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 **Multi-Armed Bandit (MAB)** over parameter space *with a lot of structure*

Policy Optimization

Vanilla:

- Continuous MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$
- **Trajectory** $\tau = s_0, a_0, r_1, s_1, \dots r_H$
- 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 exploration (Sehnke et al., 2008):

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

Multiple Importance Sampling

- Samples from many **behavioral distributions**:
  $z_0 \sim q_0, z_1 \sim q_1, \dots, z_K \sim q_K$
- **Target distribution** p
- Estimate $\mu = \mathbb{E}_{z \sim p}[f(z)]$ from available samples
- **Mixture** of behaviorals: $\Phi(z) = \frac{1}{K} \sum_{k=1}^{K} q_k(z)$
- Balance Heuristic (BH) (Veach and Guibas, 1995):

$$\widehat{\mu}_{\mathrm{BH}} := \frac{1}{K} \sum_{k=1}^{K} \frac{p(z_k)}{\Phi(z_k)} f(z_k) \qquad \text{(unbiased)}$$

- Possibly high variance:

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

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

Robust Multiple Importance Sampling Estimation

Algorithm

Regret Analysis

Implementation

Experiments

References

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.