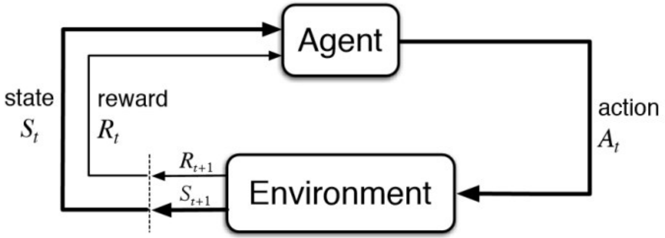
Optimistic Policy Optimization via Multiple Importance Sampling

Matteo Papini Alberto Maria Metelli
Lorenzo Lupo Marcello Restelli

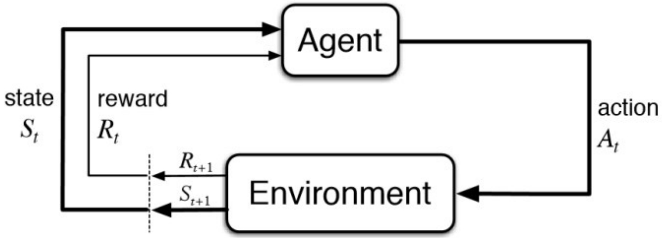
19th September 2019
Markets, Algorithms, Prediction and Learning Workshop, Politecnico di Milano, Milano, Italy

Reinforcement Learning [Sutton and Barto, 2018]



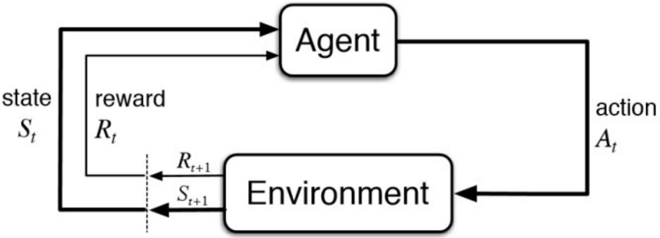
- Policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$
- Trajectories $\tau = (s_0, a_0, r_1, s_1, \dots)$
- Return $R(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$
- Goal: $\max_{\pi} \mathbb{E}_{\pi} [R(\tau)]$

Reinforcement Learning [Sutton and Barto, 2018]



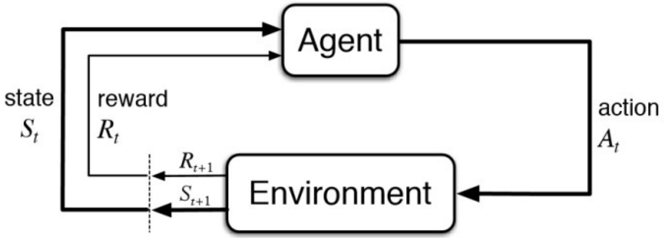
- Policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$
- Trajectories $\tau = (s_0, a_0, r_1, s_1, \dots)$
- Return $R(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$
- Goal: $\max_{\pi} \mathbb{E}_{\pi} [R(\tau)]$

Reinforcement Learning [Sutton and Barto, 2018]



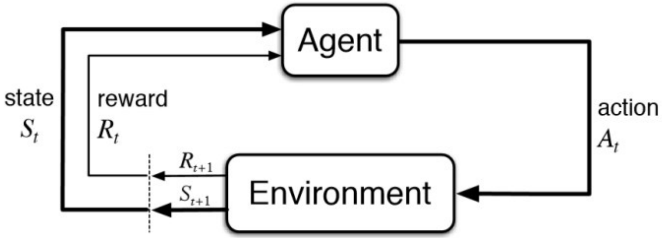
- Policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
- Trajectories $\tau = (s_0, a_0, r_1, s_1, \dots)$
- Return $R(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$
- Goal: $\max_{\pi} \mathbb{E}_{\pi} [R(\tau)]$

Reinforcement Learning [Sutton and Barto, 2018]



- 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: $\max_\pi \mathbb{E}_\pi [R(\tau)]$

Reinforcement Learning [Sutton and Barto, 2018]



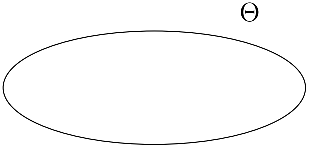
- Policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
- Trajectories $\tau = (s_0, a_0, r_1, s_1, \dots)$
- Return $R(\tau) = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$
- Goal: $\max_{\pi} \mathbb{E}_{\pi} [R(\tau)]$

Exploration vs Exploitation

Continuous Reinforcement Learning

Policy Optimization

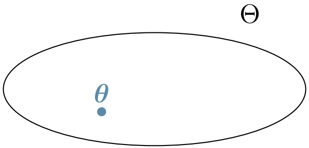
- **Parameter space** $\Theta \subseteq \mathbb{R}^d$



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

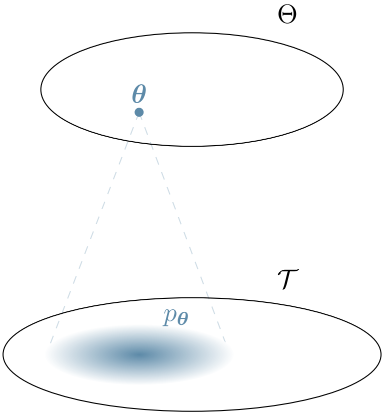
Policy Optimization

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



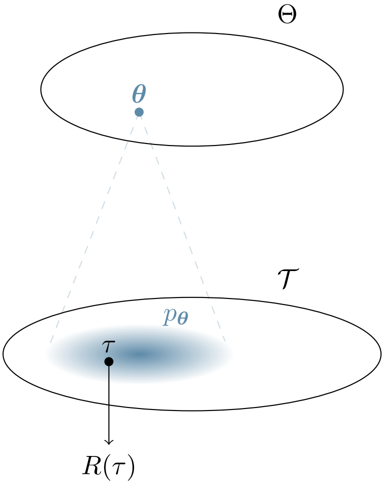
Policy Optimization

- **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: $\max\limits_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}[R(\tau)]$

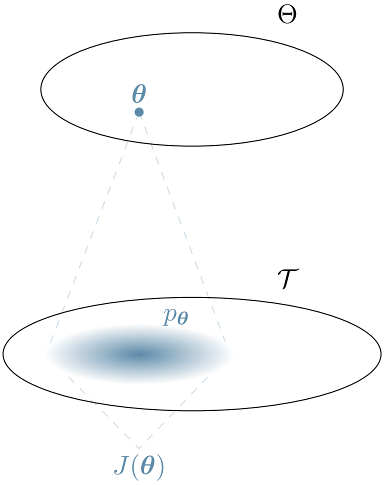


Policy Optimization

- **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: $\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}[R(\tau)]$



- **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: $\max\limits_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}} [R(\tau)]$



Policy Gradient Methods

- **Gradient ascent** on $J(\boldsymbol{\theta})$

- Popular algorithms: **REINFORCE** [Williams, 1992], **PGPE** [Sehnke et al., 2008], **TRPO** [Schulman et al., 2015], **PPO** [Schulman et al., 2017]

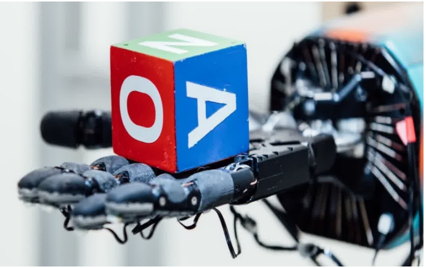
Policy Gradient Methods

- **Gradient ascent** on $J(\boldsymbol{\theta})$
- Popular algorithms: **REINFORCE** [Williams, 1992], **PGPE** [Sehnke et al., 2008], **TRPO** [Schulman et al., 2015], **PPO** [Schulman et al., 2017]

Policy Gradient Methods

- **Gradient ascent** on $J(\boldsymbol{\theta})$
- Popular algorithms: **REINFORCE** [Williams, 1992], **PGPE** [Sehnke et al., 2008], **TRPO** [Schulman et al., 2015], **PPO** [Schulman et al., 2017]

Dota 2 [OpenAI, 2018]



Manipulation [Andrychowicz et al., 2018]

Exploration in Policy Optimization

- Policy Gradient fails with **sparse rewards** [Kakade and Langford, 2002]

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

Exploration in Policy Optimization

- Policy Gradient fails with **sparse rewards** [Kakade and Langford, 2002]

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

Exploration in Policy Optimization

- Policy Gradient fails with **sparse rewards** [Kakade and Langford, 2002]
- Non-convex objective $\implies$ **local minima**

Entropy bonus [Haarnoja et al., 2018]:

- *Undirected*
- **Unsafe**
- Little theoretical understanding [Ahmed et al., 2018]

Exploration in Policy Optimization

- Policy Gradient fails with **sparse rewards** [Kakade and Langford, 2002]

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

Entropy bonus [Haarnoja et al., 2018]:

- *Undirected*
- **Unsafe**
- Little theoretical understanding [Ahmed et al., 2018]

Exploration in Policy Optimization

- Policy Gradient fails with **sparse rewards** [Kakade and Langford, 2002]

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

Entropy bonus [Haarnoja et al., 2018]:

- *Undirected*
- **Unsafe**
- Little theoretical understanding [Ahmed et al., 2018]

Exploration in Policy Optimization

- Policy Gradient fails with **sparse rewards** [Kakade and Langford, 2002]
- Non-convex objective $\implies$ **local minima**

Entropy bonus [Haarnoja et al., 2018]:

- *Undirected*
- **Unsafe**
- Little theoretical understanding [Ahmed et al., 2018]

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min Regret(T) = \sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)]$
- Wide literature on **directed exploration** [Bubeck et al., 2012, Lattimore and Szepesvári, 2019]

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min \mathit{Regret}(T) = \sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)]$
- Wide literature on **directed exploration** [Bubeck et al., 2012, Lattimore and Szepesvári, 2019]

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min \textit{Regret}(T) = \sum_{t=1}^{T} \left[\mu(a^*) - \mu(a_t)\right]$
- Wide literature on **directed exploration** [Bubeck et al., 2012, Lattimore and Szepesvári, 2019]

Multi Armed Bandit (MAB)

- Arms $a \in \mathcal{A}$
- Expected payoff $\mu(a)$
- Goal: $\min \mathit{Regret}(T) = \sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)]$
- Wide literature on **directed exploration** [Bubeck et al., 2012, Lattimore and Szepesvári, 2019]

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [Lai and Robbins, 1985]):

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

- Idea: be **optimistic** about unknown arms

- Can be applied to RL (e.g., Jaksch et al. [2010])

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [Lai and Robbins, 1985]):

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

- Idea: be **optimistic** about unknown arms

- Can be applied to RL (e.g., Jaksch et al. [2010])

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [Lai and Robbins, 1985]):

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

- Idea: be **optimistic** about unknown arms

- Can be applied to RL (e.g., Jaksch et al. [2010])

Optimism in Face of Uncertainty

- OFU strategy (e.g., UCB [Lai and Robbins, 1985]):

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

- Idea: be **optimistic** about unknown arms

- Can be applied to RL (e.g., Jaksch et al. [2010])

Policy Optimization as a MAB

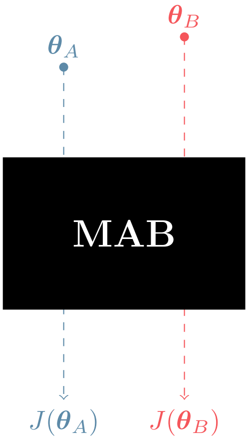
- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB**: we *need* structure [Kleinberg et al., 2013]

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

Policy Optimization as a MAB

- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB**: we *need* structure [Kleinberg et al., 2013]

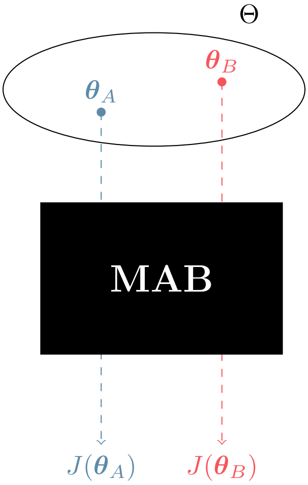
$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}\in\Theta} \quad \widehat{J}(\boldsymbol{\theta}_t) \quad + C\sqrt{\frac{\log(\frac{1}{\delta})}{\#\boldsymbol{\theta}}}$$



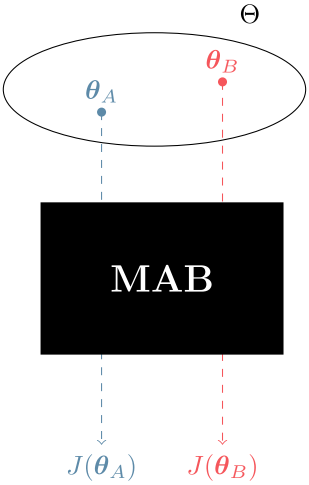
Policy Optimization as a MAB

- **Arms:** parameters $\boldsymbol{\theta}$
- **Payoff:** expected return $J(\boldsymbol{\theta})$
- **Continuous MAB**: we *need* structure [Kleinberg et al., 2013]

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



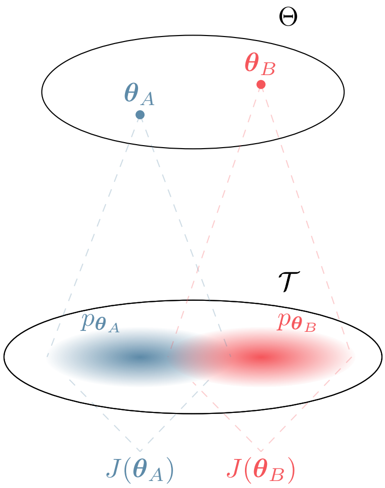
Exploiting Arm Correlation



- 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

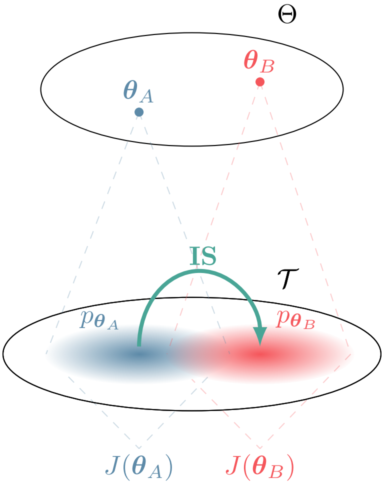


- Arms correlate through overlapping trajectory distributions

- 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

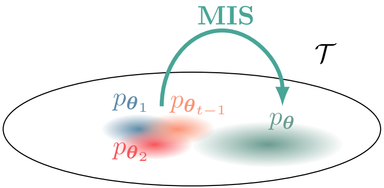
$$J(\boldsymbol{\theta}_B) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}_A}} \left[\frac{p_{\boldsymbol{\theta}_B}(\tau)}{p_{\boldsymbol{\theta}_A}(\tau)} R(\tau) \right]$$

The OPTIMIST index [Papini et al., 2019]

- A **UCB-like** index:

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

a **robust multiple** importance sampling estimator



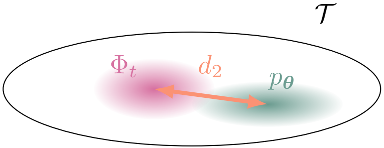
The OPTIMIST index [Papini et al., 2019]

- A **UCB-like** index:

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

ESTIMATE: a **robust multiple** importance sampling estimator

EXPLORATION BONUS: **distributional** distance from previous solutions



Robust Multiple Importance Sampling Estimator

- Use **Multiple** Importance Sampling (MIS) [Veach and Guibas, 1995] to reuse *all* past experience

- Use **dynamic truncation** to prevent **heavy-tails** [Bubeck et al., 2013, Metelli et al., 2018]

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

Robust Multiple Importance Sampling Estimator

- Use **Multiple** Importance Sampling (MIS) [Veach and Guibas, 1995] to reuse *all* past experience

- Use **dynamic truncation** to prevent **heavy-tails** [Bubeck et al., 2013, Metelli et al., 2018]

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

Exploration Bonus

- Measure novelty with the *exponentiated* **Rényi divergence** [Cortes et al., 2010, Metelli et al., 2018]

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

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

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

Exploration Bonus

- Measure novelty with the *exponentiated* **Rényi divergence** [Cortes et al., 2010, Metelli et al., 2018]

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

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

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

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* Sehnke et al. [2008]
- Difficult index optimization $\implies$ **discretization**
- Computational time can be traded-off with regret

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

OPTIMIST in Practice

- Easy implementation only for *parameter-based exploration* Sehnke et al. [2008]
- Difficult index optimization $\implies$ **discretization**
- Computational time can be traded-off with regret

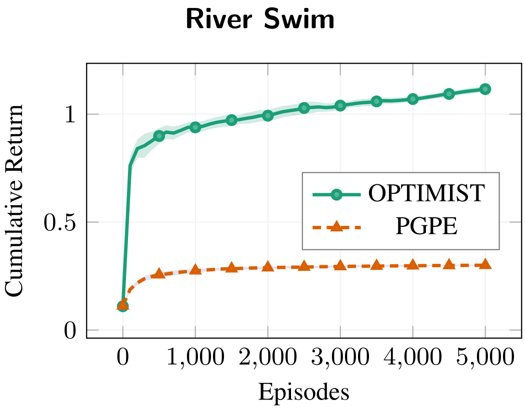
$$\widetilde{\mathcal{O}}\left(dT^{(1-\epsilon/d)}\right) \text{ regret } \implies \mathcal{O}\left(t^{(1+\epsilon)}\right) \text{ time}$$

OPTIMIST in Practice

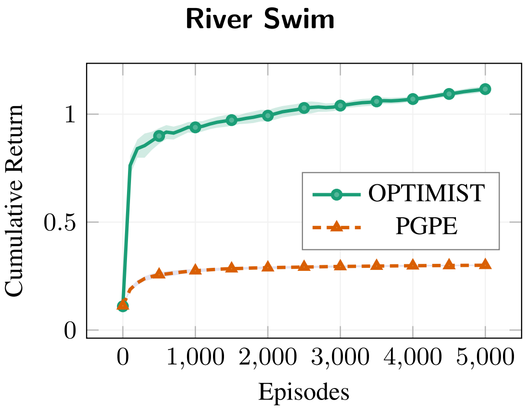
- Easy implementation only for *parameter-based exploration* Sehnke et al. [2008]
- Difficult index optimization $\implies$ **discretization**
- Computational time can be traded-off with regret

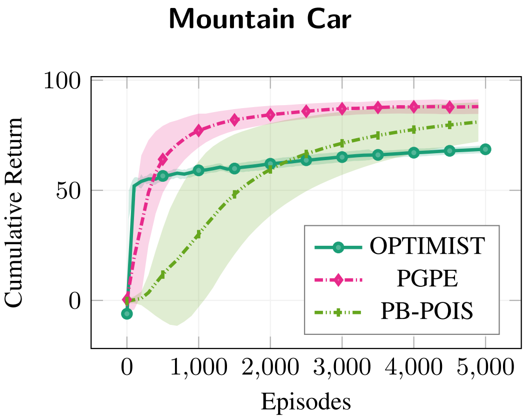
$$\widetilde{\mathcal{O}}\left(dT^{(1-\epsilon/d)}\right) \text{ regret } \implies \mathcal{O}\left(t^{(1+\epsilon)}\right) \text{ time}$$

Empirical Results



Empirical Results





Future Work

- Extend to action-based exploration
- Improve index optimization
- Posterior sampling [Thompson, 1933]

Future Work

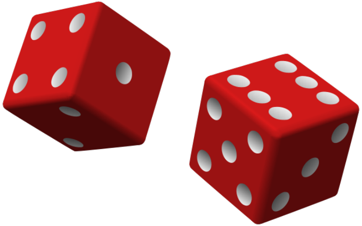
- Extend to action-based exploration
- Improve index optimization
- Posterior sampling [Thompson, 1933]

Future Work

- Extend to action-based exploration
- Improve index optimization
- Posterior sampling [Thompson, 1933]

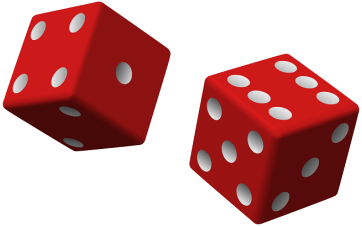
The Abstract Problem [Chu et al., 2019]

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



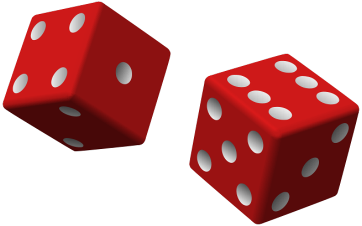
The Abstract Problem [Chu et al., 2019]

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



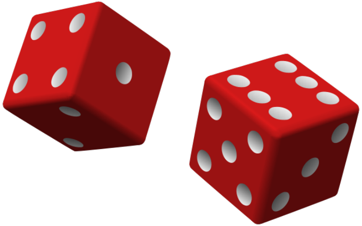
The Abstract Problem [Chu et al., 2019]

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



The Abstract Problem [Chu et al., 2019]

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



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

Ahmed, Z., Roux, N. L., Norouzi, M., and Schuurmans, D. (2018). Understanding the impact of entropy in policy learning. *arXiv preprint arXiv:1811.11214*.

Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. (2018). Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*.

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

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

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.

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.

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.

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

References (cont.)

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

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

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

Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms*. Cambridge University Press (preprint).

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.

OpenAI (2018). Openai five. `https://blog.openai.com/openai-five/`.

Papini, M., Metelli, A. M., Lupo, L., and Restelli, M. (2019). Optimistic policy optimization via multiple importance sampling. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999, Long Beach, California, USA. PMLR.

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.

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

References (cont.)

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.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

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.

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.

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