



**POLITECNICO**  
MILANO 1863

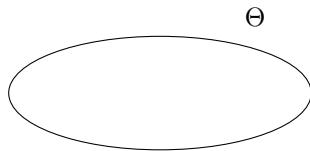
# Optimistic Policy Optimization via Multiple Importance Sampling

**Matteo Papini**    Alberto Maria Metelli  
Lorenzo Lupo    Marcello Restelli

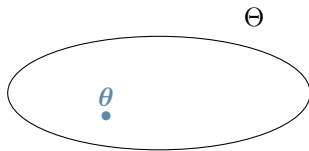
11th June 2019

Thirty-sixth International Conference on Machine Learning, Long Beach, CA, USA

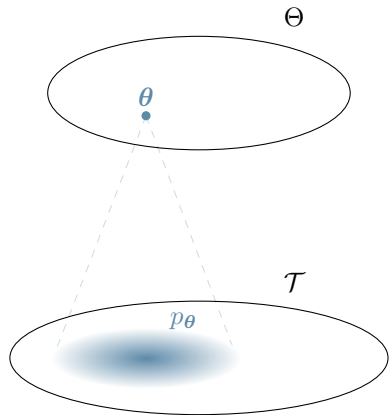
- **Parameter space**  $\Theta \subseteq \mathbb{R}^d$
- A parametric **policy** for each  $\theta \in \Theta$
- Each inducing a distribution  $p_\theta$  over **trajectories**
- A **return**  $R(\tau)$  for every trajectory  $\tau$
- **Goal:**  $\max_{\theta \in \Theta} J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)]$
- Iterative optimization (e.g., gradient ascent)



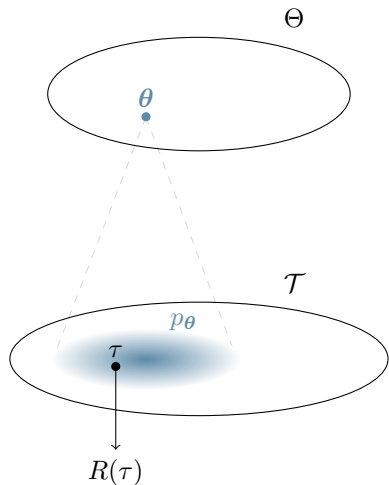
- **Parameter space**  $\Theta \subseteq \mathbb{R}^d$
- A parametric **policy** for each  $\theta \in \Theta$
- Each inducing a distribution  $p_\theta$  over **trajectories**
- A **return**  $R(\tau)$  for every trajectory  $\tau$
- **Goal:**  $\max_{\theta \in \Theta} J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)]$
- Iterative optimization (e.g., gradient ascent)



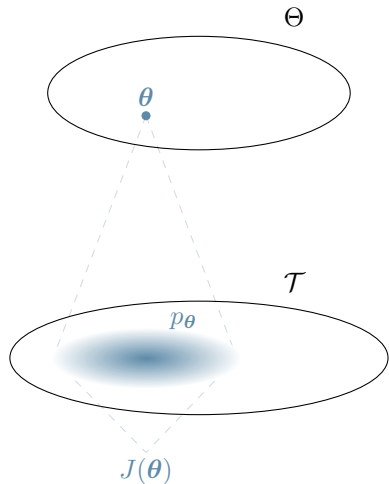
- **Parameter space**  $\Theta \subseteq \mathbb{R}^d$
- A parametric **policy** for each  $\theta \in \Theta$
- Each inducing a distribution  $p_\theta$  over **trajectories**
- A **return**  $R(\tau)$  for every trajectory  $\tau$
- **Goal:**  $\max_{\theta \in \Theta} J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)]$
- Iterative optimization (e.g., gradient ascent)



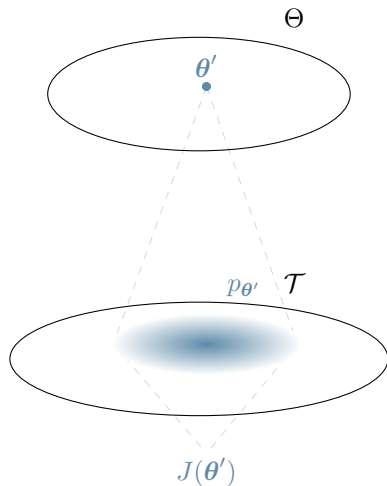
- **Parameter space**  $\Theta \subseteq \mathbb{R}^d$
- A parametric **policy** for each  $\theta \in \Theta$
- Each inducing a distribution  $p_\theta$  over **trajectories**
- A **return**  $R(\tau)$  for every trajectory  $\tau$
- **Goal:**  $\max_{\theta \in \Theta} J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)]$
- Iterative optimization (e.g., gradient ascent)



- **Parameter space**  $\Theta \subseteq \mathbb{R}^d$
- A parametric **policy** for each  $\theta \in \Theta$
- Each inducing a distribution  $p_\theta$  over **trajectories**
- A **return**  $R(\tau)$  for every trajectory  $\tau$
- **Goal:**  $\max_{\theta \in \Theta} J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)]$
- Iterative optimization (e.g., gradient ascent)



- **Parameter space**  $\Theta \subseteq \mathbb{R}^d$
- A parametric **policy** for each  $\theta \in \Theta$
- Each inducing a distribution  $p_\theta$  over **trajectories**
- A **return**  $R(\tau)$  for every trajectory  $\tau$
- **Goal:**  $\max_{\theta \in \Theta} J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)]$
- Iterative optimization (e.g., gradient ascent)



- **Continuous** decision process  $\implies$  difficult
- Policy gradient methods tend to be **greedy** (e.g., TRPO [6], PGPE [7])
- Mainly **undirected** (e.g., entropy bonus [2])
- **Lack of theoretical guarantees**



- **Continuous** decision process  $\implies$  difficult
- Policy gradient methods tend to be **greedy** (e.g., TRPO [6], PGPE [7])
- Mainly **undirected** (e.g., entropy bonus [2])
- **Lack of theoretical guarantees**

- **Continuous** decision process  $\implies$  difficult
- Policy gradient methods tend to be **greedy** (e.g., TRPO [6], PGPE [7])
- Mainly **undirected** (e.g., entropy bonus [2])
- **Lack of theoretical guarantees**

- **Continuous** decision process  $\implies$  difficult
- Policy gradient methods tend to be **greedy** (e.g., TRPO [6], PGPE [7])
- Mainly **undirected** (e.g., entropy bonus [2])
- **Lack of theoretical guarantees**

- **Continuous** decision process  $\implies$  difficult
- Policy gradient methods tend to be **greedy** (e.g., TRPO [6], PGPE [7])
- Mainly **undirected** (e.g., entropy bonus [2])
- **Lack of theoretical guarantees**

**If only this were a Multi-Armed Bandit...**

- **Continuous** decision process  $\implies$  difficult
- Policy gradient methods tend to be **greedy** (e.g., TRPO [6], PGPE [7])
- Mainly **undirected** (e.g., entropy bonus [2])
- **Lack of theoretical guarantees**

If only this were a **Correlated Multi-Armed Bandit...**

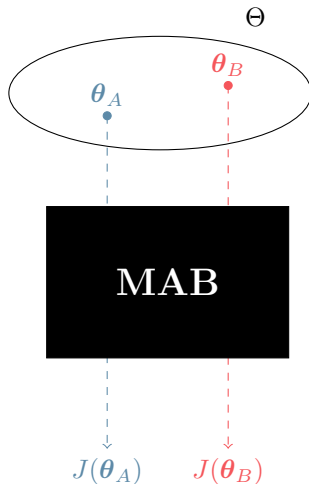
- **Arms:** parameters  $\theta$
- **Payoff:** expected return  $J(\theta)$
- **Continuous MAB** [3]: we *need* structure
- **Arm correlation** [5] through trajectory distributions
- **Importance Sampling (IS)**



- **Arms:** parameters  $\theta$
- **Payoff:** expected return  $J(\theta)$
- **Continuous MAB** [3]
- **Arm correlation** [5] through trajectory distributions
- **Importance Sampling (IS)**

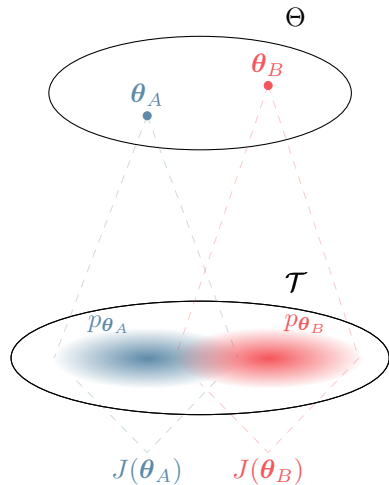


- **Arms:** parameters  $\theta$
- **Payoff:** expected return  $J(\theta)$
- **Continuous MAB** [3]
- **Arm correlation** [5] through trajectory distributions
- **Importance Sampling (IS)**

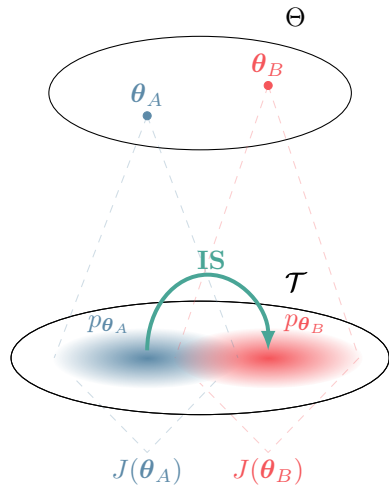




- **Arms:** parameters  $\theta$
- **Payoff:** expected return  $J(\theta)$
- **Continuous MAB** [3]
- **Arm correlation** [5] through trajectory distributions
- Importance Sampling (IS)



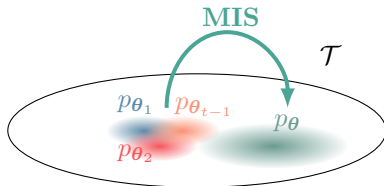
- **Arms:** parameters  $\theta$
- **Payoff:** expected return  $J(\theta)$
- **Continuous MAB** [3]
- **Arm correlation** [5] through trajectory distributions
- **Importance Sampling (IS)**



- A **UCB-like** index [4]:

$$B_t(\boldsymbol{\theta}) = \underbrace{\check{J}_t(\boldsymbol{\theta})}_{\text{ESTIMATE}}$$

a **truncated multiple**  
importance sampling estimator [8, 1]

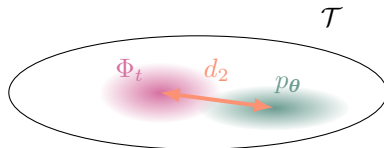


- A **UCB-like** index [4]:

$$B_t(\theta) = \underbrace{\check{J}_t(\theta)}_{\text{ESTIMATE}} + \underbrace{C \sqrt{\frac{d_2(p_\theta \| \Phi_t) \log \frac{1}{\delta_t}}{t}}}_{\text{EXPLORATION BONUS:}}$$

a **truncated multiple** importance sampling estimator [8, 1]

**distributional** distance from previous solutions



- A **UCB-like** index [4]:

$$B_t(\boldsymbol{\theta}) = \underbrace{\check{J}_t(\boldsymbol{\theta})}_{\text{ESTIMATE}} + \underbrace{C \sqrt{\frac{d_2(p_{\boldsymbol{\theta}} \parallel \Phi_t) \log \frac{1}{\delta_t}}{t}}}_{\text{EXPLORATION BONUS:}}$$

a **truncated multiple**  
importance sampling estimator [8, 1]

**distributional** distance  
from previous solutions

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

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

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

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

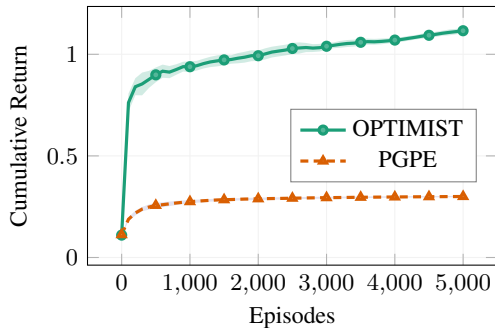
$$Regret(T) = \tilde{\mathcal{O}} \left( \sqrt{dT} \right)$$

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

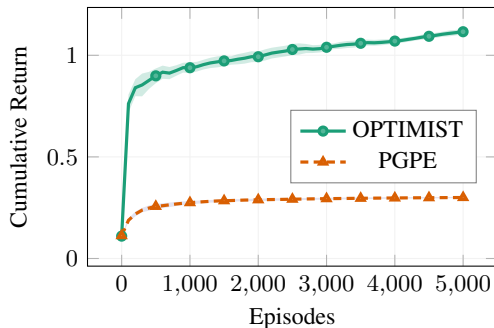
$$Regret(T) = \tilde{\mathcal{O}} \left( \sqrt{dT} \right)$$



## River Swim



## River Swim



## Caveats

- Easy implementation only for parameter-based exploration [7]
- Difficult optimization  $\implies$  discretization
- ...

# Thank You for Your Attention!

Poster **#103**

Code: [github.com/WolfLo/optimist](https://github.com/WolfLo/optimist)

Contact: [matteo.papini@polimi.it](mailto:matteo.papini@polimi.it)

Web page: [t3p.github.io/icml19](https://t3p.github.io/icml19)



- [1] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.
- [2] 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.
- [3] Kleinberg, R., Slivkins, A., and Upfal, E. (2013). Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*.
- [4] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.
- [5] 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.
- [6] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897.
- [7] Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J. (2008). Policy gradients with parameter-based exploration for control. In *International Conference on Artificial Neural Networks*, pages 387–396. Springer.
- [8] 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.