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

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

ANU

9 THEORY OF RATIONAL AGENTS

- The Bayesian Agent AIξ
- Future Value and Discounting
- Knowledge-Seeing and Optimistic Agents
- Discussion

Theory of Rational Agents: Abstract

... There are strong arguments that the resulting AIXI model is the most intelligent unbiased agent possible.

Other discussed topics are relations between problem classes, the horizon problem, and computational issues.

9.1 The Bayesian Agent AIξ: Contents

- Agents in Probabilistic Environments
- Optimal Policy and Value – AIρ Model
- The Bayes-Mixture Distribution ξ
- Questions of Interest
- Linearity and Convexity of V_ρ in ρ
- Pareto Optimality
- Self-optimizing Policies
- Environments w./ (Non)Self-Optimizing Policies

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) $\rho(x_k|y_{1:k}x_{<k})$.

Abbreviation (chain rule)

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

The average value of policy p with horizon m in environment ρ is defined as

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

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

Optimal Policy and Value – AIρ Model

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

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

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

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

Keyword: Expectimax tree/algorithm.

The Bayes-Mixture Distribution ξ

Assumption: The true environment μ is unknown.

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

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?

  $\mathcal{M} = \mathcal{M}_U$ and $w_\nu = 2^{-K(\nu)}$ $\quad\Longrightarrow\quad$ AIξ =AIXI !

Linearity and Convexity of V_ρ in ρ

Theorem 9.1 (Linearity and convexity of V_ρ in ρ)
V_ρ^p is a linear function in ρ: $V_\xi^p = \sum_\nu w_\nu V_\nu^p$
V_ρ^* is a convex function in ρ: $V_\xi^* \leq \sum_\nu w_\nu V_\nu^*$
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_ρ.

Loose interpretation: A mixture can never increase performance.

Pareto Optimality

Every policy based on an estimate ρ of μ which is closer to μ than ξ is, outperforms p^ξ in environment μ, simply because it is more tailored toward μ. On the other hand, such a system performs worse than p^ξ in other environments:

Theorem 9.2 (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.

From a practical point of view a significant increase of V for many environments ν may be desirable even if this causes a small decrease of V for a few other ν. This is impossible due to

Balanced Pareto optimality:

$\Delta_\nu := V_\nu^{p^\xi} - V_\nu^{\tilde{p}}, \quad \Delta := \sum_\nu w_\nu \Delta_\nu \quad \Rightarrow \quad \Delta \geq 0.$

Self-optimizing Policies

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

$$V_\nu^{p^\xi} \to V_\nu^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \qquad (9.3)$$

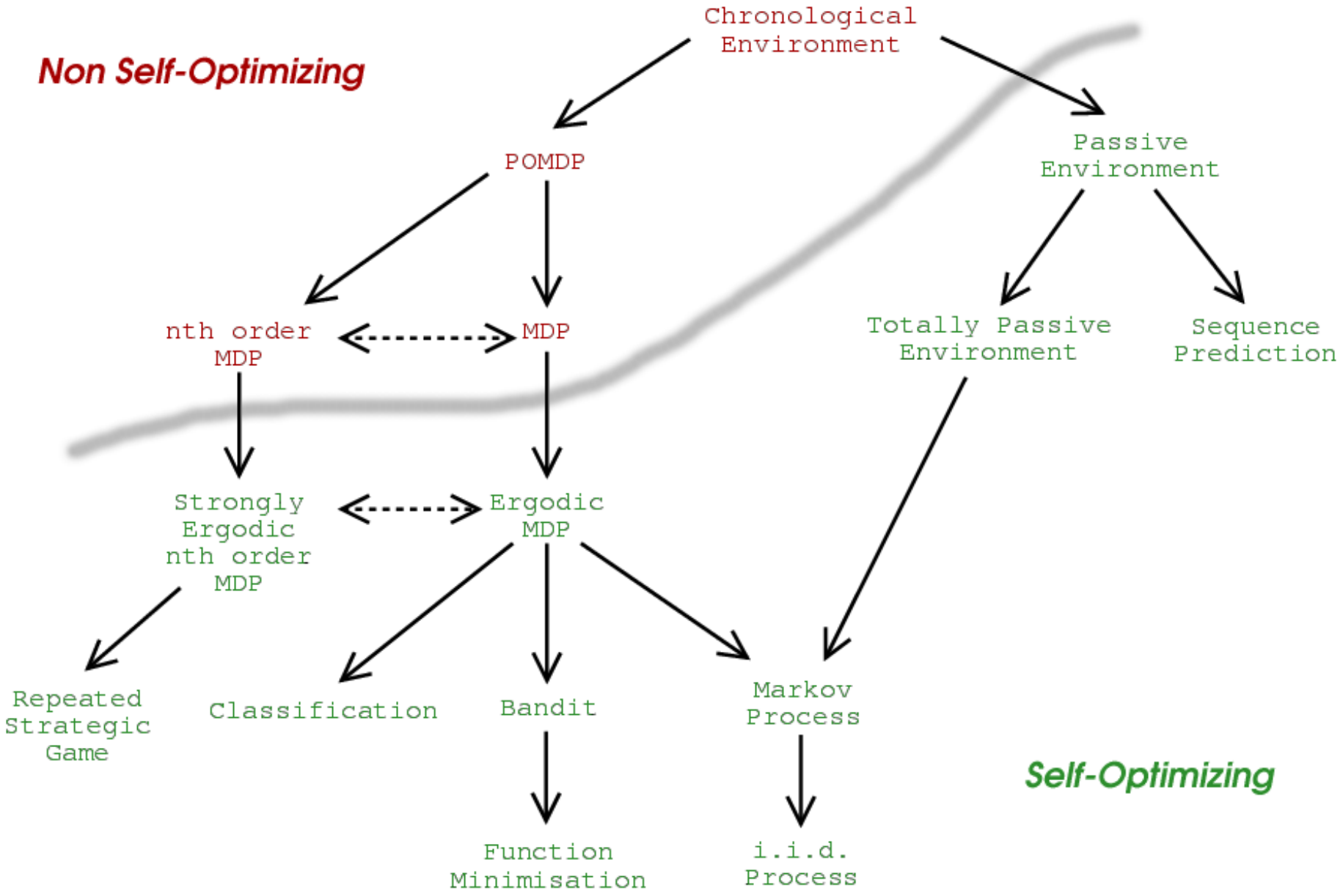
The least we must demand from $\mathcal{M}$ to have a chance that (9.3) is true is that there exists some policy $\tilde{p}$ at all with this property, i.e.

$$\exists \tilde{p}: \; V_\nu^{\tilde{p}} \to V_\nu^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \qquad (9.4)$$

Main result:

Theorem 9.5 (Self-optimizing policy p^ξ $(9.4) \Rightarrow (9.3)$)
The necessary condition of the existence of a self-optimizing policy $\tilde{p}$ is also sufficient for p^ξ to be self-optimizing.

Environments w./ (Non)Self-Optimizing Policies



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 $k \to \infty$.

9.2 Future Value and Discounting: Contents

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 $m \to \infty$.
- $V_{k\gamma}^{\pi\rho} := \frac{1}{\Gamma_k} \lim_{m\to\infty} \sum_{x_{k:m}} (\gamma_k r_k + ... + \gamma_m r_m) \rho(x_{k:m} | y_{1:m} x_{<k})_{|y_{1:m}=p(x_{<m})}$
- Further advantage: Traps (non-ergodic environments) do not necessarily prevent self-optimizing policies any more.

Results for Discounted Future Value

Theorem 9.6 (Properties of Discounted Future Value)

- $V_{k\gamma}^{\pi\rho}$ is linear in ρ: $V_{k\gamma}^{\pi\xi} = \sum_\nu w_k^\nu V_{k\gamma}^{\pi\nu}$.
- $V_{k\gamma}^{*\rho}$ is convex in ρ: $V_{k\gamma}^{*\xi} \leq \sum_\nu w_k^\nu V_{k\gamma}^{*\nu}$.
- 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 π with $V_{k\gamma}^{\pi\nu} \geq V_{k\gamma}^{p^\xi\nu}$ 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 $\exists\tilde{\pi}_k\forall\nu : V_{k\gamma}^{\tilde{\pi}_k\nu} \overset{k\to\infty}{\longrightarrow} V_{k\gamma}^{*\nu} \quad\Longrightarrow\quad V_{k\gamma}^{p^\xi\mu} \overset{k\to\infty}{\longrightarrow} V_{k\gamma}^{*\mu}$.

Continuity of Value

Theorem 9.7 (Continuity of discounted value)

The values $V_{k\gamma}^{\pi\mu}$ and $V_{k\gamma}^{*\mu}$ are continuous in μ, and $V_{k\gamma}^{p^{\hat{\mu}}\mu}$ is continuous in $\hat{\mu}$ at $\hat{\mu}=\mu$ w.r.t. a conditional 1-norm in the following sense: If $\sum_{x_k}|\mu(x_k|x_{<k}y_{1:k})-\hat{\mu}(x_k|x_{<k}y_{1:k})|\leq\varepsilon\ \forall yx_{<k}y_k\ \forall k\geq k_0$, then

$$|V_{k\gamma}^{\pi\mu}-V_{k\gamma}^{\pi\hat{\mu}}|\leq\delta(\varepsilon),\quad |V_{k\gamma}^{*\mu}-V_{k\gamma}^{*\hat{\mu}}|\leq\delta(\varepsilon),\quad |V_{k\gamma}^{*\mu}-V_{k\gamma}^{p^{\hat{\mu}}\mu}|\leq 2\delta(\varepsilon)$$

$\forall\ k\geq k_0$ and $yx_{<k}$, where $\delta(\varepsilon):=r_{max}\cdot\min_{n\geq k}\{(n-k)\varepsilon+\frac{\Gamma_n}{\Gamma_k}\}\stackrel{\varepsilon\to 0}{\longrightarrow}0$.

Warning: $V_{k\gamma}^{p^{\xi}\mu}\not\to V_{k\gamma}^{*\mu}$, since $\xi\to\mu$ does not hold for all $yx_{1:\infty}$, but only for μ-random ones.

Average Value: By setting $\gamma_k=1$ for $k\leq m$ and $\gamma_k=0$ for $k>m$ we also get continuity of $V_{km}^{\cdots}$.

Convergence of Universal to True Value

Theorem 9.8 (Convergence of universal to true value)
For a given policy p and history generated by p and μ, i.e. on-policy, the future universal value $V_{...}^{p\xi}$ converges to the true value $V_{...}^{p\mu}$:

$$V_{km_k}^{p\xi} \overset{k\to\infty}{\longrightarrow} V_{km_k}^{p\mu} \qquad \text{i.m.s.} \qquad \text{if} \quad h_{max} < \infty,$$

$$V_{k\gamma}^{p\xi} \overset{k\to\infty}{\longrightarrow} V_{k\gamma}^{p\mu} \qquad \text{i.m.} \qquad \text{for any} \;\; \gamma.$$

If the history is generated by $p = p^\xi$, this implies $V_{k\gamma}^{*\xi} \to V_{k\gamma}^{p^\xi\mu}$.

Hence the universal value $V_{k\gamma}^{*\xi}$ can be used to estimate the true value $V_{k\gamma}^{p^\xi\mu}$, without any assumptions on $\mathcal{M}$ and γ.

Nevertheless, maximization of $V_{k\gamma}^{p\xi}$ may asymptotically differ from max. of $V_{k\gamma}^{p\mu}$, since $V_{k\gamma}^{p\xi} \not\to V_{k\gamma}^{p\mu}$ for $p \neq p^\xi$ is possible (and also $V_{k\gamma}^{*\xi} \not\to V_{k\gamma}^{*\mu}$).

Markov Decision Processes (MDP)

From all possible environments, Markov (Decision) Processes are probably the most intensively studied ones.

Definition 9.9 (Ergodic MDP)
We call μ a (stationary) MDP if the probability of observing $o_k \in \mathcal{O}$ and reward $r_k \in \mathcal{R}$, only depends on the last action $y_k \in \mathcal{Y}$ and the last observation o_{k-1} (called state), i.e. if $\mu(x_k|x_{<k}y_{1:k}) = \mu(x_k|o_{k-1}y_k)$, where $x_k \equiv o_k r_k$.
An MDP μ is called ergodic if there exists a policy under which every state is visited infinitely often with probability 1.

If the transition matrix $\mu(o_k|o_{k-1}y_k)$ is independent of the action y_k, the MDP is a Markov process;

If $\mu(x_k|o_{k-1}y_k)$ is independent of o_{k-1} we have an i.i.d. process.

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 ($\sim 1/\ln\frac{1}{\gamma}$ 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: $h_k^{eff} \sim k$.

Properties of Ergodic MDPs

- Stationary MDPs μ have stationary optimal policies p^μ in case of geometric discount, mapping the same state/observation o_k always to the same action y_k.
- A mixture ξ of MDPs is itself not an MDP, i.e. $\xi \notin \mathcal{M}_{MDP} \quad \Rightarrow$ p^ξ is, in general, not a stationary policy.
- There are self-optimizing policies for the class of ergodic MDPs for the average value V_ν, and for the future value $V_{k\gamma}$ if $\frac{\gamma_{k+1}}{\gamma_k} \to 1$.
- Hence Theorems 9.5 and 9.6 imply that p^ξ is self-optimizing for ergodic MDPs (if $\frac{\gamma_{k+1}}{\gamma_k} \to 1$).
- $\frac{\gamma_{k+1}}{\gamma_k} \to 1$ for $\gamma_k = 1/k^2$, but not for $\gamma_k = \gamma^k$.
- **Fazit:** Condition $\frac{\gamma_{k+1}}{\gamma_k} \to 1$ admits *self-optimizing Bayesian policies*.

General Discounting

- Future rewards give only small contribution to $V_{k\gamma}$
  $\Rightarrow$ effective horizon.
- The only significant arbitrariness in the AIXI model lies in the choice of the horizon.
- Power damping $\gamma_k = k^{-1-\varepsilon}$ leads to horizon proportional to age k of agent.
  It does not introduce arbitrary time-scale and has natural/plausible horizon.
- Universal discount $\gamma_k = 2^{-K(k)}$ leads to largest possible horizon.
  Allows to "mimic" all other more greedy behaviors based on other discounts.

Effective Horizon

Table 9.10 (Effective horizon)

$h_k^{eff} := \min\{h \geq 0 : \Gamma_{k+h} \leq \frac{1}{2}\Gamma_k\}$ for various types of discounts γ_k

Horizons	γ_k	$\Gamma_k = \sum_{i=k}^{\infty} \gamma_i$	h_k^{eff}
finite	$\begin{matrix}1 \text{ for } k \leq m \\ 0 \text{ for } k > m\end{matrix}$	$m-k+1$	$\frac{1}{2}(m-k+1)$
geometric	$\gamma^k,\ 0 \leq \gamma < 1$	$\frac{\gamma^k}{1-\gamma}$	$\frac{\ln 2}{\ln \gamma^{-1}}$
quadratic	$\frac{1}{k(k+1)}$	$\frac{1}{k}$	k
power	$k^{-1-\varepsilon},\ \varepsilon > 0$	$\sim \frac{1}{\varepsilon} k^{-\varepsilon}$	$\sim (2^{1/\varepsilon} - 1)k$
harmonic$_{\approx}$	$\frac{1}{k \ln^2 k}$	$\sim \frac{1}{\ln k}$	$\sim k^2$
universal	$2^{-K(k)}$	decreases slower than any computable function	increases faster than any computable function

Other Attempts to Deal with Horizon Issue

- Finite horizon:
  - good if known,
  - bad if unknown and for asymptotic analysis.
- Infinite horizon:
  - Limit may not exist.
  - can delay exploitation indefinitely,
    since no finite exploration decreases value.
  - immortal agents can be lazy.
- Average reward and differential gain:
  - limit may not exist.
- Moving horizon m_k:
  - can lead to very bad time-inconsistent behavior.
- Time-inconsistent discounting ...

Time(In)Consistent Discounting

- Generalize $V_{k\gamma}^{\pi\rho} \equiv \frac{1}{\Gamma_k}\mathbb{E}^{\pi\rho}[\sum_{t=k}^{\infty}\gamma_t r_t]$ to:
  Potentially different discount sequence $d_1^k, d_2^k, d_3^k, ...$ for different k:
  Value $V_{k\gamma}^{\pi\rho} := \mathbb{E}^{\pi\rho}[\sum_{t=k}^{\infty} d_t^k r_t]$

- Leads in general to time-inconsistency,
  i.e. $\pi_k^* := \arg\max_\pi V_{k\gamma}^{\pi\rho}$ depends on k.

- Consequence: Agent plans to do one thing,
  but then changes its mind.
  Can in general lead to very bad behavior.

- Humans seem to behave time-inconsistently.
  Solution: Pre-commitment strategies.

Time(In)Consistent Discounting (ctd)

Time-consistent examples: $d_t^k = \gamma^{t-k}$ geometric discounting.

Is the only time-invariant consistent discounting

Time-inconsistent example: $d_t^k = (t-k+1)^{-1-\varepsilon}$ (≈humans)

Theorem 9.11 (Time(In)Consistent Discounting) [LH11]

d_t^k is time-consistent $\iff$ $d_{()}^k \propto d_{()}^1$ for all k.

What to do if you know you're time inconsistent?

Treat your future selves as opponents in an extensive game and follow sub-game perfect equilibrium policy.

9.3 Optimistic and Knowledge-Seeking Variations of AIξ: Contents

- Universal Knowledge-Seeking Agent
- Optimistic Agents in Deterministic Worlds
- Optimistic Agents for General Environments
- Optimism in MDPs

Universal Knowledge-Seeking Agent (KSA)

reward for exploration; goal is to learn the true environment [OLH13]

- $w_k^\nu := w_\nu \frac{\nu(x_{<k}|y_{<k})}{\xi(x_{<k}|y_{<k})}$ is the posterior belief in ν given history $yx_{<k}$.
- $w_k^{()}$ summarizes the information contained in history $yx_{<k}$.
- $w_k^{()} \rightsquigarrow w_{k+1}^{()}$ changes $\Leftrightarrow$ x_k given $yx_{<k}$ is informative about $\nu \in \mathcal{M}$
- Information gain can be quantified by KL-divergence.
- Reward agent for gained information:

$$r_k \;:=\; \mathsf{KL}(w_{k+1}^{()}||w_k^{()}) \;\equiv\; \sum_{\nu\in\mathcal{M}} w_{k+1}^\nu \log(w_{k+1}^\nu/w_k^\nu)$$

Asymptotic Optimality of Universal KSA

Theorem 9.12 (Asymptotic Optimality of Universal KSA)

- Universal π^*_ξ converges to optimal π^*_μ. More formally:
- $P^\pi_\xi(\cdot|yx_{<k})$ converges in (μ,π^*_ξ)-probability to $P^\pi_\mu(\cdot|yx_{<k})$ uniformly for all π.

Def: $P^\pi_\rho(\cdot|yx_{<k})$ is (ρ,π)-probability of future $yx_{k:\infty}$ given past $yx_{<k}$.

Note: On-policy agent π^*_ξ is able to even predict off-policy!

Remark: **No** assumption on $\mathcal{M}$ needed, i.e. Thm. applicable to $\mathcal{M}_U$.

Optimistic Agents in Deterministic Worlds

act optimally w.r.t. the most optimistic environment

until it is contradicted [SH12]

- $\pi^\circ := \pi^*_k := \arg\max_\pi \max_{\nu\in\mathcal{M}_{k-1}} V^{\pi\nu}_{k\gamma}(yx_{<k})$
- $\mathcal{M}_{k-1}$:= environments consistent with history $yx_{<k}$.
- As long as the outcome is consistent with the optimistic prediction, the return is optimal, even if the wrong environment is chosen.

Theorem 9.13 (Optimism is asymptotically optimal)

For finite $\mathcal{M} \equiv \mathcal{M}_0$,

- Asymptotic: $V^{\pi^\circ\mu}_{k\gamma} = V^{*\mu}_{k\gamma}$ for all large k.
- Errors: For geometric discount, $V^{\pi^\circ\mu}_{k\gamma} \geq V^{*\mu}_{k\gamma} - \varepsilon$ (i.e. π° ε-sub-optimal) for all but at most $|\mathcal{M}|\frac{\log\varepsilon(1-\gamma)}{\log\gamma}$ time steps k.

Optimistic Agents for General Environments

- Generalization to stochastic environments: Likelihood criterion:
  Exclude ν from $\mathcal{M}_{k-1}$ if $\nu(x_{<k}|y_{<k}) < \varepsilon_k \cdot \max_{\nu\in\mathcal{M}} \nu(x_{<k}|y_{<k})$. [SH12]
- Generalization to compact classes $\mathcal{M}$:
  Replace $\mathcal{M}$ by centers of finite ε-cover of $\mathcal{M}$ in def. of π°. [SH12]
- Use decreasing $\varepsilon_k \to 0$ to get self-optimizingness.
- There are non-compact classes for which self-optimizingness is impossible to achieve. [Ors10]
- Weaker self-optimizingness in Cesaro sense possible
  by starting with finite subset $\mathcal{M}_0 \subset \mathcal{M}$
  and adding environments ν from $\mathcal{M}$ over time to $\mathcal{M}_k$. [SH15]
- Fazit: There exist (weakly) self-optimizing policies for arbitrary (separable) /compact $\mathcal{M}$.

Optimism in MDPs

- Let $\mathcal{M}$ be the class of all MDPs with $|\mathcal{S}| < \infty$ states and $|\mathcal{A}| < \infty$ actions and geometric discount γ.
- Then $\mathcal{M}$ is continuous but compact
  $\Longrightarrow$ π° is self-optimizing by previous slide.
- But much better polynomial error bounds in this case possible:

Theorem 9.14 (PACMDP bound) $V_{k\gamma}^{\pi^\circ\mu} \leq V_{k\gamma}^{*\mu} - \varepsilon$ for at most $\tilde{O}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\log\frac{1}{\delta})$ time steps k with probability $1-\delta$. [LH12]

9.4 Discussion: Contents

- Summary
- Exercises
- Literature

Summary - Bayesian Agents

- 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.

Summary - Discounting

- Discounting: Considering future values and the right discounting γ leads to more meaningful agents and results.
- 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.
- In particular: Policy p^ξ with unbounded effective horizon is the first purely Bayesian self-optimizing consistent policy for ergodic MDPs.
- Wrong discounting leads to myopic or time-inconsistent policies (bad).

Summary - Variations of AIξ

- Use information gain as a universal choice for the rewards.
  AIξ becomes purely knowledge seeking.

- Real world has traps
  $\Longrightarrow$ no self-optimizing policy
  $\Longrightarrow$ need more explorative policies and weaker criteria like ...

- Optimistic agents: Act optimally w.r.t. the most optimistic environment until it is contradicted.

Exercises

1. [C15] Prove Pareto-optimality of p^ξ.
2. [C35] Prove Theorem 9.7 (Continuity of discounted value).
3. [C35] Prove Theorem 9.8 (Convergence of universal to true value).
4. [C15ui] Solve [Hut05, Problem 5.2]
   (Absorbing two-state environment)
5. [C25u] Derive the expressions for the effective horizons in Table 9.10.
6. [C30ui] Solve [Hut05, Problem 5.11] (Belief contamination)
7. [C20u] Solve [Hut05, Problem 5.16] (Effect of discounting)

Literature

[BT96] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

[KV86] P. R. Kumar and P. P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, Englewood Cliffs, NJ, 1986.

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

[Lat14] T. Lattimore. *Theory of General Reinforcement Learning*. PhD thesis, Research School of Computer Science, Australian National University, 2014.