Optimistic Policy Optimization via Multiple Importance Sampling

Matteo Papini Alberto Maria Metelli
Lorenzo Lupo Marcello Restelli

11th June 2019
Thirty-sixth International Conference on Machine Learning, Long Beach, CA, USA

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** $R(\tau)$ for every trajectory τ

$$\textbf{Goal:}\ \max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}} \left[R(\tau) \right]$$

- On-line, iterative optimization

Exploration in Policy Optimization

- **Exploration-exploitation** trade-off
- **Problem:** the underlying Markov process is often **continuous**
- **Undirected** exploration: entropy bonus [3]
- **Directed** exploration: pseudo-counts [1]

Lack of theoretical guarantees

Policy Optimization as a MAB

- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB** [4]: we *need* structure

Arm correlation [6] through trajectory distributions

OPTIMIST

- A **UCB-like** index [5]:

$$B_t(\boldsymbol{\theta}) = \underbrace{\check{\mu}_t(\boldsymbol{\theta})}_{\substack{\textbf{ESTIMATE:} \\ \text{a truncated multiple} \\ \text{importance sampling estimator [7, 2]}}} + \underbrace{R_{\max}\left(\sqrt{2} + \frac{4}{3}\right)\sqrt{\frac{d_2(p_{\boldsymbol{\theta}} \| \Phi_t) \log \frac{1}{\delta_t}}{t}}}_{\substack{\textbf{EXPLORATION BONUS:} \\ \text{distributional distance} \\ \text{from previous arms}}}$$

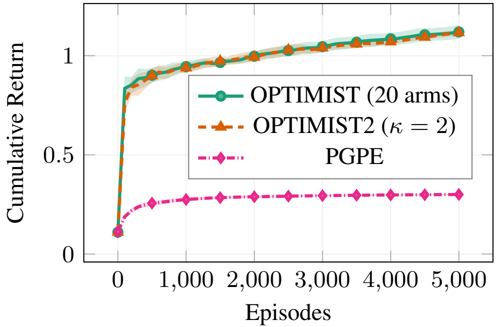
- Select $\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta} \in \Theta} B_t(\boldsymbol{\theta})$

Sublinear Regret

- Recall $Regret(T) = \sum_{t=0}^{T} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t)$
- Consider a **compact**, d-dimensional parameter space Θ
- Under **mild assumptions** on the policy class, with high probability:

$$Regret(T) = \widetilde{\mathcal{O}}\left(\sqrt{dT}\right)$$

Empirical Results



Caveats

- Easy implementation only for parameter-based exploration
- Difficult optimization $\Longrightarrow$ discretization
- ...

Thank You for Your Attention!

- Poster **#103**
- Code: `github.com/WolfLo/optimist`
- Contact: matteo.papini@polimi.it
- Web page: `t3p.github.io/icml19`

References

[1] Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479.

[2] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

[3] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865.

[4] Kleinberg, R., Slivkins, A., and Upfal, E. (2013). Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*.

[5] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

[6] Pandey, S., Chakrabarti, D., and Agarwal, D. (2007). Multi-armed bandit problems with dependent arms. In *Proceedings of the 24th international conference on Machine learning*, pages 721–728. ACM.

[7] Veach, E. and Guibas, L. J. (1995). 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.