Universal AI: Reinforcement Learning via AIXI Approximation

Cosmo Harrigan / Travis Mandel

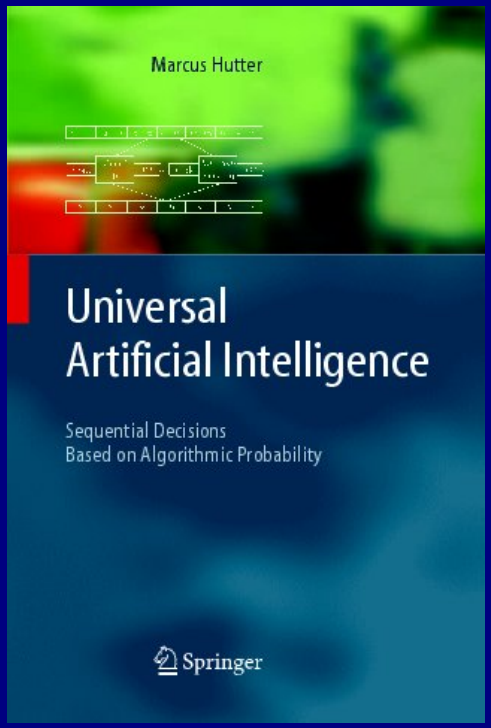
University of Washington

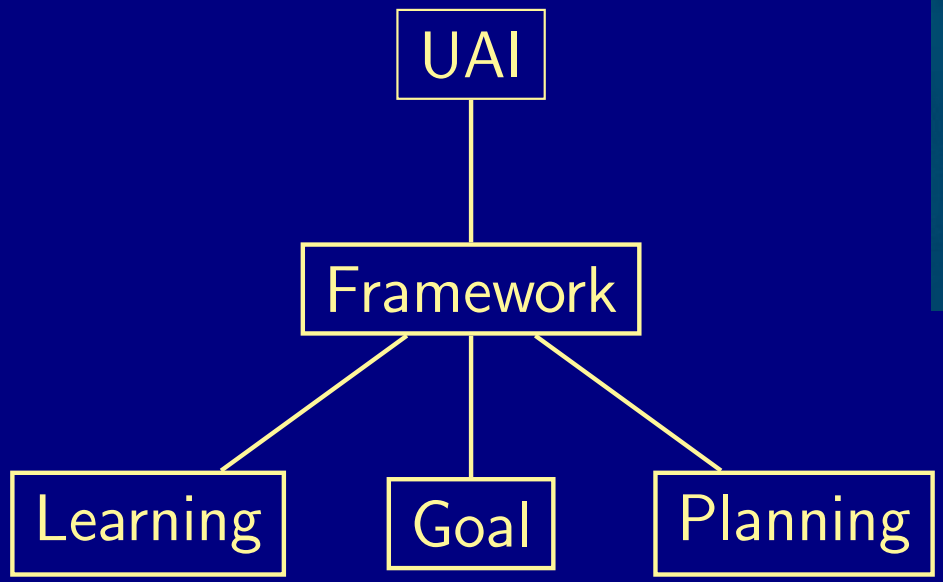
with slides from Tom Everitt (ANU) [1] [2]
`http://www.tomeveritt.se/`

August 2, 2016

Universal Artificial Intelligence (UAI)

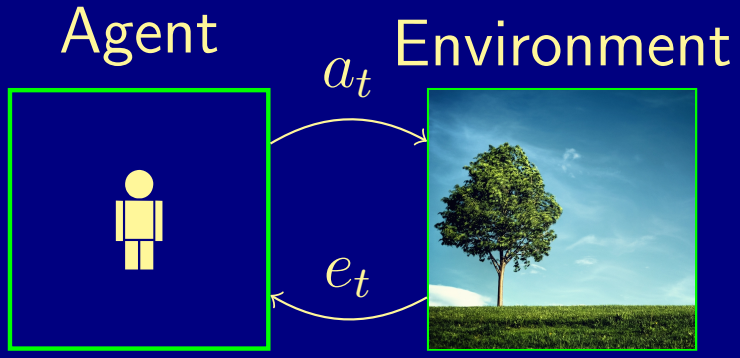
A foundational theory of AI





Answers: **What is the right thing to do?**

Framework

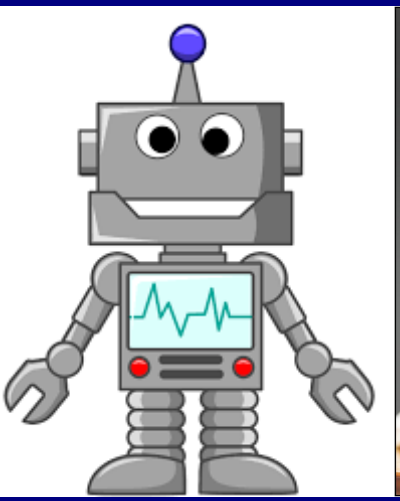


At each time step t, the agent

- submits action a_t
- receives percept e_t

History $æ_{<t} = a_1 e_1 a_2 e_2 \ldots a_{t-1} e_{t-1}$

Set of histories: $(\mathcal{A} \times \mathcal{E})^*$



AIXI takes a Model-Based Bayesian RL approach

- Start with a prior over environments
- **Learning** After gaining experience, update prior to posterior over environments
  - Sometimes convenient to think of this as single mixture environment
- **Planning** Calculate the long-term Bayes-Optimal solution
  - Expectimax search: max over actions, expectation over percepts
  - At root, probability of percept generated according to current posterior
  - At each child, update posterior given history we would have observed at that point (**very different** from planning in a known model)
  - "Solves" the exploration-exploitation dilemma: If we would gain more reward by information gathering, we will do so
  - Optimal over prior distribution

Agent and Environment

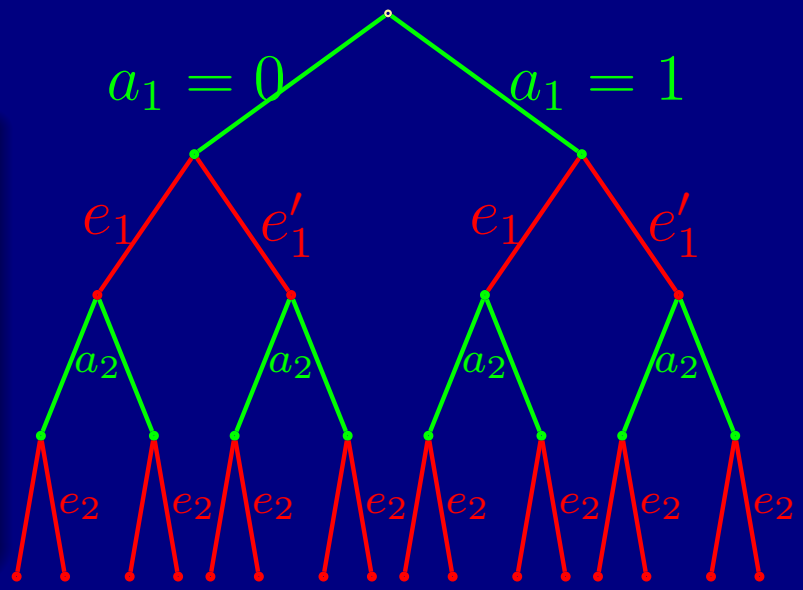
Agent

Policy

$\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

Next action

$a_t = \pi(æ_{<t})$



$a_1 = \pi(\epsilon)$

$e_1 \sim \mu(\cdot \mid a_1)$

$a_2 = \pi(a_1 e_1)$

$e_2 \sim \mu(\cdot \mid a_1 e_1 a_2)$

Environment

Distribution

$\mu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \rightsquigarrow \mathcal{E}$

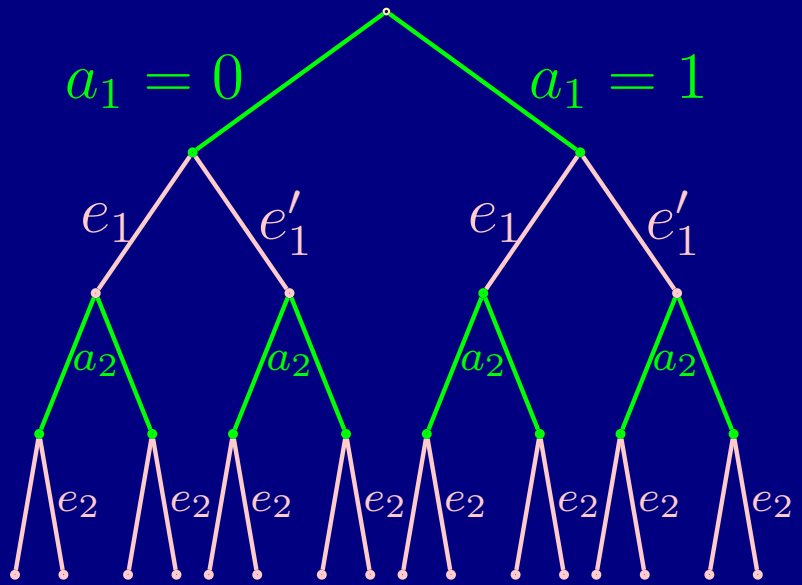
Probability of next percept:

$\mu(e_t \mid æ_{<t} a_t)$

Expectimax Planning

The *expected return* is called value: $V_\mu^\pi(æ_{<t}) = \mathbb{E}_\mu^\pi[R(æ_{1:\infty}) \mid æ_{<t}]$

$$R(æ_{1:\infty}) = \underbrace{r_1 + \gamma r_2 + \cdots \gamma^{m-1} r_m}_{\text{effective horizon}} + \underbrace{\gamma^m r_{m+1} + \cdots}_{<\epsilon} \approx R(æ_{1:m})$$



Optimal policy:
$\pi^* = \arg\max_\pi V_\mu^\pi$

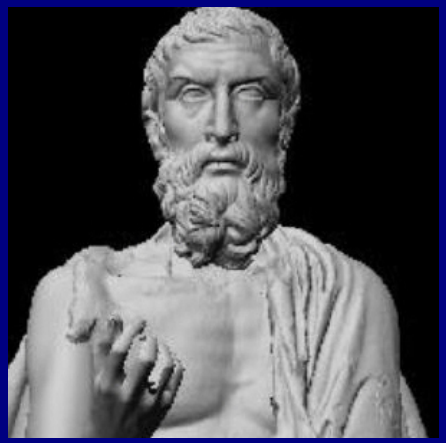
An ϵ-optimal policy can be found in any environment μ

$$a_1^* = \arg\max_{a_1} \sum_{e_1} \mu(e_1 \mid a_1) \max_{a_2} \sum_{e_2} \mu(e_2 \mid a_1 e_1 a_2) \ldots \max_{a_m} \sum_{e_m} \mu(e_m \mid æ_{<m} a_m) R(æ_{1:m})$$

Principles

Occam

Prefer the simplest consistent hypothesis



Epicurus

Keep all consistent hypotheses

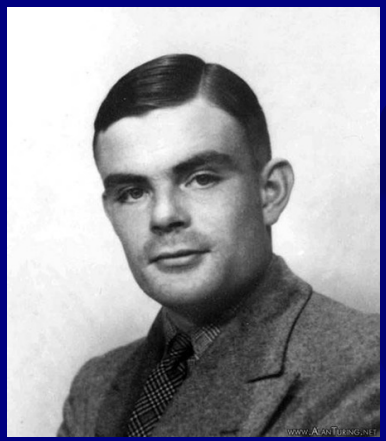
Bayes

$$\Pr(\text{Hyp} \mid \text{Data}) = \frac{\Pr(\text{Hyp}) \Pr(\text{Data} \mid \text{Hyp})}{\sum_{H_i \in \mathcal{H}} \Pr(H_i) \Pr(\text{Data} \mid H_i)}$$

Remaining questions

What is the class of hypothesis?

What is the prior?



Turing

"It is possible to invent a single machine which can be used to compute any computable sequence."

Solomonoff Induction

Use computer programs p as hypotheses/environments

Given Turing-complete programming language U, programs can

- describe essentially any environment
- be checked for consistency: is $p(a_{<t}) = e_{<t}$?
- be used for prediction: compute $p(a_{<t}a_t)$
- be ranked by simplicity: $\Pr(p) = 2^{-\ell(p)}$

Solomonoff=Epicurus+Occam+Turing+Bayes

Make a weighted prediction based on all consistent programs, with short programs weighted higher

Ray Solomonoff: A Formal Theory of Inductive Inference[1][2]

INFORMATION AND CONTROL 7, 1–22 (1964)

A Formal Theory of Inductive Inference. Part I*†

R. J. SOLOMONOFF

Rockford Research Institute, Inc., Cambridge, Massachusetts

1. SUMMARY

In Part I, four ostensibly different theoretical models of induction are presented, in which the problem dealt with is the extrapolation of a very long sequence of symbols—presumably containing all of the information to be used in the induction. Almost all, if not all problems in induction can be put in this form.

Some strong heuristic arguments have been obtained for the equivalence of the last three models. One of these models is equivalent to a Bayes formulation, in which a priori probabilities are assigned to sequences of symbols on the basis of the lengths of inputs to a universal Turing machine that are required to produce the sequence of interest as output.

[1] Ray J Solomonoff. "A formal theory of inductive inference. Part I". . In: *Information and control* 7.1 (1964), pp. 1–22.

[2] R.J. Solomonoff. "A formal theory of inductive inference. Part II". . In: *Information and Control* 7.2 (June 1964), pp. 224–254.

Solomonoff-Hutter's Universal Distribution

$$M(e_{<t} \mid a_{<t}) = \sum_{p\,:\,p(a_{<t})=e_{<t}} 2^{-\ell(p)}$$

- $a_{<t}$ action sequence
- $e_{<t}$ percept sequence
- p computer program
- $\ell(p)$ length of p

- Occam: Simpler program higher weight
- Epicurus: All consistent programs
- Bayes: Discard inconsistent programs
- Turing: Any computable environment

Predict with

$$M(e_t \mid æ_{<t} a_t) = \frac{M(e_{<t} e_t \mid a_{<t} a_t)}{M(e_{<t} \mid a_{<t})}$$

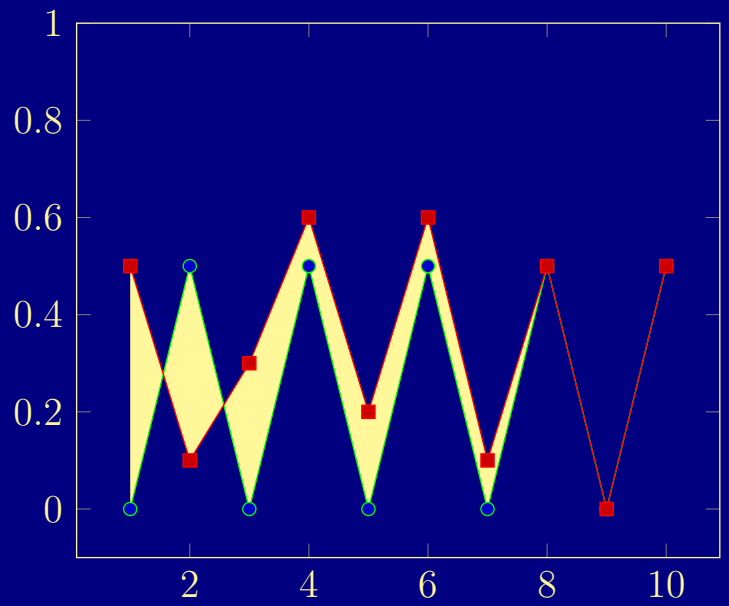
Results Solomonoff Induction

Theorem (Prediction error)

For any computable environment μ and any actions $a_{1:\infty}$:

$$\sum_{t=1}^{\infty} \mathbb{E}_\mu \Big[\underbrace{M(0 \mid æ_{<t} a_t) - \mu(0 \mid æ_{<t} a_t)}_{\text{prediction error at time } t} \Big]^2 \stackrel{+}{\leq} \frac{1}{2} \ln 2 \cdot K(\mu)$$

- Solomonoff induction only makes finitely many prediction errors
- The environment μ may be deterministic or stochastic



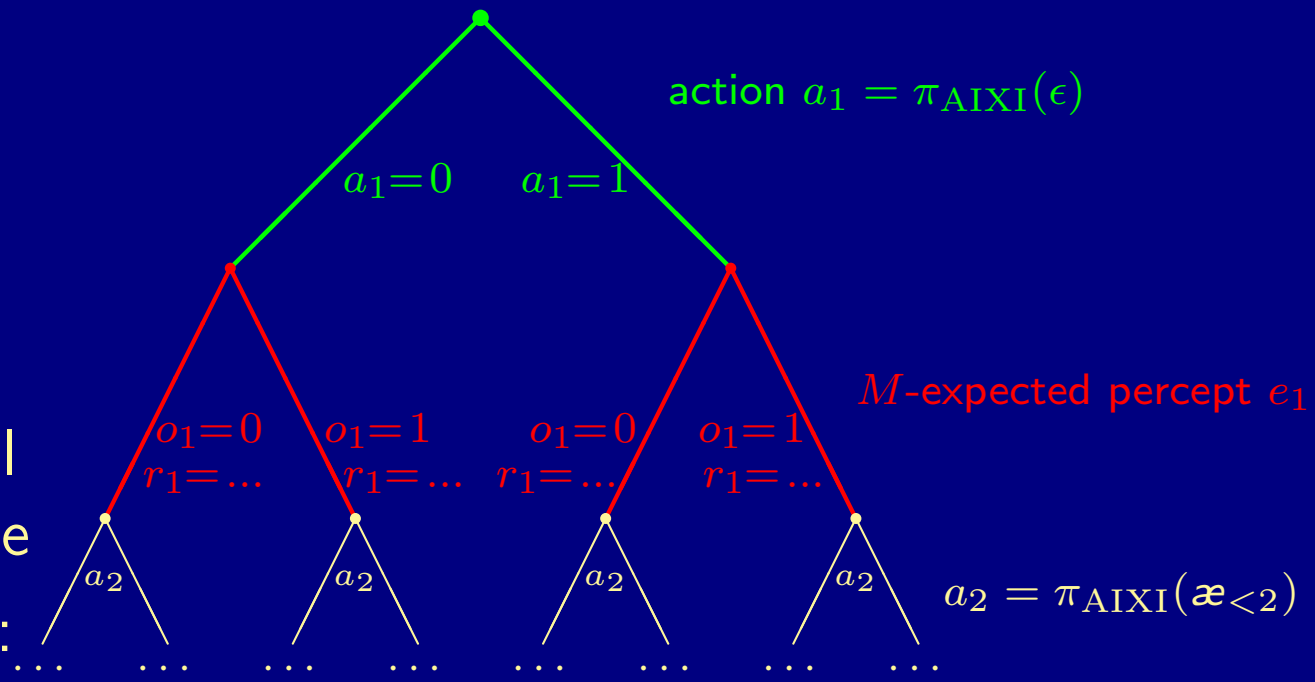
Agent can learn any computable environment

Expectimax in Unknown Environments: AIXI

AIXI replaces μ with M: $\pi_{\text{AIXI}} = \arg\max_{\pi} V_M^{\pi}$

$$a_1^* = \arg\max_{a_1} \sum_{e_1} M(e_1 \mid a_1) \max_{a_2} \sum_{e_2} M(e_2 \mid a_1 e_1 a_2) \ldots \max_{a_m} \sum_{e_m} M(e_m \mid æ_{<m} a_m) R(æ_{1:m})$$

- Learn any computable environment
- Acts Bayes-optimally
- One-equation theory for Artificial General Intelligence
- Computation time: exponential×infinite



Bayes Optimality and Optimal Exploration

- AIXI is guaranteed to learn to predict percepts it receives (on-sequence)
  - But what about those it does not receive due to actions it did not take? (off-sequence)
- AIXI is guaranteed to be Bayes-Optimal
  - Very **subjective** notion of optimality as it depends on believing the prior
  - Seems better to aim for asymptotic optimality: In the limit of data take actions optimally in any environment
  - AIXI is not asymptotically optimal, in fact the two are at odds [8]
  - **Bayes**: *Immediate, incomplete* **Asymptotic:** *Long-term, complete*
  - Very recent work suggests **optimistic approaches** or **Thompson sampling** give us asymptotic optimality [12, 6]

Benefits of a Foundational Theory of AI

AIXI/UAI provides

- (High-level) blue-print or inspiration for design
- Common terminology and goal formulation
- Understand and predict behaviour of yet-to-be-built agents
- Appreciation of fundamental challenges (e.g. exploration/exploitation)
- Definition/measure of intelligence

AIXI Approximation [15] [4]

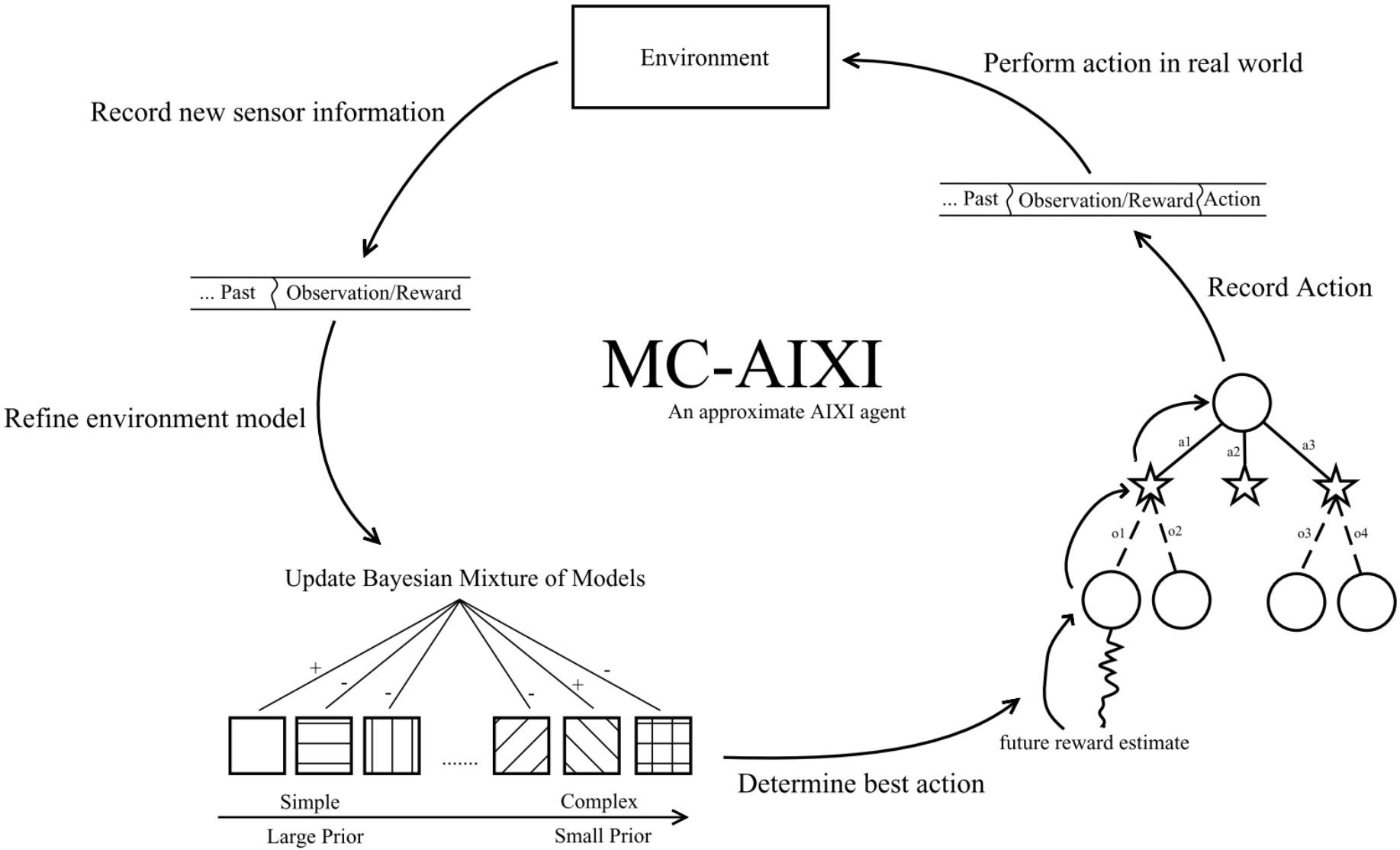
Bayesian optimality notion

"As AIXI is only asymptotically computable, it is by no means an algorithmic solution to the general reinforcement learning problem. Rather, it is best understood as a Bayesian optimality notion for decision making in general unknown environments."

Approximating AIXI

Next: How to construct tractable **approximations** of AIXI?

Monte-Carlo AIXI Framework [13]



Approximate Learning and Planning

Approximating Expectimax (§4)

Use a generalisation of Upper Confidence Bound for Trees (UCT) (Kocsis and Szepesvari 2006) to approximate the expectimax operation

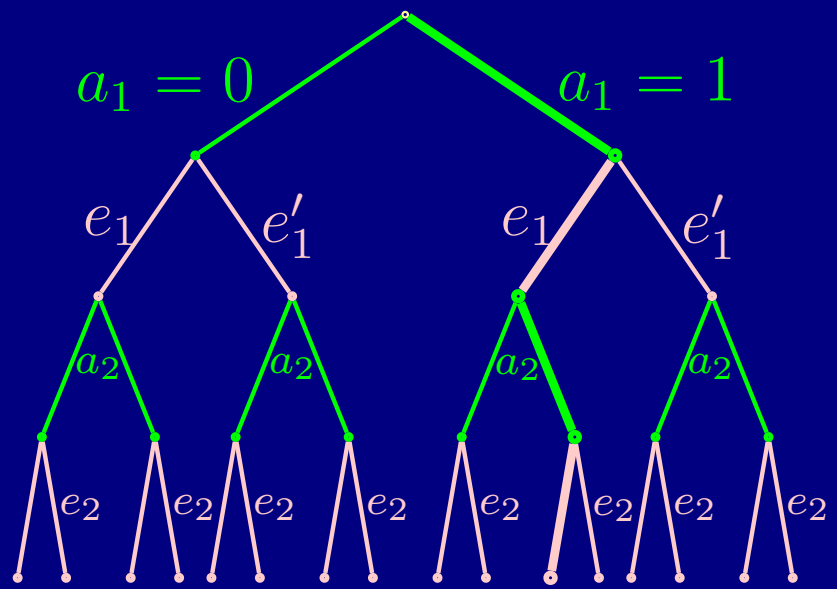
Environment Model (§5)

Use an agent-specific extension of Context Tree Weighting (CTW) (Willems, Shtarkov, and Tjalkens 1995), a Bayesian model averaging algorithm for prediction suffix trees, for prediction and learning

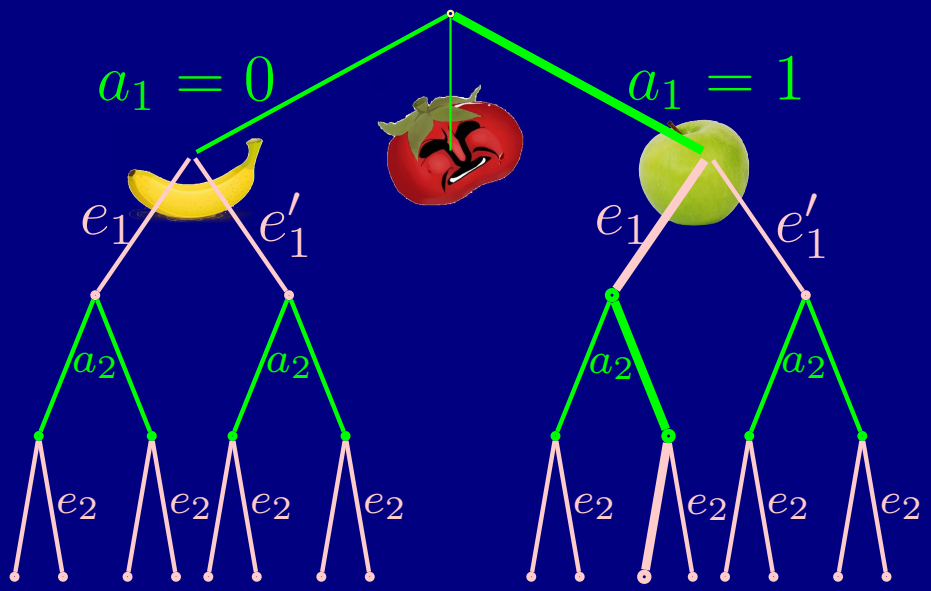
MC-AIXI-CTW: Approximating Expectimax

Planning with expectimax search takes exponential time

Sample paths in expectimax tree (anytime algorithm)



Monte Carlo Tree Search



$$a_1 = \arg\max_a V^+(a)$$

$$P(e_1 \mid a_1)$$

$$a_2 = \arg\max_a V^+(a_1 e_1 a)$$

$$P(e_2 \mid a_1 e_1 a_2)$$

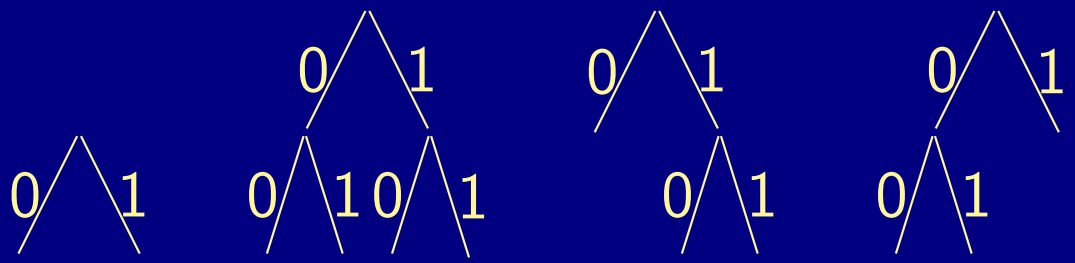
upper confidence bound

$$V^+(a) = \underbrace{\hat{V}(a)}_{\text{average}} + \underbrace{\sqrt{\log T / T(a)}}_{\text{exploration bonus}}$$

- unexplored: high $\log T/T(a)$
  $T(a) =$ times explored (a)
- promising: high $\hat{V}(a)$

MCTS famous for good performance in Go (Gelly et al., 2006)

Context Tree Weighting (CTW)



CTW "mixes" over all 2^{2^D} context trees of depth $\leq D$

$$\mathrm{CTW}(e_{<t} \mid a_{<t}) = \sum_{\Gamma} 2^{-\mathrm{CL}(\Gamma)} \Gamma(e_{<t} \mid a_{<t})$$

$$M(e_{<t} \mid a_{<t}) = \sum_{p} 2^{-\ell(p)} \, [\![\, p(a_{<t}) = e_{<t} \,]\!]$$

Computation time:

$M(e_t \mid æ_{<t} a_t)$	Infinite
$\mathrm{CTW}(e_t \mid æ_{<t} a_t)$	Constant (linear in max depth D)

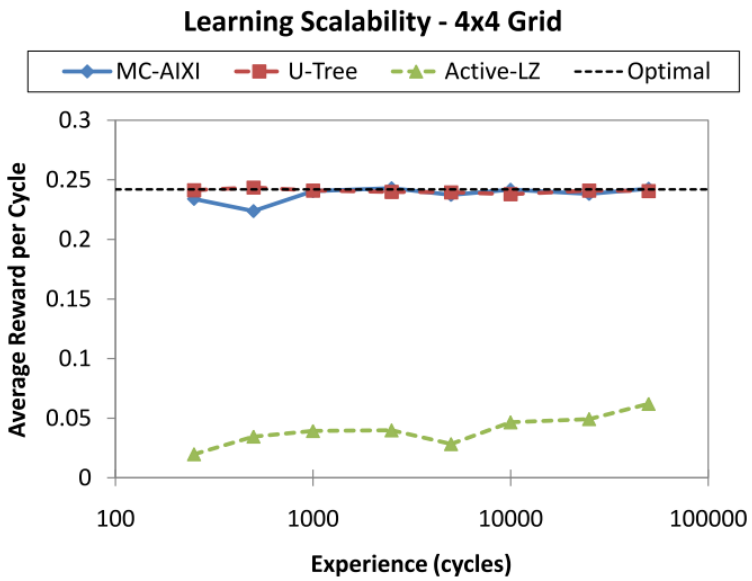
Comparison of MC-AIXI-CTW and AIXI Model Classes

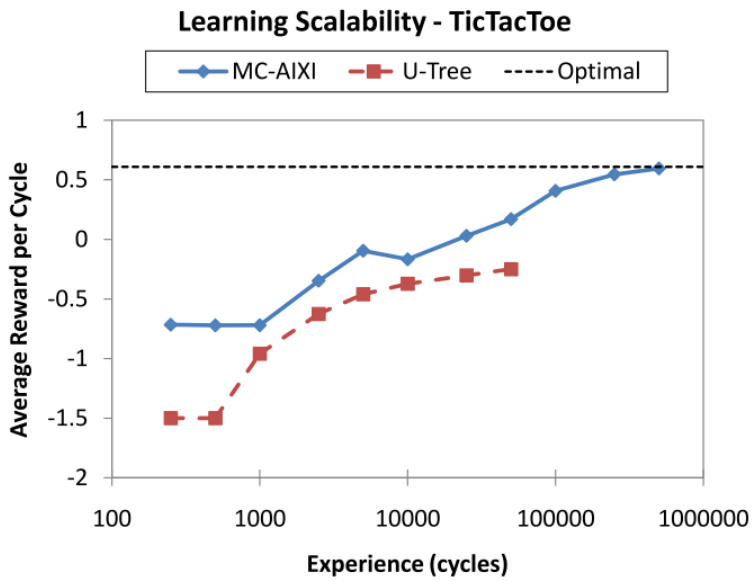
$$\arg\max_{a_t} \sum_{x_t} \cdots \max_{a_{t+m}} \sum_{x_{t+m}} \left[\sum_{i=t}^{t+m} r_i \right] \sum_{M \in \mathcal{C}_D} 2^{-\Gamma_D(\rho)} \Pr(x_{1:t+m} \mid M, a_{1:t+m}).$$

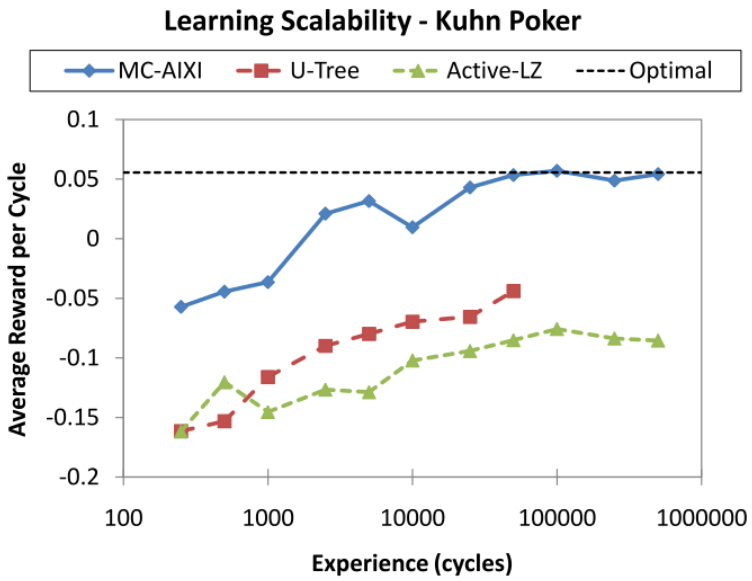
$$\arg\max_{a_t} \sum_{x_t} \ldots \max_{a_{t+m}} \sum_{x_{t+m}} \left[\sum_{i=t}^{t+m} r_i \right] \sum_{\rho \in \mathcal{M}} 2^{-K(\rho)} \rho(x_{1:t+m} \mid a_{1:t+m}),$$

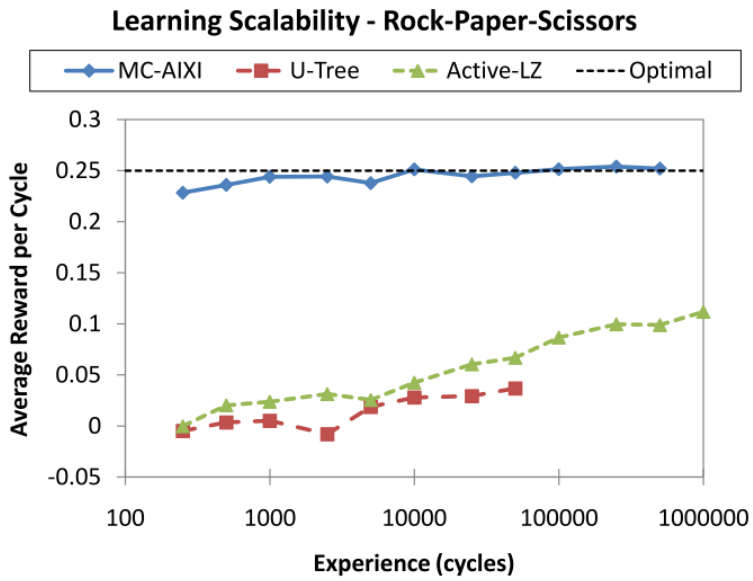
From §6: "Compare this to the action chosen by the AIXI agent, where class $\mathcal{M}$ consists of all computable environments ρ and $K(\rho)$ denotes the Kolmogorov complexity [7] of ρ. Both use a prior that favours simplicity. The main difference is in the subexpression describing the mixture over the model class. AIXI uses a mixture over all enumerable chronological semimeasures, which is completely general but incomputable. Our approximation uses a mixture of all prediction suffix trees of a certain maximum depth, which is still a rather general class, but one that is efficiently computable."

Efficiency / Experimental Results [13]









Discussion / Additional approaches to AIXI Approximation

- Optimal Ordered Problem Solver[3]
- An Approximation of the Universal Intelligence Measure[4]
- Compress and Control[5]
- Feature Reinforcement Learning[6]

[3] Jürgen Schmidhuber. "Optimal Ordered Problem Solver". In: *Machine Learning* 54.3 (Mar. 2004), pp. 211–254.

[4] Shane Legg and Joel Veness. "An Approximation of the Universal Intelligence Measure". In: (Sept. 2011), p. 14. arXiv: 1109.5951.

[5] Joel Veness et al. "Compress and Control". In: (2014). arXiv: 1411.5326.

[6] Marcus Hutter. "Feature Reinforcement Learning: Part I. Unstructured MDPs". In: *Journal of Artificial General Intelligence* 1.1 (Jan. 2009), pp. 3–24.

References I

[1] Tom Everitt. "Universal Artificial Intelligence Practical Agents and Fundamental Challenges". In: (2016). URL: http://www.tomeveritt.se/slides/AIXI-tutorial.pdf.

[2] Tom Everitt and Marcus Hutter. "Universal Artificial Intelligence: Practical Agents and Fundamental Challenges". In: (2016). URL: http://www.tomeveritt.se/papers/UAI-book-chapter.pdf.

[3] Marcus Hutter. "Feature Reinforcement Learning: Part I. Unstructured MDPs". In: *Journal of Artificial General Intelligence* 1.1 (Jan. 2009), pp. 3–24.

[4] Marcus Hutter. *Universal Artificial Intelligence*. Vol. 1. 2. 2005, pp. 1–82. ISBN: 9783540221395. DOI: 10.1145/1358628.1358961.

[5] Shane Legg and Joel Veness. "An Approximation of the Universal Intelligence Measure". In: (Sept. 2011), p. 14. arXiv: 1109.5951.

References II

[6] Jan Leike et al. "Thompson sampling is asymptotically optimal in general environments". In: *arXiv preprint arXiv:1602.07905* (2016).

[7] Ming Li and Paul M.B. Vitnyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 2008. ISBN: 0387339981, 9780387339986.

[8] Laurent Orseau. "Optimality issues of universal greedy agents with static priors". In: *International Conference on Algorithmic Learning Theory*. Springer. 2010, pp. 345–359.

[9] Jürgen Schmidhuber. "Optimal Ordered Problem Solver". In: *Machine Learning* 54.3 (Mar. 2004), pp. 211–254.

[10] Ray J Solomonoff. "A formal theory of inductive inference. Part I". In: *Information and control* 7.1 (1964), pp. 1–22.

[11] R.J. Solomonoff. "A formal theory of inductive inference. Part II". In: *Information and Control* 7.2 (June 1964), pp. 224–254.

References III

[12] Peter Sunehag and Marcus Hutter. "Rationality, optimism and guarantees in general reinforcement learning". In: *Journal of Machine Learning Research* 16 (2015), pp. 1345–1390.

[13] Joel Veness et al. "A Monte-Carlo AIXI Approximation". In: *Journal of Artificial Intelligence Research* 40.1 (2011), pp. 95–142.

[14] Joel Veness et al. "Compress and Control". In: (2014). arXiv: 1411.5326.

[15] Joel Veness et al. "Reinforcement Learning via AIXI Approximation". In: (2010).