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

POLITECNICO MILANO 1863

Motivation and Idea

Problem

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

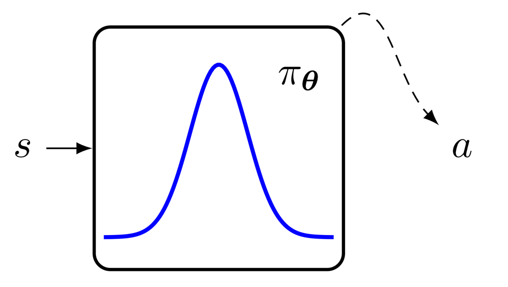
Idea

- 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)

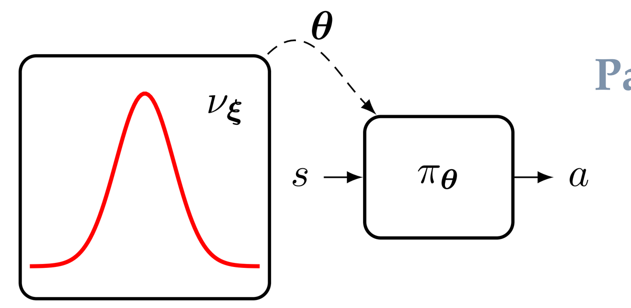


Parametric Policy
$\pi_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \Theta$

- Induced **trajectory distribution** $p_{\boldsymbol{\theta}}(\tau)$

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) := \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}[\mathcal{R}(\tau)]$$

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

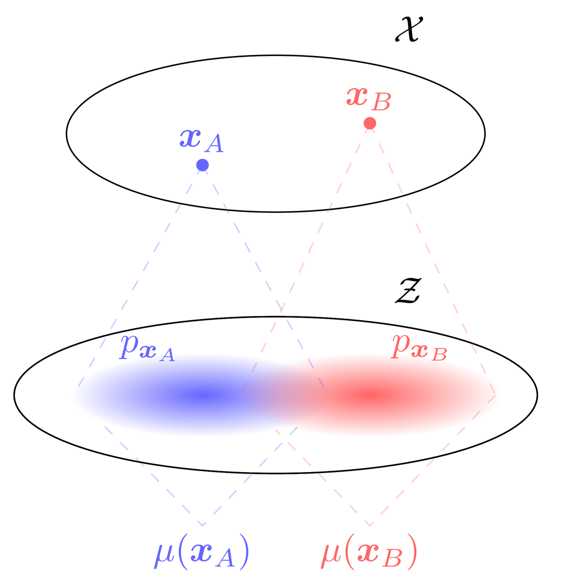


Parametric Hyperpolicy
$\nu_{\boldsymbol{\xi}}$ with $\boldsymbol{\xi} \in \Xi$

$$\boldsymbol{\xi}^* = \arg\max_{\boldsymbol{\xi} \in \Xi} J(\boldsymbol{\xi}) := \mathbb{E}_{\boldsymbol{\theta} \sim \nu_{\boldsymbol{\xi}}}[J(\boldsymbol{\theta})]$$

Policy Optimization as Correlated MAB

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



MAB jargon:

- $\boldsymbol{x}^* \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} \mu(\boldsymbol{x})$
- Gap $\Delta_t = \mu(\boldsymbol{x}^*) - \mu(\boldsymbol{x}_t)$
- $Regret(T) = \sum_{t=0}^{T} \Delta_t$

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):

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

- Unbiased, but possibly high-variance:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\mu}_{\mathrm{BH}}\right] \leqslant \|f\|_\infty^2 \frac{d_2(P\|\Phi)}{K} \leqslant \|f\|_\infty^2 \frac{1}{\sum_{k=1}^{K} \frac{1}{d_2(p\|q_k)}}$$

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

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):

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

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

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

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 $\boldsymbol{x}_0$ at random, draw outcome $z_0 \sim p_{\boldsymbol{x}_0}$ and observe payoff $f(z_0)$
- For each iteration t from 1 to T:
  - Define Upper Confidence Bound:

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

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

Regret Analysis

- Discrete arm set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_K\}$
  - Assumptions: *uniformly* bounded Rényi divergence $d_2(p_{\boldsymbol{x}}\|\Phi) \leqslant v$
  - Confidence schedule: $\delta_t = 3\delta/(t^2\pi^2 K)$

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

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

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

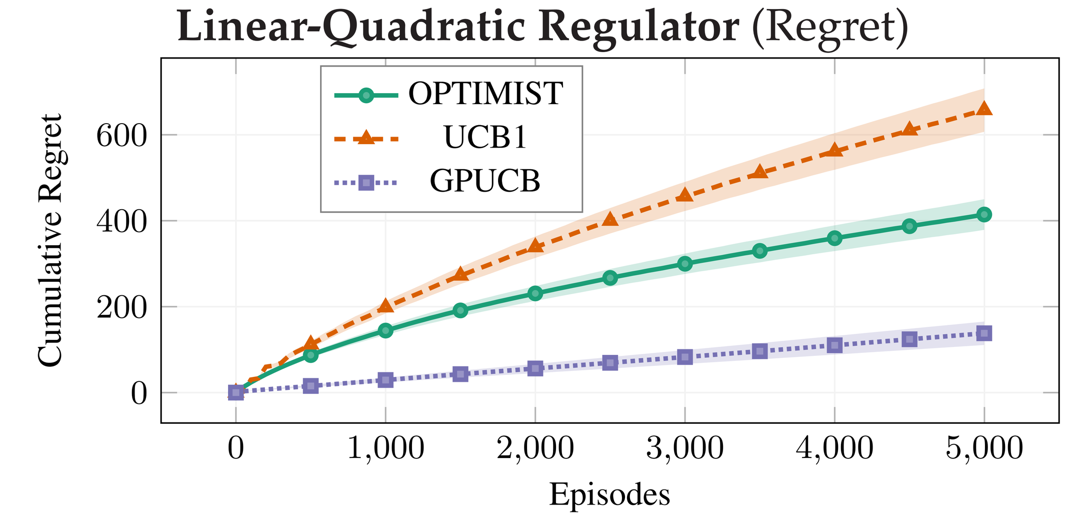
	Arm	Outcome	Induced distribution	Payoff	Objective
Correlated MAB	$\boldsymbol{x} \in \mathcal{X}$	$z \in \mathcal{Z}$	$p_{\boldsymbol{x}}(z)$	$f(z)$	$\mu(\boldsymbol{x}) = \mathop{E}_{z \sim p_{\boldsymbol{x}}}[f(z)]$
PO	$\boldsymbol{\theta} \in \Theta$	$\tau \in \mathcal{T}$	$p_{\boldsymbol{\theta}}(\tau)$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

Implementation

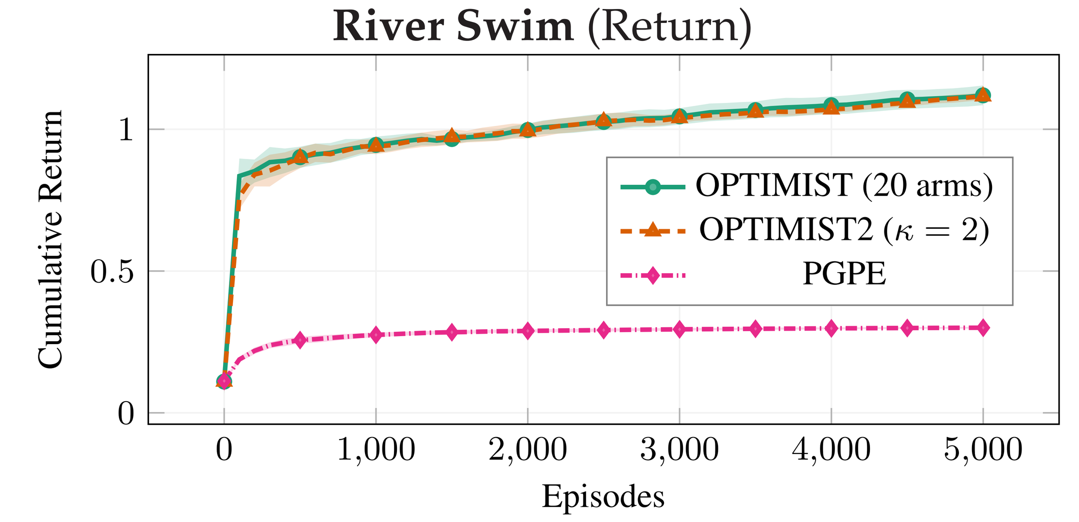
- Trajectory distributions $p_{\boldsymbol{\theta}}$ are difficult to compute $\implies$ parameter based exploration
  - Analytic hyperpolicy $\nu_{\boldsymbol{\xi}}$ (e.g., Gaussian)
  - Closed-form Rényi divergence d_2
- Difficult to optimize the UCB index on a compact space $\implies$ adaptive discretization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

Experiments

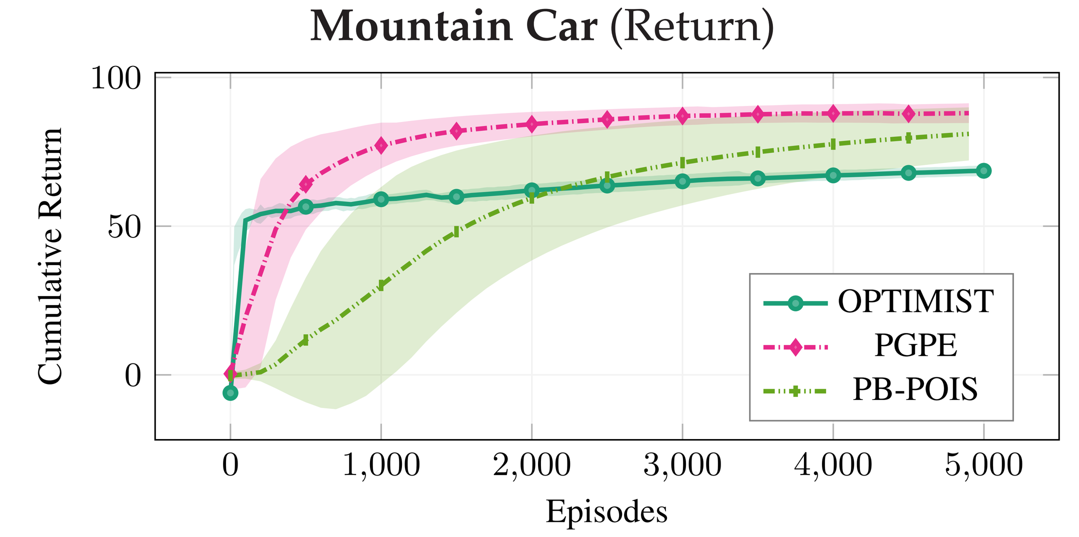
Linear-Quadratic Regulator (Regret)



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.