COMP4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU

8 UNIVERSAL RATIONAL AGENTS

- Agents in Known (Probabilistic) Environments
- The Universal Algorithmic Agent AIXI
- Important Environmental Classes
- Discussion

Universal Rational Agents: Abstract

Sequential decision theory formally solves the problem of rational agents in uncertain worlds if the true environmental prior probability distribution is known. Solomonoff's theory of universal induction formally solves the problem of sequence prediction for unknown prior distribution.

Here we combine both ideas and develop an elegant parameter-free theory of an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. The theory reduces all conceptual AI problems to pure computational ones. The resulting AIXI model is the most intelligent unbiased agent possible.

Other discussed topics are optimality notions, asymptotic consistency, and some particularly interesting environment classes.

Overview

- Decision Theory solves the problem of rational agents in uncertain worlds if the environmental probability distribution is known.
- Solomonoff's theory of Universal Induction solves the problem of sequence prediction for unknown prior distribution.
- We combine both ideas and get a parameterless model of

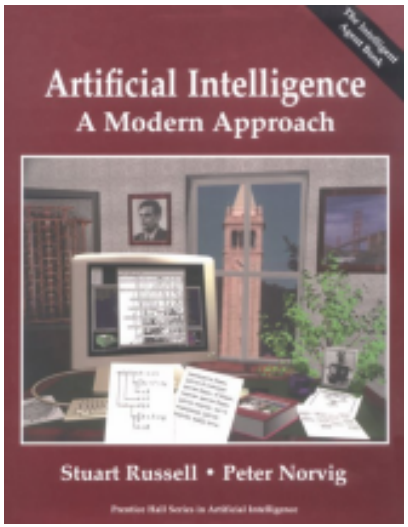
$$\begin{array}{ccc} \text{Universal Artificial Intelligence without Parameters} & & \\ = & & = \\ \text{Decision Theory} & = & \text{Probability} + \text{Utility Theory} \\ + & & + \\ \text{Universal Induction} & = & \text{Ockham} + \text{Epicurus} + \text{Bayes} \end{array}$$

Preliminary Remarks

- The goal is to mathematically define a unique model superior to any other model in any environment.
- The AIXI agent is unique in the sense that it has no parameters which could be adjusted to the actual environment in which it is used.
- In this first step toward a universal theory of AI
  we are not interested in computational aspects.
- Nevertheless, we are interested in maximizing a utility function, which means to learn in as minimal number of cycles as possible. The interaction cycle is the basic unit, not the computation time per unit.

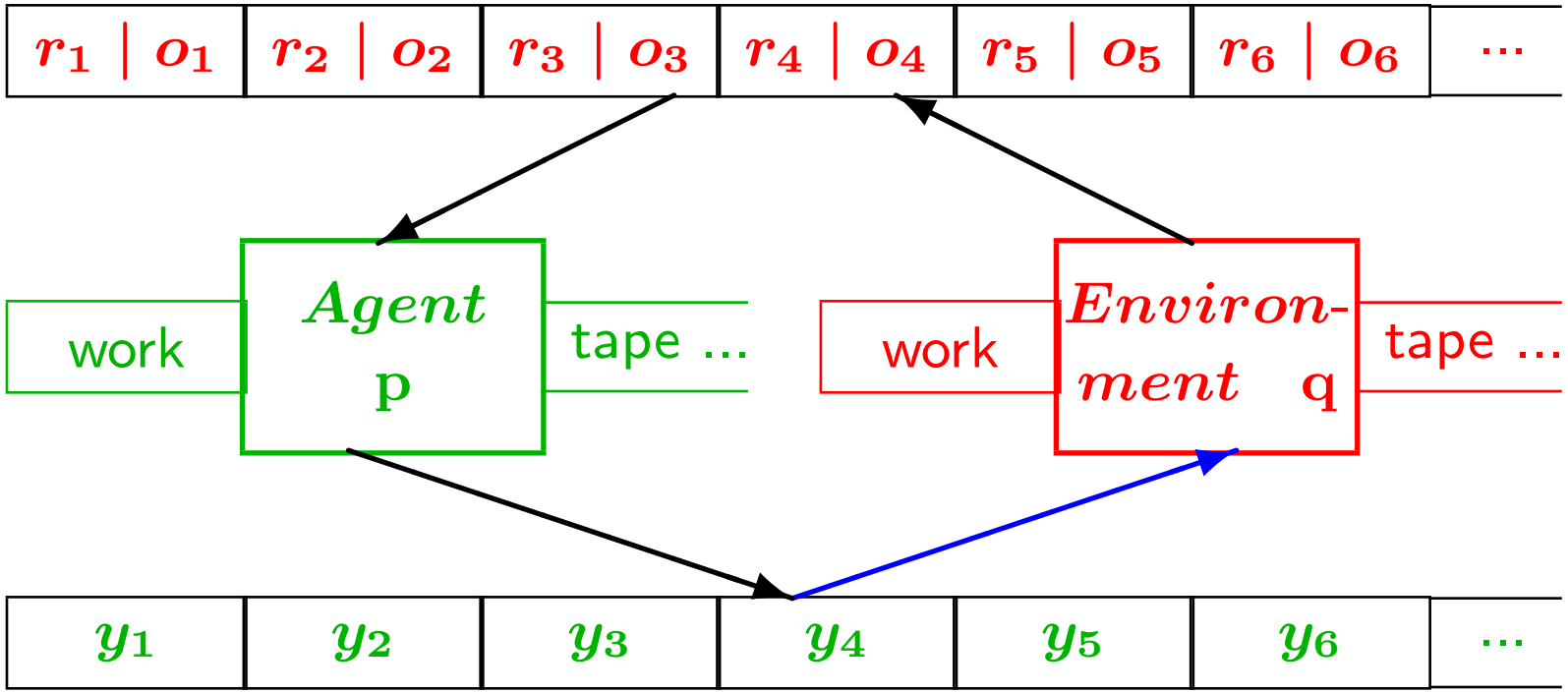
8.1 Agents in Known (Probabilistic) Environments: Contents

- The Agent-Environment Model & Interaction Cycle
- Rational Agents in Deterministic Environments
- Utility Theory for Deterministic Environments
- Emphasis in AI/ML/RL $\Leftrightarrow$ Control Theory
- Probabilistic Environment / Perceptions
- Functional$\equiv$Recursive$\equiv$Iterative AIμ Model
- Limits we are Interested in
- Relation to Bellman Equations
- (Un)Known environment μ



The Agent Model

Most if not all AI problems can be formulated within the agent framework



The Agent-Environment Interaction Cycle

for k:=1 to m do

- p thinks/computes/modifies internal state = work tape.
- p writes output $y_k \in \mathcal{Y}$.
- q reads output y_k.
- q computes/modifies internal state.
- q writes reward input $r_k \in \mathcal{R} \subset \mathbb{R}$.
- q writes regular input $o_k \in \mathcal{O}$.
- p reads input $x_k := r_k o_k \in \mathcal{X}$.

endfor

- m is lifetime of system (total number of cycles).
- Often $\mathcal{R} = \{0,1\} = \{bad, good\} = \{error, correct\}$.

Agents in Deterministic Environments

- $p: \mathcal{X}^* \to \mathcal{Y}^*$ is deterministic policy of the agent,
  $p(x_{<k}) = y_{1:k}$ with $x_{<k} \equiv x_1 ... x_{k-1}$.

- $q: \mathcal{Y}^* \to \mathcal{X}^*$ is deterministic environment,
  $q(y_{1:k}) = x_{1:k}$ with $y_{1:k} \equiv y_1 ... y_k$.

- Input $x_k \equiv r_k o_k$ consists of a regular informative part o_k
  and reward $r(x_k) := r_k \in [0..r_{max}]$.

Utility Theory for Deterministic Environments

The ($agent$,$environment$) pair (p,q) produces the unique I/O sequence

$$\omega^{pq} := y_1^{pq} x_1^{pq} y_2^{pq} x_2^{pq} y_3^{pq} x_3^{pq} ...$$

Total reward (value) in cycles k to m is defined as

$$V_{km}^{pq} := r(x_k^{pq}) + ... + r(x_m^{pq})$$

Optimal agent is policy that maximizes total reward

$$p^* := \arg\max_p V_{1m}^{pq}$$

$$\Downarrow$$

$$V_{km}^{p^*q} \geq V_{km}^{pq} \quad \forall p$$

Emphasis in AI/ML/RL $\Leftrightarrow$ Control Theory

Both fields start from Bellman-equations and aim at agents/controllers that behave optimally and are adaptive, but differ in terminology and emphasis:

agent	$\widehat{=}$	controller
environment	$\widehat{=}$	system
(instantaneous) reward	$\widehat{=}$	(immediate) cost
model learning	$\widehat{=}$	system identification
reinforcement learning	$\widehat{=}$	adaptive control
exploration$\leftrightarrow$exploitation problem	$\widehat{=}$	estimation$\leftrightarrow$control problem
qualitative solution	$\Leftrightarrow$	high precision
complex environment	$\Leftrightarrow$	simple (linear) machine
temporal difference	$\Leftrightarrow$	Kalman filtering / Ricatti eq.

AIξ is the first non-heuristic formal approach that is general enough to cover both fields. [Hut05]

Probabilistic Environment / Functional AIμ

Replace q by a prior probability distribution $\mu(q)$ over environments.

The total expected reward in cycles k to m is

$$V_{km}^{p\mu}(\dot{y}\dot{x}_{<k}) := \frac{1}{\mathcal{N}} \sum_{q:q(\dot{y}_{<k})=\dot{x}_{<k}} \mu(q) \cdot V_{km}^{pq}$$

The history is no longer uniquely determined.

$\dot{y}\dot{x}_{<k} := \dot{y}_1\dot{x}_1...\dot{y}_{k-1}\dot{x}_{k-1}$:=actual history.

AIμ maximizes expected future reward by looking $h_k \equiv m_k - k + 1$ cycles ahead (horizon). For $m_k = m$, AIμ is optimal.

$$\dot{y}_k := \arg\max_{y_k} \max_{p:p(\dot{x}_{<k})=\dot{y}_{<k}y_k} V_{km_k}^{p\mu}(\dot{y}\dot{x}_{<k})$$

Environment responds with $\dot{x}_k$ with probability determined by μ.

This functional form of AIμ is suitable for theoretical considerations.
The iterative form (next slides) is more suitable for 'practical' purpose.

Probabilistic Perceptions

The probability that the environment produces input x_k in cycle k under the condition that the history h is $y_1x_1...y_{k-1}x_{k-1}y_k$ is abbreviated by

$$\mu(x_k|yx_{<k}y_k) \equiv \mu(x_k|y_1x_1...y_{k-1}x_{k-1}y_k)$$

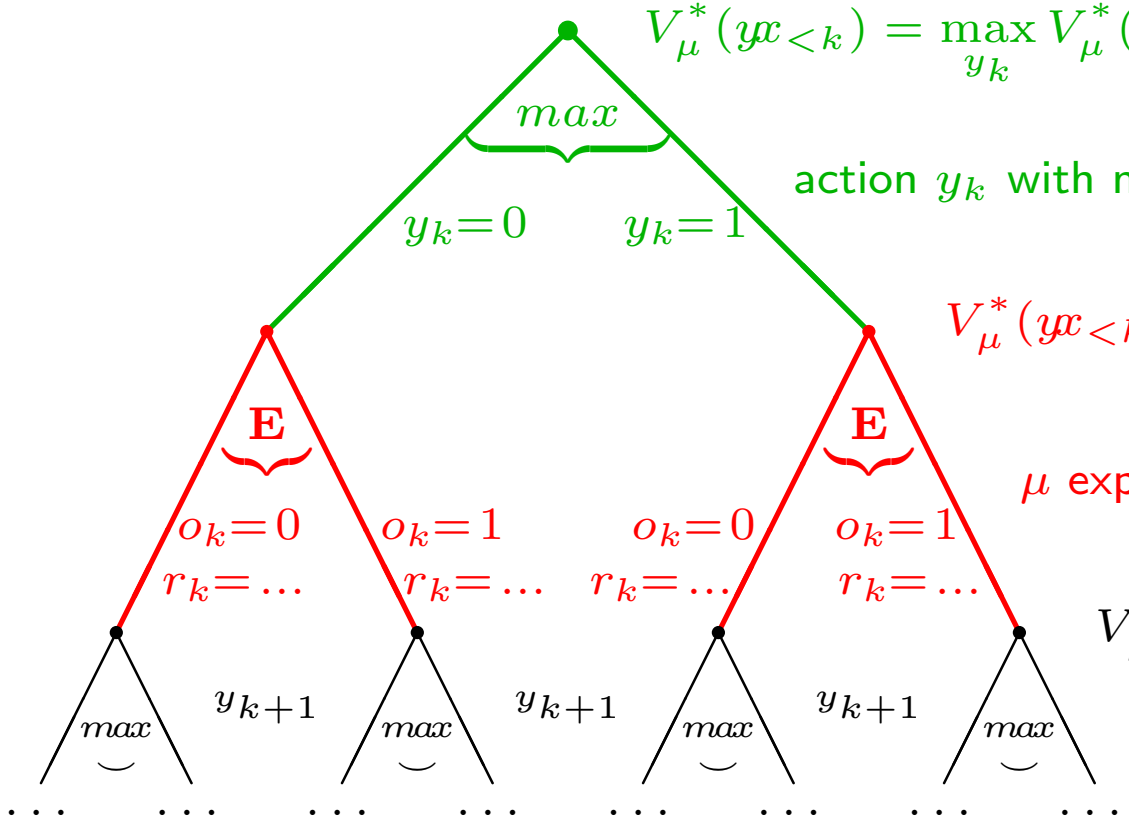
With the chain rule, the probability of input $x_1...x_k$ if system outputs $y_1...y_k$ is

$$\mu(x_1...x_k|y_1...y_k) = \mu(x_1|y_1)\cdot\mu(x_2|yx_1y_2)\cdot...\cdot\mu(x_k|yx_{<k}y_k)$$

A μ of this form is called a chronological probability distribution.

Expectimax Tree – Recursive AIμ Model

$V^*_\mu(h) \equiv V^{*\mu}_{km}(h)$ is the value (future expected reward sum) of the optimal informed agent AIμ in environment μ in cycle k given history h.



action y_k with max value.

$$V^*_\mu(yx_{<k}y_k) = \sum_{x_k}[r_k + V^*_\mu(yx_{1:k})]\mu(x_k|yx_{<k}y_k)$$

μ expected reward r_k and observation o_k.

$$V^*_\mu(yx_{1:k}) = \max_{y_{k+1}} V^*_\mu(yx_{1:k}y_{k+1})$$

Iterative AIμ Model

The Expectimax sequence/algorithm: Take reward expectation over the x_i and maximum over the y_i in chronological order to incorporate correct dependency of x_i and y_i on the history.

$$V^{*\mu}_{km}(\dot{y}\dot{x}_{<k}) \;=\; \max_{y_k}\sum_{x_k}\ldots\max_{y_m}\sum_{x_m}(r(x_k)+\ldots+r(x_m))\cdot\mu(x_{k:m}|\dot{y}\dot{x}_{<k}y_{k:m})$$

$$\dot{y}_k \;=\; \arg\max_{y_k}\sum_{x_k}\ldots\max_{y_{m_k}}\sum_{x_{m_k}}(r(x_k)+\ldots+r(x_{m_k}))\cdot\mu(x_{k:m_k}|\dot{y}\dot{x}_{<k}y_{k:m_k})$$

This is the essence of Sequential Decision Theory.

Decision Theory = Probability + Utility Theory

Functional≡Recursive≡Iterative AIμ Model

The functional and recursive/iterative AIμ models behave identically with the natural identification

$$\mu(x_{1:k}|y_{1:k}) \;=\; \sum_{q:q(y_{1:k})=x_{1:k}} \mu(q)$$

Remaining Problems:

- Computational aspects.
- The true prior probability is usually not (even approximately not) known.

Limits we are Interested in

$$\begin{array}{ccccccccccc} 1 & \ll & \langle l(y_k x_k)\rangle & \ll & k & \ll & m & \ll & |\mathcal{Y}\times\mathcal{X}| & < & \infty \\ 1 & \overset{a}{\ll} & 2^{16} & \overset{b}{\ll} & 2^{24} & \overset{c}{\ll} & 2^{32} & \overset{d}{\ll} & 2^{65536} & \overset{e}{<} & \infty \end{array}$$

(a) The agents interface is wide.

(b) The interface is sufficiently explored.

(c) The death is far away.

(d) Most input/outputs do not occur.

(e) All spaces are finite.

These limits are never used in proofs but ...

... we are only interested in theorems which do not degenerate under the above limits.

Relation to Bellman Equations

- If μ^{AI} is a completely observable Markov decision process, then AIμ reduces to the recursive Bellman equations [BT96].
- Recursive AIμ may in general be regarded as (pseudo-recursive) Bellman equation with complete history $yx_{<k}$ as environmental state.
- The AIμ model assumes neither stationarity, nor Markov property, nor complete observability of the environment.

$\Rightarrow$ every "state" occurs at most once in the lifetime of the agent. Every moment in the universe is unique!

- There is no obvious universal similarity relation on $(\mathcal{X}\times\mathcal{Y})^*$ allowing an effective reduction of the size of the state space.

Known environment μ

- Assumption: μ is the true environment in which the agent operates
- Then, policy p^μ is optimal in the sense that no other policy for an agent leads to higher μ^{AI}-expected reward.
- Special choices of μ: deterministic or adversarial environments, Markov decision processes (MDPs).
- There is no principle problem in computing the optimal action y_k as long as μ^{AI} is known and computable and $\mathcal{X}$, $\mathcal{Y}$ and m are finite.
- Things drastically change if μ^{AI} is unknown ...

Unknown environment μ

- Reinforcement learning algorithms [SB98] are commonly used in this case to learn the unknown μ or directly its value.
- They succeed if the state space is either small or has effectively been made small by so-called generalization techniques.
- Solutions are either ad hoc, or work in restricted domains only, or have serious problems with state space exploration versus exploitation, or are prone to diverge, or have non-optimal learning rate.
- We introduce a universal and optimal mathematical model now ...

8.2 The Universal Algorithmic Agent AIXI: Contents

- Formal Definition of Intelligence
- Is Universal Intelligence Υ any Good?
- Definition of the Universal AIXI Model
- Universality of M^{AI} and ξ^{AI}
- Convergence of ξ^{AI} to μ^{AI}
- Intelligence Order Relation
- On the Optimality of AIXI
- Value Bounds & Asymptotic Learnability
- The OnlyOne CounterExample
- Separability Concepts

Formal Definition of Intelligence

- Agent follows policy $\pi : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^* \rightsquigarrow \mathcal{A}$
- Environment reacts with $\mu : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^*\times\mathcal{A} \rightsquigarrow \mathcal{O}\times\mathcal{R}$
- Performance of agent π in environment μ
  $=$ expected cumulative reward $= V_\mu^\pi := \mathbb{E}_\mu^\pi[\sum_{t=1}^\infty r_t^{\pi\mu}]$
- True environment μ unknown
  $\Rightarrow$ average over wide range of environments
  (all semi-computable chronological semi-measures $\mathcal{M}_U$)
- Ockham+Epicurus: Weigh each environment with its Kolmogorov complexity $K(\mu) := \min_p\{length(p) : U(p) = \mu\}$
- Universal intelligence of agent π is $\Upsilon(\pi) := \sum_{\mu\in\mathcal{M}_U} 2^{-K(\mu)} V_\mu^\pi$.
- Compare to our informal definition: Intelligence measures an agent's ability to perform well in a wide range of environments.
- AIXI $= \arg\max_\pi \Upsilon(\pi) =$ most intelligent agent.

Is Universal Intelligence Υ any Good?

- Captures our informal definition of intelligence.
- Incorporates Occam's razor.
- Very general: No restriction on internal working of agent.
- Correctly orders simple adaptive agents.
- Agents with high Υ like AIXI are extremely powerful.
- Υ spans from very low intelligence up to ultra-high intelligence.
- Practically meaningful: High Υ = practically useful.
- Non-anthropocentric: based on information & computation theory. (unlike Turing test which measures humanness rather than int.)
- Simple and intuitive formal definition: does not rely on equally hard notions such as creativity, understanding, wisdom, consciousness.

Υ is valid, informative, wide range, general, dynamic, unbiased, fundamental, formal, objective, fully defined, universal.

Definition of the Universal AIXI Model

Universal AI = Universal Induction + Decision Theory

Replace μ^{AI} in sequential decision model AIμ by an appropriate generalization of Solomonoff's M.

$$M(x_{1:k}|y_{1:k}) := \sum_{q:q(y_{1:k})=x_{1:k}} 2^{-l(q)}$$

$$\dot{y}_k = \arg\max_{y_k} \sum_{x_k} ... \max_{y_{m_k}} \sum_{x_{m_k}} (r(x_k)+...+r(x_{m_k}))\cdot M(x_{k:m_k}|\dot{y}\dot{x}_{<k}y_{k:m_k})$$

Functional form: $\mu(q) \hookrightarrow \xi(q) := 2^{-\ell(q)}$.

Bold Claim: **AIXI is the most intelligent environmental independent agent possible.**

Universality of M^{AI} and ξ^{AI}

$$M(x_{1:n}|y_{1:n}) \stackrel{\times}{=} \xi(x_{1:n}|y_{1:n}) \geq 2^{-K(\rho)}\rho(x_{1:n}|y_{1:n}) \quad \forall \text{ chronological } \rho$$

The proof is analog as for sequence prediction. Actions y_k are pure spectators (here and below)

Convergence of ξ^{AI} to μ^{AI}

Similarly to Bayesian multistep prediction [Hut05] one can show

$$\xi^{AI}(x_{k:m_k}|x_{<k}y_{1:m_k}) \stackrel{k\to\infty}{\longrightarrow} \mu^{AI}(x_{k:m_k}|x_{<k}y_{1:m_k}) \quad \text{with } \mu \text{ prob. 1.}$$

with rapid conv. for bounded horizon $h_k \equiv m_k - k + 1 \leq h_{max} < \infty$

Does replacing μ^{AI} with ξ^{AI} lead to AIξ system with asymptotically optimal behavior with rapid convergence?

This looks promising from the analogy to the Sequence Prediction (SP) case, but is much more subtle and tricky!

Intelligence Order Relation

Definition 8.1 (Intelligence order relation) We call a policy p *more or equally intelligent* than p' and write

$$p \succeq p' \quad :\Leftrightarrow \quad \forall k \forall \dot{y}\dot{x}_{<k} : V_{km_k}^{p\xi}(\dot{y}\dot{x}_{<k}) \geq V_{km_k}^{p'\xi}(\dot{y}\dot{x}_{<k}),$$

i.e. if p yields in any circumstance higher ξ-expected reward than p'.

As the algorithm p^ξ behind the AIXI agent maximizes $V_{km_k}^{p\xi}$, we have $p^\xi \succeq p$ for all p.

The AIXI model is hence the most intelligent agent w.r.t. $\succeq$.

Relation $\succeq$ is a universal order relation in the sense that it is free of any parameters (except m_k) or specific assumptions about the environment.

On the Optimality of AIXI

- What is meant by universal optimality? Value bounds for AIXI are expected to be weaker than the SP loss bounds because problem class covered by AIXI is larger.
- The problem of defining and proving general value bounds becomes more feasible by considering, in a first step, restricted environmental classes.
- Another approach is to generalize AIXI to AIξ, where $\xi() = \sum_{\nu\in\mathcal{M}} w_\nu \nu()$ is a general Bayes mixture of distributions ν in some class $\mathcal{M}$.
- A possible further approach toward an optimality "proof" is to regard AIXI as optimal by construction. (common Bayesian perspective, e.g. Laplace rule or Gittins indices).

Value Bounds & Asymptotic Learnability

Naive value bound analogously to error bound for SP

$$V_{1m}^{p^{best}\mu} \stackrel{?}{\geq} V_{1m}^{p\mu} - o(...) \quad \forall \mu, p$$

HeavenHell Counter-Example: Set of environments $\{\mu_0, \mu_1\}$ with $\mathcal{Y}=\mathcal{R}=\{0,1\}$ and $r_k=\delta_{iy_1}$ in environment μ_i violates value bound. The first output y_1 decides whether all future $r_k=1$ or 0.

Asymptotic learnability: μ probability $D_{n\mu\xi}/n$ of suboptimal outputs of AIXI different from AIμ in the first n cycles tends to zero

$$D_{n\mu\xi}/n \to 0 \quad , \quad D_{n\mu\xi} := \mathbb{E}_\mu\Big[\sum_{k=1}^{n} 1-\delta_{\dot{y}_k^\mu,\dot{y}_k^\xi}\Big]$$

This is a weak asymptotic convergence claim.

The OnlyOne CounterExample

Let $\mathcal{R}=\{0,1\}$ and $|\mathcal{Y}|$ be large. Consider all (deterministic) environments in which a single complex output y^* is correct ($r=1$) and all others are wrong ($r=0$). The problem class is

$$\{\mu : \mu(r_k=1|x_{<k}y_{1:k})=\delta_{y_k y^*},\ K(y^*)=\lfloor \log_2 |Y| \rfloor\}$$

Problem: $D_{k\mu\xi} \leq 2^{K(\mu)}$ is the best possible error bound we can expect, which depends on $K(\mu)$ only. It is useless for $k \ll |Y| \stackrel{\times}{=} 2^{K(\mu)}$, although asymptotic convergence satisfied.

But: A bound like $2^{K(\mu)}$ reduces to $2^{K(\mu|\dot{x}_{<k})}$ after k cycles, which is $O(1)$ if enough information about μ is contained in $\dot{x}_{<k}$ in any form.

Separability Concepts

that might be useful for proving reward bounds

- Forgetful μ.
- Relevant μ.
- Asymptotically learnable μ.
- Farsighted μ.
- Uniform μ.
- (Generalized) Markovian μ.
- Factorizable μ.
- (Pseudo) passive μ.

Other concepts

- Deterministic μ.
- Chronological μ.

8.3 Important Environmental Classes: Contents

- Sequence Prediction (SP)
- Strategic Games (SG)
- Function Minimization (FM)
- Supervised Learning by Examples (EX)

In this subsection $\xi \equiv \xi^{AI} :\overset{\times}{=} M^{AI}$.

Particularly Interesting Environments

- Sequence Prediction, e.g. weather or stock-market prediction.
  Strong result: $V_\mu^* - V_\mu^{p^\xi} = O(\sqrt{\frac{K(\mu)}{m}})$, m =horizon.
- Strategic Games: Learn to play well (minimax) strategic zero-sum games (like chess) or even exploit limited capabilities of opponent.
- Optimization: Find (approximate) minimum of function with as few function calls as possible. Difficult exploration versus exploitation problem.
- Supervised learning: Learn functions by presenting $(z, f(z))$ pairs and ask for function values of z' by presenting $(z', ?)$ pairs. Supervised learning is much faster than reinforcement learning.

AIξ quickly learns to predict, play games, optimize, and learn supervised.

Sequence Prediction (SP)

SPμ Model: Binary sequence $z_1 z_2 z_3 ...$ with true prior $\mu^{SP}(z_1 z_2 z_3 ...)$.

AIμ Model: y_k = prediction for z_k; $o_{k+1} = \epsilon$.
$r_{k+1} = \delta_{y_k z_k} = 1/0$ if prediction was correct/wrong.

Correspondence:
$\mu^{AI}(r_1 ... r_k | y_1 ... y_k) = \mu^{SP}(\delta_{y_1 r_1} ... \delta_{y_k r_k}) = \mu^{SP}(z_1 ... z_k)$
For arbitrary horizon h_k: $\dot{y}_k^{AI\mu} = \arg\max_{y_k} \mu(y_k | \dot{z}_1 ... \dot{z}_{k-1}) = \dot{y}_k^{SP\Theta_\mu}$

Generalization: AIμ always reduces exactly to XXμ model if XXμ is optimal solution in domain XX.

AIξ model differs from SPξ model: Even for $h_k = 1$

$$\dot{y}_k^{AI\xi} \;=\; \arg\max_{y_k} \xi(r_k = 1 | \dot{y}\dot{r}_{<k} y_k) \;\neq\; \dot{y}_k^{SP\Theta_\xi}$$

Weak error bound: $\#\text{Errors}_{n\xi}^{AI} \stackrel{\times}{<} 2^{K(\mu)} < \infty$ for deterministic μ.

Strategic Games (SG)

- Consider strictly competitive strategic games like chess.
- Minimax is best strategy if both Players are rational with unlimited capabilities.
- Assume that the environment is a minimax player of some game $\Rightarrow$ μ^{AI} uniquely determined.
- Inserting μ^{AI} into definition of $\dot{y}_k^{AI}$ of AIμ model reduces the expecimax sequence to the minimax strategy ($\dot{y}_k^{AI} = \dot{y}_k^{SG}$).
- As $\xi^{AI} \to \mu^{AI}$ we expect AIξ to learn the minimax strategy for any game and minimax opponent.
- If there is only non-trivial reward $r_k \in \{win, loss, draw\}$ at the end of the game, repeated game playing is necessary to learn from this very limited feedback.
- AIξ can exploit limited capabilities of the opponent.

Function Maximization (FM)

Approximately maximize (unknown) functions with as few function calls as possible. Applications:

- Traveling Salesman Problem (bad example).
- Minimizing production costs.
- Find new materials with certain properties.
- Draw paintings which somebody likes.

$$\mu^{FM}(z_1...z_n|y_1...y_n) \;:=\; \sum_{f:f(y_i)=z_i\ \forall 1\leq i\leq n} \mu(f)$$

Greedily choosing y_k which maximizes f in the next cycle does not work.

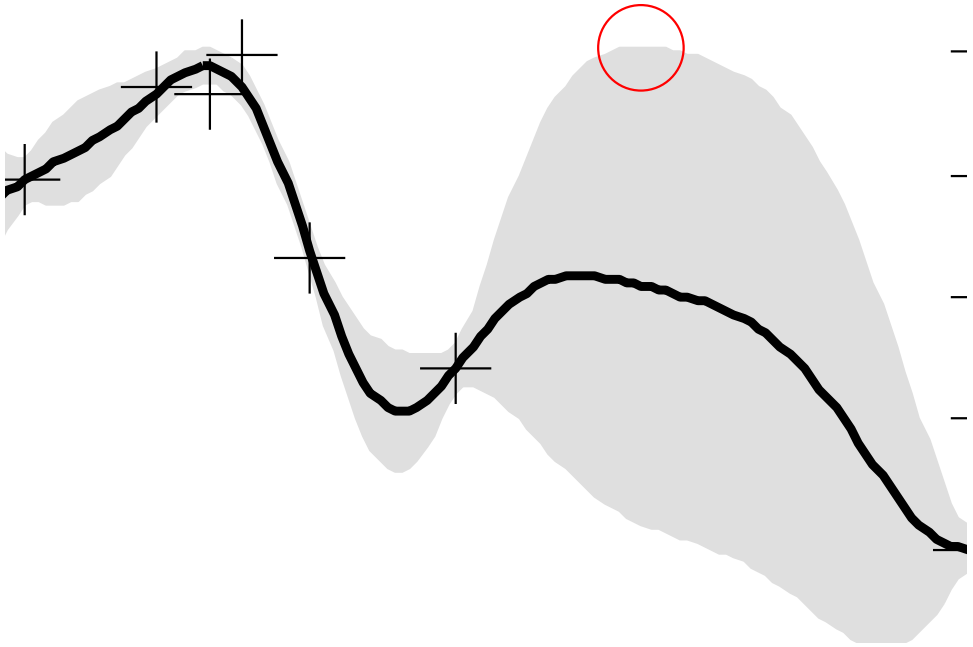
General Ansatz for FMμ/ξ:

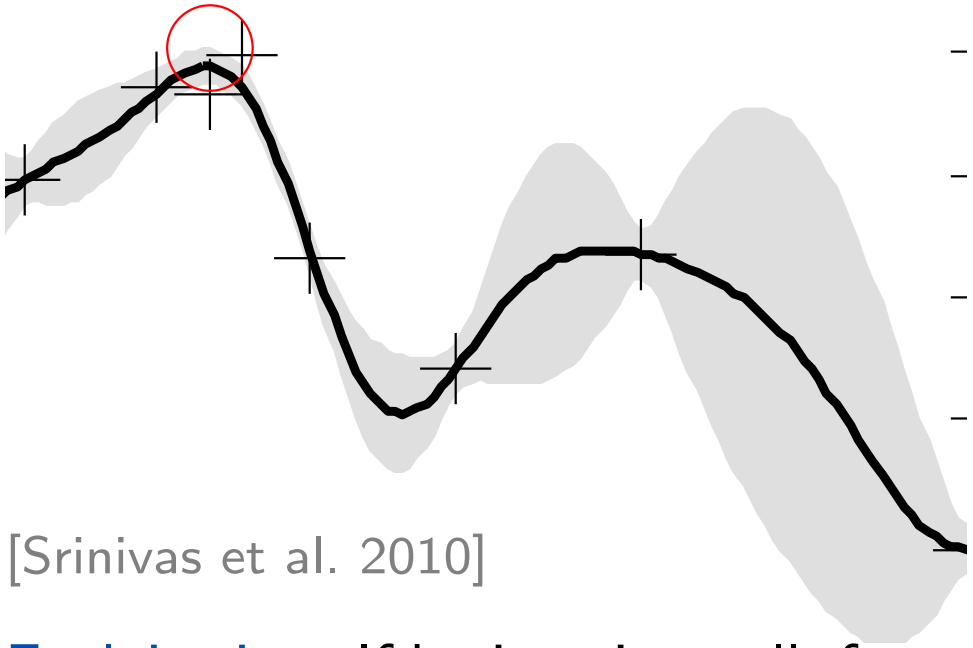
$$\dot{y}_k = \arg\max_{y_k}\sum_{z_k}...\max_{y_m}\sum_{z_m}(\alpha_1 z_1+...+\alpha_m z_m)\cdot\mu(z_m|\dot{y}\dot{z}_1...y_m)$$

Under certain weak conditions on α_i, f can be learned with AIξ.

Function Maximization – Example

Very hard problem in practice, since (unlike prediction, classification, regression) it involves the infamous exploration↔explotation problem





[Srinivas et al. 2010]

Exploration: If horizon is large, function is probed where uncertainty is large, since global maximum might be there.

Exploitation: If horizon is small, function is probed where maximum is believed to be, since agent needs/wants good results now.

Efficient and effective heuristics for special function classes available:
Extension of Upper Confidence Bound for Bandits (UCB) algorithm.

Supervised Learning by Examples (EX)

Learn functions by presenting $(z, f(z))$ pairs and ask for function values of z' by presenting $(z', ?)$ pairs.

More generally: Learn relations $R \ni (z, v)$.

Supervised learning is much faster than reinforcement learning.

The AIμ/ξ model:

$o_k \quad = (z_k, v_k) \in R \cup (Z \times \{?\}) \subset Z \times (Y \cup \{?\}) = O$

$y_{k+1} =$ guess for true v_k if actual $v_k =?$.

$r_{k+1} = 1$ iff $(z_k, y_{k+1}) \in R$

AIμ is optimal by construction.

EX is closely related to classification which itself can be phrased as sequence prediction task.

Supervised Learning – Intuition

The AIξ model:

- Inputs o_k contain much more than 1 bit feedback per cycle.
- Short codes dominate ξ.
- The shortest code of examples (z_k, v_k) is a coding of R and the indices of the (z_k, v_k) in R.
- This coding of R evolves independently of the rewards r_k.
- The system has to learn to output y_{k+1} with $(z_k, y_{k+1}) \in R$.
- As R is already coded in q, an additional algorithm of length $O(1)$ needs only to be learned.
- Rewards r_k with information content $O(1)$ are needed for this only.
- AIξ learns to learn supervised.

8.4 Discussion: Contents

- Uncovered Topics
- Remarks
- Outlook
- Exercises
- Literature

Uncovered Topics

- General and special reward bounds and convergence results for AIXI similar to SP case.
- Downscale AIXI in more detail and to more problem classes analog to the downscaling of SP to Minimum Description Length and Finite Automata.
- There is no need for implementing extra knowledge, as this can be learned by presenting it in o_k in any form.
- The learning process itself is an important aspect.
- Noise or irrelevant information in the inputs do not disturb the AIXI system.

Remarks

- We have developed a parameterless AI model based on sequential decisions and algorithmic probability.
- We have reduced the AI problem to pure computational questions.
- AIξ seems not to lack any important known methodology of AI, apart from computational aspects.
- Philosophical questions: relevance of non-computational physics (Penrose), number of wisdom Ω (Chaitin), consciousness, social consequences.

Outlook

mainly technical results for AIXI and variations

- General environment classes $\mathcal{M}_U \rightsquigarrow \mathcal{M}$.
- Results for general/universal $\mathcal{M}$ for discussed performance criteria.
- Strong guarantees for specific classes $\mathcal{M}$ by exploiting extra properties of the environments.
- Restricted policy classes.
- Universal choice of the rewards.
- Discounting future rewards and time(in)consistency.
- Approximations and algorithms.

Most of these items will be covered in the next Chapter

Exercises

1. [C30] Proof equivalence of the functional, recursive, and iterative AIμ models. Hint: Consider $k=2$ and $m=3$ first. Use $\max_{y_3(\cdot)}\sum_{x_2} f(x_2,y_3(x_2)) \equiv \sum_{x2}\max_{y_3} f(x_2,y_3)$, where $y_3(\cdot)$ is a function of x_2, and $\max_{y_3(\cdot)}$ maximizes over all such functions.
2. [C30] Show that the optimal policy $p_k^* := \arg\max_p V_{km}^{p\mu}(yx_{<k})$ is independent of k. More precisely, the actions of p_1^* and p_k^* in cycle t given history $yx_{<t}$ coincide for $k \geq t$. The derivation goes hand in hand with the derivation of Bellman's equations [BT96].

Literature

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[RN10] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition, 2010.

[LH07] S. Legg and M. Hutter. *Universal intelligence: A definition of machine intelligence.* Minds & Machines, 17(4):391–444, 2007.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. http://www.hutter1.net/ai/uaibook.htm.