Self-Optimizing and Pareto-Optimal Policies in General Environments based on Bayes-Mixtures

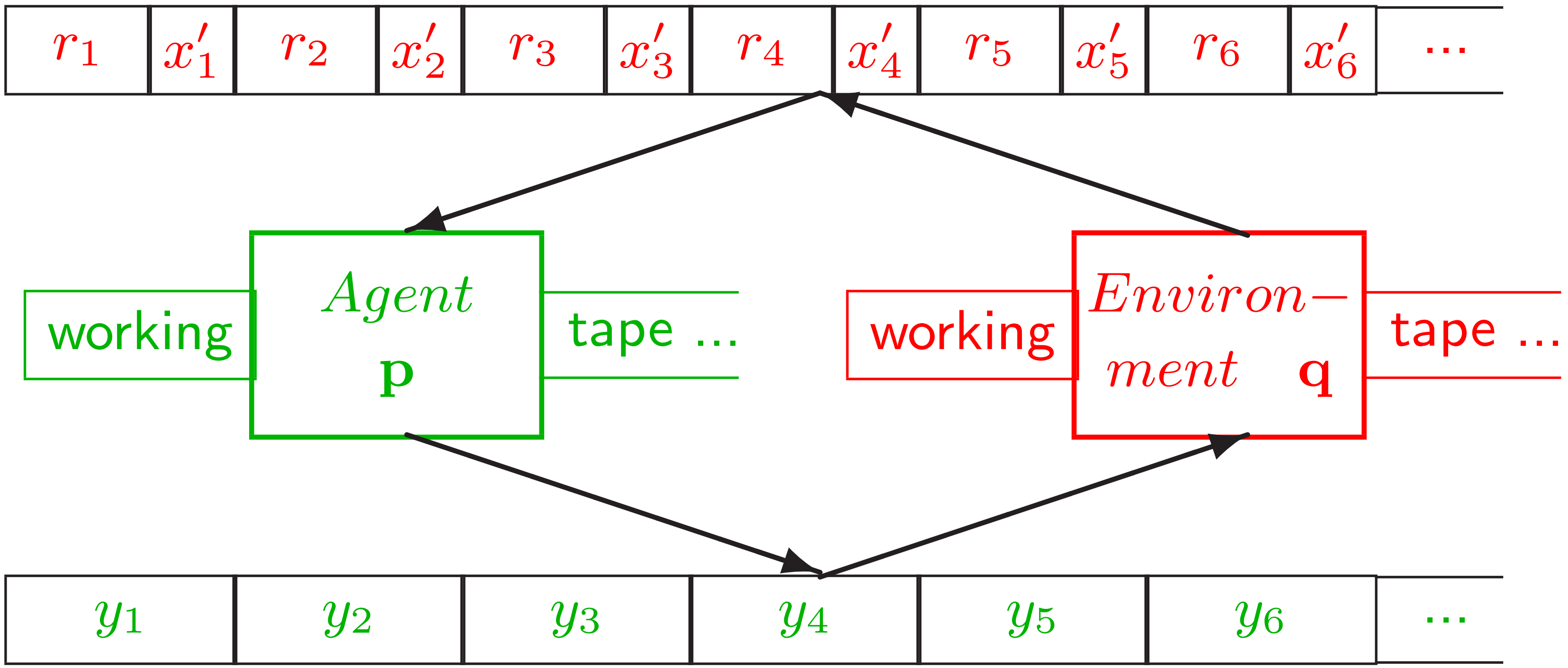
Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

COLT, 8-12 July 2002

Contents

The Agent Model



Rational 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 $x_k \equiv x'_k r_k$ consists of a regular part x'_k
  and reward $r_k \in [0..r_{max}]$.
- Value $V_{km}^{pq} := r_k + ... + r_m$,
  optimal policy $p^{best} := \arg\max_p V_{1m}^{pq}$,
  Lifespan or initial horizon m.

Agents in Probabilistic Environments

Given history $y_{1:k}x_{<k}$, the probability that the environment leads to perception x_k in cycle k is (by definition) $\sigma(x_k|y_{1:k}x_{<k})$.

Abbreviation (Bayes rule)

$$\sigma(x_{1:m}|y_{1:m}) \;=\; \sigma(x_1|y_1)\cdot\sigma(x_2|y_{1:2}x_1)\cdot\,...\,\cdot\sigma(x_m|y_{1:m}x_{<m})$$

The average value of policy p with horizon m in environment σ given history $y_{<k}x_{<k}$ is defined as

$$V_\sigma^p \;:=\; \frac{1}{m}\sum_{x_{1:m}}(r_1+...+r_m)\sigma(x_{1:m}|y_{1:m})_{|y_{1:m}=p(x_{<m})}$$

The goal of the agent should be to maximize the value.

Optimal Policy and Value

The σ-optimal policy $p^\sigma := \arg\max_p V_\sigma^p$ maximizes $V_\sigma^p \leq V_\sigma^* := V_\sigma^{p^\sigma}$.

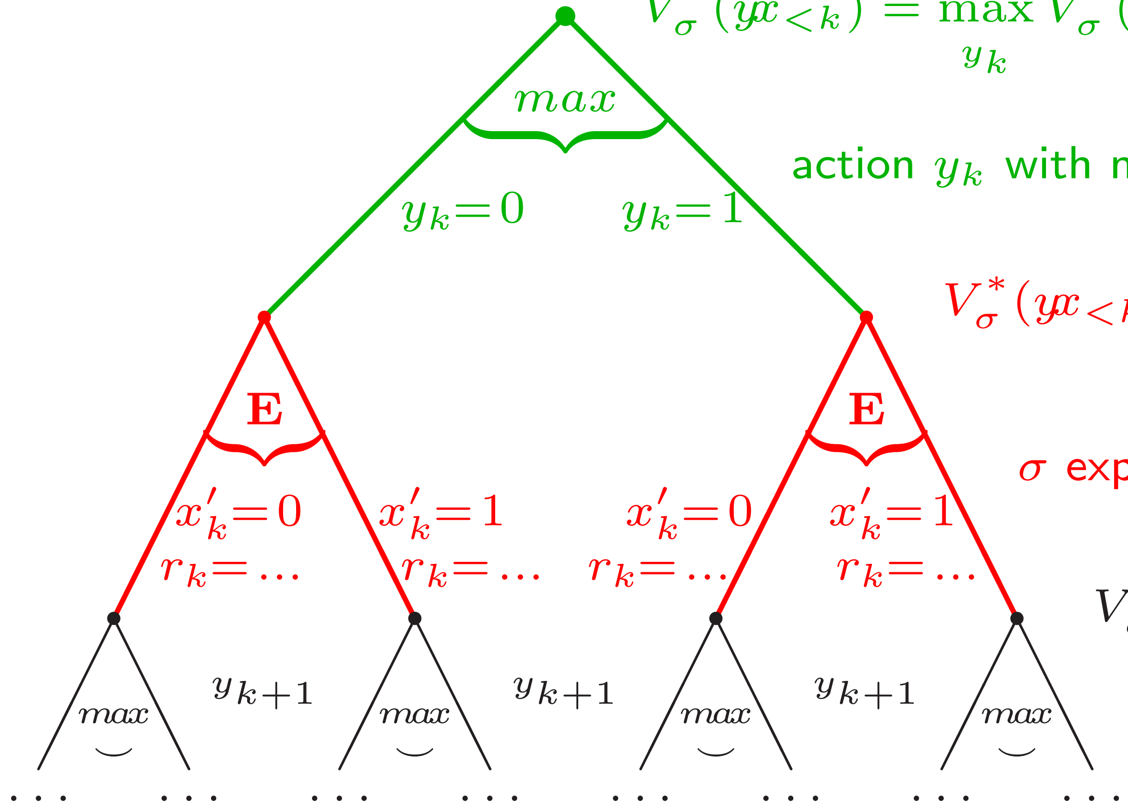
Explicit expressions for the action y_k in cycle k of the σ-optimal policy p^σ and their value V_σ^* are

$$y_k = \arg\max_{y_k}\sum_{x_k}\max_{y_{k+1}}\sum_{x_{k+1}} \dots \max_{y_m}\sum_{x_m}(r_k+\dots+r_m)\cdot\sigma(x_{k:m}|y_{1:m}x_{<k}),$$

$$V_\sigma^* = \frac{1}{m}\max_{y_1}\sum_{x_1}\max_{y_2}\sum_{x_2} \dots \max_{y_m}\sum_{x_m}(r_1+\dots+r_m)\cdot\sigma(x_{1:m}|y_{1:m}).$$

Keyword: Expectimax tree/algorithm.

Expectimax Tree/Algorithm



$$V_\sigma^*(yx_{<k}) = \max_{y_k} V_\sigma^*(yx_{<k}y_k)$$

action y_k with max value.

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

σ expected reward r_k and observation x'_k.

$$V_\sigma^*(yx_{1:k}) = \max_{y_{k+1}} V_\sigma^*(yx_{1:k}y_{k+1})$$

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-expected reward.
- Special choices of μ: deterministic environments, Markov decision processes (MDP*s*), adversarial environments.
- There is no principle problem in computing the optimal action y_k as long as μ is known and computable and $\mathcal{X}$, $\mathcal{Y}$ and m are finite.
- Things drastically change if μ is unknown ...

The Bayes-mixture distribution ξ

Assumption: The true environment μ is unknown.

Bayesian approach: The true probability distribution μ is not learned directly, but is replaced by a Bayes-mixture ξ.

Assumption: We know that the true environment μ is contained in some known (finite or countable) set $\mathcal{M}$ of environments.

The Bayes-mixture ξ is defined as

$$\xi(x_{1:m}|y_{1:m}) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x_{1:m}|y_{1:m}) \quad \text{with} \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu > 0\ \forall\nu$$

The weights w_ν may be interpreted as the prior degree of belief that the true environment is ν.

Then $\xi(x_{1:m}|y_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$, given actions $y_{1:m}$.

Questions of Interest

- It is natural to follow the policy p^ξ which maximizes V_ξ^p.
- If μ is the true environment the expected reward when following policy p^ξ will be $V_\mu^{p^\xi}$.
- The optimal (but infeasible) policy p^μ yields reward $V_\mu^{p^\mu} \equiv V_\mu^*$.
- Are there policies with uniformly larger value than $V_\mu^{p^\xi}$?
- How close is $V_\mu^{p^\xi}$ to V_μ^*?
- What is the most general class $\mathcal{M}$ and weights w_ν.

A universal choice of ξ and $\mathcal{M}$

- We have to assume the existence of some structure on the environment to avoid the No-Free-Lunch Theorems [Wolpert 96].
- We can only unravel effective structures which are describable by (semi)computable probability distributions.
- So we may include all (semi)computable (semi)distributions in $\mathcal{M}$.
- Occam's razor tells us to assign high prior belief to simple environments.
- Using Kolmogorov's universal complexity measure $K(\nu)$ for environments ν one should set $w_\nu \sim 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program on a universal TM computing ν.
- The resulting AIXI model [Hutter:00] is a unification of (Bellman's) sequential decision and Solomonoff's universal induction theory.
- In the following we consider generic $\mathcal{M}$ and w_ν.

Linearity and Convexity of V_σ in σ

V_σ^p is a linear function in σ: $V_\xi^p = \sum_\nu w_\nu V_\nu^p$

V_σ^* is a convex function in σ: $V_\xi^* \leq \sum_\nu w_\nu V_\nu^*$

where $\xi(x_{1:m}|y_{1:m}) = \sum_\nu w_\nu \, \nu(x_{1:m}|y_{1:m})$.

These are the crucial properties of the value function V_σ.

Loose interpretation: A mixture can never increase performance.

Pareto-Optimality of p^ξ

Policy p^ξ is Pareto-optimal in the sense that there is no other policy p with $V_\nu^p \geq V_\nu^{p^\xi}$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one ν.

Extension: Balanced Pareto optimality.

Self-optimizing Policies

Under which circumstances does the value of the universal policy p^ξ converge to optimum?

$$V_\nu^{p^\xi} \to V_\nu^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \qquad (1)$$

The least we must demand from $\mathcal{M}$ to have a chance that (1) is true is that there exists some policy $\tilde{p}$ at all with this property, i.e.

$$\exists \tilde{p}: \; V_\nu^{\tilde{p}} \to V_\nu^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \qquad (2)$$

Main result: $(2) \Rightarrow (1)$: The necessary condition of the existence of a self-optimizing policy $\tilde{p}$ is also sufficient for p^ξ to be self-optimizing.

Environments with Self-Optimizing Policies

- Ergodic MDPs,
- l^{th} order ergodic MDPs,
- Certain classes of POMDPs,
- Classification tasks,
- i.i.d. processes,
- Bandit problems,
- Factorizable environments,
- Repeated games,
- Prediction problems,
- ? ... ?

Discussion of Self-optimizing Property

- The beauty of this theorem is that the necessary condition of convergence is also sufficient.
- The unattractive point is that this is not an asymptotic convergence statement of a single policy p^ξ for time $k\to\infty$ for some fixed m.
- Shift focus from the total value V and horizon $m\to\infty$ to the future value (value-to-go) V and current time $k\to\infty$.

Future Value and Discounting

- Eliminate the horizon by discounting the rewards $r_k \rightsquigarrow \gamma_k r_k$ with $\Gamma_k := \sum_{i=k}^\infty \gamma_i < \infty$ and letting $m\to\infty$.
- $V_{k\gamma}^{p\sigma} := \dfrac{1}{\Gamma_k} \lim\limits_{m\to\infty} \sum\limits_{x_{k:m}} (\gamma_k r_k + ... + \gamma_m r_m) \sigma(x_{k:m}|y_{1:m} x_{<k})_{|y_{1:m}=p(x_{<m})}$
- Further advantage: Traps (non-ergodic environments) do not necessarily prevent self-optimizing policies any more.

Results for Discounted Future Value

- $V_{k\gamma}^{p\sigma}$ is linear in σ: $V_{k\gamma}^{p\xi} = \sum_\nu w_k^\nu V_{k\gamma}^{p\nu}$.

- $V_{k\gamma}^{*\sigma}$ is convex in σ: $V_{k\gamma}^{*\xi} \leq \sum_\nu w_k^\nu V_{k\gamma}^{*\nu}$.

- where $w_k^\nu := w_\nu \frac{\nu(x_{<k}|y_{<k})}{\xi(x_{<k}|y_{<k})}$ is the posterior belief in ν.

- p^ξ is Pareto-optimal in the sense that there is no other policy p with $V_{k\gamma}^{p\nu} \geq V_{k\gamma}^{p^\xi\nu}$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one ν.

- If there exists a self-optimizing policy for $\mathcal{M}$, then p^ξ is self-optimizing in the sense that
If $\exists \tilde{p}_k \forall \nu : V_{k\gamma}^{\tilde{p}_k\nu} \xrightarrow{k\to\infty} V_{k\gamma}^{*\nu} \quad \Longrightarrow \quad V_{k\gamma}^{p^\xi\mu} \xrightarrow{k\to\infty} V_{k\gamma}^{*\mu}$.

Importance of the Right Discounting

Standard geometric discounting: $\gamma_k = \gamma^k$ with $0 < \gamma < 1$.

Problem: Most environments do not possess self-optimizing policies under this discounting.

Reason: Effective horizon h_k^{eff} is finite ($\sim \ln \frac{1}{\gamma}$ for $\gamma_k = \gamma^k$).

The analogue of $m \to \infty$ is $k \to \infty$ and $h_k^{eff} \to \infty$ for $k \to \infty$.

Result: Policy p^ξ is self-optimizing for the class of (l^{th} order) ergodic MDPs if $\frac{\gamma_{k+1}}{\gamma_k} \to 1$.

Example discounting: $\gamma_k = k^{-2}$ or $\gamma_k = k^{-1-\varepsilon}$ or $\gamma_k = 2^{-K(k)}$.

Horizon is of the order of the age of the agent: $h_k^{eff} \sim k$.

Outlook

- Continuous classes $\mathcal{M}$.
- Restricted policy classes.
- Non-asymptotic bounds.
- Tighter bounds by exploiting extra properties of the environments, like the mixing rate of MDPs.
- Search for other performance criteria [Hutter:00].
- Instead of convergence of the expected reward sum, study convergence with high probability of the actually realized reward sum.

Conclusions

- Setup: Agents acting in general probabilistic environments with reinforcement feedback.
- Assumptions: True environment μ belongs to a known class of environments $\mathcal{M}$, but is otherwise unknown.
- Results: The Bayes-optimal policy p^ξ based on the Bayes-mixture $\xi = \sum_{\nu\in\mathcal{M}} w_\nu \nu$ is Pareto-optimal and self-optimizing if $\mathcal{M}$ admits self-optimizing policies.
- Application: The class of ergodic MDPs admits self-optimizing policies.
- New: Policy p^ξ with unbounded effective horizon is the first purely Bayesian self-optimizing consistent policy for ergodic MDPs.
- Learn: The combined conditions $\Gamma_k < \infty$ and $\frac{\gamma_{k+1}}{\gamma_k} \to 1$ allow a consistent self-optimizing Bayes-optimal policy based on mixtures.