COMP4620/8620: ADVANCED TOPICS IN AI
FOUNDATIONS OF ARTIFICIAL INTELLIGENCE

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/
9 THEORY OF RATIONAL AGENTS

• The Bayesian Agent Alξ

• 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 – Alρ Model
- The Bayes-Mixture Distribution ξ
- Questions of Interest
- Linearity and Convexity of Vρ 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)

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

The average value of policy $p$ with horizon $m$ in environment $\rho$ is defined as

$$V^p_\rho := \frac{1}{m} \sum_{x_{1:m}} (r_1 + \ldots + r_m) \rho(x_{1:m}|y_{1:m})|y_{1:m}=p(x_{<m})$$

The goal of the agent should be to maximize the value.
Optimal Policy and Value – $\text{AI}_\rho$ Model

The $\rho$-optimal policy $p^\rho := \arg\max_p V^p_\rho$ maximizes $V^p_\rho \leq V^*_\rho := V^{p^\rho}_\rho$.

Explicit expressions for the action $y_k$ in cycle $k$ of the $\rho$-optimal policy $p^\rho$ and their value $V^*_\rho$ are

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

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

Keyword: Expectimax tree/algorithm.
The Bayes-Mixture Distribution $\xi$

Assumption: The true environment $\mu$ is unknown.

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

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

The Bayes-mixture $\xi$ is defined as

$$\xi(x_1:m|y_1:m) := \sum_{\nu \in \mathcal{M}} w_{\nu}\nu(x_1:m|y_1:m)$$

with

$$\sum_{\nu \in \mathcal{M}} w_{\nu} = 1, \quad w_{\nu} > 0 \quad \forall \nu$$

The weights $w_{\nu}$ may be interpreted as the prior degree of belief that the true environment is $\nu$.

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 $p^\xi$ which maximizes $V^p_\xi$.

- If $\mu$ is the true environment the expected reward when following policy $p^\xi$ will be $V^{p^\xi}_\mu$.

- The optimal (but infeasible) policy $p^\mu$ yields reward $V^{p^\mu}_\mu \equiv V^*_\mu$.

- Are there policies with uniformly larger value than $V^{p^\xi}_\mu$?

- How close is $V^{p^\xi}_\mu$ to $V^*_\mu$?

- What is the most general class $\mathcal{M}$ and weights $w_\nu$?

$$\mathcal{M} = \mathcal{M}_U \text{ and } w_\nu = 2^{-K(\nu)} \implies \text{AI}_\xi = \text{AIXI}!$$
Linearity and Convexity of $V_\rho$ in $\rho$

**Theorem 9.1 (Linearity and convexity of $V_\rho$ in $\rho$)**

- $V_\rho^p$ is a **linear** function in $\rho$:  
  $$V_\xi^p = \sum_\nu w_\nu V_\nu^p$$

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

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 $V_\rho$.

**Loose interpretation:** A mixture can never increase performance.
Pareto Optimality

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

\begin{center}
\textbf{Theorem 9.2 (Pareto optimality of $p^\xi$)} Policy $p^\xi$ is Pareto-optimal in the sense that there is no other policy $p$ with $V^p_\nu \geq V^{p^\xi}_\nu$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one $\nu$. \\
From a practical point of view a significant increase of $V$ for many environments $\nu$ may be desirable even if this causes a small decrease of $V$ for a few other $\nu$. This is impossible due to \\
Balanced Pareto optimality: \\
$\Delta_\nu := V^{p^\xi}_\nu - V^\tilde{p}_\nu$, \quad $\Delta := \sum_\nu w_\nu \Delta_\nu$ \quad $\Rightarrow$ \quad $\Delta \geq 0$.
\end{center}
Self-optimizing Policies

Under which circumstances does the value of the universal policy \( p^\xi \) converge to optimum?

\[
V_{\nu}^{p^\xi} \rightarrow V_{\nu}^{*} \quad \text{for horizon} \quad m \rightarrow \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \tag{9.3}
\]

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.

\[
\exists \tilde{p} : V_{\nu}^{\tilde{p}} \rightarrow V_{\nu}^{*} \quad \text{for horizon} \quad m \rightarrow \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \tag{9.4}
\]

Main result:

**Theorem 9.5 (Self-optimizing policy \( p^\xi \) (9.4) \( \Rightarrow \) (9.3))**

The necessary condition of the existence of a self-optimizing policy \( \tilde{p} \) is also sufficient for \( p^\xi \) to be self-optimizing.
Environments w./ (Non)Self-Optimizing Policies

Non Self-Optimizing

- POMDP
- nth order MDP
- Strongly Ergodic nth order MDP
- Repeated Strategic Game
- Classification
- Bandit

Self-Optimizing

- Chronological Environment
- Passive Environment
- Totally Passive Environment
- Sequence Prediction
- Markov Process
- i.i.d. Process
- Function Minimisation
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 $p^\xi$ 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

- Results for Discounted Future Value
- Continuity of Value
- Convergence of Universal to True Value
- Markov Decision Processes (MDP)
- Importance of the Right Discounting
- Properties of Ergodic MDPs
- General Discounting
- Effective Horizon
- Other Attempts to Deal with the Horizon Issue
- Time(In)Consistent Discounting
Future Value and Discounting

- Eliminate the horizon by discounting the rewards $r_k \sim \gamma_k r_k$ with $\Gamma_k := \sum_{i=k}^{\infty} \gamma_i < \infty$ and letting $m \to \infty$.

- $V^{\pi\rho}_{k\gamma} := \frac{1}{\Gamma_k} \lim_{m \to \infty} \sum_{x_k:m} (\gamma_k r_k + \ldots + \gamma_m r_m) \rho(x_k:m | y_1:m x < k) | y_1:m = p(x < m)$

- Further advantage: Traps (non-ergodic environments) do not necessarily prevent self-optimizing policies any more.
Theorem 9.6 (Properties of Discounted Future Value)

- \( V_{k\gamma}^{\pi\rho} \) is linear in \( \rho \):
  \[ V_{k\gamma}^{\pi\xi} = \sum_{\nu} w_{k}^{\nu} V_{k\gamma}^{\pi\nu}. \]

- \( V_{k\gamma}^{*\rho} \) is convex in \( \rho \):
  \[ V_{k\gamma}^{*\xi} \leq \sum_{\nu} w_{k}^{\nu} V_{k\gamma}^{*\nu}. \]

- where \( w_{k}^{\nu} := w_{\nu} \frac{\nu(x < k | y < k)}{\xi(x < k | y < k)} \) is the posterior belief in \( \nu \).

- \( p^{\xi} \) is Pareto-optimal in the sense that there is no other policy \( \pi \) 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 \( \nu \).

- If there exists a self-optimizing policy for \( \mathcal{M} \), then \( p^{\xi} \) is self-optimizing in the sense that
  \[ \exists \tilde{\pi}_{k} \forall \nu : V_{k\gamma}^{\tilde{\pi}_{k}\nu} k \rightarrow \infty V_{k\gamma}^{*\nu} \Rightarrow V_{k\gamma}^{p^{\xi}\mu} k \rightarrow \infty V_{k\gamma}^{*\mu}. \]
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 $\mu$, 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 $\delta(\varepsilon) := r_{max} \cdot \min_{n \geq k}\{(n - k)\varepsilon + \frac{\Gamma_n}{\Gamma_k}\} \xrightarrow{\varepsilon \to 0} 0$.

Warning: $V_{k\gamma}^{p\xi\mu} \rightarrow V_{k\gamma}^{*\mu}$, since $\xi \rightarrow \mu$ does not hold for all $yx_{1:\infty}$, but only for $\mu$-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 $V_{km}$.
**Theorem 9.8 (Convergence of universal to true value)**

For a given policy \(p\) and history generated by \(p\) and \(\mu\), i.e. on-policy, the future universal value \(V^{p\xi}_{k} \) converges to the true value \(V^{p\mu}_{k}\):

\[
V^{p\xi}_{km_k} \xrightarrow{k \to \infty} V^{p\mu}_{km_k} \quad \text{i.m.s. if } h_{\text{max}} \leq \infty,
\]

\[
V^{p\xi}_{k\gamma} \xrightarrow{k \to \infty} V^{p\mu}_{k\gamma} \quad \text{i.m. for any } \gamma.
\]

If the history is generated by \(p = p^{\xi}\), this implies \(V^{*\xi}_{k\gamma} \to V^{p\xi\mu}_{k\gamma}\).

Hence the universal value \(V^{*\xi}_{k\gamma}\) can be used to estimate the true value \(V^{p\xi\mu}_{k\gamma}\), without any assumptions on \(M\) and \(\gamma\).

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

An MDP $\mu$ 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 \( h_k^{\text{eff}} \) is finite (\( \sim 1 / \ln \frac{1}{\gamma} \) for \( \gamma_k = \gamma^k \)).

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

Result: Policy \( p^\xi \) is self-optimizing for the class of (\( l^{\text{th}} \) 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: \( h_k^{\text{eff}} \sim k \).
Properties of Ergodic MDPs

- Stationary MDPs $\mu$ have stationary optimal policies $p^\mu$ in case of geometric discount, mapping the same state/observation $o_k$ always to the same action $y_k$.

- A mixture $\xi$ of MDPs is itself not an MDP, i.e. $\xi \notin \mathcal{M}_{MDP} \Rightarrow p^\xi$ is, in general, not a stationary policy.

- There are self-optimizing policies for the class of ergodic MDPs for the average value $V_\nu$, 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 $p^\xi$ 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)

\[ h_k^{\text{eff}} := \min\{h \geq 0 : \Gamma_{k+h} \leq \frac{1}{2} \Gamma_k\} \] for various types of discounts \( \gamma_k \)

<table>
<thead>
<tr>
<th>Horizons</th>
<th>( \gamma_k )</th>
<th>( \Gamma_k = \sum_{i=k}^{\infty} \gamma_i )</th>
<th>( h_k^{\text{eff}} )</th>
</tr>
</thead>
</table>
| finite     | \( 1 \) for \( k \leq m \)
             | \( 0 \) for \( k > m \)                       | \( m - k + 1 \)         | \( \frac{1}{2} (m - k + 1) \) |
| 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   | \( \frac{1}{k \ln^2 k} \)                     | \( \sim \frac{1}{\ln k} \)                | \( \sim k^2 \)           |
| 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 $d_1^k, d_2^k, d_3^k, \ldots$ 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: \( d_t^k = (t - k + 1)^{-1-\varepsilon} \) (\( \approx \) humans)

**Theorem 9.11 (Time(In)Consistent Discounting)**

\[ d_t^k \text{ is time-consistent} \iff d_t^k \propto d_1^k \text{ for all } k. \]  

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 $\nu$ given history $yx^<_k$.

- $w_k()$ summarizes the information contained in history $yx^<_k$.

- $w_k() \leadsto w_{k+1}()$ changes $\iff$ $x_k$ given $yx^<_k$ is informative about $\nu \in \mathcal{M}$.

- Information gain can be quantified by KL-divergence.

- Reward agent for gained information:
  \[ r_k := \text{KL}(w_{k+1}()|w_k()) \equiv \sum_{\nu \in \mathcal{M}} w_{k+1}^\nu \log \left( \frac{w_{k+1}^\nu}{w_k^\nu} \right) \]
Asymptotic Optimality of Universal KSA

Theorem 9.12 (Asymptotic Optimality of Universal KSA)

• Universal $\pi_\xi^*$ converges to optimal $\pi_\mu^*$. More formally:
• $P^\pi_{\xi}(\cdot|yx<k)$ converges in $(\mu, \pi_\xi^*)$-probability to $P^\pi_{\mu}(\cdot|yx<k)$ uniformly for all $\pi$.

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

Note: On-policy agent $\pi_\xi^*$ 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]

- $\pi^o := \pi^*_k := \arg \max_{\pi} \max_{\nu \in \mathcal{M}_{k-1}} V_{k\gamma}^{\pi\nu}(y_x < k)$
- $\mathcal{M}_{k-1} := \text{environments consistent with history } y_x < k$
- 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: $V_{k\gamma}^{\pi^o\mu} = V_{k\gamma}^{*\mu}$ for all large $k$.
- Errors: For geometric discount, $V_{k\gamma}^{\pi^o\mu} \geq V_{k\gamma}^{*\mu} - \varepsilon$ (i.e. $\pi^o \varepsilon$-sub-optimal) for all but at most $|\mathcal{M}| \frac{\log \frac{\varepsilon}{1-\gamma}}{\log \gamma}$ time steps $k$. 
Optimistic Agents for General Environments

• **Generalization to stochastic environments:** Likelihood criterion:
  Exclude $\nu$ 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 $\varepsilon$-cover of $\mathcal{M}$ in def. of $\pi^\circ$. [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 $\nu$ 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 $|S| < \infty$ states and $|A| < \infty$ actions and geometric discount $\gamma$.

- Then $\mathcal{M}$ is continuous but compact
  $$\Rightarrow \pi^{\circ} \text{ is self-optimizing by previous slide.}$$

- But much better polynomial error bounds in this case possible:

\[
\tilde{O}(\frac{|S|^2|A|}{\varepsilon^2(1-\gamma)^3} \log \frac{1}{\delta}) \text{ time steps } k \text{ with probability } 1 - \delta.
\]

[Theorem 9.14 (PACMDP bound)]
9.4 Discussion: Contents

- Summary
- Exercises
- Literature
Summary - Bayesian Agents

- **Setup:** Agents acting in general probabilistic environments with reinforcement feedback.

- **Assumptions:** True environment $\mu$ belongs to a known class of environments $\mathcal{M}$, but is otherwise unknown.

- **Results:** The Bayes-optimal policy $p^\xi$ 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 $\gamma$ 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 $p^\xi$ 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 $\text{AI}^\xi$

- Use information gain as a universal choice for the rewards. $\text{AI}^\xi$ becomes purely knowledge seeking.

- Real world has traps
  $\implies$ no self-optimizing policy
  $\implies$ 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 $p^\xi$.

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


