

Dynamic Programming and Optimal Policies for MDPs under Epistemic Uncertainty

Axel Benyamine

CMAP, École polytechnique · Inria Paris

joint work with Alain Durmus, Julien Grand-Clément, Marek Petrik,
and Michael I. Jordan

Dynamic Programming for Epistemic Uncertainty in Markov Decision Processes, ICML 2026

On the Existence of Uniformly Optimal Policies for MDPs under Epistemic Uncertainty, under review

Discounted Markov Decision Processes

Discounted MDP:

$$M_P = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

Solve:

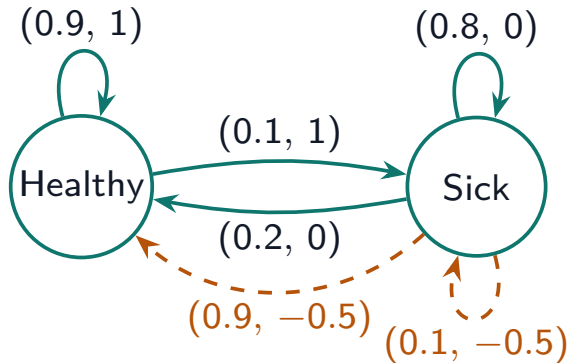
$$\sup_{\pi \in \Pi_H} \mathbb{E}_{\mu}^{\pi, P} \left[\sum_{t \geq 0} \gamma^t r(\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}) \right],$$

with $\mu \in \Delta(\mathcal{S})$ the initial state distribution.

Sets of policies:

$$\Pi_S \subset \Pi_H$$

stationary $\subset$ history-dependent



arcs: (probability, reward)

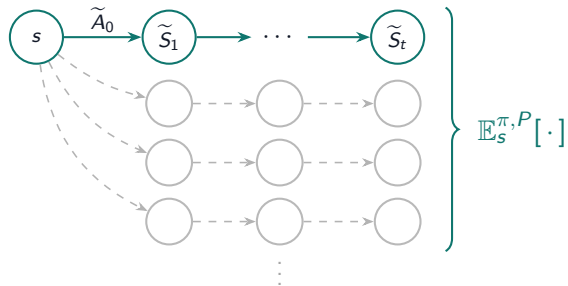
— wait - - - treat

Aleatoric randomness

P known, π fixed:
the trajectory $(\tilde{S}_t, \tilde{A}_t)$ is still random.

Average it out:

$$V^{\pi, P}(s) = \mathbb{E}_s^{\pi, P} \left[\sum_{t \geq 0} \gamma^t r(\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}) \right]$$



one model, many trajectories

Value function: $V^{\pi, P} \in \mathbb{R}^{\mathcal{S}}$ (deterministic).

Optimal Value function: $V^P = (\sup_{\pi \in \Pi_H} V^{\pi, P}(s))_{s \in \mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ (deterministic).

Expected return when starting from $\mu \in \Delta(\mathcal{S})$: $\mu^\top V^{\pi, P}$.

Dynamic programming

Bellman operator $T^{\pi,P} : \mathbb{R}^{\mathcal{S}} \rightarrow \mathbb{R}^{\mathcal{S}}$:

$$\forall v \in \mathbb{R}^{\mathcal{S}}, \quad T^{\pi,P} v(s) = \sum_a \pi(s, a) \sum_{s'} P(s, a, s') (r(s, a, s') + \gamma v(s')),$$

Optimal Bellman operator $T^P : \mathbb{R}^{\mathcal{S}} \rightarrow \mathbb{R}^{\mathcal{S}}$:

$$\forall v \in \mathbb{R}^{\mathcal{S}}, \quad T^P v(s) = \sup_{\pi \in \Pi_{\mathcal{S}}} T^{\pi,P} v(s).$$

Theorem (Bellman; Puterman, 1994)

For any discounted MDP and any policy $\pi \in \Pi_{\mathcal{S}}$:

$$V^{\pi,P} = T^{\pi,P} V^{\pi,P}$$

$$V^P = T^P V^P$$

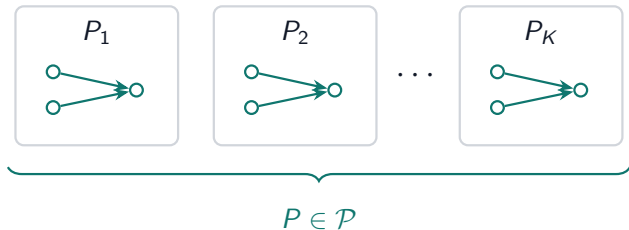
Consequences:

- Algorithms with linear rate: Policy Iteration & Value Iteration
- $\exists \pi^* \in \Pi_{\mathcal{S}}$ optimal for every μ (*uniformly optimal*).

Epistemic uncertainty

P is estimated from data.

$\hat{P} \neq P \implies$ policy degrades



Aleatoric noise inside a model.

Epistemic which model?

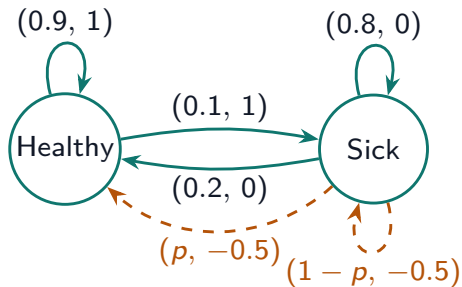
Robust MDPs

Maximize the worst-case return:

$$\sup_{\pi \in \Pi_H} \inf_{P \in \mathcal{P}} \mu^\top V^{\pi, P}$$

$\mathcal{P}$ compact convex (confidence region)

Iyengar (2005); Nilim & El Ghaoui (2005);
Wiesemann et al. (2013)



$$p \in \mathcal{P} = [0.6, 0.9]$$

evaluate at the worst admissible p

— wait - - - treat

Robust dynamic programming

Robust Bellman operator $T^{\pi, \mathcal{P}}, T^{\mathcal{P}} : \mathbb{R}^{\mathcal{S}} \rightarrow \mathbb{R}^{\mathcal{S}}$:

$$\forall v \in \mathbb{R}^{\mathcal{S}}, \quad T^{\pi, \mathcal{P}} v(s) = \inf_{P \in \mathcal{P}} T^{\pi, P} v(s), \quad T^{\mathcal{P}} v(s) = \sup_{\pi \in \Pi_{\mathcal{S}}} T^{\pi, \mathcal{P}} v(s),$$

Robust value functions $V^{\pi, \mathcal{P}}, V^{\mathcal{P}} \in \mathbb{R}^{\mathcal{S}}$:

$$V^{\pi, \mathcal{P}}(s) = \inf_{P \in \mathcal{P}} V^{\pi, P}(s), \quad V^{\mathcal{P}}(s) = \sup_{\pi \in \Pi_{\mathcal{H}}} V^{\pi, \mathcal{P}}(s).$$

Property

If $V^{\pi, \mathcal{P}} = T^{\pi, \mathcal{P}} V^{\pi, \mathcal{P}}$ and $V^{\mathcal{P}} = T^{\mathcal{P}} V^{\mathcal{P}}$, then:

- value iteration, policy iteration: linear rate.
- $\exists$ uniformly optimal *stationary* policy.

s-rectangularity

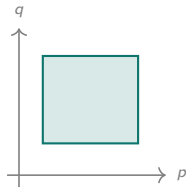
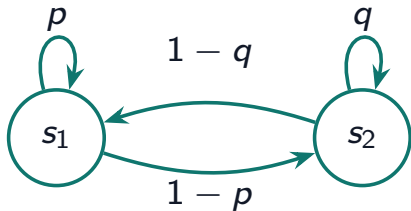
Definition: $\mathcal{P}$ is *s-rectangular* if

$$\mathcal{P} = \bigtimes_{s \in \mathcal{S}} \mathcal{P}_s$$

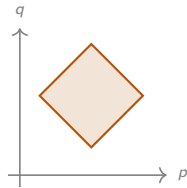
what nature may do in s does not constrain s'

Examples:

- $|p - \hat{p}| + |q - \hat{q}| \leq \kappa$ (L^1 -ball)
 $\Rightarrow$ ~~s-rectangular~~.
- $\max\{|p - \hat{p}|, |q - \hat{q}|\} \leq \kappa$ (L^∞ -ball)
 $\Rightarrow$ s-rectangular.



product set



budget: coupled

Robust DP holds *exactly* on rectangular sets

Theorem (Wiesemann et al., 2013; Grand-Clément et al., 2024)

Robust Bellman equations $\iff \mathcal{P}$ is s -rectangular

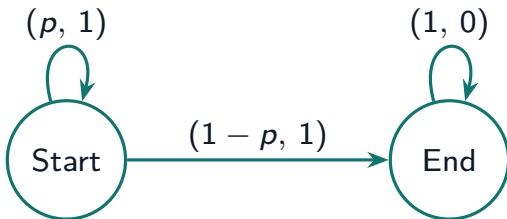
Good: as tractable as nominal MDPs.

Bad: strong assumption, very conservative.

Can we be less pessimistic and keep Dynamic Programming?

Average possible models $P \in \mathcal{P}$:

$$\sup_{\pi \in \Pi_H} \int_{\mathcal{P}} \mu^\top V^{\pi, P} dP$$



$$p \in [0, 1], \quad \gamma = \frac{1}{2}$$

$$V(\text{Start}) = \int_{[0,1]} \frac{1}{1 - \gamma p} dp = 2 \log 2 \quad \neq \quad TV(\text{Start}) = 1 + \frac{\log 2}{2}$$

One action, and Bellman equations already fail.

Which models are compatible with Dynamic Programming?

Model	Uncertainty aggregated by	DP?
Robust	worst plausible kernel	yes
Optimistic	best plausible kernel	yes
Multi-model	average over kernels	no
Percentile	a quantile of the return	no
CVaR, entropic risk, ...	softer risk criteria	?

One language to rule them all, and what else can work?

Risk measures

A map from random variables to real numbers:

$$\rho : \{\text{bounded random variables}\} \longrightarrow \mathbb{R}$$

Examples: $\mathbb{E}$ ess inf ess sup quantiles

Standard assumptions

- **Law-invariant** depends only on the distribution
- **Monotone** $\tilde{X} \geq \tilde{Y}$ a.s. $\Rightarrow \rho(\tilde{X}) \geq \rho(\tilde{Y})$
- **Translation-invariant** $\rho(\tilde{X} + c) = \rho(\tilde{X}) + c$

Ambiguity-averse MDPs

P modeled as a **random variable**:

$$\tilde{P} \sim \nu \implies \mu^\top V^{\pi, \tilde{P}} \text{ random}$$

$$\boxed{\sup_{\pi \in \Pi_H} \rho(\mu^\top V^{\pi, \tilde{P}})}$$

ess inf $\mapsto$ Robust MDP

$\mathbb{E} \mapsto$ multi-model MDP

Value functions and operators, for $\pi \in \Pi_S$:

$$V^{\pi, \nu, \rho}(s) = \rho(V^{\pi, \tilde{P}}(s)),$$

$$V^{\nu, \rho}(s) = \sup_{\pi \in \Pi_H} V^{\pi, \nu, \rho}(s),$$

$$T^{\pi, \nu, \rho} v(s) = \rho(T^{\pi, \tilde{P}} v(s)),$$

$$T^{\nu, \rho} v(s) = \sup_{\pi \in \Pi_S} T^{\pi, \nu, \rho} v(s).$$

Bellman Equations

$$V^{\pi, \nu, \rho} = T^{\pi, \nu, \rho} V^{\pi, \nu, \rho}, \quad V^{\nu, \rho} = T^{\nu, \rho} V^{\nu, \rho}$$

What Dynamic Programming buys us

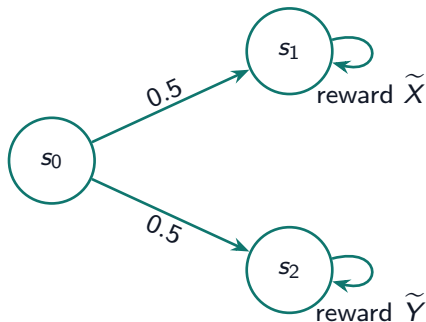
Property

If Bellman equations hold,

- value iteration, policy iteration: linear rate;
- $\exists \pi^* \in \Pi_S$ **uniformly optimal** (optimal for every starting distribution μ).

So: **which ρ ?**

One tiny instance already constrains ρ



one action

$$V(s_0) \approx \rho\left(\frac{\tilde{X} + \tilde{Y}}{2}\right)$$

$$TV(s_0) \approx \rho\left(\frac{\rho(\tilde{X}) + \rho(\tilde{Y})}{2}\right)$$

$V = TV$ forces

$$\rho(\tilde{X} + \tilde{Y}) = \rho(\tilde{X}) + \rho(\tilde{Y})$$

With richer branches: other restrictive conditions on ρ .

Dynamic Programming forces algebraic identities

Proposition (ρ law-invariant, monotone)

If ρ satisfies Bellman equations,

- $\rho(x) = x$ on constants;
- $\rho(\tilde{X} + \tilde{Y}) = \rho(\tilde{X}) + \rho(\tilde{Y})$ for $\tilde{X} \perp\!\!\!\perp \tilde{Y}$;
- $\rho(\tilde{X}\tilde{Y}) = \rho(\tilde{X})\rho(\tilde{Y})$ for $\tilde{X} \perp\!\!\!\perp \tilde{Y} \geq 0$.

Out: $\tilde{X} \mapsto \log \mathbb{E}[e^{\beta\tilde{X}}]$ (homogeneity)

Out: $\text{CVaR}_\alpha, \text{VaR}_\alpha$ (additivity)

Still standing: $\mathbb{E}$, ess inf , ess sup

Is the list complete? Yes !

Characterisation

Theorem (ρ law-invariant, monotone)

Bellman equations for s-rectangular MDPs $\iff \rho \in \{\text{ess inf}, \text{ess sup}\}$

Dynamic programming = pure pessimism or pure optimism.

Is Dynamic Programming the real obstruction?

Can we use another form of Dynamic Programming ? Or something else?

Another DP?

augmented states, nested risk measures

Another method?

policy gradient on $\rho(\mu^\top V^{\pi, \tilde{P}})$

Ask for the property directly:

Can other methods give the same guarantee? But first, does the guarantee even exist?

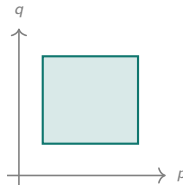
Robust MDPs: rectangularity is also *necessary*

Theorem ($\mathcal{P}$ compact convex)

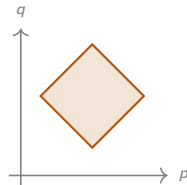
For a Robust MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P})$,

For all (r, γ) , $\exists \pi^*$ uniformly optimal and stationary $\iff \mathcal{P}$ is s -rectangular
optimal for all starting distributions μ

- Worst cases cannot be optimized simultaneously if $\mathcal{P}$ is not rectangular.



product set



budget: coupled

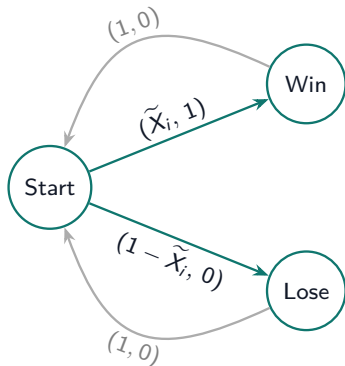
Ambiguity-averse MDPs: same verdict

Theorem (ρ law-invariant, monotone, non-expansive)

$$\text{uniformly optimal stationary } \pi^* \iff \rho \in \{\text{ess inf}, \text{ess sup}\}$$

- Excludes multi-model MDPs and percentile optimisation.

Why: explore, then exploit



2 actions; $\tilde{X}_1, \tilde{X}_2$ i.i.d., drawn once

stationary:

$$\sup_{\lambda \in [0,1]} \rho(\lambda \tilde{X}_1 + (1 - \lambda) \tilde{X}_2) \approx \max_i \rho(\tilde{X}_i)$$

history-dependent: $\rho(\max_i \tilde{X}_i)$

Forced

$$\rho(\tilde{X}_1) = \max_i \rho(\tilde{X}_i) = \rho(\max_i \tilde{X}_i)$$

m i.i.d. copies: $\max_i \tilde{X}_i \rightarrow \text{ess sup}(\tilde{X})$ a.s.

Takeaways

ρ	Dynamic Programming	uniformly optimal $\pi^* \in \Pi_S$
ess inf, ess sup	yes	yes
$\mathbb{E}$	no	no
anything else	no	no

- One framework: $\rho(\mu^\top V^{\pi, \tilde{P}})$ for robust, multi-model, percentile, ...
- DP and uniformly optimal stationary policies: the *same* restriction.
- Beyond the worst case: give up stationarity, or give up uniformity.

Dynamic Programming for Epistemic Uncertainty in Markov Decision Processes, ICML 2026
On the Existence of Uniformly Optimal Policies for MDPs under Epistemic Uncertainty, under review