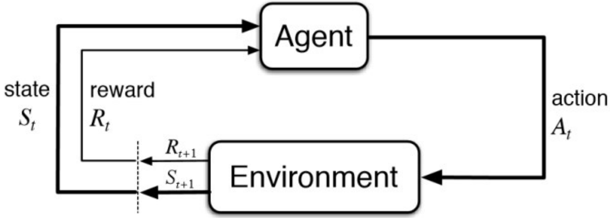
Optimistic Policy Optimization via Multiple Importance Sampling

Matteo Papini Alberto Maria Metelli
Lorenzo Lupo Marcello Restelli

19-20th September 2019
Markets, Algorithms, Prediction and Learning Workshop, Politecnico di Milano, Milano, Italy

Reinforcement Learning



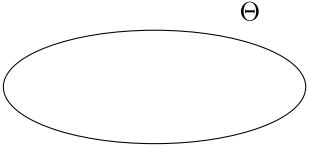
- Policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
- Trajectories $\tau = (s_0, a_0, r_1, s_1, \dots)$
- Return $R(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$
- Goal: $\max_{\pi} \mathbb{E}_{\pi} [R(\tau)]$

Exploration vs Exploitation

Continuous Reinforcement Learning

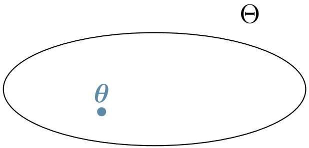
Policy Optimization

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



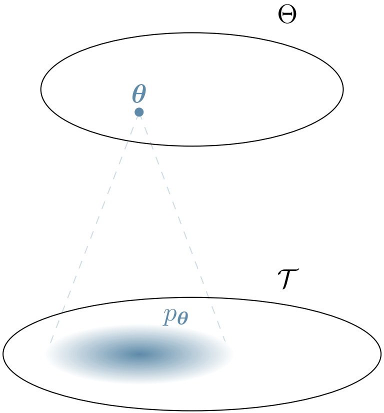
Policy Optimization

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



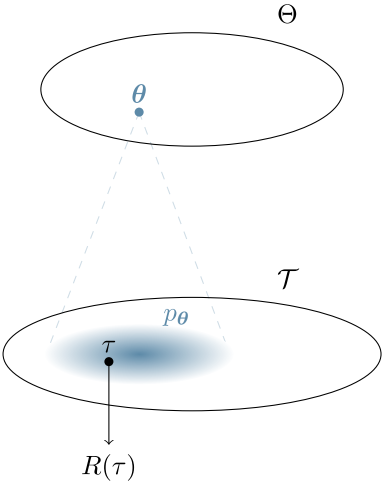
Policy Optimization

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



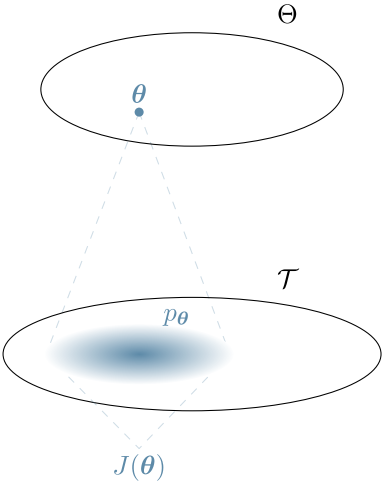
Policy Optimization

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



Policy Optimization

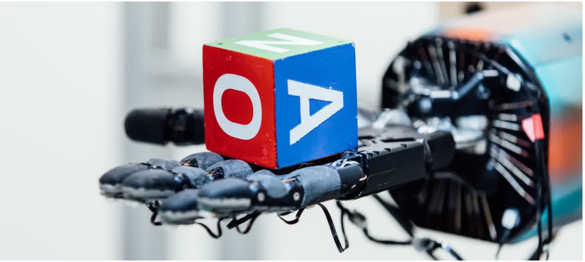
- **Parameter space** $\Theta \subseteq \mathbb{R}^d$
- A **parametric policy** $\pi_{\boldsymbol{\theta}}$ for each $\boldsymbol{\theta} \in \Theta$
- Each inducing a distribution $p_{\boldsymbol{\theta}}$ over **trajectories**
- A **return** $R(\tau)$ for every trajectory τ
- Goal: $\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}\left[R(\tau)\right]$



Policy Gradient Methods

- **Gradient ascent** on $J(\boldsymbol{\theta})$
- Popular algorithms: REINFORCE [15], PGPE [12], TRPO [10], PPO [11]

OpenAI Five, 2018



Learning Dexterous In-Hand Manipulation, OpenAI, 2019

Exploration in Policy Optimization

- Policy Gradient struggles in presence of **sparse rewards** [6]

- Non-convex objective $\implies$ **local minima**

Exploration in Policy Optimization

- Policy Gradient struggles in presence of **sparse rewards** [6]

- Non-convex objective $\implies$ **local minima**

Exploration in Policy Optimization

- Policy Gradient struggles in presence of **sparse rewards** [6]
- Non-convex objective $\implies$ **local minima**

Remedy: inject stochasticity (e.g., entropy bonus [4])

- *Undirected*
- **Unsafe**
- Lack of theoretical understanding

- Policy Gradient struggles in presence of **sparse rewards** [6]
- Non-convex objective $\implies$ **local minima**

Remedy: inject stochasticity (e.g., entropy bonus [4])

- *Undirected*
- **Unsafe**
- Lack of theoretical understanding

Exploration in Policy Optimization

- Policy Gradient struggles in presence of **sparse rewards** [6]
- Non-convex objective $\implies$ **local minima**

Remedy: inject stochasticity (e.g., entropy bonus [4])

- *Undirected*
- **Unsafe**
- Lack of theoretical understanding

- Policy Gradient struggles in presence of **sparse rewards** [6]

- Non-convex objective $\Longrightarrow$ **local minima**

Remedy: inject stochasticity (e.g., entropy bonus [4])

- *Undirected*

- **Unsafe**

- Lack of theoretical understanding

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min \mathit{Regret}(T) = \sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)]$

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min \mathit{Regret}(T) = \sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)]$

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min \mathit{Regret}(T) = \sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)]$

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [8]):

$$a_t = \arg\max_{a \in \mathcal{A}} \underbrace{\widehat{\mu}(a)}_{\textbf{ESTIMATE}}$$

- Idea: be **optimistic** about unknown arms
- Can be applied to RL (e.g., [5])

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [8]):

$$a_t = \arg\max_{a \in \mathcal{A}} \underbrace{\widehat{\mu}(a)}_{\textbf{ESTIMATE}} + \underbrace{C\sqrt{\frac{\log(\frac{1}{\delta})}{\#a}}}_{\textbf{EXPLORATION BONUS}}$$

- Idea: be **optimistic** about unknown arms
- Can be applied to RL (e.g., [5])

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [8]):

$$a_t = \arg\max_{a \in \mathcal{A}} \underbrace{\widehat{\mu}(a)}_{\textbf{ESTIMATE}} + \underbrace{C\sqrt{\frac{\log(\frac{1}{\delta})}{\#a}}}_{\textbf{EXPLORATION BONUS}}$$

- Idea: be **optimistic** about unknown arms
- Can be applied to RL (e.g., [5])

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [8]):

$$a_t = \arg\max_{a \in \mathcal{A}} \underbrace{\widehat{\mu}(a)}_{\textbf{ESTIMATE}} + \underbrace{C\sqrt{\frac{\log(\frac{1}{\delta})}{\#a}}}_{\textbf{EXPLORATION BONUS}}$$

- Idea: be **optimistic** about unknown arms
- Can be applied to RL (e.g., [5])

Policy Optimization as a MAB

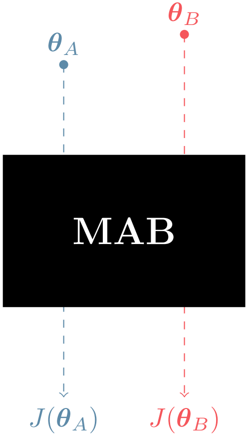
- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB**: *need* structure [7]

$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}\in\Theta} \widehat{J}(\boldsymbol{\theta}_t) + C\sqrt{\frac{\log(\frac{1}{\delta})}{\#\boldsymbol{\theta}}}$$

Policy Optimization as a MAB

- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB**: *need* structure [7]

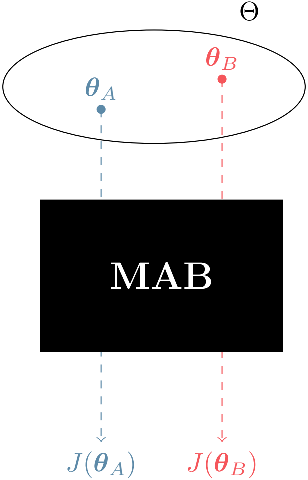
$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}\in\Theta} \widehat{J}(\boldsymbol{\theta}_t) + C\sqrt{\frac{\log(\frac{1}{\delta})}{\#\boldsymbol{\theta}}}$$



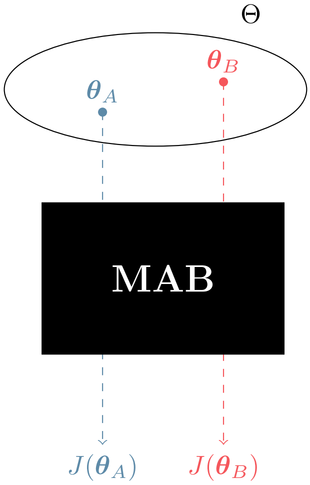
Policy Optimization as a MAB

- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB**: *need* structure [7]

$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}\in\Theta} \widehat{J}(\boldsymbol{\theta}_t) + C\sqrt{\frac{\log(\frac{1}{\delta})}{\#\boldsymbol{\theta}}}$$



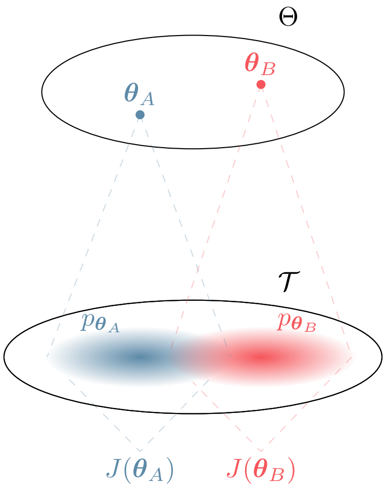
Exploiting Arm Correlation



- Arms correlate through overlapping trajectory distributions
- Use **Importance Sampling (IS)** to transfer information

$$J(\boldsymbol{\theta}_B) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}_A}} \left[\frac{p_{\boldsymbol{\theta}_B}(\tau)}{p_{\boldsymbol{\theta}_A}(\tau)} R(\tau) \right]$$

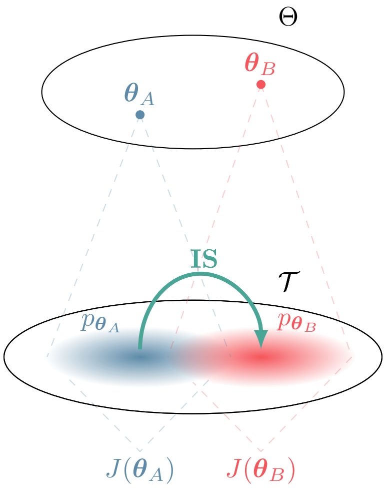
Exploiting Arm Correlation



- Arms correlate through overlapping trajectory distributions
- Use **Importance Sampling (IS)** to transfer information

$$J(\boldsymbol{\theta}_B) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}_A}} \left[\frac{p_{\boldsymbol{\theta}_B}(\tau)}{p_{\boldsymbol{\theta}_A}(\tau)} R(\tau) \right]$$

Exploiting Arm Correlation



- Arms correlate through overlapping trajectory distributions
- Use **Importance Sampling (IS)** to transfer information

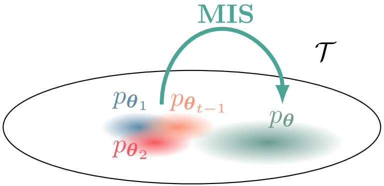
$$J(\boldsymbol{\theta}_B) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}_A}} \left[\frac{p_{\boldsymbol{\theta}_B}(\tau)}{p_{\boldsymbol{\theta}_A}(\tau)} R(\tau) \right]$$

The OPTIMIST index

- A **UCB-like** index:

$$\boldsymbol{\theta}_t \quad = \quad \arg\max_{\boldsymbol{\theta}\in\Theta} \qquad \underbrace{\check{J}_t(\boldsymbol{\theta})}_{\textbf{ESTIMATE}}$$

a **robust multiple** importance sampling estimator



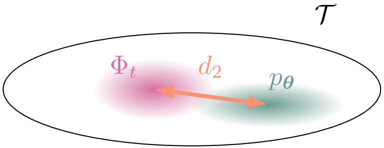
The OPTIMIST index

- A **UCB-like** index:

$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}\in\Theta} \underbrace{\check{J}_t(\boldsymbol{\theta})}_{\textbf{ESTIMATE}} + \underbrace{C\sqrt{\frac{d_2(p_{\boldsymbol{\theta}}\|\Phi_t)\log\frac{1}{\delta_t}}{t}}}_{\textbf{EXPLORATION BONUS:}}$$

a **robust multiple** importance sampling estimator

distributional distance from previous solutions



Robust Multiple Importance Sampling Estimator

- Use Multiple Importance Sampling (MIS) [14] to reuse *all* past experience

- Use **dynamic truncation** to prevent **heavy-tails** [1, 9]

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t}\sum_{k=0}^{t-1} \underbrace{\frac{p_{\boldsymbol{\theta}}(\tau_k)}{\Phi_t(\tau_k)}}_{\textbf{MIS weight}} R(\tau_k), \qquad \underbrace{\Phi_t(\tau) = \frac{1}{\tau}\sum_{k=0}^{t-1} p_{\boldsymbol{\theta}_k}(\tau)}_{\textbf{mixture}}$$

Robust Multiple Importance Sampling Estimator

- Use **Multiple** Importance Sampling (MIS) [14] to reuse *all* past experience

- Use **dynamic truncation** to prevent **heavy-tails** [1, 9]

$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t}\sum_{k=0}^{t-1} \min\left\{M_t, \frac{p_{\boldsymbol{\theta}}(\tau_k)}{\Phi_t(\tau_k)}\right\} R(\tau_k), \qquad \underbrace{M_t = \sqrt{\frac{t d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}{\log(1/\delta_t)}}}_{\textbf{threshold}}$$

Exploration Bonus

- Measure novelty with the *exponentiated* Rényi divergence [3]

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

- Used to **upper bound** the true value (OFU):

$$J(\boldsymbol{\theta}) \leqslant \check{J}_t(\boldsymbol{\theta}) + C \sqrt{\frac{d_2(p_{\boldsymbol{\theta}} \| \Phi_t) \log \frac{1}{\delta_t}}{t}} \qquad \text{with high probability}$$

Sublinear Regret

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

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

Sublinear Regret

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

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

Sublinear Regret

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

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

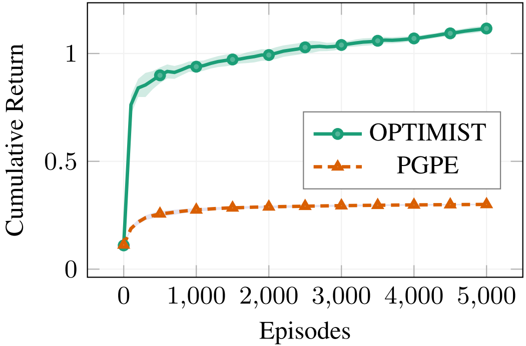
OPTIMIST in Practice

- Easy implementation only for *parameter-based exploration* [12]
- Difficult index optimization $\implies$ **discretization**
- Computational time can be traded-off with regret

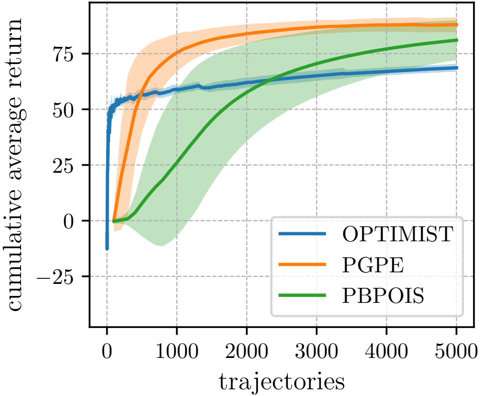
$$\widetilde{\mathcal{O}}\left(dT^{(1-\epsilon/d)}\right) \text{ regret} \implies \mathcal{O}\left(t^{(1+\epsilon)}\right) \text{ time}$$

Empirical Results

River Swim



Mountain Car



Future Work

- Extend to action-based exploration
- Improve index optimization
- Posterior sampling [13]

The Abstract Problem [2]

- Outcome space $\mathcal{Z}$
- Decision set: $\mathcal{P} \in \Delta(\mathcal{Z})$
- Payoff $f : \mathcal{Z} \to \mathbb{R}$
- $\max\limits_{p \in \mathcal{P}} \mathbb{E}_{z \sim p}\left[f(z)\right]$

Thank You for Your Attention!

Papini, Matteo, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. "Optimistic Policy Optimization via Multiple Importance Sampling." In International Conference on Machine Learning, pp. 4989-4999. 2019.

Code: `github.com/WolfLo/optimist`

Contact: matteo.papini@polimi.it

Web page: `t3p.github.io/icml19`

References

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

[2] Chu, C., Blanchet, J., and Glynn, P. (2019). Probability functional descent: A unifying perspective on gans, variational inference, and reinforcement learning. In *International Conference on Machine Learning*, pages 1213–1222.

[3] Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. Curran Associates, Inc.

[4] 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.

[5] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

[6] Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning.

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

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

References (cont.)

[9] Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. (2018). Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, pages 5447–5459.

[10] 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.

[11] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[12] 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.

[13] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

[14] 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.

[15] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.