Policy Optimization as Online Learning with Mediator Feedback

Alberto Maria Metelli*, Matteo Papini*, Pierluca D'Oro, and Marcello Restelli
{albertomaria.metelli, matteo.papini, marcello.restelli}@polimi.it, pierluca.doro@mail.polimi.it
POLITECNICO MILANO 1863

Motivation and Idea

Problem: How to deal with exploration in Policy Optimization (PO)?
Idea: exploit the inherent structure of the PO problem via multiple importance sampling

Policy Optimization

- Parameter space $\Theta \subseteq \mathbb{R}^d$
- A parametric policy for each $\boldsymbol{\theta} \in \Theta$
- Each inducing a distribution $p_{\boldsymbol{\theta}}$ over trajectories
- A return $\mathcal{R}(\tau)$ for every trajectory τ
- Goal: maximize the expected return (Deisenroth et al., 2013)

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

Policy Optimization as Online Learning

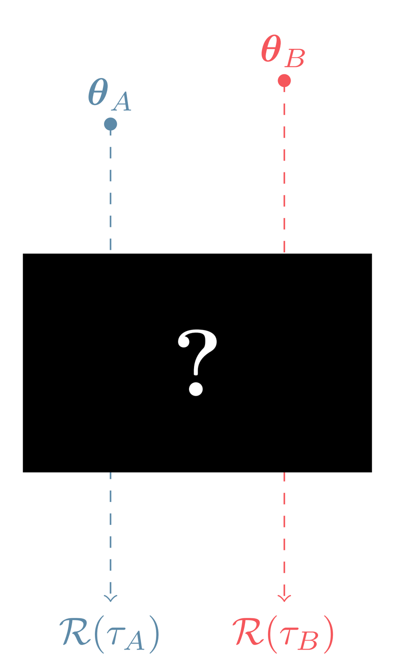
For $t = 1, 2, \dots$

- Select parameter $\boldsymbol{\theta}_t$ and run $\pi_{\boldsymbol{\theta}_t}$
- Observe the trajectory τ_t and the return $\mathcal{R}(\tau_t)$

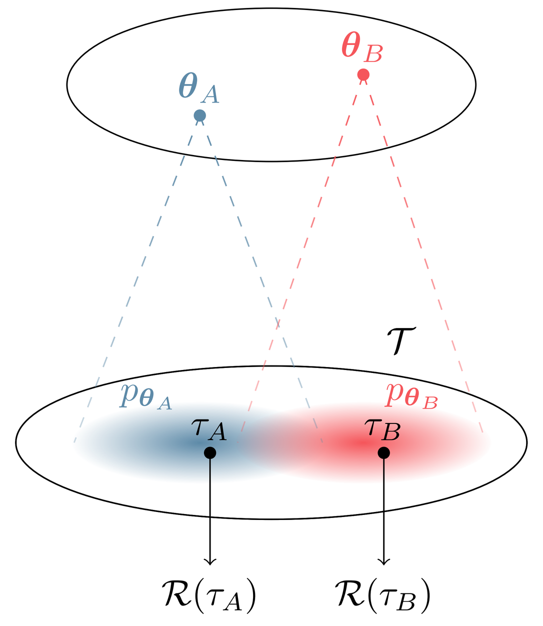
Goal: minimize the regret (Auer et al., 2002)

$$\text{Regret}(n) = \sum_{t=1}^{n} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t) = \sum_{t=1}^{n} \Delta(\boldsymbol{\theta}_t)$$

Bandit Feedback



Mediator Feedback



Regret Lower Bounds

- Two-parameter space $\Theta = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_B\}$
- Performance gap $\Delta = J(\boldsymbol{\theta}_A) - J(\boldsymbol{\theta}_B)$
- If $D_{KL}(p_{\boldsymbol{\theta}_A}\|p_{\boldsymbol{\theta}_B}) < \infty$ and $D_{KL}(p_{\boldsymbol{\theta}_B}\|p_{\boldsymbol{\theta}_A}) < \infty \implies$ constant regret

$$\mathbb{E}\,\text{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\right)$$

- If $D_{KL}(p_{\boldsymbol{\theta}_A}\|p_{\boldsymbol{\theta}_B}) = \infty$ or $D_{KL}(p_{\boldsymbol{\theta}_B}\|p_{\boldsymbol{\theta}_A}) = \infty \implies$ logarithmic regret

$$\mathbb{E}\,\text{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\log\left(\Delta^2 n\right)\right)$$

Importance Sampling for Mediator Feedback

- Idea: use **all** the samples to estimate the expected return of **any** policy $\mathcal{H}_t = \{(\boldsymbol{\theta}_i, \tau_i, \mathcal{R}(\tau_i))\}_{i\in[t-1]}$
- Mixture distribution Φ_t of the behavioral policies played and relative multiple importance weight with *balance heuristic* (Veach and Guibas, 1995) w.r.t. the target policy $p_{\boldsymbol{\theta}}(\tau_i)$:

$$\Phi_t = \frac{1}{t-1}\sum_{j=1}^{t-1} p_{\boldsymbol{\theta}_j}(\tau_i) \implies \frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)} = \frac{p_{\boldsymbol{\theta}}(\tau_i)}{\frac{1}{t-1}\sum_{j=1}^{t-1} p_{\boldsymbol{\theta}_j}(\tau_i)} = \frac{\prod_{h=0}^{H-1}\pi_{\boldsymbol{\theta}}(a_{ih}|s_{ih})}{\frac{1}{t-1}\sum_{j=1}^{t-1}\prod_{h=0}^{H-1}\pi_{\boldsymbol{\theta}_j}(a_{ih}|s_{ih})}$$

- Vanilla importance weight leads to heavy-tailed estimator (Metelli et al., 2018) $\implies$ employ a *time-variant* weight truncation threshold $M_t(\boldsymbol{\theta})$ (Ionides, 2008)

$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1}\sum_{i=1}^{t-1}\min\left\{\frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}, M_t(\boldsymbol{\theta})\right\}\mathcal{R}(\tau_i)$$

Truncation threshold

$$M_t(\boldsymbol{\theta}) = \sqrt{\frac{(t-1)\,d_2(p_{\boldsymbol{\theta}}\|\Phi_t)}{\log\frac{1}{\delta}}}$$

Renyi divergence

$$d_2(p_{\boldsymbol{\theta}}\|\Phi_t) = \int \frac{p_{\boldsymbol{\theta}}(\tau_i)^2}{\Phi_t(\tau_i)}\mathrm{d}\tau$$

- We obtain exponential concentration (Papini et al., 2019; Metelli et al., 2020):

$$\check{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant 2.75\sqrt{\frac{\log\frac{1}{\delta}}{\eta_t(\boldsymbol{\theta})}}$$

Effective sample size

$$\eta_t(\boldsymbol{\theta}) = \frac{t-1}{d_2(p_{\boldsymbol{\theta}}\|\Phi_t)}$$

Algorithms

Execute $\pi_{\boldsymbol{\theta}_1}$, observe $\tau_1 \sim p_{\boldsymbol{\theta}_1}$ and $\mathcal{R}(\tau_1)$
for $t = 2, \dots, n$ **do**
  Compute expected return estimate $\check{J}_t(\boldsymbol{\theta})$
  Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta}\in\Theta} B_t(\boldsymbol{\theta})$
  Execute $\pi_{\boldsymbol{\theta}_t}$, observe $\tau_t \sim p_{\boldsymbol{\theta}_t}$ and $\mathcal{R}(\tau_t)$
end for

OPTIMIST (Papini et al., 2019)
Compute upper confidence bound:

$$B_t(\boldsymbol{\theta}) = \check{J}_t(\boldsymbol{\theta}) + 2.42\sqrt{\frac{\alpha\log t}{\eta_t(\boldsymbol{\theta})}}$$

RANDOMIST (new!)
Generate perturbation:

$$U_t(\boldsymbol{\theta}) = \frac{1}{\eta_t(\boldsymbol{\theta})}\sum_{l=1}^{a\eta_t(\boldsymbol{\theta})}\tau_l + b, \text{ with } \tau_l \sim \text{Ber}(1/2)$$

Compute index $B_t(\boldsymbol{\theta}) = \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

Regret Upper Bounds Comparison

Finite Policy Space

$$v = \max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta} d_2(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\theta}'}) \text{ and } \Delta = \min_{\boldsymbol{\theta}\neq\boldsymbol{\theta}^*} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta})$$

Algorithm	Exploration	$\mathbb{E}\,\text{Regret}(n)$ $v=\infty$	$\mathbb{E}\,\text{Regret}(n)$ $v<\infty$
Greedy	-	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{v}{\Delta}\log\frac{v}{\Delta^2}\right)$
UCB1	deterministic	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$
OPTIMIST	deterministic	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta}\log\frac{v}{\Delta^2}\right)$
RANDOMIST	randomized	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta}\log\frac{v}{\Delta^2}\right)$

Compact Policy Space

$$\Theta = [-D, D]^d \text{ and } v = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta} d_2(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\theta}'})$$

Algorithm	Complexity	$\mathbb{E}\,\text{Regret}(n)$
OPTIMIST	$t^{1+\frac{d}{2}}$	$\mathcal{O}\left(\sqrt{vdn}\right)$
RANDOMIST	dt^2	?

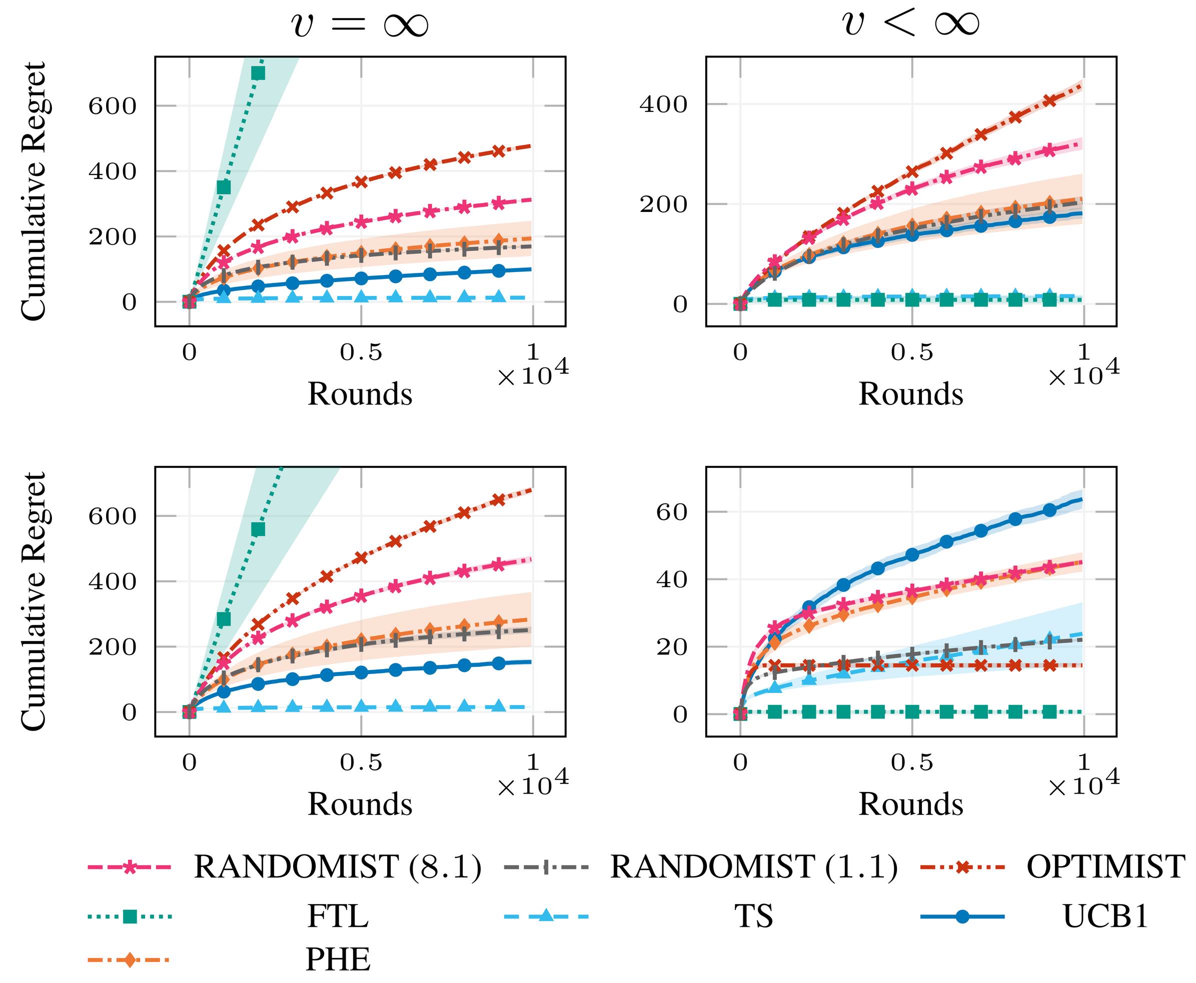
RANDOMIST replaces the discretization of OPTIMIST with MCMC sampling:

$$\boldsymbol{\theta}_t \sim \Pr\left(\check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}'\in\Theta}\check{J}_t(\boldsymbol{\theta}') + U_t(\boldsymbol{\theta}')\right)$$

Experiments

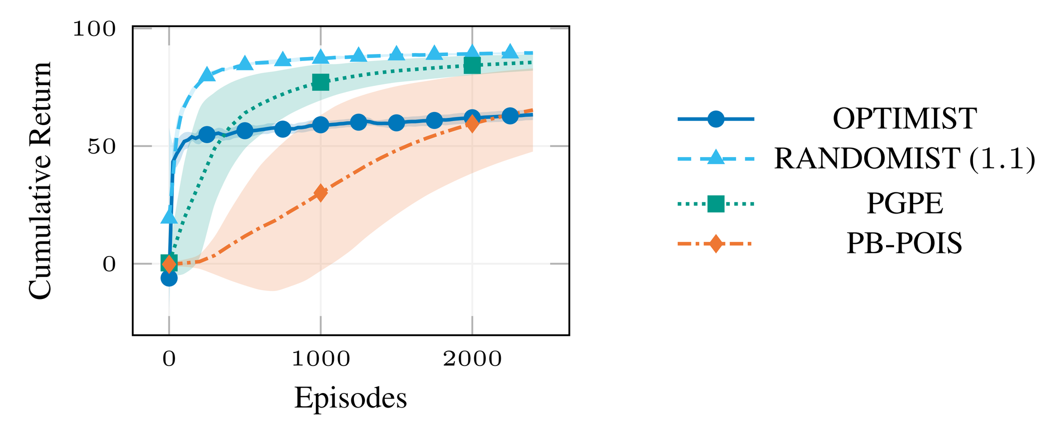
Finite Policy Spaces

Illustrative Examples (Regret)

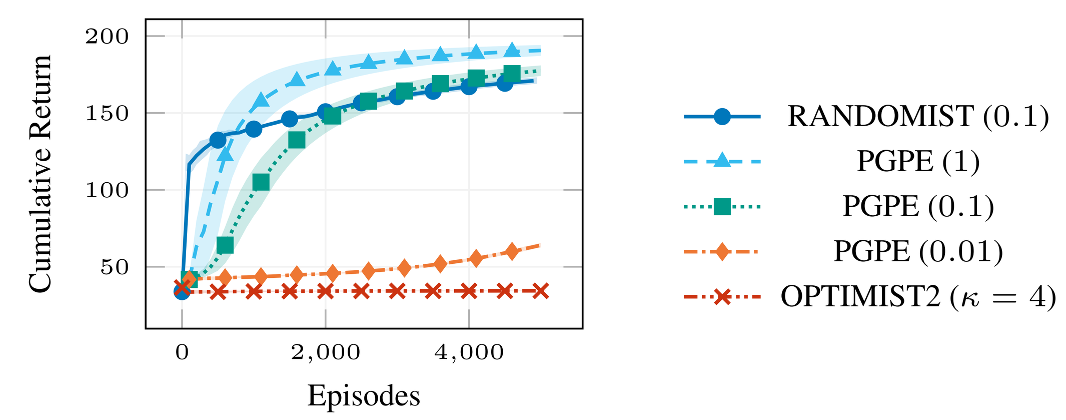


Compact Policy Spaces

Mountain Car



Cartpole



References

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.
M. P. Deisenroth, G. Neumann, and J. Peters. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2013.
E. L. Ionides. Truncated importance sampling. *JCGS*, 17(2):295–311, 2008.
A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *NeurIPS*, 2018.
A. M. Metelli, M. Papini, N. Montali, and M. Restelli. Importance sampling techniques for policy optimization. *JMLR*, 21(141):1–75, 2020.
M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In *ICML*, 2019.
E. Veach and L. J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In S. G. Mair and R. Cook, editors, *SIGGRAPH*, pages 419–428. ACM, 1995.