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 **Multi-Armed Bandit (MAB)** over parameter space *with arm correlation*
- Apply **Optimism in Face of Uncertainty (OFU)**

Policy Optimization

Vanilla action-based PO (Peters and Schaal, 2008)

- Continuous MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$
- **Trajectory** $\tau = s_0, a_0, r_1, s_1, \dots r_H \quad \in \quad \mathcal{T}$
- Return $\mathcal{R}(\tau) = \sum_{h=0}^{H-1} \gamma^h r_{h+1}$
- Parametric policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ with $\boldsymbol{\theta} \in \Theta$
- Induced **trajectory distribution** $p_{\boldsymbol{\theta}}(\tau)$
- **Performance** $J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}[\mathcal{R}(\tau)]$
- Find $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta})$

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

- Hyperpolicy $\nu_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ with $\boldsymbol{\xi} \in \Xi$ (e.g., Gaussian)
- Find $\boldsymbol{\xi}^* = \arg\max_{\boldsymbol{\xi} \in \Xi} \mathbb{E}_{\boldsymbol{\theta} \sim \nu_{\boldsymbol{\xi}}}[J(\boldsymbol{\theta})]$

Policy Optimization as Correlated MAB

	Correlated MAB	PO	PB-PO
Arm	$\boldsymbol{x} \in \mathcal{X}$	$\boldsymbol{\theta} \in \Theta$	$\boldsymbol{\xi} \in \Xi$
Outcome	$z \in \mathcal{Z}$	$\tau \in \mathcal{T}$	$\boldsymbol{\theta} \in \Theta$
Induced distribution	$p_{\boldsymbol{x}}(z)$	$p_{\boldsymbol{\theta}}(\tau)$	$\nu_{\boldsymbol{\xi}}(\boldsymbol{\theta})$
Payoff	$f(z)$	$\mathcal{R}(\tau)$	$J(\boldsymbol{\theta})$

Multiple Importance Sampling

- 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):

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

- Unbiased, but possibly high-variance:

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

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

Robust 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., 2012):

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

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

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

OPTIMIST Algorithm

Regret Analysis

Implementation

Experiments

References

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

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.