Feature Markov Decision Processes

Marcus Hutter

Canberra, ACT, 0200, Australia

http://www.hutter1.net/

ANU

RSISE

NICTA

AGI, 6–9 March 2009, Washington DC

Abstract

General purpose intelligent learning agents cycle through (complex,non-MDP) sequences of observations, actions, and rewards. On the other hand, reinforcement learning is well-developed for small finite state Markov Decision Processes (MDPs). It is an art performed by human designers to extract the right state representation out of the bare observations, i.e. to reduce the agent setup to the MDP framework. Before we can think of mechanizing this search for suitable MDPs, we need a formal objective criterion. The main contribution in these slides is to develop such a criterion. I also integrate the various parts into one learning algorithm. Extensions to more realistic dynamic Bayesian networks are briefly discussed.

Contents

Universal AI in Perspective

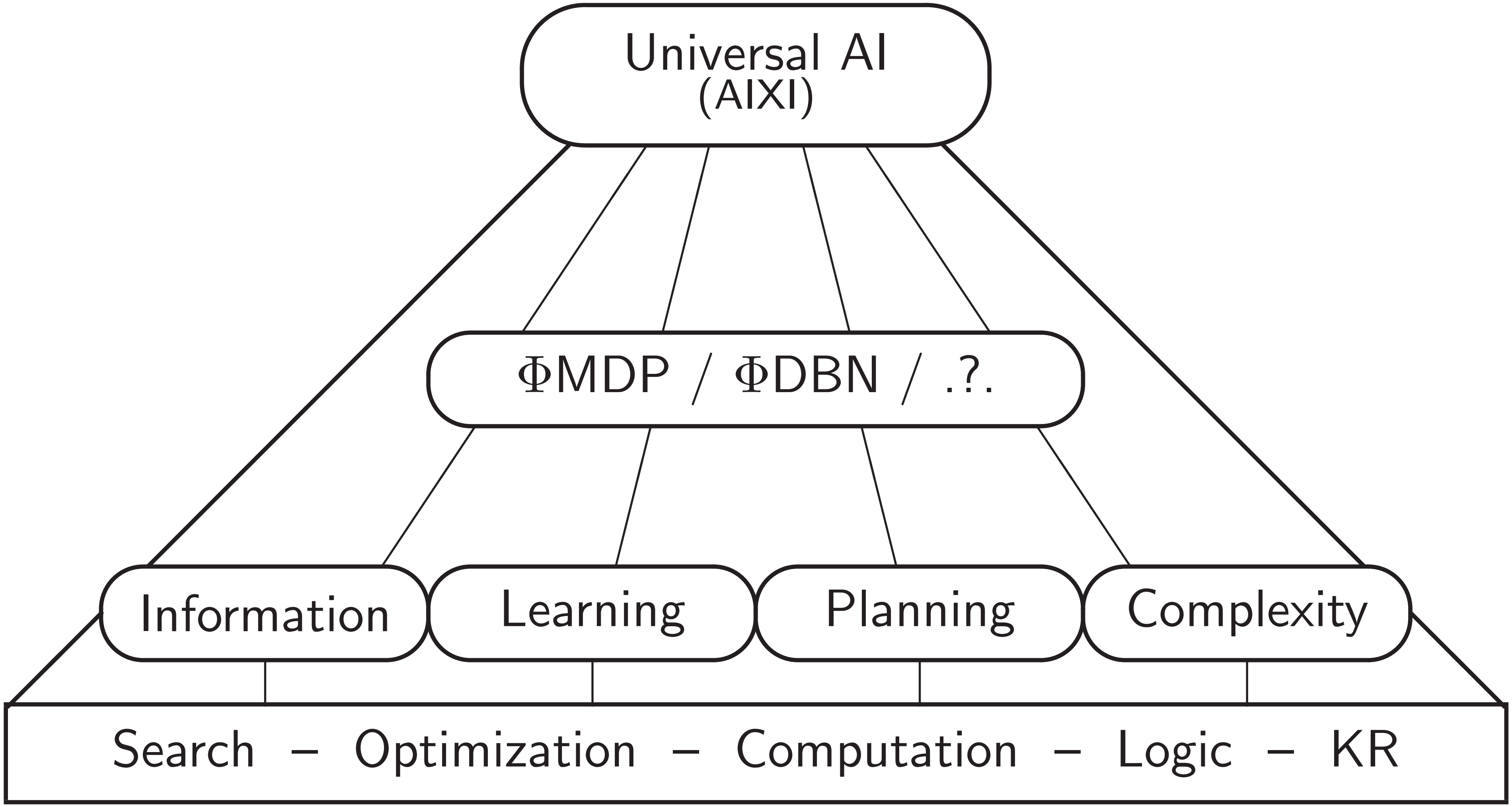
What is A(G)I?	Thinking	Acting
humanly	Cognitive Science	Turing Test
rationally	Laws of Thought	Doing the right thing

Difference matters until systems reach self-improvement threshold

- Universal AI: analytically analyzable generic learning systems
- Real world is nasty: partially unobservable, uncertain, unknown, non-ergodic, reactive, vast but luckily structured, ...
- Dealing properly with uncertainty and learning is crucial.
- Never trust a ~~theory~~ experiment if it is not supported by an ~~experiment~~ theory

Progress is achieved by an interplay between theory and experiment !

ΦMDP in Perspective



Agents = General Framework, Interface = Robots,Vision,Language

ΦMDP Overview in 1 Slide

Goal: Develop efficient general purpose intelligent agent.

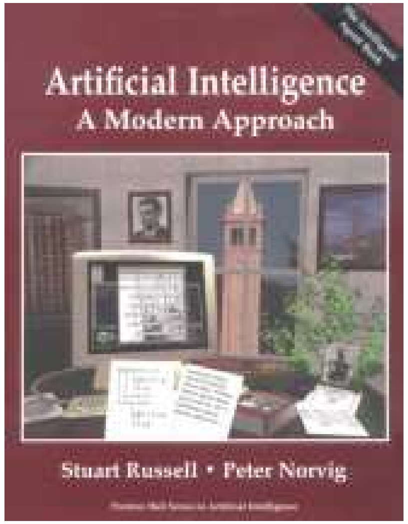
State-of-the-art: (a) AIXI: Incomputable theoretical solution.
(b) MDP: Efficient limited problem class.
(c) POMDP: Notoriously difficult. (d) PSRs: Underdeveloped.

Idea: ΦMDP reduces real problem to MDP automatically by learning.

Accomplishments so far: (i) Criterion for evaluating quality of reduction.
(ii) Integration of the various parts into one learning algorithm.
(iii) Generalization to structured MDPs (DBNs)

ΦMDP is promising path towards the grand goal & alternative to (a)-(d)

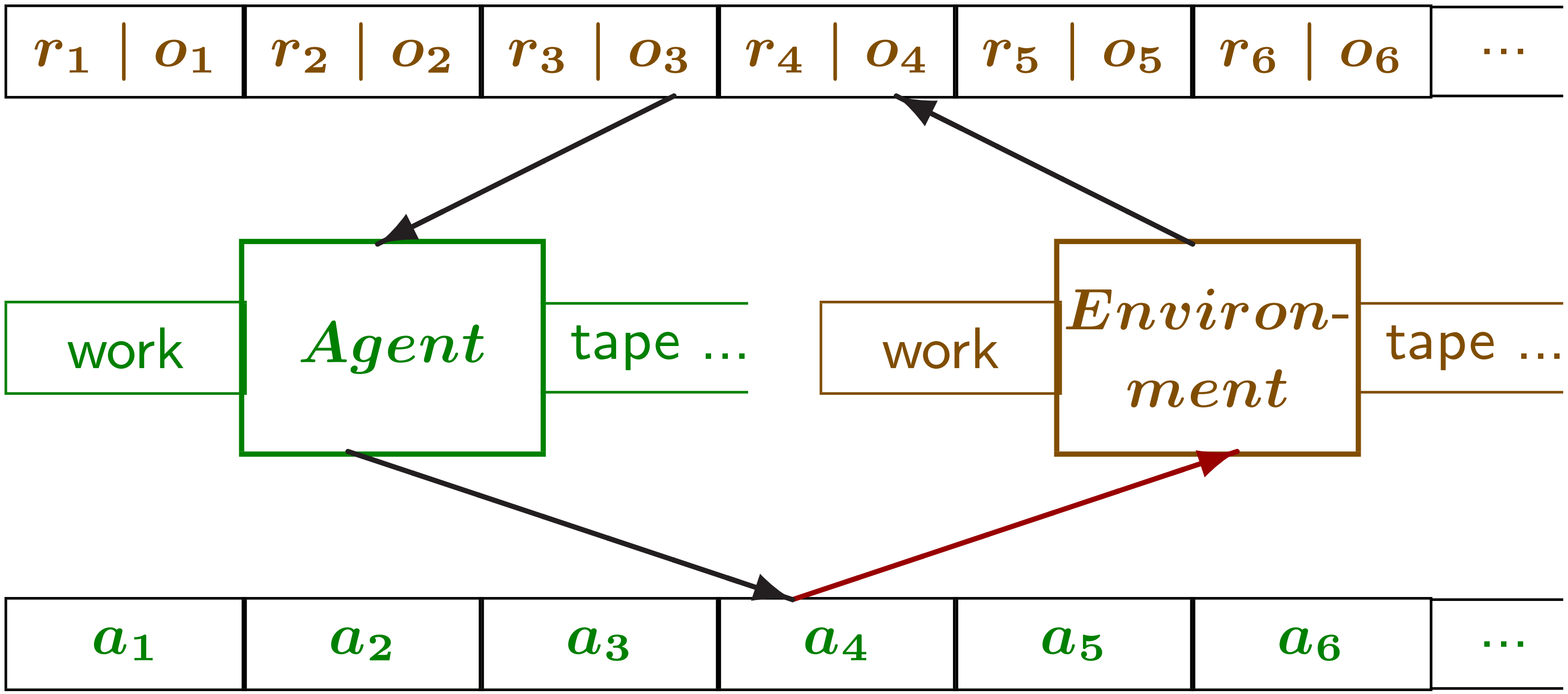
Problem: Find reduction Φ efficiently (generic optimization problem?)



Agent Model with Reward

Framework for all AI problems!

Is there a universal solution?



Types of Environments / Problems

all fit into the general Agent setup but few are MDPs

---

		sequential (prediction)	⇔	i.i.d (classification/regression)
supervised	⇔	unsupervised	⇔	reinforcement learning
		known environment	⇔	unknown environment
		planning	⇔	learning
		exploitation	⇔	exploration
		passive prediction	⇔	active learning
		Fully Observable MDP	⇔	Partially Observed MDP
		Unstructured (MDP)	⇔	Structured (DBN)
		Competitive (Multi-Agents)	⇔	Stochastic Env (Single Agent)
		Games	⇔	Optimization

Universal Artificial Intelligence

Key idea: Optimal action/plan/policy based on the simplest world model consistent with history. Formally ...

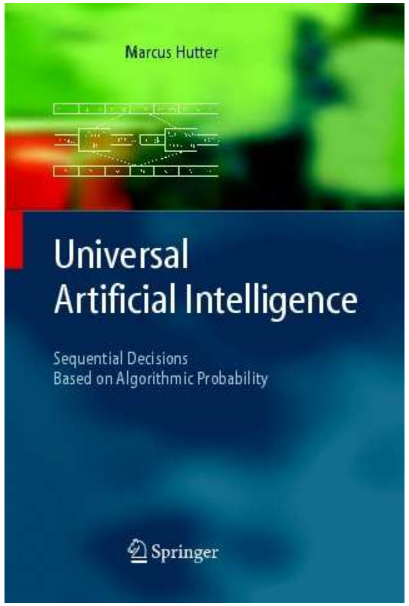
$$\text{AIXI:} \quad a_k := \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{p\,:\,U(p,a_1..a_m)=o_1 r_1..o_m r_m} 2^{-\ell(p)}$$

$a$ction, $r$eward, $o$bservation, Universal TM, $p$rogram, k=now

AIXI is an elegant, complete, essentially unique, and limit-computable mathematical theory of AI.

Claim: AIXI is the most intelligent environmental independent, i.e. universally optimal, agent possible.

Proof: For formalizations, quantifications, proofs see ⇒



Problem: Computationally intractable.

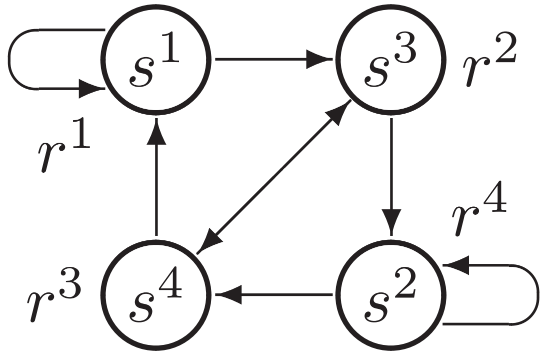
Achievement: Well-defines AGI. Gold standard to aim at. Inspired practical algorithms. Cf. infeasible exact minimax.

Markov Decision Processes (MDPs)

a computationally tractable class of problems

- MDP Assumption: State $s_t := o_t$ and r_t are probabilistic functions of o_{t-1} and a_{t-1} only.

Example MDP



- Further Assumption:
  State=observation space $\mathcal{S}$ is finite and small.

- Goal: Maximize long-term expected reward.

- Learning: Probability distribution is unknown but can be learned.

- Exploration: Optimal exploration is intractable
  but there are polynomial approximations.

- Problem: Real problems are not of this simple form.

Map Real Problem to MDP

Map history $h_t := o_1 a_1 r_1 ... o_{t-1}$ to state $s_t := \Phi(h_t)$, for example:

Games: Full-information with static opponent: $\Phi(h_t) = o_t$.

Classical physics: Position+velocity of objects = position at two time-slices: $s_t = \Phi(h_t) = o_t o_{t-1}$ is (2nd order) Markov.

I.i.d. processes of unknown probability (e.g. clinical trials $\simeq$ Bandits), Frequency of obs. $\Phi(h_n) = (\sum_{t=1}^n \delta_{o_t o})_{o\in\mathcal{O}}$ is sufficient statistic.

Identity: $\Phi(h) = h$ is always sufficient, but not learnable.

Find/Learn Map Automatically

$$\Phi^{best} := \arg\min_\Phi \text{Cost}(\Phi|h_t)$$

- What is the best map/MDP? (i.e. what is the right Cost criterion?)
- Is the best MDP good enough? (i.e. is reduction always possible?)
- How to find the map Φ (i.e. minimize Cost) efficiently?

ΦMDP Cost Criterion

Reward↔State Trade-Off

- $\text{CL}(r_{1:n}|s_{1:n},a_{1:n}) :=$ optimal MDP code length of $r_{1:n}$ given $s_{1:n}$.
- Needs $\text{CL}(s_{1:n}|a_{1:n}) :=$ optimal MDP code length of $s_{1:n}$.
- Small state space $\mathcal{S}$ has short $\text{CL}(s_{1:n}|a_{1:n})$ but obscures structure of reward sequence $\Rightarrow$ $\text{CL}(r_{1:n}|s_{1:n}a_{1:n})$ large.
- Large $\mathcal{S}$ usually makes predicting=compressing $r_{1:n}$ easier, but a large model is hard to learn, i.e. the code for $s_{1:n}$ will be large

$$\text{Cost}(\Phi|h_n) := \text{CL}(s_{1:n}|a_{1:n}) + \text{CL}(r_{1:n}|s_{1:n},a_{1:n})$$

is minimized for Φ that keeps all and only relevant information for predicting rewards.

- Recall $s_t := \Phi(h_t)$ and $h_t := a_1 o_1 r_1 ... o_t$.

Cost(Φ) Minimization

- Minimize Cost($\Phi|h$) by search: random, blind, informed, adaptive, local, global, population based, exhaustive, heuristic, other search.
- Most algs require a neighborhood relation between candidate Φ.
- Φ is equivalent to a partitioning of $(\mathcal{O} \times \mathcal{A} \times \mathcal{R})^*$.
- Example partitioners: Decision trees/lists/grids/etc.
- Example neighborhood: Subdivide=split or merge partitions.

Stochastic Φ-Search (Monte Carlo)

- Randomly choose a neighbor Φ' of Φ (by splitting or merging states)
- Replace Φ by Φ' for sure if Cost gets smaller or with some small probability if Cost gets larger. Repeat.

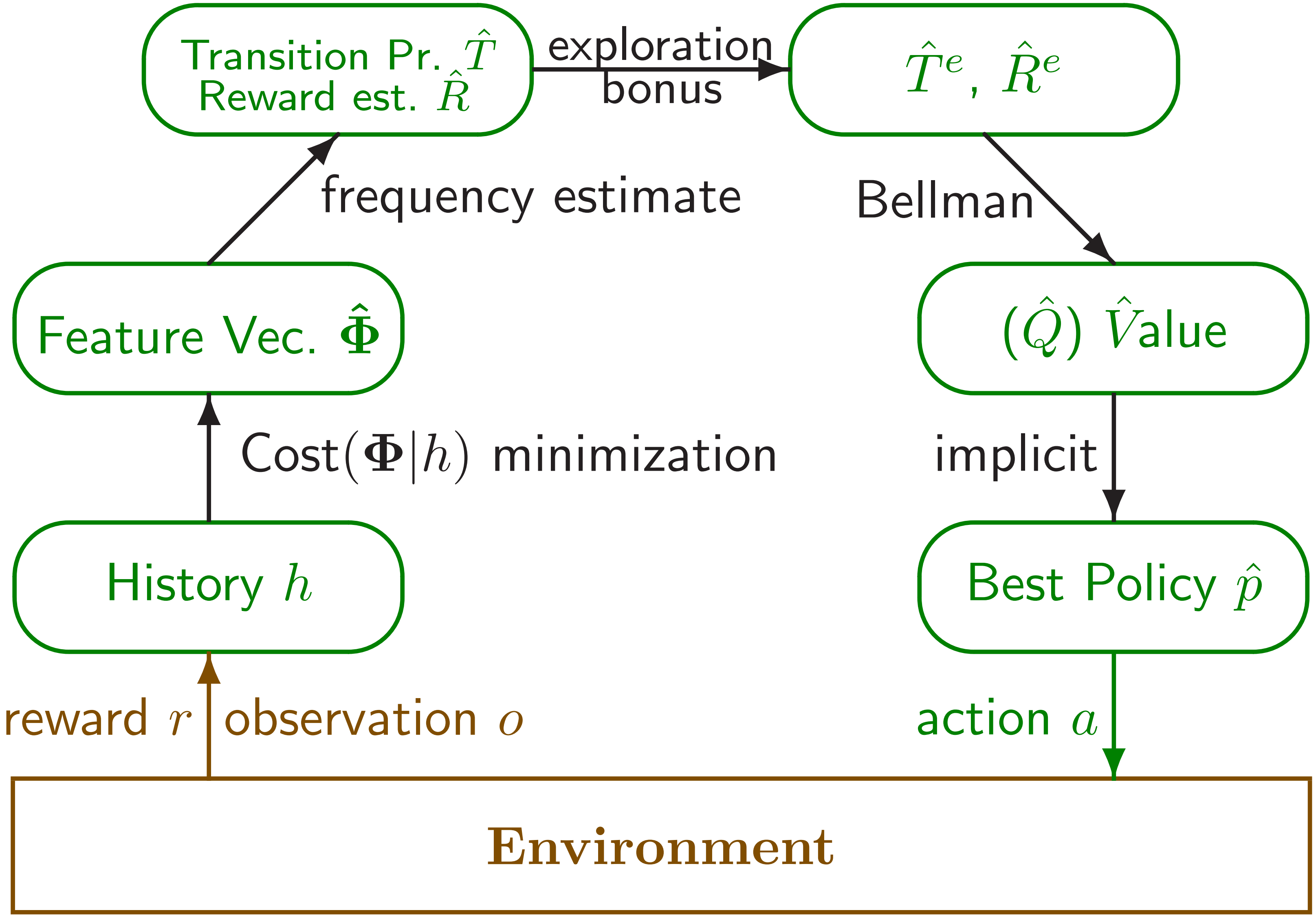
Optimal Action

- Let $\hat{\Phi}$ be a good estimate of Φ^{best}.
  $\Rightarrow$ Compressed history: $s_1 a_1 r_1 ... s_n a_n r_n \approx$ MDP sequence.
- For a finite MDP with known transition probabilities, optimal action a_{n+1} follows from Bellman equations.
- Use simple frequency estimate of transition probability and reward function $\Rightarrow$ Infamous problem ...

Exploration & Exploitation

- Polynomially optimal solutions: Rmax, E3, OIM [KS98,SL08].
- Main idea: Motivate agent to explore by pretending high-reward for unexplored state-action pairs.
- Now compute the agent's action based on modified rewards.

Computational Flow



Extension to Dynamic Bayesian Networks

- Unstructured MDPs are only suitable for relatively small problems.
- Dynamic Bayesian Networks = Structured MDPs for large problems.
- $\mathbf{\Phi}(h)$ is now vector of (loosely coupled binary) features=nodes.
- Assign global reward to local nodes by linear regression.
- $\text{Cost}(\mathbf{\Phi}, \text{Structure}|h)$ = sum of local node Costs.
- Learn optimal DBN structure in pseudo-polynomial time.
- Search for approximation $\hat{\mathbf{\Phi}}$ of $\mathbf{\Phi}^{best}$ as before.
  Neighborhood = adding/removing features.
- Use local linear value function approximation.
- Optimal action/policy by combining [KK99,KP00,SDL07,SL09].

Conclusion

Goal: Develop efficient general purpose intelligent agent.

State-of-the-art: (a) AIXI: Incomputable theoretical solution.
(b) MDP: Efficient limited problem class.
(c) POMDP: Notoriously difficult. (d) PSRs: Underdeveloped.

Idea: ΦMDP reduces real problem to MDP automatically by learning.

Accomplishments so far: (i) Criterion for evaluating quality of reduction.
(ii) Integration of the various parts into one learning algorithm.
(iii) Generalization to structured MDPs (DBNs)

ΦMDP is promising path towards the grand goal & alternative to (a)-(d)

Problem: Find reduction Φ efficiently (generic optimization problem?)

Connection to AI SubFields

- Agents: ΦMDP is a reinforcement learning (single) agent.
- Search & Optimization: Minimizing Cost$(\Phi,\text{Structure}|h)$ is a well-defined but hard (non-continuous, non-convex) optimization problem.
- Planning: More sophisticated planners needed for large DBNs.
- Information Theory: Needed for analyzing&improving Cost criterion.
- Learning: So far mainly reinforcement learning, but others relevant.
- Logic/Reasoning: For agents that reason, rule-based logical recursive partitions of domain $(\mathcal{O}\times\mathcal{A}\times\mathcal{R})^*$ are predestined.
- Knowledge Representation (KR): Searching for Φ^{best} is actually a search for the best KR. Restrict search space to reasonable KR Φ.
- Complexity Theory: We need polynomial time and ultimately linear-time approximation algorithms for all building blocks.
- Application dependent interfaces: Robotics, Vision, Language.

AGI versus NAI

Artificial General Intelligence $\leftrightarrow$ Narrow Artificial Intelligence

Lesson for NAI Students & Researchers:

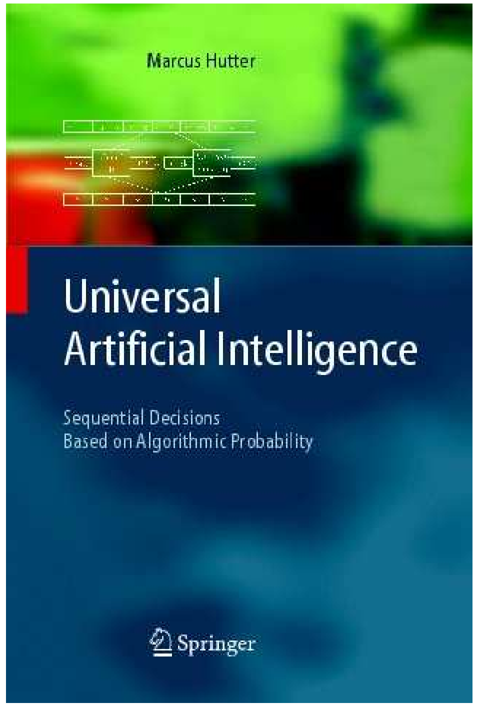
- Don't lose the big picture (if you care about real AI)
- GOFAI: Everything is uncertain ! Learning is key !
- SML: The world is not i.i.d. !

Lesson for AGI Students & Researchers:

- Do your homework (if you want to have any chance of succeeding)
- Minimum reading: Russell&Norvig (2003) book *word-by-word*.
  All references in Brian Milch (AGI'2008).
  (whatever approach you personally take)

Thanks! Questions? Details:

- M.H., *Feature Markov Decision Processes.* (AGI 2009)
- M.H., *Feature Dynamic Bayesian Networks.* (AGI 2009)
- Human Knowledge Compression Contest: (50'000€)
- M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* EATCS, Springer, 300 pages, 2005. http://www.idsia.ch/~marcus/ai/uaibook.htm



Decision Theory = Probability + Utility Theory
\+ +
Universal Induction = Ockham + Bayes + Turing
= =
A Unified View of Artificial Intelligence