofs

▷ Linearized notation for sequent-style ND proofs

1. $\mathcal{H}_1 \vdash \mathbf{A}_1 \quad (\mathcal{J}_1)$
2. $\mathcal{H}_2 \vdash \mathbf{A}_2 \quad (\mathcal{J}_2)$
3. $\mathcal{H}_3 \vdash \mathbf{A}_3 \quad (\mathcal{J}_3 1, 2)$

corresponds to $\frac{\mathcal{H}_1 \vdash \mathbf{A}_1 \quad \mathcal{H}_2 \vdash \mathbf{A}_2}{\mathcal{H}_3 \vdash \mathbf{A}_3}\mathcal{R}$

▷ **Example 10.5.10.** We show a linearized version of the $\mathcal{ND}_0$ examples Example 10.5.8

#	hyp	⊢	formula	NDjust
1.	1	⊢	$\mathbf{A} \wedge \mathbf{B}$	Ax
2.	1	⊢	$\mathbf{B}$	$\wedge E_r$ 1
3.	1	⊢	$\mathbf{A}$	$\wedge E_l$ 1
4.	1	⊢	$\mathbf{B} \wedge \mathbf{A}$	$\wedge I$ 2, 3
5.		⊢	$\mathbf{A} \wedge \mathbf{B} \Rightarrow \mathbf{B} \wedge \mathbf{A}$	$\Rightarrow I$ 4

#	hyp	⊢	formula	NDjust
1.	1	⊢	$\mathbf{A}$	Ax
2.	2	⊢	$\mathbf{B}$	Ax
3.	1, 2	⊢	$\mathbf{A}$	weaken 1, 2
4.	1	⊢	$\mathbf{B} \Rightarrow \mathbf{A}$	$\Rightarrow I$ 3
5.		⊢	$\mathbf{A} \Rightarrow \mathbf{B} \Rightarrow \mathbf{A}$	$\Rightarrow I$ 4

Michael Kohlhase: Artificial Intelligence 1 341 2023-09-20

Each row in the table represents one inference step in the proof. It consists of line number (for referencing), a formula for the asserted property, a justification via a ND rules (and the rows this one is derived from), and finally a list of row numbers of proof steps that are local hypotheses in effect for the current row.

10.6 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25027`.

Summary

▷ Sometimes, it pays off to think before acting.

▷ In AI, "thinking" is implemented in terms of reasoning in order to deduce new knowledge from a knowledge base represented in a suitable logic.

▷ Logic prescribes a syntax for formulas, as well as a semantics prescribing which interpretations satisfy them. $\mathbf{A}$ entails $\mathbf{B}$ if all interpretations that satisfy $\mathbf{A}$ also satisfy $\mathbf{B}$. deduction is the process of deriving new entailed formulas.

▷ Propositional logic formulas are built from atomic propositions, with the connectives *and*, *or*, *not*.

Issues with Propositional Logic

▷ **Awkward to write for humans:** E.g., to model the Wumpus world we had to make a copy of the rules for every cell ...

$R_1{:=}\neg S_{1,1} \Rightarrow \neg W_{1,1} \wedge \neg W_{1,2} \wedge \neg W_{2,1}$
$R_2{:=}\neg S_{2,1} \Rightarrow \neg W_{1,1} \wedge \neg W_{2,1} \wedge \neg W_{2,2} \wedge \neg W_{3,1}$
$R_3{:=}\neg S_{1,2} \Rightarrow \neg W_{1,1} \wedge \neg W_{1,2} \wedge \neg W_{2,2} \wedge \neg W_{1,3}$

Compared to

Cell adjacent to Wumpus: Stench (else: None)

that is not a very nice description language ...

▷ **Time:** For things that change (e.g., Wumpus moving according to certain rules), we need time-indexed propositions (like, $S_{2,1}^{t=7}$) to represent validity over time $\leadsto$ further expansion of the rules.

▷ **Can we design a more human-like logic?:** Yep

- ▷ Predicate logic: quantification of variables ranging over individuals. (cf. chapter 14 and chapter 15)
- ▷ ...and a whole zoo of logics much more powerful still.
- ▷ Note: In applications, propositional CNF encodings are generated by computer programs. This mitigates (but does not remove!) the inconveniences of propositional modeling.

Suggested Reading:

- *Chapter 7: Logical Agents*, Sections 7.1 – 7.5 [RN09].
  - Sections 7.1 and 7.2 roughly correspond to my "Introduction", Section 7.3 roughly corresponds to my "Logic (in AI)", Section 7.4 roughly corresponds to my "Propositional Logic", Section 7.5 roughly corresponds to my "Resolution" and "Killing a Wumpus".
  - Overall, the content is quite similar. I have tried to add some additional clarifying illustrations. RN gives many complementary explanations, nice as additional background reading.

- I would note that RN's presentation of resolution seems a bit awkward, and Section 7.5 contains some additional material that is imho not interesting (alternate inference rules, forward and backward chaining). Horn clauses and unit resolution (also in Section 7.5), on the other hand, are quite relevant.

Chapter 11

Machine-Oriented Calculi for Propositional Logic

A Video Nugget covering this chapter can be found at `https://fau.tv/clip/id/22531`.

Automated Deduction as an Agent Inference Procedure

- ▷ **Recall:** Our knowledge of the cave entails a definite Wumpus position!(slide 308)
- ▷ **Problem:** That was human reasoning, can we build an agent function that does this?
- ▷ **Answer:** As for constraint networks, we use inference, here resolution/tableaux.

The following theorem is simple, but will be crucial later on.

Unsatisfiability Theorem

- ▷ **Theorem 11.0.1 (Unsatisfiability Theorem).** *$\mathcal{H} \models \mathbf{A}$ iff $\mathcal{H} \cup \{\neg \mathbf{A}\}$ is unsatisfiable.*
- ▷ *Proof:* We prove both directions separately
  1. "$\Rightarrow$": Say $\mathcal{H} \models \mathbf{A}$
     1.1. For any φ with $\varphi \models \mathcal{H}$ we have $\varphi \models \mathbf{A}$ and thus $\varphi \not\models \neg \mathbf{A}$.
  2. "$\Leftarrow$": Say $\mathcal{H} \cup \{\neg \mathbf{A}\}$ is unsatisfiable.
     2.1. For any φ with $\varphi \models \mathcal{H}$ we have $\varphi \not\models \neg \mathbf{A}$ and thus $\varphi \models \mathbf{A}$.
- ▷ **Observation 11.0.2.** *Entailment can be tested via satisfiability.*

Test Calculi: A Paradigm for Automating Inference

- ▷ **Definition 11.0.3.** Given a formal system $\langle \mathcal{L}, \mathcal{K}, \models, \mathcal{C} \rangle$, the task of theorem proving consists in determining whether $\mathcal{H} \vdash_{\mathcal{C}} C$ for a conjecture $C \in \mathcal{L}$ and hypotheses $\mathcal{H} \subseteq$

$\mathcal{L}$.

▷ **Definition 11.0.4.** Given a logical system $\mathcal{L}:=\langle\mathcal{L},\mathcal{K},\models\rangle$, the task of automated theorem proving (ATP) consists of developing calculi for $\mathcal{L}$ and programs – called (automated) theorem provers – that given a set $\mathcal{H}\subseteq\mathcal{L}$ of hypotheses and a conjecture $\mathbf{A}\in\mathcal{L}$ determine whether $\mathcal{H}\models\mathbf{A}$ (usually by searching for $\mathcal{C}$-derivations $\mathcal{H}\vdash_{\mathcal{C}}\mathbf{A}$ in a calculus $\mathcal{C}$).

▷ **Idea:** ATP with a calculus $\mathcal{C}$ for $\langle\mathcal{L},\mathcal{K},\models\rangle$ induces a search problem $\Pi:=\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$, where the states $\mathcal{S}$ are sets of formulae in $\mathcal{L}$, the actions $\mathcal{A}$ are the inference rules from $\mathcal{C}$, the initial state $\mathcal{I}=\{\mathcal{H}\}$, and the goal states are those with $\mathbf{A}\in\mathcal{S}$.

▷ **Problem:** ATP as a search problem does not admit good heuristics, since these need to take the conjecture $\mathcal{A}$ into account.

▷ **Idea:** Turn the search around – using the unsatisfiability theorem (Theorem 11.0.1).

▷ **Definition 11.0.5.** For a given conjecture A and hypotheses $\mathcal{H}$ a test calculus $\mathcal{T}$ tries to derive $\mathcal{H},\overline{A}\vdash_{\mathcal{T}}\bot$ instead of $\mathcal{H}\vdash A$, where $\overline{A}$ is unsatisfiable iff A is valid and $\bot$, an "obviously" unsatisfiable formula.

A derivation $\mathcal{H},\overline{A}\vdash_{\mathcal{T}}\bot$ is called a refutation of A (from $\mathcal{H}$, if $\mathcal{H}\neq\emptyset$).

▷ **Observation:** A test calculus $\mathcal{C}$ induces a search problem where the initial state is $\mathcal{H}\cup\{\neg\mathbf{A}\}$ and $S\in\mathcal{S}$ is a goal state iff $\bot\in S$.(proximity of $\bot$ easier for heuristics)

11.1 Normal Forms

Before we can start, we will need to recap some nomenclature on formulae.

Recap: Atoms and Literals

▷ **Definition 11.1.1.** A formula is called atomic (or an atom) if it does not contain logical constants, else it is called complex.

▷ **Definition 11.1.2.** If $\mathbf{A}$ be a formula, then we call a pair $\mathbf{A}^{\alpha}$ a labeled formula, if $\alpha\in\{\mathsf{T},\mathsf{F}\}$. For a set Φ of formulae we use $\Phi^{\alpha}:=\{\mathbf{A}^{\alpha}|\mathbf{A}\in\Phi\}$.

▷ **Definition 11.1.3.** A labeled atom $\mathbf{A}^{\alpha}$ is called a (positive if $\alpha=\mathsf{T}$, else negative) literal.

▷ **Intuition:** To satisfy a formula, we make it "true". To satisfy a labeled formula $\mathbf{A}^{\alpha}$, it must have the truth value α.

▷ **Definition 11.1.4.** For a literal $\mathbf{A}^{\alpha}$, we call the literal $\mathbf{A}^{\beta}$ with $\alpha\neq\beta$ the opposite literal (or partner literal).

The idea about literals is that they are atoms (the simplest formulae) that carry around their intended truth value.

Alternative Definition: Literals

▷ **Note:** Literals are often defined without recurring to labeled formulae:

▷ **Definition 11.1.5.** A literal is an atoms $\mathbf{A}$ (positive literal) or negated atoms $\neg\mathbf{A}$ (negative literal). $\mathbf{A}$ and $\neg\mathbf{A}$ are opposite literals.

▷ **Note:** This notion of literal is equivalent to the labeled formulae-notion of literal, but does not generalize as well to logics with more than two truth values.

Normal Forms

▷ There are two quintessential normal forms for propositional formulae: (there are others as well)

▷ **Definition 11.1.6.** A formula is in conjunctive normal form (CNF) if it is a conjunction of disjunctions of literals: i.e. if it is of the form $\bigwedge_{i=1}^{n}\bigvee_{j=1}^{m_i} l_{ij}$

▷ **Definition 11.1.7.** A formula is in disjunctive normal form (DNF) if it is a disjunction of conjunctions of literals: i.e. if it is of the form $\bigvee_{i=1}^{n}\bigwedge_{j=1}^{m_i} l_{ij}$

▷ **Observation 11.1.8.** *Every formula has equivalent formulae in CNF and DNF.*

Video Nuggets covering this chapter can be found at `https://fau.tv/clip/id/23705` and `https://fau.tv/clip/id/23708`.

11.2 Analytical Tableaux

Test Calculi: Tableaux and Model Generation

▷ **Idea:** A tableau calculus is a test calculus that

- ▷ analyzes a labeled formulae in a tree to determine satisfiability,
- ▷ its branches correspond to valuations ($\rightsquigarrow$ models).

▷ **Example 11.2.1.** Tableau calculi try to construct models for labeled formulae:

<table>
<tr><th>Tableau refutation (Validity)</th><th>Model generation (Satisfiability)</th></tr>
<tr><td>$\models P\wedge Q\Rightarrow Q\wedge P$</td><td>$\models P\wedge(Q\vee\neg R)\wedge\neg Q$</td></tr>
<tr><td>$\begin{array}{c}(P\wedge Q\Rightarrow Q\wedge P)^{\mathsf{F}}\\ (P\wedge Q)^{\mathsf{T}}\\ (Q\wedge P)^{\mathsf{F}}\\ P^{\mathsf{T}}\\ Q^{\mathsf{T}}\\ \begin{array}{c|c} P^{\mathsf{F}} & Q^{\mathsf{F}}\\ \bot & \bot\end{array}\end{array}$</td><td>$\begin{array}{c}(P\wedge(Q\vee\neg R)\wedge\neg Q)^{\mathsf{T}}\\ (P\wedge(Q\vee\neg R))^{\mathsf{T}}\\ \neg Q^{\mathsf{T}}\\ Q^{\mathsf{F}}\\ P^{\mathsf{T}}\\ (Q\vee\neg R)^{\mathsf{T}}\\ \begin{array}{c|c} Q^{\mathsf{T}} & \neg R^{\mathsf{T}}\\ \bot & R^{\mathsf{F}}\end{array}\end{array}$</td></tr>
<tr><td>No Model</td><td>Herbrand Model $\{P^{\mathsf{T}},Q^{\mathsf{F}},R^{\mathsf{F}}\}$
$\varphi{:=}\{P\mapsto\mathsf{T},Q\mapsto\mathsf{F},R\mapsto\mathsf{F}\}$</td></tr>
</table>

- **Idea:** Open branches in saturated tableaux yield models.
- **Algorithm:** Fully expand all possible tableaux, (no rule can be applied)
  - Satisfiable, iff there are open branches (correspond to models)

Tableau calculi develop a formula in a tree-shaped arrangement that represents a case analysis on when a formula can be made true (or false). Therefore the formulae are decorated with exponents that hold the intended truth value.

On the left we have a refutation tableau that analyzes a negated formula (it is decorated with the intended truth value F). Both branches contain an elementary contradiction $\perp$.

On the right we have a model generation tableau, which analyzes a positive formula (it is decorated with the intended truth value T. This tableau uses the same rules as the refutation tableau, but makes a case analysis of when this formula can be satisfied. In this case we have a closed branch and an open one, which corresponds a model).

Now that we have seen the examples, we can write down the tableau rules formally.

Analytical Tableaux (Formal Treatment of $\mathcal{T}_0$)

- **Idea:** A test calculus where
  - A labeled formula is analyzed in a tree to determine satisfiability,
  - branches correspond to valuations (models)
- **Definition 11.2.2.** The propositional tableau calculus $\mathcal{T}_0$ has two inference rules per connective (one for each possible label)

$$\frac{(\mathbf{A}\wedge\mathbf{B})^{\mathsf{T}}}{\begin{array}{c}\mathbf{A}^{\mathsf{T}}\\ \mathbf{B}^{\mathsf{T}}\end{array}}\mathcal{T}_0\wedge \qquad \frac{(\mathbf{A}\wedge\mathbf{B})^{\mathsf{F}}}{\mathbf{A}^{\mathsf{F}}\;\Big|\;\mathbf{B}^{\mathsf{F}}}\mathcal{T}_0\vee \qquad \frac{\neg\mathbf{A}^{\mathsf{T}}}{\mathbf{A}^{\mathsf{F}}}\mathcal{T}_0\neg^{\mathsf{T}} \qquad \frac{\neg\mathbf{A}^{\mathsf{F}}}{\mathbf{A}^{\mathsf{T}}}\mathcal{T}_0\neg^{\mathsf{F}} \qquad \frac{\begin{array}{c}\mathbf{A}^{\alpha}\\ \mathbf{A}^{\beta}\end{array}\quad \alpha\neq\beta}{\perp}\mathcal{T}_0\perp$$

  Use rules exhaustively as long as they contribute new material ($\rightsquigarrow$ termination)
- **Definition 11.2.3.** We call any tree ($|$ introduces branches) produced by the $\mathcal{T}_0$ inference rules from a set Φ of labeled formulae a tableau for Φ.
- **Definition 11.2.4.** Call a tableau saturated, iff no rule adds new material and a branch closed, iff it ends in $\perp$, else open. A tableau is closed, iff all of its branches are.

These inference rules act on tableaux have to be read as follows: if the formulae over the line appear in a tableau branch, then the branch can be extended by the formulae or branches below the line. There are two rules for each primary connective, and a branch closing rule that adds the special symbol $\perp$ (for unsatisfiability) to a branch.

We use the tableau rules with the convention that they are only applied, if they contribute new material to the branch. This ensures termination of the tableau procedure for propositional logic (every rule eliminates one primary connective).

Definition 11.2.5. We will call a closed tableau with the labeled formula $\mathbf{A}^{\alpha}$ at the root a tableau refutation for $\mathcal{A}^{\alpha}$.

The saturated tableau represents a full case analysis of what is necessary to give $\mathbf{A}$ the truth value α; since all branches are closed (contain contradictions) this is impossible.

Analytical Tableaux ($\mathcal{T}_0$ continued)

▷ **Definition 11.2.6 ($\mathcal{T}_0$-Theorem/Derivability).** $\mathbf{A}$ is a $\mathcal{T}_0$-theorem ($\vdash_{\mathcal{T}_0}\mathbf{A}$), iff there is a closed tableau with $\mathbf{A}^{\mathsf{F}}$ at the root.

$\Phi \subseteq \mathit{wff}_0(\mathcal{V}_0)$ derives $\mathbf{A}$ in $\mathcal{T}_0$ ($\Phi\vdash_{\mathcal{T}_0}\mathbf{A}$), iff there is a closed tableau starting with $\mathbf{A}^{\mathsf{F}}$ and Φ^{T}. The tableau with only a branch of $\mathbf{A}^{\mathsf{F}}$ and Φ^{T} is called initial for $\Phi\vdash_{\mathcal{T}_0}\mathbf{A}$.

Definition 11.2.7. We will call a tableau refutation for $\mathbf{A}^{\mathsf{F}}$ a tableau proof for $\mathbf{A}$, since it refutes the possibility of finding a model where $\mathbf{A}$ evaluates to F. Thus $\mathbf{A}$ must evaluate to T in all models, which is just our definition of validity.

Thus the tableau procedure can be used as a calculus for propositional logic. In contrast to the propositional Hilbert calculus it does not prove a theorem $\mathbf{A}$ by deriving it from a set of axioms, but it proves it by refuting its negation. Such calculi are called negative or test calculi. Generally negative calculi have computational advantages over positive ones, since they have a built-in sense of direction.

We have rules for all the necessary connectives (we restrict ourselves to $\wedge$ and $\neg$, since the others can be expressed in terms of these two via the propositional identities above. For instance, we can write $\mathbf{A} \vee \mathbf{B}$ as $\neg(\neg\mathbf{A} \wedge \neg\mathbf{B})$, and $\mathbf{A} \Rightarrow \mathbf{B}$ as $\neg\mathbf{A} \vee \mathbf{B}$,....)

We now look at a formulation of propositional logic with fancy variable names. Note that $\mathrm{loves}(\mathrm{mary}, \mathrm{bill})$ is just a variable name like P or X, which we have used earlier.

A Valid Real-World Example

▷ **Example 11.2.8.** *If Mary loves Bill and John loves Mary, then John loves Mary*

$$
\begin{array}{c}
(\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary}) \Rightarrow \mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{F}} \\
\neg(\neg\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary})) \wedge \neg\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{F}} \\
(\neg\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary})) \wedge \neg\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{T}} \\
\neg\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{T}} \\
\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{F}} \\
(\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{T}} \\
\neg\mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{T}} \\
\mathrm{loves}(\mathrm{mary},\mathrm{bill})^{\mathsf{T}} \\
\mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{T}} \\
\mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{F}} \\
\bot
\end{array}
$$

This is a closed tableau, so the $\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary}) \Rightarrow \mathrm{loves}(\mathrm{john},\mathrm{mary})$ is a $\mathcal{T}_0$-theorem.

As we will see, $\mathcal{T}_0$ is sound and complete, so

$$\mathrm{loves}(\mathrm{mary},\mathrm{bill}) \wedge \mathrm{loves}(\mathrm{john},\mathrm{mary}) \Rightarrow \mathrm{loves}(\mathrm{john},\mathrm{mary})$$

is valid.

We could have used the unsatisfiability theorem (Theorem 11.0.1) here to show that *If Mary loves Bill and John loves Mary* entails *John loves Mary*. But there is a better way to show entailment: we directly use derivability in $\mathcal{T}_0$.

Deriving Entailment in $\mathcal{T}_0$

▷ **Example 11.2.9.** *Mary loves Bill* and *John loves Mary* together entail that *John loves Mary*

$$\begin{array}{c} \mathrm{loves}(\mathrm{mary},\mathrm{bill})^{\mathsf{T}} \\ \mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{T}} \\ \mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{F}} \\ \bot \end{array}$$

This is a closed tableau, so $\{\mathrm{loves}(\mathrm{mary},\mathrm{bill}),\mathrm{loves}(\mathrm{john},\mathrm{mary})\}\vdash_{\mathcal{T}_0}\mathrm{loves}(\mathrm{john},\mathrm{mary})$.

Again, as $\mathcal{T}_0$ is sound and complete we have

$$\{\mathrm{loves}(\mathrm{mary},\mathrm{bill}),\mathrm{loves}(\mathrm{john},\mathrm{mary})\}\models\mathrm{loves}(\mathrm{john},\mathrm{mary})$$

Note: We can also use the tableau calculus to try and show entailment (and fail). The nice thing is that the failed proof, we can see what went wrong.

A Falsifiable Real-World Example

▷ **Example 11.2.10.** * *If Mary loves Bill or John loves Mary, then John loves Mary*
Try proving the implication (this fails)

$$\begin{array}{c} ((\mathrm{loves}(\mathrm{mary},\mathrm{bill})\vee\mathrm{loves}(\mathrm{john},\mathrm{mary}))\Rightarrow\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{F}} \\ \neg(\neg\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill})\vee\mathrm{loves}(\mathrm{john},\mathrm{mary}))\wedge\neg\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{F}} \\ (\neg\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill})\vee\mathrm{loves}(\mathrm{john},\mathrm{mary}))\wedge\neg\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{T}} \\ \neg\mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{T}} \\ \mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{F}} \\ \neg\neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill})\vee\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{T}} \\ \neg(\mathrm{loves}(\mathrm{mary},\mathrm{bill})\vee\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{F}} \\ (\mathrm{loves}(\mathrm{mary},\mathrm{bill})\vee\mathrm{loves}(\mathrm{john},\mathrm{mary}))^{\mathsf{T}} \\ \begin{array}{c|c} \mathrm{loves}(\mathrm{mary},\mathrm{bill})^{\mathsf{T}} & \mathrm{loves}(\mathrm{john},\mathrm{mary})^{\mathsf{T}} \\ & \bot \end{array} \end{array}$$

Indeed we can make $\mathcal{I}_\varphi(\mathrm{loves}(\mathrm{mary},\mathrm{bill}))=\mathsf{T}$ but $\mathcal{I}_\varphi(\mathrm{loves}(\mathrm{john},\mathrm{mary}))=\mathsf{F}$.

Obviously, the tableau above is saturated, but not closed, so it is not a tableau proof for our initial entailment conjecture. We have marked the literal on the open branch green, since they allow us to read of the conditions of the situation, in which the entailment fails to hold. As we intuitively argued above, this is the situation, where *Mary loves Bill*. In particular, the open branch gives us a variable assignment (marked in green) that satisfies the initial formula. In this case, *Mary loves Bill*, which is a situation, where the entailment fails.

Again, the derivability version is much simpler:

Testing for Entailment in $\mathcal{T}_0$

▷ **Example 11.2.11.** Does *Mary loves Bill or John loves Mary* entail that *John loves Mary*?

$$\begin{array}{c} (\text{loves}(\text{mary},\text{bill}) \vee \text{loves}(\text{john},\text{mary}))^{\mathsf{T}} \\ \text{loves}(\text{john},\text{mary})^{\mathsf{F}} \\ \begin{array}{c|c} \text{loves}(\text{mary},\text{bill})^{\mathsf{T}} & \text{loves}(\text{john},\text{mary})^{\mathsf{T}} \\ & \bot \end{array} \end{array}$$

This saturated tableau has an open branch that shows that the interpretation with $\mathcal{I}_\varphi(\text{loves}(\text{mary},\text{bill})) = \mathsf{T}$ but $\mathcal{I}_\varphi(\text{loves}(\text{john},\text{mary})) = \mathsf{F}$ falsifies the derivability/entailment conjecture.

We have seen in the examples above that while it is possible to get by with only the connectives $\vee$ and $\neg$, it is a bit unnatural and tedious, since we need to eliminate the other connectives first. In this chapter, we will make the calculus less frugal by adding rules for the other connectives, without losing the advantage of dealing with a small calculus, which is good making statements about the calculus itself.

11.3 Practical Enhancements for Tableaux

The main idea here is to add the new rules as derivable inference rules, i.e. rules that only abbreviate derivations in the original calculus. Generally, adding derivable inference rules does not change the derivation relation of the calculus, and is therefore a safe thing to do. In particular, we will add the following rules to our tableau calculus.

We will convince ourselves that the first rule is derivable, and leave the other ones as an exercise.

Derived Rules of Inference

▷ **Definition 11.3.1.** An inference rule $\dfrac{\mathbf{A}_1 \quad \ldots \quad \mathbf{A}_n}{\mathbf{C}}$ is called derivable (or a derived rule) in a calculus $\mathcal{C}$, if there is a $\mathcal{C}$ derivation $\mathbf{A}_1, \ldots, \mathbf{A}_n \vdash_{\mathcal{C}} \mathbf{C}$.

▷ **Definition 11.3.2.** We have the following derivable inference rules in $\mathcal{T}_0$:

$$\frac{(\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{T}}}{\mathbf{A}^{\mathsf{F}} \mid \mathbf{B}^{\mathsf{T}}} \qquad \frac{(\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{F}}}{\begin{array}{c}\mathbf{A}^{\mathsf{T}} \\ \mathbf{B}^{\mathsf{F}}\end{array}} \qquad \frac{\begin{array}{c}\mathbf{A}^{\mathsf{T}} \\ (\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{T}}\end{array}}{\mathbf{B}^{\mathsf{T}}}$$

$$\frac{(\mathbf{A} \vee \mathbf{B})^{\mathsf{T}}}{\mathbf{A}^{\mathsf{T}} \mid \mathbf{B}^{\mathsf{T}}} \qquad \frac{(\mathbf{A} \vee \mathbf{B})^{\mathsf{F}}}{\begin{array}{c}\mathbf{A}^{\mathsf{F}} \\ \mathbf{B}^{\mathsf{F}}\end{array}} \qquad \frac{\mathbf{A} \Leftrightarrow \mathbf{B}^{\mathsf{T}}}{\begin{array}{c|c}\mathbf{A}^{\mathsf{T}} & \mathbf{A}^{\mathsf{F}} \\ \mathbf{B}^{\mathsf{T}} & \mathbf{B}^{\mathsf{F}}\end{array}} \qquad \frac{\mathbf{A} \Leftrightarrow \mathbf{B}^{\mathsf{F}}}{\begin{array}{c|c}\mathbf{A}^{\mathsf{T}} & \mathbf{A}^{\mathsf{F}} \\ \mathbf{B}^{\mathsf{F}} & \mathbf{B}^{\mathsf{T}}\end{array}}$$

$$\begin{array}{c} \mathbf{A}^{\mathsf{T}} \\ (\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{T}} \\ (\neg\mathbf{A} \vee \mathbf{B})^{\mathsf{T}} \\ \neg(\neg\neg\mathbf{A} \wedge \neg\mathbf{B})^{\mathsf{T}} \\ (\neg\neg\mathbf{A} \wedge \neg\mathbf{B})^{\mathsf{F}} \\ \begin{array}{c|c} \neg\neg\mathbf{A}^{\mathsf{F}} & \neg\mathbf{B}^{\mathsf{F}} \\ \neg\mathbf{A}^{\mathsf{T}} & \mathbf{B}^{\mathsf{T}} \\ \mathbf{A}^{\mathsf{F}} & \\ \bot & \end{array} \end{array}$$

With these derived rules, theorem proving becomes quite efficient. With these rules, the tableau (Example 11.2.8) would have the following simpler form:

Tableaux with derived Rules (example)

Example 11.3.3.

$$\begin{array}{c}
(\mathsf{loves}(\mathsf{mary},\mathsf{bill}) \wedge \mathsf{loves}(\mathsf{john},\mathsf{mary}) \Rightarrow \mathsf{loves}(\mathsf{john},\mathsf{mary}))^{\mathsf{F}} \\
(\mathsf{loves}(\mathsf{mary},\mathsf{bill}) \wedge \mathsf{loves}(\mathsf{john},\mathsf{mary}))^{\mathsf{T}} \\
\mathsf{loves}(\mathsf{john},\mathsf{mary})^{\mathsf{F}} \\
\mathsf{loves}(\mathsf{mary},\mathsf{bill})^{\mathsf{T}} \\
\mathsf{loves}(\mathsf{john},\mathsf{mary})^{\mathsf{T}} \\
\bot
\end{array}$$

11.4 Soundness and Termination of Tableaux

As always we need to convince ourselves that the calculus is sound, otherwise, tableau proofs do not guarantee validity, which we are after. Since we are now in a refutation setting we cannot just show that the inference rules preserve validity: we care about unsatisfiability (which is the dual notion to validity), as we want to show the initial labeled formula to be unsatisfiable. Before we can do this, we have to ask ourselves, what it means to be (un)-satisfiable for a labeled formula or a tableau.

Soundness (Tableau)

- **Idea:** A test calculus is refutation sound, iff its inference rules preserve satisfiability and the goal formulae are unsatisfiable.
- **Definition 11.4.1.** A labeled formula $\mathbf{A}^\alpha$ is valid under φ, iff $\mathcal{I}_\varphi(\mathbf{A}) = \alpha$.
- **Definition 11.4.2.** A tableau $\mathcal{T}$ is satisfiable, iff there is a satisfiable branch $\mathcal{P}$ in $\mathcal{T}$, i.e. if the set of formulae on $\mathcal{P}$ is satisfiable.
- **Lemma 11.4.3.** *$\mathcal{T}_0$ rules transform satisfiable tableaux into satisfiable ones.*
- **Theorem 11.4.4 (Soundness).** *$\mathcal{T}_0$ is sound, i.e. $\Phi \subseteq wff_0(\mathcal{V}_0)$ valid, if there is a closed tableau $\mathcal{T}$ for Φ^{F}.*
- *Proof:* by contradiction
  1. Suppose Φ isfalsifiable $\widehat{=}$ not valid.
  2. Then the initial tableau is satisfiable, (Φ^{F} satisfiable)
  3. so $\mathcal{T}$ is satisfiable, by Lemma 11.4.3.
  4. Thus there is a satisfiable branch (by definition)
  5. but all branches are closed ($\mathcal{T}$ closed)
- **Theorem 11.4.5 (Completeness).** *$\mathcal{T}_0$ is complete, i.e. if $\Phi \subseteq wff_0(\mathcal{V}_0)$ is valid, then there is a closed tableau $\mathcal{T}$ for Φ^{F}.*

  Proof sketch: Proof difficult/interesting; see Corollary A.2.2

Thus we only have to prove Lemma 11.4.3, this is relatively easy to do. For instance for the first rule: if we have a tableau that contains $(\mathbf{A} \wedge \mathbf{B})^{\mathsf{T}}$ and is satisfiable, then it must have a satisfiable branch. If $(\mathbf{A} \wedge \mathbf{B})^{\mathsf{T}}$ is not on this branch, the tableau extension will not change satisfiability, so we can assume that it is on the satisfiable branch and thus $\mathcal{I}_\varphi(\mathbf{A} \wedge \mathbf{B}) = \mathsf{T}$ for some variable assignment φ. Thus $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{T}$ and $\mathcal{I}_\varphi(\mathbf{B}) = \mathsf{T}$, so after the extension (which adds the formulae $\mathbf{A}^{\mathsf{T}}$ and $\mathbf{B}^{\mathsf{T}}$ to the branch), the branch is still satisfiable. The cases for the other rules are similar.

The next result is a very important one, it shows that there is a procedure (the tableau procedure) that will always terminate and answer the question whether a given propositional formula is valid or not. This is very important, since other logics (like the often-studied first-order logic) does not enjoy this property.

▷ Termination for Tableaux

▷ **Lemma 11.4.6.** *$\mathcal{T}_0$ terminates, i.e. every $\mathcal{T}_0$ tableau becomes saturated after finitely many rule applications.*

▷ *Proof:* By examining the rules wrt. a measure μ

1. Let us call a labeled formulae $\mathbf{A}^\alpha$ worked off in a tableau $\mathcal{T}$, if a $\mathcal{T}_0$ rule has already been applied to it.
2. It is easy to see that applying rules to worked off formulae will only add formulae that are already present in its branch.
3. Let $\mu(\mathcal{T})$ be the number of connectives in labeled formulae in $\mathcal{T}$ that are not worked off.
4. Then each rule application to a labeled formula in $\mathcal{T}$ that is not worked off reduces $\mu(\mathcal{T})$ by at least one. (inspect the rules)
5. At some point the tableau only contains worked off formulae and literals.
6. Since there are only finitely many literals in $\mathcal{T}$, so we can only apply $\mathcal{T}_0\bot$ a finite number of times.

▷ **Corollary 11.4.7.** *$\mathcal{T}_0$ induces a decision procedure for validity in PL^0.*

Proof: We combine the results so far

▷ 1. By Lemma 11.4.6 it is decidable whether $\vdash_{\mathcal{T}_0}\mathbf{A}$
2. By soundness (Theorem 11.4.4) and completeness (Theorem 11.4.5), $\vdash_{\mathcal{T}_0}\mathbf{A}$ iff $\mathbf{A}$ is valid.

Note: The proof above only works for the "base $\mathcal{T}_0$" because (only) there the rules do not "copy". A rule like

$$\frac{\mathbf{A} \Leftrightarrow \mathbf{B}^{\mathsf{T}}}{\begin{array}{c|c} \mathbf{A}^{\mathsf{T}} & \mathbf{A}^{\mathsf{F}} \\ \mathbf{B}^{\mathsf{T}} & \mathbf{B}^{\mathsf{F}} \end{array}}$$

does, and in particular the number of non-worked-off variables below the line is larger than above the line. For such rules, we would have a more intricate version of μ which – instead of returning a natural number – returns a more complex object; a multiset of numbers. would work here. In our proof we are just assuming that the defined connectives have already eliminated. The tableau calculus basically computes the disjunctive normal form: every branch is a disjunct that is a conjunction of literals. The method relies on the fact that a DNF is unsatisfiable, iff each literal is, i.e. iff each branch contains a contradiction in form of a pair of opposite literals.

11.5 Resolution for Propositional Logic

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/23712`.

The next calculus is a test calculus based on the conjunctive normal form: the resolution calculus. In contrast to the tableau method, it does not compute the normal form as it goes along, but has a pre-processing step that does this and a single inference rule that maintains the normal form. The goal of this calculus is to derive the empty clause, which is unsatisfiable.

Another Test Calculus: Resolution

- **Definition 11.5.1.** A clause is a disjunction $l_1^{\alpha_1} \vee \ldots \vee l_n^{\alpha_n}$ of literals. We will use $\Box$ for the "empty" disjunction (no disjuncts) and call it the empty clause. A clause with exactly one literal is called a unit clause.
- **Definition 11.5.2 (Resolution Calculus).** The resolution calculus $\mathcal{R}_0$ operates a clause sets via a single inference rule:

$$\frac{P^{\mathsf{T}} \vee \mathbf{A} \quad P^{\mathsf{F}} \vee \mathbf{B}}{\mathbf{A} \vee \mathbf{B}}\mathcal{R}$$

  This rule allows to add the resolvent (the clause below the line) to a clause set which contains the two clauses above. The literals P^{T} and P^{F} are called cut literals.
- **Definition 11.5.3 (Resolution Refutation).** Let S be a clause set, then we call an $\mathcal{R}_0$-derivation of $\Box$ from S $\mathcal{R}_0$-refutation and write $\mathcal{D}\colon S \vdash_{\mathcal{R}_0} \Box$.

Clause Normal Form Transformation (A calculus)

- **Definition 11.5.4.** We will often write a clause set $\{C_1, \ldots, C_n\}$ as $C_1 ; \ldots ; C_n$, use $S ; T$ for the union of the clause sets S and T, and $S ; C$ for the extension by a clause C.
- **Definition 11.5.5 (Transformation into Clause Normal Form).** The CNF transformation calculus CNF_0 consists of the following four inference rules on sets of labeled formulae.

$$\frac{\mathbf{C} \vee (\mathbf{A} \vee \mathbf{B})^{\mathsf{T}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{T}} \vee \mathbf{B}^{\mathsf{T}}} \quad \frac{\mathbf{C} \vee (\mathbf{A} \vee \mathbf{B})^{\mathsf{F}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{F}} ; \mathbf{C} \vee \mathbf{B}^{\mathsf{F}}} \qquad \frac{\mathbf{C} \vee \neg \mathbf{A}^{\mathsf{T}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{F}}} \quad \frac{\mathbf{C} \vee \neg \mathbf{A}^{\mathsf{F}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{T}}}$$

- **Definition 11.5.6.** We write $CNF_0(\mathbf{A}^\alpha)$ for the set of all clauses derivable from $\mathbf{A}^\alpha$ via the rules above.

that the **C**-terms in the definition of the inference rules are necessary, since we assumed that the assumptions of the inference rule must match full clauses. The **C** terms are used with the convention that they are optional. So that we can also simplify $(\mathbf{A} \vee \mathbf{B})^{\mathsf{T}}$ to $\mathbf{A}^{\mathsf{T}} \vee \mathbf{B}^{\mathsf{T}}$.

Background: The background behind this notation is that $\mathbf{A}$ and $T \vee \mathbf{A}$ are equivalent for any $\mathbf{A}$. That allows us to interpret the **C**-terms in the assumptions as T and thus leave them out.

The clause normal form translation as we have formulated it here is quite frugal; we have left out rules for the connectives $\vee$, $\Rightarrow$, and $\Leftrightarrow$, relying on the fact that formulae containing these

connectives can be translated into ones without before CNF transformation. The advantage of having a calculus with few inference rules is that we can prove meta properties like soundness and completeness with less effort (these proofs usually require one case per inference rule). On the other hand, adding specialized inference rules makes proofs shorter and more readable.
Fortunately, there is a way to have your cake and eat it. Derivable inference rules have the property that they are formally redundant, since they do not change the expressive power of the calculus. Therefore we can leave them out when proving meta-properties, but include them when actually using the calculus.

Derived Rules of Inference

- **Definition 11.5.7.** An inference rule $\frac{\mathbf{A}_1 \quad \ldots \quad \mathbf{A}_n}{\mathbf{C}}$ is called derivable (or a derived rule) in a calculus $\mathcal{C}$, if there is a $\mathcal{C}$ derivation $\mathbf{A}_1, \ldots, \mathbf{A}_n \vdash_{\mathcal{C}} \mathbf{C}$.
- **Idea:** Derived rules make proofs shorter.
- **Example 11.5.8.**

$$\dfrac{\dfrac{\dfrac{\mathbf{C} \vee (\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{T}}}{\mathbf{C} \vee (\neg \mathbf{A} \mathbf{B})^{\mathsf{T}}}}{\mathbf{C} \vee \neg \mathbf{A}^{\mathsf{T}} \vee \mathbf{B}^{\mathsf{T}}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{F}} \vee \mathbf{B}^{\mathsf{T}}} \quad \rightsquigarrow \quad \frac{\mathbf{C} \vee (\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{T}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{F}} \vee \mathbf{B}^{\mathsf{T}}}$$

- **Other Derived CNF Rules:**

$$\frac{\mathbf{C} \vee (\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{T}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{F}} \vee \mathbf{B}^{\mathsf{T}}} \quad \frac{\mathbf{C} \vee (\mathbf{A} \Rightarrow \mathbf{B})^{\mathsf{F}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{T}} \,;\, \mathbf{C} \vee \mathbf{B}^{\mathsf{F}}} \qquad \frac{\mathbf{C} \vee (\mathbf{A} \wedge \mathbf{B})^{\mathsf{T}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{T}} \,;\, \mathbf{C} \vee \mathbf{B}^{\mathsf{T}}} \quad \frac{\mathbf{C} \vee (\mathbf{A} \wedge \mathbf{B})^{\mathsf{F}}}{\mathbf{C} \vee \mathbf{A}^{\mathsf{F}} \vee \mathbf{B}^{\mathsf{F}}}$$

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 363 2023-09-20

With these derivable rules, theorem proving becomes quite efficient. To get a better understanding of the calculus, we look at an example: we prove an axiom of the Hilbert Calculus we have studied above.

Example: Proving Axiom S with Resolution

- **Example 11.5.9.** Clause Normal Form transformation

$$\dfrac{\dfrac{\dfrac{((P \Rightarrow Q \Rightarrow R) \Rightarrow (P \Rightarrow Q) \Rightarrow P \Rightarrow R)^{\mathsf{F}}}{(P \Rightarrow Q \Rightarrow R)^{\mathsf{T}} \,;\, ((P \Rightarrow Q) \Rightarrow P \Rightarrow R)^{\mathsf{F}}}}{P^{\mathsf{F}} \vee (Q \Rightarrow R)^{\mathsf{T}} \,;\, (P \Rightarrow Q)^{\mathsf{T}} \,;\, (P \Rightarrow R)^{\mathsf{F}}}}{P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} \,;\, P^{\mathsf{F}} \vee Q^{\mathsf{T}} \,;\, P^{\mathsf{T}} \,;\, R^{\mathsf{F}}}$$

Result $\{P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}}, P^{\mathsf{F}} \vee Q^{\mathsf{T}}, P^{\mathsf{T}}, R^{\mathsf{F}}\}$

- **Example 11.5.10.** Resolution Proof

1	$P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}}$	initial
2	$P^{\mathsf{F}} \vee Q^{\mathsf{T}}$	initial
3	P^{T}	initial
4	R^{F}	initial
5	$P^{\mathsf{F}} \vee Q^{\mathsf{F}}$	resolve 1.3 with 4.1
6	Q^{F}	resolve 5.1 with 3.1
7	P^{F}	resolve 2.2 with 6.1
8	$\square$	resolve 7.1 with 3.1

Clause Set Simplification

▷ **Observation:** Let Δ be a clause set, l a literal, and Δ' be Δ where

- ▷ all clauses $l \vee C$ have been removed and
- ▷ and all clauses $\bar{l} \vee C$ have been shortened to C.

Then Δ is satisfiable, iff Δ' is. We call Δ' the clause set simplification of Δ wrt. l.

▷ **Corollary 11.5.11.** *Adding clause set simplification wrt. unit clauses to $\mathcal{R}_0$ does not affect soundness and completeness.*

▷ This is almost always a good idea! (clause set simplification is cheap)

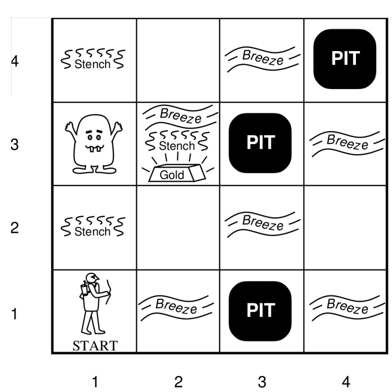
11.6 Killing a Wumpus with Propositional Inference

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/23713`.

Let us now consider an extended example, where we also address the question how inference in PL^0 – here resolution is embedded into the rational agent metaphor we use in AI-1: we come back to the Wumpus world.

Applying Propositional Inference: Where is the Wumpus?

▷ **Example 11.6.1 (Finding the Wumpus).** The situation and what the agent knows



1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 $\boxed{A}$ S OK	2,2	3,2	4,2
1,1 V OK	2,1 B V OK	3,1	4,1

- What should the agent do next and why?
- **One possibility**: Convince yourself that the Wumpus is in $[1,3]$ and shoot it.
- What is the general mechanism here? (for the agent function)

Before we come to the general mechanism, we will go into how we would "convince ourselves that the Wumpus is in $[1,3]$.

Where is the Wumpus? Our Knowledge

- **Idea:** We formalize the knowledge about the Wumpus world in PL^0 and use a test calculus to check for entailment.
- **Simplification:** We worry only about the Wumpus and stench: $S_{i,j} \mathrel{\widehat{=}}$ *stench in* $[i,j]$, $W_{i,j} \mathrel{\widehat{=}}$ Wumpus in $[i,j]$.
- **Propositions whose value we know:** $\neg S_{1,1}$, $\neg W_{1,1}$, $\neg S_{2,1}$, $\neg W_{2,1}$, $S_{1,2}$, $\neg W_{1,2}$.
- **Knowledge about the Wumpus and smell:** From *Cell adjacent to Wumpus: Stench (else: None)*, we get

$$R_1{:=}\neg S_{1,1} \Rightarrow \neg W_{1,1} \wedge \neg W_{1,2} \wedge \neg W_{2,1}$$
$$R_2{:=}\neg S_{2,1} \Rightarrow \neg W_{1,1} \wedge \neg W_{2,1} \wedge \neg W_{2,2} \wedge \neg W_{3,1}$$
$$R_3{:=}\neg S_{1,2} \Rightarrow \neg W_{1,1} \wedge \neg W_{1,2} \wedge \neg W_{2,2} \wedge \neg W_{1,3}$$
$$R_4{:=}S_{1,2} \Rightarrow (W_{1,3} \vee W_{2,2} \vee W_{1,1})$$
$$\vdots$$

- **To show:**

$R_1, R_2, R_3, R_4 \models W_{1,3}$ (we will use resolution)

The first in is to compute the clause normal form of the relevant knowledge.

And Now Using Resolution Conventions

- We obtain the clause set Δ composed of the following clauses:
  - **Propositions whose value we know**: $S_{1,1}{}^{\mathsf{F}}$, $W_{1,1}{}^{\mathsf{F}}$, $S_{2,1}{}^{\mathsf{F}}$, $W_{2,1}{}^{\mathsf{F}}$, $S_{1,2}{}^{\mathsf{T}}$,

$W_{1,2}{}^{\mathsf{F}}$

- **Knowledge about the Wumpus and smell**:

from	clauses
R_1	$S_{1,1}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{F}}$, $S_{1,1}{}^{\mathsf{T}} \vee W_{1,2}{}^{\mathsf{F}}$, $S_{1,1}{}^{\mathsf{T}} \vee W_{2,1}{}^{\mathsf{F}}$
R_2	$S_{2,1}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{F}}$, $S_{2,1}{}^{\mathsf{T}} \vee W_{2,1}{}^{\mathsf{F}}$, $S_{2,1}{}^{\mathsf{T}} \vee W_{2,2}{}^{\mathsf{F}}$, $S_{2,1}{}^{\mathsf{T}} \vee W_{3,1}{}^{\mathsf{F}}$
R_3	$S_{1,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{F}}$, $S_{1,2}{}^{\mathsf{T}} \vee W_{1,2}{}^{\mathsf{F}}$, $S_{1,2}{}^{\mathsf{T}} \vee W_{2,2}{}^{\mathsf{F}}$, $S_{1,2}{}^{\mathsf{T}} \vee W_{1,3}{}^{\mathsf{F}}$
R_4	$S_{1,2}{}^{\mathsf{F}} \vee W_{1,3}{}^{\mathsf{T}} \vee W_{2,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{T}}$

- **Negated goal formula**: $W_{1,3}{}^{\mathsf{F}}$

Given this clause normal form, we only need to find generate empty clause via repeated applications of the resolution rule.

Resolution Proof Killing the Wumpus!

- **Example 11.6.2 (Where is the Wumpus).** We show a derivation that proves that he is in $(1,3)$.
  - *Assume the Wumpus is not in $(1,3)$. Then either there's no stench in $(1,2)$, or the Wumpus is in some other neigbor cell of $(1,2)$.*
    - Parents: $W_{1,3}{}^{\mathsf{F}}$ and $S_{1,2}{}^{\mathsf{F}} \vee W_{1,3}{}^{\mathsf{T}} \vee W_{2,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{T}}$.
    - Resolvent: $S_{1,2}{}^{\mathsf{F}} \vee W_{2,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{T}}$.
  - *There's a stench in $(1,2)$, so it must be another neighbor.*
    - Parents: $S_{1,2}{}^{\mathsf{T}}$ and $S_{1,2}{}^{\mathsf{F}} \vee W_{2,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{T}}$.
    - Resolvent: $W_{2,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{T}}$.
  - *We've been to $(1,1)$, and there's no Wumpus there, so it can't be $(1,1)$.*
    - Parents: $W_{1,1}{}^{\mathsf{F}}$ and $W_{2,2}{}^{\mathsf{T}} \vee W_{1,1}{}^{\mathsf{T}}$.
    - Resolvent: $W_{2,2}{}^{\mathsf{T}}$.
  - *There is no stench in $(2,1)$ so it can't be $(2,2)$ either, in contradiction.*
    - Parents: $S_{2,1}{}^{\mathsf{F}}$ and $S_{2,1}{}^{\mathsf{T}} \vee W_{2,2}{}^{\mathsf{F}}$.
    - Resolvent: $W_{2,2}{}^{\mathsf{F}}$.
    - Parents: $W_{2,2}{}^{\mathsf{F}}$ and $W_{2,2}{}^{\mathsf{T}}$.
    - Resolvent: $\square$.

  As resolution is sound, we have shown that indeed $R_1, R_2, R_3, R_4 \models W_{1,3}$.

Now that we have seen how we can use propositional inference to derive consequences of the percepts and world knowledge, let us come back to the question of a general mechanism for agent functions with propositional inference.

Where does the Conjecture $W_{1,3}{}^{\mathsf{F}}$ come from?

- **Question:** Where did the $W_{1,3}{}^{\mathsf{F}}$ come from?

▷ **Observation 11.6.3.** *We need a general mechanism for making conjectures.*

▷ **Idea:** Interpret the Wumpus world as a search problem $\mathcal{P}:=\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ where

- ▷ the states $\mathcal{S}$ are given by the cells (and agent orientation) and
- ▷ the actions $\mathcal{A}$ by the possible actions of the agent.

Use tree search as the main agent function and a test calculus for testing all dangers (pits), opportunities (gold) and the Wumpus.

▷ **Example 11.6.4 (Back to the Wumpus).** In Example 11.6.1, the agent is in $[1,2]$, it has perceived stench, and the possible actions include shoot, and goForward. Evaluating either of these leads to the conjecture $W_{1,3}$. And since $W_{1,3}$ is entailed, the action shoot probably comes out best, heuristically.

▷ **Remark:** Analogous to the backtracking with inference algorithm from CSP.

Admittedly, the search framework from chapter 6 does not quite cover the agent function we have here, since that assumes that the world is fully observable, which the Wumpus world is emphatically not. But it already gives us a good impression of what would be needed for the "general mechanism".

Summary

▷ Every propositional formula can be brought into conjunctive normal form (CNF), which can be identified with a set of clauses.

▷ The tableau and resolution calculi are deduction procedures based on trying to derive a contradiction from the negated theorem (a closed tableau or the empty clause). They are refutation complete, and can be used to prove $\mathsf{KB}\models\mathbf{A}$ by showing that $\mathsf{KB}\cup\{\neg\mathbf{A}\}$ is unsatisfiable.

Excursion: A full analysis of any calculus needs a completeness proof. We will not cover this in AI-1, but provide one for the calculi introduced so far in**??**.

Chapter 12

Formal Systems: Syntax, Semantics, Entailment, and Derivation in General

We will now take a more abstract view and introduce the necessary prerequisites of abstract rule systems. We will also take the opportunity to discuss the quality criteria for calculi.

Recap: General Aspects of Propositional Logic

- **There are many ways to define Propositional Logic:**
  - We chose $\wedge$ and $\neg$ as primitive, and many others as defined.
  - We could have used $\vee$ and $\neg$ just as well.
  - We could even have used only one connective e.g. negated conjunction $\uparrow$ or disjunction NOR and defined $\wedge$, $\vee$, and $\neg$ via $\uparrow$ and NOR respectively.

$\uparrow$	$\top$	$\bot$
$\top$	F	T
$\bot$	T	T

NOR	$\top$	$\bot$
$\top$	F	F
$\bot$	F	T

$\neg a$	$a \uparrow a$	a NOR a
ab	$a \uparrow b \uparrow a \uparrow b$	a NOR ab NOR b
ab	$a \uparrow a \uparrow b \uparrow b$	a NOR b NOR a NOR b

- **Observation:** The set $wff_0(\mathcal{V}_0)$ of well-formed propositional formulae is a formal language over the alphabet given by $\mathcal{V}_0$, the connectives, and brackets.
- **Recall:** We are mostly interested in
  - satisfiability i.e. whether $\mathcal{M} \models^{\varphi} \mathbf{A}$, and
  - entailment i.e whether $\mathbf{A} \models \mathbf{B}$.
- **Observation:** In particular, the inductive/compositional nature of $wff_0(\mathcal{V}_0)$ and $\mathcal{I}_{\varphi} \colon wff_0(\mathcal{V}_0) \rightarrow \mathcal{D}_0$ are secondary.
- **Idea:** Concentrate on language, models $(\mathcal{M}, \varphi)$, and satisfiability.

The notion of a logical system is at the basis of the field of logic. In its most abstract form, a logical system consists of a formal language, a class of models, and a satisfaction relation between models and expressions of the formal language. The satisfaction relation tells us when an expression is deemed true in this model.

Logical Systems

- **Definition 12.0.1.** A logical system (or simply a logic) is a triple $\mathcal{L}:=\langle\mathcal{L},\mathcal{K},\models\rangle$, where $\mathcal{L}$ is a formal language, $\mathcal{K}$ is a set and $\models\subseteq\mathcal{K}\times\mathcal{L}$. Members of $\mathcal{L}$ are called formulae of $\mathcal{L}$, members of $\mathcal{K}$ models for $\mathcal{L}$, and $\models$ the satisfaction relation.
- **Example 12.0.2 (Propositional Logic).**
  $\langle wff(\Sigma_{PL^0},\mathcal{V}_{PL^0}),\mathcal{K},\models\rangle$ is a logical system, if we define $\mathcal{K}:=\mathcal{V}_0\rightharpoonup\mathcal{D}_0$ (the set of variable assignments) and $\varphi\models\mathbf{A}$ iff $\mathcal{I}_\varphi(\mathbf{A})=\mathsf{T}$.
- **Definition 12.0.3.** Let $\langle\mathcal{L},\mathcal{K},\models\rangle$ be a logical system, $\mathcal{M}\in\mathcal{K}$ be a model and $\mathbf{A}\in\mathcal{L}$ a formula, then we say that $\mathbf{A}$ is
  - satisfied by $\mathcal{M}$, iff $\mathcal{M}\models\mathbf{A}$.
  - falsified by $\mathcal{M}$, iff $\mathcal{M}\not\models\mathbf{A}$.
  - satisfiable in $\mathcal{K}$, iff $\mathcal{M}\models\mathbf{A}$ for some $\mathcal{M}\in\mathcal{K}$.
  - valid in $\mathcal{K}$ (write $\models\mathcal{M}$), iff $\mathcal{M}\models\mathbf{A}$ for all $\mathcal{M}\in\mathcal{K}$.
  - falsifiable in $\mathcal{K}$, iff $\mathcal{M}\not\models\mathbf{A}$ for some $\mathcal{M}\in\mathcal{K}$.
  - unsatisfiable in $\mathcal{K}$, iff $\mathcal{M}\not\models\mathbf{A}$ for all $\mathcal{M}\in\mathcal{K}$.

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 373 2023-09-20

Let us now turn to the syntactical counterpart of the entailment relation: derivability in a calculus. Again, we take care to define the concepts at the general level of logical systems.
The intuition of a calculus is that it provides a set of syntactic rules that allow to reason by considering the form of propositions alone. Such rules are called inference rules, and they can be strung together to derivations — which can alternatively be viewed either as sequences of formulae where all formulae are justified by prior formulae or as trees of inference rule applications. But we can also define a calculus in the more general setting of logical systems as an arbitrary relation on formulae with some general properties. That allows us to abstract away from the homomorphic setup of logics and calculi and concentrate on the basics.

Derivation Relations and Inference Rules

- **Definition 12.0.4.** Let $\mathcal{L}:=\langle\mathcal{L},\mathcal{K},\models\rangle$ be a logical system, then we call a relation $\vdash\subseteq\mathcal{P}(\mathcal{L})\times\mathcal{L}$ a derivation relation for $\mathcal{L}$, if
  - $\mathcal{H}\vdash\mathbf{A}$, if $\mathbf{A}\in\mathcal{H}$ ($\vdash$ is proof reflexive),
  - $\mathcal{H}\vdash\mathbf{A}$ and $\mathcal{H}'\cup\{\mathbf{A}\}\vdash\mathbf{B}$ imply $\mathcal{H}\cup\mathcal{H}'\vdash\mathbf{B}$ ($\vdash$ is proof transitive),
  - $\mathcal{H}\vdash\mathbf{A}$ and $\mathcal{H}\subseteq\mathcal{H}'$ imply $\mathcal{H}'\vdash\mathbf{A}$ ($\vdash$ is monotonic or admits weakening).
- **Definition 12.0.5.** We call $\langle\mathcal{L},\mathcal{K},\models,\mathcal{C}\rangle$ a formal system, iff $\mathcal{L}:=\langle\mathcal{L},\mathcal{K},\models\rangle$ is a logical system, and $\mathcal{C}$ a calculus for $\mathcal{L}$.
- **Definition 12.0.6.** Let $\mathcal{L}$ be the formal language of a logical system, then an inference rule over $\mathcal{L}$ is a decidable $n+1$ ary relation on $\mathcal{L}$. Inference rules are

traditionally written as

$$\frac{\mathbf{A}_1 \quad \ldots \quad \mathbf{A}_n}{\mathbf{C}}\mathcal{N}$$

where $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $\mathbf{C}$ are formula schemata for $\mathcal{L}$ and $\mathcal{N}$ is a name.

The $\mathbf{A}_i$ are called assumptions of $\mathcal{N}$, and $\mathbf{C}$ is called its conclusion.

▷ **Definition 12.0.7.** An inference rule without assumptions is called an axiom.

▷ **Definition 12.0.8.** Let $\mathcal{L}:=\langle\mathcal{L}, \mathcal{K}, \models\rangle$ be a logical system, then we call a set $\mathcal{C}$ of inference rules over $\mathcal{L}$ a calculus (or inference system) for $\mathcal{L}$.

With formula schemata we mean representations of sets of formulae, we use boldface uppercase letters as (meta)-variables for formulae, for instance the formula schema $\mathbf{A} \Rightarrow \mathbf{B}$ represents the set of formulae whose head is $\Rightarrow$.

Derivations

▷ **Definition 12.0.9.**Let $\mathcal{L}:=\langle\mathcal{L}, \mathcal{K}, \models\rangle$ be a logical system and $\mathcal{C}$ a calculus for $\mathcal{L}$, then a $\mathcal{C}$-derivation of a formula $\mathbf{C}\in\mathcal{L}$ from a set $\mathcal{H} \subseteq \mathcal{L}$ of hypotheses (write $\mathcal{H}\vdash_{\mathcal{C}}\mathbf{C}$) is a sequence $\mathbf{A}_1, \ldots, \mathbf{A}_m$ of $\mathcal{L}$-formulae, such that

- ▷ $\mathbf{A}_m = \mathbf{C}$, (derivation culminates in $\mathbf{C}$)
- ▷ for all $1\leq i\leq m$, either $\mathbf{A}_i\in\mathcal{H}$, or (hypothesis)
- ▷ there is an inference rule $\frac{\mathbf{A}_{l_1} \quad \ldots \quad \mathbf{A}_{l_k}}{\mathbf{A}_i}$ in $\mathcal{C}$ with $l_j<i$ for all $j\leq k$. (rule application)

We can also see a derivation as a derivation tree, where the $\mathbf{A}_{l_j}$ are the children of the node $\mathbf{A}_k$.

▷ **Example 12.0.10.**

In the propositional Hilbert calculus $\mathcal{H}^0$ we have the derivation $P\vdash_{\mathcal{H}^0}Q\Rightarrow P$: the sequence is $P\Rightarrow Q\Rightarrow P, P, Q\Rightarrow P$ and the corresponding tree on the right.

$$\dfrac{\dfrac{}{P \Rightarrow Q \Rightarrow P}K \qquad P}{Q \Rightarrow P}MP$$

Inference rules are relations on formulae represented by formula schemata (where boldface, uppercase letters are used as meta-variables for formulae). For instance, in Example 12.0.10 the inference rule $\frac{\mathbf{A} \Rightarrow \mathbf{B} \quad \mathbf{A}}{\mathbf{B}}$ was applied in a situation, where the meta-variables $\mathbf{A}$ and $\mathbf{B}$ were instantiated by the formulae P and $Q \Rightarrow P$.

As axioms do not have assumptions, they can be added to a derivation at any time. This is just what we did with the axioms in Example 12.0.10.

Formal Systems

▷ Let $\langle\mathcal{L}, \mathcal{K}, \models\rangle$ be a logical system and $\mathcal{C}$ a calculus, then $\vdash_{\mathcal{C}}$ is a derivation relation and thus $\langle\mathcal{L}, \mathcal{K}, \models, \vdash_{\mathcal{C}}\rangle$ a derivation system.

▷ Therefore we will sometimes also call $\langle \mathcal{L}, \mathcal{K}, \models, \mathcal{C} \rangle$ a formal system, iff $\mathcal{L} := \langle \mathcal{L}, \mathcal{K}, \models \rangle$ is a logical system, and $\mathcal{C}$ a calculus for $\mathcal{L}$.

▷ **Definition 12.0.11.** Let $\mathcal{C}$ be a calculus, then a $\mathcal{C}$-derivation $\emptyset \vdash_{\mathcal{C}} \mathbf{A}$ is called a proof of $\mathbf{A}$ and if one exists (write $\vdash_{\mathcal{C}} \mathbf{A}$) then $\mathbf{A}$ is called a $\mathcal{C}$-theorem.

Definition 12.0.12. The act of finding a proof for a formula $\mathbf{A}$ is called proving $\mathbf{A}$.

▷ **Definition 12.0.13.** An inference rule $\mathcal{I}$ is called admissible in a calculus $\mathcal{C}$, if the extension of $\mathcal{C}$ by $\mathcal{I}$ does not yield new theorems.

▷ **Definition 12.0.14.** An inference rule $\frac{\mathbf{A}_1 \quad \ldots \quad \mathbf{A}_n}{\mathbf{C}}$ is called derivable (or a derived rule) in a calculus $\mathcal{C}$, if there is a $\mathcal{C}$ derivation $\mathbf{A}_1, \ldots, \mathbf{A}_n \vdash_{\mathcal{C}} \mathbf{C}$.

▷ **Observation 12.0.15.** *Derivable inference rules are admissible, but not the other way around.*

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 376 2023-09-20

The notion of a formal system encapsulates the most general way we can conceptualize a system with a calculus, i.e. a system in which we can do "formal reasoning".

Chapter 13

Propositional Reasoning: SAT Solvers

13.1 Introduction

A Video Nugget covering this section can be found at https://fau.tv/clip/id/25019.

Reminder: Our Agenda for Propositional Logic

- ▷ **chapter 10**: Basic definitions and concepts; machine-oriented calculi
  - ▷ Sets up the framework. Tableaux and resolution are the quintessential reasoning procedure underlying most successful SAT solvers.
- ▷ **This chapter**: The Davis Putnam procedure and clause learning.
  - ▷ State-of-the-art algorithms for reasoning about propositional logic, and an important observation about how they behave.

SAT: The Propositional Satisfiability Problem

- ▷ **Definition 13.1.1.** The SAT problem (SAT): Given a propositional formula $\mathbf{A}$, decide whether or not $\mathbf{A}$ is satisfiable. We denote the class of all SAT problems with SAT
- ▷ The SAT problem was the first problem proved to be **NP**-complete!
- ▷ $\mathbf{A}$ is commonly assumed to be in CNF. This is without loss of generality, because any $\mathbf{A}$ can be transformed into a satisfiability-equivalent CNF formula (cf. chapter 10) in polynomial time.
- ▷ Active research area, annual SAT conference, lots of tools etc. available: http://www.satlive.org/
- ▷ **Definition 13.1.2.** Tools addressing SAT are commonly referred to as SAT solvers.

▷ **Recall:** To decide whether KB $\models \mathbf{A}$, decide satisfiability of $\theta := \mathrm{KB} \cup \{\neg\mathbf{A}\}$: θ is unsatisfiable iff KB $\models \mathbf{A}$.

▷ **Consequence:** Deduction can be performed using SAT solvers.

SAT vs. CSP

▷ **Recall:** Constraint network $\langle V, D, C\rangle$ has variables $v \in V$ with finite domains $D_v \in D$, and binary constraints $C_{uv} \in C$ which are relations over u, v specifying the permissible combined assignments to u and v. One extension is to allow constraints of higher arity.

▷ **Observation 13.1.3 (SAT: A kind of CSP).** *SAT can be viewed as a CSP problem in which all variable domains are Boolean, and the constraints have unbounded arity.*

▷ **Theorem 13.1.4 (Encoding CSP as SAT).** *Given any constraint network $\mathcal{C}$, we can in low order polynomial time construct a CNF formula $\mathbf{A}(\mathcal{C})$ that is satisfiable iff $\mathcal{C}$ is solvable.*

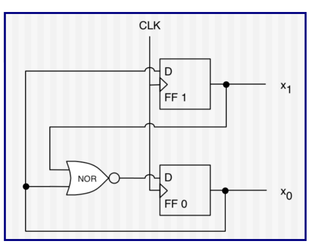
▷ *Proof:* We design a formula, relying on known transformation to CNF

1. encode multi-XOR for each variable
2. encode each constraint by DNF over relation
3. Running time: $\mathcal{O}(nd^2 + md^2)$ where n is the number of variables, d the domain size, and m the number of constraints.

▷ **Upshot:** Anything we can do with CSP, we can (in principle) do with SAT.

Example Application: Hardware Verification Verification

▷ **Example 13.1.5 (Hardware Verification).**



▷ Counter, repeatedly from $c = 0$ to $c = 2$.

▷ 2 bits x_1 and x_0; $c = 2 * x_1 + x_0$.

▷ (FF$\widehat{=}$ Flip-Flop, D $\widehat{=}$ Data IN, CLK $\widehat{=}$ Clock)

▷ **To Verify**: If $c < 3$ in current clock cycle, then $c < 3$ in next clock cycle.

▷ **Step 1:** Encode into propositional logic.

- ▷ **Propositions**: x_1, x_0; and y_1, y_0 (value in next cycle).
- ▷ **Transition relation**: $y_1 \Leftrightarrow y_0$; $y_0 \Leftrightarrow (\neg(x_1 \vee x_0))$.
- ▷ **Initial state**: $\neg(x_1 \wedge x_0)$.
- ▷ **Error property**: $x_1 \wedge y_0$.

▷ **Step 2:** Transform to CNF, encode as a clause set Δ.

- **Clauses**: ${y_1}^{\mathsf{F}} \vee {x_0}^{\mathsf{T}}$, ${y_1}^{\mathsf{T}} \vee {x_0}^{\mathsf{F}}$, ${y_0}^{\mathsf{T}} \vee {x_1}^{\mathsf{T}} \vee {x_0}^{\mathsf{T}}$, ${y_0}^{\mathsf{F}} \vee {x_1}^{\mathsf{F}}$, ${y_0}^{\mathsf{F}} \vee {x_0}^{\mathsf{F}}$, ${x_1}^{\mathsf{F}} \vee {x_0}^{\mathsf{F}}$, ${y_1}^{\mathsf{T}}$, ${y_0}^{\mathsf{T}}$.

- **Step 3:** Call a SAT solver (up next).

Our Agenda for This Chapter

- **The Davis-Putnam (Logemann-Loveland) Procedure:** How to systematically test satisfiability?
  - The quintessential SAT solving procedure, DPLL.
- **DPLL is (A Restricted Form of) Resolution:** How does this relate to what we did in the last chapter?
  - mathematical understanding of DPLL.
- **Why Did Unit Propagation Yield a Conflict?:** How can we analyze which mistakes were made in "dead" search branches?
  - Knowledge is power, see next.
- **Clause Learning:** How can we learn from our mistakes?
  - One of the key concepts, perhaps *the* key concept, underlying the success of SAT.
- **Phase Transitions – Where the Really Hard Problems Are:** Are *all* formulas "hard" to solve?
  - The answer is "no". And in some cases we can figure out exactly when they are/aren't hard to solve.

13.2 The Davis-Putnam (Logemann-Loveland) Procedure

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25026`.

The DPLL Procedure

- **Definition 13.2.1.** The Davis Putnam procedure (DPLL) is a SAT solver called on a clause set Δ and the empty assignment ϵ. It interleaves unit propagation (UP) and splitting:

function DPLL(Δ,I) **returns** a partial assignment I, or "unsatisfiable"
  /* Unit Propagation (UP) Rule: */
  Δ' := a copy of Δ; I' := I
  while Δ' contains a unit clause $C = P^\alpha$ **do**
    extend I' with $[\alpha/P]$, clause−set simplify Δ'
   /* Termination Test: */
   if $\square\in\Delta'$ **then return** "unsatisfiable"

if $\Delta' = \{\}$ **then return** I'
/* Splitting Rule: */
select some proposition P **for** which I' is not defined
I'' := I' extended with one truth value **for** P; Δ'' := a copy of Δ'; simplify Δ''
if I''' := DPLL(Δ'',I'') $\neq$ ''unsatisfiable'' **then return** I'''
I'' := I' extended with the other truth value **for** P; Δ'' := Δ'; simplify Δ''
return DPLL(Δ'',I'')

▷ In practice, of course one uses flags etc. instead of "copy".

DPLL: Example (Vanilla1)

▷ **Example 13.2.2 (UP and Splitting).** Let $\Delta := (P^{\mathsf{T}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}}; P^{\mathsf{F}} \vee Q^{\mathsf{F}}; R^{\mathsf{T}}; P^{\mathsf{T}} \vee Q^{\mathsf{F}})$

1. UP Rule: $R \mapsto \mathsf{T}$
   $P^{\mathsf{T}} \vee Q^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} ; P^{\mathsf{T}} \vee Q^{\mathsf{F}}$
2. Splitting Rule:

2a. $P \mapsto \mathsf{F}$ $Q^{\mathsf{T}} ; Q^{\mathsf{F}}$	2b. $P \mapsto \mathsf{T}$ Q^{F}
3a. UP Rule: $Q \mapsto \mathsf{T}$ $\square$ **returning "unsatisfiable"**	3b. UP Rule: $Q \mapsto \mathsf{F}$ clause set empty **returning "**$R \mapsto \mathsf{T}, P \mapsto \mathsf{T}, Q \mapsto \mathsf{F}$

DPLL: Example (Vanilla2)

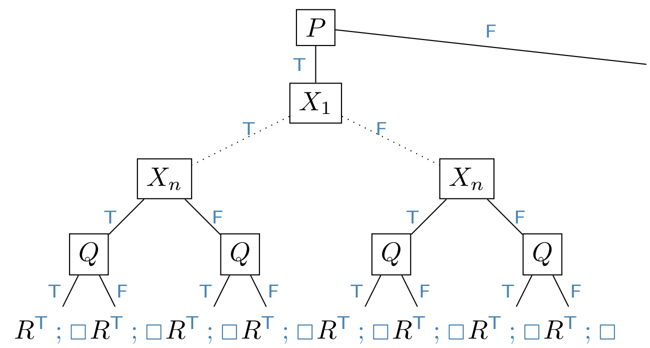
▷ **Observation:** Sometimes UP is all we need.

▷ **Example 13.2.3.** Let $\Delta := (Q^{\mathsf{F}} \vee P^{\mathsf{F}} ; P^{\mathsf{T}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} \vee S^{\mathsf{F}} ; Q^{\mathsf{T}} \vee S^{\mathsf{F}} ; R^{\mathsf{T}} \vee S^{\mathsf{F}} ; S^{\mathsf{T}})$

1. UP Rule: $S \mapsto \mathsf{T}$
   $Q^{\mathsf{F}} \vee P^{\mathsf{F}} ; P^{\mathsf{T}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; Q^{\mathsf{T}} ; R^{\mathsf{T}}$
2. UP Rule: $Q \mapsto \mathsf{T}$
   $P^{\mathsf{F}} ; P^{\mathsf{T}} \vee R^{\mathsf{F}} ; R^{\mathsf{T}}$
3. UP Rule: $R \mapsto \mathsf{T}$
   $P^{\mathsf{F}} ; P^{\mathsf{T}}$
4. UP Rule: $P \mapsto \mathsf{T}$
   $\square$

DPLL: Example (Redundance1)

▷ **Example 13.2.4.** We introduce some nasty redundance to make DPLL slow.
$\Delta := (P^F \vee Q^F \vee R^T ; P^F \vee Q^F \vee R^F ; P^F \vee Q^T \vee R^T ; P^F \vee Q^T \vee R^F)$
DPLL on $\Delta ; \Theta$ with $\Theta := ({X_1}^T \vee \ldots \vee {X_n}^T ; {X_1}^F \vee \ldots \vee {X_n}^F)$



Properties of DPLL

▷ **Unsatisfiable case:** What can we say if "unsatisfiable" is returned?

▷ In this case, we know that Δ is unsatisfiable: Unit propagation is *sound*, in the sense that it does not reduce the set of solutions.

▷ **Satisfiable case:** What can we say when a partial interpretation I is returned?

▷ Any extension of I to a complete interpretation satisfies Δ. (By construction, I suffices to satisfy all clauses.)

▷ Déjà Vu, Anybody?

▷ DPLL $\widehat{=}$ backtracking with inference, where inference $\widehat{=}$ unit propagation.

▷ Unit propagation is sound: It does not reduce the set of solutions.

▷ Running time is exponential in worst case, good variable/value selection strategies required.

13.3 DPLL $\widehat{=}$ (A Restricted Form of) Resolution

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/27022`.

In the last slide we have discussed the semantic properties of the DPLL procedure: DPLL is (refutation) sound and complete. Note that this is a theoretical resultin the sense that the algorithm is, but that does not mean that a particular implementation of DPLL might not contain bugs that affect sounds and completeness.

In the satisfiable case, DPLL returns a satisfying variable assignment, which we can check (in low-order polynomial time) but in the unsatisfiable case, it just reports on the fact that it has tried all branches and found nothing. This is clearly unsatisfactory, and we will address this situation now by presenting a way that DPLL can output a resolution proof in the unsatisfiable case.

UP $\widehat{=}$ Unit Resolution

▷ **Observation:** The unit propagation (UP) rule corresponds to a calculus:

while Δ' contains a unit clause $\{l\}$ **do**
  extend I' with the respective truth value **for** the proposition underlying l
  simplify Δ' /* remove false literals */

▷ **Definition 13.3.1 (Unit Resolution).** Unit resolution (UR) is the test calculus consisting of the following inference rule:

$$\frac{C \vee P^{\alpha} \quad P^{\beta} \quad \alpha \neq \beta}{C}\,\mathrm{UR}$$

▷ Unit propagation $\widehat{=}$ resolution restricted to cases where one parent is unit clause.

▷ **Observation 13.3.2 (Soundness).** *UR is refutation sound.* *(since resolution is)*

▷ **Observation 13.3.3 (Completeness).** *UR is not refutation complete (alone).*

▷ **Example 13.3.4.** $P^{\mathsf{T}} \vee Q^{\mathsf{T}}\,;\, P^{\mathsf{T}} \vee Q^{\mathsf{F}}\,;\, P^{\mathsf{F}} \vee Q^{\mathsf{T}}\,;\, P^{\mathsf{F}} \vee Q^{\mathsf{F}}$ is satisfiable but UR cannot derive the empty clause $\square$.

▷ UR makes only limited inferences, as long as there are unit clauses. It does not guarantee to infer everything that can be inferred.

DPLL vs. Resolution

▷ **Definition 13.3.5.** We define the number of decisions of a DPLL run as the total number of times a truth value was set by either unit propagation or splitting.

▷ **Theorem 13.3.6.** *If DPLL returns "unsatisfiable" on Δ, then $\Delta \vdash_{\mathcal{R}_0} \square$ with a resolution proof whose length is at most the number of decisions.*

▷ *Proof:* Consider first DPLL without UP

1. Consider any leaf node N, for proposition X, both of whose truth values directly result in a clause C that has become empty.
2. Then for $X = \mathsf{F}$ the respective clause C must contain X^{T}; and for $X = \mathsf{T}$ the respective clause C must contain X^{F}. Thus we can resolve these two clauses to a clause $C(N)$ that does not contain X.
3. $C(N)$ can contain only the negations of the decision literals $l_1, \ldots, l_k$ above N. Remove N from the tree, then iterate the argument. Once the tree is empty, we have derived the empty clause.
4. Unit propagation can be simulated via applications of the splitting rule, choosing a proposition that is constrained by a unit clause: One of the two truth values then immediately yields an empty clause.

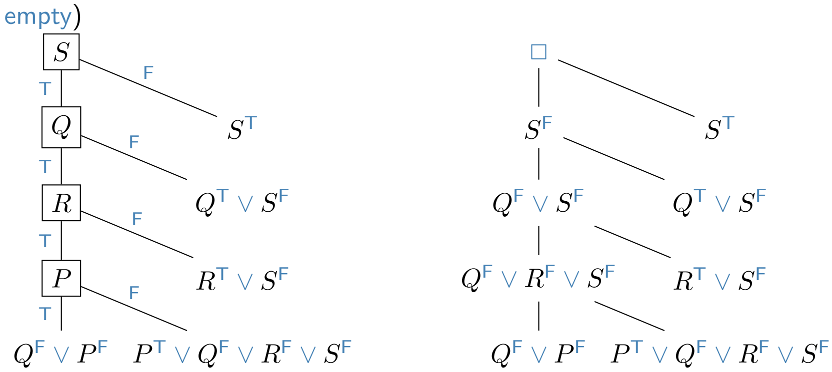
DPLL vs. Resolution: Example (Vanilla2)

▷ **Observation:** The proof of Theorem 13.3.6 is constructive, so we can use it as a method to read of a resolution proof from a DPLL trace.

▷ **Example 13.3.7.** We follow the steps in the proof of Theorem 13.3.6 for $\Delta := (Q^{\mathsf{F}} \vee P^{\mathsf{F}} \,;\, P^{\mathsf{T}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} \vee S^{\mathsf{F}} \,;\, Q^{\mathsf{T}} \vee S^{\mathsf{F}} \,;\, R^{\mathsf{T}} \vee S^{\mathsf{F}} \,;\, S^{\mathsf{T}})$

DPLL: (Without UP; leaves annotated with clauses that became empty)

Resolution proof from that DPLL tree:



▷ **Intuition:** From a (top-down) DPLL tree, we generate a (bottom-up) resolution proof.

For reference, we give the full proof here.

Theorem 13.3.8. *If DPLL returns "unsatisfiable" on Δ, then $S\colon \Box \vdash_{\mathcal{R}_0} \Box$ with a $\mathcal{R}_0$-derivation whose length is at most the number of decisions.*

Proof: Consider first DPLL with no unit propagation.

1. If the search tree is not empty, then there exists a leaf node N, i.e., a node associated to proposition X so that, for each value of X, the partial assignment directly results in an empty clause.
2. Denote the parent decisions of N by $L_1, \ldots, L_k$, where L_i is a literal for proposition X_i and the search node containing X_i is N_i.
3. Denote the empty clause for X by $C(N,X)$, and denote the empty clause for X^{F} by $C(N,X^{\mathsf{F}})$.
4. For each $x \in \{X^{\mathsf{T}}, X^{\mathsf{F}}\}$ we have the following properties:
   1. $x^{\mathsf{F}} \in C(N,x)$; and
   2. $C(N,x) \subseteq \{x^{\mathsf{F}}, \overline{L_1}, \ldots, \overline{L_k}\}$.

   Due to , we can resolve $C(N,X)$ with $C(N,X^{\mathsf{F}})$; denote the outcome clause by $C(N)$.
5. We obviously have that (1) $C(N) \subseteq \{\overline{L_1}, \ldots, \overline{L_k}\}$.
6. The proof now proceeds by removing N from the search tree and attaching $C(N)$ at the L_k branch of N_k, in the role of $C(N_k, L_k)$ as above. Then we select the next leaf node N' and iterate the argument; once the tree is empty, by (1) we have derived the empty clause. What we need to show is that, in each step of this iteration, we preserve the properties (a) and (b) for all leaf nodes. Since we did not change anything in other parts of the tree, the only node we need to show this for is $N' := N_k$.
7. Due to (1), we have (b) for N_k. But we do not necessarily have (a): $C(N) \subseteq \{\overline{L_1}, \ldots, \overline{L_k}\}$, but there are cases where $\overline{L_k} \notin C(N)$ (e.g., if X_k is not contained in any clause and thus

branching over it was completely unnecessary). If so, however, we can simply remove N_k and all its descendants from the tree as well. We attach $C(N)$ at the $L_{(k-1)}$ branch of $N_{(k-1)}|$, in the role of $C(N_{(k-1)}, L_{(k-1)})$. If $\overline{L_{(k-1)}} \in C(N)$ then we have (a) for $N' := N_{(k-1)}$ and can stop. If ${L_{(k-1)}}^{\mathsf{F}} \notin C(N)$, then we remove $N_{(k-1)}$ and so forth, until either we stop with (a), or have removed N_1 and thus must already have derived the empty clause (because $C(N) \subseteq \{\overline{L_1}, \ldots, \overline{L_k}\} \backslash \{\overline{L_1}, \ldots, \overline{L_k}\}$).

8. Unit propagation can be simulated via applications of the splitting rule, choosing a proposition that is constrained by a unit clause: One of the two truth values then immediately yields an empty clause.

DPLL vs. Resolution: Discussion

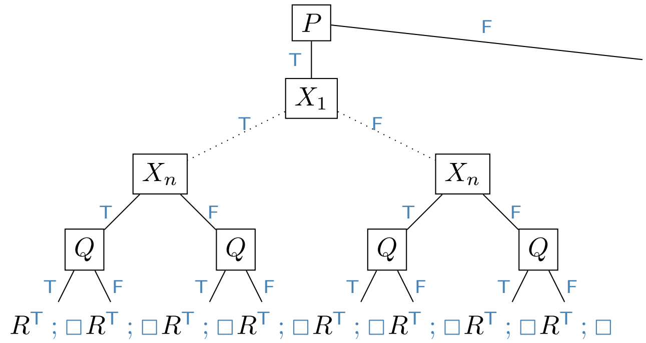
- **So What?:** The theorem we just proved helps to *understand* DPLL: DPLL is an effective practical method for conducting resolution proofs.
- **In fact:** DPLL $\widehat{=}$ tree resolution.
- **Definition 13.3.9.** In a tree resolution, each derived clause C is used only once (at its parent).
- **Problem:** The same C must be derived anew every time it is used!
- **This is a fundamental weakness:** There are inputs Δ whose shortest tree resolution proof is exponentially longer than their shortest (general) resolution proof.
- **Intuitively:** DPLL makes the same mistakes over and over again.
- **Idea:** DPLL should learn from its mistakes on one search branch, and apply the learned knowledge to other branches.
- **To the rescue:** clause learning (up next)

13.4 *Why* Did Unit Propagation Yield a Conflict?

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/27026`.

DPLL: Example (Redundance1)

- **Example 13.4.1.** We introduce some nasty redundance to make DPLL slow.
  $\Delta := (P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}})$
  DPLL on $\Delta ; \Theta$ with $\Theta := ({X_1}^{\mathsf{T}} \vee \ldots \vee {X_n}^{\mathsf{T}} ; {X_1}^{\mathsf{F}} \vee \ldots \vee {X_n}^{\mathsf{F}})$



How To *Not* Make the Same Mistakes Over Again?

▷ **It's not that difficult, really:**

(A) Figure out what went wrong.

(B) Learn to not do that again in the future.

▷ **And now for DPLL:**

(A) *Why* did unit propagation yield a Conflict?

▹ This Section. We will capture the "what went wrong" in terms of graphs over literals set during the search, and their dependencies.

▷ **What can we learn from that information?:**

▹ A new clause! Next section.

Implication Graphs for DPLL

▷ **Definition 13.4.2.** Let β be a branch in a DPLL derivation and P a variable on β then we call

▹ P^α a choice literal if its value is set to α by the splitting rule.

▹ P^α an implied literal, if the value of P is set to α by the UP rule.

▹ P^α a conflict literal, if it contributes to a derivation of the empty clause.

▷ **Definition 13.4.3 (Implication Graph).**

Let Δ be a clause set, β a DPLL search branch on Δ. The implication graph G_β^{impl} is the directed graph whose vertices are labeled with the choice and implied literals along β, as well as a separate conflict vertex $\Box_C$ for every clause C that became empty on β.

Whereever a clause $l_1, \ldots, l_k \vee l' \in \Delta$ became unit with implied literal l', G_β^{impl} includes the edges $(\overline{l_i}, l')$.

Where $C = l_1 \vee \ldots \vee l_k \in \Delta$ became empty, G_β^{impl} includes the edges $(\overline{l_i}, \square_C)$.

▷ **Question:** How do we know that $\overline{l_i}$ are vertices in G_β^{impl}?

▷ **Answer:** Because $l_1 \vee \ldots \vee l_k \vee l'$ became unit/empty.

▷ **Observation 13.4.4.** G_β^{impl} *is acyclic.*

▷ *Proof sketch:* UP can't derive l' whose value was already set beforehand.

▷ **Intuition:** The initial vertices are the choice literals and unit clauses of Δ.

Implication Graphs: Example (Vanilla1) in Detail

▷ **Example 13.4.5.** Let $\Delta := (P^{\mathsf{T}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} ; R^{\mathsf{T}} ; P^{\mathsf{T}} \vee Q^{\mathsf{F}})$.

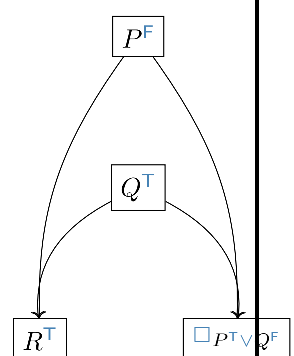
We look at the left branch of the derivation from Example 13.2.2:

1. UP Rule: $R \mapsto \mathsf{T}$
   Implied literal R^{T}.
   $P^{\mathsf{T}} \vee Q^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} ; P^{\mathsf{T}} \vee Q^{\mathsf{F}}$
2. Splitting Rule:

2a. $P \mapsto \mathsf{F}$
   Choice literal P^{F}.
   $Q^{\mathsf{T}} ; Q^{\mathsf{F}}$

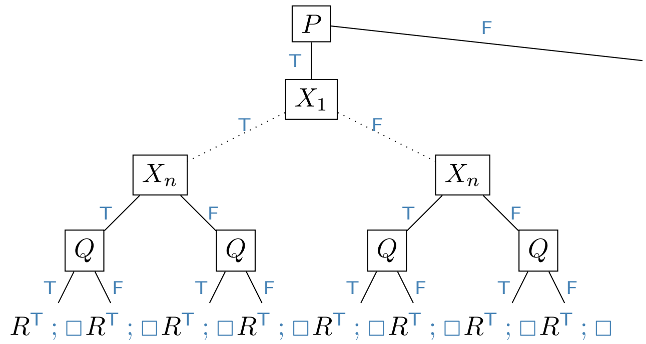
3a. UP Rule: $Q \mapsto \mathsf{T}$
   Implied literal Q^{T}
   edges $(R^{\mathsf{T}}, Q^{\mathsf{T}})$ and $(P^{\mathsf{F}}, Q^{\mathsf{T}})$.
   $\square$
   Conflict vertex $\square_{P^{\mathsf{T}} \vee Q^{\mathsf{F}}}$
   edges $(P^{\mathsf{F}}, \square_{P^{\mathsf{T}} \vee Q^{\mathsf{F}}})$ and $(Q^{\mathsf{T}}, \square_{P^{\mathsf{T}} \vee Q^{\mathsf{F}}})$.

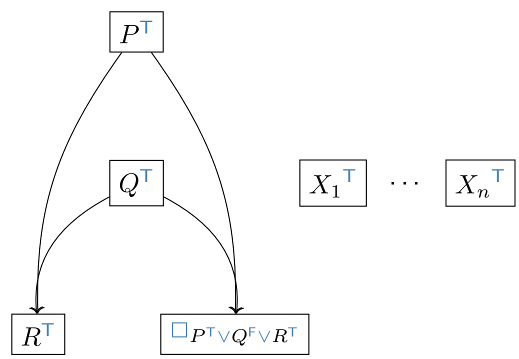
Implication graph:



Implication Graphs: Example (Redundance1)

▷ **Example 13.4.6.** Continuing from Example 13.4.5: $\Delta := (P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}})$
DPLL on $\Delta ; \Theta$ with $\Theta := (X_1{}^{\mathsf{T}} \vee \ldots \vee X_n{}^{\mathsf{T}} ; X_1{}^{\mathsf{F}} \vee \ldots \vee X_n{}^{\mathsf{F}})$
Choice literals: P^{T}, $(X_1{}^{\mathsf{T}}), \ldots, (X_n{}^{\mathsf{T}})$, Q^{T}. Implied literal: R^{T}.



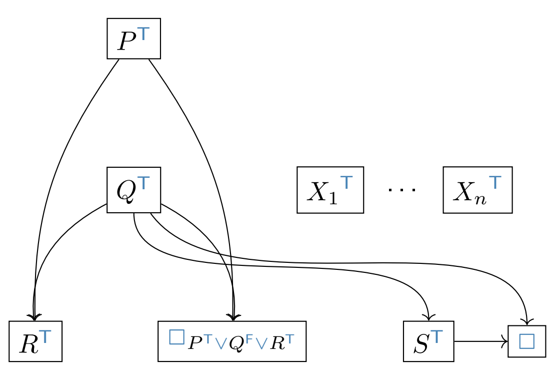


Implication Graphs: Example (Redundance2)

▷ **Example 13.4.7.** Continuing from Example 13.4.1:

$$\begin{aligned}\Delta &:= P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}} \\ \Theta &:= {X_1}^{\mathsf{T}} \vee \ldots \vee {X_n}^{\mathsf{T}} ; {X_1}^{\mathsf{F}} \vee \ldots \vee {X_n}^{\mathsf{F}}\end{aligned}$$

DPLL on $\Delta ; \Theta ; \Phi$ with $\Phi := (Q^{\mathsf{F}} \vee S^{\mathsf{T}} ; Q^{\mathsf{F}} \vee S^{\mathsf{F}})$
Choice literals: P^{T}, $({X_1}^{\mathsf{T}}), \ldots, ({X_n}^{\mathsf{T}})$, Q^{T}. Implied literals:

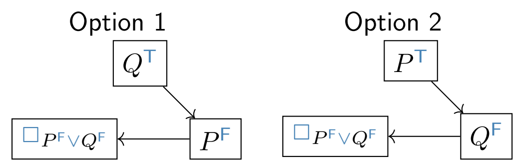


Implication Graphs: A Remark

▷ The implication graph is *not* uniquely determined by the Choice literals.

▷ It depends on "ordering decisions" during UP: Which unit clause is picked first.

▷ **Example 13.4.8.** $\Delta = P^{\mathsf{F}} \vee Q^{\mathsf{F}} \,;\, Q^{\mathsf{T}} \,;\, P^{\mathsf{T}}$



Conflict Graphs

▷ A conflict graph captures "what went wrong" in a failed node.

▷ **Definition 13.4.9 (Conflict Graph).** Let Δ be a clause set, and let $G^{\mathsf{impl}}_{\beta}$ be the implication graph for some search branch β of DPLL on Δ. A subgraph C of $G^{\mathsf{impl}}_{\beta}$ is a conflict graph if:

(i) C contains exactly one conflict vertex $\Box_C$.

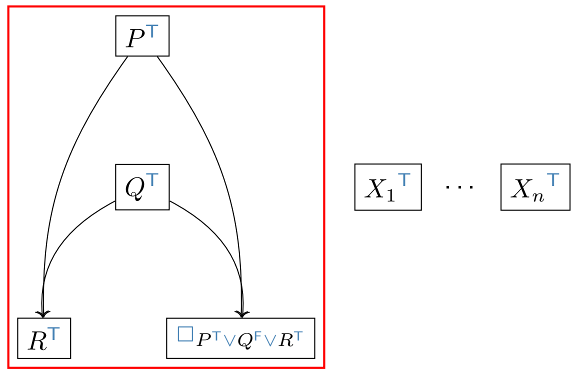
(ii) If l' is a vertex in C, then all parents of l', i.e. vertices $\overline{l_i}$ with a I edge $(\overline{l_i},l')$, are vertices in C as well.

(iii) All vertices in C have a path to $\Box_C$.

▷ Conflict graph $\widehat{=}$ Starting at a conflict vertex, backchain through the implication graph until reaching choice literals.

Conflict-Graphs: Example (Redundance1)

▷ **Example 13.4.10.** Continuing from Example 13.4.6: $\Delta{:=}(P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} \,;\, P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} \,;\, P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} \,;\, P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}})$
DPLL on $\Delta \,;\, \Theta$ with $\Theta{:=}({X_1}^{\mathsf{T}} \vee \ldots \vee {X_{100}}^{\mathsf{T}} \,;\, {X_1}^{\mathsf{F}} \vee \ldots \vee {X_{100}}^{\mathsf{F}})$
Choice literals: P^{T}, $({X_1}^{\mathsf{T}}), \ldots, ({X_{100}}^{\mathsf{T}})$, Q^{T}. Implied literals: R^{T}.

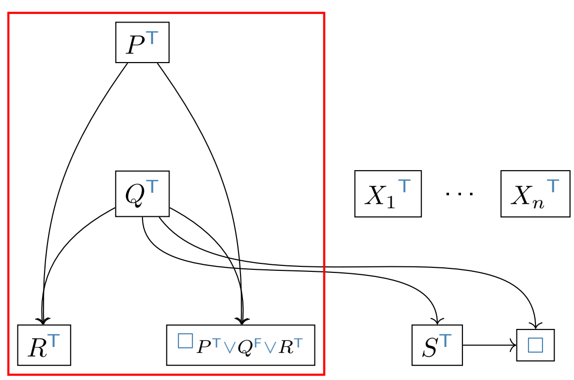


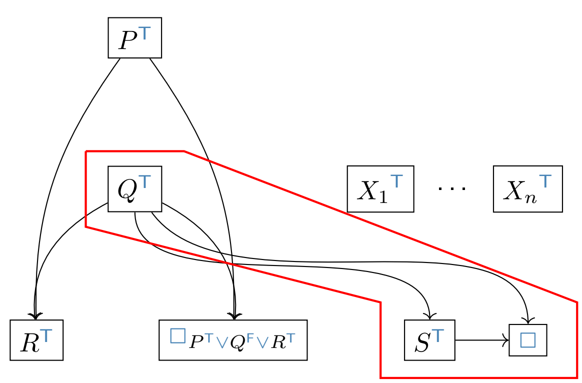
Conflict Graphs: Example (Redundance2)

▷ **Example 13.4.11.** Continuing from Example 13.4.7 and Example 13.4.10:

$$\Delta := P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}}$$
$$\Theta := {X_1}^{\mathsf{T}} \vee \ldots \vee {X_n}^{\mathsf{T}} ; {X_1}^{\mathsf{F}} \vee \ldots \vee {X_n}^{\mathsf{F}}$$

DPLL on $\Delta ; \Theta ; \Phi$ with $\Phi := (Q^{\mathsf{F}} \vee S^{\mathsf{T}} ; Q^{\mathsf{F}} \vee S^{\mathsf{F}})$
Choice literals: P^{T}, $({X_1}^{\mathsf{T}}), \ldots, ({X_n}^{\mathsf{T}})$, Q^{T}. Implied literals: R^{T}.





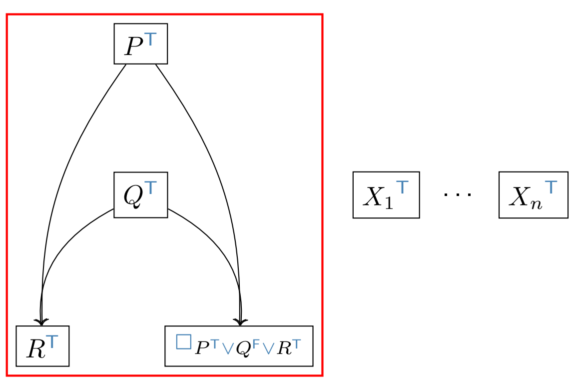
13.5 Clause Learning

Clause Learning

- **Observation:** Conflict graphs encode the entailment relation.
- **Definition 13.5.1.** Let Δ be a clause set, C be a conflict graph at some time point during a run of DPLL on Δ, and L be the choice literals in C, then we call $c{:=}\bigvee_{l\in L}\overline{l}$ the learned clause for C.
- **Theorem 13.5.2.** *Let Δ, C, and c as in Definition 13.5.1, then $\Delta \models c$.*
- **Idea:** We can add learned clauses to DPLL derivations at any time without losing soundness. (maybe this helps, if we have a good notion of learned clauses)
- **Definition 13.5.3.** Clause learning is the process of adding learned clauses to DPLL clause sets at specific points. (details coming up)

Clause Learning: Example (Redundance1)

- **Example 13.5.4.** Continuing from Example 13.4.10:
  $\Delta{:=}(P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}})$
  DPLL on $\Delta ; \Theta$ with $\Theta{:=}({X_1}^{\mathsf{T}} \vee \ldots \vee {X_n}^{\mathsf{T}} ; {X_1}^{\mathsf{F}} \vee \ldots \vee {X_n}^{\mathsf{F}})$
  Choice literals: P^{T}, $({X_1}^{\mathsf{T}}), \ldots, ({X_n}^{\mathsf{T}})$, Q^{T}. Implied literals: R^{T}.



Learned clause: $P^{\mathsf{F}} \vee Q^{\mathsf{F}}$

The Effect of Learned Clauses (in Redundance1)

- What happens after we learned a new clause C?

1. We add C into Δ. e.g. $C = P^{\mathsf{F}} \vee Q^{\mathsf{F}}$.
2. We retract the last choice l'. e.g. the choice $l' = Q$.

▷ **Observation:** Let C be a learned clause, i.e. $C = \bigvee_{l \in L} \bar{l}$, where L is the set of conflict literals in a conflict graph G.

Before we learn C, G must contain the most recent choice l': otherwise, the conflict would have occured earlier on.

So $C = {l_1}^{\mathsf{T}} \vee \ldots \vee {l_k}^{\mathsf{T}} \vee \overline{l'}$ where $l_1, \ldots, l_k$ are earlier choices.

▷ **Example 13.5.5.** $l_1 = P$, $C = P^{\mathsf{F}} \vee Q^{\mathsf{F}}$, $l' = Q$.

▷ **Observation:** Given the earlier choices $l_1, \ldots, l_k$, after we learned the new clause $C = \overline{l_1} \vee \ldots \vee \overline{l_k} \vee \overline{l'}$, the value of $\overline{l'}$ is now set by UP!

▷ So we can continue:

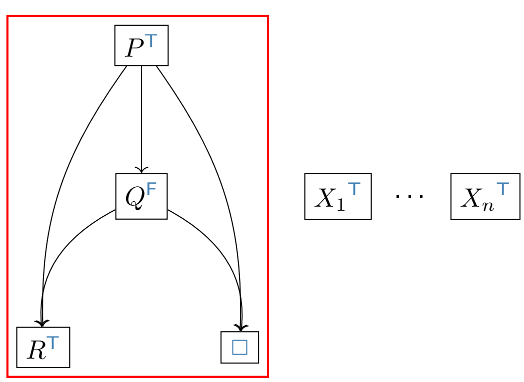
3. We set the opposite choice $\overline{l'}$ as an implied literal.
   e.g. Q^{F} as an implied literal.
4. We run UP and analyze conflicts.
   Learned clause: earlier choices only! e.g. $C = P^{\mathsf{F}}$, see next slide.

The Effect of Learned Clauses: Example (Redundance1)

▷ **Example 13.5.6.** Continuing from Example 13.5.4:

$$\begin{aligned} \Delta &:= P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}} \\ \Theta &:= {X_1}^{\mathsf{T}} \vee \ldots \vee {X_{100}}^{\mathsf{T}} ; {X_1}^{\mathsf{F}} \vee \ldots \vee {X_{100}}^{\mathsf{F}} \end{aligned}$$

DPLL on $\Delta ; \Theta ; \Phi$ with $\Phi := (P^{\mathsf{F}} \vee Q^{\mathsf{F}})$
Choice literals: P^{T}, $({X_1}^{\mathsf{T}}), \ldots, ({X_{100}}^{\mathsf{T}})$, Q^{T}. Implied literals: $Q^{\mathsf{F}}, R^{\mathsf{T}}$.



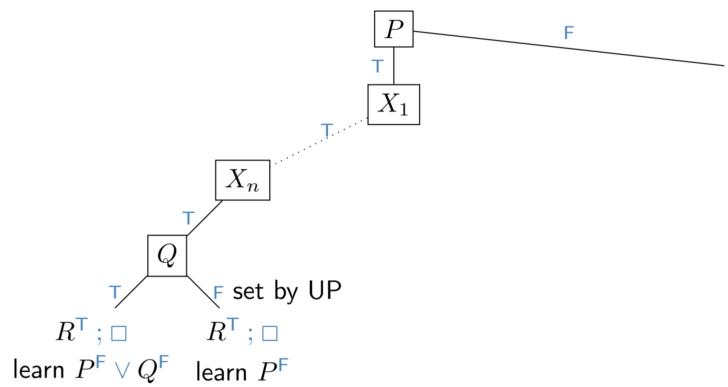
Learned clause: P^{F}

NOT the same Mistakes over Again: (Redundance1)

▷ **Example 13.5.7.** Continuing from Example 13.4.10:

$$\Delta := (P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{T}} ; P^{\mathsf{F}} \vee Q^{\mathsf{T}} \vee R^{\mathsf{F}})$$

DPLL on $\Delta ; \Theta$ with $\Theta := ({X_1}^{\mathsf{T}} \vee \ldots \vee {X_n}^{\mathsf{T}} ; {X_1}^{\mathsf{F}} \vee \ldots \vee {X_n}^{\mathsf{F}})$



▷ **Note:** Here, the problem could be avoided by splitting over different variables.

▷ **Problem:** This is not so in general! (see next slide)

Clause Learning vs. Resolution

▷ **Recall:** DPLL $\widehat{=}$ tree resolution (from slide 390)

1. **in particular**: each derived clause C (not in Δ) is derived anew every time it is used.
2. **Problem**: there are Δ whose shortest tree resolution proof is exponentially longer than their shortest (general) resolution proof.

▷ **Good News:** This is no longer the case with clause learning!

1. We add each learned clause C to Δ, can use it as often as we like.
2. Clause learning renders DPLL equivalent to full resolution [BKS04; PD09]. (Inhowfar exactly this is the case was an open question for ca. 10 years, so it's not as easy as I made it look here . . .)

▷ **In particular:** Selecting different variables/values to split on can *provably* not bring DPLL up to the power of DPLL+Clause Learning. (cf. slide 405, and previous slide)

"DPLL + Clause Learning"?

- **Disclaimer:** We have only seen *how to learn a clause from a conflict*.
- We will *not* cover how the overall DPLL algorithm changes, given this learning. Slides 403 – 405 are merely meant to give a *rough intuition* on "backjumping".
- **Definition 13.5.8 (Just for the record).** (not exam or exercises relevant)
  - One *could* run "DPLL + Clause Learning" by always backtracking to the maximal-level choice variable contained in the learned clause.
  - The actual algorithm is called Conflict Directed Clause Learning (CDCL), and differs from DPLL more radically:

let $L := 0$; $I := \emptyset$
repeat
  execute UP
  if a conflict was reached **then** /* learned clause $C = \overline{l_1} \vee \ldots \vee \overline{l_k} \vee \overline{l'}$ */
    if $L = 0$ **then return** UNSAT
    $L := \max_{i=1}^{k}$ level(l_i); erase I below L
    add C into Δ; add $\overline{l'}$ **to** I at level L
  else
    if I is a total interpretation **then return** I
    choose a new decision literal l; add l **to** I at level L
    $L := L + 1$

Remarks

- **Which clause(s) to learn?:**
  - While we only select choice literals, much more can be done.
  - For any cut through the conflict graph, with Choice literals on the "left hand" side of the cut and the conflict literals on the right-hand side, the literals on the left border of the cut yield a learnable clause.
  - Must take care to *not learn too many clauses* ...
- **Origins of clause learning:**
  - Clause learning originates from "explanation-based (no-good) learning" developed in the CSP community.
  - The distinguishing feature here is that the "no-good" is a clause:
    - The exact same type of constraint as the rest of Δ.

13.6 Phase Transitions: Where the *Really* Hard Problems Are

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25088`.

Where Are the Hard Problems?

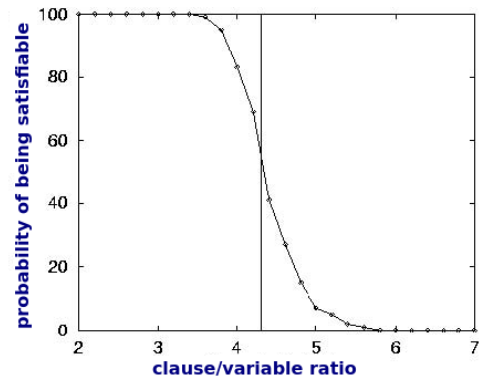
- SAT is **NP** hard. Worst case for DPLL is $\mathcal{O}(2^n)$, with n propositions.
- Imagine I gave you as homework to make a formula family $\{\varphi\}$ where DPLL running time necessarily is in the order of $\mathcal{O}(2^n)$.
  - I promise you're not gonna find this easy …(although it is of course possible: e.g., the "Pigeon Hole Problem").
- People noticed by the early 90s that, in practice, the DPLL worst case does not tend to happen.
- Modern SAT solvers successfully tackle practical instances where $n > 1.000.000$.

Where Are the Hard Problems?

- **So, what's the problem:** Science is about *understanding the world*.
  - Are "hard cases" just pathological outliers?
  - Can we say something about the *typical case*?
- **Difficulty 1:** What is the "typical case" in applications? E.g., what is the "average" hardware verification instance?
  - Consider precisely defined random distributions instead.
- **Difficulty 2:** Search trees get very complex, and are difficult to analyze mathematically, even in trivial examples. Never mind examples of practical relevance …
  - The most successful works are empirical. (Interesting theory is mainly concerned with *hand-crafted* formulas, like the Pigeon Hole Problem.)

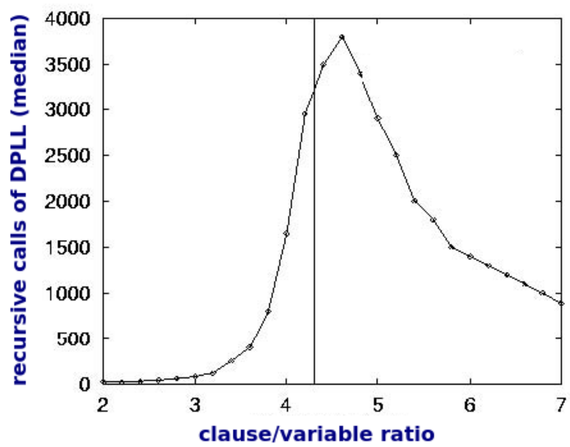
Phase Transitions in SAT [MSL92]

- **Fixed clause length model:** Fix clause length k; n variables.
  Generate m clauses, by uniformly choosing k variables P for each clause C, and for each variable P deciding uniformly whether to add P or P^{F} into C.
- **Order parameter:** Clause/variable ratio $\frac{m}{n}$.
- **Phase transition:** (Fixing $k = 3$, $n = 50$)



Does DPLL Care?

▷ **Oh yes, it does:** Extreme running time peak at the phase transition!



Why Does DPLL Care?

▷ **Intuition:**

Under-Constrained: Satisfiability likelihood close to 1. Many solutions, first DPLL search path usually successful. ("Deep but narrow")

Over-Constrained: Satisfiability likelihood close to 0. Most DPLL search paths short, conflict reached after few applications of splitting rule. ("Broad but shallow")

Critically Constrained: At the phase transition, many *almost-successful* DPLL search paths. ("Close, but no cigar")

The Phase Transition Conjecture

- **Definition 13.6.1.** We say that a class P of problems exhibits a phase transition, if there is an order parameter o, i.e. a structural parameter of P, so that almost all the hard problems of P cluster around a critical value c of o and c separates one region of the problem space from another, e.g. over-constrained and under-constrained regions.
- All **NP**-complete problems exhibit at least one phase transition.
- [CKT91] confirmed this for Graph Coloring and Hamiltonian Circuits. Later work confirmed it for SAT (see previous slides), and for numerous other **NP**-complete problems.

Why Should *We* Care?

- **Enlightenment:**
  - Phase transitions contribute to the fundamental understanding of the behavior of search, even if it's only in random distributions.
  - There are interesting theoretical connections to phase transition phenomena in physics. (See [GS05] for a short summary.)
- **Ok, but what can we use these results for?:**
  - **Benchmark design**: Choose instances from phase transition region.
    - Commonly used in competitions etc. (In SAT, random phase transition formulas are the most difficult for DPLL style searches.)
  - **Predicting solver performance**: Yes, but very limited because:
- All this works only for the particular considered *distributions of instances*! Not meaningful for any other instances.

13.7 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25090`.

Summary

- SAT solvers decide satisfiability of CNF formulas. This can be used for deduction, and is highly successful as a general problem solving technique (e.g., in verification).
- DPLL $\widehat{=}$ backtracking with inference performed by unit propagation (UP), which iteratively instantiates unit clauses and simplifies the formula.

▷ DPLL proofs of unsatisfiability correspond to a restricted form of resolution. The restriction forces DPLL to "makes the same mistakes over again".

▷ Implication graphs capture how UP derives conflicts. Their analysis enables us to do clause learning. DPLL with clause learning is called CDCL. It corresponds to full resolution, not "making the same mistakes over again".

▷ CDCL is state of the art in applications, routinely solving formulas with millions of propositions.

▷ In particular random formula distributions, typical problem hardness is characterized by phase transitions.

State of the Art in SAT

▷ **SAT competitions:**

- ▷ Since beginning of the 90s `http://www.satcompetition.org/`
- ▷ *random* vs. *industrial* vs. *handcrafted* benchmarks.
- ▷ Largest industrial instances: $> 1.000.000$ propositions.

▷ **State of the art is CDCL:**

- ▷ Vastly superior on handcrafted and industrial benchmarks.
- ▷ Key techniques: clause learning! Also: Efficient implementation (UP!), good branching heuristics, random restarts, portfolios.

▷ **What about local search?:**

- ▷ Better on random instances.
- ▷ No "dramatic" progress in last decade.
- ▷ Parameters are difficult to adjust.

But – What About Local Search for SAT?

▷ **There's a wealth of research on local search for SAT, e.g.:**

▷ **Definition 13.7.1.** The GSAT algorithm **OUTPUT**: a satisfying truth assignment of Δ, if found

```
function GSAT (Δ, MaxFlips MaxTries
  for i :=1 to MaxTries
    I := a randomly−generated truth assignment
    for j :=1 to MaxFlips
    if I satisfies Δ then return I
        X:= a proposition reversing whose truth assignment gives
        the largest increase in the number of satisfied clauses
      I := I with the truth assignment of X reversed
```

```
    end for
  end for
  return ''no satisfying assignment found''
```

▷ local search is not as successful in SAT applications, and the underlying ideas are very similar to those presented in section 6.6 (Not covered here)

Topics We Didn't Cover Here

▷ **Variable/value selection heuristics**: A whole zoo is out there.

▷ **Implementation techniques**: One of the most intensely researched subjects. Famous "watched literals" technique for UP had huge practical impact.

▷ **Local search**: In space of all truth value assignments. GSAT (slide 418) had huge impact at the time (1992), caused huge amount of follow-up work. Less intensely researched since clause learning hit the scene in the late 90s.

▷ **Portfolios**: How to combine several SAT solvers effectively?

▷ **Random restarts**: Tackling heavy-tailed runtime distributions.

▷ **Tractable SAT**: Polynomial-time sub-classes (most prominent: 2-SAT, Horn formulas).

▷ **MaxSAT**: Assign weight to each clause, maximize weight of satisfied clauses (= optimization version of SAT).

▷ **Resolution special cases**: There's a universe in between unit resolution and full resolution: trade off inference vs. search.

▷ **Proof complexity**: Can one resolution special case X simulate another one Y polynomially? Or is there an exponential separation (example families where X is exponentially less effective than Y)?

Suggested Reading:

- *Chapter 7: Logical Agents*, Section 7.6.1 [RN09].
  - Here, RN describe DPLL, i.e., basically what I cover under "The Davis-Putnam (Logemann-Loveland) Procedure".
  - That's the only thing they cover of this Chapter's material. (And they even mark it as "can be skimmed on first reading".)
  - This does not do the state of the art in SAT any justice.
- *Chapter 7: Logical Agents*, Sections 7.6.2, 7.6.3, and 7.7 [RN09].
  - Sections 7.6.2 and 7.6.3 say a few words on local search for SAT, which I recommend as additional background reading. Section 7.7 describes in quite some detail how to build an agent using propositional logic to take decisions; nice background reading as well.

Chapter 14

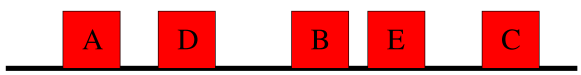
First-Order Predicate Logic

14.1 Motivation: A more Expressive Language

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25091`.

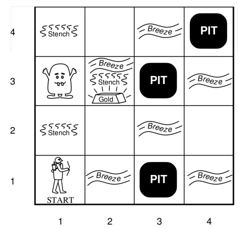
Let's Talk About Blocks, Baby ...

- ▷ **Question:** What do you see here?



- ▷ **You say:** "All blocks are red"; "All blocks are on the table"; "A is a block".
- ▷ **And now:** Say it in propositional logic!
- ▷ **Answer:** "isRedA","isRedB", ..., "onTableA", "onTableB", ..., "isBlockA", ...
- ▷ **Wait a sec!:** Why don't we just say, e.g., "AllBlocksAreRed" and "isBlockA"?
- ▷ **Problem:** Could we conclude that A is red? (No)
  These statements are atomic (just strings); their inner structure ("all blocks", "is a block") is not captured.
- ▷ **Idea:** Predicate Logic (PL^1) extends propositional logic with the ability to explicitly speak about objects and their properties.
- ▷ **How?:** Variables ranging over objects, predicates describing object properties, ...
- ▷ **Example 14.1.1.** "$\forall x.\mathrm{block}(x) \Rightarrow \mathrm{red}(x)$"; "$\mathrm{block}(\mathbf{A})$"

Let's Talk About the Wumpus Instead?

▷ 

Percepts: $[\mathit{Stench}, \mathit{Breeze}, \mathit{Glitter}, \mathit{Bump}, \mathit{Scream}]$

- ▹ Cell adjacent to Wumpus: Stench (else: None).
- ▹ Cell adjacent to Pit: Breeze (else: None).
- ▹ Cell that contains gold: $\mathit{Glitter}$ (else: None).
- ▹ You walk into a wall: Bump (else: None).
- ▹ Wumpus shot by arrow: Scream (else: None).

▷ Say, in propositional logic: "Cell adjacent to Wumpus: *Stench*."

- ▹ $W_{1,1} \Rightarrow S_{1,2} \wedge S_{2,1}$
- ▹ $W_{1,2} \Rightarrow S_{2,2} \wedge S_{1,1} \wedge S_{1,3}$
- ▹ $W_{1,3} \Rightarrow S_{2,3} \wedge S_{1,2} \wedge S_{1,4}$
- ▹ ...

▷ **Note:** Even when we *can* describe the problem suitably, for the desired reasoning, the propositional formulation typically is way too large to write (by hand).

▷ **PL1 solution:** "$\forall x.\mathrm{Wumpus}(x) \Rightarrow (\forall y.\mathrm{adj}(x,y) \Rightarrow \mathrm{stench}(y))$"

Blocks/Wumpus, Who Cares? Let's Talk About Numbers!

▷ Even worse!

▷ **Example 14.1.2 (Integers).** A limited vocabulary to talk about these

- ▹ The objects: $\{1, 2, 3, \ldots\}$.
- ▹ Predicate 1: "$\mathrm{even}(x)$" should be true iff x is even.
- ▹ Predicate 2: "$\mathrm{eq}(x,y)$" should be true iff $x = y$.
- ▹ Function: $\mathrm{succ}(x)$ maps x to $x + 1$.

▷ **Old problem:** Say, in propositional logic, that "$1 + 1 = 2$".

- ▹ Inner structure of vocabulary is ignored (cf. "AllBlocksAreRed").
- ▹ PL1 solution: "$\mathrm{eq}(\mathrm{succ}(1), 2)$".

▷ **New Problem:** Say, in propositional logic, "if x is even, so is $x + 2$".

- ▹ It is impossible to speak about infinite sets of objects!
- ▹ PL1 solution: "$\forall x.\mathrm{even}(x) \Rightarrow \mathrm{even}(\mathrm{succ}(\mathrm{succ}(x)))$".

Now We're Talking

▷ **Example 14.1.3.**

$$\forall n.\mathsf{gt}(n,2) \Rightarrow \neg(\exists a,b,c.\mathsf{eq}(\mathsf{plus}(\mathsf{pow}(a,n),\mathsf{pow}(b,n)),\mathsf{pow}(c,n)))$$

Read: *Forall $n > 2$, there are a, b, c, such that $a^n + b^n = c^n$* (Fermat's last theorem)

▷ **Theorem proving in PL1:** Arbitrary theorems, in principle.

- ▹ Fermat's last theorem is of course infeasible, but interesting theorems can and have been proved automatically.
- ▹ See http://en.wikipedia.org/wiki/Automated_theorem_proving.
- ▹ **Note**: Need to axiomatize "Plus", "PowerOf", "Equals". See http://en.wikipedia.org/wiki/Peano_axioms

What Are the Practical Relevance/Applications?

▷ **. . . even asking this question is a sacrilege:**

▷ (Quotes from Wikipedia)

- ▹ *"In Europe, logic was first developed by Aristotle. Aristotelian logic became widely accepted in science and mathematics."*
- ▹ *"The development of logic since Frege, Russell, and Wittgenstein had a profound influence on the practice of philosophy and the perceived nature of philosophical problems, and Philosophy of mathematics."*
- ▹ *"During the later medieval period, major efforts were made to show that Aristotle's ideas were compatible with Christian faith."*
- ▹ (In other words: the church issued for a long time that Aristotle's ideas were *in*compatible with Christian faith.)

What Are the Practical Relevance/Applications?

▷ **You're asking it anyhow:**

- ▹ **Logic programming**. Prolog et al.
- ▹ **Databases**. Deductive databases where elements of logic allow to conclude additional facts. Logic is tied deeply with database theory.
- ▹ Semantic technology. Mega-trend since > a decade. Use PL1 fragments to annotate data sets, facilitating their use and analysis.

▷ Prominent PL1 fragment: Web Ontology Language OWL.

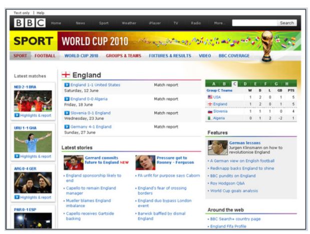
▷ Prominent data set: The WWW. (semantic web)

▷ **Assorted quotes on Semantic Web and OWL:**

- *The brain of humanity.*
- *The Semantic Web will never work.*
- *A TRULY meaningful way of interacting with the Web may finally be here: the Semantic Web. The idea was proposed 10 years ago. A triumvirate of internet heavyweights – Google, Twitter, and Facebook – are making it real.*

(A Few) Semantic Technology Applications

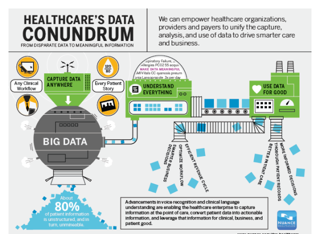
Web Queries



Jeopardy (IBM Watson)

Context-Aware Apps

Healthcare



Our Agenda for This Topic

- **This Chapter**: Basic definitions and concepts; normal forms.
  - Sets up the framework and basic operations.
  - **Syntax**: How to write PL1 formulas? (Obviously required)
  - **Semantics**: What is the meaning of PL1 formulas? (Obviously required.)
  - **Normal Forms**: What are the basic normal forms, and how to obtain them? (Needed for algorithms, which are defined on these normal forms.)
- **Next Chapter**: Compilation to propositional reasoning; unification; lifted resolution/tableau.
  - Algorithmic principles for reasoning about predicate logic.

14.2 First-Order Logic

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25093`.

First-order logic is the most widely used formal systems for modelling knowledge and inference processes. It strikes a very good bargain in the trade-off between expressivity and conceptual and computational complexity. To many people first-order logic is "the logic", i.e. the only logic worth considering, its applications range from the foundations of mathematics to natural language semantics.

First-Order Predicate Logic (PL^1)

- ▷ **Coverage:** We can talk about (*All humans are mortal*)
  - ▷ individual things and denote them by variables or constants
  - ▷ properties of individuals, (e.g. being human or mortal)
  - ▷ relations of individuals, (e.g. $sibling\_of$ relationship)
  - ▷ functions on individuals, (e.g. the $father\_of$ function)

  We can also state the existence of an individual with a certain property, or the universality of a property.
- ▷ But we cannot state assertions like
  - ▷ *There is a surjective function from the natural numbers into the reals.*
- ▷ First-Order Predicate Logic has many good properties (complete calculi, compactness, unitary, linear unification,...)
- ▷ But too weak for formalizing: (at least directly)
  - ▷ natural numbers, torsion groups, calculus, ...
  - ▷ generalized quantifiers (*most*, *few*,...)

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 428 2023-09-20

14.2.1 First-Order Logic: Syntax and Semantics

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/25094`.

The syntax and semantics of first-order logic is systematically organized in two distinct layers: one for truth values (like in propositional logic) and one for individuals (the new, distinctive feature of first-order logic).

The first step of defining a formal language is to specify the alphabet, here the first-order signatures and their components.

PL^1 Syntax (Signature and Variables)

- ▷ **Definition 14.2.1.** First-order logic (PL^1), is a formal system extensively used in mathematics, philosophy, linguistics, and computer science. It combines propositional logic with the ability to quantify over individuals.
- ▷ PL^1 talks about two kinds of objects: (so we have two kinds of symbols)
  - ▷ truth values by reusing PL^0

- ▹ individuals, e.g. numbers, foxes, Pokémon,. . .

▷ **Definition 14.2.2.** A first-order signature consists of (all disjoint; $k \in \mathbb{N}$)

- ▹ connectives: $\Sigma^o = \{T, F, \neg, \vee, \wedge, \Rightarrow, \Leftrightarrow, \ldots\}$ (functions on truth values)
- ▹ function constants: $\Sigma^f_k = \{f, g, h, \ldots\}$ (functions on individuals)
- ▹ predicate constants: $\Sigma^p_k = \{p, q, r, \ldots\}$ (relationships among individuals.)
- ▹ (Skolem constants: $\Sigma^{sk}_k = \{f^1_k, f^2_k, \ldots\}$) (witness constructors; countably ∞)
- ▹ We take Σ_ι to be all of these together: $\Sigma_\iota := \Sigma^f \cup \Sigma^p \cup \Sigma^{sk}$, where $\Sigma^* := \bigcup_{k \in \mathbb{N}} \Sigma^*_k$ and define $\Sigma := \Sigma_\iota \cup \Sigma^o$.

▷ **Definition 14.2.3.** We assume a set of individual variables: $\mathcal{V}_\iota := \{X, Y, Z, \ldots\}$. (countably ∞)

We make the deliberate, but non-standard design choice here to include Skolem constants into the signature from the start. These are used in inference systems to give names to objects and construct witnesses. Other than the fact that they are usually introduced by need, they work exactly like regular constants, which makes the inclusion rather painless. As we can never predict how many Skolem constants we are going to need, we give ourselves countably infinitely many for every arity. Our supply of individual variables is countably infinite for the same reason. The formulae of first-order logic is built up from the signature and variables as terms (to represent individuals) and propositions (to represent propositions). The latter include the propositional connectives, but also quantifiers.

PL¹ Syntax (Formulae)

▷ **Definition 14.2.4.** Terms: $\mathbf{A} \in wff_\iota(\Sigma_\iota, \mathcal{V}_\iota)$ (denote individuals)

- ▹ $\mathcal{V}_\iota \subseteq wff_\iota(\Sigma_\iota, \mathcal{V}_\iota)$,
- ▹ if $f \in \Sigma^f_k$ and $\mathbf{A}^i \in wff_\iota(\Sigma_\iota, \mathcal{V}_\iota)$ for $i \leq k$, then $f(\mathbf{A}^1, \ldots, \mathbf{A}^k) \in wff_\iota(\Sigma_\iota, \mathcal{V}_\iota)$.

▷ **Definition 14.2.5.** if Propositions: $\mathbf{A} \in wff_o(\Sigma_\iota, \mathcal{V}_\iota)$: (denote truth values)

- ▹ if $p \in \Sigma^p_k$ and $\mathbf{A}^i \in wff_\iota(\Sigma_\iota, \mathcal{V}_\iota)$ for $i \leq k$, then $p(\mathbf{A}^1, \ldots, \mathbf{A}^k) \in wff_o(\Sigma_\iota, \mathcal{V}_\iota)$,
- ▹ if $\mathbf{A}, \mathbf{B} \in wff_o(\Sigma_\iota, \mathcal{V}_\iota)$ and $X \in \mathcal{V}_\iota$, then $T, \mathbf{A} \wedge \mathbf{B}, \neg \mathbf{A}, \forall X.\mathbf{A} \in wff_o(\Sigma_\iota, \mathcal{V}_\iota)$. $\forall$ is a binding operator called the universal quantifier.

▷ **Definition 14.2.6.** We define the connectives $F, \vee, \Rightarrow, \Leftrightarrow$ via the abbreviations $\mathbf{A} \vee \mathbf{B} := \neg(\neg \mathbf{A} \wedge \neg \mathbf{B})$, $\mathbf{A} \Rightarrow \mathbf{B} := \neg \mathbf{A} \vee \mathbf{B}$, $\mathbf{A} \Leftrightarrow \mathbf{B} := (\mathbf{A} \Rightarrow \mathbf{B}) \wedge (\mathbf{B} \Rightarrow \mathbf{A})$, and $F := \neg T$. We will use them like the primary connectives $\wedge$ and $\neg$

▷ **Definition 14.2.7.** We use $\exists X.\mathbf{A}$ as an abbreviation for $\neg(\forall X.\neg \mathbf{A})$. $\exists$ is a binding operator called the existential quantifier.

▷ **Definition 14.2.8.** Call formulae without connectives or quantifiers atomic else complex.

Note: that we only need e.g. conjunction, negation, and universal quantification, all other logical constants can be defined from them (as we will see when we have fixed their interpretations).

Alternative Notations for Quantifiers

Here	Elsewhere
$\forall x.\mathbf{A}$	$\bigwedge x.\mathbf{A}$ $(x)\mathbf{A}$
$\exists x.\mathbf{A}$	$\bigvee x.\mathbf{A}$

The introduction of quantifiers to first-order logic brings a new phenomenon: variables that are under the scope of a quantifiers will behave very differently from the ones that are not. Therefore we build up a vocabulary that distinguishes the two.

Free and Bound Variables

- ▷ **Definition 14.2.9.** We call an occurrence of a variable X bound in a formula $\mathbf{A}$, iff it occurs in a sub-formula $\forall X.\mathbf{B}$ of $\mathbf{A}$. We call a variable occurrence free otherwise.

  For a formula $\mathbf{A}$, we will use $\mathsf{BVar}(\mathbf{A})$ (and $\mathsf{free}(\mathbf{A})$) for the set of bound (free) variables of $\mathbf{A}$, i.e. variables that have a free/bound occurrence in $\mathbf{A}$.

- ▷ **Definition 14.2.10.** We define the set $\mathsf{free}(\mathbf{A})$ of frees variable of a formula $\mathbf{A}$:

$$
\begin{aligned}
&\mathsf{free}(X){:=}\{X\}\\
&\mathsf{free}(f(\mathbf{A}_1,\ldots,\mathbf{A}_n)){:=}\textstyle\bigcup_{1\leq i\leq n}\mathsf{free}(\mathbf{A}_i)\\
&\mathsf{free}(p(\mathbf{A}_1,\ldots,\mathbf{A}_n)){:=}\textstyle\bigcup_{1\leq i\leq n}\mathsf{free}(\mathbf{A}_i)\\
&\mathsf{free}(\neg\mathbf{A}){:=}\mathsf{free}(\mathbf{A})\\
&\mathsf{free}(\mathbf{A}\wedge\mathbf{B}){:=}\mathsf{free}(\mathbf{A})\cup\mathsf{free}(\mathbf{B})\\
&\mathsf{free}(\forall X.\mathbf{A}){:=}\mathsf{free}(\mathbf{A})\backslash\{X\}
\end{aligned}
$$

- ▷ **Definition 14.2.11.** We call a formula $\mathbf{A}$ closed or ground, iff $\mathsf{free}(\mathbf{A}) = \emptyset$. We call a closed proposition a sentence, and denote the set of all ground terms with $\mathit{cwff}_\iota(\Sigma_\iota)$ and the set of sentences with $\mathit{cwff}_o(\Sigma_\iota)$.

- ▷ **Axiom 14.2.12.** *Bound variables can be renamed, i.e. any subterm $\forall X.\mathbf{B}$ of a formula $\mathbf{A}$ can be replaced by $\mathbf{A}'{:=}(\forall Y.\mathbf{B}')$, where $\mathbf{B}'$ arises from $\mathbf{B}$ by replacing all $X{\in}\mathsf{free}(\mathbf{B})$ with a new variable Y that does not occur in $\mathbf{A}$. We call $\mathbf{A}$ an alphabetical variant of $\mathbf{A}$.*

We will be mainly interested in (sets of) sentences – i.e. closed propositions – as the representations of meaningful statements about individuals. Indeed, we will see below that free variables do not gives us expressivity, since they behave like constants and could be replaced by them in all situations, except the recursive definition of quantified formulae. Indeed in all situations where variables occur freely, they have the character of meta-variables, i.e. syntactic placeholders that can be instantiated with terms when needed in an inference calculus.
The semantics of first-order logic is a Tarski-style set-theoretic semantics where the atomic syntactic entities are interpreted by mapping them into a well-understood structure, a first-order universe that is just an arbitrary set.

Semantics of PL^1 (Models)

▷ **Definition 14.2.13.** We inherit the universe $\mathcal{D}_0 = \{\mathsf{T}, \mathsf{F}\}$ of truth values from PL^0 and assume an arbitrary universe $\mathcal{D}_\iota \neq \emptyset$ of individuals (this choice is a parameter to the semantics)

▷ **Definition 14.2.14.** An interpretation $\mathcal{I}$ assigns values to constants, e.g.

- ▷ $\mathcal{I}(\neg)\colon \mathcal{D}_0 \to \mathcal{D}_0$ with $\mathsf{T} \mapsto \mathsf{F}$, $\mathsf{F} \mapsto \mathsf{T}$, and $\mathcal{I}(\wedge) = \ldots$ (as in PL^0)
- ▷ $\mathcal{I}\colon \Sigma_k^f \to {\mathcal{D}_\iota}^k \to \mathcal{D}_\iota$ (interpret function symbols as arbitrary functions)
- ▷ $\mathcal{I}\colon \Sigma_k^p \to \mathcal{P}({\mathcal{D}_\iota}^k)$ (interpret predicates as arbitrary relations)

▷ **Definition 14.2.15.** A variable assignment $\varphi\colon \mathcal{V}_\iota \to \mathcal{D}_\iota$ maps variables into the universe.

▷ **Definition 14.2.16.** A model $\mathcal{M} = \langle \mathcal{D}_\iota, \mathcal{I} \rangle$ of PL^1 consists of a universe $\mathcal{D}_\iota$ and an interpretation $\mathcal{I}$.

We do not have to make the universe of truth values part of the model, since it is always the same; we determine the model by choosing a universe and an interpretation function.
Given a first-order model, we can define the evaluation function as a homomorphism over the construction of formulae.

Semantics of PL^1 (Evaluation)

▷ **Definition 14.2.17.** Given a model $\langle \mathcal{D}, \mathcal{I} \rangle$, the value function $\mathcal{I}_\varphi$ is recursively defined: (two parts: terms & propositions)

- ▷ $\mathcal{I}_\varphi\colon \mathit{wff}_\iota(\Sigma_\iota, \mathcal{V}_\iota) \to \mathcal{D}_\iota$ assigns values to terms.
  - ▷ $\mathcal{I}_\varphi(X) := \varphi(X)$ and
  - ▷ $\mathcal{I}_\varphi(f(\mathbf{A}_1, \ldots, \mathbf{A}_k)) := \mathcal{I}(f)(\mathcal{I}_\varphi(\mathbf{A}_1), \ldots, \mathcal{I}_\varphi(\mathbf{A}_k))$
- ▷ $\mathcal{I}_\varphi\colon \mathit{wff}_o(\Sigma_\iota, \mathcal{V}_\iota) \to \mathcal{D}_0$ assigns values to formulae:
  - ▷ $\mathcal{I}_\varphi(T) = \mathcal{I}(T) = \mathsf{T}$,
  - ▷ $\mathcal{I}_\varphi(\neg\mathbf{A}) = \mathcal{I}(\neg)(\mathcal{I}_\varphi(\mathbf{A}))$
  - ▷ $\mathcal{I}_\varphi(\mathbf{A} \wedge \mathbf{B}) = \mathcal{I}(\wedge)(\mathcal{I}_\varphi(\mathbf{A}), \mathcal{I}_\varphi(\mathbf{B}))$ (just as in PL^0)
  - ▷ $\mathcal{I}_\varphi(p(\mathbf{A}_1, \ldots, \mathbf{A}_k)) := \mathsf{T}$, iff $\langle \mathcal{I}_\varphi(\mathbf{A}_1), \ldots, \mathcal{I}_\varphi(\mathbf{A}_k) \rangle \in \mathcal{I}(p)$
  - ▷ $\mathcal{I}_\varphi(\forall X.\mathbf{A}) := \mathsf{T}$, iff $\mathcal{I}_{\varphi,[\mathsf{a}/X]}(\mathbf{A}) = \mathsf{T}$ for all $\mathsf{a} \in \mathcal{D}_\iota$.

▷ **Definition 14.2.18 (Assignment Extension).** Let φ be a variable assignment into D and $a \in D$, then $\varphi,[a/X]$ is called the extension of φ with $[a/X]$ and is defined as $\{(Y,a) \in \varphi | Y \neq X\} \cup \{(X,a)\}$: $\varphi,[a/X]$ coincides with φ off X, and gives the result a there.

The only new (and interesting) case in this definition is the quantifier case, there we define the value of a quantified formula by the value of its scope – *but with an extension of the incoming variable assignment*. Note that by passing to the scope $\mathbf{A}$ of $\forall x.\mathbf{A}$, the occurrences of the variable x in $\mathbf{A}$ that were bound in $\forall x.\mathbf{A}$ become free and are amenable to evaluation by the variable assignment $\psi := \varphi,[\mathsf{a}/X]$. Note that as an extension of φ, the assignment ψ supplies exactly the right value for x in $\mathbf{A}$. This variability of the variable assignment in the definition of the value function justifies the somewhat complex setup of first-order evaluation, where we have the (static) interpretation function for the symbols from the signature and the (dynamic) variable assignment for the variables.

Note furthermore, that the value $\mathcal{I}_\varphi(\exists x.\mathbf{A})$ of $\exists x.\mathbf{A}$, which we have defined to be $\neg(\forall x.\neg\mathbf{A})$ is true, iff it is not the case that $\mathcal{I}_\varphi(\forall x.\neg\mathbf{A}) = \mathcal{I}_\psi(\neg\mathbf{A}) = \mathsf{F}$ for all $\mathsf{a}\in\mathcal{D}_\iota$ and $\psi:=\varphi,[\mathsf{a}/X]$. This is the case, iff $\mathcal{I}_\psi(\mathbf{A}) = \mathsf{T}$ for some $\mathsf{a}\in\mathcal{D}_\iota$. So our definition of the existential quantifier yields the appropriate semantics.

Semantics Computation: Example

▷ **Example 14.2.19.** We define an instance of first-order logic:

- ▷ Signature: Let $\Sigma_0^f:=\{j,m\}$, $\Sigma_1^f:=\{f\}$, and $\Sigma_2^p:=\{o\}$
- ▷ Universe: $\mathcal{D}_\iota:=\{J,M\}$
- ▷ Interpretation: $\mathcal{I}(j):=J$, $\mathcal{I}(m):=M$, $\mathcal{I}(f)(J):=M$, $\mathcal{I}(f)(M):=M$, and $\mathcal{I}(o):=\{(M,J)\}$.

Then $\forall X.o(f(X),X)$ is a sentence and with $\psi:=\varphi,[\mathsf{a}/X]$ for $\mathsf{a}\in\mathcal{D}_\iota$ we have

$$\begin{array}{lll} \mathcal{I}_\varphi(\forall X.o(f(X),X)) = \mathsf{T} & \text{iff} & \mathcal{I}_\psi(o(f(X),X)) = \mathsf{T} \text{ for all } \mathsf{a}\in\mathcal{D}_\iota \\ & \text{iff} & (\mathcal{I}_\psi(f(X)),\mathcal{I}_\psi(X))\in\mathcal{I}(o) \text{ for all } \mathsf{a}\in\{J,M\} \\ & \text{iff} & (\mathcal{I}(f)(\mathcal{I}_\psi(X)),\psi(X))\in\{(M,J)\} \text{ for all } \mathsf{a}\in\{J,M\} \\ & \text{iff} & (\mathcal{I}(f)(\psi(X)),a) = (M,J) \text{ for all } \mathsf{a}\in\{J,M\} \\ & \text{iff} & \mathcal{I}(f)(a) = M \text{ and } a = J \text{ for all } \mathsf{a}\in\{J,M\} \end{array}$$

But $\mathsf{a} \neq J$ for $\mathsf{a} = M$, so $\mathcal{I}_\varphi(\forall X.o(f(X),X)) = \mathsf{F}$ in the model $\langle\mathcal{D}_\iota,\mathcal{I}\rangle$.

14.2.2 First-Order Substitutions

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/25156`.

We will now turn our attention to substitutions, special formula-to-formula mappings that operationalize the intuition that (individual) variables stand for arbitrary terms.

Substitutions on Terms

▷ **Intuition:** If $\mathbf{B}$ is a term and X is a variable, then we denote the result of systematically replacing all occurrences of X in a term $\mathbf{A}$ by $\mathbf{B}$ with $[\mathbf{B}/X](\mathbf{A})$.

▷ **Problem:** What about $[Z/Y],[Y/X](X)$, is that Y or Z?

▷ **Folklore:** $[Z/Y],[Y/X](X) = Y$, but $[Z/Y]([Y/X](X)) = Z$ of course. (Parallel application)

▷ **Definition 14.2.20.**[for=sbstListfromto,sbstListdots,sbst]

Let $wfe(\Sigma,\mathcal{V})$ be an expression language, then we call $\sigma\colon\mathcal{V}\to wfe(\Sigma,\mathcal{V})$ a substitution, iff the support $\operatorname{supp}(\sigma):=\{X|(X,\mathbf{A})\in\sigma, X\neq\mathbf{A}\}$ of σ is finite. We denote the empty substitution with ϵ.

▷ **Definition 14.2.21 (Substitution Application).** We define substitution application by

- ▷ $\sigma(c) = c$ for $c \in \Sigma$
- ▷ $\sigma(X) = \mathbf{A}$, iff $\mathbf{A} \in \mathcal{V}$ and $(X, \mathbf{A}) \in \sigma$.
- ▷ $\sigma(f(\mathbf{A}_1, \ldots, \mathbf{A}_n)) = f(\sigma(\mathbf{A}_1), \ldots, \sigma(\mathbf{A}_n))$,
- ▷ $\sigma(\beta X.\mathbf{A}) = \beta X.\sigma_{-X}(\mathbf{A})$.

▷ **Example 14.2.22.** $[a/x], [f(b)/y], [a/z]$ instantiates $g(x, y, h(z))$ to $g(a, f(b), h(a))$.

▷ **Definition 14.2.23.** Let σ be a substitution then we call $\mathrm{intro}(\sigma) := \bigcup_{X \in \mathrm{supp}(\sigma)} \mathrm{free}(\sigma(X))$ the set of variables introduced by σ.

The extension of a substitution is an important operation, which you will run into from time to time. Given a substitution σ, a variable x, and an expression $\mathbf{A}$, $\sigma,[\mathbf{A}/x]$ extends σ with a new value for x. The intuition is that the values right of the comma overwrite the pairs in the substitution on the left, which already has a value for x, even though the representation of σ may not show it.

Substitution Extension

▷ **Definition 14.2.24 (Substitution Extension).**
Let σ be a substitution, then we denote the extension of σ with $[\mathbf{A}/X]$ by $\sigma,[\mathbf{A}/X]$ and define it as $\{(Y,\mathbf{B}) \in \sigma \mid Y \neq X\} \cup \{(X,\mathbf{A})\}$: $\sigma,[\mathbf{A}/X]$ coincides with σ off X, and gives the result $\mathbf{A}$ there.

▷ **Note:** If σ is a substitution, then $\sigma,[\mathbf{A}/X]$ is also a substitution.

▷ We also need the dual operation: removing a variable from the support:

▷ **Definition 14.2.25.** We can discharge a variable X from a substitution σ by setting $\sigma_{-X} := \sigma,[X/X]$.

Note that the use of the comma notation for substitutions defined in **??** is consistent with substitution extension. We can view a substitution $[a/x], [f(b)/y]$ as the extension of the empty substitution (the identity function on variables) by $[f(b)/y]$ and then by $[a/x]$. Note furthermore, that substitution extension is not commutative in general.
For first-order substitutions we need to extend the substitutions defined on terms to act on propositions. This is technically more involved, since we have to take care of bound variables.

Substitutions on Propositions

▷ **Problem:** We want to extend substitutions to propositions, in particular to quantified formulae: What is $\sigma(\forall X.\mathbf{A})$?

▷ **Idea:** σ should not instantiate bound variables. ($[\mathbf{A}/X](\forall X.\mathbf{B}) = \forall \mathbf{A}.\mathbf{B}'$ ill-formed)

▷ **Definition 14.2.26.** $\sigma(\forall X.\mathbf{A}) := (\forall X.\sigma_{-X}(\mathbf{A}))$.

▷ **Problem:** This can lead to variable capture: $[f(X)/Y](\forall X.p(X, Y))$ would evaluate to $\forall X.p(X, f(X))$, where the second occurrence of X is bound after instantiation, whereas it was free before. **Solution:** Rename away the bound variable X in $\forall X.p(X,Y)$ before applying the substitution.

▷ **Definition 14.2.27 (Capture-Avoiding Substitution Application).** Let σ be a substitution, $\mathbf{A}$ a formula, and $\mathbf{A}'$ an alphabetical variant of $\mathbf{A}$, such that $\mathrm{intro}(\sigma) \cap \mathrm{BVar}(\mathbf{A}) = \emptyset$. Then we define $\sigma(\mathbf{A}) := \sigma(\mathbf{A}')$.

We now introduce a central tool for reasoning about the semantics of substitutions: the "substitution value Lemma", which relates the process of instantiation to (semantic) evaluation. This result will be the motor of all soundness proofs on axioms and inference rules acting on variables via substitutions. In fact, any logic with variables and substitutions will have (to have) some form of a substitution value Lemma to get the meta-theory going, so it is usually the first target in any development of such a logic. We establish the substitution-value Lemma for first-order logic in two steps, first on terms, where it is very simple, and then on propositions.

Substitution Value Lemma for Terms

▷ **Lemma 14.2.28.** *Let $\mathbf{A}$ and $\mathbf{B}$ be terms, then $\mathcal{I}_\varphi([\mathbf{B}/X]\mathbf{A}) = \mathcal{I}_\psi(\mathbf{A})$, where $\psi = \varphi,[\mathcal{I}_\varphi(\mathbf{B})/X]$.*

▷ *Proof:* by induction on the depth of $\mathbf{A}$:

1. depth=0 *Then $\mathbf{A}$ is a variable (say Y), or constant, so we have three cases*
   1.1. $\mathbf{A} = Y = X$
      1.1.1. then $\mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A})) = \mathcal{I}_\varphi([\mathbf{B}/X](X)) = \mathcal{I}_\varphi(\mathbf{B}) = \psi(X) = \mathcal{I}_\psi(X) = \mathcal{I}_\psi(\mathbf{A})$.
   1.2. $\mathbf{A} = Y \neq X$
      1.2.1. then $\mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A})) = \mathcal{I}_\varphi([\mathbf{B}/X](Y)) = \mathcal{I}_\varphi(Y) = \varphi(Y) = \psi(Y) = \mathcal{I}_\psi(Y) = \mathcal{I}_\psi(\mathbf{A})$.
   1.3. $\mathbf{A}$ is a constant
      1.3.1. Analogous to the preceding case ($Y \neq X$).
   1.4. This completes the base case (depth = 0).
2. depth> 0
   2.1. then $\mathbf{A} = f(\mathbf{A}_1,\ldots,\mathbf{A}_n)$ and we have

$$\begin{aligned}\mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A})) &= \mathcal{I}(f)(\mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A}_1)),\ldots,\mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A}_n))) \\ &= \mathcal{I}(f)(\mathcal{I}_\psi(\mathbf{A}_1),\ldots,\mathcal{I}_\psi(\mathbf{A}_n)) \\ &= \mathcal{I}_\psi(\mathbf{A}).\end{aligned}$$

      by inductive hypothesis
   2.2. This completes the inductive case, and we have proven the assertion.

Substitution Value Lemma for Propositions

▷ **Lemma 14.2.29.** $\mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A})) = \mathcal{I}_\psi(\mathbf{A})$, *where* $\psi = \varphi,[\mathcal{I}_\varphi(\mathbf{B})/X]$.

▷ *Proof:* by induction on the number n of connectives and quantifiers in $\mathbf{A}$:

1. $n = 0$
   1.1. then $\mathbf{A}$ is an atomic proposition, and we can argue like in the inductive case of the substitution value lemma for terms.
2. $n>0$ and $\mathbf{A} = \neg\mathbf{B}$ or $\mathbf{A} = \mathbf{C} \circ \mathbf{D}$
   2.1. Here we argue like in the inductive case of the term lemma as well.
3. $n>0$ and $\mathbf{A} = \forall Y.\mathbf{C}$ where (WLOG) $X \neq Y$ *(otherwise rename)*
   3.1. then $\mathcal{I}_\psi(\mathbf{A}) = \mathcal{I}_\psi(\forall Y.\mathbf{C}) = \mathsf{T}$, iff $\mathcal{I}_{\psi,[a/Y]}(\mathbf{C}) = \mathsf{T}$ for all $a \in \mathcal{D}_\iota$.
   3.2. But $\mathcal{I}_{\psi,[a/Y]}(\mathbf{C}) = \mathcal{I}_{\varphi,[a/Y]}([\mathbf{B}/X](\mathbf{C})) = \mathsf{T}$, by inductive hypothesis.
   3.3. So $\mathcal{I}_\psi(\mathbf{A}) = \mathcal{I}_\varphi(\forall Y.[\mathbf{B}/X](\mathbf{C})) = \mathcal{I}_\varphi([\mathbf{B}/X](\forall Y.\mathbf{C})) = \mathcal{I}_\varphi([\mathbf{B}/X](\mathbf{A}))$

To understand the proof fully, you should think about where the *WLOG* – it stands for without loss of generality comes from.

14.3 First-Order Natural Deduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/25157`.

In this section, we will introduce the first-order natural deduction calculus. Recall from section 10.5 that natural deduction calculus have introduction and elimination for every logical constant (the connectives in PL^0). Recall furthermore that we had two styles/notations for the calculus, the classical ND calculus and the Sequent-style notation. These principles will be carried over to natural deduction in PL^1.

This allows us to introduce the calculi in two stages, first for the (propositional) connectives and then extend this to a calculus for first-order logic by adding rules for the quantifiers. In particular, we can define the first-order calculi simply by adding (introduction and elimination) rules for the (universal and existential) quantifiers to the calculus $\mathcal{ND}_0$ defined in section 10.5.

To obtain a first-order calculus, we have to extend $\mathcal{ND}_0$ with (introduction and elimination) rules for the quantifiers.

First-Order Natural Deduction ($\mathcal{ND}^1$; Gentzen [Gen34])

▷ Rules for connectives just as always

▷ **Definition 14.3.1 (New Quantifier Rules).** The first-order natural deduction calculus $\mathcal{ND}^1$ extends $\mathcal{ND}_0$ by the following four rules:

$$\frac{\mathbf{A}}{\forall X.\mathbf{A}}\forall I^* \qquad \frac{\forall X.\mathbf{A}}{[\mathbf{B}/X](\mathbf{A})}\forall E$$

$$\frac{[\mathbf{B}/X](\mathbf{A})}{\exists X.\mathbf{A}}\exists I \qquad \frac{\exists X.\mathbf{A} \quad \begin{array}{c}[[c/X](\mathbf{A})]^1\\ \vdots \\ \mathbf{C}\end{array} \quad c \in \Sigma_0^{sk} \text{ new}}{\mathbf{C}}\exists E^1$$

* means that $\mathbf{A}$ does not depend on any hypothesis in which X is free.

The intuition behind the rule $\forall I$ is that a formula $\mathbf{A}$ with a (free) variable X can be generalized to $\forall X.\mathbf{A}$, if X stands for an arbitrary object, i.e. there are no restricting assumptions about X. The $\forall E$ rule is just a substitution rule that allows to instantiate arbitrary terms $\mathbf{B}$ for X in $\mathbf{A}$.

The $\exists I$ rule says if we have a witness $\mathbf{B}$ for X in $\mathbf{A}$ (i.e. a concrete term $\mathbf{B}$ that makes $\mathbf{A}$ true), then we can existentially close $\mathbf{A}$. The $\exists E$ rule corresponds to the common mathematical practice, where we give objects we know exist a new name c and continue the proof by reasoning about this concrete object c. Anything we can prove from the assumption $[c/X](\mathbf{A})$ we can prove outright if $\exists X.\mathbf{A}$ is known.

A Complex $\mathcal{ND}^1$ Example

▷ **Example 14.3.2.** We prove $\neg(\forall X.P(X)) \vdash_{\mathcal{ND}^1} \exists X.\neg P(X)$.

$$\dfrac{\dfrac{\dfrac{\neg(\forall X.P(X)) \quad \dfrac{\dfrac{\dfrac{\dfrac{[\neg(\exists X.\neg P(X))]^1 \quad \dfrac{[\neg P(X)]^2}{\exists X.\neg P(X)}\exists I}{F}FI}{\neg\neg P(X)}\neg I^2}{P(X)}\neg E}{\forall X.P(X)}\forall I}{F}FI}{\neg\neg(\exists X.\neg P(X))}\neg I^1}{\exists X.\neg P(X)}\neg E$$

Now we reformulate the classical formulation of the calculus of natural deduction as a sequent calculus by lifting it to the "judgements level" as we die for propositional logic. We only need provide new quantifier rules.

First-Order Natural Deduction in Sequent Formulation

▷ Rules for connectives from $\mathcal{ND}^0_\vdash$

▷ **Definition 14.3.3 (New Quantifier Rules).** The inference rules of the first-order sequent calculus $\mathcal{ND}^1_\vdash$ consist of those from $\mathcal{ND}^0_\vdash$ plus the following quantifier rules:

$$\dfrac{\Gamma\vdash\mathbf{A} \quad X\notin\mathrm{free}(\Gamma)}{\Gamma\vdash\forall X.\mathbf{A}}\forall I \qquad\qquad \dfrac{\Gamma\vdash\forall X.\mathbf{A}}{\Gamma\vdash[\mathbf{B}/X](\mathbf{A})}\forall E$$

$$\dfrac{\Gamma\vdash[\mathbf{B}/X](\mathbf{A})}{\Gamma\vdash\exists X.\mathbf{A}}\exists I \qquad\qquad \dfrac{\Gamma\vdash\exists X.\mathbf{A} \quad \Gamma,[c/X](\mathbf{A})\vdash\mathbf{C} \quad c\in\Sigma_0^{sk}\ \text{new}}{\Gamma\vdash\mathbf{C}}\exists E$$

Natural Deduction with Equality

▷ **Definition 14.3.4 (First-Order Logic with Equality).** We extend PL^1 with a new logical symbol for equality $= \in\Sigma_2^p$ and fix its semantics to $\mathcal{I}(=):=\{(x,x)|x\in\mathcal{D}_\iota\}$. We call the extended logic first-order logic with equality ($\mathrm{PL}^1_=$)

- ▷ We now extend natural deduction as well.
- ▷ **Definition 14.3.5.** For the calculus of natural deduction with equality ($\mathcal{ND}^1_=$) we add the following two rules to $\mathcal{ND}^1$ to deal with equality:

$$\frac{}{\mathbf{A}=\mathbf{A}}{=}I \qquad \frac{\mathbf{A}=\mathbf{B} \quad \mathbf{C}\,[\mathbf{A}]_p}{[\mathbf{B}/p]\mathbf{C}}{=}E$$

where $\mathbf{C}\,[\mathbf{A}]_p$ if the formula $\mathbf{C}$ has a subterm $\mathbf{A}$ at position p and $[\mathbf{B}/p]\mathbf{C}$ is the result of replacing that subterm with $\mathbf{B}$.

- ▷ In many ways equivalence behaves like equality, we will use the following rules in $\mathcal{ND}^1$
- ▷ **Definition 14.3.6.** $\Leftrightarrow I$ is derivable and $\Leftrightarrow E$ is admissible in $\mathcal{ND}^1$:

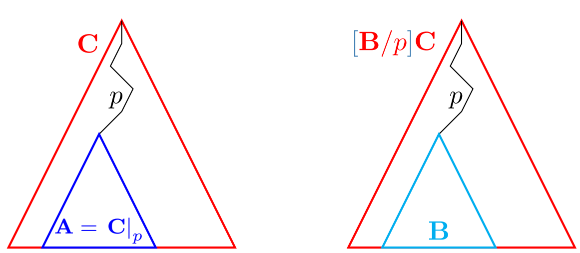
$$\frac{}{\mathbf{A}\Leftrightarrow\mathbf{A}}{\Leftrightarrow}I \qquad \frac{\mathbf{A}\Leftrightarrow\mathbf{B} \quad \mathbf{C}\,[\mathbf{A}]_p}{[\mathbf{B}/p]\mathbf{C}}{\Leftrightarrow}E$$

Again, we have two rules that follow the introduction/elimination pattern of natural deduction calculi. To make sure that we understand the constructions here, let us get back to the "replacement at position" operation used in the equality rules.

Positions in Formulae

- ▷ **Idea:** Formulae are (naturally) trees, so we can use tree positions to talk about subformulae
- ▷ **Definition 14.3.7.** A position p is a tuple of natural numbers that in each node of a expression (tree) specifies into which child to descend. For a expression $\mathbf{A}$ we denote the subexpression at p with $\mathbf{A}|_p$.

  We will sometimes write a expression $\mathbf{C}$ as $\mathbf{C}\,[\mathbf{A}]_p$ to indicate that $\mathbf{C}$ the subexpression $\mathbf{A}$ at position p.
- ▷ **Definition 14.3.8.** Let p be a position, then $[\mathbf{A}/p]\mathbf{C}$ is the expression obtained from $\mathbf{C}$ by replacing the subexpression at p by $\mathbf{A}$.
- ▷ **Example 14.3.9 (Schematically).**



The operation of replacing a subformula at position p is quite different from e.g. (first-order) substitutions:

- We are replacing subformulae with subformulae instead of instantiating variables with terms.
- substitutions replace all occurrences of a variable in a formula, whereas formula replacement only affects the (one) subformula at position p.

We conclude this section with an extended example: the proof of a classical mathematical result in the natural deduction calculus with equality. This shows us that we can derive strong properties about complex situations (here the real numbers; an uncountably infinite set of numbers).

$\mathcal{ND}^1_=$ Example: $\sqrt{2}$ is Irrational

▷ We can do real mathematics with $\mathcal{ND}^1_=$:

▷ **Theorem 14.3.10.** *$\sqrt{2}$ is irrational*

Proof: We prove the assertion by contradiction

1. Assume that $\sqrt{2}$ is rational.
2. Then there are numbers p and q such that $\sqrt{2} = p/q$.
3. So we know $2q^2 = p^2$.
4. But $2q^2$ has an odd number of prime factors while p^2 an even number.
5. This is a contradiction (since they are equal), so we have proven the assertion

If we want to formalize this into $\mathcal{ND}^1$, we have to write down all the assertions in the proof steps in PL^1 syntax and come up with justifications for them in terms of $\mathcal{ND}^1$ inference rules. The next two slides show such a proof, where we write $\prime n$ to denote that n is prime, use $\#(n)$ for the number of prime factors of a number n, and write $\mathrm{irr}(r)$ if r is irrational.

$\mathcal{ND}^1_=$ Example: $\sqrt{2}$ is Irrational (the Proof)

#	hyp	formula	NDjust
1		$\forall n, m.\neg(2n+1) = (2m)$	lemma
2		$\forall n, m.\#(n^m) = m\#(n)$	lemma
3		$\forall n, p.\mathsf{prime}(p) \Rightarrow \#(pn) = (\#(n)+1)$	lemma
4		$\forall x.\mathsf{irr}(x) \Leftrightarrow (\neg(\exists p, q.x = p/q))$	definition
5		$\mathsf{irr}(\sqrt{2}) \Leftrightarrow (\neg(\exists p, q.\sqrt{2} = p/q))$	$\forall E(4)$
6	6	$\neg\mathsf{irr}(\sqrt{2})$	Ax
7	6	$\neg\neg(\exists p, q.\sqrt{2} = p/q)$	$\Leftrightarrow E(6,5)$
8	6	$\exists p, q.\sqrt{2} = p/q$	$\neg E(7)$
9	6,9	$\sqrt{2} = p/q$	Ax
10	6,9	$2q^2 = p^2$	arith(9)
11	6,9	$\#(p^2) = 2\#(p)$	$\forall E^2(2)$
12	6,9	$\mathsf{prime}(2) \Rightarrow \#(2q^2) = (\#(q^2)+1)$	$\forall E^2(1)$

Lines 6 and 9 are local hypotheses for the proof (they only have an implicit counterpart in the inference rules as defined above). Finally we have abbreviated the arithmetic simplification of line 9 with the justification "arith" to avoid having to formalize elementary arithmetic.

$\mathcal{ND}^1_=$ Example: $\sqrt{2}$ is Irrational (the Proof continued)

13		$\mathrm{prime}(2)$	lemma
14	6,9	$\#(2q^2) = \#(q^2) + 1$	$\Rightarrow E(13, 12)$
15	6,9	$\#(q^2) = 2\#(q)$	$\forall E^2(2)$
16	6,9	$\#(2q^2) = 2\#(q) + 1$	$=E(14, 15)$
17		$\#(p^2) = \#(p^2)$	$=I$
18	6,9	$\#(2q^2) = \#(q^2)$	$=E(17, 10)$
19	6.9	$2\#(q) + 1 = \#(p^2)$	$=E(18, 16)$
20	6.9	$2\#(q) + 1 = 2\#(p)$	$=E(19, 11)$
21	6.9	$\neg(2\#(q) + 1) = (2\#(p))$	$\forall E^2(1)$
22	6,9	F	$FI(20, 21)$
23	6	F	$\exists E^6(22)$
24		$\neg\neg\mathrm{irr}(\sqrt{2})$	$\neg I^6(23)$
25		$\mathrm{irr}(\sqrt{2})$	$\neg E^2(23)$

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 448 2023-09-20

We observe that the $\mathcal{ND}^1$ proof is much more detailed, and needs quite a few Lemmata about $\#$ to go through. Furthermore, we have added a definition of irrationality (and treat definitional equality via the equality rules). Apart from these artefacts of formalization, the two representations of proofs correspond to each other very directly.

14.4 Conclusion

Summary (Predicate Logic)

- *Predicate logic* allows to explicitly speak about objects and their properties. It is thus a more natural and compact representation language than propositional logic; it also enables us to speak about infinite sets of objects.
- Logic has thousands of years of history. A major current application in AI is *Semantic Technology*.
- *First-order predicate logic (PL1)* allows *universal* and *existential quantification* over objects.
- A PL1 *interpretation* consists of a *universe* U and a function I mapping *constant symbols*/*predicate symbols*/*function symbols* to elements/relations/functions on U.

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 449 2023-09-20

Suggested Reading:

- *Chapter 8: First-Order Logic*, Sections 8.1 and 8.2 in [RN09]
  - A less formal account of what I cover in “Syntax” and “Semantics”. Contains different examples, and complementary explanations. Nice as additional background reading.
- Sections 8.3 and 8.4 provide additional material on using PL1, and on modeling in PL1, that I don’t cover in this lecture. Nice reading, not required for exam.
- *Chapter 9: Inference in First-Order Logic*, Section 9.5.1 in [RN09]

  - A very brief (2 pages) description of what I cover in “Normal Forms”. Much less formal; I couldn’t find where (if at all) RN cover transformation into prenex normal form. Can serve as additional reading, can’t replace the lecture.

- **Excursion:** A full analysis of any calculus needs a completeness proof. We will not cover this in AI-1, but provide one for the calculi introduced so far in**??**.

Chapter 15

Automated Theorem Proving in First-Order Logic

In this chapter, we take up the machine-oriented calculi for propositional logic from chapter 11 and extend them to the first-order case. While this has been relatively easy for the natural deduction calculus – we only had to introduce the notion of substitutions for the elimination rule for the universal quantifier we have to work much more here to make the calculi effective for implementation.

15.1 First-Order Inference with Tableaux

15.1.1 First-Order Tableau Calculi

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/25156`.

Test Calculi: Tableaux and Model Generation

- ▷ **Idea:** A tableau calculus is a test calculus that
  - ▷ analyzes a labeled formulae in a tree to determine satisfiability,
  - ▷ its branches correspond to valuations ($\rightsquigarrow$ models).
- ▷ **Example 15.1.1.** Tableau calculi try to construct models for labeled formulae:

Tableau refutation (Validity)	Model generation (Satisfiability)
$\models P \wedge Q \Rightarrow Q \wedge P$	$\models P \wedge (Q \vee \neg R) \wedge \neg Q$
$(P \wedge Q \Rightarrow Q \wedge P)^{\mathsf{F}}$ $(P \wedge Q)^{\mathsf{T}}$ $(Q \wedge P)^{\mathsf{F}}$ P^{T} Q^{T} $P^{\mathsf{F}} \mid Q^{\mathsf{F}}$ $\bot \mid \bot$	$(P \wedge (Q \vee \neg R) \wedge \neg Q)^{\mathsf{T}}$ $(P \wedge (Q \vee \neg R))^{\mathsf{T}}$ $\neg Q^{\mathsf{T}}$ Q^{F} P^{T} $(Q \vee \neg R)^{\mathsf{T}}$ $Q^{\mathsf{T}} \mid \neg R^{\mathsf{T}}$ $\bot \mid R^{\mathsf{F}}$
No Model	Herbrand Model $\{P^{\mathsf{T}}, Q^{\mathsf{F}}, R^{\mathsf{F}}\}$ $\varphi := \{P \mapsto \mathsf{T}, Q \mapsto \mathsf{F}, R \mapsto \mathsf{F}\}$

- ▷ **Idea:** Open branches in saturated tableaux yield models.

▷ **Algorithm:** Fully expand all possible tableaux, (no rule can be applied)

▷ Satisfiable, iff there are open branches (correspond to models)

Tableau calculi develop a formula in a tree-shaped arrangement that represents a case analysis on when a formula can be made true (or false). Therefore the formulae are decorated with exponents that hold the intended truth value.

On the left we have a refutation tableau that analyzes a negated formula (it is decorated with the intended truth value F). Both branches contain an elementary contradiction $\perp$.

On the right we have a model generation tableau, which analyzes a positive formula (it is decorated with the intended truth value T. This tableau uses the same rules as the refutation tableau, but makes a case analysis of when this formula can be satisfied. In this case we have a closed branch and an open one, which corresponds a model).

Now that we have seen the examples, we can write down the tableau rules formally.

Analytical Tableaux (Formal Treatment of $\mathcal{T}_0$)

▷ **Idea:** A test calculus where

- ▷ A labeled formula is analyzed in a tree to determine satisfiability,
- ▷ branches correspond to valuations (models)

▷ **Definition 15.1.2.** The propositional tableau calculus $\mathcal{T}_0$ has two inference rules per connective (one for each possible label)

$$\frac{(\mathbf{A} \wedge \mathbf{B})^{\mathsf{T}}}{\begin{array}{c}\mathbf{A}^{\mathsf{T}}\\ \mathbf{B}^{\mathsf{T}}\end{array}}\mathcal{T}_0\wedge \qquad \frac{(\mathbf{A} \wedge \mathbf{B})^{\mathsf{F}}}{\mathbf{A}^{\mathsf{F}} \;\Big|\; \mathbf{B}^{\mathsf{F}}}\mathcal{T}_0\vee \qquad \frac{\neg\mathbf{A}^{\mathsf{T}}}{\mathbf{A}^{\mathsf{F}}}\mathcal{T}_0\neg^{\mathsf{T}} \qquad \frac{\neg\mathbf{A}^{\mathsf{F}}}{\mathbf{A}^{\mathsf{T}}}\mathcal{T}_0\neg^{\mathsf{F}} \qquad \frac{\begin{array}{c}\mathbf{A}^{\alpha}\\ \mathbf{A}^{\beta}\end{array} \quad \alpha \neq \beta}{\perp}\mathcal{T}_0\perp$$

Use rules exhaustively as long as they contribute new material ($\leadsto$ termination)

▷ **Definition 15.1.3.** We call any tree ($\mid$ introduces branches) produced by the $\mathcal{T}_0$ inference rules from a set Φ of labeled formulae a tableau for Φ.

▷ **Definition 15.1.4.** Call a tableau saturated, iff no rule adds new material and a branch closed, iff it ends in $\perp$, else open. A tableau is closed, iff all of its branches are.

These inference rules act on tableaux have to be read as follows: if the formulae over the line appear in a tableau branch, then the branch can be extended by the formulae or branches below the line. There are two rules for each primary connective, and a branch closing rule that adds the special symbol $\perp$ (for unsatisfiability) to a branch.

We use the tableau rules with the convention that they are only applied, if they contribute new material to the branch. This ensures termination of the tableau procedure for propositional logic (every rule eliminates one primary connective).

Definition 15.1.5. We will call a closed tableau with the labeled formula $\mathbf{A}^{\alpha}$ at the root a tableau refutation for $\mathcal{A}^{\alpha}$.

The saturated tableau represents a full case analysis of what is necessary to give $\mathbf{A}$ the truth value α; since all branches are closed (contain contradictions) this is impossible.

Analytical Tableaux ($\mathcal{T}_0$ continued)

▷ **Definition 15.1.6 ($\mathcal{T}_0$-Theorem/Derivability).** $\mathbf{A}$ is a $\mathcal{T}_0$-theorem ($\vdash_{\mathcal{T}_0}\mathbf{A}$), iff there is a closed tableau with $\mathbf{A}^{\mathsf{F}}$ at the root.

$\Phi \subseteq \mathit{wff}_0(\mathcal{V}_0)$ derives $\mathbf{A}$ in $\mathcal{T}_0$ ($\Phi\vdash_{\mathcal{T}_0}\mathbf{A}$), iff there is a closed tableau starting with $\mathbf{A}^{\mathsf{F}}$ and Φ^{T}. The tableau with only a branch of $\mathbf{A}^{\mathsf{F}}$ and Φ^{T} is called initial for $\Phi\vdash_{\mathcal{T}_0}\mathbf{A}$.

Definition 15.1.7. We will call a tableau refutation for $\mathbf{A}^{\mathsf{F}}$ a tableau proof for $\mathbf{A}$, since it refutes the possibility of finding a model where $\mathbf{A}$ evaluates to F. Thus $\mathbf{A}$ must evaluate to T in all models, which is just our definition of validity.

Thus the tableau procedure can be used as a calculus for propositional logic. In contrast to the propositional Hilbert calculus it does not prove a theorem $\mathbf{A}$ by deriving it from a set of axioms, but it proves it by refuting its negation. Such calculi are called negative or test calculi. Generally negative calculi have computational advantages over positive ones, since they have a built-in sense of direction.

We have rules for all the necessary connectives (we restrict ourselves to $\wedge$ and $\neg$, since the others can be expressed in terms of these two via the propositional identities above. For instance, we can write $\mathbf{A} \vee \mathbf{B}$ as $\neg(\neg\mathbf{A} \wedge \neg\mathbf{B})$, and $\mathbf{A} \Rightarrow \mathbf{B}$ as $\neg\mathbf{A} \vee \mathbf{B}$,$\ldots$)

We will now extend the propositional tableau techniques to first-order logic. We only have to add two new rules for the universal quantifiers (in positive and negative polarity).

First-Order Standard Tableaux ($\mathcal{T}_1$)

▷ **Definition 15.1.8.** The standard tableau calculus ($\mathcal{T}_1$) extends $\mathcal{T}_0$ (propositional tableau calculus) with the following quantifier rules:

$$\frac{(\forall X.\mathbf{A})^{\mathsf{T}} \quad \mathbf{C}\in \mathit{cwff}_{\iota}(\Sigma_{\iota})}{([\mathbf{C}/X](\mathbf{A}))^{\mathsf{T}}}\,\mathcal{T}_1\forall \qquad \frac{(\forall X.\mathbf{A})^{\mathsf{F}} \quad c\in\Sigma_0^{sk} \text{ new}}{([c/X](\mathbf{A}))^{\mathsf{F}}}\,\mathcal{T}_1\exists$$

▷ **Problem:** The rule $\mathcal{T}_1\forall$ displays a case of "don't know indeterminism": to find a refutation we have to guess a formula $\mathbf{C}$ from the (usually infinite) set $\mathit{cwff}_{\iota}(\Sigma_{\iota})$.

For proof search, this means that we have to systematically try all, so $\mathcal{T}_1\forall$ is infinitely branching in general.

The rule $\mathcal{T}_1\forall$ operationalizes the intuition that a universally quantified formula is true, iff all of the instances of the scope are. To understand the $\mathcal{T}_1\exists$ rule, we have to keep in mind that $\exists X.\mathbf{A}$ abbreviates $\neg(\forall X.\neg\mathbf{A})$, so that we have to read $(\forall X.\mathbf{A})^{\mathsf{F}}$ existentially — i.e. as $(\exists X.\neg\mathbf{A})^{\mathsf{T}}$, stating that there is an object with property $\neg\mathbf{A}$. In this situation, we can simply give this object a name: c, which we take from our (infinite) set of witness constants Σ_0^{sk}, which we have given ourselves expressly for this purpose when we defined first-order syntax. In other words $([c/X](\neg\mathbf{A}))^{\mathsf{T}} = ([c/X](\mathbf{A}))^{\mathsf{F}}$ holds, and this is just the conclusion of the $\mathcal{T}_1\exists$ rule.

Note that the $\mathcal{T}_1\forall$ rule is computationally extremely inefficient: we have to guess an (i.e. in a search setting to systematically consider all) instance $\mathbf{C}\in \mathit{wff}_{\iota}(\Sigma_{\iota}, \mathcal{V}_{\iota})$ for X. This makes the rule infinitely branching.

In the next calculus we will try to remedy the computational inefficiency of the $\mathcal{T}_1\forall$ rule. We do this by delaying the choice in the universal rule.

Free variable Tableaux ($\mathcal{T}_1^f$)

- **Definition 15.1.9.** The free variable tableau calculus ($\mathcal{T}_1^f$) extends $\mathcal{T}_0$ (propositional tableau calculus) with the quantifier rules:

$$\frac{(\forall X.\mathbf{A})^{\mathsf{T}} \quad Y \text{ new}}{([Y/X](\mathbf{A}))^{\mathsf{T}}}\mathcal{T}_1^f\forall \qquad \frac{(\forall X.\mathbf{A})^{\mathsf{F}} \quad \text{free}(\forall X.\mathbf{A}) = \{X^1, \ldots, X^k\} \quad f\in\Sigma_k^{sk} \text{ new}}{([f(X^1,\ldots,X^k)/X](\mathbf{A}))^{\mathsf{F}}}\mathcal{T}_1^f\exists$$

  and generalizes its cut rule $\mathcal{T}_0\bot$ to:

$$\frac{\begin{array}{l}\mathbf{A}^\alpha\\ \mathbf{B}^\beta\end{array} \quad \alpha \neq \beta \quad \sigma(\mathbf{A}) = \sigma(\mathbf{B})}{\bot : \sigma}\mathcal{T}_1^f\bot$$

  $\mathcal{T}_1^f\bot$ instantiates the whole tableau by σ.

- **Advantage:** No guessing necessary in $\mathcal{T}_1^f\forall$-rule!
- **New Problem:** find suitable substitution (most general unifier) (later)

Metavariables: Instead of guessing a concrete instance for the universally quantified variable as in the $\mathcal{T}_1\forall$ rule, $\mathcal{T}_1^f\forall$ instantiates it with a new meta-variable Y, which will be instantiated by need in the course of the derivation.
Skolem terms as witnesses: The introduction of meta-variables makes is necessary to extend the treatment of witnesses in the existential rule. Intuitively, we cannot simply invent a new name, since the meaning of the body $\mathbf{A}$ may contain meta-variables introduced by the $\mathcal{T}_1^f\forall$ rule. As we do not know their values yet, the witness for the existential statement in the antecedent of the $\mathcal{T}_1^f\exists$ rule needs to depend on that. So witness it using a witness term, concretely by applying a Skolem function to the meta-variables in $\mathbf{A}$.
Instantiating Metavariables: Finally, the $\mathcal{T}_1^f\bot$ rule completes the treatment of meta-variables, it allows to instantiate the whole tableau in a way that the current branch closes. This leaves us with the problem of finding substitutions that make two terms equal.

Free variable Tableaux ($\mathcal{T}_1^f$): Derivable Rules

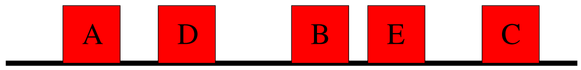
- **Definition 15.1.10.** Derivable quantifier rules in $\mathcal{T}_1^f$:

$$\frac{(\exists X.\mathbf{A})^{\mathsf{T}} \quad \text{free}(\forall X.\mathbf{A}) = \{X^1, \ldots, X^k\} \quad f\in\Sigma_k^{sk} \text{ new}}{([f(X^1,\ldots,X^k)/X](\mathbf{A}))^{\mathsf{T}}}$$

$$\frac{(\exists X.\mathbf{A})^{\mathsf{F}} \quad Y \text{ new}}{([Y/X](\mathbf{A}))^{\mathsf{F}}}$$

Tableau Reasons about Blocks

▷ **Example 15.1.11 (Reasoning about Blocks).** Returing to slide 420



Can we prove $\mathsf{red}(\mathbf{A})$ from $\forall x.\mathsf{block}(x) \Rightarrow \mathsf{red}(x)$ and $\mathsf{block}(\mathbf{A})$?

$$\begin{array}{c|c}
\multicolumn{2}{c}{(\forall X.\mathsf{block}(X) \Rightarrow \mathsf{red}(X))^{\mathsf{T}}} \\
\multicolumn{2}{c}{\mathsf{block}(\mathbf{A})^{\mathsf{T}}} \\
\multicolumn{2}{c}{\mathsf{red}(\mathbf{A})^{\mathsf{F}}} \\
\multicolumn{2}{c}{(\mathsf{block}(Y) \Rightarrow \mathsf{red}(Y))^{\mathsf{T}}} \\
\mathsf{block}(Y)^{\mathsf{F}} & \mathsf{red}(\mathbf{A})^{\mathsf{T}} \\
\bot : [\mathbf{A}/Y] & \bot
\end{array}$$

15.1.2 First-Order Unification

Video Nuggets covering this subsection can be found at `https://fau.tv/clip/id/26810` and `https://fau.tv/clip/id/26811`.

We will now look into the problem of finding a substitution σ that make two terms equal (we say it unifies them) in more detail. The presentation of the unification algorithm we give here "transformation-based" this has been a very influential way to treat certain algorithms in theoretical computer science.

A transformation-based view of algorithms: The "transformation-based" view of algorithms divides two concerns in presenting and reasoning about algorithms according to Kowalski's slogan [Kow97]

algorithm = logic + control

The computational paradigm highlighted by this quote is that (many) algorithms can be thought of as manipulating representations of the problem at hand and transforming them into a form that makes it simple to read off solutions. Given this, we can simplify thinking and reasoning about such algorithms by separating out their "logical" part, which deals with is concerned with how the problem representations can be manipulated in principle from the "control" part, which is concerned with questions about when to apply which transformations.

It turns out that many questions about the algorithms can already be answered on the "logic" level, and that the "logical" analysis of the algorithm can already give strong hints as to how to optimize control.

In fact we will only concern ourselves with the "logical" analysis of unification here.

The first step towards a theory of unification is to take a closer look at the problem itself. A first set of examples show that we have multiple solutions to the problem of finding substitutions that make two terms equal. But we also see that these are related in a systematic way.

Unification (Definitions)

▷ **Definition 15.1.12.** For given terms $\mathbf{A}$ and $\mathbf{B}$, unification is the problem of finding a substitution σ, such that $\sigma(\mathbf{A}) = \sigma(\mathbf{B})$.

▷ **Notation:** We write term pairs as $\mathbf{A} =^? \mathbf{B}$ e.g. $f(X) =^? f(g(Y))$.

- **Definition 15.1.13.** Solutions (e.g. $[g(a)/X], [a/Y]$, $[g(g(a))/X], [g(a)/Y]$, or $[g(Z)/X], [Z/Y]$) are called unifiers, $\mathbf{U}(\mathbf{A}{=}^{?}\mathbf{B}){:=}\{\sigma|\sigma(\mathbf{A}) = \sigma(\mathbf{B})\}$.
- **Idea:** Find representatives in $\mathbf{U}(\mathbf{A}{=}^{?}\mathbf{B})$, that generate the set of solutions.
- **Definition 15.1.14.** Let σ and θ be substitutions and $W \subseteq \mathcal{V}_\iota$, we say that a substitution σ is more general than θ (on W; write $\sigma{\leq}\theta[W]$), iff there is a substitution ρ, such that $\theta{=}(\rho \circ \sigma)[W]$, where $\sigma{=}\rho[W]$, iff $\sigma(X) = \rho(X)$ for all $X{\in}W$.
- **Definition 15.1.15.** σ is called a most general unifier (mgu) of $\mathbf{A}$ and $\mathbf{B}$, iff it is minimal in $\mathbf{U}(\mathbf{A}{=}^{?}\mathbf{B})$ wrt. $\leq[(\mathrm{free}(\mathbf{A}) \cup \mathrm{free}(\mathbf{B}))]$.

The idea behind a most general unifier is that all other unifiers can be obtained from it by (further) instantiation. In an automated theorem proving setting, this means that using most general unifiers is the least committed choice — any other choice of unifiers (that would be necessary for completeness) can later be obtained by other substitutions.
Note that there is a subtlety in the definition of the ordering on substitutions: we only compare on a subset of the variables. The reason for this is that we have defined substitutions to be total on (the infinite set of) variables for flexibility, but in the applications (see the definition of most general unifiers), we are only interested in a subset of variables: the ones that occur in the initial problem formulation. Intuitively, we do not care what the unifiers do off that set. If we did not have the restriction to the set W of variables, the ordering relation on substitutions would become much too fine-grained to be useful (i.e. to guarantee unique most general unifiers in our case).

Now that we have defined the problem, we can turn to the unification algorithm itself. We will define it in a way that is very similar to logic programming: we first define a calculus that generates "solved forms" (formulae from which we can read off the solution) and reason about control later. In this case we will reason that control does not matter.

Unification Problems ($\widehat{=}$ Equational Systems)

- **Idea:** Unification is equation solving.
- **Definition 15.1.16.**
  We call a formula $\mathbf{A}^1{=}^{?}\mathbf{B}^1 \wedge \ldots \wedge \mathbf{A}^n{=}^{?}\mathbf{B}^n$ an unification problem iff $\mathbf{A}^i, \mathbf{B}^i{\in}\mathit{wff}_\iota(\Sigma_\iota, \mathcal{V}_\iota)$.
- **Note:** We consider unification problems as sets of equations ($\wedge$ is ACI), and equations as two-element multisets (${=}^{?}$ is C).
- **Definition 15.1.17.** A substitution is called a unifier for a unification problem $\mathcal{E}$ (and thus $\mathcal{D}$ unifiable), iff it is a (simultaneous) unifier for all pairs in $\mathcal{E}$.

In principle, unification problems are sets of equations, which we write as conjunctions, since all of them have to be solved for finding a unifier. Note that it is not a problem for the "logical view" that the representation as conjunctions induces an order, since we know that conjunction is associative, commutative and idempotent, i.e. that conjuncts do not have an intrinsic order or multiplicity, if we consider two equational problems as equal, if they are equivalent as propositional formulae. In the same way, we will abstract from the order in equations, since we know that the equality relation is symmetric. Of course we would have to deal with this somehow in the implementation (typically, we would implement equational problems as lists of pairs), but that belongs into the "control" aspect of the algorithm, which we are abstracting from at the moment.

Solved forms and Most General Unifiers

▷ **Definition 15.1.18.** We call a pair $\mathbf{A}=^?\mathbf{B}$ solved in a unification problem $\mathcal{E}$, iff $\mathbf{A} = X$, $\mathcal{E} = X=^?\mathbf{A}\wedge\mathcal{E}$, and $X\notin(\mathrm{free}(\mathbf{A})\cup\mathrm{free}(\mathcal{E}))$. We call an unification problem $\mathcal{E}$ a solved form, iff all its pairs are solved.

▷ **Lemma 15.1.19.** *Solved forms are of the form $X^1=^?\mathbf{B}^1 \wedge \ldots \wedge X^n=^?\mathbf{B}^n$ where the X^i are distinct and $X^i\notin\mathrm{free}(\mathbf{B}^j)$.*

▷ **Definition 15.1.20.** Any substitution $\sigma = [\mathbf{B}^1/X^1],\ldots,[\mathbf{B}^n/X^n]$ induces a solved unification problem $\mathcal{E}_\sigma{:=}(X^1=^?\mathbf{B}^1 \wedge \ldots \wedge X^n=^?\mathbf{B}^n)$.

▷ **Lemma 15.1.21.** *If $\mathcal{E} = X^1=^?\mathbf{B}^1 \wedge \ldots \wedge X^n=^?\mathbf{B}^n$ is a solved form, then $\mathcal{E}$ has the unique most general unifier $\sigma_\mathcal{E}{:=}[\mathbf{B}^1/X^1],\ldots,[\mathbf{B}^n/X^n]$.*

▷ *Proof:* Let $\theta\in\mathbf{U}(\mathcal{E})$

1. then $\theta(X^i) = \theta(\mathbf{B}^i) = \theta\circ\sigma_\mathcal{E}(X^i)$
2. and thus $\theta=(\theta\circ\sigma_\mathcal{E})[\mathrm{supp}(\sigma)]$.

▷ **Note:** We can rename the introduced variables in most general unifiers!

It is essential to our "logical" analysis of the unification algorithm that we arrive at unification problems whose unifiers we can read off easily. Solved forms serve that need perfectly as Lemma 15.1.21 shows.

Given the idea that unification problems can be expressed as formulae, we can express the algorithm in three simple rules that transform unification problems into solved forms (or unsolvable ones).

Unification Algorithm

▷ **Definition 15.1.22.** The inference system $\mathcal{U}$ consists of the following rules:

$$\frac{\mathcal{E}\wedge f(\mathbf{A}^1,\ldots,\mathbf{A}^n)=^?f(\mathbf{B}^1,\ldots,\mathbf{B}^n)}{\mathcal{E}\wedge\mathbf{A}^1=^?\mathbf{B}^1\wedge\ldots\wedge\mathbf{A}^n=^?\mathbf{B}^n}\mathcal{U}\mathrm{dec}\qquad\frac{\mathcal{E}\wedge\mathbf{A}=^?\mathbf{A}}{\mathcal{E}}\mathcal{U}\mathrm{triv}$$

$$\frac{\mathcal{E}\wedge X=^?\mathbf{A}\quad X\notin\mathrm{free}(\mathbf{A})\quad X\in\mathrm{free}(\mathcal{E})}{[\mathbf{A}/X](\mathcal{E})\wedge X=^?\mathbf{A}}\mathcal{U}\mathrm{elim}$$

▷ **Lemma 15.1.23.** *$\mathcal{U}$ is correct: $\mathcal{E}\vdash_\mathcal{U}\mathcal{F}$ implies $\mathbf{U}(\mathcal{F})\subseteq\mathbf{U}(\mathcal{E})$.*

▷ **Lemma 15.1.24.** *$\mathcal{U}$ is complete: $\mathcal{E}\vdash_\mathcal{U}\mathcal{F}$ implies $\mathbf{U}(\mathcal{E})\subseteq\mathbf{U}(\mathcal{F})$.*

▷ **Lemma 15.1.25.** *$\mathcal{U}$ is confluent: the order of derivations does not matter.*

▷ **Corollary 15.1.26.** *First-order unification is unitary: i.e. most general unifiers are unique up to renaming of introduced variables.*

▷ *Proof sketch:* $\mathcal{U}$ is trivially branching.

The decomposition rule $\mathcal{U}$dec is completely straightforward, but note that it transforms one unification pair into multiple argument pairs; this is the reason, why we have to directly use unification problems with multiple pairs in $\mathcal{U}$.

Note furthermore, that we could have restricted the $\mathcal{U}$triv rule to variable-variable pairs, since for any other pair, we can decompose until only variables are left. Here we observe, that constant-constant pairs can be decomposed with the $\mathcal{U}$dec rule in the somewhat degenerate case without arguments.

Finally, we observe that the first of the two variable conditions in $\mathcal{U}$elim (the "occurs-in-check") makes sure that we only apply the transformation to unifiable unification problems, whereas the second one is a termination condition that prevents the rule to be applied twice.

The notion of completeness and correctness is a bit different than that for calculi that we compare to the entailment relation. We can think of the "logical system of unifiability" with the model class of sets of substitutions, where a set satisfies an equational problem $\mathcal{E}$, iff all of its members are unifiers. This view induces the soundness and completeness notions presented above.

The three meta-properties above are relatively trivial, but somewhat tedious to prove, so we leave the proofs as an exercise to the reader.

We now fortify our intuition about the unification calculus by two examples. Note that we only need to pursue one possible $\mathcal{U}$ derivation since we have confluence.

Unification Examples

▷ **Example 15.1.27.** Two similar unification problems:

$\begin{array}{ll} f(g(X,X),h(a))=^?f(g(a,Z),h(Z)) & \\ \hline g(X,X)=^?g(a,Z) \wedge h(a)=^?h(Z) & \mathcal{U}\text{dec} \\ \hline X=^?a \wedge X=^?Z \wedge h(a)=^?h(Z) & \mathcal{U}\text{dec} \\ \hline X=^?a \wedge X=^?Z \wedge a=^?Z & \mathcal{U}\text{dec} \\ \hline X=^?a \wedge a=^?Z \wedge a=^?Z & \mathcal{U}\text{elim} \\ \hline X=^?a \wedge Z=^?a \wedge a=^?a & \mathcal{U}\text{elim} \\ \hline X=^?a \wedge Z=^?a & \mathcal{U}\text{triv} \end{array}$	$\begin{array}{ll} f(g(X,X),h(a))=^?f(g(b,Z),h(Z)) & \\ \hline g(X,X)=^?g(b,Z) \wedge h(a)=^?h(Z) & \mathcal{U}\text{dec} \\ \hline X=^?b \wedge X=^?Z \wedge h(a)=^?h(Z) & \mathcal{U}\text{dec} \\ \hline X=^?b \wedge X=^?Z \wedge a=^?Z & \mathcal{U}\text{dec} \\ \hline X=^?b \wedge b=^?Z \wedge a=^?Z & \mathcal{U}\text{elim} \\ \hline X=^?b \wedge Z=^?b \wedge a=^?b & \mathcal{U}\text{elim} \end{array}$
MGU: $[a/X], [a/Z]$	$a=^?b$ not unifiable

Michael Kohlhase: Artificial Intelligence 1 461 2023-09-20

We will now convince ourselves that there cannot be any infinite sequences of transformations in $\mathcal{U}$. Termination is an important property for an algorithm.

The proof we present here is very typical for termination proofs. We map unification problems into a partially ordered set $\langle S, \prec\rangle$ where we know that there cannot be any infinitely descending sequences (we think of this as measuring the unification problems). Then we show that all transformations in $\mathcal{U}$ strictly decrease the measure of the unification problems and argue that if there were an infinite transformation in $\mathcal{U}$, then there would be an infinite descending chain in S, which contradicts our choice of $\langle S, \prec\rangle$.

The crucial step in in coming up with such proofs is finding the right partially ordered set. Fortunately, there are some tools we can make use of. We know that $\langle \mathbb{N}, <\rangle$ is terminating, and there are some ways of lifting component orderings to complex structures. For instance it is well-known that the lexicographic ordering lifts a terminating ordering to a terminating ordering on finite dimensional Cartesian spaces. We show a similar, but less known construction with multisets for our proof.

Unification (Termination)

▷ **Definition 15.1.28.** Let S and T be multisets and $\leq$ a partial ordering on $S \cup T$. Then we define $S \prec^m S$, iff $S = C \uplus T'$ and $T = C \uplus \{t\}$, where $s \leq t$ for all $s \in S'$. We call $\leq^m$ the multiset ordering induced by $\leq$.

▷ **Definition 15.1.29.** We call a variable X solved in an unification problem $\mathcal{E}$, iff $\mathcal{E}$ contains a solved pair $X =^? \mathbf{A}$.

▷ **Lemma 15.1.30.** *If $\prec$ is linear/terminating on S, then $\prec^m$ is linear/terminating on $\mathcal{P}(S)$.*

▷ **Lemma 15.1.31.** *$\mathcal{U}$ is terminating.* *(any $\mathcal{U}$-derivation is finite)*

▷ *Proof:* We prove termination by mapping $\mathcal{U}$ transformation into a Noetherian space.

1. Let $\mu(\mathcal{E}) := \langle n, \mathcal{N} \rangle$, where
   - ▷ n is the number of unsolved variables in $\mathcal{E}$
   - ▷ $\mathcal{N}$ is the multiset of term depths in $\mathcal{E}$
2. The lexicographic order $\prec$ on pairs $\mu(\mathcal{E})$ is decreased by all inference rules.
   - 2.1. $\mathcal{U}$dec and $\mathcal{U}$triv decrease the multiset of term depths without increasing the unsolved variables.
   - 2.2. $\mathcal{U}$elim decreases the number of unsolved variables (by one), but may increase term depths.

But it is very simple to create terminating calculi, e.g. by having no inference rules. So there is one more step to go to turn the termination result into a decidability result: we must make sure that we have enough inference rules so that any unification problem is transformed into solved form if it is unifiable.

First-Order Unification is Decidable

▷ **Definition 15.1.32.** We call an equational problem $\mathcal{E}$ $\mathcal{U}$-reducible, iff there is a $\mathcal{U}$-step $\mathcal{E} \vdash_{\mathcal{U}} \mathcal{F}$ from $\mathcal{E}$.

▷ **Lemma 15.1.33.** *If $\mathcal{E}$ is unifiable but not solved, then it is $\mathcal{U}$-reducible.*

▷ *Proof:* We assume that $\mathcal{E}$ is unifiable but unsolved and show the $\mathcal{U}$ rule that applies.

1. There is an unsolved pair $\mathbf{A} =^? \mathbf{B}$ in $\mathcal{E} = \mathcal{E} \wedge \mathbf{A} =^? \mathbf{B}'$.

   we have two cases
2. $\mathbf{A}, \mathbf{B} \notin \mathcal{V}_\iota$
   - 2.1. then $\mathbf{A} = f(\mathbf{A}^1 \ldots \mathbf{A}^n)$ and $\mathbf{B} = f(\mathbf{B}^1 \ldots \mathbf{B}^n)$, and thus $\mathcal{U}$dec is applicable
3. $\mathbf{A} = X \in \mathrm{free}(\mathcal{E})$
   - 3.1. then $\mathcal{U}$elim (if $\mathbf{B} \neq X$) or $\mathcal{U}$triv (if $\mathbf{B} = X$) is applicable.

▷ **Corollary 15.1.34.** *First-order unification is decidable in PL^1.*

Proof:

▷ 1. $\mathcal{U}$-irreducible unification problems can be reached in finite time by Lemma 15.1.31
2. They are either solved or unsolvable by Lemma 15.1.33, so they provide the answer.

15.1.3 Efficient Unification

Now that we have seen the basic ingredients of an unification algorithm, let us as always consider complexity and efficiency issues.
We start with a look at the complexity of unification and – somewhat surprisingly – find exponential time/space complexity based simply on the fact that the results – the unifiers – can be exponentially large.

Complexity of Unification

▷ **Observation:** Naive implementations of unification are exponential in time and space.

▷ **Example 15.1.35.** Consider the terms

$$\begin{aligned} s_n &= f(f(x_0,x_0),f(f(x_1,x_1),f(\ldots,f(x_{n-1},x_{n-1}))\ldots)) \\ t_n &= f(x_1,f(x_2,f(x_3,f(\cdots,x_n)\cdots))) \end{aligned}$$

▷ The most general unifier of s_n and t_n is

$\sigma_n := [f(x_0,x_0)/x_1],[f(f(x_0,x_0),f(x_0,x_0))/x_2],[f(f(f(x_0,x_0),f(x_0,x_0)),f(f(x_0,x_0),f(x_0,x_0)))/x_3],\ldots$

▷ It contains $\sum_{i=1}^{n} 2^i = 2^{n+1}-2$ occurrences of the variable x_0. (exponential)

▷ **Problem:** The variable x_0 has been copied too often.

▷ **Idea:** Find a term representation that re-uses subterms.

Indeed, the only way to escape this combinatorial explosion is to find representations of substitutions that are more space efficient.

Directed Acyclic Graphs (DAGs) for Terms

▷ **Recall:** Terms in first-order logic are essentially trees.

▷ **Concrete Idea:** Use directed acyclic graphs for representing terms:

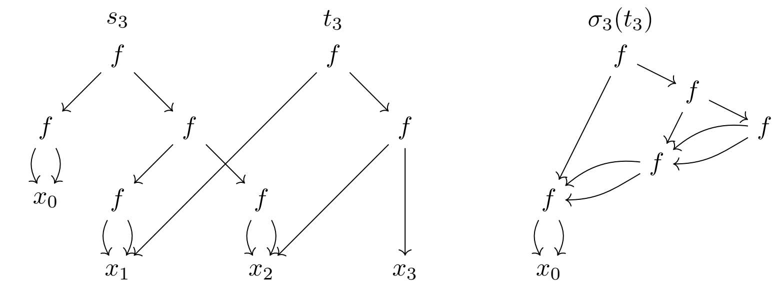
▷ variables my only occur once in the DAG.

▷ subterms can be referenced multiply. (subterm sharing)

▷ we can even represent multiple terms in a common DAG

▷ **Observation 15.1.36.** *Terms can be transformed into DAGs in linear time.*

▷ **Example 15.1.37.** Continuing from Example 15.1.35 … s_3, t_3, and $\sigma_3(s_3)$ as DAGs:



In general: s_n, t_n, and $\sigma_n(s_n)$ only need space in $\mathcal{O}(n)$. (just count)

If we look at the unification algorithm from Definition 15.1.22 and the considerations in the termination proof (Lemma 462) with a particular focus on the role of copying, we easily find the culprit for the exponential blowup: $\mathcal{U}$elim, which applies solved pairs as substitutions.

DAG Unification Algorithm

- **Observation:** In $\mathcal{U}$, the $\mathcal{U}$elim rule applies solved pairs $\leadsto$ subterm duplication.
- **Idea:** Replace $\mathcal{U}$elim the notion of solved forms by something better.
- **Definition 15.1.38.** We say that $X^1 =^? \mathbf{B}^1 \wedge \ldots \wedge X^n =^? \mathbf{B}^n$ is a DAG solved form, iff the X^i are distinct and $X^i \notin \mathrm{free}(\mathbf{B}^j)$ for $i \leq j$.
- **Definition 15.1.39.** The inference system $\mathcal{DU}$ contains rules $\mathcal{U}$dec and $\mathcal{U}$triv from $\mathcal{U}$ plus the following:

$$\frac{\mathcal{E} \wedge X =^? \mathbf{A} \wedge X =^? \mathbf{B} \quad \mathbf{A}, \mathbf{B} \notin \mathcal{V}_\iota \quad |\mathbf{A}| \leq |\mathbf{B}|}{\mathcal{E} \wedge X =^? \mathbf{A} \wedge \mathbf{A} =^? \mathbf{B}} \mathcal{DU}\mathrm{merge}$$

$$\frac{\mathcal{E} \wedge X =^? Y \quad X \neq Y \quad X, Y \in \mathrm{free}(\mathcal{E})}{[Y/X](\mathcal{E}) \wedge X =^? Y} \mathcal{DU}\mathrm{evar}$$

  where $|\mathbf{A}|$ is the number of symbols in $\mathbf{A}$.
- The analysis for $\mathcal{U}$ applies mutatis mutandis.

We will now turn the ideas we have developed in the last couple of slides into a usable functional algorithm. The starting point is treating terms as DAGs. Then we try to conduct the transformation into solved form without adding new nodes.

Unification by DAG-chase

- **Idea:** Extend the Input-DAGs by edges that represent unifiers.
- write $n.a$, if a is the symbol of node n.
- (standard) auxiliary procedures: (all constant or linear time in DAGs)
  - find(n) follows the path from n and returns the end node.

- union(n, m) adds an edge between n and m.
- occur(n, m) determines whether $n.x$ occurs in the DAG with root m.

Algorithm dag−unify

- Input: symmetric pairs of nodes in DAGs

```
fun dag−unify(n,n) = true
  | dag−unify(n.x,m) = if occur(n,m) then true else union(n,m)
  | dag−unify(n.f,m.g) =
      if g!=f then false
      else
         forall (i,j) => dag−unify(find(i),find(j)) (chld m,chld n)
      end
```

- **Observation 15.1.40.** dag−unify *uses linear space, since no new nodes are created, and at most one link per variable.*
- **Problem:** dag−unify still uses exponential time.
- **Example 15.1.41.** Consider terms $f(s_n, f(t'_n, x_n)), f(t_n, f(s'_n, y_n)))$, where $s'_n = [y_i/x_i](s_n)$ und $t'_n = [y_i/x_i](t_n)$.

  dag−unify needs exponentially many recursive calls to unify the nodes x_n and y_n. (they are unified after n calls, but checking needs the time)

Algorithm uf−unify

- **Recall:** dag−unify still uses exponential time.
- **Idea:** Also bind the function nodes, if the arguments are unified.

```
uf−unify(n.f,m.g) =
  if g!=f then false
  else union(n,m);
    forall (i,j) => uf−unify(find(i),find(j)) (chld m,chld n)
  end
```

- This only needs linearly many recursive calls as it directly returns with true or makes a node inaccessible for find.
- Linearly many calls to linear procedures give quadratic running time.
- **Remark:** There are versions of uf−unify that are linear in time and space, but for most purposes, our algorithm suffices.

15.1.4 Implementing First-Order Tableaux

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/26797`.

We now come to some issues (and clarifications) pertaining to implementing proof search for free variable tableaux. They all have to do with the – often overlooked – fact that $\mathcal{T}_1^f\bot$ instantiates the whole tableau.

The first question one may ask for implementation is whether we expect a terminating proof search; after all, $\mathcal{T}_0$ terminated. We will see that the situation for $\mathcal{T}_1^f$ is different.

Termination and Multiplicity in Tableaux

▷ **Recall:**
In $\mathcal{T}_0$, all rules only needed to be applied once.
$\rightsquigarrow$ $\mathcal{T}_0$ terminates and thus induces a decision procedure for PL^0.

▷ **Observation 15.1.42.** *All $\mathcal{T}_1^f$ rules except $\mathcal{T}_1^f\forall$ only need to be applied once.*

▷ **Example 15.1.43.** A tableau proof for $(p(a) \vee p(b)) \Rightarrow (\exists.p())$.

<table>
<tr><th>Start, close left branch</th><th>use $\mathcal{T}_1^f\forall$ again (right branch)</th></tr>
<tr><td>$((p(a) \vee p(b)) \Rightarrow (\exists.p()))^{\mathsf{F}}$
$(p(a) \vee p(b))^{\mathsf{T}}$
$(\exists x.p(x))^{\mathsf{F}}$
$(\forall x.\neg p(x))^{\mathsf{T}}$
$\neg p(y)^{\mathsf{T}}$
$p(y)^{\mathsf{F}}$
$\begin{array}{c|c} p(a)^{\mathsf{T}} & p(b)^{\mathsf{T}} \\ \bot : [a/y] & \end{array}$</td><td>$((p(a) \vee p(b)) \Rightarrow (\exists.p()))^{\mathsf{F}}$
$(p(a) \vee p(b))^{\mathsf{T}}$
$(\exists x.p(x))^{\mathsf{F}}$
$(\forall x.\neg p(x))^{\mathsf{T}}$
$\neg p(a)^{\mathsf{T}}$
$p(a)^{\mathsf{F}}$
$\begin{array}{c|c} p(a)^{\mathsf{T}} & p(b)^{\mathsf{T}} \\ \bot : [a/y] & \neg p(z)^{\mathsf{T}} \\ & p(z)^{\mathsf{F}} \\ & \bot : [b/z] \end{array}$</td></tr>
</table>

After we have used up $p(y)^{\mathsf{F}}$ by applying $[a/y]$ in $\mathcal{T}_1^f\bot$, we have to get a new instance $p(z)^{\mathsf{F}}$ via $\mathcal{T}_1^f\forall$.

▷ **Definition 15.1.44.** Let $\mathcal{T}$ be a tableau for $\mathbf{A}$, and a positive occurrence of $\forall x.\mathbf{B}$ in $\mathbf{A}$, then we call the number of applications of $\mathcal{T}_1^f\forall$ to $\forall x.\mathbf{B}$ its multiplicity.

▷ **Observation 15.1.45.** *Given a prescribed multiplicity for each positive $\forall$, saturation with $\mathcal{T}_1^f$ terminates.*

▷ *Proof sketch:* All $\mathcal{T}_1^f$ rules reduce the number of connectives and negative $\forall$ or the multiplicity of positive $\forall$.

▷ **Theorem 15.1.46.** *$\mathcal{T}_1^f$ is only complete with unbounded multiplicities.*

▷ *Proof sketch:* Replace $p(a) \vee p(b)$ with $p(a_1) \vee \ldots \vee p(a_n)$ in Example 15.1.43.

▷ **Remark:** Otherwise validity in PL^1 would be decidable.

▷ **Implementation:** We need an iterative multiplicity deepening process.

The other thing we need to realize is that there may be multiple ways we can use $\mathcal{T}_1^f\bot$ to close a branch in a tableau, and – as $\mathcal{T}_1^f\bot$ instantiates the whole tableau and not just the branch itself – this choice matters.

Treating $\mathcal{T}_1^f\bot$

- **Recall:** The $\mathcal{T}_1^f\bot$ rule instantiates the whole tableau.
- **Problem:** There may be more than one $\mathcal{T}_1^f\bot$ opportunity on a branch.
- **Example 15.1.47.** Choosing which matters – this tableau does not close!

$$
\begin{array}{c}
(\exists x.(p(a) \wedge p(b) \Rightarrow p()) \wedge (q(b) \Rightarrow q(x)))^{\mathsf{F}} \\
((p(a) \wedge p(b) \Rightarrow p()) \wedge (q(b) \Rightarrow q(y)))^{\mathsf{F}} \\
\begin{array}{c|c}
(p(a) \Rightarrow p(b) \Rightarrow p())^{\mathsf{F}} & (q(b) \Rightarrow q(y))^{\mathsf{F}} \\
p(a)^{\mathsf{T}} & q(b)^{\mathsf{T}} \\
p(b)^{\mathsf{T}} & q(y)^{\mathsf{F}} \\
p(y)^{\mathsf{F}} & \\
\bot : [a/y] &
\end{array}
\end{array}
$$

  choosing the other $\mathcal{T}_1^f\bot$ in the left branch allows closure.

- **Idea:** Two ways of systematic proof search in $\mathcal{T}_1^f$:
  - backtracking search over $\mathcal{T}_1^f\bot$ opportunities
  - saturate without $\mathcal{T}_1^f\bot$ and find spanning matings (next slide)

The method of spanning matings follows the intuition that if we do not have good information on how to decide for a pair of opposite literals on a branch to use in $\mathcal{T}_1^f\bot$, we delay the choice by initially disregarding the rule altogether during saturation and then – in a later phase– looking for a configuration of cuts that have a joint overall unifier. The big advantage of this is that we only need to know that one exists, we do not need to compute or apply it, which would lead to exponential blow-up as we have seen above.

Spanning Matings for $\mathcal{T}_1^f\bot$

- **Observation 15.1.48.** *$\mathcal{T}_1^f$ without $\mathcal{T}_1^f\bot$ is terminating and confluent for given multiplicities.*
- **Idea:** Saturate without $\mathcal{T}_1^f\bot$ and treat all cuts at the same time (later).
- **Definition 15.1.49.**

  Let $\mathcal{T}$ be a $\mathcal{T}_1^f$ tableau, then we call a unification problem $\mathcal{E}:=\mathbf{A}_1=^?\mathbf{B}_1 \wedge \ldots \wedge \mathbf{A}_n=^?\mathbf{B}_n$ a mating for $\mathcal{T}$, iff $\mathbf{A}_i{}^{\mathsf{T}}$ and $\mathbf{B}_i{}^{\mathsf{F}}$ occur in the same branch in $\mathcal{T}$.

  We say that $\mathcal{E}$ is a spanning mating, if $\mathcal{E}$ is unifiable and every branch $\mathcal{B}$ of $\mathcal{T}$ contains $\mathbf{A}_i{}^{\mathsf{T}}$ and $\mathbf{B}_i{}^{\mathsf{F}}$ for some i.
- **Theorem 15.1.50.** *A $\mathcal{T}_1^f$-tableau with a spanning mating induces a closed $\mathcal{T}_1$ tableau.*

▷ *Proof sketch:* Just apply the unifier of the spanning mating.

▷ **Idea:** Existence is sufficient, we do not need to compute the unifier.

▷ **Implementation:** Saturate without $\mathcal{T}_1^f\bot$, backtracking search for spanning matings with $\mathcal{DU}$, adding pairs incrementally.

Excursion: Now that we understand basic unification theory, we can come to the meta-theoretical properties of the tableau calculus. We delegate this discussion to**??**.

15.2 First-Order Resolution

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26817`.

First-Order Resolution (and CNF)

▷ **Definition 15.2.1.** The first-order CNF calculus CNF_1 is given by the inference rules of CNF_0 extended by the following quantifier rules:

$$\frac{(\forall X.\mathbf{A})^{\mathsf{T}} \vee \mathbf{C} \quad Z \notin (\mathsf{free}(\mathbf{A}) \cup \mathsf{free}(\mathbf{C}))}{([Z/X](\mathbf{A}))^{\mathsf{T}} \vee \mathbf{C}}$$

$$\frac{(\forall X.\mathbf{A})^{\mathsf{F}} \vee \mathbf{C} \quad \{X_1, \ldots, X_k\} = \mathsf{free}(\forall X.\mathbf{A}) \quad f \in \Sigma_k^{sk} \text{ new}}{([f(X_1, \ldots, X_k)/X](\mathbf{A}))^{\mathsf{F}} \vee \mathbf{C}}$$

$CNF_1(\Phi)$ is the set of all clauses that can be derived from Φ.

▷ **Definition 15.2.2 (First-Order Resolution Calculus).** The First-order resolution calculus ($\mathcal{R}_1$) is a test calculus that manipulates formulae in conjunctive normal form. $\mathcal{R}_1$ has two inference rules:

$$\frac{\mathbf{A}^{\mathsf{T}} \vee \mathbf{C} \quad \mathbf{B}^{\mathsf{F}} \vee \mathbf{D} \quad \sigma = \mathbf{mgu}(\mathbf{A}, \mathbf{B})}{(\sigma(\mathbf{C})) \vee (\sigma(\mathbf{D}))} \qquad \frac{\mathbf{A}^{\alpha} \vee \mathbf{B}^{\alpha} \vee \mathbf{C} \quad \sigma = \mathbf{mgu}(\mathbf{A}, \mathbf{B})}{(\sigma(\mathbf{A})) \vee (\sigma(\mathbf{C}))}$$

First-Order CNF – Derived Rules

▷ **Definition 15.2.3.** The following inference rules are derivable from the ones above via $(\exists X.\mathbf{A}) = \neg(\forall X.\neg\mathbf{A})$:

$$\frac{(\exists X.\mathbf{A})^{\mathsf{T}} \vee \mathbf{C} \quad \{X_1, \ldots, X_k\} = \mathsf{free}(\forall X.\mathbf{A}) \quad f \in \Sigma_k^{sk} \text{ new}}{([f(X_1, \ldots, X_k)/X](\mathbf{A}))^{\mathsf{T}} \vee \mathbf{C}}$$

$$\frac{(\exists X.\mathbf{A})^{\mathsf{F}} \vee \mathbf{C} \quad Z \notin (\mathsf{free}(\mathbf{A}) \cup \mathsf{free}(\mathbf{C}))}{([Z/X](\mathbf{A}))^{\mathsf{F}} \vee \mathbf{C}}$$

Excursion: Again, we relegate the meta-theoretical properties of the first-order resolution calculus to??.

15.2.1 Resolution Examples

Col. West, *a Criminal?*

▷ **Example 15.2.4.** From [RN09]

> The law says it is a crime for an American to sell weapons to hostile nations. The country Nono, an enemy of America, has some missiles, and all of its missiles were sold to it by Colonel West, who is American.

Prove that Col. West is a criminal.

▷ **Remark:** Modern resolution theorem provers prove this in less than 50ms.

▷ **Problem:** That is only true, if we only give the theorem prover exactly the right laws and background knowledge. If we give it all of them, it drowns in the combinatory explosion.

▷ Let us build a resolution proof for the claim above.

▷ **But first** we must translate the situation into first-order logic clauses.

▷ **Convention:** In what follows, for better readability we will sometimes write implications $P \wedge Q \wedge R \Rightarrow S$ instead of clauses $P^{\mathsf{F}} \vee Q^{\mathsf{F}} \vee R^{\mathsf{F}} \vee S^{\mathsf{T}}$.

Col. West, *a Criminal?*

▷ *It is a crime for an American to sell weapons to hostile nations*:
Clause: $\mathsf{ami}(X_1) \wedge \mathsf{weap}(Y_1) \wedge \mathsf{sell}(X_1, Y_1, Z_1) \wedge \mathsf{host}(Z_1) \Rightarrow \mathsf{crook}(X_1)$

▷ *Nono has some missiles*: $\exists X.\mathsf{own}(\mathsf{NN}, X) \wedge \mathsf{mle}(X)$
Clauses: $\mathsf{own}(\mathsf{NN}, c)^{\mathsf{T}}$ and $\mathsf{mle}(c)$ (c is Skolem constant)

▷ *All of Nono's missiles were sold to it by Colonel West.*
Clause: $\mathsf{mle}(X_2) \wedge \mathsf{own}(\mathsf{NN}, X_2) \Rightarrow \mathsf{sell}(\mathsf{West}, X_2, \mathsf{NN})$

▷ *Missiles are weapons*:
Clause: $\mathsf{mle}(X_3) \Rightarrow \mathsf{weap}(X_3)$

▷ *An enemy of America counts as "hostile"*:
Clause: $\mathsf{enmy}(X_4, \mathsf{USA}) \Rightarrow \mathsf{host}(X_4)$

▷ *West is an American:*
Clause: $\mathsf{ami}(\mathsf{West})$

▷ *The country Nono is an enemy of America*:
$\mathsf{enmy}(\mathsf{NN}, \mathsf{USA})$

Col. West, *a Criminal!* PL1 Resolution Proof

$\mathsf{ami}(X_1)^F \vee \mathsf{weap}(Y_1)^F \vee \mathsf{sell}(X_1,Y_1,Z_1)^F \vee \mathsf{host}(Z_1)^F \vee \mathsf{crook}(X_1)^T$ $\quad\mathsf{crook}(\mathsf{West})^F$

$[\mathsf{West}/X_1]$

$\mathsf{ami}(\mathsf{West})^T$ $\quad\mathsf{ami}(\mathsf{West})^F \vee \mathsf{weap}(Y_1)^F \vee \mathsf{sell}(\mathsf{West},Y_1,Z_1)^F \vee \mathsf{host}(Z_1)^F$

$\mathsf{mle}(X_3)^F \vee \mathsf{weap}(X_3)^T$ $\quad\mathsf{weap}(Y_1)^F \vee \mathsf{sell}(\mathsf{West},Y_1,Z_1)^F \vee \mathsf{host}(Z_1)^F$

$[Y_1/X_3]$

$\mathsf{mle}(c)^T$ $\quad\mathsf{mle}(Y_1)^F \vee \mathsf{sell}(\mathsf{West},Y_1,Z_1)^F \vee \mathsf{host}(Z_1)^F$

$[c/Y_1]$ $\quad\mathsf{mle}(X_2)^F \vee \mathsf{own}(\mathsf{NN},X_2)^F \vee \mathsf{sell}(\mathsf{West},X_2,\mathsf{NN})^T$

$\mathsf{sell}(\mathsf{West},c,Z_1)^F \vee \mathsf{host}(Z_1)^F$ $\quad[c/X_2]$

$[\mathsf{NN}/Z_1]$

$\mathsf{mle}(c)^T$ $\quad\mathsf{mle}(c)^F \vee \mathsf{own}(\mathsf{NN},c)^F \vee \mathsf{host}(\mathsf{NN})^F$

$\mathsf{own}(\mathsf{NN},c)^T$ $\quad\mathsf{own}(\mathsf{NN},c)^F \vee \mathsf{host}(\mathsf{NN})^F$

$\mathsf{enmy}(X_4,\mathsf{USA})^F \vee \mathsf{host}(X_4)^T$ $\quad\mathsf{host}(\mathsf{NN})^F$

$[\mathsf{NN}/X_4]$

$\mathsf{enmy}(\mathsf{NN},\mathsf{USA})^T$ $\quad\mathsf{enmy}(\mathsf{NN},\mathsf{USA})^F$

$\Box$

Curiosity Killed the Cat?

▷ **Example 15.2.5.** From [RN09]

Everyone who loves all animals is loved by someone.
Anyone who kills an animal is loved by noone.
Jack loves all animals.
Cats are animals.
Either Jack or curiosity killed the cat (whose name is "Garfield").

Prove that curiosity killed the cat.

Curiosity Killed the Cat? Clauses

▷ *Everyone who loves all animals is loved by someone*:
$\forall X.(\forall Y.\mathsf{animal}(Y) \Rightarrow \mathsf{love}(X,Y)) \Rightarrow (\exists.\mathsf{love}(Z,X))$
Clauses: $\mathsf{animal}(g(X_1))^T \vee \mathsf{love}(g(X_1),X_1)^T$ and $\mathsf{love}(X_2,f(X_2))^F \vee \mathsf{love}(g(X_2),X_2)^T$

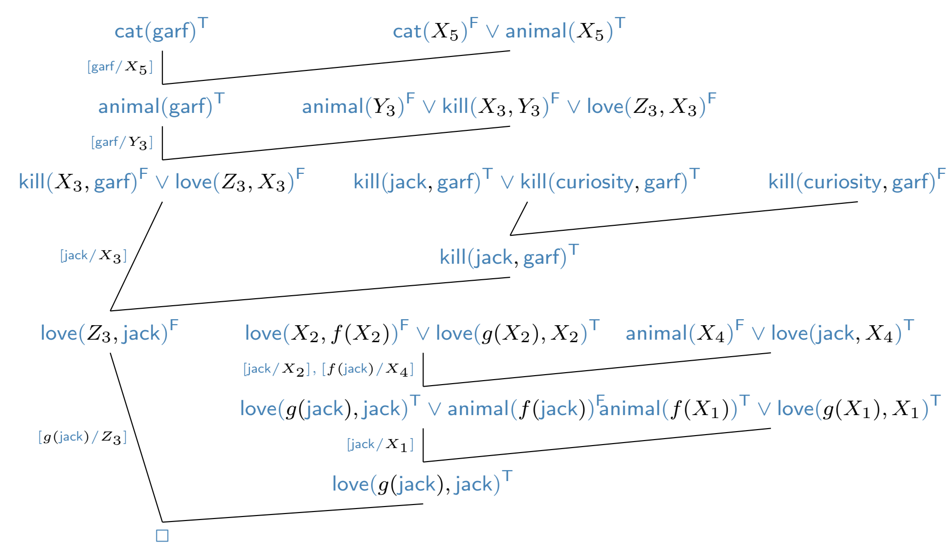
▷ *Anyone who kills an animal is loved by noone*:
$\forall X.\exists Y.\mathsf{animal}(Y) \wedge \mathsf{kill}(X,Y) \Rightarrow (\forall.\neg\mathsf{love}(Z,X))$
Clause: $\mathsf{animal}(Y_3)^F \vee \mathsf{kill}(X_3,Y_3)^F \vee \mathsf{love}(Z_3,X_3)^F$

▷ *Jack loves all animals*:

Clause: $\mathsf{animal}(X_4)^{\mathsf{F}} \vee \mathsf{love}(\mathsf{jack}, X_4)^{\mathsf{T}}$

- ▷ *Cats are animals*:
  Clause: $\mathsf{cat}(X_5)^{\mathsf{F}} \vee \mathsf{animal}(X_5)^{\mathsf{T}}$
- ▷ *Either Jack or curiosity killed the cat (whose name is "Garfield")*:
  Clauses: $\mathsf{kill}(\mathsf{jack}, \mathsf{garf})^{\mathsf{T}} \vee \mathsf{kill}(\mathsf{curiosity}, \mathsf{garf})^{\mathsf{T}}$ and $\mathsf{cat}(\mathsf{garf})^{\mathsf{T}}$

Curiosity Killed the Cat! PL1 Resolution Proof



Excursion: A full analysis of any calculus needs a completeness proof. We will not cover this in the course, but provide one for the calculi introduced so far in**??**.

15.3 Logic Programming as Resolution Theorem Proving

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26820`. To understand Prolog better, we can interpret the language of Prolog as resolution in PL[1].

We know all this already

- ▷ Goals, goal sets, rules, and facts are just clauses. (called Horn clauses)
- ▷ **Observation 15.3.1 (Rule).** *H:–B_1,...,B_n. corresponds to* $H^{\mathsf{T}} \vee B_1{}^{\mathsf{F}} \vee \ldots \vee B_n{}^{\mathsf{F}}$ *(head the only positive literal)*
- ▷ **Observation 15.3.2 (Goal set).** ?– G_1,...,G_n. *corresponds to* $G_1{}^{\mathsf{F}} \vee \ldots \vee G_n{}^{\mathsf{F}}$
- ▷ **Observation 15.3.3 (Fact).** *F. corresponds to the unit clause* F^{T}.

▷ **Definition 15.3.4.** A Horn clause is a clause with at most one positive literal.

▷ **Recall:** Backchaining as search:

- ▹ state = tuple of goals; goal state = empty list (of goals).
- ▹ $next(\langle G, R_1, \ldots, R_l\rangle){:=}\langle\sigma(B_1), \ldots, \sigma(B_m), \sigma(R_1), \ldots, \sigma(R_l)\rangle$ if there is a rule $H{:-}B_1, \ldots, B_m$. and a substitution σ with $\sigma(H) = \sigma(G)$.

▷ **Note:** Backchaining becomes resolution

$$\frac{P^{\mathsf{T}} \vee \mathbf{A} \quad P^{\mathsf{F}} \vee \mathbf{B}}{\mathbf{A} \vee \mathbf{B}}$$

positive, unit-resulting hyperresolution (PURR)

This observation helps us understand Prolog better, and use implementation techniques from theorem proving.

PROLOG (Horn Logic)

▷ **Definition 15.3.5.** A clause is called a Horn clause, iff contains at most one positive literal, i.e. if it is of the form $B_1{}^{\mathsf{F}} \vee \ldots \vee B_n{}^{\mathsf{F}} \vee A^{\mathsf{T}}$ – i.e. A:$-B_1, \ldots, B_n$. in Prolog notation.

- ▹ Rule clause: general case, e.g. fallible(X) : human(X).
- ▹ Fact clause: no negative literals, e.g. human(sokrates).
- ▹ Program: set of rule and fact clauses.
- ▹ Query: no positive literals: e.g. ?– fallible(X),greek(X).

▷ **Definition 15.3.6.** Horn logic is the formal system whose language is the set of Horn clauses together with the calculus $\mathcal{H}$ given by MP, $\wedge I$, and Subst.

▷ **Definition 15.3.7.** A logic program P entails a query Q with answer substitution σ, iff there is a $\mathcal{H}$ derivation D of Q from P and σ is the combined substitution of the Subst instances in D.

PROLOG: Our Example

▷ **Program:**

```
human(leibniz).
human(sokrates).
greek(sokrates).
fallible(X):—human(X).
```

▷ **Example 15.3.8 (Query).** ?– fallible(X),greek(X).

▷ **Answer substitution:** [sokrates/X]

To gain an intuition for this quite abstract definition let us consider a concrete knowledge base about cars. Instead of writing down everything we know about cars, we only write down that cars are motor vehicles with four wheels and that a particular object c has a motor and four wheels. We can see that the fact that c is a car can be derived from this. Given our definition of a knowledge base as the deductive closure of the facts and rule explicitly written down, the assertion that c is a car is in the induced knowledge base, which is what we are after.

Knowledge Base (Example)

▷ **Example 15.3.9.** car(c). is in the knowlege base generated by

```
has_motor(c).
has_wheels(c,4).
car(X):— has_motor(X),has_wheels(X,4).
```

$$\dfrac{\dfrac{m(c) \quad w(c,4)}{m(c) \wedge w(c,4)}\wedge I \qquad \dfrac{m(x) \wedge w(x,4) \Rightarrow car()}{m(c) \wedge w(c,4) \Rightarrow car()}\text{Subst}}{car(c)}\text{MP}$$

In this very simple example $car(c)$ is about the only fact we can derive, but in general, knowledge bases can be infinite (we will see examples below).

Why Only Horn Clauses?

▷ General clauses of the form A1,. . .,An : B1,. . .,Bn.

▷ e.g. greek(sokrates),greek(perikles)

- ▷ **Question**: Are there fallible greeks?
- ▷ **Indefinite answer**: Yes, Perikles or Sokrates
- ▷ Warning: how about Sokrates and Perikles?

▷ e.g. greek(sokrates),roman(sokrates):—.

- ▷ **Query**: Are there fallible greeks?
- ▷ **Answer**: Yes, Sokrates, if he is not a roman
- ▷ Is this abduction?????

Three Principal Modes of Inference

▷ **Definition 15.3.10.** Deduction $\widehat{=}$ knowledge extension

▷ **Example 15.3.11.** $\dfrac{rains \Rightarrow wet\_street \quad rains}{wet\_street}D$

▷ **Definition 15.3.12.** Abduction $\widehat{=}$ explanation

▷ **Example 15.3.13.** $\dfrac{rains \Rightarrow wet\_street \quad wet\_street}{rains}A$

▷ **Definition 15.3.14.** Induction $\widehat{=}$ learning general rules from examples

▷ **Example 15.3.15.** $\dfrac{wet\_street \quad rains}{rains \Rightarrow wet\_street}I$

Chapter 16

Knowledge Representation and the Semantic Web

The field of "Knowledge Representation" is usually taken to be an area in Artificial Intelligence that studies the representation of knowledge in formal systems and how to leverage inference techniques to generate new knowledge items from existing ones. Note that this definition coincides with with what we know as logical systems in this course. This is the view taken by the subfield of "description logics", but restricted to the case, where the logical systems have an entailment relation to ensure applicability. This chapter is organized as follows. We will first give a general introduction to the concepts of knowledge representation using semantic networks – an early and very intuitive approach to knowledge representation – as an object-to-think-with. In section 16.2 we introduce the principles and services of logic-based knowledge-representation using a non-standard interpretation of propositional logic as the basis, this gives us a formal account of the taxonomic part of semantic networks. In ?? we introduce the logic $\mathcal{ALC}$ that adds relations (called "roles") and restricted quantification and thus gives us the full expressive power of semantic networks. Thus $\mathcal{ALC}$ can be seen as a prototype description logic. In section 16.4 we show how description logics are applied as the basis of the "semantic web".

16.1 Introduction to Knowledge Representation

A Video Nugget covering the introduction to knowledge representation can be found at `https://fau.tv/clip/id/27279`.

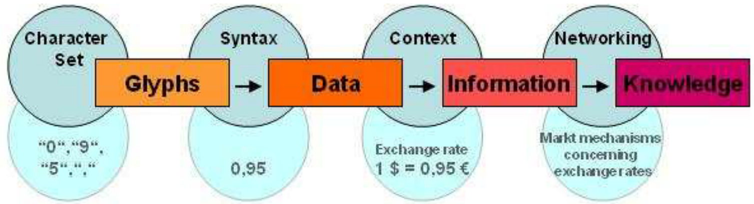
Before we start into the development of description logics, we set the stage by looking into alternatives for knowledge representation.

16.1.1 Knowledge & Representation

To approach the question of knowledge representation, we first have to ask ourselves, what knowledge might be. This is a difficult question that has kept philosophers occupied for millennia. We will not answer this question in this course, but only allude to and discuss some aspects that are relevant to our cause of knowledge representation.

What is knowledge? Why Representation?

- ▷ Lots/all of (academic) disciplines deal with knowledge!
- ▷ According to Probst/Raub/Romhardt [PRR97]



▷ **For the purposes of this course:** Knowledge is the information necessary to support intelligent reasoning!

representation	can be used to determine
set of words	whether a word is admissible
list of words	the rank of a word
a lexicon	translation and/or grammatical function
structure	function

According to an influential view of [PRR97], knowledge appears in layers. Staring with a character set that defines a set of glyphs, we can add syntax that turns mere strings into data. Adding context information gives information, and finally, by relating the information to other information allows to draw conclusions, turning information into knowledge.

Note that we already have aspects of representation and function in the diagram at the top of the slide. In this, the additional functionaltiy added in the successive layers gives the representations more and more functions, until we reach the knowledge level, where the function is given by inferencing. In the second example, we can see that representations determine possible functions.

Let us now strengthen our intuition about knowledge by contrasting knowledge representations from "regular" data structures in computation.

Knowledge Representation vs. Data Structures

▷ **Idea:** Representation as structure and function.

- ▹ the representation determines the content theory (what is the data?)
- ▹ the function determines the process model (what do we do with the data?)

▷ **Question:** Why do we use the term "knowledge representation" rather than

- ▹ data structures? (sets, lists, ... above)
- ▹ information representation? (it is information)

▷ **Answer:**

No good reason other than AI practice, with the intuition that

- ▹ data is simple and general (supports many algorithms)
- ▹ knowledge is complex (has distinguished process model)

As knowledge is such a central notion in artificial intelligence, it is not surprising that there are multiple approaches to dealing with it. We will only deal with the first one and leave the others to self-study.

Some Paradigms for Knowledge Representation in AI/NLP

- GOFAI (good old-fashioned AI)
  - symbolic knowledge representation, process model based on heuristic search
- Statistical, corpus-based approaches.
  - symbolic representation, process model based on machine learning
  - knowledge is divided into symbolic- and statistical (search) knowledge
- The connectionist approach
  - sub-symbolic representation, process model based on primitive processing elements (nodes) and weighted links
  - knowledge is only present in activation patters, etc.

When assessing the relative strengths of the respective approaches, we should evaluate them with respect to a pre-determined set of criteria.

KR Approaches/Evaluation Criteria

- **Definition 16.1.1.** The evaluation criteria for knowledge representation approaches are:
  - Expressive adequacy: What can be represented, what distinctions are supported.
  - Reasoning efficiency: Can the representation support processing that generates results in acceptable speed?
  - Primitives: What are the primitive elements of representation, are they intuitive, cognitively adequate?
  - Meta representation: Knowledge about knowledge
  - Completeness: The problems of reasoning with knowledge that is known to be incomplete.

16.1.2 Semantic Networks

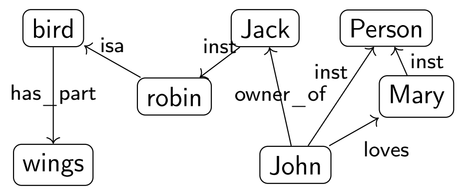
A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27280`.
To get a feeling for early knowledge representation approaches from which description logics developed, we take a look at "semantic networks" and contrast them to logical approaches.
Semantic networks are a very simple way of arranging knowledge about objects and concepts and their relationships in a graph.

Semantic Networks [CQ69]

- **Definition 16.1.2.** A semantic network is a directed graph for representing knowledge:

- nodes represent objects and concepts (classes of objects) (e.g. John (object) and bird (concept))
- edges (called links) represent relations between these (isa, father_of, belongs_to)

▷ **Example 16.1.3.** A semantic network for birds and persons:



▷ **Problem:** How do we derive new information from such a network?

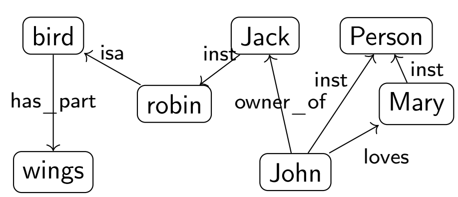
▷ **Idea:** Encode taxonomic information about objects and concepts in special links ("isa" and "inst") and specify property inheritance along them in the process model.

Even though the network in Example 16.1.3 is very intuitive (we immediately understand the concepts depicted), it is unclear how we (and more importantly a machine that does not associate meaning with the labels of the nodes and edges) can draw inferences from the "knowledge" represented.

Deriving Knowledge Implicit in Semantic Networks

▷ **Observation 16.1.4.** *There is more knowledge in a semantic network than is explicitly written down.*

▷ **Example 16.1.5.** In the network below, we "know" that *robins have wings* and in particular, *Jack has wings*.



▷ **Idea:** Links labeled with "isa" and "inst" are special: they propagate properties encoded by other links.

▷ **Definition 16.1.6.** We call links labeled by

- "isa" an inclusion or isa link (inclusion of concepts)
- "inst" instance or inst link (concept membership)

We now make the idea of "propagating properties" rigorous by defining the notion of derived relations, i.e. the relations that are left implicit in the network, but can be added without changing its meaning.

Deriving Knowledge Semantic Networks

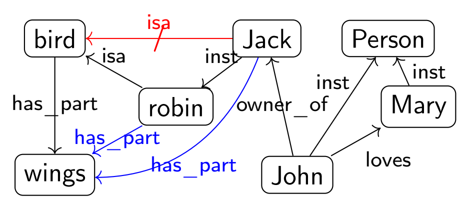
- **Definition 16.1.7 (Inference in Semantic Networks).** We call all link labels except "inst" and "isa" in a semantic network relations.

  Let N be a semantic network and R a relation in N such that $A \xrightarrow{\text{isa}} B \xrightarrow{R} C$ or $A \xrightarrow{\text{inst}} B \xrightarrow{R} C$, then we can derive a relation $A \xrightarrow{R} C$ in N.

  The process of deriving new concepts and relations from existing ones is called inference and concepts/relations that are only available via inference implicit (in a semantic network).

- **Intuition:** Derived relations represent knowledge that is implicit in the network; they could be added, but usually are not to avoid clutter.

- **Example 16.1.8.** Derived relations in Example 16.1.5



- **Slogan:** Get out more knowledge from a semantic networks than you put in.

Note that Definition 16.1.7 does not quite allow to derive that *Jack is a bird* (did you spot that "isa" is not a relation that can be inferred?), even though we know it is true in the world. This shows us that inference in semantic networks has be to very carefully defined and may not be "complete", i.e. there are things that are true in the real world that our inference procedure does not capture.

Dually, if we are not careful, then the inference procedure might derive properties that are not true in the real world even if all the properties explicitly put into the network are. We call such an inference procedure unsound or incorrect.

These are two general phenomena we have to keep an eye on.

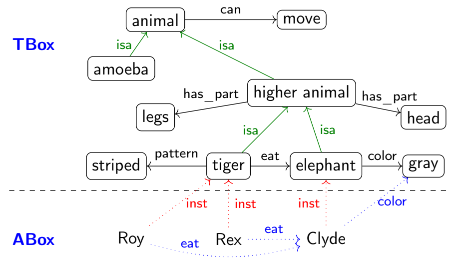
Another problem is that semantic nets (e.g. in in Example 16.1.3) confuse two kinds of concepts: individuals (represented by proper names like *John* and *Jack*) and concepts (nouns like *robin* and *bird*). Even though the isa and inst link already acknowledge this distinction, the "has_part" and "loves" relations are at different levels entirely, but not distinguished in the networks.

Terminologies and Assertions

- *Remark 16.1.9.* We should distinguish concepts from objects.

- **Definition 16.1.10.** We call the subgraph of a semantic network N spanned by the isa links and relations between concepts the terminology (or TBox, or the famous Isa Hierarchy) and the subgraph spanned by the inst links and relations between objects, the assertions (or ABox) of N.

- **Example 16.1.11.** In this network we keep objects concept apart notationally:



In particular we have objects "Rex", "Roy", and "Clyde", which have (derived) relations (e.g. *Clyde* is *gray*).

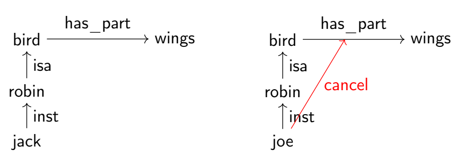
Michael Kohlhase: Artificial Intelligence 1 494 2023-09-20

But there are severe shortcomings of semantic networks: the suggestive shape and node names give (humans) a false sense of meaning, and the inference rules are only given in the process model (the implementation of the semantic network processing system).

This makes it very difficult to assess the strength of the inference system and make assertions e.g. about completeness.

Limitations of Semantic Networks

- What is the meaning of a link?
  - link labels are very suggestive (misleading for humans)
  - meaning of link types defined in the process model (no denotational semantics)
- **Problem:** No distinction of optional and defining traits!
- **Example 16.1.12.** Consider a robin that has lost its wings in an accident:



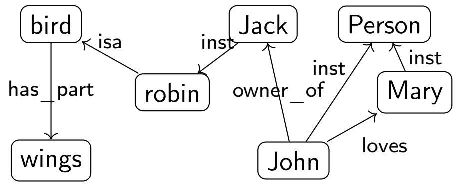
"Cancel-links" have been proposed, but their status and process model are debatable.

Michael Kohlhase: Artificial Intelligence 1 495 2023-09-20

To alleviate the perceived drawbacks of semantic networks, we can contemplate another notation that is more linear and thus more easily implemented: function/argument notation.

Another Notation for Semantic Networks

- **Definition 16.1.13.** Function/argument notation for semantic networks
  - interprets nodes as arguments (reification to individuals)
  - interprets links as functions (predicates actually)
- **Example 16.1.14.**



isa(robin,bird)
haspart(bird,wings)
inst(Jack,robin)
owner_of(John, robin)
loves(John,Mary)

▷ **Evaluation:**

+ linear notation (equivalent, but better to implement on a computer)

+ easy to give process model by deduction (e.g. in Prolog)

– worse locality properties (networks are associative)

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 496 2023-09-20 SOME RIGHTS RESERVED

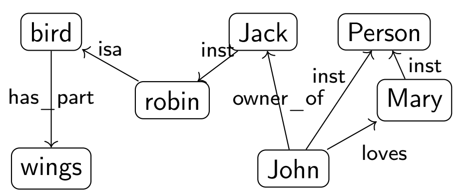
Indeed the function/argument notation is the immediate idea how one would naturally represent semantic networks for implementation.

This notation has been also characterized as subject/predicate/object triples, alluding to simple (English) sentences. This will play a role in the "semantic web" later.

Building on the function/argument notation from above, we can now give a formal semantics for semantic network: we translate them into first-order logic and use the semantics of that.

A Denotational Semantics for Semantic Networks

▷ **Observation:** If we handle isa and inst links specially in function/argument notation



robin $\subseteq$ bird
haspart(bird,wings)
Jack$\in$robin
owner_of(John, Jack)
loves(John,Mary)

it looks like first-order logic, if we take

- ▹ $a{\in}S$ to mean $S(a)$ for an object a and a concept S.
- ▹ $A \subseteq B$ to mean $\forall X.A(X) \Rightarrow B(X)$ and concepts A and B
- ▹ $R(A, B)$ to mean $\forall X.A(X) \Rightarrow (\exists Y.B(Y) \wedge R(X, Y))$ for a relation R.

▷ **Idea:** Take first-order deduction as process model (gives inheritance for free)

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 497 2023-09-20 SOME RIGHTS RESERVED

Indeed, the semantics induced by the translation to first-order logic, gives the intuitive meaning to the semantic networks. Note that this only holds only for the features of semantic networks that are representable in this way, e.g. the "cancel links" shown above are not (and that is a feature, not a bug).

But even more importantly, the translation to first-order logic gives a first process model: we can use first-order inference to compute the set of inferences that can be drawn from a semantic network.

Before we go on, let us have a look at an important application of knowledge representation technologies: the semantic web.

16.1.3 The Semantic Web

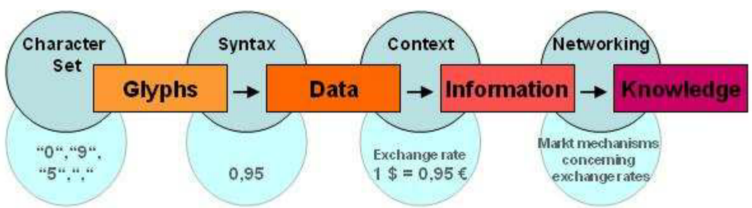
A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27281`.
We will now define the term semantic web and discuss the pertinent ideas involved. There are two central ones, we will cover here:

- Information and data come in different levels of explicitness; this is usually visualized by a "ladder" of information.
- if information is sufficiently machine-understandable, then we can automate drawing conclusions.

The Semantic Web

▷ **Definition 16.1.15.** The semantic web is the result including of semantic content in web pages with the aim of converting the WWW into a machine-understandable "web of data", where inference based services can add value to the ecosystem.

▷ **Idea:** Move web content up the ladder, use inference to make connections.



▷ **Example 16.1.16.** Information not explicitly represented (in one place)

Query: *Who was US president when Barak Obama was born?*

Google: *...BIRTH DATE: August 04, 1961...*

Query: *Who was US president in 1961?*

Google: *President: Dwight D. Eisenhower [...] John F. Kennedy (starting Jan. 20.)*

Humans understand the text and combine the information to get the answer. Machines need more than just text $\rightsquigarrow$ semantic web technology.

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 498 2023-09-20

The term "semantic web" was coined by Tim Berners Lee in analogy to semantic networks, only applied to the world wide web. And as for semantic networks, where we have inference processes that allow us the recover information that is not explicitly represented from the network (here the world-wide-web).

To see that problems have to be solved, to arrive at the semantic web, we will now look at a concrete example about the "semantics" in web pages. Here is one that looks typical enough.

What is the Information a User sees?

▷ **Example 16.1.17.** Take the following web-site with a conference announcement

WWW2002
The eleventh International World Wide Web Conference
Sheraton Waikiki Hotel
Honolulu, Hawaii, USA

7-11 May 2002

Registered participants coming from
Australia, Canada, Chile Denmark, France, Germany, Ghana, Hong Kong, India,
Ireland, Italy, Japan, Malta, New Zealand, The Netherlands, Norway,
Singapore, Switzerland, the United Kingdom, the United States, Vietnam, Zaire

On the 7th May Honolulu will provide the backdrop of the eleventh
International World Wide Web Conference.

Speakers confirmed
Tim Berners-Lee: Tim is the well known inventor of the Web,
Ian Foster: Ian is the pioneer of the Grid, the next generation internet.

But as for semantic networks, what you as a human can see ("understand" really) is deceptive, so let us obfuscate the document to confuse your "semantic processor". This gives an impression of what the computer "sees".

What the machine sees

▷ **Example 16.1.18.** Here is what the machine "sees" from the conference announcement:

𝒲𝒲𝒲∈′′∈
𝒯⟨⌉⌉↕⌉⊑⌉\⊔⟨ℐ\⊔⌉∇\⊣⊔⟩≀\⊣↕𝒲≀∇↕⌈𝒲⟩⌈⌉𝒲⌉⌊𝒞≀\{⌉∇⌉\⌋⌉
𝒮⟨⌉∇⊣⊔≀\𝒲⊣⟩‖⟩‖⟩ℋ≀⊔⌉↕
ℋ≀\≀↕⊓↕⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜
↖∞∞ℳ⊣†∈′′∈

ℛ⌉}⟩∫⊔⌉∇⌉⌈√⊣∇⊔⟩⌋⟩√⊣\⊔∫⌋≀⇕⟩\}{∇≀⇕
𝒜⊓∫⊔∇⊣↕⟩⊣⇔𝒞⊣\⊣⌈⊣⇔𝒞⟨⟩↕⌉𝒟⌉\⇕⊣∇‖⇔ℱ∇⊣\⌋⌉⇔𝒢⌉∇⇕⊣\†⇔𝒢⟨⊣\⊣⇔ℋ≀\}𝒦≀\}⇔ℐ\⌈⟩⊣⇔
ℐ∇⌉↕⊣\⌈⇔ℐ⊔⊣↕†⇔𝒥⊣√⊣\⇔ℳ⊣↕⊔⊣⇔𝒩⌉⊒𝒵⌉⊣↕⊣\⌈⇔𝒯⟨⌉𝒩⌉⊔⟨⌉∇↕⊣\⌈∫⇔𝒩≀∇⊒⊣†⇔
𝒮⟩\}⊣√≀∇⌉⇔𝒮⊒⟩⊔‡⌉∇↕⊣\⌈⇔⊔⟨⌉𝒰\⟩⊔⌉⌈𝒦⟩\}⌈≀⇕⇔⊔⟨⌉𝒰\⟩⊔⌉⌈𝒮⊔⊣⊔⌉∫⇔𝒱⟩⌉⊔\⊣⇕⇔𝒵⊣⟩∇⌉

𝒪\⊔⟨⌉⊔⟨ℳ⊣†ℋ≀\≀↕⊓↕⊓⊒⟩↕↕√∇≀⊑⟩⌈⌉⊔⟨⌉⌊⊣⌋‖⌈∇≀√≀{⊔⟨⌉⌉↕⌉⊑⌉\⊔⟨
ℐ\⊔⌉∇\⊣⊔⟩≀\⊣↕𝒲≀∇↕⌈𝒲⟩⌈⌉𝒲⌉⌊𝒞≀\{⌉∇⌉\⌋⌉↙

𝒮√⌉⊣‖⌉∇∫⌋≀\{⟩∇⇕⌉⌈
𝒯⟩⇕ℬ⌉∇\⌉∇∫↖ℒ⌉⌉¬𝒯⟩⇕⟩∫⊔⟨⌉⊒⌉↕↕‖\≀⊒\⟩\⊑⌉\⊔≀∇≀{⊔⟨⌉𝒲⌉⌊⇔
ℐ⊣\ℱ≀∫⊔⌉∇¬ℐ⊣\⟩∫⊔⟨⌉√⟩≀\⌉⌉∇≀{⊔⟨⌉𝒢∇⟩⌈⇔⊔⟨⌉\⌉§⊔}⌉\⌉∇⊣⊔⟩≀\⟩\⊔⌉∇\⌉⊔↙

Obviously, there is not much the computer understands, and as a consequence, there is not a lot the computer can support the reader with. So we have to "help" the computer by providing some meaning. Conventional wisdom is that we add some semantic/functional markup. Here we pick XML without loss of generality, and characterize some fragments of text e.g. as dates.

Solution: XML markup with "meaningful" Tags

▷ **Example 16.1.19.** Let's annotate (parts of) the meaning via XML markup

```
<title>𝒲𝒲𝒲∈′′∈
𝒯⟨]]‡⊑]\⊔⟨ℐ\⊔]∇\⊣⊔⟩≀\⊣‡𝒲≀∇‡⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\]</title>
<place>𝒮⟨]∇⊣⊔≀\𝒲⊣⟩∥⟩∥⟩ℋ≀⊔]‡ℋ≀\≀‡⊓‡⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜</place>
<date>↖∞∞ℳ⊣†∈′′∈</date>
<participants>ℛ]}⟩∫⊔]∇]⌈√⊣∇⊔⟩⌋⟩√⊣\⊔∫]≀‡⟩\}{∇≀‡
𝒜⊓∫⊔∇⊣‡⟩⊣⇔𝒞⊣\⊣⌈⊣⇔𝒞⟨⟩‡]𝒟]\‡⊣∇∥⇔ℱ∇⊣\]]⇔𝒢]∇‡⊣\†⇔𝒢⟨⊣\⊣⇔ℋ≀\}𝒦≀\}⇔ℐ\⌈⟩⊣⇔
ℐ∇]‡⊣\⌈⇔ℐ⊔⊣‡†⇔𝒥⊣√⊣\⇔ℳ⊣‡⊔⊣⇔𝒩]⊒𝒵]⊣‡⊣\⌈⇔𝒯⟨]𝒩]⊔⟨]∇‡⊣\⌈∫⇔𝒩≀∇⊒⊣†⇔
𝒮⟩\}⊣√≀∇]⇔𝒮⊒⟩⊔‡]∇‡⊣\⌈⇔⊔⟨]𝒰\⟩⊔]⌈𝒦⟩\}⌈≀‡⇔⊔⟨]𝒰\⟩⊔]⌈𝒮⊔⊣⊔]∫⇔𝒱]⊔\⊣‡⇔𝒵⊣⟩∇]
</participants>
<introduction>𝒪\⊔⟨]⊔⟨ℳ⊣†ℋ≀\≀‡⊓‡⊓⊒⟩‡‡√∇≀⊑⟩⌈]⊔⟨]⌊⊣∥⌈∇≀√{⊔⟨]]‡⊑]\⊔⟨ℐ\⊔]∇↖
\⊣⊔⟩≀\⊣‡𝒲≀∇‡⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\]↙</introduction>
<program>𝒮√]⊣∥]∇∫]≀\{⟩∇‡]⌈
<speaker>𝒯⟩‡ℬ]∇\]∇∫↖ℒ]]¬𝒯⟩‡⟩∫⊔⟨]⊒]‡‡∥\≀⊒\⟩\⊑]\⊔≀∇≀{⊔⟨]𝒲]⌊</speaker>
<speaker>ℐ⊣\ℱ≀∫⊔]∇¬ℐ⊣\⟩∫⊔⟨]√≀\]]∇≀{⊔⟨]𝒢∇⟩⌈⇔⊔⟨]\§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇\]⊔<speaker>
</program>
```

But does this really help? Is conventional wisdom correct?

What can we do with this?

▷ **Example 16.1.20.** Consider the following fragments:

```
ℜ⊔⟩⊔‡]⊤𝒲𝒲𝒲∈′′∈
𝒯⟨]]‡⊑]\⊔⟨ℐ\⊔]∇\⊣⊔⟩≀\⊣‡𝒲≀∇‡⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\]ℜ∝⊔⟩⊔‡]⊤
ℜ√‡⊣⌉]⊤𝒮⟨]∇⊣⊔≀\𝒲⊣⟩∥⟩∥⟩ℋ≀⊔]‡ℋ≀\≀‡⊓‡⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜ℜ∝√‡⊣⌉]⊤
ℜ⌈⊣⊔]⊤↖∞∞ℳ⊣†∈′′∈ℜ∝⌈⊣⊔]⊤
```

Given the markup above, a machine agent can

- ▷ parse ∞∞ℳ⊣†∈′′∈ as the date May 7 11 2002 and add this to the user's calendar,
- ▷ parse 𝒮⟨]∇⊣⊔≀\𝒲⊣⟩∥⟩∥⟩ℋ≀⊔]‡ℋ≀\≀‡⊓‡⊓⇔ℋ⊣⊒⊣⟩⟩⇔𝒰𝒮𝒜 as a destination and find flights.

▷ **But:** do not be deceived by your ability to understand English!

To understand what a machine can understand we have to obfuscate the markup as well, since it does not carry any intrinsic meaning to the machine either.

What the machine sees of the XML

▷ **Example 16.1.21.** Here is what the machine sees of the XML

```
<title>𝒲𝒲𝒲∈′′∈
𝒯⟨]]‡⊑]\⊔⟨ℐ\⊔]∇\⊣⊔⟩≀\⊣‡𝒲≀∇‡⌈𝒲⟩⌈]𝒲]⌊𝒞≀\{]∇]\]</⊔⟩⊔‡]>
```

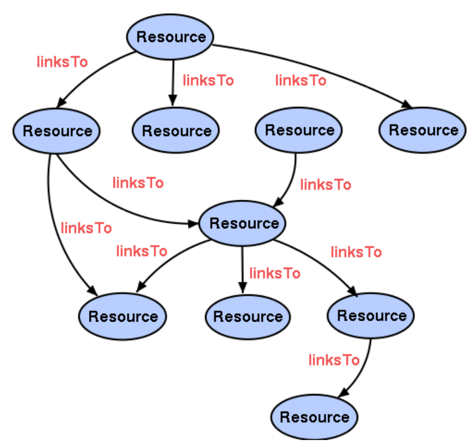
< √↕⊣|]>S⟨]∇⊣⊔≀\W⊣⟩∥⟩H≀⊔]↕H≀\≀↕⊓↕⊓⇔H⊣⊒⊣⟩⟩⇔USA</ √↕⊣|]>
<⌈⊣⊔]>↖∞∞M⊣†∈′′∈</⌈⊣⊔]>
< √⊣∇⊔⟩⌋⟩ √⊣\⊔∫>R]}⟩∫⊔]∇]⌈ √⊣∇⊔⟩⌋⟩ √⊣\⊔∫]≀↕⟩\}{∇≀↕
A⊓∫⊔∇⊣↕⟩⊣⇔C⊣\⊣⌈⊣⇔C⟨⟩↕]D]\↕⊣∇∥⇔F∇⊣\⌋]⇔G]∇↕⊣\†⇔G⟨⊣\⊣⇔H≀\}K≀\}⇔I\⌈⟩⊣⇔
I∇]↕⊣\⌈⇔I⊔⊣↕†⇔J⊣ √⊣\⇔M⊣↕⊔⊣⇔N]⊒Z]⊣↕⊣\⌈⇔T⟨]N]⊔⟨]∇↕⊣\⌈∫⇔N≀∇⊒⊣†⇔
S⟩\}⊣ √≀∇]⇔S⊒⟩⊔‡]∇↕⊣\⌈⇔⊔⟨]U\⟩⊔]⌈K⟩\}⌈≀↕⇔⊔⟨]U\⟩⊔]⌈S⊔⊣⊔]∫⇔V⟩]⊔\⊣↕⇔Z⊣⟩∇]
</ √⊣∇⊔⟩⌋⟩ √⊣\⊔∫>
<⟩\⊔∇≀⌈⊓⌋⊔⟩≀\>O\⊔⟨]⊔⟨M⊣†H≀\≀↕⊓↕⊓⊒⟩↕↕ √∇≀⊑⟩⌈]⊔⟨]⌊⊣]∥⌈∇≀ √≀{⊔⟨]]↕]⊑]\⊔⟨I\⊔]∇\⊣↖
⊔⟩≀\⊣↕W≀∇↕⌈W⟩⌈]W]⌊C≀\{]∇]\⌋]↙</⟩\⊔∇≀⌈⊓⌋⊔⟩≀\>
< √∇≀}∇⊣↕>S √]⊣∥]∇∫]≀\{⟩∇↕]⌈
<∫ √]⊣∥]∇>T⟩↕B]∇\]∇∫↖L]]¬T⟩↕⟩∫⊔⟨]⊒]↕↕∥\≀⊒\⟩\⊑]\⊔≀∇≀{⊔⟨]W]⌊</∫ √]⊣∥]∇>
<∫ √]⊣∥]∇>I⊣\F⟩∫⊔]∇¬I⊣\⟩∫⊔⟨] √⟩≀\]]∇≀{⊔⟨]G∇⟩⌈⇔⊔⟨]\§⊔}]\]∇⊣⊔⟩≀\⟩\⊔]∇\]⊔<∫ √]⊣∥]∇>
</ √∇≀}∇⊣↕>

So we have not really gained much either with the markup, we really have to give meaning to the markup as well, this is where techniques from semenatic web come into play.

To understand how we can make the web more semantic, let us first take stock of the current status of (markup on) the web. It is well-known that world-wide-web is a hypertext, where multimedia documents (text, images, videos, etc. and their fragments) are connected by hyperlinks. As we have seen, all of these are largely opaque (non-understandable), so we end up with the following situation (from the viewpoint of a machine).

The Current Web

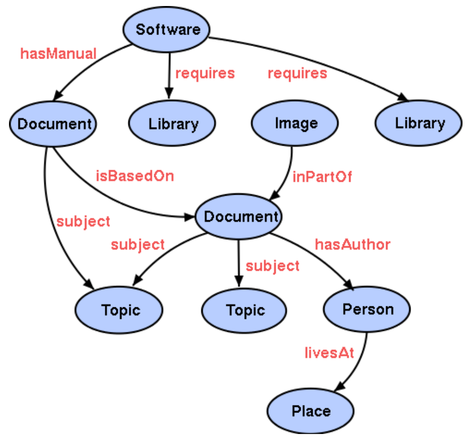
- ▷ **Resources:** identified by URIs, untyped
- ▷ **Links:** href, src, . . . limited, non-descriptive
- ▷ **User:** Exciting world - semantics of the resource, however, gleaned from content
- ▷ **Machine:** Very little information available - significance of the links only evident from the context around the anchor.



Let us now contrast this with the envisioned semantic web.

The Semantic Web

- ▷ **Resources:** Globally identified by URIs or Locally scoped (Blank), Extensible, Relational
- ▷ **Links:** Identified by URIs, Extensible, Relational
- ▷ **User:** Even more exciting world, richer user experience
- ▷ **Machine:** More processable information is available (Data Web)
- ▷ **Computers and people:** Work, learn and exchange knowledge effectively



Essentially, to make the web more machine-processable, we need to classify the resources by the concepts they represent and give the links a meaning in a way, that we can do inference with that. The ideas presented here gave rise to a set of technologies jointly called the "semantic web", which we will now summarize before we return to our logical investigations of knowledge representation techniques.

Towards a "Machine-Actionable Web"

- ▷ **Recall:** We need external agreement on meaning of annotation tags.
- ▷ **Idea:** standardize them in a community process (e.g. DIN or ISO)
- ▷ **Problem:** Inflexible, Limited number of things can be expressed
- ▷ **Better:** Use ontologies to specify meaning of annotations
  - ▷ Ontologies provide a vocabulary of terms
  - ▷ New terms can be formed by combining existing ones
  - ▷ Meaning (semantics) of such terms is formally specified
  - ▷ Can also specify relationships between terms in multiple ontologies
- ▷ Inference with annotations and ontologies (get out more than you put in!)
  - ▷ Standardize annotations in RDF [KC04] or RDFa [Her+13b] and ontologies on OWL [OWL09]
  - ▷ Harvest RDF and RDFa in to a triplestore or OWL reasoner.
  - ▷ Query that for implied knowledge (e.g. chaining multiple facts from Wikipedia)

    SPARQL: Who was US President when Barack Obama was Born?

    DBPedia: John F. Kennedy (was president in August 1961)

16.1.4 Other Knowledge Representation Approaches

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27282`.
Now that we know what semantic networks mean, let us look at a couple of other approaches that were influential for the development of knowledge representation. We will just mention them for reference here, but not cover them in any depth.

Frame Notation as Logic with Locality

- Predicate Logic: (where is the locality?)

$catch\_22 \in catch\_object$	There is an instance of catching
$catcher(catch\_22, jack\_2)$	Jack did the catching
$caught(catch\_22, ball\_5)$	He caught a certain ball

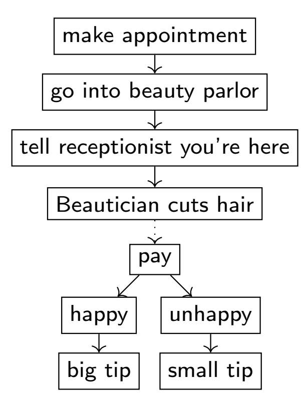
- **Definition 16.1.22.** Frames (group everything around the object)

```
(catch_object catch_22
              (catcher jack_2)
              (caught ball_5))
```

+ Once you have decided on a frame, all the information is local
+ easy to define schemes for concepts (aka. types in feature structures)
– how to determine frame, when to choose frame (log/chair)

KR involving Time (Scripts [Shank '77])

- **Idea:** Organize typical event sequences, actors and props into representation.
  - **Definition 16.1.23.** A script is a structured representation describing a stereotyped sequence of events in a particular context. Structurally, scripts are very much like frames, except the values that fill the slots must be ordered.
  - **Example 16.1.24.** getting your hair cut (at a beauty parlor)
    - props, actors as "script variables"
    - events in a (generalized) sequence
  - use script material for
    - anaphora, bridging references
    - default common ground
    - to fill in missing material into situations



Other Representation Formats (not covered)

- ▷ Procedural Representations (production systems)
- ▷ Analogical representations (interesting but not here)
- ▷ Iconic representations (interesting but very difficult to formalize)
- ▷ If you are interested, come see me off-line

16.2 Logic-Based Knowledge Representation

A Video Nugget covering this section can be found at https://fau.tv/clip/id/27297.

We now turn to knowledge representation approaches that are based on some kind of logical system. These have the advantage that we know exactly what we are doing: as they are based on symbolic representations and declaratively given inference calculi as process models, we can inspect them thoroughly and even prove facts about them.

Logic-Based Knowledge Representation

- ▷ Logic (and related formalisms) have a well-defined semantics
  - ▷ explicitly (gives more understanding than statistical/neural methods)
  - ▷ transparently (symbolic methods are monotonic)
  - ▷ systematically (we can prove theorems about our systems)
- ▷ Problems with logic-based approaches
  - ▷ Where does the world knowledge come from? (Ontology problem)
  - ▷ How to guide search induced by logical calculi (combinatorial explosion)
- ▷ **One possible answer:** description logics. (next couple of times)

But of course logic-based approaches have big drawbacks as well. The first is that we have to obtain the symbolic representations of knowledge to do anything – a non-trivial challenge, since most knowledge does not exist in this form in the wild, to obtain it, some agent has to experience the word, pass it through its cognitive apparatus, conceptualize the phenomena involved, systematize them sufficiently to form symbols, and then represent those in the respective formalism at hand.

The second drawback is that the process models induced by logic-based approaches (inference with calculi) are quite intractable. We will see that all inferences can be played back to satisfiability tests in the underlying logical system, which are exponential at best, and undecidable or even incomplete at worst.

Therefore a major thrust in logic-based knowledge representation is to investigate logical systems that are expressive enough to be able to represent most knowledge, but still have a decidable – and maybe even tractable in practice – satisfiability problem. Such logics are called "description logics". We will study the basics of such logical systems and their inference procedures in the following.

16.2.1 Propositional Logic as a Set Description Language

Before we look at "real" description logics in ??, we will make a "dry run" with a logic we already understand: propositional logic, which we will re-interpret the system as a set description language by giving a new, non-standard semantics. This allows us to already preview most of the inference procedures and knowledge services of knowledge representation systems in the next subsection.
To establish propositional logic as a set description language, we use a different interpretation than usual. We interpret propositional variables as names of sets and the connectives as set operations, which is why we give them a different – more suggestive – syntax.

Propositional Logic as Set Description Language

- **Idea:** Use propositional logic as a set description language: (variant syntax/semantics)
- **Definition 16.2.1.** Let $\mathrm{PL}^0_{\mathrm{DL}}$ be given by the following grammar for the $\mathrm{PL}^0_{\mathrm{DL}}$ concepts. (formulae)

$$\mathcal{L}::=C \mid \top \mid \bot \mid \overline{\mathcal{L}} \mid \mathcal{L}\sqcap\mathcal{L} \mid \mathcal{L}\sqcup\mathcal{L} \mid \mathcal{L}\sqsubseteq\mathcal{L} \mid \mathcal{L}\equiv\mathcal{L}$$

  i.e. $\mathrm{PL}^0_{\mathrm{DL}}$ formed from
  - atomic ($\widehat{=}$ propositional variables)
  - concept intersection ($\sqcap$) ($\widehat{=}$ conjunction $\wedge$)
  - concept complement ($\overline{\cdot}$) ($\widehat{=}$ negation $\neg$)
  - concept union ($\sqcup$), subsumption ($\sqsubseteq$), and equality ($\equiv$) defined from these. ($\widehat{=}$ $\vee$, $\Rightarrow$, and $\Leftrightarrow$)
- **Definition 16.2.2 (Formal Semantics).**
  Let $\mathcal{D}$ be a given set (called the domain) and $\varphi\colon \mathcal{V}_0\to\mathcal{P}(\mathcal{D})$, then we define
  - $[\![P]\!]:=\varphi(P)$, (remember $\varphi(P)\subseteq\mathcal{D}$).
  - $[\![\mathbf{A}\sqcap\mathbf{B}]\!]:=[\![\mathbf{A}]\!]\cap[\![\mathbf{B}]\!]$ and $[\![\overline{\mathbf{A}}]\!]:=\mathcal{D}\backslash[\![\mathbf{A}]\!]$ ...
- **Note:** $\langle \mathrm{PL}^0_{\mathrm{DL}}, \mathcal{S}, [\![\cdot]\!]\rangle$, where $\mathcal{S}$ is the class of possible domains forms a logical system.

The main use of the set-theoretic semantics for PL^0 is that we can use it to give meaning to concept axioms, which we use to describe the "world".

Concept Axioms

- **Observation:** Set-theoretic semantics of 'true' and 'false'($\top:=\varphi\sqcup\overline{\varphi}$ $\quad\bot:=\varphi\sqcap\overline{\varphi}$)

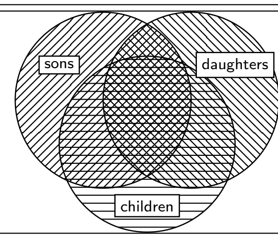
$$[\![\top]\!]=[\![p]\!]\cup[\![\overline{p}]\!]=[\![p]\!]\cup(\mathcal{D}\backslash[\![p]\!])=\mathcal{D} \qquad \text{Analogously: } [\![\bot]\!]=\emptyset$$

- **Idea:** Use logical axioms to describe the world (Axioms restrict the class of admissible domain structures)

▷ **Definition 16.2.3.** A concept axiom is a $\mathrm{PL}^0_{\mathrm{DL}}$ formula $\mathbf{A}$ that is assumed to be true in the world.

▷ **Definition 16.2.4 (Set-Theoretic Semantics of Axioms).** $\mathbf{A}$ is true in domain $\mathcal{D}$ iff $[\![\mathbf{A}]\!] = \mathcal{D}$.

▷ **Example 16.2.5.** A world with three concepts and no concept axioms

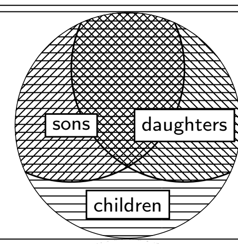
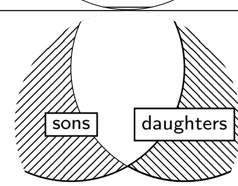
concepts	Set Semantics
child daughter son	

Concept axioms are used to restrict the set of admissible domains to the intended ones. In our situation, we require them to be true – as usual – which here means that they denote the whole domain $\mathcal{D}$.

Let us fortify our intuition about concept axioms with a simple example about the sibling relation. We give four concept axioms and study their effect on the admissible models by looking at the respective Venn diagrams. In the end we see that in all admissible models, the denotations of the concepts son and daughter are disjoint, and child is the union of the two – just as intended.

Effects of Axioms to Siblings

▷ **Example 16.2.6.** We can use concept axioms to describe the world from Example 16.2.5.

Axioms	Semantics
$\text{son} \sqsubseteq \text{child}$ iff $\overline{[\![\text{son}]\!]} \cup [\![\text{child}]\!] = \mathcal{D}$ iff $[\![\text{son}]\!] \subseteq [\![\text{child}]\!]$ $\text{daughter} \sqsubseteq \text{child}$ iff $\overline{[\![\text{daughter}]\!]} \cup [\![\text{child}]\!] = \mathcal{D}$ iff $[\![\text{daughter}]\!] \subseteq [\![\text{child}]\!]$	
$\overline{\text{son} \sqcap \text{daughter}}$ $\text{child} \sqsubseteq \text{son} \sqcup \text{daughter}$	

The set-theoretic semantics introduced above is compatible with the regular semantics of propositional logic, therefore we have the same propositional identities. Their validity can be established directly from the settings in Definition 16.2.2.

Propositional Identities

Name	for $\sqcap$	for $\sqcup$
Idenpot.	$\varphi \sqcap \varphi = \varphi$	$\varphi \sqcup \varphi = \varphi$
Identity	$\varphi \sqcap \top = \varphi$	$\varphi \sqcup \bot = \varphi$
Absorpt.	$\varphi \sqcup \top = \top$	$\varphi \sqcap \bot = \bot$
Commut.	$\varphi \sqcap \psi = \psi \sqcap \varphi$	$\varphi \sqcup \psi = \psi \sqcup \varphi$
Assoc.	$\varphi \sqcap \psi \sqcap \theta = \varphi \sqcap \psi \sqcap \theta$	$\varphi \sqcup \psi \sqcup \theta = \varphi \sqcup \psi \sqcup \theta$
Distrib.	$\varphi \sqcap (\psi \sqcup \theta) = \varphi \sqcap \psi \sqcup \varphi \sqcap \theta$	$\varphi \sqcup \psi \sqcap \theta = (\varphi \sqcup \psi) \sqcap (\varphi \sqcup \theta)$
Absorpt.	$\varphi \sqcap (\varphi \sqcup \theta) = \varphi$	$\varphi \sqcup \varphi \sqcap \theta = \varphi \sqcap \theta$
Morgan	$\overline{\varphi \sqcap \psi} = \overline{\varphi} \sqcup \overline{\psi}$	$\overline{\varphi \sqcup \psi} = \overline{\varphi} \sqcap \overline{\psi}$
dneg	$\overline{\overline{\varphi}} = \varphi$	

Michael Kohlhase: Artificial Intelligence 1 514 2023-09-20

There is another way we can approach the set description interpretation of propositional logic: by translation into a logic that can express knowledge about sets – first-order logic.

Set-Theoretic Semantics and Predicate Logic

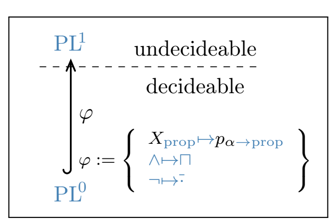
▷ **Definition 16.2.7.** Translation into PL^1 (borrow semantics from that)

- ▷ recursively add argument variable x
- ▷ change back $\sqcap, \sqcup, \sqsubseteq, \equiv$ to $\wedge, \vee, \Rightarrow, \Leftrightarrow$
- ▷ universal closure for x at formula level.

Definition	Comment
$\overline{p}^{fo(x)} := p(x)$	
$\overline{\overline{\mathbf{A}}}^{fo(x)} := \neg\overline{\mathbf{A}}^{fo(x)}$	
$\overline{\mathbf{A} \sqcap \mathbf{B}}^{fo(x)} := \overline{\mathbf{A}}^{fo(x)} \wedge \overline{\mathbf{B}}^{fo(x)}$	$\wedge$ vs. $\sqcap$
$\overline{\mathbf{A} \sqcup \mathbf{B}}^{fo(x)} := \overline{\mathbf{A}}^{fo(x)} \vee \overline{\mathbf{B}}^{fo(x)}$	$\vee$ vs. $\sqcup$
$\overline{\mathbf{A} \sqsubseteq \mathbf{B}}^{fo(x)} := \overline{\mathbf{A}}^{fo(x)} \Rightarrow \overline{\mathbf{B}}^{fo(x)}$	$\Rightarrow$ vs. $\sqsubseteq$
$\overline{\mathbf{A} = \mathbf{B}}^{fo(x)} := \overline{\mathbf{A}}^{fo(x)} \Leftrightarrow \overline{\mathbf{B}}^{fo(x)}$	$\Leftrightarrow$ vs. $=$
$\overline{\mathbf{A}}^{fo} := (\forall x.\overline{\mathbf{A}}^{fo(x)})$	for formulae

Michael Kohlhase: Artificial Intelligence 1 515 2023-09-20

Normally, we embed PL^0 into PL^1 by mapping propositional variables to atomic predicates and the connectives to themselves. The purpose of this embedding is to "talk about truth/falsity of assertions". For "talking about sets" we use a non-standard embedding: propositional variables in PL^0 are mapped to first-order predicates, and the connectives to corresponding set operations. This uses the convention that a set S is represented by a unary predicate p_S (its characteristic predicate), and set membership $a \in S$ as $p_S(a)$.



Translation Examples

▷ **Example 16.2.8.** We translate the concept axioms from Example 16.2.6 to fortify our intuition:

$$\overline{\mathsf{son} \sqsubseteq \mathsf{child}}^{fo} = \forall x.\mathsf{son}(x) \Rightarrow \mathsf{child}(x)$$
$$\overline{\mathsf{daughter} \sqsubseteq \mathsf{child}}^{fo} = \forall x.\mathsf{daughter}(x) \Rightarrow \mathsf{child}(x)$$
$$\overline{\overline{\mathsf{son} \sqcap \mathsf{daughter}}}^{fo} = \forall x.\mathsf{son}(x) \wedge \overline{\mathsf{daughter}(x)}$$
$$\overline{\mathsf{child} \sqsubseteq \mathsf{son} \sqcup \mathsf{daughter}}^{fo} = \forall x.\mathsf{child}(x) \Rightarrow (\mathsf{son}(x) \vee \mathsf{daughter}(x))$$

▷ What are the advantages of translation to PL^1?
- ▷ **theoretically**: A better understanding of the semantics
- ▷ **computationally**: Description Logic Framework, but NOTHING for PL^0
  - ▷ we can follow this pattern for richer description logics.
  - ▷ many tests are decidable for PL^0, but not for PL^1. (Description Logics?)

16.2.2 Ontologies and Description Logics

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27298`. We have seen how sets of concept axioms can be used to describe the "world" by restricting the set of admissible models. We want to call such concept descriptions "ontologies" – formal descriptions of (classes of) objects and their relations.

Ontologies aka. "World Descriptions"

▷ **Definition 16.2.9 (Classical).** An ontology is a representation of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.

▷ **Remark:** Definition 16.2.9 is very general, and depends on what we mean by "representation", "entities", "types", and "interrelationships".

This may be a feature, and not a bug, since we can use the same intuitions across a variety of representations.

▷ **Definition 16.2.10.** An ontology consists of a logical system $\langle \mathcal{L}, \mathcal{K}, \models \rangle$ and concept axioms (expressed in $\mathcal{L}$) about
- ▷ individuals: concrete instances of object in the domain,
- ▷ concepts: classes of individuals that share properties and aspects, and
- ▷ relations: ways in which concept and individuals can be related to one another.

▷ **Example 16.2.11.** Semantic networks are ontologies. (relatively informal)

▷ **Example 16.2.12.** $\mathrm{PL}^0_{\mathrm{DL}}$ is an ontology format. (formal, but relatively weak)

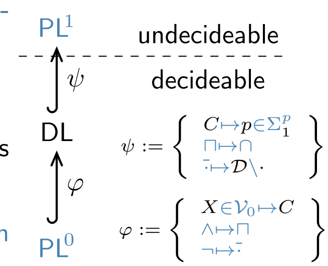
▷ **Example 16.2.13.** PL^1 is an ontology format as well. (formal, expressive)

As we will see, the situation for $\mathrm{PL}^0_{\mathrm{DL}}$ is typical for formal ontologies (even though it only offers concepts), so we state the general description logic paradigm for ontologies. The important idea is that having a formal system as an ontology format allows us to capture, study, and implement ontological inference.

The Description Logic Paradigm

- **Idea:** Build a whole family of logics for describing sets and their relations. (tailor their expressivity and computational properties)
- **Definition 16.2.14.** A description logic is a formal system for talking about collections of objects and their relations that is at least as expressive as PL^0 with set-theoretic semantics and offers individuals and relations.

  A description logic has the following four components:
  - a formal language $\mathcal{L}$ with logical constants $\sqcap$, $\bar{\cdot}$, $\sqcup$, $\sqsubseteq$, and $\equiv$,
  - a set-theoretic semantics $\langle \mathcal{D}, [\![\cdot]\!] \rangle$,
  - a translation into first-order logic that is compatible with $\langle \mathcal{D}, [\![\cdot]\!] \rangle$, and
  - a calculus for $\mathcal{L}$ that induces a decision procedure for $\mathcal{L}$-satisfiability.



- **Definition 16.2.15.** Given a description logic $\mathcal{D}$, a $\mathcal{D}$ ontology consists of
  - a terminology (or TBox): concepts and roles and a set of concept axioms that describe them, and
  - sassertion (or ABox): a set of individuals and statements about concept membership and role relationships for them.

For convenience we add concept definitions as a mechanism for defining new concepts from old ones. The so-defined concepts inherit the properties from the concepts they are defined from.

TBoxes in Description Logics

- Let $\mathcal{D}$ be a description logic with concepts $\mathcal{C}$.
- **Definition 16.2.16.** A concept definition is a pair $c = \mathbf{C}$, where c is a new concept name and $\mathbf{C} \in \mathcal{C}$ is a $\mathcal{D}$-formula.
- **Definition 16.2.17.** A concept definition $c = C$ is called recursive, iff c occurs in C.
- **Example 16.2.18.** We can define mother$=$woman $\sqcap$ has_child.
- **Definition 16.2.19.** An TBox is a finite set of concept definitions and concept axioms. It is called acyclic, iff it does not contain recursive definitions.

- **Definition 16.2.20.** A formula $\mathbf{A}$ is called normalized wrt. an TBox $\mathcal{T}$, iff it does not contain concepts defined in $\mathcal{T}$. (convenient)
- **Definition 16.2.21 (Algorithm).** (for arbitrary DLs)
  Input: A formula $\mathbf{A}$ and a TBox $\mathcal{T}$.
  - While [$\mathbf{A}$ contains concept c and $\mathcal{T}$ a concept definition $c{=}\mathbf{C}$]
    - substitute c by $\mathbf{C}$ in $\mathbf{A}$.
- **Lemma 16.2.22.** *This algorithm terminates for acyclic TBoxes, but results can be exponentially large.*

Michael Kohlhase: Artificial Intelligence 1 519 2023-09-20

As $\mathrm{PL}^0_{\mathrm{DL}}$ does not offer any guidance on this, we will leave the discussion of ABoxes to subsection 16.3.3 when we have introduced our first proper description logic $\mathcal{ALC}$.

16.2.3 Description Logics and Inference

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27299`.

Now that we have established the description logic paradigm, we will have a look at the inference services that can be offered on this basis.
Before we go into details of particular description logics, we must ask ourselves what kind of inference support we would want for building systems that support knowledge workers in building, maintaining and using ontologies. An example of such a system is the Protégé system [Pro], which can serve for guiding our intuition.

Kinds of Inference in Description Logics

- **Definition 16.2.23.** Ontology systems employ three main reasoning services:
  - Consistency test: is a concept definition satisfiable?
  - Subsumption test: does a concept subsume another?
  - Instance test: is an individual an example of a concept?
- **Problem:** decidability, complexity, algorithm

Michael Kohlhase: Artificial Intelligence 1 520 2023-09-20

We will now through these inference-based tests separately.
The consistency test checks for concepts that do not/cannot have instances. We want to avoid such concepts in our ontologies, since they clutter the namespace and do not contribute any meaningful contribution.

Consistency Test

- **Example 16.2.24 (T-Box).**

man	$=$	person $\sqcap$ has_Y	person with y-chromosome
woman	$=$	person $\sqcap$ $\overline{\text{has\_Y}}$	person without y-chromosome
hermaphrodite	$=$	man $\sqcap$ woman	man and woman

- This specification is inconsistent, i.e. $[\![\text{hermaphrodite}]\!] = \emptyset$ for all $\mathcal{D}$, φ.
- **Algorithm:** Propositional satisfiability test (NP complete)
  we know how to do this, e.g. tableau, resolution.

Even though consistency in our example seems trivial, large ontologies can make machine support necessary. This is even more true for ontologies that change over time. Say that an ontology initially has the concept definitions woman=person⊓long_hair and man=person⊓bearded, and then is modernized to a more biologically correct state. In the initial version the concept hermaphrodite is consistent, but becomes inconsistent after the renovation; the authors of the renovation should be made aware of this by the system.
The subsumption test determines whether the sets denoted by two concepts are in a subset relation. The main justification for this is that humans tend to be aware of concept subsumption, and tend to think in taxonomytaxonomic hierarchies. To cater to this, the subsumption test is useful.

Subsumption Test

- **Example 16.2.25.** In this case trivial

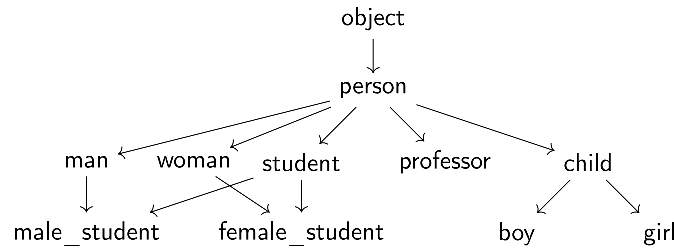
axiom	entailed subsumption relation
man = person ⊓ has_Y	man ⊑ person
woman = person ⊓ $\overline{\text{has\_Y}}$	woman ⊑ person

- **Reduction to consistency test:** (need to implement only one)
  $\mathit{Axioms} \Rightarrow (\mathbf{A} \Rightarrow \mathbf{B})$ is valid iff $\mathit{Axioms} \wedge \mathbf{A} \wedge \neg\mathbf{B}$ is consistentin.
- **Definition 16.2.26.** $\mathbf{A}$ subsumes $\mathbf{B}$ (modulo an axiom set $\mathcal{A}$)
  iff $[\![\mathbf{B}]\!] \subseteq [\![\mathbf{A}]\!]$ for all interpretations $\mathcal{D}$, that satisfy $\mathcal{A}$
  iff $\mathcal{A} \Rightarrow \mathbf{B} \Rightarrow \mathbf{A}$ is valid.
- **In our example:** person subsumes woman and man

The good news is that we can reduce the subsumption test to the consistency test, so we can re-use our existing implementation.
The main user-visible service of the subsumption test is to compute the actual taxonomy induced by an ontology.

Classification

- The subsumption relation among all concepts (subsumption graph)
- Visualization of the subsumption graph for inspection (plausibility)
- **Definition 16.2.27.** Classification is the computation of the subsumption graph.
- **Example 16.2.28.** (not always so trivial)



If we take stock of what we have developed so far, then we can see $\mathrm{PL}^0_{\mathrm{DL}}$ as a rational reconstruction of semantic networks restricted to the "isa" relation. We relegate the "instance" relation to subsection 16.3.3.

This reconstruction can now be used as a basis on which we can extend the expressivity and inference procedures without running into problems.

16.3 A simple Description Logic: ALC

In this section, we instantiate the description-logic paradigm further with the prototypical logic $\mathcal{ALC}$, which we will introduce now.

16.3.1 Basic ALC: Concepts, Roles, and Quantification

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27300`.

In this subsection, we instantiate the description-logic paradigm with the prototypical logic $\mathcal{ALC}$, which we will develop now.

Motivation for $\mathcal{ALC}$ (Prototype Description Logic)

- ▷ Propositional logic (PL^0) is not expressive enough
- ▷ **Example 16.3.1.** "mothers are women that have a child"
- ▷ **Reason:** there are no quantifiers in PL^0 (existential ($\exists$) and universal ($\forall$))
- ▷ **Idea:** Use first-order predicate logic (PL^1)

$$\forall x.mother(x) \Leftrightarrow (woman(x) \wedge (\exists y.has\_child(x,y)))$$

- ▷ **Problem:** Complex algorithms, non termination (PL^1 is too expressive)
- ▷ **Idea:** Try to travel the middle ground
  More expressive than PL^0 (quantifiers) but weaker than PL^1. (still tractable)
- ▷ **Technique:** Allow only "restricted quantification", where quantified variables only range over values that can be reached via a binary relation like has_child.

$\mathcal{ALC}$ extends the concept operators of $\mathrm{PL}^0_{\mathrm{DL}}$ with binary relations (called "roles" in $\mathcal{ALC}$). This gives $\mathcal{ALC}$ the expressive power we had for the basic semantic networks from **??**.

Syntax of $\mathcal{ALC}$

▷ **Definition 16.3.2 (Concepts).** (aka. "predicates" in PL^1 or "propositional variables" in $\mathrm{PL}^0_{\mathrm{DL}}$) concepts in DLs name classes of objects like in OOP.

▷ **Definition 16.3.3 (Special concepts).** The top concept $\top$ (for "true" or "all") and the bottom concept $\bot$ (for "false" or "none").

▷ **Example 16.3.4.** person, woman, man, mother, professor, student, car, BMW, computer, computer program, heart attack risk, furniture, table, leg of a chair, ...

▷ **Definition 16.3.5.** Roles name binary relations (like in PL^1)

▷ **Example 16.3.6.** has_child, has_son, has_daughter, loves, hates, gives_course, executes_computer_program, has_leg_of_table, has_wheel, has_motor, ...

$\mathcal{ALC}$ restricts the quantifications to range all individuals reachable as role successors. The distinction between universal and existential quantifiers clarifies an implicit ambiguity in semantic networks.

Syntax of $\mathcal{ALC}$: Formulae $F_{\mathcal{ALC}}$

▷ **Definition 16.3.7 (Grammar).** $F_{\mathcal{ALC}} ::= C \mid \top \mid \bot \mid \overline{F_{\mathcal{ALC}}} \mid F_{\mathcal{ALC}} \sqcap F_{\mathcal{ALC}} \mid F_{\mathcal{ALC}} \sqcup F_{\mathcal{ALC}} \mid \exists \mathrm{R}.F_{\mathcal{ALC}} \mid \forall \mathrm{R}.F_{\mathcal{ALC}}$

▷ **Example 16.3.8.**

- ▷ person ⊓ ∃has_child.student (parents of students) (The set of persons that have a child which is a student)
- ▷ person ⊓ ∃has_child.∃has_child.student (grandparents of students)
- ▷ person ⊓ ∃has_child.∃has_child.(student ⊔ teacher) (grandparents of students or teachers)
- ▷ person ⊓ ∀has_child.student (parents whose children are all students)
- ▷ person ⊓ ∀haschild.∃has_child.student (grandparents, whose children all have at least one child that is a student)

More $\mathcal{ALC}$ Examples

▷ **Example 16.3.9.** car ⊓ ∃has_part.∃made_in.$\overline{\text{EU}}$ (cars that have at least one part that has not been made in the EU)

▷ **Example 16.3.10.** student ⊓ ∀audits_course.graduatelevelcourse (students, that only audit graduate level courses)

▷ **Example 16.3.11.** house ⊓ ∀has_parking.off_street (houses with off-street parking)

▷ **Note:** $p \sqsubseteq q$ can still be used as an abbreviation for $\overline{p} \sqcup q$.

▷ **Example 16.3.12.** student $\sqcap$ $\forall$audits_course.($\exists$hastutorial.$\top \sqsubseteq \forall$has_TA.woman) (students that only audit courses that either have no tutorial or tutorials that are TAed by women)

As before we allow concept definitions so that we can express new concepts from old ones, and obtain more concise descriptions.

$\mathcal{ALC}$ Concept Definitions

▷ **Idea:** Define new concepts from known ones.

▷ **Definition 16.3.13.** A concept definition is a pair consisting of a new concept name (the definiendum) and an $\mathcal{ALC}$ formula (the definiens). Concept names are not definienda are called primitive.

▷ We extend the $\mathcal{ALC}$ grammar from Definition 16.3.7 by the production $CD_{\mathcal{ALC}} ::= C = F_{\mathcal{ALC}}$.

▷ **Example 16.3.14.**

Definition	rec?
man = person $\sqcap$ $\exists$has_chrom.Y_chrom	-
woman = person $\sqcap$ $\forall$has_chrom.$\overline{\text{Y\_chrom}}$	-
mother = woman $\sqcap$ $\exists$has_child.person	-
father = man $\sqcap$ $\exists$has_child.person	-
grandparent = person $\sqcap$ $\exists$has_child.(mother $\sqcup$ father)	-
german = person $\sqcap$ $\exists$has_parents.german	+
number_list = empty_list $\sqcup$ $\exists$is_first.number $\sqcap$ $\exists$is_rest.number_list	+

As before, we can normalize a TBox by definition expansion if it is acyclic. With the introduction of roles and quantification, concept definitions in $\mathcal{ALC}$ have a more "interesting" way to be cyclic as Observation 16.3.19 shows.

TBox Normalization in $\mathcal{ALC}$

▷ **Definition 16.3.15.** We call an $\mathcal{ALC}$ formula φ normalized wrt. a set of concept definitions, iff all concept names occurring in φ are primitive.

▷ **Definition 16.3.16.** Given a set $\mathcal{D}$ of concept definitions, normalization is the process of replacing in an $\mathcal{ALC}$ formula φ all occurrences of definienda in $\mathcal{D}$ with their definientia.

▷ **Example 16.3.17 (Normalizing grandparent).**

grandparent
$\mapsto$ person $\sqcap$ $\exists$has_child.(mother $\sqcup$ father)
$\mapsto$ person $\sqcap$ $\exists$has_child.(woman $\sqcap$ $\exists$has_child.person $\sqcap$ man $\sqcap$ $\exists$has_child.person)
$\mapsto$ person $\sqcap$ $\exists$has_child.(person $\sqcap$ $\exists$has_chrom.Y_chrom $\sqcap$ $\exists$has_child.person $\sqcap$ person $\sqcap$ $\exists$has_chrom.Y_chrom $\sqcap$ $\exists$has_child.person)

▷ **Observation 16.3.18.** *Normalization results can be exponential.* *(contain redundancies)*

▷ **Observation 16.3.19.** *Normalization need not terminate on cyclic TBoxes.*

▷ **Example 16.3.20.**

$$\begin{array}{rcl} \mathsf{german} & \mapsto & \mathsf{person} \sqcap \exists \mathsf{has\_parents}.\mathsf{german} \\ & \mapsto & \mathsf{person} \sqcap \exists \mathsf{has\_parents}.(\mathsf{person} \sqcap \exists \mathsf{has\_parents}.\mathsf{german}) \\ & \mapsto & \ldots \end{array}$$

Now that we have motivated and fixed the syntax of $\mathcal{ALC}$, we will give it a formal semantics. The semantics of $\mathcal{ALC}$ is an extension of the set-theoretic semantics for PL^0, thus the interpretation $[\![\cdot]\!]$ assigns subsets of the domain to concepts and binary relations over the domain to roles.

Semantics of $\mathcal{ALC}$

▷ $\mathcal{ALC}$ semantics is an extension of the set-semantics of propositional logic.

▷ **Definition 16.3.21.** A model for $\mathcal{ALC}$ is a pair $\langle \mathcal{D}, [\![\cdot]\!] \rangle$, where $\mathcal{D}$ is a non-empty set called the domain and $[\![\cdot]\!]$ a mapping called the interpretation, such that

Op.	formula semantics
	$[\![c]\!] \subseteq \mathcal{D} = [\![\top]\!]$ $\quad [\![\bot]\!] = \emptyset$ $\quad [\![r]\!] \subseteq \mathcal{D} \times \mathcal{D}$
$\overline{\cdot}$	$[\![\overline{\varphi}]\!] = \overline{[\![\varphi]\!]} = \mathcal{D} \backslash [\![\varphi]\!]$
$\sqcap$	$[\![\varphi \sqcap \psi]\!] = [\![\varphi]\!] \cap [\![\psi]\!]$
$\sqcup$	$[\![\varphi \sqcup \psi]\!] = [\![\varphi]\!] \cup [\![\psi]\!]$
$\exists \mathsf{R}.$	$[\![\exists \mathsf{R}.\varphi]\!] = \{x \in \mathcal{D} \mid \exists y.\langle x, y \rangle \in [\![\mathsf{R}]\!] \text{ and } y \in [\![\varphi]\!]\}$
$\forall \mathsf{R}.$	$[\![\forall \mathsf{R}.\varphi]\!] = \{x \in \mathcal{D} \mid \forall y.\text{if } \langle x, y \rangle \in [\![\mathsf{R}]\!] \text{ then } y \in [\![\varphi]\!]\}$

▷ Alternatively we can define the semantics of $\mathcal{ALC}$ by translation into PL^1.

▷ **Definition 16.3.22.** The translation of $\mathcal{ALC}$ into PL^1 extends the one from Definition 16.2.7 by the following quantifier rules:

$$\overline{\forall \mathsf{R}.\varphi}^{fo(x)} := (\forall y.\mathsf{R}(x, y) \Rightarrow \overline{\varphi}^{fo(y)}) \qquad \overline{\exists \mathsf{R}.\varphi}^{fo(x)} := (\exists y.\mathsf{R}(x, y) \wedge \overline{\varphi}^{fo(y)})$$

▷ **Observation 16.3.23.** *The set-theoretic semantics from Definition 16.3.21 and the "semantics-by-translation" from Definition 16.3.22 induce the same notion of satisfiability.*

We can now use the $\mathcal{ALC}$ identities above to establish a useful normal form for $\mathcal{ALC}$. This will play a role in the inference procedures we study next.
The following identitieswill be useful later on. They can be proven directly with the settings from Definition 16.3.21; we carry this out for one of them below.

$\mathcal{ALC}$ Identities

▷

1	$\overline{\exists \mathsf{R}.\varphi} = \forall \mathsf{R}.\overline{\varphi}$	3	$\overline{\forall \mathsf{R}.\varphi} = \exists \mathsf{R}.\overline{\varphi}$
2	$\forall \mathsf{R}.(\varphi \sqcap \psi) = \forall \mathsf{R}.\varphi \sqcap \forall \mathsf{R}.\psi$	4	$\exists \mathsf{R}.(\varphi \sqcup \psi) = \exists \mathsf{R}.\varphi \sqcup \exists \mathsf{R}.\psi$

▷ Proof of 1

$$\begin{aligned}
[\![\overline{\exists \mathsf{R}.\varphi}]\!] = \mathcal{D} \backslash [\![\exists \mathsf{R}.\varphi]\!] &= \mathcal{D} \backslash \{x \in \mathcal{D} \mid \exists y.(\langle x,y\rangle \in [\![\mathsf{R}]\!]) \text{ and } (y \in [\![\varphi]\!])\} \\
&= \{x \in \mathcal{D} \mid \text{not } \exists y.(\langle x,y\rangle \in [\![\mathsf{R}]\!]) \text{ and } (y \in [\![\varphi]\!])\} \\
&= \{x \in \mathcal{D} \mid \forall y.\text{if } (\langle x,y\rangle \in [\![\mathsf{R}]\!]) \text{ then } (y \notin [\![\varphi]\!])\} \\
&= \{x \in \mathcal{D} \mid \forall y.\text{if } (\langle x,y\rangle \in [\![\mathsf{R}]\!]) \text{ then } (y \in (\mathcal{D} \backslash [\![\varphi]\!]))\} \\
&= \{x \in \mathcal{D} \mid \forall y.\text{if } (\langle x,y\rangle \in [\![\mathsf{R}]\!]) \text{ then } (y \in [\![\overline{\varphi}]\!])\} \\
&= [\![\forall \mathsf{R}.\overline{\varphi}]\!]
\end{aligned}$$

The form of the identities (interchanging quantification with connectives) is reminiscient of identities in PL^1; this is no coincidence as the "semantics by translation" of Definition 16.3.22 shows.

Negation Normal Form

▷ **Definition 16.3.24 (NNF).** An $\mathcal{ALC}$ formula is in negation normal form (NNF), iff $\overline{\cdot}$ is only applied to concept names.

▷ Use the $\mathcal{ALC}$ identities as rules to compute it. (in linear time)

example	by rule
$\overline{\exists \mathsf{R}.(\forall \mathsf{S}.e \sqcap \overline{\forall \mathsf{S}.d})}$	
$\mapsto \forall \mathsf{R}.\overline{\forall \mathsf{S}.e \sqcap \overline{\forall \mathsf{S}.d}}$	$\overline{\exists \mathsf{R}.\varphi} \mapsto \forall \mathsf{R}.\overline{\varphi}$
$\mapsto \forall \mathsf{R}.(\overline{\forall \mathsf{S}.e} \sqcup \overline{\overline{\forall \mathsf{S}.d}})$	$\overline{\varphi \sqcap \psi} \mapsto \overline{\varphi} \sqcup \overline{\psi}$
$\mapsto \forall \mathsf{R}.(\exists \mathsf{S}.\overline{e} \sqcup \overline{\overline{\forall \mathsf{S}.d}})$	$\overline{\forall \mathsf{R}.\varphi} \mapsto \exists \mathsf{R}.\overline{\varphi}$
$\mapsto \forall \mathsf{R}.(\exists \mathsf{S}.\overline{e} \sqcup \forall \mathsf{S}.d)$	$\overline{\overline{\varphi}} \mapsto \varphi$

Finally, we extend $\mathcal{ALC}$ with an ABox component. This mainly means that we define two new assertions in $\mathcal{ALC}$ and specify their semantics and PL^1 translation.

$\mathcal{ALC}$ with Assertions about Individuals

▷ **Definition 16.3.25.** We define the assertions for $\mathcal{ALC}$

- ▷ $a{:}\varphi$ (a is a φ)
- ▷ $a\ \mathsf{R}\ b$ (a stands in relation R to b)

assertions make up the ABox in $\mathcal{ALC}$.

▷ **Definition 16.3.26.** Let $\langle \mathcal{D}, [\![\cdot]\!] \rangle$ be a model for $\mathcal{ALC}$, then we define

- ▷ $[\![a{:}\varphi]\!] = \mathsf{T}$, iff $[\![a]\!] \in [\![\varphi]\!]$, and

- ▷ $[\![a\ \mathsf{R}\ b]\!] = \mathsf{T}$, iff $([\![a]\!], [\![b]\!]) \in [\![\mathsf{R}]\!]$.

▷ **Definition 16.3.27.** We extend the PL^1 translation of $\mathcal{ALC}$ to $\mathcal{ALC}$ assertions:

- ▷ $\overline{a{:}\varphi}^{fo} := \overline{\varphi}^{fo(a)}$, and
- ▷ $\overline{a\ \mathsf{R}\ b}^{fo} := \mathsf{R}(a, b)$.

If we take stock of what we have developed so far, then we see that $\mathcal{ALC}$ as a rational reconstruction of semantic networks restricted to the "isa" and "instance" relations – which are the only ones that can really be given a denotational and operational semantics.

16.3.2 Inference for ALC

Video Nuggets covering this subsection can be found at `https://fau.tv/clip/id/27301` and `https://fau.tv/clip/id/27302`.

In this subsection we make good on the motivation from **??** that description logics enjoy tractable inference procedures: We present a tableau calculus for $\mathcal{ALC}$, show that it is a decision procedure, and study its complexity.

$\mathcal{T}_{\mathcal{ALC}}$: A Tableau-Calculus for $\mathcal{ALC}$

▷ **Recap Tableaux:** A tableau calculus develops an initial tableau in a tree-formed scheme using tableau extension rules.

A saturated tableau (no rules applicable) constitutes a refutation, if all branches are closed (end in $\perp$).

▷ **Definition 16.3.28.** The $\mathcal{ALC}$ tableau calculus $\mathcal{T}_{\mathcal{ALC}}$ acts on assertions

- ▷ $x{:}\varphi$ (x inhabits concept φ)
- ▷ $x\ \mathsf{R}\ y$ (x and y are in relation R)

where φ is a normalized $\mathcal{ALC}$ concept in negation normal form with the following rules:

$$\frac{\begin{array}{c}x{:}c\\ x{:}\overline{c}\end{array}}{\perp}\mathcal{T}_{\perp} \qquad \frac{x{:}\varphi \sqcap \psi}{\begin{array}{c}x{:}\varphi\\ x{:}\psi\end{array}}\mathcal{T}_{\sqcap} \qquad \frac{x{:}\varphi \sqcup \psi}{x{:}\varphi \mid x{:}\psi}\mathcal{T}_{\sqcup} \qquad \frac{\begin{array}{c}x{:}\forall \mathsf{R}.\varphi\\ x\ \mathsf{R}\ y\end{array}}{y{:}\varphi}\mathcal{T}_{\forall} \qquad \frac{x{:}\exists \mathsf{R}.\varphi}{\begin{array}{c}x\ \mathsf{R}\ y\\ y{:}\varphi\end{array}}\mathcal{T}_{\exists}$$

▷ To test consistency of a concept φ, normalize φ to ψ, initialize the tableau with $x{:}\psi$, saturate. Open branches $\rightsquigarrow$ consistent. (x arbitrary)

In contrast to the tableau calculi for theorem proving we have studied earlier, $\mathcal{T}_{\mathcal{ALC}}$ is run in "model generation mode". Instead of initializing the tableau with the axioms and the negated conjecture and hope that all branches will close, we initialize the $\mathcal{T}_{\mathcal{ALC}}$ tableau with axioms and the "conjecture" that a given concept φ is satisfiable – i.e. φ h as a member x, and hope for branches that are open, i.e. that make the conjecture true (and at the same time give a model).

Let us now work through two very simple examples; one unsatisfiable, and a satisfiable one.

$\mathcal{T}_{\mathcal{ALC}}$ Examples

▷ **Example 16.3.29.** We have two similar conjectures about children.

- ▷ $x{:}\forall \text{has\_child}.\text{man} \sqcap \exists \text{has\_child}.\overline{\text{man}}$ (all sons, but a daughter)
- ▷ $x{:}\forall \text{has\_child}.\text{man} \sqcap \exists \text{has\_child}.\text{man}$ (only sons, and at least one)

▷ **Example 16.3.30 (Tableau Proof).**

1	$x{:}\forall \text{has\_child}.\text{man} \sqcap \exists \text{has\_child}.\overline{\text{man}}$	initial	$x{:}\forall \text{has\_child}.\text{man} \sqcap \exists \text{has\_child}.\text{man}$	initial
2	$x{:}\forall \text{has\_child}.\text{man}$	$\mathcal{T}_{\sqcap}$	$x{:}\forall \text{has\_child}.\text{man}$	$\mathcal{T}_{\sqcap}$
3	$x{:}\exists \text{has\_child}.\overline{\text{man}}$	$\mathcal{T}_{\sqcap}$	$x{:}\exists \text{has\_child}.\text{man}$	$\mathcal{T}_{\sqcap}$
4	x has_child y	$\mathcal{T}_{\exists}$	x has_child y	$\mathcal{T}_{\exists}$
5	$y{:}\overline{\text{man}}$	$\mathcal{T}_{\exists}$	$y{:}\text{man}$	$\mathcal{T}_{\exists}$
6	$y{:}\text{man}$	$\mathcal{T}_{\forall}$	open	
7	$\bot$	$\mathcal{T}_{\bot}$		
	inconsistent			

The right tableau has a model: there are two persons, x and y. y is the only child of x, y is a man

Another example: this one is more complex, but the concept is satisfiable.

Another $\mathcal{T}_{\mathcal{ALC}}$ Example

▷ **Example 16.3.31.** $\forall \text{has\_child}.(\text{ugrad} \sqcup \text{grad}) \sqcap \exists \text{has\_child}.\overline{\text{ugrad}}$ is satisfiable.

- ▷ Let's try it on the board
- ▷ Tableau proof for the notes

1	$x{:}\forall \text{has\_child}.(\text{ugrad} \sqcup \text{grad}) \sqcap \exists \text{has\_child}.\overline{\text{ugrad}}$		
2	$x{:}\forall \text{has\_child}.(\text{ugrad} \sqcup \text{grad})$		$\mathcal{T}_{\sqcap}$
3	$x{:}\exists \text{has\_child}.\overline{\text{ugrad}}$		$\mathcal{T}_{\sqcap}$
4	x has_child y		$\mathcal{T}_{\exists}$
5	$y{:}\overline{\text{ugrad}}$		$\mathcal{T}_{\exists}$
6	$y{:}\text{ugrad} \sqcup \text{grad}$		$\mathcal{T}_{\forall}$
7	$y{:}\text{ugrad}$	$y{:}\text{grad}$	$\mathcal{T}_{\sqcup}$
8	$\bot$	open	

The left branch is closed, the right one represents a model: y is a child of x, y is a graduate student, x hat exactly one child: y.

After we got an intuition about $\mathcal{T}_{\mathcal{ALC}}$, we can now study the properties of the calculus to determine that it is a decision procedure for $\mathcal{ALC}$.

Properties of Tableau Calculi

▷ We study the following properties of a tableau calculus $\mathcal{C}$:

▷ **Termination**: there are no infinite sequences of rule applications.

▷ **Correctness**: If φ is satisfiable, then $\mathcal{C}$ terminates with an open branch.

▷ **Completeness**: If φ is in unsatisfiable, then $\mathcal{C}$ terminates and all branches are closed.

▷ **Complexity** of the algorithm (time and space complexity).

▷ Additionally, we are interested in the complexity of the satisfiability itself (as a benchmark)

The correctness result for $\mathcal{T}_{\mathcal{ALC}}$ is as usual: we start with a model of $x{:}\varphi$ and show that an $\mathcal{T}_{\mathcal{ALC}}$ tableau must have an open branch.

Correctness

▷ **Lemma 16.3.32.** *If φ satisfiable, then $\mathcal{T}_{\mathcal{ALC}}$ terminates on $x{:}\varphi$ with open branch.*

▷ *Proof:* Let $\mathcal{M}{:=}\langle\mathcal{D},[\![\cdot]\!]\rangle$ be a model for φ and $w{\in}[\![\varphi]\!]$.

1. We define $[\![x]\!]:=w$ and $\begin{array}{lll}\mathcal{I}{\models}(x{:}\psi) & \text{iff} & [\![x]\!]\in[\![\psi]\!]\\ \mathcal{I}{\models}x\ \mathsf{R}\ y & \text{iff} & \langle x,y\rangle{\in}[\![\mathsf{R}]\!]\\ \mathcal{I}{\models}S & \text{iff} & \mathcal{I}{\models}c \text{ for all } c{\in}S\end{array}$
2. This gives us $\mathcal{M}{\models}(x{:}\varphi)$ (base case)
3. If the branch is satisfiable, then either
   ▷ no rule applicable to leaf, (open branch)
   ▷ or rule applicable and one new branch satisfiable. (inductive case)
4. There must be an open branch. (by termination)

We complete the proof by looking at all the $\mathcal{T}_{\mathcal{ALC}}$ inference rules in turn.

Case analysis on the rules

$\mathcal{T}_{\sqcap}$ **applies** then $\mathcal{I}{\models}(x{:}\varphi\sqcap\psi)$, i.e. $[\![x]\!]\in[\![\varphi\sqcap\psi]\!]$
so $[\![x]\!]\in[\![\varphi]\!]$ and $[\![x]\!]\in[\![\psi]\!]$, thus $\mathcal{I}{\models}(x{:}\varphi)$ and $\mathcal{I}{\models}(x{:}\psi)$.

$\mathcal{T}_{\sqcup}$ **applies** then $\mathcal{I}{\models}(x{:}\varphi\sqcup\psi)$, i.e $[\![x]\!]\in[\![\varphi\sqcup\psi]\!]$
so $[\![x]\!]\in[\![\varphi]\!]$ or $[\![x]\!]\in[\![\psi]\!]$, thus $\mathcal{I}{\models}(x{:}\varphi)$ or $\mathcal{I}{\models}(x{:}\psi)$,
wlog. $\mathcal{I}{\models}(x{:}\varphi)$.

$\mathcal{T}_{\forall}$ **applies** then $\mathcal{I}{\models}(x{:}\forall\mathsf{R}.\varphi)$ and $\mathcal{I}{\models}x\ \mathsf{R}\ y$, i.e. $[\![x]\!]\in[\![\forall\mathsf{R}.\varphi]\!]$ and $\langle x,y\rangle{\in}[\![\mathsf{R}]\!]$, so $[\![y]\!]\in[\![\varphi]\!]$

$\mathcal{T}_{\exists}$ **applies** then $\mathcal{I}{\models}(x{:}\exists\mathsf{R}.\varphi)$, i.e $[\![x]\!]\in[\![\exists\mathsf{R}.\varphi]\!]$,
so there is a $v{\in}D$ with $\langle[\![x]\!],v\rangle{\in}[\![\mathsf{R}]\!]$ and $v{\in}[\![\varphi]\!]$.
We define $[\![y]\!]:=v$, then $\mathcal{I}{\models}x\ \mathsf{R}\ y$ and $\mathcal{I}{\models}(y{:}\varphi)$

For the completeness result for $\mathcal{T}_{\mathcal{ALC}}$ we have to start with an open tableau branch and construct at model that satisfies all judgements in the branch. We proceed by building a Herbrand model, whose domain consists of all the individuals mentioned in the branch and which interprets all concepts and roles as specified in the branch. Not surprisingly, the model thus constructed satisfies the branch.

Completeness of the Tableau Calculus

▷ **Lemma 16.3.33.** *Open saturated tableau branches for φ induce models for φ.*

▷ *Proof:* construct a model for the branch and verify for φ

1. Let $\mathcal{B}$ be an open saturated branch
   ▷ we define

$$\begin{aligned}\mathcal{D} &:= \{x|x{:}\psi\in\mathcal{B} \text{ or } z \mathrel{\mathsf{R}} x\in\mathcal{B}\}\\ [\![c]\!] &:= \{x|x{:}c\in\mathcal{B}\}\\ [\![\mathsf{R}]\!] &:= \{\langle x,y\rangle|x \mathrel{\mathsf{R}} y\in\mathcal{B}\}\end{aligned}$$

   ▷ well-defined since never $x{:}c, x{:}\overline{c}\in\mathcal{B}$ (otherwise $\mathcal{T}_\bot$ applies)
   ▷ $\mathcal{M}$ satisfies all constraints $x{:}c$, $x{:}\overline{c}$ and $x \mathrel{\mathsf{R}} y$, (by construction)
2. $\mathcal{M}\models(y{:}\psi)$, for all $y{:}\psi\in\mathcal{B}$ (on $k = size(\psi)$ next slide)
3. $\mathcal{M}\models(x{:}\varphi)$.

We complete the proof by looking at all the $\mathcal{T}_{\mathcal{ALC}}$ inference rules in turn.

Case Analysis for Induction

case $y{:}\psi = y{:}\psi_1 \sqcap \psi_2$ Then $\{y{:}\psi_1, y{:}\psi_2\} \subseteq \mathcal{B}$ ($\mathcal{T}_\sqcap$-rule, saturation) so $\mathcal{M}\models(y{:}\psi_1)$ and $\mathcal{M}\models(y{:}\psi_2)$ and $\mathcal{M}\models(y{:}\psi_1 \sqcap \psi_2)$ (IH, Definition)

case $y{:}\psi = y{:}\psi_1 \sqcup \psi_2$ Then $y{:}\psi_1\in\mathbf{B}$ or $y{:}\psi_2\in\mathbf{B}$ ($\mathcal{T}_\sqcup$, saturation) so $\mathcal{M}\models(y{:}\psi_1)$ or $\mathcal{M}\models(y{:}\psi_2)$ and $\mathcal{M}\models(y{:}\psi_1 \sqcup \psi_2)$ (IH, Definition)

case $y{:}\psi = y{:}\exists\mathbf{R}.\theta$ then $\{y \mathrel{\mathsf{R}} z, z{:}\theta\} \subseteq \mathbf{B}$ (z new variable) ($\mathcal{T}_\exists$-rules, saturation) so $\mathcal{M}\models(z{:}\theta)$ and $\mathcal{M}\models y \mathrel{\mathsf{R}} z$, thus $\mathcal{M}\models(y{:}\exists\mathsf{R}.\theta)$. (IH, Definition)

case $y{:}\psi = y{:}\forall\mathbf{R}.\theta$ Let $\langle [\![y]\!], v\rangle\in [\![\mathsf{R}]\!]$ for some $r\in\mathcal{D}$ then $v = z$ for some variable z with $y \mathrel{\mathsf{R}} z\in\mathbf{B}$ (construction of $[\![\mathsf{R}]\!]$) So $z{:}\theta\in\mathcal{B}$ and $\mathcal{M}\models(z{:}\theta)$. ($\mathcal{T}_\forall$-rule, saturation, Def) Since v was arbitrary we have $\mathcal{M}\models(y{:}\forall\mathsf{R}.\theta)$.

Termination

▷ **Theorem 16.3.34.** $\mathcal{T}_{\mathcal{ALC}}$ *terminates*

▷ To prove termination of a tableau algorithm, find a well-founded measure (function) that is decreased by all rules

$$\frac{\begin{array}{c}x{:}c\\x{:}\overline{c}\end{array}}{\bot}\mathcal{T}_\bot \qquad \frac{x{:}\varphi\sqcap\psi}{\begin{array}{c}x{:}\varphi\\x{:}\psi\end{array}}\mathcal{T}_\sqcap \qquad \frac{x{:}\varphi\sqcup\psi}{x{:}\varphi \mid x{:}\psi}\mathcal{T}_\sqcup \qquad \frac{\begin{array}{c}x{:}\forall \mathsf{R}.\varphi\\x\ \mathsf{R}\ y\end{array}}{y{:}\varphi}\mathcal{T}_\forall \qquad \frac{x{:}\exists \mathsf{R}.\varphi}{\begin{array}{c}x\ \mathsf{R}\ y\\y{:}\varphi\end{array}}\mathcal{T}_\exists$$

▷ *Proof:* Sketch (full proof very technical)

1. Any rule except $\mathcal{T}_\forall$ can only be applied once to $x{:}\psi$.
2. Rule $\mathcal{T}_\forall$ applicable to $x{:}\forall\mathsf{R}.\psi$ at most as the number of R-successors of x. (those y with $x\ \mathsf{R}\ y$ above)
3. The R-successors are generated by $x{:}\exists\mathsf{R}.\theta$ above, (number bounded by size of input formula)
4. Every rule application to $x{:}\psi$ generates constraints $z{:}\psi'$, where ψ' a proper sub-formula of ψ.

We can turn the termination result into a worst-case complexity result by examining the sizes of branches.

Complexity

▷ **Idea:** Work of tableau branches one after the other. (Branch size $\widehat{=}$ space complexity)

▷ **Observation 16.3.35.** *The size of the branches is* polynomial *in the size of the input formula:*

$$\textit{branch size} = |\textit{input formulae}| + \#(\exists\textit{-formulae}) \cdot \#(\forall\textit{-formulae})$$

▷ *Proof sketch:* Re-examine the termination proof and count: the first summand comes from Proof step 4., the second one from Proof step 3. and Proof step 2.

▷ **Theorem 16.3.36.** *The satisfiability problem for $\mathcal{ALC}$ is in* ***PSPACE****.*

▷ **Theorem 16.3.37.** *The satisfiability problem for $\mathcal{ALC}$ is* ***PSPACE****-Complete.*

▷ *Proof sketch:* Reduce a **PSPACE**-complete problem to $\mathcal{ALC}$-satisfiability

▷ **Theorem 16.3.38 (Time Complexity).** *The $\mathcal{ALC}$ satisfiability problem is in* ***EXPTIME****.*

▷ *Proof sketch:* There can be exponentially many branches (already for PL^0)

In summary, the theoretical complexity of $\mathcal{ALC}$ is the same as that for PL^0, but in practice $\mathcal{ALC}$ is much more expressive. So this is a clear win.

But the description of the tableau algorithm $\mathcal{T}_{\mathcal{ALC}}$ is still quite abstract, so we look at an exemplary implementation in a functional programmingfunctional programming language.

The functional Algorithm for $\mathcal{ALC}$

▷ **Observation:** (leads to a better treatment for $\exists$)

- the $\mathcal{T}_\exists$-rule generates the constraints x R y and $y{:}\psi$ from $x{:}\exists$R.ψ
- this triggers the $\mathcal{T}_\forall$-rule for $x{:}\forall$R.θ_i, which generate $y{:}\theta_1, \ldots, y{:}\theta_n$
- for y we have $y{:}\psi$ and $y{:}\theta_1, \ldots, y{:}\theta_n$. (do all of this in a single step)
- we are only interested in non-emptiness, not in particular witnesses (leave them out)

▷ **Definition 16.3.39.** The functional algorithm for $\mathcal{T}_{\mathcal{ALC}}$ is

consistent(S) =
  if $\{c, \overline{c}\} \subseteq S$ **then false**
  elif '$\varphi \sqcap \psi$'$\in S$ **and** ('φ'$\notin S$ or 'ψ'$\notin S$)
    then consistent($S \cup \{\varphi, \psi\}$)
  elif '$\varphi \sqcup \psi$'$\in S$ **and** $\{\varphi, \psi\} \notin S$
    then consistent($S \cup \{\varphi\}$) or consistent($S \cup \{\psi\}$)
  elif forall '$\exists$R.ψ'$\in S$
   consistent($\{\psi\} \cup \{\theta \in \theta \mid$ '$\forall$R.θ'$\in S\}$)
  else true

▷ Relatively simple to implement. (good implementations optimized)

▷ **But:** This is restricted to $\mathcal{ALC}$. (extension to other DL difficult)

Note that we have (so far) only considered an empty TBox: we have initialized the tableau with a normalized concept; so we did not need to include the concept definitions. To cover "real" ontologies, we need to consider the case of concept axioms as well.
We now extend $\mathcal{T}_{\mathcal{ALC}}$ with concept axioms. The key idea here is to realize that the concept axioms apply to all individuals. As the individuals are generated by the $\mathcal{T}_\exists$ rule, we can simply extend that rule to apply all the concepts axioms to the newly introduced individual.

Extending the Tableau Algorithm by Concept Axioms

▷ Concept axioms, e.g. child $\sqsubseteq$ son $\sqcup$ daughter cannot be handled in $\mathcal{T}_{\mathcal{ALC}}$ yet.

▷ **Idea:** Whenever a new variable y is introduced (by $\mathcal{T}_\exists$-rule) add the information that axioms hold for y.

- Initialize tableau with $\{x{:}\varphi\} \cup \mathcal{CA}$ ($\mathcal{CA}{:} =$ set of concept axioms)
- New rule for $\exists$: $\dfrac{x{:}\exists\mathsf{R}.\varphi \quad \mathcal{CA} = \{\alpha_1, \ldots, \alpha_n\}}{\begin{array}{c} y{:}\varphi \\ x\ \mathsf{R}\ y \\ y{:}\alpha_1 \\ \vdots \\ y{:}\alpha_n \end{array}}\mathcal{T}^\exists_{\mathcal{CA}}$ (instead of $\mathcal{T}_\exists$)

▷ **Problem:** $\mathcal{CA}{:=}\{\exists\mathsf{R}.c\}$ and start tableau with $x{:}d$ (non-termination)

The problem of this approach is that it spoils termination, since we cannot control the number of rule applications by (fixed) properties of the input formulae. The example shows this very nicely.

We only sketch a path towards a solution.

Non-Termination of $\mathcal{T}_{\mathcal{ALC}}$ with Concept Axioms

▷ **Problem:** $\mathcal{CA}{:=}\{\exists \mathsf{R}.c\}$ and start tableau with $x{:}d$. (non-termination)

$x{:}d$	start
$x{:}\exists \mathsf{R}.c$	in $\mathcal{CA}$
$x\ \mathsf{R}\ y_1$	$\mathcal{T}_{\exists}$
$y_1{:}c$	$\mathcal{T}_{\exists}$
$y_1{:}\exists \mathsf{R}.c$	$\mathcal{T}^{\exists}_{\mathcal{CA}}$
$y_1\ \mathsf{R}\ y_2$	$\mathcal{T}_{\exists}$
$y_2{:}c$	$\mathcal{T}_{\exists}$
$y_2{:}\exists \mathsf{R}.c$	$\mathcal{T}^{\exists}_{\mathcal{CA}}$
...	

Solution: Loop-Check:

▷ Instead of a new variable y take an old variable z, if we can guarantee that whatever holds for y already holds for z.

▷ We can only do this, iff the $\mathcal{T}_{\forall}$-rule has been exhaustively applied.

▷ **Theorem 16.3.40.** *The consistency problem of $\mathcal{ALC}$ with concept axioms is decidable.*

Proof sketch: $\mathcal{T}_{\mathcal{ALC}}$ with a suitable loop check terminates.

16.3.3 ABoxes, Instance Testing, and ALC

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27303`.

Now that we have a decision problem for $\mathcal{ALC}$ with concept axioms, we can go the final step to the general case of inference in description logics: we add an ABox with assertional axioms that describe the individuals.
We will now extend the description logic $\mathcal{ALC}$ with assertions that

▷ Instance Test: Concept Membership

▷ **Definition 16.3.41.** An instance test computes whether given an $\mathcal{ALC}$ ontology an individual is a member of a given class.

▷ **Example 16.3.42 (An Ontology).**

TBox (terminological Box)			ABox (assertional Box, data base)	
woman	=	person ⊓ has_Y	tony:person	Tony is a person
man	=	person ⊓ has_Y	tony:has_Y	Tony has a y-chrom

This entails: tony:man (Tony is a man).

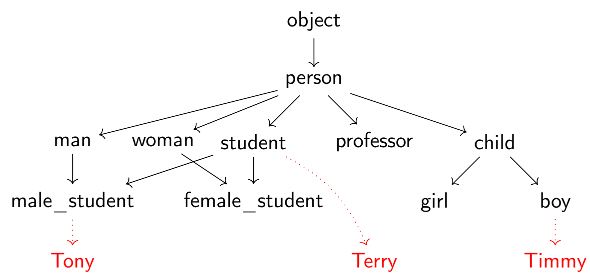
▷ **Problem:** Can we compute this?

If we combine classification with the instance test, then we get the full picture of how concepts and individuals relate to each other. We see that we get the full expressivity of semantic networks in $\mathcal{ALC}$.

Realization

▷ **Definition 16.3.43.** Realization is the computation of all instance relations between ABox objects and TBox concepts.

▷ **Observation:** It is sufficient to remember the lowest concepts in the subsumption graph. (rest by subsumption)



▷ **Example 16.3.44.** If tony:male_student is known, we do not need tony:man.

Let us now get an intuition on what kinds of interactions between the various parts of an ontology.

ABox Inference in $\mathcal{ALC}$: Phenomena

▷ There are different kinds of interactions between TBox and ABox in $\mathcal{ALC}$ and in description logics in general.

▷ **Example 16.3.45.**

property	example
internally inconsistent	tony:student, tony:$\overline{\text{student}}$
inconsistent with a TBox	TBox: student $\sqcap$ prof ABox: tony:student, tony:prof
implicit info that is not explicit	ABox: tony:$\forall$has_grad.genius tony has_grad mary $\models$ mary:genius
information that can be combined with TBox info	TBox: happy_prof = prof $\sqcap$ $\forall$has_grad.genius ABox: tony:happy_prof, tony has_grad mary $\models$ mary:genius

Again, we ask ourselves whether all of these are computable.
Fortunately, it is very simple to add assertions to $\mathcal{T}_{\mathcal{ALC}}$. In fact, we do not have to change anything, as the judgments used in the tableau are already of the form of ABox assertionss.

Tableau-based Instance Test and Realization

▷ **Query:** Do the ABox and TBox together entail $a{:}\varphi$? $(a{\in}\varphi?)$

▷ **Algorithm:** Test $a{:}\overline{\varphi}$ for consistency with ABox and TBox. (use our tableau)

- **Necessary changes:** (no big deal)
  - Normalize ABox wrt. TBox. (definition expansion)
  - Initialize the tableau with ABox in NNF. (so it can be used)
- **Example 16.3.46.**

Example: add mary:$\overline{\text{genius}}$ to determine $ABox, TBox \models$ mary:genius

TBox	happy_prof = prof ⊓ ∀has_grad.genius	tony:prof ⊓ ∀has_grad.genius	TBox
		tony has_grad mary	ABox
		mary:$\overline{\text{genius}}$	Query
		tony:prof	$\mathcal{T}_\sqcap$
ABox	tony:happy_prof tony has_grad mary	tony:∀has_grad.genius	$\mathcal{T}_\sqcap$
		mary:genius	$\mathcal{T}_\forall$
		$\bot$	$\mathcal{T}_\bot$

- **Note:** The instance test is the base for realization. (remember?)
- **Idea:** Extend to more complex ABox queries. (e.g. give me all instances of φ)

This completes our investigation of inference for $\mathcal{ALC}$. We summarize that $\mathcal{ALC}$ is a logic-based ontology language where the inference problems are all decidable/computable via $\mathcal{T}_{\mathcal{ALC}}$. But of course, while we have reached the expressivity of basic semantic networks, there are still things that we cannot express in $\mathcal{ALC}$, so we will try to extend $\mathcal{ALC}$ without losing decidability/computability.

16.4 Description Logics and the Semantic Web

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/27289`.

In this section we discuss how we can apply description logics in the real world, in particular, as a conceptual and algorithmic basis of the semantic web. That tries to transform the World Wide Web from a human-understandable web of multimedia documents into a "web of machine-understandable data". In this context, "machine-understandable" means that machines can draw inferences from data they have access to.

Note that the discussion in this digression is not a full-blown introduction to RDF and OWL, we leave that to [SR14; Her+13a; Hit+12] and the respective W3C recommendations. Instead we introduce the ideas behind the mappings from a perspective of the description logics we have discussed above.

The most important component of the semantic web is a standardized language that can represent "data" about information on the Web in a machine-oriented way.

Resource Description Framework

- **Definition 16.4.1.** The Resource Description Framework (RDF) is a framework for describing resources on the web. It is an XML vocabulary developed by the W3C.
- **Note:** RDF is designed to be read and understood by computers, not to be displayed to people. (it shows)
- **Example 16.4.2.** RDF can be used for describing (all "objects on the WWW")
  - properties for shopping items, such as price and availability

- ▷ time schedules for web events
- ▷ information about web pages (content, author, created and modified date)
- ▷ content and rating for web pictures
- ▷ content for search engines
- ▷ electronic libraries

Note that all these examples have in common that they are about "objects on the Web", which is an aspect we will come to now.

"Objects on the Web" are traditionally called "resources", rather than defining them by their intrinsic properties – which would be ambitious and prone to change – we take an external property to define them: everything that has a URI is a web resource. This has repercussions on the design of RDF.

Resources and URIs

- ▷ RDF describes resources with properties and property values.
- ▷ RDF uses Web identifiers (URIs) to identify resources.
- ▷ **Definition 16.4.3.** A resource is anything that can have a URI, such as `http://www.fau.de`.
- ▷ **Definition 16.4.4.** A property is a resource that has a name, such as *author* or *homepage*, and a property value is the value of a property, such as *Michael Kohlhase* or `http://kwarc.info/kohlhase`. (a property value can be another resource)
- ▷ **Definition 16.4.5.** A RDF statement s (also known as a triple) consists of a resource (the subject of s), a property (the predicate of s), and a property value (the object of s). A set of RDF triples is called an RDF graph.
- ▷ **Example 16.4.6.** Statement: *[This slide]^{subj} has been [author]^{pred}ed by [Michael Kohlhase]^{obj}*

The crucial observation here is that if we map "subjects" and "objects" to "individuals", and "predicates" to "relations", the RDF triples are just relational ABox statements of description logics. As a consequence, the techniques we developed apply.

Note:

Actually, a RDF graph is technically a labeled multigraph, which allows multiple edges between any two nodes (the resources) and where nodes and edges are labeled by URIs.

We now come to the concrete syntax of RDF. This is a relatively conventional XML syntax that combines RDF statements with a common subject into a single "description" of that resource.

XML Syntax for RDF

- ▷ RDF is a concrete XML vocabulary for writing statements
- ▷ **Example 16.4.7.** The following RDF document could describe the slides as a resource

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22−rdf−syntax−ns#"
         xmlns:dc= "http://purl.org/dc/elements/1.1/">
  <rdf:Description about="https://.../CompLog/kr/en/rdf.tex">
    <dc:creator>Michael Kohlhase</dc:creator>
    <dc:source>http://www.w3schools.com/rdf</dc:source>
  </rdf:Description>
</rdf:RDF>
```

This RDF document makes two statements:

- ▷ The subject of both is given in the about attribute of the rdf:Description element
- ▷ The predicates are given by the element names of its children
- ▷ The objects are given in the elements as URIs or literal content.

▷ **Intuitively:** RDF is a web scalable way to write down ABox information.

Note that XML namespaces play a crucial role in using element to encode the predicate URIs. Recall that an element name is a qualified name that consists of a namespace URI and a proper element name (without a colon character). Concatenating them gives a URI in our example the predicate URI induced by the dc:creator element is `http://purl.org/dc/elements/1.1/creator`. Note that as URIs go RDF URIs do not have to be URLs, but this one is and it references (is redirected to) the relevant part of the Dublin Core elements specification [DCM12].
RDF was deliberately designed as a standoff markup format, where URIs are used to annotate web resources by pointing to them, so that it can be used to give information about web resources without having to change them. But this also creates maintenance problems, since web resources may change or be deleted without warning.

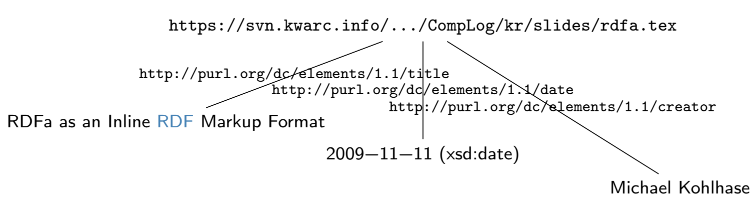
RDFa gives authors a way to embed RDF triples into web resources and make keeping RDF statements about them more in sync.

RDFa as an Inline RDF Markup Format

▷ **Problem:** RDF is a standoff markup format (annotate by URIs pointing into other files)

▷ **Example 16.4.8.**

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/" id="address">
    <h2 about="#address" property="dc:title">RDF as an Inline RDF Markup Format</h2>
    <h3 about="#address" property="dc:creator">Michael Kohlhase</h3>
    <em about="#address" property="dc:date" datatype="xsd:date"
        content="2009−11−11">November 11., 2009</em>
</div>
```



In the example above, the about and property attribute are reserved by RDFa and specify the subject and predicate of the RDF statement. The object consists of the body of the element, unless otherwise specified e.g. by the resource attribute.
Let us now come back to the fact that RDF is just an XML syntax for ABox statements.

RDF as an ABox Language for the Semantic Web

- **Idea:** RDF triples are ABox entries h R s or h:φ.
- **Example 16.4.9.** h is the resource for Ian Horrocks, s is the resource for Ulrike Sattler, R is the relation "hasColleague", and φ is the class foaf:Person

```
<rdf:Description about="some.uri/person/ian_horrocks">
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
  <hasColleague resource="some.uri/person/uli_sattler"/>
</rdf:Description>
```

- **Idea:** Now, we need an similar language for TBoxes (based on $\mathcal{ALC}$)

In this situation, we want a standardized representation language for TBox information; OWL does just that: it standardizes a set of knowledge representation primitives and specifies a variety of concrete syntaxes for them. OWL is designed to be compatible with RDF, so that the two together can form an ontology language for the web.

OWL as an Ontology Language for the Semantic Web

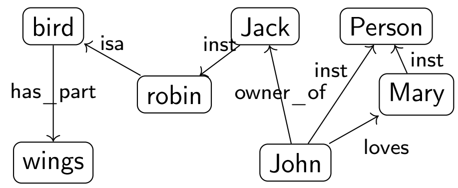
- **Task:** Complement RDF (ABox) with a TBox language.
- **Idea:** Make use of resources that are values in rdf:type. (called Classes)
- **Definition 16.4.10.** OWL (the ontology web language) is a language for encoding TBox information about RDF classes.
- **Example 16.4.11 (A concept definition for "Mother").** Mother=Woman $\sqcap$ Parent is represented as

<table>
<tr><th>XML Syntax</th><th>Functional Syntax</th></tr>
<tr><td><pre>
<EquivalentClasses>
   <Class IRI="Mother"/>
   <ObjectIntersectionOf>
      <Class IRI="Woman"/>
      <Class IRI="Parent"/>
   </ObjectIntersectionOf>
</EquivalentClasses>
</pre></td><td><pre>
EquivalentClasses(
  :Mother
  ObjectIntersectionOf(
    :Woman
    :Parent
)
)
</pre></td></tr>
</table>

But there are also other syntaxes in regular use. We show the functional syntax which is inspired by the mathematical notation of relations.

Extended OWL Example in Functional Syntax

▷ **Example 16.4.12.** The semantic network from Example 16.1.5 can be expressed in OWL (in functional syntax)




```
ClassAssertion (:Jack :robin)
ClassAssertion(:John :person)
ClassAssertion (:Mary :person)
ObjectPropertyAssertion(:loves :John :Mary)
ObjectPropertyAssertion(:owner :John :Jack)
SubClassOf(:robin :bird)
SubClassOf (:bird ObjectSomeValuesFrom(:hasPart :wing))
```

- ▷ ClassAssertion formalizes the "inst" relation,
- ▷ ObjectPropertyAssertion formalizes relations,
- ▷ SubClassOf formalizes the "isa" relation,
- ▷ for the "has_part" relation, we have to specify that *all birds have a part that is a wing* or equivalently *the class of birds is a subclass of all objects that have some wing*.

We have introduced the ideas behind using description logics as the basis of a "machine-oriented web of data". While the first OWL specification (2004) had three sublanguages "OWL Lite", "OWL DL" and "OWL Full", of which only the middle was based on description logics, with the OWL2 Recommendation from 2009, the foundation in description logics was nearly universally accepted.

The semantic web hype is by now nearly over, the technology has reached the "plateau of productivity" with many applications being pursued in academia and industry. We will not go into these, but briefly instroduce one of the tools that make this work.

SPARQL an RDF Query language

▷ **Definition 16.4.13.** SPARQL, the "SPARQL Protocol and RDF Query Language" is an RDF query language, able to retrieve and manipulate data stored in RDF. The SPARQL language was standardized by the World Wide Web Consortium in 2008 [PS08].

▷ SPARQL is pronounced like the word "*sparkle*".

▷ **Definition 16.4.14.** A system is called a SPARQL endpoint, iff it answers SPARQL queries.

▷ **Example 16.4.15.** Query for person names and their e-mails from a triplestore with FOAF data.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
  ?person a foaf:Person.
  ?person foaf:name ?name.
  ?person foaf:mbox ?email.
}
```


SPARQL end-points can be used to build interesting applications, if fed with the appropriate data. An interesting – and by now paradigmatic – example is the DBPedia project, which builds a large ontology by analyzing Wikipedia fact boxes. These are in a standard HTML form which can be analyzed e.g. by regular expressions, and their entries are essentially already in triple form: The subject is the Wikipedia page they are on, the predicate is the key, and the object is either the URI on the object value (if it carries a link) or the value itself.

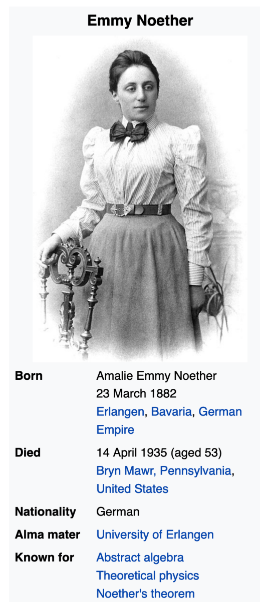
SPARQL Applications: DBPedia

▷ **Typical Application:** DBPedia screen-scrapes Wikipedia fact boxes for RDF triples and uses SPARQL for querying the induced triplestore.

▷ **Example 16.4.16 (DBPedia Query).** People who were born in Erlangen before 1900 (`http://dbpedia.org/snorql`)

```
SELECT ?name ?birth ?death ?person WHERE {
     ?person dbo:birthPlace :Erlangen .
     ?person dbo:birthDate ?birth .
     ?person foaf:name ?name .
     ?person dbo:deathDate ?death .
     FILTER (?birth < "1900−01−01"^^xsd:date) .
}
ORDER BY ?name
```

▷ The answers include Emmy Noether and Georg Simon Ohm.



A more complex DBPedia Query

▷ **Demo:** DBPedia `http://dbpedia.org/snorql/`
Query: Soccer players born in a country with more than 10 M inhabitants, who play as goalie in a club that has a stadium with more than 30.000 seats.
Answer: computed by DBPedia from a SPARQL query

```
SELECT distinct ?soccerplayer ?countryOfBirth ?team ?countryOfTeam ?stadiumcapacity
{
?soccerplayer a dbo:SoccerPlayer ;
   dbo:position|dbp:position <http://dbpedia.org/resource/Goalkeeper_(association_football)> ;
   dbo:birthPlace/dbo:country* ?countryOfBirth ;
   #dbo:number 13 ;
   dbo:team ?team .
   ?team dbo:capacity ?stadiumcapacity ; dbo:ground ?countryOfTeam .
   ?countryOfBirth a dbo:Country ; dbo:populationTotal ?population .
   ?countryOfTeam a dbo:Country .
FILTER (?countryOfTeam != ?countryOfBirth)
FILTER (?stadiumcapacity > 30000)
FILTER (?population > 10000000)
} order by ?soccerplayer
```

Results: Browse Go! Reset

SPARQL results:

soccerplayer	countryOfBirth	team	countryOfTeam	stadiumcapacity
:Abdesslam_Benabdellah	:Algeria	:Wydad_Casablanca	:Morocco	67000
:Airton_Moraes_Michellon	:Brazil	:FC_Red_Bull_Salzburg	:Austria	31000
:Alain_Gouaméné	:Ivory_Coast	:Raja_Casablanca	:Morocco	67000
:Allan_McGregor	:United_Kingdom	:Beşiktaş_J.K.	:Turkey	41903
:Anthony_Scribe	:France	:FC_Dinamo_Tbilisi	:Georgia_(country)	54549
:Brahim_Zaari	:Netherlands	:Raja_Casablanca	:Morocco	67000
:Bréiner_Castillo	:Colombia	:Deportivo_Táchira	:Venezuela	38755
:Carlos_Luis_Morales	:Ecuador	:Club_Atlético_Independiente	:Argentina	48069
:Carlos_Navarro_Montoya	:Colombia	:Club_Atlético_Independiente	:Argentina	48069
:Cristián_Muñoz	:Argentina	:Colo-Colo	:Chile	47000
:Daniel_Ferreyra	:Argentina	:FBC_Melgar	:Peru	60000
:David_Bičík	:Czech_Republic	:Karşıyaka_S.K.	:Turkey	51295
:David_Loria	:Kazakhstan	:Karşıyaka_S.K.	:Turkey	51295
:Denys_Boyko	:Ukraine	:Beşiktaş_J.K.	:Turkey	41903
:Eddie_Gustafsson	:United_States	:FC_Red_Bull_Salzburg	:Austria	31000
:Emilian_Dolha	:Romania	:Lech_Poznań	:Poland	43269
:Eusebio_Acasuzo	:Peru	:Club_Bolívar	:Bolivia	42000
:Faryd_Mondragón	:Colombia	:Real_Zaragoza	:Spain	34596
:Faryd_Mondragón	:Colombia	:Club_Atlético_Independiente	:Argentina	48069
:Federico_Vilar	:Argentina	:Club_Atlas	:Mexico	54500
:Fernando_Martinuzzi	:Argentina	:Real_Garcilaso	:Peru	45000
:Fábio_André_da_Silva	:Portugal	:Servette_FC	:Switzerland	30084
:Gerhard_Tremmel	:Germany	:FC_Red_Bull_Salzburg	:Austria	31000
:Gift_Muzadzi	:United_Kingdom	:Lech_Poznań	:Poland	43269
:Günay_Güvenç	:Germany	:Beşiktaş_J.K.	:Turkey	41903
:Hugo_Marques	:Portugal	:C.D._Primeiro_de_Agosto	:Angola	48500
:Héctor_Landazuri	:Colombia	:La_Paz_F.C.	:Bolivia	42000

We conclude our survey of the semantic web technology stack with the notion of a triplestore, which refers to the database component, which stores vast collections of ABox triples.

Triple Stores: the Semantic Web Databases

▷ **Definition 16.4.17.** A triplestore or RDF store is a purpose-built database for the storage RDF graphs and retrieval of RDF triples usually through variants of SPARQL.

▷ Common triplestores include

- ▷ Virtuoso: `https://virtuoso.openlinksw.com/` (used in DBpedia)
- ▷ GraphDB: `http://graphdb.ontotext.com/` (often used in WissKI)
- ▷ blazegraph: `https://blazegraph.com/` (open source; used in WikiData)

▷ **Definition 16.4.18.** A description logic reasoner implements of reaonsing services based on a satisfiabiltiy test for description logics.

▷ Common description logic reasoners include

- ▷ FACT++: `http://owl.man.ac.uk/factplusplus/`
- ▷ HermiT: `http://www.hermit-reasoner.com/`

▷ **Intuition:** Triplestores concentrate on querying very large ABoxes with partial consideration of the TBox, while DL reasoners concentrate on the full set of ontology inference services, but fail on large ABoxes.

Part IV

Planning & Acting

This part covers the AI subfield of “planning”, i.e. search-based problem solving with a structured representation language for environment state and actions — in planning, the focus is on the latter.

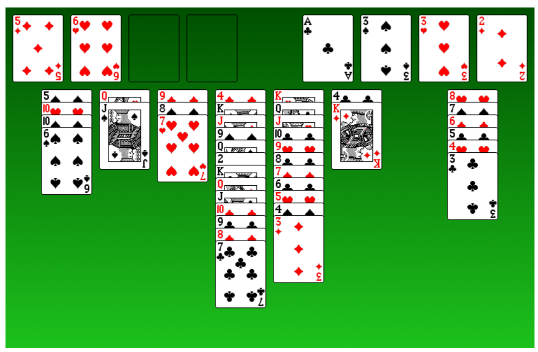
We first introduce the framework of planning (structured representation languages for problems and actions) and then present algorithms and complexity results. Finally, we lift some of the simplifying assumptions – deterministic, fully observable environments – we made in the previous parts of the course.

Chapter 17

Planning I: Framework

Reminder: Classical Search Problems

▷ **Example 17.0.1 (Solitaire as a Search Problem).**



- ▷ States: Card positions (e.g. position_Jspades=Qhearts).
- ▷ Actions: Card moves (e.g. move_Jspades_Qhearts_freecell4).
- ▷ Initial state: Start configuration.
- ▷ Goal states: All cards "home".
- ▷ Solutions: Card moves solving this game.

Planning

▷ **Ambition:** Write one program that can solve all classical search problems.

▷ **Idea:** For CSP, going from "state/action-level search" to "problem-description level search" did the trick.

▷ **Definition 17.0.2.** Let Π be a search problem (see chapter 6)

- ▷ The blackbox description of Π is an API providing functionality allowing to construct the state space: InitialState(), GoalTest(s), ...

- ▷ "Specifying the problem" $\widehat{=}$ programming the API.
- ▷ The declarative description of Π comes in a problem description language. This allows to implement the API, and much more.
  - ▷ "Specifying the problem" $\widehat{=}$ writing a problem description.
- ▷ Here, "problem description language" $\widehat{=}$ planning language. (up next)
- ▷ **But Wait:** Didn't we do this already in the last chapter with logics? (For the Wumpus?)

17.1 Logic-Based Planning

Before we go into the planning framework and its particular methods, let us see what we would do with the methods from Part III if we were to develop a "logic-based language" for describing states and actions. We will use the Wumpus world from section 10.1 as a running example.

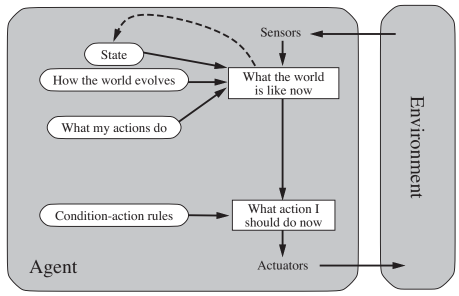
Fluents: Time-Dependent Knowledge in Planning

- ▷ **Recall from section 10.1:** We can represent the Wumpus rules in logical systems (propositional/first-order/ALC)
  - ▷ Use inference systems to deduce new world knowledge from percepts and actions.
- ▷ **Problem:** Representing (changing) percepts immediately leads to contradictions!
- ▷ **Example 17.1.1.** If the agent moves and a cell with a draft (a perceived breeze) is followed by one without.
- ▷ **Obvious Idea:** Make representations of percepts time-dependent
- ▷ **Example 17.1.2.** D^t for $t\in\mathbb{N}$ for PL^0 and $draft(t)$ in PL^1 and PL^{nq}.
- ▷ **Definition 17.1.3.** We use the word fluent to refer an aspect of the world that changes, all others we call atemporal.

Let us recall the agent-based setting we were using for the inference procedures from Part III. We will elaborate this further in this section.

Recap: Logic-Based Agents

- ▷ **Recall:** A model-based agent uses inference to model the environment, percept, and actions.




```
function KB−AGENT (percept) returns an action
  persistent: KB, a knowledge base
              t, a counter, initially 0, indicating time
  TELL(KB, MAKE−PERCEPT−SENTENCE(percept,t))
  action := ASK(KB, MAKE−ACTION−QUERY(t))
  TELL(KB, MAKE−ACTION−SENTENCE(action,t))
   t := t+1
  return action
```

▷ **Still Unspecified:** (up next)

- ▹ MAKE−PERCEPT−SENTENCE: the effects of percepts.
- ▹ MAKE−ACTION−QUERY: what is the best next action?
- ▹ MAKE−ACTION−SENTENCE: the effects of that action.

In particular, we will look at the effect of time/change. (neglected so far)

Now that we have the notion of fluents to represent the percepts at a given time point, let us try to model how they influence the agent's world model.

Fluents: Modeling the Agent's Sensors

▷ **Idea:** Relate percept fluents to atemporal cell attributes.

▷ **Example 17.1.4.** E.g., if the agent perceives a draft at time t, when it is in cell $[x, y]$, then there must be a breeze there: $\forall t, x, y.\mathrm{Ag@}(t, x, y) \Rightarrow \mathrm{draft}(t) \Leftrightarrow \mathrm{breeze}(x, y)$.

▷ Axiom like these model the agent's sensors – here that they are totally reliable: there is a breeze, iff the agent feels a draft.

▷ **Definition 17.1.5.** We call fluents that describe the agent's sensors sensor axioms.

▷ **Problem:** Where do fluents like $\mathrm{Ag@}(t, x, y)$ come from?

You may have noticed that for the sensor axioms we have only used first-order logic. There is a general story to tell here: if we have finite domains (as we do in the Wumpus cave) we can always "compile first-order logic" into propositional logic. We will develop this here before we go on with the Wumpus models.

Digression: Fluents and Finite Temporal Domains

- **Observation:** Fluents like $\forall t,x,y.\mathsf{Ag@}(t,x,y) \Rightarrow \mathsf{draft}(t) \Leftrightarrow \mathsf{breeze}(x,y)$ from Example 17.1.4 are best represented in first-order logic. In PL^0 and $\mathrm{PL}^{\mathrm{nq}}$ we would have to use concrete instances like $\mathsf{Ag@}(7,2,1) \Rightarrow \mathsf{draft}(7) \Leftrightarrow \mathsf{breeze}(2,1)$ for all suitable t, x, and y.
- **Problem:** Unless we restrict ourselves to finite domains and an end time t_{end} we have infinitely many axioms. Even then, formalization in PL^0 and $\mathrm{PL}^{\mathrm{nq}}$ is very tedious.
- **Solution:** Formalize in first-order logic and then compile down:
  1. enumerate ranges of bound variables, instantiate body, ($\rightsquigarrow \mathrm{PL}^{\mathrm{nq}}$)
  2. translate $\mathrm{PL}^{\mathrm{nq}}$ atoms to propositional variables. ($\rightsquigarrow \mathrm{PL}^0$)
- **In Practice:** The choice of domain, end time, and logic is up to agent designer, weighing expressivity vs. efficiency of inference.
- WLOG: We will use PL^1 in the following. (easier to read)

We now continue to our logic-based agent models: Now we focus on effect axioms to model the effects of an agent's actions.

Fluents: Effect Axioms for the Transition Model

- **Problem:** Where do fluents like $\mathsf{Ag@}(t,x,y)$ come from?
- **Thus:** We also need fluents to keep track of the agent's actions. (The transition model of the underlying search problem).
- **Idea:** We also use fluents for the representation of actions.
- **Example 17.1.6.** The action of "going forward" at time t is captured by the fluent $\mathsf{forw}(t)$.
- **Definition 17.1.7.** Effect axioms describe how the environment change under an agent's actions.
- **Example 17.1.8.** If the agent is in cell $[1,1]$ facing east at time 0 and goes forwardq, she is in cell $[2,1]$ and no longer in $[1,1]$:

  $$\mathsf{Ag@}(0,1,1) \wedge \mathsf{faceeast}(0) \wedge \mathsf{forw}(0) \Rightarrow \mathsf{Ag@}(1,2,1) \wedge \neg\mathsf{Ag@}(1,1,1)$$

  Generally: (barring exceptions for domain border cells)

  $$\forall t,x,y.\mathsf{Ag@}(t,x,y) \wedge \mathsf{faceeast}(t) \wedge \mathsf{forw}(t) \Rightarrow \mathsf{Ag@}(t+1,x+1,y) \wedge \neg\mathsf{Ag@}(t+1,x,y)$$

  This compiles down to $16 \cdot t_{\mathrm{end}}$ $\mathrm{PL}^{\mathrm{nq}}/\mathrm{PL}^0$ axioms.

Unfortunately, the percept fluents, sensor axioms, and effect axioms are not enough, as we will show in Example 17.1.9. We will see that this is a more general problem – the famous frame problem that needs to be considered whenever we deal with change in environments.

Frame and Frame Axioms

- ▷ **Problem:** Effect axioms are not enough.
- ▷ **Example 17.1.9.** Say that the agent has an arrow at time 0, and then moves forward into $[2,1]$, perceives a glitter, and knows that the Wumpus is ahead.

  To evaluate the action shoot(1), it the corresponding effect axiom needs to know havarrow(1), but cannot prove it from havarrow(0).

  Problem: The information of having an arrow has been lost in the move forward.
- ▷ **Definition 17.1.10.** The frame problem describes that for a representation of actions we need to formalize the not their effects on the aspects they change, but also their non-effect on the static frame of reference.
- ▷ **Partial Solution:** (there are many many more; some better)

  Frame axioms formalize that particular fluents are invariant under a given action.
- ▷ **Problem:** For an agent with n actions and an environment with m fluents, we need $\mathcal{O}(nm)$ frame axioms.

  Representing and reasoning with them easily drowns out the sensor and transition models.

We conclude our discussion with a rellatively complete implementation of a logic-based Wumpus agent, building on the schema from slide 565.

A Hybrid Agent for the Wumpus World

- ▷ **Example 17.1.11 (A Hybrid Agent).** This agent uses
  - ▷ logic inference for sensor and transition modeling,
  - ▷ special code and A^* for action selection & route planning.

function HYBRID−WUMPUS−AGENT($percept$) **returns** an action
  inputs: $percept$, a list, [stench,breeze,glitter,bump,scream]
  persistent: KB , a knowledge base, initially the atemporal "wumpus physics"
    t, a counter, initially 0, indicating time
    $plan$, an action sequence, initially empty
  TELL(KB, MAKE−PERCEPT−SENTENCE($percept$,t))
  TELL(KB, the temporal "physics" sentences **for** time t)

then some special code for action selection (up next)

  TELL(KB, MAKE−ACTION−SENTENCE($action$,t))
  $t := t + 1$
  return $action$

So far, not much new over our original version.

Now look at the “special code” we have promised.

A Hybrid Agent: Custom Action Selection

▷ **Example 17.1.12 (A Hybrid Agent).**

$safe$:= $\{[x,y]|$ASK(KB,OK(t,x,y))=T$\}$
if ASK(KB,glitter(t)) = T **then**
  $plan$:= [grab] + PLAN−ROUTE($current$,$\{[1,1]\}$,$safe$) + [exit]
if $plan$ is empty **then**
  $unvisited$:= $\{[x,y]|$ASK(KB,Ag@(t',x,y))=F$\}$ **for** all $t'\leq t$
  $plan$:= PLAN−ROUTE($current$,$unvisited \cup safe$,$safe$)
if $plan$ is empty and ASK(KB,havarrow(t)) = T **then**
  $possible\_wumpus$:= $\{x,y|[x,y]\}$ASK(KB,$\neg Wx,y$) = F
  $plan$:= PLAN−SHOT($current$,$possible\_wumpus$,$safe$)
if $plan$ is empty **then** // no choice but **to** take a risk
  not_unsafe := $\{[x,y]|$ASK(KB,$\neg OKtx,y$) = F$\}$
  $plan$:= PLAN−ROUTE($current$,$unvisited \cup not\_unsafe$,$safe$)
if $plan$ is empty **then**
  $plan$:= PLAN−ROUTE($current$,$\{[1,1]\}$,$safe$) + [exit]
  $action$:= POP($plan$)

Note that OK and glitter are fluents, since the Wumpus might have died or the gold might have been grabbed.

And finally the route planning part of the code. This is essentially just A^* search.

A Hybrid Agent: Custom Action Selection

▷ **Example 17.1.13.** And the code for PLAN−ROUTE (PLAN−SHOT similar)

function PLAN−ROUTE(current,goals,allowed) **returns** an action sequence
  inputs: current, the agent's current position
    goals, a set of squares; try **to** plan a route **to** one of them
    allowed, a set of squares that can form part of the route
  problem := ROUTE−PROBLEM(current,goals,allowed)
  return A^*(problem)

▷ **Evaluation:** Even though this works for the Wumpus world, it is not the "universal, logic-based problem solver" we dreamed of!

▷ Planning tries to solve this with another representation of actions. (up next)

17.2 Planning: Introduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26892`.

How does a planning language describe a problem?

▷ **Definition 17.2.1.** A planning language is a logical language for the components of a search problem; in particular a *logical description* of the

- ▷ possible states (vs. blackbox: data structures). (E.g.: predicate $Eq(.,.)$.)
- ▷ initial state I (vs. data structures). (E.g.: $Eq(x,1)$.)
- ▷ goal test G (vs. a goal test function). (E.g.: $Eq(x,2)$.)
- ▷ set A of actions in terms of preconditions and effects (vs. functions returning applicable actions and successor states). (E.g.: "increment x: pre $Eq(x,1)$, eff $Eq(x \wedge 2) \wedge \neg Eq(x,1)$".)

A logical description of all of these is called a planning task.

▷ **Definition 17.2.2.** Solution (plan) $\hat{=}$ sequence of actions from $\mathcal{A}$, transforming $\mathcal{I}$ into a state that satisfies $\mathcal{G}$. (E.g.: "increment x".)

The process of finding a plan given a planning task is called planning.

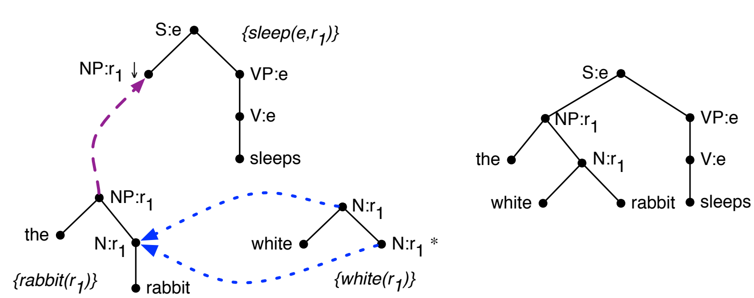
Planning Language Overview

▷ **Disclaimer:** Planning languages go way beyond classical search problems. There are variants for inaccessible, stochastic, dynamic, continuous, and multi-agent settings.

▷ We focus on classical search for simplicity (and practical relevance).

▷ For a comprehensive overview, see [GNT04].

Application: Natural Language Generation

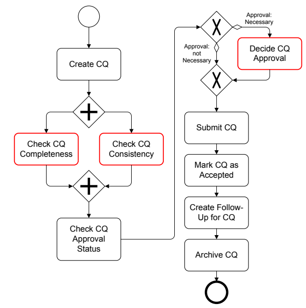


▷ **Input**: Tree-adjoining grammar, intended meaning.

▷ **Output**: Sentence expressing that meaning.

Application: Business Process Templates at SAP

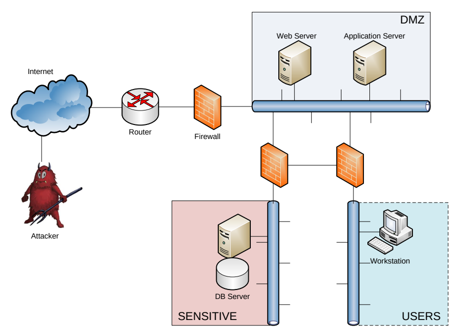
Action name	precondition	effect
Check CQ Completeness	CQ.archiving:notArchived	CQ.completeness:complete OR CQ.completeness:notComplete
Check CQ Consistency	CQ.archiving:notArchived	CQ.consistency:consistent OR CQ.consistency:notConsistent
Check CQ Approval Status	CQ.archiving:notArchived AND CQ.approval:notChecked AND CQ.completeness:complete AND CQ.consistency:consistent	CQ.approval:necessary OR CQ.approval:notNecessary
Decide CQ Approval	CQ.archiving:notArchived AND CQ.approval:necessary	CQ.approval:granted OR CQ.approval:notGranted
Submit CQ	CQ.archiving:notArchived AND (CQ.approval:notNecessary OR CQ.approval:granted)	CQ.submission:submitted
Mark CQ as Accepted	CQ.archiving:notArchived AND CQ.submission:submitted	CQ.acceptance:accepted
Create Follow-Up for CQ	CQ.archiving:notArchived AND CQ.acceptance:accepted	CQ.followUp:documentCreated
Archive CQ	CQ.archiving:notArchived	CQ.archiving:archived

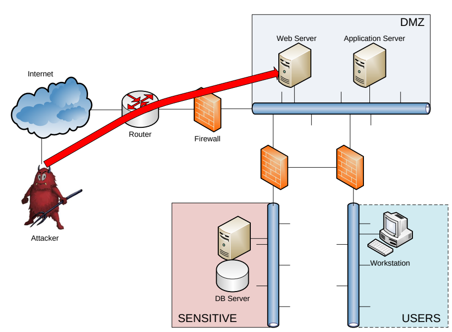


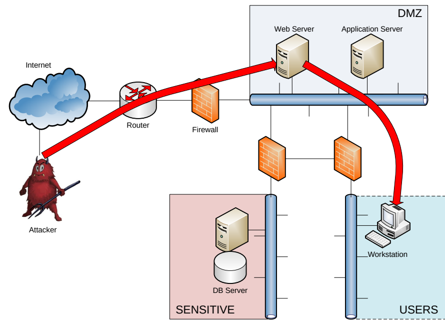
▷ **Input**: model of behavior of activities on business objects, process endpoint.

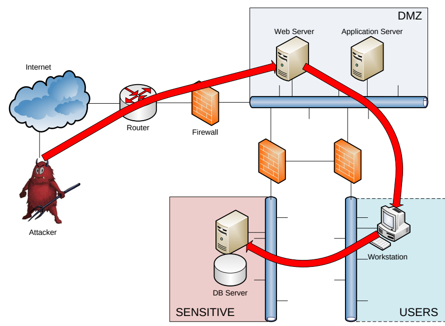
▷ **Output**: Process template leading to this point.

Application: Automatic Hacking









▷ **Input**: Network configuration, location of sensible data.

▷ **Output**: Sequence of exploits giving access to that data.

Reminder: General Problem Solving, Pros and Cons

▷ **Powerful:** In some applications, generality is absolutely necessary. (E.g. SAP)

▷ **Quick:** Rapid prototyping: 10s lines of problem description vs. 1000s lines of C++ code. (E.g. language generation)

▷ **Flexible:** Adapt/maintain *the description*. (E.g. network security)

▷ **Intelligent:** Determines automatically how to solve a complex problem effectively! (The ultimate goal, no?!)

▷ **Efficiency loss:** Without any domain-specific knowledge about chess, you don't beat Kasparov . . .

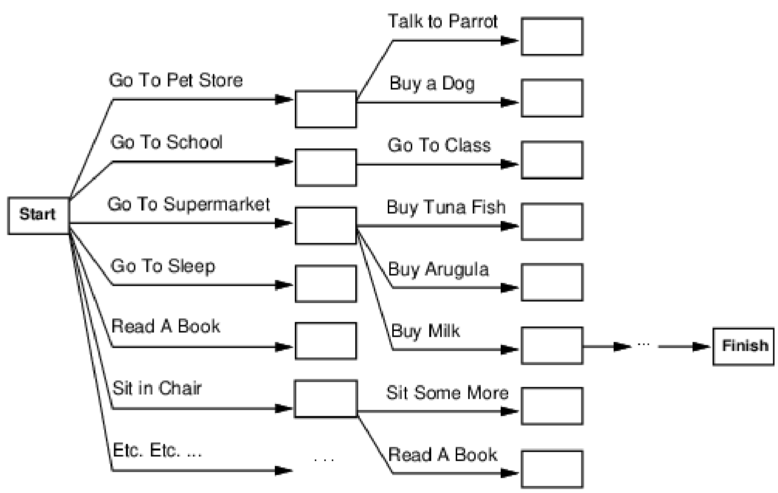
  ▷ Trade-off between "automatic and general" vs. "manual work but effective".

▷ **Research Question:** How to make fully automatic algorithms effective?

Search vs. planning

▷ Consider the task *get milk, bananas, and a cordless drill*

▷ Standard search algorithms seem to fail miserably:



After-the-fact heuristic/goal test inadequate

▷ Planning systems do the following:

1. open up action and goal representation to allow selection
2. divide-and-conquer by subgoaling

▷ relax requirement for sequential construction of solutions

	Search	Planning
States	Lisp data structures	Logical sentences
Actions	Lisp code	Preconditions/outcomes
Goal	Lisp code	Logical sentence (conjunction)
Plan	Sequence from S_0	Constraints on actions

Reminder: Greedy Best-First Search and A^*

▷ Duplicate elimination omitted for simplicity:

function Greedy_Best−First_Search [A^*](problem) **returns** a solution, or failure
  node := a node n with n.state=problem.InitialState
  frontier := a priority queue ordered by ascending h [$g+h$], only element n
  loop do
    if Empty?(frontier) **then return** failure
    n := Pop(frontier)
    if problem.GoalTest(n.State) **then return** Solution(n)
    for each action a **in** problem.Actions(n.State) **do**
      n' := ChildNode(problem,n,a)

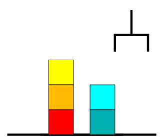
Insert(n', $h(n')$ $[g(n') + h(n')]$, frontier)

▷ Is Greedy Best-First Search optimal? No $\leadsto$ *satisficing* planning.

▷ Is A^* optimal? Yes, but only if h is admissible $\leadsto$ *optimal* planning, **with such** h.

ps. "Making Fully Automatic Algorithms Effective"

▷ **Example 17.2.3.**



▷ n blocks, 1 hand.

▷ A single action either takes a block with the hand or puts a block we're holding onto some other block/the table.

blocks	states	blocks	states
1	1	9	4596553
2	3	10	58941091
3	13	11	824073141
4	73	12	12470162233
5	501	13	202976401213
6	4051	14	3535017524403
7	37633	15	65573803186921
8	394353	16	1290434218669921

▷ **Observation 17.2.4.** *State spaces typically are huge even for simple problems.*

▷ **In other words:** Even solving "simple problems" automatically (without help from a human) requires a form of intelligence.

▷ With blind search, even the largest super computer in the world won't scale beyond 20 blocks!

Algorithmic Problems in Planning

▷ **Definition 17.2.5.** We speak of satisficing planning if
Input: A planning task Π.
Output: A plan for Π, or "unsolvable" if no plan for Π exists.
and of optimal planning if
Input: A planning task Π.
Output: An optimal plan for Π, or "unsolvable" if no plan for Π exists.

▷ The techniques successful for either one of these are almost disjoint. And satisficing planning is *much* more effective in practice.

▷ **Definition 17.2.6.** Programs solving these problems are called (optimal) planner, planning system, or planning tool.

Our Agenda for This Topic

- ▷ **Now:** Background, planning languages, complexity.
  - ▷ Sets up the framework. Computational complexity is essential to distinguish different algorithmic problems, and for the design of heuristic functions. (see next)
- ▷ **Next:** How to automatically generate a heuristic function, given planning language input?
  - ▷ Focussing on heuristic search as the solution method, this is the main question that needs to be answered.

Our Agenda for This Chapter

1. **The History of Planning**: How did this come about?
   - ▷ Gives you some background, and motivates our choice to focus on heuristic search.
2. **The STRIPS Planning Formalism**: Which concrete planning formalism will we be using?
   - ▷ Lays the framework we'll be looking at.
3. **The PDDL Language**: What do the input files for off-the-shelf planning software look like?
   - ▷ So you can actually play around with such software. (Exercises!)
4. **Planning Complexity**: How complex is planning?
   - ▷ The price of generality is complexity, and here's what that "price" is, exactly.

17.3 The History of Planning

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26894`.

Planning History: In the Beginning . . .

- ▷ **In the beginning: Man invented Robots:**
  - ▷ "Planning" as in "the making of plans by an autonomous robot".
  - ▷ Shakey the Robot (Full video here)
- ▷ **In a little more detail:**
  - ▷ [NS63] introduced *general problem solving*.

▷ ... *not much happened (well not much we still speak of today)* ...

▷ 1966-72, Stanford Research Institute developed a robot named "Shakey".

▷ They needed a "planning" component taking decisions.

▷ They took inspiration from general problem solving and theorem proving, and called the resulting algorithm STRIPS.

History of Planning Algorithms

▷ **Compilation into Logics/Theorem Proving:**

▷ e.g. $\exists s_0, a, s_1.at(A, s_0) \wedge execute(s_0, a, s_1) \wedge at(B, s_1)$

▷ **Popular when**: Stone Age – 1990.

▷ **Approach**: *From planning task description, generate PL1 formula φ that is satisfiable iff there exists a plan; use a theorem prover on φ.*

▷ **Keywords/cites**: Situation calculus, frame problem, ...

▷ Partial order planning

▷ e.g. $open = \{at(B)\}$; apply $move(A, B)$; $\rightsquigarrow open = \{at(A)\}$ ...

▷ **Popular when**: 1990 – 1995.

▷ **Approach**: *Starting at goal, extend partially ordered set of actions by inserting achievers for open sub-goals, or by adding ordering constraints to avoid conflicts.*

▷ **Keywords/cites**: **UCPOP** [PW92], causal links, flaw selection strategies, ...

History of Planning Algorithms, ctd.

▷ GraphPlan

▷ e.g. $F_0 = at(A); A_0 = \{move(A, B)\}; F_1 = \{at(B)\}$;
mutex $A_0 = \{move(A, B), move(A, C)\}$.

▷ **Popular when**: 1995 – 2000.

▷ **Approach**: *In a forward phase, build a layered "planning graph" whose "time steps" capture which pairs of actions can achieve which pairs of facts; in a backward phase, search this graph starting at goals and excluding options proved to not be feasible.*

▷ **Keywords/cites**: [BF95; BF97; Koe+97], action/fact mutexes, step-optimal plans, ...

▷ **Planning as SAT:**

▷ SAT variables $at(A)_0$, $at(B)_0$, $move(A, B)_0$, $move(A, C)_0$, $at(A)_1$, $at(B)_1$; clauses to encode transition behavior e.g. ${at(B)_1}^{\mathsf{F}} \vee {move(A, B)_0}^{\mathsf{T}}$; unit clauses to encode initial state ${at(A)_0}^{\mathsf{T}}$, ${at(B)_0}^{\mathsf{T}}$; unit clauses to encode goal ${at(B)_1}^{\mathsf{T}}$.

- **Popular when**: 1996 – today.
- **Approach**: *From planning task description, generate propositional CNF formula φ_k that is satisfiable iff there exists a plan with k steps; use a SAT solver on φ_k, for different values of k.*
- **Keywords/cites**: [KS92; KS98; RHN06; Rin10], SAT encoding schemes, BlackBox, ...

History of Planning Algorithms, ctd.

- **Planning as Heuristic Search:**
  - init $at(A)$; apply $move(A, B)$; generates state $at(B)$; ...
  - **Popular when**: 1999 – today.
  - **Approach**: Devise a method $\mathcal{R}$ to simplify ("relax") any planning task Π; given Π, solve $\mathcal{R}(\Pi)$ to generate a heuristic function h for informed search.
  - **Keywords/cites**: [BG99; HG00; BG01; HN01; Ede01; GSS03; Hel06; HHH07; HG08; KD09; HD09; RW10; NHH11; KHH12a; KHH12b; KHD13; DHK15], critical path heuristics, ignoring delete lists, relaxed plans, landmark heuristics, abstractions, partial delete relaxation, ...

The International Planning Competition (IPC)

- **Definition 17.3.1.** The International Planning Competition (IPC) is an event for benchmarking planners (`http://ipc.icapsconference.org/`)
  - **How**: Run competing planners on a set of benchmarks.
  - **When**: Runs every two years since 2000, annually since 2014.
  - **What**: Optimal track vs. satisficing track; others: uncertainty, learning, ...
- **Prerequisite/Result:**
  - Standard representation language: PDDL [McD+98; FL03; HE05; Ger+09]
  - Problem Corpus: ≈ 50 **domains**, $\gg 1000$ **instances**, 74 (!!) planners in 2011

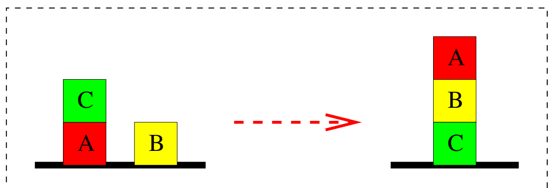
International Planning Competition

- **Question:** If planners x and y compete in IPC'YY, and x wins, is x "better than" y?
- **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

▷ **Generally:** reserved for the plenary sessions ⇝ be there!

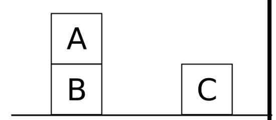
Planning History, p.s.: Planning is Non-Trivial!

▷ **Example 17.3.2.** The Sussman anomaly is a simple blocksword planning problem:

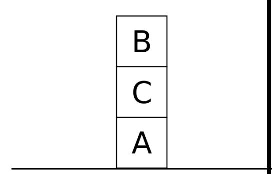


Simple planners that split the goal into subgoals on(A, B) and on(B, C) fail:

- ▷ If we pursue on(A, B) by unstacking C, and moving A onto B, we achieve the first subgoal, but cannot achieve the second without undoing the first.



- ▷ If we pursue on(B, C) by moving B onto C, we achieve the second subgoal, but cannot achieve the first without undoing the second.



17.4 The STRIPS Planning Formalism

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26896`.

STRIPS Planning

▷ **Definition 17.4.1.** STRIPS = Stanford Research Institute Problem Solver.

STRIPS is the simplest possible (reasonably expressive) logics based planning language.

▷ STRIPS has only propositional variables as atomic formulae.

▷ Its preconditions/effects/goals are as canonical as imaginable:

- ▷ Preconditions, goals: conjunctions of atoms.
- ▷ Effects: conjunctions of literals

▷ We use the common special-case notation for this simple formalism.

▷ I'll outline some extensions beyond STRIPS later on, when we discuss PDDL.

▷ **Historical note:** STRIPS [FN71] was originally a planner (cf. Shakey), whose language actually wasn't quite that simple.

STRIPS Planning: Syntax

▷ **Definition 17.4.2.** A STRIPS task is a quadruple $\langle P, A, I, G\rangle$ where:

- ▷ P is a finite set of facts (aka proposition).
- ▷ A is a finite set of actions; each $a \in A$ is a triple $a = \langle \mathrm{pre}_a, \mathrm{add}_a, \mathrm{del}_a\rangle$ of subsets of P referred to as the action's precondition, add list, and delete list respectively; we require that $\mathrm{add}_a \cap \mathrm{del}_a = \emptyset$.
- ▷ $I \subseteq P$ is the initial state.
- ▷ $G \subseteq P$ is the goal.

We will often give each action $a \in A$ a name (a string), and identify a with that name.

▷ **Note:** We assume, for simplicity, that every action has cost 1. (Unit costs, cf. chapter 6)

"TSP" in Australia

▷ **Example 17.4.3 (Salesman Travelling in Australia).**

Strictly speaking, this is not actually a **TSP** problem instance; simplified/adapted for illustration.

STRIPS Encoding of "TSP"

▷ **Example 17.4.4 (continuing).**

- ▷ Facts P: $\{\mathsf{at}(x), \mathsf{vis}(x) | x \in \{\mathsf{Sy}, \mathsf{Ad}, \mathsf{Br}, \mathsf{Pe}, \mathsf{Da}\}\}$.
- ▷ Initial state I: $\{\mathsf{at}(\mathsf{Sy}), \mathsf{vis}(\mathsf{Sy})\}$.
- ▷ Goal G:$\{\mathsf{at}(\mathsf{Sy})\} \cup \{\mathsf{vis}(x) | x \in \{\mathsf{Sy}, \mathsf{Ad}, \mathsf{Br}, \mathsf{Pe}, \mathsf{Da}\}\}$.
- ▷ Actions $a \in A$: $\mathsf{drv}(x, y)$ where x and y have a road.
  Preconditions pre_a: $\{\mathsf{at}(x)\}$.
  Add list add_a: $\{\mathsf{at}(y), \mathsf{vis}(y)\}$.
  Delete list del_a: $\{\mathsf{at}(x)\}$.
- ▷ Plan: $\langle \mathsf{drv}(\mathsf{Sy}, \mathsf{Br}), \mathsf{drv}(\mathsf{Br}, \mathsf{Sy}), \mathsf{drv}(\mathsf{Sy}, \mathsf{Ad}), \mathsf{drv}(\mathsf{Ad}, \mathsf{Pe}), \mathsf{drv}(\mathsf{Pe}, \mathsf{Ad}), \ldots$
  $\ldots, \mathsf{drv}(\mathsf{Ad}, \mathsf{Da}), \mathsf{drv}(\mathsf{Da}, \mathsf{Ad}), \mathsf{drv}(\mathsf{Ad}, \mathsf{Sy})\rangle$

STRIPS Planning: Semantics

▷ **Idea:** We define a plan for a STRIPS task Π as a solution to an induced search problem Θ_Π. (save work by reduction)

▷ **Definition 17.4.5.** Let $\Pi := \langle P, A, I, G\rangle$ be a STRIPS task. The search problem induced by Π is $\Theta_\Pi = \langle S_P, A, T_A, I, S_G\rangle$ where:

- ▷ The states (also world state) $S_G := \mathcal{P}(P)$ are the subsets of P.
- ▷ A is just Π's action set. (so we can define plans easily)
- ▷ The transition model T_A is $\{s \xrightarrow{a} \mathsf{apply}(s, a) | \mathsf{pre}_a \subseteq s\}$.
  If $\mathsf{pre}_a \subseteq s$, then $a \in A$ is applicable in s and $\mathsf{apply}(s, a) := s \cup \mathsf{add}_a \backslash \mathsf{del}_a$. If $\mathsf{pre}_a \not\subseteq s$, then $\mathsf{apply}(s, a)$ is undefined.
- ▷ I is Π's initial state.
- ▷ The goal states $S_G = \{s \in S_G | G \subseteq s\}$ are those that satisfy Π's goal.

An (optimal) plan for Π is an (optimal) solution Θ_Π, i.e., a path from s to some $s' \in S_G$. A solution for I is called a plan for Π. Π is solvable if a plan for Π exists.

For $a = \langle a_1, \ldots, a_n\rangle$, $\mathsf{apply}(s, a) := \mathsf{apply}(s, \mathsf{apply}(s, a_2 \ldots \mathsf{apply}(s, a_n)))$ if each a_i is applicable in the respective state; else, $\mathsf{apply}(s, a)$ is undefined.

STRIPS Encoding of Simplified "TSP"

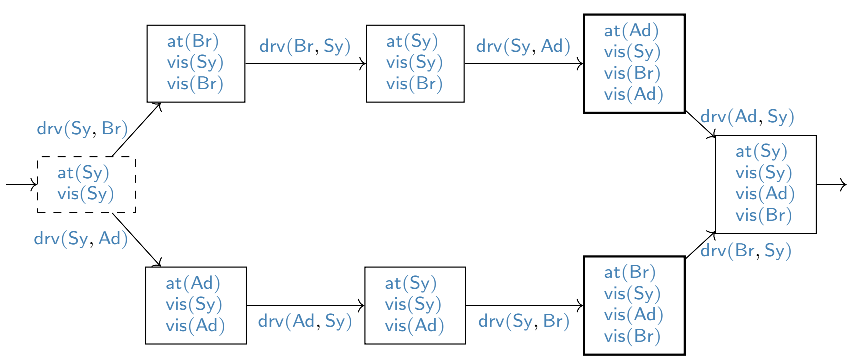
▷ **Example 17.4.6 (Simplified Traveling Salesman Problem in Australia).**

Let TSP_ be the STRIPS task, $\langle P, A, I, G\rangle$, where

- ▷ Facts P: $\{\mathrm{at}(x), \mathrm{vis}(x) | x \in \{\mathrm{Sy}, \mathrm{Ad}, \mathrm{Br}\}\}$.
- ▷ Initial state I: $\{\mathrm{at}(\mathrm{Sy}), \mathrm{vis}(\mathrm{Sy})\}$.
- ▷ Goal G: $\{\mathrm{vis}(x) | x \in \{\mathrm{Sy}, \mathrm{Ad}, \mathrm{Br}\}\}$ (note: noat(Sy))
- ▷ Actions A: $a \in A$: $\mathrm{drv}(x, y)$ where x y have a road.
  - ▷ preconditions pre_a: $\{\mathrm{at}(x)\}$.
  - ▷ add list add_a: $\{\mathrm{at}(y), \mathrm{vis}(y)\}$.
  - ▷ delete list del_a: $\{\mathrm{at}(x)\}$.

Questionaire: State Space of TSP_

▷ The state space of the search problem $\Theta_{\mathrm{TSP\_}}$ induced by TSP_ from Example 17.4.6 is



▷ **Question:** Are there any plans for TSP_ in this graph?

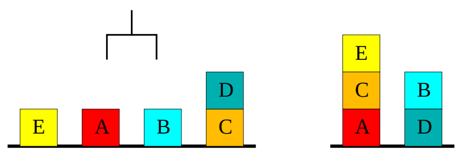
▷ **Answer:** Yes, two – plans for TSP_ are solutions for $\Theta_{\mathrm{TSP\_}}$, dashed node $\widehat{=}$ I, thick nodes $\widehat{=}$ G:

- ▷ $\mathrm{drv}(\mathrm{Sy}, \mathrm{Br}), \mathrm{drv}(\mathrm{Br}, \mathrm{Sy}), \mathrm{drv}(\mathrm{Sy}, \mathrm{Ad})$ (upper path)
- ▷ $\mathrm{drv}(\mathrm{Sy}, \mathrm{Ad}), \mathrm{drv}(\mathrm{Ad}, \mathrm{Sy}), \mathrm{drv}(\mathrm{Sy}, \mathrm{Br})$. (lower path)

- **Question:** Is the graph above actually the state space induced by ?
- **Answer:** No, only the part reachable from I. The state space of $\Theta_{\mathsf{TSP\_}}$ also includes e.g. the states $\{\mathsf{vis}(\mathsf{Sy})\}$ and $\{\mathsf{at}(\mathsf{Sy}), \mathsf{at}(\mathsf{Br})\}$.

The Blocksworld

- **Definition 17.4.7.** The blocks world is a simple planning domain: a set of wooden blocks of various shapes and colors sit on a table. The goal is to build one or more vertical stacks of blocks. Only one block may be moved at a time: it may either be placed on the table or placed atop another block.
- **Example 17.4.8.**



Initial State Goal State

  - Facts: $\mathsf{on}(x,y)$, $\mathsf{onTable}(x)$, $\mathsf{clear}(x)$, $\mathsf{holding}(x)$, $\mathsf{armEmpty}$.
  - Initial state: $\{\mathsf{onTable}(E), \mathsf{clear}(E), \ldots, \mathsf{onTable}(C), \mathsf{on}(D,C), \mathsf{clear}(D), \mathsf{armEmpty}\}$.
  - Goal: $\{\mathsf{on}(E,C), \mathsf{on}(C,A), \mathsf{on}(B,D)\}$.
  - Actions: $\mathsf{stack}(x,y)$, $\mathsf{unstack}(x,y)$, $\mathsf{putdown}(x)$, $\mathsf{pickup}(x)$.
  - $\mathsf{stack}(x,y)$?

    pre : $\{\mathsf{holding}(x), \mathsf{clear}(y)\}$
    add : $\{\mathsf{on}(x,y), \mathsf{armEmpty}\}$
    del : $\{\mathsf{holding}(x), \mathsf{clear}(y)\}$.

STRIPS for the Blocksworld

- **Question:** Which are correct encodings (ones that are part of some correct overall model) of the STRIPS Blocksworld $\mathsf{pickup}(x)$ action schema?

(A)	$\{\mathsf{onTable}(x), \mathsf{clear}(x), \mathsf{armEmpty}\}$ $\{\mathsf{holding}(x)\}$ $\{\mathsf{onTable}(x)\}$	(B)	$\{\mathsf{onTable}(x), \mathsf{clear}(x), \mathsf{armEmpty}\}$ $\{\mathsf{holding}(x)\}$ $\{\mathsf{armEmpty}\}$
(C)	$\{\mathsf{onTable}(x), \mathsf{clear}(x), \mathsf{armEmpty}\}$ $\{\mathsf{holding}(x)\}$ $\{\mathsf{onTable}(x), \mathsf{armEmpty}, \mathsf{clear}(x)\}$	(D)	$\{\mathsf{onTable}(x), \mathsf{clear}(x), \mathsf{armEmpty}\}$ $\{\mathsf{holding}(x)\}$ $\{\mathsf{onTable}(x), \mathsf{armEmpty}\}$

Recall: an actions a represented by a tuple $\langle \mathsf{pre}_a, \mathsf{add}_a, \mathsf{del}_a \rangle$ of lists of facts.

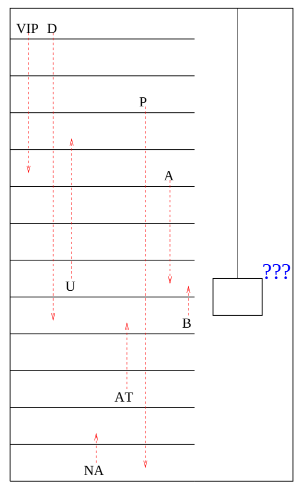
- **Hint:** The only differences between them are the delete lists
- **Answer:** reserved for the plenary sessions $\leadsto$ be there!

The next example for a planning problem is not obvious at first sight, but has been quite influential, showing that many industry problems can be specified declaratively by formalizing the domain and the particular planning problems in PDDL and then using off-the-shelf planners to solve them. [KS00] reports that this has significantly reduced labor costs and increased maintainability of the implementation.

Miconic-10: A Real-World Example

▷ **Example 17.4.9.** Elevator control as a planning problem; details at [KS00]
Specify mobility needs before boarding, let a planner schedule/otimize trips

- ▷ VIP: Served first.
- ▷ D: Lift may only go *down* when inside; similar for U.
- ▷ NA: Never-alone
- ▷ AT: Attendant.
- ▷ A, B: Never together in the same elevator
- ▷ P: Normal passenger



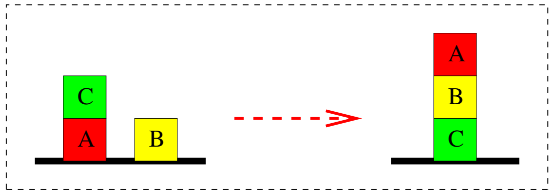
17.5 Partial Order Planning

In this section we introduce a new and different planning algorithm: partial order planning that works on several subgoals independently without having to specify in which order they will be pursued and later combines them into a global plan. **A Video Nugget** covering this section can be found at `https://fau.tv/clip/id/28843`.

To fortify our intuitions about partial order planning let us have another look at the Sussman anomaly, where pursuing two subgoals independently and then reconciling them is a prerequisite.

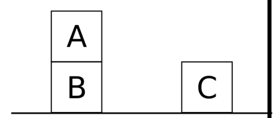
Planning History, p.s.: Planning is Non-Trivial!

▷ **Example 17.5.1.** The Sussman anomaly is a simple blocksworld planning problem:

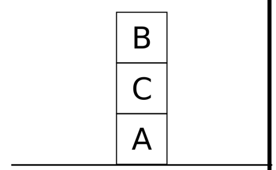


Simple planners that split the goal into subgoals on(A, B) and on(B, C) fail:

- If we pursue on(A, B) by unstacking C, and moving A onto B, we achieve the first subgoal, but cannot achieve the second without undoing the first.



- If we pursue on(B, C) by moving B onto C, we achieve the second subgoal, but cannot achieve the first without undoing the second.



Before we go into the details, let us try to understand the main ideas of partial order planning.

Partial Order Planning

- **Definition 17.5.2.** Any algorithm that can place two actions into a plan without specifying which comes first is called as partial order planning.
- **Ideas** for partial order planning:
  - Organize the planning steps in a DAG that supports multiple paths from initial to goal state
    - nodes (steps) are labeled with actions (actions can occur multiply)
    - edges with propositions added by source and presupposed by target

    acyclicity of the graph induces a partial ordering on steps. q
  - additional temporal constraints resolve subgoal interactions and induce a linear order.
- **Advantages** of partial order planning:
  - problems can be decomposed $\rightsquigarrow$ can work well with non-cooperative environments.
  - efficient by least-commitment strategy
  - causal links (edges) pinpoint unworkable subplans early.

We now make the ideas discussed above concrete by giving a mathematical formulation. It is advantageous to cast a partially ordered plan as a labeled DAG rather than a partial ordering since it draws the attention to the difference between actions and steps.

Partially Ordered Plans

- **Definition 17.5.3.** Let $\langle P, A, I, G\rangle$ be a STRIPS task, then a partially ordered plan $\mathcal{P} = \langle V, E\rangle$ is a labeled DAG, where the nodes in V (called steps) are labeled with actions from A, or are a
  - start step, which has label effect I, or a
  - finish step, which has label precondition G.

  Every edge $(S,T) \in E$ is either labeled by:
  - A non-empty set $p \subseteq P$ of facts that are effects of the action of S and the preconditions of that of T. We call such a labeled edge a causal link and write it $S \xrightarrow{p} T$.
  - $\prec$, then call it a temporal constraint and write it as $S \prec T$.

  An open condition is a precondition of a step not yet causally linked.
- **Definition 17.5.4.** Let be a partially ordered plan Π, then we call a step U possibly intervening in a causal link $S \xrightarrow{p} T$, iff $\Pi \cup \{S \prec U, U \prec T\}$ is acyclic.
- **Definition 17.5.5.** A precondition is achieved iff it is the effect of an earlier step and no possibly intervening step undoes it.
- **Definition 17.5.6.** A partially ordered plan Π is called complete iff every precondition is achieved.
- **Definition 17.5.7.** Partial order planning is the process of computing complete and acyclic partially ordered plans for a given planning task.

A Notation for STRIPS Actions

- **Notation:** Write STRIPS actions into boxes with preconditions above and effects below.
- **Example 17.5.8.**
  - Actions: $Buy(x)$
  - Preconditions: $At(p)$, $Sells(p, x)$
  - Effects: $Have(x)$

  $$\begin{array}{c} At(p)\ Sells(p,x) \\ \boxed{Buy(x)} \\ Have(x) \end{array}$$
- **Notation:** A causal link $S \xrightarrow{p} T$ can also be denoted by a direct arrow between the effects p of S and the preconditions p of T in the STRIPS action notation above.

  Show temporal constraints as dashed arrows.

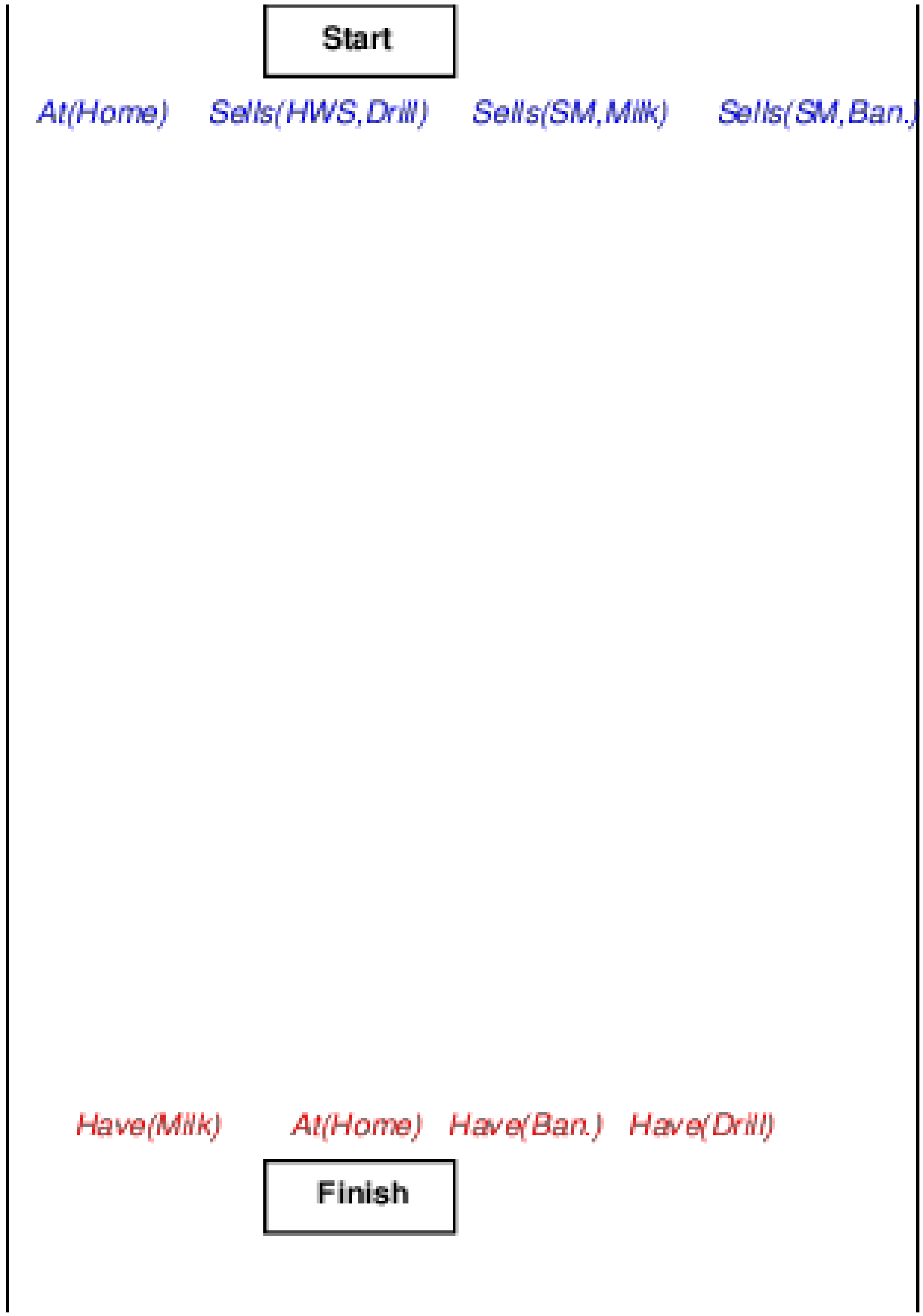
Planning Process

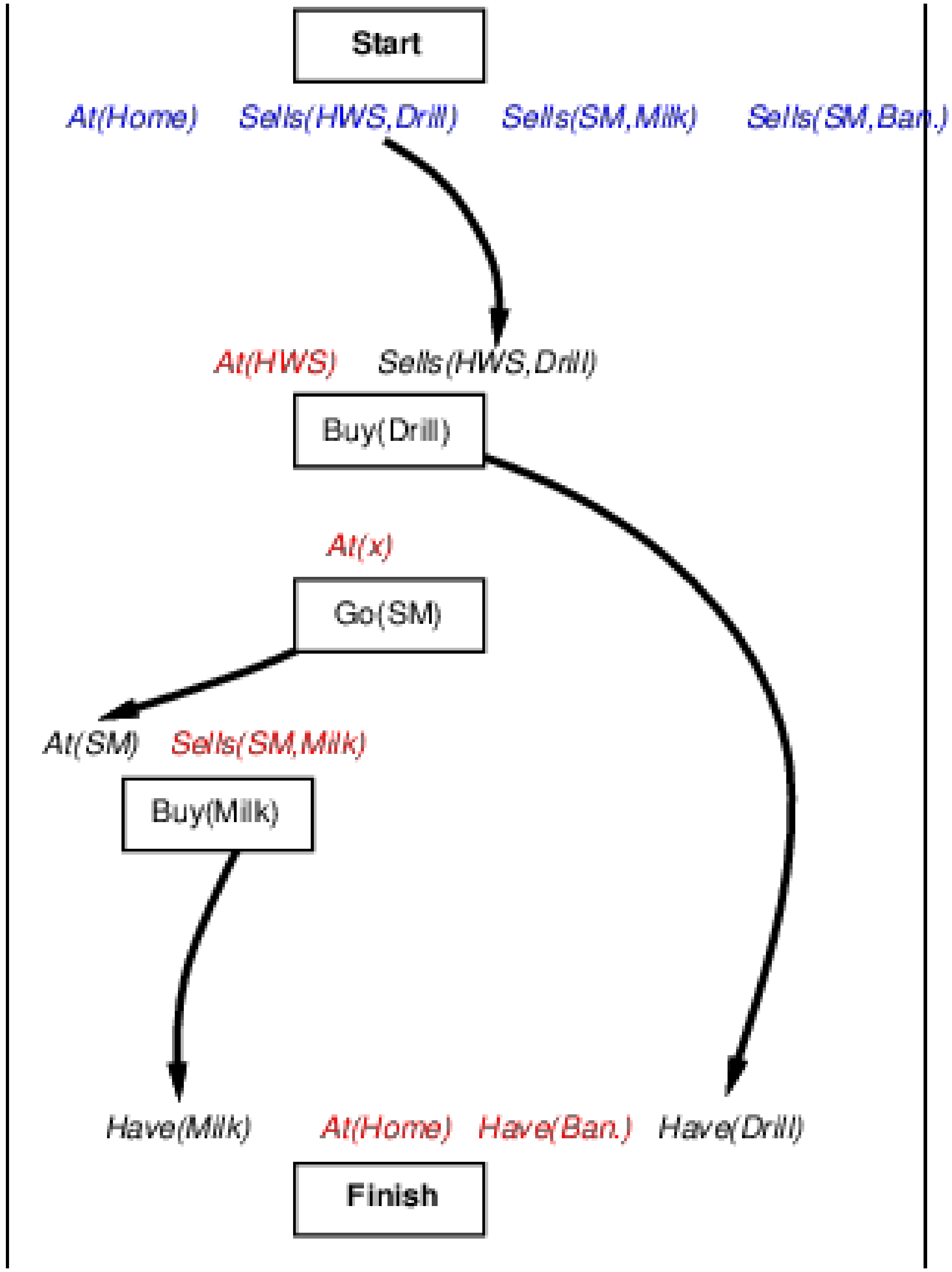
▷ **Definition 17.5.9.** Partial order planning is search in the space of partial plans via the following operations:

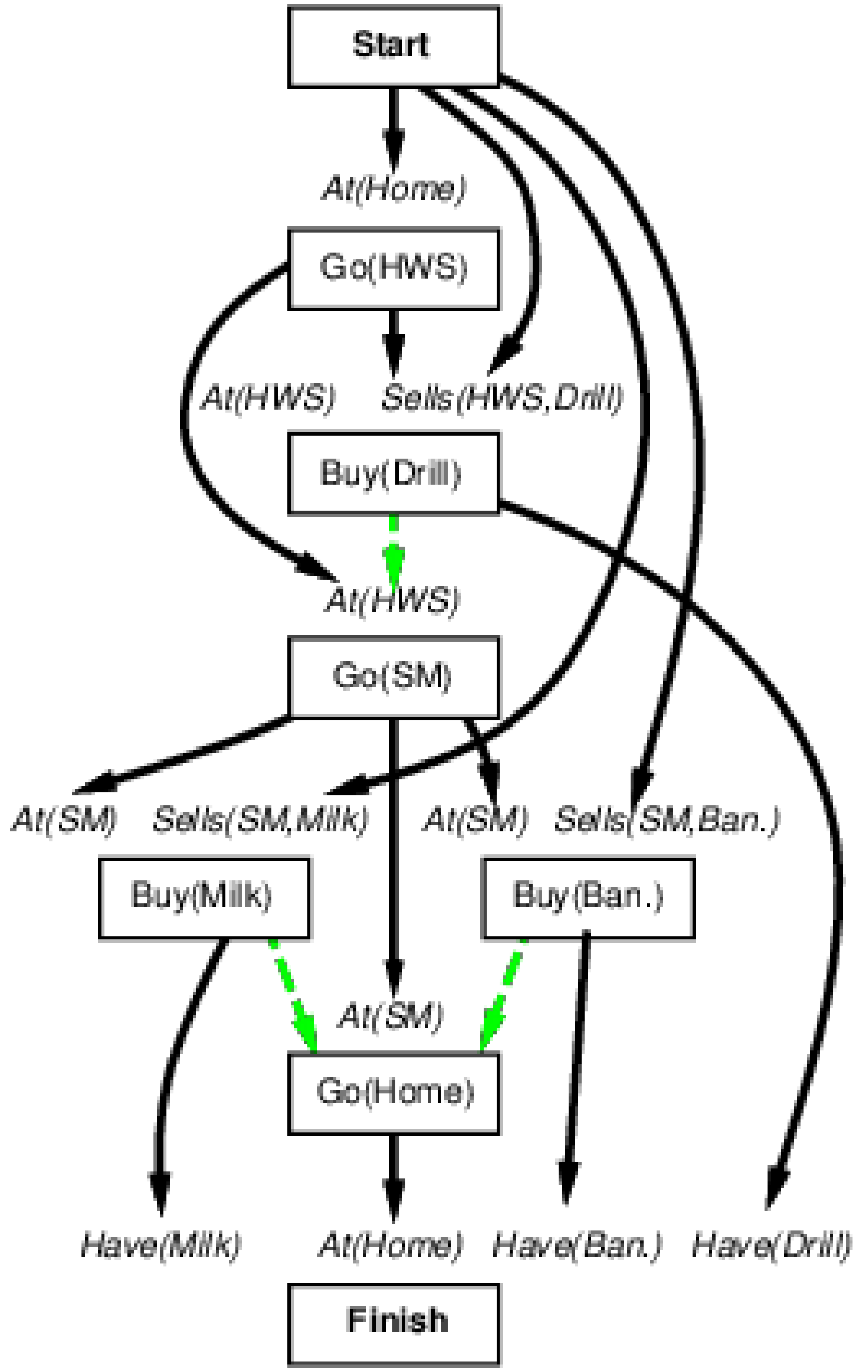
- ▷ add link from an existing action to an open precondition,
- ▷ add step (an action with links to other steps) to fulfil an open condition,
- ▷ order one step wrt. another to remove possible conflicts.

▷ **Idea:** Gradually move from incomplete/vague plans to complete, correct plans. Backtrack if an open condition is unachievable or if a conflict is unresolvable.

Example: Shopping for Bananas, Milk, and a Cordless Drill


Start
At(Home)
Sells(HWS,Drill)
Sells(SM,Milk)
Sells(SM,Ban.)
Have(Milk)
At(Home)
Have(Ban.)
Have(Drill)
Finish


Start
At(Home)
Sells(HWS,Drill)
Sells(SM,Milk)
Sells(SM,Ban.)
At(HWS)
Sells(HWS,Drill)
Buy(Drill)
At(x)
Go(SM)
At(SM)
Sells(SM,Milk)
Buy(Milk)
Have(Milk)
At(Home)
Have(Ban.)
Have(Drill)
Finish



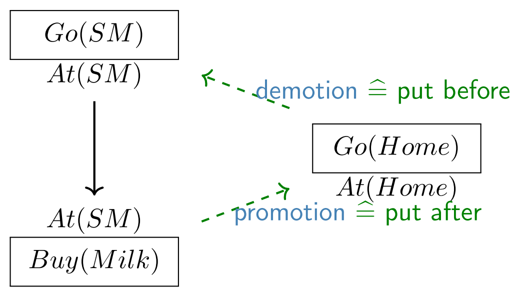
Clobbering and Promotion/Demotion

▷ **Definition 17.5.10.** In a partially ordered plan, a step C clobbers a causal link $L{:=}S\xrightarrow{p}T$, iff it destroys the condition p achieved by L.

▷ **Definition 17.5.11.** If C clobbers $S \xrightarrow{p} T$ in a partially ordered plan Π, then we can solve the induced conflict by

- ▷ demotion: add a temporal constraint $C \prec S$ to Π, or
- ▷ promotion: add $T \prec C$ to Π.

▷ **Example 17.5.12.** $Go(Home)$ clobbers $At(Supermarket)$:



POP algorithm sketch

▷ **Definition 17.5.13.** The POP algorithm for constructing complete partially ordered plans:

function POP (initial, goal, operators) : plan
    plan:= Make—Minimal—Plan(initial, goal)
    loop do
        if Solution?(plan) **then return** plan
        S_{need}, c := Select—Subgoal(plan)
        Choose—Operator(plan, operators, S_{need},c)
        Resolve—Threats(plan)
    end

function Select—Subgoal (plan, S_{need}, c)
    pick a plan step S_{need} from Steps(plan)
        with a precondition c that has not been achieved
    return S_{need}, c

POP algorithm contd.

▷ **Definition 17.5.14.** The missing parts for the POP algorithm.

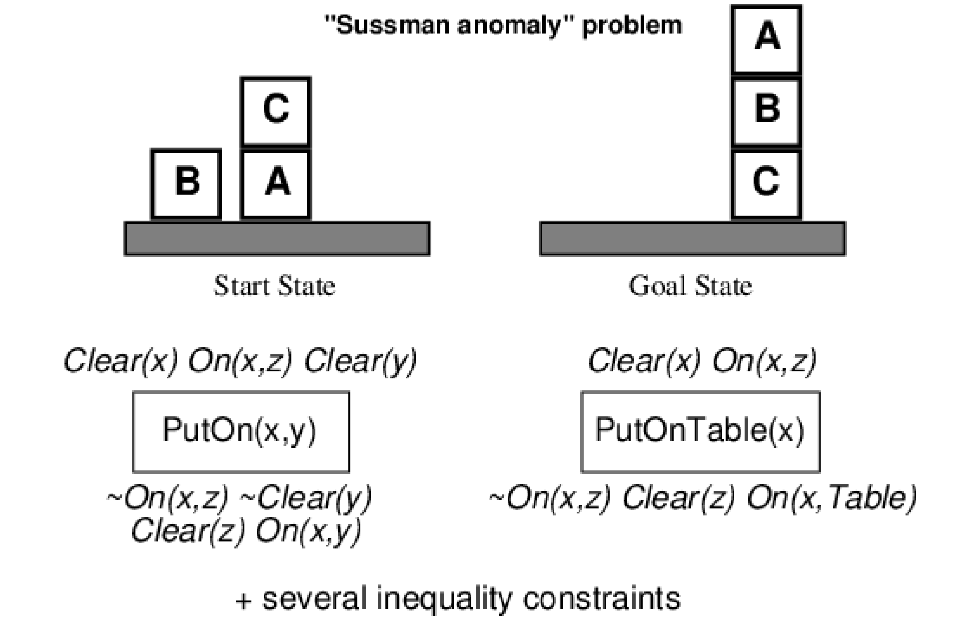
function Choose—Operator (plan, operators, S_{need}, c)
    choose a step S_{add} from operators or Steps(plan) that has c as an effect
    if there is no such step **then fail**
    add the ausal—link $S_{add} \xrightarrow{c} S_{need}$ **to** Links(plan)

```
    add the temporal—constraint S_add ≺ S_need to Orderings(plan)
    if S_add is a newly added \step from operators then
        add S_add to Steps(plan)
        add Start ≺ S_add ≺ Finish to Orderings(plan)

function Resolve—Threats (plan)
    for each S_threat that threatens a causal—link S_i —c→ S_j in Links(plan) do
        choose either
            demotion: Add S_threat ≺ S_i to Orderings(plan)
            promotion: Add S_j ≺ S_threat to Orderings(plan)
        if not Consistent(plan) then fail
```


Properties of POP

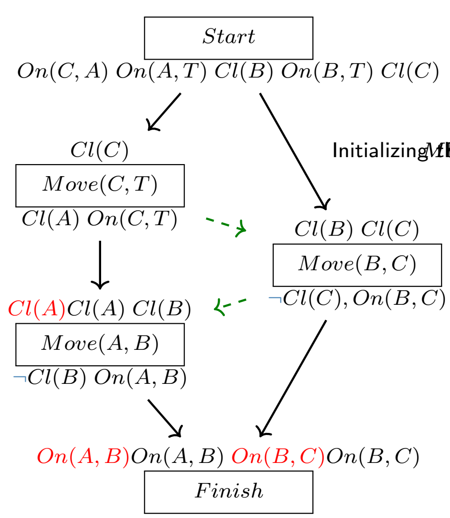
- Nondeterministic algorithm: backtracks at choice points on failure:
  - choice of S_{add} to achieve S_{need},
  - choice of demotion or promotion for clobberer,
  - selection of S_{need} is irrevocable.
- **Observation 17.5.15.** *POP is sound, complete, and systematic i.e. no repetition*
- There are extensions for disjunction, universals, negation, conditionals.
- It can be made efficient with good heuristics derived from problem description.
- Particularly good for problems with many loosely related subgoals.

Example: Solving the Sussman Anomaly



Example: Solving the Sussman Anomaly (contd.)

▷ **Example 17.5.16.** Solving the Sussman anomaly



17.6 The PDDL Language

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26897`.

PDDL: Planning Domain Description Language

▷ **Definition 17.6.1.** The Planning Domain Description Language (PDDL) is a standardized representation language for planning benchmarks in various extensions of the STRIPS formalism.

▷ **Definition 17.6.2.** PDDL is not a propositional language

- ▷ Representation is lifted, using object variables to be instantiated from a finite set of objects. (Similar to predicate logic)
- ▷ Action schemas parameterized by objects.
- ▷ Predicates to be instantiated with objects.

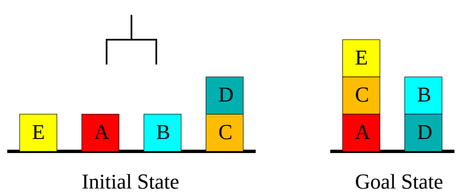
▷ **Definition 17.6.3.** A PDDL planning task comes in two pieces

- ▷ The problem file gives the objects, the initial state, and the goal state.
- ▷ The domain file gives the predicates and the actions.

History and Versions:

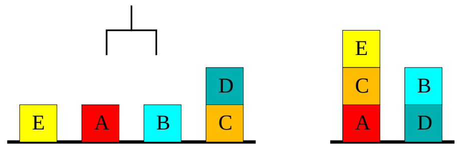
- Used in the International Planning Competition (IPC).
- 1998: PDDL [McD+98].
- 2000: "PDDL subset for the 2000 competition" [Bac00].
- 2002: PDDL2.1, Levels 1-3 [FL03].
- 2004: PDDL2.2 [HE05].
- 2006: PDDL3 [Ger+09].

The Blocksworld in PDDL: Domain File




```
(define (domain blocksworld)
  (:predicates (clear ?x) (holding ?x) (on ?x ?y)
               (on−table ?x) (arm−empty))
  (:action stack
   :parameters (?x ?y)
   :precondition (and (clear ?y) (holding ?x))
   :effect (and (arm−empty) (on ?x ?y)
                (not (clear ?y)) (not (holding ?x))))
  ...)
```

The Blocksworld in PDDL: Problem File



Initial State Goal State

```
(define (problem bw—abcde)
  (:domain blocksworld)
  (:objects a b c d e)
  (:init (on—table a) (clear a)
    (on—table b) (clear b)
    (on—table e) (clear e)
    (on—table c) (on d c) (clear d)
    (arm—empty))
  (:goal (and (on e c) (on c a) (on b d))))
```


Miconic-ADL "Stop" Action Schema in PDDL

```
(:action stop
 :parameters (?f — floor)
 :precondition (and (lift—at ?f)
  (imply
   (exists
    (?p — conflict—A)
    (or (and (not (served ?p))
               (origin ?p ?f))
          (and (boarded ?p)
               (not (destin ?p ?f)))))
   (forall
    (?q — conflict—B)
    (and (or (destin ?q ?f)
               (not (boarded ?q)))
          (or (served ?q)
               (not (origin ?q ?f))))))
  (imply (exists
           (?p — conflict—B)
           (or (and (not (served ?p))
                      (origin ?p ?f))
                 (and (boarded ?p)
                      (not (destin ?p ?f)))))
          (forall
           (?q — conflict—A)
           (and (or (destin ?q ?f)
                      (not (boarded ?q)))
                 (or (served ?q)
                      (not (origin ?q ?f))))))
  (imply
   (exists
    (?p — never—alone)
    (or (and (origin ?p ?f)
               (not (served ?p)))
          (and (boarded ?p)
               (not (destin ?p ?f)))))
   (exists
    (?q — attendant)
    (or (and (boarded ?q)
               (not (destin ?q ?f)))
          (and (not (served ?q))
               (origin ?q ?f)))))
  (forall
   (?p — going—nonstop)
   (imply (boarded ?p) (destin ?p ?f)))
  (or (forall
       (?p — vip) (served ?p))
      (exists
       (?p — vip)
       (or (origin ?p ?f) (destin ?p ?f))))
  (forall
   (?p — passenger)
   (imply
    (no—access ?p ?f) (not (boarded ?p)))))
 )
```


Planning Domain Description Language

▷ **Question:** What is PDDL good for?

(A) Nothing.

(B) Free beer.

(C) Those AI planning guys.

(D) Being lazy at work.

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

17.7 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26900`.

Summary

▷ General problem solving attempts to develop solvers that perform well across a large class of problems.

▷ Planning, as considered here, is a form of general problem solving dedicated to the class of classical search problems. (Actually, we also address inaccessible, stochastic, dynamic, continuous, and multi-agent settings.)

▷ Heuristic search planning has dominated the International Planning Competition (IPC). We focus on it here.

▷ STRIPS is the simplest possible, while reasonably expressive, language for our purposes. It uses Boolean variables (facts), and defines actions in terms of precondition, add list, and delete list.

▷ PDDL is the de-facto standard language for describing planning problems.

▷ Plan existence (bounded or not) is **PSPACE**-complete to decide for STRIPS. If we bound plans polynomially, we get down to **NP**-completeness.

Suggested Reading:

- Chapters 10: *Classical Planning* and 11: *Planning and Acting in the Real World* in [RN09].
  - Although the book is named "*A Modern Approach*", the planning section was written long before the IPC was even dreamt of, before PDDL was conceived, and several years before heuristic search hit the scene. As such, what we have right now is the attempt of two outsiders trying in vain to catch up with the dramatic changes in planning since 1995.
  - Chapter 10 is Ok as a background read. Some issues are, imho, misrepresented, and it's far from being an up-to-date account. But it's Ok to get some additional intuitions in words different from my own.
  - Chapter 11 is useful in our context here because we don't cover any of it. If you're interested in extended/alternative planning paradigms, do read it.
- A good source for modern information (some of which we covered in the lecture) is Jörg Hoffmann's *Everything You Always Wanted to Know About Planning (But Were Afraid to Ask)* [Hof11] which is available online at `http://fai.cs.uni-saarland.de/hoffmann/papers/ki11.pdf`

Chapter 18

Planning II: Algorithms

18.1 Introduction

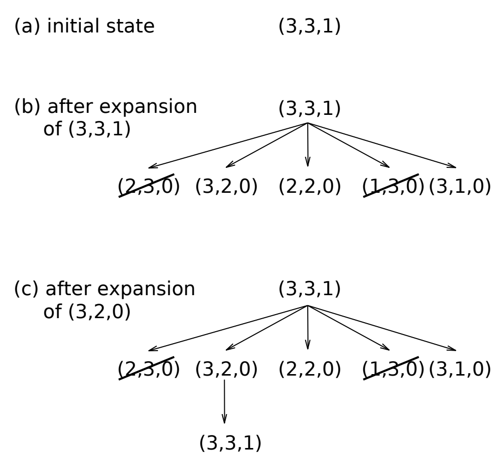
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26901`.

Reminder: Our Agenda for This Topic

- ▷ **chapter 17**: Background, planning languages, complexity.
  - ▹ Sets up the framework. computational complexity is essential to distinguish different algorithmic problems, and for the design of heuristic functions.
- ▷ **This Chapter:** How to automatically generate a heuristic function, given planning language input?
  - ▹ Focussing on heuristic search as the solution method, this is the main question that needs to be answered.

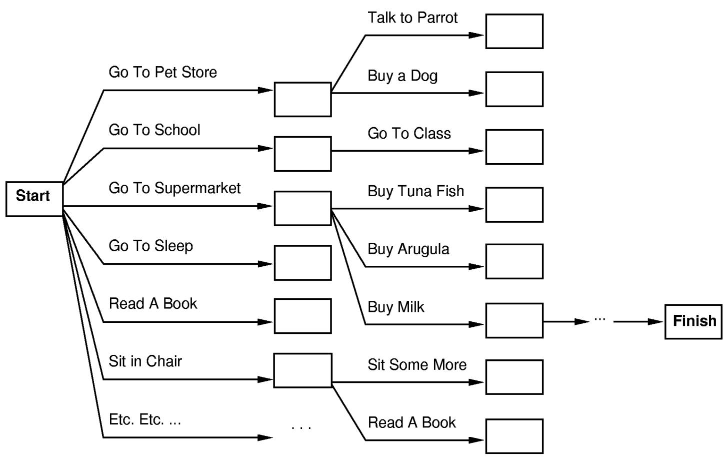
Reminder: Search

- ▷ Starting at initial state, produce all successor states step by step:



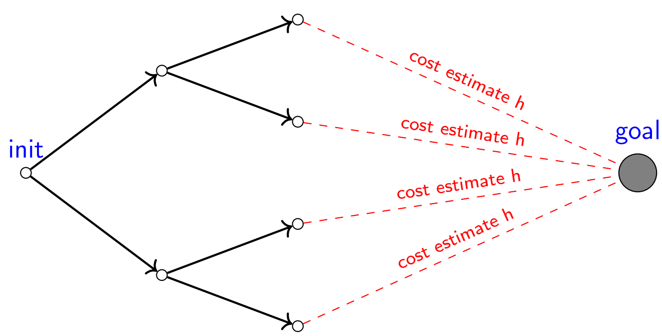
In planning, this is referred to as forward search, or forward state-space search.

Search in the State Space?



▷ Use heuristic function to guide the search towards the goal!

Reminder: Informed Search



▷ Heuristic function h estimates the cost of an optimal path from a state s to the goal; search prefers to expand states s with small $h(s)$.

▷ Live Demo vs. Breadth-First Search:

http://qiao.github.io/PathFinding.js/visual/

Reminder: Heuristic Functions

- **Definition 18.1.1.** Let Π be a STRIPS task with states S. A heuristic function, short heuristic, for Π is a function $h\colon S \to \mathbb{N} \cup \{\infty\}$ so that $h(s) = 0$ whenever s is a goal state.
- Exactly like our definition from chapter 6. Except, because we assume unit costs here, we use $\mathbb{N}$ instead of $\mathbb{R}^+$.
- **Definition 18.1.2.** Let Π be a STRIPS task with states S. The perfect heuristic h^* assigns every $s \in S$ the length of a shortest path from s to a goal state, or ∞ if no such path exists. A heuristic function h for Π is admissible if, for all $s \in S$, we have $h(s) \leq h^*(s)$.
- Exactly like our definition from chapter 6, except for path *length* instead of path *cost* (cf. above).
- In all cases, we attempt to approximate $h^*(s)$, the length of an optimal plan for s. Some algorithms guarantee to lower bound $h^*(s)$.

Our (Refined) Agenda for This Chapter

- **How to Relax:** How to relax a problem?
  - Basic principle for generating heuristic functions.
- **The Delete Relaxation:** How to relax a planning problem?
  - The delete relaxation is the most successful method for the *automatic* generation of heuristic functions. It is a key ingredient to almost all IPC winners of the last decade. It relaxes STRIPS tasks by ignoring the delete lists.
- **The h^+ Heuristic:** What is the resulting heuristic function?
  - h^+ is the "ideal" delete relaxation heuristic.
- **Approximating h^+:** How to actually compute a heuristic?
  - Turns out that, in practice, we must approximate h^+.

18.2 How to Relax in Planning

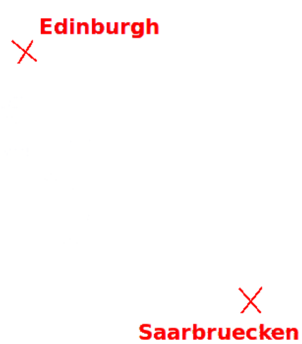
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26902`.

We will now instantiate our general knowledge about heuristic search to the planning domain. As always, the main problem is to find good heuristics. We will follow the intuitions of our discussion in subsection 6.5.4 and consider full solutions to relaxed problems as a source for heuristics.

Reminder: Heuristic Functions from Relaxed Problems

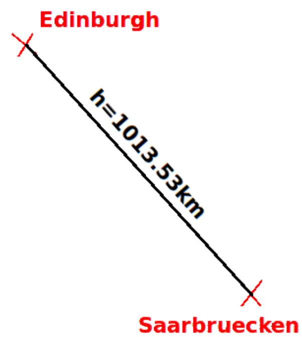
▷ Problem Π: Find a route from Saarbrücken to Edinburgh.

Reminder: Heuristic Functions from Relaxed Problems



▷ Relaxed Problem Π′: Throw away the map.

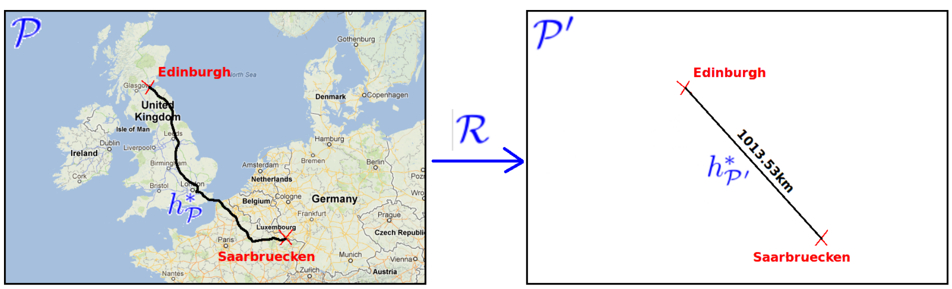
Reminder: Heuristic Functions from Relaxed Problems



▷ Heuristic function h: Straight line distance.

Michael Kohlhase: Artificial Intelligence 1 628 2023-09-20

Relaxation in Route-Finding



▷ **Problem class** $\mathcal{P}$: Route finding.

▷ **Perfect heuristic** $h^*_{\mathcal{P}}$ **for** $\mathcal{P}$: Length of a shortest route.

▷ **Simpler problem class** $\mathcal{P}'$: Route finding on an empty map.

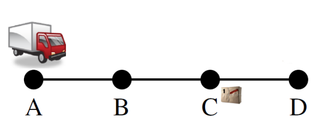
▷ **Perfect heuristic** $h^*_{\mathcal{P}'}$ **for** $\mathcal{P}'$: Straight-line distance.

▷ **Transformation** $\mathcal{R}$: Throw away the map.

Michael Kohlhase: Artificial Intelligence 1 629 2023-09-20

How to Relax in Planning? (A Reminder!)

▷ **Example 18.2.1 (Logistics).**



- facts P: $\{\text{truck}(x)|x\in\{A,B,C,D\}\}\cup\{\text{pack}(x)|x\in\{A,B,C,D,T\}\}$.
- initial state I: $\{\text{truck}(A),\text{pack}(C)\}$.
- goal G: $\{\text{truck}(A),\text{pack}(D)\}$.
- actions A: (Notated as "precondition $\Rightarrow$ adds, $\neg$ deletes")
  - drive(x,y), where x and y have a road: "truck(x) $\Rightarrow$ truck(y), $\neg$truck(x)".
  - load(x): "truck(x), pack(x) $\Rightarrow$ pack(T), $\neg$pack(x)".
  - unload(x): "truck(x), pack(T) $\Rightarrow$ pack(x), $\neg$pack(T)".

▷ **Example 18.2.2 ("Only-Adds" Relaxation).** Drop the preconditions and deletes.

- "drive(x,y): $\Rightarrow$ truck(y)";
- "load(x): $\Rightarrow$ pack(T)";
- "unload(x): $\Rightarrow$ pack(x)".

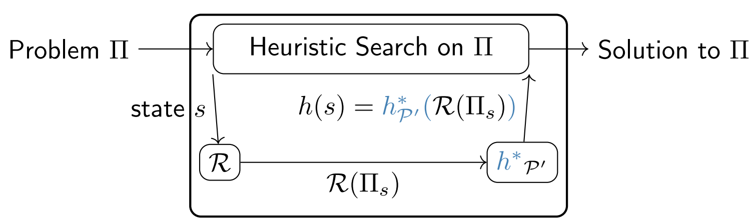
▷ Heuristic value for I is?

▷ $h^{\mathcal{R}}(I)=1$: A plan for the relaxed task is $\langle\text{unload}(D)\rangle$.

We will start with a very simple relaxation, which could be termed "positive thinking": we do not consider preconditions of actions and leave out the delete lists as well.

How to Relax During Search: Overview

▷ **Attention:** Search uses the real (un-relaxed) Π. The relaxation is applied (e.g., in Only-Adds, the simplified actions are used) **only within the call to** $h(s)$**!!!**



- Here, Π_s is Π with initial state replaced by s, i.e., $\Pi{:=}\langle P,A,I,G\rangle$ changed to $\Pi^s{:=}\langle P,A,\{s\},G\rangle$: The task of finding a plan for search state s.
- A common student mistake is to instead apply the relaxation once to the whole problem, then doing the whole search "within the relaxation".
- The next slide illustrates the correct search process in detail.

How to Relax During Search: Only-Adds

Real problem:

- ▷ Initial state I: AC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $drXY, loX, ulX$.

Relaxed problem:

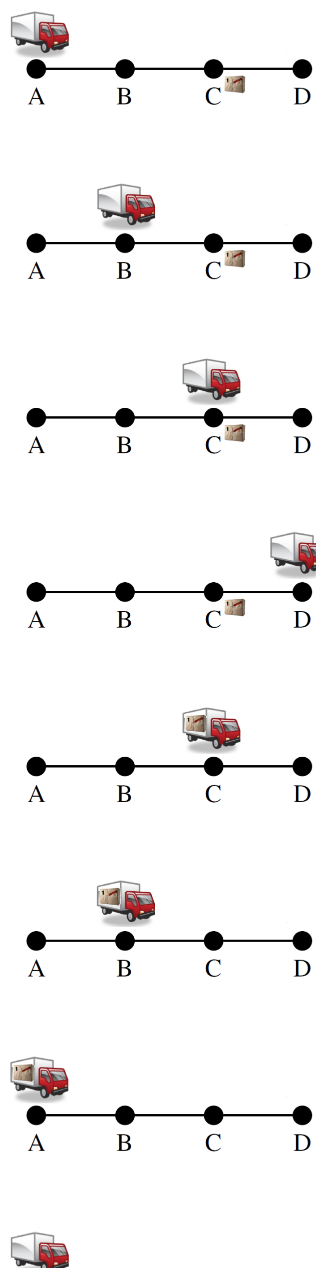
- ▷ State s: AC; goal G: AD.
- ▷ Actions A: add.
- ▷ $h^{\mathcal{R}}(s)$ =1: $\langle ulD \rangle$.

Real problem:

- ▷ State s: BC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $AC \xrightarrow{drAB} BC$.

Relaxed problem:

- ▷ State s: BC; goal G: AD.
- ▷ Actions A: add.
- ▷ $h^{\mathcal{R}}(s)$ =2: $\langle drBA, ulD \rangle$.



Real problem:

- ▷ State s: CC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $BC \xrightarrow{drBC} CC$.

Relaxed problem:

- ▷ State s: CC; goal G: AD.
- ▷ Actions A: add.
- ▷ $h^{\mathcal{R}}(s)$ =2: $\langle drBA, ulD \rangle$.

Real problem:

- ▷ State s: AC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $BC \xrightarrow{drBA} AC$.

Real problem:

- ▷ State s: AC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ Duplicate state, prune.

Real problem:

- ▷ State s: DC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $CC \xrightarrow{drCD} DC$.

Relaxed problem:

- ▷ State s: DC; goal G: AD.
- ▷ Actions A: add.
- ▷ $h^{\mathcal{R}}(s)$ =2: $\langle drBA, ulD \rangle$.

Real problem:

- ▷ State s: CT; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $CC \xrightarrow{loC} CT$.

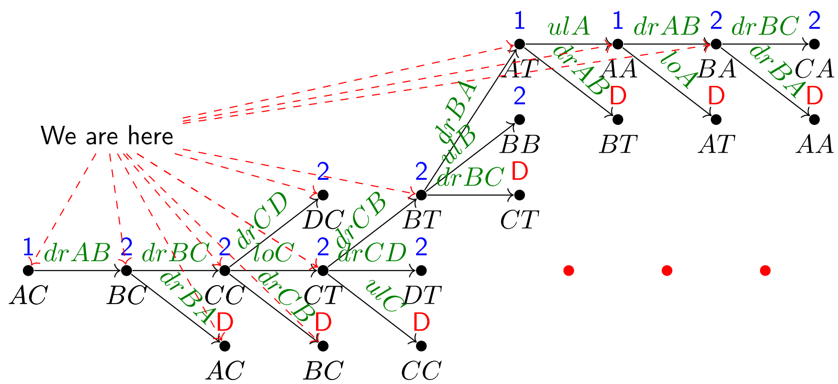
Relaxed problem:

- ▷ State s: CT; goal G: AD.
- ▷ Actions A: add.
- ▷ $h^{\mathcal{R}}(s)$ =2: $\langle drBA, ulD \rangle$.

Real problem:

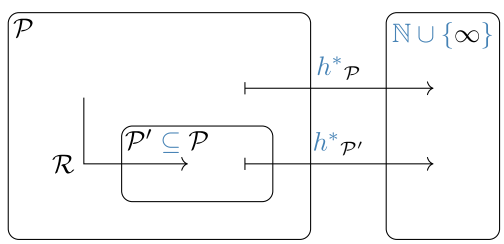
- ▷ State s: BC; goal G: AD.
- ▷ Actions A: pre, add, del.
- ▷ $CC \xrightarrow{drCB} BC$.

▷ **Greedy best-first search:** (tie-breaking: alphabetic)



Only-Adds is a "Native" Relaxation

▷ **Definition 18.2.3 (Native Relaxations).** Confusing special case where $\mathcal{P}' \subseteq \mathcal{P}$.



- ▷ **Problem class $\mathcal{P}$:** STRIPS tasks.
- ▷ **Perfect heuristic $h^*{}_{\mathcal{P}}$ for $\mathcal{P}$:** Length h^* of a shortest plan.
- ▷ **Transformation $\mathcal{R}$:** Drop the preconditions and delete lists.
- ▷ **Simpler problem class $\mathcal{P}'$** is a special case of $\mathcal{P}$, $\mathcal{P}' \subseteq \mathcal{P}$: STRIPS tasks with empty preconditions and delete lists.
- ▷ Perfect heuristic for $\mathcal{P}'$: Shortest plan for only-adds STRIPS task.

18.3 The Delete Relaxation

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26903`.

We turn to a more realistic relaxation, where we only disregard the delete list.

How the Delete Relaxation Changes the World

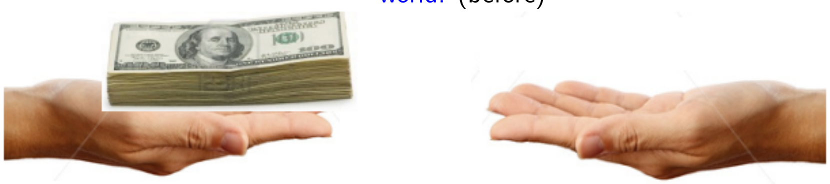
▷ Relaxation mapping $\mathcal{R}$ saying that:

"When the world changes, its previous state remains true as well."

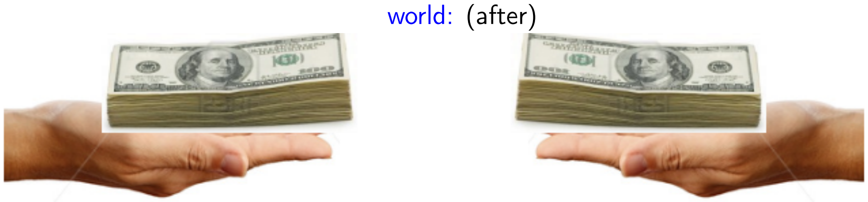
Real world: (before)

Real world: (after)

Relaxed world: (before)



Relaxed world: (after)



Real world: (before)

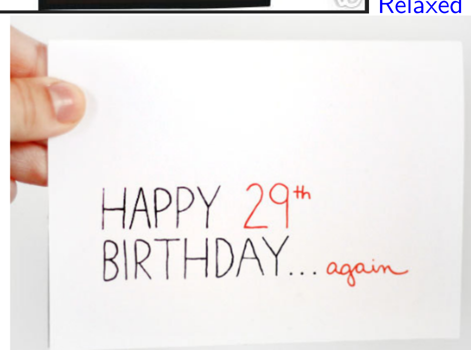
Real world: (after)

Relaxed world: (before)

Relaxed world: (after)

Real world:

Relaxed world:



The Delete Relaxation

▷ **Definition 18.3.1 (Delete Relaxation).** Let $\Pi{:=}\langle P, A, I, G\rangle$ be a STRIPS task. The delete relaxation of Π is the task $\Pi^+ = \langle P, A^+, I, G\rangle$ where $A^+{:=}\{a^+ | a{\in}A\}$ with $\mathrm{pre}_{a^+}{:=}\mathrm{pre}_a$, $\mathrm{add}_{a^+}{:=}\mathrm{add}_a$, and $\mathrm{del}_{a^+}{:=}\emptyset$.

▷ In other words, the class of simpler problems $\mathcal{P}'$ is the set of all STRIPS tasks with empty delete lists, and the relaxation mapping $\mathcal{R}$ drops the delete lists.

▷ **Definition 18.3.2 (Relaxed Plan).** Let $\Pi := \langle P, A, I, G \rangle$ be a STRIPS task, and let s be a state. A relaxed plan for s is a plan for $\langle P, A, s, G \rangle^+$. A relaxed plan for I is called a relaxed plan for Π.

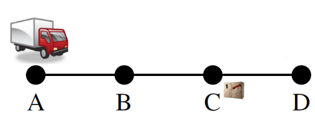
▷ A relaxed plan for s is an action sequence that solves s when pretending that all delete lists are empty.

▷ Also called "delete-relaxed plan": "relaxation" is often used to mean "delete-relaxation" by default.

A Relaxed Plan for "TSP" in Australia

1. **Initial state:** $\{\text{at}(\text{Sy}), \text{vis}(\text{Sy})\}$.
2. $\textbf{drv}(\textbf{Sy}, \textbf{Br})^+$: $\{\text{at}(\text{Br}), \text{vis}(\text{Br}), \text{at}(\text{Sy}), \text{vis}(\text{Sy})\}$.
3. $\textbf{drv}(\textbf{Sy}, \textbf{Ad})^+$: $\{\text{at}(\text{Ad}), \text{vis}(\text{Ad}), \text{at}(\text{Br}), \text{vis}(\text{Br}), \text{at}(\text{Sy}), \text{vis}(\text{Sy})\}$.
4. $\textbf{drv}(\textbf{Ad}, \textbf{Pe})^+$: $\{\text{at}(\text{Pe}), \text{vis}(\text{Pe}), \text{at}(\text{Ad}), \text{vis}(\text{Ad}), \text{at}(\text{Br}), \text{vis}(\text{Br}), \text{at}(\text{Sy}), \text{vis}(\text{Sy})\}$.
5. $\textbf{drv}(\textbf{Ad}, \textbf{Da})^+$: $\{\text{at}(\text{Da}), \text{vis}(\text{Da}), \text{at}(\text{Pe}), \text{vis}(\text{Pe}), \text{at}(\text{Ad}), \text{vis}(\text{Ad}), \text{at}(\text{Br}), \text{vis}(\text{Br}), \text{at}(\text{Sy}), \text{vis}(\text{Sy})\}$.

A Relaxed Plan for "Logistics"



▷ Facts P: $\{\text{truck}(x) | x \in \{A, B, C, D\}\} \cup \{\text{pack}(x) | x \in \{A, B, C, D, T\}\}$.

▷ Initial state I: $\{\text{truck}(A), \text{pack}(C)\}$.

▷ Goal G: $\{\text{truck}(A), \text{pack}(D)\}$.

▷ Relaxed actions A^+: (Notated as "precondition $\Rightarrow$ adds")

- ▹ $\text{drive}(x, y)^+$: "$\text{truck}(x) \Rightarrow \text{truck}(y)$".

- $\mathsf{load}(x)^+$: "$\mathsf{truck}(x), \mathsf{pack}(x) \Rightarrow \mathsf{pack}(T)$".
- $\mathsf{unload}(x)^+$: "$\mathsf{truck}(x), \mathsf{pack}(T) \Rightarrow \mathsf{pack}(x)$".

Relaxed plan:

$$\langle \mathsf{drive}(A,B)^+, \mathsf{drive}(B,C)^+, \mathsf{load}(C)^+, \mathsf{drive}(C,D)^+, \mathsf{unload}(D)^+ \rangle$$

▷ We don't need to drive the truck back, because "it is still at A".

PlanEx$^+$

▷ **Definition 18.3.3 (Relaxed Plan Existence Problem).** By **PlanEx$^+$**, we denote the problem of deciding, given a STRIPS task $\Pi{:=}\langle P, A, I, G\rangle$, whether or not there exists a relaxed plan for Π.

▷ This is easier than **PlanEx** for general STRIPS!

▷ **PlanEx$^+$** is in **P**.

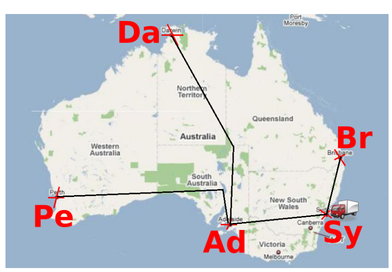
▷ *Proof:* The following algorithm decides **PlanEx$^+$**

1.

   var $F := I$
   while $G \not\subseteq F$ **do**
   $\quad F' := F \cup \bigcup_{a \in A: \mathsf{pre}_a \subseteq F} \mathsf{add}_a$
   $\quad$ **if** $F' = F$ **then return** "unsolvable" endif $\qquad (*)$
   $\quad F := F'$
   endwhile
   return "solvable"

2. The algorithm terminates after at most $|P|$ iterations, and thus runs in polynomial time.
3. Correctness: See slide 641

Deciding PlanEx$^+$ in "TSP" in Australia

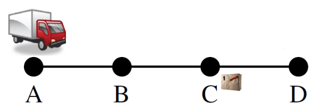


Iterations on F:

1. $\{$at(Sy), vis(Sy)$\}$
2. $\cup$ $\{$at(Ad), vis(Ad), at(Br), vis(Br)$\}$
3. $\cup$ $\{$at(Da), vis(Da), at(Pe), vis(Pe)$\}$

Deciding PlanEx$^+$ in "Logistics"

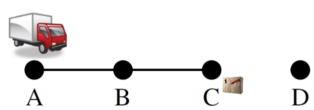
▷ **Example 18.3.4 (The solvable Case).**



Iterations on F:

1. $\{$truck(A), pack$(C)\}$
2. $\cup\{$truck$(B)\}$
3. $\cup\{$truck$(C)\}$
4. $\cup\{$truck(D), pack$(T)\}$
5. $\cup\{$pack(A), pack(B), pack$(D)\}$

▷ **Example 18.3.5 (The unsolvable Case).**



Iterations on F:

1. $\{$truck(A), pack$(C)\}$
2. $\cup\{$truck$(B)\}$
3. $\cup\{$truck$(C)\}$
4. $\cup\{$pack$(T)\}$
5. $\cup\{$pack(A), pack$(B)\}$
6. $\cup\emptyset$

PlanEx$^+$ Algorithm: Proof

Proof: To show: The algorithm returns "solvable" iff there is a relaxed plan for Π.

1. Denote by F_i the content of F after the ith iteration of the while-loop,
2. All $a \in A_0$ are applicable in I, all $a \in A_1$ are applicable in apply(I, A_0^+), and so forth.
3. Thus $F_i = $ apply$(I, \langle A_0^+, \ldots, A_{i-1}^+ \rangle)$. (Within each A_j^+, we can sequence the actions in any order.)
4. Direction "$\Rightarrow$" *If "solvable" is returned after iteration n then $G \subseteq F_n = $ apply$(I, \langle A_0^+, \ldots, A_{n-1}^+ \rangle)$ so $\langle A_0^+, \ldots, A_{n-1}^+ \rangle$ can be sequenced to a relaxed plan which shows the claim.*
5. Direction "$\Leftarrow$"
   5.1. Let $\langle a_0^+, \ldots, a_{n-1}^+ \rangle$ be a relaxed plan, hence $G \subseteq$ apply$(I, \langle a_0^+, \ldots, a_{n-1}^+ \rangle)$.
   5.2. Assume, for the moment, that we drop line (*) from the algorithm. It is then easy to see that $a_i \in A_i$ and apply$(I, \langle a_0^+, \ldots, a_{i-1}^+ \rangle) \subseteq F_i$, for all i.

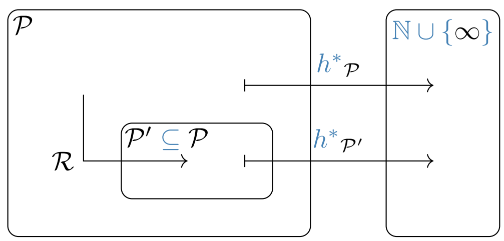
5.3. We get $G \subseteq \text{apply}(I, \langle a_0^+, \ldots, a_{n-1}^+ \rangle) \subseteq F_n$, and the algorithm returns "solvable" as desired.

5.4. Assume to the contrary of the claim that, in an iteration $i < n$, (*) fires. Then $G \not\subseteq F$ and $F = F'$. But, with $F = F'$, $F = F_j$ for all $j > i$, and we get $G \not\subseteq F_n$ in contradiction.

18.4 The h^+Heuristic

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26905`.

Hold on a Sec – Where are we?



- $\mathcal{P}$: STRIPS tasks; $h^*_{\mathcal{P}}$: Length h^* of a shortest plan.
- $\mathcal{P}' \subseteq \mathcal{P}$: STRIPS tasks with empty delete lists.
- $\mathcal{R}$: Drop the delete lists.
- Heuristic function: Length of a shortest *relaxed* plan ($h^* \circ \mathcal{R}$).
- $\mathbf{PlanEx}^+$ is not actually what we're looking for. $\mathbf{PlanEx}^+ \mathrel{\widehat{=}}$ relaxed plan *existence*; we want relaxed plan *length* $h^* \circ \mathcal{R}$.

h^+: The Ideal Delete Relaxation Heuristic

- **Definition 18.4.1 (Optimal Relaxed Plan).** Let $\langle P, A, I, G \rangle$ be a STRIPS task, and let s be a state. A optimal relaxed plan for s is an optimal plan for $\langle P, A, \{s\}, G \rangle^+$.
- Same as slide 635, just adding the word "optimal".
- Here's what we're looking for:
- **Definition 18.4.2.** Let $\Pi := \langle P, A, I, G \rangle$ be a STRIPS task with states S. The ideal delete relaxation heuristic h^+ for Π is the function $h^+ : S \rightarrow \mathbb{N} \cup \{\infty\}$ where $h^+(s)$ is the length of an optimal relaxed plan for s if a relaxed plan for s exists, and $h^+(s) = \infty$ otherwise.
- In other words, $h^+ = h^* \circ \mathcal{R}$, cf. previous slide.

h^+ is Admissible

▷ **Lemma 18.4.3.** *Let $\Pi{:=}\langle P, A, I, G\rangle$ be a STRIPS task, and let s be a state. If $\langle a_1, \ldots, a_n\rangle$ is a plan for $\Pi_s{:=}\langle P, A, \{s\}, G\rangle$, then $\langle a_1^+, \ldots, a_n^+\rangle$ is a plan for Π^+.*

▷ *Proof sketch:* Show by induction over $0 \leq i \leq n$ that $\text{apply}(s, \langle a_1, \ldots, a_i\rangle) \subseteq \text{apply}(s, \langle a_1^+, \ldots, a_i^+\rangle)$.

▷ If we ignore deletes, the states along the plan can only get bigger.

▷ **Theorem 18.4.4.** *h^+ is Admissible.*

▷ *Proof:*

1. Let $\Pi{:=}\langle P, A, I, G\rangle$ be a STRIPS task with states P, and let $s \in P$.
2. $h^+(s)$ is defined as optimal plan length in Π_s^+.
3. With the lemma above, any plan for Π also constitutes a plan for Π_s^+.
4. Thus optimal plan length in Π_s^+ can only be shorter than that in $\Pi_s i$, and the claim follows.

How to Relax During Search: Ignoring Deletes

Real problem:

- Initial state I: AC; goal G: AD.
- Actions A: pre, add, del.
- $drXY, loX, ulX$.

Relaxed problem:

- State s: AC; goal G: AD.
- Actions A: pre, add.
- $h^+(s)$ =5: e.g. $\langle drAB, drBC, drCD, loC, ulD \rangle$.

Real problem:

- State s: BC; goal G: AD.
- Actions A: pre, add, del.
- $AC \xrightarrow{drAB} BC$.

Relaxed problem:

- State s: BC; goal G: AD.
- Actions A: pre, add.
- $h^+(s)$ =5: e.g. $\langle drBA, drBC, drCD, loC, ulD \rangle$.

Real problem:

- State s: CC; goal G: AD.
- Actions A: pre, add, del.
- $BC \xrightarrow{drBC} CC$.

Relaxed problem:

- State s: CC; goal G: AD.
- Actions A: pre, add.
- $h^+(s)$ =5: e.g. $\langle drCB, drBA, drCD, loC, ulD \rangle$.

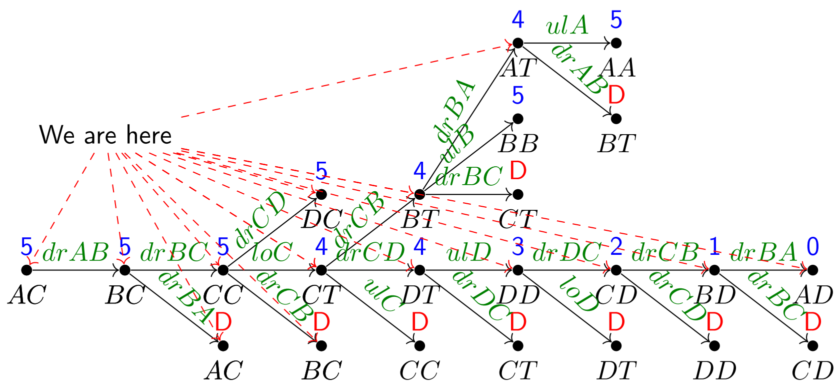
Real problem:

- State s: AC; goal G: AD.
- Actions A: pre, add, del.
- $BC \xrightarrow{drBA} AC$.

Real problem:

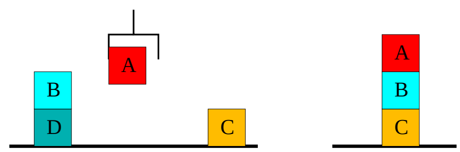
- State s: AC; goal G: AD.

Greedy best-first search: (tie-breaking: alphabetic)



Of course there are also bad cases. Here is one.

h^+ in the Blocksworld



▷ Initial State Goal State

- ▷ **Optimal plan**: ⟨putdown(A), unstack(B, D), stack(B, C), pickup(A), stack(A, B)⟩.
- ▷ **Optimal relaxed plan**: ⟨stack(A, B), unstack(B, D), stack(B, C)⟩.
- ▷ **Observation:** What can we say about the "search space surface" at the initial state here?
- ▷ The initial state lies on a local minimum under h^+, together with the successor state s where we stacked A onto B. All direct other neighbours of these two states have a strictly higher h^+ value.

18.5 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26906`.

Summary

- ▷ Heuristic search on classical search problems relies on a function h mapping states s to an estimate $h(s)$ of their goal distance. Such functions h are derived by solving relaxed problems.
- ▷ In planning, the relaxed problems are generated and solved automatically. There are four known families of suitable relaxation methods: *abstractions*, *landmarks*,

critical paths, and *ignoring deletes* (aka delete relaxation).

▷ The delete relaxation consists in dropping the deletes from STRIPS tasks. A relaxed plan is a plan for such a relaxed task. $h^+(s)$ is the length of an optimal relaxed plan for state s. h^+ is **NP**-hard to compute.

▷ h^{FF} approximates h^+ by computing some, not necessarily optimal, relaxed plan. That is done by a forward pass (building a *relaxed planning graph*), followed by a backward pass (*extracting a relaxed plan*).

Topics We Didn't Cover Here

▷ **Abstractions, Landmarks, Critical-Path Heuristics, Cost Partitions, Compilability between Heuristic Functions, Planning Competitions:**

▷ **Tractable fragments:** Planning sub-classes that can be solved in polynomial time. Often identified by properties of the "causal graph" and "domain transition graphs".

▷ **Planning as SAT:** Compile length-k bounded plan existence into satisfiability of a CNF formula φ. Extensive literature on how to obtain small φ, how to schedule different values of k, how to modify the underlying SAT solver.

▷ **Compilations:** Formal framework for determining whether planning formalism X is (or is not) at least as expressive as planning formalism Y.

▷ **Admissible pruning/decomposition methods:** Partial-order reduction, symmetry reduction, simulation-based dominance pruning, factored planning, decoupled search.

▷ **Hand-tailored planning:** Automatic planning is the extreme case where the computer is given no domain knowledge other than "physics". We can instead allow the user to provide search control knowledge, trading off modeling effort against search performance.

▷ **Numeric planning**, **temporal planning**, **planning under uncertainty** . . .

Suggested Reading (RN: Same As Previous Chapter):

- Chapters 10: *Classical Planning* and 11: *Planning and Acting in the Real World* in [RN09].
  - Although the book is named "*A Modern Approach*", the planning section was written long before the IPC was even dreamt of, before PDDL was conceived, and several years before heuristic search hit the scene. As such, what we have right now is the attempt of two outsiders trying in vain to catch up with the dramatic changes in planning since 1995.
  - Chapter 10 is Ok as a background read. Some issues are, imho, misrepresented, and it's far from being an up-to-date account. But it's Ok to get some additional intuitions in words different from my own.
  - Chapter 11 is useful in our context here because we don't cover any of it. If you're interested in extended/alternative planning paradigms, do read it.

- A good source for modern information (some of which we covered in the lecture) is Jörg Hoffmann's *Everything You Always Wanted to Know About Planning (But Were Afraid to Ask)* [Hof11] which is available online at `http://fai.cs.uni-saarland.de/hoffmann/papers/ki11.pdf`

Chapter 19

Searching, Planning, and Acting in the Real World

Outline

▷ **So Far:** we made idealizing/simplifying assumptions:
The environment is fully observable and deterministic.

▷ **Outline:** In this chapter we will lift some of them

- ▷ The real world (things go wrong)
- ▷ Agents and Belief States
- ▷ Conditional planning
- ▷ Monitoring and replanning

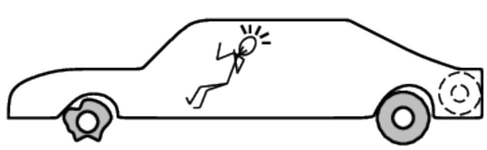
▷ **Note:** The considerations in this chapter apply to both search and planning.

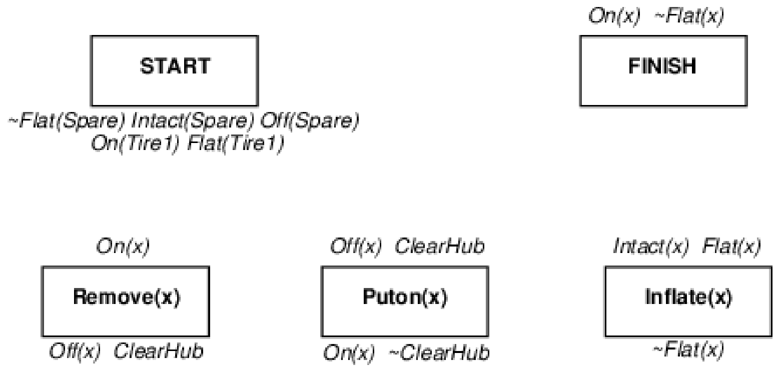
19.1 Introduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26908`.

The real world

▷ **Example 19.1.1.** We have a flat tire – what to do?





Generally: Things go wrong (in the real world)

- ▷ **Example 19.1.2 (Incomplete Information).**
  - ▷ Unknown preconditions, e.g., $Intact(Spare)$?
  - ▷ Disjunctive effects, e.g., $Inflate(x)$ causes $Inflated(x) \vee SlowHiss(x) \vee Burst(x) \vee BrokenPump \vee \ldots$
- ▷ **Example 19.1.3 (Incorrect Information).**
  - ▷ Current state incorrect, e.g., spare NOT intact
  - ▷ Missing/incorrect effects in actions.
- ▷ **Definition 19.1.4.** The qualification problem in planning is that we can never finish listing all the required preconditions and possible conditional effects of actions.
- ▷ **Root Cause:** The environment is partially observable and/or non-deterministic.
- ▷ **Technical Problem:** We cannot know the "current state of the world", but search/-planning algorithms are based on this assumption.
- ▷ **Idea:** Adapt search/planning algorithms to work with "sets of possible states".

What can we do if things (can) go wrong?

- ▷ **One Solution:** Sensorless planning: plans that work regardless of state/outcome.
- ▷ **Problem:** Such plans may not exist! (but they often do in practice)
- ▷ **Another Solution:** Conditional plans:
  - ▷ Plan to obtain information, (observation actions)
  - ▷ Subplan for each contingency.

▷ **Example 19.1.5 (A conditional Plan).** (AAA $\widehat{=}$ ADAC)
$[Check(T1),$ if $Intact(T1)$ then $Inflate(T1)$ else $CallAAA$ fi$]$

▷ **Problem:** Expensive because it plans for many unlikely cases.

▷ **Still another Solution:** Execution monitoring/replanning

- ▷ Assume normal states/outcomes, check progress *during execution*, replan if necessary.

▷ **Problem:** Unanticipated outcomes may lead to failure. (e.g., no AAA card)

▷ **Observation 19.1.6.** *We really need a combination; plan for likely/serious eventualities, deal with others when they arise, as they must eventually.*

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 652 2023-09-20

19.2 The Furniture Coloring Example

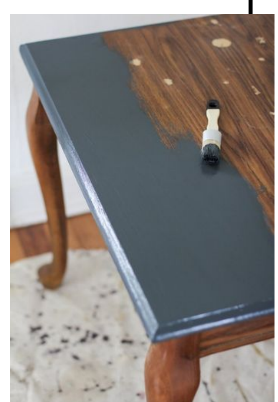
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29180`. We now introduce a planning example that shows off the various features.

The Furniture-Coloring Example: Specification

▷ **Example 19.2.1 (Coloring Furniture).**

Paint a chair and a table in matching colors.

- ▷ The initial state is:
  - ▷ we have two cans of paint of unknown color,
  - ▷ the color of the furniture is unknown as well,
  - ▷ only the table is in the agent's field of view.
- ▷ Actions:
  - ▷ remove lid from can
  - ▷ paint object with paint from open can.



FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 1 653 2023-09-20

We formalize the example in PDDL for simplicity. Note that the :percept scheme is not part of the official PDDL, but fits in well with the design.

The Furniture-Coloring Example: PDDL

▷ **Example 19.2.2 (Formalization in PDDL).**

- ▷ The PDDL domain file is as expected (actions below)

```
(define (domain furniture−coloring)
  (:predicates (object ?x) (can ?x) (inview ?x) (color ?x ?y))
  ...)
```

▷ The PDDL problem file has a "free" variable ?c for the (undetermined) joint color.

```
(define (problem tc−coloring)
  (:domain furniture−objects)
  (:objects table chair c1 c2)
  (:init (object table) (object chair) (can c1) (can c2) (inview table))
  (:goal (color chair ?c) (color table ?c)))
```

▷ Two action schemata: *remove can lid to open* and *paint with open can*

```
(:action remove−lid
        :parameters (?x)
        :precondition (can ?x)
        :effect (open can))
(:action paint
        :parameters (?x ?y)
        :precondition (and (object ?x) (can ?y) (color ?y ?c) (open ?y))
        :effect (color ?x ?c))
```

has a universal variable ?c for the paint action ⇜ we cannot just give paint a color argument in a partially observable environment.

▷ **Sensorless Plan**: Open one can, paint chair and table in its color.

▷ **Note**: Contingent planning can create better plans, but needs perception

▷ Two percept schemata: *color of an object* and *color in a can*

```
(:percept color
        :parameters (?x ?c)
        :precondition (and (object ?x) (inview ?x)))
(:percept can−color
        :parameters (?x ?c)
        :precondition (and (can ?x) (inview ?x) (open ?x)))
```

To perceive the color of an object, it must be in view, a can must also be open.
Note: In a fully observable world, the percepts would not have preconditions.

▷ An action schema: *look at an object* that causes it to come into view.

```
(:action lookat
        :parameters (?x)
        :precond: (and (inview ?y) and (notequal ?x ?y))
        :effect (and (inview ?x) (not (inview ?y))))
```

▷ **Contingent Plan**:

1. look at furniture to determine color, if same ⇝ done.
2. else, look at open and look at paint in cans
3. if paint in one can is the same as an object, paint the other with this color
4. else paint both in any color

19.3 Searching/Planning with Non-Deterministic Actions

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29181`.

Conditional Plans

▷ **Definition 19.3.1.** Conditional plans extend the possible actions in plans by conditional steps that execute sub plans conditionally whether $K+P \models C$, where $K+P$ is the current knowledge base + the percepts.

▷ **Example 19.3.2.** $[\ldots,$ if C then $Plan_A$ else $Plan_B$ fi$,\ldots]$

▷ **Definition 19.3.3.** If the possible percepts are limited to determining the current state in a conditional plan, then we speak of a contingency plan.

▷ **Note:** Need *some plan* for *every possible percept*! Compare to

game playing: *some response* for *every opponent* move.

backchaining: *some rule* such that *every premise* satisfied.

▷ **Idea:** Use an AND–OR tree search (very similar to backward chaining algorithm)

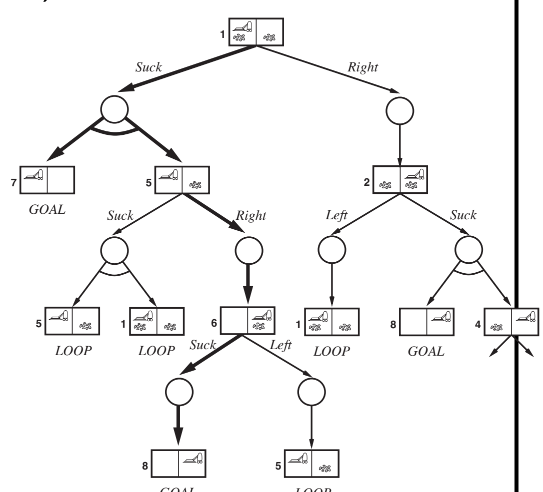
Contingency Planning: The Erratic Vacuum Cleaner

▷ **Example 19.3.4 (Erratic vacuum world).**

A variant $suck$ action:
if square is

▷ $dirty$: clean the square, sometimes remove dirt in adjacent square.

▷ $clean$: sometimes deposits dirt on the carpet.



Solution: $[suck,$ if $State=5$ then $[right, suck]$ else $[]$ fi$]$

Conditional AND/OR Search (Data Structure)

▷ **Idea:** Use OR trees as data structures for representing problems (or goals) that can be reduced to to conjunctions and disjunctions of subproblems (or subgoals).

▷ **Definition 19.3.5.** An OR graph is a is a graph whose non-terminal nodes are partitioned into AND nodes and OR nodes. A valuation of an OR graph T is an assignment of T or F to the nodes of T. A valuation of the terminal nodes of T can be extended by all nodes recursively: Assign T to an

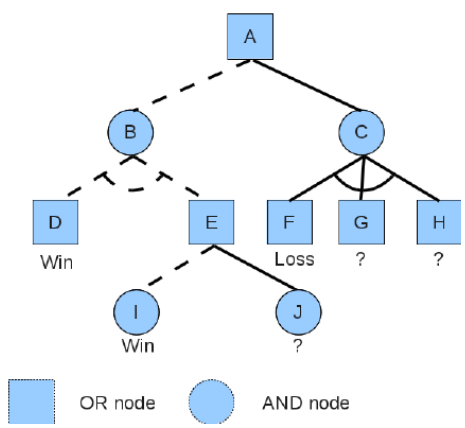
- OR node, iff at least one of its children is T.
- AND node, iff all of its children are T.

A solution for T is a valuation that assigns T to the initial nodes of T.

- **Idea:** A planning task with non deterministic actions generates a OR graph T. A solution that assigns T to a terminal node, iff it is a goal node. Corresponds to a conditional plan.

Conditional AND/OR Search (Example)

- **Definition 19.3.6.** An OR tree is a OR graph that is also a tree.
  Notation: AND nodes are written with arcs connecting the child edges.
- **Example 19.3.7 (An AND/OR-tree).**



Conditional AND/OR Search (Algorithm)

- **Definition 19.3.8.** OR search is an algorithm for searching AND–OR graphs generated by nondeterministic environments.

function AND/OR–GRAPH–SEARCH($prob$) **returns** a conditional plan, or **fail**
  OR–SEARCH($prob$.INITIAL–STATE, $prob$, [])
function OR–SEARCH($state$,$prob$,$path$) **returns** a conditional plan, or **fail**
  if $prob$.GOAL–TEST($state$) **then return** the empty plan
  if $state$ is on $path$ **then return fail**
  for each $action$ **in** $prob$.ACTIONS($state$) **do**
    $plan$:= AND–SEARCH(RESULTS($state$,$action$),$prob$,[$state$ | $path$])
    if $plan \neq$ **fail then return** [$action$ | $plan$]
  return fail
function AND–SEARCH($states$,$prob$,$path$) **returns** a conditional plan, or **fail**

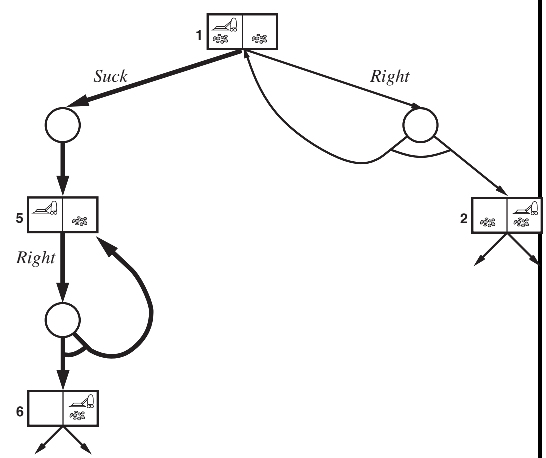
for each s_i in $states$ do
  p_i := OR−SEARCH(s_i,$prob$,$path$)
  if p_i = fail then return fail
  return [if s_1 then p_1 else if s_2 then p_2 else ... if s_{n-1} then p_{n-1} else p_n]

▷ **Cycle Handling:** If a state has been seen before ⇝ **fail**
  ▷ **fail** does not mean *there is no solution*, but
  ▷ *if there is a non-cyclic solution, then it is reachable by an earlier incarnation*!

The Slippery Vacuum Cleaner (try, try, try, ... try again)

▷ **Example 19.3.9 (Slippery Vacuum World).**

Moving sometimes fails
⇝ OR graph



Two possible solutions (depending on what our plan language allows)

  ▷ $[L_1 : left$, if AtR then L_1 else [if $CleanL$ then $\emptyset$ else $suck$ fi] fi] or
  ▷ [while AtR do [$left$] done, if $CleanL$ then $\emptyset$ else $suck$ fi]

▷ We have an infinite loop but plan eventually works unless action always fails.

19.4 Agent Architectures based on Belief States

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29182`.

We are now ready to proceed to environments which can only partially observed and where are our actions are non deterministic. Both sources of uncertainty conspire to allow us only partial knowledge about the world, so that we can only optimize "expected utility" instead of "actual utility" of our actions.

World Models for Uncertainty

▷ **Problem:** We do not know with certainty what state the world is in!

▷ **Idea:** Just keep track of all the possible states it could be in.

▷ **Definition 19.4.1.** A model-based agent has a world model consisting of

- ▷ a belief state that has information about the possible states the world may be in, and
- ▷ a sensor model that updates the belief state based on sensor information
- ▷ a transition model that updates the belief state based on actions.

▷ **Idea:** The agent environment determines what the world model can be.

▷ In a fully observable, deterministic environment,

- ▷ we can observe the initial state and subsequent states are given by the actions alone.
- ▷ thus the belief state is a singleton (we call its member the world state) and the transition model is a function from states and actions to states: a transition function.

That is exactly what we have been doing until now: we have been studying methods that build on descriptions of the “actual” world, and have been concentrating on the progression from atomic to factored and ultimately structured representations. Tellingly, we spoke of “world states” instead of “belief states”; we have now justified this practice in the brave new belief-based world models by the (re-) definition of “world states” above. To fortify our intuitions, let us recap from a belief-state-model perspective.

World Models by Agent Type in AI-1

▷ **Note:** All of these considerations only give requirements to the world model. What we can do with it depends on representation and inference.

▷ **Search-based Agents:** In a fully observable, deterministic environment

- ▷ goal-based agent with world state $\widehat{=}$ “current state”
- ▷ no inference. (goal $\widehat{=}$ goal state from search problem)

▷ **CSP-based Agents:** In a fully observable, deterministic environment

- ▷ goal-based agent withworld state $\widehat{=}$ constraint network,
- ▷ inference $\widehat{=}$ constraint propagation. (goal $\widehat{=}$ satisfying assignment)

▷ **Logic-based Agents:** In a fully observable, deterministic environment

- ▷ model-based agent with world state $\widehat{=}$ logical formula
- ▷ inference $\widehat{=}$ e.g. DPLL or resolution. (no decision theory covered in AI-1)

▷ **Planning Agents:** In a fully observable, deterministic, environment

- ▷ goal-based agent with world state $\widehat{=}$ PL0, transition model $\widehat{=}$ STRIPS,
- ▷ inference $\widehat{=}$ state/plan space search. (goal: complete plan/execution)

Let us now see what happens when we lift the restrictions of total observability and determinism.

World Models for Complex Environments

- In a fully observable, but stochastic environment,
  - the belief state must deal with a set of possible states.
  - $\leadsto$ generalize the transition function to a transition relation.
- **Note:** This even applies to online problem solving, where we can just perceive the state. (e.g. when we want to optimize utility)
- In a deterministic, but partially observable environment,
  - the belief state must deal with a set of possible states.
  - we can use transition functions.
  - We need a sensor model, which predicts the influence of percepts on the belief state – during update.
- In a stochastic, partially observable environment,
  - mix the ideas from the last two. (sensor model + transition relation)

Preview: New World Models (Belief) $\leadsto$ new Agent Types

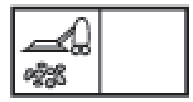
- **Probabilistic Agents:** In a partially observable environment
  - belief state $\widehat{=}$ Bayesian networks,
  - inference $\widehat{=}$ probabilistic inference.
- **Decision-Theoretic Agents:**
  In a partially observable, stochastic environment
  - belief state + transition model $\widehat{=}$ decision networks,
  - inference $\widehat{=}$ maximizing expected utility.
- We will study them in detailin the coming semester.

19.5 Searching/Planning without Observations

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29183`.

Conformant/Sensorless Planning

▷ **Definition 19.5.1.** Conformant or sensorless planning tries to find plans that work without any sensing. (not even the initial state)



▷ **Example 19.5.2 (Sensorless Vacuum Cleaner World).**

States	integer dirt and robot locations
Actions	$left, right, suck, noOp$
Goal tests	$notdirty?$

▷ **Observation 19.5.3.** *In a sensorless world we do not know the initial state. (or any state after)*

▷ **Observation 19.5.4.** *Sensorless planning must search in the space of belief states (sets of possible actual states).*

▷ **Example 19.5.5 (Searching the Belief State Space).**

- Start in $\{1, 2, 3, 4, 5, 6, 7, 8\}$
- Solution: $[right, suck, left, suck]$

$$\begin{array}{ll} right & \rightarrow \{2, 4, 6, 8\} \\ suck & \rightarrow \{4, 8\} \\ left & \rightarrow \{3, 7\} \\ suck & \rightarrow \{7\} \end{array}$$

Search in the Belief State Space: Let's Do the Math

▷ **Recap:** We describe an search problem $\Pi{:=}\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ via its states $\mathcal{S}$, actions $\mathcal{A}$, and transition model $\mathcal{T}\colon\mathcal{A}{\times}\mathcal{S}{\rightarrow}\mathcal{P}(\mathcal{A})$, goal states $\mathcal{G}$, and initial state $\mathcal{I}$.

▷ **Problem:** What is the corresponding sensorless problem?

▷ **Let' think:** Let $\Pi{:=}\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$ be a (physical) problem

- States $\mathcal{S}^b$: The belief states are the $2^{|\mathcal{S}|}$ subsets of $\mathcal{S}$.
- The initial state $\mathcal{I}^b$ is just $\mathcal{S}$ (no information)
- Goal states $\mathcal{G}^b{:=}\{S{\in}\mathcal{S}^b \mid S \subseteq \mathcal{G}\}$ (all possible states must be physical goal states)
- Actions $\mathcal{A}^b$: we just take $\mathcal{A}$. (that's the point!)
- Transition model $\mathcal{T}^b\colon\mathcal{A}^b{\times}\mathcal{S}^b{\rightarrow}\mathcal{P}(\mathcal{A}^b)$: i.e. what is $\mathcal{T}^b(a, S)$ for $a{\in}\mathcal{A}$ and $S \subseteq \mathcal{S}$? This is slightly tricky as a need not be applicable to all $s{\in}S$.
  1. if actions are harmless to the environment, take $\mathcal{T}^b(a, S){:=}\bigcup_{s\in S}\mathcal{T}(a, s)$.
  2. if not, better take $\mathcal{T}^b(a, S){:=}\bigcap_{s\in S}\mathcal{T}(a, s)$. (the safe bet)

▷ **Observation 19.5.6.** *In belief-state space the problem is always fully observable!*

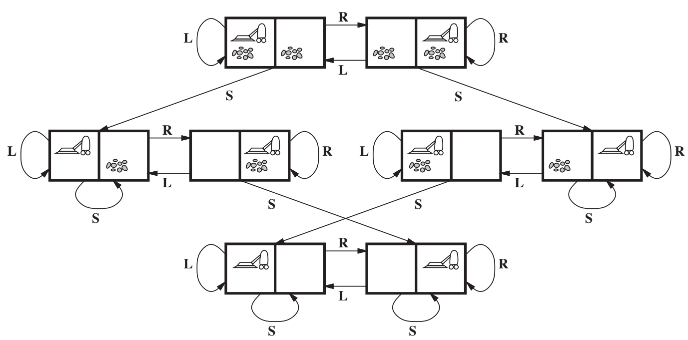
Let us see if we can understand the options for $\mathcal{T}^b(a, S)$ a bit better. The first question is when we want an action a to be applicable to a belief state $S \subseteq \mathcal{S}$, i.e. when should $\mathcal{T}^b(a, S)$ be non-empty.

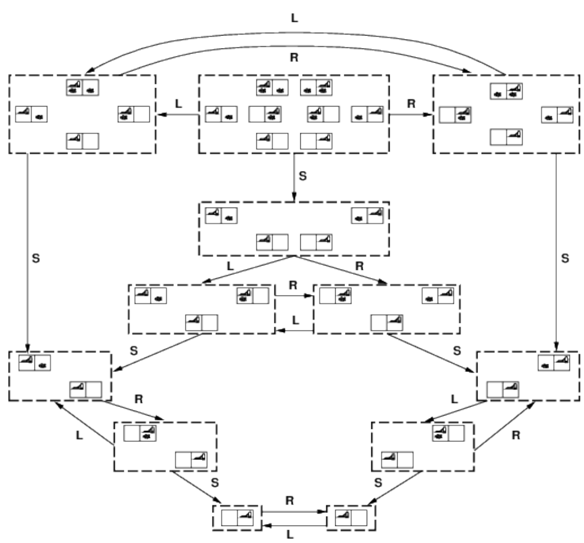
In the first case, a^b would be applicable iff a is applicable to some $s \in S$, in the second case if a is applicable to all $s \in S$. So we only want to choose the first case if actions are harmless.

The second question we ask ourselves is what should be the results of applying a to $S \subseteq \mathcal{S}$?, again, if actions are harmless, we can just collect the results, otherwise, we need to make sure that all members of the result a^b are reached for all possible states in S.

State Space vs. Belief State Space

▷ **Example 19.5.7 (State/Belief State Space in the Vacuum World).** In the vacuum world all actions are always applicable (1./2. equal)





Evaluating Conformant Planning

▷ **Upshot:** We can build belief-space problem formulations automatically,

- ▷ but they are exponentially bigger in theory, in practice they are often similar;
- ▷ e.g. 12 reachable belief states out of $2^8 = 256$ for vacuum example.

▷ **Problem:** Belief states are HUGE; e.g. initial belief state for the 10×10 vacuum world contains $100 \cdot 2^{100} \approx 10^{32}$ physical states

▷ **Idea:** Use planning techniques: compact descriptions for

- ▷ belief states; e.g. *all* for initial state or *not leftmost column* after $\mathcal{A}left$.
- ▷ actions as belief state to belief state operations.

▷ **This actually works:** Therefore we talk about conformant planning!

19.6 Searching/Planning with Observation

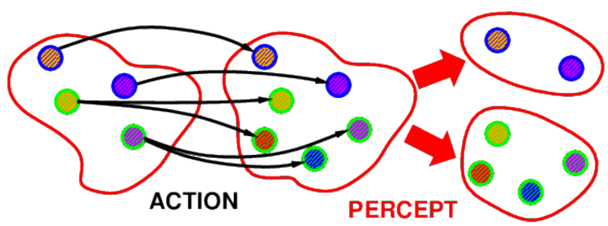
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29184`.

Conditional planning (Motivation)

▷ **Note:** So far, we have never used the agent's sensors.

- ▷ In chapter 6, since the environment was observable and deterministic we could just use offline planning.
- ▷ In section 19.5 because we chose to.

▷ **Note:** If the world is nondeterministic or partially observable then percepts usually provide information, i.e., split up the belief state



▷ **Idea:** This can systematically be used in search/planning via belief-state search, but we need to rethink/specialize the Transition model.

A Transition Model for Belief-State Search

▷ We extend the ideas from slide 666 to include partial observability.

▷ **Definition 19.6.1.** Given a (physical) sproblem $\Pi:=\langle\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{I},\mathcal{G}\rangle$, we define the belief state search problem induced by Π to be $\langle\mathcal{P}(\mathcal{S}),\mathcal{A},\mathcal{T}^b,\mathcal{S},\{S\in\mathcal{S}^b\,|\,S\subseteq\mathcal{G}\}\rangle$, where the transition model $\mathcal{T}^b$ is constructed in three stages:

- ▷ The prediction stage: given a belief state b and an action a we define $\widehat{b}:=\mathrm{PRED}(b,a)$ for some function $\mathrm{PRED}\colon\mathcal{P}(\mathcal{S})\times\mathcal{A}\to\mathcal{P}(\mathcal{S})$.
- ▷ The observation prediction stage determines the set of possible percepts that could be observed in the predicted belief state: $\mathrm{PossPERC}(\widehat{b})=\{\mathrm{PERC}(s)|s\in\widehat{b}\}$.

▷ The update stage determines, for each possible percept, the resulting belief state: $\mathsf{UPDATE}(\widehat{b}, o) := \{s | o = \mathsf{PERC}(s) \text{ and } s \in \widehat{b}\}$

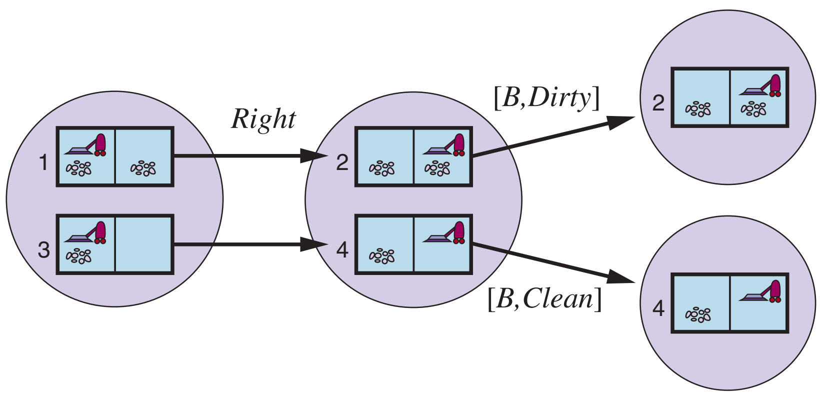
The functions PRED and PERC are the main parameters of this model. We define $\mathsf{RESULT}(b, a) := \{\mathsf{UPDATE}(\mathsf{PRED}(b, a), o) | \mathsf{PossPERC}(\mathsf{PRED}(b, a))\}$

▷ **Observation 19.6.2.** *We always have* $\mathsf{UPDATE}(\widehat{b}, o) \subseteq \widehat{b}$.

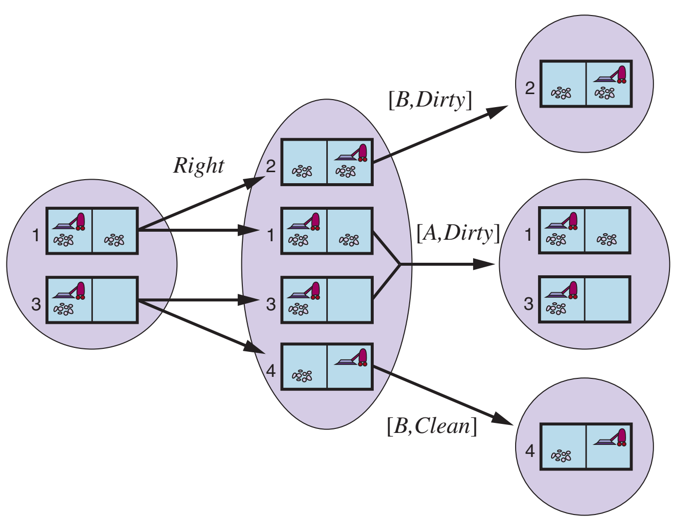
▷ **Observation 19.6.3.** *If sensing is deterministic, belief states for different possible percepts are disjoint, forming a partition of the original predicted belief state.*

Example: Local Sensing Vacuum Worlds

▷ **Example 19.6.4 (Transitions in the Vacuum World).** Deterministic World:



The action $Right$ is deterministic, sensing disambiguates to singletons Slippery World:



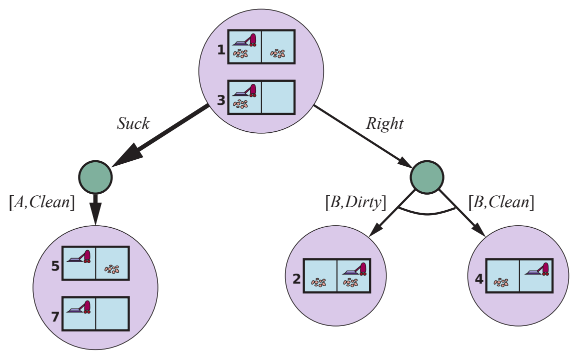
The action $Right$ is non-deterministic, sensing disambiguates somewhat

Belief-State Search with Percepts

▷ **Observation:** The belief-state transition model induces an OR graph.

▷ **Idea:** Use OR search in non deterministic environments.

▷ **Example 19.6.5.** OR graph for initial percept $[A, Dirty]$.



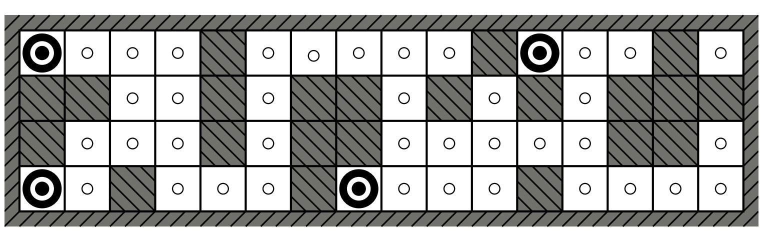
Solution: $[Suck, Right,$ if $Bstate = \{6\}$ then $Suck$ else $[]$ fi$]$

▷ **Note:** Belief-state-problem $\rightsquigarrow$ conditional step tests on belief-state percept (plan would not be executable in a partially observable environment otherwise)

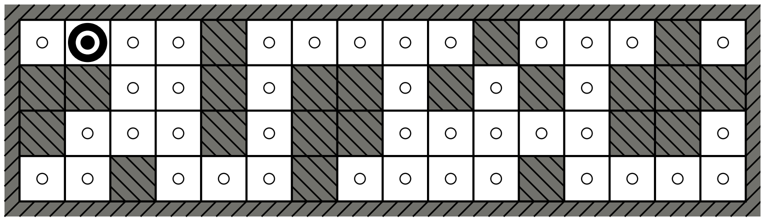
Example: Agent Localization

▷ **Example 19.6.6.** An agent inhabits a maze of which it has an accurate map. It has four sensors that can (reliably) detect walls. The $Move$ action is non-deterministic, moving the agent randomly into one of the adjacent squares.

1. Initial belief state $\rightsquigarrow \widehat{b}_1$ all possible locations.
2. Initial percept: NWS (walls north, west, and south) $\rightsquigarrow \widehat{b}_2 =$ UPDATE$(\widehat{b}_1, NWS)$



3. Agent executes $Move \rightsquigarrow \widehat{b}_3 =$ PRED$(\widehat{b}_2, Move) =$ *one step away from these*.
4. Next percept: $NS \rightsquigarrow \widehat{b}_4 =$ UPDATE$(\widehat{b}_3, NS)$



All in all, $\widehat{b}_4 = \mathrm{UPDATE}(\mathrm{PRED}(\mathrm{UPDATE}(\widehat{b}_1, NWS), Move), NS)$ localizes the agent.

▷ **Observation:** PRED enlarges the belief state, while UPDATE shrinks it again.

Contingent Planning

▷ **Definition 19.6.7.** The generation of plan with conditional branching based on percepts is called contingent planning, solutions are called contingent plans.

▷ Appropriate for partially observable or non-deterministic environments.

▷ **Example 19.6.8.** Continuing Example 19.2.1.
One of the possible contingent plan is
((lookat table) (lookat chair)
  (**if** (**and** (color table c) (color chair c)) (**noop**)
    ((removelid c1) (lookat c1) (removelid c2) (lookat c2)
      (**if** (**and** (color table c) (color can c)) ((paint chair can))
        (**if** (**and** (color chair c) (color can c)) ((paint table can))
          ((paint chair c1) (paint table c1)))))))

▷ **Note:** Variables in this plan are existential; e.g. in

- ▷ line 2: If there is come joint color c of the table and chair $\leadsto$ done.
- ▷ line 4/5: Condition can be satisfied by $[c_1/can]$ or $[c_2/can]$ $\leadsto$ instantiate accordingly.

▷ **Definition 19.6.9.** During plan execution the agent maintains the belief state b, chooses the branch depending on whether $b \models c$ for the condition c.

▷ **Note:** The planner must make sure $b \models c$ can always be decided.

Contingent Planning: Calculating the Belief State

▷ **Problem:** How do we compute the belief state?

▷ **Recall:** Given a belief state b, the new belief state $\widehat{b}$ is computed based on prediction with the action a and the refinement with the percept p.

▷ **Here:**

Given an action a and percepts $p = p_1 \wedge \ldots \wedge p_n$, we have

- $\widehat{b} = (b \backslash \mathrm{del}_a) \cup \mathrm{add}_a$ (as for the sensorless agent)
- If $n = 1$ and (:percept p_1 :precondition c) is the only percept axiom, also add p and c to $\widehat{b}$. (add c as otherwise p impossible)
- If $n > 1$ and (:percept p_i :precondition c_i) are the percept axioms, also add p and $c_1 \vee \ldots \vee c_n$ to $\widehat{b}$. (belief state no longer conjunction of literals ☹)

▷ **Idea:** Given such a mechanism for generating (exact or approximate) updated belief states, we can generate contingent plans with an extension of OR search over belief states.

▷ **Extension:** This also works for non-deterministic actions: we extend the representation of effects to disjunctions.

19.7 Online Search

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29185`.

Online Search and Replanning

▷ **Note:** So far we have concentrated on offline problem solving, where the agent only acts (plan execution) after search/planning terminates.

▷ **Recall:** In online problem solving an agent interleaves computation and action: it computes one action at a time based on incoming perceptions.

▷ Online problem solving is helpful in

- dynamic or semidynamic environments. (long computation times can be harmful)
- stochastic environments. (solve contingencies only when they arise)

▷ Online problem solving is necessary in unknown environments $\rightsquigarrow$ exploration problem.

Online Search Problems

▷ **Observation:** Online problem solving even makes sense in deterministic, fully observable environments.

▷ **Definition 19.7.1.** A online search problem consists of a set S of states, and

- a function Actions(s) that returns a list of actions allowed in state s.
- the step cost function c, where $c(s,a,s')$ is the cost of executing action a in state s with outcome s'. (cost unknown before executing a)
- a goal test Goal Test.

▷ **Note:** We can only determine RESULT(s,a) by being in s and executing a.

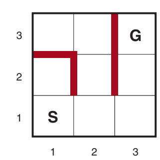
▷ **Definition 19.7.2.** The competitive ratio of an online problem solving agent is the quotient of

- ▷ offline performance, i.e. cost of optimal solutions with full information and
- ▷ online performance, i.e. the actual cost induced by online problem solving.

Online Search Problems (Example)

▷ **Example 19.7.3 (A simple maze problem).**
The agent starts at S and must reach G but knows nothing of the environment. In particular not that

- ▷ Up$(1,1)$ results in $(1,2)$ and
- ▷ Down$(1,1)$ results in $(1,1)$ (i.e. back)



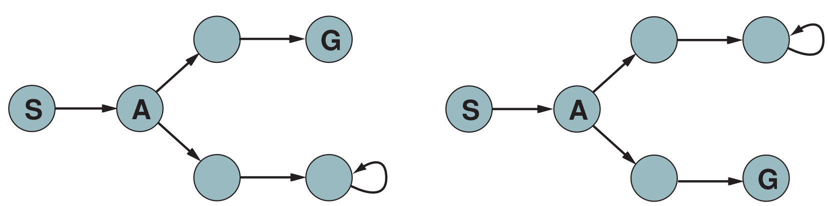
Online Search Obstacles (Dead Ends)

▷ **Definition 19.7.4.** We call a state a dead end, iff no state is reachable from it by an action. An action that leads to a dead end is called irreversible.

▷ **Note:** With irreversible actions the competitive ratio can be infinite.

▷ **Observation 19.7.5.** *No online algorithm can avoid dead ends in all state spaces.*

▷ **Example 19.7.6.** Two state spaces that lead an online agent into dead ends:

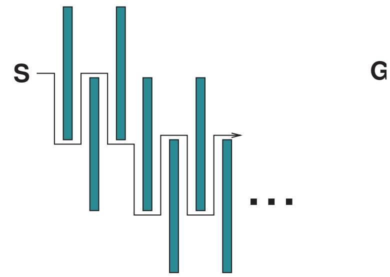


Any agent will fail in at least one of the spaces.

▷ **Definition 19.7.7.** We call Example 19.7.6 an adversary argument.

▷ **Example 19.7.8.** Forcing an online agent into an arbitrarily inefficient route:

Whichever choice the agent makes the adversary can block with a long, thin wall



▷ **Observation:** Dead ends are a real problem for robots: ramps, stairs, cliffs, ...

▷ **Definition 19.7.9.** A state space is called safely explorable, iff a goal state is reachable from every reachable state.

▷ We will always assume this in the following.

Online Search Agents

▷ **Observation:** Online and offline search algorithms differ considerably:

- ▷ For an offline agent, the environment is visible a priori.
- ▷ An online agent builds a "map" of the environment from percepts in visited states.

Therefore, e.g. A^* can expand any node in the fringe, but an online agent must go there to explore it.

▷ **Intuition:** It seems best to expand nodes in "local order" to avoid spurious travel.

▷ **Idea:** Depth first search seems a good fit. (must only travel for backtracking)

Online DFS Search Agent

▷ **Definition 19.7.10.** The :

function ONLINE–DFS–AGENT(s') **returns** an action
  inputs: s', a percept that identifies the current state
  persistent: $result$, a table mapping (s,a) **to** s', initially empty
    $untried$, a table mapping s **to** a list of untried actions
    $unbacktracked$, a table mapping s **to** a list backtracks not tried
    s, a, the previous state and action, initially null
  if Goal-Test(s') **then return** stop
  if $s' \notin untried$ **then** $untried[s'] :=$ Actions(s')
  if s is not null **then**
    $result[s,a] := s'$
    add s **to** the front of $unbacktracked[s']$
  if $untried[s']$ is empty **then**
    if $unbacktracked[s']$ is empty **then return** stop

else $a :=$ an action b such that $result[s', b] = pop(unbacktracked[s'])$
else $a := pop(untried[s'])$
$s := s'$
return a

▷ **Note:** $result$ is the "environment map" constructed as the agent explores.

19.8 Replanning and Execution Monitoring

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29186`.

Replanning (Ideas)

▷ **Idea:** We can turn a planner P into an online problem solver by adding an action RePlan(g) without preconditions that re-starts P in the current state with goal g.

▷ **Observation:** Replanning induces a tradeoff between pre planning and re-planning.

▷ **Example 19.8.1.** The plan [RePlan(g)] is a (trivially) complete plan for any goal g. (not helpful)

▷ **Example 19.8.2.** A plan with sub-plans for every contingency (e.g. what to do if a meteor strikes) may be too costly/large. (wasted effort)

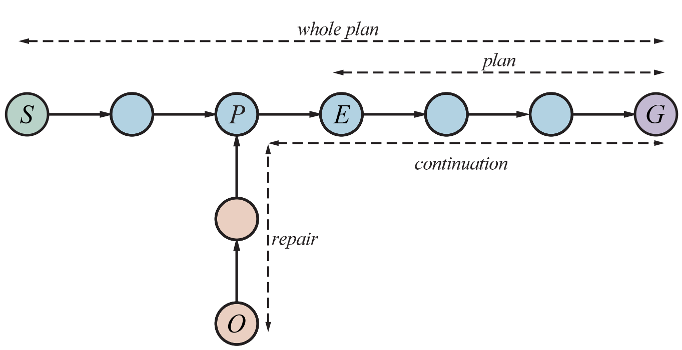
▷ **Example 19.8.3.** But when a tire blows while driving into the desert, we want to have water pre-planned. (due diligence against catastrophies)

▷ **Observation:** In stochastic or partially observable environments we also need some form of execution monitoring to determine the need for replanning (plan repair).

Replanning for Plan Repair

▷ **Generally:** Replanning when the agent's model of the world is incorrect.

▷ **Example 19.8.4 (Plan Repair by Replanning).** Given a plan from S to G.



- ▷ The agent executes $wholeplan$ step by step, monitoring the rest ($plan$).
- ▷ After a few steps the agent expects to be in E, but observes state O.
- ▷ **Replanning**: by calling the planner recursively
  - ▷ find state P in $wholeplan$ and a plan $repair$ from O to P. (P may be G)
  - ▷ minimize the cost of $repair + continuation$

Factors in World Model Failure ⇝ Monitoring

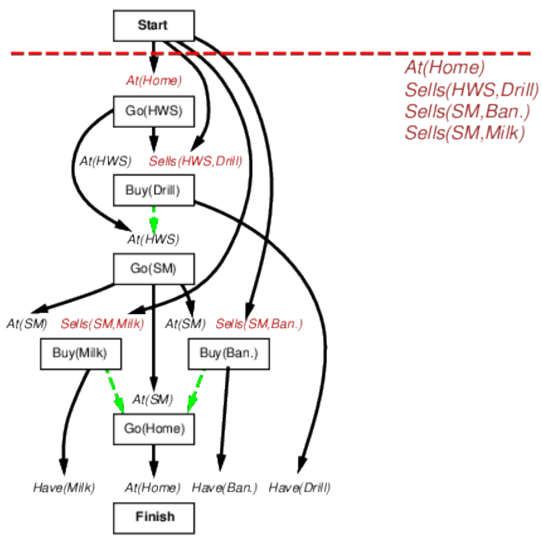
- ▷ **Generally:** The agent's world model can be incorrect, because
  - ▷ an action has a missing precondition (need a screwdriver for remove—lid)
  - ▷ an action misses an effect (painting a table gets paint on the floor)
  - ▷ it is missing a state variable (amount of paint in a can: no paint ⇝ no color)
  - ▷ no provisions for exogenous events (someone knocks over a paint can)
- ▷ **Observation:** Without a way for monitoring for these, planning is very brittle.
- ▷ **Definition 19.8.5.** There are three levels of execution monitoring: before executing an action
  - ▷ action monitoring checks whether all preconditions still hold.
  - ▷ plan monitoring checks that the remaining plan will still succeed.
  - ▷ goal monitoring checks whether there is a better set of goals it could try to achieve.
- ▷ **Note:** Example 19.8.4 was a case of action monitoring leading to replanning.

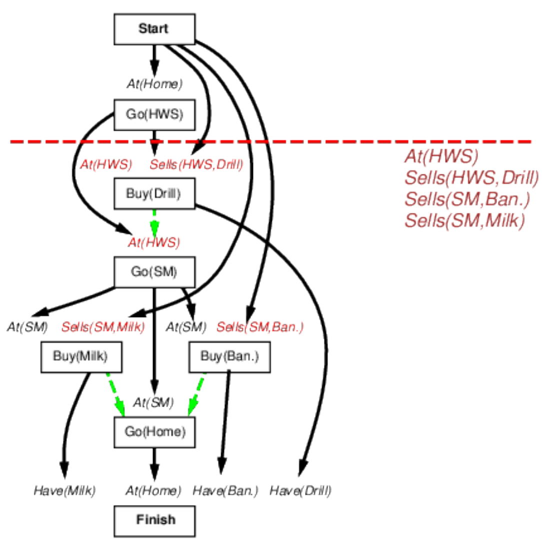
Integrated Execution Monitoring and Planning

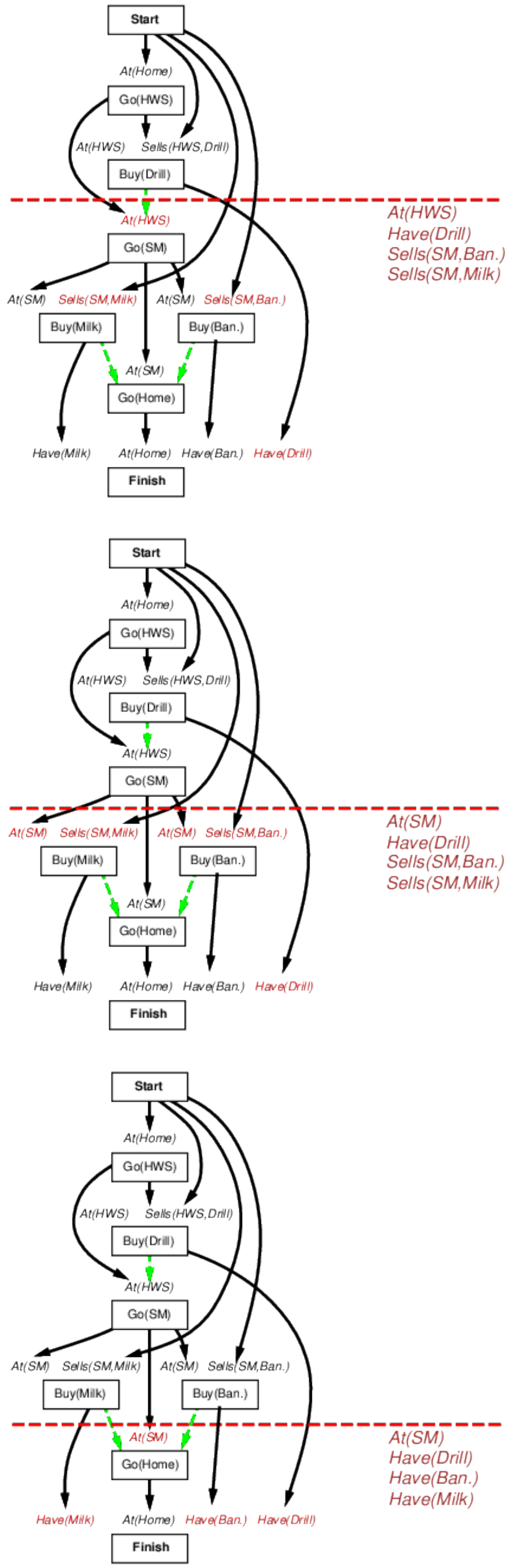
- ▷ **Problem:** Need to upgrade planing data structures by bookkeeping for execution monitoring.
- ▷ **Observation:** With their causal links, partially ordered plans already have most of the infrastructure for action monitoring:
  Preconditions of remaining plan
  $\widehat{=}$ all preconditions of remaining steps not achieved by remaining steps
  $\widehat{=}$ all causal link "crossing current time point"
- ▷ **Idea:** On failure, resume planning (e.g. by POP) to achieve open conditions from current state.
- ▷ **Definition 19.8.6.** IPEM (Integrated Planning, Execution, and Monitoring):
  - ▷ keep updating $Start$ to match current state
  - ▷ links from actions replaced by links from $Start$ when done

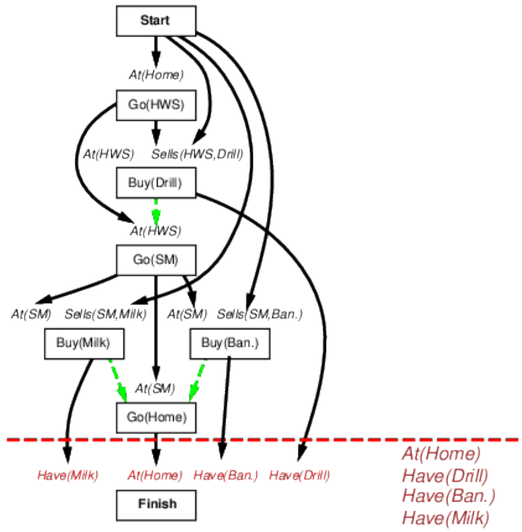
Execution Monitoring Example

▷ **Example 19.8.7 (Shopping for a drill, milk, and bananas).** Start/end at home, drill sold by hardware store, milk/bananas by supermarket.






Start
At(Home)
Go(HWS)
At(HWS)
Sells(HWS,Drill)
Buy(Drill)
At(HWS)
Go(SM)
At(SM)
Sells(SM,Milk)
At(SM)
Sells(SM,Ban.)
Buy(Milk)
Buy(Ban.)
At(SM)
Go(Home)
Have(Milk)
At(Home)
Have(Ban.)
Have(Drill)
Finish
At(HWS)
Have(Drill)
Sells(SM,Ban.)
Sells(SM,Milk)
Start
At(Home)
Go(HWS)
At(HWS)
Sells(HWS,Drill)
Buy(Drill)
At(HWS)
Go(SM)
At(SM)
Sells(SM,Milk)
At(SM)
Sells(SM,Ban.)
Buy(Milk)
Buy(Ban.)
At(SM)
Go(Home)
Have(Milk)
At(Home)
Have(Ban.)
Have(Drill)
Finish
At(SM)
Have(Drill)
Sells(SM,Ban.)
Sells(SM,Milk)
Start
At(Home)
Go(HWS)
At(HWS)
Sells(HWS,Drill)
Buy(Drill)
At(HWS)
Go(SM)
At(SM)
Sells(SM,Milk)
At(SM)
Sells(SM,Ban.)
Buy(Milk)
Buy(Ban.)
At(SM)
Go(Home)
Have(Milk)
At(Home)
Have(Ban.)
Have(Drill)
Finish
At(SM)
Have(Drill)
Have(Ban.)
Have(Milk)


Start
At(Home)
Go(HWS)
At(HWS)
Sells(HWS,Drill)
Buy(Drill)
At(HWS)
Go(SM)
At(SM)
Sells(SM,Milk)
At(SM)
Sells(SM,Ban.)
Buy(Milk)
Buy(Ban.)
At(SM)
Go(Home)
Have(Milk)
At(Home)
Have(Ban.)
Have(Drill)
Finish
At(Home)
Have(Drill)
Have(Ban.)
Have(Milk)

Chapter 20

Semester Change-Over

20.1 What did we learn in AI 1?

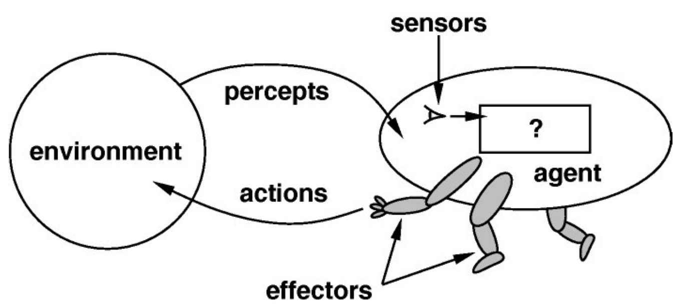
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/26916`.

Topics of AI-1 (Winter Semester)

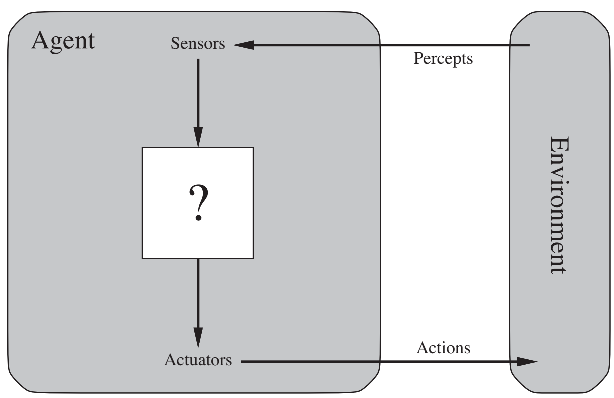
- ▷ Getting Started
  - ▹ What is Artificial Intelligence? (situating ourselves)
  - ▹ Logic programming in Prolog (An influential paradigm)
  - ▹ Intelligent Agents (a unifying framework)
- ▷ Problem Solving
  - ▹ Problem Solving and search (Black Box World States and Actions)
  - ▹ Adversarial Search (Game playing) (A nice application of Search)
  - ▹ constraint satisfaction problems (Factored World States)
- ▷ Knowledge and Reasoning
  - ▹ Formal Logic as the mathematics of Meaning
  - ▹ Propositional logic and satisfiability (Atomic Propositions)
  - ▹ First-order logic and theorem proving (Quantification)
  - ▹ Logic programming (Logic + Search$\leadsto$ Programming)
  - ▹ Description logics and semantic web
- ▷ Planning
  - ▹ Planning Frameworks
  - ▹ Planning Algorithms
  - ▹ Planning and Acting in the real world

Rational Agents as an Evaluation Framework for AI

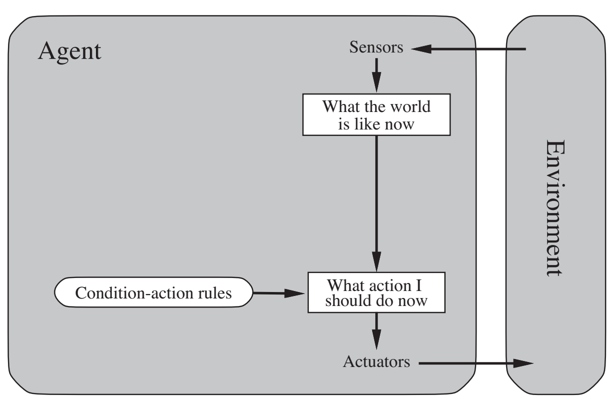
▷ Agents interact with the environment



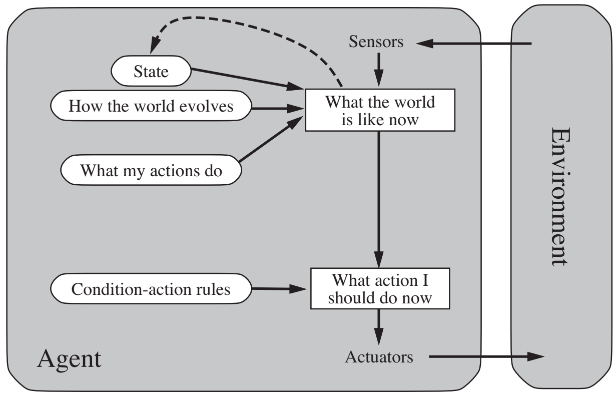
General agent schema



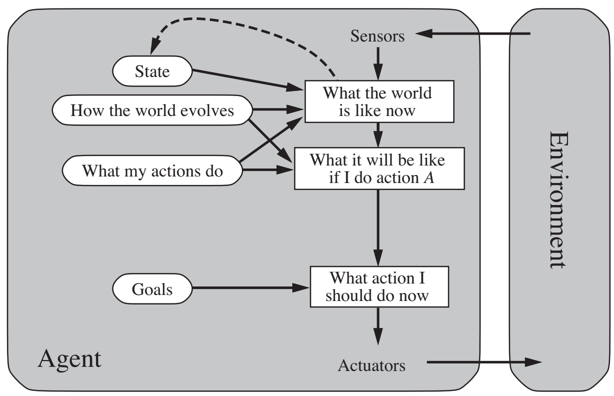
Simple Reflex Agents



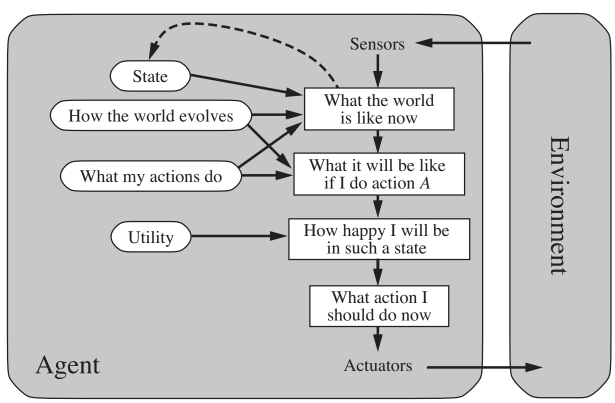
Reflex Agents with State



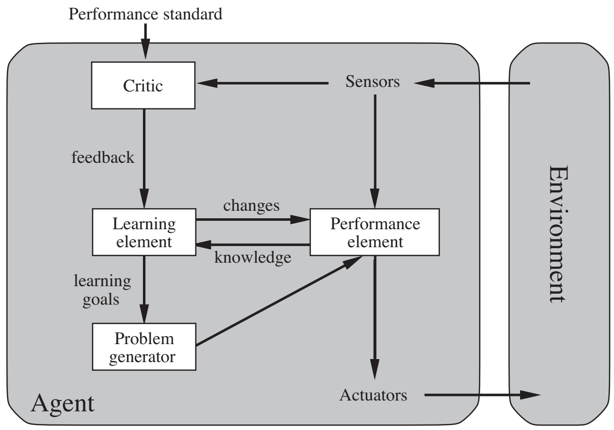
Goal-Based Agents



Utility-Based Agent



Learning Agents



Rational Agent

- **Idea:** Try to design agents that are successful (do the right thing)
- **Definition 20.1.1.** An agent is called rational, if it chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date. This is called the MEU principle.
- **Note:** A rational agent need not be perfect
  - only needs to maximize expected value (rational $\neq$ omniscient)
    - need not predict e.g. very unlikely but catastrophic events in the future
  - percepts may not supply all relevant information (Rational $\neq$ clairvoyant)
    - if we cannot perceive things we do not need to react to them.
    - but we may need to try to find out about hidden dangers (exploration)
  - action outcomes may not be as expected (rational $\neq$ successful)
    - but we may need to take action to ensure that they do (more often) (learning)
- Rational $\rightsquigarrow$ exploration, learning, autonomy

Symbolic AI: Adding Knowledge to Algorithms

- Problem Solving (Black Box States, Transitions, Heuristics)
  - **Framework**: Problem Solving and Search (basic tree/graph walking)
  - **Variant**: Game playing (Adversarial Search) (Minimax + $\alpha\beta$-Pruning)
- Constraint Satisfaction Problems (heuristic search over partial assignments)
  - States as partial variable assignments, transitions as assignment

  - ▹ Heuristics informed by current restrictions, constraint graph
  - ▹ Inference as constraint propagation (transferring possible values across arcs)
- ▷ Describing world states by formal language (and drawing inferences)
  - ▹ Propositional logic and DPLL (deciding entailment efficiently)
  - ▹ First-order logic and ATP (reasoning about infinite domains)
  - ▹ **Digression**: Logic programming (logic + search)
  - ▹ Description logics as moderately expressive, but decidable logics
- ▷ Planning: Problem Solving using white-box world/action descriptions
  - ▹ **Framework**: describing world states in logic as sets of propositions and actions by preconditions and add/delete lists
  - ▹ **Algorithms**: e.g heuristic search by problem relaxations

Topics of AI-2 (Summer Semester)

- ▷ Uncertain Knowledge and Reasoning
  - ▹ Uncertainty
  - ▹ Probabilistic reasoning
  - ▹ Making Decisions in Episodic Environments
  - ▹ Problem Solving in Sequential Environments
- ▷ Foundations of machine learning
  - ▹ Learning from Observations
  - ▹ Knowledge in Learning
  - ▹ Statistical Learning Methods
- ▷ Communication (If there is time)
  - ▹ Natural Language Processing
  - ▹ Natural Language for Communication

Artificial Intelligence I/II

Prof. Dr. Michael Kohlhase
Professur für Wissensrepräsentation und -verarbeitung
Informatik, FAU Erlangen-Nürnberg
Michael.Kohlhase@FAU.de

20.2 Administrativa

We will now go through the ground rules for the course. This is a kind of a social contract between the instructor and the students. Both have to keep their side of the deal to make learning as efficient and painless as possible.

Prerequisites for AI-2

- **Content Prerequisites:** the mandatory courses in CS@FAU; Sem 1-4, in particular:
  - course "Mathematik C4" (InfMath4). (for stochastics)
  - (very) elementary complexity theory. (big Oh and friends)

  also AI-1 ("Artificial Intelligence I") (of course)
- **Intuition:** (take them with a kilo of salt)
  - This is what I assume you know! (I have to assume something)
  - In many cases, the dependency of AI-2 on these is partial and "in spirit".
  - If you have not taken these (or do not remember), read up on them as needed!
- **The real Prerequisite:** Motivation, Interest, Curiosity, hard work. (AI-2 is non-trivial)
- You can do this course if you want! (and I hope you are successful)

Now we come to a topic that is always interesting to the students: the grading scheme.

Assessment, Grades

- **Overall (Module) Grade:**
  - Grade via the exam (Klausur) $\leadsto$ 100% of the grade.
  - Up to 10% bonus on-top for an exam with $\geq 50\%$ points.($\leq 50\%$ $\leadsto$ no bonus)
  - Bonus points $\widehat{=}$ percentage sum of the best 10 tuesday quizzes divided by 100.
- **Exam:** 90 minutes exam conducted in presence on paper ($\sim$ Oct. 1. 2023)
- **Retake Exam:** 90 min exam six months later ($\sim$ April 1. 2024)
- ⚠ You have to register for exams in campo in the first month of classes.
- **Note:** You can de-register from an exam on campo up to three working days before.
- **Tuesday Quizzes:** Every tuesday we start the lecture with a 10 min online quiz – the tuesday quiz – about the material from the previous week. (starts in week 2)

AI-2 Homework Assignments

- **Homework Assignments:** Small individual problem/programming/proof task
  - but take time to solve (at least read them directly ⇝ questions)
- ⚠ Homeworks give no bonus points, but without trying you are unlikely to pass the exam.
- **Homework/Tutorial Discipline:**
  - Start early! (many assignments need more than one evening's work)
  - Don't start by sitting at a blank screen (talking & study group help)
  - Humans will be trying to understand the text/code/math when grading it.
  - Go to the tutorials, discuss with your TA! (they are there for you!)
- ⚠ We will not be able to grade all homework assignments!
- **Graded Assignments:** To keep things running smoothly
  - Homeworks will be posted on StudOn.
  - Sign up for AI-2 under `https://www.studon.fau.de/crs4941850.html`.
  - Homeworks are handed in electronically there. (plain text, program files, PDF)
  - Do not sign up for the "AI-2 Übungen" on StudOn (we do not use them)
- **Ungraded Assignments:** Are peer-feedbacked in ALeA (see below)

It is very well-established experience that without doing the homework assignments (or something similar) on your own, you will not master the concepts, you will not even be able to ask sensible questions, and take very little home from the course. Just sitting in the course and nodding is not enough! If you have questions please make sure you discuss them with the instructor, the teaching assistants, or your fellow students. There are three sensible venues for such discussions: online in the lecture, in the tutorials, which we discuss now, or in the course forum – see below. Finally, it is always a very good idea to form study groups with your friends.

Tutorials for Artificial Intelligence 1

- **Approach:** Weekly tutorials and homework assignments (first one in week two)
- **Goal 1:** Reinforce what was taught in class. (you need practice)
- **Goal 2:** Allow you to ask any question you have in a protected environment.
- **Instructor/Lead TA:** Florian Rabe (KWARC Postdoc)
  - Room: 11.137 @ Händler building, `florian.rabe@fau.de`
- **Tutorials:** one each taught by Florian Rabe, ...
- **Life-saving Advice:** Go to your tutorial, and prepare for it by having looked at the slides and the homework assignments!

▷ **Caveat:** We cannot grade all submissions with 5 TAs and ∼1000 students.

▷ **Also:** Group submission has not worked well in the past! (too many freeloaders)

One special case of academic rules that affects students is the question of cheating, which we will cover next.

Cheating [adapted from CMU:15-211 (P. Lee, 2003)]

▷ There is no need to cheat in this course!! (hard work will usually do)

▷ **Note:** Cheating prevents you from learning (you are cutting into your own flesh)

▷ We expect you to know what is useful collaboration and what is cheating.

- ▷ You have to hand in your own original code/text/math for all assignments
- ▷ You may discuss your homework assignments with others, but if doing so impairs your ability to write truly original code/text/math, you will be cheating
- ▷ Copying from peers, books or the Internet is plagiarism unless properly attributed (even if you change most of the actual words)

▷ I am aware that there may have been different standards about this at your previous university! (these are the ground rules here)

▷ ⚠ There are data mining tools that monitor the originality of text/code. ⚠

▷ **Procedure:** If we catch you at cheating... (correction: if we suspect cheating)

- ▷ We will confront you with the allegation and impose a grade sanction.
- ▷ If you have a reasonable explanation we lift that. (you have to convince us)

▷ **Note:** Both active (copying from others) and passive cheating (allowing others to copy) are penalized equally.

We are fully aware that the border between cheating and useful and legitimate collaboration is difficult to find and will depend on the special case. Therefore it is very difficult to put this into firm rules. We expect you to develop a firm intuition about behavior with integrity over the course of stay at FAU. Do use the opportunity to discuss the AI-2 topics with others. After all, one of the non-trivial skills you want to learn in the course is how to talk about Artificial Intelligence topics. And that takes practice, practice, and practice.

Due to the current AI hype, the course Artificial Intelligence is very popular and thus many degree programs at FAU have adopted it for their curricula. Sometimes the course setup that fits for the CS program does not fit the other's very well, therefore there are some special conditions. I want to state here.

⚠ Special Admin Conditions ⚠

▷ Some degree programs do not "import" the course Artificial Intelligence, and thus you may not be able to register for the exam via `https://campus.fau.de`.

- ▷ Just send me an e-mail and come to the exam, we will issue a "Schein".
- ▷ Tell your program coordinator about AI-1/2 so that they remedy this situation

▷ In "Wirtschafts-Informatik" you can only take AI-1 and AI-2 together in the "Wahlpflichtbereich".

- ▷ ECTS credits need to be divisible by five ⇝ $7.5 + 7.5 = 15$.

Michael Kohlhase: Artificial Intelligence 2 697 2023-09-20

I can only warn of what I am aware, so if your degree program lets you jump through extra hoops, please tell me and then I can mention them here.

20.3 Overview over AI and Topics of AI-II

We restart the new semester by reminding ourselves of (the problems, methods, and issues of) Artificial Intelligence, and what has been achived so far.

20.3.1 What is Artificial Intelligence?

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/21701`.
The first question we have to ask ourselves is "What is Artificial Intelligence?", i.e. how can we define it. And already that poses a problem since the natural definition *like human intelligence, but artificially realized* presupposes a definition of Intelligence, which is equally problematic; even Psychologists and Philosophers – the subjects nominally "in charge" of human intelligence – have problems defining it, as witnessed by the plethora of theories e.g. found at [WHI].

What is Artificial Intelligence? Definition

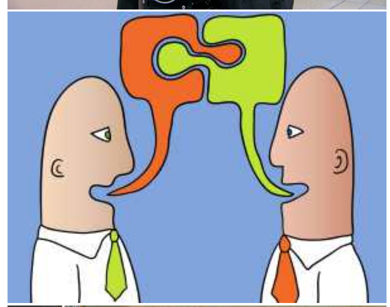
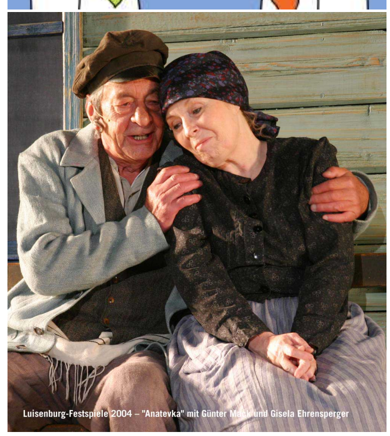
- ▷ **Definition 20.3.1 (According to Wikipedia).** Artificial Intelligence (AI) is intelligence exhibited by machines
- ▷ **Definition 20.3.2 (also).** Artificial Intelligence (AI) is a sub-field of computer science that is concerned with the automation of intelligent behavior.
- ▷ **BUT:** it is already difficult to define "Intelligence" precisely
- ▷ **Definition 20.3.3 (Elaine Rich).** Artificial Intelligence (AI) studies how we can make the computer do things that humans can still do better at the moment.

Michael Kohlhase: Artificial Intelligence 2 698 2023-09-20

Maybe we can get around the problems of defining "what Artificial intelligence is", by just describing the necessary components of AI (and how they interact). Let's have a try to see whether that is more informative.

What is Artificial Intelligence? Components


- **Elaine Rich:** AI studies how we can make the computer do things that humans can still do better at the moment.
- This needs a combination of
  - the ability to learn
  - inference
  - perception
  - language understanding
  - emotion

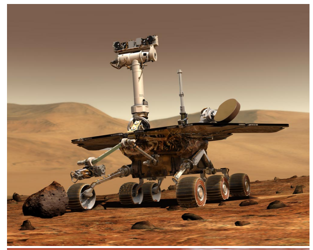




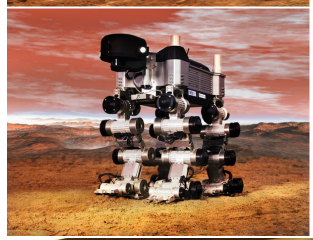
20.3.2 Artificial Intelligence is here today!

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/21697`.

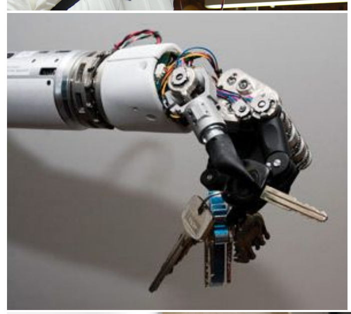
The components of Artificial Intelligence are quite daunting, and none of them are fully understood, much less achieved artificially. But for some tasks we can get by with much less. And indeed that is what the field of Artificial Intelligence does in practice – but keeps the lofty ideal around. This practice of "trying to achieve AI in selected and restricted domains" (cf. the discussion starting with slide 29) has borne rich fruits: systems that meet or exceed human capabilities in such areas. Such systems are in common use in many domains of application.

Artificial Intelligence is here today!





▷ in outer space

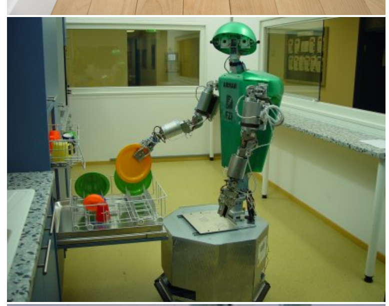


- ▹ in outer space systems need autonomous control:
- ▹ remote control impossible due to time lag

▷ in artificial limbs

- ▹ the user controls the prosthesis via existing nerves, can e.g. grip a sheet of paper.

▷ in household appliances

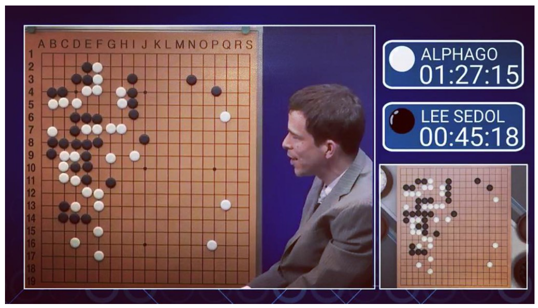


- ▹ The iRobot Roomba vacuums, mops, and sweeps in corners, . . . , parks, charges, and discharges.
- ▹ general robotic household help is on the horizon.

▷ in hospitals

- ▹ in the USA 90% of the prostate operations are carried out by RoboDoc
- ▹ Paro is a cuddly robot that eases solitude in nursing homes.

And here's what you all have been waiting for . . .



CC-BY-SA: Buster Benson@ https://www.flickr.com/photos/erikbenson/25717574115

- ▷ AlphaGo is a program by Google DeepMind to play the board game go.
- ▷ In March 2016, it beat Lee Sedol in a five-game match, the first time a go program has beaten a 9 dan professional without handicaps. In December 2017 AlphaZero, a successor of AlphaGo "learned" the games go, chess, and shogi in 24 hours, achieving a superhuman level of play in these three games by defeating world-champion programs. By September 2019, AlphaStar, a variant of AlphaGo, attained "grandmaster level" in Starcraft II, a real time strategy game with partially observable state. AlphaStar now among the top 0.2% of human players.

We will conclude this subsection with a note of caution.

The AI Conundrum

- ▷ **Observation:** Reserving the term "Artificial Intelligence" has been quite a land grab!
- ▷ **But:** researchers at the Dartmouth Conference (1956) really thought they would solve/reach AI in two/three decades.
- ▷ **Consequence:** AI still asks the big questions.
- ▷ **Another Consequence:** AI as a field is an incubator for many innovative technologies.
- ▷ **AI Conundrum:** Once AI solves a subfield it is called "computer science". (becomes a separate subfield of CS)
- ▷ **Example 20.3.4.** Functional/Logic Programming, automated theorem proving, Planning, machine learning, Knowledge Representation, . . .
- ▷ **Still Consequence:** AI research was alternatingly flooded with money and cut off brutally.

20.3.3 Ways to Attack the AI Problem

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/21717`.
There are currently three main avenues of attack to the problem of building artificially intelligent systems. The (historically) first is based on the symbolic representation of knowledge about the world and uses inference-based methods to derive new knowledge on which to base action decisions. The second uses statistical methods to deal with uncertainty about the world state and learning methods to derive new (uncertain) world assumptions to act on.

Three Main Approaches to Artificial Intelligence

- ▷ **Definition 20.3.5.** Symbolic AI is based on the assumption that many aspects of intelligence can be achieved by the manipulation of symbols, combining them into structures (expressions) and manipulating them (using processes) to produce new expressions.
- ▷ **Definition 20.3.6.** Statistical AI remedies the two shortcomings of symbolic AI approaches: that all concepts represented by symbols are crisply defined, and that all aspects of the world are knowable/representable in principle. Statistical AI adopts sophisticated mathematical models of uncertainty and uses them to create more accurate world models and reason about them.
- ▷ **Definition 20.3.7.** Subsymbolic AI attacks the assumption of symbolic and statistical AI that intelligence can be achieved by reasoning about the state of the world. Instead it posits that intelligence must be embodied i.e. situated in the world, equipped with a "body" that can interact with it via sensors and actuators. The main method for realizing intelligent behavior is by learning from the world, i.e. machine learning.

As a consequence, the field of Artificial Intelligence (AI) is an engineering field at the intersection of computer science (logic, programming, applied statistics), cognitive science (psychology, neuroscience), philosophy (can machines think, what does that mean?), linguistics (natural language understanding), and mechatronics (robot hardware, sensors).
Subsymbolic AI and in particular machine learning is currently hyped to such an extent, that many people take it to be synonymous with "Artificial Intelligence". It is one of the goals of this course to show students that this is a very impoverished view.

Two ways of reaching Artificial Intelligence?

- ▷ We can classify the AI approaches by their coverage and the analysis depth (they are complementary)

Deep	symbolic AI-1	not there yet cooperation?
Shallow	no-one wants this	statistical/sub symbolic AI-2
Analysis ↑ vs. Coverage →	Narrow	Wide

- ▷ **This semester** we will cover foundational aspects of symbolic AI (deep/narrow processing)
- ▷ **next semester** concentrate on statistical/subsymbolic AI. (shallow/wide-coverage)

Michael Kohlhase: Artificial Intelligence 2 704 2023-09-20

We combine the topics in this way in this course, not only because this reproduces the historical development but also as the methods of statistical and subsymbolic AI share a common basis.
It is important to notice that all approaches to AI have their application domains and strong points. We will now see that exactly the two areas, where symbolic AI and statistical/subsymbolic AI have their respective fortes correspond to natural application areas.

Environmental Niches for both Approaches to AI

- ▷ **Observation:** There are two kinds of applications/tasks in AI
  - ▷ Consumer tasks: consumer grade applications have tasks that must be fully generic and wide coverage. (e.g. machine translation like Google Translate)
  - ▷ Producer tasks: producer grade applications must be high-precision, but can be domain-specific (e.g. multilingual documentation, machinery-control, program verification, medical technology)

Precision		
100%	Producer Tasks	
50%		Consumer Tasks
	$10^{3\pm1}$ Concepts	$10^{6\pm1}$ Concepts **Coverage**

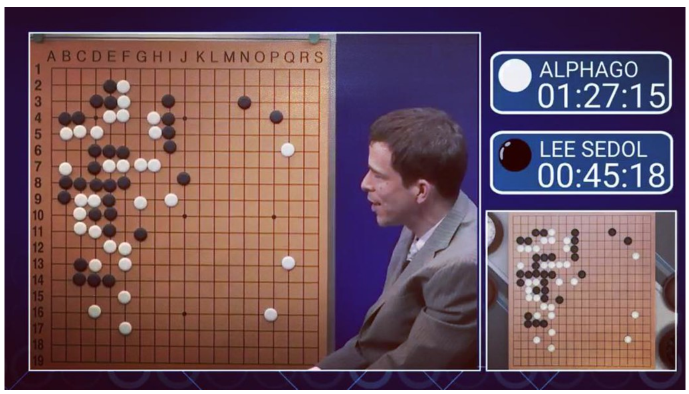
- ▷ **General Rule:** Subsymbolic AI is well suited for consumer tasks, while symbolic AI is better suited for producer tasks.
- ▷ A domain of producer tasks I am interested in: mathematical/technical documents.

Michael Kohlhase: Artificial Intelligence 2 705 2023-09-20

An example of a producer task – indeed this is where the name comes from – is the case of a machine tool manufacturer T, which produces digitally programmed machine tools worth multiple million Euro and sells them into dozens of countries. Thus T must also comprehensive machine operation manuals, a non-trivial undertaking, since no two machines are identical and they must be translated into many languages, leading to hundreds of documents. As those manual share a lot of semantic content, their management should be supported by AI techniques. It is critical that these methods maintain a high precision, operation errors can easily lead to very costly machine damage and loss of production. On the other hand, the domain of these manuals is quite restricted. A machine tool has a couple of hundred components only that can be described by a comple of thousand attribute only.

Indeed companies like T employ high-precision AI techniques like the ones we will cover in this course successfully; they are just not so much in the public eye as the consumer tasks.

To get this out of the way ...



CC-BY-SA: Buster Benson@ https://www.flickr.com/photos/erikbenson/25717574115

- ▷ AlphaGo = search + neural networks (symbolic + subsymbolic AI)
  - ▷ we do search this semester and cover neural networks in AI-2.
  - ▷ I will explain AlphaGo a bit in chapter 7.

Michael Kohlhase: Artificial Intelligence 2 706 2023-09-20

20.3.4 AI in the KWARC Group

The KWARC Research Group

- ▷ **Observation:** The ability to represent knowledge about the world and to draw logical inferences is one of the central components of intelligent behavior.
- ▷ **Thus:** reasoning components of some form are at the heart of many AI systems.
- ▷ **KWARC Angle:** Scaling up (web-coverage) without dumbing down (too much)
  - ▷ Content markup instead of full formalization (too tedious)
  - ▷ User support and quality control instead of "The Truth" (elusive anyway)
  - ▷ use Mathematics as a test tube (Mathematics $\widehat{=}$ Anything Formal)
  - ▷ care more about applications than about philosophy (we cannot help getting this right anyway as logicians)
- ▷ The KWARC group was established at Jacobs Univ. in 2004, moved to FAU Erlangen in 2016

▷ see http://kwarc.info for projects, publications, and links

Overview: KWARC Research and Projects

Applications: eMath 3.0, Active Documents, Active Learning, Semantic Spreadsheets/CAD/CAM, Change Mangagement, Global Digital Math Library, Math Search Systems, SMGloM: Semantic Multilingual Math Glossary, Serious Games, ...

Foundations of Math:	**KM & Interaction**:	**Semantization**:
▷ MathML, OpenMath ▷ advanced Type Theories ▷ MMT: Meta Meta Theory ▷ Logic Morphisms/Atlas ▷ Theorem Prover/CAS Interoperability ▷ Mathematical Models/Simulation	▷ Semantic Interpretation (aka. Framing) ▷ math-literate interaction ▷ MathHub: math archives & active docs ▷ Active documents: embedded semantic services ▷ Model-based Education	▷ LaTeXML: LaTeX → XML ▷ sTeX: Semantic LaTeX ▷ invasive editors ▷ Context-Aware IDEs ▷ Mathematical Corpora ▷ Linguistics of Math ▷ ML for Math Semantics Extraction

Foundations: Computational Logic, Web Technologies, OMDoc/MMT

Research Topics in the KWARC Group

▷ We are always looking for bright, motivated KWARCies.

▷ We have topics in for all levels! (Enthusiast, Bachelor, Master, Ph.D.)

▷ List of current topics: https://gl.kwarc.info/kwarc/thesis-projects/

  ▷ Automated Reasoning: Maths Representation in the Large
  ▷ Logics development, (Meta)^n-Frameworks
  ▷ Math Corpus Linguistics: Semantics Extraction
  ▷ Serious Games, Cognitive Engineering, Math Information Retrieval, Legal Reasoning, ...

▷ We always try to find a topic at the intersection of your and our interests.

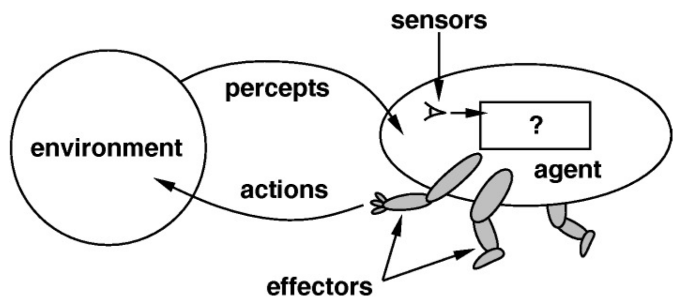
▷ We also often have positions!. (HiWi, Ph.D.: $\frac{1}{2}$, PostDoc: full)

20.3.5 AI-II: Advanced Rational Agents

Remember the conceptual framework we gave ourselves in chapter 5: we posited that all (artificial and natural) intelligence is situated in an agent that interacts with a given environment, and postulated that what we experience as "intelligence" in a (natural or artificial) agent can be ascribed to the agent behaving rationally, i.e. optimizing the expected utility of its actions given the (current) environment.

Agents and Environments

- **Definition 20.3.8.** An agent is anything that
  - perceives its environment via sensors (a means of sensing the environment)
  - acts on it with actuators (means of changing the environment).



- **Example 20.3.9.** Agents include humans, robots, softbots, thermostats, etc.

In the last semester we restricted ourselves to fully observable, deterministic, episodic environments, where optimizing utility is easy in principle – but may still be computationally intractable, since we have full information about the world

Artificial Intelligence II Overview

- We construct rational agents.
- An agent is an entity that perceives its environment through sensors and acts upon that environment through actuators.
- A rational agent is an agent maximizing its expected performance measure.
- In AI-1 we dealt mainly with a logical approach to agent design (no uncertainty).
- We ignored
  - interface to environment (sensors, actuators)
  - uncertainty
  - the possibility of self-improvement (learning)

This semester we want to alleviate all these restrictions and study rationality in more realistic circumstances, i.e. environments which need only be partially observe and where our actions can be non deterministic. Both of these extensions conspire to allow us only partial knowledge about the world, so that we can only optimize "expected utility" instead of " actual utility" of our actions. This directly leads to the first topic.

The second topic is motivated by the fact that environments can change and and are initially unknown, and therefore the agent must obtain and/or update parameters like utilities and world knowledge by observing the environment.

Topics of AI-2 (Summer Semester)

- ▷ Uncertain Knowledge and Reasoning
  - ▷ Uncertainty
  - ▷ Probabilistic reasoning
  - ▷ Making Decisions in Episodic Environments
  - ▷ Problem Solving in Sequential Environments
- ▷ Foundations of machine learning
  - ▷ Learning from Observations
  - ▷ Knowledge in Learning
  - ▷ Statistical Learning Methods
- ▷ Communication (If there is time)
  - ▷ Natural Language Processing
  - ▷ Natural Language for Communication

The last topic (which we will only attack if we have time) is motivated by multi agent environments, where multiple agents have to collaborate for problem solving. Note that even though the adversarial search methods discussed in chapter 7 were essentially single agent as both opponents optimized the utility of their actions alone.

In true multi agent environments we have to also optimize collaboration between agents, and that is usually radially more efficient if agents can communicate.

Part V

Reasoning with Uncertain Knowledge

This part of the course notes addresses inference and agent decision making in partially observable environments, i.e. where we only know probabilities instead of certainties whether propositions are true/false. We cover basic probability theory and – based on that – Bayesian Networks and simple decision making in such environments. Finally we extend this to probabilistic temporal models and their decision theory.

Chapter 21

Quantifying Uncertainty

In this chapter we develop a machinery for dealing with uncertainty: Instead of thinking about what we know to be true, we must think about what is likely to be true.

21.1 Dealing with Uncertainty: Probabilities

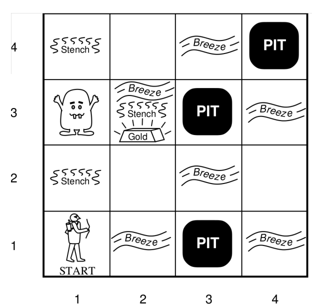
Before we go into the technical machinery in section 21.1, let us contemplate the sources of uncertainty our agents might have to deal with (subsection 21.1.1) and how the agent models need to be extended to cope with that (section 19.4).

21.1.1 Sources of Uncertainty

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27582`.

Sources of Uncertainty in Decision-Making

Where's that d. . . Wumpus?
And where am I, anyway??



- **Non-deterministic actions:**
  - "When I try to go forward in this dark cave, I might actually go forward-left or forward-right."
- **Partial observability with unreliable sensors:**
  - "Did I feel a breeze right now?";
  - "I think I might smell a Wumpus here, but I got a cold and my nose is blocked."
  - "According to the heat scanner, the Wumpus is probably in cell [2,3]."
- **Uncertainty about the domain behavior:**

▷ "Are you *sure* the Wumpus never moves?"

Unreliable Sensors

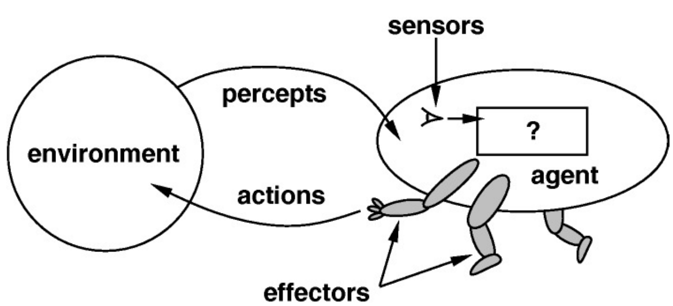
- **Robot Localization:** Suppose we want to support localization using landmarks to narrow down the area.
- **Example 21.1.1.** *If you see the Eiffel tower, then you're in Paris.*
- **Difficulty:** Sensors can be imprecise.
  - Even if a landmark is perceived, we cannot conclude with certainty that the robot is at that location.
  - *This is the half-scale Las Vegas copy, you dummy.*
  - Even if a landmark is *not* perceived, we cannot conclude with certainty that the robot is *not* at that location.
  - *Top of Eiffel tower hidden in the clouds.*
- Only the probability of being at a location increases or decreases.

21.1.2 Recap: Rational Agents as a Conceptual Framework

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27585`.

Agents and Environments

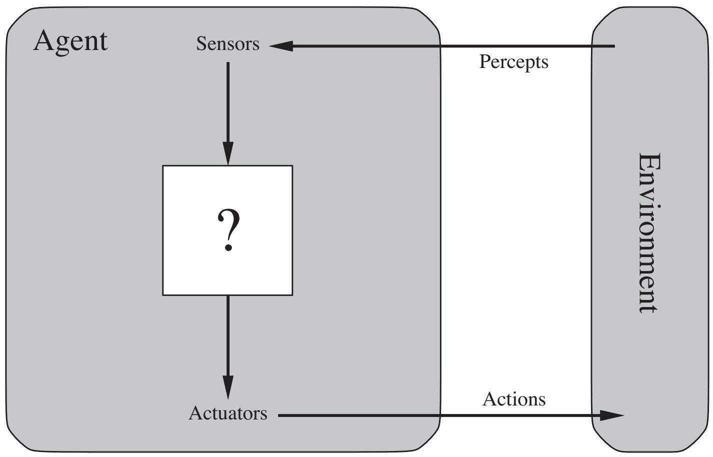
- **Definition 21.1.2.** An agent is anything that
  - perceives its environment via sensors (a means of sensing the environment)
  - acts on it with actuators (means of changing the environment).



- **Example 21.1.3.** Agents include humans, robots, softbots, thermostats, etc.

Agent Schema: Visualizing the Internal Agent Structure

▷ **Agent Schema:** We will use the following kind of schema to visualize the internal structure of an agent:



Different agents differ on the contents of the white box in the center.

Rationality

▷ **Idea:** Try to design agents that are successful! (aka. "do the right thing")

▷ **Definition 21.1.4.** A performance measure is a function that evaluates a sequence of environments.

▷ **Example 21.1.5.** A performance measure for the vacuum cleaner world could

- ▷ award one point per square cleaned up in time T?
- ▷ award one point per clean square per time step, minus one per move?
- ▷ penalize for $> k$ dirty squares?

▷ **Definition 21.1.6.** An agent is called rational, if it chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date.

▷ **Question:** Why is rationality a good quality to aim for?

Consequences of Rationality: Exploration, Learning, Autonomy

▷ **Note:** a rational agent need not be perfect

- ▷ only needs to maximize expected value (rational $\neq$ omniscient)
  - ▷ need not predict e.g. very unlikely but catastrophic events in the future
- ▷ percepts may not supply all relevant information (rational $\neq$ clairvoyant)
  - ▷ if we cannot perceive things we do not need to react to them.
  - ▷ but we may need to try to find out about hidden dangers (exploration)
- ▷ action outcomes may not be as expected (rational $\neq$ successful)
  - ▷ but we may need to take action to ensure that they do (more often) (learning)

▷ **Note:** rational $\leadsto$ exploration, learning, autonomy

▷ **Definition 21.1.7.** An agent is called autonomous, if it does not rely on the prior knowledge about the environment of the designer.

▷ Autonomy avoids fixed behaviors that can become unsuccessful in a changing environment. (anything else would be irrational)

▷ The agent has to learning agentlearn all relevant traits, invariants, properties of the environment and actions.

PEAS: Describing the Task Environment

▷ **Observation:** To design a rational agent, we must specify the task environment in terms of performance measure, environment, actuators, and sensors, together called the PEAS components.

▷ **Example 21.1.8.** When designing an automated taxi:

- ▷ **Performance measure:** safety, destination, profits, legality, comfort, ...
- ▷ **Environment:** US streets/freeways, traffic, pedestrians, weather, ...
- ▷ **Actuators:** steering, accelerator, brake, horn, speaker/display, ...
- ▷ **Sensors:** video, accelerometers, gauges, engine sensors, keyboard, GPS, ...

▷ **Example 21.1.9 (Internet Shopping Agent).**
The task environment:

- ▷ Performance measure: price, quality, appropriateness, efficiency
- ▷ Environment: current and future WWW sites, vendors, shippers
- ▷ Actuators: display to user, follow URL, fill in form
- ▷ Sensors: HTML pages (text, graphics, scripts)

Environment types

▷ **Observation 21.1.10.** *Agent design is largely determined by the type of environment it is intended for.*

▷ **Problem:** There is a vast number of possible kinds of environments in AI.

▷ **Solution:** Classify along a few "dimensions". (independent characteristics)

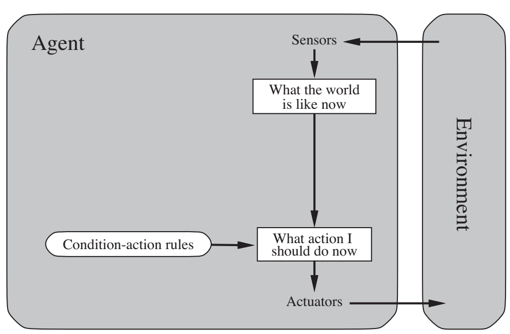
▷ **Definition 21.1.11.** For an agent a we classify the environment e of a by its type, which is one of the following. We call e

1. fully observable, iff the a's sensors give it access to the complete state of the environment at any point in time, else partially observable.
2. deterministic, iff the next state of the environment is completely determined by the current state and a's action, else stochastic.
3. episodic, iff a's experience is divided into atomic episodes, where it perceives and then performs a single action. Crucially the next episode does not depend on previous ones. Non-episodic environments are called sequential.
4. dynamic, iff the environment can change without an action performed by a, else static. If the environment does not change but a's performance measure does, we call e semidynamic.
5. discrete, iff the sets of e's state and a's actions are countable, else continuous.
6. single agent, iff only a acts on e; else multi agent (when must we count parts of e as agents?)

Simple reflex agents

▷ **Definition 21.1.12.** A simple reflex agent is an agent a that only bases its actions on the last percept: so the agent function simplifies to $f_a : \mathcal{P} \to \mathcal{A}$.

▷ **Agent Schema:**

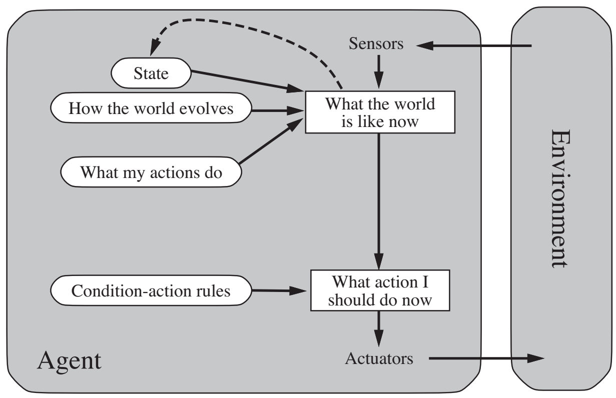


▷ **Example 21.1.13 (Agent Program).**

procedure Reflex−Vacuum−Agent [location,status] **returns** an action
  if status = Dirty **then** . . .

Model-based Reflex Agents: Idea

- **Idea:** Keep track of the state of the world we cannot see in an internal model.
- **Agent Schema:**



Model-based Reflex Agents: Definition

- **Definition 21.1.14.** A model-based agent (also called reflex agent with state) is an agent whose function depends on
  - a world model: a set $\mathcal{S}$ of possible states.
  - a sensor model S that given a state s and percepts determines a new state s'.
  - (optionally) a transition model T, that predicts a new state s'' from a state s' and an action a .
  - An action function f that maps (new) states to actions.

  The agent function is iteratively computed via $e \mapsto f(S(s,e))$.
- **Note:** As different percept sequences lead to different states, so the agent function $f_a : \mathcal{P}^* \to \mathcal{A}$ no longer depends only on the last percept.
- **Example 21.1.15 (Tail Lights Again).** Model-based agents can do the 98 if the states include a concept of tail light brightness.

21.1.3 Agent Architectures based on Belief States

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/29041`.

We are now ready to proceed to environments which can only partially observed and where are our actions are non deterministic. Both sources of uncertainty conspire to allow us only partial knowledge about the world, so that we can only optimize "expected utility" instead of "actual utility" of our actions.

World Models for Uncertainty

- ▷ **Problem:** We do not know with certainty what state the world is in!
- ▷ **Idea:** Just keep track of all the possible states it could be in.
- ▷ **Definition 21.1.16.** A model-based agent has a world model consisting of
  - ▷ a belief state that has information about the possible states the world may be in, and
  - ▷ a sensor model that updates the belief state based on sensor information
  - ▷ a transition model that updates the belief state based on actions.
- ▷ **Idea:** The agent environment determines what the world model can be.
- ▷ In a fully observable, deterministic environment,
  - ▷ we can observe the initial state and subsequent states are given by the actions alone.
  - ▷ thus the belief state is a singleton (we call its member the world state) and the transition model is a function from states and actions to states: a transition function.

That is exactly what we have been doing until now: we have been studying methods that build on descriptions of the "actual" world, and have been concentrating on the progression from atomic to factored and ultimately structured representations. Tellingly, we spoke of "world states" instead of "belief states"; we have now justified this practice in the brave new belief-based world models by the (re-) definition of "world states" above. To fortify our intuitions, let us recap from a belief-state-model perspective.

World Models by Agent Type in AI-1

- ▷ **Note:** All of these considerations only give requirements to the world model. What we can do with it depends on representation and inference.
- ▷ **Search-based Agents:** In a fully observable, deterministic environment
  - ▷ goal-based agent with world state $\widehat{=}$ "current state"
  - ▷ no inference. (goal $\widehat{=}$ goal state from search problem)
- ▷ **CSP-based Agents:** In a fully observable, deterministic environment
  - ▷ goal-based agent withworld state $\widehat{=}$ constraint network,
  - ▷ inference $\widehat{=}$ constraint propagation. (goal $\widehat{=}$ satisfying assignment)
- ▷ **Logic-based Agents:** In a fully observable, deterministic environment
  - ▷ model-based agent with world state $\widehat{=}$ logical formula
  - ▷ inference $\widehat{=}$ e.g. DPLL or resolution. (no decision theory covered in AI-1)
- ▷ **Planning Agents:** In a fully observable, deterministic, environment

- goal-based agent with world state $\widehat{=}$ PL0, transition model $\widehat{=}$ STRIPS,
- inference $\widehat{=}$ state/plan space search. (goal: complete plan/execution)

Let us now see what happens when we lift the restrictions of total observability and determinism.

World Models for Complex Environments

- In a fully observable, but stochastic environment,
  - the belief state must deal with a set of possible states.
  - $\leadsto$ generalize the transition function to a transition relation.
- **Note:** This even applies to online problem solving, where we can just perceive the state. (e.g. when we want to optimize utility)
- In a deterministic, but partially observable environment,
  - the belief state must deal with a set of possible states.
  - we can use transition functions.
  - We need a sensor model, which predicts the influence of percepts on the belief state – during update.
- In a stochastic, partially observable environment,
  - mix the ideas from the last two. (sensor model + transition relation)

Preview: New World Models (Belief) $\leadsto$ new Agent Types

- **Probabilistic Agents:** In a partially observable environment
  - belief state $\widehat{=}$ Bayesian networks,
  - inference $\widehat{=}$ probabilistic inference.
- **Decision-Theoretic Agents:**
  In a partially observable, stochastic environment
  - belief state + transition model $\widehat{=}$ decision networks,
  - inference $\widehat{=}$ maximizing expected utility.
- We will study them in detailthis semester.

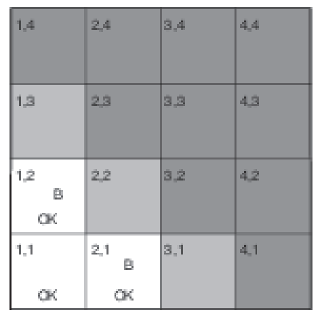
21.1.4 Modeling Uncertainty

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/29043`.

So we have extended the agent's world models to use sets of possible worlds instead of single (deterministic) world states. Let us evaluate whether this is enough for them to survive in the world.

Wumpus World Revisited

- **Recall:** We have updated agents with world/transition models with possible worlds.
- **Problem:** But pure sets of possible worlds are not enough
- **Example 21.1.17 (Beware of the Pit).**
  - We have a maze with pits that are detected in neighbouring squares via breeze (Wumpus and gold will not be assumed now).
  - Where does the agent should go, if there is breeze at (1,2) and (2,1)?
  - **Problem**: (1.3), (2,2), and (3.1) are all unsafe! (there are possible worlds with pits in any of them)



- **Idea:** We need world models that estimate the pit-likelyhood in cells!

Uncertainty and Logic

- **Example 21.1.18 (Diagnosis).** We want to build an expert dental diagnosis system, that deduces the cause (the disease) from the symptoms.
- Can we base this on logic?
- **Attempt 1:** Say we have a toothache. How's about:

$$\forall p.\mathrm{Symptom}(p, \mathrm{toothache}) \Rightarrow \mathrm{Disease}(p, \mathrm{cavity})$$

  - Is this rule correct?
  - No, toothaches may have different causes ("cavity" $\widehat{=}$ "Loch im Zahn").
- **Attempt 2:** So what about this:

$$\forall p.\mathrm{Symptom}(p, \mathrm{toothache}) \Rightarrow (\mathrm{Disease}(p, \mathrm{cavity}) \vee \mathrm{Disease}(p, \mathrm{gingivitis}) \vee \ldots)$$

  - We don't know all possible causes.
  - And we'd like to be able to deduce which causes are more plausible!

Uncertainty and Logic, ctd.

▷ **Attempt 3:** Perhaps a "causal" rule is better?

$$\forall p.\mathsf{Disease}(p, \mathsf{cavity}) \Rightarrow \mathsf{Symptom}(p, \mathsf{toothache})$$

▷ **Question:** Is this rule correct?

▷ **Answer:** No, not all cavities cause toothaches.

▷ **Question:** Does this rule allow to deduce a cause from a symptom?

▷ **Answer:** No, setting $\mathsf{Symptom}(p, \mathsf{toothache})$ to true here has no consequence on the truth of $\mathsf{Disease}(p, \mathsf{cavity})$.

▷ **Note:** If $\mathsf{Symptom}(p, \mathsf{toothache})$ is *false*, we would conclude $\neg\mathsf{Disease}(p, \mathsf{cavity})$ ...which would be incorrect, cf. previous question.

▷ Anyway, this still doesn't allow to compare the plausibility of different causes.

▷ **Summary:** Logic does not allow to weigh different alternatives, and it does not allow to express incomplete knowledge ("cavity does not always come with a toothache, nor vice versa").

Beliefs and Probabilities

▷ **Question:** What do we model with probabilities?

▷ **Answer:** Incomplete knowledge!

- ▷ We are certain, but we *believe to a certain degree* that something is true.
- ▷ Probability $\widehat{=}$ Our degree of belief, given our current knowledge.

▷ **Example 21.1.19 (Diagnosis).**

- ▷ $\mathsf{Symptom}(p, \mathsf{toothache}) \Rightarrow \mathsf{Disease}(p, \mathsf{cavity})$ with 80% probability.
- ▷ But, for any given p, in reality we do, or do not, have cavity: 1 or 0!
- ▷ The "probability" depends on our knowledge!
- ▷ The "80%" refers to the fraction of cavities within the set of all p' that are indistinguishable from p based on our knowledge.
- ▷ If we receive new knowledge (e.g., $\mathsf{Disease}(p, \mathsf{gingivitis})$), the probability changes!

▷ Probabilities represent and measure the uncertainty that stems from lack of knowledge.

How to Obtain Probabilities?

▷ **Assessing probabilities through statistics:**

- ▷ The agent is 90% convinced by its sensor information. (in 9 out of 10 cases, the information is correct)

- Disease(p, cavity) $\Rightarrow$ Symptom(p, toothache) with 80% probability
  :$\widehat{=}$ 8 out of 10 persons with a cavity have toothache.

- **Definition 21.1.20.** The process of estimating a probability P using statistics is called assessing P.
- **Observation:** Assessing even a single P can require huge effort!
- **Example 21.1.21.** The likelihood of making it to the university within 10 minutes.
- **What is probabilistic reasoning?** Deducing probabilities from knowledge about *other* probabilities.
- **Idea:** Probabilistic reasoning determines, based on probabilities that are (relatively) easy to assess, probabilities that are difficult to assess.

21.1.5 Acting under Uncertainty

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/29044`.

Decision-Making Under Uncertainty

- **Example 21.1.22 (Giving a lecture).**
  - **Goal**: Be in HS002 at 10:15 to give a lecture.
  - **Possible plans**:
    - P_1: *Get up at 8:00, leave at 8:40, arrive at 9:00.*
    - P_2: *Get up at 9:50, leave at 10:05, arrive at 10:15.*
  - **Decision**: Both plans are correct, but P_2 succeeds only with probability 50%, and giving a lecture is important, so P_1 is the plan of choice.
- **Example 21.1.23 (Better Example).** Which train to take to Frankfurt airport?

Uncertainty and Rational Decisions

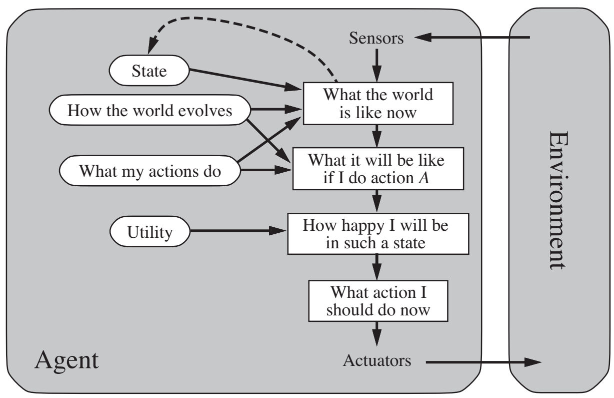
- **Here:** We're only concerned with deducing the likelihood of facts, not with action choice. In general, selecting actions is of course important.
- **Rational Agents:**
  - We have a choice of actions: go to FRA early or go to FRA just in time.
  - These can lead to different solutions with different probabilities.
  - The actions have different costs.
  - The results have different utilities (safe timing/dislike airport food).
- A rational agent chooses the action with the maximum expected utility.

▷ Decision Theory $\widehat{=}$ Utility Theory + Probability Theory.

Utility-based Agents

▷ **Definition 21.1.24.** A utility-based agent uses a world model along with a utility function that models its preferences among the states of that world. It chooses the action that leads to the best expected utility.

▷ **Agent Schema:**



Decision-Theoretic Agent

▷ **Example 21.1.25 (A particular kind of utility-based agent).**

function DT-AGENT(*percept*) **returns** an *action*
  persistent: *belief_state*, probabilistic beliefs about the current state of the world
    action, the agent's action

  update *belief_state* based on *action* and *percept*
  calculate outcome probabilities for actions,
    given action descriptions and current *belief_state*
  select *action* with highest expected utility
    given probabilities of outcomes and utility information
  return *action*

21.1.6 Agenda for this Chapter: Basics of Probability Theory

A Video Nugget covering this subsection can be found at https://fau.tv/clip/id/29046.

Our Agenda for This Topic

- ▷ Our treatment of the topic "probabilistic reasoning" consists of this Chapter and the next.
  - ▷ This Chapter: All the basic machinery at use in Bayesian networks.
  - ▷ chapter 22: Bayesian networks: What they are, how to build them, how to use them.
- ▷ Bayesian networks are the most widespread and successful practical framework for probabilistic reasoning.

Our Agenda for This Chapter

- ▷ **Unconditional Probabilities and Conditional Probabilities:** Which concepts and properties of probabilities will be used?
  - ▷ Mostly a recap of things you're familiar with from school.
- ▷ **Independence and Basic Probabilistic Reasoning Methods:** What simple methods are there to avoid enumeration and to deduce probabilities from other probabilities?
  - ▷ A basic tool set we'll need. (Still familiar from school?)
- ▷ **Bayes' Rule:** What's that "Bayes"? How is it used and why is it important?
  - ▷ The basic insight about how to invert the "direction" of conditional probabilities.
- ▷ **Conditional Independence:** How to capture and exploit complex relations between random variables?
  - ▷ Explains the difficulties arising when using Bayes' rule on multiple evidences. conditional independence is used to ameliorate these difficulties.

21.2 Unconditional Probabilities

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/29047` and `https://fau.tv/clip/id/29048`.

Probabilistic Models

- ▷ **Definition 21.2.1.** A probability theory is an assertion language for talking about possible worlds and an inference method for quantifying the degree of belief in such assertions.
- ▷ **Remark:** Like logic, but for non binary belief degree.

- The possible worlds are
  - mutually exclusive: different possible worlds cannot both be the case and
  - exhaustive: one possible world must be the case.
- This determines the set of possible worlds.
- **Example 21.2.2.** If we roll two (distinguishable) dice with six sides, then we have 36 possible worlds: $(1,1)$, $(2,1)$, $\ldots$, $(6,6)$.
- We will restrict ourselves to a discrete, countable sample space. (others more complicated, less useful in AI)
- **Definition 21.2.3.** A probability model $\langle\Omega, P\rangle$ consists of a countable set Ω of possible worlds called the sample space and a probability function $P \colon \Omega \rightarrow \mathbb{R}$, such that $0 \leq P(\omega) \leq 1$ for all $\omega \in \Omega$ and $\sum_{\omega\in\Omega} P(\omega) = 1$.

Unconditional Probabilities, Random Variables, and Events

- **Definition 21.2.4.** A random variable (also called random quantity, aleatory variable, or stochastic variable) is a variable quantity whose value depends on possible outcomes of unknown variables and processes we do not understand.
- **Definition 21.2.5.** If X is a random variable and x a possible value, we will refer to the fact $X = x$ as an outcome and a set of outcomes as an event. The set of possible outcomes of X is called the domain of X.
- The notation *uppercase "X"* for a random variable, and *lowercase "x"* for one of its values will be used frequently. (following Russel/Norvig)
- **Definition 21.2.6.** Given a random variable X, $P(X = x)$ denotes the prior probability, or unconditional probability, that X has value x in the absence of any other information.
- **Example 21.2.7.** $P(\text{Cavity} = \mathsf{T}) = 0.2$, where Cavity is a random variable whose value is true iff some given person has a cavity.

Types of Random Variables

- **Definition 21.2.8.** We say that a random variable X is finite domain, iff the domain D of X is finite and Boolean, iff $D = \{\mathsf{T}, \mathsf{F}\}$.
- **Note:** In general, random variables can have arbitrary domains. In AI-2, we restrict ourselves to finite domain and Boolean random variables.

▷ **Example 21.2.9.** Some prior probabilities

$$P(\text{Weather} = \text{sunny}) = 0.7$$
$$P(\text{Weather} = \text{rain}) = 0.2$$
$$P(\text{Weather} = \text{cloudy}) = 0.08$$
$$P(\text{Weather} = \text{snow}) = 0.02$$
$$P(\text{Headache} = \mathsf{T}) = 0.1$$

Unlike us, Russel and Norvig live in California . . . :-(:-(

▷ **Convenience Notations:**

- ▷ By convention, we denote Boolean random variables with A, B, and more general finite domain random variables with X, Y.
- ▷ For a Boolean random variable Name, we write name for the outcome Name $= \mathsf{T}$ and $\neg$name for Name $= \mathsf{F}$. (Follows Russel/Norvig as well)

Probability Distributions

▷ **Definition 21.2.10.** The probability distribution for a random variable X, written $\mathbf{P}(X)$, is the vector of probabilities for the (ordered) domain of X.

▷ **Example 21.2.11.** Probability distributions for finite domain and Boolean random variables

$$\mathbf{P}(\text{Headache}) = \langle 0.1, 0.9\rangle$$
$$\mathbf{P}(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02\rangle$$

define the probability distribution for the random variables Headache and Weather.

▷ **Definition 21.2.12.** Given a subset $\mathbf{Z} \subseteq \{X_1, \ldots, X_n\}$ of random variables, an event is an assignment of values to the variables in $\mathbf{Z}$. The joint probability distribution, written $\mathbf{P}(\mathbf{Z})$, lists the probabilities of all events.

▷ **Example 21.2.13.** $\mathbf{P}(\text{Headache}, \text{Weather})$ is

	Headache $= \mathsf{T}$	$Headache = \mathsf{F}$
Weather = sunny	$P(W = \text{sunny} \wedge \text{headache})$	$P(W = \text{sunny} \wedge \neg\text{headache})$
Weather = rain		
Weather = cloudy		
Weather = snow		

The Full Joint Probability Distribution

▷ **Definition 21.2.14.**

Given random variables $\{X_1, \ldots, X_n\}$, an atomic event is an assignment of values to all variables.

▷ **Example 21.2.15.** If A and B are Boolean random variables, then we have four atomic events: $a \wedge b$, $a \wedge \neg b$, $\neg a \wedge b$, $\neg a \wedge \neg b$.

▷ **Definition 21.2.16.**

Given random variables $\{X_1, \ldots, X_n\}$, the full joint probability distribution, denoted $\mathbf{P}(X_1, \ldots, X_n)$, lists the probabilities of all atomic events.

▷ **Observation:**

Given random variables $X_1, \ldots, X_n$ with domains $D_1, \ldots, D_n$, the full joint probability distribution is an n-dimensional array of size $\langle D_1, \ldots, D_n \rangle$.

▷ **Example 21.2.17.** $\mathbf{P}(Cavity, Toothache)$

	toothache	¬toothache
cavity	0.12	0.08
¬cavity	0.08	0.72

▷ **Note:** All atomic events are disjoint (their pairwise conjunctions all are equivalent to F); the sum of all fields is 1 (the disjunction over all atomic events is T).

Probabilities of Propositional Formulae

▷ **Definition 21.2.18.** Given random variables $\{X_1, \ldots, X_n\}$, a proposition is a PL^0 wff over the atoms $X_i = x_i$ where the x_i are values in the domains of X_i.

A function P that maps propositions into $[0,1]$ is a probability measure if

1. $P(\top) = 1$ and
2. for all propositions A, $P(A) = \sum_{e \models A} P(e)$ where e is an atomic event.

▷ Propositions represent sets of atomic events: the interpretations satisfying the formula.

▷ **Example 21.2.19.** $P(\text{cavity} \wedge \text{toothache}) = 0.12$ is the probability that some given person has both a cavity and a toothache. (Note the use of cavity for Cavity $=$ T and toothache for Toothache $=$ T.)

▷ **Notes:**

- Instead of $P(a \wedge b)$, we often write $P(a, b)$.
- Propositions can be viewed as Boolean random variables; we will denote them with A, B as well.

The role of clause 2 in Definition 21.2.18 is for P to "make sense": intuitively, the probability weight of a formula should be the sum of the weights of the interpretations satisfying it. Imagine this was not so; then, for example, we could have $P(A) = 0.2$ and $P(A \wedge B) = 0.8$. The role of 1 here is to "normalize" P so that the maximum probability is 1. (The minimum probability is 0 simply because of 1: the empty sum has weight 0).

Kolmogorov and Negation

- **Theorem 21.2.20 (Kolmogorow).** *A function P that maps propositions into $[0,1]$ is a probability measure if and only if*

  i $P(\top) = 1$ *and*

  ii' *for all propositions A, B: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.*

- **Observation:** We can equivalently replace

  ii for all propositions A, $P(A) = \sum_{I \models A} P(I)$ with Kolmogorow's (*ii'*).

- **Question:** Assume we have

  iii $P(\bot) = 0$.

  How to derive from (*i*), (*ii'*), and (*iii*) that, for all propositions A, $P(\neg a) = 1 - P(a)$?

- **Answer:** reserved for the plenary sessions $\leadsto$ be there!

Believing in Kolmogorov?

- **Reminder 1:** (*i*) $P(\top) = 1$; (*ii'*) $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.
- **Reminder 2:** "Probabilities model our belief."
  - If P represents an objectively observable probability, the axioms clearly make sense.
  - But why should an agent respect these axioms, when modeling its subjective own belief?
- **Question:** Do you believe in Kolmogorow's axioms?
- **Answer:** reserved for the plenary sessions $\leadsto$ be there!

21.3 Conditional Probabilities

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29049`.

Conditional Probabilities: Intuition

- Do probabilities change as we gather new knowledge?
- Yes! Probabilities model our *belief*, thus they depend on our knowledge.
- **Example 21.3.1.** Your "probability of missing the connection train" increases when you are informed that your current train has 30 minutes delay.

▷ **Example 21.3.2.** The "probability of cavity" increases when the doctor is informed that the patient has a toothache.

▷ In the presence of additional information, we can no longer use the unconditional (*prior!*) probabilities.

▷ Given propositions A and B, $P(a|b)$ denotes the conditional probability of a (i.e., $A = \mathsf{T}$) given that all we know is b (i.e., $B = \mathsf{T}$).

▷ **Example 21.3.3.** $P(\text{cavity}) = 0.2$ vs. $P(\text{cavity}|\text{toothache}) = 0.6$.

▷ **Example 21.3.4.** $P(\text{cavity}|\text{toothache} \wedge \neg\text{cavity}) = 0$

Conditional Probabilities: Definition

▷ **Definition 21.3.5.** Given propositions A and B where $P(b) \neq 0$, the conditional probability, or posterior probability, of a given b, written $P(a|b)$, is defined as:

$$P(a|b) := \frac{P(a \wedge b)}{P(b)}$$

▷ **Intuition:** The likelihood of having a and b, within the set of outcomes where we have b.

▷ **Example 21.3.6.** $P(\text{cavity} \wedge \text{toothache}) = 0.12$ and $P(\text{toothache}) = 0.2$ yield $P(\text{cavity}|\text{toothache}) = 0.6$.

Conditional Probability Distributions

▷ **Definition 21.3.7.** Given random variables X and Y, the conditional probability distribution of X given Y, written $\mathbf{P}(X|Y)$, i.e. with a boldface P, is the table of all conditional probabilities of values of X given values of Y.

▷ For sets of variables: $\mathbf{P}(X_1, \ldots, X_n|Y_1, \ldots, Y_m)$.

▷ **Example 21.3.8.** $\mathbf{P}(\text{Weather}|\text{Headache}) =$

	Headache $= \mathsf{T}$	$Headache = \mathsf{F}$
Weather $=$ sunny	$P(W = \text{sunny}|\text{headache})$	$P(W = \text{sunny}|\neg\text{headache})$
Weather $=$ rain		
Weather $=$ cloudy		
Weather $=$ snow		

What is *The probability of sunshine given that I have a headache?*

▷ If you're susceptible to headaches depending on weather conditions, this makes sense. Otherwise, the two variables are independent. (see next section)

21.4 Independence

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29050`.

Working with the Full Joint Probability Distribution

▷ **Example 21.4.1.** Consider the following full joint probability distribution:

	toothache	¬toothache
cavity	0.12	0.08
¬cavity	0.08	0.72

- ▷ How to compute $P(\text{cavity})$?
- ▷ Sum across the row:

$$P(\text{cavity} \wedge \text{toothache}) + P(\text{cavity} \wedge \neg\text{toothache}) = 0.2$$

- ▷ How to compute $P(\text{cavity} \vee \text{toothache})$?
- ▷ Sum across atomic events:

$$P(\text{cavity} \wedge \text{toothache}) + P(\neg\text{cavity} \wedge \text{toothache}) + P(\text{cavity} \wedge \neg\text{toothache}) = 0.28$$

- ▷ How to compute $P(\text{cavity}|\text{toothache})$?
- ▷ $\frac{P(\text{cavity} \wedge \text{toothache})}{P(\text{toothache})}$
- ▷ All relevant probabilities can be computed using the full joint probability distribution, by expressing propositions as disjunctions of atomic events.

Working with the Full Joint Probability Distribution??

▷ **Question:** Is it a good idea to use the full joint probability distribution?

▷ **Answer:** No:

- ▷ Given n random variables with k values each, the full joint probability distribution contains k^n probabilities.
- ▷ Computational cost of dealing with this size.
- ▷ Practically impossible to assess all these probabilities.

▷ **Question:** So, is there a compact way to represent the full joint probability distribution? Is there an efficient method to work with that representation?

▷ **Answer:** Not in general, but it works in many cases. We can work directly with conditional probabilities, and exploit conditional independence.

▷ **Eventually:** Bayesian networks. (First, we do the simple case)

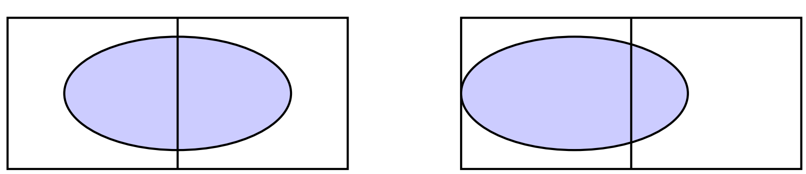
Independence of Events and Random Variables

- **Definition 21.4.2.** Events a and b are independent if $P(a \wedge b) = P(a) \cdot P(b)$.
- Given independent events a and b where $P(b) \neq 0$, we have $P(a|b) = P(a)$.
- *Proof:*
  1. By definition, $P(a|b) = \frac{P(a \wedge b)}{P(b)}$,
  2. which by independence is equal to $\frac{P(a) \cdot P(b)}{P(b)} = P(a)$.
- Similarly, if $P(a) \neq 0$, we have $P(b|a) = P(b)$.
- **Definition 21.4.3.** Random variables X and Y are independent if $\mathbf{P}(X,Y) = \mathbf{P}(X) \otimes \mathbf{P}(Y)$. (System of equations given by outer product!)

Independence (Examples)

- **Example 21.4.4.**
  - $P(\text{Die1} = 6 \wedge \text{Die2} = 6) = 1/36$.
  - $P(W = \text{sunny}|\text{headache}) = P(W = \text{sunny})$ (unless you're weather-sensitive; cf. slide 754)
  - But toothache and cavity are NOT independent.
  - The fraction of "cavity" is higher within "toothache" than within "¬toothache". $P(\text{toothache}) = 0.2$ and $P(\text{cavity}) = 0.2$, but $P(\text{toothache} \wedge \text{cavity}) = 0.12 > 0.04$.
- **Intuition:**

Independent Dependent



Oval independent of rectangle, iff split equally

Illustration: Exploiting Independence

- **Example 21.4.5.** Consider (again) the following full joint probability distribution:

	toothache	¬toothache
cavity	0.12	0.08
¬cavity	0.08	0.72

Adding variable Weather with values sunny, rain, cloudy, snow, the full joint probability distribution contains 16 probabilities.

But your teeth do not influence the weather, nor vice versa!

- ▷ Weather is independent of each of Cavity and Toothache: For all value combinations (c,t) of Cavity and Toothache, and for all values w of Weather, we have $P(c \wedge t \wedge w) = P(c \wedge t) \cdot P(w)$.
- ▷ $\mathbf{P}(\text{Cavity}, \text{Toothache}, \text{Weather})$ can be reconstructed from the separate tables $\mathbf{P}(\text{Cavity}, \text{Toothache})$ and $\mathbf{P}(\text{Weather})$. (8 probabilities)

▷ Independence can be exploited to represent the full joint probability distribution more compactly.

▷ Sometimes, variables are independent only under particular conditions: conditional independence. (see later)

21.5 Basic Probabilistic Reasoning Methods

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29051`.

The Product Rule

▷ **Definition 21.5.1.** The following identity is called the product rule:
Given propositions a and b, $P(a \wedge b) = P(a|b) \cdot P(b)$.

▷ **Note:** The product rule is a direct consequence of the definition of conditional probability

▷ **Example 21.5.2.** $P(\text{cavity} \wedge \text{toothache}) = P(\text{toothache}|\text{cavity}) \cdot P(\text{cavity})$.

▷ If we know the values of $P(a|b)$ and $P(b)$, then we can compute $P(a \wedge b)$.

▷ Similarly, $P(a \wedge b) = P(b|a) \cdot P(a)$.

▷ **Definition 21.5.3.** We use the component wise array product (bold dot)
$\mathbf{P}(X,Y) = \mathbf{P}(X|Y) \boldsymbol{\cdot} \mathbf{P}(Y)$ as a summary notation for the equation system $\mathbf{P}(x_i, y_j) = \mathbf{P}(x_i|y_j) \cdot \mathbf{P}(y_j)$ where i, j range over domain sizes of X and Y.

▷ **Example 21.5.4.** $\mathbf{P}(\text{Weather}, \text{Ache}) = \mathbf{P}(\text{Weather}|\text{Ache}) \boldsymbol{\cdot} \mathbf{P}(\text{Ache})$ is

$$\begin{aligned} P(W = \text{sunny} \wedge \text{ache}) &= P(W = \text{sunny}|\text{ache}) \cdot P(\text{ache}) \\ P(W = \text{rain} \wedge \text{ache}) &= P(W = \text{rain}|\text{ache}) \cdot P(\text{ache}) \\ \cdots &= \cdots \\ P(W = \text{snow} \wedge \neg\text{ache}) &= P(W = \text{snow}|\neg\text{ache}) \cdot P(\neg\text{ache}) \end{aligned}$$

▷ **Note:** The outer product in $\mathbf{P}(X,Y) = \mathbf{P}(X) \boldsymbol{\cdot} \mathbf{P}(Y)$ is just by conincidence, we will use $\mathbf{P}(X,Y) = \mathbf{P}(X) \boldsymbol{\cdot} \mathbf{P}(Y)$ instead.

The component wise array product from Definition 21.5.3 is something that Russell/Norvig (and the literature in general) glosses over and sweeps under the rug. The problem is that it is not a real mathematical operator, that can be defined notation independently, because it depends on the indices in the representation. But the notation is just too convenient to bypass.
It is just a coincidence that we can use the outer product in probability distributions $\mathbf{P}(X,Y) = \mathbf{P}(X)\cdot\mathbf{P}(Y)$. Here, the outer product and component wise array product co-incide.

The Chain Rule

▷ **Lemma 21.5.5 (Chain Rule).** *Given random variables $X_1,\ldots,X_n$, we have*

$$\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1)\cdot\mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1)\cdot\ldots\cdot\mathbf{P}(X_2|X_1)\cdot\mathbf{P}(X_1)$$

This identity is called the chain rule.

▷ **Example 21.5.6.**

$$\begin{aligned} & P(\neg\mathsf{brush}\wedge\mathsf{cavity}\wedge\mathsf{toothache}) \\ = \; & P(\mathsf{toothache}|\mathsf{cavity},\neg\mathsf{brush})\cdot P(\mathsf{cavity},\neg\mathsf{brush}) \\ = \; & P(\mathsf{toothache}|\mathsf{cavity},\neg\mathsf{brush})\cdot P(\mathsf{cavity}|\neg\mathsf{brush})\cdot P(\neg\mathsf{brush}) \end{aligned}$$

▷ *Proof:* Iterated application of the product rule

1. $\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1)\cdot\mathbf{P}(X_{n-1},\ldots,X_1)$ by the product rule.
2. In turn, $\mathbf{P}(X_{n-1},\ldots,X_1) = \mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1)\cdot\mathbf{P}(X_{n-2},\ldots,X_1)$, etc.

▷ **Note:** This works *for any ordering* of the variables.

- ▷ We can recover the probability of atomic events from sequenced conditional probabilities for any ordering of the variables.
- ▷ First of the four basic techniques in Bayesian networks.

Marginalization

▷ Extracting a sub-distribution from a larger joint distribution:

▷ Given sets $\mathbf{X}$ and $\mathbf{Y}$ of random variables, we have:

$$\mathbf{P}(\mathbf{X}) = \sum_{y\in\mathbf{Y}}\mathbf{P}(\mathbf{X},y)$$

where $\sum_{\mathbf{y}\in\mathbf{Y}}$ sums over all possible value combinations of $\mathbf{Y}$.

▷ **Example 21.5.7.** (Note: Equation system!)

$$\mathbf{P}(\mathsf{Cavity}) = \sum_{y \in \mathsf{Toothache}} \mathbf{P}(\mathsf{Cavity}, y)$$
$$P(\mathsf{cavity}) = P(\mathsf{cavity}, \mathsf{toothache}) + P(\mathsf{cavity}, \neg\mathsf{toothache})$$
$$P(\neg\mathsf{cavity}) = P(\neg\mathsf{cavity}, \mathsf{toothache}) + P(\neg\mathsf{cavity}, \neg\mathsf{toothache})$$

Questionnaire: Rules of Probabilistic Reasoning

▷ Say $P(\mathsf{dog}) = 0.4$, $(\neg\mathsf{dog}) \Leftrightarrow \mathsf{cat}$, and $P(\mathsf{likeslasagna}|\mathsf{cat}) = 0.5$.

▷ **Question:** Is $P(\mathsf{likeslasagna} \wedge \mathsf{cat})$ is A: 0.2, B: 0.5, C: 0.475, D: 0.3

▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

▷ **Question:** Can we compute the value of $P(\mathsf{likeslasagna})$, given the above informations?

▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

We now come to a very important technique of computing unknown probabilities, which looks almost like magic. Before we formally define it on the next slide, we will get an intuition by considering it in the context of our dentistry example.

Normalization: Idea

▷ **Problem:** We know $P(\mathsf{cavity} \wedge \mathsf{toothache})$ but don't know $P(\mathsf{toothache})$.

▷ **Step 1:** Case distinction over values of Cavity: ($P(\mathsf{toothache})$ as an unknown)

$$\begin{aligned} P(\mathsf{cavity}|\mathsf{toothache}) &= \frac{P(\mathsf{cavity} \wedge \mathsf{toothache})}{P(\mathsf{toothache})} = \frac{0.12}{P(\mathsf{toothache})} \\ P(\neg\mathsf{cavity}|\mathsf{toothache}) &= \frac{P(\neg\mathsf{cavity} \wedge \mathsf{toothache})}{P(\mathsf{toothache})} = \frac{0.08}{P(\mathsf{toothache})} \end{aligned}$$

▷ **Step 2:** Assuming placeholder $\alpha{:=}1/P(\mathsf{toothache})$:

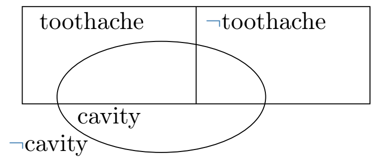
$$\begin{aligned} P(\mathsf{cavity}|\mathsf{toothache}) &= \alpha P(\mathsf{cavity} \wedge \mathsf{toothache}) = \alpha 0.12 \\ P(\neg\mathsf{cavity}|\mathsf{toothache}) &= \alpha P(\neg\mathsf{cavity} \wedge \mathsf{toothache}) = \alpha 0.08 \end{aligned}$$

▷ **Step 3:** Fixing toothache to be true, view $P(\mathsf{cavity} \wedge \mathsf{toothache})$ vs. $P(\neg\mathsf{cavity} \wedge \mathsf{toothache})$ as the "relative weights of $P(\mathsf{cavity})$ vs. $P(\neg\mathsf{cavity})$ within toothache".

Then normalize their summed-up weight to 1:
$1 = \alpha(0.12 + 0.08) \rightsquigarrow \alpha = \frac{1}{0.12+0.08} = \frac{1}{0.2} = 5$

▷ α is a normalization constant scaling the sum of relative weights to 1.

To understand what is going on, consider the situation in the following diagram:



Now consider the areas of $A_1 = \text{toothache}\wedge\text{cavity}$ and $A_2 = \text{toothache}\wedge\neg\text{cavity}$ then $A_1\cup A_2 = \text{toothache}$; this is exactly what we will exploit (see next slide), but we notate it slightly differently in what will be a convenient manner in step 1.

In step 2 we only introduce a convenient placeholder α that makes subsequent argumentation easier.

In step 3, we view A_1 and A_2 as "relative weights"; say that we perceive the left half as "1" (because we already know toothache and don't need to worry about ¬toothache), and we re-normalize to get the desired sum $\alpha A_1 + \alpha A_2 = 1$.

Normalization

▷ **Question:** Say we know $P(\text{likeschappi}\wedge\text{dog}) = 0.32$ and $P(\neg\text{likeschappi}\wedge\text{dog}) = 0.08$. Can we compute $P(\text{likeschappi}|\text{dog})$? (Chappi $\widehat{=}$ popular dog food)

▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

▷ **Question:** So what is $P(\text{likeschappi}|\text{dog})$?

▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

Normalization: Formal

▷ **Definition 21.5.8.** Given a vector $\langle w_1,\ldots,w_k\rangle$ of numbers in $[0,1]$ where $\sum_{i=1}^k w_i\leq 1$, the normalization constant α is $\alpha\langle w_1,\ldots,w_k\rangle := \frac{1}{\sum_{i=1}^k w_i}$.

▷ **Note:** The condition $\sum_{i=1}^k w_i\leq 1$ is needed because these will be relative weights, i.e. case distinction over a subset of all worlds (the one fixed by the knowledge in our conditional probability).

▷ **Example 21.5.9.** $\alpha\langle 0.12, 0.08\rangle = 5\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$.

▷ Given a random variable X and an event $\mathbf{e}$, we have $\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e})$.

Proof:

1. For each value x of X, $P(X = x|\mathbf{e}) = P(X = x \wedge \mathbf{e})/P(\mathbf{e})$.
2. So all we need to prove is that $\alpha = 1/P(\mathbf{e})$.

3. By definition, $\alpha = 1/\sum_x P(X = x \wedge \mathbf{e})$, so we need to prove

$$P(\mathbf{e}) = \sum_x P(X = x \wedge \mathbf{e})$$

which holds by marginalization.

Normalization: Formal

- **Example 21.5.10.** $\alpha\langle P(\text{cavity}\wedge\text{toothache}), P(\neg\text{cavity}\wedge\text{toothache})\rangle = \alpha\langle 0.12, 0.08\rangle$, so $P(\text{cavity}|\text{toothache}) = 0.6$, and $P(\neg\text{cavity}|\text{toothache}) = 0.4$.
- Another way of saying this is: "We use α as a placeholder for $1/P(\mathbf{e})$, which we compute using the sum of relative weights by Marginalization."
- **Computation Rule: Normalization+Marginalization**
  Given "query variable" X, "observed event" $\mathbf{e}$, and "hidden variables" set $\mathbf{Y}$:

$$\mathbf{P}(X|\mathbf{e}) = \alpha \cdot \mathbf{P}(X, \mathbf{e}) = \alpha \cdot (\sum_{\mathbf{y}\in\mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}))$$

- Second of the four basic techniques in Bayesian networks.

21.6 Bayes' Rule

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29053`.

Bayes' Rule

- **Definition 21.6.1 (Bayes' Rule).** Given propositions A and B where $P(a) \neq 0$ and $P(b) \neq 0$, we have:

$$P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)}$$

  This equation is called Bayes' rule.
- *Proof:*
  1. By definition, $P(a|b) = \frac{P(a\wedge b)}{P(b)}$
  2. by the product rule $P(a \wedge b) = P(b|a) \cdot P(a)$ is equal to the claim.
- **Notation:** This is a system of equations!

$$\mathbf{P}(X|Y) = \frac{\mathbf{P}(Y|X) \cdot \mathbf{P}(X)}{\mathbf{P}(Y)}$$

Applying Bayes' Rule

▷ **Example 21.6.2.** Say we know that $P(\text{toothache}|\text{cavity}) = 0.6$, $P(\text{cavity}) = 0.2$, and $P(\text{toothache}) = 0.2$.

We can we compute $P(\text{cavity}|\text{toothache})$: By Bayes' rule, $P(\text{cavity}|\text{toothache}) = \frac{P(\text{toothache}|\text{cavity})\cdot P(\text{cavity})}{P(\text{toothache})} = \frac{0.6\cdot 0.2}{0.2} = 0.6$.

▷ **Ok, but:** Why don't we simply assess $P(\text{cavity}|\text{toothache})$ directly?

▷ **Definition 21.6.3.** We have to take cause and effect into account (cavities cause toothache)

- ▷ $P(\text{toothache}|\text{cavity})$ is causal,
- ▷ $P(\text{cavity}|\text{toothache})$ is diagnostic.

▷ **Intuition:** Causal dependencies are robust over frequency of the causes.

▷ **Example 21.6.4.** If there is a cavity epidemic then $P(\text{cavity}|\text{toothache})$ increases, but $P(\text{toothache}|\text{cavity})$ remains the same. (only depends on how cavities "work")

▷ Also, causal dependencies are often easier to assess.

▷ **Intuition:** "reason about causes in order to draw conclusions about symptoms".

▷ Bayes' rule allows to perform diagnosis (observing a symptom, what is the cause?) based on prior probabilities and causal dependencies.

Extended Example: Bayes' Rule and Meningitis

▷ **Facts known to doctors:**

- ▷ The prior probabilities of meningitis (m) and stiff neck (s) are $P(m) = 0.00002$ and $P(s) = 0.01$.
- ▷ Meningitis causes a stiff neck 70% of the time: $P(s|m) = 0.7$.

▷ **Doctor d uses Bayes' Rule:** $P(m|s) = \frac{P(s|m)\cdot P(m)}{P(s)} = \frac{0.7\cdot 0.00002}{0.01} = 0.0014 \sim \frac{1}{700}$.

- ▷ Even though stiff neck is strongly indicated by meningitis ($P(s|m) = 0.7$)
- ▷ the probability of meningitis in the patient remains small.
- ▷ The prior probability of stiff necks is much higher than that of meningitis.

▷ Doctor d' knows $P(m|s)$ from observation; she does not need Bayes' rule!

▷ Indeed, but what if a meningitis epidemic erupts

▷ Then d knows that $P(m|s)$ grows proportionally with $P(m)$ (d' clueless)

Bayes Rule for Dogs

- Say $P(\text{dog}) = 0.4$, $P(\text{likeschappi}|\text{dog}) = 0.8$, and $P(\text{likeschappi}) = 0.5$.
- **Question:** What is $P(\text{dog}|\text{likeschappi})$?
  A: 0.8 B: 0.64 C: 0.9 D: 0.32?
- **Answer:** reserved for the plenary sessions $\leadsto$ be there!
- **Question:** Is $P(\text{dog}|\text{likeschappi})$ causal or diagnostic?
- **Answer:** reserved for the plenary sessions $\leadsto$ be there!
- **Question:** Is $P(\text{likeschappi}|\text{dog})$ causal or diagnostic?
- **Answer:** reserved for the plenary sessions $\leadsto$ be there!

21.7 Conditional Independence

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29054`.

Bayes' Rule with Multiple Evidence

- **Example 21.7.1.** Say we know from medicinical studies that $P(\text{cavity}) = 0.2$, $P(\text{toothache}|\text{cavity}) = 0.6$, $P(\text{toothache}|\neg\text{cavity}) = 0.1$, $P(\text{catch}|\text{cavity}) = 0.9$, and $P(\text{catch}|\neg\text{cavity}) = 0.2$.

  Now, in case we did observe the symptoms toothache and catch (the dentist's probe catches in the aching tooth), what would be the likelihood of having a cavity? What is $P(\text{cavity}|\text{toothache} \wedge \text{catch})$?

  - **Trial 1:** Bayes' rule

$$P(\text{cavity}|\text{toothache} \wedge \text{catch}) = \frac{P(\text{toothache} \wedge \text{catch}|\text{cavity}) \cdot P(\text{cavity})}{P(\text{toothache} \wedge \text{catch})}$$

  - **Trial 2:** Normalization $\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X, \mathbf{e})$ then Product Rule $\mathbf{P}(X, \mathbf{e}) = \mathbf{P}(\mathbf{e}|X) \cdot \mathbf{P}(X)$, with $X = \text{Cavity}$, $\mathbf{e} = \text{toothache} \wedge \text{catch}$:

$$\begin{aligned}
\mathbf{P}(\text{Cavity}|\text{catch} \wedge \text{toothache}) &= \alpha \cdot \mathbf{P}(\text{toothache} \wedge \text{catch}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity}) \\
P(\text{cavity}|\text{catch} \wedge \text{toothache}) &= \alpha \cdot P(\text{toothache} \wedge \text{catch}|\text{cavity}) \cdot P(\text{cavity}) \\
P(\neg\text{cavity}|\text{catch} \wedge \text{toothache}) &= \alpha P(\text{toothache} \wedge \text{catch}|\neg\text{cavity}) P(\neg\text{cavity})
\end{aligned}$$

Bayes' Rule with Multiple Evidence, ctd.

- $\mathbf{P}(\text{Cavity}|\text{toothache} \wedge \text{catch}) = \alpha\mathbf{P}(\text{toothache} \wedge \text{catch}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity})$
- **Question:** So, is everything fine?

▷ **Answer:** No! We need $\mathbf{P}(\text{toothache} \wedge \text{catch}|\text{Cavity})$, i.e. causal dependencies for all combinations of symptoms! ($\gg 2$, in general)

▷ **Question:** Are Toothache and Catch independent?

▷ **Answer:** No. If a probe catches, we probably have a cavity which probably causes toothache.

▷ **But:** They are conditionally independent given the presence or absence of a cavity!

Conditional Independence

▷ **Definition 21.7.2.** Given sets of random variables $\mathbb{Z}_1$, $\mathbb{Z}_2$, and $\mathbf{Z}$, we say that $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are conditionally independent given $\mathbf{Z}$ if:

$$\mathbf{P}(\mathbb{Z}_1, \mathbb{Z}_2|\mathbf{Z}) = \mathbf{P}(\mathbb{Z}_1|\mathbf{Z}) \cdot \mathbf{P}(\mathbb{Z}_2|\mathbf{Z})$$

We alternatively say that $\mathbb{Z}_1$ is conditionally independent of $\mathbb{Z}_2$ given $\mathbf{Z}$.

▷ **Example 21.7.3.** Catch and Toothache are conditionally independent given Cavity.

- ▷ For cavity: this may cause both, but they don't influence each other.
- ▷ For $\neg$cavity: something else causes catch and/or toothache.

So we have:

$$\mathbf{P}(\text{Toothache}, \text{Catch}|\text{cavity}) = \mathbf{P}(\text{Toothache}|\text{cavity}) \cdot \mathbf{P}(\text{Catch}|\text{cavity})$$
$$\mathbf{P}(\text{Toothache}, \text{Catch}|\neg\text{cavity}) = \mathbf{P}(\text{Toothache}|\neg\text{cavity}) \cdot \mathbf{P}(\text{Catch}|\neg\text{cavity})$$

▷ **Note:** The definition is symmetric regarding the roles of $\mathbb{Z}_1$ and $\mathbb{Z}_2$: Toothache is conditionally independent of Catch given Cavity.

▷ But there may be dependencies *within* $\mathbb{Z}_1$ or $\mathbb{Z}_2$, e.g. $\mathbb{Z}_2 = \{\text{Toothache}, \text{Sleeplessness}\}$.

Conditional Independence, ctd.

▷ If $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are conditionally independent given $\mathbf{Z}$, then $\mathbf{P}(\mathbb{Z}_1|\mathbb{Z}_2, \mathbf{Z}) = \mathbf{P}(\mathbb{Z}_1|\mathbf{Z})$.

▷ *Proof:*

1. By definition, $\mathbf{P}(\mathbb{Z}_1|\mathbb{Z}_2, \mathbf{Z}) = \frac{\mathbf{P}(\mathbb{Z}_1, \mathbb{Z}_2, \mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})}$
2. which by product rule is equal to $\frac{\mathbf{P}(\mathbb{Z}_1, \mathbb{Z}_2|\mathbf{Z}) \cdot \mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})}$
3. which by conditional independence is equal to $\frac{\mathbf{P}(\mathbb{Z}_1|\mathbf{Z}) \cdot \mathbf{P}(\mathbb{Z}_2|\mathbf{Z}) \cdot \mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})}$.
4. Since $\frac{\mathbf{P}(\mathbb{Z}_2|\mathbf{Z}) \cdot \mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})} = 1$ this proves the claim.

▷ **Example 21.7.4.** Using {Toothache} as $\mathbf{Z_1}$, {Catch} as $\mathbf{Z_2}$, and {Cavity} as $\mathbf{Z}$: $\mathbf{P}(\text{Toothache}|\text{Catch}, \text{Cavity}) = \mathbf{P}(\text{Toothache}|\text{Cavity})$.

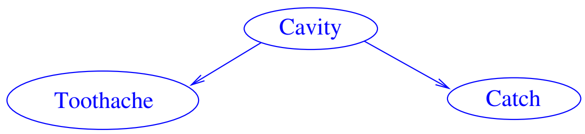
▷ In the presence of conditional independence, we can drop variables from the right-hand side of conditional probabilities.

▷ Third of the four basic techniques in Bayesian networks.

▷ **Last missing technique:** "Capture variable dependencies in a graph"; illustration see next slide, details see chapter 22

Exploiting Conditional Independence: Overview

▷ **1. Graph captures variable dependencies:** (Variables $X_1, \ldots, X_n$)



- ▷ Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X|e)$.
- ▷ Remaining vars: $\mathbf{Y}$.

▷ **2. Normalization+Marginalization:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha \cdot \mathbf{P}(X, \mathbf{e}); \text{ if } \mathbf{Y} \neq \emptyset \text{ then } \mathbf{P}(X|\mathbf{e}) = \alpha \cdot (\sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}))$$

- ▷ A sum over atomic events!

▷ **3. Chain rule:** Order $X_1, \ldots, X_n$ consistently with dependency graph.

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n|X_{n-1}, \ldots, X_1) \cdot \mathbf{P}(X_{n-1}|X_{n-2}, \ldots, X_1) \cdot \ldots \cdot \mathbf{P}(X_1)$$

▷ **4. Exploit Conditional Independence:** Instead of $\mathbf{P}(X_i|X_{i-1}, \ldots, X_1)$, with previous slide we can use $\mathbf{P}(X_i|\text{Parents}(X_i))$.

- ▷ Bayesian networks!

Exploiting Conditional Independence: Example

▷ **1. Graph captures variable dependencies:** (See previous slide.)

- ▷ Given toothache, catch, want $\mathbf{P}(\text{Cavity}|\text{toothache}, \text{catch})$. Remaining vars: $\emptyset$.

▷ **2. Normalization+Marginalization:**

$$\mathbf{P}(\text{Cavity}|\text{toothache}, \text{catch}) = \alpha \cdot \mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch})$$

▷ **3. Chain rule:**
Order X_1 = Cavity, X_2 = Toothache, X_3 = Catch.

$$\mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch}) = \mathbf{P}(\text{catch}|\text{toothache}, \text{Cavity}) \cdot \mathbf{P}(\text{toothache}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity})$$

▷ **4. Exploit Conditional independence:**

Instead of $\mathbf{P}(\text{catch}|\text{toothache}, \text{Cavity})$ use $\mathbf{P}(\text{catch}|\text{Cavity})$.

▷ **Thus:**

$$\begin{aligned}&\mathbf{P}(\text{Cavity}|\text{toothache}, \text{catch})\\ &= \alpha \cdot \mathbf{P}(\text{catch}|\text{Cavity}) \cdot \mathbf{P}(\text{toothache}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity})\\ &= \alpha \cdot \langle 0.9 \cdot 0.6 \cdot 0.2, 0.2 \cdot 0.1 \cdot 0.8\rangle\\ &= \alpha \cdot \langle 0.108, 0.016\rangle\end{aligned}$$

▷ **So:** $\alpha \approx 8.06$ and $\mathbf{P}(\text{cavity}|\text{toothache} \wedge \text{catch}) \approx 0.87$.

Naive Bayes Models

▷ **Definition 21.7.5.** A Bayesian network in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause is called a naive Bayes model or Bayesian classifier.

▷ **Observation 21.7.6.** *In a naive Bayes model, the full joint probability distribution can be written as*

$$\mathbf{P}(cause|effect_1, \ldots, effect_n) = \alpha\langle effect_1, \ldots, effect_n\rangle \cdot \mathbf{P}(cause) \cdot \prod_i \mathbf{P}(effect_i|cause)$$

▷ **Note:** This kind of model is called "naive" since it is often used as a simplifying model if the effects are not conditionally independent after all.

▷ It is also called idiot Bayes model by Bayesian fundamentalists.

▷ In practice, naive Bayes models can work surprisingly well, even when the conditional independence assumption is not true.

▷ **Example 21.7.7.** The dentistry example is a (true) naive Bayes model.

Questionnaire

▷ Consider the random variables X_1 = Animal, X_2 = LikesChappi, and X_3 = LoudNoise, and X_1 has values $\{\text{dog}, \text{cat}, \text{other}\}$, X_2 and X_3 are Boolean.

▷ **Question:** Which statements are correct?

(A) Animal is independent of LikesChappi.

(B) LoudNoise is independent of LikesChappi.

(C) Animal is conditionally independent of LikesChappi given LoudNoise.

(D) LikesChappi is conditionally independent of LoudNoise given Animal.

Think about this intuitively: Given both values for variable X, are the chances of Y being true higher for one of these (fixing value of the third variable where specified)?

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 2 774 2023-09-20

21.8 The Wumpus World Revisited

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29055`.

We will fortify our intuition about naive Bayes models with a variant of the Wumpus world we looked at Example 21.1.17 to understand whether logic was up to the job of guiding an agent in the Wumpus cave.

Wumpus World Revisited

▷ **Example 21.8.1 (The Wumpus is Back).**

- ▷ We have a maze where
  - ▷ pits cause a breeze in neighboring cells
  - ▷ Every cell except $[1,1]$ has a 20% pit probability. (unfair otherwise)
  - ▷ we forget the wumpus and the gold for now (simpler)
- ▷ Where does the agent should go, if there is breeze at $[1,2]$ and $[2,1]$?
- ▷ Pure logical inference can conclude nothing about which square is most likely to be safe!

▷ **Idea:** Let's evaluate our probabilistic reasoning machinery, if that can help!

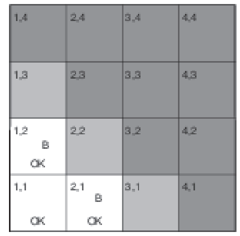
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 2 775 2023-09-20

Wumpus: Probabilistic Model

▷ Boolean random variables (only for the observed squares)

▷ $P_{i,j}$: pit at square $[i,j]$

▷ $B_{i,j}$: breeze at square $[i,j]$



▷ Full joint probability distribution

1. $\mathbf{P}(P_{1,1},\ldots,P_{4,4},B_{1,1},B_{1,2},B_{2,1}) = \mathbf{P}(B_{1,1},B_{1,2},B_{2,1}|P_{1,1},\ldots,P_{4,4}) \cdot \mathbf{P}(P_{1,1},\ldots,P_{4,4})$ (Product Rule)
2. $\mathbf{P}(P_{1,1},\ldots P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j})$ (pits are spread independently)
3. For a particular configuration $p_{1,1},\ldots,p_{4,4}$ with $p_{i,j} \in \{\mathsf{T},\mathsf{F}\}$, n pits, and $P(p_{i,j}) = 0.2$ we have $P(p_{1,1},\ldots,p_{4,4}) = 0.2^n \cdot 0.8^{16-n}$

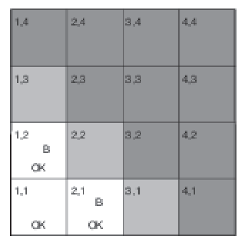
Wumpus: Query and Simple Reasoning

▷ We have evidence in our example:

▷ $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$ and

▷ $\kappa = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

We are interested in answering queries such as $P(P_{1,3}|\kappa,b)$. (pit in $(1,3)$ given evidence)



▷ **Observation:** The answer can be computed by enumeration of the full joint probability distribution.

▷ **Standard Approach:** Let U be the variables $P_{i,j}$ except $P_{1,3}$ and κ, then

$$P(P_{1,3}|\kappa,b) = \sum_{u \in U} \mathbf{P}(P_{1,3},u,\kappa,b)$$

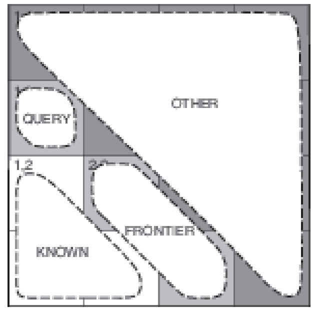
▷ **Problem:** Need to explore all possible values of variables in U ($2^{12} = 4096$ terms!)

▷ Can we do better? (faster; with less computation)

Wumpus: Conditional Independence

▷ **Observation 21.8.2.**

The observed breezes are conditionally independent of the other variables given the known, frontier, and query variables.



- ▷ We split the set of hidden variables into fringe and other variables: $U = F \cup O$ where F is the fringe and O the rest.
- ▷ **Corollary 21.8.3.** $P(b|P_{1,3}, \kappa, U) = P(b|P_{1,3}, \kappa, F)$ (*by conditional independence*)
- ▷ **Now:** let us exploit this formula.

Wumpus: Reasoning

- ▷ We calculate:

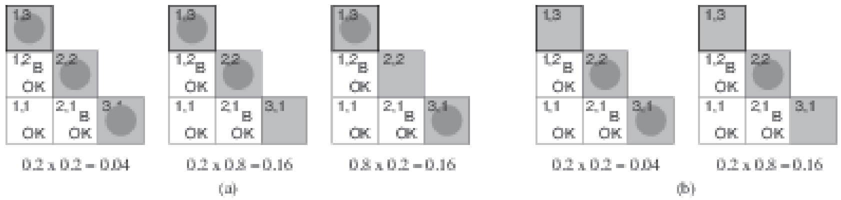
$$\begin{aligned}
P(P_{1,3}|\kappa, b) &= \alpha(\sum_{u \in U} \mathbf{P}(P_{1,3}, u, \kappa, b)) \\
&= \alpha(\sum_{u \in U} \mathbf{P}(b|P_{1,3}, \kappa, u) \cdot \mathbf{P}(P_{1,3}, \kappa, u)) \\
&= \alpha(\sum_{f \in F} \sum_{o \in O} \mathbf{P}(b|P_{1,3}, \kappa, f, o) \cdot \mathbf{P}(P_{1,3}, \kappa, f, o)) \\
&= \alpha(\sum_{f \in F} \mathbf{P}(b|P_{1,3}, \kappa, f) \cdot (\sum_{o \in O} \mathbf{P}(P_{1,3}, \kappa, f, o))) \\
&= \alpha(\sum_{f \in F} \mathbf{P}(b|P_{1,3}, \kappa, f) \cdot (\sum_{o \in O} \mathbf{P}(P_{1,3}) \cdot P(\kappa) \cdot P(f) \cdot P(o))) \\
&= \alpha \mathbf{P}(P_{1,3}) P(\kappa)(\sum_{f \in F} \mathbf{P}(b|P_{1,3}, \kappa, f) \cdot P(f) \cdot (\sum_{o \in O} P(o))) \\
&= \alpha' P(P_{1,3})(\sum_{f \in F} \mathbf{P}(b|P_{1,3}, \kappa, f) \cdot P(f))
\end{aligned}$$

for $\alpha' := \alpha P(\kappa)$ as $\sum_{o \in O} P(o) = 1$.

Wumpus: Solution

- ▷ We calculate using the product rule and conditional independence (see above)
$P(P_{1,3}|\kappa, b) = \alpha' \cdot P(P_{1,3}) \cdot (\sum_{f \in F} \mathbf{P}(b|P_{1,3}, \kappa, f) \cdot P(f))$

▷ Let us explore possible models (values) of Fringe that are F compatible with observation b.



▷ $\mathbf{P}(P_{1,3}|\kappa, b) = \alpha' \cdot \langle 0.2 \cdot (0.04 + 0.16 + 0.16), 0.8 \cdot (0.04 + 0.16) \rangle = \langle 0.31, 0.69 \rangle$

▷ $\mathbf{P}(P_{3,1}|\kappa, b) = \langle 0.31, 0.69 \rangle$ by symmetry

▷ $\mathbf{P}(P_{2,2}|\kappa, b) = \langle 0.86, 0.14 \rangle$ (definitely avoid)

21.9 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29056`.

Summary

▷ Uncertainty is unavoidable in many environments, namely whenever agents do not have perfect knowledge.

▷ Probabilities express the degree of belief of an agent, given its knowledge, into an event.

▷ Conditional probabilities express the likelihood of an event given observed evidence.

▷ Assessing a probability $\widehat{=}$ use statistics to approximate the likelihood of an event.

▷ Bayes' rule allows us to derive, from probabilities that are easy to assess, probabilities that aren't easy to assess.

▷ Given multiple evidence, we can exploit conditional independence.

▷ Bayesian networks (up next) do this, in a comprehensive, computational manner.

Reading: *Chapter 13: Quantifying Uncertainty* [RN03].

Content: Sections 13.1 and 13.2 roughly correspond to my "Introduction" and "Probability Theory Concepts". Section 13.3 and 13.4 roughly correspond to my "Basic Probabilistic Inference". Section 13.5 roughly corresponds to my "Bayes' Rule" and "Multiple Evidence".

In Section 13.6, RN go back to the Wumpus world and discuss some inferences in a probabilistic version thereof.

Overall, the content is quite similar. I have added some examples, have tried to make a few subtle points more explicit, and I indicate already how these techniques will be used in Bayesian networks. RN gives many complementary explanations, nice as additional background reading.

Chapter 22

Probabilistic Reasoning: Bayesian Networks

22.1 Introduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29218`.

Reminder: Our Agenda for This Topic

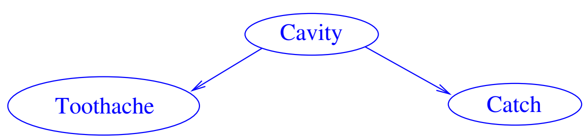
▷ Our treatment of the topic "probabilistic reasoning" consists of this and last section.

- ▷ chapter 21: All the basic machinery at use in Bayesian networks.
- ▷ **This section**: Bayesian networks: What they are, how to build them, how to use them.
- ▷ The most wide-spread and successful practical framework for probabilistic reasoning.

Michael Kohlhase: Artificial Intelligence 2 782 2023-09-20

Reminder: Our Machinery

1. **Graph captures variable dependencies:** (Variables $X_1, \ldots, X_n$)



▷ Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X|e)$. Remaining vars: $\mathbf{Y}$.

2. **Normalization+Marginalization:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$

▷ A sum over atomic events!

3. **Chain rule:** $X_1, \ldots, X_n$ consistently with dependency graph.

$$\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1) \cdot \mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1) \cdot \ldots \cdot \mathbf{P}(X_1)$$

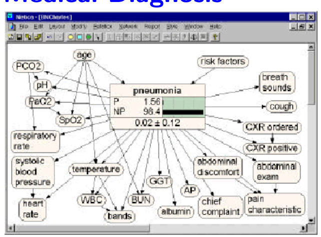
4. **Exploit conditional independence:** Instead of $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$, we can use $\mathbf{P}(X_i|\mathrm{Parents}(X_i))$.

   ▷ Bayesian networks!

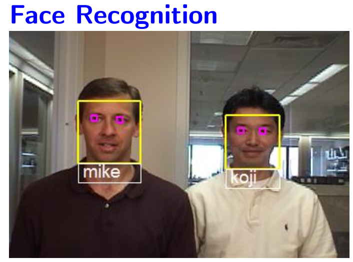
Some Applications

▷ A ubiquitous problem: Observe "symptoms", need to infer "causes".

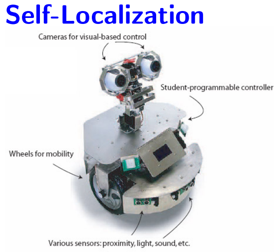
Medical Diagnosis



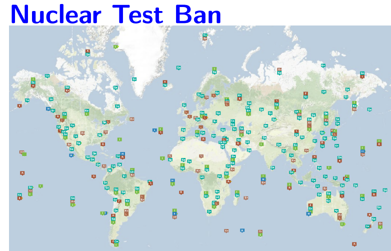
Face Recognition



Self-Localization



Nuclear Test Ban



Our Agenda for This Chapter

▷ **What is a Bayesian Network?:** i.e. What is the syntax?

  ▷ Tells you what Bayesian networks look like.

▷ **What is the Meaning of a Bayesian Network?:** What is the semantics?

  ▷ Makes the intuitive meaning precise.

▷ **Constructing Bayesian Networks:** How do we design these networks? What effect do our choices have on their size?

  ▷ Before you can start doing inference, you need to model your domain.

▷ **Inference in Bayesian Networks:** How do we use these networks? What is the associated complexity?

▷ Inference is our primary purpose. It is important to understand its complexities and how it can be improved.

22.2 What is a Bayesian Network?

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29221`.

What is a Bayesian Network? (Short: BN)

▷ What do the others say?

▷ *"A Bayesian network is a methodology for representing the full joint probability distribution. In some cases, that representation is compact."*

▷ *"A Bayesian network is a graph whose nodes are random variables X_i and whose edges $\langle X_j, X_i \rangle$ denote a direct influence of X_j on X_i. Each node X_i is associated with a conditional probability table (CPT), specifying $\mathbf{P}(X_i|Parents(X_i))$."*

▷ *"A Bayesian network is a graphical way to depict conditional independence relations within a set of random variables."*

▷ A Bayesian network (BN) represents the structure of a given domain. Probabilistic inference exploits that structure for improved efficiency.

▷ BN inference: Determine the distribution of a query variable X given observed evidence $\mathbf{e}$: $\mathbf{P}(X|\mathbf{e})$.

John, Mary, and My Brand-New Alarm

▷ **Example 22.2.1 (From Russell/Norvig).**

▷ I got very valuable stuff at home. So I bought an alarm. Unfortunately, the alarm just rings at home, doesn't call me on my mobile.

▷ I've got two neighbors, Mary and John, who'll call me if they hear the alarm.

▷ The problem is that, sometimes, the alarm is caused by an earthquake.

▷ Also, John might confuse the alarm with his telephone, and Mary might miss the alarm altogether because she typically listens to loud music.

▷ **Question:** Given that both John and Mary call me, what is the probability of a burglary?

John, Mary, and My Alarm: Designing the Network

▷ **Cooking Recipe:**

(1) Design the random variables $X_1, \ldots, X_n$;

(2) Identify their dependencies;

(3) Insert the conditional probability tables $\mathbf{P}(X_i|\text{Parents}(X_i))$.

▷ **Example 22.2.2 (Let's cook!).** Using this recipe on Example 22.2.1, . . .

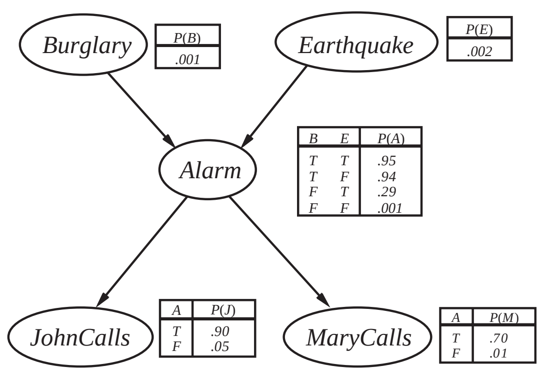
(1) Random variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls.

(2) **Dependencies**: Burglaries and earthquakes are independent. (this is actually debatable ⇝ design decision!)
The alarm might be activated by either. John and Mary call if and only if they hear the alarm. (they don't care about earthquakes)

(3) **Conditional probability tables**: Assess the probabilities, see next slide.

John, Mary, and My Alarm: The Bayesian network

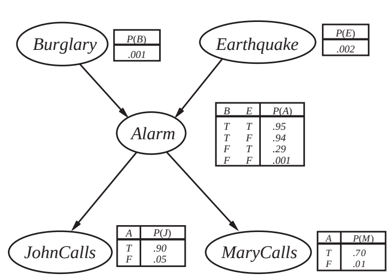
▷ **Example 22.2.3.** Continuing Example 22.2.2 we obtain



▷ **Note:** In each $\mathbf{P}(X_i|\text{Parents}(X_i))$, we show only $\mathbf{P}(X_i = \mathsf{T}|\text{Parents}(X_i))$. We don't show $\mathbf{P}(X_i = \mathsf{F}|\text{Parents}(X_i))$ which is $1 - \mathbf{P}(X_i = \mathsf{T}|\text{Parents}(X_i))$.

The Syntax of Bayesian Networks

▷ To fix the exact definition of Bayesian networks recall the **??**:



▷ **Definition 22.2.4 (Bayesian Network).** Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$, a Bayesian network (also belief network or probabilistic network) is a node labeled DAG $\mathcal{B} := \langle\{X_1, \ldots, X_n\}, E, \mathsf{CPT}\rangle$.

Each X_i is labeled with a function

$$\mathsf{CPT}(X_i) \colon D_i \times \prod_{X_j \in \mathsf{Parents}(X_i)} D_j \to [0,1]$$

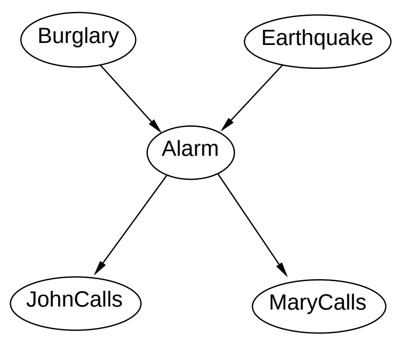
where $\mathsf{Parents}(X_i) := \{X_j \mid (X_j, X_i) \in E\}$ it is called the conditional probability table at X_i.

▷ **Definition 22.2.5.** Bayesian networks and related formalisms summed up under the term graphical models.

22.3 What is the Meaning of a Bayesian Network?

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29223`.

The Semantics of Bayesian Networks: Illustration



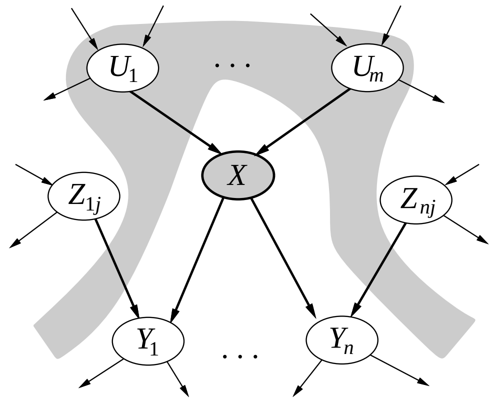
▷ Alarm depends on Burglary and Earthquake.

▷ MaryCalls only depends on Alarm. $\mathbf{P}(\mathsf{MaryCalls} \mid \mathsf{Alarm}, \mathsf{Burglary}) = \mathbf{P}(\mathsf{MaryCalls} \mid \mathsf{Alarm})$

▷ Bayesian networks represent sets of conditional independence assumptions.

The Semantics of Bayesian Networks: Illustration, ctd.

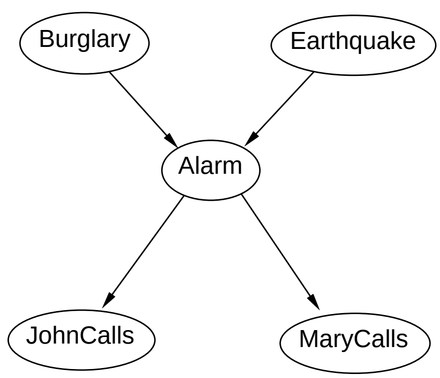
▷ **Observation 22.3.1.** *Each node X in a BN is conditionally independent of its non-descendants given its parents $\mathit{Parents}(X)$.*



▷ **Question:** Why non-descendants of X?

▷ **Intuition:** Given that BNs are acyclic, these are exactly those nodes that *could* have an edge into X.

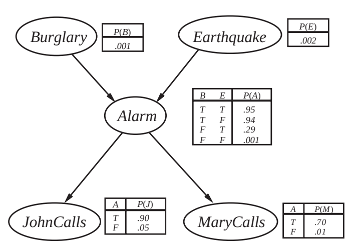
The Semantics of BNs



▷ **Question:** Given the value of Alarm, MaryCalls is independent of?

▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

The Semantics of Bayesian Networks: Formal



▷ **Definition 22.3.2.** Let $\langle\mathcal{X},E\rangle$ be a Bayesian network, $X\in\mathcal{X}$, and E^* the transitive reflexive closure of E, then $\mathrm{NonDesc}(X):=\{Y|(X,Y)\notin E^*\}\backslash\mathrm{Parents}(X)$ is the set of non-descendents of X.

▷ **Definition 22.3.3.** Given a Bayesian network $\mathcal{B}:=\langle\mathcal{X},E\rangle$, we identify $\mathcal{B}$ with the following two assumptions:

(A) $X\in\mathcal{X}$ is conditionally independent of $\mathrm{NonDesc}(X)$ given $\mathrm{Parents}(X)$.

(B) For all values x of $X\in\mathcal{X}$, and all value combinations of $\mathrm{Parents}(X)$, we have $P(x|\mathrm{Parents}(X))=\mathrm{CPT}(x,\mathrm{Parents}(X))$.

Recovering the Full Joint Probability Distribution

▷ **Intuition:** A Bayesian network is a methodology for representing the full joint probability distribution.

▷ **Problem:** How to recover the full joint probability distribution $\mathbf{P}(X_1,\ldots,X_n)$ from $\mathcal{B}:=\langle\{X_1,\ldots,X_n\},E\rangle$?

▷ **Chain Rule:** For *any* ordering $X_1,\ldots,X_n$, we have:

$$\mathbf{P}(X_1,\ldots,X_n)=\mathbf{P}(X_n|X_{n-1},\ldots,X_1)\cdot\mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1)\cdot\ldots\cdot\mathbf{P}(X_1)$$

Choose $X_1,\ldots,X_n$ consistent with $\mathcal{B}$: $X_j\in\mathrm{Parents}(X_i)\rightsquigarrow j<i$.

▷ **Observation 22.3.4 (Exploiting Conditional Independence).** *With Definition 22.3.3 (A), we can use $\mathbf{P}(X_i|\mathit{Parents}(X_i))$ instead of $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$:*

$$\mathbf{P}(X_1,\ldots,X_n)=\prod_{i=1}^{n}\mathbf{P}(X_i|\mathit{Parents}(X_i))$$

The distributions $\mathbf{P}(X_i|\mathit{Parents}(X_i))$ are given by Definition 22.3.3 (B).

▷ Same for atomic events $P(X_1,\ldots,X_n)$.

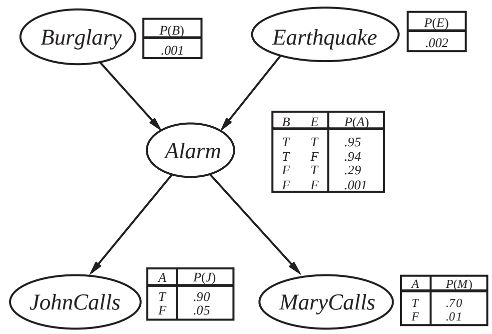
▷ **Observation 22.3.5 (Why "acyclic"?).** *For cyclic $\mathcal{B}$, this does NOT hold, indeed cyclic BNs may be self contradictory.* (*need a consistent ordering*)

Note: If there is a cycle, then any ordering $X_1,\ldots,X_n$ will not be consistent with the BN; so in the chain rule on $X_1,\ldots,X_n$ there comes a point where we have $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$ in the chain but $\mathbf{P}(X_i|\mathrm{Parents}(X_i))$ in the definition of distribution, and $\mathrm{Parents}(X_i)\not\subseteq\{X_{i-1},\ldots,X_1\}$ but then the products are different. So the chain rule can no longer be used to prove that we can reconstruct the full joint probability distribution. In fact, cyclic Bayesian network contain ambiguities (several interpretations possible) and may be self-contradictory (no probability distribution matches the Bayesian network).

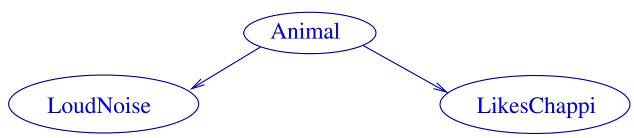
Recovering a Probability for John, Mary, and the Alarm

▷ **Example 22.3.6.** John and Mary called because there was an alarm, but no earthquake or burglary

$$\begin{aligned} P(j,m,a,\neg b,\neg e) &= P(j|a)\cdot P(m|a)\cdot P(a|\neg b,\neg e)\cdot P(\neg b)\cdot P(\neg e) \\ &= 0.9*0.7*0.001*0.999*0.998 \\ &= 0.00062 \end{aligned}$$



Meaning of Bayesian Networks

▷ 

Say $\mathcal{B}$ is the Bayesian network above. Which statements are correct?

(A) Animal is independent of LikesChappi.

(B) LoudNoise is independent of LikesChappi.

(C) Animal is conditionally independent of LikesChappi given LoudNoise.

(D) LikesChappi is conditionally independent of LoudNoise given Animal.

Think about this intuitively: Given both values for variable X, is the chances of Y being true higher for one of these (fixing value of third var where specified)?

▷ **Answers:** reserved for the plenary sessions $\rightsquigarrow$ be there!

22.4 Constructing Bayesian Networks

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/29224` and `https://fau.tv/clip/id/29226`.

Constructing Bayesian Networks

▷ **BN construction algorithm:**

1. Initialize $BN{:=}\langle\{X_1,\ldots,X_n\},E\rangle$ where $E=\emptyset$.

2. Fix any order of the variables, $X_1, \ldots, X_n$.
3. **for** $i := 1, \ldots, n$ **do**
   a. Choose a minimal set $\mathrm{Parents}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ so that

   $$\mathbf{P}(X_i|X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i|\mathrm{Parents}(X_i))$$

   b. For each $X_j \in \mathrm{Parents}(X_i)$, insert (X_j, X_i) into E.
   c. Associate X_i with $\mathrm{CPT}(X_i)$ corresponding to $\mathbf{P}(X_i|\mathrm{Parents}(X_i))$.

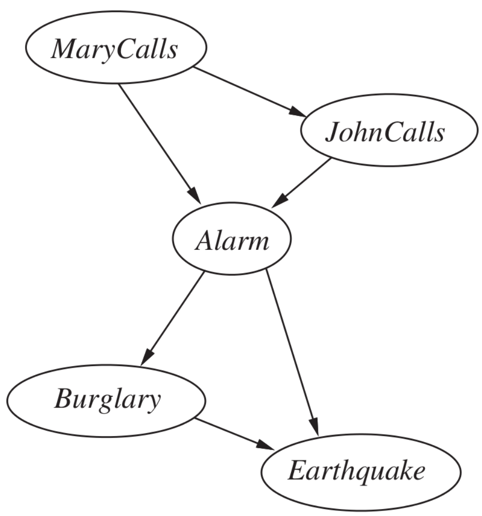
▷ **Attention:** *Which* variables we need to include into $\mathrm{Parents}(X_i)$ depends on what "$\{X_1, \ldots, X_{i-1}\}$" is ... !

▷ **Thus:** The size of the resulting BN depends on the chosen order $X_1, \ldots, X_n$.

▷ **In Particular:** The size of a Bayesian network is *not* a fixed property of the domain. It depends on the skill of the designer.

John and Mary Depend on the Variable Order!

▷ **Example 22.4.1.** MaryCalls, JohnCalls, Alarm, Burglary, Earthquake.



Note: For **??** we try to determine whether – given different value assignments to potential parents – the probability of X_i being true differs? If yes, we include these parents. In the particular case:

1. M to J yes because the common cause may be the alarm.

2. M, J to A yes because they may have heard alarm.

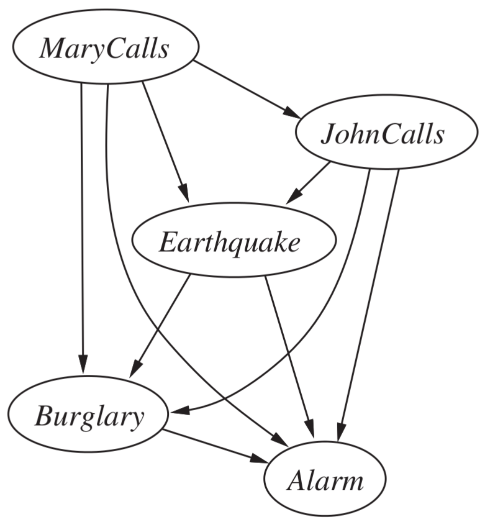
3. A to B yes because if A then higher chance of B.

4. However, M/J to B no because M/J only react to the alarm so if we have the value of A then values of M/J don't provide more information about B.

5. A to E yes because if A then higher chance of E.
6. B to E yes because, if A and not B then chances of E are higher than if A and B.

John and Mary Depend on the Variable Order! Ctd.

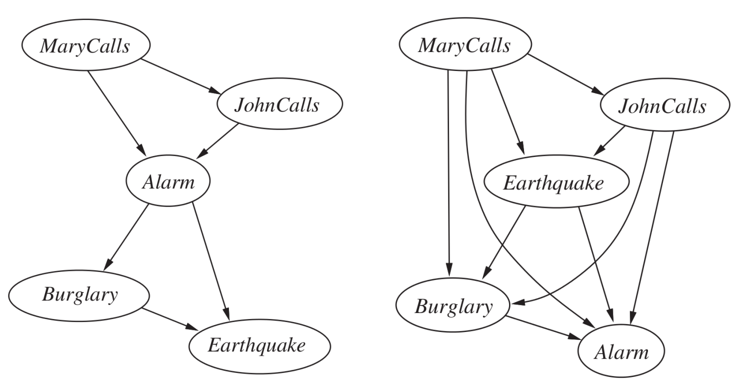
▷ **Example 22.4.2.** MaryCalls, JohnCalls, Earthquake, Burglary, Alarm.



Again: Given different value assignments to potential parents, does the probability of X_i being true differ? If yes, include these parents.

1. M to J as before.
2. M, J to E as probability of E is higher if M/J is true.
3. Same for B; E to B because, given M and J are true, if E is true as well then prob of B is lower than if E is false.
4. $M/J/B/E$ to A because if $M/J/B/E$ is true (even when changing the value of just one of these) then probability of A is higher.

John and Mary, What Went Wrong?



▷ **Intuition:** These BNs link from symptoms to causes! ($\mathbf{P}(\text{Cavity}|\text{Toothache})$) Even though M and J are conditionally independent given A, they are *not* independent without any additional evidence; thus we don't "see" their conditional independence unless we ordered A before M and J! $\leadsto$ We organized the domain in the wrong way here.

We fail to identify many conditional independence relations (e.g., get dependencies between conditionally independent symptoms).

▷ **Also recall:** Conditional probabilities $\mathbf{P}(\text{Symptom}|\text{Cause})$ are more robust and often easier to assess than $\mathbf{P}(\text{Cause}|\text{Symptom})$.

▷ **Rule of Thumb:** We should order causes before symptoms.

Compactness of Bayesian Networks

▷ **Definition 22.4.3.** Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$, the size of $\mathcal{B}:=\langle\{X_1, \ldots, X_n\}, E\rangle$ is defined as

$$\text{size}(\mathcal{B}):=\sum_{i=1}^{n}\#(D_i)\cdot\prod_{X_j\in\text{Parents}(X_i)}\#(D_j)$$

▷ **Note:** $\text{size}(\mathcal{B}) \mathrel{\widehat{=}}$ The total number of entries in the CPTs.

▷ **Note:** Smaller BN $\leadsto$ need to assess less probabilities, more efficient inference.

▷ **Observation 22.4.4.** *Explicit full joint probability distribution has size* $\prod_{i=1}^{n}\#(D_i)$.

▷ **Observation 22.4.5.** *If* $\#(\text{Parents}(X_i)) \leq k$ *for every* X_i, *and* $D_{\max}$ *is the largest random variable domain, then* $\text{size}(\mathcal{B}) \leq n\#(D_{\max})^{k+1}$.

▷ **Example 22.4.6.** For $\#(D_{\max}) = 2$, $n = 20$, $k = 4$ we have $2^{20} = 1048576$ probabilities, but a Bayesian network of size $\leq 20\cdot 2^5 = 640\ldots$!

▷ In the *worst case*, $\text{size}(\mathcal{B}) = n\cdot\prod_{i=1}^{n}\#(D_i)$, namely if every variable depends on all its predecessors in the chosen order.

▷ **Intuition:** BNs are compact – i.e. of small size – if each variable is directly influenced only by few of its predecessor variables.

Constructing Bayesian Networks

▷ **Question:** What is the Bayesian network we get by constructing according to the ordering

1. $X_1 = \text{LoudNoise}, X_2 = \text{Animal}, X_3 = \text{LikesChappi}$?
2. $X_1 = \text{LoudNoise}, X_2 = \text{LikesChappi}, X_3 = \text{Animal}$?

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

22.5 Modeling Simple Dependencies

Representing Conditional Distributions: Deterministic Nodes

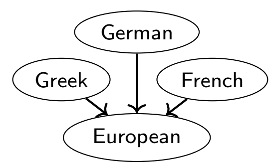
▷ **Problem:** Even if max(Parents) is small, the CPT has 2^k entries. (worst-case)

▷ **Idea:** Usually CPTs follow standard patterns called canonical distributions.

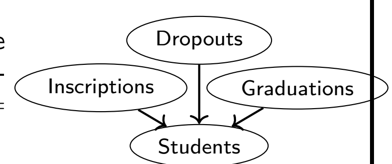
▷ only need to determine pattern and some values.

▷ **Definition 22.5.1.** A node X in a Bayesian network is called deterministic, if its value is completely determined by the values of Parents(X).

▷ **Example 22.5.2 (Logical Dependencies).**
In the network on the right, the node $European$ is deterministic, the CPT corresponds to a logical disjunction, i.e. P(european) = P(greek $\vee$ german $\vee$ french).



▷ **Example 22.5.3 (Numerical Dependencies).**
In the network on the right, the node $Students$ is deterministic, the CPT corresponds to a sum, i.e. $P(S = i - d - g) = P(I = i) + P(D = d) + P(G = g)$.



▷ **Intuition:** Deterministic nodes model direct, causal relationships.

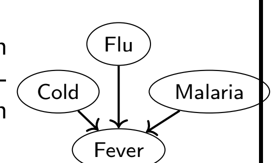
Representing Conditional Distributions: Noisy Nodes

▷ **Problem:** Sometimes, values of nodes are only "almost deterministic". (uncertain, but mostly logical)

▷ **Idea:** Use "noisy" logical relationships. (generalize logical ones softly to $[0,1]$)

▷ **Example 22.5.4 (Inhibited Causal Dependencies).**

In the network on the right, deterministic disjunction for the node Fever is incorrect, since the diseases sometimes fail to develop fever. The causal relation between parent and child is inhibited.



▷ **Assumptions:** We make the following assumptions for modeling Example 22.5.4:

1. Cold, Flu, and Malaria is a complete list of fever causes (add a leak node for the others otherwise).
2. Inhibitions of the parents are independent.

Thus we can model the inhibitions by individual inhibition factors q_d.

▷ **Definition 22.5.5.** The CPT of a noisy disjunction node X in a Bayesian network is given by $P(X_i|\text{Parents}(X_i)) = 1 - \prod_{\{j|X_j=\mathsf{T}\}} q_j$, where the q_i are the inhibition factors of $X_i \in \text{Parents}(X)$.

Representing Conditional Distributions: Noisy Nodes

▷ **Example 22.5.6.** We have the following inhibition factors for Example 22.5.4:

$$\begin{aligned} q_{\text{cold}} &= P(\neg\text{fever}|\text{cold}, \neg\text{flu}, \neg\text{malaria}) = 0.6 \\ q_{\text{flu}} &= P(\neg\text{fever}|\neg\text{cold}, \text{flu}, \neg\text{malaria}) = 0.2 \\ q_{\text{malaria}} &= P(\neg\text{fever}|\neg\text{cold}, \neg\text{flu}, \text{malaria}) = 0.1 \end{aligned}$$

If we model Fever as a noisy disjunction node, then the general rule $P(X_i|\text{Parents}(X_i)) = \prod_{\{j|X_j=\mathsf{T}\}} q_j$ for the CPT gives the following table:

Cold	Flu	Malaria	$P(\text{Fever})$	$P(\neg\text{Fever})$
F	F	F	0.0	1.0
F	F	T	0.9	$\mathbf{0.1}$
F	T	F	0.8	$\mathbf{0.2}$
F	T	T	0.98	$0.02 = 0.2 \cdot 0.1$
T	F	F	0.4	$\mathbf{0.6}$
T	F	T	0.94	$0.06 = 0.6 \cdot 0.1$
T	T	F	0.88	$0.12 = 0.6 \cdot 0.2$
T	T	T	0.988	$0.012 = 0.6 \cdot 0.2 \cdot 0.1$

Representing Conditional Distributions: Noisy Nodes

▷ **Observation 22.5.7.** *In general, noisy logical relationships in which a variable depends on k parents can be described by $\mathcal{O}(k)$ parameters instead of $\mathcal{O}(2^k)$ for the full conditional probability table. This can make assessment (and learning) tractable.*

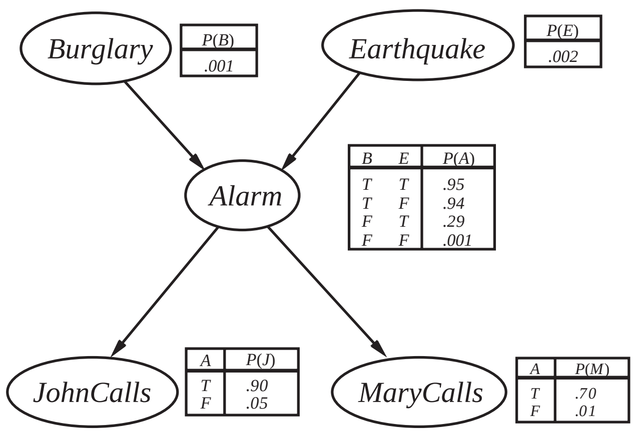
▷ **Example 22.5.8.** The CPCS network [Pra+94] uses noisy-OR and noisy-MAX distributions to model relationships among diseases and symptoms in internal medicine. With 448 nodes and 906 links, it requires only 8,254 values instead of 133,931,430 for a network with full CPTs.

22.6 Inference in Bayesian Networks

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29227`.

Inference for Mary and John

- **Intuition:** Observe evidence variables and draw conclusions on query variables.
- **Example 22.6.1.**



  - What is $\mathbf{P}(\text{Burglary}|\text{johncalls})$?
  - What is $\mathbf{P}(\text{Burglary}|\text{johncalls}, \text{marycalls})$?

Probabilistic Inference Tasks in Bayesian Networks

- **Definition 22.6.2 (Probabilistic Inference Task).** Given random variables $X_1, \ldots, X_n$, a probabilistic inference task consists of a set $\mathbf{X} \subseteq \{X_1, \ldots, X_n\}$ of query variables, a set $\mathbf{E} \subseteq \{X_1, \ldots, X_n\}$ of evidence variables, and an event $\mathbf{e}$ that assigns values to $\mathbf{E}$. We wish to compute the conditional probability distribution $\mathbf{P}(\mathbf{X}|\mathbf{e})$.

  $\mathbf{Y} := \{X_1, \ldots, X_n\} \backslash \mathbf{X} \cup \mathbf{E}$ are the hidden variables.

- **Notes:**
  - We assume that a Bayesian network $\mathcal{B}$ for $X_1, \ldots, X_n$ is given.
  - In the remainder, for simplicity, $\mathbf{X} = \{X\}$ is a singleton.
- **Example 22.6.3.** In $\mathbf{P}(\text{Burglary}|\text{johncalls}, \text{marycalls})$, $X = \text{Burglary}$, $\mathbf{e} = \text{johncalls}, \text{marycalls}$, and $\mathbf{Y} = \{\text{Alarm}, \text{EarthQuake}\}$.

Inference by Enumeration: The Principle (A Reminder!)

▷ **Problem:** Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X|\mathbf{e})$.
Hidden variables: $\mathbf{Y}$.

▷ **1. Bayesian network:** Construct a Bayesian network $\mathcal{B}$ that captures variable dependencies.

▷ **2. Normalization+Marginalization:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e}); \text{ if } \mathbf{Y} \neq \emptyset \text{ then } \mathbf{P}(X|\mathbf{e}) = \alpha(\sum_{\mathbf{y}\in\mathbf{Y}} \mathbf{P}(X,\mathbf{e},\mathbf{y}))$$

▷ Recover the summed-up probabilities $\mathbf{P}(X,\mathbf{e},\mathbf{y})$ from $\mathcal{B}$!

▷ **3. Chain Rule:** Order $X_1,\ldots,X_n$ consistent with $\mathcal{B}$.

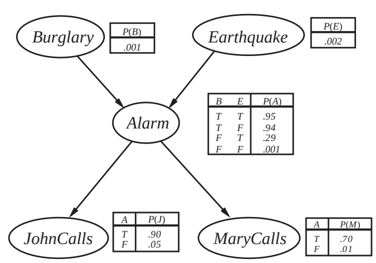
$$\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1) \cdot \mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1) \cdot \ldots \cdot \mathbf{P}(X_1)$$

▷ **4. Exploit conditional independence:** Instead of $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$, use $\mathbf{P}(X_i|\text{Parents}(X_i))$.

▷ Given a Bayesian network $\mathcal{B}$, probabilistic inference tasks can be solved as sums of products of conditional probabilities from $\mathcal{B}$.

▷ Sum over all value combinations of hidden variables.

Inference by Enumeration: John and Mary



▷ **Want:** $\mathbf{P}(\text{Burglary}|\text{johncalls},\text{marycalls})$.
Hidden variables: $\mathbf{Y} = \{\text{Earthquake},\text{Alarm}\}$.

▷ **Normalization+Marginalization:**

$$\mathbf{P}(B|j,m) = \alpha\mathbf{P}(B,j,m) = \alpha(\sum_{v_E}\sum_{v_A}\mathbf{P}(B,j,m,v_E,v_A))$$

▷ **Order:** $X_1 = B$, $X_2 = E$, $X_3 = A$, $X_4 = J$, $X_5 = M$.

▷ **Chain rule and conditional independence:**

$$\mathbf{P}(B|j,m) = \alpha(\sum_{v_E}\sum_{v_A}\mathbf{P}(B) \cdot P(v_E) \cdot \mathbf{P}(v_A|B,v_E) \cdot P(j|v_A) \cdot P(m|v_A))$$

Inference by Enumeration: John and Mary, ctd.

- **Move variables outwards** until we hit the first parent:

$$\mathbf{P}(B|j,m) = \alpha \cdot \mathbf{P}(B) \cdot (\sum_{v_E} P(v_E) \cdot (\sum_{v_A} \mathbf{P}(v_A|B,v_E) \cdot P(j|v_A) \cdot P(m|v_A)))$$

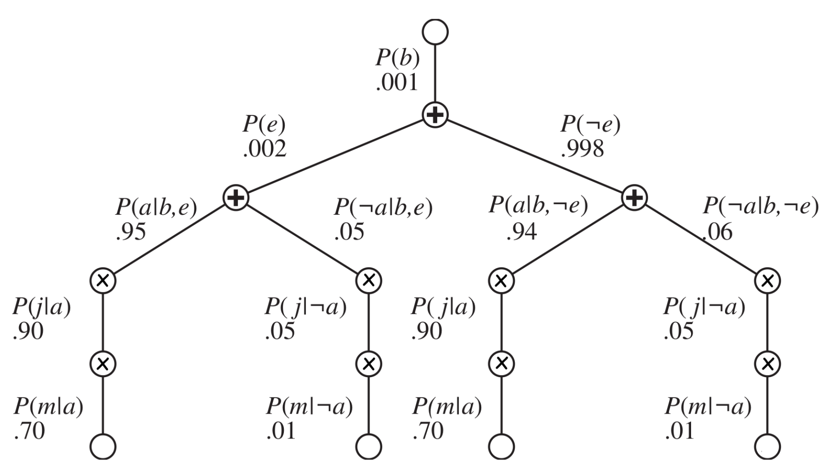
  Note: This step *is* actually done by the pseudo-code, implicitly in the sense that in the recursive calls to enumerate-all we multiply our own prob with all the rest. That is valid because, the variable ordering being consistent, all our parents are already here which is just another way of saying "my own prob does not depend on the variables in the rest of the order".

- The probabilities of the outside-variables multiply the entire "rest of the sum"
- **Chain rule and conditional independence, ctd.:**

$$\begin{aligned}
& \mathbf{P}(B|j,m) \\
= \; & \alpha \mathbf{P}(B)(\sum_{v_E} P(v_E)(\sum_{v_A} \mathbf{P}(v_A|B,v_E)P(j|v_A)P(m|v_A))) \\
= \; & \alpha \cdot P(b) \cdot \left(\begin{array}{l} P(e) \cdot \left(\begin{array}{l} \overbrace{P(a|b,e)P(j|a)P(m|a)}^{a} \\ + \underbrace{P(\neg a|b,e)P(j|\neg a)P(m|\neg a)}_{\neg a} \end{array} \right) \Bigg\} e \\ + \; P(\neg e) \cdot \left(\begin{array}{l} \overbrace{P(a|b,\neg e)P(j|a)P(m|a)}^{a} \\ + \underbrace{P(\neg a|b,\neg e)P(j|\neg a)P(m|\neg a)}_{\neg a} \end{array} \right) \Bigg\} \neg e \end{array} \right) \\
= \; & \alpha \langle 0.00059224, 0.0014919 \rangle \approx \langle 0.284, 0.716 \rangle
\end{aligned}$$

This computation can be viewed as a "search tree"! (see next slide)

The Evaluation of $P(b|j,m)$, as a "Search Tree"



- Inference by enumeration = a tree with "sum nodes" branching over values of hidden variables, and with non-branching "multiplication nodes".

Inference by Enumeration: Variable Elimination

- **Inference by Enumeration:**
  - Evaluates the tree in a depth-first manner.
  - space complexity: linear in the number of variables.
  - time complexity: exponential in the number of hidden variables, e.g. $\mathcal{O}(2^{\#(\mathbf{Y})})$ in case these variables are Boolean.
- Can we do better than this?
- **Definition 22.6.4.** Variable elimination is a BNI algorithm that avoids
  - repeated computation, and (see below)
  - irrelevant computation. (see below)
- In some special cases, variable elimination runs in polynomial time.

Variable Elimination: Sketch of Ideas

- **Avoiding repeated computation:** Evaluate expressions from right to left, storing all intermediate results.
- For query $P(B|j,m)$:
  1. CPTs of BN yield *factors* (probability tables):

$$\mathbf{P}(B|j,m) = \alpha \cdot \underbrace{\mathbf{P}(B)}_{\mathbf{f}_1(B)} \cdot (\sum_{v_E} \underbrace{P(v_E)}_{\mathbf{f}_2(E)} \sum_{v_A} \underbrace{\mathbf{P}(v_A|B,v_E)}_{\mathbf{f}_3(A,B,E)} \cdot \underbrace{P(j|v_A)}_{\mathbf{f}_4(A)} \cdot \underbrace{P(m|v_A)}_{\mathbf{f}_5(A)})$$

  2. Then the computation is performed in terms of *factor product* and *summing out variables* from factors:

$$\mathbf{P}(B|j,m) = \alpha \cdot \mathbf{f}_1(B) \cdot (\sum_{v_E} \mathbf{f}_2(E) \cdot (\sum_{v_A} \mathbf{f}_3(A,B,E) \cdot \mathbf{f}_4(A) \cdot \mathbf{f}_5(A)))$$

- **Avoiding irrelevant computation:** Repeatedly remove hidden variables that are leaf nodes.
- For query $P(\text{JohnCalls}|\text{burglary})$:

$$\mathbf{P}(J|b) = \alpha \cdot P(b) \cdot (\sum_{v_E} P(v_E) \cdot (\sum_{v_A} P(v_A|b,v_E) \cdot \mathbf{P}(J|v_A) \cdot (\sum_{v_M} P(v_M|v_A))))$$

▷ The rightmost sum equals 1 and can be dropped.

The Complexity of Exact Inference

- ▷ **Definition 22.6.5.** A graph G is called singly connected, or a polytree (otherwise multiply connected), if there is at most one undirected path between any two nodes in G.
- ▷ **Theorem 22.6.6 (Good News).** *On singly connected Bayesian networks, variable elimination runs in polynomial time.*
- ▷ Is our BN for Mary & John a polytree? (Yes.)
- ▷ **Theorem 22.6.7 (Bad News).** *For multiply connected Bayesian networks, probabilistic inference is* **#P***-hard.* (**#P** *is harder than* **NP***, i.e.* **NP** $\subseteq$ **#P***)*
- ▷ **So?:** Life goes on ... In the hard cases, if need be we can throw exactitude to the winds and approximate.
- ▷ **Example 22.6.8.** Sampling techniques as in MCTS.

22.7 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29228`.

Summary

- ▷ Bayesian networks (BN) are a wide-spread tool to model uncertainty, and to reason about it. A BN represents conditional independence relations between random variables. It consists of a graph encoding the variable dependencies, and of conditional probability tables (CPTs).
- ▷ Given a variable order, the BN is small if every variable depends on only a few of its predecessors.
- ▷ Probabilistic inference requires to compute the probability distribution of a set of query variables, given a set of evidence variables whose values we know. The remaining variables are hidden.
- ▷ Inference by enumeration takes a BN as input, then applies Normalization+Marginalization, the chain rule, and exploits conditional independence. This can be viewed as a tree search that branches over all values of the hidden variables.
- ▷ Variable elimination avoids unnecessary computation. It runs in polynomial time for poly-tree BNs. In general, exact probabilistic inference is **#P**-hard. Approximate probabilistic inference methods exist.

Topics We Didn't Cover Here

▷ **Inference by sampling**: A whole zoo of methods for doing this exists.

▷ **Clustering**: Pre-combining subsets of variables to reduce the running time of inference.

▷ **Compilation to SAT**: More precisely, to "weighted model counting" in CNF formulas. Model counting extends DPLL with the ability to determine the number of satisfying interpretations. Weighted model counting allows to define a mass for each such interpretation (= the probability of an atomic event).

▷ **Dynamic BN**: BN with one slice of variables at each "time step", encoding probabilistic behavior over time.

▷ **Relational BN**: BN with predicates and object variables.

▷ **First-order BN**: Relational BN with quantification, i.e. probabilistic logic. E.g., the BLOG language developed by Stuart Russel and co-workers.

Reading:

- *Chapter 14: Probabilistic Reasoning* of [RN03].
  - Section 14.1 roughly corresponds to my "What is a Bayesian Network?".
  - Section 14.2 roughly corresponds to my "What is the Meaning of a Bayesian Network?" and "Constructing Bayesian Networks".The main change I made here is to *define* the semantics of the BN in terms of the conditional independence relations, which I find clearer than RN's definition that uses the reconstructed full joint probability distribution instead.
  - Section 14.4 roughly corresponds to my "Inference in Bayesian Networks". RN give full details on variable elimination, which makes for nice ongoing reading.
  - Section 14.3 discusses how CPTs are specified in practice.
  - Section 14.5 covers approximate sampling-based inference.
  - Section 14.6 briefly discusses relational and first-order BNs.
  - Section 14.7 briefly discusses other approaches to reasoning about uncertainty.

  All of this is nice as additional background reading.

Chapter 23

Making Simple Decisions Rationally

23.1 Introduction

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30338`.

Decision Theory

▷ **Definition 23.1.1.** Decision theory investigates decision problems, i.e. how an agent a deals with choosing among actions based on the desirability of their outcomes given by a real-valued utility function u on states $s \in S$: i.e. $u: S \to \mathbb{R}$.

▷ **Wait:** Isn't that what we did in section 6.1?

▷ **Yes, but:** Now we do it for stochastic (i.e. non-deterministic), partially observable environments.

▷ **Recall:** We call the environment of an agent A

- ▹ fully observable, iff the A's sensors give it access to the complete state of the environment at any point in time, else partially observable.
- ▹ deterministic, iff the next state of the environment is completely determined by the current state and A's action, else stochastic.
- ▹ episodic, iff A's experience is divided into atomic episodes, where it perceives and then performes a single action. Crucially the next episode does not depend on previous ones. Non-episodic environments are called sequential.

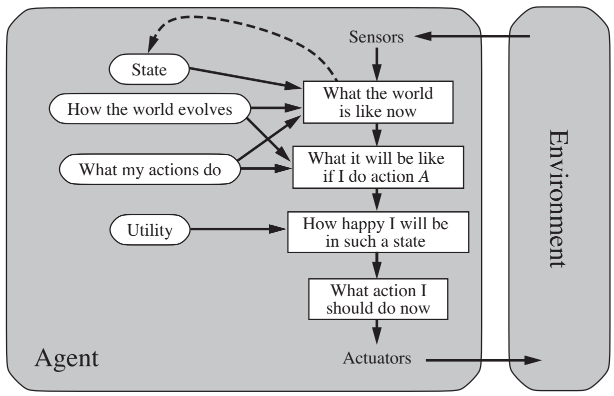
▷ **For now:** We restrict ourselves to episodic decision theory, which deals with choosing among actions based on the desirability of their *immediate* outcomes. (no need to treat time explicitly)

▷ **Later:** We will study sequential decision problems, where the agent's utility depends on a sequence of decisions. (chapter 25)

Utility-based Agents

▷ **Definition 23.1.2.** A utility-based agent uses a world model along with a utility function that models its preferences among the states of that world. It chooses the action that leads to the best expected utility.

- **Agent Schema:**



Maximizing Expected Utility (Ideas)

- **Definition 23.1.3 (MEU principle for Rationality).** We call an action rational if it maximizes expected utility. An utility-based agent is called rational, iff it always chooses a rational action.
- **Note:** An agent can be entirely rational (consistent with MEU) without ever representing or manipulating utilities and probabilities.
- **Example 23.1.4.** A simple reflex agent for tic tac toe based on a perfect lookup table is rational if we take "winning/drawing in n steps" as the utility function.
- **But we will see:** An observer can construct a value function V by observing the agent's preferences. (even if the agent does not know V)
- **Before we go on:** Let us understand how this meshes with AI-1 content!

World Models by Agent Type in AI-1

- **Note:** All of these considerations only give requirements to the world model. What we can do with it depends on representation and inference.
- **Search-based Agents:** In a fully observable, deterministic environment
  - goal-based agent with world state $\widehat{=}$ "current state"
  - no inference. (goal $\widehat{=}$ goal state from search problem)
- **CSP-based Agents:** In a fully observable, deterministic environment

- ▹ goal-based agent withworld state $\widehat{=}$ constraint network,
- ▹ inference $\widehat{=}$ constraint propagation. (goal $\widehat{=}$ satisfying assignment)

▷ **Logic-based Agents:** In a fully observable, deterministic environment

- ▹ model-based agent with world state $\widehat{=}$ logical formula
- ▹ inference $\widehat{=}$ e.g. DPLL or resolution. (no decision theory covered in AI-1)

▷ **Planning Agents:** In a fully observable, deterministic, environment

- ▹ goal-based agent with world state $\widehat{=}$ PL0, transition model $\widehat{=}$ STRIPS,
- ▹ inference $\widehat{=}$ state/plan space search. (goal: complete plan/execution)

Recap: Episodic Decision Theory in AI-1

▷ **Observation:** In AI-1, the environment of an agent was

- ▹ fully observable, so the sensor model S and transition model T are functions.
- ▹ deterministic, so we know the result states of all actions.
- ▹ The "expected utility" of an action is just the utility of the result state.

So maximizing expected utility is easy: $\mathrm{EU}(a) = U(T(S(s,e),a))$, where e the most recent percept, and s the current state.

▷ **Intuition:** Utility functions can be handled like heuristics were in AI-1.

Preview: Episodic Decision Theory in AI-1/2

▷ **Problem:** In AI-2, the environment may be

- ▹ partially observable, so we do not know the "current state".
- ▹ stochastic, so we do not know the result state of an action.

▷ **Idea:** Treat the result state of an action a as a random variable R_a.

- ▹ Study $P(R_a = s'|a, \mathbf{e})$ given evidence observations $\mathbf{e}$.
- ▹ The expected utility $\mathrm{EU}(a)$ of an action a is then

$$\mathrm{EU}(a|\mathbf{e}) = \sum_{s'} P(R_a = s'|a, \mathbf{e}) \cdot U(s')$$

▷ **Intuitively:** A formalization of what it means to "do the right thing".

▷ **Hooray:** This solves all of the AI problem. (in principle)

▷ **Problem:** There is a long long way towards an operationalization. (do that now)

Outline of this Chapter

- ▷ Rational preferences
- ▷ Utilities and Money
- ▷ Multi attribute utility functions
- ▷ Decision networks
- ▷ Value of information

23.2 Rational Preferences

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/32525`.

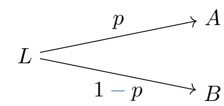
Preferences in Deterministic Environments

- ▷ **Problem:** We cannot directly measure utility of (or satisfaction/happiness in) a state.
- ▷ **Example 23.2.1.** I have to decide whether to go to class today (or sleep in). What is the utility of this lecture. (obviously 42)
- ▷ **Idea:** We can let people/agents choose between two states! (subjective preference)
- ▷ **Example 23.2.2.** *Give me your cell-phone or I will give you a bloody nose.* $\rightsquigarrow$ To make a decision in a deterministic environment, the agent must determine whether it prefers a state without phone to one with a bloody nose?
- ▷ **Definition 23.2.3.** Given states A and B (we call them prizes) and agent can express preferences of the form
  - ▷ $A \succ B$ A preferred over B
  - ▷ $A \sim B$ indifference between A and B
  - ▷ $A \succeq B$ B not preferred over A

Preferences in Non-Deterministic Environments

- ▷ **Problem:** In nondeterministic environments we do not have full information about the states we choose between.
- ▷ **Example 23.2.4 (Airline Food).** *Do you want chicken or pasta* (but we cannot see through the tin foil)

▷ **Definition 23.2.5.** Given prizes A_i and probabilities p_i with $\sum_{i=1}^{n} p_i = 1$, a lottery $[p_1,A_1;\ldots;p_n,A_n]$ represents the result of a nondeterministic action that can have outcomes A_i with prior probability p_i. For the binary case, we use $[p,A;1-p,B]$.



▷ We extend preferences to include lotteries for nondeterministic environments.

Rational Preferences

▷ **Idea:** Preferences of a rational agent must obey constraints:
Rational preferences $\rightsquigarrow$ behavior describable as .

▷ **Definition 23.2.6.** We call a set $\succ$ of preferences rational, iff the following constraints hold:

Orderability	$A \succ B \vee B \succ A \vee A \sim B$
Transitivity	$A \succ B \wedge B \succ C \Rightarrow A \succ C$
Continuity	$A \succ B \succ C \Rightarrow (\exists p.[p,A;1-p,C] \sim B)$
Substitutability	$A \sim B \Rightarrow [p,A;1-p,C] \sim [p,B;1-p,C]$
Monotonicity	$A \succ B \Rightarrow (p > q) \Leftrightarrow [p,A;1-p,B] \succ [q,A;1-q,B]$
Decomposability	$[p,A;1-p,[q,B;1-q,C]] \sim [p,A\,;\,((1-p)q),B\,;\,((1-p)(1-q)),C]$

The rationality constraints can be understood as follows:

Orderability: $A \succ B \vee B \succ A \vee A \sim B$ Given any two prizes or lotteries, a rational agent must either prefer one to the other or else rate the two as equally preferable. That is, the agent cannot avoid deciding. Refusing to bet is like refusing to allow time to pass.

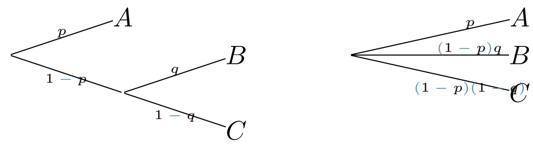
Transitivity: $A \succ B \wedge B \succ C \Rightarrow A \succ C$

Continuity: $A \succ B \succ C \Rightarrow (\exists p.[p,A;1-p,C] \sim B)$ If some lottery B is between A and C in preference, then there is some probability p for which the rational agent will be indifferent between getting B for sure and the lottery that yields A with probability p and C with probability $1-p$.

Substitutability: $A \sim B \Rightarrow [p,A;1-p,C] \sim [p,B;1-p,C]$ If an agent is indifferent between two lotteries A and B, then the agent is indifferent between two more complex lotteries that are the same except that B is substituted for A in one of them. This holds regardless of the probabilities and the other outcome(s) in the lotteries.

Monotonicity: $A \succ B \Rightarrow (p > q) \Leftrightarrow [p,A;1-p,B] \succ [q,A;1-q,B]$ Suppose two lotteries have the same two possible outcomes, A and B. If an agent prefers A to B, then the agent must prefer the lottery that has a higher probability for A (and vice versa).

Decomposability: $[p,A;1-p,[q,B;1-q,C]] \sim [p,A\,;\,((1-p)q),B\,;\,((1-p)(1-q)),C]$ Compound lotteries can be reduced to simpler ones using the laws of probability. This has been called the “no fun in gambling” rule because it says that two consecutive lotteries can be compressed into a single equivalent lottery: the following two are equivalent:

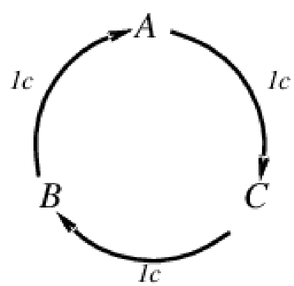


Rational preferences contd.

▷ Violating the rationality constraints from Definition 23.2.6 leads to self-evident irrationality.

▷ **Example 23.2.7.** An agent with intransitive preferences can be induced to give away all its money:

- ▷ If $B \succ C$, then an agent who has C would pay (say) 1 cent to get B
- ▷ If $A \succ B$, then an agent who has B would pay (say) 1 cent to get A
- ▷ If $C \succ A$, then an agent who has A would pay (say) 1 cent to get C



23.3 Utilities and Money

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30341` and `https://fau.tv/clip/id/30342`.

Ramseys Theorem and Value Functions

▷ **Theorem 23.3.1.** *(Ramsey, 1931; von Neumann and Morgenstern, 1944)*

Given a rational set of preferences there exists a real valued function U such that $U(A) \geq U(B)$, iff $A \succeq B$ and $U([p_1,S_1;\ldots;p_n,S_n]) = \sum_i p_i U(S_i)$

▷ This is an existence theorem, uniqueness not guaranteed.

▷ **Note:** Agent behavior is *invariant* w.r.t. positive linear transformations, i.e. an agent with utility function $U'(x) = k_1 U(x) + k_2$ where $k_1 > 0$ behaves exactly like one with U.

▷ **Observation:** With deterministic prizes only (no lottery choices), only a total ordering on prizes can be determined.

▷ **Definition 23.3.2.** We call a total ordering on states a value function or ordinal utility function.

Maximizing Expected Utility (Definitions)

▷ We first formalize the notion of expectation of a random variable.

▷ **Definition 23.3.3.** Given a probability model $\langle\Omega, P\rangle$ and a $X\colon\Omega\to\mathbb{R}_0^+$ a random variable, then $E(X) := \sum_{x\in\Omega} P(X = x)\cdot x$ is called the expected value (or expectation) of X.

▷ **Idea:** Apply this idea to get the expected utility of an action, this is stochastic:

- ▹ In partially observable environments, we do not know the current state.
- ▹ In nondeterministic environments, we cannot be sure of the result of an action.

▷ **Definition 23.3.4.** Let $\mathcal{A}$ be an agent with a set Ω of states and a utility function $U\colon\Omega\to\mathbb{R}_0^+$, then for each action a, we define a random variable R_a whose values are the results of performing a in the current state.

▷ **Definition 23.3.5.** The expected utility $\mathrm{EU}(a|\mathbf{e})$ of an action a (given evidence $\mathbf{e}$) is

$$\mathrm{EU}(a|\mathbf{e}) := \sum_{s\in\Omega} P(R_a = s|a, \mathbf{e})\cdot U(s)$$

Utilities

▷ **Intuition:** Utilities map states to real numbers.

▷ **Question:** Which numbers exactly?

▷ **Definition 23.3.6 (Standard approach to assessment of human utilities).** Compare a given state A to a standard lottery L_p that has

- ▹ "best possible prize" $u_\top$ with probability p
- ▹ "worst possible catastrophe" $u_\perp$ with probability $1-p$

adjust lottery probability p until $A \sim L_p$. Then $U(A) = p$.

▷ **Example 23.3.7.** Choose $u_\top \mathrel{\widehat{=}}$ current state, $u_\perp \mathrel{\widehat{=}}$ instant death

pay \$30 $\sim L$ $\xrightarrow{0.999999}$ continue as before

pay \$30 $\sim L$ $\xrightarrow{0.000001}$ instant death

Measuring Utility

▷ **Definition 23.3.8.** Normalized utilities: $u_\top = 1$, $u_\bot = 0$.

▷ **Definition 23.3.9.** Micromorts: one millionth chance of instant death.

▷ Micromorts are useful for Russian roulette, paying to reduce product risks, etc.

▷ **Problem:** What is the value of a micromort?

▷ **Ask them directly:** What would you pay to avoid playing Russian roulette with a million-barrelled revolver? (very large numbers)

▷ **But their behavior suggests a lower price:**

- ▷ Driving in a car for 370km incurs a risk of one micromort;
- ▷ Over the life of your car – say, $150,000$km that's 400 micromorts.
- ▷ People appear to be willing to pay about $10,000$€ more for a safer car that halves the risk of death. ($\leadsto$ 25€ per micromort)

▷ This figure has been confirmed across many individuals and risk types.

▷ Of course, this argument holds only for small risks. Most people won't agree to kill themselves for 25M€.

▷ **Definition 23.3.10.** QALYs: quality adjusted life years

▷ **Application:** QALYs are useful for medical decisions involving substantial risk.

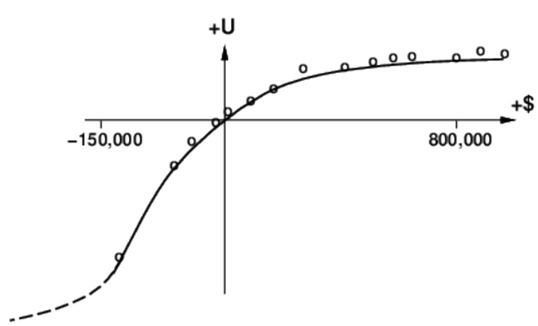
Money vs. Utility

▷ Money does *not* behave as a utility function should.

▷ Given a lottery L with expected monetary value EMV(L), usually $U(L) < U(\mathrm{EMV}(L))$, i.e., people are risk averse.

▷ **Utility curve:** For what probability p am I indifferent between a prize x and a lottery $[p,M\$;1-p,0\$]$ for large numbers M?

▷ Typical empirical data, extrapolated with risk prone behavior for debitors:



▷ **Empirically:** Comes close to the logarithm on the positive numbers.

23.4 Multi-Attribute Utility

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30343` and `https://fau.tv/clip/id/30344`.

In this section we will make the ideas introduced above more practical. The discussion above conceived utility functions as functions on atomic states, which were good enough for introducing the theory. But when we build decision models for utility-based agent we want to characterize states by attributes that are already random variables in the Bayesian network we use to represent the belief state. For factored states, the utility function can be expressed as a multivariate function on attribute values.

Utility Functions on Attributes

▷ **Recap:** So far we understand how to obtain utility functions $u\colon S \to \mathbb{R}$ on states $s \in S$ from (rational) preferences.

▷ **But** in a partially observable, stochastic environment, we cannot know the current state. (utilities/preferences useless?)

▷ **Idea:** Base utilities/preferences on random variables that we can model.

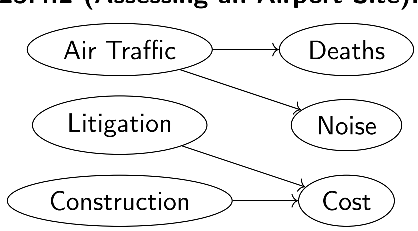
▷ **Definition 23.4.1.** Let $X_1, \ldots, X_n$ be random variables with domains $D_1, \ldots, D_n$. Then we call a function $u\colon D_1 \times \ldots \times D_n \to \mathbb{R}$ a (multi-attribute) utility function on attributes $X_1, \ldots, X_n$.

▷ **Intuition:** Given a probabilistic belief state that includes random variables $X_1, \ldots, X_n$, and a utility function on attributes $X_1, \ldots, X_n$, we can still maximize expected utility! (MEU principle)

▷ **Preview:** Understand multi attribute utility functions and use Bayesian networks as representations of belief states.

Multi-Attribute Utility: Example

▷ **Example 23.4.2 (Assessing an Airport Site).**



▷ Attributes: Deaths, Noise, Cost.

▷ **Question**: What is U(Deaths, Noise, Cost) for a projected airport?

▷ How can complex utility functions be assessed from preference behaviour?

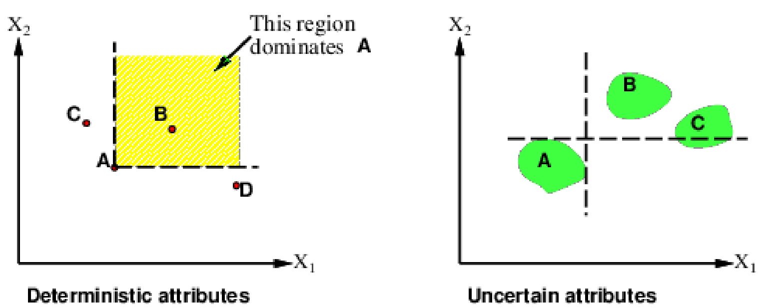
▷ **Idea 1:** Identify conditions under which decisions can be made without complete identification of $U(X_1, \ldots, X_n)$.

▷ **Idea 2:** Identify various types of *independence* in preferences and derive consequent canonical forms for $U(X_1, \ldots, X_n)$.

Strict Dominance

▷ Typically define attributes such that U is monotone in each argument. (wlog. growing)

▷ **Definition 23.4.3.** Choice B strictly dominates choice A iff $X_i(B) \geq X_i(A)$ for all i (and hence $U(B) \geq U(A)$)



▷ **Observation:** Strict dominance seldom holds in practice (life is difficult) but is useful for narrowing down the field of contenders.

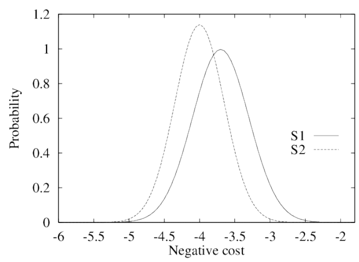
▷ For uncertain attributes strict dominance is even more unlikely.

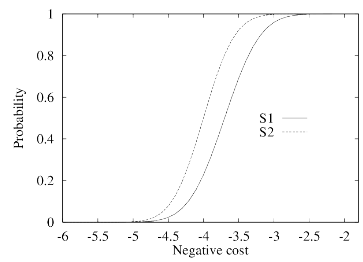
Stochastic Dominance

▷ **Definition 23.4.4.** A distribution p_2 stochastically dominates distribution p_1 iff the cummulative distribution of p_2 strictly dominates that for p_1 for all t, i.e.

$$\int_t^{-\infty} p_1(x)dx \leq \int_t^{-\infty} p_2(x)dx$$

▷ **Example 23.4.5.** Even if the distributions (left) overlap considerably the cummulative distribution (right) strictly dominates.





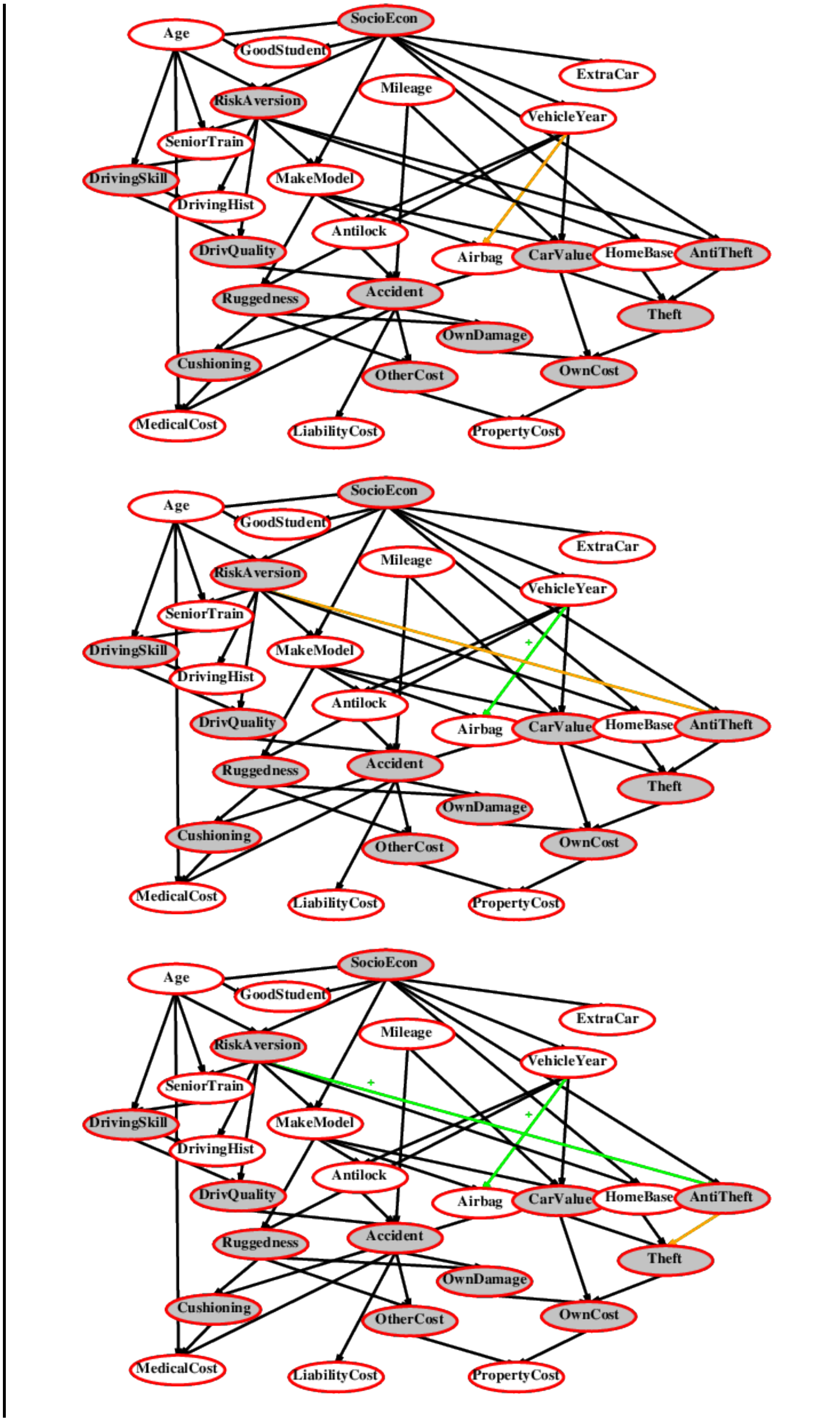
Stochastic dominance contd.

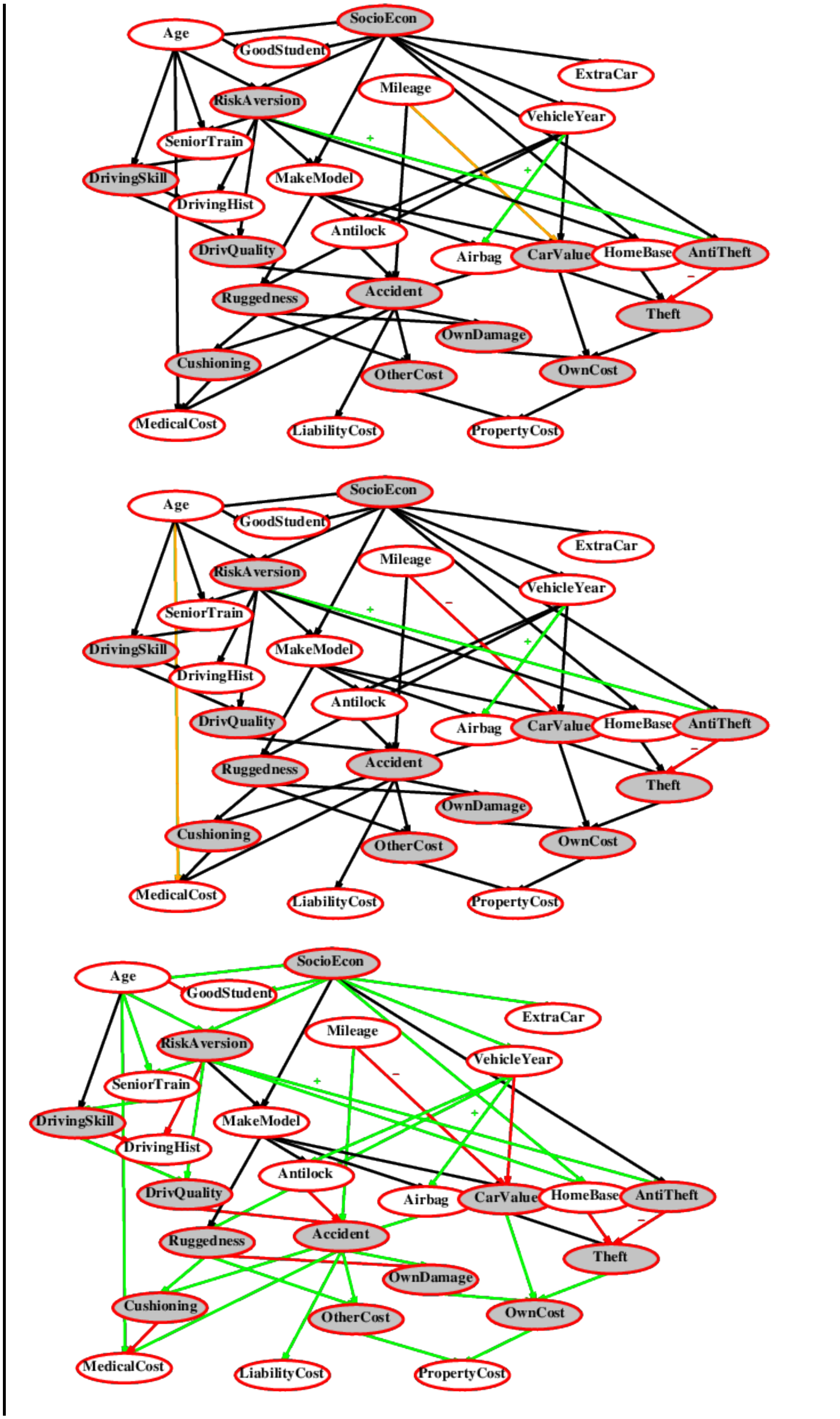
- **Observation 23.4.6.** *If U is monotone in x, then A_1 with outcome distribution p_1 stochastically dominates A_2 with outcome distribution p_2:*

$$\int_{\infty}^{-\infty} p_1(x)U(x)dx \geq \int_{\infty}^{-\infty} p_2(x)U(x)dx$$

- Multi-attribute case: stochastic dominance on all attributes $\rightsquigarrow$ optimal.
- **Observation:** Stochastic dominance can often be determined without exact distributions using *qualitative* reasoning.
- **Example 23.4.7 (Construction cost increases with distance).** If airport location S_1 is closer to the city than S_2 $\rightsquigarrow$ S_1 stochastically dominates S_2 on cost.q
- **Example 23.4.8.** Injury increases with collision speed.
- **Idea:** Annotate Bayesian networks with stochastic dominance information.
- **Definition 23.4.9.** $X \xrightarrow{+} Y$ (X positively influences Y) means that $\mathbf{P}(Y|X_1,\mathbf{z})$ stochastically dominates $\mathbf{P}(Y|X_2,\mathbf{z})$ for every value $\mathbf{z}$ of Y's other parents $\mathbf{Z}$ and all X_1 and X_2 with $X_1 \geq X_2$.

Label the arcs + or – for influence in a Bayesian Network


SocioEcon
Age
GoodStudent
ExtraCar
Mileage
RiskAversion
VehicleYear
SeniorTrain
DrivingSkill
MakeModel
DrivingHist
Antilock
DrivQuality
Airbag
CarValue
HomeBase
AntiTheft
Ruggedness
Accident
Theft
OwnDamage
Cushioning
OwnCost
OtherCost
MedicalCost
LiabilityCost
PropertyCost


SocioEcon
Age
GoodStudent
ExtraCar
Mileage
RiskAversion
VehicleYear
SeniorTrain
DrivingSkill
MakeModel
DrivingHist
Antilock
DrivQuality
Airbag
CarValue
HomeBase
AntiTheft
Ruggedness
Accident
Theft
OwnDamage
Cushioning
OwnCost
OtherCost
MedicalCost
LiabilityCost
PropertyCost

We have seen how we can do inference with attribute-based utility functions, let us consider the computational implications. We observe that we have just replaced one evil – exponentially many states (in terms of the attributes) – by another – exponentially many parameters of the utility functions.

Wo we do what we always do in AI-2: we look for structure in the domain, do more theory to be able to turn such structures into computationally improved representations.

Preference Structure and Multi-Attribute Utility

- ▷ **Observation 23.4.10.** *With n attributes with d values each ⇝ need d^n parameters for the utility function $U(X_1, \ldots, X_n)$.* (worst case)
- ▷ **Assumption:** Preferences of real agents have much more structure.
- ▷ **Approach:** Identify regularities and prove representation theorems based on these:

$$U(X_1, \ldots, X_n) = F(f_1(X_1), \ldots, f_n(f_n)X_n)$$

  where F is simple, e.g. addition.
- ▷ Note the similarity to Bayesian networks that decompose the full joint probability distribution.

Preference structure: Deterministic

- ▷ **Recall:** In deterministic environments an agent has a value function.
- ▷ **Definition 23.4.11.** X_1 and X_2 preferentially independent of X_3 iff preference between $\langle x_1, x_2, z\rangle$ and $\langle x'_1, x'_2, z\rangle$ does not depend on z.
- ▷ **Example 23.4.12.** E.g., ⟨Noise, Cost, Safety⟩: are preferentially independent ⟨20,000 suffer, 4.6 G$, 0.06 deaths/mpm⟩ vs.⟨70,000 suffer, 4.2 G$, 0.06 deaths/mpm⟩
- ▷ **Theorem 23.4.13 (Leontief, 1947).** *If every pair of attributes is preferentially independent of its complement, then every subset of attributes is preferentially independent of its complement: mutual preferential independence.*
- ▷ **Theorem 23.4.14 (Debreu, 1960).** *Mutual preferential independence implies that there is an additive value function: $V(S) = \sum_i V_i(X_i(S))$, where V_i is a value function referencing just one variable X_i.*
- ▷ Hence assess n single-attribute functions. (often a good approximation)
- ▷ **Example 23.4.15.** The value function for the airport decision might be

$$V(noise, cost, deaths) = -noise \cdot 10^4 - cost - deaths \cdot 10^{12}$$

Preference structure: Stochastic

- ▷ Need to consider preferences over lotteries and real utility functions(not just value functions)
- ▷ **Definition 23.4.16.** $\mathbf{X}$ is utility independent of $\mathbf{Y}$ iff preferences over lotteries in $\mathbf{X}$ do not depend on particular values in $\mathbf{Y}$.
- ▷ **Definition 23.4.17.** A set $\mathbf{X}$ is mutually utility independent (MUI), iff each subset is utility independent of its complement.
- ▷ **Example 23.4.18.** Arguably, the attributes of Example 23.4.2 are MUI.
- ▷ **Theorem 23.4.19.** *For MUI sets of attributes, there is a multiplicative utility function:* *[Kee74]*
- ▷ **Definition 23.4.20.** We "define" a multiplicative utility function by example: For three attributes we have:

$$U = k_1U_1+k_2U_2+k_3U_3+k_1k_2U_1U_2+k_2k_3U_2U_3+k_3k_1U_3U_1+k_1k_2k_3U_1U_2U_3$$

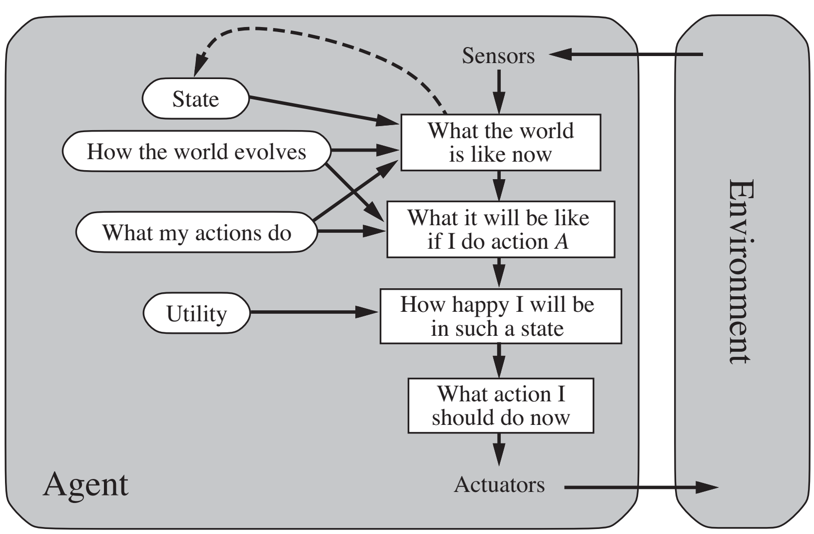
- ▷ **System Support:** Routine procedures and software packages for generating preference tests to identify various canonical families of utility functions.

23.5 Decision Networks

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30345`.

Now that we understand multi-attribute utility functions, we can complete our design of a utility-based agent, which we now recapitulate as a refresher.

Utility-Based Agents (Recap)



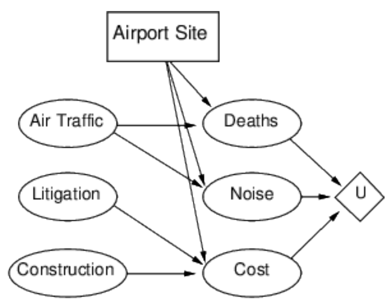
▷

As we already use Bayesian networks for the belief state of an utility-based agent, integrating utilities and possible actions into the network suggests itself naturally. This leads to the notion of a decision network.

Decision networks

▷ **Definition 23.5.1.** A decision network is a Bayesian network with added action nodes and utility nodes (also called value node) that enable decision making.

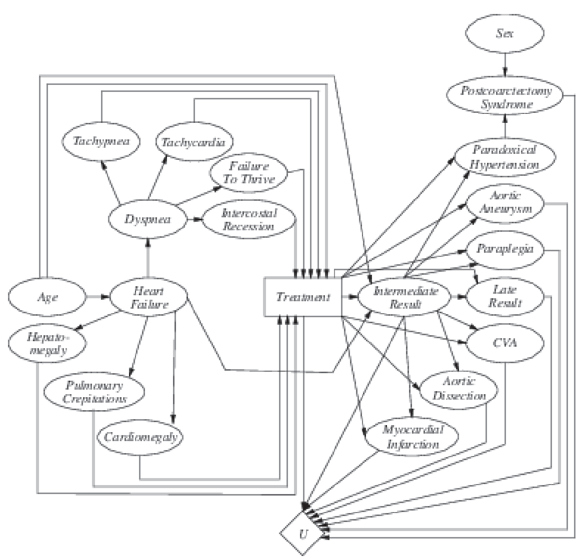
▷ **Example 23.5.2 (Choosing an Airport Site).**



▷ **Algorithm:** For each value of action node
compute expected value of utility node given action, evidence
Return MEU action (via argmax)

Decision Networks: Example

▷ **Example 23.5.3 (A Decision-Network for Aortic Coarctation).** from [Luc96]



Knowledge Eng. for Decision-Theoretic Expert Systems

▷ **Question:** How do you create a model like the one from Example 23.5.3?

▷ **Answer:** By a systematic process of the form: (after [Luc96])

1. **Create a causal model**: a graph with nodes for symptoms, disorders, treatments, outcomes, and their influences (edges).
2. **Simplify to a qualitative decision model**: remove random variables not involved in treatment decisions.
3. **Assign probabilities**: (⇝ Bayesian network)
   e.g. from patient databases, literature studies, or the expert's subjective assessments
4. **Assign utilities**. (e.g. in QALYs or micromorts)
5. **Verify and refine the model** wrt. a gold standard given by experts
   e.g. refine by "running the model backwards" and compare with the literature.
6. **Perform sensitivity analysis**: (important step in practice)
   - ▷ is the optimal treatment decision robust against small changes in the parameters? (if yes ⇝ great! if not, collect better data)

23.6 The Value of Information

Video Nuggets covering this section can be found at https://fau.tv/clip/id/30346 and https://fau.tv/clip/id/30347.

So far we have tacitly been concentrating on actions that directly affect the environment. We will now come to a type of action we have hypothesized in the beginning of the course, but have completely ignored up to now: information gathering actions.

What if we do not have all information we need?

▷ **It is Well-Known:** One of the most important parts of decision making is knowing what questions to ask.

▷ **Example 23.6.1 (Medical Diagnosis).**

- ▷ We do not expect a doctor to already know the results of the diagnostic tests when the patient comes in.
- ▷ Tests are often expensive, and sometimes hazardous. (directly or by delaying treatment)
- ▷ **Therefore**: Only test, if
  - ▷ knowing the results lead to a significantly better treatment plan,
  - ▷ information from test results is not drowned out by a-priori likelihood.

▷ **Definition 23.6.2.** Information value theory enables the agent to make decisions on information gathering rationally.

▷ **Intuition:** Simple form of sequential decision making. (action only impacts belief state).

▷ **Intuition:** With the new information, we can base the action choice to the *actual* information, rather than the average.

Value of Information by Example

▷ **Idea:** Compute value of acquiring each possible piece of evidence.

▷ **We will see:** This can be done directly from a decision network.

▷ **Example 23.6.3 (Buying Oil Drilling Rights).** There are n blocks of rights, exactly one has oil, worth k€, in particular

- ▹ Prior probabilities $1/n$ each, mutually exclusive.
- ▹ Current price of each block is k/n€.
- ▹ "Consultant" offers accurate survey of block 3. What's a fair price?

▷ **Solution:** Compute expected value of information $\widehat{=}$ expected value of best action given the information minus expected value of best action without information.

▷ **Example 23.6.4 (Oil Drilling Rights contd.).**

- ▹ Survey may say *oil in block 3 with probability* $1/n$ $\rightsquigarrow$ buy block 3 for k/n€ make profit of $(k-k/n)$€.
- ▹ Survey may say *no oil in block 3 with probability* $(n-1)/n$ $\rightsquigarrow$ buy another block, make profit of $k/(n-1)-k/n$€.
- ▹ Expected profit is $\frac{1}{n}\cdot\frac{(n-1)k}{n}+\frac{n-1}{n}\cdot\frac{k}{n(n-1)}=\frac{k}{n}$.
- ▹ So, we should pay up to k/n€ for the information. (as much as block 3 is worth)

General formula (VPI)

▷ Given current evidence E, possible actions $a\in A$ with outcomes in S_a, and current best action α

$$\mathrm{EU}(\alpha|E)=\max_{a\in A}\left(\sum_{s\in S_a}U(s)\cdot P(s|E,a)\right)$$

▷ Suppose we knew $F=f$ (new evidence), then we would choose α_f s.t.

$$\mathrm{EU}(\alpha_f|E,F=f)=\max_{a\in A}\left(\sum_{s\in S_a}U(s)\cdot P(s|E,a,F=f)\right)$$

here, F is a random variable with domain D whose value is *currently* unknown.

▷ **Idea:** So we must compute the expected gain over all possible values $f \in D$.

▷ **Definition 23.6.5.** Let F be a random variable with domain D, then the value of perfect information (VPI) on F given evidence E is defined as

$$\mathsf{VPI}_E(F) := \left(\sum_{f \in D} P(F=f|E) \cdot \mathsf{EU}(\alpha_f|E, F=f)\right) - \mathsf{EU}(\alpha|E)$$

where $\alpha_f = \underset{a \in A}{\mathsf{argmax}}\, \mathsf{EU}(a|E, F=f)$ and A the set of possible actions.

Properties of VPI

▷ **Observation 23.6.6 (VPI is Non-negative).**
$VPI_E(F) \geq 0$ for all j and E *(in expectation, not post hoc)*

▷ **Observation 23.6.7 (VPI is Non-additive).**
$VPI_E(F, G) \neq VPI_E(F) + VPI_E(G)$ *(consider, e.g., obtaining F twice)*

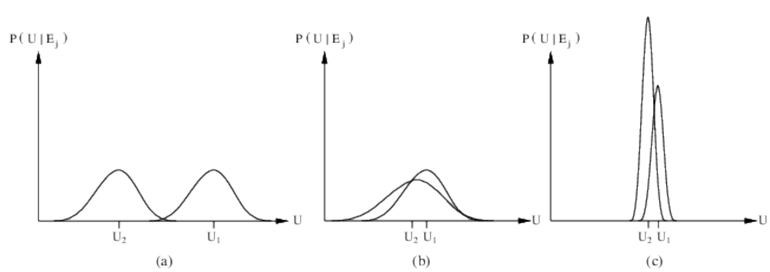
▷ **Observation 23.6.8 (VPI is Order-independent).**

$$VPI_E(F, G) = VPI_E(F) + VPI_{E,F}(G) = VPI_E(G) + VPI_{E,G}(F)$$

▷ **Note:** When more than one piece of evidence can be gathered, maximizing VPI for each to select one is not always optimal
⇝ evidence-gathering becomes a sequential decision problem.

Qualitative behavior of VPI

▷ **Question:** Say we have three distributions for $P(U|E_j)$



Qualitatively: What is the value of information (VPI) in these three cases?

▷ **Answers:** reserved for the plenary sessions ⇝ be there!

We will now use information value theory to specialize our utility-based agent from above.

A simple Information-Gathering Agent

▷ **Definition 23.6.9.** A simple information gathering agent. (gathers info before acting)

function Information—Gathering—Agent (percept) **returns** an action
  persistent: D, a decision network
  integrate percept into D
  $j :=$ $\underset{k}{\mathrm{argmax}}\ \mathsf{VPI}_E(E_k)/Cost(E_k)$
  if $\mathsf{VPI}_E(E_j) > Cost(E_j)$ **return** Request(E_j)
  else return the best action from D

The next percept after Request(E_j) provides a value for E_j.

▷ **Problem:** The information gathering implemented here is myopic, i.e. calculating VPI as if only a single evidence variable will be acquired. (cf. greedy search)

▷ But it works relatively well in practice. (e.g. outperforms humans for selecting diagnostic tests)

Chapter 24

Temporal Probability Models

Outline of this chapter

- ▷ Modeling time and uncertainty for sequential environments.
- ▷ Markov inference: Filtering, prediction, smoothing, and most likely explanation.
- ▷ Hidden Markov models
- ▷ Dynamic Bayesian networks
- ▷ Particle filtering?
- ▷ Further algorithms and Topics?

24.1 Modeling Time and Uncertainty

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/32520`.

Time and uncertainty

- ▷ **Observation 24.1.1.** *The world changes; we need to track and predict it!*
- ▷ **Example 24.1.2.** Consider the following decision problems:
  - ▷ Vehicle diagnosis: car state constant during diagnosis $\leadsto$ episodic!
  - ▷ Diabetes management: patient state can quickly deteriorate $\leadsto$ sequential!
- ▷ Here we lay the mathematical foundations for the latter.
- ▷ **Definition 24.1.3.** A temporal probability model is a probability model, where possible worlds are indexed by a time structure $\langle S, \preceq\rangle$.
- ▷ We restrict ourselves to linear, discrete time structures, i.e. $\langle S, \preceq\rangle = \langle\mathbb{N}, \leq\rangle$.(Step size irrelevant for theory, depends on problem in practice)
- ▷ **Definition 24.1.4 (Basic Setup).** A temporal probability model has two sets of random variables indexed by $\mathbb{N}$.

- $\mathbf{X}_t \mathrel{\widehat{=}}$ set of (unobservable) state variables at time $t{\geq}0$
  e.g., $\mathsf{BloodSugar}_t$, $\mathsf{StomachContents}_t$, etc.
- $\mathbf{E}_t \mathrel{\widehat{=}}$ set of (observable) evidence variables at time $t{>}0$
  e.g., $\mathsf{MeasuredBloodSugar}_t$, $\mathsf{PulseRate}_t$, $\mathsf{FoodEaten}_t$

▷ **Notation:** $\mathbf{X}_{a:b} = \mathbf{X}_a, \mathbf{X}_{a+1}, \ldots, \mathbf{X}_{b-1}, \mathbf{X}_b$

Time and uncertainty (Running Example)

▷ **Example 24.1.5 (Umbrellas).** You are a security guard in a secret underground facility, want to know it if is raining outside. Your only source of information is whether the director comes in with an umbrella.

- State variables: $\mathsf{R}_0, \mathsf{R}_1, \mathsf{R}_2, \ldots,$
- Observations (evidence variables): $\mathsf{U}_1, \mathsf{U}_2, \mathsf{U}_3, \ldots$

Markov Processes

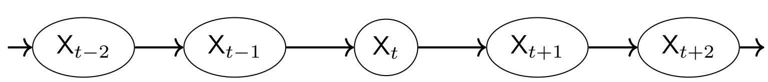
▷ **Idea:** Construct a Bayesian network from these variables. (parents?)

▷ **Definition 24.1.6.** Markov property: $\mathbf{X}_t$ only depends on a bounded subset of $\mathbf{X}_{0:t-1}$. (in particular not on $\mathbf{E}_{1:t}$)

▷ **Definition 24.1.7.** A (discrete-time) Markov process is a sequence of random variables with the Markov property.

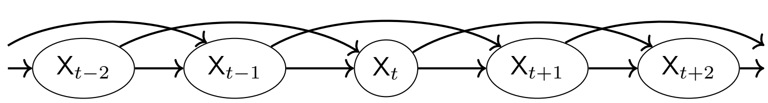
▷ **Definition 24.1.8.** We say that a Markov process has the nth order Markov property for $n{\in}\mathbb{N}^+$, iff $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-n:t-1})$. Special Cases

- First-order Markov property: $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$



  A first order Markov process is called a Markov chain.

- Second-order Markov property: $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-2}, \mathbf{X}_{t-1})$

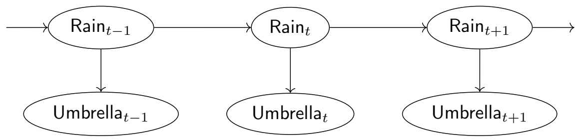


▷ **Intuition:** Increasing the order adds “memory” to the process, Markov chains have none.

▷ **Preview:** We will use Markov processess to model sequential environments.

Markov Process Example: The Umbrella

▷ **Example 24.1.9 (Umbrellas continued).** We model the situation in a Bayesian network:



▷ **Problem:** First-order Markov property not exactly true in real world!

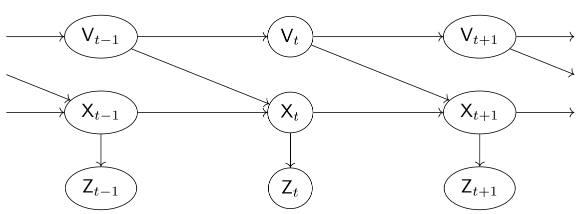
▷ **Possible fixes:**

1. Increase the order of the Markov process. (more dependencies)
2. Add state variables, e.g., add Temp_t, Pressure_t. (more information sources)

We will see the second in another example: tracking robot motion.

Markov Process Example: Robot Motion

▷ **Example 24.1.10 (Random Robot Motion).** To track a robot wandering randomly on the X/Y plane, use the following Markov chain



We use Newton's laws to calculate the new position

- ▷ the velocity V_i may change unpredictably.
- ▷ the position X_i depends on previous position X_{i-1} and velocity V_{i-1}
- ▷ the position X_i influences the observed position Z_i.

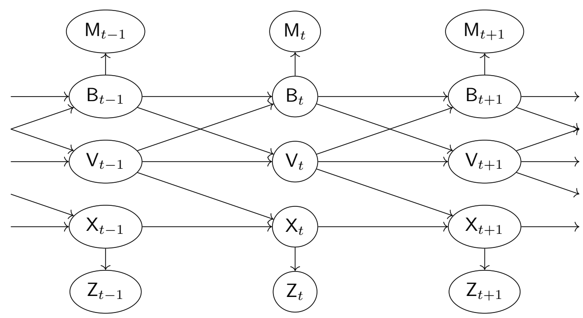
▷ **Example 24.1.11 (Battery Powered Robot).** Markov property violated!

- ▷ Battery exhaustion has a systematic effect on the change in velocity.
- ▷ This depends on how much power was used by all previous manoeuvres.

▷ **Idea:** We can restore the Markov property by including a state variable for the charge level B_t. (Better still: Battery level sensor)

Markov Process Example: Robot Motion

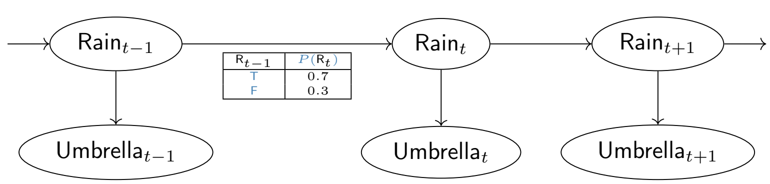
- **Example 24.1.12 (Battery Powered Robot Motion).**



  - Battery level B_i is influenced by previous level $B_{i-1}$and velocity V_{i-1}.
  - Velocity V_i is influenced by previous level $B_{i-1}$and velocity V_{i-1} as well.
  - Battery meter M_i is only influenced by Battery level B_i.

Stationary Markov Processes as Transition Models

- **Theorem 24.1.13.** *Let M be a Markov chain with state variables $\mathbf{X}_t$ evidence variables $\mathbf{E}_t$; then $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$ is the transition model and $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_{0:t},\mathbf{E}_{1:t-1})$ the sensor model of M.*
- **Problem:** Even with Markov property the transition model is infinite. ($t\in\mathbb{N}$)
- **Definition 24.1.14.** A Markov chain is called stationary if the transition model is independent of time, i.e. $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$ is the same for all t.
- **Example 24.1.15 (Umbrellas are stationary).** $\mathbf{P}(R_t|R_{t-1})$ does not depend on t. (need only one table)



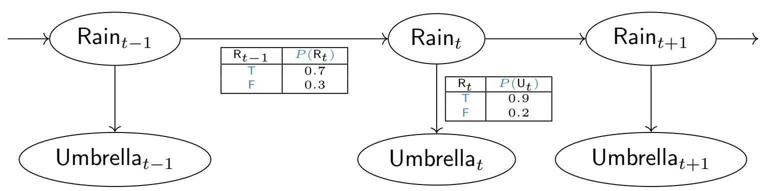
- ⚠ Don't confuse "stationary" (Markov processes) with "static" (environments).
- We restrict ourselves to stationary Markov processes in AI-2.

Markov Sensor Models

- **Recap:** The sensor model predicts the influence of percepts (and the world state) on the belief state. (used during update)
- **Problem:** The evidence variables $\mathbf{E}_t$ could depend on previous variables as well as the current state.
- We restrict dependency to current state. (otherwise state repn. deficient)
- **Definition 24.1.16.** We say that a sensor model has the sensor Markov property, iff $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_{0:t}, \mathbf{E}_{1:t-1}) = \mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$
- **Assumptions on Sensor Models:** We usually assume the sensor Markov property and make it stationary as well: $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$ is fixed for all t.

Umbrellas, the full Story

- **Example 24.1.17 (Umbrellas, Transition & Sensor Models).**



Note that influence goes from Rain_t to Umbrella_t. (causal dependency)

- **Observation 24.1.18.** *If we additionally know the initial prior probabilities* $\mathbf{P}(\mathbf{X}_0)$ *(*$\widehat{=}$ *time* $t = 0$*), then we can compute the full joint probability distribution as*

$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \mathbf{P}(\mathbf{X}_0) \cdot \prod_{i=1}^{t} \mathbf{P}(\mathbf{X}_i|\mathbf{X}_{i-1}) \cdot \mathbf{P}(\mathbf{E}_i|\mathbf{X}_i)$$

24.2 Inference: Filtering, Prediction, and Smoothing

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30350`, `https://fau.tv/clip/id/30351`, and `https://fau.tv/clip/id/3052`.

Inference tasks

- **Definition 24.2.1.** The Markov inference tasks consist of filtering, prediction, smoothing, and most likely explanation as sdefined below.
- **Definition 24.2.2.** Filtering (or monitoring): $\mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$ computing the belief state input to the decision process of a rational agent.
- **Definition 24.2.3.** Prediction (or state estimation): $\mathbf{P}(\mathbf{X}_{t+k}|\mathbf{e}_{1:t})$ for $k > 0$

evaluation of possible action sequences. ($\widehat{=}$ filtering without the evidence)

▷ **Definition 24.2.4.** Smoothing (or hindsight): $\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t})$ for $0 \leq k < t$ better estimate of past states. (essential for learning)

▷ **Definition 24.2.5.** Most likely explanation: $\underset{\mathbf{x}_{1:t}}{\operatorname{argmax}}\,(P(\mathbf{x}_{1:t}|\mathbf{e}_{1:t}))$ speech recognition, decoding with a noisy channel.

Filtering (Computing the Belief State given Evidence)

▷ **Aim:** Recursive state estimation: $\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) = f(\mathbf{e}_{t+1:},\mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t}))$

▷ Project the current distribution forward from t to $t+1$:

$$\begin{aligned} \mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) &= \mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t},\mathbf{e}_{t+1}) && \text{(dividing up evidence)}\\ &= \alpha\cdot\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1},\mathbf{e}_{1:t})\cdot\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}) && \text{(using Bayes' rule)}\\ &= \alpha\cdot\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\cdot\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}) && \text{(sensor Markov property)} \end{aligned}$$

▷ **Note:** $\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})$ can be obtained directly from the sensor model.

▷ Continue by conditioning on the current state $\mathbf{X}_t$:

$$\begin{aligned} &\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1})\\ &\quad= \alpha\cdot\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\cdot(\sum_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t,\mathbf{e}_{1:t})\cdot P(\mathbf{x}_t|\mathbf{e}_{1:t}))\\ &\quad= \alpha\cdot\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\cdot(\sum_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t)\cdot P(\mathbf{x}_t|\mathbf{e}_{1:t})) \end{aligned}$$

▷ $\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{X}_t)$ is simply the transition model, $P(\mathbf{x}_t|\mathbf{e}_{1:t})$ the "recursive call".

▷ So $\mathbf{f}_{1:t+1} = \alpha\cdot\mathrm{FORWARD}(\mathbf{f}_{1:t},\mathbf{e}_{t+1})$ where $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$ and FORWARD is the update shown above. (Time and space *constant* (independent of t))

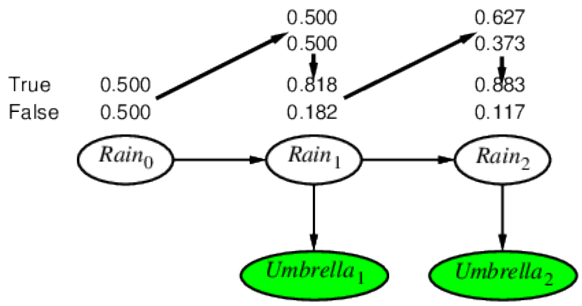
Filtering the Umbrellas

▷ **Example 24.2.6.** Say the guard believes $\mathbf{P}(\mathsf{R}_0) = \langle 0.5, 0.5\rangle$. On day 1 and 2 the umbrella appears.

$$\mathbf{P}(\mathsf{R}_1) = \sum_{r_0}\mathbf{P}(\mathsf{R}_1|r_0)\cdot P(r_0) = \langle 0.7, 0.3\rangle\cdot 0.5 + \langle 0.3, 0.7\rangle\cdot 0.5 = \langle 0.5, 0.5\rangle$$

Update with evidence for $t = 1$ gives:

$$\mathbf{P}(\mathsf{R}_1|\mathsf{u}_1) = \alpha\cdot\mathbf{P}(\mathsf{u}_1|\mathsf{R}_1)\cdot\mathbf{P}(\mathsf{R}_1) = \alpha\cdot\langle 0.9, 0.2\rangle\langle 0.5, 0.5\rangle = \alpha\cdot\langle 0.45, 0.1\rangle \approx \langle 0.818, 0.182\rangle$$

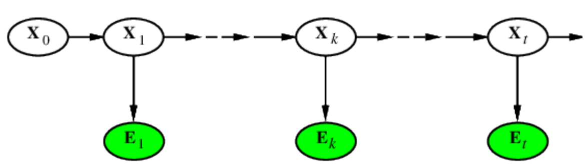


Prediction in Markov Chains

- Prediction computes future $k > 0$ state distributions: $\mathbf{P}(\mathbf{X}_{t+k}|\mathbf{e}_{1:t})$.
- **Intuition:** Prediction is filtering without new evidence.
- **Lemma 24.2.7.** $\mathbf{P}(\mathbf{X}_{t+k+1}|\mathbf{e}_{1:t}) = \sum_{\mathbf{x}_{t+k}} \mathbf{P}(\mathbf{X}_{t+k+1}|\mathbf{x}_{t+k}) \cdot P(\mathbf{x}_{t+k}|\mathbf{e}_{1:t})$
- *Proof sketch:* Using the same reasoning as for the FORWARD algorithm for filtering.
- **Observation 24.2.8.** *As $k \to \infty$, $P(\mathbf{x}_{t+k}|\mathbf{e}_{1:t})$ tends to the stationary distribution of the Markov chain, i.e. the a fixed point under prediction.*
- **Intuition:** The mixing time, i.e. the time until prediction reaches the stationary distribution depends on how "stochastic" the chain is.

Smoothing

- Smoothing estimates past states by computing $\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t})$ for $0 \leq k < t$



- Divide evidence $\mathbf{e}_{1:t}$ into $\mathbf{e}_{1:k}$ (before k) and $\mathbf{e}_{k+1:t}$ (after k):

$$\begin{aligned}
\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t}) &= \mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k}, \mathbf{e}_{k+1:t}) \\
&= \alpha \cdot \mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k}) \cdot \mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k, \mathbf{e}_{1:k}) & \text{(Bayes Rule)} \\
&= \alpha \cdot \mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k}) \cdot \mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) & \text{(cond. independence)} \\
&= \alpha \cdot \mathbf{f}_{1:k} \cdot \mathbf{b}_{k+1:t}
\end{aligned}$$

Smoothing (continued)

▷ Backward message $\mathbf{b}_{k+1:t} = \mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k)$ computed by a backwards recursion:

$$\begin{aligned}
\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) &= \sum_{\mathbf{x}_{k+1}} \mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k, \mathbf{x}_{k+1}) \cdot \mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\
&= \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1:t}|\mathbf{x}_{k+1}) \cdot \mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\
&= \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1}, \mathbf{e}_{k+2:t}|\mathbf{x}_{k+1}) \cdot \mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\
&= \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1}|\mathbf{x}_{k+1}) \cdot P(\mathbf{e}_{k+2:t}|\mathbf{x}_{k+1}) \cdot \mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k)
\end{aligned}$$

$P(\mathbf{e}_{k+1}|\mathbf{x}_{k+1})$ and $\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k)$ can be directly obtained from the model, $P(\mathbf{e}_{k+2:t}|\mathbf{x}_{k+1})$ is the "recursive call" ($\mathbf{b}_{k+2:t}$).

▷ In message notation: $\mathbf{b}_{k+1:t} = \mathsf{BACKWARD}(\mathbf{b}_{k+2:t}, \mathbf{e}_{k+1:t})$ where BACKWARD is the update shown above. (time and space constant (independent of t))

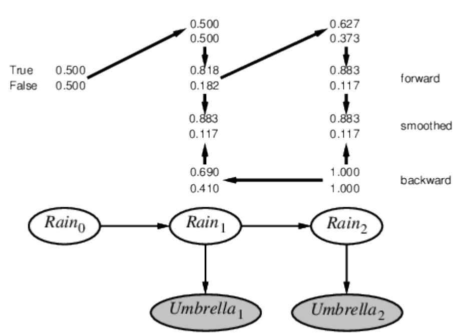
Smoothing example

▷ **Example 24.2.9 (Smoothing Umbrellas).** Umbrella appears on days 1/2.

- ▷ $\mathbf{P}(\mathsf{R}_1|\mathsf{u}_1, \mathsf{u}_2) = \alpha \cdot \mathbf{P}(\mathsf{R}_1|\mathsf{u}_1) \cdot \mathbf{P}(\mathsf{u}_2|\mathsf{R}_1) = \alpha \cdot \langle 0.818, 0.182\rangle \cdot \mathbf{P}(\mathsf{u}_2|\mathsf{R}_1)$
- ▷ Compute $\mathbf{P}(\mathsf{u}_2|\mathsf{R}_1)$ by backwards recursion:

$$\begin{aligned}
\mathbf{P}(\mathsf{u}_2|\mathsf{R}_1) &= \sum_{\mathsf{r}_2} P(\mathsf{u}_2|\mathsf{r}_2) \cdot P(|\mathsf{r}_2) \cdot \mathbf{P}(\mathsf{r}_2|\mathsf{R}_1) \\
&= 0.9 \cdot 1 \cdot \langle 0.7, 0.3\rangle + 0.2 \cdot 1 \cdot \langle 0.3, 0.7\rangle = \langle 0.69, 0.41\rangle
\end{aligned}$$

- ▷ So $\mathbf{P}(\mathsf{R}_1|\mathsf{u}_1, \mathsf{u}_2) = \alpha \cdot \langle 0.818, 0.182\rangle \cdot \langle 0.69, 0.41\rangle \approx 0.883, 0.117$



Smoothing gives a higher probability for rain on day 1

- ▷ umbrella on day 2
- ▷ $\leadsto$ rain more likely on day 2
- ▷ $\leadsto$ rain more likely on day 1.

Forward/Backward Algorithm for Smoothing

▷ **Definition 24.2.10.** Forward backward algorithm: cache forward messages along the way:

function Forward–Backward (ev,$prior$)
  returns: a vector of probability distributions
  inputs: ev, a vector of evidence evidence values **for** steps 1, . . . ,t
    $prior$, the prior distribution on the initial state, $P(\mathrm{X}_0)$
  local: fv, a vector of forward messages **for** steps 0, . . . , t
    b, a representation of the backward message, initially all 1s
    sv, a vector of smoothed estimates **for** steps 1, . . . ,t
  $fv[0] := prior$
  for $i = 1$ **to** t **do**
    $fv[i]$:= FORWARD($fv[i-1]$, $ev[i]$)
  for i = t downto k **do**
    $sv[i]$:= NORMALIZE($fv[i]b$)
    b := BACKWARD(b, $ev[i]$)
  return sv

▷ Time complexity linear in t (polytree inference), Space complexity $\mathcal{O}(t \cdot \#(\mathbf{f}))$.

Most Likely Explanation

▷ **Observation 24.2.11.** *Most likely sequence $\neq$ sequence of most likely states!*

▷ **Example 24.2.12.** Suppose the umbrella sequence is T, T, F, T, T what is the most likely weather sequence?

▷ **Prominent Application:** In speech recognition, we want to find the most likely word sequence, given what we have heard. (can be quite noisy)

▷ **Idea:** Use smoothing to find posterior distribution in each time step, construct sequence of most likely states.

▷ **Problem:** These posterior distributions range over a single time step. (and this difference matters)

Most Likely Explanation (continued)

▷ Most likely path to each $\mathbf{x}_{t+1}$ = most likely path to *some* $\mathbf{x}_t$ plus one more step

$$\max_{\mathbf{x}_1,\ldots,\mathbf{x}_t}(\mathbf{P}(\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}))$$
$$= \mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}) \cdot \max_{\mathbf{x}_t}(\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t) \cdot \max_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-1}}(P(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t|\mathbf{e}_{1:t})))$$

▷ Identical to filtering, except $\mathbf{f}_{1:t}$ replaced by

$$\mathbf{m}_{1:t} = \max_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-1}} (\mathbf{P}(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{X}_t|\mathbf{e}_{1:t}))$$

I.e., $\mathbf{m}_{1:t}(i)$ gives the probability of the most likely path to state i.
Update has sum replaced by max, giving the Viterbi algorithm:

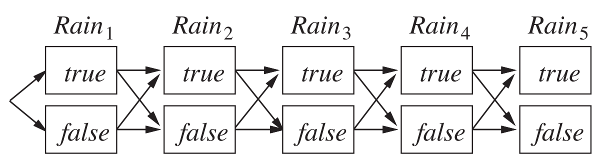
$$\mathbf{m}_{1:t+1} = \mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}) \cdot \max_{\mathbf{x}_t} (\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t, \mathbf{m}_{1:t}))$$

▷ **Observation 24.2.13.** *Viterbi has linear time complexity (like filtering), but linear space complexity (needs to keep a pointer to most likely sequence leading to each state).*

Viterbi example

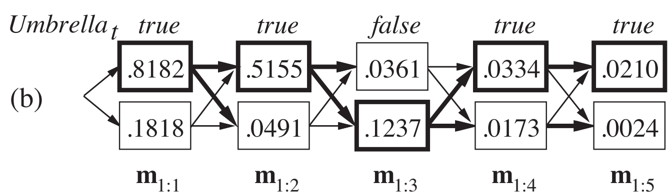
▷ **Example 24.2.14 (Viterbi for Umbrellas).** View the possible state sequences for Rain_t as paths through state graph.

(a)



Operation of the Viterbi algorithm for the sequence [T, T, F, T, T]:

(b)



▷ values are $\mathbf{m}_{1:t}$ (probability of best sequence reaching state at time t)

▷ bold arrows: best predecessor measured by "best preceding sequence probability × transition probability"

To find "most likely sequence", follow bold arrows back from "most likely state $\mathbf{m}_{1:5}$.

24.3 Hidden Markov Models

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30353` and `https://fau.tv/clip/id/30354`.

The preceding section developed algorithms for temporal probabilistic reasoning using a general framework that was independent of the specific form of the transition and sensor models. In this section, we discuss more concrete models and applications that illustrate the power of the basic algorithms and implementation issues.

In particular, we will introduce hidden Markov models, special simple Markov chains where Markov inference can be expressed in terms of matrix calculations.

Hidden Markov Models

- **Definition 24.3.1.** A hidden Markov model (HMM) is a Markov chain with a single, discrete state variable X_t with domain $\{1,\ldots,S\}$ and a single, discrete evidence variable.
- **Example 24.3.2.** The umbrella example from Example 24.1.5 is an HMM.
- **Observation:** Transition model $\mathbf{P}(X_t|X_{t1}) \mathrel{\hat{=}}$ a single $S \times s$ matrix.
- **Definition 24.3.3.** Transition matrix: $\mathbf{T}_{ij} := P(X_t = j | X_{t1} = i)$
- **Example 24.3.4 (Umbrellas).** $\mathbf{T} = P(X_t|X_{t-1}) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$.
- **Observation:** Sensor model $\mathbf{P}(e_t|X_t = i) \mathrel{\hat{=}}$ an S-vector.
- **Idea:** Re-cast Markov inference as matrix calculations.

  This works best, if we make the sensor model into a diagonal matrix. (see below)
- **Definition 24.3.5.** Sensor matrix $\mathbf{O}_t$ for each time step $\mathrel{\hat{=}}$ diagonal matrix with $\mathbf{O}_{tii} = P(e_t|X_t = i)$.
- **Example 24.3.6 (Umbrellas).** With $U_1 = T$ and $U_3 = F$ we have

$$\mathbf{O}_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.2 \end{pmatrix} \quad \text{and} \quad \mathbf{O}_3 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.8 \end{pmatrix}$$

HMM Algorithm

- **Idea:** The forward and backward messages are column vectors in HMMs.
- **Definition 24.3.7.** Recasting the Markov inference as matrix computation, gives us two identities:

HMM filtering equation:	$\mathbf{f}_{1:t+1}$	$=$	$\alpha \cdot \mathbf{O}_{t+1}\mathbf{T}^t\mathbf{f}_{1:t}$
HMM smoothing equation:	$\mathbf{b}_{k+1:t}$	$=$	$\mathbf{T}\mathbf{O}_{k+1}\mathbf{b}_{k+2:t}$

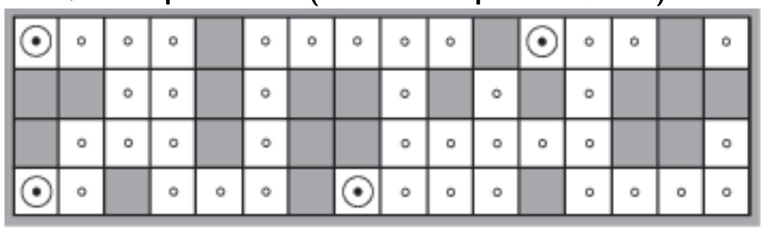
- **Observation 24.3.8.** *The forward backward algorithm for HMMs has time complexity $\mathcal{O}(S^2 t)$ and space complexity $\mathcal{O}(St)$.*

Example: Robot Localization using Common Sense

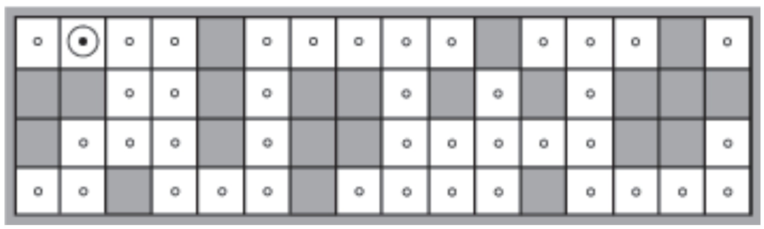
- **Example 24.3.9 (Robot Localization in a Maze).** A robot has four sonar sensors that tell it about obstacles in four directions: N, S, W, E.

▷ **Notation:** We write the result where the sensor that detects obstacles in the north, south, and east as N S E.

▷ **Example 24.3.10 (Filter out Impossible States).**



a) Possible robot locations after $e_1 =$ N S W



b) Possible robot locations after $e_1 =$ N S W and $e_2 =$ N S

▷ *Remark 24.3.11.* This only works for perfect sensors. (else no impossible states)

HMM Example: Robot Localization (Modeling)

▷ **Example 24.3.12 (HMM-based Robot Localization).**

- ▷ Random variable X_t for robot location (domain: 42 empty squares)
- ▷ Transition matrix for the move action: ($\mathbf{T}$ has $42^2 = 1764$ entries)

$$P(X_{t+1} = j|X_t = i) = \mathbf{T}_{ij} = \begin{cases} \frac{1}{\#(N(i))} & \text{if } j \in N(i) \\ 0 & \text{else} \end{cases}$$

where $N(i)$ is the set of neighboring fields of state i.

- ▷ We do not know where the robot starts: $P(X_0) = \frac{1}{n}$ (here $n = 42$)
- ▷ Evidence variable E_t: four bit presence/absence of obstacles in N, S, W, E. Let d_{it} be the number of wrong bits and ϵ the error rate of the sensor.

$$P(E_t = e_t|X_t = i) = \mathbf{O}_{tii} = (1-\epsilon)^{4-d_{it}} \cdot \epsilon^{d_{it}}$$

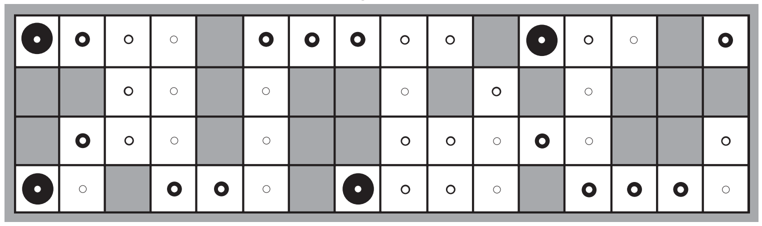
- ▷ For instance, the probability that the sensor on a square with obstacles in north and south would produce N S E is $(1-\epsilon)^3 \cdot \epsilon^1$.

▷ **Idea:** Use the HMM filtering equation $\mathbf{f}_{1:t+1} = \alpha \cdot \mathbf{O}_{t+1}\mathbf{T}^t\mathbf{f}_{1:t}$ for localization. (next)

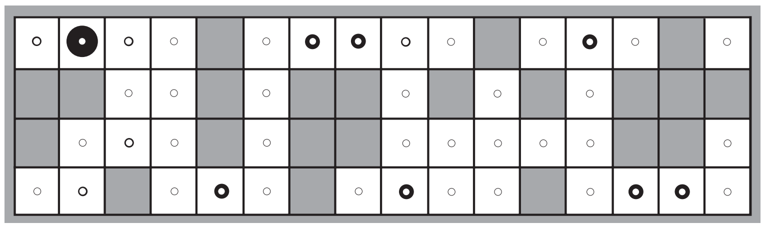
HMM Example: Robot Localization

▷ **Idea:** Use HMM filtering equation $\mathbf{f}_{1:t+1} = \alpha \cdot \mathbf{O}_{t+1}\mathbf{T}^t\mathbf{f}_{1:t}$ to compute posterior distribution over locations. (i.e. robot localization)

▷ **Example 24.3.13.** Redoing Example 24.3.9, with $\epsilon = 0.2$.



a) Posterior distribution over robot location after E_1 = N S W



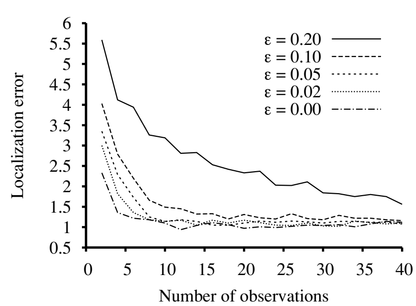
b) Posterior distribution over robot location after E_1 = N S W and E_2 = N S

Still the same locations as in the "perfect sensing" case, but now other locations have non-zero probability.

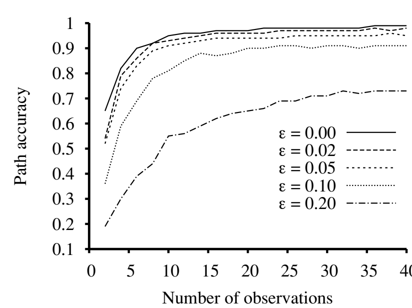
HMM Example: Further Inference Applications

▷ **Idea:** Use smoothing: $\mathbf{b}_{k+1:t} = \mathbf{TO}_{k+1}\mathbf{b}_{k+2:t}$ to find out where it started and the Viterbi algorithm to find the most likely path it took.

▷ **Example 24.3.14.** Performance of HMM localization vs. observation length (various error rates ϵ)



Localization error (Manhattan distance from true location)



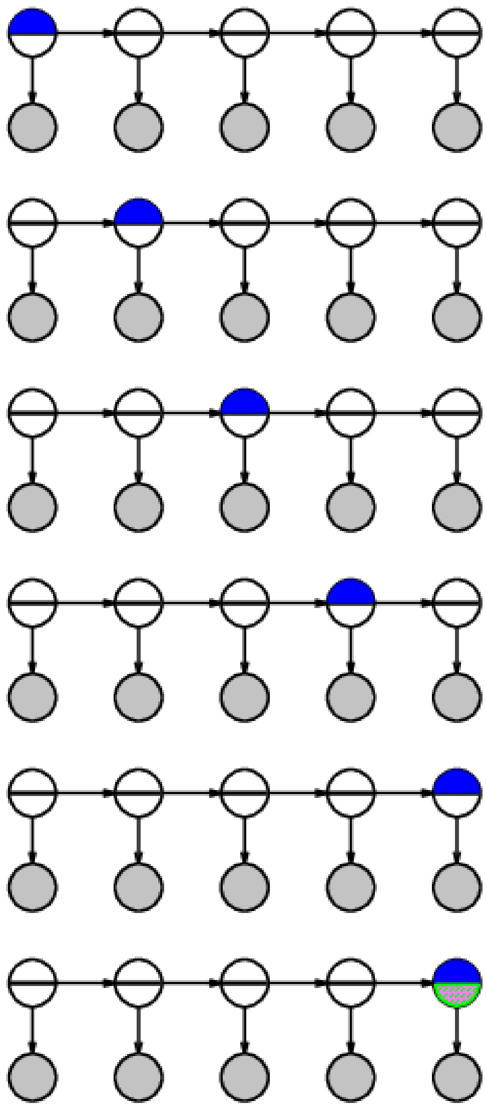
Viterbi path accuracy (fraction of correct states on Viterbi path)

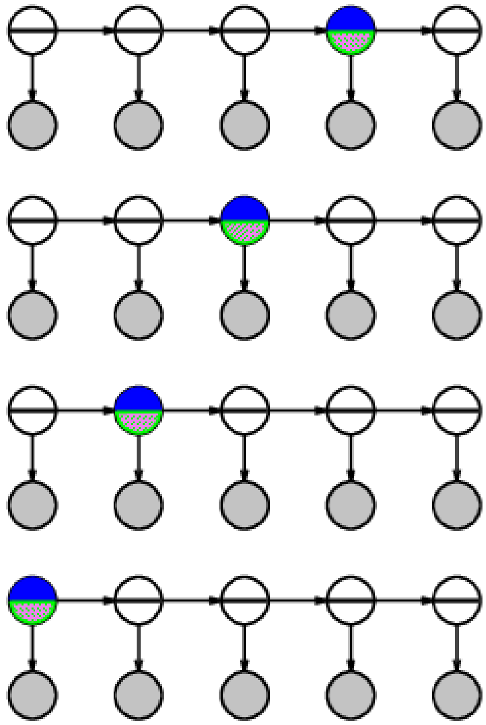
Country dance algorithm

▷ **Idea:** We can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned} \mathbf{f}_{1:t+1} &= \alpha \cdot \mathbf{O}_{t+1}\mathbf{T}^t\mathbf{f}_{1:t} \\ {\mathbf{O}_{t+1}}^{-1}\mathbf{f}_{1:t+1} &= \alpha \cdot \mathbf{T}^t\mathbf{f}_{1:t} \\ \alpha \cdot {\mathbf{T}'^{t}}^{-1}{\mathbf{O}_{t+1}}^{-1}\mathbf{f}_{1:t+1} &= \mathbf{f}_{1:t} \end{aligned}$$

▷ **Algorithm:** Forward pass computes $\mathbf{f}_{1:t}$, backward pass does $\mathbf{f}_{1:i}$, $\mathbf{b}_{t-i:t}$.





- **Observation:** Backwards pass only stores one copy of $\mathbf{f}_{1:i}$, $\mathbf{b}_{t:t-i}$ $\leadsto$ constant space.
- **Problem:** Algorithm is severely limited: transition matrix must be invertible and sensor matrix cannot have zeroes – that is, that every observation be possible in every state.

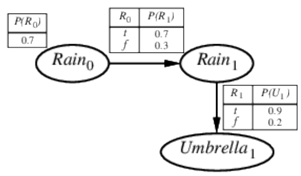
24.4 Dynamic Bayesian Networks

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30355`.

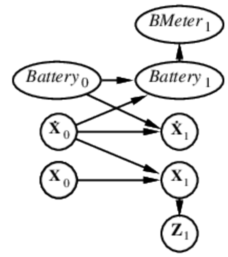
Dynamic Bayesian networks

- **Definition 24.4.1.** A Bayesian network $\mathcal{D}$ is called dynamic (a DBN), iff its random variables are indexed by a time structure. We assume that $\mathcal{D}$ is
  - time sliced, i.e. that the time slices $\mathcal{D}_t$ – the subgraphs of t-indexed random variables and the edges between them – are isomorphic.
  - a stationary Markov chain, i.e. that variables X_t can only have parents in $\mathcal{D}_t$ and $\mathcal{D}_{t-1}$.
- $\mathbf{X}_t$, $\mathbf{E}_t$ contain arbitrarily many variables in a replicated Bayesian network.
- **Example 24.4.2.**

Umbrellas



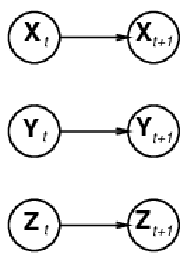
Robot Motion

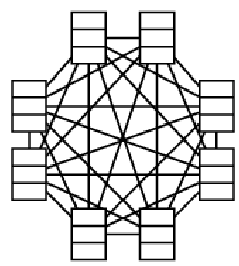


DBNs vs. HMMs

▷ **Observation 24.4.3.**

- ▷ *Every HMM is a single-variable DBN.* (*trivially*)
- ▷ *Every discrete DBN is an HMM.* (*combine variables into tuple*)
- ▷ *DBNs have sparse dependencies* $\rightsquigarrow$ *exponentially fewer parameters;*

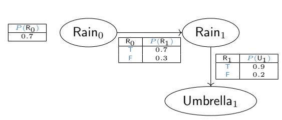


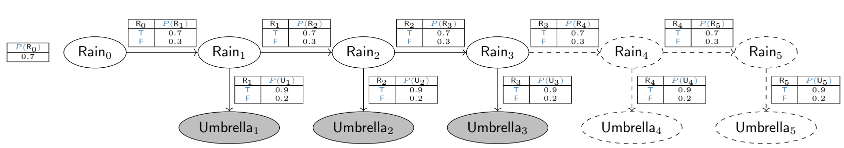


▷ **Example 24.4.4 (Sparse Dependencies).** With 20 Boolean state variables, three parents each, a DBN has $20 \cdot 2^3 = 160$ parameters, the corresponding HMM has $2^{20} \cdot 2^{20} \approx 10^{12}$.

Exact inference in DBNs

▷ **Definition 24.4.5 (Naive method).** Unroll the network and run any exact algorithm.





▷ **Problem:** Inference cost for each update grows with t.

▷ **Definition 24.4.6.** Rollup filtering: add slice $t+1$, "sum out" slice t using variable elimination.

▷ **Observation:** Largest factor is $\mathcal{O}(d^{n+1})$, update cost $\mathcal{O}(d^{n+2})$, where d is the maximal domain size.

▷ **Note:** Much better than the HMM update cost of $\mathcal{O}(d^{2n})$

Summary

▷ Temporal probability models use state and evidence variables replicated over time.

▷ Markov property and stationarity assumption, so we need both

  ▷ a transition model and $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$

  ▷ a sensor model $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$.

▷ Tasks are filtering, prediction, smoothing, most likely sequence; (all done recursively with constant cost per time step)

▷ Hidden Markov models have a single discrete state variable; (used for speech recognition)

▷ DBNs subsume HMMs, exact update intractable.

▷ Particle filtering is a good approximate filtering algorithm for DBNs.

Chapter 25

Making Complex Decisions

A Video Nugget covering the introduction to this chapter can be found at https://fau.tv/clip/id/30356.

We will now pick up the thread from chapter 23 but using temporal models instead of simply probabilistic ones. We will first look at a sequential decision theory in the special case, where the environment is stochastic, but fully observable (Markov decision processes) and then lift that to obtain POMDPs and present an agent design based on that.

Outline

▷ Markov decision processes (MDPs) for sequential environments.

▷ Value/policy iteration for computing utilities in MDPs.

▷ Partially observable MDP (POMDPs).

▷ Decision theoretic agents for POMDPs.

25.1 Sequential Decision Problems

A Video Nugget covering this section can be found at https://fau.tv/clip/id/30357.

Sequential Decision Problems

▷ **Definition 25.1.1.** In sequential decision problems, the agent's utility depends on a sequence of decisions (or their result states).

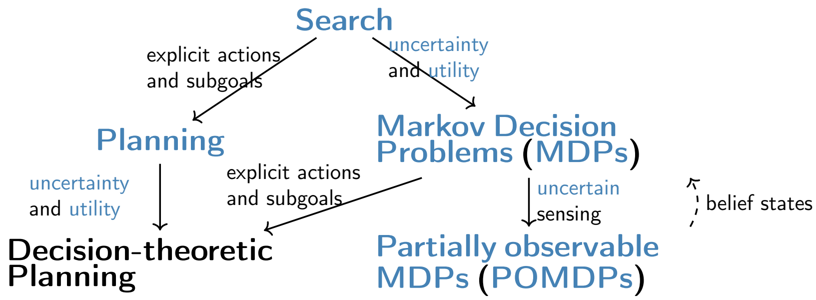
▷ **Definition 25.1.2.** Utility functions on action sequences are often expressed in terms of immediate rewards that are incurred upon reaching a (single) state.

▷ **Methods:** depend on the environment:

- ▷ If it is fully observable $\rightsquigarrow$ Markov decision process (MDPs)
- ▷ else $\rightsquigarrow$ partially observable MDP (POMDP).

▷ Sequential decision problems incorporate utilities, uncertainty, and sensing.

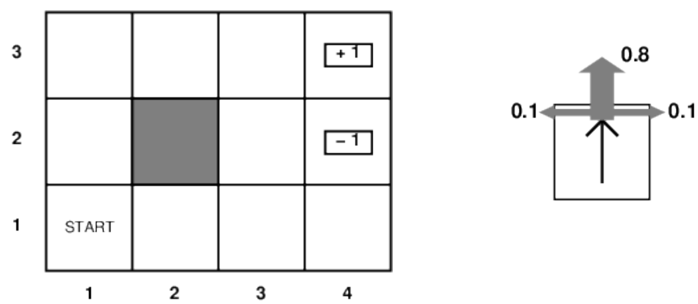
▷ **Preview:** Search problems and planning tasks are special cases.



We will fortify our intuition by an example. It is specifically chosen to be very simple, but to exhibit all the peculiarities of Markov decision problems, which we will generalize from this example.

Markov Decision Problem: Running Example

▷ **Example 25.1.3 (Running Example: The 4x3 World).** A (fully observable) 4×3 environment with non-deterministic actions:



- ▷ States $s \in S$, actions $a \in A$.
- ▷ Transition model: $P(s'|s,a)$ $\widehat{=}$ probability that a in s leads to s'.
- ▷ Reward function:

$$R(s,a,s') := \begin{cases} -0.04 & \text{if (small penalty) for nonterminal states} \\ \pm 1 & \text{if for terminal states} \end{cases}$$

Perhaps what is more interesting than the components of an MDP is that is *not* a component: a belief and/or sensor model. Recall that MDPs are for fully observable environments.

Markov Decision Process

▷ **Motivation:** We are interested in sequential decision problems in a fully observable, stochastic environment with Markovian transition models and additive reward functions.

▷ **Definition 25.1.4.** A Markov decision process (MDP) consists of

- ▷ a set of S of states (with initial state state $s_0 \in S$),
- ▷ sets Actions(s) of actions for each state s.
- ▷ a transition model $P(s'|s,a)$, and
- ▷ a reward function $R\colon S \to \mathbb{R}$ we call $R(s)$ a reward.

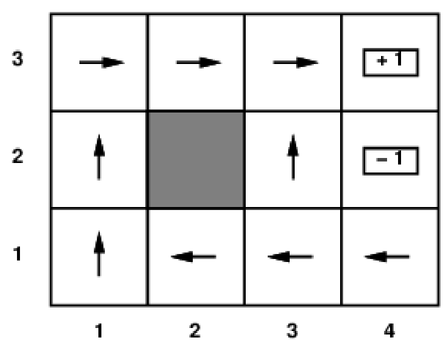
Solving MDPs

▷ **Recall:** In search problems, the aim is to find an optimal sequence of actions.

▷ In MDPs, the aim is to find an optimal policy $\pi(s)$ i.e., best action for every possible state s. (because can't predict where one will end up)

▷ **Definition 25.1.5.** In an MDP, a policy is a mapping from states to actions. An optimal policy maximizes (say) the expected sum of rewards. (MEU)

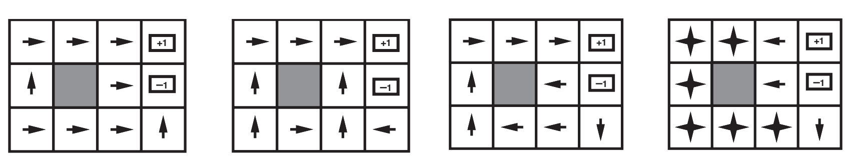
▷ **Example 25.1.6.** Optimal policy when state penalty $R(s)$ is 0.04:



Note: When you run against a wall, you stay in your square.

Risk and Reward

▷ **Example 25.1.7.** Optimal policy depends on the reward $R(s)$.



$R(s) < -1.6284$ $-0.4278 < R(s) < -0.0850$ $-0.0221 < R(s) < 0$ $R(s) > 0$

▷ **Question:** Explain what you see in a qualitative manner!

▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

25.2 Utilities over Time

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30358`.

In this section we address the problem that even if the transition models are stationary, the utilities may not be. In fact we generally have to take the utilities of state sequences into account in sequential decision problems. If we can derive a notion of the utility of a (single) state from that, we may be able to reuse the machinery we introduced above, so that is exactly what we will attempt.

Utility of state sequences

- **Recall:** We cannot observe/assess utility functions, only preferences $\leftrightsquigarrow$ induce utility functions from rational preferences
- **Problem:** In MDPs we need to understand preferences between sequences of states.
- **Definition 25.2.1.** We call preferences on reward sequences stationary, iff

$$[r, r_0, r_1, r_2, \ldots] \succ [r, r'_0, r'_1, r'_2, \ldots] \Leftrightarrow [r_0, r_1, r_2, \ldots] \succ [r'_0, r'_1, r'_2, \ldots]$$

- **Theorem 25.2.2.** *For stationary preferences, there are only two ways to combine rewards over time.*
  - *additive rewards:* $U([s_0, s_1, s_2, \ldots]) = R(s_0) + R(s_1) + R(s_2) + \cdots$
  - *discounted rewards:* $U([s_0, s_1, s_2, \ldots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots$ *where* γ *is called discount factor.*

Utilities of State Sequences

- **Problem:** Infinite lifetimes $\rightsquigarrow$ additive utilities become infinite.
- **Possible Solutions:**

1. Finite horizon: terminate utility computation at a fixed time T

$$U([s_0, \ldots, s_\infty]) = R(s_0) + \cdots + R(s_T)$$

   $\rightsquigarrow$ nonstationary policy: $\pi(s)$ depends on time left.

2. If there are absorbing states: for any policy π agent eventually "dies" with probability 1 $\rightsquigarrow$ expected utility of every state is finite.
3. Discounting: assuming $\gamma < 1$, $R(s) \leq R_{\max}$,

$$U([s_0, \ldots, s_\infty]) = \sum_{t=0}^{\infty} \gamma^t R(s_t) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = R_{\max}/(1 - \gamma)$$

   Smaller $\gamma \rightsquigarrow$ shorter horizon.

▷ **Idea:** Maximize system gain $\widehat{=}$ average reward per time step.

▷ **Theorem 25.2.3.** *The optimal policy has constant gain after initial transient.*

▷ **Example 25.2.4.** Taxi driver's daily scheme cruising for passengers.

Utility of States

▷ **Intuition:** Utility of a state $\widehat{=}$ *expected (discounted) sum of rewards (until termination) assuming optimal actions.*

▷ **Definition 25.2.5.** Given a policy π, let s_t be the state the agent reaches at time t starting at state s_0. Then the expected utility obtained by executing π starting in s is given by

$$U^\pi(s) := E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)\right]$$

we define $\pi^*_s := \underset{\pi}{\operatorname{argmax}}\, U^\pi(s)$.

▷ **Observation 25.2.6.** *π^*_s is independent of the state s.*

▷ *Proof sketch:* If π^*_a and π^*_b reach point c, then there is no reason to disagree – or with π^*_c

▷ **Definition 25.2.7.** We call $\pi^* := \pi^*_s$ for some s the optimal policy.

▷ ⚠ Observation 25.2.6 does not hold for finite horizon policies.

▷ **Definition 25.2.8.** The utility $U(s)$ of a state s is $U^{\pi^*}(s)$.

Utility of States (continued)

▷ **Remark:** $R(s)$ $\widehat{=}$ "short-term reward", whereas U $\widehat{=}$ "long-term reward".

▷ Given the utilities of the states, choosing the best action is just MEU:

- ▷ maximize the expected utility of the immediate successor states

$$\pi^*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \left(\sum_{s'} P(s'|s,a) \cdot U(s')\right)$$

▷ **Example 25.2.9 (Running Example Continued).**

Expected Utility

3	0.812	0.868	0.912	+1
2	0.762		0.660	−1
1	0.705	0.655	0.611	0.388
	1	2	3	4

Optimal Policy

3	→	→	→	+1
2	↑		↑	−1
1	↑	←	←	←
	1	2	3	4

▷ **Question:** Why do we go left in $(3,1)$ and not up? (follow the utility)

Michael Kohlhase: Artificial Intelligence 2 895 2023-09-20

25.3 Value/Policy Iteration

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30359`.

Dynamic programming: the Bellman equation

▷ **Problem:** We have defined $U(s)$ via the optimal policy: $U(s){:=}U^{\pi^*}(s)$, but how to compute it without knowing π^*?

▷ **Observation:** A simple relationship among utilities of neighboring states:

expected sum of rewards = current reward + γ $\cdot$ exp. reward sum after best action

▷ **Theorem 25.3.1 (Bellman equation (1957)).**

$$U(s) = R(s) + \gamma \cdot \max_{a \in A(s)} \sum_{s'} U(s') \cdot P(s'|s,a)$$

We call this equation the Bellman equation

▷ **Example 25.3.2.** $U(1,1) = -0.04$
$+\ \gamma\ \max\{0.8U(1,2) + 0.1U(2,1) + 0.1U(1,1),$ *up*
$0.9U(1,1) + 0.1U(1,2)$ *left*
$0.9U(1,1) + 0.1U(2,1)$ *down*
$0.8U(2,1) + 0.1U(1,2) + 0.1U(1,1)\}$ *right*

▷ **Problem:** One equation/state $\leadsto$ n nonlinear (max isn't) equations in n unknowns.
$\leadsto$ cannot use linear algebra techniques for solving them.

Michael Kohlhase: Artificial Intelligence 2 896 2023-09-20

Value Iteration Algorithm

▷ **Idea:** We use a simple iteration scheme to find a fixpoint:

1. start with arbitrary utility values,

2. update to make them locally consistent with the Bellman equation,
3. everywhere locally consistent $\rightsquigarrow$ global optimality.

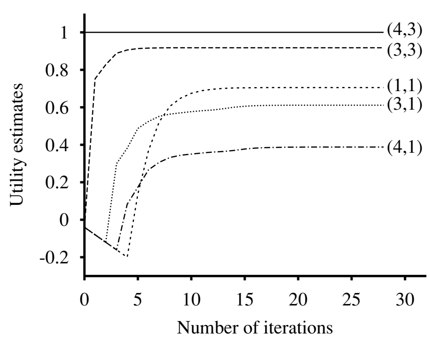
▷ **Definition 25.3.3.** The value iteration algorithm for utility functions is given by

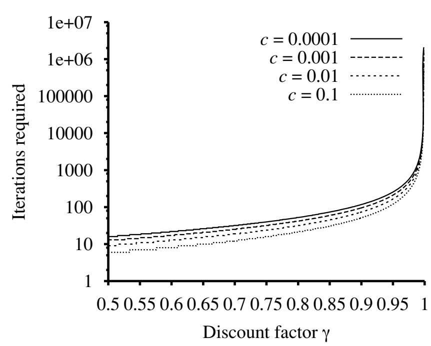
function VALUE—ITERATION (mdp,ϵ) **returns** a utility fn.
  inputs: mdp, an MDP with states S, actions $A(s)$, transition model $P(s'|s,a)$,
        rewards $R(s)$, and discount γ
      ϵ, the maximum error allowed **in** the utility of any state
  local variables: U, U', vectors of utilities **for** states **in** S, initially zero
        δ, the maximum change **in** the utility of any state **in** an iteration
  repeat
    $U := U'$; $\delta := 0$
    for each state s **in** S **do**
      $U'[s] := R(s) + \gamma \cdot \max_a \left(\sum_{s'} U[s'] \cdot P(s'|s,a)\right)$
      if $|U'[s] - U[s]| > \delta$ **then** $\delta := |U'[s] - U[s]|$
    until $\delta < \epsilon(1-\gamma)/\gamma$
    return U

▷ **Remark:** Retrieve the optimal policy with $\pi[s] := \operatorname{argmax}_a \left(\sum_{s'} U[s'] \cdot P(s'|s,a)\right)$

Value Iteration Algorithm (Example)

▷ **Example 25.3.4 (Iteration on 4x3).**





Convergence

▷ **Definition 25.3.5.** The maximum norm $\|U\| = \max_s |U(s)|$, so $\|U - V\|$ = maximum difference between U and V.

▷ Let U^t and U^{t+1} be successive approximations to the true utility U.

▷ **Theorem 25.3.6.** *For any two approximations U^t and V^t*

$$\left\|U^{t+1} - V^{t+1}\right\| \leq \gamma \left\|U^t - V^t\right\|$$

I.e., any distinct approximations must get closer to each other *so, in particular, any approximation must get closer to the true U and value iteration converges to a unique, stable, optimal solution.*

▷ **Theorem 25.3.7.** *If $\left\|U^{t+1}-U^t\right\|<\epsilon$, then $\left\|U^{t+1}-U\right\|<2\epsilon\gamma/1-\gamma$ I.e., once the change in U^t becomes small, we are almost done.*

▷ **Remark:** MEU policy using U^t may be optimal long before convergence of values.

So we see that iteration with Bellman updates will always converge towards the utility of a state, even without knowing the optimal policy. That gives us a first way of dealing with sequential decision problems: we compute utility functions based on states and then use the standard MEU machinery. We have seen above that optimal policies and state utilities are essentially interchangeable: we can compute one from the other. This leads to another approach to computing state utilities: policy iteration, which we will discuss now.

Policy Iteration

▷ **Recap:** Value iteration computes utilities $\leadsto$ optimal policy by MEU.

▷ This even works if the utility estimate is inaccurate. ($\leftsquigarrow$ policy loss small)

▷ **Idea:** Search for optimal policy and state utilityutility values simultaneously [How60]: Iterate

- ▷ policy evaluation: given policy π_i, calculate $U_i=U^{\pi_i}$, the utility of each state were π_i to be executed.
- ▷ policy improvement: calculate a new MEU policy π_{i+1} using 1 lookahead

Terminate if policy improvement yields no change in computed utilities.

▷ **Observation 25.3.8.** *Upon termination U_i is a fixpoint of Bellman update $\leadsto$ Solution to Bellman equation $\leadsto$ π_i is an optimal policy.*

▷ **Observation 25.3.9.** *Policy improvement improves policy and policy space is finite $\leadsto$ termination.*

Policy Iteration Algorithm

▷ **Definition 25.3.10.** The policy iteration algorithm is given by the following pseudocode:

function POLICY−ITERATION(mdp) **returns** a policy
  inputs: mdp, and MDP with states S, actions $A(s)$, transition model $P(s'|s,a)$
  local variables: U a vector of utilities **for** states **in** S, initially zero
    π a policy indexed by state, initially random,
  repeat
    U := POLICY−EVALUATION(π,U,mdp)
    unchanged? := **true**
    foreach state s **in** X **do**
      if $\max_{a\in A(s)}\left(\sum_{s'}P(s'|s,a)\cdot U(s')\right)>\sum_{s'}P(s'|s,\pi[s'])\cdot U(s')$ **then do**
        $\pi[s]$:= $\operatorname{argmax}_{b\in A(s)}\left(\sum_{s'}P(s'|s,b)\cdot U(s')\right)$
        unchanged? := **false**

until unchanged?
return π

Policy Evaluation

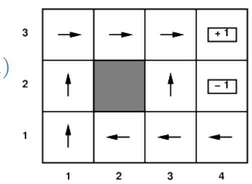
▷ **Problem:** How to implement the POLICY—EVALUATION algorithm?

▷ **Solution:** To compute utilities given a fixed π: For all s we have

$$U(s) = R(s) + \gamma(\sum_{s'} U(s') \cdot P(s'|s, \pi(s)))$$

▷ **Example 25.3.11 (Simplified Bellman Equations for π).**

$$\begin{aligned} U_i(1,1) &= -0.04 + 0.8U_i(1,2) + 0.1U_i(1,1) + 0.1U_i(2,1) \\ U_i(1,2) &= -0.04 + 0.8U_i(1,3) + 0.1U_i(1,2) \\ &\vdots \end{aligned}$$



▷ **Observation 25.3.12.** *n simultaneous linear equations in n unknowns, solve in $\mathcal{O}(n^3)$ with standard linear algebra methods.*

Modified Policy Iteration

▷ Policy iteration often converges in few iterations, but each is expensive.

▷ **Idea:** Use a few steps of value iteration (but with π fixed)
starting from thevalue function produced the last time
to produce an approximate value determination step.

▷ Often converges much faster than pure VI or PI.

▷ Leads to much more general algorithms where Bellman value updates and Howard policy updates can be performed locally in any order.

▷ **Remark:** Reinforcement learning algorithms operate by performing such updates based on the observed transitions made in an initially unknown environment.

25.4 Partially Observable MDPs

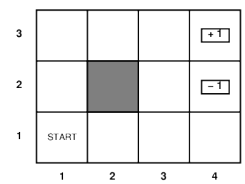
We will now lift the last restriction we made in the decision problems for our agents: in the definition of Markov decision processes we assumed that the environment was fully observable. As we have seen Observation 25.2.6 this entails that the optimal policy only depends on the current state. **A Video Nugget** covering this section can be found at `https://fau.tv/clip/id/30360`.

Partial Observability

▷ **Definition 25.4.1.** A partially observable MDP (a POMDP for short) is a MDP together with an sensor model O that has the sensor Markov property and is stationary: $O(s, e) = P(e|s)$.

▷ **Example 25.4.2 (Noisy 4x3 World).**

Add a partial and/or noisy sensor.
e.g. count number of adjacent walls $(1 \leq w \leq 2)$
with 0.1 error (noise)
If sensor reports 1, we are in $(3, ?)$ (probably)



▷ **Problem:** Agent does not know which state it is in $\rightsquigarrow$ makes no sense to talk about policy $\pi(s)$!

▷ **Theorem 25.4.3 (Astrom 1965).** *The optimal policy in a POMDP is a function $\pi(b)$ where b is the belief state (probability distribution over states).*

▷ **Idea:** Convert a POMDP into an MDP in belief state space, where $\mathcal{T}(b, a, b')$ is the probability that the new belief state is b' given that the current belief state is b and the agent does a. I.e., essentially a filtering update step.

POMDP: Filtering at the Belief State Level

▷ **Recap:** Filtering updates the belief state for new evidence.

▷ For POMDPs, we also need to consider actions. (but the effect is the same)

▷ If b is the previous belief state and agent does action a and then perceives e, then the new belief state is

$$b'(s') = \alpha \cdot P(e|s') \cdot (\sum_s P(s'|s, a) \cdot b(s))$$

We write $b' = \text{FORWARD}(b, a, e)$ in analogy to recursive state estimation.

▷ **Fundamental Insight for POMDPs:** The optimal action only depends on the agent's current belief state. (good, it does not know the state!)

▷ **Consequence:** The optimal policy can be written as a function $\pi^*(b)$ from belief states to actions.

▷ **Definition 25.4.4.** The POMDP decision cycle is to iterate over

1. Given the current belief state b, execute the action $a = \pi^*(b)$
2. Receive percept e.
3. Set the current belief state to $\text{FORWARD}(b, a, e)$ and repeat.

▷ **Intuition:** POMDP decision cycle is search in belief state space.

Partial Observability contd.

▷ **Recap:** The POMDP decision cycle is search in belief state space.

▷ **Observation 25.4.5.** *Actions change the belief state, not just the (physical) state.*

▷ **Thus** POMDP solutions automatically include information gathering behavior.

▷ **Problem:** The belief state is continuous: If there are n states, b is an n-dimensional real-valued vector.

▷ **Example 25.4.6.** The belief state of the 4x3 world is a 11 dimensional continuous space. (11 states)

▷ **Theorem 25.4.7.** *Solving POMDPs is very hard!* (*actually, **PSPACE** hard*)

▷ **In particular,** none of the algorithms we have learned applies. (discreteness assumption)

▷ The real world is a POMDP (with initially unknown transition model T and sensor model O)

Reducing POMDPs to Belief-State MDPs

▷ **Idea:** Calculating the probability that an agent in belief state b reaches belief-state b' after executing action a.

- ▷ if we knew the action and the subsequent percept, then $b' = \text{FORWARD}(b, a, e)$. (deterministic update to the belief state)
- ▷ but we don't, so b' depends on e. (let's calculate $P(e|a, b)$)

▷ **Idea:** To compute $P(e|a, b)$ — the probability that e is perceived after executing a in belief state b — sum up over all actual states the agent might reach:

$$\begin{aligned}
P(e|a,b) &= \sum_{s'} P(e|a,s',b) \cdot P(s'|a,b) \\
&= \sum_{s'} P(e|s') \cdot P(s'|a,b) \\
&= \sum_{s'} P(e|s') \cdot (\sum_{s} P(s'|s,a), b(s))
\end{aligned}$$

Write the probability of reaching b' from b, given action a, as $P(b'|b,a)$, then

$$\begin{aligned} P(b'|b,a) &= P(b'|a,b) = \sum_e P(b'|e,a,b) \cdot P(e|a,b) \\ &= \sum_e P(b'|e,a,b) \cdot (\sum_{s'} P(e|s') \cdot (\sum_s P(s'|s,a), b(s))) \end{aligned}$$

where $P(b'|e,a,b)$ is 1 if $b' = \text{FORWARD}(b,a,e)$ and 0 otherwise.

▷ **Observation:** This equation defines a transition model for belief state space!

▷ **Idea:** We can also define a reward function for belief states:

$$\rho(b) := \sum_s b(s) \cdot R(s)$$

i.e., the expected reward for the actual states the agent might be in.

▷ Together, $P(b'|b,a)$ and $\rho(b)$ define an (observable) MDP on the space of belief states.

▷ **Theorem 25.4.8.** *An optimal policy $\pi^*(b)$ for this MDP, is also an optimal policy for the original POMDP.*

▷ **Upshot:** Solving a POMDP on a physical state space can be reduced to solving an MDP on the corresponding belief-state space.

▷ **Remember:** The belief state is always observable to the agent, by definition.

Ideas towards Value-Iteration on POMDPs

▷ **Recap:** The value iteration algorithm from ?? computes one utility value per state.

▷ **Problem:** We have infinitely many belief states $\rightsquigarrow$ be more creative!

▷ **Observation:** Consider an optimal policy π^*

- ▷ applied in a specific belief state b: π^* generates an action,
- ▷ for each subsequent percept, the belief state is updated and a new action is generated ...

For this specific b: $\pi^* \widehat{=}$ a conditional plan!

▷ **Idea:** Think about conditional plans and how the expected utility of executing a fixed conditional plan varies with the initial belief state.(instead of optimal policies)

Expected Utilities of Conditional Plans on Belief States

▷ **Observation 1:** Let p be a conditional plan and $\alpha_p(s)$ the utility of executing p

in state s.

- ▷ the expected utility of p in belief state b is $\sum_s b(s) \cdot \alpha_p(s) \widehat{=} b \cdot \alpha_p$ as vectors.
- ▷ the expected utility of a fixed conditional plan varies linearly with b
- ▷ $\leadsto$ it corresponds to a hyperplane in belief state space.

▷ **Observation 2:** Let π^* be the optimal policy. At any given belief state b,

- ▷ π^* will choose to execute the conditional plan with highest expected utility
- ▷ the expected utility of b under the π^* is the utility of that plan:

$$U(b) = U^{\pi^*}(b) = \max_b (b \cdot \alpha_p)$$

- ▷ If the optimal policy π^* chooses to execute p starting at b, then it is reasonable to expect that it might choose to execute p in belief states that are very close to b;
- ▷ if we bound the depth of the conditional plans, then there are only finitely many such plans
- ▷ the continuous space of belief states will generally be divided into regions, each corresponding to a particular conditional plan that is optimal in that region.

▷ **Observation 3 (conbined):** The utility function $U(b)$ on belief states, being the maximum of a collection of hyperplanes, is piecewise linear and convex.

A simple Illustrating Example

▷ **Example 25.4.9.** A world with states 0 and 1, where $R(0) = 0$ and $R(1) = 1$ and two actions:

- ▷ "Stay" stays put with probability 0.9
- ▷ "Go" switches to the other state with probability 0.9.
- ▷ The sensor reports the correct state with probability 0.6.

Obviously, the agent should "Stay" when it thinks it's in state 1 and "Go" when it thinks it's in state 0.

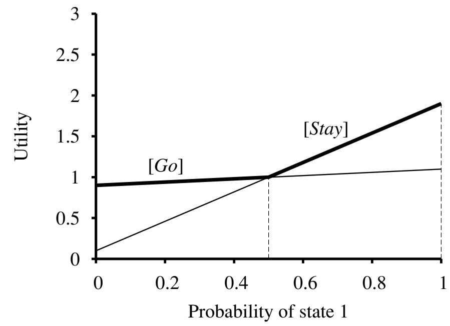
▷ The belief state has dimension 1. (the two probabilities sum up to 1)

▷ Consider the one-step plans $[Stay]$ and $[Go]$ and their (discounted) rewards:

$$\begin{aligned}
\alpha_{([Stay])}(0) &= R(0) + \gamma(0.9r(0) + 0.1r(1)) = 0.1 \\
\alpha_{([stay])}(1) &= r(1) + \gamma(0.9r(1) + 0.1r(0)) = 1.9 \\
\alpha_{([go])}(0) &= r(0) + \gamma(0.9r(1) + 0.1r(0)) = 0.9 \\
\alpha_{([go])}(1) &= r(1) + \gamma(0.9r(0) + 0.1r(1)) = 1.1
\end{aligned}$$

for now we will assume the discount factor $\gamma = 1$.

▷ Let us visualize the hyperplanes $b \cdot \alpha_{([Stay])}$ and $b \cdot \alpha_{([Go])}$.



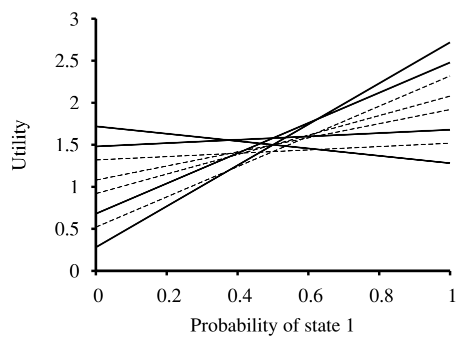
- ▷ The maximum represents the represents the utility function for the finite-horizon problem that allows just one action
- ▷ in each "piece" the optimal action is the first action of the corresponding plan.
- ▷ Here the optimal one-step policy is to "Stay" when $b(1) > 0.5$ and "Go" otherwise.

▷ compute the utilities for conditional plans of depth 2 by considering

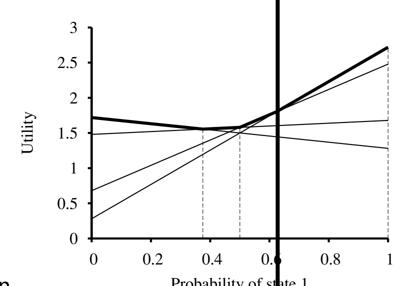
- ▷ each possible first action,
- ▷ each possible subsequent percept, and then
- ▷ each way of choosing a depth-1 plan to execute for each percept:

There are eight of depth 2:

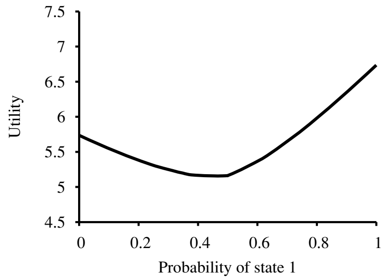
[$Stay$, if $P = 0$ then $Stay$ else $Stay$ fi], [$Stay$, if $P = 0$ then $Stay$ else Go fi], ...



Four of them (dashed lines) are suboptimal for the whole belief space
We call them dominated (they can be ignored)



▷ There are four undominated plans, each optimal in their region



- **Idea:** Repeat for depth 3 and so on.
- **Theorem 25.4.10 (POMDP Plan Utility).** *Let p be a depth-d conditional plan whose initial action is a and whose depth-$d-1$-subplan for percept e is $p.e$, then*

$$\alpha_p(s) = R(s) + \gamma(\sum_{s'} P(s'|s,a)(\sum_{e} P(e|s') \cdot \alpha_{p.e}(s')))$$

- This recursion naturally gives us a value iteration algorithm,

A Value Iteration Algorithm for POMDPs

- **Definition 25.4.11.** The POMDP value iteration algorithm for POMDPs is given by

function POMDP−VALUE−ITERATION($pomdp$, ϵ) **returns** a utility **function**
  inputs: $pomdp$, a POMDP with states S, actions $A(s)$, transition model $P(s'|s,a)$,
    sensor model $P(e|s)$, rewards $R(s)$, discount γ
    ϵ the maximum error allowed **in** the utility of any state
  local variables: U , U', sets of plans p with associated utility vectors α_p
    U' := a set containing just the empty plan $[]$, with $\alpha_{([])}(s) = R(s)$
    repeat
      $U := U'$
      U' := the set of all plans consisting of an action and, **for** each possible next percept,
        a plan **in** U with utility vectors computed via the POMDP Plan Utility Theorem
      U' := REMOVE−DOMPLANS(U')
    until MAX−DIFF(U,U') $< \epsilon(1-\gamma)/\gamma$
  return U

Where REMOVE−DOMPLANS and MAX−DIFF are implemented as linear programs.

- **Observations:** The complexity depends primarily on the generated plans:
  - Given $\#(A)$ actions and $\#(E)$ possible observations, there are are $\mathcal{O}(\#(A)^{\#(E)^{d-1}})$ distinct depth-d plans.
  - Even for the example with $d = 8$, we have 2255 (144 undominated)
  - The elimination of dominated plans is essential for reducing this doubly exponential growth (but they are already constructed)
- Hopelessly inefficient in practice – even the 3x4 POMDP is too hard!

25.5 Online Agents with POMDPs

In the last section we have seen that even though we can in principle compute utilities of states – and thus use the MEU principle – to make decisions in sequential decision problems, all methods based on the "lifting idea" are hopelessly inefficient.

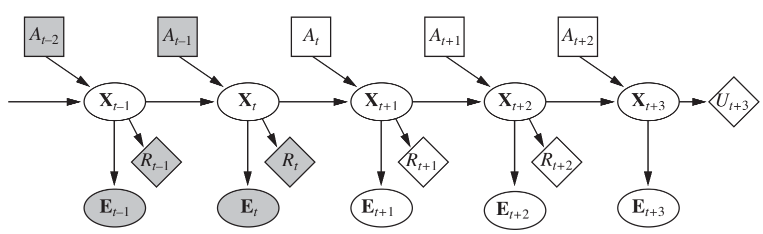
This section describes a different, approximate method for solving POMDPs, one based on look-ahead search. **A Video Nugget** covering this section can be found at `https://fau.tv/clip/id/30361`.

DDN: Decision Networks for POMDPs

- **Idea:** Let's try to use the computationally efficient representations (dynamic Bayesian networks and decision networks) for POMDPs.
- **Definition 25.5.1.** A dynamic decision network (DDN) is a graph-based representation of a POMDP, where
  - Transition and sensor model are represented as a DBN.
  - Action nodes and utility nodes are added as in decision networks.
- In a DDN, a filtering algorithm is used to incorporate each new percept and action and to update the belief state representation.
- Decisions are made in DDN by projecting forward possible action sequences and choosing the best one.
- DDNs – like the DBNs they are based on – are factored representations
  $\rightsquigarrow$ typically exponential complexity advantages!

Structure of DDNs for POMDPs

- **DDN for POMDPs:** The generic structure of a dymamic decision network at time t is

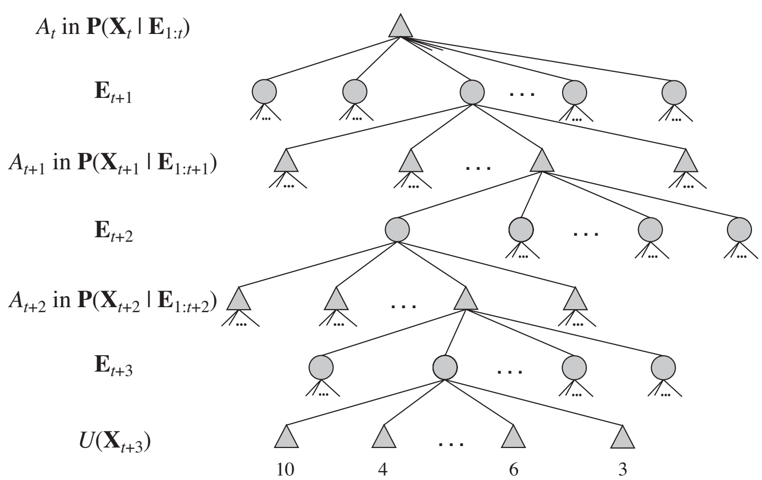


  - POMDP state S_t becomes a set of random variables X_t
  - there may be multiple evidence variables E_t
  - Action at time t denoted by A_t. agent must choose a value for A_t.
  - Transition model: $P(\mathsf{X}_{t+1}|\mathsf{X}_t, A_t)$; sensor model: $P(\mathsf{E}_t|\mathsf{X}_t)$.
  - Reward functions R_t and utility U_t of state S_t.
  - Variables with known values are gray, rewards for $t = 0, \ldots, t+2$, but utility for $t+3$ ($\widehat{=}$ discounted sum of rest)

- **Problem:** How do we compute with that?
- **Answer:** All POMDP algorithms can be adapted to DDNs! (only need CPTs)

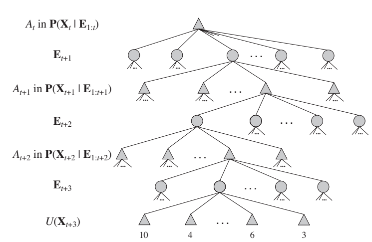
Lookahead: Searching over the Possible Action Sequences

- **Idea:** Search over the tree of possible action sequences (like in game-play)
- Part of the lookahead solution of the DDN above (three steps lookahead)



- circle $\widehat{=}$ chance nodes (the environment decides)
- triangle $\widehat{=}$ belief state (each action decision is taken there)

Designing Online Agents for POMDPs



- **Note:** belief state update is deterministic irrespective of the action outcome
  $\leadsto$ no chance nodes for action outcomes
- Belief state at triangle computed by filtering with actions/percepts leading to it
  - for decision A_{t+i} will use percepts $\mathbf{E}_{t+1:t+i}$ (even if values at time t unknown)

- ▹ thus a POMDP agent automatically takes into account the value of information and executes information-gathering actions where appropriate.

▷ **Observation:** Time complexity for exhaustive search up to depth d is $\mathcal{O}(|A|^d \cdot |\mathbf{E}|^d)$ ($|A| \widehat{=}$ number of actions, $|\mathbf{E}| \widehat{=}$ number of percepts)

▷ **Upshot:** Much better than POMDP value iteration with $\mathcal{O}(\#(A)^{\#(E)^{d-1}})$.

▷ **Empirically:** For problems in which the discount factor γ is not too close to 1, a shallow search is often good enough to give near-optimal decisions.

Summary

▷ Decision theoretic agents for sequential environments

▷ Building on temporal, probabilistic models/inference (dynamic Bayesian networks)

▷ MDPs for fully observable case.

▷ Value/Policy Iteration for MDPs $\rightsquigarrow$ optimal policies.

▷ POMDPs for partially observable case.

▷ POMDPs$\widehat{=}$ MDP on belief state space.

▷ The world is a POMDP with (initially) unknown transition and sensor models.

Part VI

Machine Learning

This part introduces the foundations of machine learning methods in AI. We discuss the problem learning from observations in general, study inference-based techniques, and then go into elementary statistical methods for learning.

The current hype topics of deep learning, reinforcement learning, and large language models are only very superficially covered, leaving them to specialized lectures.

Chapter 26

Learning from Observations

A Video Nugget covering the introduction to this chapter can be found at `https://fau.tv/clip/id/30369`.

In this chapter we introduce the concepts, methods, and limitations of inductive learning, i.e. learning from a set of given examples.

Outline

- ▷ Learning agents
- ▷ Inductive learning
- ▷ Decision tree learning
- ▷ Measuring learning performance
- ▷ Computational Learning Theory
- ▷ Linear regression and classification
- ▷ Neural Networks
- ▷ Support Vector Machines

26.1 Forms of Learning

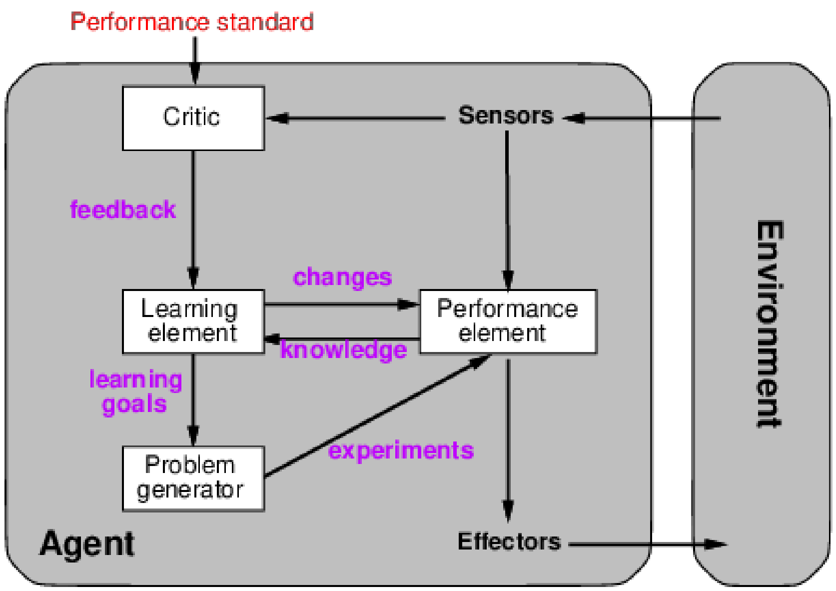
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30370`.

Learning (why is this a good idea)

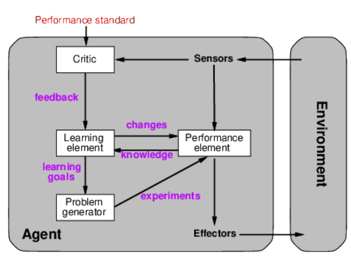
- ▷ Learning is essential for unknown environments:
  - ▷ i.e., when designer lacks omniscience.
  - ▷ The world is a POMDP with (initially) unknown transition and sensor models.
- ▷ Learning is useful as a system construction method.
  - ▷ i.e., expose the agent to reality rather than trying to write it down

▷ Learning modifies the agent's decision mechanisms to improve performance.

Recap: Learning Agents



Recap: Learning Agents (continued)



▷ **Definition 26.1.1.** Performance element is what we called "agent" up to now.

▷ **Definition 26.1.2.** Critic/learning element/problem generator do the "improving".

▷ **Definition 26.1.3.** Performance standard is fixed; (outside the environment)

- ▷ We can't adjust performance standard to flatter own behaviour!
- ▷ No standard *in the environment*: e.g. ordinary chess and suicide chess look identical.
- ▷ Essentially, certain kinds of percepts are "hardwired" as good/bad (e.g.,pain, hunger)

▷ **Definition 26.1.4.** Learning element may use knowledge already acquired in the performance element.

▷ **Definition 26.1.5.** Learning may require experimentation actions an agent might not normally consider such as dropping rocks from the Tower of Pisa.

Learning Element

▷ **Observation:** The design of learning element is dictated by

- ▷ what type of performance element is used,
- ▷ which functional component is to be learned,
- ▷ how that functional component is represented,
- ▷ what kind of feedback is available.

▷ **Example 26.1.6 (Learning Scenarios).**

Performance Elt.	Component	Representation	Feedback
Alpha-beta search	Evaluation fn.	Weighted linear fn.	Win/loss
Logical agent	transition model	Successor state ax.	Outcome
Utility-based agent	transition model	Dynamic Bayes net	Outcome
Simple reflex agent	Percept action fn.	Neural net	Corr. Action

▷ **Preview:**

- ▷ Supervised learning: correct answers for each instance
- ▷ Reinforcement learning: occasional rewards

Note:

1. Learning transition models is "supervised" if observable.
2. Supervised learning of correct actions requires "teacher".
3. Reinforcement learning is harder, but requires no teacher.

26.2 Inductive Learning

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30371`.

Inductive learning (a.k.a. Science)

▷ **Simplest form:** Learn a function from arg/value examples. (tabula rasa)

▷ **Definition 26.2.1.** An example is a pair (x,y) of an input sample x and a classification y. We call a set S of examples consistent, iff S is a function.

▷ **Example 26.2.2 (Examples in Tic-Tac-Toe).** $\left(\begin{array}{c|c|c} O & O & X \\ \hline & X & \\ \hline X & & \end{array}, +1\right)$

▷ **Definition 26.2.3.** The inductive learning problem $\mathcal{P}{:=}\langle\mathcal{H}, f, \approxeq\rangle$ consists in finding a hypothesis $h{\in}\mathcal{H}$ such that $f{\approxeq}(h|_{\mathbf{dom}(f)})$ for a consistent training set f of examples and a hypothesis space $\mathcal{H}$. We also call f the target function.

Inductive learning algorithms solve this problem.

▷ **Definition 26.2.4.** Inductive learning algorithms solve inductive learning problems.

▷ **Note:** This is a highly simplified model of what a learning agent does: it

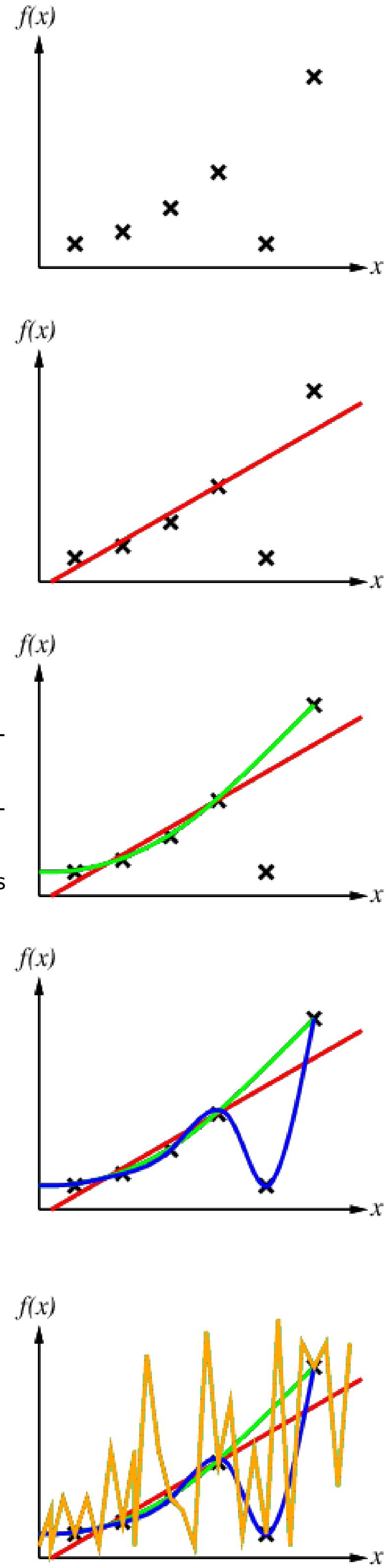
- ▷ ignores prior knowledge.
- ▷ assumes deterministic, observable environments.
- ▷ assumes examples are *given*.
- ▷ assumes that the agent *wants* to learn f. (why?)

Inductive Learning Method

▷ **Idea:** Construct/adjust hypothesis $h{\in}\mathcal{H}$ to agree with a training set f.

▷ **Definition 26.2.5.** We call h consistent with f (on a set $T \subseteq \mathbf{dom}(f)$), if it agrees with f on all examples in T.

▷ **Example 26.2.6 (Curve Fitting).**



Training SetLinear Hypothesis

partially, approximatively consistent Quadratic Hypothesis

partially consistent Degree-4 Hypothesis

consistent High-degree Hypothesis

consistent

- ▷ **Ockham's-razor:** maximize a combination of consistency and simplicity.

Choosing the Hypothesis Space

- ▷ **Observation:** Whether we can find a consistent hypothesis for a given training set depends on the chosen hypothesis space.
- ▷ **Definition 26.2.7.** We say that an inductive learning problem $\langle \mathcal{H}, f, \approxeq \rangle$ is realizable, iff there is a $h \in \mathcal{H}$ consistent with f.
- ▷ **Problem:** We do not know whether a given learning problem is realizable, unless we have prior knowledge.
- ▷ **Solution:** Make $\mathcal{H}$ large, e.g. the class of all Turing machines.
- ▷ **Tradeoff:** The computational complexity of the inductive learning problem is tied to the size of the hypothesis space. E.g. consistency is not even decidable for general Turing machines.
- ▷ Much of the research in machine learning has concentrated on simple hypothesis spaces.
- ▷ **Preview:** We will concentrate on propositional logic and related languages first.

26.3 Learning Decision Trees

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30372`.

Attribute-based Representations

- ▷ **Definition 26.3.1.** In attribute-based representations, examples are described by
  - ▷ attributes: (simple) functions on input samples, (think pre classifiers on examples)
  - ▷ their value, and (classify by attributes)
  - ▷ classifications. (Boolean, discrete, continuous, etc.)
- ▷ **Example 26.3.2 (In a Restaurant).** Situations where I will/won't wait for a table:

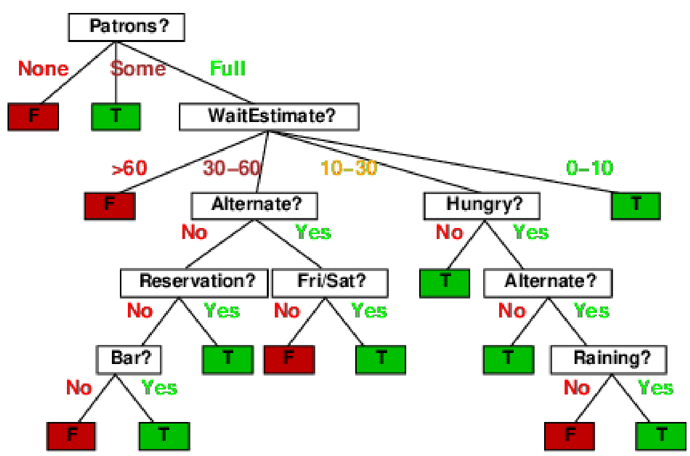
Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	WillWait
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

▷ **Definition 26.3.3.** Classification of examples is positive (T) or negative (F).

Decision Trees

▷ Decision trees are one possible representation for hypotheses.

▷ **Example 26.3.4 (Restaurant continued).** Here is the "true" tree for deciding whether to wait:



We evaluate the tree by going down the tree from the top, and always take the branch whose attribute matches the situation; we will eventually end up with a Boolean value; the result. Using the attribute values from X_3 in Example 26.3.2 to descend through the tree in Example 26.3.4 we indeed end up with the result "true". Note that

1. some of the original set of attributes X_3 are irrelevant.

2. the training set in Example 26.3.2 is realizable – i.e. the target is definable in hypothesis class of decision trees.

Decision Trees (Definition)

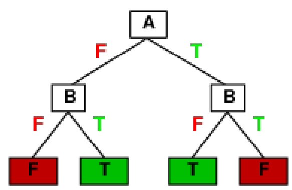
▷ **Definition 26.3.5.** A decision tree for a given attribute-based representation is a tree, where the non-leaf nodes are labeled by attributes, their outgoing edges by the corresponding attribute values, and the leaf nodes are labeled by the classifications.

Expressiveness

▷ Decision trees can express any function of the input attributes.

▷ **Example 26.3.6.** for Boolean functions, truth table row $\rightsquigarrow$ path to leaf:

A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F



▷ Trivially, for any training set there is a consistent hypothesis with one path to leaf for each example (unless f nondeterministic in x) but it probably won't generalize to new examples.

▷ **Solution:** Prefer to find more *compact* decision trees.

Hypothesis Spaces

▷ **Question:** How many distinct decision trees are there with n Boolean attributes?

▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

▷ **Question:** How many purely conjunctive hypotheses? (e.g., Hungry $\wedge \neg$Rain)

▷ **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

Decision Tree learning

▷ **Aim:** Find a small decision tree consistent with the training examples.

▷ **Idea:** (recursively) choose "most significant" attribute as root of (sub)tree.

▷ **Definition 26.3.7.** The following algorithm performs decision tree learning (DTL)

function DTL($examples$, $attributes$, $default$) **returns** a decision tree

```
if examples is empty then return default
  else if all examples have the same classification then return the classification
  else if attributes is empty then return MODE(examples)
  else
     best := Choose−Attribute(attributes, examples)
     tree := a new decision tree with root test best
     m := MODE(examples)
     for each value v_i of best do
        examples_i := {elements of examples with best = v_i}
        subtree := DTL(examples_i, attributes \ best, m)
        add a branch to tree with label v_i and subtree subtree
   return tree
```

MODE($examples$)= most frequent value in $example$.

Michael Kohlhase: Artificial Intelligence 2 934 2023-09-20

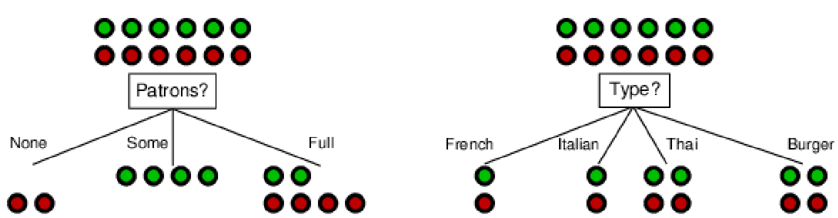
Note: We have three base cases:

1. empty examples $\rightsquigarrow$ arises for empty branches of non Boolean parent attribute.
2. uniform example classifications $\rightsquigarrow$ this is "normal" leaf.
3. $attributes$ empty $\rightsquigarrow$ target is not deterministic in input $attributes$.

The recursive step steps pick an attribute and then subdivides the examples.

Choosing an Attribute

- **Idea:** A good attribute splits the examples into subsets that are (ideally) "all positive" or "all negative".
- **Example 26.3.8.**



Attribute "Patrons?" is a better choice, it gives gives information about the classification.

- Can we make this more formal? $\rightsquigarrow$ Use information theory! (up next)

Michael Kohlhase: Artificial Intelligence 2 935 2023-09-20

26.4 Using Information Theory

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/20373` and `https://fau.tv/clip/id/30374`.

Information Entropy

▷ **Intuition:** Information answers questions.

▷ The more clueless I am about the answer initially, the more information is contained in the answer.

▷ **Scale:** $1\mathrm{b} \mathrel{\widehat{=}} 1$ bit $\mathrel{\widehat{=}}$ answer to Boolean question with prior probability $(0.5,0.5)$.

▷ **Definition 26.4.1.**

If the prior probability is $\langle P_1,\ldots,P_n\rangle$, then the information in an answer (also called entropy of the prior) is

$$I(\langle P_1,\ldots,P_n\rangle) := \sum_{i=1}^{n} -P_i \cdot \log_2(P_i)$$

Note: The case $P_i = 0$ requires special treatment. ($\log_2(0)$ is undefined)

▷ **Example 26.4.2 (Information of a Coin Toss).**

- ▷ For a fair coin toss we have $I(\langle\frac{1}{2},\frac{1}{2}\rangle) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1\mathrm{b}$.
- ▷ With a loaded coin (99% heads) we have $I(\langle\frac{1}{100},\frac{99}{100}\rangle) = 0.08\mathrm{b}$.
- ▷ **Intuition**: Information goes to 0 as head probability goes to 1.

Information Gain in Decision Trees

▷ Suppose we have p examples classified as positive and n examples as negative.

▷ **Idea:** We can estimate the probability distribution of the classification C with $\mathbf{P}(C) = \langle p/(p+n), n/(p+n)\rangle$.

▷ Then $I(\mathbf{P}(C))$ bits are needed to classify a new example.

▷ **Example 26.4.3.** For 12 restaurant examples, $p = n = 6$ so we need $I(\mathbf{P}(\mathsf{WillWait})) = I(\langle\frac{6}{12},\frac{6}{12}\rangle) = 1\mathrm{b}$ of information.

▷ Treating attributes also as random variables, we can compute how much information is needed *after* knowing the value for one attribute.

▷ **Example 26.4.4.** If we know $\mathsf{Pat} = \mathsf{Full}$, we only need $I(\mathbf{P}(\mathsf{WillWait}|\mathsf{Pat} = \mathsf{Full})) = I(\langle\frac{4}{6},\frac{2}{6}\rangle)$ bits of information.

▷ **Note:** The expected number of bits needed after an attribute test on A is

$$\sum_a P(A = a) \cdot I(\mathbf{P}(C|A = a))$$

▷ **Definition 26.4.5.** The information gain from an attribute test A is

$$\mathsf{Gain}(A) := I(\mathbf{P}(C)) - \sum_a P(A = a) \cdot I(\mathbf{P}(C|A = a))$$

Information Gain (continued)

▷ **Definition 26.4.6.** Assume we know the results of some attribute tests $b{:=}B_1 = b_1 \wedge \ldots \wedge B_n = b_n$. Then the conditional information gain from an attribute test A is

$$\mathrm{Gain}(A|b){:=}I(\mathbf{P}(C|b)) - \sum_a P(A = a|b) \cdot I(\mathbf{P}(C|a, b))$$

▷ **Example 26.4.7.** If the classification C is Boolean and we have p positive and n negative examples, the information gain is

$$\mathrm{Gain}(A) = I(\langle\frac{p}{p+n}, \frac{n}{p+n}\rangle) - \sum_a \frac{p_a + n_a}{p+n} I(\langle\frac{p_a}{p_a+n_a}, \frac{n_a}{p_a+n_a}\rangle)$$

where p_a and n_a are the positive and negative examples with $A = a$.

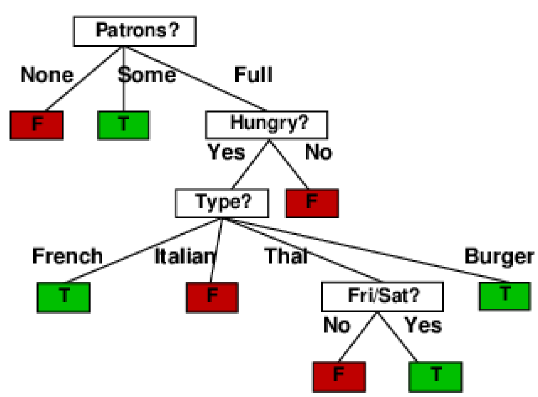
▷ **Example 26.4.8.**

$$\begin{aligned}
\mathrm{Gain}(Patrons?) &= 1 - \frac{2}{12}I(\langle 0,1\rangle) + \frac{4}{12}I(\langle 1,0\rangle) + \frac{6}{12}I(\langle\frac{2}{6},\frac{4}{6}\rangle) \\
&\approx 0.541\mathrm{b} \\
\mathrm{Gain}(Type) &= 1 - \frac{2}{12}I(\langle\frac{1}{2},\frac{1}{2}\rangle) + \frac{2}{12}I(\langle\frac{1}{2},\frac{1}{2}\rangle) + \frac{4}{12}I(\langle\frac{2}{4},\frac{2}{4}\rangle) + \frac{4}{12}I(\langle\frac{2}{4},\frac{2}{4}\rangle) \\
&\approx 0\mathrm{b}
\end{aligned}$$

▷ **Idea:** Choose the attribute that maximizes information gain.

Restaurant Example contd.

▷ **Example 26.4.9.** Decision tree learned by DTL from the 12 examples using information gain maximization for Choose−Attribute:



▷ **Result:** Substantially simpler than "true" tree – a more complex hypothesis isn't justified by small amount of data.

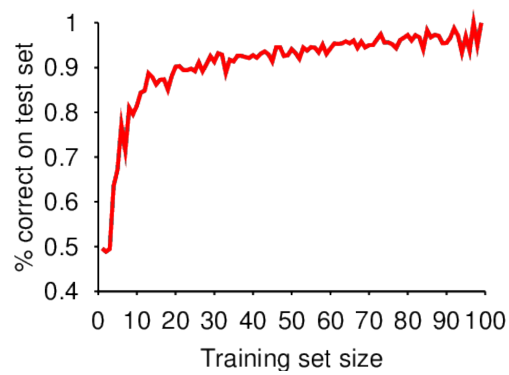
Performance measurement

▷ **Question:** How do we know that $h \approx f$? (Hume's *Problem of Induction*)

1. Use theorems of computational/statistical learning theory.
2. Try h on a new test set of examples. (use *same distribution over example space* as training set)

▷ **Definition 26.4.10.** The learning curve $\widehat{=}$ percentage correct on test set as a function of training set size.

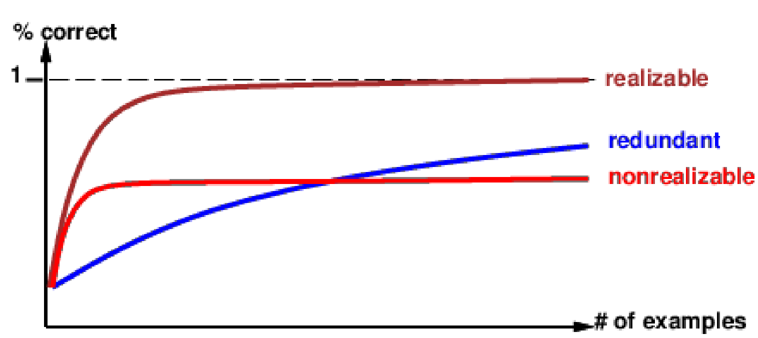
▷ **Example 26.4.11.** Restaurant data; graph averaged over 20 trials



Performance measurement contd.

▷ **Observation 26.4.12.** *The learning curve depends on*

- ▷ *realizable (can express target function) vs. non-realizable non-realizability can be due to missing attributes or restricted hypothesis class (e.g., thresholded linear function)*
- ▷ *redundant expressiveness (e.g., lots of irrelevant attributes)*



Generalization and Overfitting

- **Observation:** Sometimes a learned hypothesis is more specific than the experiments warrant.
- **Definition 26.4.13.** We speak of overfitting, if a hypothesis h describes random error in the (limited) training set rather than the underlying relationship. Underfitting occurs when h cannot capture the underlying trend of the data.
- **Qualitatively:** Overfitting increases with the size of hypothesis space and the number of attributes, but decreases with number of examples.
- **Idea:** Combat overfitting by "generalizing" decision trees computed by DTL.

Decision Tree Pruning

- **Idea:** Combat overfitting by "generalizing" decision trees $\leadsto$ prune "irrelevant" nodes.
- **Definition 26.4.14.** For decision tree pruning repeat the following on a learned decision tree:
  - Find a terminal test node n (only result leaves as children)
  - If test is irrelevant, i.e. has low information gain, prune it by replacing n by with a leaf node.
- **Question:** How big should the information gain be to split ($\leadsto$ keep) a node?
- **Idea:** Use a statistical significance test.
- **Definition 26.4.15.** A result has statistical significance, if the probability they could arise from the null hypothesis (i.e. the assumption that there is no underlying pattern) is very low (usually 5%).

Determining Attribute Irrelevance

- For decision tree pruning, the null hypothesis is that the attribute is irrelevant.
- Compute the probability that the example distribution (p positive, n negative) for a terminal node deviates from the expected distribution under the null hypothesis.
- For an attribute A with d values, compare the actual numbers p_k and n_k in each subset s_k with the expected numbers (expected if A is irrelevant) $\widehat{p}_k = p \cdot \frac{p_k+n_k}{p+n}$ and $\widehat{n}_k = n \cdot \frac{p_k+n_k}{p+n}$.

▷ A convenient measure of the total deviation is (sum of squared errors)

$$\Delta = \sum_{k=1}^{d} \frac{(p_k - \widehat{p}_k)^2}{\widehat{p}_k} + \frac{(n_k - \widehat{n}_k)^2}{\widehat{n}_k}$$

▷ **Lemma 26.4.16 (Neyman-Pearson).** *Under the null hypothesis, the value of Δ is distributed according to the χ^2 distribution with $d-1$ degrees of freedom. [JN33]*

▷ **Definition 26.4.17.** Decision tree pruning with Pearson's χ^2 with $d-1$ degrees of freedom for Δ is called χ^2 pruning. (χ^2 values from stats library.)

▷ **Example 26.4.18.** The $type$ attribute has four values, so three degrees of freedom, so $\Delta = 7.82$ would reject the null hypothesis at the 5% level.

26.5 Evaluating and Choosing the Best Hypothesis

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30375` and `https://fau.tv/clip/id/30376`.

Independent and Identically Distributed

▷ **Problem:** We want to learn a hypothesis that fits the future data best.

▷ **Intuition:** This only works, if the training set is "representative" for the underlying process.

▷ **Idea:** We think of examples (seen and unseen) as a sequence, and express the "representativeness" as a *stationarity assumption* for the probability distribution.

▷ **Method:** Each example before we see it is a random variable E_j, the observed value $e_j = (x_j, y_j)$ samples its distribution.

▷ **Definition 26.5.1.** A sequence of $E_1, \ldots, E_n$ of random variables is independent and identically distributed (short IID), iff they are

- ▷ independent, i.e. $\mathbf{P}(E_j | E_{(j-1)}, E_{(j-2)}, \ldots) = \mathbf{P}(E_j)$ and
- ▷ identically distributed, i.e. $\mathbf{P}(E_i) = \mathbf{P}(E_j)$ for all i and j.

▷ **Example 26.5.2.** A sequence of die tosses is IID. (fair or loaded does not matter)

▷ **Stationarity Assumption:** We assume that the set $\mathcal{E}$ of examples is IID in the future.

Error Rates and Cross-Validation

▷ **Recall:** We want to learn a hypothesis that fits the future data best.

▷ **Definition 26.5.3.** Given an inductive learning problem $\langle \mathcal{H}, f, \approxeq \rangle$, we define the error rate of a hypothesis $h \in \mathcal{H}$ as the fraction of errors:

$$\frac{\#(\{x \in \mathbf{dom}(f) \,|\, h(x) \neq f(x)\})}{\#(\mathbf{dom}(f))}$$

▷ **Caveat:** A low error rate on the training set does not mean that a hypothesis generalizes well.

▷ **Idea:** Do not use homework questions in the exam.

▷ **Definition 26.5.4.** The practice of splitting the data available for learning into

1. a training set from which the learning algorithm produces a hypothesis h and
2. a test set, which is used for evaluating h

is called holdout cross validation. (no peeking at test set allowed)

Error Rates and Cross-Validation

▷ **Question:** What is a good ratio between training set and test set size?

- ▷ small training set ⇝ poor hypothesis.
- ▷ small test set ⇝ poor estimate of the accuracy.

▷ **Definition 26.5.5.** In k fold cross validation, we perform k rounds of learning, each with $1/k$ of the data as test set and average over the k error rates.

▷ **Intuition:** Each example does double duty: for training and testing.

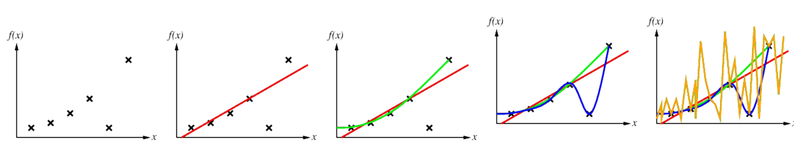
▷ $k=5$ and $k=10$ are popular ⇝ good accuracy at k times computation time.

▷ **Definition 26.5.6.** If $k = \#(\mathbf{dom}(f))$, then k fold cross validation is called leave one out cross validation (LOOCV).

Model Selection

▷ **Definition 26.5.7.** The model selection problem is to determine – given data – a good hypothesis space.

▷ **Example 26.5.8.** What is the best polynomial degree to fit the data



▷ **Observation 26.5.9.** *We can solve the problem of "learning from observations f" in a two-part process:*

1. *model selection determines a hypothesis space $\mathcal{H}$,*
2. *optimization solves the induced inductive learning problem $\langle\mathcal{H}, f, \approxeq\rangle$.*

▷ **Idea:** Solve the two parts together by iteration over "size". (they inform each other)

▷ **Problem:** Need a notion of "size" ⇝ e.g. number of nodes in a decision tree.

▷ **Concrete Problem:** Find the "size" that best balances overfitting and underfitting to optimize test set accuracy.

Model Selection Algorithm (Wrapper)

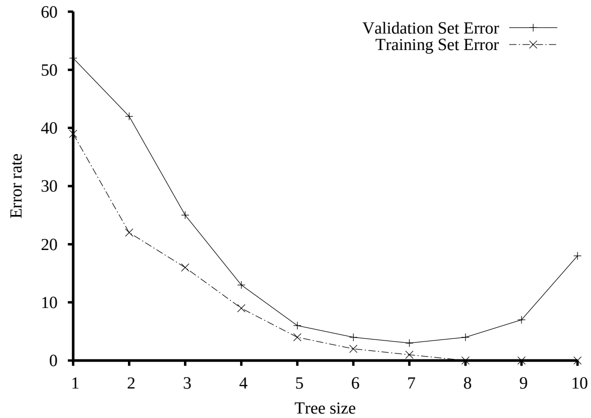
function CROSS–VALIDATION–WRAPPER($Learner$,k,$examples$) **returns** a hypothesis
  local variables: $errT$, an array, indexed by size, storing training–set error rates
      $errV$, an array, indexed by size, storing validation–set error rates
  for size = 1 **to** ∞ **do**
   $errT[size]$, $errV[size]$:= CROSS–VALIDATION($Learner$,$size$,k,$examples$)
   if $errT$ has converged **then do**
    $best\_size$:= the value of size with minimum $errV[size]$
    return Learner($best\_size$,$examples$)

function CROSS–VALIDATION($Learner$,$size$,k,$examples$) **returns** two values:
    average training set error rate, average validation set error rate
  $fold\_errT$:= 0; $fold\_errV$:= 0
  for fold = 1 **to** k **do**
    $training\_set$, $validation\_set$:= PARTITION($examples$,$fold$,k)
    h := Learner($size$,$training\_set$)
    $fold\_errT$:= $fold\_errT$ + ERROR–RATE(h,$training\_set$)
    $fold\_errV$:= $fold\_errV$ + ERROR–RATE(h,$validation\_set$)
   return $fold\_errT/k$, $fold\_errV/k$

function PARTITION($examples$,$fold$,k) **returns** two sets:
  a validation set of size $|examples|/k$ and the rest; the split is different **for** each $fold$ value

Error Rates on Training/Validation Data

▷ **Example 26.5.10 (An Error Curve for Restaurant Decision Trees).** Modify DTL to be breadth-first, information gain sorted, stop after k nodes.



Stops when training set error rate converges, choose optimal tree for validation curve. (here a tree with 7 nodes)

From Error Rates to Loss Functions

- **So far** we have been minimizing error rates. (better than maximizing ☺)
- **Example 26.5.11 (Classifying Spam).** It is much worse to classify ham (legitimate mails) as spam than vice versa. (message loss)
- **Recall Rationality:** Decision-makers should maximize expected utility (MEU).
- **So:** Machine learning should maximize "utility". (not only minimize error rates)
- machine learning traditionally deals with utilities in form of "loss functions".
- **Definition 26.5.12.** The loss function L is defined by setting $L(x,y,\widehat{y})$ to be the amount of utility lost by prediction $h(x) = \widehat{y}$ instead of $f(x) = y$. If L is independent of x, we often use $L(y,\widehat{y})$.
- **Example 26.5.13.** $L(spam,ham) = 1$, while $L(ham,spam) = 10$.

Generalization Loss

- **Note:** $L(y,y) = 0$. (no loss if you are exactly correct)
- **Definition 26.5.14 (Popular general loss functions).**

absolute value loss	$L_1(y,\widehat{y}) := \lvert y-\widehat{y}\rvert$	small errors are good
squared error loss	$L_2(y,\widehat{y}) := (y-\widehat{y})^2$	dito
0/1 loss	$L_{0/1}(y,\widehat{y}) := 0$, if $y=\widehat{y}$, else 1	error rate

- **Idea:** Maximize expected utility by choosing hypothesis h that minimizes expectationexpected loss over all $(x,y)\in f$.

▷ **Definition 26.5.15.** Let $\mathcal{E}$ be the set of all possible examples and $\mathbf{P}(X,Y)$ the prior probability distribution over its components, then the expected generalization loss for a hypothesis h with respect to a loss function L is

$$\mathrm{GenLoss}_L(h) := \sum_{(x,y)\in\mathcal{E}} L(y,h(x))\cdot P(x,y)$$

and the best hypothesis $h^* := \underset{h\in\mathcal{H}}{\mathrm{argmin}}\ \mathrm{GenLoss}_L(h)$.

Empirical Loss

▷ **Problem:** $\mathbf{P}(X,Y)$ is unknown $\leadsto$ learner can only estimate generalization loss:

▷ **Definition 26.5.16.** Let L be a loss function and E a set of examples with $\#(E) = N$, then we call

$$\mathrm{EmpLoss}_{L,E}(h) := \frac{1}{N}\Big(\sum_{(x,y)\in E} L(y,h(x))\Big)$$

the empirical loss and $\widehat{h}^* := \underset{h\in\mathcal{H}}{\mathrm{argmin}}\ \mathrm{EmpLoss}_{L,E}(h)$ the estimated best hypothesis.

▷ There are four reasons why $\widehat{h}^*$ may differ from f:

1. Realizablility: then we have to settle for an approximation $\widehat{h}^*$ of f.
2. Variance: different subsets of f give different $\widehat{h}^* \leadsto$ more examples.
3. Noise: if f is non deterministic, then we cannot expect perfect results.
4. Computational complexity: if $\mathcal{H}$ is too large to systematically explore, we make due with subset and get an approximation.

Regularization

▷ **Idea:** Directly use empirical loss to solve model selection. (finding a good $\mathcal{H}$)

Minimize the weighted sum of empirical loss and hypothesis complexity. (to avoid overfitting).

▷ **Definition 26.5.17.** Let $\lambda\in\mathbb{R}$, $h\in\mathcal{H}$, and E a set of examples, then we call

$$\mathrm{Cost}_{L,E}(h) := \mathrm{EmpLoss}_{L,E}(h) + \lambda\mathrm{Complexity}(h)$$

the total cost of h on E.

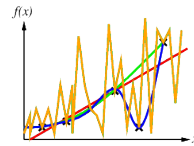
▷ **Definition 26.5.18.** The process of finding a total cost minimizing hypothesis

$$\widehat{h}^* := \underset{h\in\mathcal{H}}{\mathrm{argmin}}\ \mathrm{Cost}_{L,E}(h)$$

is called regularization; Complexity is called the regularization function or hypothesis complexity.

▷ **Example 26.5.19 (Regularization for Polynomials).**

A good regularization function for polynomials is the sum of squares of exponents. $\rightsquigarrow$ keep away from wriggly curves!



Minimal Description Length

▷ **Remark:** In regularization, empirical loss and hypothesis complexity are not measured in the same scale $\rightsquigarrow$ λ mediates between scales.

▷ **Idea:** Measure both in the same scale $\rightsquigarrow$ use information content, i.e. in bits.

▷ **Definition 26.5.20.** Let $h{\in}\mathcal{H}$ be a hypothesis and E a set of examples, then the description length of (h,E) is computed as follows:

1. encode the hypothesis as a Turing machine program, count bits.
2. count data bits:
   - ▷ correctly predicted example $\rightsquigarrow$ 0b
   - ▷ incorrectly predicted example $\rightsquigarrow$ according to size of error.

The minimum description length or MDL hypothesis minimizes the total number of bits required.

▷ This works well in the limit, but for smaller problems there is a difficulty in that the choice of encoding for the program affects the outcome.

- ▷ e.g., how best to encode a decision tree as a bit string?

The Scale of Machine Learning

▷ Traditional methods in statistics and early machine learning concentrated on small-scale learning (50-5000 examples)

- ▷ Generalization error mostly comes from
  - ▷ approximation error of not having the true f in the hypothesis space
  - ▷ estimation error of too few training examples to limit variance.

▷ In recent years there has been more emphasis on large-scale learning. (millions of examples)

- Generalization error is dominated by limits of computation
  - there is enough data and a rich enough model that we could find an h that is very close to the true f,
  - but the computation to find it is too complex, so we settle for a sub-optimal approximation.
- Hardware advances (GPU farms, Amazon EC2, Google Data Centers, …) help.

26.6 Computational Learning Theory

Video Nuggets covering this section can be found at https://fau.tv/clip/id/30377 and https://fau.tv/clip/id/30378.

A (General) Theory of Learning?

- **Main Question:** How can we be sure that our learning algorithm has produced a hypothesis that will predict the correct value for previously unseen inputs?
- **Formally:** How do we know that the hypothesis h is close to the target function f if we don't know what f is?
- **Other - more recent - Questions:**
  - How many examples do we need to get a good h?
  - What hypothesis space $\mathcal{H}$ should we use?
  - If the $\mathcal{H}$ is very complex, can we even find the best h, or do we have to settle for a local maximum in $\mathcal{H}$.
  - How complex should h be?
  - How do we avoid overfitting?
- "Computational Learning Theory" tries to answer these using concepts from AI, statistics, and theoretical CS.

PAC Learning

- **Basic idea of Computational Learning Theory:**
  - Any hypothesis h that is seriously wrong will almost certainly be "found out" with high probability after a small number of examples, because it will make an incorrect prediction.
  - Thus, if h is consistent with a sufficiently large set of training examples is unlikely to be seriously wrong.
  - $\leadsto$ h is probably approximately correct.
- **Definition 26.6.1.** Any learning algorithm that returns hypotheses that are probably approximately correct is called a PAC learning algorithm.

▷ Derive performance bounds for PAC learning algorithms in general, using the

▷ **Stationarity Assumption (again):** We assume that the set $\mathcal{E}$ of possible examples is IID $\rightsquigarrow$ we have a fixed distribution $\mathbf{P}(E) = \mathbf{P}(X, Y)$ on examples.

▷ **Simplifying Assumptions:** f is a function (deterministic) and $f \in \mathcal{H}$.

PAC Learning

▷ Start with PAC theorems for Boolean functions, for which $L_{0/1}$ is appropriate.

▷ **Definition 26.6.2.** The error rate error(h) of a hypothesis h is the probability that h misclassifies a new example.

$$\text{error}(h) := \text{GenLoss}_{L_{0/1}}(h) = \sum_{(x,y)\in\mathcal{E}} L_{0/1}(y, h(x)) \cdot P(x, y)$$

▷ **Intuition:** error(h) is the probability that h misclassifies a new example.

▷ This is the same quantity as measured in the learning curves above.

▷ **Definition 26.6.3.** A hypothesis h is called approximatively correct, iff error$(h) \leq \epsilon$ for some small $\epsilon > 0$.

We write $\mathcal{H}_b := \{h \in \mathcal{H} \,|\, \text{error}(h) > \epsilon\}$ for the "seriously bad" hypotheses.

Sample Complexity

▷ Let's compute the probability that $h_b \in \mathcal{H}_b$ is consistent with the first N examples.

▷ We know error$(h_b) > \epsilon$
$\rightsquigarrow$ $P(h_b$ agrees with N examples$) \leq (1-\epsilon)^N$. (independence)
$\rightsquigarrow$ $P(\mathcal{H}_b$ contains consistent hyp.$) \leq \#(\mathcal{H}_b) \cdot (1-\epsilon)^N \leq \#(\mathcal{H}) \cdot (1-\epsilon)^N$.
($\mathcal{H}_b \subseteq \mathcal{H}$)
$\rightsquigarrow$ to bound this by a small δ, show the algorithm $N \geq \frac{1}{\epsilon} \cdot (\log_2(\frac{1}{\delta}) + \log_2(\#(\mathcal{H})))$ examples.

▷ **Definition 26.6.4.** The number of required examples as a function of ϵ and δ is called the sample complexity of $\mathcal{H}$.

▷ **Example 26.6.5.** If $\mathcal{H}$ is the set of n-ary Boolean functions, then $\#(\mathcal{H}) = 2^{2^n}$.
$\rightsquigarrow$ sample complexity grows with $\mathcal{O}(\log_2(2^{2^n})) = \mathcal{O}(2^n)$.
There are 2^n possible examples,
$\rightsquigarrow$ PAC learning for Boolean functions needs to see (nearly) all examples.

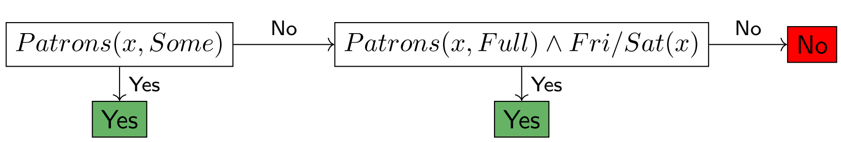
Escaping Sample Complexity

- **Problem:** PAC learning for Boolean functions needs to see (nearly) all examples.
  - $\mathcal{H}$ contains enough hypotheses to classify any given set of examples in all possible ways.
  - In particular, for any set of N examples, the set of hypotheses consistent with those examples contains equal numbers of hypotheses that predict x_{N+1} to be positive and hypotheses that predict x_{N+1} to be negative.
- **Idea/Problem:** restrict the $\mathcal{H}$ in some way (but we may lose realizability)
- **Three Ways out of this Dilemma:**

1. bring prior knowledge into the problem. (section 28.4)
2. prefer simple hypotheses. (e.g. decision tree pruning)
3. focus on "learnable subsets" of $\mathcal{H}$. (next)

PAC Learning: Decision Lists

- **Idea:** Apply PAC learning to a "learnable hypothesis space".
- **Definition 26.6.6.** A decision list consists of a sequence of tests, each of which is a conjunction of literals.
  - If a test succeeds when applied to an example description, the decision list specifies the value to be returned.
  - If the test fails, processing continues with the next test in the list.
- **Remark:** Like decision trees, but restricted branching, but more complex tests.
- **Example 26.6.7 (A decision list for the Restaurant Problem).**



- **Lemma 26.6.8.** *Given arbitrary size conditions, decision lists can represent arbitrary Boolean functions.*
- This directly defeats our purpose of finding a "learnable subset" of $\mathcal{H}$.

Decision Lists: Learnable Subsets (Size-Restricted Cases)

- **Definition 26.6.9.** The set of decision lists where tests are of conjunctions of at most k literals is denoted by k-**DL**.

▷ **Example 26.6.10.** The decision list from Example 26.6.7 is in 2-**DL**.

▷ **Observation 26.6.11.** *k-**DL** contains k-**DT**, the set of decision trees of depth at most k.*

▷ **Definition 26.6.12.** We denote the set of k-**DL** decision lists with at most n Boolean attributes with k-**DL**(n). The language of conjunctions of at most k literals using n attributes is written as $\mathrm{Conj}(k,n)$.

▷ Decision lists are constructed of optional yes/no tests, so there are at most $3^{|\mathrm{Conj}(k,n)|}$ distinct sets of component tests. Each of these sets of tests can be in any order, so $|k\text{-}\mathbf{DL}(n)| \leq 3^{|\mathrm{Conj}(k,n)|} \cdot |\mathrm{Conj}(k,n)|!$

Decision Lists: Learnable Subsets (Sample Complexity)

▷ The number of conjunctions of k literals from n attributes is given by

$$|\mathrm{Conj}(k,n)| = \sum_{i=1}^{k} \binom{2n}{i}$$

thus $|\mathrm{Conj}(k,n)| = \mathcal{O}(n^k)$. Hence, we obtain (after some work)

$$|k\text{-}\mathbf{DL}(n)| = 2^{\mathcal{O}(n^k \log_2(n^k))}$$

▷ Plug this into the equation for the sample complexity: $N \geq \frac{1}{\epsilon} \cdot (\log_2(\frac{1}{\delta}) + \log_2(|\mathcal{H}|))$ to obtain

$$N \geq \frac{1}{\epsilon} \cdot (\log_2(\frac{1}{\delta}) + \log_2(\mathcal{O}(n^k \log_2(n^k))))$$

▷ **Intuitively:** Any algorithm that returns a consistent decision list will PAC learn a k-**DL** function in a reasonable number of examples, for small k.

Decision Lists Learning

▷ **Idea:** Use a greedy search algorithm that repeats

1. **find** test that agrees exactly with some subset E of the training set,
2. **add** it to the decision list under construction and removes E,
3. **construct** the remainder of the DL using just the remaining examples,

until there are no examples left.

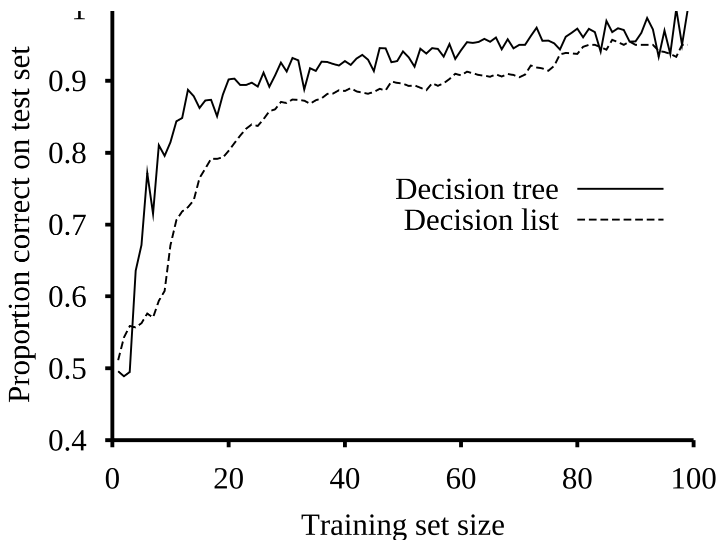
▷ **Definition 26.6.13.** The following algorithm performs decision list learning

function DLL(E) **returns** a decision list, or failure
  if E is empty **then return** (the trivial decision list) No

$t :=$ a test that matches a nonempty subset E_t of E
such that the members of E_t are all positive or all negative
if there is no such t **then return** failure
if the examples **in** E_t are positive **then** $o :=$ Yes **else** $o :=$ No
return a decision list with initial test t and outcome o and remaining tests given by
DLL($E \backslash E_t$)

Decision Lists Learning in Comparison

▷ **Learning curves:** for DLL (and DTL for comparison)



▷ **Upshot:** The simpler DLL works quite well!

26.7 Regression and Classification with Linear Models

Video Nuggets covering this section can be found at https://fau.tv/clip/id/30379, https://fau.tv/clip/id/30380, and https://fau.tv/clip/id/30381.

Linear Regression and Classification

▷ We pass on to another hypothesis space: linear functions over continuous-valued inputs.

▷ **Definition 26.7.1.** We call an inductive learning problem $\langle \mathcal{H}, f, \approxeq \rangle$

- ▹ a classification problem, iff $\mathbf{codom}(f)$ is countable; the members of $\mathbf{codom}(f)$ are called classes, and
- ▹ a regression problem if $\mathbf{codom}(f)$ is uncountable (usually real valued).

▷ **Definition 26.7.2.** An algorithm that solves a classification problem is called a classifier.

Univariate Linear Regression

▷ **Definition 26.7.3.** A univariate or unary function is a function with one argument.

▷ **Recall:** A mapping between vector spaces is called linear, iff it preserves plus and scalar multiplication.

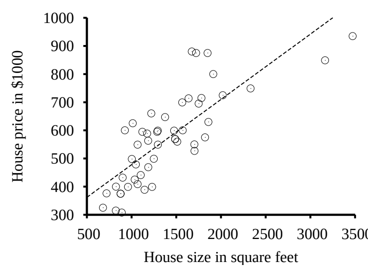
▷ **Observation 26.7.4.** *A univariate, linear function $f : \mathbb{R} \to \mathbb{R}$ is of the form $f(x) = \mathbf{w}_1 x + \mathbf{w}_0$ for some $\mathbf{w}_i \in \mathbb{R}$.*

▷ **Definition 26.7.5.** Given a vector $\mathbf{w} := (\mathbf{w}_0, \mathbf{w}_1)$, we define $h_{\mathbf{w}}(x) := \mathbf{w}_1 x + \mathbf{w}_0$.

▷ **Definition 26.7.6.** Given a set of examples $E \subseteq \mathbb{R} \times \mathbb{R}$, the task of finding $h_{\mathbf{w}}$ that best fits E is called linear regression.

▷ **Example 26.7.7.**

Examples of house price vs. square feet in houses sold in Berkeley in July 2009.
Also: linear function hypothesis that minimizes squared error loss $y = 0.232x + 246$.



Univariate Linear Regression by Loss Minimization

▷ **Idea:** Minimize squared error loss over $\{(x_i, y_i) | i \leq N\}$ (used already by Gauss)

$$\mathrm{Loss}(h_{\mathbf{w}}) = \sum_{j=1}^{N} L_2(y_j, h_{\mathbf{w}}(x_j)) = \sum_{j=1}^{N} (y_j - h_{\mathbf{w}}(x_j))^2 = \sum_{j=1}^{N} (y_j - (\mathbf{w}_1 x_j + \mathbf{w}_0))^2$$

Task: find $\mathbf{w}^* := \underset{\mathbf{w}}{\mathrm{argmin}}\ \mathrm{Loss}(h_{\mathbf{w}})$.

▷ **Recall:** $\sum_{j=1}^{N} (y_j - (\mathbf{w}_1 x_j + \mathbf{w}_0))^2$ is minimized, when the partial derivatives wrt. the $\mathbf{w}_i$ are zero, i.e. when

$$\frac{\partial}{\partial \mathbf{w}_0}\left(\sum_{j=1}^{N} (y_j - (\mathbf{w}_1 x_j + \mathbf{w}_0))^2\right) = 0 \quad \text{and} \quad \frac{\partial}{\partial \mathbf{w}_1}\left(\sum_{j=1}^{N} (y_j - (\mathbf{w}_1 x_j + \mathbf{w}_0))^2\right) = 0$$

▷ **Observation:** These equations have a unique solution:

$$\mathbf{w}_1 = \frac{N(\sum_j x_j y_j) - (\sum_j x_j)(\sum_j y_j)}{N(\sum_j {x_j}^2) - (\sum_j x_j)^2} \qquad \mathbf{w}_0 = \frac{(\sum_j y_j) - \mathbf{w}_1(\sum_j x_j)}{N}$$

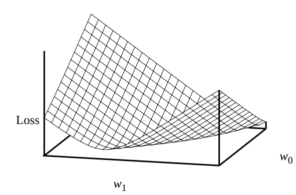
▷ **Remark:** Closed-form solutions only exist for linear regression, for other (differentiable) hypothesis spaces use gradient descent methods for adjusting/learning weights.

A Picture of the Weight Space

▷ **Remark:** Many forms of learning involve adjusting weights to minimize loss.

▷ **Definition 26.7.8.** The weight space is the space of all possible combinations of weights. Loss minimization in a weight space is called weight fitting.

▷ The weight space of univariate linear regression is $\mathbb{R}^2$.
$\rightsquigarrow$ graph the loss function over $\mathbb{R}^2$.
Note: it is convex.



▷ **Observation 26.7.9.** *The squared error loss function is convex for any linear regression problem $\rightsquigarrow$ there are no local minimumlocal minima.*

Gradient Descent Methods

▷ If we do not have closed form solutions for minimizing loss, we need to search.

▷ **Idea:** Use local search (hill climbing) methods.

▷ **Definition 26.7.10.** The gradient descent algorithm for finding a minimum of a continuous function f is hill climbing in the direction of the steepest descent, which can be computed by the partial derivatives of f.

function gradient−descent(f,$\mathbf{w}$,α) **returns** a **local** minimum of f
  inputs: a differentiable **function** f and initial weights $\mathbf{w} = (\mathbf{w}_0,\mathbf{w}_1)$.
  loop until $\mathbf{w}$ converges **do**
    for each $\mathbf{w}_i$ **do**
      $\mathbf{w}_i \longleftarrow \mathbf{w}_i - \alpha\frac{\partial}{\partial \mathbf{w}_i}(f(\mathbf{w}))$
    end for
  end loop

The parameter α is called the learning rate. It can be a fixed constant or it can decay as learning proceeds.

Gradient-Descent for Loss

- Let's try gradient descent for Loss.
- Work out the partial derivatives for one example (x,y):

$$\frac{\partial \mathrm{Loss}(\mathbf{w})}{\partial \mathbf{w}_i} = \frac{\partial (y - h_{\mathbf{w}}(x))^2}{\partial \mathbf{w}_i} = 2(y - h_{\mathbf{w}}(x))\frac{\partial (y - (\mathbf{w}_1 x + w))}{\partial \mathbf{w}_i}$$

and thus

$$\frac{\partial \mathrm{Loss}(\mathbf{w})}{\partial \mathbf{w}_0} = -2(y - h_{\mathbf{w}}(x)) \qquad \frac{\partial \mathrm{Loss}(\mathbf{w})}{\partial \mathbf{w}_1} = -2(y - h_{\mathbf{w}}(x))x$$

Plug this into the gradient descent updates:

$$\mathbf{w}_0 \longleftarrow \mathbf{w}_0 - \alpha - 2(y - h_{\mathbf{w}}(x)) \qquad \mathbf{w}_1 \longleftarrow \mathbf{w}_1 - \alpha - 2(y - h_{\mathbf{w}}(x))x$$

Gradient-Descent for Loss (continued)

- Analogously for n training examples (x_j,y_j):
- **Definition 26.7.11.**

$$\mathbf{w}_0 \longleftarrow \mathbf{w}_0 - \alpha(\sum_j -2(y_j - h_{\mathbf{w}}(x_j))) \quad \mathbf{w}_1 \longleftarrow \mathbf{w}_1 - \alpha(\sum_j -2(y_j - h_{\mathbf{w}}(x_n))x_n)$$

These updates constitute the batch gradient descent learning rule for univariate linear regression.

- Convergence to the unique global loss minimum is guaranteed (as long as we pick α small enough) but may be very slow.

Multivariate Linear Regression

- **Definition 26.7.12.** A multivariate or n ary function is a function with one or more arguments.
- We can use it for multivariate linear regression.
- **Idea:** Every example $\vec{x}_j$ is an n element vector and the hypothesis space is the set of functions

$$h_{sw}(\vec{x}_j) = \mathbf{w}_0 + \mathbf{w}_1 x_{j,1} + \ldots + \mathbf{w}_n x_{j,n} = \mathbf{w}_0 + \sum_i \mathbf{w}_i x_{j,i}$$

▷ **Trick:** Invent $x_{j,0}:=1$ and use matrix notation:

$$h_{sw}(\vec{x}_j) = \vec{w}\cdot\vec{x}_j = \vec{w}^t\vec{x}_j = \sum_i \mathbf{w}_i x_{j,i}$$

▷ **Definition 26.7.13.** The best vector of weights, $\mathbf{w}^*$, minimizes squared-error loss over the examples: $\mathbf{w}^*:=\underset{\mathbf{w}}{\text{argmin}}\,(\sum_j L_2(y_j)(\mathbf{w}\cdot\vec{x}_j))$.

▷ Gradient descent will reach the (unique) minimum of the loss function; the update equation for each weight $\mathbf{w}_i$ is

$$\mathbf{w}_i \longleftarrow \mathbf{w}_i - \alpha(\sum_j x_{j,i}(y_j - h_{\mathbf{w}}(\vec{x}_j)))$$

Multivariate Linear Regression (Analytic Solutions)

▷ We can also solve analytically for the $\mathbf{w}^*$ that minimizes loss.

▷ Let $\vec{y}$ be the vector of outputs for the training examples, and $\mathbf{X}$ be the data matrix, i.e., the matrix of inputs with one n-dimensional example per row.

Then the solution $\mathbf{w}^* = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{y}$ minimizes the squared error.

Multivariate Linear Regression (Regularization)

▷ **Remark:** Univariate linear regression does not overfit, but in the multivariate case there might be "redundant dimensions" that result in overfitting.

▷ **Idea:** Use regularization with a complexity function based on weights.

▷ **Definition 26.7.14.** Complexity$(h_{\mathbf{w}}) = L_q(\mathbf{w}) = \sum_i |\mathbf{w}_i|^q$

▷ **Caveat:** Do not confuse this with the loss functions L_1 and L_2.

▷ **Problem:** Which q should be pick? (L_1 and L_2 minimize sum of absolute values/squares)

▷ **Answer:** It depends on the application.

▷ **Remark:** L_1-regularization tends to produce a sparse model, i.e. it sets many weights to 0, effectively declaring the corresponding attributes to be irrelevant.

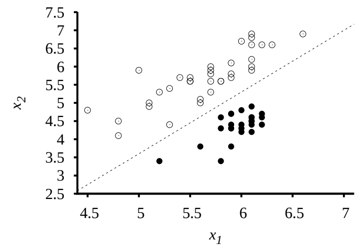
Hypotheses that discard attributes can be easier for a human to understand, and may be less likely to overfit. (see [RN03, Section 18.6.2])

Linear Classifiers with a hard Threshold

▷ **Idea:** The result of linear regression can be used for classification.

▷ **Example 26.7.15 (Nuclear Test Ban Verification).**

Plots of seismic data parameters: body wave magnitude x_1 vs. surface wave magnitude x_2. White: earthquakes, black: underground explosions
Also: $h_{\mathbf{w}^*}$ as a decision boundary $x_2 = 17x_1 - 4.9$.



▷ **Definition 26.7.16.** A decision boundary is a line (or a surface, in higher dimensions) that separates two classes of points. A linear decision boundary is called a linear separator and data that admits one are called linearly separable.

▷ **Example 26.7.17 (Nuclear Tests continued).** The linear separator for Example 26.7.15is defined by $-4.9 + 1.7x_1 - x_2 = 0$, explosions are characterized by $-4.9 + 1.7x_1 - x_2 > 0$, earthquakes by $-4.9 + 1.7x_1 - x_2 < 0$.

▷ **Useful Trick:** If we introduce dummy coordinate $x_0 = 1$, then we can write the classification hypothesis as $h_{\mathbf{w}}(\mathrm{x}) = 1$ if $\mathbf{w}\cdot\mathrm{x} > 0$ and 0 otherwise.

Linear Classifiers with a hard Threshold (Perceptron Rule)

▷ So $h_{\mathbf{w}}(\mathrm{x}) = 1$ if $\mathbf{w}\cdot\mathrm{x}>0$ and 0 otherwise is well-defined, how to choose $\mathbf{w}$?

▷ Think of $h_{\mathbf{w}}(\mathrm{x}) = \mathcal{T}(\mathbf{w}\cdot\mathrm{x})$, where $\mathcal{T}(z) = 1$, if $z>0$ and $\mathcal{T}(z) = 0$ otherwise. We call $\mathcal{T}$ a threshold function.

▷ **Problem:** $\mathcal{T}$ is not differentiable and $\frac{\partial\mathcal{T}}{\partial z} = 0$ where defined $\rightsquigarrow$

- ▷ No closed-form solutions by setting $\frac{\partial\mathcal{T}}{\partial z} = 0$ and solving.
- ▷ Gradient-descent methods in weight-space do not work either.

▷ We can learn weights by iterating over the following rule:

▷ **Definition 26.7.18.** Given an example (x,y), the perceptron learning rule is

$$\mathbf{w}_i \longleftarrow \mathbf{w}_i + \alpha\cdot(y - h_{\mathbf{w}}(\mathrm{x}))\cdot x_i$$

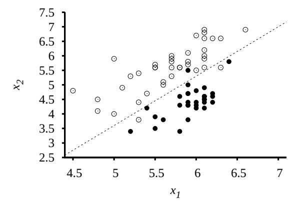
▷ as we are considering 0/1 classification, there are three possibilities:

1. If $y = h_{\mathbf{w}}(\mathrm{x})$, then $\mathbf{w}_i$ remains unchanged.
2. If $y = 1$ and $h_{\mathbf{w}}(\mathrm{x}) = 0$, then $\mathbf{w}_i$ is in/decreased if x_i is positive/negative. (we want to make $\mathbf{w}\cdot\mathrm{x}$ bigger so that $\mathcal{T}(\mathbf{w}\cdot\mathrm{x}) = 1$)

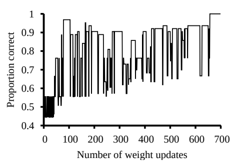
3. If $y = 0$ and $h_{\mathbf{w}}(\mathrm{x}) = 1$, then $\mathbf{w}_i$ is de/increased if x_i is positive/negative. (we want to make $\mathbf{w}\cdot\mathrm{x}$ smaller so that $\mathcal{T}(\mathbf{w}\cdot\mathrm{x}) = 0$)

Learning Curves for Linear Classifiers (Perceptron Rule)

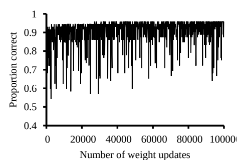
▷ **Example 26.7.19.** Learning curves (plots of total training set accuracy vs. number of iterations) for the perceptron rule on the earthquake/explosions data.



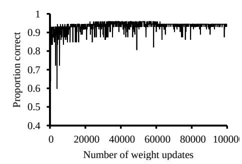
original data	noisy, non-separable data	learning rate decay $\alpha(t) = 1000/(1000+t)$



messy convergence
700 iterations



convergence failure
100,000 iterations



slow convergence
100,000 iterations

▷ **Theorem 26.7.20.** *Finding the minimal-error hypothesis is NP hard, but possible with learning rate decay.*

Linear Classification with Logistic Regression

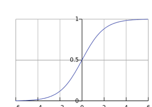
▷ **So far:** Passing the output of a linear function through a threshold function $\mathcal{T}$ yields a linear classifier.

▷ **Problem:** The hard nature of $\mathcal{T}$ brings problems:

- ▹ $\mathcal{T}$ is not differentiable nor continuous $\leadsto$ learning via perceptron rule becomes unpredictable.
- ▹ $\mathcal{T}$ is "overly precise" near the boundary $\leftrightsquigarrow$ need more graded judgements.

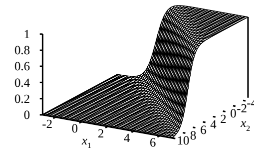
▷ **Idea:** Soften the threshold, approximate it with a differentiable function.

We use the standard logistic function $l(x) = \frac{1}{1+e^{-x}}$
So we have $h_{\mathbf{w}}(\mathrm{x}) = l(\mathbf{w}\cdot\mathrm{x}) = \frac{1}{1+e^{-(\mathbf{w}\cdot\mathrm{x})}}$



▷ **Example 26.7.21 (Logistic Regression Hypothesis in Weight Space).**

Plot of a logistic regression hypothesis for the earthquake/explosion data.
The value at $(\mathbf{w}_0,\mathbf{w}_1)$ is the probability of belonging to the class labeled 1.



We speak of the cliff in the classifier intuitively.

Logistic Regression

▷ **Definition 26.7.22.** The process of weight fitting in $h_{\mathbf{w}}(\mathrm{x}) = \frac{1}{1+e^{-(\mathbf{w}\cdot\mathrm{x})}}$ is called logistic regression.

▷ There is no easy closed form solution, but gradient descent is straightforward,

▷ As our hypotheses have continuous output, use the squared error loss function L_2.

▷ For an example (x,y) we compute the partial derivatives: (via chain rule)

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}_i}(L_2(\mathbf{w})) &= \frac{\partial}{\partial \mathbf{w}_i}(y - h_{\mathbf{w}}(\mathrm{x})^2) \\
&= 2 \cdot h_{\mathbf{w}}(\mathrm{x}) \cdot \frac{\partial}{\partial \mathbf{w}_i}(y - h_{\mathbf{w}}(\mathrm{x})) \\
&= -2 \cdot h_{\mathbf{w}}(\mathrm{x}) \cdot l'(\mathbf{w}\cdot\mathrm{x}) \cdot \frac{\partial}{\partial \mathbf{w}_i}(\mathbf{w}\cdot\mathrm{x}) \\
&= -2 \cdot h_{\mathbf{w}}(\mathrm{x}) \cdot l'(\mathbf{w}\cdot\mathrm{x}) \cdot x_i
\end{aligned}$$

Logistic Regression (continued)

▷ The derivative of the logistic function satisfies $l'(z) = l(z)(1 - l(z))$, thus

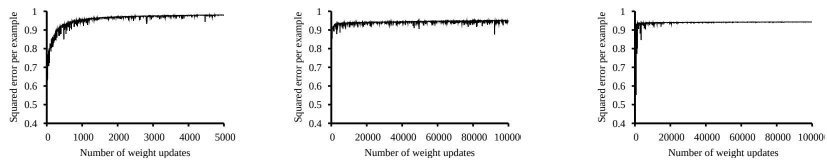
$$l'(\mathbf{w}\cdot\mathrm{x}) = l(\mathbf{w}\cdot\mathrm{x})(1 - l(\mathbf{w}\cdot\mathrm{x})) = h_{\mathbf{w}}(\mathrm{x})(1 - h_{\mathbf{w}}(\mathrm{x}))$$

▷ **Definition 26.7.23.** The rule for logistic update (weight update for minimizing the loss) is

$$\mathbf{w}_i \longleftarrow \mathbf{w}_i + \alpha \cdot (y - h_{\mathbf{w}}(\mathrm{x})) \cdot h_{\mathbf{w}}(\mathrm{x}) \cdot (1 - h_{\mathbf{w}}(\mathrm{x})) \cdot x_i$$

▷ **Example 26.7.24 (Redoing the Training Curves).**

original data	noisy, non-separable data	learning rate decay $\alpha(t) = 1000/(1000 + t)$



messy convergence 5000 iterations	convergence failure 100,000 iterations	slow convergence 100,000 iterations

▷ **Upshot:** Logistic update seems to perform better than perceptron update.

Michael Kohlhase: Artificial Intelligence 2 982 2023-09-20

26.8 Artificial Neural Networks

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30382`, `https://fau.tv/clip/id/30383`, `https://fau.tv/clip/id/30384`, and `https://fau.tv/clip/id/30386`.

Outline

▷ Brains

▷ Neural networks

▷ Perceptrons

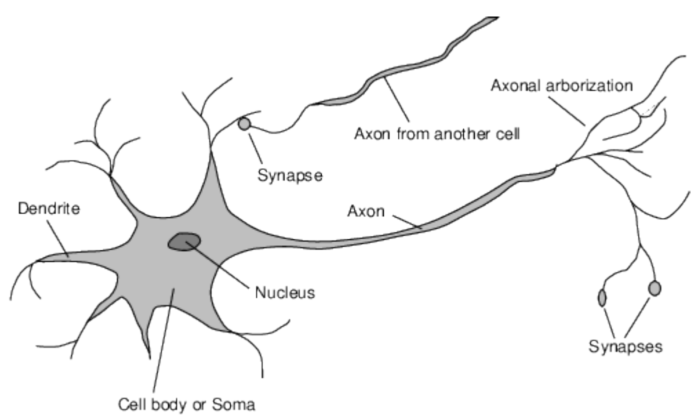
▷ Multilayer perceptrons

▷ Applications of neural networks

Michael Kohlhase: Artificial Intelligence 2 983 2023-09-20

Brains

▷ **Axiom 26.8.1 (Neuroscience Hypothesis).** *Mental activity consists consists primarily of electrochemical activity in networks of brain cells called neurons.*



▷ **Definition 26.8.2.** The animal brain is a biological neural network

- ▹ with 10^{11} neurons of > 20 types, 10^{14} synapses, $(1\text{ms}) – (10\text{ms})$ cycle time.
- ▹ Signals are noisy "spike trains" of electrical potential.

Neural Networks as an approach to Artificial Intelligence

▷ One approach to Artificial Intelligence is to model and simulate brains. (and hope that AI comes along naturally)

▷ **Definition 26.8.3.** The AI sub field of neural networks (also called connectionism, parallel distributed processing, and neural computation) studies computing systems inspired by the biological neural networks that constitute brains.

▷ Neural networks are attractive computational devices, since they perform important AI tasks most importantly learning and distributed, noise-tolerant computation – naturally and efficiently.

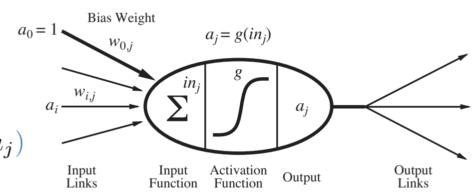
Neural Networks – McCulloch-Pitts "unit"

▷ **Definition 26.8.4.** An artificial neural network is a directed graph of units and links. A link from unit i to unit j propagates the activation a_i from unit i to unit j, it has a weight $w_{i,j}$ associated with it.

▷ In 1943 McCulloch and Pitts proposed a simple model for a neuron/brain.

▷ **Definition 26.8.5.** A McCulloch-Pitts unit first computes a weighted sum of all inputs and then applies an activation function g to it.

$$\begin{aligned} \mathrm{in}_i &= \sum_j \mathbf{w}_{j,i} a_j \\ a_i &\leftarrow g(\mathrm{in}_i) = g(\sum_j \mathbf{w}_{j,i} a_j) \end{aligned}$$

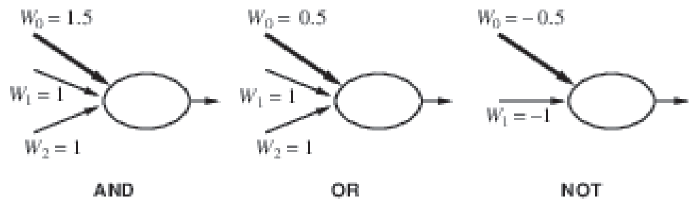


If g is a threshold function, we call the unit a perceptron unit, if g is a logistic function a sigmoid perceptron unit.

A McCulloch-Pitts network is a neural network with McCulloch-Pitts units.

Implementing Logical Functions as Units

▷ McCulloch-Pitts units are a gross oversimplification of real neurons, but its purpose is to develop understanding of what neural networks of simple units can do.

- **Theorem 26.8.6 (McCulloch and Pitts).** *Every Boolean function can be implemented as McCulloch-Pitts networks.*
- *Proof:* by construction
  1. Recall that $a_i \longleftarrow g(\sum_j \mathbf{w}_{j,i} a_j)$.
  2. As for linear regression we use $a_0 = 1 \rightsquigarrow \mathbf{w}_{0,i}$ as a bias weight (or intercept) (determines the threshold)
  3. 

  4. Any Boolean function can be implemented as a DAG of McCulloch-Pitts units.

Network Structures: Feed-Forward Networks

- We have models for neurons $\rightsquigarrow$ connect them to neural networks.
- **Definition 26.8.7.** A neural network is called a feed-forward network, if it is acyclic.
- **Intuition:** Feed-forward networks implement functions, they have no internal state.
- **Definition 26.8.8.** Feed-forward networks are usually organized in layers: a n layer network has a partition $\{L_0, \ldots, L_n\}$ of the nodes, such that edges only connect nodes from subsequent layer.

  L_0 is called the input layer and its members input units, and L_n the output layer and its members output units. Any unit that is not in the input layer or the output layer is called hidden.

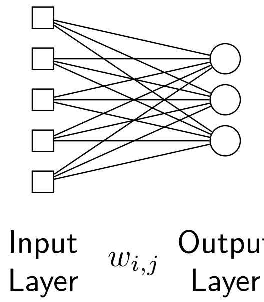
Network Structures: Recurrent Networks

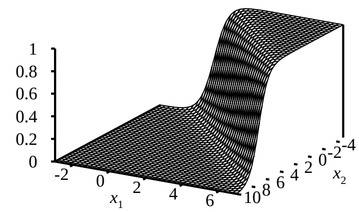
- **Definition 26.8.9.** A neural network is called recurrent (RNNs, iff it has cycles.
  - Hopfield networks have symmetric weights ($\mathbf{w}_{i,j} = \mathbf{w}_{j,i}$) $g(x) = \text{sign}(x)$, $a_i = \pm 1$; (holographic associative memory)
  - Boltzmann machines use stochastic activation functions.
- Recurrent neural networks have cycles with delays $\rightsquigarrow$ have internal state (like flip-flops), can oscillate etc.

Single-layer Perceptrons

▷ **Definition 26.8.10.** A perceptron network is a feed-forward network of perceptron units. A single layer perceptron network is called a perceptron.

▷ **Example 26.8.11.**





▷ All input units are directly connected to output units.

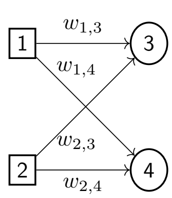
▷ Output units all operate separately, no shared weights $\rightsquigarrow$ treat as the combination of n perceptron units.

▷ Adjusting weights moves the location, orientation, and steepness of cliff.

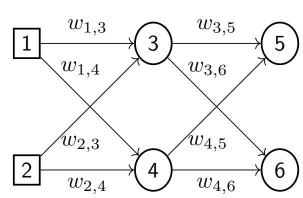
Feed-forward Neural Networks (Example)

▷ Feed-forward network $\widehat{=}$ a parameterized family of nonlinear functions:

▷ **Example 26.8.12.** We show two feed-forward networks:



a) single layer (perceptron network)



b) 2 layer feed-forward network

$$
\begin{aligned}
a_5 &= g(\mathbf{w}_{3,5} \cdot a_3 + \mathbf{w}_{4,5} \cdot a_4) \\
&= g(\mathbf{w}_{3,5} \cdot g(\mathbf{w}_{1,3} \cdot a_1 + \mathbf{w}_{2,3} a_2) + \mathbf{w}_{4,5} \cdot g(\mathbf{w}_{1,4} \cdot a_1 + \mathbf{w}_{2,4} a_2))
\end{aligned}
$$

▷ **Idea:** Adjusting weights changes the function: do learning this way!

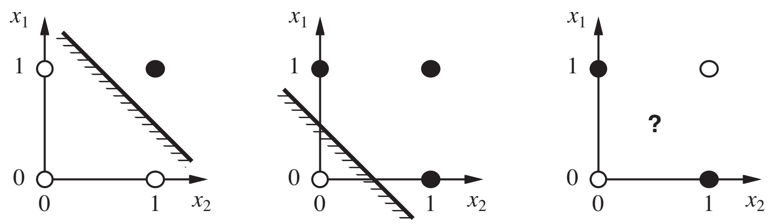
Expressiveness of Perceptrons

▷ Consider a perceptron with $g =$ step function (Rosenblatt, 1957, 1960)

▷ Can represent AND, OR, NOT, majority, etc., but not XOR (and thus no adders)

▷ Represents a linear separator in input space:

$$\sum_j \mathbf{w}_j x_j > 0 \quad \text{or} \quad \mathbf{W}, \mathbf{x}\cdot > 0$$



(a) x_1 **and** x_2 (b) x_1 **or** x_2 (c) x_1 **xor** x_2

▷ Minsky & Papert (1969) pricked the first neural network balloon!

Perceptron Learning

▷ **Idea:** Wlog. treat only single-output perceptrons $\leadsto$ $\mathbf{w}$ is a "weight vector".
Learn by adjusting weights in $\mathbf{w}$ to reduce generalization loss on training set.

▷ Let us compute with the squared error loss of a weight vector $\mathbf{w}$ for an example $(\mathbf{x},y)$.

$$Loss(\mathbf{w}) = Err^2 = (y - h_{\mathbf{w}}(\mathbf{x}))^2$$

▷ Perform optimization search by gradient descent for any weight $\mathbf{w}_i$:

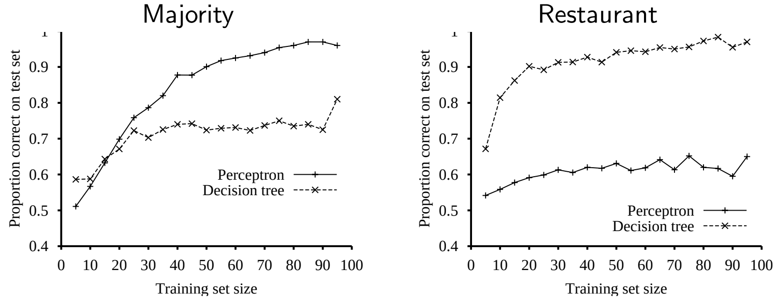
$$\begin{aligned}\frac{\partial Loss(\mathbf{w})}{\partial \mathbf{w}_j} &= 2 \cdot Err \cdot \frac{\partial Err}{\partial \mathbf{w}_j} = 2 \cdot Err \cdot \frac{\partial}{\partial \mathbf{w}_j}(y - g(\sum_{j=0}^{n} \mathbf{w}_j x_j)) \\ &= -2 \cdot Err \cdot g'(\mathrm{in}_j) \cdot x_j\end{aligned}$$

▷ Simple weight update rule:

$$\mathbf{w}_{j,k} \leftarrow \mathbf{w}_{j,k} + \alpha \cdot Err \cdot g'(\mathrm{in}_j) \cdot x_j$$

Perceptron learning contd.

▷ Perceptron learning rule converges to a consistent function
for any linearly separable data set

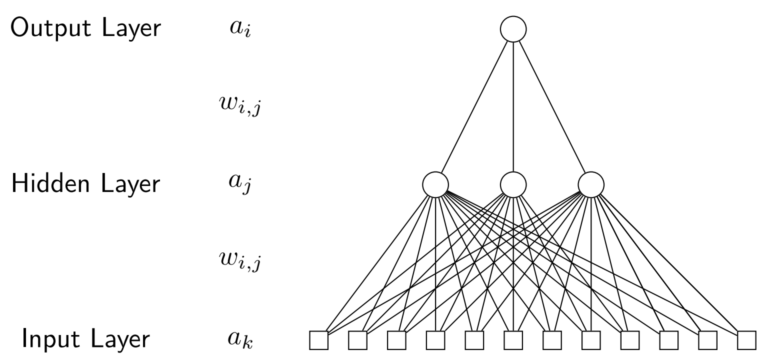


▷ Perceptron learns majority function easily, DTL is hopeless.

▷ DTL learns restaurant function easily, perceptron cannot represent it.

Multilayer perceptrons

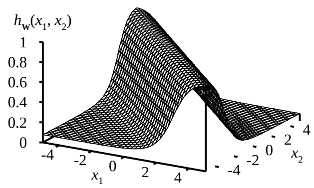
▷ **Definition 26.8.13.** In multi layer perceptron (MLPs), layers are usually fully connected;
numbers of hidden units typically chosen by hand.

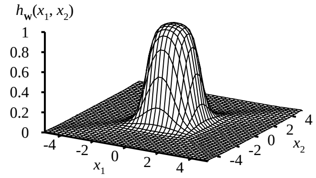


▷ **Definition 26.8.14.** Some MLPs have residual connections, i.e. connections that skip layers.

Expressiveness of MLPs

▷ All continuous functions w/ 2 layers, all functions w/ 3 layers.





- ▷ Combine two opposite-facing threshold functions to make a ridge.
- ▷ Combine two perpendicular ridges to make a bump.
- ▷ Add bumps of various sizes and locations to fit any surface.
- ▷ Proof requires exponentially many hidden units. (cf. DTL proof)

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 2 996 2023-09-20

Learning in Multilayer Networks (Output Layer)

- ▷ **Idea:** Learn by adjusting weights to reduce error on training set.
- ▷ **Problem:** Neural networks have multiple outputs.
- ▷ **Idea:** We use $\mathbf{h_w}$ with output vector y.
- ▷ **Observation:** The squared error loss of a weight matrix $\mathbf{w}$ for an example (x,y) is

$$Loss(\mathbf{w}) = \|(\mathrm{y} - \mathbf{h_w}(\mathrm{x}))\|_2{}^2 = \sum_{k=1}^{n} (y_k - a_k)^2$$

- ▷ **Output layer:** Analogous to that for single-layer perceptron, but multiple output units

$$\mathbf{w}_{j,i} \leftarrow \mathbf{w}_{j,i} + \alpha \cdot a_j \cdot \Delta_i$$

where $\Delta_i = Err_i \cdot g'(\mathrm{in}_i)$ and $Err = \mathrm{y} - \mathbf{h_w}(\mathrm{x})$. (error vector)

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 2 997 2023-09-20

Learning in Multilayer Networks (Hidden Layers)

- ▷ **Problem:** The error Err is well-defined only for the output layer. ⇝ The examples do not say anything about the hidden layers.
- ▷ **Idea:** Back-propagate the error from the output layer; actually back-propagate Δ_k.
The hidden node j is "responsible" for some fraction of Δ_k.(by connection weight)
- ▷ **Definition 26.8.15.** The back-propagation rule for hidden nodes of a multilayer

perceptron is

$$\Delta_j \leftarrow g'(\mathrm{in}_j) \cdot (\sum_i \mathbf{w}_{j,i}\Delta_i)$$

▷ Update rule for weights in hidden layer:

$$\mathbf{w}_{k,j} \leftarrow \mathbf{w}_{k,j} + \alpha \cdot a_k \cdot \Delta_j$$

▷ **Remark:** Most neuroscientists deny that back-propagation occurs in the brain.

Back-Propagation Process

▷ The back-propagation process can be summarized as follows:

1. Compute the Δ values for the output units, using the observed error.
2. Starting with output layer, repeat the following for each layer in the network, until the earliest hidden layer is reached:
   (a) Propagate the Δ values back to the previous (hidden) layer.
   (b) Update the weights between the two layers.

▷ Details (algorithm) later.

Backprogagation Learning Algorithm

▷ **Definition 26.8.16.** The back-propagation learning algorithm is given the following pseudocode

```
function BACK−PROP−LEARNING(examples,network) returns a neural network
  inputs: examples, a set of examples, each with input vector x and output vector y
          network, a multilayer network with L layers, weights w_{i,j}, activation function g
  local variables: Δ, a vector of errors, indexed by network node
  foreach weight w_{i,j} in network do
    w_{i,j} := a small random number
  repeat
    foreach example (x,y) in examples do
      /∗ Propagate the inputs forward to compute the outputs ∗/
      foreach node i in the input layer do a_i := x_i
        for l = 2 to L do
          foreach node j in layer l do
            in_j := ∑_i w_{i,j} a_i
            a_j := g(in_j)
      /∗ Propagate deltas backward from output layer to input layer ∗/
      foreach node j in the output layer do Δ[j] := actfun!'(in_j) · (y_j − a_j)
      for l = L − 1 to 1 do
        foreach node i in layer l do Δ[i] := actfun!'(in_i) · (∑_j w_{i,j} Δ[j])
      /∗ Update every weight in network using deltas ∗/
      foreach weight w_{i,j} in network do w_{i,j} := w_{i,j} + α · a_i · Δ[j]
  until some stopping criterion is satisfied
  return network
```

Back-Propagation Derivation from First Principles

- ▷ This is very similar to the gradient calculation for logistic regression
- ▷ Compute the loss gradient wrt. the weights between the output and hidden layers:

$$\begin{aligned}
\frac{\partial Loss_k}{\partial \mathrm{w}_{j,k}} &= -2(y_k - a_k)\frac{\partial a_k}{\partial \mathrm{w}_{j,k}} = -2(y_k - a_k)\frac{\partial g(\mathrm{in}_k)}{\partial \mathrm{w}_{j,k}} \\
&= -2(y_k - a_k)g'(\mathrm{in}_k)\frac{\partial \mathrm{in}_k}{\partial \mathrm{w}_{j,k}} \\
&= -2(y_k - a_k)g'(\mathrm{in}_k)\frac{\partial}{\partial \mathrm{w}_{j,k}}(\sum_j \mathrm{w}_{j,k} a_j) \\
&= -2(y_k - a_k)g'(\mathrm{in}_k)a_j = -2a_j\Delta_k
\end{aligned}$$

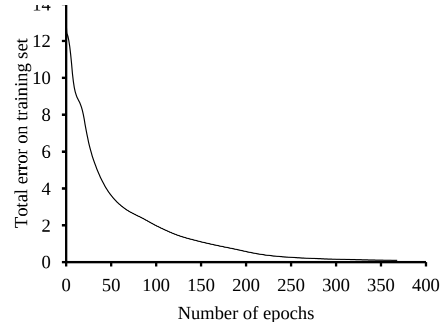
Back-propagation derivation contd.

- ▷ We continue computing where we left off

$$\begin{aligned}
\frac{\partial Loss_k}{\partial \mathrm{w}_{i,j}} &= -2(y_k - a_k)\frac{\partial a_k}{\partial \mathrm{w}_{i,j}} = -2(y_k - a_k)\frac{\partial g(\mathrm{in}_k)}{\partial \mathrm{w}_{i,j}} \\
&= -2(y_k - a_k)g'(\mathrm{in}_k)\frac{\partial \mathrm{in}_k}{\partial \mathrm{w}_{i,j}} = -2\Delta_k\frac{\partial}{\partial \mathrm{w}_{i,j}}(\sum_j \mathrm{w}_{j,i} a_j) \\
&= -2\Delta_k \mathrm{w}_{j,k}\frac{\partial a_j}{\partial \mathrm{w}_{i,j}} = -2\Delta_k \mathrm{w}_{j,k}\frac{\partial g(\mathrm{in}_j)}{\partial \mathrm{w}_{i,j}} \\
&= -2\Delta_k \mathrm{w}_{j,k} g'(\mathrm{in}_j)\frac{\partial \mathrm{in}_j}{\partial \mathrm{w}_{i,j}} \\
&= -2\Delta_k \mathrm{w}_{j,k} g'(\mathrm{in}_j)\frac{\partial}{\partial \mathrm{w}_{i,j}}(\sum_k \mathrm{w}_{i,j} a_k) \\
&= -2\Delta_k \mathrm{w}_{j,k} g'(\mathrm{in}_j)a_k = -a_k\Delta_j
\end{aligned}$$

Back-Propagation – Properties

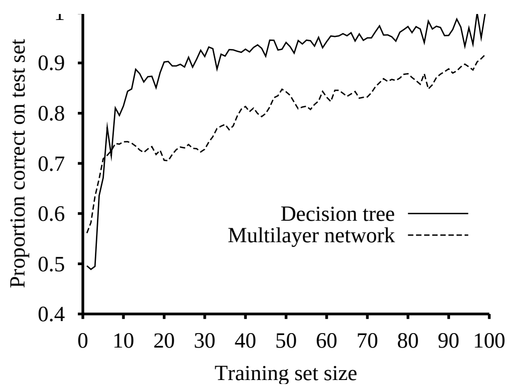
- ▷ At each epoch, sum gradient updates for all examples and apply.
- ▷ Training curve for 100 restaurant examples: finds exact fit.



▷ **Typical problems:** slow convergence, local minima.

Back-Propagation – Properties (contd.)

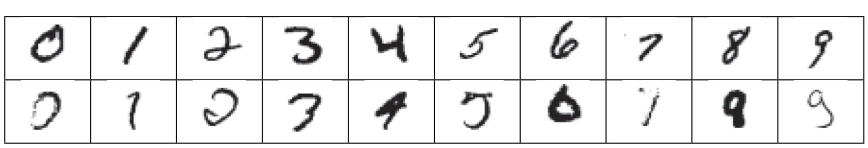
▷ **Example 26.8.17.** Learning curve for MLPs with 4 hidden units:



▷ **Experience shows:** MLPs are quite good for complex pattern recognition tasks, but resulting hypotheses cannot be understood easily.

▷ This makes MLPs ineligible for some tasks, such as credit card and loan approvals, where law requires clear unbiased criteria.

Handwritten digit recognition



▷ 400–300–10 unit MLP = 1.6% error

▷ LeNet: 768–192–30–10 unit MLP = 0.9% error

▷ Current best (kernel machines, vision algorithms) ≈ 0.6% error

Michael Kohlhase: Artificial Intelligence 2 1005 2023-09-20

Summary

▷ Most brains have lots of neurons; each neuron ≈ linear-threshold unit (?)

▷ Perceptrons (one-layer networks) insufficiently expressive

▷ Multi-layer networks are sufficiently expressive; can be trained by gradient descent, i.e., error back-propagation

▷ Many applications: speech, driving, handwriting, fraud detection, etc.

▷ Engineering, cognitive modelling, and neural system modelling subfields have largely diverged

Michael Kohlhase: Artificial Intelligence 2 1006 2023-09-20

XKCD on Machine Learning

▷ **A Skepticists View:** see `https://xkcd.com/1838/`

Michael Kohlhase: Artificial Intelligence 2 1007 2023-09-20

26.9 Support Vector Machines

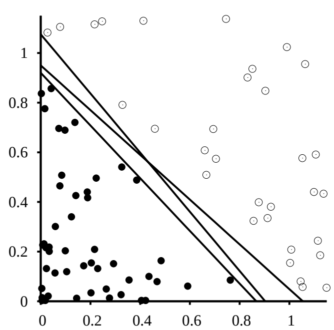
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30386`.

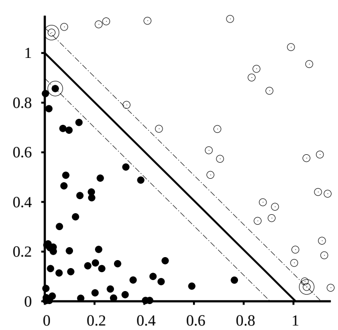
Support Vector Machines

- **Definition 26.9.1.** Support-vector machines (SVMs also support-vector networks) are supervised learning models for classification and regression. SVMs
  - construct a maximum margin separator, i.e. a decision boundary with the largest possible distance to example points. This helps them generalize well.
  - can embed data into a higher-dimensional space, where it is linearly separable by the *kernel trick* $\leadsto$ the separating hyperplane is a hyper-surface in original data.
  - prioritize critical examples (support vector). (better generalization)
- Currently the most popular approach for "off-the-shelf" supervised learning.

Support Vector Machines (Separation with Margin)

- **Definition 26.9.2.** Given a linearly separable data set E the maximum margin separator is the linear separator s that maximizes the margin, i.e. the distance of the E from s.
- **Example 26.9.3.** All lines on the left are valid linear separators:





We expect the maximum margin separator on the right to generalize better!

- **Idea:** Minimize the expected generalization loss instead of the empirical loss.

Finding the Maximum Margin Separator

- Before we see how to find the maximum margin separator, ...
- **We have** a training $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where

- ▷ $y_i \in \{-1, 1\}$ (instead of $\{1, 0\}$)
- ▷ $\mathbf{x}_i \in \mathbb{R}^p$ (multi-linear classification)

▷ **We want** to find a hyperplane that maximally separates the points $\mathbf{x}_i$ with $y_i = -1$ from those with $y_i = 1$.

▷ **Recall:** Any hyperplane (in particular any linear separator) is represented as the set $\{\mathbf{x} | (\mathbf{w} \cdot \mathbf{x}) + b = 0\}$, where $\mathbf{w}$ is the (not necessarily normalized) normal vector the hyperplane.

The parameter $\frac{b}{\|\mathbf{w}\|_2}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

▷ **Idea:** Use gradient descent to search the space of all $\mathbf{w}$ and b for maximizing combinations. (works, but SVMs follow a different route)

Finding the Maximum Margin Separator (Separable Case)

▷ **Idea:** The margin is bounded by the two hyperplanes described by $(\mathbf{w} \cdot \mathbf{x}) + b = -1$ (lower boundary) and $(\mathbf{w} \cdot \mathbf{x}) + b = 1$ (upper boundary).

The distance between them is $\frac{2}{\|\mathbf{w}\|_2}$.

$\rightsquigarrow$ to maximize the margin, minimize $\|\mathbf{w}\|_2$ while keeping $\mathbf{x}_i$ out of the margin.

▷ **Constraints:** $(\mathbf{w} \cdot \mathbf{x}_i) + b \geq 1$ for $y_i = 1$ and $(\mathbf{w} \cdot \mathbf{x}_i) + b \leq -1$ for $y_i = -1$ or simply $y_i((\mathbf{w} \cdot \mathbf{x}_i) - b) \geq 1$ for $1 \leq i \leq n$.

▷ **Optimization Problem:** Minimize $\|\mathbf{w}\|_2$ while $y_i((\mathbf{w} \cdot \mathbf{x}_i) - b) \geq 1$ for $1 \leq i \leq n$.

Finding the Maximum Margin Separator (Separable Case)

▷ After a bit of mathematical magic (solving for the Lagrangian dual) we get

▷ **Alternative Representation:** Find the optimal solution by solving the SVM equation

$$\underset{\alpha}{\text{argmax}} \left(\sum_j \alpha_j - \frac{1}{2}\left(\sum_{j,k} \alpha_j \alpha_k y_j y_k (\mathbf{x}_j \cdot \mathbf{x}_k)\right)\right)$$

under the constraints $\alpha_j \geq 0$ and $\sum_j \alpha_j y_j = 0$.

▷ **Observations:** This equation has three important properties:

1. The expression is convex $\rightsquigarrow$ the single global maximum can found efficiently.
2. Data enter the expression only in the form of dot products of point pairs$\rightsquigarrow$ once the optimal α_i have been calculated, we have

$$h(\mathbf{x}) = \text{sign}\left(\sum_j \alpha_j y_j (\mathbf{x} \cdot \mathbf{x}_j) - b\right)$$

3. The weights α_j associated with each data point are zero except at the support vectors the points closest to the separator.

▷ There are good software packages for solving such quadratic programming optimizations

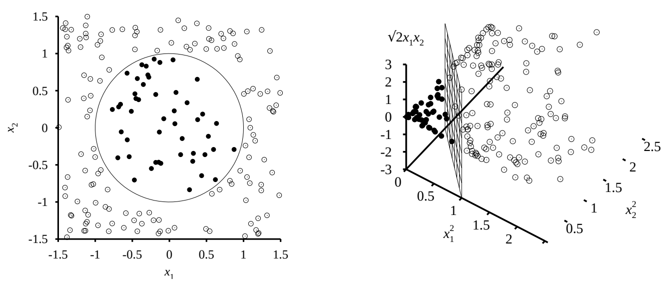
▷ Once we found an optimal vector α, use $\mathbf{w} = \sum_j \alpha_j x_j$.

Support Vector Machines (Kernel Trick)

▷ **Problem:** What if the data is not linearly separable?

▷ **Idea:** Transform the data into a higher-dimensional space, where they are.

▷ **Example 26.9.4 (Projecting Up a Non-Separable Data Set). left**: The true decision boundary is ${x_1}^2 + {x_2}^2 \leq 1$.



right: mapping into a three-dimensional input space $\langle {x_1}^2, {x_2}^2, \sqrt{2}x_1x_2\rangle \rightsquigarrow$ separable by a hyperplane.

▷ **Upshot:** We map each input vector x to a $F(\mathrm{x})$ with $f_1 = {x_1}^2$, $f_2 = {x_2}^2$, and $f_3 = \sqrt{2}x_1x_2$.

Support Vector Machines (Kernel Trick continued)

▷ **Idea:** Replace $\mathrm{x}_j \cdot \mathrm{x}_j$ by $F(\mathrm{x}_j) \cdot F(\mathrm{x}_j)$ in the SVM equation.(compute in high dim space.)

▷ Often we can compute $F(\mathrm{x}_j) \cdot F(\mathrm{x}_j)$ without computing F everywhere.

▷ **Example 26.9.5.** If $F(\mathrm{x}) = \langle {x_1}^2, {x_2}^2, \sqrt{2}x_1x_2\rangle$, then $F(\mathrm{x}_j) \cdot F(\mathrm{x}_j) = (\mathrm{x}_j \cdot \mathrm{x}_j)^2$ (have added the $\sqrt{2}$ in F so that this works)

▷ We call the function $(\mathrm{x}_j \cdot \mathrm{x}_j)^2$ a kernel function. (there are others; next)

▷ **Definition 26.9.6.** Let X be a nonempty set, sometimes referred to as the index set. A symmetric function $K : X \times X \rightarrow \mathbb{R}$ is called a (positive definite) kernel function on X, iff $\sum_{i,j=1}^{n} c_i c_j K(\mathrm{x}_i, \mathrm{x}_j) \geq 0$ for any $x_i \in X$, $n \in \mathbb{N}$, and $c_i \in \mathbb{R}$.

Support Vector Machines (Kernel Trick continued)

▷ **Generally:** We can learn non-linear separators by solving

$$\operatorname*{argmax}_{\alpha} \left(\sum_j \alpha_j - \frac{1}{2}\left(\sum_{j,k} \alpha_j \alpha_k y_j y_k K(\mathrm{x}_j, \mathrm{x}_k)\right)\right)$$

where K is a kernel function

▷ **Definition 26.9.7.** The function $K(\mathrm{x}_j, \mathrm{x}_k) = (1 + (\mathrm{x}_j \cdot \mathrm{x}_j))^d$ is a kernel function corresponding to a feature space whose dimension is exponential in d. It is called the polynomial kernel.

▷ **Theorem 26.9.8 (Mercer's Theorem).** *Every kernel function K where $K(\mathrm{x}_j, \mathrm{x}_k)$ is positive definite corresponds to some feature space.*

Summary of Inductive Learning

▷ Learning needed for unknown environments, lazy designers.

▷ Learning agent = performance element + learning element.

▷ Learning method depends on type of performance element, available feedback, type of component to be improved, and its representation.

▷ For supervised learning, the aim is to find a simple hypothesis that is approximately consistent with training examples

▷ Decision tree learning using information gain.

▷ Learning performance = prediction accuracy measured on test set

▷ PAC learning as a general theory of learning boundaries.

▷ Linear regression (hypothesis space of univariate linear functions).

▷ Linear classification by linear regression with hard and soft thresholds.

Chapter 27

Statistical Learning

Part V we learned how to reason in non-deterministic, partially observable environments by quantifying uncertainty and reasoning with it. The key resource there were probabilistic models and their efficient representations: Bayesian networks.

Part V we assumed that these models were given, perhaps designed by the agent developer. We will now learn how these models can – at least partially – be learned from observing the environment.

Statistical Learning: Outline

- ▷ Bayesian learning, i.e. learning probabilistic models (e.g. Bayesian networks) from observations.
- ▷ Maximum *a posteriori* and maximum likelihood learning
- ▷ Bayes network learning
  - ▷ ML Parameter Learning with Complete Data
  - ▷ Linear regression
  - ▷ Naive Bayes Models/Learning

27.1 Full Bayesian Learning

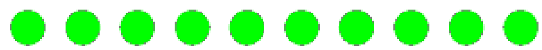
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30388`.

The Candy Flavors Example

▷ **Example 27.1.1.** Suppose there are five kinds of bags of candies:

1. 10% are h_1: 100% cherry candies
2. 20% are h_2: 75% cherry candies + 25% lime candies
3. 40% are h_3: 50% cherry candies + 50% lime candies
4. 20% are h_4: 25% cherry candies + 75% lime candies
5. 10% are h_5: 100% lime candies

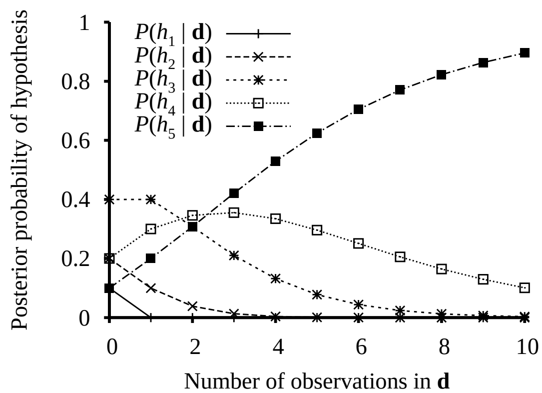
Then we observe candies drawn from some bag:



What kind of bag is it? What flavour will the next candy be?

Candy Flavors: Posterior probability of hypotheses

▷ **Example 27.1.2.** The probability of hypothesis h_i after n limes are observed $\widehat{=}$



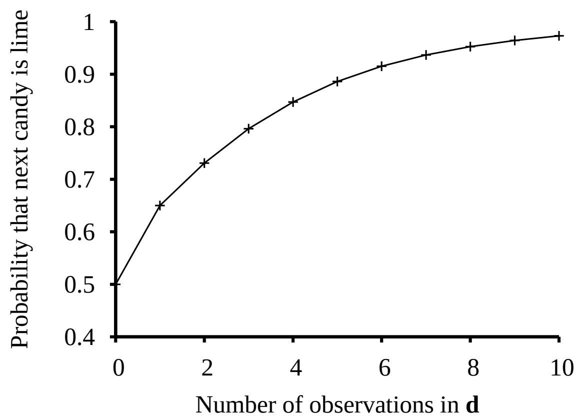
if the observation are IID, i.e. $P(\mathbf{d}|h_i) = \prod_j P(d_j|h_i)$ and the hypothesis prior is as advertised. (e.g. $P(\mathbf{d}|h_3) = 0.5^{10} = 0.1\%$)

The posterior probabilities start with the hypothesis priors, change with data.

Candy Flavors: Prediction Probability

▷ We calculate that the $n+1$-th candy is lime:

$$\mathbf{P}(d_{n+1} = \text{lime}|\mathbf{d}) = \sum_i \mathbf{P}(d_{n+1} = \text{lime}|h_i) \cdot P(h_i|\mathbf{d})$$



Full Bayesian Learning

- **Idea:** View learning as Bayesian updating of a probability distribution over the hypothesis space:
  - H is the hypothesis variable with values $h_1, h_2, \ldots$ and prior $\mathbf{P}(H)$.
  - jth observation d_j gives the outcome of random variable D_j.
  - $\mathbf{d}{:=}d_1, \ldots, d_N$ constitutes the training set of a inductive learning problem.
- **Definition 27.1.3.** Bayesian learning calculates the probability of each hypothesis and makes predictions based on this:
  - Given the data so far, each hypothesis has a posterior probability:

    $$P(h_i|\mathbf{d}) = \alpha \cdot P(\mathbf{d}|h_i) \cdot P(h_i)$$

    where $P(\mathbf{d}|h_i)$ is called the likelihood (of the data under each hypothesis) and $P(h_i)$ the hypothesis prior.
  - Bayesian predictions use a likelihood-weighted average over the hypotheses:

    $$\mathbf{P}(\mathsf{X}|\mathbf{d}) = \sum_i \mathbf{P}(\mathsf{X}|\mathbf{d}, h_i) \cdot P(h_i|\mathbf{d}) = \sum_i \mathbf{P}(\mathsf{X}|h_i) \cdot P(h_i|\mathbf{d})$$
- **Observation:** No need to pick one best-guess hypothesis for Bayesian predictions! (and that is all an agent cares about)

Full Bayesian Learning: Properties

- **Observation:** The Bayesian prediction eventually agrees with the true hypothesis.
  - The probability of generating "uncharacteristic" data indefinitely is vanishingly small.

- *Proof sketch:* Argument analogous to PAC learning.

- **Problem:** Summing over the hypothesis space is often intractable.

- **Example 27.1.4.** There are $2^{2^6} = 18,446,744,073,709,551,616$ Boolean functions of 6 arguments.

- **Solution:** Approximate the learning methods to simplify.

27.2 Approximations of Bayesian Learning

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30389`.

Maximum A Posteriori (MAP) Approximation

- **Goal:** Get rid of summation over the space of all hypotheses in predictions.

- **Idea:** Make predictions wrt. the "most probable hypothesis"!

- **Definition 27.2.1.** For maximum a posteriori learning (MAP learning) choose the MAP hypothesis h_{MAP} that maximizes $P(h_i|\mathbf{d})$.

  I.e., maximize $P(\mathbf{d}|h_i) \cdot P(h_i)$ or (even better) $\log_2(P(\mathbf{d}|h_i)) + \log_2(P(h_i))$.

- Predictions made according to a MAP hypothesis h_{MAP} are approximately Bayesian to the extent that $\mathbf{P}(X|\mathbf{d}) \approx \mathbf{P}(X|h_{\mathsf{MAP}})$.

- **Example 27.2.2.** In our candy example, $h_{\mathsf{MAP}} = h_5$ after three limes in a row
  - a MAP learner then predicts that candy 4 is lime with probability 1.
  - compare with Bayesian prediction of 0.8. (see prediction curves above)

- As more data arrive, the MAP and Bayesian predictions become closer, because the competitors to the MAP hypothesis become less and less probable.

- For deterministic hypotheses, $P(\mathbf{d}|h_i)$ is 1 if consistent, 0 otherwise
  $\rightsquigarrow$ MAP = simplest consistent hypothesis. (cf. science)

- **Remark:** Finding MAP hypotheses is often much easier than Bayesian learning, because it requires solving an optimization problem instead of a large summation (or integration) problem.

Digression From MAP-learning to MDL-learning

- **Idea:** Reinterpret the log terms $\log_2(P(\mathbf{d}|h_i)) + \log_2(P(h_i))$ in MAP learning:
  - Maximizing $P(\mathbf{d}|h_i) \cdot P(h_i)$ $\widehat{=}$ minimizing $-\log_2(P(\mathbf{d}|h_i)) - \log_2(P(h_i))$.
  - $-\log_2(P(\mathbf{d}|h_i))$ $\widehat{=}$ number of bits to encode data given hypothesis.
  - $-\log_2(P(h_i))$ $\widehat{=}$ additional bits to encode hypothesis. (section 26.4)

- ▷ **Indeed** if hypothesis predicts the data exactly – e.g. h_5 in candy example – then $\log_2(1) = 0 \rightsquigarrow$ preferred hypothesis.
- ▷ This is more directly modeled by the following approximation to Bayesian learning:
- ▷ **Definition 27.2.3.** In minimum description length learning (MDL learning) the MDL hypothesis h_{MDL} minimizes the information entropy of the hypothesis likelihood.

Maximum Likelihood (ML) approximation

- ▷ **Observation:** For large data sets, the prior becomes irrelevant. (we might not trust it anyways)
- ▷ **Idea:** Use this to simplify learning.
- ▷ **Definition 27.2.4.** Maximum likelihood learning (ML learning): choose the ML hypothesis h_{ML} maximizing $P(\mathbf{d}|h_i)$. (simply get the best fit to the data)
- ▷ **Remark:** ML learning $\widehat{=}$ MAP learning for a uniform prior. (reasonable if all hypotheses are of the same complexity)
- ▷ ML learning is the "standard" (non Bayesian) statistical learning method.

27.3 Parameter Learning for Bayesian Networks

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30390`.

ML Parameter Learning in Bayesian Nets

- ▷ **Example 27.3.1.** Bag from a new manufacturer; fraction θ of cherry candies?

$P(F = cherry)$
θ

Flavor

- ▷ **New Facet:** Any θ is possible: continuum of hypotheses h_θ
  θ is a parameter for this simple (binomial) family of models.
- ▷ Suppose we unwrap N candies, c cherries and $\ell = N - c$ limes.
- ▷ **Lemma 27.3.2.** *These are IID observations, so the likelihood is*

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^{N} P(d_j|h_\theta) = \theta^c \cdot (1-\theta)^\ell$$

▷ **Trick:** When optimizing a product, optimize the logarithm instead! ($\log_2(!)$ is monotone and turns products into sums)

▷ **Definition 27.3.3.** The log likelihood is just the binary logarithm of the likelihood.

$$L(\mathbf{d}|h) := \log_2(P(\mathbf{d}|h))$$

ML Parameter Learning in Bayes Nets

▷ Compute the log likelihood as (using Lemma 27.3.2)

$$\begin{aligned} L(\mathbf{d}|h_\theta) &= \log_2(P(\mathbf{d}|h_\theta)) \\ &= \sum_{j=1}^{N} \log_2(P(\mathsf{d}_j|h_\theta)) \\ &= c\log_2(\theta) + \ell\log_2(1-\theta) \end{aligned}$$

▷ Maximize this w.r.t. θ

$$\frac{\partial}{\partial\theta}(L(\mathbf{d}|h_\theta)) = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0$$

$\leadsto \theta = \frac{c}{c+\ell} = \frac{c}{N}$

▷ **In English:** h_θ asserts that the actual proportion of cherries in the bag is equal to the observed proportion in the candies unwrapped so far!

▷ Seems sensible, but causes problems with 0 counts!

▷ **Question:** Haven't we done a lot of work to obtain the obvious?

▷ **Answer:** So far yes, but this is a general method of broad applicability!

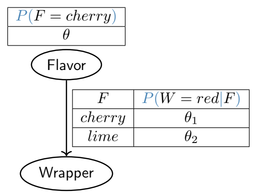
ML Learning for Multiple Parameters in Bayesian Networks

▷ **Cooking Recipe:**

1. Write down an expression for the likelihood of the data as a function of the parameter(s).
2. Write down the derivative of the log likelihood with respect to each parameter.
3. Find the parameter values such that the derivatives are zero

Multiple Parameters Example

▷ **Example 27.3.4.** Red/green wrapper depends probabilistically on flavour:



▷ Likelihood for, e.g., cherry candy in green wrapper:

$$\begin{aligned} & P(F = cherry, W = green|h_{\theta,\theta_1,\theta_2}) \\ = \;& P(F = cherry|h_{\theta,\theta_1,\theta_2}) \cdot P(W = green|F = cherry, h_{\theta,\theta_1,\theta_2}) \\ = \;& \theta \cdot (1 - \theta_1) \end{aligned}$$

▷ **Ovservation:** For N candies, r_c red-wrapped cherry candies, etc. we have

$$P(\mathbf{d}|h_{\theta,\theta_1,\theta_2}) = \theta^c \cdot (1-\theta)^\ell \cdot {\theta_1}^{r_c} \cdot (1-\theta_1)^{g_c} \cdot {\theta_2}^{r_\ell} \cdot (1-\theta_2)^{g_\ell}$$

Multiple Parameters Example (contd.)

▷ Minimize the log likelihood:

$$\begin{aligned} L &= c\log_2(\theta) + \ell\log_2(1-\theta) \\ &+ r_c\log_2(\theta_1) + g_c\log_2(1-\theta_1) \\ &+ r_\ell\log_2(\theta_2) + g_\ell\log_2(1-\theta_2) \end{aligned}$$

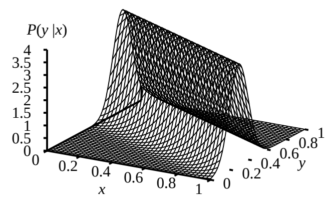
▷ Derivatives of L contain only the relevant parameter:

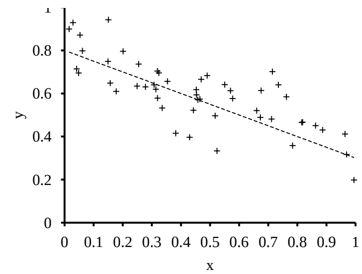
$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \rightsquigarrow \quad \theta = \frac{c}{c+\ell} \\ \frac{\partial L}{\partial \theta_1} &= \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \rightsquigarrow \quad \theta_1 = \frac{r_c}{r_c+g_c} \\ \frac{\partial L}{\partial \theta_2} &= \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \rightsquigarrow \quad \theta_2 = \frac{r_\ell}{r_\ell+g_\ell} \end{aligned}$$

▷ **Upshot:** With complete data, *parameters can be learned separately* in Bayesian networks.

▷ **Remaining Problem:** Have to be careful with zero values! (division by zero)

Example: Linear Gaussian Model





- ▷ Maximizing $P(y|x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$ w.r.t. θ_1, θ_2

  $=$ minimizing $E = \sum_{j=1}^{N} (y_j(\theta_1 x_j + \theta_2))^2$

- ▷ That is, minimizing the sum of squared errors gives the ML solution for a linear fit *assuming Gaussian noise of fixed variance.*

27.4 Naive Bayes Models

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30391`.

Naive Bayes Models

- ▷ **Definition 27.4.1.** A Bayesian network in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause is called a naive Bayes model or Bayesian classifier.

- ▷ **Observation 27.4.2.** *In a naive Bayes model, the full joint probability distribution can be written as*

$$\mathbf{P}(\mathit{cause}|\mathit{effect}_1, \ldots, \mathit{effect}_n) = \alpha\langle \mathit{effect}_1, \ldots, \mathit{effect}_n\rangle \cdot \mathbf{P}(\mathit{cause}) \cdot \prod_i \mathbf{P}(\mathit{effect}_i|\mathit{cause})$$

- ▷ **Note:** This kind of model is called "naive" since it is often used as a simplifying model if the effects are not conditionally independent after all.

- ▷ It is also called idiot Bayes model by Bayesian fundamentalists.

- ▷ In practice, naive Bayes models can work surprisingly well, even when the conditional independence assumption is not true.

- ▷ **Example 27.4.3.** The dentistry example is a (true) naive Bayes model.

Naive Bayes Models for Learning (continued)

- ▷ Naive Bayes models are probably the most commonly used Bayesian network model in machine learning.
  - ▷ The "class" variable C (which is to be predicted) is the root.

- The "attribute" variables X_i are the leaves.

▷ **Observation:** The Example 27.3.4 is a (true) naive Bayes model.(only one effect)

- Assuming Boolean variables, the parameters are: $\theta = P(c = \mathsf{T})$, $\theta_{i1} = P(X_i = \mathsf{T}|C = \mathsf{T})$, and $\theta_{i2} = P(X_i = \mathsf{T}|C = \mathsf{F})$
- then the maximum likelihood parameters can be found exactly like above.

▷ **Idea:** Once trained, use this model to classify new examples, where C is unobserved:

- With observed values $x_1, \ldots x_n$, the probability of each class is given by

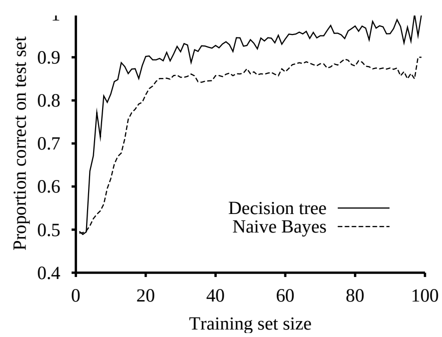
$$\mathbf{P}(C|x_1, \ldots, x_n) = \alpha \cdot \mathbf{P}(C) \cdot \prod_i \mathbf{P}(x_i|C)$$

- A deterministic prediction can be obtained by choosing the most likely class.

Naive Bayes Models for Learning (Properties)

▷ Naive Bayes learning turns out to do surprisingly well in a wide range of applications.

▷ **Example 27.4.4.** Learning curve for naive Bayes learning on the restaurant example



▷ **Naive Bayes learning scales well:** with n Boolean attributes, there are just $2n + 1$ parameters, and no search is required to find h_{ML}.

▷ Naive Bayes learning systems have no difficulty with noisy or missing data and can give probabilistic predictions when appropriate.

Statistical Learning: Summary

▷ Full Bayesian learning gives best possible predictions but is intractable.

▷ MAP learning balances complexity with accuracy on training data.

▷ Maximum likelihood learning assumes uniform prior, OK for large data sets:

1. Choose a parameterized family of models to describe the data.
   $\rightsquigarrow$ requires substantial insight and sometimes new models.
2. Write down the likelihood of the data as a function of the parameters.
   $\rightsquigarrow$ may require summing over hidden variables, i.e., inference.
3. Write down the derivative of the log likelihood w.r.t. each parameter.
4. Find the parameter values such that the derivatives are zero.
   $\rightsquigarrow$ may be hard/impossible; modern optimization techniques help.

▷ Naive Bayes models as a fall-back solution for machine learning:

- ▷ conditional independence of all attributes as simplifying assumption.

Chapter 28

Knowledge in Learning

28.1 Logical Formulations of Learning

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/30392` and `https://fau.tv/clip/id/30393`.

Knowledge in Learning: Motivation

- ▷ **Recap:** Learning from examples. (last chapter)
- ▷ **Idea:** Construct a function with the input/output behavior observed in data.
- ▷ **Method:** Search for suitable functions in the hypothesis space. (e.g. decision trees)
- ▷ **Observation 28.1.1.** *Every learning task begins from zero. (except for the choice of hypothesis space)*
- ▷ **Problem:** We have to forget everything before we can learn something new.
- ▷ **Idea:** Utilize prior knowledge about the world! (represented e.g. in logic)

A logical Formulation of Learning

- ▷ **Recall:** Examples are composed of descriptions (of the input sample) and classifications.
- ▷ **Idea:** Represent examples and hypotheses as logical formulae.
- ▷ **Example 28.1.2.** For attribute-based representations, we can use PL^1: we use predicate constants for Boolean attributes and classification and function constants for the other attributes.
- ▷ **Definition 28.1.3.** Logic based inductive learning tries to learn an hypothesis h that explains the classifications of the examples given their description, i.e. $h, \mathcal{D} \models \mathcal{C}$ (the explanation constraint), where

- ▷ $\mathcal{D}$ is the conjunction of the descriptions, and
- ▷ $\mathcal{C}$ the conjunction of their classifications.

▷ **Idea:** We solve the explanation constraint $h, \mathcal{D} \models \mathcal{C}$ for h where h ranges over some hypothesis space.

▷ **Refinement:** Use Occam's razor or additional constraints to avoid $h = \mathcal{C}$. (too easy otherwise/boring; see below)

A logical Formulation of Learning (Restaurant Examples)

▷ **Example 28.1.4 (Restaurant Example again).** Descriptions are conjunctions of literals built up from

- ▷ predicates Alt, Bar, Fri/Sat, Hun, Rain, and Res
- ▷ equations about the functions Pat, Price, Type, and Est.

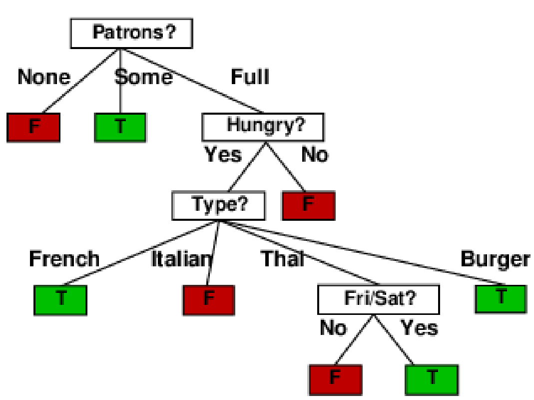
For instance the first example X_1 from Example 26.3.2, can be described as

$$\text{Alt}(X_1) \wedge \neg\text{Bar}(X_1) \wedge \text{Fri/Sat}(X_1) \wedge \text{Hun}(X_1) \wedge \ldots$$

The classification is given by the goal predicate WillWait, in this case $\text{WillWait}(X_1)$ or $\neg\text{WillWait}(X_1)$.

A logical Formulation of Learning (Restaurant Tree)

▷ **Example 28.1.5 (Restaurant Example again; Tree).** The induced decision tree from Example 26.4.9



can be represented as

$$\begin{array}{rcl} \forall r.\mathrm{WillWait}(r) & \Leftrightarrow & \mathrm{Pat}(r, \mathrm{Some}) \\ & \vee & \mathrm{Pat}(r, \mathrm{Full}) \wedge \mathrm{Hun}(r) \wedge \mathrm{Type}(r, \mathrm{French}) \\ & \vee & \mathrm{Pat}(r, \mathrm{Full}) \wedge \mathrm{Hun}(r) \wedge \mathrm{Type}(r, \mathrm{Thai}) \wedge \mathrm{Fri/Sat}(r) \\ & \vee & \mathrm{Pat}(r, \mathrm{Full}) \wedge \mathrm{Hun}(r) \wedge \mathrm{Type}(r, \mathrm{Burger}) \end{array}$$

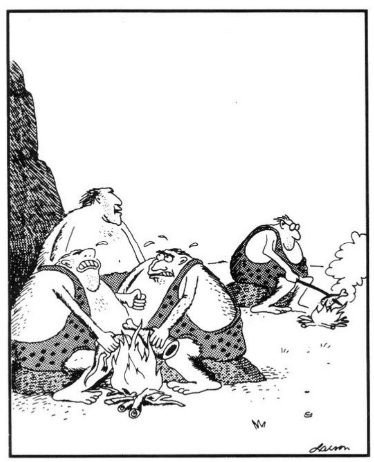
Method: Construct a disjunction of all the paths from the root to the positive leaves interpreted as conjunctions of the attributes on the path.

Note: The equivalence takes care of positive and negative examples.

Cumulative Development

▷ **Example 28.1.6.** Learning from very few examples using background knowledge:

1. Caveman Zog and the fish on a stick:



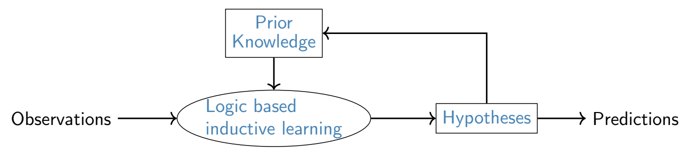
"Hey! Look what Zog do!"

2. Generalizing from one Brazilian:
   Upon meeting her first Brazilian – Fernando – who speaks Portugese, Sarah
   - ▷ learns/generalizes that all Brazilians speak Portugese,
   - ▷ but not that all Brazilians are called Fernando.
3. General rules about effectiveness of antibiotics:
   When Sarah – gifted in diagnostics, but clueless in pharmacology – observes a doctor prescribing the antibiotic Proxadone for an inflamed foot, she learns/infers that Proxadone is effective against this ailment.

▷ **Observation:** The methods/algorithms from section 26.2 cannot replicate this. (why?)

▷ **Missing Piece:** The background knowledge!

- ▷ **Problem:** To use background knowledge, need a method to obtain it. (use learning)
- ▷ **Question:** How to use knowledge to learn more efficiently?
- ▷ **Answer:** Cumulative development: collect knowledge and use it in learning!



- ▷ **Definition 28.1.7.** We call the body of knowledge accumulated by (a group of) agents their background knowledge. It acts as prior knowledge in logic based learning processes.

Adding Background Knowledge to Learning: Overview

- ▷ Explanation based learning (EBL)
- ▷ Relevance based learning (RBL)
- ▷ Knowledge based inductive learning (KBIL)

Explanation-based Learning

- ▷ **Idea:** Use explanation of success to infer a general rule.
- ▷ **Example 28.1.8 (Caveman Zog).** Cavemen generalize by explaining the success of the pointed stick: it supports the lizard while keeping hand away from fire.

  From this explanation, they can infer a *general rule*: any long, rigid, sharp object can be used to toast small, soft-bodied edibles.
- ▷ **Definition 28.1.9.** Explanation based learning (EBL) refines the explanation constraint to the EBL constraints:

$$\begin{aligned} Hypothesis \wedge Descriptions &\models Classifications \\ Background &\models Hypothesis \end{aligned}$$

- ▷ **Intuition:** Converting first-principles theories into useful, special purpose knowledge.
- ▷ **Observation:** General rule follows logically from the background knowledge.

Relevance-based Learning

- **Idea:** Use the prior knowledge to determine the relevance of a set of features to the goal predicate. (reduce the hypothesis space to the relevant ones)
- **Example 28.1.10.** In a given country most people speak the same language, but do not have the same name:
- **Definition 28.1.11.** Relevance based learning (RBL) refines the explanation constraint to the RBL constraints:

$$\begin{aligned} Hypothesis \wedge Descriptions &\models Classifications \\ Background \wedge Descriptions \wedge Classifications &\models Hypothesis \end{aligned}$$

The second constraint only allows hypotheses that are relevant.

Deductive Learning

- **Definition 28.1.12.** We call a procedure a deductive learning algorithm, if it makes use of the observations, but does not produce hypothesis beyond the background knowledge and the observations.
- **Example 28.1.13.** EBL and RBL are deductive learning processes.
- **Problem:** Deductive learning processes do not learn anything factually new!

  The cavemen could have learned without Zog by thinking (Zog did!)
- **Idea:** Replace the explanation constraint by something stronger.

Three Principal Modes of Inference

- **Definition 28.1.14.** Deduction $\widehat{=}$ knowledge extension
- **Example 28.1.15.** $\dfrac{rains \Rightarrow wet\_street \quad rains}{wet\_street} D$
- **Definition 28.1.16.** Abduction $\widehat{=}$ explanation
- **Example 28.1.17.** $\dfrac{rains \Rightarrow wet\_street \quad wet\_street}{rains} A$
- **Definition 28.1.18.** Induction $\widehat{=}$ learning general rules from examples
- **Example 28.1.19.** $\dfrac{wet\_street \quad rains}{rains \Rightarrow wet\_street} I$

Knowledge-based Inductive Learning

▷ **Idea:** Background knowledge and new hypothesis combine to explain the examples.

▷ **Example 28.1.20.** Inferring disease D from the symptoms is not enough to explain the prescription of medicine M.

Need a new general rule: M *is effective against* D (induction from example)

▷ **Definition 28.1.21.** Knowledge based inductive learning (KBIL) replaces the explanation constraint by the KBIL constraint:

$$Background \wedge Hypothesis \wedge Descriptions \models Classifications$$

Inductive Logic Programming

▷ **Definition 28.1.22.** Inductive logic programming (ILP) is logic based inductive learning method that uses logic programming as a uniform representation for examples, background knowledge and hypotheses.

Given an encoding of the known background knowledge and a set of examples represented as a logical knowledge base of facts, an ILP system will derive a hypothesised logic program which entails all the positive and none of the negative examples.

▷ Main field of study for KBIL algorithms.

▷ Prior knowledge plays two key roles:

1. The effective hypothesis space is reduced to include only those theories that are consistent with what is already known.
2. Prior knowledge can be used to reduce the size of the hypothesis explaining the observations.
   ▷ Smaller hypotheses are easier to find.

▷ **Observation:** ILP systems can formulate hypotheses in first-order logic.

$\rightsquigarrow$ Can learn in environments not understood by simpler systems.

28.2 Explanation-Based Learning

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30394`.

Explanation-Based Learning

▷ **Intuition:** EBL $\widehat{=}$ Extracting general rules from individual observations.

▷ **Example 28.2.1.** Differentiating and simplifying algebraic expressions

1. Differentiate X^2 with respect to X to get $2X$.
2. Logical reasoning system $\mathsf{ask}(\mathsf{Deriv}(X^2, X) = d, KB)$ with solution $d = 2X$.
3. Solving this for the first time using standard rules of differentiation gives $1 \times (2 \times (X^{2-1}))$.
4. This takes a first-time program 136 proof steps with 99 dead end branches.

▷ **Idea:** Use memoization:

- ▷ Speed up by saving the results of computation.
- ▷ Create a database of input/output pairs.

Creating general rules

▷ Memoization in explanation-based learning

- ▷ Create general rules that cover an entire class of cases

▷ **Example 28.2.2.** Extract the general rule $\mathsf{ArithVar}(u) \Rightarrow \mathsf{Deriv}(u^2, u) = 2u$.

▷ Once something is understood, it can be generalized and reused in other circumstances.

- ▷ *Civilization advances by extending the number of important operations that we can do without thinking about them.* (Alfred North Whitehead)

▷ Explaining why something is a good idea is much easier than coming up with the idea in the first place:

- ▷ Watch caveman Zog roast his lizard vs. thinking about putting the fish on a stick.

Extracting rules from examples

▷ Basic idea behind EBL:

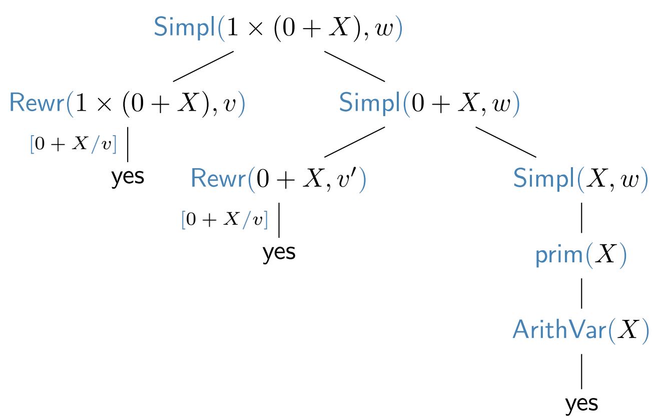
1. Construct an explanation of the observation using prior knowledge.
2. Establish a definition of the class of cases for which the same explanation can be used.

▷ **Example 28.2.3.** Simplifying $1 \times (0 + X)$ using a knowledge base with the following rules:

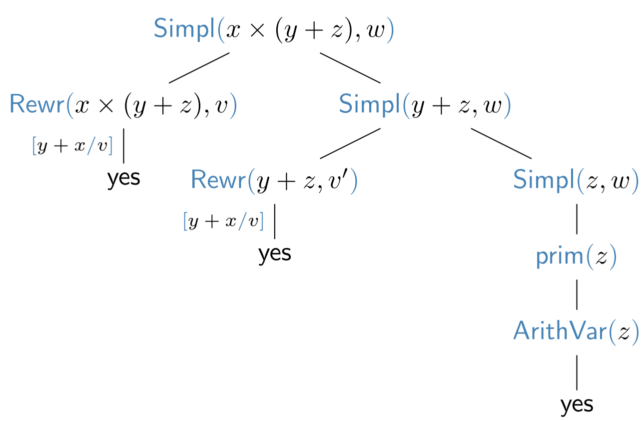
- ▷ $\mathsf{Rewr}(u, v) \wedge \mathsf{Simpl}(v, w) \Rightarrow \mathsf{Simpl}(u, w)$
- ▷ $\mathsf{prim}(u) \Rightarrow \mathsf{Simpl}(u, u)$
- ▷ $\mathsf{ArithVar}(u) \Rightarrow \mathsf{prim}(u)$
- ▷ $\mathsf{Num}(u) \Rightarrow \mathsf{prim}(u)$

- $\triangleright$ $\mathsf{Rewr}(1 \times u, u)$
- $\triangleright$ $\mathsf{Rewr}(0 + u, u)$
- $\triangleright$ ...

Proof Tree for the Original Problem



Generalized Proof Tree



Generalizing proofs in EBL

- $\triangleright$ The variabilized proof proceeds using exactly the same rule applications.
  - $\triangleright$ This may lead to variable instantiation.

▷ **Example 28.2.4.** Take the leaves of the generalized proof tree to get the general rule

$$\mathsf{Rewr}(1 \times (0+z), 0+z) \wedge \mathsf{Rewr}(0+z, z) \wedge \mathsf{ArithVar}(z) \Rightarrow \mathsf{Simpl}(1 \times (0+z), z)$$

▷ The first two conditions are true independently of z, so this becomes

$$\mathsf{ArithVar}(z) \Rightarrow \mathsf{Simpl}(1 \times (0+z), z)$$

▷ **Recap:**

- ▷ Use background knowledge to construct a proof for the example.
- ▷ In parallel, construct a generalized proof tree.
- ▷ New rule is the conjunction of the leaves of the proof tree and the variabilized goal.
- ▷ Drop conditions that are true regardless of the variables in the goal.

Improving Efficiency of EBL

▷ **Idea:** Pruning the proof tree to get more general rules.

▷ **Example 28.2.5.**

$$\mathsf{prim}(z) \Rightarrow \mathsf{Simpl}(1 \times (0+z), z)$$

$$\mathsf{Simpl}(y+z, w) \Rightarrow \mathsf{Simpl}(1 \times (y+z), w)$$

▷ **Problem:** Which rules to choose?

- ▷ Adding large numbers of rules to the knowledge base slows down the reasoning process (increases the branching factor of the search space).
- ▷ To compensate, the derived rules must offer significant speed increases.
- ▷ Derived rules should be as general as possible to apply to the largest possible set of cases.

Improving efficiency in EBL (continued)

▷ Operationality of subgoals in the rule:

- ▷ A subgoal must be "easy" to solve.
- ▷ $\mathsf{prim}(z)$ is easy to solve, but $\mathsf{Simpl}(y+z, w)$ leads to an arbitrary amount of inference.
- ▷ Keep operational subgoals and prune the rest of the tree.

▷ Trade-off between operationality and generality:

- ▷ More specific subgoals are easier to solve but cover fewer cases.

- ▷ How many steps are still called operational?
- ▷ Cost of a subgoal depends on the rules in the knowledge base.

Maximizing the efficiency of an initial knowledge base is a complex optimization problem.

Improving efficiency of EBL (Analysis)

- ▷ Empirical analysis of efficiency:
  - ▷ Average-case complexity on a population of problems that needs to be solved.
- ▷ By generalizing from past example problems, EBL makes the knowledge base more efficient for the kind of problems that it is reasonable to expect.
  - ▷ Works if the distribution of past problems is roughly the same as for future problems.
  - ▷ Can lead to great improvements
    - ▷ Swedish to English translator was made 1200 times faster by using EBL [SR91].

28.3 Relevance-Based Learning

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30394`.

Recap: Relevance-based Learning

- ▷ The prior knowledge concerns the relevance of a set of features to the goal predicate.
- ▷ **Example 28.3.1.** In a given country most people speak the same language, but do not have the same name

$$\begin{aligned} Hypothesis \wedge Descriptions &\models Classifications \\ Background \wedge Descriptions \wedge Classifications &\models Hypothesis \end{aligned}$$

- ▷ Deductive learning: Makes use of the observations, but does not produce hypothesis beyond the background knowledge and the observations.

Relevance-based Learning: Determinations

- ▷ **Example 28.3.2 (Background knowledge in Brazil).**

$$\forall x, y, n, l. \mathsf{Nationality}(x,n) \wedge \mathsf{Nationality}(y,n) \wedge \mathsf{Language}(x,l) \Rightarrow \mathsf{Language}(y,l)$$

So

$$\text{Nationality}(Fernando, Brazil) \wedge \text{Language}(Fernando, Portuguese)$$

entails

$$\forall x.\text{Nationality}(x, Brazil) \Rightarrow \text{Language}(x, Portugese)$$

Special syntax: $\text{Nationality}(x, n) \succ \text{Language}(x, l)$

▷ **Definition 28.3.3.** If $\forall v, w.\forall x, y.P(x, v) \wedge P(y, v) \wedge Q(x, w) \Rightarrow Q()$, then we say that P determines Q and write $P \succ Q$; we call this formula a determination or functional dependency.

Here x and y range over all examples; v and w range over the possible values of attributes P and Q, respectively.

▷ **Intuition:** If we know the values of P and Q for one example x, e.g., $P(x, a)$ and $Q(x, b)$, we can use the determination $P \succ Q$ and to infer $\forall y.P(y, a) \Rightarrow Q(y, b)$.

Determining the Hypothesis Space

▷ Determinations limit the hypothesis space.

- ▷ Only consider the important features (i.e. not day of the week, hair style of David Beckham).

▷ Determinations specify a sufficient basis vocabulary from which to construct hypotheses.

▷ Reduction of the hypothesis space makes it easier to learn the target predicate:

- ▷ Learning Boolean functions of n variables in CNF: Size of the hypothesis space $\#(H) = \mathcal{O}(2^{2^n})$.
- ▷ For Boolean functions $\log_2(\#(H))$ examples are needed in a $\#(H)$ size hypothesis space: Without restrictions, this is $\mathcal{O}(2^n)$ examples.
- ▷ If the determination contains d predicates on the left, only $\mathcal{O}(2^d)$ examples are needed.
- ▷ Reduction of size by $\mathcal{O}(2^{n-d})$.

Learning Relevance Information

▷ **Observation:** Prior knowledge also needs to be learned.

▷ **Idea:** Learning algorithms for determinations:

- ▷ Find the simplest determination consistent with the observations.
- ▷ A determination $P \succ Q$ says: if examples match P they must also match Q.

▷ **Definition 28.3.4.** A determination $P \succ Q$ is consistent with a set of examples E, if every pair in E that matches on the predicates in P also matches on the target predicate.

A consistent determination $P \succ Q$ is minimal, iff there is no consistent determination $P' \succ Q$ with fewer atoms in P'.

Learning relevance information

Example 28.3.5.

Sample	Mass	Temp	Material	Size	Conductance
S1	12	26	Copper	3	0.59
S1	12	100	Copper	3	0.57
S2	24	26	Copper	6	0.59
S3	12	26	Lead	2	0.05
S3	12	100	Lead	2	0.04
S4	24	26	Lead	4	0.05

▷ Minimal consistent determination $(Material \wedge Temperature) \succ Conductance$

▷ Non-minimal consistent determination $(Mass \wedge Size \wedge Temperature) \succ Conductance$

Learning Relevance Information (Algorithm)

▷ **Definition 28.3.6.** The MCD algorithm is a simple $\emptyset$-up generate and test algorithm over subsets:

```
function MCD(E,A) returns a determination
  inputs: E, a set of examples
          A, a set of attributes, of size n
  for i := 1, ..., n do
        for each subset A_i of A of size i do
            if ConsDet?(A_i,E) then return A_i
        end
  end

function ConsDet?(A,E) returns a truth—value
  inputs: A, a set of attributes
          E, a set of examples
  local variables: H, a hash table
  for each example e in E do
      if some h∈H has the same A—value as e but different class
      then return False
      store the class of e in H, indexed by the A—values of e
  end
  return True
```

Complexity of the MCD Algorithm

- ▷ Time complexity depends on the size of the minimal consistent determinations.
- ▷ In case of p attributes and a total of n attributes, the algorithm has to search all subsets of A of size p.
- ▷ There are $\mathcal{O}(\binom{n}{p}) = \mathcal{O}(n^p)$ of these, so the MCD algorithm is exponential.
- ▷ **Theorem 28.3.7.** *The general problem of finding minimal consistent determinations is NP complete.*
- ▷ **Good News:** In most domains there is sufficient local structure to make p small.

Deriving Hypotheses

- ▷ Given an algorithm for learning determinations, a learning agent has a way to construct a minimal hypothesis within which to learn the target predicate.
- ▷ **Idea:** Use decision tree learning for computing hypotheses.
- ▷ **Goal:** Minimize size of hypotheses.
- ▷ **Result:** Relevance based decision tree learning.

Relevance-based Decision Tree Learning

- ▷ **Idea:** Use determinations to tune attribute selection in decision tree learning.
- ▷ **Definition 28.3.8.** The relevance based decision tree learning algorithm (RBDTL) first determines a relevant set of of attributes by MCD.

  function RBDTL(E,A,v) **returns** a decision tree
  return DTL(E, MCD(E,A),v)

  Then uses it by "regular" DTL: (with adapted recursive call)

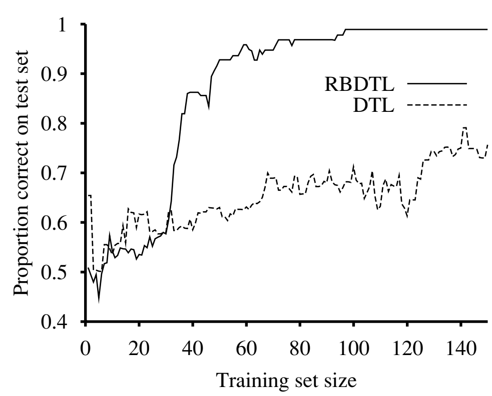
  function DTL($examples$,$attributes$,$default$) **returns** a decision tree
  if $examples$ is empty **then return** $default$
  else if all examples have the same classification **then return** the classification
  else if $attributes$ is empty **then return** Majority($examples$)
  else
  $best$:= Choose−Attribute($attributes$,$examples$)
  $tree$:= a new decision tree with root test $best$
  for each value v_i of best **do**
  examples$_i$:= {elements of examples with $best = v_i$}
  subtree := RBDTL($examples_i$,$attributes - best$, Mode($examples$))
  add a branch **to** tree with label v_i and subtree $subtree$
  return tree

Exploiting Knowledge

- RBDTL simultaneously learns and uses relevance information to minimize its hypothesis space.
- Declarative bias
  - How can prior knowledge be used to identify the appropriate hypothesis space to search for the correct target definition?
- Unanswered questions: (engineering needed to make ideas practical)
  - How to handle noise?
  - How to use other kinds of prior knowledge besides determinations?
  - How can the algorithms be generalized to cover any first-order theory?

RBDTL vs. DTL

- **Observation:** RBDTL does rather well on well-structured domains.
- **Example 28.3.9.** A performance comparison between DTL and RBDTL on randomly generated data for a target function that depends on only 5 of 16 attributes.



28.4 Inductive Logic Programming

Inductive Logic Programming

- Combines inductive methods with the power of first-order representations.
- Offers a rigorous approach to the general KBIL problem.

▷ Offers complete algorithms for inducing general, first-order theories from examples.

28.4.1 An Example

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/30396`.

ILP: An example

▷ General knowledge-based induction problem

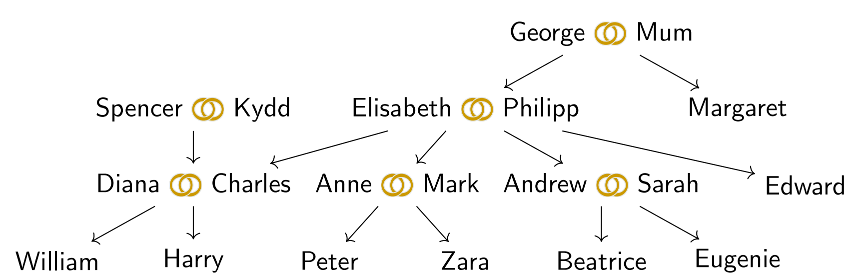
$$Background \wedge Descriptions \wedge Classifications \models Hypothesis$$

▷ **Example 28.4.1 (Learning family relations from examples).**

- ▷ Observations are an extended family tree
  - ▷ mother, father and married relations
  - ▷ male and female properties
- ▷ Target predicates: grandparent, BrotherInLaw, Ancestor

British Royalty Family Tree (not quite not up to date)

▷ The facts about kinship and relations can be visualized as a family tree:



Example

▷ Descriptions include facts like

- ▷ father$(Philip, Charles)$
- ▷ mother$(Mum, Margaret)$
- ▷ married$(Diana, Charles)$
- ▷ male$(Philip)$
- ▷ female$(Beatrice)$

▷ Sentences in classifcations depend on the target concept being learned (in the example: 12 positive, 388 negative)

▹ grandparent$(Mum, Charles)$

▹ ¬grandparent$(Mum, Harry)$

▷ **Goal:** Find a set of sentences for hypothesis such that the entailment constraint is satisfied.

▷ **Example 28.4.2.** Without background knowledge, define grandparent in terms of mother and father.

$$\mathrm{grandparent}(x,y)\Leftrightarrow(\exists z.\mathrm{mother}(x,z)\wedge\mathrm{mother}(z,y))\vee(\exists z.\mathrm{mother}(x,z)\wedge\mathrm{father}(z,y))\vee\ldots\vee(\exists z.\mathrm{father}(x,z)\wedge\mathrm{father}(z,y))$$

Why Attribute-based Learning Fails

▷ **Observation:** Decision tree learning will get nowhere!

▹ To express Grandparent as a (Boolean) attribute, pairs of people need to be objects $Grandparent(\langle Mum, Charles\rangle)$.

▹ But then the example descriptions can not be represented

$$FirstElementIsMotherOfElizabeth(\langle Mum, Charles\rangle)$$

▹ A large disjunction of specific cases without any hope of generalization to new examples.

▷ **Generally:** Attribute-based learning algorithms are incapable of learning relational predicates.

Background knowledge

▷ **Observation:** A little bit of background knowledge helps a lot.

▷ **Example 28.4.3.** If the background knowledge contains

$$\mathrm{parent}(x,y)\Leftrightarrow\mathrm{mother}(x,y)\vee\mathrm{father}(x,y)$$

then Grandparent can be reduced to

$$\mathrm{grandparent}(x,y)\Leftrightarrow(\exists z.\mathrm{parent}(x,z)\wedge\mathrm{parent}(z,y))$$

▷ **Definition 28.4.4.** A constructive induction algorithm creates new predicates to facilitate the expression of explanatory hypotheses.

▷ **Example 28.4.5.** Use constructive induction to introduce a predicate parent to simplify the definitions of the target predicates.

28.4.2 Top-Down Inductive Learning: FOIL

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/30397`.

Top-Down Inductive Learning

- ▷ Top-down learning method
  - ▷ Decision-tree learning: start from the observations and work backwards.
  - ▷ Decision tree is gradually grown until it is consistent with the observations.
  - ▷ Top-down learning: start from a general rule and specialize it.

Top-Down Inductive Learning: FOIL

- ▷ Split positive and negative examples
  - ▷ Positive: $\langle George, Anne\rangle$, $\langle Philip, Peter\rangle$, $\langle Spencer, Harry\rangle$
  - ▷ Negative: $\langle George, Elizabeth\rangle$, $\langle Harry, Zara\rangle$, $\langle Charles, Philip\rangle$
- ▷ Construct a set of Horn clauses with head grandfather(x, y) such that the positive examples are instances of the grandfather relationship.
  - ▷ Start with a clause with an empty body $\Rightarrow$grandfather(x, y).
  - ▷ All examples are now classified as positive, so specialize to rule out the negative examples: Here are 3 potential additions:
    1. father(x, y) $\Rightarrow$ grandfather(x, y)
    2. parent(x, z) $\Rightarrow$ grandfather(x, y)
    3. father(x, z) $\Rightarrow$ grandfather(x, y)
  - ▷ The first one incorrectly classifies the 12 positive examples.
  - ▷ The second one is incorrect on a larger part of the negative examples.
  - ▷ Prefer the third clause and specialize to father$(x, z)\wedge$parent$(z, y)\Rightarrow$grandfather(x, y).

FOIL

```
function Foil(examples,target) returns a set of Horn clauses
   inputs: examples, set of examples
           target, a literal for the goal predicate
   local variables: clauses, set of clauses, initially empty
   while examples contains positive examples do
     clause := New−Clause(examples,target)
     remove examples covered by clause from examples
     add clause to clauses
   return clauses
```

FOIL

function New—Clause($examples$,$target$) **returns** a Horn clause
  local variables: $clause$, a clause with target as head and an empty body
    l, a literal **to** be added **to** the clause
    $extendedExamples$, a set of examples with values **for** new variables
  $extendedExamples := examples$
  while $extendedExamples$ contains negative examples **do**
    $l :=$ Choose—Literal(New—Literals($clause$),$extendedExamples$)
    append l **to** the body of $clause$
    $extendedExamples :=$ map Extend—Example over $extendedExamples$
  return $clause$

function Extend—Example($example$,$literal$) **returns** a new example
  if $example$ satisfies $literal$
  then return the set of examples created by extending $example$ with each
    possible constant value **for** each new variable **in** $literal$
  else return the empty set

function New—Literals($clause$) **returns** a set of possibly ‘‘useful’’ literals
function Choose—Literal($literals$) **returns** the ‘‘best’’ literal from $literals$

FOIL: Choosing Literals

- ▷ New-Literals: Takes a clause and constructs all possibly “useful” literals
- ▷ father$(x, z) \Rightarrow$ grandfather(x, y)
- ▷ Add literals using predicates
  - ▷ Negated or unnegated
  - ▷ Use any existing predicate (including the goal)
  - ▷ Arguments must be variables
  - ▷ Each literal must include at least one variable from an earlier literal or from the head of the clause
    - ▷ Valid: $Mother(z, u), Married(z, z)$, grandfather$(v, x)$
    - ▷ Invalid: $Married(u, v)$
- ▷ Equality and inequality literals
  - ▷ E.g. $z \neq x$, empty list
- ▷ Arithmetic comparisons
  - ▷ E.g. $x > y$, threshold values

FOIL: Choosing Literals

- The way New-Literal changes the clauses leads to a very large branching factor.
- Improve performance by using type information:
  - E.g., $\mathsf{parent}(x, n)$ where x is a person and n is a number
- Choose-Literal uses a heuristic similar to information gain.
- Ockham's razor to eliminate hypotheses.
  - If the clause becomes longer than the total length of the positive examples that the clause explains, this clause is not a valid hypothesis.
- Most impressive demonstration
  - Learn the correct definition of list-processing functions in Prolog from a small set of examples, using previously learned functions as background knowledge.

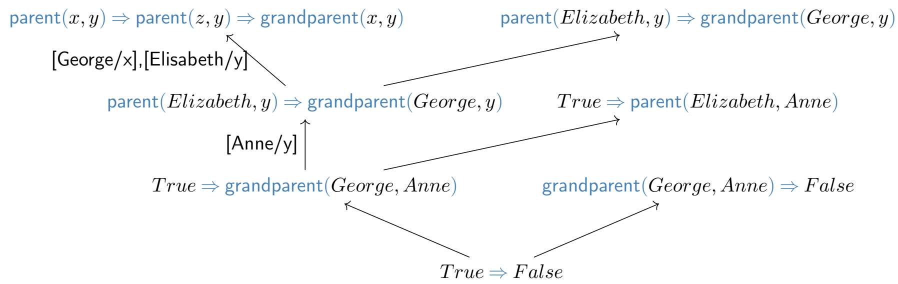
28.4.3 Inverse Resolution

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/30398`.

Inverse Resolution

- **Inverse resolution in a nutshell:**
  - Classifications follows from $Background \wedge Hypothesis \wedge Descriptions$.
  - This can be proven by resolution
  - Run the proof backwards to find hypothesis
- **Problem:** How to run the proof backwards?
- **Recap:** In ordinary resolution we take two clauses $C_1 = L \vee R_1$ and $C_2 = \neg L \vee R_2$ and resolve them to produce the resolvent $C = R_1 \vee R_2$.
- **Idea:** Two possible variants of inverse resolution:
  - Take resolvent C and produce two clauses C_1 and C_2.
  - Take C and C_1 and produce C_2.

Generating Inverse Proofs (Example)



Generating Inverse Proofs

- Inverse resolution is a search algorithm: For any C and C_1 there can be several or even an infinite number of clauses C_2.
- **Example 28.4.6.** Instead of $\mathsf{parent}(Elizabeth,y) \Rightarrow \mathsf{grandparent}(George,y)$ there were numerous alternatives:
  - $\mathsf{parent}(Elizabeth,Anne) \Rightarrow \mathsf{grandparent}(George,Anne)$
  - $\mathsf{parent}(z,Anne) \Rightarrow \mathsf{grandparent}(George,Anne)$
  - $\mathsf{parent}(z,y) \Rightarrow \mathsf{grandparent}(George,y)$
- The clauses C_1 that participate in each step can be chosen from Background, Descriptions, Classifications or from hypothesized clauses already generated.
- ILP needs restrictions to make the search manageable
  - Eliminate function symbols
  - Generate only the most specific hypotheses
  - Use Horn clauses
  - All hypothesized clauses must be consistent with each other
  - Each hypothesized clause must agree with the observations

New Predicates and New Knowledge

- An inverse resolution procedure is a complete algorithm for learning first-order logicfirst-order theories:
  - If some unknown hypothesis generates a set of examples, then an inverse resolution procedure can generate hypothesis from the examples.
- Can inverse resolution infer the law of gravity from examples of falling bodies?
  - Yes, given suitable background mathematics!

▷ Monkey and typewriter problem: How to overcome the large branching factor and the lack of structure in the search space?

New Predicates and New Knowledge

▷ Inverse resolution is capable of generating new predicates:

- ▷ Resolution of C_1 and C_2 into C eliminates a literal that C_1 and C_2 share.
- ▷ This literal might contain a predicate that does not appear in C.
- ▷ When working backwards, one possibility is to generate a new predicate from which to construct the missing literal.

New Predicates and New Knowledge

▷ **Example 28.4.7.**

$$\begin{array}{ccc} Father(George; y) \Rightarrow P(x, y) & & P(George; y) \Rightarrow Ancestor(George, y) \\ & [George/x] & \\ & Father(George; y) \Rightarrow Ancestor(George, y) & \end{array}$$

P can be used in later inverse resolution steps.

▷ **Example 28.4.8.** $mother(x, y) \Rightarrow P(x, y)$ or $father(x, y) \Rightarrow P(x, y)$ leading to the "Parent" relationship.

▷ Inventing new predicates is important to reduce the size of the definition of the goal predicate.

▷ Some of the deepest revolutions in science come from the invention of new predicates. (e.g. Galileo's invention of acceleration)

Applications of ILP

▷ ILP systems have outperformed knowledge free methods in a number of domains.

▷ Molecular biology: the GOLEM system has been able to generate high-quality predictions of protein structures and the therapeutic efficacy of various drugs.

▷ GOLEM is a completely general-purpose program that is able to make use of background knowledge about any domain.

Knowledge in Learning: Summary

- ▷ Cumulative learning: Improve learning ability as new knowledge is acquired.
- ▷ Prior knowledge helps to eliminate hypothesis and fills in explanations, leading to shorter hypotheses.
- ▷ Entailment constraints: Logical definition of different learning types.
- ▷ Explanation based learning (EBL): Explain the examples and generalize the explanation.
- ▷ Relevance base learning (RBL): Use prior knowledge in the form of determinations to identify the relevant attributes.
- ▷ Knowledge based inductive learning (KBIL): Finds inductive hypotheses that explain sets of observations.
- ▷ Inductive logic programming (ILP):
  - ▷ Perform KBIL using knowledge expressed in first-order logic.
  - ▷ Generates new predicates with which concise new theories can be expressed.

Chapter 29

Reinforcement Learning

29.1 Reinforcement Learning: Introduction & Motivation

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30399`.

Unsupervised Learning

- **So far:** we have studied "learning from examples". (functions, logical theories, probability models)
- **Now:** How can agents learn "what to do" in the absence of labeled examples of "what to do". We call this problem unsupervised learning.
- **Example 29.1.1 (Playing Chess).** Learn transition models for own moves and maybe predict opponent's moves.
- **Problem:** The agent needs to have some feedback about what is good/bad $\leadsto$ cannot decide "what to do" otherwise. (recall: external performance standard for learning agents)
- **Example 29.1.2.** The ultimate feedback in chess is whether you win, lose, or draw.
- **Definition 29.1.3.** We call a learning situation where there are no labeled examples unsupervised learning and the feedback involved a reward or reinforcement.
- **Example 29.1.4.** In soccer, there are intermediate reinforcements in the shape of goals, penalties, . . .

Reinforcement Learning as Policy Learning

- **Definition 29.1.5.** Reinforcement learning is a type of unsupervised learning where an agent learn how to behave in a environment by performing actions and seeing the results.
- **Recap:** In section 25.1 we introduced rewards as parts of MDPs (Markov decision processes) to define optimal policies.

- an optimal policy maximizes the expected total reward.

- **Idea:** The task of reinforcement learning is to use observed rewards to come up with an optimal policy.
- In MDPs, the agent has total knowledge about the environment *and the reward function*, in reinforcement learning we do not assume this. ($\rightsquigarrow$ POMDPs+reward-learning)
- **Example 29.1.6.** You play a game without knowing the rules, and at some time the opponent shouts *you lose!*

Scope and Forms of Reinforcement Learning

- **Reinforcement Learning solves all of AI:** An agent is placed in an environment and must learn to behave successfully therein.
- **KISS:** We will only look at simple environments and simple agent designs:
  - A utility-based agent learns a utility function on states and uses it to select actions that maximize the expected outcome utility. (passive learning)
  - A Q-learning agent learns an action-utility function, or Q-function, giving the expected utility of taking a given action in a given state. (active learning)
  - A reflex agent learns a policy that maps directly from states to actions.

29.2 Passive Learning

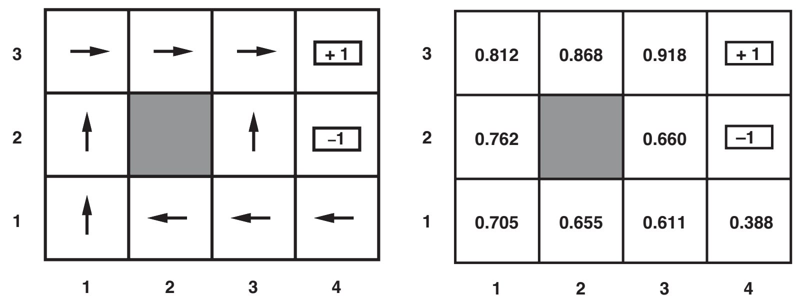
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/30400`.

Passive Learning

- **Definition 29.2.1 (To keep things simple).** Agent uses a state-based representation in a fully observable environment:
  - In passive learning, the agent's policy π is fixed: in state s, it always executes the action $\pi(s)$.
  - Its goal is simply to learn how good the policy is – that is, to learn the utility function $U^{\pi}(s)$.
- The passive learning task is similar to the policy evaluation task (part of the policy iteration algorithm) but the agent does not know
  - the transition model $P(s|s,a)$, which specifies the probability of reaching state s' from state s after doing action a,
  - the reward function $R(s)$, which specifies the reward for each state.

Passive Learning by Example

▷ **Example 29.2.2 (Passive Learning).** We use the 4×3 world introduced above



Optimal Policy π Utilities, given π

▷ The agent executes a set of in the environment using its policy π.

▷ In each trial, the agent starts in state (1,1) and experiences a sequence of state transitions until it reaches one of the terminal states, (4,2) or (4,3).

▷ Its percepts supply both the current state and the reward received in that state.

Passive Learning by Example

▷ **Example 29.2.3.** Typical trials might look like this:

1. $(1,1)_{-0.4} \rightsquigarrow (1,2)_{-0.4} \rightsquigarrow (1,3)_{-0.4} \rightsquigarrow (1,2)_{-0.4} \rightsquigarrow (1,3)_{-0.4} \rightsquigarrow (2,3)_{-0.4} \rightsquigarrow (3,3)_{-0.4} \rightsquigarrow (4,3)_{+1}$
2. $(1,1)_{-0.4} \rightsquigarrow (1,2)_{-0.4} \rightsquigarrow (1,3)_{-0.4} \rightsquigarrow (2,3)_{-0.4} \rightsquigarrow (3,3)_{-0.4} \rightsquigarrow (3,2)_{-0.4} \rightsquigarrow (3,3)_{-0.4} \rightsquigarrow (4,3)_{+1}$
3. $(1,1)_{-0.4} \rightsquigarrow (2,1)_{-0.4} \rightsquigarrow (3,1)_{-0.4} \rightsquigarrow (3,2)_{-0.4} \rightsquigarrow (4,2)_{-1}$.

▷ **Definition 29.2.4.** The utility is defined to be the expected sum of (discounted) rewards obtained if policy π is followed.

$$U^\pi(s) := E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t)\right]$$

where $R(s)$ is the reward for a state, S_t (a random variable) is the state reached at time t when executing policy π, and $S_0 = s$. (for 4×3 we take the discount factor $\gamma = 1$)

Direct Utility Estimation

▷ A simple method for direct utility estimation was invented in the late 1950s in the area of adaptive control theory.

▷ **Definition 29.2.5.** The utility of a state is the expected total reward from that state onward (called the expected reward to go).

▷ **Idea:** Each trial provides a sample of the reward to go for each state visited.

▷ **Example 29.2.6.** The first trial in Example 29.2.3 provides a sample total reward of 0.72 for state (1,1), two samples of 0.76 and 0.84 for (1,2), two samples of 0.80 and 0.88 for (1,3), ...

▷ **Definition 29.2.7.** The direct utility estimation algorithm cycles over trials, calculates the reward to go for each state, and updates the estimated utility for that state by keeping the running average for that for each state in a table.

▷ **Observation 29.2.8.** *In the limit, the sample average will converge to the true expectation (utility) from Definition 29.2.4.*

▷ *Remark 29.2.9.* Direct utility estimation is just supervised learning, where each example has the state as input and the observed reward to go as output.

▷ **Upshot:** We have reduced reinforcement learning to an inductive learning problem.

Adaptive Dynamic Programming

▷ **Problem:** The utilities of states are not independent in direct utility estimation!

▷ The utility of each state equals its own reward plus the expected utility of its successor states.

▷ **So:** The utility values obey a Bellman equation for a fixed policy π.

$$U^\pi(s) = R(s) + \gamma \cdot (\sum_{s'} P(s'|s,\pi(s)) \cdot U^\pi(s'))$$

▷ **Observation 29.2.10.** *By ignoring the connections between states, direct utility estimation misses opportunities for learning.*

▷ **Example 29.2.11.** Recall trial 2 in Example 29.2.3; state (3,3) is new.

2 $(1,1)_{-0.4} \rightsquigarrow (1,2)_{-0.4} \rightsquigarrow (1,3)_{-0.4} \rightsquigarrow (2,3)_{-0.4} \rightsquigarrow (3,3)_{-0.4} \rightsquigarrow (3,2)_{-0.4} \rightsquigarrow (3,3)_{-0.4} \rightsquigarrow (4,3)_{+1}$

- ▷ The next transition reaches (3,3), (known high utility from trial 1)
- ▷ Bellman equation: $\rightsquigarrow$ high $U^\pi(3,2)$ because $(3,2)_{-0.4} \rightsquigarrow (3,3)$
- ▷ But direct utility estimation learns nothing until the end of the trial.

▷ **Intuition:** Direct utility estimation searches for U in a hypothesis space that too large $\leftsquigarrow$ many functions that violate the Bellman equations.

▷ Thus the algorithm often converges very slowly.

Adaptive Dynamic Programming

▷ **Idea:** Take advantage of the constraints among the utilities of states by

- ▷ learning the transition model that connects them,
- ▷ solving the corresponding Markov decision process using a dynamic programming method.

This means plugging the learned transition model $P(s'|s,\pi(s))$ and the observed rewards $R(s)$ into the Bellman equations to calculate the utilities of the states.

▷ **As above:** These equations are linear (no maximization involved) (solve with any any linear algebra package).

▷ **Observation 29.2.12.** *Learning the model itself is easy, because the environment is fully observable.*

▷ **Corollary 29.2.13.** *We have a supervised learning task where the input is a state–action pair and the output is the resulting state.*

- ▷ *In the simplest case, we can represent the transition model as a table of probabilities.*
- ▷ *Count how often each action outcome occurs and estimate the transition probability $P(s'|s,a)$ from the frequency with which s' is reached by action a in s.*

▷ **Example 29.2.14.** In the 3 trials from Example 29.2.3, $Right$ is executed 3 times in $(1,3)$ and 2 times the result is $(2,3)$, so $P((2,3)|(1,3),Right)$ is estimated to be $2/3$.

Passive ADP Learning Algorithm

▷ **Definition 29.2.15.** The passive ADP algorithm is given by

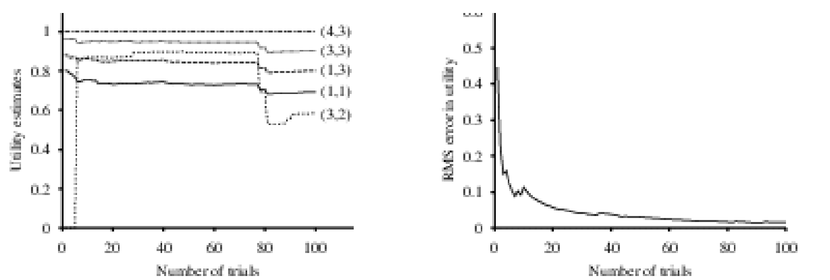
function PASSIVE−ADP−AGENT(percept) **returns** an action
  inputs: percept, a percept indicating the current state s' and reward signal r'
  persistent: π a fixed policy
    mdp, an MDP with model P, rewards R, discount γ
    U, a table of utilities, initially empty
    N_{sa}, a table of frequencies **for** state−action pairs, initially zero
    $N_{s'|sa}$, a table of outcome frequencies given state−action pairs, initially zero
    s, a, the previous state and action, initially null
  if s' is new **then** $U[s'] := r'$; $R[s'] := r'$
  if s is not null **then**
    increment $N_{sa}[s,a]$ and $N_{s'|sa}[s',s,a]$
    for each t such that $N_{s]|sa}[t,s,a]$ is nonzero **do**

$P(t|s,a) := N_{s'|sa}[t,s,a]/N_{sa}[s,a]$
U := POLICY−EVALUATION(π,mdp)
if s'.TERMINAL? **then** s,a := null **else** $s,a := s',\pi[s']$
return a

POLICY−EVALUATION computes $U^\pi(s) := E\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)\right]$ in a MDP.

Passive ADP Convergence

▷ **Example 29.2.16 (Passive ADP learning curves for the 4x3 world).** Given the optimal policy from Example 29.2.2



utility estimates/trials error for $U(1,1)$: 20 runs of 100 trials

Note the large changes occurring around the 78th trial – this is the first time that the agent falls into the -1 terminal state at (4,2).

▷ **Observation 29.2.17.** *The ADP agent is limited only by its ability to learn the transition model.* (*intractable for large state spaces*)

▷ **Example 29.2.18.** In backgammon, roughly 10^{50} equations in 10^{50} unknowns.

▷ **Idea:** Use this as a baseline to compare passive learning algorithms

29.3 Active Reinforcement Learning

Active Reinforcement Learning

▷ **Recap:** A passive learning agent has a fixed policy that determines its behavior.

▷ An active agent must also decide what actions to take.

▷ **Idea:** Adapt the passive ADP algorithm to handle this new freedom.

- ▷ learn a complete model with outcome probabilities for all actions, rather than just the model for the fixed policy. (use PASSIVE-ADP-AGENT)
- ▷ choose actions; the utilities to learn are defined by the optimal policy, they obey

the Bellman equation:

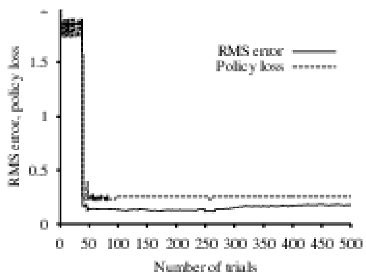
$$U(s) = R(s) + \gamma \cdot \max_{a \in A(s)} (\sum_{s'} U(s') \cdot P(s'|s,a))$$

- ▷ solve with value/policy iteration techniques from section 25.3.
- ▷ choose a good action, e.g.
  - ▷ by one-step lookahead to maximize expected utility, or
  - ▷ if agent uses policy iteration and has optimal policy, execute that.

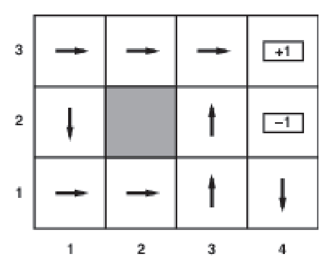
This agent/algorithm is greedy, since it only optimizes the next step!

Greedy ADP Learning (Evaluation)

▷ **Example 29.3.1 (Greedy ADP learning curves for the 4x3 world).**



average error/loss



suboptimal policy involved

The agent follows the optimal policy for the learned model at each step.

- ▷ It does not learn the true utilities or the true optimal policy!
- ▷ instead, in the 39th trial, it finds a policy that reaches the +1 reward along the lower route via (2,1), (3,1), (3,2), and (3,3).
- ▷ After experimenting with minor variations, from the 276th trial onward it sticks to that policy, never learning the utilities of the other states and never finding the optimal route via (1,2), (1,3), and (2,3).

Exploration in Active Reinforcement Learning

▷ **Observation 29.3.2.** *Greedy active ADP learning agents very seldom converge against the optimal solution*

- ▷ *The learned model is not the same as the true environment,*
- ▷ *What is optimal in the learned model need not be in the true environment.*

▷ What can be done? The agent does not know the true environment.

- **Idea:** Actions do more than provide rewards according to the learned model
  - they also contribute to learning the true model by affecting the percepts received.
  - By improving the model, the agent may reap greater rewards in the future.
- **Observation 29.3.3.** *An agent must make a tradeoff between*
  - *exploitation to maximize its reward as reflected in its current utility estimates and*
  - *exploration to maximize its long term well-being.*

  Pure exploitation risks getting stuck in a rut. Pure exploration to improve one's knowledge is of no use if one never puts that knowledge into practice.
- Compare with the information gathering agent from section 23.6.

Part VII

Natural Language

A Video Nugget covering this part can be found at `https://fau.tv/clip/id/35294`.

This part introduces the basics of natural language processing and the use of natural language for communication with humans.

Fascination of (Natural) Language

- **Definition 29.3.4.** A natural language is any form of spoken or signed means communication that has evolved naturally in humans through use and repetition without conscious planning or premeditation.
- **In other words:** the language you use all day long, e.g. English, German, ...
- **Why Should we care about natural language?:**
  - Even more so than thinking, language is a skill that only humans have.
  - It is a miracle that we can express complex thoughts in a sentence in a matter of seconds.
  - It is no less miraculous that a child can learn tens of thousands of words and a complex grammar in a matter of a few years.

Natural Language and AI

- Without natural language capabilities (understanding and generation) no AI!
- Ca. 100.000 years ago, humans learned to speak, ca. 7.000 years ago, to write.
- Alan Turing based his test on natural language: (for good reason)
  - We want AI agents to be able to communicate with humans.
  - We want AI agents to be able to acquire knowledge from written documents.
- In this part, we analyze the problem with specific information-seeking tasks:
  - Language models (Which strings are English/Spanish/etc.)
  - Text classification (E.g. spam detection)
  - Information retrieval (aka. Search Engines)
  - Information extraction (finding objects and their relations in texts)

Chapter 30

Natural Language Processing

30.1 Introduction to NLP

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/35295`.
The general context of AI-2 is natural language processing (NLP), and in particular natural language understanding (NLU). The dual side of NLU: natural language generation (NLG) requires similar foundations, but different techniques is less relevant for the purposes of this course.

What is Natural Language Processing?

- ▷ **Generally:** Studying of natural languages and development of systems that can use/generate these.
- ▷ **Definition 30.1.1.** Natural language processing (NLP) is an engineering field at the intersection of computer science, artificial intelligence, and linguistics which is concerned with the interactions between computers and human (natural) languages. Most challenges in NLP involve:
  - ▷ Natural language understanding (NLU) that is, enabling computers to derive meaning (representations) from human or natural language input.
  - ▷ Natural language generation (NLG) which aims at generating natural language or speech from meaning representation.
- ▷ For communication with/among humans we need both NLU and NLG.

Language Technology

- ▷ Language Assistance:
  - ▷ written language: Spell/grammar/style-checking,
  - ▷ spoken language: dictation systems and screen readers,
  - ▷ multilingual text: machine-supported text and dialog translation, eLearning.
- ▷ Information management:

- search and classification of documents, (e.g. Google/Bing)
- information extraction, question answering. (e.g. http://ask.com)

- Dialog Systems/Interfaces:
  - information systems: at airport, tele-banking, e-commerce, call centers,
  - dialog interfaces for computers, robots, cars. (e.g. Siri/Alexa)

- **Observation:** The earlier technologies largely rely on pattern matching, the latter ones need to compute the meaning of the input utterances, e.g. for database lookups in information systems.

30.2 Natural Language and its Meaning

A Video Nugget covering this section can be found at https://fau.tv/clip/id/35295.

Before we embark on the journey into understanding the meaning of natural language, let us get an overview over what the concept of “semantics” or “meaning” means in various disciplines.

What is (NL) Semantics? Answers from various Disciplines!

- **Observation:** Different (academic) disciplines specialize the notion of semantics (of natural language) in different ways.

- **Philosophy:** has a long history of trying to answer it, e.g.
  - Platon ⇝ cave allegory, Aristotle ⇝ Syllogisms.
  - Frege/Russell ⇝ sense vs. referent. (*Michael Kohlhase* vs. *Odysseus*)

- **Linguistics/Language Philosophy:** We need semantics e.g. in translation
  Der Geist ist willig aber das Fleisch ist schwach! vs.
  Der Schnaps ist gut, aber der Braten ist verkocht! (meaning counts)

- **Psychology/Cognition:** Semantics $\widehat{=}$ “what is in our brains” (⇝ mental models)

- **Mathematics** has driven much of modern logic in the quest for foundations.
  - Logic as “foundation of mathematics” solved as far as possible
  - In daily practice syntax and semantics are not differentiated (much).

- **Logic@AI/CS** tries to define meaning and compute with them. (applied semantics)
  - makes syntax explicit in a formal language (formulae, sentences)
  - defines truth/validity by mapping sentences into “world” (interpretation)
  - gives rules of truth-preserving reasoning (inference)

A good probe into the issues involved in natural language understanding is to look at translations between natural language utterances – a task that arguably involves understanding the utterances first.

Meaning of Natural Language; e.g. Machine Translation

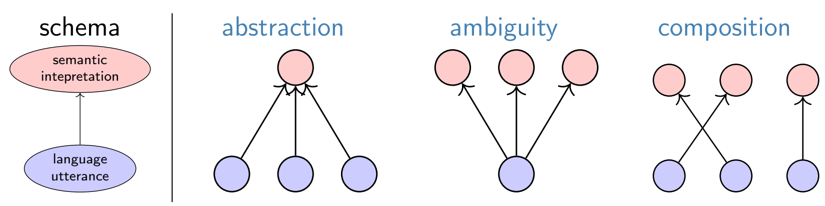
- **Idea:** Machine Translation is very simple! (we have good lexica)
- **Example 30.2.1.** *Peter liebt Maria.* ⇝ *Peter loves Mary.*
- ⚠ this only works for simple examples!
- **Example 30.2.2.** *Wirf der Kuh das Heu über den Zaun.* ⇝̸ *Throw the cow the hay over the fence.* (differing grammar; Google Translate)
- **Example 30.2.3.** ⚠ Grammar is not the only problem
  - *Der Geist ist willig, aber das Fleisch ist schwach!*
  - *Der Schnaps ist gut, aber der Braten ist verkocht!*
- **Observation 30.2.4.** *We have to understand the meaning for high-quality translation!*

If it is indeed the meaning of natural language, we should look further into how the form of the utterances and their meaning interact.

Language and Information

- **Observation:** Humans use words (sentences, texts) in natural languages to represent and communicate information.
- **But:** What really counts is not the words themselves, but the meaning information they carry.
- **Example 30.2.5 (Word Meaning).** *Newspaper* ⇝

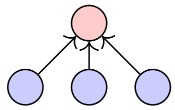
- For questions/answers, it would be very useful to find out what words (sentences/texts) mean.
- Interpretation of natural language utterances: three problems



Let us support the last claim a couple of initial examples. We will come back to these phenomena again and again over the course of the course and study them in detail.

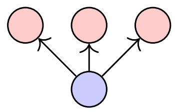
Language and Information (Examples)

▷ **Example 30.2.6 (Abstraction).**



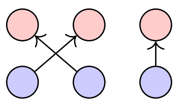
Car and *automobile* have the same meaning

▷ **Example 30.2.7 (Ambiguity).**



A *bank* can be a financial institution or a geographical feature

▷ **Example 30.2.8 (Composition).**



Every student sleeps $\leadsto \forall x.student(x) \Rightarrow sleep(x)$

But there are other phenomena that we need to take into account when compute the meaning of NL utterances.

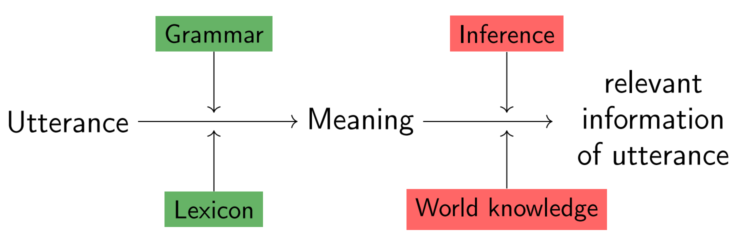
Context Contributes to the Meaning of NL Utterances

▷ **Observation:** Not all information conveyed is linguistically realized in an utterance.

▷ **Example 30.2.9.** *The lecture begins at 11:00 am.* What lecture? Today?

▷ **Definition 30.2.10.** We call a piece i of information linguistically realized in an utterance U, iff, we can trace i to a fragment of U.

▷ **Definition 30.2.11 (Possible Mechanism).** Inferring the missing pieces from the context and world knowledge:

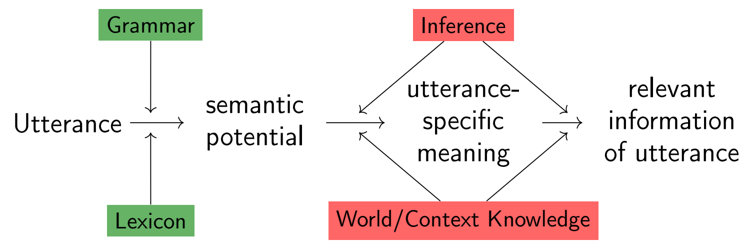


We call this process pragmatic analysis.

We will look at another example, that shows that the situation with pragmatic analysis is even more complex than we thought. Understanding this is one of the prime objectives of the AI-2 lecture.

Context Contributes to the Meaning of NL Utterances

- ▷ **Example 30.2.12.** *It starts at eleven.* What starts?
- ▷ Before we can resolve the time, we need to resolve the anaphor *it*.
- ▷ **Possible Mechanism:** More Inference!



⇝ Pragmatic analysis is quite complex! (prime topic of AI-2)

Example 30.2.12 is also a very good example for the claim Observation 30.2.4 that even for high-quality (machine) translation we need semantics. We end this very high-level introduction with a caveat.

Semantics is not a Cure-It-All!

How many animals of each species did Moses take onto the ark?

- ▷ Actually, it was Noah (But you understood the question anyways)

But Semantics works in some cases

- ▷ The only thing that currently really helps is a restricted domain:
  - ▷ I. e. a restricted vocabulary and world model.
- ▷ **Demo:**

  DBPedia `http://dbpedia.org/snorql/`
  Query: Soccer players, who are born in a country with more than 10 million inhabitants, who played as goalkeeper for a club that has a stadium with more than 30.000 seats and the club country is different from the birth country

But Semantics works in some cases

- ▷ **Answer:**

  (is computed by DBPedia from a SPARQL query)

```
SELECT distinct ?soccerplayer ?countryOfBirth ?team ?countryOfTeam ?stadiumcapacity
{
?soccerplayer a dbo:SoccerPlayer ;
   dbo:position|dbp:position <http://dbpedia.org/resource/Goalkeeper_(association_football)> ;
   dbo:birthPlace/dbo:country* ?countryOfBirth ;
   #dbo:number 13 ;
   dbo:team ?team .
   ?team dbo:capacity ?stadiumcapacity ; dbo:ground ?countryOfTeam .
   ?countryOfBirth a dbo:Country ; dbo:populationTotal ?population .
   ?countryOfTeam a dbo:Country .
FILTER (?countryOfTeam != ?countryOfBirth)
FILTER (?stadiumcapacity > 30000)
FILTER (?population > 10000000)
} order by ?soccerplayer
```

Results: Browse Go! Reset

SPARQL results:

soccerplayer	countryOfBirth	team	countryOfTeam	stadiumcapacity
:Abdesslam_Benabdellah	:Algeria	:Wydad_Casablanca	:Morocco	67000
:Airton_Moraes_Michellon	:Brazil	:FC_Red_Bull_Salzburg	:Austria	31000
:Alain_Gouaméné	:Ivory_Coast	:Raja_Casablanca	:Morocco	67000
:Allan_McGregor	:United_Kingdom	:Beşiktaş_J.K.	:Turkey	41903
:Anthony_Scribe	:France	:FC_Dinamo_Tbilisi	:Georgia_(country)	54549
:Brahim_Zaari	:Netherlands	:Raja_Casablanca	:Morocco	67000
:Bréiner_Castillo	:Colombia	:Deportivo_Táchira	:Venezuela	38755
:Carlos_Luis_Morales	:Ecuador	:Club_Atlético_Independiente	:Argentina	48069
:Carlos_Navarro_Montoya	:Colombia	:Club_Atlético_Independiente	:Argentina	48069
:Cristián_Muñoz	:Argentina	:Colo-Colo	:Chile	47000
:Daniel_Ferreyra	:Argentina	:FBC_Melgar	:Peru	60000
:David_Bičík	:Czech_Republic	:Karşıyaka_S.K.	:Turkey	51295
:David_Loria	:Kazakhstan	:Karşıyaka_S.K.	:Turkey	51295
:Denys_Boyko	:Ukraine	:Beşiktaş_J.K.	:Turkey	41903
:Eddie_Gustafsson	:United_States	:FC_Red_Bull_Salzburg	:Austria	31000
:Emilian_Dolha	:Romania	:Lech_Poznań	:Poland	43269
:Eusebio_Acasuzo	:Peru	:Club_Bolívar	:Bolivia	42000
:Faryd_Mondragón	:Colombia	:Real_Zaragoza	:Spain	34596
:Faryd_Mondragón	:Colombia	:Club_Atlético_Independiente	:Argentina	48069
:Federico_Vilar	:Argentina	:Club_Atlas	:Mexico	54500
:Fernando_Martinuzzi	:Argentina	:Real_Garcilaso	:Peru	45000
:Fábio_André_da_Silva	:Portugal	:Servette_FC	:Switzerland	30084
:Gerhard_Tremmel	:Germany	:FC_Red_Bull_Salzburg	:Austria	31000
:Gift_Muzadzi	:United_Kingdom	:Lech_Poznań	:Poland	43269
:Günay_Güvenç	:Germany	:Beşiktaş_J.K.	:Turkey	41903
:Hugo_Marques	:Portugal	:C.D._Primeiro_de_Agosto	:Angola	48500
:Héctor_Landazuri	:Colombia	:La_Paz_F.C.	:Bolivia	42000

Even if we can get a perfect grasp of the semantics (aka. meaning) of NL utterances, their structure and context dependency – we will try this in this lecture, but of course fail, since the issues are much too involved and complex for just one lecture – then we still cannot account for all the human mind does with language. But there is hope, for limited and well-understood domains, we can to amazing things. This is what this course tries to show, both in theory as well as in practice.

30.3 Looking at Natural Language

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/35296`.
The next step will be to make some observations about natural language and its meaning, so that we get an intuition of what problems we will have to overcome on the way to modeling natural language.

Fun with Diamonds (are they real?) [Dav67]

▷ **Example 30.3.1.** We study the truth conditions of adjectival complexes:

- ▷ *This is a diamond.* ($\models$ *diamond*)
- ▷ *This is a blue diamond.* ($\models$ *diamond*, $\models$ *blue*)
- ▷ *This is a big diamond.* ($\models$ *diamond*, $\not\models$ *big*)
- ▷ *This is a fake diamond.* ($\models \neg$*diamond*)
- ▷ *This is a fake blue diamond.* ($\models$ *blue*?, $\models$ *diamond*?)
- ▷ *Mary knows that this is a diamond.* ($\models$ *diamond*)
- ▷ *Mary believes that this is a diamond.* ($\not\models$ *diamond*)

Logical analysis vs. conceptual analysis: These examples — mostly borrowed from Davidson:tam67 — help us to see the difference between "'logical-analysis' and "'conceptual-analysis'.

We observed that from *This is a big diamond.* we cannot conclude *This is big*. Now consider the sentence *Jane is a beautiful dancer*. Similarly, it does not follow from this that Jane is beautiful, but only that she dances beautifully. Now, what it is to be beautiful or to be a beautiful dancer is a complicated matter. To say what these things are is a problem of conceptual analysis. The job of semantics is to uncover the logical form of these sentences. Semantics should tell us that the two sentences have the same logical forms; and ensure that these logical forms make the right predictions about the entailments and truth conditions of the sentences, specifically, that they don't entail that the object is big or that Jane is beautiful. But our semantics should provide a distinct logical form for sentences of the type: *This is a fake diamond.* From which it follows that the thing is fake, but not that it is a diamond.

Ambiguity: The dark side of Meaning

▷ **Definition 30.3.2.** We call an utterance ambiguous, iff it has multiple meanings, which we call readings.

▷ **Example 30.3.3.** All of the following sentences are ambiguous:

- ▷ *John went to the bank.* (river or financial?)
- ▷ *You should have seen the bull we got from the pope.* (three readings!)
- ▷ *I saw her duck.* (animal or action?)
- ▷ *John chased the gangster in the red sports car.* (three-way too!)

One way to think about the examples of ambiguity on the previous slide is that they illustrate a certain kind of indeterminacy in sentence meaning. But really what is indeterminate here is what sentence is represented by the physical realization (the written sentence or the phonetic string). The symbol *duck* just happens to be associated with two different things, the noun and the verb. Figuring out how to interpret the sentence is a matter of deciding which item to select. Similarly for the syntactic ambiguity represented by PP attachment. Once you, as interpreter, have selected one of the options, the interpretation is actually fixed. (This doesn't mean, by the way, that as an interpreter you necessarily do select a particular one of the options, just that you can.) **A brief digression:** Notice that this discussion is in part a discussion about compositionality, and gives us an idea of what a non-compositional account of meaning could look like. The Radical Pragmatic View is a non-compositional view: it allows the information content of a sentence to be fixed by something that has no linguistic reflex.

To help clarify what is meant by compositionality, let me just mention a couple of other ways in which a semantic account could fail to be compositional.

- Suppose your syntactic theory tells you that S has the structure $[a[bc]]$ but your semantics computes the meaning of S by first combining the meanings of a and b and then combining the result with the meaning of c. This is non-compositional.
- Recall the difference between:

  1. Jane knows that George was late.
  2. Jane believes that George was late.

  Sentence 1. entails that George was late; sentence 2. doesn't. We might try to account for this by saying that in the environment of the verb *believe*, a clause doesn't mean what it usually means, but something else instead. Then the clause *that George was late* is assumed to contribute different things to the informational content of different sentences. This is a non-compositional account.

Quantifiers, Scope and Context

- ▷ **Example 30.3.4.** *Every man loves a woman.* (Keira Knightley or his mother!)
- ▷ **Example 30.3.5.** *Every car has a radio.* (only one reading!)
- ▷ **Example 30.3.6.** *Some student in every course sleeps in every class at least some of the time.* (how many readings?)
- ▷ **Example 30.3.7.** *The president of the US is having an affair with an intern.* (2002 or 2000?)
- ▷ **Example 30.3.8.** *Everyone is here.* (who is everyone?)

Observation: If we look at the first sentence, then we see that it has two readings:

1. there is one woman who is loved by every man.
2. for each man there is one woman whom that man loves.

These correspond to distinct situations (or possible worlds) that make the sentence true.
Observation: For the second example we only get one reading: the analogue of 2. The reason for this lies not in the logical structure of the sentence, but in concepts involved. We interpret the meaning of the word *has* as the relation "has as physical part", which in our world carries a certain uniqueness condition: If a is a physical part of b, then it cannot be a physical part of c,

unless b is a physical part of c or vice versa. This makes the structurally possible analogue to 1. impossible in our world and we discard it.

Observation: In the examples above, we have seen that (in the worst case), we can have one reading for every ordering of the quantificational phrases in the sentence. So, in the third example, we have four of them, we would get $4! = 24$ readings. It should be clear from introspection that we (humans) do not entertain 12 readings when we understand and process this sentence. Our models should account for such effects as well.

Context and Interpretation: It appears that the last two sentences have different informational content on different occasions of use. Suppose I say *Everyone is here.* at the beginning of class. Then I mean that everyone who is meant to be in the class is here. Suppose I say it later in the day at a meeting; then I mean that everyone who is meant to be at the meeting is here. What shall we say about this? Here are three different kinds of solution:

Radical Semantic View On every occasion of use, the sentence literally means that everyone in the world is here, and so is strictly speaking false. An interpreter recognizes that the speaker has said something false, and uses general principles to figure out what the speaker actually meant.

Radical Pragmatic View What the semantics provides is in some sense incomplete. What the sentence means is determined in part by the context of utterance and the speaker's intentions. The differences in meaning are entirely due to extra-linguistic facts which have no linguistic reflex.

The Intermediate View The logical form of sentences with the quantifier *every* contains a slot for information which is contributed by the context. So extra-linguistic information is required to fix the meaning; but the contribution of this information is mediated by linguistic form.

More Context: Anaphora

- ▷ *John is a bachelor. His wife is very nice.* (Uh, what?, who?)
- ▷ *John likes his dog Spiff even though he bites him sometimes.* (who bites?)
- ▷ *John likes Spiff. Peter does too.* (what to does Peter do?)
- ▷ *John loves his wife. Peter does too.* (whom does Peter love?)
- ▷ *John loves golf, and Mary too.* (who does what?)

Context is Personal and keeps changing

- ▷ *The king of America is rich.* (true or false?)
- ▷ *The king of America isn't rich.* (false or true?)
- ▷ *If America had a king, the king of America would be rich.* (true or false!)
- ▷ *The king of Buganda is rich.* (Where is Buganda?)
- ▷ *. . . Joe Smith. . . The CEO of Westinghouse announced budget cuts.* (CEO=J.S.!)

30.4 Language Models

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/35200`.

Natural Languages vs. Formal Language

- **Recap:** A formal language is a set of strings.
- **Example 30.4.1.** Programming languages like `Java` or $\mathtt{C}^{++}$ are formal languages.
- *Remark 30.4.2.* Natural languages like English, German, or Spanish are not.
- **Example 30.4.3.** Let us look at concrete examples
  - *Not to be invited is sad!* (definitely English)
  - *To not be invited is sad!* (controversial)
- **Idea:** Let's be lenient, instead of a hard set, use a probability distribution.
- **Definition 30.4.4.** A (statistical) language model is a probability distribution over sequences of characters or words.
- **Idea:** Try to learn/derive language models from text corpora.
- **Definition 30.4.5.** A text corpus (or simply corpus; plural corpora) is a large and structured collection of natural language texts.
- **Definition 30.4.6.** In corpus linguistics, corpora are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific natural language.

N-gram Character Models

- Written text is composed of characters letters, digits, punctuation, and spaces.
- **Idea:** Let's study language models for sequences of characters.
- As for Markov processes, we write $P(\mathbf{c}_{1:N})$ for the probability of a character sequence $\mathtt{c_1 \ldots c_n}$ of length N.
- **Definition 30.4.7.** We call an character sequence of length n an n gram (unigram, bigram, trigram for $n = 1, 2, 3$).
- **Definition 30.4.8.** An n gram model is a Markov process of order $n - 1$.
- *Remark 30.4.9.* For a trigram model, $P(\mathsf{c}_i|\mathbf{c}_{1:i-1}) = P(\mathsf{c}_i|\mathsf{c}_{(i-2)}, \mathsf{c}_{(i-1)})$. Factoring with the chain rule and then using the Markov property, we obtain

$$P(\mathbf{c}_{1:N}) = \prod_{i=1}^{N} P(\mathsf{c}_i|\mathbf{c}_{1:i-1}) = \prod_{i=1}^{N} P(\mathsf{c}_i|\mathsf{c}_{(i-2)}, \mathsf{c}_{(i-1)})$$

▷ **Thus,** a trigram model for a language with 100 characters, $\mathbf{P}(c_i|\mathbf{c}_{i-2:i-1})$ has 1.000.000 entries. It can be estimated from a corpus with 10^7 characters.

Applications of N-Gram Models of Character Sequences

▷ What can we do with N gram models?

▷ **Definition 30.4.10.** The problem of language identification is given a text, determine the natural language it is written in.

▷ *Remark 30.4.11.* Current technology can classify even short texts like *Hello, world*, or *Wie geht es Dir* correctly with more than 99% accuracy.

▷ **One approach:** Build a trigram language model $\mathbf{P}(c_i|\mathbf{c}_{i-2:i-1},\ell)$ for each candidate language ℓ by counting trigrams in a ℓ-corpus.

Apply Bayes' rule and the Markov property to get the most likely language:

$$\begin{aligned}
\ell^* &= \underset{\ell}{\text{argmax}}\,(P(\ell|\mathbf{c}_{1:N})) \\
&= \underset{\ell}{\text{argmax}}\,(P(\ell)\cdot P(\mathbf{c}_{1:N}|\ell)) \\
&= \underset{\ell}{\text{argmax}}\,(P(\ell)\cdot \prod_{i=1}^{N} P(c_i|\mathbf{c}_{i-2:i-1},\ell))
\end{aligned}$$

The prior probability $P(\ell)$ can be estimated, it is not a critical factor, since the trigram language models are extremely sensitive.

Other Applications of Character N-Gram Models

▷ Spelling correction is a direct application of a single-language language model: Estimate the probability of a word and all off-by-one variants.

▷ **Definition 30.4.12.** Genre classification means deciding whether a text is a news story, a legal document, a scientific article, etc.

▷ *Remark 30.4.13.* While many features help make this classification, counts of punctuation and other character n-gram features go a long way [KNS97].

▷ **Definition 30.4.14.** Named entity recognition (NER) is the task of finding names of things in a document and deciding what class they belong to.

▷ **Example 30.4.15.** In *Mr. Sopersteen was prescribed aciphex.* NER should recognize that *Mr. Sopersteen* is the name of a person and *aciphex* is the name of a drug.

▷ *Remark 30.4.16.* Character-level language models are good for this task because they can associate the character sequence *ex* with a drug name and *steen* with a person name, and thereby identify words that they have never seen before.

N-Grams over Word Sequences

- **Idea:** n gram models apply to word sequences as well.
- **Problems:** The method works identically, but

1. There are many more words than characters. (100 vs. 10^5 in Englisch)
2. And what is a word anyways? (space/punctuation-delimited substrings?)
3. Data sparsity: we do not have enough data! For a language model| for (10^5) words in English, we have 10^{15} trigrams.
4. Most training corpora do not have all words.

Word N-Grams: Out-of-Vocab Words

- **Definition 30.4.17.** Out of vocabulary (OOV) words are unknown words that appear in the test corpus but not training corpus.
- *Remark 30.4.18.* OOV words are usually content words such as names and locations which contain information crucial to the success of NLP tasks.
- **Idea:** Model OOV words by

1. adding a new word token, e.g. <UNK> to the vocabulary,
2. in the training corpus, replacing the respective first occurrence of a previously unknown word by <UNK>,
3. counting n grams as usual, treating <UNK> as a regular word.

This trick can be refined if we have a word classifier, then use a new token per class, e.g. <EMAIL> or <NUM>.

What can Word N-Gram Models do?

- **Example 30.4.19 (Test n-grams).** Build unigram, bigram, and trigram language models over the words [RN03], randomly sample sequences from the models.

1. Unigram: *logical are as are confusion a may right tries agent goal the was . . .*
2. Bigram: *systems are very similar computational approach would be represented . . .*
3. Trigram: *planning and scheduling are integrated the success of naive bayes model . . .*

- **Clearly** there are differences, how can we measure them to evaluate the models?
- **Definition 30.4.20.** The perplexity of a sequence $\mathbf{c}_{1:N}$ is defined as

$$\text{Perplexity}(\mathbf{c}_{1:N}):=P(\mathbf{c}_{1:N})^{-(\frac{1}{N})}$$

▷ **Intuition:** The reciprocal of probability, normalized by sequence length.

▷ **Example 30.4.21.** For a language with n characters or words and a language model that predicts that all are equally likely, the perplexity of any sequence is n.

If some characters or words are more likely than others, and the model reflects that, then the perplexity of correct sequences will be less than n.

▷ **Example 30.4.22.** In Example 30.4.19, the perplexity was 891 for the unigram model, 142 for the bigram model and 91 for the trigram model.

30.5 Part of Speech Tagging

Language Models and Generalization

▷ **Recall:** n-grams can predict that a word sequence like *a black cat* is more likely than *cat black a*. (as trigram 1. appears 0.000014% in a corpus and 2. never)

▷ **Native Speakers However:** Will tell you that *a black cat* matches a familiar pattern: article-adjective-noun, while *cat black a* does not!

▷ **Example 30.5.1.** Consider *the fulvous kitten* a native speaker reasons that it

- ▷ follows the article-adjective-noun pattern
- ▷ *fulvous* ($\widehat{=}$ brownish yellow) ends in *ous* $\leadsto$ adjective

So by generalization this is (probably) correct English.

▷ **Observation:** The order syntactical categories of words plays a role in English!

▷ **Problem:** How can we compute them? (up next)

Part-of-Speech Tagging

▷ **Definition 30.5.2.** Part-of-speech tagging (also POS tagging, POST, or grammatical tagging is the process of marking up a word in corpus with tags (called POS tags) as corresponding to a particular part of speech (a category of words with similar syntactic properties) based on both its definition and its context.

▷ **Example 30.5.3.** A sentence tagged with POS tags from the the Penn Treebank: (see below)

From	the	start	,	it	took	a	person	with	great	qualities	to	succeed
IN	DT	NN	,	PRP	VBD	DT	NN	IN	JJ	NNS	TO	VB

1. *From* is tagged as a preposition (IN)
2. *the* as a determiner (DT)
3. . . .

▷ **Observation:** Even though POS tagging is uninteresting in its own right, it is useful as a first step in many NLP tasks.

▷ **Example 30.5.4.** In text-to-speech synthesis, a POS tag of "noun" for *record* helps determine the correct pronunciation (as opposed to the tag "verb")

The Penn Treebank POS tags

▷ **Example 30.5.5.** The following 45 POS tags are used by the Penn Treebank.

Tag	Word	Description	Tag	Word	Description
CC	*and*	Coordinating conjunction	PRP$	*your*	Possessive pronoun
CD	*three*	Cardinal number	RB	*quickly*	Adverb
DT	*the*	Determiner	RBR	*quicker*	Adverb, comparative
EX	*there*	Existential there	RBS	*quickest*	Adverb, superlative
FW	*per se*	Foreign word	RP	*off*	Particle
IN	*of*	Preposition	SYM	+	Symbol
JJ	*purple*	Adjective	TO	*to*	to
JJR	*better*	Adjective, comparative	UH	*eureka*	Interjection
JJS	*best*	Adjective, superlative	VB	*talk*	Verb, base form
LS	*1*	List item marker	VBD	*talked*	Verb, past tense
MD	*should*	Modal	VBG	*talking*	Verb, gerund
NN	*kitten*	Noun, singular or mass	VBN	*talked*	Verb, past participle
NNS	*kittens*	Noun, plural	VBP	*talk*	Verb, non-3rd-sing
NNP	*Ali*	Proper noun, singular	VBZ	*talks*	Verb, 3rd-sing
NNPS	*Fords*	Proper noun, plural	WDT	*which*	Wh-determiner
PDT	*all*	Predeterminer	WP	*who*	Wh-pronoun
POS	*'s*	Possessive ending	WP$	*whose*	Possessive wh-pronoun
PRP	*you*	Personal pronoun	WRB	*where*	Wh-adverb
$	$	Dollar sign	#	#	Pound sign
"	'	Left quote	"	'	Right quote
(	[	Left parenthesis	)	]	Right parenthesis
,	,	Comma	.	!	Sentence end
:	;	Mid-sentence punctuation			

Computing Part of Speech Tags

▷ **Idea:** Treat the POS tags in a sentence as state variables $\mathbf{C}_{1:n}$ in a HMM: the words are the evidence variables $\mathbf{W}_{1:n}$, use prediction for POS tagging.

▷ The HMM is a generative model that

- ▷ starts in the tag predicted by the prior probability (usually IN) (problematic!)
- ▷ and then, for each step makes two choices:
  - ▷ what word – e.g. *From* – should be emitted
  - ▷ what state – e.g. DT – should come next

▷ **This works, but** there are problems

- ▷ the HMM does not consider context other than the current state (Markov property)
- ▷ it does not have any idea what the sentence is trying to convey

▷ **Idea:** Use the Viterbi algorithm to find the most probable sequence of hidden states (POS tags)

▷ POS taggers based on the Viterbi algorithm can reach an F_1 score of up to 97%.

The Viterbi algorithm for POS tagging – Details

▷ We need a transition model $P(\mathsf{C}_t|\mathsf{C}_{t-1})$: the probability of one POS tag following another.

▷ **Example 30.5.6.** $P(\mathsf{C}_t = VB|\mathsf{C}_{t-1} = MD) = 0.8$ means that given a modal verb (e.g. *would*) the following word is a verb (e.g. *think*) with probability 0.8.

▷ **Question:** Where does the number 0.8 come from?

▷ **Answer:** From counts in the corpus – with appropriate smoothing!
There are 13124 instances of MD in the Penn Treebank and 10471 are followed by a VB.

▷ For the sensor model $P(\mathsf{W}_t = would|\mathsf{C}_t = MD) = 0.1$ means that if we choose a modal verb, we will choose *would* 10% of the time.

▷ These numbers also come from the corpus with appropriate smoothing.

▷ **Limitations:** HMM models only know about he transition and sensor models
In particular, we cannot take into account that e.g. words ending in *ous* are likely adjectives.

▷ We will see methods based on neural networks later.

30.6 Text Classification

Text Classification as a NLP Task

▷ **Problem:** Often we want to (ideally) automatically see who can best deal with a given document (e.g. e-mails in customer service)

▷ **Definition 30.6.1.** Given a set of categories the task of deciding which one a given document belongs to is called text classification or categorization.

▷ **Example 30.6.2.** Language identification and genre classification are examples of text classification.

▷ **Example 30.6.3.** Sentiment analysis – classifying a product review as positive or negative.

▷ **Example 30.6.4.** Spam detection – classifying an email message as spam or ham (i.e. non-spam).

Spam Detection

- **Definition 30.6.5.** Spam detection – classifying an email message as spam or ham (i.e. non-spam)
- **General Idea:** Use NLP/machine learning techniques to learn the categories.
- **Example 30.6.6.** We have lots of examples of spam/ham, e.g.

Spam (from my spam folder)	Ham (in my inbox)
Wholesale Fashion Watches -57% today. Designer watches for cheap …	The practical significance of hypertree width in identifying more …
You can buy ViagraFr$1.85 All Medications at unbeatable prices! …	Abstract: We will motivate the problem of social identity clustering: …
WE CAN TREAT ANYTHING YOU SUFFER FROM JUST TRUST US …	Good to see you my friend. Hey Peter, It was good to hear from you. …
Sta.rt earn*ing the salary yo,u d-eserve by o'btaining the prope,r crede'ntials!	PDS implies convexity of the resulting optimization problem (Kernel Ridge …

- **Specifically:** What are good features to classify e-mails by?
  - n-grams like *for cheap* and *You can buy* indicate spam(but also occur in ham)
  - character-level features: capitalization, punctuation (e.g. in *yo,u d-eserve*)
- **Note:** We have two complementary ways of talking about classification: (up next)
  - using language models
  - using machine learning

Spam Detection as Language Modeling

- **Idea:** Define two n-gram language models:

1. one for $\mathbf{P}(\text{Message}|\text{spam})$ by training on the spam folder
2. one for $\mathbf{P}(\text{Message}|\text{ham})$ by training on the inbox

Then we can classify a new message m with an application of Bayes' rule:

$$\underset{c\in\{\text{spam},\text{ham}\}}{\text{argmax}}(P(c|m)) \quad \underset{c\in\{\text{spam},\text{ham}\}}{\text{argmax}}(P(m|c)P(c))$$

where $P(c)$ is estimated just by counting the total number of spam and ham messages.

- This approach works well for spam detection, just as it did for language identification.

Spam Detection as Language Modeling

- **Idea:** Define two n-gram language models:
  1. one for $\mathbf{P}(\text{Message}|\text{spam})$ by training on the spam folder
  2. one for $\mathbf{P}(\text{Message}|\text{ham})$ by training on the inbox

  Then we can classify a new message m with an application of Bayes' rule:

$$\underset{c\in\{\text{spam},\text{ham}\}}{\text{argmax}}(P(c|m))\ \underset{c\in\{\text{spam},\text{ham}\}}{\text{argmax}}(P(m|c)P(c))$$

  where $P(c)$ is estimated just by counting the total number of spam and ham messages.
- This approach works well for spam detection, just as it did for language identification.

Classifier Success Measures: Precision, Recall, and F_1 score

- We need a way to measure success in classification tasks.
- **Definition 30.6.7.** Let $f_C\colon S\to\mathbb{B}$ be a binary classifier for a class $C\subseteq S$, then we call $a\in S$ with $f_C(a)=\mathsf{T}$ a false positive, iff $a\notin C$ and $f_C(a)=\mathsf{F}$ a false negative, iff $a\in C$. False positives and negatives are erros of f_C. True positives and negatives occur when f_C correctly indicates actual membership in S.
- **Definition 30.6.8.** The precision of f_C is defined as $\frac{\#(TP)}{\#(TP)+\#(FN)}$ and the recall is $\frac{\#(TP)}{\#(TP)+\#(FP)}$, where TP is the set of true positives and $FN/FP$the sets of false negatives and false positives of f_C.
- **Intuitively** these measure the rates of:
  - true positives in class C. (precision high, iff few false positives)
  - true positives in ${f_C}^{-1}(\mathsf{T})$. (recall high, iff few true positives forgotten)
- **Definition 30.6.9.** The F_1 score combines precision and recall into a single number: (harmonic mean)

$$2\frac{\text{precision}\cdot\text{recall}}{(\text{precision}+\text{recall})}$$

- **Observation:** Classifiers try to reach precision and recall $\rightsquigarrow$ F_1 score of 1.
  - if that is impossible, compromize on one $\rightsquigarrow$ F_β score . (application-dependent)
  - The F_β score generalizes the F_1 score by weighing the precision β times as important as recall.

30.7 Information Retrieval

A Video Nugget covering this section can be found at https://fau.tv/clip/id/35274.

Information Retrieval

▷ **Definition 30.7.1.** Information retrieval (IR) deals with the representation, organization, storage, and maintenance of information objects that provide users with easy access to the relevant information and satisfy their various information needs.

▷ **Definition 30.7.2.** An information need is an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need.

▷ **Definition 30.7.3.** An information object is medium that is mainly used for its information content.

▷ **Observation (Hjørland 1997):** Information need is closely related to relevance: If something is relevant for a person in relation to a given task, we might say that the person needs the information for that task.

▷ **Definition 30.7.4.** Relevance denotes how well an information object meets the information need of the user. Relevance may include concerns such as timeliness, authority or novelty of the object.

▷ **Observation:** We normally come in contact with IR in the form of web search.

▷ **Definition 30.7.5.** Web search is a fully automatic process that responds to a user query by returning a sorted document list relevant to the user requirements expressed in the query.

▷ **Example 30.7.6.** Google and Bing are web search engines, their query is a bag of words and documents are web pages, PDFs, images, videos, shopping portals.

Vector Space Models for IR

▷ **Idea:** For web search, we usually represent documents and queries as bags of words over a fixed vocabulary V. Given a query Q, we return all documents that are "similar".

▷ **Definition 30.7.7.** Given a vocabulary (a list) V of words, a word $w \in V$, and a document d, then we define the raw term frequency (often just called the term frequency) of w in d as the number of occurrences of w in d.

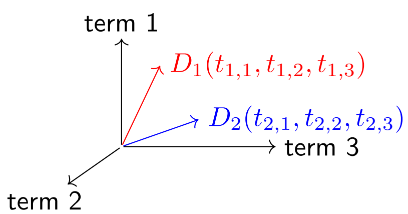
▷ **Definition 30.7.8.** A multiset of words in $V = \{t_1, \ldots, t_n\}$ is called a bag of words (BOW), and can be represented as a word frequency vectors in $\mathbb{N}^{|V|}$: the vector of raw word frequencies.

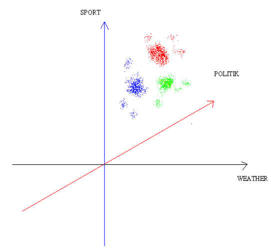
▷ **Example 30.7.9.** If we have two documents: $d_1 =$ *Have a good day!* and $d_2 =$ *Have a great day!*, then we can use $V =$ *Have, a, good, great, day* and can represent *good* as $\langle 0, 0, 1, 0, 0\rangle$, *great* as $\langle 0, 0, 0, 1, 0\rangle$, and d_1 a $\langle 1, 1, 1, 0, 1\rangle$.

Words outside the vocabulary are ignored in the BOW approach. So the document $d_3 =$ *What a day, a good day* is represented as $\langle 0, 2, 1, 0, 2\rangle$.

Vector Space Models for IR

- **Idea:** Query and document are similar, iff the angle between their word frequency vectors is small.





- **Lemma 30.7.10 (Euclidean Dot Product Formula).** $A \cdot B = \|A\|_2 \|B\|_2 \cos\theta$, *where θ is the angle between A and B.*
- **Definition 30.7.11.** The cosine similarity of A and B is $\cos\theta = \frac{A \cdot B}{\|A\|_2 \|B\|_2}$.

TF-IDF: Term Frequency/Inverse Document Frequency

- **Problem:** Word frequency vectors treat all the words equally.
- **Example 30.7.12.** In an query *the brown cow*, the *the* is less important than *brown cow*. (because *the* is less specific)
- **Idea:** Introduce a weighting factor for the word frequency vector that de-emphasizes the dimension of the more (globally) frequent words.
- We need to normalize the word frequency vectors first:
- **Definition 30.7.13.** Given a document d and a vocabulary word $t \in V$, the normalized term frequency (also usually called just term frequency) $\mathrm{tf}(t, d)$ is the raw term frequency divided by $|d|$.
- **Definition 30.7.14.** Given a document collection $D = \{d_1, \ldots, d_N\}$ and a word t the inverse document frequency is given by $\mathrm{idf}(t, D) := \log_{10}(\frac{N}{|\{d \in D \mid t \in d\}|})$.
- **Definition 30.7.15.** We define $\mathrm{tfidf}(t, d, D) := \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$.
- **Idea:** Use the tfidf-vector with cosine similarity for information retrieval instead.

TF-IDF Example

- ▷ Let $D{:=}\{d_1, d_2\}$ be a document corpus over the vocabulary

$$V = \{this, is, a, sample, another, example\}$$

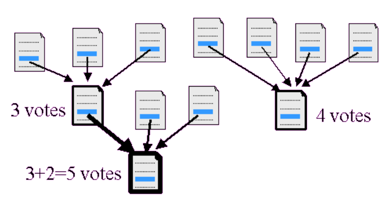
with word frequency vectors $\langle 1,1,1,2,0,0\rangle$ and $\langle 1,1,0,0,2,3\rangle$.

- ▷ Then we compute for the word *this*
  - ▷ $\mathrm{tf}(this, d_1) = \frac{1}{5} = 0.2$ and $\mathrm{tf}(this, d_2) = \frac{1}{7} \approx 0.14$,
  - ▷ idf is constant over D, we have $\mathrm{idf}(this, D) = \log_{10}(\frac{2}{2}) = 0$,
  - ▷ thus $\mathrm{tfidf}(this, d_1, D) = 0 = \mathrm{tfidf}(this, d_2, D)$. (*this* occurs in both)
- ▷ The word *example* is more interesting, since it occurs only in d_2 (thrice)
  - ▷ $\mathrm{tf}(example, d_1) = \frac{0}{5} = 0$ and $\mathrm{tf}(example, d_2) = \frac{3}{7} \approx 0.429$.
  - ▷ $\mathrm{idf}(example, D) = \log_{10}(\frac{2}{1}) \approx 0.301$,
  - ▷ thus $\mathrm{tfidf}(example, d_1, D) = 0 \cdot 0.301 = 0$ and $\mathrm{tfidf}(example, d_2, D) \approx 0.429 \cdot 0.301 = 0.129$.

Once an answer set has been determined, the results have to be sorted, so that they can be presented to the user. As the user has a limited attention span – users will look at most at three to eight results before refining a query, it is important to rank the results, so that the hits that contain information relevant to the user's information need early. This is a very difficult problem, as it involves guessing the intentions and information context of users, to which the search engine has no access.

Ranking Search Hits: e.g. Google's Page Rank

- ▷ **Problem:** There are many hits, need to sort them (e.g. by importance)
- ▷ **Idea:** A web site is important, ... if many other hyperlink to it.



- ▷ **Refinement:** ..., if many important web pages hyperlink to it.
- ▷ **Definition 30.7.16.** Let A be a web page that is hyperlinked from web pages $S_1, \ldots, S_n$, then the page rank PR of A is defined as

$$\mathrm{PR}(A) = 1 - d + d\left(\frac{\mathrm{PR}(S_1)}{C(S_1)} + \cdots + \frac{\mathrm{PR}(S_n)}{C(S_n)}\right)$$

where $C(W)$ is the number of links in a page W and $d = 0.85$.

- ▷ *Remark 30.7.17.* $\mathrm{PR}(A)$ is the probability of reaching A by random browsing.

Getting the ranking right is a determining factor for success of a search engine. In fact, the early of Google was based on the pagerank algorithm discussed above (and the fact that they figured out a revenue stream using text ads to monetize searches).

30.8 Information Extraction

Information Extraction

- **Definition 30.8.1.** Information extraction is the process of acquiring information by skimming a text and looking for occurrences of a particular class of object and for relationships among objects.
- **Example 30.8.2.** Extracting instances of addresses from web pages, with attributes for street, city, state, and zip code;
- **Example 30.8.3.** Extracting instances of storms from weather reports, with attributes for temperature, wind speed, and precipitation.
- **Observation:** In a limited domain, this can be done with high accuracy.

Attribute-Based Information Extraction

- **Definition 30.8.4.** In attribute-based information extraction we assume that the text refers to a single object and the task is to extract a factored representation.
- **Example 30.8.5 (Computer Prices).** Extracting from the text *IBM ThinkBook 970. Our price: $399.00* the attribute-based representation \{Manufacturer=IBM, Model=ThinkBook970,Price=$399.00\}.
- **Idea:** Try a template-based approach for each attribute.
- **Definition 30.8.6.** A template is a finite automaton that recognizes the information to be extracted. The template often consists of three sub-automata per attribute: the prefix pattern followed by the target pattern (it matches the attribute value) and the postfix pattern.
- **Example 30.8.7 (Extracing Prices with Regular Expressions).** When we want to extract computer price information, we could use regular expressions for the automata, concretely, the
  - prefix pattern: .∗price[:]?
  - target pattern: [$][0−9]+([.][0−9][0−9])?
  - postfix pattern: + shipping|
- **Alternative:** take all the target matches and choose among them.
- **Example 30.8.8.** For *List price $99.00, special sale price $78.00, shipping $3.00.* take the lowest price that is within 50% of the highest price. $\rightsquigarrow$ *$78.00*

Relational Information Extraction

- **Question:** Can we also do structured representations?
- **Answer:** That is the next step up from attribute-based information extraction.
- **Definition 30.8.9.** The task of a relational extraction system is to extract multiple objects and the relationships among them from a text.
- **Example 30.8.10.** When these systems see the text *$249.99,* they need to determine not just that it is a price, but also which object has that price.
- **Example 30.8.11.** FASTUS is a typical relational extraction system, which handles news stories about corporate mergers and acquisitions. It can read the story

  > Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

  and extract the relations:

$$e{\in}JointVentures \wedge Product(e,"golfclubs) \wedge Date(e,"Friday")$$

$$Member(e,"BridgestoneSportsCo") \wedge Member(e,"alocalconcern")$$

$$Member(e,"aJapanesetradinghouse")$$

30.9 Grammar

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/35581`.

Phrase Structure Grammars (Motivation)

- **Problem Recap:** We do not have enough text data to build word sequence language models ⇜ data sparsity.
- **Idea:** Categorize words into classes and then generalize "acceptable word sequences" into "acceptable word class sequences" ⇝ phrase structure grammars.
- **Advantage:** We can get by with much less information.
- **Example 30.9.1 (Generative Capacity).** 10^3 structural rules over a lexicon of 10^5 words generate most German sentences.
- Vervet monkeys, antelopes etc. use isolated symbols for sentences.
  ⇝ restricted set of communicable propositions, no generative capacity.
- **Disadvantage:** Grammars may over generalize or under generalize.
- The formal study of grammars was introduced by Noam Chomsky in 1957 [Cho65].

We fortify our intuition about these – admittedly very abstract – constructions by an example and introduce some more vocabulary.

Phrase Structure Grammars (cont.)

▷ **Example 30.9.2.** A simple phrase structure grammar G:

$$\begin{array}{rcl} S & \rightarrow & NP; Vi \\ NP & \rightarrow & Article; N \\ Article & \rightarrow & \textbf{the} \mid \textbf{a} \mid \textbf{an} \\ N & \rightarrow & \textbf{dog} \mid \textbf{teacher} \mid \ldots \\ Vi & \rightarrow & \textbf{sleeps} \mid \textbf{smells} \mid \ldots \end{array}$$

Here S, is the start symbol, NP, VP, $Article$, N, and Vi are nonterminals.

▷ **Definition 30.9.3.** The subset of lexical rules, i.e. those whose body consists of a single terminal is called its lexicon and the set of body symbols the alphabet. The nonterminals in their heads are called lexical categories.

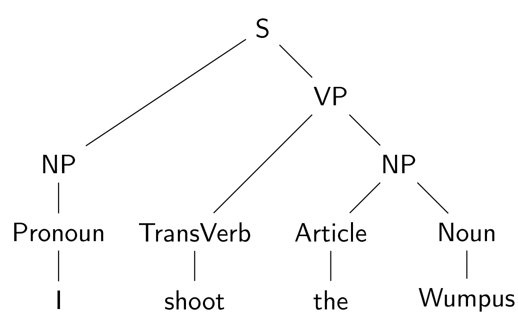
▷ **Definition 30.9.4.** The non-lexicon production rules are called structural, and the nonterminals in the heads are called phrasal categories.

Context-Free Parsing

▷ **Recall:** The sentences accepted by a grammar are defined "top-down" as those the start symbol can be rewritten into.

▷ **Definition 30.9.5.** Bottom up parsing works by replacing any substring that matches the body of a production rule with its head.

▷ **Example 30.9.6.** Using the Wumpus grammar (below), we get the following parse trees in bottom up parsing:



Traditional linear notation: Also write this as:

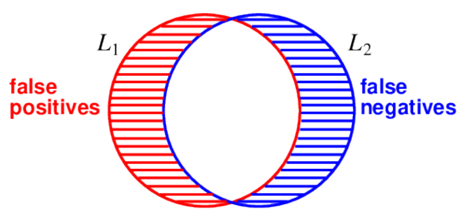
$[S[NP[Pronoun\ \mathbf{I}]][VP[TransVerb\ \mathbf{shoot}][NP[Article\ \mathbf{the}][Noun\ \mathbf{Wumpus}]]]]$

▷ Bottom up parsing algorithms tend to be more efficient than top-down ones.

- Efficient context-free parsing algorithms run in $\mathcal{O}(n^3)$, run at several thousand words/sec for real grammars.
- **Theorem 30.9.7.** *Context-free parsing $\widehat{=}$ Boolean matrix multiplication!*
- $\leadsto$ unlikely to find faster practical algorithms. (details in [Lee02])

Grammaticality Judgements

- **Problem:** The formal language L_G accepted by a grammar G may differ from the natural language L_n it supposedly models.
- **Definition 30.9.8.** We say that a grammar G over generates, iff it accepts strings outside of L_n (false positives) and under generates, iff there are L_n strings (false negatives) that L_G does not accept.



- Adjusting L_G to agree with L_N is a learning problem!
  - * *the gold grab the wumpus*
  - * *I smell the wumpus the gold*
  - *I give the wumpus the gold*
  - * *I donate the wumpus the gold*
- Intersubjective agreement somewhat reliable, independent of semantics!
- Real grammars (100–5000 rules) are insufficient even for "proper" English.

Probabilistic, Context-Free Grammars

- **Recall:** We introduced grammars as an efficient substitute for language models.
- **Problem (Poor Subsitute):** Grammars are deterministic language models.
- **Idea:** Add a probabilistic component to grammars.
- **Definition 30.9.9.** A probabilistic context-free grammar (PCFG) is a phrase structure grammar, where every production rule is associated with a probability.
- **Idea:** A PCFG induces a language model by assigning probabilities to its sentences.

▷ **Definition 30.9.10.** Let G be a PCFG, S a sentence of G, and D a G-derivation of S, then the probability of D is the product of the probabilites of the production rules in all steps of D. The probability of S is the sum of the probabilities of all its derivations.

Example: The Wumpus Grammar (Lexicon)

▷ **Example 30.9.11 (Wumpus Grammar Lexicon).**

$$\begin{array}{rcl}
\mathit{Noun} & \rightarrow & \textbf{stench}[.05] \mid \textbf{breeze}[.01] \mid \textbf{wumpus}[.15] \mid \textbf{pits}[.05] \mid \ldots \\
\mathit{Verb} & \rightarrow & \textbf{is}[.1] \mid \textbf{feel}[.1] \mid \textbf{smells}[.05] \mid \textbf{stinks}[.05] \mid \ldots \\
\mathit{TransVerb} & \rightarrow & \textbf{see}[.1] \mid \textbf{shoot}[.1] \mid \ldots \\
\mathit{Adjective} & \rightarrow & \textbf{right}[.1] \mid \textbf{dead}[.05] \mid \textbf{smelly}[.02] \mid \textbf{breezy}[.02] \mid \ldots \\
\mathit{Adverb} & \rightarrow & \textbf{here}[.05] \mid \textbf{ahead}[.05] \mid \textbf{nearby}[.02] \mid \ldots \\
\mathit{Pronoun} & \rightarrow & \textbf{me}[.1] \mid \textbf{you}[.03] \mid \textbf{I}[.1] \mid \textbf{it}[.1] \mid \ldots \\
\mathit{RelPron} & \rightarrow & \textbf{that}[.4] \mid \textbf{which}[.15] \mid \textbf{who}[.2] \mid \textbf{whom}[.02] \mid \ldots \\
\mathit{Name} & \rightarrow & \textbf{John}[.01] \mid \textbf{Mary}[.01] \mid \textbf{Boston}[.01] \mid \ldots \\
\mathit{Article} & \rightarrow & \textbf{the}[.4] \mid \textbf{a}[.3] \mid \textbf{an}[.1] \mid \textbf{every}[.05] \mid \ldots \\
\mathit{Preposition} & \rightarrow & \textbf{to}[.2] \mid \textbf{in}[.1] \mid \textbf{on}[.05] \mid \textbf{near}[.1] \mid \ldots \\
\mathit{Conjunction} & \rightarrow & \textbf{and}[.5] \mid \textbf{or}[.1] \mid \textbf{but}[.2] \mid \textbf{yet}[.2] \mid \ldots \\
\mathit{Digit} & \rightarrow & \textbf{0}[.2] \mid \textbf{1}[.2] \mid \textbf{2}[.2] \mid \textbf{3}[.2] \mid \textbf{4}[.2] \mid \textbf{5}[.2] \mid \ldots
\end{array}$$

Divided into closed and open classes

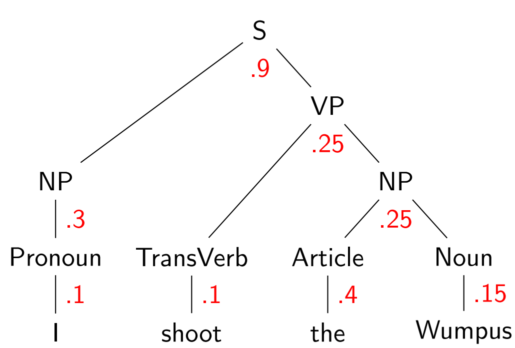
Wumpus grammar

▷

S	$\rightarrow$	$NP;VP$	[.9]	I + feel a breeze
	$\mid$	$S;Conjunction;S$	[.1]	I feel a breeze + and + I smell a wumpus
NP	$\rightarrow$	$Pronoun$	[.3]	I
	$\mid$	$Name$	[.1]	John
	$\mid$	$Noun$	[.1]	pits
	$\mid$	$Article;Noun$	[.25]	the + wumpus
	$\mid$	$Article;Adjs;Noun$	[.05]	the + smelly dead + wumpus
	$\mid$	$Digit;Digit$	[.05]	3 4
	$\mid$	$NP;PP$	[.1]	the wumpus + in 1 3
	$\mid$	$NP;RelClause$	[.05]	the wumpus + that is smelly
VP	$\rightarrow$	$Verb$	[.25]	stinks
	$\mid$	$TransVerb;NP$	[.25]	see + the Wumpus
	$\mid$	$VP;NP$	[.25]	feel + a breeze
	$\mid$	$VP;Adjective$	[.05]	is + smelly
	$\mid$	$VP;PP$	[.1]	turn + to the east
	$\mid$	$VP;Adverb$	[.1]	go + ahead
$Adjs$	$\rightarrow$	$Adjective$	[.8]	smelly
	$\mid$	$Adjective;Adjs$	[.2]	smelly + dead
PP	$\rightarrow$	$Prep;NP$	[1]	to + the east
$RelClause$	$\rightarrow$	**that**; VP	[1]	that + is smelly

PCFG Parsing

▷ **Example 30.9.12.** Reconsidering Example 30.9.6 with the Wumpus grammar above, we get the PCFG parse tree:



It has the probability $.9 \cdot .3 \cdot .1 \cdot .25 \cdot .1 \cdot .25 \cdot .4 \cdot .15 = 1.013 \times 10^{-5}$.

As this is the only derivation, this is also the probability of the sentence.

Learning PCFG Probabilities from Data

▷ **Recall:** A PCFG has many rules, each with a probability.

▷ **Problem:** Where do they come from?

▷ **Idea:** Learn/sample them from data. (but what data?)

▷ **Definition 30.9.13.** A treebank is a parsed text corpus that annotates syntactic or semantic sentence structure.

▷ **Idea:** To learn the probability for the rule $R{:=}H{\to}B$, look for parse subtrees that match R (i.e. with root category H) in the treebank.

If has N subtrees and n match R, then annotate R with probability n/N.

▷ Treebanks have revolutionized practical linguistics.

▷ There are many treebanks: `https://en.wikipedia.org/wiki/Treebank`

- ▷ multiple languages (from Abaza to Yoruba)
- ▷ multiple language types (newswire to spontaneous speech)

▷ **Outlook:** Treebank less grammar learning (rules and probabilitiess) is possible, but much more difficult.

The Penn Treebank

▷ **Definition 30.9.14.** The Penn treebank [MMS93] is a treebank of newswire texts (nearly 5 million words) annotated with part of speech and parse tree structure, using human labor assisted by some automated tools.

▷ **Example 30.9.15.** A tree from the Penn treebank for the sentence *Her eyes were glazed as if she didn't hear or even see him.*:

```
[ [S [NP—SBJ—2 Her eyes]
     [VP were
         [VP glazed
             [NP ∗—2]
             [SBAR—ADV as if
                 [S [NP—SBJ she]
                    [VP did n't
                        [VP [VP hear [NP ∗—1]]
                            or
                            [VP [ADVP even] see [NP ∗—1]]
                            [NP—1 him]]]]]]]]
 .]
```

Note: two S-rooted subtrees, one with NP—SBJ—2 child and one with NP SBJ.

Advertisement: Logic-Based Natural Language Semantics

▷ **Advanced Course:** "Logic-Based Natural Language Semantics" (next semester)

- ▷ Wed. 10:15-11:50 and Thu 12:15-13:50 (expected: ≤ 10 Students)

▷ **Contents:** (Alternating Lectures and hands-on Lab Sessions)

- ▷ Foundations of Natural Language Semantics (NLS)
- ▷ Montague's Method of Fragments (Grammar, Semantics Constr., Logic)
- ▷ Implementing Fragments in GLF (Grammatical Framework and MMT)
- ▷ Inference Systems for Natural Language Pragmatics (tableau machine)
- ▷ Advanced logical systems for NLS (modal, higher-order, dynamic Logics)

▷ **Grading:** Attendance & Wakefulness, Project/Homework, Oral Exam.

▷ **Course Intent:** Groom students for Bachelor/Master Theses and as KWARC research assistants.

Chapter 31

Deep Learning for NLP

Deep Learning for NLP: Agenda

▷ **Observation:** Symbolic and statistical systems have demonstrated success on many NLP tasks, but their performance is limited by the endless complexity of natural language.

▷ **Idea:** Given the vast amount of text in machine-readable form, can data-driven machine-learning base approaches do better?

▷ In this chapter, we explore this idea, using – and extending – the methods from Part VI.

▷ Overview:

1. Word embeddings
2. Recurrent neural networks for NLP
3. Sequence-to-sequence models
4. Transformer Architecture
5. Pretraining and transfer learning.

31.1 Word Embeddings

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/35276`.

Word Embeddings

▷ **Problem:** For ML methods in NLP, we need numerical data. (not words)

▷ **Idea:** Embed words or word sequences into real vector spaces.

▷ **Definition 31.1.1.** A word embedding is a mapping from words in context into a real vector space $\mathbb{R}^n$ used for natural language processing.

▷ **Definition 31.1.2.** A vector is called one hot, iff all components are 0 except for one 1. We call a word embedding one hot, iff all of its vectors are.

▷ **Example 31.1.3 (Vector Space Methods in Information Retrieval).** Word frequency vectors are induced by adding up one hot word embeddings.

▷ **Example 31.1.4.** Given a corpus D – the context – the tf idf word embedding is given by $e\colon t \mapsto \langle \text{tfidf}(t, d_1, D), \ldots, \text{tfidf}(t, d_{\#(D)}, D) \rangle$.

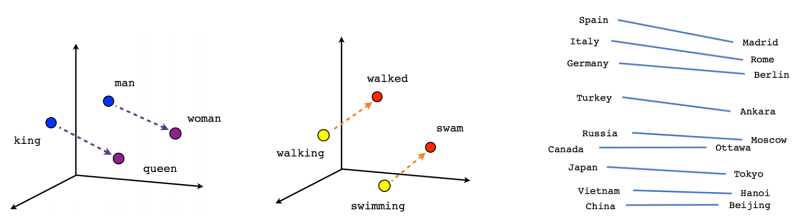
▷ **Intuition behind these two:** Words that occur in similar documents are similar.

Word2Vec: A Popular, Semantic Word Embedding

▷ **Distributional Semantics:** "a word is characterized by the company it keeps".

▷ **Idea:** Find word embeddings that take context into account.

▷ **Result Preview:** Semantic word embeddings



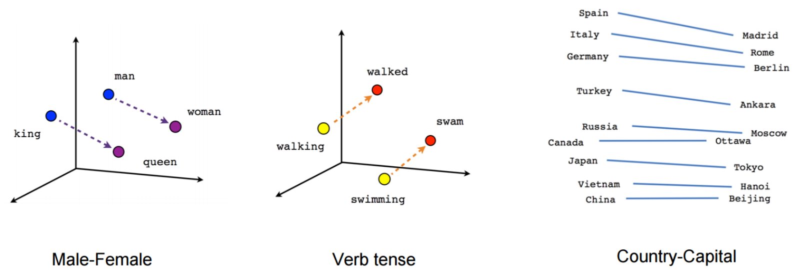
Male-Female | Verb tense | Country-Capital

- ▷ **as before**: Words that occur in similar documents are similar.
- ▷ **also in Word2vec**: Vector differences encode word relations.

▷ **Algorithm Preview:** Use a neural network to predict the word corresponding to an input context.

The Common Bag Of Words (CBOW) Algorithm I

▷ **Idea:** For the intended behavior



we need to maintain linear regularities, i.e. additive vector properties like:

$$V(King) - V(Man) + V(Queen) \quad \text{is close to} \quad V(Woman)$$

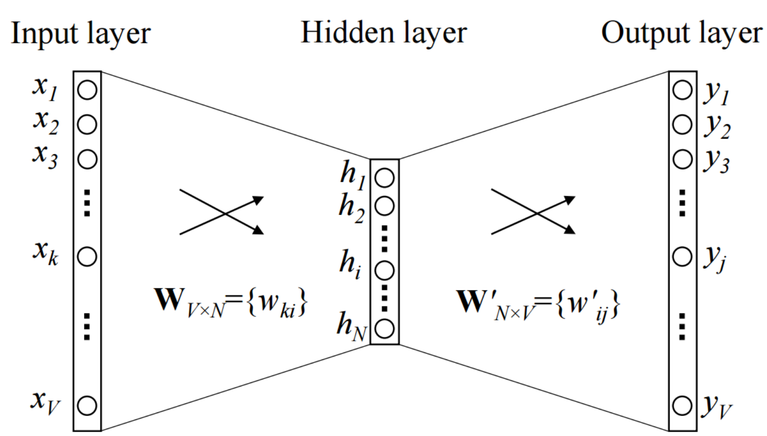
▷ **Example 31.1.5.** For the text *watch movies rather than read books* context size: 2, target *rather*, we have

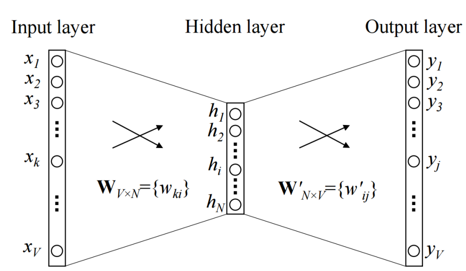
- ▷ context: $C{:=}\{watch, movies, than, read\}$
- ▷ Vocabulary: $V{:=}\{watch, movies, rather, than, read, books\}$

So in CBOW, build a neural network that
given the input $\{watch, movies, than, read\}$ produces *rather*.

The Common Bag Of Words (CBOW) Algorithm II

▷ A CBOW network for a single word





- ▷ The hidden layer neurons just copy the weighted sum of inputs to the next layer (no threshold)
- ▷ The output layer computes the softmax of the hidden nodes

Weighted sums and softmax maintain linear regularities. (as intended)

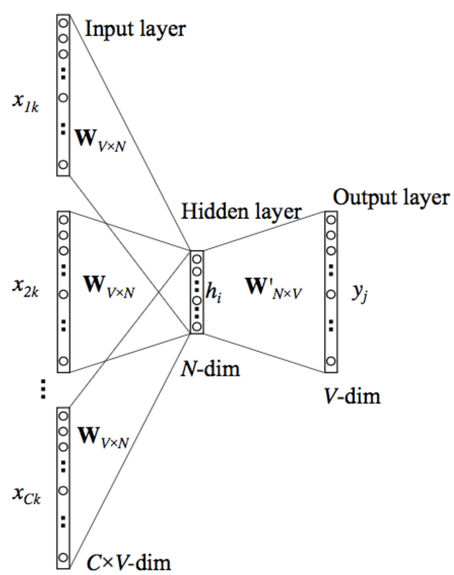
▷ **Definition 31.1.6.** The softmax function, (also normalized exponential function) $\sigma\colon \mathbb{R}\times\mathbb{R}^k\to\mathbb{R}^k$ is defined by

$$\sigma_\beta(z)_i := \frac{e^{\beta z_i}}{\sum_{j=1}^{k} e^{\beta z_j}}$$

where $z = \langle z_1, \ldots, z_k\rangle$, $i\in\{1,\ldots,k\}$. e^β is called the base of σ. For $\beta = 1$ we call σ the standard softmax function.

The Common Bag Of Words (CBOW) Algorithm III

▷ A neural network for a multiple words



Common Word Embeddings

▷ **Observation:** Word embeddings are crucial as first steps in any NN-based NLP methods.

▷ In practice it is often sufficient to use generic, pretrained word embeddings

▷ **Definition 31.1.7.** Common pretrained – i.e. trained for generic NLP applications word embeddings include

- ▷ Word2vec: the original system that established the concept (see above)
- ▷ GloVe (Global Vectors)
- ▷ FASTTEXT (embeddings for 157 languages)

▷ But we can also train our own word embedding (together with main task) (up next)

Learning POS tags and Word embeddings simultaneously

▷ Specific word embeddings are trained on a carefully selected corpus and tend to emphasize the characteristics of the task.

▷ **Example 31.1.8.** POS tagging – even though simple – is a good but non-trivial example.

Recall that many words can have multiple POS tags, e.g. *cut* can be

- ▷ a present-tense verb (transitive or intransitive)
- ▷ a past-tense verb
- ▷ a infinitive verb
- ▷ a past participle
- ▷ an adjective
- ▷ a noun.

If a nearby temporal adverb refers to the past $\leadsto$ this occurrence may be a past-tense verb.

The POS/Embedding: Setup

▷ We use the following process for learning a POS tagger including word embeddings from a POS tagged corpus:

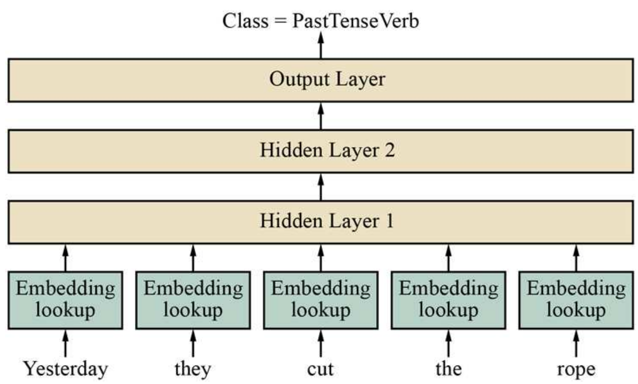
▷ **Example 31.1.9 (Preprocessing/Setup).**

1. Choose a width w – an odd number of words for the prediction window.
   A choice of $w = 5$ means that the tag is predicted with a context of 2 words before/after the word.
2. Split every sentence in the corpus into prediction windows for each word; together with the tags these constitute the training examples.
3. Create a vocabulary V of the unique, sufficiently common word tokens in the corpus $v{:=}\#(V)$.
4. Sort V in any order
5. Choose a value d for the size of the word embedding vector.

6. Create a new $v \times d$ matrix $\mathbf{E}$; the word embedding matrix has the word embedding of the ith word in V in row.
7. Initialize $\mathbf{E}$ randomly (or from pretrained vectors).

The POS/Embedding: Network

▷ **Example 31.1.10.** The POS/Embedding network has the following form/layers:



1. The input to the network is a word sequence of length w. (here 5; but see below)
2. The first layer consists of w copies $\mathbf{E}$.
3. The second (hidden) layer z_1 with weight matrix $\mathbf{W}_1$ computes $z_1 = l(\mathbf{W}_1 x)$
4. The third (hidden) layer z_2 with weight matrix $\mathbf{W}_2$ computes $z_2 = l(\mathbf{W}_2 z_1)$
5. The output layer computes the probability distribution $y = \sigma(\mathbf{W}_{out} z_2)$ over the possible POS tags for the middle word with a weight matrix $\mathbf{W}_{out}$.

The POS/Embedding: Computation

▷ To encode a word sequence $\mathbf{w}$ concatenate the encodings of each word $\leadsto$ input vector x of length wd.

▷ **Problem:** Every word in $\mathbf{w}$ will have the same encoding irrespective of its place in $\mathbf{w}$.

▷ **Answer:** But it will be treated differently in the first hidden layer (by a different part).

▷ Train the weights in $\mathbf{E}$, $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_{out}$ using gradient descent.

▷ **Example 31.1.11.** If all goes well in training, *cut* will be labeld as a past-tense verb given the context which includes the temporal past word *yesterday*.

31.2 Word Embeddings

Recurrent Neural Networks in NLP

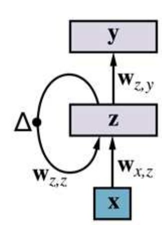
- word embeddings give a good representation of words in isolation.
- But natural language of word sequences ⇝ surrounding words provide context!
- For simple tasks like POS tagging, a fixed-size window of e.g. 5 words is sufficient.
- **Observation:** For advanced tasks like question answering we need more context!
- **Example 31.2.1.** In the sentence *Eduardo told me that Miguel was very sick so I took him to the hospital*, the pronoun *him* refers to *Miguel* and not *Eduardo*. (14 words of context)
- **Observation:** Language models with n-grams or n-word feed-forward networks have problems:
  Either the context is too small or the model has too many parameters! (or both)
- **Observation:** Feed-forward networks N also have the problem of asymmetry: whatever N learns about a word w at position n, it has to relearn about w at position $m \neq n$.
- **Idea:** What about recurrent neural networks – nets with cycles? (up next)

RNNs for Time Series

- **Idea:** RNNs – neural networks with cycles – have memory
  ⇝ use that for more context in neural NLP.
- **Example 31.2.2 (A simple RNN).**
  It has an input layer $\mathbf{x}$, a hidden layer $\mathbf{z}$ with recurrent connections and delay Δ, and an output layer $\mathbf{y}$ as shown on the right.
  Defining Equations for time step t:

$$\begin{aligned} \mathbf{z}_t &= \mathbf{g}_\mathbf{z}(\mathbf{W}_{\mathbf{z},\mathbf{z}}\mathbf{z}_{t-1} + \mathbf{W}_{\mathbf{x},\mathbf{z}}\mathbf{x}_t) \\ \mathbf{y}_t &= \mathbf{g}_\mathbf{y}(\mathbf{W}_{\mathbf{z},\mathbf{y}}\mathbf{z}_t) \end{aligned}$$

  where $\mathbf{g}_\mathbf{z}$ and $\mathbf{g}_\mathbf{y}$ are the activation functions for the hidden and output layers.



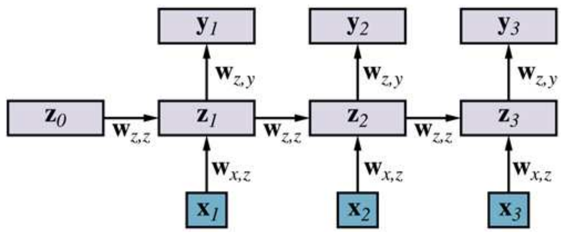
- **Intuition:** RNNs are a bit like HMMs and dynamic Bayesian Networks:
  They make a Markov assumption: the hidden state $\mathbf{z}$ suffices to capture the input from all previous inputs.

▷ **Side Benefit:** RNNs solve the asymmetry problem ⇝, the $\mathbf{W}_{\mathbf{z},\mathbf{z}}$ are the same at every step.

Training RNNs for NLP

▷ **Idea:** For training, unroll a RNN into a feed-forward network ⇝ back-propagation.

▷ **Example 31.2.3.** The RNN from Example 31.2.2 unrolled three times.



Problem: The weight matrices $\mathbf{W}_{\mathbf{x},\mathbf{z}}$, $\mathbf{W}_{\mathbf{z},\mathbf{z}}$, and $\mathbf{W}_{\mathbf{z},\mathbf{y}}$ are shared over all time slides.

▷ **Definition 31.2.4.** The back-propagation through time algorithm carefully maintains the identity of $\mathbf{W}_{\mathbf{z},\mathbf{z}}$ over all steps

Bidirectional RNN for more Context

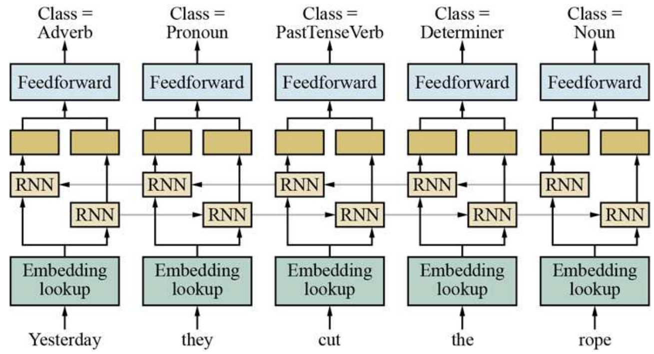
▷ **Observation:** RNNs only take "left context" – i.e. words before – into account, but we may also need "right context".

▷ **Example 31.2.5.** For *Eduardo told me that Miguel was very sick so I took him to the hospital* the pronoun *him* resolves to *Miguel* with high probability.

If the sentence ended with *to see Miguel*, then it should be *Eduardo*.

▷ **Definition 31.2.6.** A bidirectional RNN concatenates a separate right-to-left model onto a left-to-right model

▷ **Example 31.2.7.** Bidirectional RNNs can be used for POS tagging, extending the network from Example 31.1.10



Long Short-Term Memory RNNs

- **Problem:** When training a vanilla RNN using back-propagation through time, the long-term gradients which are back-propagated can "vanish" – tend to zero – or "explode" – tend to infinity.
- **Definition 31.2.8.** LSTMs provide a short-term memory for RNN that can last thousands of time steps, thus the name "long short-term memory". A LSTM can learn when to remember and when to forget pertinent information,
- **Example 31.2.9.** In NLP LSTMs can learn grammatical dependencies.

  An LSTM might process the sentence *Dave, as a result of his controversial claims, is now a pariah* by
  - remembering the (statistically likely) grammatical gender and number of the subject *Dave*,
  - note that this information is pertinent for the pronoun *his* and
  - note that this information is no longer important after the verb *is*.

31.3 Sequence-to-Sequence Models

Neural Machine Translation

- **Question:** Machine translation (MT) is an important task in NLP, can we do it with neural networks?
- **Observation:** If there were a one-to-one correspondence between source words and target words MT would be a simple tagging task. But
  - the three Spanish words *caballo de mar* translate to the English *seahorse* and
  - the two Spanish words *perro grande* translate to English as *big dog*.
  - in English, the subject is usually first and in Fijian last.

▷ **Idea:** For MT, generate one word at a time, but keep track of the context, so that

- ▹ we can remember parts of the source we have not translated yet
- ▹ we remember what we already translated so we do not repeat ourselves.

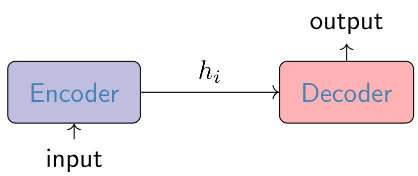
We may have to process the whole source sentence before generating the target!

▷ **Remark:** This smells like we need LSTMs.

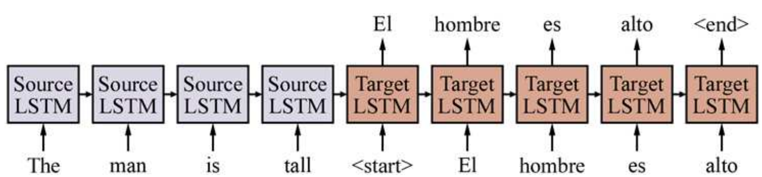
Sequence-To-Sequence Models

▷ **Idea:** Use two coupled RNNs, one for the source, and one for the target.

▷ **Definition 31.3.1.** A sequence-to-sequence (seq2seq) model is a neural model for translating an input sequence x into an output sequence y by an encoder followed by a decoder generates y.



▷ **Example 31.3.2.** A simple seq2seq model (without embedding and outut layers)



Each block represents one LSTM time step; inputs are fed successively followed by the token `<start>` to start the decoder.

Seq2Seq Evaluation

▷ **Remark:** Seq2seq models were a major breakthrough in NLP and MT. But they have three major shortcomings:

- ▹ **nearby context bias**: RNNs remember with their hidden state, which has more information about a word in – say – step 56 than in step 5. BUT long-distance context can also be important.
- ▹ **fixed context size**: the entire information about the source sentence must be compressed into the fixed-dimensional – typically 1024 – vector. Larger vectors ⇝ slow training and overfitting.

▷ **Idea:** Concatenate all source RNN hidden vectors to use all of them to mitigate the nearby context bias.

▷ **Problem:** Huge increase of weights ⇝ slow training and overfitting.

Attention

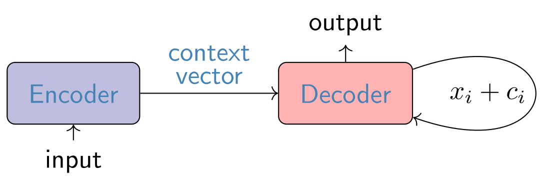
▷ **Bad Idea:** Concatenate all source RNN hidden vectors to use all of them to mitigate the nearby context bias.

▷ **Better Idea:** The decoder generates the target sequence one word at a time. ⇝ Only a small part of the source is actually relevant.
the decoder must focus on different parts of the source for every word.

▷ **Idea:** We need a neural component that does context-free summarization.

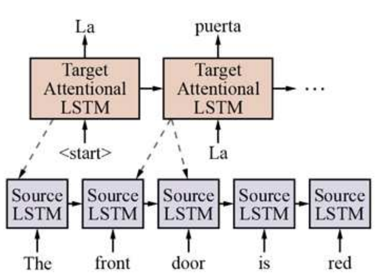
▷ **Definition 31.3.3.** An attentional seq2seq model is a seq2seq that passes along a context vector c_i in the decoder. If $h_i = RNN(h_{i-1}, x_i)$ is the standard decoder, then the decoder with attention is given by $h_i = RNN(h_{i-1}, x_i + c_i)$, where $x_i + c_i$ is the concatenation of the input x_i and context vectors c_i with

$$\begin{array}{rcll} r_{ij} & = & h_{i-1} \cdot s_j & \text{raw attention score} \\ a_{ij} & = & e^{r_{ij}} / (\sum_k e^{r_{ij}}) & \text{attention probability matrix} \\ c_i & = & \sum_j a_{ij} \cdot s_j & \text{context vector} \end{array}$$

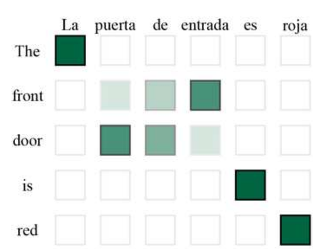


Attention: English to Spanish Translation

▷ **Example 31.3.4.** An attentional seq2seq model for English-to-Spanish translation



dashed lines represent attention



attention probablity matrix
darker colors ⇝ higher probabilities

▷ **Remarks:** The attention

- ▷ component learns no weights and supports variable-length sequences.
- ▷ is entirely latent – the developer does not influence it.

Attention: Greedy Decoding

▷ During training, a seq2seq model tries to maximize the probability of each word in the training sequence, conditioned on the source and the previous target words.

▷ **Definition 31.3.5.** The procedure that generates the target one word at a time and feeds it back at the next time step is called decoding.

▷ **Definition 31.3.6.** Always selecting the highest probability word is called greedy decoding.

▷ **Problem:** This may not always maximize the probability of the whole sequence

▷ **Example 31.3.7.** Let's use a greedy decoder on *The front door is red*.

- ▷ The correct translation is *La puerta de entrada es roja*.
- ▷ suppose we have generated the first word *La* for *The*.
- ▷ A greedy decoder might propose *entrada* for *front*.

▷ Greedy decoding is fast, but has no mechanism for correcting mistakes.

▷ **Solution:** Use an optimizing search algorithm (e.g. beam search)

Decoding with Beam Search

▷ **Recall:** Greedy decoding is not optimal!

▷ **Idea:** Search for an optimal decoding (or at least a good one) using one of the search algorithms from chapter 6.

▷ Local beam search is a common choice in machine translation. Concretely:

- ▷ keep the top k hypotheses at each stage,
- ▷ extending each by one word using the top k choices of words,
- ▷ then chooses the best k of the resulting k^2 new hypotheses.

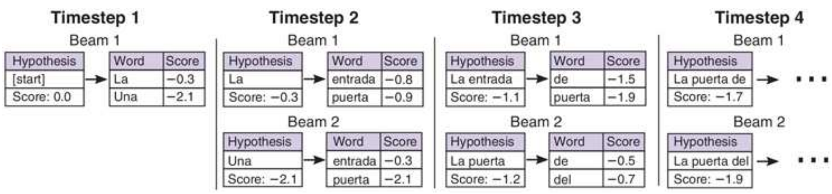
When all hypotheses in the beam generate the special <end> token, the algorithm outputs the highest scoring hypothesis.

▷ **Observation:** The better the seq2seq models get, the smaller we can keep beam size

Today beams of $b = 4$ are sufficient after $b = 100$ a decade ago.

Decoding with Beam Search

▷ **Example 31.3.8.** A local beam search with beam size $b = 2$



- ▷ Word scores are log-probabilities generated by the decoder softmax
- ▷ hypothesis score is the sum of the word scores.

At time step 3, the highest scoring hypothesis *La entrada* can only generate low-probability continuations, so it "falls off the beam". (as intended)

31.4 The Transformer Architecture

Self-Attention

▷ **Idea:** "Attention is all you need!" (see [Vas+17])

▷ So far, attention was used from the encoder to the decoder.

▷ Self-attention extends this so that each hidden states sequence also attends to itself. (*coder to *coder)

▷ **Idea:** Just use the dot product of the input vectors

▷ **Problem:** Always high, so each hidden state will be biased towards attending to itself.

▷ Self-attention solves this by first projecting the input into three different representations using three different weight matrices:

- ▷ the query vector $\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i \widehat{=}$ standard attention
- ▷ key vector $\mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i \widehat{=}$ the source in seq2seq
- ▷ value vector $\mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i$ is the context being generated

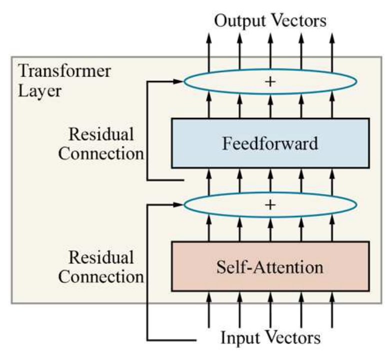
$$
\begin{aligned}
r_{ij} &= (\mathbf{q}_i \cdot \mathbf{k}_i)/\sqrt{d} \\
a_{ij} &= e^{r_{ij}}/(\sum_k e^{r_{ij}}) \\
c_i &= \sum_j a_{ij} \cdot \mathbf{v}_j
\end{aligned}
$$

where d is the dimension of $\mathbf{k}$ and $\mathbf{q}$.

The Transformer Architecture

- **Definition 31.4.1.** The transformer architecture uses neural blocks called transformers, which are built up from multiple transformer layers.
- **Remark:** The context modeled in self-attention is agnostic to word order ⇝ transformers use positional embedding to cope with that.
- **Example 31.4.2.**

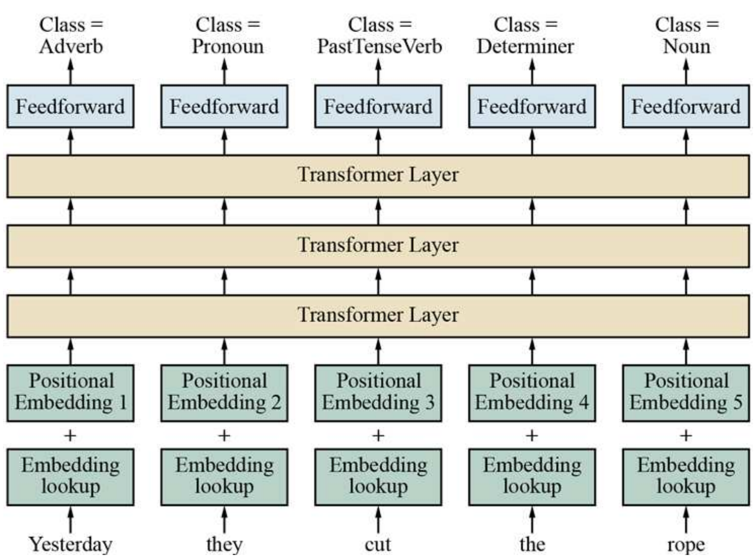
A single-layer transformer consists of self-attention, a feed-forward network, and residual connections to cope with the vanishing gradient problem.



- In practice transformers consist of 6-7 transformer layers.

A Transformer for POS tagging

- **Example 31.4.3.** A transformers for POS tagging:



31.5 Pretraining and Transfer Learning

Pretraining and Transfer Learning

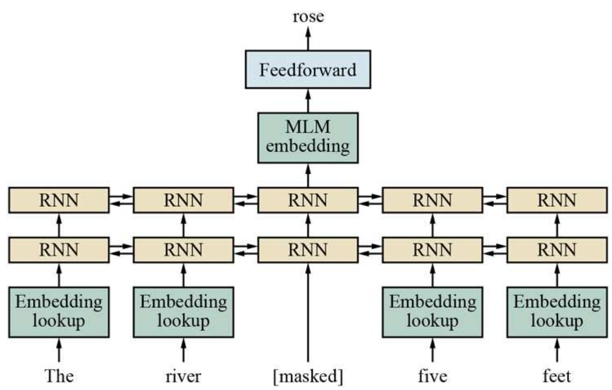
- Getting enough data to build a robust model can be a challenge.
- In NLP we often work with unlabeled data
  - syntactic/semantic labeling is much more difficult $\rightsquigarrow$ costly than image labeling.
  - the Internet has lots of texts (adds $\sim 10^{11}$ words/day)
- **Idea:** Why not let other's do this work and re-use their training efforts.
- **Definition 31.5.1.** In pretraining we use
  - a large amount of shared general-domain language data to train an initial version of an NLP model.
  - a smaller amount of domain-specific data (perhaps labeled) to refine it to the vocabulary, idioms, syntactic structures, and other linguistic phenomena that are specific to the new domain.
- Pretraining is a form of transfer learning:
- **Definition 31.5.2.** In Transfer learning (TL), knowledge learned from a task is re-used in order to boost performance on a related task.
- **Idea:** Take a pretrained neural network and randomly overwrite the weights in chosen layers, and then train on your own corpus.
- **Observation:** Simple but surprisingly effective!

Masked Language Models

- **Recall:** Standard language models such as n-gram models is that the contextualization of each word is based only on the previous words of the sentence. Predictions are made from left to right.
- But sometimes context from later in a sentence helps to clarify earlier words.
- **Example 31.5.3.** *feet* in the phrase *rose five feet* makes the "flower reading" for *rose* less likely.
- **Definition 31.5.4.** A masked language model (MLM) is trained by masking (hiding) a word in the input and asking the model to predict it.
- **Idea:** Use a single sentence with different masks in MLM training.
- **Remark:** MLM does not need labels $\leftrightsquigarrow$, the rest sentence acts as one
- **Observation:** If MLM are trained large corpora, they generate pretrained representations that perform well across a wide variety of NLP tasks (machine translation, question answering, summarization, grammaticality judgments, and others).

A Masked Language Model

▷ **Example 31.5.5.** We mask the middle word in *The river rose five feet* to obtain *The river five feet*.



Deep Learning for NLP: Evaluation

- ▷ Deep learning methods are currently dominant in NLP! (think ChatGPT)
  - ▷ Data-driven methods are easier to develop and maintain than symbolic ones
  - ▷ also perform better models crafted by humans (with reasonable effort)
- ▷ But problems remain;
  - ▷ DL methods work best on immense amounts of data. (small languages?)
  - ▷ LLM contain knowledge, but integration with symbolic methods elusive.
- ▷ **Question:** Why did we learn statistical NLP if it is obsoleted by DL methods?
- ▷ **Answer:** We do not know how the future will go! It may be that
  - ▷ DL-methods learn the latent characteristics embodied in symbolic models
  - ▷ breakthroughs in grammatical/semantic models reverse the pendulum
  - ▷ combinations of both will eventually do best (first examples exist)
- ▷ DL4NLP methods do very well, but only after processing orders of magnitude more data than humans do for learning language.
- ▷ This suggests that there is of scope for new insigths from all areas.

Chapter 32

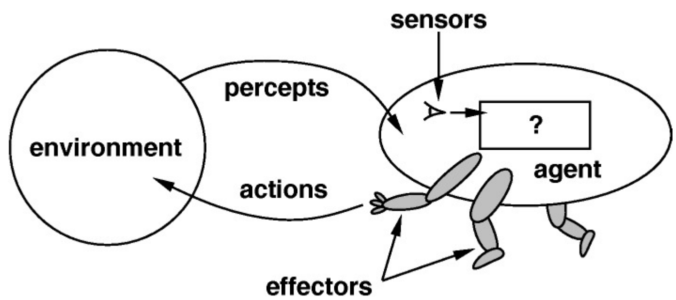
What did we learn in AI 1/2?

Topics of AI-1 (Winter Semester)

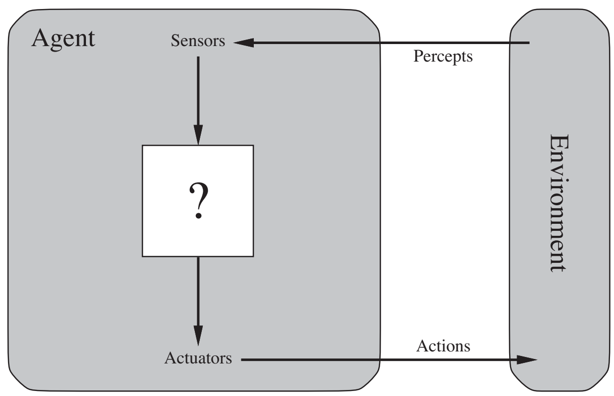
- ▷ Getting Started
  - ▷ What is Artificial Intelligence? (situating ourselves)
  - ▷ Logic programming in Prolog (An influential paradigm)
  - ▷ Intelligent Agents (a unifying framework)
- ▷ Problem Solving
  - ▷ Problem Solving and search (Black Box World States and Actions)
  - ▷ Adversarial Search (Game playing) (A nice application of Search)
  - ▷ constraint satisfaction problems (Factored World States)
- ▷ Knowledge and Reasoning
  - ▷ Formal Logic as the mathematics of Meaning
  - ▷ Propositional logic and satisfiability (Atomic Propositions)
  - ▷ First-order logic and theorem proving (Quantification)
  - ▷ Logic programming (Logic + Search $\leadsto$ Programming)
  - ▷ Description logics and semantic web
- ▷ Planning
  - ▷ Planning Frameworks
  - ▷ Planning Algorithms
  - ▷ Planning and Acting in the real world

Rational Agents as an Evaluation Framework for AI

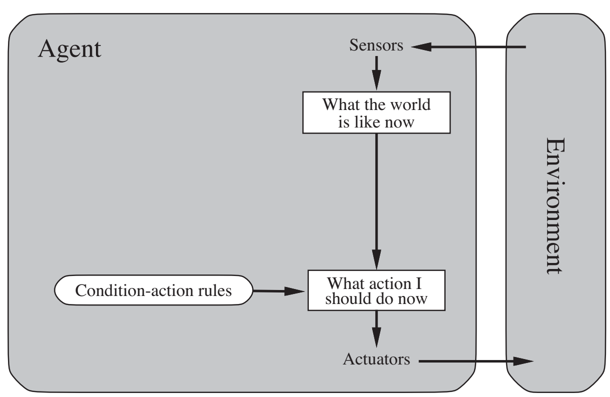
- ▷ Agents interact with the environment



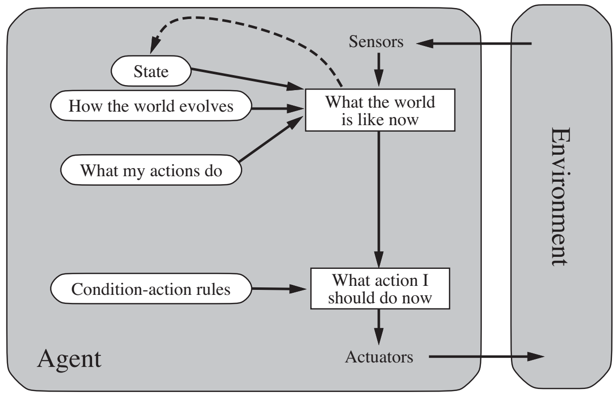
General agent schema



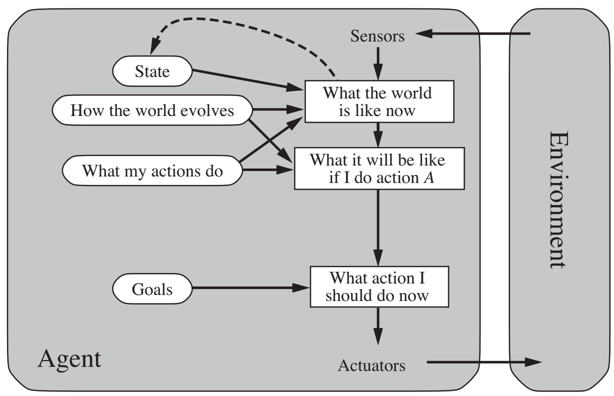
Simple Reflex Agents



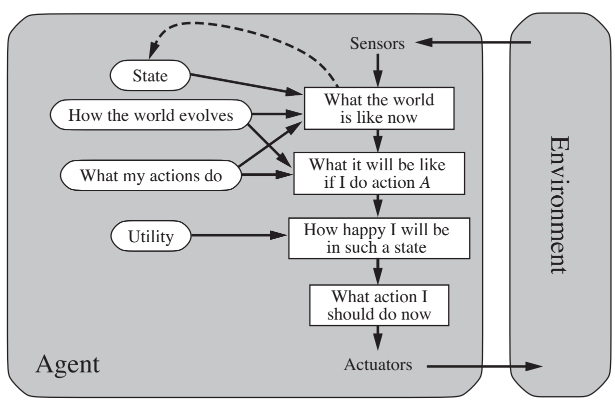
Reflex Agents with State



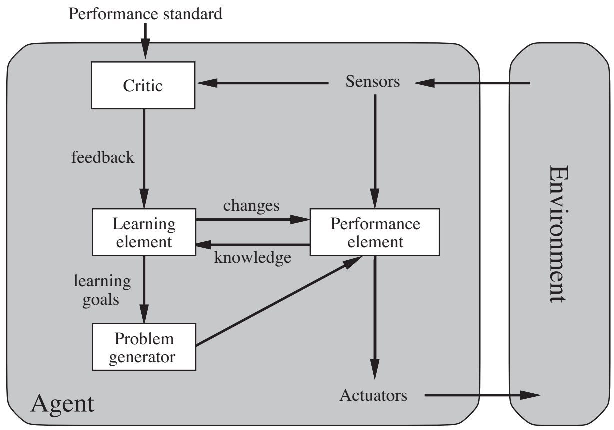
Goal-Based Agents



Utility-Based Agent



Learning Agents



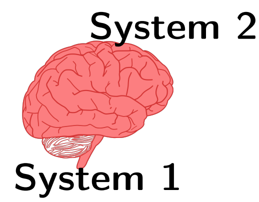
Rational Agent

- ▷ **Idea:** Try to design agents that are successful (do the right thing)
- ▷ **Definition 32.0.1.** An agent is called rational, if it chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date. This is called the MEU principle.
- ▷ **Note:** A rational agent need not be perfect
  - ▷ only needs to maximize expected value (rational $\neq$ omniscient)
    - ▷ need not predict e.g. very unlikely but catastrophic events in the future
  - ▷ percepts may not supply all relevant information (Rational $\neq$ clairvoyant)
    - ▷ if we cannot perceive things we do not need to react to them.
    - ▷ but we may need to try to find out about hidden dangers (exploration)
  - ▷ action outcomes may not be as expected (rational $\neq$ successful)
    - ▷ but we may need to take action to ensure that they do (more often) (learning)
- ▷ Rational $\rightsquigarrow$ exploration, learning, autonomy

Thinking, Fast and Slow (two Brain systems)

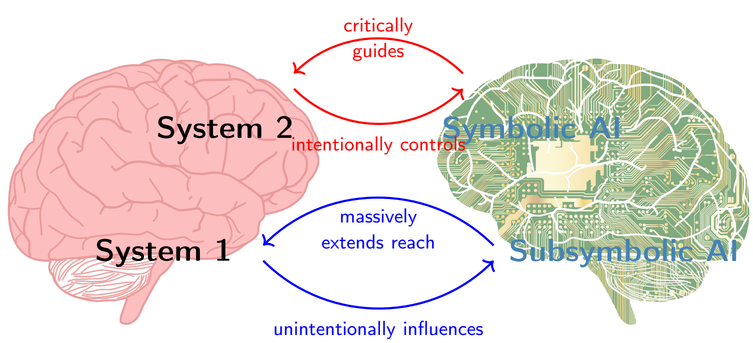
- ▷ In his 2011 Bestseller *Thinking, fast and slow* [Kah11], David Kahnemann posits a dichotomy between two modes of thought:
  - ▷ "System 1" is fast, instinctive and emotional;
  - ▷ "System 2" is slower, more deliberative, and more logical.

- System 1 can
  - see whether an object is near or far
  - complete the phrase *war and ...*
  - display disgust when seeing a gruesome image
  - solve 2+2=?
  - read text on a billboard
- System 2 can
  - look out for the woman with the grey hair
  - sustain a higher than normal walking rate
  - count the number of A's in a certain text
  - give someone your phone number
  - park into a tight parking space
  - solve 17×24
  - determine the validity of a complex argument



Thinking, Fast and Slow (two AI systems)

- System 1 and subsymbolic AI interface well. System 2 and symbolic AI interface well.



System 1

- low attention level
- short term desires
- little to no reflection
- microdecisions
- unintended influence

subsymbolic AI

- low transparency
- low interactivity
- low accountability
- rudimentary theory of mind

System 2	symbolic AI
▷ high attention level	▷ high transparency
▷ stable convictions	▷ high interactivity
▷ high level of reflection	▷ high accountability
▷ macro-decisions	▷ advanced theory of mind

Symbolic AI: Adding Knowledge to Algorithms

- ▷ Problem Solving (Black Box States, Transitions, Heuristics)
  - ▷ **Framework**: Problem Solving and Search (basic tree/graph walking)
  - ▷ **Variant**: Game playing (Adversarial Search) (Minimax + $\alpha\beta$-Pruning)
- ▷ Constraint Satisfaction Problems (heuristic search over partial assignments)
  - ▷ States as partial variable assignments, transitions as assignment
  - ▷ Heuristics informed by current restrictions, constraint graph
  - ▷ Inference as constraint propagation (transferring possible values across arcs)
- ▷ Describing world states by formal language (and drawing inferences)
  - ▷ Propositional logic and DPLL (deciding entailment efficiently)
  - ▷ First-order logic and ATP (reasoning about infinite domains)
  - ▷ **Digression**: Logic programming (logic + search)
  - ▷ Description logics as moderately expressive, but decidable logics
- ▷ Planning: Problem Solving using white-box world/action descriptions
  - ▷ **Framework**: describing world states in logic as sets of propositions and actions by preconditions and add/delete lists
  - ▷ **Algorithms**: e.g heuristic search by problem relaxations

Topics of AI-2 (Summer Semester)

- ▷ Uncertain Knowledge and Reasoning
  - ▷ Uncertainty
  - ▷ Probabilistic reasoning
  - ▷ Making Decisions in Episodic Environments
  - ▷ Problem Solving in Sequential Environments

- ▷ Foundations of machine learning
  - ▷ Learning from Observations
  - ▷ Knowledge in Learning
  - ▷ Statistical Learning Methods
- ▷ Communication (If there is time)
  - ▷ Natural Language Processing
  - ▷ Natural Language for Communication

Statistical AI: Adding uncertainty and Learning

- ▷ Problem Solving under uncertainty(non-observable environment, stochastic states)
  - ▷ **Framework**: Probabilistic Inference: Conditional Probabilities/Independence
  - ▷ **Intuition**: Reasoning in Belief Space instead of State Space!
  - ▷ **Implementation**: Bayesian Networks (exploit conditional independence)
  - ▷ **Extension**: Utilities and Decision Theory (for static/episodic environments)
- ▷ Problem Solving in Sequential Worlds:
  - ▷ **Framework**: Markov Processes, transition models
  - ▷ **Extension**: MDPs, POMDPs (+ utilities/decisions)
  - ▷ **Implementation**: Dynamic Bayesian Networks
- ▷ Machine learning: adding optimization in changing environments (unsupervised)
  - ▷ **Framework**: Learning from Observations (positive/negative examples)
  - ▷ **Intuitions**: finding consistent/optimal hypotheses in a hypothesis space
  - ▷ **Problems**: consistency, expressivity, under/overfitting, computational/data resources.
  - ▷ Extensions
    - ▷ knowledge in learning (based on logical methods)
    - ▷ statistical learning (optimizing the probability distribution over hypspace, learning BNs)
  - ▷ Communication
    - ▷ Phenomena of natural language (NL is interesting/complex)
    - ▷ symbolic/statistical NLP (historic/as a backup)
    - ▷ Deep Learning for NLP (the current hype/solution)

Topics of AI-3 – A Course not taught at FAU ☹

- ▷ Machine Learning
  - ▷ Theory and Practice of Deep Learning
  - ▷ More Reinforcement Learning
- ▷ Communicating, Perceiving, and Acting
  - ▷ More NLP, dialogue, speech acts, ...
  - ▷ Natural Language Semantics/Pragmatics
  - ▷ Perception
  - ▷ Robotics
  - ▷ Emotions, Sentiment Analysis
- ▷ **The Good News:** All is not lost
  - ▷ There are tons of specialized courses at FAU (more as we speak)
  - ▷ Russell/Norvig's AIMA [RN09] cover some of them as well!

Bibliography

[Bac00] Fahiem Bacchus. *Subset of PDDL for the AIPS2000 Planning Competition.* The AIPS-00 Planning Competition Comitee. 2000.

[BF95] Avrim L. Blum and Merrick L. Furst. “Fast planning through planning graph analysis”. In: *Proceedings of the 14^{th} International Joint Conference on Artificial Intelligence (IJCAI).* Ed. by Chris S. Mellish. Montreal, Canada: Morgan Kaufmann, San Mateo, CA, 1995, pp. 1636–1642.

[BF97] Avrim L. Blum and Merrick L. Furst. “Fast planning through planning graph analysis”. In: *Artificial Intelligence* 90.1-2 (1997), pp. 279–298.

[BG01] Blai Bonet and Héctor Geffner. “Planning as Heuristic Search”. In: *Artificial Intelligence* 129.1–2 (2001), pp. 5–33.

[BG99] Blai Bonet and Héctor Geffner. “Planning as Heuristic Search: New Results”. In: *Proceedings of the 5th European Conference on Planning (ECP’99).* Ed. by S. Biundo and M. Fox. Springer-Verlag, 1999, pp. 60–72.

[BKS04] Paul Beame, Henry A. Kautz, and Ashish Sabharwal. “Towards Understanding and Harnessing the Potential of Clause Learning”. In: *Journal of Artificial Intelligence Research* 22 (2004), pp. 319–351.

[Bon+12] Blai Bonet et al., eds. *Proceedings of the 22nd International Conference on Automated Planning and Scheduling (ICAPS’12).* AAAI Press, 2012.

[Bro90] Rodney Brooks. In: *Robotics and Autonomous Systems* 6.1–2 (1990), pp. 3–15. DOI: `10.1016/S0921-8890(05)80025-9`.

[Cho65] Noam Chomsky. *Syntactic structures.* Den Haag: Mouton, 1965.

[CKT91] Peter Cheeseman, Bob Kanefsky, and William M. Taylor. “Where the *Really* Hard Problems Are”. In: *Proceedings of the 12^{th} International Joint Conference on Artificial Intelligence (IJCAI).* Ed. by John Mylopoulos and Ray Reiter. Sydney, Australia: Morgan Kaufmann, San Mateo, CA, 1991, pp. 331–337.

[CM85] Eugene Charniak and Drew McDermott. *Introduction to Artificial Intelligence.* Addison Wesley, 1985.

[CQ69] Allan M. Collins and M. Ross Quillian. “Retrieval time from semantic memory”. In: *Journal of verbal learning and verbal behavior* 8.2 (1969), pp. 240–247. DOI: `10.1016/S0022-5371(69)80069-1`.

[Dav67] Donald Davidson. “Truth and Meaning”. In: *Synthese* 17 (1967).

[DCM12] DCMI Usage Board. *DCMI Metadata Terms.* DCMI Recommendation. Dublin Core Metadata Initiative, June 14, 2012. URL: `http://dublincore.org/documents/2012/06/14/dcmi-terms/`.

[DF31] B. De Finetti. “Sul significato soggettivo della probabilita”. In: *Fundamenta Mathematicae* 17 (1931), pp. 298–329.

[DHK15] Carmel Domshlak, Jörg Hoffmann, and Michael Katz. "Red-Black Planning: A New Systematic Approach to Partial Delete Relaxation". In: *Artificial Intelligence* 221 (2015), pp. 73–114.

[Ede01] Stefan Edelkamp. "Planning with Pattern Databases". In: *Proceedings of the 6th European Conference on Planning (ECP'01)*. Ed. by A. Cesta and D. Borrajo. Springer-Verlag, 2001, pp. 13–24.

[FD14] Zohar Feldman and Carmel Domshlak. "Simple Regret Optimization in Online Planning for Markov Decision Processes". In: *Journal of Artificial Intelligence Research* 51 (2014), pp. 165–205.

[Fis] John R. Fisher. *prolog :- tutorial.* URL: `https://www.cpp.edu/~jrfisher/www/prolog_tutorial/` (visited on 10/10/2019).

[FL03] Maria Fox and Derek Long. "PDDL2.1: An Extension to PDDL for Expressing Temporal Planning Domains". In: *Journal of Artificial Intelligence Research* 20 (2003), pp. 61–124.

[Fla94] Peter Flach. Wiley, 1994. ISBN: 0471 94152 2. URL: `https://github.com/simply-logical/simply-logical/releases/download/v1.0/SL.pdf`.

[FN71] Richard E. Fikes and Nils Nilsson. "STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving". In: *Artificial Intelligence* 2 (1971), pp. 189–208.

[Gen34] Gerhard Gentzen. "Untersuchungen über das logische Schließen I". In: *Mathematische Zeitschrift* 39.2 (1934), pp. 176–210.

[Ger+09] Alfonso Gerevini et al. "Deterministic planning in the fifth international planning competition: PDDL3 and experimental evaluation of the planners". In: *Artificial Intelligence* 173.5-6 (2009), pp. 619–668.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability—A Guide to the Theory of NP-Completeness.* BN book: Freeman, 1979.

[Glo] *Grundlagen der Logik in der Informatik*. Course notes at `https://www8.cs.fau.de/_media/ws16:gloin:skript.pdf`. URL: `https://www8.cs.fau.de/_media/ws16:gloin:skript.pdf` (visited on 10/13/2017).

[GNT04] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated Planning: Theory and Practice*. Morgan Kaufmann, 2004.

[GS05] Carla Gomes and Bart Selman. "Can get satisfaction". In: *Nature* 435 (2005), pp. 751–752.

[GSS03] Alfonso Gerevini, Alessandro Saetti, and Ivan Serina. "Planning through Stochastic Local Search and Temporal Action Graphs". In: *Journal of Artificial Intelligence Research* 20 (2003), pp. 239–290.

[Hau85] John Haugeland. *Artificial intelligence: the very idea.* Massachusetts Institute of Technology, 1985.

[HD09] Malte Helmert and Carmel Domshlak. "Landmarks, Critical Paths and Abstractions: What's the Difference Anyway?" In: *Proceedings of the 19th International Conference on Automated Planning and Scheduling (ICAPS'09)*. Ed. by Alfonso Gerevini et al. AAAI Press, 2009, pp. 162–169.

[HE05] Jörg Hoffmann and Stefan Edelkamp. "The Deterministic Part of IPC-4: An Overview". In: *Journal of Artificial Intelligence Research* 24 (2005), pp. 519–579.

[Hel06] Malte Helmert. "The Fast Downward Planning System". In: *Journal of Artificial Intelligence Research* 26 (2006), pp. 191–246.

[Her+13a] Ivan Herman et al. *RDF 1.1 Primer (Second Edition). Rich Structured Data Markup for Web Documents*. W3C Working Group Note. World Wide Web Consortium (W3C), 2013. URL: `http://www.w3.org/TR/rdfa-primer`.

[Her+13b] Ivan Herman et al. *RDFa 1.1 Primer – Second Edition. Rich Structured Data Markup for Web Documents*. W3C Working Goup Note. World Wide Web Consortium (W3C), Apr. 19, 2013. URL: `http://www.w3.org/TR/xhtml-rdfa-primer/`.

[HG00] Patrik Haslum and Hector Geffner. "Admissible Heuristics for Optimal Planning". In: *Proceedings of the 5th International Conference on Artificial Intelligence Planning Systems (AIPS'00)*. Ed. by S. Chien, R. Kambhampati, and C. Knoblock. Breckenridge, CO: AAAI Press, Menlo Park, 2000, pp. 140–149.

[HG08] Malte Helmert and Hector Geffner. "Unifying the Causal Graph and Additive Heuristics". In: *Proceedings of the 18th International Conference on Automated Planning and Scheduling (ICAPS'08)*. Ed. by Jussi Rintanen et al. AAAI Press, 2008, pp. 140–147.

[HHH07] Malte Helmert, Patrik Haslum, and Jörg Hoffmann. "Flexible Abstraction Heuristics for Optimal Sequential Planning". In: *Proceedings of the 17th International Conference on Automated Planning and Scheduling (ICAPS'07)*. Ed. by Mark Boddy, Maria Fox, and Sylvie Thiebaux. Providence, Rhode Island, USA: Morgan Kaufmann, 2007, pp. 176–183.

[Hit+12] Pascal Hitzler et al. *OWL 2 Web Ontology Language Primer (Second Edition)*. W3C Recommendation. World Wide Web Consortium (W3C), 2012. URL: `http://www.w3.org/TR/owl-primer`.

[HN01] Jörg Hoffmann and Bernhard Nebel. "The FF Planning System: Fast Plan Generation Through Heuristic Search". In: *Journal of Artificial Intelligence Research* 14 (2001), pp. 253–302.

[Hof11] Jörg Hoffmann. "Every806thing You Always Wanted to Know about Planning (But Were Afraid to Ask)". In: *Proceedings of the 34th Annual German Conference on Artificial Intelligence (KI'11)*. Ed. by Joscha Bach and Stefan Edelkamp. Vol. 7006. Lecture Notes in Computer Science. Springer, 2011, pp. 1–13. URL: `http://fai.cs.uni-saarland.de/hoffmann/papers/ki11.pdf`.

[How60] R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

[ILD] *7. Constraints: Interpreting Line Drawings*. URL: `https://www.youtube.com/watch?v=l-tzjenXrvI&t=2037s` (visited on 11/19/2019).

[JN33] E. S. Pearson J. Neyman. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 231.694-706 (1933), pp. 289–337. DOI: `10.1098/rsta.1933.0009`.

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Penguin Books, 2011. ISBN: 9780141033570.

[KC04] Graham Klyne and Jeremy J. Carroll. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C Recommendation. World Wide Web Consortium (W3C), Feb. 10, 2004. URL: `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`.

[KD09] Erez Karpas and Carmel Domshlak. "Cost-Optimal Planning with Landmarks". In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*. Ed. by C. Boutilier. Pasadena, California, USA: Morgan Kaufmann, July 2009, pp. 1728–1733.

[Kee74] R. L. Keeney. "Multiplicative utility functions". In: *Operations Research* 22 (1974), pp. 22–34.

[KHD13] Michael Katz, Jörg Hoffmann, and Carmel Domshlak. "Who Said We Need to Relax *all* Variables?" In: *Proceedings of the 23rd International Conference on Automated Planning and Scheduling (ICAPS'13)*. Ed. by Daniel Borrajo et al. Rome, Italy: AAAI Press, 2013, pp. 126–134.

[KHH12a] Michael Katz, Jörg Hoffmann, and Malte Helmert. "How to Relax a Bisimulation?" In: *Proceedings of the 22nd International Conference on Automated Planning and Scheduling (ICAPS'12)*. Ed. by Blai Bonet et al. AAAI Press, 2012, pp. 101–109.

[KHH12b] Emil Keyder, Jörg Hoffmann, and Patrik Haslum. "Semi-Relaxed Plan Heuristics". In: *Proceedings of the 22nd International Conference on Automated Planning and Scheduling (ICAPS'12)*. Ed. by Blai Bonet et al. AAAI Press, 2012, pp. 128–136.

[KNS97] B. Kessler, G. Nunberg, and H. Schütze. "Automatic detection of text genre". In: *CoRR* cmp-lg/9707002 (1997).

[Koe+97] Jana Koehler et al. "Extending Planning Graphs to an ADL Subset". In: *Proceedings of the 4th European Conference on Planning (ECP'97)*. Ed. by S. Steel and R. Alami. Springer-Verlag, 1997, pp. 273–285. URL: `ftp://ftp.informatik.uni-freiburg.de/papers/ki/koehler-etal-ecp-97.ps.gz`.

[Koh08] Michael Kohlhase. "Using LaTeX as a Semantic Markup Format". In: *Mathematics in Computer Science* 2.2 (2008), pp. 279–304. URL: `https://kwarc.info/kohlhase/papers/mcs08-stex.pdf`.

[Kow97] Robert Kowalski. "Algorithm = Logic + Control". In: *Communications of the Association for Computing Machinery* 22 (1997), pp. 424–436.

[KS00] Jana Köhler and Kilian Schuster. "Elevator Control as a Planning Problem". In: *AIPS 2000 Proceedings*. AAAI, 2000, pp. 331–338. URL: `https://www.aaai.org/Papers/AIPS/2000/AIPS00-036.pdf`.

[KS06] Levente Kocsis and Csaba Szepesvári. "Bandit Based Monte-Carlo Planning". In: *Proceedings of the 17th European Conference on Machine Learning (ECML 2006)*. Ed. by Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou. Vol. 4212. LNCS. Springer-Verlag, 2006, pp. 282–293.

[KS92] Henry A. Kautz and Bart Selman. "Planning as Satisfiability". In: *Proceedings of the 10th European Conference on Artificial Intelligence (ECAI'92)*. Ed. by B. Neumann. Vienna, Austria: Wiley, Aug. 1992, pp. 359–363.

[KS98] Henry A. Kautz and Bart Selman. "Pushing the Envelope: Planning, Propositional Logic, and Stochastic Search". In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence AAAI-96*. MIT Press, 1998, pp. 1194–1201.

[Kur90] Ray Kurzweil. *The Age of Intelligent Machines*. MIT Press, 1990. ISBN: 0-262-11121-7.

[Lee02] Lillian Lee. "Fast context-free grammar parsing requires fast Boolean matrix multiplication". In: *Journal of the ACM* 49.1 (2002), pp. 1–15.

[LPN] *Learn Prolog Now!* URL: `http://lpn.swi-prolog.org/` (visited on 10/10/2019).

[LS93] George F. Luger and William A. Stubblefield. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. World Student Series. The Benjamin/Cummings, 1993. ISBN: 9780805347852.

[Luc96] Peter Lucas. "Knowledge Acquisition for Decision-theoretic Expert Systems". In: *AISB Quarterly* 94 (1996), pp. 23–33. URL: `https://www.researchgate.net/publication/2460438_Knowledge_Acquisition_for_Decision-theoretic_Expert_Systems`.

[McD+98] Drew McDermott et al. *The PDDL Planning Domain Definition Language*. The AIPS-98 Planning Competition Comitee. 1998.

[Met+53] N. Metropolis et al. "Equations of state calculations by fast computing machines". In: *Journal of Chemical Physics* 21 (1953), pp. 1087–1091.

[Min] *Minion - Constraint Modelling*. System Web page at `http://constraintmodelling.org/minion/`. URL: `http://constraintmodelling.org/minion/`.

[MMS93] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: the penn treebank". In: *Computational Linguistics* 19.2 (1993), pp. 313–330.

[MR91] John Mylopoulos and Ray Reiter, eds. Sydney, Australia: Morgan Kaufmann, San Mateo, CA, 1991.

[MSL92] David Mitchell, Bart Selman, and Hector J. Levesque. "Hard and Easy Distributions of SAT Problems". In: *Proceedings of the 10th National Conference of the American Association for Artificial Intelligence (AAAI'92)*. San Jose, CA: MIT Press, 1992, pp. 459–465.

[NHH11] Raz Nissim, Jörg Hoffmann, and Malte Helmert. "Computing Perfect Heuristics in Polynomial Time: On Bisimulation and Merge-and-Shrink Abstraction in Optimal Planning". In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*. Ed. by Toby Walsh. AAAI Press/IJCAI, 2011, pp. 1983–1990.

[Nor+18a] Emily Nordmann et al. *Lecture capture: Practical recommendations for students and lecturers*. 2018. URL: `https://osf.io/huydx/download`.

[Nor+18b] Emily Nordmann et al. *Vorlesungsaufzeichnungen nutzen: Eine Anleitung für Studierende*. 2018. URL: `https://osf.io/e6r7a/download`.

[NS63] Allen Newell and Herbert Simon. "GPS, a program that simulates human thought". In: *Computers and Thought*. Ed. by E. Feigenbaum and J. Feldman. McGraw-Hill, 1963, pp. 279–293.

[NS76] Alan Newell and Herbert A. Simon. "Computer Science as Empirical Inquiry: Symbols and Search". In: *Communications of the ACM* 19.3 (1976), pp. 113–126. DOI: `10.1145/360018.360022`.

[OWL09] OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation. World Wide Web Consortium (W3C), Oct. 27, 2009. URL: `http://www.w3.org/TR/2009/REC-owl2-overview-20091027/`.

[PD09] Knot Pipatsrisawat and Adnan Darwiche. "On the Power of Clause-Learning SAT Solvers with Restarts". In: *Proceedings of the 15th International Conference on Principles and Practice of Constraint Programming (CP'09)*. Ed. by Ian P. Gent. Vol. 5732. Lecture Notes in Computer Science. Springer, 2009, pp. 654–668.

[Pól73] George Pólya. *How to Solve it. A New Aspect of Mathematical Method*. Princeton University Press, 1973.

[Pra+94] Malcolm Pradhan et al. "Knowledge Engineering for Large Belief Networks". In: *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*. UAI'94. Seattle, WA: Morgan Kaufmann Publishers Inc., 1994, pp. 484–490. ISBN: 1-55860-332-8. URL: `http://dl.acm.org/citation.cfm?id=2074394.2074456`.

[Pro] *Protégé*. Project Home page at `http://protege.stanford.edu`. URL: `http://protege.stanford.edu`.

[PRR97] G. Probst, St. Raub, and Kai Romhardt. *Wissen managen*. 4 (2003). Gabler Verlag, 1997.

[PS08] Eric Prud'hommeaux and Andy Seaborne. *SPARQL Query Language for RDF*. W3C Recommendation. World Wide Web Consortium (W3C), Jan. 15, 2008. URL: `http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/`.

[PW92] J. Scott Penberthy and Daniel S. Weld. "UCPOP: A Sound, Complete, Partial Order Planner for ADL". In: *Principles of Knowledge Representation and Reasoning: Proceedings of the 3rd International Conference (KR-92)*. Ed. by B. Nebel, W. Swartout, and C. Rich. Cambridge, MA: Morgan Kaufmann, Oct. 1992, pp. 103–114. URL: `ftp://ftp.cs.washington.edu/pub/ai/ucpop-kr92.ps.Z`.

[RHN06] Jussi Rintanen, Keijo Heljanko, and Ilkka Niemelä. "Planning as satisfiability: parallel plans and algorithms for plan search". In: *Artificial Intelligence* 170.12-13 (2006), pp. 1031–1080.

[Rin10] Jussi Rintanen. "Heuristics for Planning with SAT". In: *Proceeedings of the 16th International Conference on Principles and Practice of Constraint Programming*. 2010, pp. 414–428.

[RN03] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 2nd ed. Pearso n Education, 2003. ISBN: 0137903952.

[RN09] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. Prentice Hall Press, 2009. ISBN: 0136042597, 9780136042594.

[RN95] Stuart J. Russell and Peter Norvig. *Artificial Intelligence — A Modern Approach*. Upper Saddle River, NJ: Prentice Hall, 1995.

[RW10] Silvia Richter and Matthias Westphal. "The LAMA Planner: Guiding Cost-Based Anytime Planning with Landmarks". In: *Journal of Artificial Intelligence Research* 39 (2010), pp. 127–177.

[RW91] S. J. Russell and E. Wefald. *Do the Right Thing — Studies in limited Rationality*. MIT Press, 1991.

[Sil+16] David Silver et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529 (2016), pp. 484–503. URL: `http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html`.

[Smu63] Raymond M. Smullyan. "A Unifying Principle for Quantification Theory". In: *Proc. Nat. Acad Sciences* 49 (1963), pp. 828–832.

[SR14] Guus Schreiber and Yves Raimond. *RDF 1.1 Primer*. W3C Working Group Note. World Wide Web Consortium (W3C), 2014. URL: `http://www.w3.org/TR/rdf-primer`.

[SR91] C. Samuelsson and M. Rayner. "Quantitative evaluation of explanation-based learning as an optimization tool for a large-scale natural language system". In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*. Ed. by John Mylopoulos and Ray Reiter. Sydney, Australia: Morgan Kaufmann, San Mateo, CA, 1991, pp. 609–615.

[sTeX] *sTeX: A semantic Extension of TeX/LaTeX*. URL: `https://github.com/sLaTeX/sTeX` (visited on 05/11/2020).

[SWI] *SWI Prolog Reference Manual*. URL: `https://www.swi-prolog.org/pldoc/refman/` (visited on 10/10/2019).

[Tur50] Alan Turing. "Computing Machinery and Intelligence". In: *Mind* 59 (1950), pp. 433–460.

[Vas+17] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[Wal75] David Waltz. "Understanding Line Drawings of Scenes with Shadows". In: *The Psychology of Computer Vision*. Ed. by P. H. Winston. McGraw-Hill, 1975, pp. 1–19.

[WHI] *Human intelligence — Wikipedia The Free Encyclopedia*. URL: `https://en.wikipedia.org/w/index.php?title=Human_intelligence` (visited on 04/09/2018).

Part VIII

Excursions

As this course is predominantly an overview over the topics of Artificial Intelligence, and not about the theoretical underpinnings, we give the discussion about these as a “suggested readings” part here.

Appendix A

Completeness of Calculi for Propositional Logic

The next step is to analyze the two calculi for completeness. For that we will first give ourselves a very powerful tool: the "model existence theorem" (**??**), which encapsulates the model-theoretic part of completeness theorems. With that, completeness proofs – which are quite tedious otherwise – become a breeze.

A.1 Abstract Consistency and Model Existence

We will now come to an important tool in the theoretical study of reasoning calculi: the "abstract consistency"/"model existence" method. This method for analyzing calculi was developed by Jaako Hintikka, Raymond Smullyan, and Peter Andrews in 1950-1970 as an encapsulation of similar constructions that were used in completeness arguments in the decades before. The basis for this method is Smullyan's Observation [Smu63] that completeness proofs based on Hintikka sets only certain properties of consistency and that with little effort one can obtain a generalization "Smullyan's Unifying Principle".

The basic intuition for this method is the following: typically, a logical system $\mathcal{L} = \langle \mathcal{L}, \mathcal{K}, \models \rangle$ has multiple calculi, human-oriented ones like the natural deduction calculi and machine-oriented ones like the automated theorem proving calculi. All of these need to be analyzed for completeness (as a basic quality assurance measure).

A completeness proof for a calculus $\mathcal{C}$ for $\mathcal{S}$ typically comes in two parts: one analyzes $\mathcal{C}$-consistency (sets that cannot be refuted in $\mathcal{C}$), and the other construct $\mathcal{K}$-models for $\mathcal{C}$-consistent sets.

In this situtation the "abstract consistency"/"model existence" method encapsulates the model construction process into a meta-theorem: the "model existence" theorem. This provides a set of syntactic ("abstract consistency") conditions for calculi that are sufficient to construct models.

With the model existence theorem it suffices to show that $\mathcal{C}$-consistency is an abstract consistency property (a purely syntactic task that can be done by a $\mathcal{C}$-proof transformation argument) to obtain a completeness result for $\mathcal{C}$.

Model Existence (Overview)

- ▷ **Definition:** Abstract consistency
- ▷ **Definition:** Hintikka set (maximally abstract consistent)
- ▷ **Theorem:** Hintikka sets are satisfiable

- ▷ **Theorem:** If Φ is abstract consistent, then Φ can be extended to a Hintikka set.
- ▷ **Corollary:** If Φ is abstract consistent, then Φ is satisfiable.
- ▷ **Application:** Let $\mathcal{C}$ be a calculus, if Φ is $\mathcal{C}$-consistent, then Φ is abstract consistent.
- ▷ **Corollary:** $\mathcal{C}$ is complete.

The proof of the model existence theorem goes via the notion of a Hintikka set, a set of formulae with very strong syntactic closure properties, which allow to read off models. Jaako Hintikka's original idea for completeness proofs was that for every complete calculus $\mathcal{C}$ and every $\mathcal{C}$-consistent set one can induce a Hintikka set, from which a model can be constructed. This can be considered as a first model existence theorem. However, the process of obtaining a Hintikka set for a $\mathcal{C}$-consistent set Φ of sentences usually involves complicated calculus dependent constructions.

In this situation, Raymond Smullyan was able to formulate the sufficient conditions for the existence of Hintikka sets in the form of "abstract consistency properties" by isolating the calculus independent parts of the Hintikka set construction. His technique allows to reformulate Hintikka sets as maximal elements of abstract consistency classes and interpret the Hintikka set construction as a maximizing limit process.

To carry out the "model-existence"/"abstract consistency" method, we will first have to look at the notion of consistency.

Consistency and refutability are very important notions when studying the completeness for calculi; they form syntactic counterparts of satisfiability.

Consistency

- ▷ Let $\mathcal{C}$ be a calculus,. . .
- ▷ **Definition A.1.1.** Let $\mathcal{C}$ be a calculus, then a formula set Φ is called $\mathcal{C}$-, if there is a refutation, i.e. a derivation of a contradiction from Φ. The act of finding a refutation for Φ is called refuting Φ.
- ▷ **Definition A.1.2.** We call a pair of formulae $\mathbf{A}$ and $\neg\mathbf{A}$ a contradiction.
- ▷ So a set Φ is $\mathcal{C}$-refutable, if $\mathcal{C}$ canderive a contradiction from it.
- ▷ **Definition A.1.3.** Let $\mathcal{C}$ be a calculus, then a formula set Φ is called $\mathcal{C}$-, iff there is a formula $\mathbf{B}$, that is not derivable from Φ in $\mathcal{C}$.
- ▷ **Definition A.1.4.** We call a calculus $\mathcal{C}$ reasonable, iff implication elimination and conjunction introduction are admissible in $\mathcal{C}$ and $\mathbf{A} \wedge \neg\mathbf{A} \Rightarrow \mathbf{B}$ is a $\mathcal{C}$-theorem.
- ▷ **Theorem A.1.5.** *$\mathcal{C}$-inconsistency and $\mathcal{C}$-refutability coincide for reasonable calculi.*

It is very important to distinguish the syntactic $\mathcal{C}$-refutability and $\mathcal{C}$-consistency from satisfiability, which is a property of formulae that is at the heart of semantics. Note that the former have the calculus (a syntactic device) as a parameter, while the latter does not. In fact we should actually say $\mathcal{S}$-satisfiability, where $\langle\mathcal{L}, \mathcal{K}, \models\rangle$ is the current logical system.

Even the word "contradiction" has a syntactical flavor to it, it translates to "saying against each other" from its Latin root.

Abstract Consistency

- ▷ **Definition A.1.6.** Let ∇ be a collection of sets. We call ∇ closed under subsets, iff for each $\Phi\in\nabla$, all subsets $\Psi \subseteq \Phi$ are elements of ∇.
- ▷ **Notation:** We will use $\Phi * \mathbf{A}$ for $\Phi \cup \{\mathbf{A}\}$.
- ▷ **Definition A.1.7.** A collection ∇ of sets of propositional formulae is called an abstract consistency class, iff it is closed under subsets, and for each $\Phi\in\nabla$

  ∇_c) $P\notin\Phi$ or $\neg P\notin\Phi$ for $P\in\mathcal{V}_0$

  $\nabla_\neg$) $\neg\neg\mathbf{A}\in\Phi$ implies $\Phi * \mathbf{A}\in\nabla$

  $\nabla_\vee$) $\mathbf{A} \vee \mathbf{B}\in\Phi$ implies $\Phi * \mathbf{A}\in\nabla$ or $\Phi * \mathbf{B}\in\nabla$

  $\nabla_\wedge$) $\neg(\mathbf{A} \vee \mathbf{B})\in\Phi$ implies $\Phi \cup \{\neg\mathbf{A}, \neg\mathbf{B}\}\in\nabla$
- ▷ **Example A.1.8.** The empty set is an abstract consistency class
- ▷ **Example A.1.9.** The set $\{\emptyset, \{Q\}, \{P \vee Q\}, \{P \vee Q, Q\}\}$ is an abstract consistency class
- ▷ **Example A.1.10.** The family of satisfiable sets is an abstract consistency class.

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG Michael Kohlhase: Artificial Intelligence 2 1199 2023-09-20

So a family of sets (we call it a family, so that we do not have to say "set of sets" and we can distinguish the levels) is an abstract consistency class, iff it fulfills five simple conditions, of which the last three are closure conditions.

Think of an abstract consistency class as a family of "consistent" sets (e.g. $\mathcal{C}$-consistent for some calculus $\mathcal{C}$), then the properties make perfect sense: They are naturally closed under subsets — if we cannot derive a contradiction from a large set, we certainly cannot from a subset, furthermore,

∇_c) If both $P\in\Phi$ and $\neg P\in\Phi$, then Φ cannot be "consistent".

$\nabla_\neg$) If we cannot derive a contradiction from Φ with $\neg\neg\mathbf{A}\in\Phi$ then we cannot from $\Phi * \mathbf{A}$, since they are logically equivalent.

The other two conditions are motivated similarly. We will carry out the proof here, since it gives us practice in dealing with the abstract consistency properties.

The main result here is that abstract consistency classes can be extended to compact ones. The proof is quite tedious, but relatively straightforward. It allows us to assume that all abstract consistency classes are compact in the first place (otherwise we pass to the compact extension).

Actually we are after abstract consistency classes that have an even stronger property than just being closed under subsets. This will allow us to carry out a limit construction in the Hintikka set extension argument later.

Compact Collections

- ▷ **Definition A.1.11.** We call a collection ∇ of sets compact, iff for any set Φ we have
  $\Phi\in\nabla$, iff $\Psi\in\nabla$ for every finite subset Ψ of Φ.
- ▷ **Lemma A.1.12.** *If ∇ is compact, then ∇ is closed under subsets.*
- ▷ *Proof:*

1. Suppose $S \subseteq T$ and $T \in \nabla$.
2. Every finite subset A of S is a finite subset of T.
3. As ∇ is compact, we know that $A \in \nabla$.
4. Thus $S \in \nabla$.

The property of being closed under subsets is a "downwards-oriented" property: We go from large sets to small sets, compactness (the interesting direction anyways) is also an "upwards-oriented" property. We can go from small (finite) sets to large (infinite) sets. The main application for the compactness condition will be to show that infinite sets of formulae are in a collection ∇ by testing all their finite subsets (which is much simpler).

Compact Abstract Consistency Classes

▷ **Lemma A.1.13.** *Any abstract consistency class can be extended to a compact one.*

▷ *Proof:*

1. We choose $\nabla' := \{\Phi \subseteq wff_0(\mathcal{V}_0) | \text{every finite subset of } \Phi \text{ is in } \nabla\}$.
2. Now suppose that $\Phi \in \nabla$. ∇ is closed under subsets, so every finite subset of Φ is in ∇ and thus $\Phi \in \nabla'$. Hence $\nabla \subseteq \nabla'$.
3. Next let us show that each ∇ is compact.'
   3.1. Suppose $\Phi \in \nabla'$ and Ψ is an arbitrary finite subset of Φ.
   3.2. By definition of ∇' all finite subsets of Φ are in ∇ and therefore $\Psi \in \nabla'$.
   3.3. Thus all finite subsets of Φ are in ∇' whenever Φ is in ∇'.
   3.4. On the other hand, suppose all finite subsets of Φ are in ∇'.
   3.5. Then by the definition of ∇' the finite subsets of Φ are also in ∇, so $\Phi \in \nabla'$. Thus ∇' is compact.
4. Note that ∇' is closed under subsets by the Lemma above.
5. Now we show that if ∇ satisfies ∇_*, then ∇ satisfies ∇_*.'
   5.1. To show ∇_c, let $\Phi \in \nabla'$ and suppose there is an atom $\mathbf{A}$, such that $\{\mathbf{A}, \neg\mathbf{A}\} \subseteq \Phi$. Then $\{\mathbf{A}, \neg\mathbf{A}\} \in \nabla$ contradicting ∇_c.
   5.2. To show $\nabla_\neg$, let $\Phi \in \nabla'$ and $\neg\neg\mathbf{A} \in \Phi$, then $\Phi * \mathbf{A} \in \nabla'$.
      5.2.1. Let Ψ be any finite subset of $\Phi * \mathbf{A}$, and $\Theta := (\Psi \backslash \{\mathbf{A}\}) * \neg\neg\mathbf{A}$.
      5.2.2. Θ is a finite subset of Φ, so $\Theta \in \nabla$.
      5.2.3. Since ∇ is an abstract consistency class and $\neg\neg\mathbf{A} \in \Theta$, we get $\Theta * \mathbf{A} \in \nabla$ by $\nabla_\neg$.
      5.2.4. We know that $\Psi \subseteq \Theta * \mathbf{A}$ and ∇ is closed under subsets, so $\Psi \in \nabla$.
      5.2.5. Thus every finite subset Ψ of $\Phi * \mathbf{A}$ is in ∇ and therefore by definition $\Phi * \mathbf{A} \in \nabla'$.
   5.3. the other cases are analogous to $\nabla_\neg$.

Hintikka sets are sets of sentences with very strong analytic closure conditions. These are motivated as maximally consistent sets i.e. sets that already contain everything that can be consistently added to them.

∇-Hintikka Set

▷ **Definition A.1.14.** Let ∇ be an abstract consistency class, then we call a set $\mathcal{H}\in\nabla$ a ∇ Hintikka Set, iff $\mathcal{H}$ is maximal in ∇, i.e. for all $\mathbf{A}$ with $\mathcal{H}*\mathbf{A}\in\nabla$ we already have $\mathbf{A}\in\mathcal{H}$.

▷ **Theorem A.1.15 (Hintikka Properties).** *Let ∇ be an abstract consistency class and $\mathcal{H}$ be a ∇-Hintikka set, then*

$\mathcal{H}_c$) *For all $\mathbf{A}\in wff_0(\mathcal{V}_0)$ we have $\mathbf{A}\notin\mathcal{H}$ or $\neg\mathbf{A}\notin\mathcal{H}$*

$\mathcal{H}_\neg$) *If $\neg\neg\mathbf{A}\in\mathcal{H}$ then $\mathbf{A}\in\mathcal{H}$*

$\mathcal{H}_\vee$) *If $\mathbf{A}\vee\mathbf{B}\in\mathcal{H}$ then $\mathbf{A}\in\mathcal{H}$ or $\mathbf{B}\in\mathcal{H}$*

$\mathcal{H}_\wedge$) *If $\neg(\mathbf{A}\vee\mathbf{B})\in\mathcal{H}$ then $\neg\mathbf{A},\neg\mathbf{B}\in\mathcal{H}$*

∇-Hintikka Set

▷ *Proof:*

We prove the properties in turn

1. $\mathcal{H}_c$ *by induction on the structure of* $\mathbf{A}$
   1.1. $\mathbf{A}\in\mathcal{V}_0$ Then $\mathbf{A}\notin\mathcal{H}$ or $\neg\mathbf{A}\notin\mathcal{H}$ by ∇_c.
   1.2. $\mathbf{A}=\neg\mathbf{B}$
      1.2.1. Let us assume that $\neg\mathbf{B}\in\mathcal{H}$ and $\neg\neg\mathbf{B}\in\mathcal{H}$,
      1.2.2. then $\mathcal{H}*\mathbf{B}\in\nabla$ by $\nabla_\neg$, and therefore $\mathbf{B}\in\mathcal{H}$ by maximality.
      1.2.3. So both $\mathbf{B}$ and $\neg\mathbf{B}$ are in $\mathcal{H}$, which contradicts the inductive hypothesis.
   1.3. $\mathbf{A}=\mathbf{B}\vee\mathbf{C}$ *similar to the previous case*
2. We prove $\mathcal{H}_\neg$ by maximality of $\mathcal{H}$ in ∇.
   2.1. If $\neg\neg\mathbf{A}\in\mathcal{H}$, then $\mathcal{H}*\mathbf{A}\in\nabla$ by $\nabla_\neg$.
   2.2. The maximality of $\mathcal{H}$ now gives us that $\mathbf{A}\in\mathcal{H}$.

Proof sketch: *other $\mathcal{H}_*$ are similar*

The following theorem is one of the main results in the "abstract consistency"/"model existence" method. For any abstract consistent set Φ it allows us to construct a Hintikka set $\mathcal{H}$ with $\Phi\in\mathcal{H}$.

Extension Theorem

▷ **Theorem A.1.16.** *If ∇ is an abstract consistency class and $\Phi\in\nabla$, then there is a ∇-Hintikka set $\mathcal{H}$ with $\Phi\subseteq\mathcal{H}$.*

▷ *Proof:*

1. Wlog. we assume that ∇ is compact (otherwise pass to compact extension)
2. We choose an enumeration $\mathbf{A}_1,\ldots$ of the set $wff_0(\mathcal{V}_0)$
3. and construct a sequence of sets $\mathbf{H}_i$ with $\mathbf{H}_0:=\Phi$ and

$$\mathbf{H}_{n+1}:=\begin{cases}\mathbf{H}_n & \text{if } \mathbf{H}_n*\mathbf{A}_n\notin\nabla\\ \mathbf{H}_n*\mathbf{A}_n & \text{if } \mathbf{H}_n*\mathbf{A}_n\in\nabla\end{cases}$$

4. Note that all $\mathbf{H}_i\in\nabla$, choose $\mathcal{H}:=\bigcup_{i\in\mathbb{N}}\mathbf{H}_i$

5. $\Psi \subseteq \mathcal{H}$ finite implies there is a $j \in \mathbb{N}$ such that $\Psi \subseteq \mathbf{H}_j$,
6. so $\Psi \in \nabla$ as ∇ is closed under subsets and $\mathcal{H} \in \nabla$ as ∇ is compact.
7. Let $\mathcal{H} * \mathbf{B} \in \nabla$, then there is a $j \in \mathbb{N}$ with $\mathbf{B} = \mathbf{A}_j$, so that $\mathbf{B} \in \mathbf{H}_{j+1}$ and $\mathbf{H}_{j+1} \subseteq \mathcal{H}$
8. Thus $\mathcal{H}$ is ∇-maximal

Note that the construction in the proof above is non-trivial in two respects. First, the limit construction for $\mathcal{H}$ is not executed in our original abstract consistency class ∇, but in a suitably extended one to make it compact — the original would not have contained $\mathcal{H}$ in general. Second, the set $\mathcal{H}$ is not unique for Φ, but depends on the choice of the enumeration of $\mathit{wff}_0(\mathcal{V}_0)$. If we pick a different enumeration, we will end up with a different $\mathcal{H}$. Say if $\mathbf{A}$ and $\neg\mathbf{A}$ are both ∇-consistent[1] with Φ, then depending on which one is first in the enumeration $\mathcal{H}$, will contain that one; with all the consequences for subsequent choices in the construction process.

Valuation

▷ **Definition A.1.17.** A function $\nu \colon \mathit{wff}_0(\mathcal{V}_0) \to \mathcal{D}_o$ is called a valuation, iff

- ▷ $\nu(\neg\mathbf{A}) = \mathsf{T}$, iff $\nu(\mathbf{A}) = \mathsf{F}$
- ▷ $\nu(\mathbf{A} \wedge \mathbf{B}) = \mathsf{T}$, iff $\nu(\mathbf{A}) = \mathsf{T}$ and $\nu(\mathbf{B}) = \mathsf{T}$

▷ **Lemma A.1.18.** *If $\nu \colon \mathit{wff}_0(\mathcal{V}_0) \to \mathcal{D}_o$ is a valuation and $\Phi \subseteq \mathit{wff}_0(\mathcal{V}_0)$ with $\nu(\Phi) = \{\mathsf{T}\}$, then Φ is satisfiable.*

▷ *Proof sketch:* $\nu|_{\mathcal{V}_0} : \mathcal{V}_0 \to \mathcal{D}_o$ is a satisfying variable assignment.

▷ **Lemma A.1.19.** *If $\varphi \colon \mathcal{V}_0 \to \mathcal{D}_o$ is a variable assignment, then $\mathcal{I}_\varphi \colon \mathit{wff}_0(\mathcal{V}_0) \to \mathcal{D}_o$ is a valuation.*

Now, we only have to put the pieces together to obtain the model existence theorem we are after.

Model Existence

▷ **Lemma A.1.20 (Hintikka-Lemma).** *If ∇ is an abstract consistency class and $\mathcal{H}$ a ∇-Hintikka set, then $\mathcal{H}$ is satisfiable.*

▷ *Proof:*

1. We define $\nu(\mathbf{A}) := \mathsf{T}$, iff $\mathbf{A} \in \mathcal{H}$
2. then ν is a valuation by the Hintikka properties
3. and thus $\nu|_{\mathcal{V}_0}$ is a satisfying assignment.

▷ **Theorem A.1.21 (Model Existence).** *If ∇ is an abstract consistency class and $\Phi \in \nabla$, then Φ is satisfiable.*

Proof:

▷ 1. There is a ∇-Hintikka set $\mathcal{H}$ with $\Phi \subseteq \mathcal{H}$ (Extension Theorem)
2. We know that $\mathcal{H}$ is satisfiable. (Hintikka-Lemma)
3. In particular, $\Phi \subseteq \mathcal{H}$ is satisfiable.

[1] EdNote: introduce this above

A.2 A Completeness Proof for Propositional Tableaux

With the model existence proof we have introduced in the last section, the completeness proof for first-order natural deduction is rather simple, we only have to check that Tableaux-consistency is an abstract consistency property.
We encapsulate all of the technical difficulties of the problem in a technical Lemma. From that, the completeness proof is just an application of the high-level theorems we have just proven.

Abstract Completeness for $\mathcal{T}_0$

▷ **Lemma A.2.1.** *$\{\Phi|\Phi^{\mathsf{T}}$ has no closed tableau$\}$ is an abstract consistency class.*

▷ *Proof:* Let's call the set above ∇

We have to convince ourselves of the abstract consistency properties

1. $\nabla_c P, \neg P \in \Phi$ *implies* $P^{\mathsf{F}}, P^{\mathsf{T}} \in \Phi^{\mathsf{T}}$.
2. $\nabla_\neg$ *Let* $\neg\neg\mathbf{A} \in \Phi$.
   2.1. For the proof of the contrapositive we assume that $\Phi * \mathbf{A}$ has a closed tableau $\mathcal{T}$ and show that already Φ has one:
   2.2. applying each of $\mathcal{T}_0\neg^{\mathsf{T}}$ and $\mathcal{T}_0\neg^{\mathsf{F}}$ once allows to extend any tableau with $\neg\neg\mathbf{B}^\alpha$ by $\mathbf{B}^\alpha$.
   2.3. any path in $\mathcal{T}$ that is closed with $\neg\neg\mathbf{A}^\alpha$, can be closed by $\mathbf{A}^\alpha$.
3. $\nabla_\vee$ *Suppose* $\mathbf{A} \vee \mathbf{B} \in \Phi$ *and both* $\Phi * \mathbf{A}$ *and* $\Phi * \mathbf{B}$ *have closed tableaux*
   3.1. consider the tableaux:

$$\begin{array}{c} \Phi^{\mathsf{T}} \\ \mathbf{A}^{\mathsf{T}} \\ Rest^1 \end{array} \qquad \begin{array}{c} \Phi^{\mathsf{T}} \\ \mathbf{B}^{\mathsf{T}} \\ Rest^2 \end{array} \qquad \begin{array}{c} \Psi^{\mathsf{T}} \\ (\mathbf{A} \vee \mathbf{B})^{\mathsf{T}} \\ \begin{array}{c|c} \mathbf{A}^{\mathsf{T}} & \mathbf{B}^{\mathsf{T}} \\ Rest^1 & Rest^2 \end{array} \end{array}$$

4. $\nabla_\wedge$ *suppose,* $\neg(\mathbf{A} \vee \mathbf{B}) \in \Phi$ *and* $\Phi\{\neg\mathbf{A}, \neg\mathbf{B}\}$ *have closed tableau* $\mathcal{T}$.
   4.1. We consider

$$\begin{array}{c} \Phi^{\mathsf{T}} \\ \mathbf{A}^{\mathsf{F}} \\ \mathbf{B}^{\mathsf{F}} \\ Rest \end{array} \qquad \begin{array}{c} \Psi^{\mathsf{T}} \\ (\mathbf{A} \vee \mathbf{B})^{\mathsf{F}} \\ \mathbf{A}^{\mathsf{F}} \\ \mathbf{B}^{\mathsf{F}} \\ Rest \end{array}$$

where $\Phi = \Psi * \neg(\mathbf{A} \vee \mathbf{B})$.

Observation: If we look at the completeness proof below, we see that the Lemma above is the only place where we had to deal with specific properties of the $\mathcal{T}_0$.

So if we want to prove completeness of any other calculus with respect to propositional logic, then we only need to prove an analogon to this lemma and can use the rest of the machinery we have already established "off the shelf".

This is one great advantage of the "abstract consistency method"; the other is that the method can be extended transparently to other logics.

Completeness of $\mathcal{T}_0$

▷ **Corollary A.2.2.** *$\mathcal{T}_0$ is complete.*

▷ *Proof:* by contradiction

1. We assume that $\mathbf{A} \in wff_0(\mathcal{V}_0)$ is valid, but there is no closed tableau for $\mathbf{A}^{\mathsf{F}}$.
2. We have $\{\neg\mathbf{A}\} \in \nabla$ as $\neg\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{\mathsf{F}}$.
3. so $\neg\mathbf{A}$ is satisfiable by the model existence theorem (which is applicable as ∇ is an abstract consistency class by our Lemma above)
4. this contradicts our assumption that $\mathbf{A}$ is valid.

Appendix B

Completeness of Calculi for First-Order Logic

We will now analyze the first-order calculi for completeness. Just as in the case of the propositional calculi, we prove a model existence theorem for the first-order model theory and then use that for the completeness proofs[2]. The proof of the first-order model existence theorem is completely analogous to the propositional one; indeed, apart from the model construction itself, it is just an extension by a treatment for the first-order quantifiers.[3] EdN:2 EdN:3

B.1 Abstract Consistency and Model Existence

We will now come to an important tool in the theoretical study of reasoning calculi: the "abstract consistency"/"model existence" method. This method for analyzing calculi was developed by Jaako Hintikka, Raymond Smullyan, and Peter Andrews in 1950-1970 as an encapsulation of similar constructions that were used in completeness arguments in the decades before. The basis for this method is Smullyan's Observation [Smu63] that completeness proofs based on Hintikka sets only certain properties of consistency and that with little effort one can obtain a generalization "Smullyan's Unifying Principle".

The basic intuition for this method is the following: typically, a logical system $\mathcal{L} = \langle \mathcal{L}, \mathcal{K}, \models \rangle$ has multiple calculi, human-oriented ones like the natural deduction calculi and machine-oriented ones like the automated theorem proving calculi. All of these need to be analyzed for completeness (as a basic quality assurance measure).

A completeness proof for a calculus $\mathcal{C}$ for $\mathcal{S}$ typically comes in two parts: one analyzes $\mathcal{C}$-consistency (sets that cannot be refuted in $\mathcal{C}$), and the other construct $\mathcal{K}$-models for $\mathcal{C}$-consistent sets.

In this situtation the "abstract consistency"/"model existence" method encapsulates the model construction process into a meta-theorem: the "model existence" theorem. This provides a set of syntactic ("abstract consistency") conditions for calculi that are sufficient to construct models.

With the model existence theorem it suffices to show that $\mathcal{C}$-consistency is an abstract consistency property (a purely syntactic task that can be done by a $\mathcal{C}$-proof transformation argument) to obtain a completeness result for $\mathcal{C}$.

Model Existence (Overview)

▷ **Definition:** Abstract consistency

[2] EdNote: reference the theorems
[3] EdNote: MK: what about equality?

- ▷ **Definition:** Hintikka set (maximally abstract consistent)
- ▷ **Theorem:** Hintikka sets are satisfiable
- ▷ **Theorem:** If Φ is abstract consistent, then Φ can be extended to a Hintikka set.
- ▷ **Corollary:** If Φ is abstract consistent, then Φ is satisfiable.
- ▷ **Application:** Let $\mathcal{C}$ be a calculus, if Φ is $\mathcal{C}$-consistent, then Φ is abstract consistent.
- ▷ **Corollary:** $\mathcal{C}$ is complete.

The proof of the model existence theorem goes via the notion of a Hintikka set, a set of formulae with very strong syntactic closure properties, which allow to read off models. Jaako Hintikka's original idea for completeness proofs was that for every complete calculus $\mathcal{C}$ and every $\mathcal{C}$-consistent set one can induce a Hintikka set, from which a model can be constructed. This can be considered as a first model existence theorem. However, the process of obtaining a Hintikka set for a $\mathcal{C}$-consistent set Φ of sentences usually involves complicated calculus dependent constructions.

In this situation, Raymond Smullyan was able to formulate the sufficient conditions for the existence of Hintikka sets in the form of "abstract consistency properties" by isolating the calculus independent parts of the Hintikka set construction. His technique allows to reformulate Hintikka sets as maximal elements of abstract consistency classes and interpret the Hintikka set construction as a maximizing limit process.

To carry out the "model-existence"/"abstract consistency" method, we will first have to look at the notion of consistency.

Consistency and refutability are very important notions when studying the completeness for calculi; they form syntactic counterparts of satisfiability.

Consistency

- ▷ Let $\mathcal{C}$ be a calculus,...
- ▷ **Definition B.1.1.** Let $\mathcal{C}$ be a calculus, then a formula set Φ is called $\mathcal{C}$-, if there is a refutation, i.e. a derivation of a contradiction from Φ. The act of finding a refutation for Φ is called refuting Φ.
- ▷ **Definition B.1.2.** We call a pair of formulae $\mathbf{A}$ and $\neg\mathbf{A}$ a contradiction.
- ▷ So a set Φ is $\mathcal{C}$-refutable, if $\mathcal{C}$ canderive a contradiction from it.
- ▷ **Definition B.1.3.** Let $\mathcal{C}$ be a calculus, then a formula set Φ is called $\mathcal{C}$-, iff there is a formula $\mathbf{B}$, that is not derivable from Φ in $\mathcal{C}$.
- ▷ **Definition B.1.4.** We call a calculus $\mathcal{C}$ reasonable, iff implication elimination and conjunction introduction are admissible in $\mathcal{C}$ and $\mathbf{A} \wedge \neg\mathbf{A} \Rightarrow \mathbf{B}$ is a $\mathcal{C}$-theorem.
- ▷ **Theorem B.1.5.** *$\mathcal{C}$-inconsistency and $\mathcal{C}$-refutability coincide for reasonable calculi.*

It is very important to distinguish the syntactic $\mathcal{C}$-refutability and $\mathcal{C}$-consistency from satisfiability, which is a property of formulae that is at the heart of semantics. Note that the former have the calculus (a syntactic device) as a parameter, while the latter does not. In fact we should actually say $\mathcal{S}$-satisfiability, where $\langle\mathcal{L}, \mathcal{K}, \models\rangle$ is the current logical system.

Even the word "contradiction" has a syntactical flavor to it, it translates to "saying against each other" from its Latin root.
The notion of an "abstract consistency class" provides the a calculus-independent notion of consistency: A set Φ of sentences is considered "consistent in an abstract sense", iff it is a member of an abstract consistency class ∇.

Abstract Consistency

- **Definition B.1.6.** Let ∇ be a collection of sets. We call ∇ closed under subsets, iff for each $\Phi\in\nabla$, all subsets $\Psi\subseteq\Phi$ are elements of ∇.
- **Notation:** We will use $\Phi*\mathbf{A}$ for $\Phi\cup\{\mathbf{A}\}$.
- **Definition B.1.7.** A family $\nabla\subseteq wff_o(\Sigma_\iota,\mathcal{V}_\iota)$ of sets of formulae is called a (first-order) abstract consistency class, iff it is closed under subsets, and for each $\Phi\in\nabla$

  ∇_c) $\mathbf{A}\notin\Phi$ or $\neg\mathbf{A}\notin\Phi$ for atomic $\mathbf{A}\in wff_o(\Sigma_\iota,\mathcal{V}_\iota)$.

  $\nabla_\neg$) $\neg\neg\mathbf{A}\in\Phi$ implies $\Phi*\mathbf{A}\in\nabla$

  $\nabla_\wedge$) $\mathbf{A}\wedge\mathbf{B}\in\Phi$ implies $\Phi\cup\{\mathbf{A},\mathbf{B}\}\in\nabla$

  $\nabla_\vee$) $\neg(\mathbf{A}\wedge\mathbf{B})\in\Phi$ implies $\Phi*\neg\mathbf{A}\in\nabla$ or $\Phi*\neg\mathbf{B}\in\nabla$

  $\nabla_\forall$) If $\forall X.\mathbf{A}\in\Phi$, then $\Phi*([\mathbf{B}/X](\mathbf{A}))\in\nabla$ for each closed term $\mathbf{B}$.

  $\nabla_\exists$) If $\neg(\forall X.\mathbf{A})\in\Phi$ and c is an individual constant that does not occur in Φ, then $\Phi*\neg([c/X](\mathbf{A}))\in\nabla$

The conditions are very natural: Take for instance ∇_c, it would be foolish to call a set Φ of sentences "consistent under a complete calculus", if it contains an elementary contradiction. The next condition $\nabla_\neg$ says that if a set Φ that contains a sentence $\neg\neg\mathbf{A}$ is "consistent", then we should be able to extend it by $\mathbf{A}$ without losing this property; in other words, a complete calculus should be able to recognize $\mathbf{A}$ and $\neg\neg\mathbf{A}$ to be equivalent. We will carry out the proof here, since it gives us practice in dealing with the abstract consistency properties.
The main result here is that abstract consistency classes can be extended to compact ones. The proof is quite tedious, but relatively straightforward. It allows us to assume that all abstract consistency classes are compact in the first place (otherwise we pass to the compact extension).
Actually we are after abstract consistency classes that have an even stronger property than just being closed under subsets. This will allow us to carry out a limit construction in the Hintikka set extension argument later.

Compact Collections

- **Definition B.1.8.** We call a collection ∇ of sets compact, iff for any set Φ we have
  $\Phi\in\nabla$, iff $\Psi\in\nabla$ for every finite subset Ψ of Φ.
- **Lemma B.1.9.** *If ∇ is compact, then ∇ is closed under subsets.*
- *Proof:*
  1. Suppose $S\subseteq T$ and $T\in\nabla$.
  2. Every finite subset A of S is a finite subset of T.
  3. As ∇ is compact, we know that $A\in\nabla$.

4. Thus $S \in \nabla$.

The property of being closed under subsets is a "downwards-oriented" property: We go from large sets to small sets, compactness (the interesting direction anyways) is also an "upwards-oriented" property. We can go from small (finite) sets to large (infinite) sets. The main application for the compactness condition will be to show that infinite sets of formulae are in a collection ∇ by testing all their finite subsets (which is much simpler).

Compact Abstract Consistency Classes

▷ **Lemma B.1.10.** *Any first-order abstract consistency class can be extended to a compact one.*

▷ *Proof:*

1. We choose $\nabla' := \{\Phi \subseteq cwff_o(\Sigma_\iota) | \text{every finite subset of } \Phi \text{is in } \nabla\}$.
2. Now suppose that $\Phi \in \nabla$. ∇ is closed under subsets, so every finite subset of Φ is in ∇ and thus $\Phi \in \nabla'$. Hence $\nabla \subseteq \nabla'$.
3. Let us now show that each ∇ is compact.'
   3.1. Suppose $\Phi \in \nabla'$ and Ψ is an arbitrary finite subset of Φ.
   3.2. By definition of ∇' all finite subsets of Φ are in ∇ and therefore $\Psi \in \nabla'$.
   3.3. Thus all finite subsets of Φ are in ∇' whenever Φ is in ∇'.
   3.4. On the other hand, suppose all finite subsets of Φ are in ∇'.
   3.5. Then by the definition of ∇' the finite subsets of Φ are also in ∇, so $\Phi \in \nabla'$. Thus ∇' is compact.
4. Note that ∇' is closed under subsets by the Lemma above.
5. Next we show that if ∇ satisfies ∇_*, then ∇ satisfies ∇_*.'
   5.1. To show ∇_c, let $\Phi \in \nabla'$ and suppose there is an atom $\mathbf{A}$, such that $\{\mathbf{A}, \neg\mathbf{A}\} \subseteq \Phi$. Then $\{\mathbf{A}, \neg\mathbf{A}\} \in \nabla$ contradicting ∇_c.
   5.2. To show $\nabla_\neg$, let $\Phi \in \nabla'$ and $\neg\neg\mathbf{A} \in \Phi$, then $\Phi * \mathbf{A} \in \nabla'$.
      5.2.1. Let Ψ be any finite subset of $\Phi * \mathbf{A}$, and $\Theta := (\Psi \backslash \{\mathbf{A}\}) * \neg\neg\mathbf{A}$.
      5.2.2. Θ is a finite subset of Φ, so $\Theta \in \nabla$.
      5.2.3. Since ∇ is an abstract consistency class and $\neg\neg\mathbf{A} \in \Theta$, we get $\Theta * \mathbf{A} \in \nabla$ by $\nabla_\neg$.
      5.2.4. We know that $\Psi \subseteq \Theta * \mathbf{A}$ and ∇ is closed under subsets, so $\Psi \in \nabla$.
      5.2.5. Thus every finite subset Ψ of $\Phi * \mathbf{A}$ is in ∇ and therefore by definition $\Phi * \mathbf{A} \in \nabla'$.
   5.3. the other cases are analogous to $\nabla_\neg$.

Hintikka sets are sets of sentences with very strong analytic closure conditions. These are motivated as maximally consistent sets i.e. sets that already contain everything that can be consistently added to them.

∇-Hintikka Set

▷ **Definition B.1.11.** Let ∇ be an abstract consistency class, then we call a set $\mathcal{H} \in \nabla$ a ∇ Hintikka Set, iff $\mathcal{H}$ is maximal in ∇, i.e. for all $\mathbf{A}$ with $\mathcal{H} * \mathbf{A} \in \nabla$ we already have $\mathbf{A} \in \mathcal{H}$.

▷ **Theorem B.1.12 (Hintikka Properties).** *Let ∇ be an abstract consistency class and $\mathcal{H}$ be a ∇-Hintikka set, then*

$\mathcal{H}_c$) *For all* $\mathbf{A}\in wf\!f_o(\Sigma_\iota, \mathcal{V}_\iota)$ *we have* $\mathbf{A}\notin\mathcal{H}$ *or* $\neg\mathbf{A}\notin\mathcal{H}$.

$\mathcal{H}_\neg$) *If* $\neg\neg\mathbf{A}\in\mathcal{H}$ *then* $\mathbf{A}\in\mathcal{H}$.

$\mathcal{H}_\wedge$) *If* $\mathbf{A}\wedge\mathbf{B}\in\mathcal{H}$ *then* $\mathbf{A}, \mathbf{B}\in\mathcal{H}$.

$\mathcal{H}_\vee$) *If* $\neg(\mathbf{A}\wedge\mathbf{B})\in\mathcal{H}$ *then* $\neg\mathbf{A}\in\mathcal{H}$ *or* $\neg\mathbf{B}\in\mathcal{H}$.

$\mathcal{H}_\forall$) *If* $\forall X.\mathbf{A}\in\mathcal{H}$, *then* $[\mathbf{B}/X](\mathbf{A})\in\mathcal{H}$ *for each closed term* $\mathbf{B}$.

$\mathcal{H}_\exists$) *If* $\neg(\forall X.\mathbf{A})\in\mathcal{H}$ *then* $\neg([\mathbf{B}/X](\mathbf{A}))\in\mathcal{H}$ *for some term closed term* $\mathbf{B}$.

▷ *Proof:*

We prove the properties in turn $\mathcal{H}_c$ *goes by induction on the structure of* $\mathbf{A}$

1. $\mathbf{A}$ atomic
   1.1. Then $\mathbf{A}\notin\mathcal{H}$ or $\neg\mathbf{A}\notin\mathcal{H}$ by ∇_c.
2. $\mathbf{A} = \neg\mathbf{B}$
   2.1. Let us assume that $\neg\mathbf{B}\in\mathcal{H}$ and $\neg\neg\mathbf{B}\in\mathcal{H}$,
   2.2. then $\mathcal{H}*\mathbf{B}\in\nabla$ by $\nabla_\neg$, and therefore $\mathbf{B}\in\mathcal{H}$ by maximality.
   2.3. So $\{\mathbf{B}, \neg\mathbf{B}\}\subseteq\mathcal{H}$, which contradicts the inductive hypothesis.
3. $\mathbf{A} = \mathbf{B}\vee\mathbf{C}$ *similar to the previous case*
4. We prove $\mathcal{H}_\neg$ by maximality of $\mathcal{H}$ in ∇.
   4.1. If $\neg\neg\mathbf{A}\in\mathcal{H}$, then $\mathcal{H}*\mathbf{A}\in\nabla$ by $\nabla_\neg$.
   4.2. The maximality of $\mathcal{H}$ now gives us that $\mathbf{A}\in\mathcal{H}$.
5. The other $\mathcal{H}_*$ are similar

The following theorem is one of the main results in the "abstract consistency"/"model existence" method. For any abstract consistent set Φ it allows us to construct a Hintikka set $\mathcal{H}$ with $\Phi\in\mathcal{H}$.

Extension Theorem

▷ **Theorem B.1.13.** *If* ∇ *is an abstract consistency class and* $\Phi\in\nabla$ *finite, then there is a* ∇*-Hintikka set* $\mathcal{H}$ *with* $\Phi\subseteq\mathcal{H}$.

▷ *Proof:*

1. Wlog. assume that ∇ compact (else use compact extension)
2. Choose an enumeration $\mathbf{A}_1, \ldots$ of $cwf\!f_o(\Sigma_\iota)$ and $c_1, \ldots$ of Σ_0^{sk}.
3. and construct a sequence of sets $\mathbf{H}_i$ with $\mathbf{H}_0 := \Phi$ and

$$\mathbf{H}_{n+1} := \begin{cases} \mathbf{H}_n & \text{if } \mathbf{H}_n*\mathbf{A}_n\notin\nabla \\ \mathbf{H}_n\cup\{\mathbf{A}_n, \neg([c_n/X](\mathbf{B}))\} & \text{if } \mathbf{H}_n*\mathbf{A}_n\in\nabla \text{ and } \mathbf{A}_n = \neg(\forall X.\mathbf{B}) \\ \mathbf{H}_n*\mathbf{A}_n & \text{else} \end{cases}$$

4. Note that all $\mathbf{H}_i\in\nabla$, choose $\mathcal{H} := \bigcup_{i\in\mathbb{N}}\mathbf{H}_i$
5. $\Psi\subseteq\mathcal{H}$ finite implies there is a $j\in\mathbb{N}$ such that $\Psi\subseteq\mathbf{H}_j$,
6. so $\Psi\in\nabla$ as ∇ closed under subsets and $\mathcal{H}\in\nabla$ as ∇ is compact.
7. Let $\mathcal{H}*\mathbf{B}\in\nabla$, then there is a $j\in\mathbb{N}$ with $\mathbf{B} = \mathbf{A}_j$, so that $\mathbf{B}\in\mathbf{H}_{j+1}$ and $\mathbf{H}_{j+1}\subseteq\mathcal{H}$
8. Thus $\mathcal{H}$ is ∇-maximal

Note that the construction in the proof above is non-trivial in two respects. First, the limit construction for $\mathcal{H}$ is not executed in our original abstract consistency class ∇, but in a suitably extended one to make it compact — the original would not have contained $\mathcal{H}$ in general. Second, the set $\mathcal{H}$ is not unique for Φ, but depends on the choice of the enumeration of $cwff_o(\Sigma_\iota)$. If we pick a different enumeration, we will end up with a different $\mathcal{H}$. Say if $\mathbf{A}$ and $\neg\mathbf{A}$ are both ∇-consistent[4] with Φ, then depending on which one is first in the enumeration $\mathcal{H}$, will contain that one; with all the consequences for subsequent choices in the construction process.

Valuations

▷ **Definition B.1.14.** A function $\mu\colon cwff_o(\Sigma_\iota)\to\mathcal{D}_0$ is called a (first-order) valuation, iff

- ▷ $\mu(\neg\mathbf{A})=\mathsf{T}$, iff $\mu(\mathbf{A})=\mathsf{F}$
- ▷ $\mu(\mathbf{A}\wedge\mathbf{B})=\mathsf{T}$, iff $\mu(\mathbf{A})=\mathsf{T}$ and $\mu(\mathbf{B})=\mathsf{T}$
- ▷ $\mu(\forall X.\mathbf{A})=\mathsf{T}$, iff $\mu([\mathbf{B}/X](\mathbf{A}))=\mathsf{T}$ for all closed terms $\mathbf{B}$.

▷ **Lemma B.1.15.** *If $\varphi\colon\mathcal{V}_\iota\to D$ is a variable assignment, then $\mathcal{I}_\varphi\colon cwff_o(\Sigma_\iota)\to\mathcal{D}_0$ is a valuation.*

▷ *Proof sketch:* Immediate from the definitions

Thus a valuation is a weaker notion of evaluation in first-order logic; the other direction is also true, even though the proof of this result is much more involved: The existence of a first-order valuation that makes a set of sentences true entails the existence of a model that satisfies it.[5]

Valuation and Satisfiability

▷ **Lemma B.1.16.** *If $\mu\colon cwff_o(\Sigma_\iota)\to\mathcal{D}_0$ is a valuation and $\Phi\subseteq cwff_o(\Sigma_\iota)$ with $\mu(\Phi)=\{\mathsf{T}\}$, then Φ is satisfiable.*

▷ *Proof:* We construct a model for Φ.

1. Let $\mathcal{D}_\iota:=cwff_\iota(\Sigma_\iota)$, and
   - ▷ $\mathcal{I}(f)\colon{\mathcal{D}_\iota}^k\to\mathcal{D}_\iota;\langle\mathbf{A}_1,\ldots,\mathbf{A}_k\rangle\mapsto f(\mathbf{A}_1,\ldots,\mathbf{A}_k)$ for $f\in\Sigma^f$
   - ▷ $\mathcal{I}(p)\colon{\mathcal{D}_\iota}^k\to\mathcal{D}_0;\langle\mathbf{A}_1,\ldots,\mathbf{A}_k\rangle\mapsto\mu(p(\mathbf{A}_1,\ldots,\mathbf{A}_k))$ for $p\in\Sigma^p$.
2. Then variable assignments into $\mathcal{D}_\iota$ are ground substitutions.
3. We show $\mathcal{I}_\varphi(\mathbf{A})=\varphi(\mathbf{A})$ for $\mathbf{A}\in wff_\iota(\Sigma_\iota,\mathcal{V}_\iota)$ by induction on $\mathbf{A}$:

   3.1. $\mathbf{A}=X$

   3.1.1. then $\mathcal{I}_\varphi(\mathbf{A})=\varphi(X)$ by definition.

   3.2. $\mathbf{A}=f(\mathbf{A}_1,\ldots,\mathbf{A}_k)$

   3.2.1. then $\mathcal{I}_\varphi(\mathbf{A})=\mathcal{I}(f)(\mathcal{I}_\varphi(\mathbf{A}_1),\ldots,\mathcal{I}_\varphi(\mathbf{A}_n))=\mathcal{I}(f)(\varphi(\mathbf{A}_1),\ldots,\varphi(\mathbf{A}_n))=f(\varphi(\mathbf{A}_1),\ldots,\varphi(\mathbf{A}_n))=\varphi(f(\mathbf{A}_1,\ldots,\mathbf{A}_k))=\varphi(\mathbf{A})$

   We show $\mathcal{I}_\varphi(\mathbf{A})=\mu(\varphi(\mathbf{A}))$ for $\mathbf{A}\in wff_o(\Sigma_\iota,\mathcal{V}_\iota)$ by induction on $\mathbf{A}$.

   3.3. $\mathbf{A}=p(\mathbf{A}_1,\ldots,\mathbf{A}_k)$

   3.3.1. then $\mathcal{I}_\varphi(\mathbf{A})=\mathcal{I}(p)(\mathcal{I}_\varphi(\mathbf{A}_1),\ldots,\mathcal{I}_\varphi(\mathbf{A}_n))=\mathcal{I}(p)(\varphi(\mathbf{A}_1),\ldots,\varphi(\mathbf{A}_n))=\mu(p(\varphi(\mathbf{A}_1),\ldots,\varphi(\mathbf{A}_n)))=\mu(\varphi(p(\mathbf{A}_1,\ldots,\mathbf{A}_k)))=\mu(\varphi(\mathbf{A}))$

[4] EdNote: introduce this above
[5] EdNote: I think that we only get a semivaluation, look it up in Andrews.

3.4. $\mathbf{A} = \neg\mathbf{B}$
3.4.1. then $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{T}$, iff $\mathcal{I}_\varphi(\mathbf{B}) = \mu(\varphi(\mathbf{B})) = \mathsf{F}$, iff $\mu(\varphi(\mathbf{A})) = \mathsf{T}$.
3.5. $\mathbf{A} = \mathbf{B} \wedge \mathbf{C}$
3.5.1. similar
3.6. $\mathbf{A} = \forall X.\mathbf{B}$
3.6.1. then $\mathcal{I}_\varphi(\mathbf{A}) = \mathsf{T}$, iff $\mathcal{I}_\psi(\mathbf{B}) = \mu(\psi(\mathbf{B})) = \mathsf{T}$, for all $\mathbf{C}{\in}\mathcal{D}_\iota$, where $\psi = \varphi,[\mathbf{C}/X]$. This is the case, iff $\mu(\varphi(\mathbf{A})) = \mathsf{T}$.
4. Thus $\mathcal{I}_\varphi(\mathbf{A})\mu(\varphi(\mathbf{A})) = \mu(\mathbf{A}) = \mathsf{T}$ for all $\mathbf{A}{\in}\Phi$.
5. Hence $\mathcal{M}{\models}\mathbf{A}$ for $\mathcal{M}{:=}\langle\mathcal{D}_\iota,\mathcal{I}\rangle$.

Now, we only have to put the pieces together to obtain the model existence theorem we are after.

Model Existence

▷ **Theorem B.1.17 (Hintikka-Lemma).** *If ∇ is an abstract consistency class and $\mathcal{H}$ a ∇-Hintikka set, then $\mathcal{H}$ is satisfiable.*

▷ *Proof:*

1. we define $\mu(\mathbf{A}){:=}\mathsf{T}$, iff $\mathbf{A}{\in}\mathcal{H}$,
2. then μ is a valuation by the Hintikka set properties.
3. We have $\mu(\mathcal{H}) = \{\mathsf{T}\}$, so $\mathcal{H}$ is satisfiable.

▷ **Theorem B.1.18 (Model Existence).** *If ∇ is an abstract consistency class and $\Phi{\in}\nabla$, then Φ is satisfiable.*

Proof:

▷ 1. There is a ∇-Hintikka set $\mathcal{H}$ with $\Phi \subseteq \mathcal{H}$ (Extension Theorem)
2. We know that $\mathcal{H}$ is satisfiable. (Hintikka-Lemma)
3. In particular, $\Phi \subseteq \mathcal{H}$ is satisfiable.

B.2 A Completeness Proof for First-Order ND

With the model existence proof we have introduced in the last section, the completeness proof for first-order natural deduction is rather simple, we only have to check that ND-consistency is an abstract consistency property.

Consistency, Refutability and Abstract Consistency

▷ **Theorem B.2.1 (Non-Refutability is an Abstract Consistency Property).** $\Gamma{:=}\{\Phi \subseteq cwff_o(\Sigma_\iota) | \Phi \text{ not } \mathcal{ND}^1\text{−refutable}\}$ *is an abstract consistency class.*

▷ *Proof:* We check the properties of an ACC

1. If Φ is non-refutable, then any subset is as well, so Γ is closed under subsets.
We show the abstract consistency conditions ∇_* for $\Phi{\in}\Gamma$.
2. ∇_c
2.1. We have to show that $\mathbf{A}{\notin}\Phi$ or $\neg\mathbf{A}{\notin}\Phi$ for atomic $\mathbf{A}{\in}wff_o(\Sigma_\iota,\mathcal{V}_\iota)$.
2.2. Equivalently, we show the contrapositive: If $\{\mathbf{A},\neg\mathbf{A}\} \subseteq \Phi$, then $\Phi{\notin}\Gamma$.

2.3. So let $\{\mathbf{A}, \neg\mathbf{A}\} \subseteq \Phi$, then Φ is $\mathcal{ND}^1$-refutable by construction.
2.4. So $\Phi \notin \Gamma$.

3. $\nabla_\neg$ *We show the contrapositive again*
   3.1. Let $\neg\neg\mathbf{A} \in \Phi$ and $\Phi * \mathbf{A} \notin \Gamma$
   3.2. Then we have a refutation $\mathcal{D}\colon \Phi * \mathbf{A} \vdash_{\mathcal{ND}^1} F$
   3.3. By prepending an application of $\neg E$ for $\neg\neg\mathbf{A}$ to $\mathcal{D}$, we obtain a refutation $\mathcal{D}\colon \Phi \vdash_{\mathcal{ND}^1} F'$.
   3.4. Thus $\Phi \notin \Gamma$.

Proof sketch: *other* ∇_* *similar*

This directly yields two important results that we will use for the completeness analysis.

Henkin's Theorem

▷ **Corollary B.2.2 (Henkin's Theorem).** *Every $\mathcal{ND}^1$-consistent set of sentences has a model.*

▷ *Proof:*
1. Let Φ be a $\mathcal{ND}^1$-consistent set of sentences.
2. The class of sets of $\mathcal{ND}^1$-consistent propositions constitute an abstract consistency class.
3. Thus the model existence theorem guarantees a model for Φ.

▷ **Corollary B.2.3 (Löwenheim&Skolem Theorem).** *Satisfiable set Φ of first-order sentences has a countable model.*

Proof sketch: The model we constructed is countable, since the set of ground terms is.

Now, the completeness result for first-order natural deduction is just a simple argument away. We also get a compactness theorem (almost) for free: logical systems with a complete calculus are always compact.

▷ Completeness and Compactness

▷ **Theorem B.2.4 (Completeness Theorem for $\mathcal{ND}^1$).** *If $\Phi \models \mathbf{A}$, then $\Phi \vdash_{\mathcal{ND}^1} \mathbf{A}$.*

▷ *Proof:* We prove the result by playing with negations.
1. If $\mathbf{A}$ is valid in all models of Φ, then $\Phi * \neg\mathbf{A}$ has no model
2. Thus $\Phi * \neg\mathbf{A}$ is inconsistent by (the contrapositive of) Henkins Theorem.
3. So $\Phi \vdash_{\mathcal{ND}^1} \neg\neg\mathbf{A}$ by $\neg I$ and thus $\Phi \vdash_{\mathcal{ND}^1} \mathbf{A}$ by $\neg E$.

▷ **Theorem B.2.5 (Compactness Theorem for first-order logic).** *If $\Phi \models \mathbf{A}$, then there is already a finite set $\Psi \subseteq \Phi$ with $\Psi \models \mathbf{A}$.*

Proof: This is a direct consequence of the completeness theorem

▷ 1. We have $\Phi \models \mathbf{A}$, iff $\Phi \vdash_{\mathcal{ND}^1} \mathbf{A}$.

2. As a proof is a finite object, only a finite subset $\Psi \subseteq \Phi$ can appear as leaves in the proof.

B.3 Soundness and Completeness of First-Order Tableaux

The soundness of the first-order free-variable tableaux calculus can be established a simple induction over the size of the tableau.

Soundness of $\mathcal{T}_1^f$

▷ **Lemma B.3.1.** *Tableau rules transform satisfiable tableaux into satisfiable ones.*

▷ *Proof:*

we examine the tableau rules in turn

1. propositional rules *as in propositional tableaux*
2. $\mathcal{T}_1^f\exists$ *by* **??**
3. $\mathcal{T}_1^f\bot$ *by* **??** *(substitution value lemma)*
4. $\mathcal{T}_1^f\forall$
   4.1. $\mathcal{I}_\varphi(\forall X.\mathbf{A}) = \mathsf{T}$, iff $\mathcal{I}_\psi(\mathbf{A}) = \mathsf{T}$ for all $a \in \mathcal{D}_\iota$
   4.2. so in particular for some $a \in \mathcal{D}_\iota \neq \emptyset$.

▷ **Corollary B.3.2.** $\mathcal{T}_1^f$ *is correct.*

The only interesting steps are the cut rule, which can be directly handled by the substitution value lemma, and the rule for the existential quantifier, which we do in a separate lemma.

Soundness of $\mathcal{T}_1^f\exists$

▷ **Lemma B.3.3.** $\mathcal{T}_1^f\exists$ *transforms satisfiable tableaux into satisfiable ones.*

▷ *Proof:* Let $\mathcal{T}'$ be obtained by applying $\mathcal{T}_1^f\exists$ to $(\forall X.\mathbf{A})^{\mathsf{F}}$ in $\mathcal{T}$, extending it with $([f(X^1, \ldots, X^k)/X](\mathbf{A}))^{\mathsf{F}}$, where $W{:=}\mathrm{free}(\forall X.\mathbf{A}) = \{X^1, \ldots, X^k\}$

1. Let $\mathcal{T}$ be satisfiable in $\mathcal{M}{:=}\langle\mathcal{D}, \mathcal{I}\rangle$, then $\mathcal{I}_\varphi(\forall X.\mathbf{A}) = \mathsf{F}$.
   We need to find a model $\mathcal{M}'$ that satisfies $\mathcal{T}'$ (*find interpretation for f*)
2. By definition $\mathcal{I}_{\varphi,[a/X]}(\mathbf{A}) = \mathsf{F}$ for some $a \in \mathcal{D}$ (depends on $\varphi|_W$)
3. Let $g \colon \mathcal{D}^k \to \mathcal{D}$ be defined by $g(a_1, \ldots, a_k){:=}a$, if $\varphi(X^i) = a_i$
4. choose $\mathcal{M} = \langle\mathcal{D}, \mathcal{I}'\rangle'$ with $\mathcal{I}'{:=}\mathcal{I},[g/f]$, then by subst. value lemma

$$\begin{aligned}\mathcal{I}'_\varphi([f(X^1, \ldots, X^k)/X](\mathbf{A})) &= \mathcal{I}'_{(\varphi,[\mathcal{I}'_\varphi(f(X^1,\ldots,X^k))/X])}(\mathbf{A}) \\ &= \mathcal{I}'_{(\varphi,[a/X])}(\mathbf{A}) = \mathsf{F}\end{aligned}$$

5. So $([f(X^1, \ldots, X^k)/X](\mathbf{A}))^{\mathsf{F}}$ satisfiable in $\mathcal{M}'$

This proof is paradigmatic for soundness proofs for calculi with Skolemization. We use the axiom of choice at the meta-level to choose a meaning for the Skolem function symbol.

Armed with the Model Existence Theorem for first-order logic (Theorem B.1.18), the completeness of first-order tableaux is similarly straightforward. We just have to show that the collection of tableau-irrefutable sentences is an abstract consistency class, which is a simple proof-transformation exercise in all but the universal quantifier case, which we postpone to its own Lemma (Theorem B.3.5).

Completeness of $(\mathcal{T}_1^f)$

▷ **Theorem B.3.4.** *$\mathcal{T}_1^f$ is refutation complete.*

▷ *Proof:* We show that $\nabla{:=}\{\Phi|\Phi^{\mathsf{T}}$ has no closed Tableau$\}$ is an abstract consistency class

1. as for propositional case.
2. by the lifting lemma below
3. Let $\mathcal{T}$ be a closed tableau for $\neg(\forall X.\mathbf{A})\in\Phi$ and $\Phi^{\mathsf{T}}*([c/X](\mathbf{A}))^{\mathsf{F}}\in\nabla$.

$$\begin{array}{c} \Psi^{\mathsf{T}} \\ (\forall X.\mathbf{A})^{\mathsf{F}} \\ ([c/X](\mathbf{A}))^{\mathsf{F}} \\ Rest \end{array} \qquad \begin{array}{c} \Psi^{\mathsf{T}} \\ (\forall X.\mathbf{A})^{\mathsf{F}} \\ ([f(X_1,\ldots,X_k)/X](\mathbf{A}))^{\mathsf{F}} \\ [f(X_1,\ldots,X_k)/c](Rest) \end{array}$$

So we only have to treat the case for the universal quantifier. This is what we usually call a "lifting argument", since we have to transform ("lift") a proof for a formula $\theta(\mathbf{A})$ to one for $\mathbf{A}$. In the case of tableaux we do that by an induction on the tableau refutation for $\theta(\mathbf{A})$ which creates a tableau-isomorphism to a tableau refutation for $\mathbf{A}$.

Tableau-Lifting

▷ **Theorem B.3.5.** *If $\mathcal{T}_\theta$ is a closed tableau for a set $\theta(\Phi)$ of formulae, then there is a closed tableau $\mathcal{T}$ for Φ.*

▷ *Proof:* by induction over the structure of $\mathcal{T}_\theta$ we build an isomorphic tableau $\mathcal{T}$, and a tableau-isomorphism $\omega\colon\mathcal{T}\to\mathcal{T}_\theta$, such that $\omega(\mathbf{A})=\theta(\mathbf{A})$.

only the tableau-substitution rule is interesting.

1. Let $(\theta(\mathbf{A}_i))^{\mathsf{T}}$ and $(\theta(\mathbf{B}_i))^{\mathsf{F}}$ cut formulae in the branch Θ_θ^i of $\mathcal{T}_\theta$
2. there is a joint unifier σ of $(\theta(\mathbf{A}_1))=^?(\theta(\mathbf{B}_1))\wedge\ldots\wedge(\theta(\mathbf{A}_n))=^?(\theta(\mathbf{B}_n))$
3. thus $\sigma\circ\theta$ is a unifier of $\mathbf{A}$ and $\mathbf{B}$
4. hence there is a most general unifier ρ of $\mathbf{A}_1=^?\mathbf{B}_1\wedge\ldots\wedge\mathbf{A}_n=^?\mathbf{B}_n$
5. so Θ is closed.

Again, the "lifting lemma for tableaux" is paradigmatic for lifting lemmata for other refutation calculi.

B.4 Soundness and Completeness of First-Order Resolution

Correctness (CNF)

▷ **Lemma B.4.1.** *A set Φ of sentences is satisfiable, iff $CNF_1(\Phi)$ is.*

▷ *Proof:* propositional rules and ∀-rule are trivial; do the ∃-rule

1. Let $(\forall X.\mathbf{A})^{\mathsf{F}}$ satisfiable in $\mathcal{M}{:=}\langle\mathcal{D},\mathcal{I}\rangle$ and $\mathrm{free}(\mathbf{A}) = \{X^1,\ldots,X^n\}$
2. $\mathcal{I}_\varphi(\forall X.\mathbf{A}) = \mathsf{F}$, so there is an $a{\in}\mathcal{D}$ with $\mathcal{I}_{\varphi,[a/X]}(\mathbf{A}) = \mathsf{F}$ (only depends on $\varphi|_{\mathrm{free}(\mathbf{A})}$)
3. let $g\colon \mathcal{D}^n{\to}\mathcal{D}$ be defined by $g(a_1,\ldots,a_n){:=}a$, iff $\varphi(X^i) = a_i$.
4. choose $\mathcal{M}'{:=}\langle\mathcal{D},\mathcal{I}'\rangle$ with $\mathcal{I}(f)'{:=}g$, then $\mathcal{I}'_\varphi([f(X^1,\ldots,X^k)/X](\mathbf{A})) = \mathsf{F}$
5. Thus $([f(X^1,\ldots,X^k)/X](\mathbf{A}))^{\mathsf{F}}$ is satisfiable in $\mathcal{M}'$

Resolution (Correctness)

▷ **Definition B.4.2.** A clause is called satisfiable, iff $\mathcal{I}_\varphi(\mathbf{A}) = \alpha$ for one of its literals $\mathbf{A}^\alpha$.

▷ **Lemma B.4.3.** *□ is unsatisfiable*

▷ **Lemma B.4.4.** *CNF transformations preserve satisfiability* (see above)

▷ **Lemma B.4.5.** *Resolution and factorization too!*

Completeness ($\mathcal{R}_1$)

▷ **Theorem B.4.6.** *$\mathcal{R}_1$ is refutation complete.*

▷ *Proof:* $\nabla{:=}\{\Phi|\Phi^{\mathsf{T}}$ has no closed tableau$\}$ is an abstract consistency class

1. as for propositional case.
2. by the lifting lemma below
3. Let $\mathcal{T}$ be a closed tableau for $\neg(\forall X.\mathbf{A}){\in}\Phi$ and $\Phi^{\mathsf{T}}{*}([c/X](\mathbf{A}))^{\mathsf{F}}{\in}\nabla$.
4. $CNF_1(\Phi^{\mathsf{T}}) = CNF_1(\Psi^{\mathsf{T}}) \cup CNF_1((([f(X_1,\ldots,X_k)/X](\mathbf{A}))^{\mathsf{F}})$
5. $([f(X_1,\ldots,X_k)/c](CNF_1(\Phi^{\mathsf{T}}))){*}([c/X](\mathbf{A}))^{\mathsf{F}} = CNF_1(\Phi^{\mathsf{T}})$
6. so $\mathcal{R}_1 : CNF_1(\Phi^{\mathsf{T}}){\vdash_{\mathcal{D}'}}\square$, where $\mathcal{D} = [f(X_1',\ldots,X_k')/c](\mathcal{D})$.

Clause Set Isomorphism

▷ **Definition B.4.7.** Let $\mathbf{B}$ and $\mathbf{C}$ be clauses, then a clause isomorphism $\omega\colon \mathbf{C}{\to}\mathbf{D}$

is a bijection of the literals of $\mathbf{C}$ and $\mathbf{D}$, such that $\omega(\mathbf{L})^{\alpha} = \mathbf{M}^{\alpha}$ (conserves labels)
We call ω θ compatible, iff $\omega(\mathbf{L}^{\alpha}) = (\theta(\mathbf{L}))^{\alpha}$

▷ **Definition B.4.8.** Let Φ and Ψ be clause sets, then we call a bijection $\Omega\colon \Phi \to \Psi$ a clause set isomorphism, iff there is a clause isomorphism $\omega\colon \mathbf{C} \to \Omega(\mathbf{C})$ for each $\mathbf{C} \in \Phi$.

▷ **Lemma B.4.9.** *If $\theta(\Phi)$ is set of formulae, then there is a θ-compatible clause set isomorphism $\Omega\colon CNF_1(\Phi) \to CNF_1(\theta(\Phi))$.*

▷ *Proof sketch:* by induction on the CNF derivation of $CNF_1(\Phi)$.

Lifting for $\mathcal{R}_1$

▷ **Theorem B.4.10.** *If $\mathcal{R}_1 \colon (\theta(\Phi)) \vdash_{\mathcal{D}_\theta} \Box$ for a set $\theta(\Phi)$ of formulae, then there is a $\mathcal{R}_1$-refutation for Φ.*

▷ *Proof:* by induction over $\mathcal{D}_\theta$ we construct a $\mathcal{R}_1$-derivation $\mathcal{R}_1 \colon \Phi \vdash_{\mathcal{D}} \mathbf{C}$ and a θ-compatible clause set isomorphism $\Omega\colon \mathcal{D} \to \mathcal{D}_\theta$

1. If $\mathcal{D}_\theta$ ends in $$\frac{\dfrac{\mathcal{D}'_\theta}{((\theta(\mathbf{A})) \vee (\theta(\mathbf{C})))^{\mathsf{T}}} \quad \dfrac{\mathcal{D}''_\theta}{(\theta(\mathbf{B}))^{\mathsf{F}} \vee (\theta(\mathbf{D}))}}{(\sigma(\theta(\mathbf{C}))) \vee (\sigma(\theta(\mathbf{B})))}\, res$$

   then we have (IH) clause isormorphisms $\omega'\colon \mathbf{A}^{\mathsf{T}} \vee \mathbf{C} \to (\theta(\mathbf{A}))^{\mathsf{T}} \vee (\theta(\mathbf{C}))$ and $\omega'\colon \mathbf{B}^{\mathsf{T}} \vee \mathbf{D} \to (\theta(\mathbf{B}))^{\mathsf{T}}, \theta(\mathbf{D})$
2. thus $\dfrac{\mathbf{A}^{\mathsf{T}} \vee \mathbf{C} \quad \mathbf{B}^{\mathsf{F}} \vee \mathbf{D}}{(\rho(\mathbf{C})) \vee (\rho(\mathbf{B}))} Res$ where $\rho = \mathbf{mgu}(\mathbf{A}, \mathbf{B})$ (exists, as $\sigma \circ \theta$ unifier)