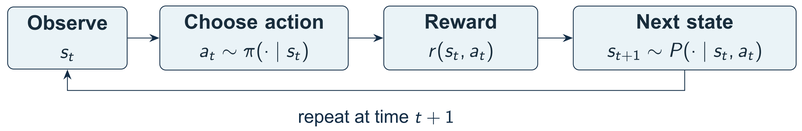
# Robust Average-Reward Markov Decision Processes:
# Minimax-Optimal Learning via Plug-in Reductions

**Yuepeng Yang** (Yale, UPenn)  Yuxin Chen (UPenn)  Yuejie Chi (Yale)

Paper link

## 1 Distributionally robust Markov decision processes

A **Markov decision process** $(\mathcal{S}, \mathcal{A}, P, r)$ consists of a state space $\mathcal{S}$ of size $S$, an action space $\mathcal{A}$ of size $A$, a transition kernel $P$, and a reward function $r$. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ selects actions using the current state:




### Uncertainty set and robust objective

For a nominal kernel $P^0$ and known radii $0 \le \sigma_{s,a} \le \sigma$, consider the total-variation uncertainty set with independent perturbations at each state-action pair:

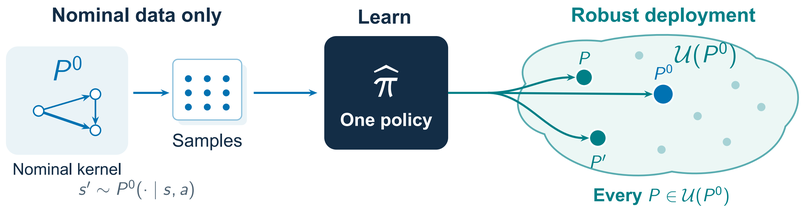
$$\mathcal{U}(P^0) = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left\{ P_{s,a} \in \Delta(\mathcal{S}) : \|P_{s,a} - P^0_{s,a}\|_{\mathrm{TV}} \le \sigma_{s,a} \right\}.$$

We maximize the **worst-case long-run average reward**:

$$\rho^{\pi,\sigma}(s) = \inf_{P\in\mathcal{U}(P^0)} \lim_{T\to\infty} \mathbb{E}^\pi_P \left[ \frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) \,\middle|\, s_0 = s \right].$$

### Generative model and sample complexity

- Assume access to a simulator of the nominal transition kernel $P^0$.
- Collect $N$ samples per state-action pair, for a total of $NSA$ samples.




For $\varepsilon > 0$ and $\delta \in (0,1)$, how can we learn $\widehat{\pi}$ such that, with probability at least $1-\delta$,

$$\rho^{\star,\sigma}(s) - \rho^{\widehat{\pi},\sigma}(s) \le \varepsilon \qquad \text{for every } s \in \mathcal{S}\,?$$

How many samples are needed?

### Difficulty parameters: bias spans

- Starting states share the same average reward, but their early rewards can differ.
- The **nominal** and **robust bias functions**, $h^\pi_{P^0}$ and $h^{\pi,\sigma}$, measure the expected cumulative (robust) reward surplus above the long-run average under policy $\pi$.
- With $\|h\|_{\mathrm{span}} = \max_s h(s) - \min_s h(s)$, the **nominal** and **robust optimal bias spans** are

$$H_0 = \max\left\{1, \|h^\star_{P^0}\|_{\mathrm{span}}\right\}, \qquad H_\sigma = \max\left\{1, \|h^{\star,\sigma}\|_{\mathrm{span}}\right\}.$$

- These spans measure difficulty, analogous to $(1-\gamma)^{-1}$ in discounted MDPs.

## 2 Learning through discounted plug-in reductions

### Discounted values approximate average rewards

- For discount factor $0 < \gamma < 1$, the **robust discounted value** is

$$V^{\pi,\sigma}_\gamma(s) = \inf_{P\in\mathcal{U}(P^0)} \mathbb{E}^\pi_P \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right].$$

- The optimal value $V^{\star,\sigma}_\gamma(s) = \max_\pi V^{\pi,\sigma}_\gamma(s)$ satisfies

$$\left|(1-\gamma)V^{\star,\sigma}_\gamma(s) - \rho^{\star,\sigma}\right| \le (1-\gamma)H_\sigma.$$

As $\gamma$ approaches 1, the effective horizon $1/(1-\gamma)$ increases

**Better approximation**
Smaller approximation error

**Harder discounted problem**
Need more samples; slower iteration

- The returned policy $\widehat{\pi}$ satisfies

$$\rho^{\star,\sigma} - \rho^{\widehat{\pi},\sigma} \le 2(1-\gamma)H_\sigma + (1-\gamma)\left\|V^{\star,\sigma}_\gamma - V^{\widehat{\pi},\sigma}_\gamma\right\|_\infty.$$

approximation + discounted policy error

- **Ideal choice of $\gamma$:** $1-\gamma \asymp \varepsilon/H_\sigma$ gives an approximation error of $O(\varepsilon)$.

### Nominal and robust discounted MDP solvers

- Estimate $\widehat{P}^0$ from observed transition frequencies and initialize $\widehat{V}^\sigma_{\gamma,0}(s) = 0$.
- **Robust value iteration**:

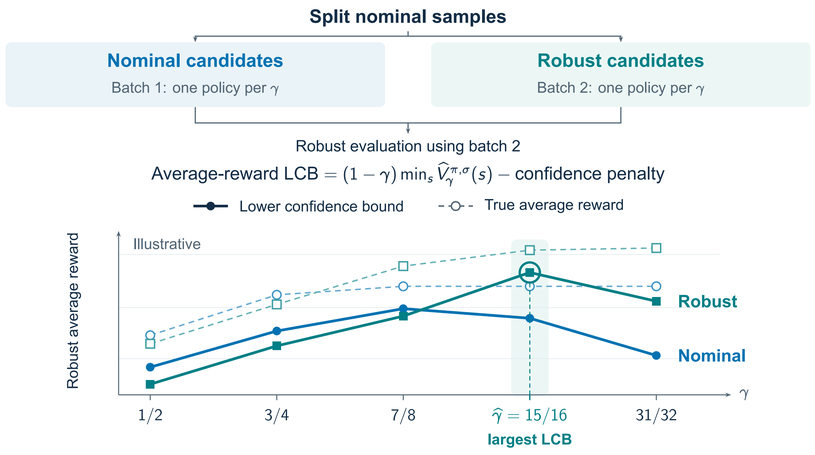
$$\widehat{V}^\sigma_{\gamma,k+1}(s) \leftarrow \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \min_{q\in\mathcal{U}_{s,a}(\widehat{P}^0)} \sum_{s'\in\mathcal{S}} q(s')\, \widehat{V}^\sigma_{\gamma,k}(s') \right\}.$$

immediate reward + $\gamma\times$ robust next-state value

- The nominal variant uses $\widehat{P}^0$ directly in the Bellman update.

### Span-agnostic reduction

When the spans are unknown, try **both reductions** over a finite grid $\gamma = 1 - 2^{-k}$:




**Why both reductions?** The nominal branch is better when $H_0 < H_\sigma$ and $\varepsilon \gtrsim \sigma H_0$.

## 3 Main result: the statistical price of robustness

### Minimax sample complexity rates in two regimes

| Regime | Condition | Total samples |
|---|---|---|
| **High tolerance** | $\varepsilon \gtrsim \sigma H_0$ | $NSA \asymp \dfrac{SA\min\{H_0, H_\sigma\}}{\varepsilon^2}$ |
| **Low tolerance** | $\varepsilon \lesssim \sigma H_0$ | $NSA \asymp \dfrac{SA\left(\min\{H_0, H_\sigma\} + \boldsymbol{\sigma H_\sigma^2}\right)}{\varepsilon^2}$ |

The span-agnostic reductions achieve these rates up to logarithmic factors.

### Why the scale $\sigma H_0$?

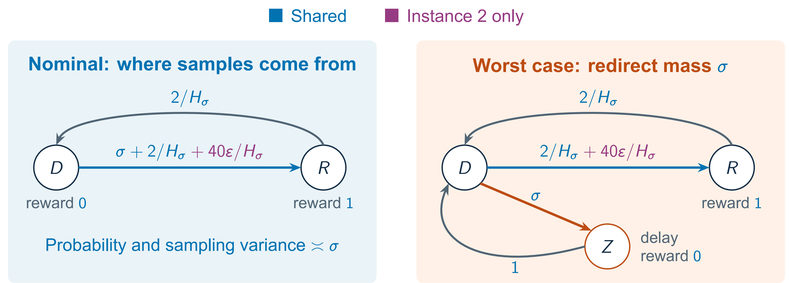
There exists a nominal optimal policy $\pi^\star$ whose robust performance satisfies

$$\rho^{\star,\sigma} - \rho^{\pi^\star,\sigma} \le \sigma H_0.$$

A suitable nominal optimal policy therefore meets the robust target when $\varepsilon \gtrsim \sigma H_0$.

### Analysis overview

- **Span-aware variance analysis. ($H_\sigma^2 \to H_\sigma + \sigma H_\sigma^2$)**
  - ✗ Bound all value-function spans by $(1-\gamma)^{-1}$.
  - ✓ Bound a core set of value spans by $O(H_\sigma)$ plus estimation errors.
- **Using a nominal optimal policy as a reference point. ($H_\sigma \to \min\{H_0, H_\sigma\}$)**
- **Hard instances with amplified robust gaps.**




transition gap $\asymp \varepsilon/H_\sigma$ and sampling variance $\asymp \sigma \implies N \gtrsim \sigma H_\sigma^2/\varepsilon^2$

### Adaptive choice of reduction and discount factor