Dynamic Programming for Epistemic Uncertainty in Markov Decision Processes

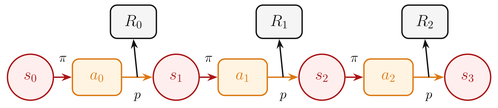
Axel Benyamine [1,2] Julien Grand-Clément [3] Marek Petrik [4] Michael I. Jordan [2,5] Alain Durmus [1]

[1]École polytechnique, IP Paris [2]Inria Paris [3]HEC Paris [4]University of New Hampshire [5]UC Berkeley

TL;DR: *(i)* Treating transition probabilities as random variables gives a **unified language for epistemic uncertainty**; *(ii)* **Bellman Dynamic Programming** is **only** compatible with **pessimistic, optimistic and neutral** risk-attitudes.

Model Estimation Errors in MDPs

In nominal discounted MDPs $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, the transition kernel P **is fixed and known**, and the objective is to maximize the expected return:



$$\sup_{\pi \in \Pi_{\mathsf{H}}} \mathbb{E}_{\mu}^{\pi, P}\Big[\sum_{t \geqslant 0} \gamma^t r(\widetilde{S}_t, \widetilde{A}_t, \widetilde{S}_{t+1})\Big].$$

In practice, **the kernel is estimated from limited data**, which can be critical in high-stakes applications:

Kernel estimated from limited data ⟹ Epistemic uncertainty (i.e., on the model parameters) ⟹ Unable to find the true optimal policy

Two types of uncertainty coexist:

- aleatoric (on the trajectories)
- epistemic (on the model parameters)

Many existing Models for Epistemic Uncertainty

Model	Treatment of uncertainty	DP Compatible?
Robust MDP	**worst** plausible transition kernel	yes
Multi-model MDP	**average** over plausible kernels	sometimes yes
Percentile optimization	optimize a **quantile** of return	no
CVaR, entropic risk	smoother risk-aware objectives	?

Problem: All these models were **introduced independently**, with ***ad hoc*** analyses.

Main Reseach Questions:

Can we unify these approaches? Which epistemic-uncertainty objectives are compatible with Bellman-style dynamic programming?

Ambiguity-Averse MDPs

Core modeling steps:

- Treat the transition kernel as a **random variable**: $\widetilde{P} \sim \nu$.
- Policies π induce **(random) expected returns** $\mu^\top V^{\pi, \widetilde{P}} = \mathbb{E}_{\mu}^{\pi, \widetilde{P}}\big(\sum_{t=0}^{\infty} \gamma^t r(\widetilde{S}_t, \widetilde{A}_t, \widetilde{S}_{t+1})\big)$.
- Use a **risk measure** $\rho : \{\text{RVs}\} \to \mathbb{R}$ to evaluate them and solve $\sup_{\pi \in \Pi_{\mathsf{H}}} \rho(\mu^\top V^{\pi, \widetilde{P}})$.

Definition

An **ambiguity-averse MDP** $(\mathcal{S}, \mathcal{A}, \nu, r, \gamma, \mu)$ is an MDP with a random transition kernel $\widetilde{P} \sim \nu$, where ν is a distribution over transition kernels.

Objective. For a risk measure $\rho : \{\text{RVs}\} \to \mathbb{R}$, find

$$\pi^\star \in \arg\sup_{\pi \in \Pi_{\mathsf{H}}} \rho(\mu^\top V^{\pi, \widetilde{P}}).$$

Ambiguity-averse MDPs recover most epistemic-uncertainty models in the litterature:

- $\nu = \delta_P$: nominal MDP.
- $\rho = \operatorname{ess\,inf}$: robust MDP.
- $\rho = \mathrm{VaR}_\alpha$: percentile objective.
- $\rho = \ldots$: any other risk attitude.
- $\rho = \mathbb{E}^\nu$: multi-model MDP.
- $\rho = \operatorname{ess\,sup}$: optimistic MDP.
- $\rho = \mathrm{CVaR}_\alpha$: tail-risk objective.

Two ways Epistemic Uncertainty Evolves

Static kernels. One kernel is sampled from ν at $t = 0$ and remains fixed:
$\widetilde{P}_0 = \widetilde{P}_1 = \widetilde{P}_2 = \cdots$.

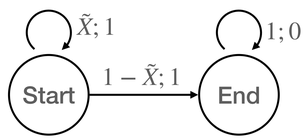
Resampled kernels. A fresh plausible kernel is sampled from ν at each period $t \in \mathbb{N}$:
$\widetilde{P}_0, \widetilde{P}_1, \widetilde{P}_2, \ldots$ i.i.d.

A Static Kernel Example

MDP with two states, a **unique action** and $\widetilde{X} \sim \mathrm{Unif}([0, 1])$.

For $\rho = \mathbb{E}^\nu$:

$$\mathbb{E}\Big[V_{\text{Start}}^{\pi, \widetilde{P}}\Big] = \mathbb{E}\left[\frac{1}{1 - \gamma\widetilde{X}}\right] = \frac{-\log(1-\gamma)}{\gamma}.$$



Edges (s, a, s') are labeled with pairs $(P(s, a, s'), r(s, a, s'))$.

Ambiguity-Averse Dynamic Programming

For each state s, define the **ambiguity-averse value** of a policy and the **optimal value**:

$$V^{\pi,\nu,\rho}(s) = \rho\Big(V^{\pi, \widetilde{P}}(s)\Big), \qquad V^{\nu,\rho}(s) = \sup_{\pi \in \Pi_{\mathsf{H}}} V^{\pi,\nu,\rho}(s).$$

For a stationary policy $\pi \in \Pi_{\mathsf{S}}$, define the **ambiguity-averse Bellman operator** and the **optimal operator**:

$$T^{\pi,\nu,\rho} v(s) = \rho\Big(T^{\pi, \widetilde{P}} v(s)\Big), \qquad T^{\nu,\rho} v(s) = \sup_{\pi \in \Pi_{\mathsf{S}}} T^{\pi,\nu,\rho} v(s).$$

We say that ρ satisfies the **Dynamic Programming principles** if the value functions are fixed points of the Bellman operators:

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

Proposition. If ρ satisfies the Dynamic Programming principles, then

- there exists a stationary optimal policy $\pi^\star \in \Pi_{\mathsf{S}}$.
- value and policy iteration algorithms inherit the usual convergence guarantees.

Main Results

Proposition. If ρ is monotone, law-invariant and satisfies the DP principles, then

- $\rho(\widetilde{X} + \widetilde{Y}) = \rho(\widetilde{X}) + \rho(\widetilde{Y})$ for $\widetilde{X} \perp\!\!\!\perp \widetilde{Y}$.
- $\rho(\alpha\widetilde{X}) = \alpha\rho(\widetilde{X})$ for $\alpha \geqslant 0$.

Theorem. A monotone, law-invariant ρ satisfies the DP principles if and **only if**

- $\rho \in \{\operatorname{ess\,inf}, \operatorname{ess\,sup}\}$ (Static kernel case).
- $\rho \in \{\operatorname{ess\,inf}, \operatorname{ess\,sup}, \mathbb{E}\}$ (Resampled kernel case).

Takeaways

- Model uncertainty leads to misspecified policies.
- A full characterization for risk-attitudes compatible with DP.
- Ambiguity-Averse MDPs: Unified framework for epistemic uncertainty.
- Need to go beyond classical DP to use richer risk-attitudes.

arXiv:2602.03381 ICML 2026

axel.benyamine@polytechnique.edu