



POLITECNICO  
MILANO 1863

# OPTIMISTIC POLICY OPTIMIZATION VIA MULTIPLE IMPORTANCE SAMPLING

MATTEO PAPINI, ALBERTO M. METELLI, LORENZO LUPO AND MARCELLO RESTELLI

{matteo.papini, albertomaria.metelli, marcello.restelli}@polimi.it, lorenzo.lupo@mail.polimi.it



## MOTIVATION AND IDEA

### Problem:

- Policy Optimization (PO) methods **neglect exploration**
- Existing exploration strategies are **undirected**
- Lack of **provably efficient** solutions

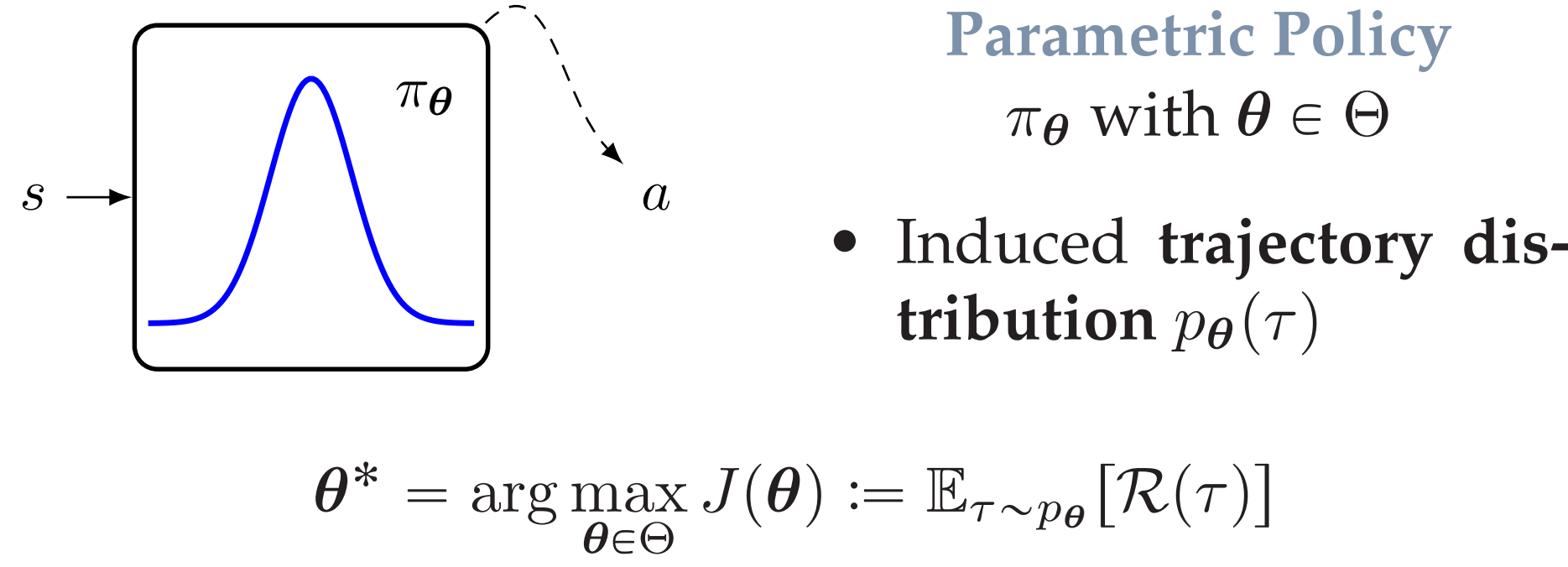
### Idea:

- Frame PO as a continuous **Multi-Armed Bandit (MAB)**
- Use **Multiple Importance Sampling (MIS)** to exploit natural **arm correlation**
- Apply **Optimism in Face of Uncertainty (OFU)**

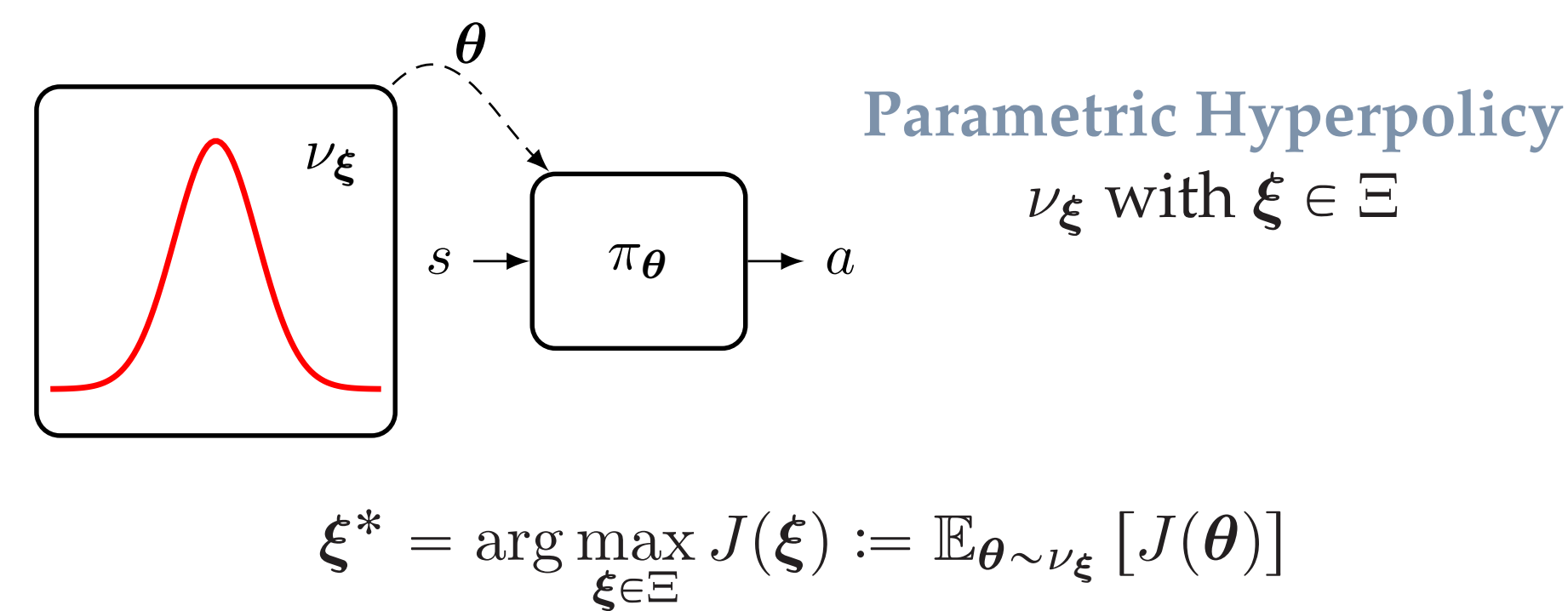
## POLICY OPTIMIZATION

- Continuous** MDP  $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$
- Trajectories**  $\tau = s_0, a_0, r_1, s_1, \dots, r_H \in \mathcal{T}$
- Return  $\mathcal{R}(\tau) = \sum_{h=0}^{H-1} \gamma^h r_{h+1}$

### Action-based PO (Peters and Schaal, 2008)

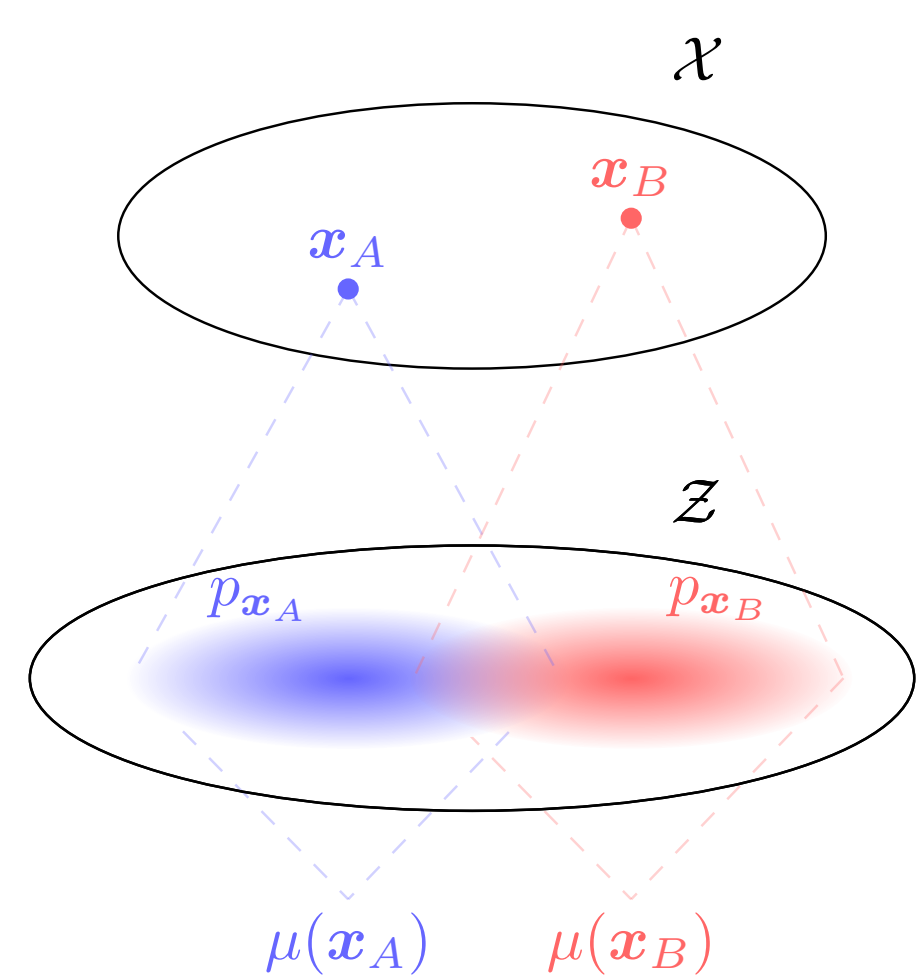


### Parameter-based PO (Sehnke et al., 2008)



## POLICY OPTIMIZATION AS CORRELATED MAB

- (Hyper)parameters as arms  $\Rightarrow$  **continuous** MAB
- Arms **correlate** through common outcome space



## MULTIPLE IMPORTANCE SAMPLING (MIS)

- Samples from several **behavioral** distributions:  
 $z_0 \sim q_0, z_1 \sim q_1, \dots, z_K \sim q_K$
- Estimate  $\mu := \mathbb{E}_{z \sim p} [f(z)]$  under **target** distribution  $p$
- Balance Heuristic (BH)** (Veach and Guibas, 1995):

$$\hat{\mu}_{\text{BH}} := \frac{1}{K} \sum_{k=1}^K \underbrace{\frac{p(z_k)}{\Phi(z_k)}}_{\text{Importance Weight (IW)}} f(z_k), \quad \Phi(z) = \underbrace{\frac{1}{K} \sum_{k=1}^K q_k(z)}_{\text{mixture}}$$

- Unbiased**, but possibly **high-variance**:

$$\mathbb{V}\text{ar} [\hat{\mu}_{\text{BH}}] \leq \|f\|_\infty^2 \frac{d_2(P\|\Phi)}{K} \leq \|f\|_\infty^2 \frac{1}{\sum_{k=1}^K \frac{1}{d_2(p\|q_k)}}$$

$$d_2(p\|q) := \int_{\mathcal{Z}} \left( \frac{p(z)}{q(z)} \right)^2 dz \quad (\text{Rényi divergence})$$

## ROBUST MIS ESTIMATOR

- Importance Sampling estimators are **heavy-tailed** (Metelli et al., 2018)
- This prevents the formation of *exponential* **Upper Confidence Bounds (UCB)**
- Robust estimation via **adaptive truncation** (Bubeck et al., 2013):

$$\check{\mu}_{\text{BH}} := \frac{1}{K} \sum_{k=1}^K \min \left\{ \underbrace{\sqrt{\frac{K d_2(p\|\Phi)}{\log \frac{1}{\delta}}}}_{\text{truncation}}, \underbrace{\frac{p(z_k)}{\Phi(z_k)}}_{\text{IW}} \right\} f(z_k)$$

- Thanks to truncation, with probability at least  $1 - 2\delta$ :

$$|\check{\mu}_{\text{BH}} - \mu| \leq \|f\|_\infty \left( \sqrt{2} + \frac{4}{3} \right) \sqrt{\frac{d_2(p\|\Phi) \log \frac{1}{\delta}}{K}}$$

## OPTIMIST ALGORITHM

A UCB-like algorithm based on the **Optimism in Face of Uncertainty** principle:

- Select *confidence schedule*  $(\delta_t)_{t=0}^T$
- Select initial arm  $x_0$  at random, draw outcome  $z_0 \sim p_{x_0}$  and observe payoff  $f(z_0)$
- For each iteration  $t$  from 1 to  $T$ :

- Define **Upper Confidence Bound**:

$$B_t(x, \delta_t) := \underbrace{\check{\mu}_t(x)}_{\text{Robust MIS Estimator}} + \underbrace{\|f\|_\infty \left( \sqrt{2} + \frac{4}{3} \right) \sqrt{\frac{d_{1+\epsilon}(p_x\|\Phi_t) \log \frac{1}{\delta_t}}{t}}}_{\text{Exploration Bonus}}$$

- Select arm  $x_t = \arg \max_{x \in \mathcal{X}} B_t(x, \delta_t)$ , draw outcome  $z_t \sim p_{x_t}$  and observe payoff  $f(z_t)$

## REGRET ANALYSIS

- Discrete** arm set  $\mathcal{X} = \{x_1, \dots, x_K\}$ 
  - Assumptions: *uniformly* bounded Rényi divergence  $d_2(p_x\|\Phi) \leq v$
  - Confidence schedule:  $\delta_t = 3\delta / (t^2 \pi^2 K)$

$$\text{Regret}(T) \leq \Delta_0 + \left( 4\sqrt{2} + \frac{10}{3} \right) \|f\|_\infty \sqrt{Tv \left( 2 \log T + \log \frac{\pi^2 K}{3\delta} \right)} = \tilde{\mathcal{O}}(\sqrt{T})$$

- Compact** arm space  $\mathcal{X} \subseteq [-D, D]^d$ 
  - Assumptions: *uniformly* bounded Rényi divergence  $d_2(p_x\|\Phi) \leq v$ ,  $L$ -Lipschitz objective  $\mu$
  - Confidence schedule:  $\delta_t = 6\delta / (\pi^2 t^2 (1 + d^d t^{2d}))$

$$\text{Regret}(T) \leq \Delta_0 + \frac{\pi^2 L D}{6} + \left( 4\sqrt{2} + \frac{10}{3} \right) \|f\|_\infty \sqrt{Tv \left( 2(d+1) \log T + d \log d + \log \frac{\pi^2}{3\delta} \right)} = \tilde{\mathcal{O}}(\sqrt{dT})$$

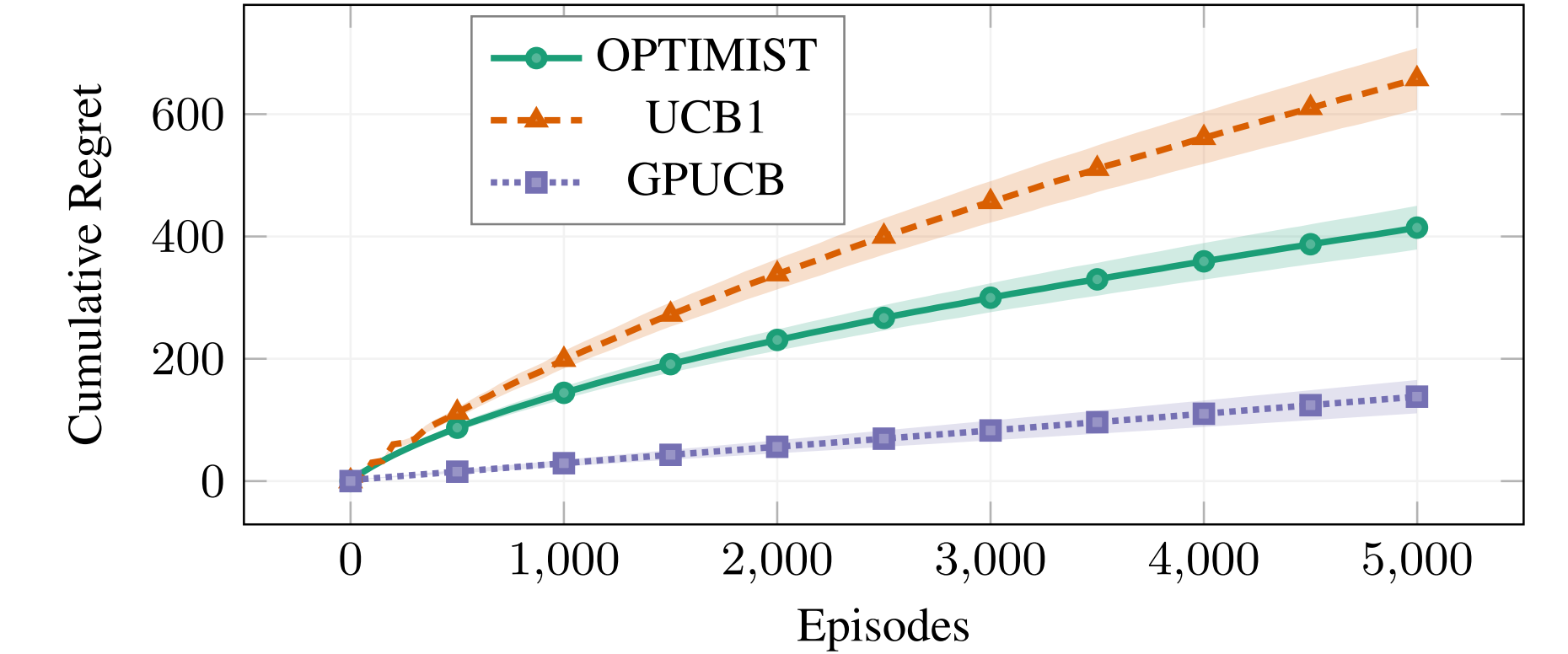
	Arm	Outcome	Induced distribution	Payoff	Objective
<b>Correlated MAB</b>	$x \in \mathcal{X}$	$z \in \mathcal{Z}$	$p_x(z)$	$f(z)$	$\mu(x) = \mathbb{E}_{z \sim p_x} [f(z)]$
<b>PO</b>	$\theta \in \Theta$	$\tau \in \mathcal{T}$	$p_\theta(\tau)$	$\mathcal{R}(\tau)$	$J(\theta)$
<b>PB-PO</b>	$\xi \in \Xi$	$\theta \in \Theta$	$\nu_\xi(\theta)$	$J(\theta)$	$J(\xi)$

## IMPLEMENTATION

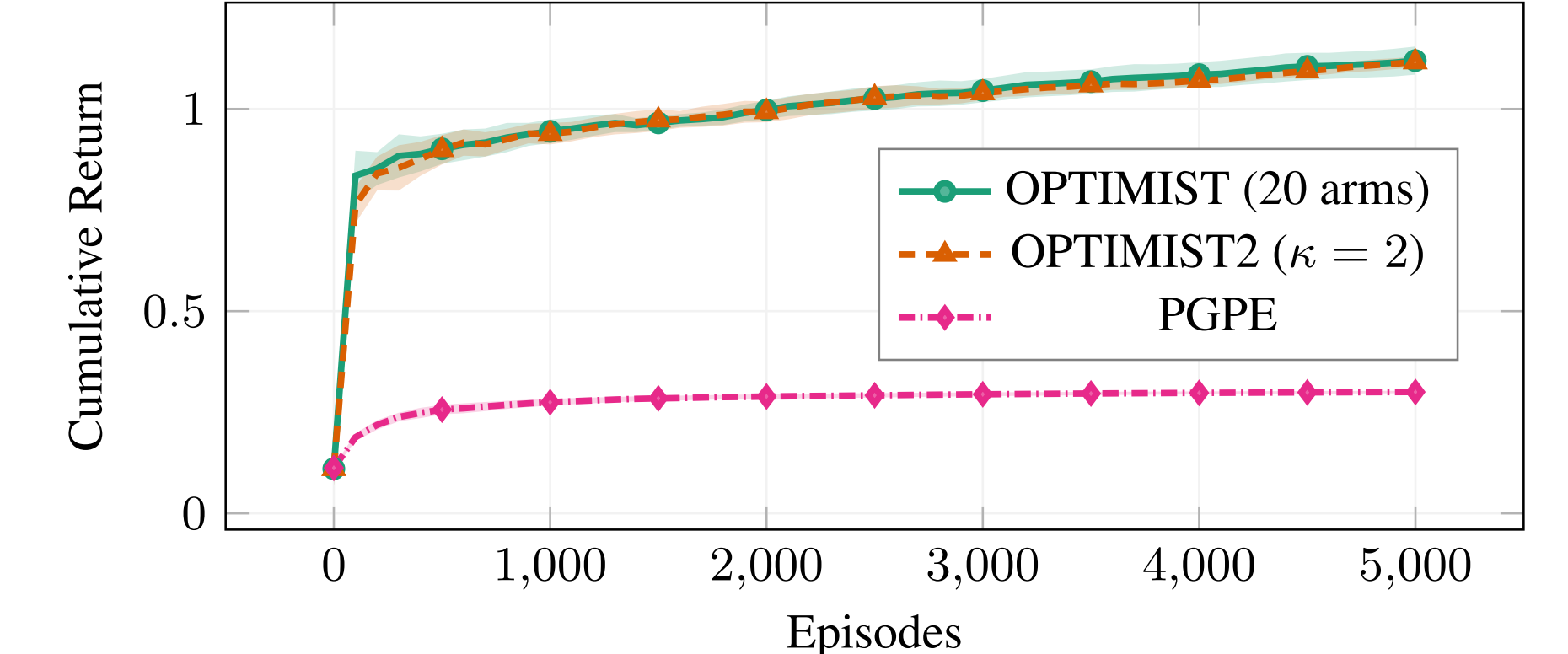
- Trajectory distributions**  $p_\theta$  are difficult to compute  
 $\Rightarrow$  **parameter based exploration**
  - Analytic hyperpolicy  $\nu_\xi$  (e.g., Gaussian)
  - Closed-form Rényi divergence  $d_2$
- Difficult to optimize** the UCB index on a compact space  
 $\Rightarrow$  **adaptive discretization**
  - Use finer and finer grid of  $\left\lceil t^{1/\kappa} \right\rceil^d$  points
  - Confidence schedule:  $\delta_t = 6\delta / (\pi^2 t^2 (1 + t^{d/\kappa}))$
  - Meta-parameter  $\kappa \geq 2$  allows to trade-off regret  $\tilde{\mathcal{O}}(dT^{1-\frac{1}{\kappa}})$  with time  $\mathcal{O}(t^{1+\frac{d}{\kappa}})$  per iteration.
  - $k = 2$  recovers the  $\tilde{\mathcal{O}}(\sqrt{dT})$  regret at the cost of **exponential** time
  - $k = d$  yields **sublinear** regret in **polynomial** time

## EXPERIMENTS

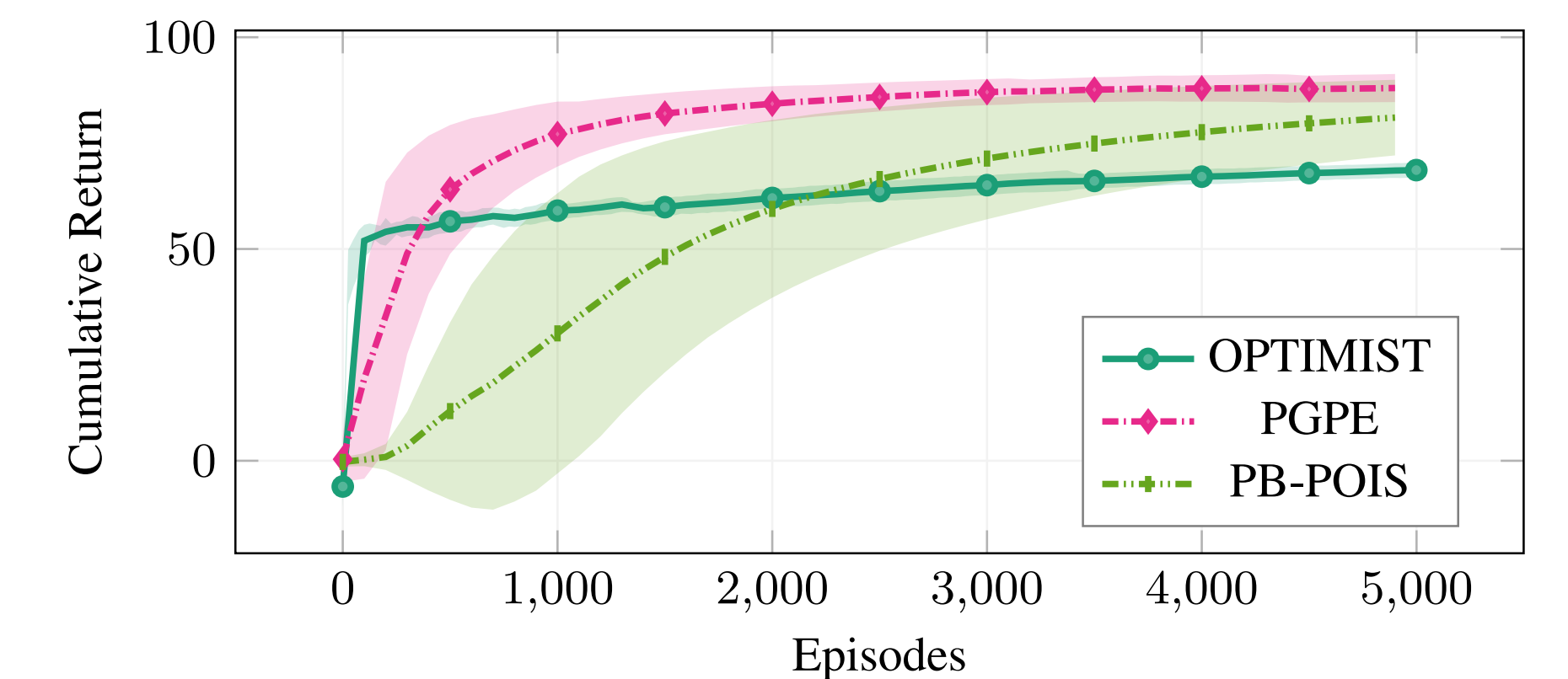
### Linear-Quadratic Regulator (Regret)



### River Swim (Return)



### Mountain Car (Return)



## REFERENCES

- S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.
- A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, pages 5447–5459, 2018.
- J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.
- F. Sehnke, C. Osendorfer, T. Rückstieß, A. Graves, J. Peters, and J. Schmidhuber. Policy gradients with parameter-based exploration for control. In *International Conference on Artificial Neural Networks*, pages 387–396. Springer, 2008.
- E. Veach and L. J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95*, pages 419–428. ACM Press, 1995.