Policy Optimization as Online Learning with Mediator Feedback

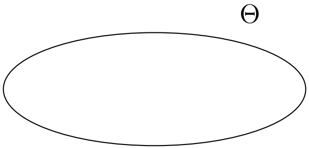
Alberto Maria Metelli* Matteo Papini* Pierluca D'Oro Marcello Restelli

February 2021
35th AAAI Conference on Artificial Intelligence

Policy Optimization (PO)

- **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** $\mathcal{R}(\tau)$ for every trajectory τ
- **Goal**: maximize the **expected return** (Deisenroth et al., 2013)

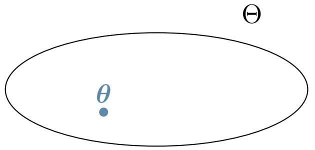
$$J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}} [\mathcal{R}(\tau)]$$



Policy Optimization (PO)

- **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** $\mathcal{R}(\tau)$ for every trajectory τ
- **Goal**: maximize the **expected return** (Deisenroth et al., 2013)

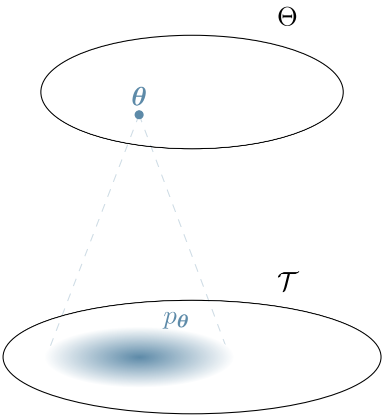
$$J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}} [\mathcal{R}(\tau)]$$



Policy Optimization (PO)

- **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** $\mathcal{R}(\tau)$ for every trajectory τ
- **Goal**: maximize the **expected return** (Deisenroth et al., 2013)

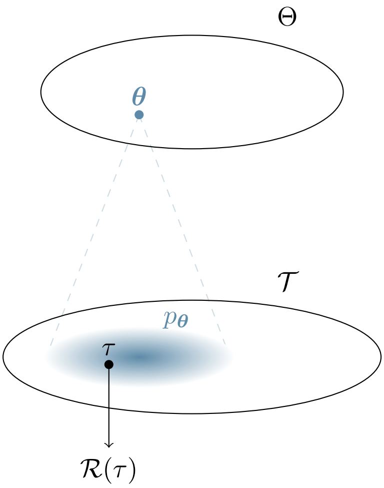
$$J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}} [\mathcal{R}(\tau)]$$



Policy Optimization (PO)

- **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** $\mathcal{R}(\tau)$ for every trajectory τ
- **Goal**: maximize the **expected return** (Deisenroth et al., 2013)

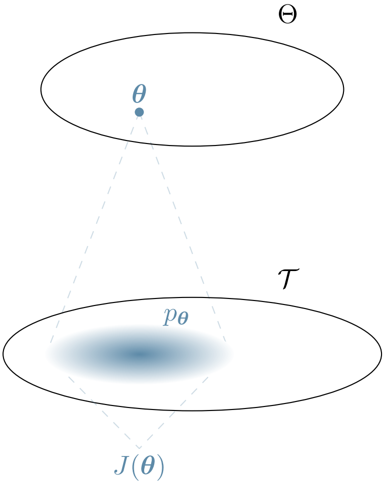
$$J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}} [\mathcal{R}(\tau)]$$



Policy Optimization (PO)

- **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** $\mathcal{R}(\tau)$ for every trajectory τ
- **Goal**: maximize the **expected return** (Deisenroth et al., 2013)

$$J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\boldsymbol{\theta}}} [\mathcal{R}(\tau)]$$



Online PO with Bandit Feedback

- **Select** parameter $\boldsymbol{\theta}$ and run $\pi_{\boldsymbol{\theta}}$
- **Observe** the trajectory τ
- **Observe** the return $\mathcal{R}(\tau)$

- **Goal**: minimize the **regret** (Auer et al., 2002)

$$\mathrm{Regret}(n) = \sum_{t=1}^{n} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t) = \sum_{t=1}^{n} \Delta(\boldsymbol{\theta}_t)$$

- The trajectory τ **mediates** between the parameter $\boldsymbol{\theta}$ and the return $\mathcal{R}(\tau)$ (Papini et al., 2019)

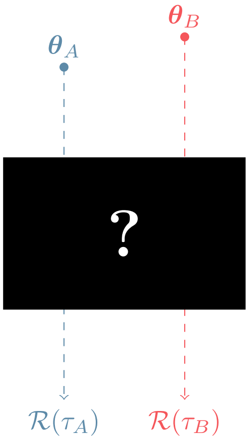
Online PO with Bandit Feedback

- **Select** parameter $\boldsymbol{\theta}$ and run $\pi_{\boldsymbol{\theta}}$
- **Observe** the trajectory τ
- **Observe** the return $\mathcal{R}(\tau)$

- **Goal**: minimize the **regret** (Auer et al., 2002)

$$\text{Regret}(n) = \sum_{t=1}^{n} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t) = \sum_{t=1}^{n} \Delta(\boldsymbol{\theta}_t)$$

- The trajectory τ **mediates** between the parameter $\boldsymbol{\theta}$ and the return $\mathcal{R}(\tau)$ (Papini et al., 2019)



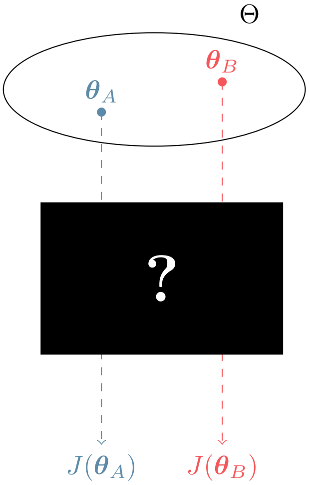
Online PO with Bandit Feedback

- **Select** parameter $\boldsymbol{\theta}$ and run $\pi_{\boldsymbol{\theta}}$
- **Observe** the trajectory τ
- **Observe** the return $\mathcal{R}(\tau)$

- **Goal**: minimize the **regret** (Auer et al., 2002)

$$\mathrm{Regret}(n) = \sum_{t=1}^{n} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t) = \sum_{t=1}^{n} \Delta(\boldsymbol{\theta}_t)$$

- The trajectory τ **mediates** between the parameter $\boldsymbol{\theta}$ and the return $\mathcal{R}(\tau)$ (Papini et al., 2019)



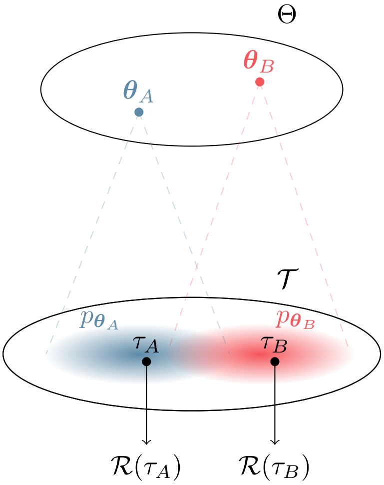
Online PO with Mediator Feedback

- **Select** parameter $\boldsymbol{\theta}$ and run $\pi_{\boldsymbol{\theta}}$
- **Observe** the trajectory τ
- **Observe** the return $\mathcal{R}(\tau)$

- **Goal**: minimize the **regret** (Auer et al., 2002)

$$\text{Regret}(n) = \sum_{t=1}^{n} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t) = \sum_{t=1}^{n} \Delta(\boldsymbol{\theta}_t)$$

- The trajectory τ **mediates** between the parameter θ and the return $\mathcal{R}(\tau)$ (Papini et al., 2019)



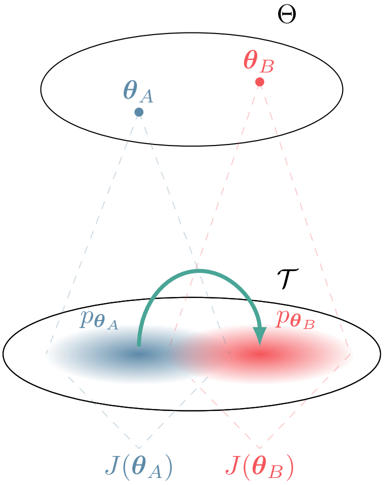
Online PO with Mediator Feedback

- **Select** parameter $\boldsymbol{\theta}$ and run $\pi_{\boldsymbol{\theta}}$
- **Observe** the trajectory τ
- **Observe** the return $\mathcal{R}(\tau)$

- **Goal**: minimize the **regret** (Auer et al., 2002)

$$\text{Regret}(n) = \sum_{t=1}^{n} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_t) = \sum_{t=1}^{n} \Delta(\boldsymbol{\theta}_t)$$

- The trajectory τ **mediates** between the parameter $\boldsymbol{\theta}$ and the return $\mathcal{R}(\tau)$ (Papini et al., 2019)



Contributions

- Regret Lower Bounds with Mediator feedback
- Importance Sampling for Mediator feedback
- New Randomized Algorithm: RANDOMIST and Regret Analysis
- Numerical Simulations

Contributions

- Regret Lower Bounds with Mediator feedback
- Importance Sampling for Mediator feedback
- New Randomized Algorithm: RANDOMIST and Regret Analysis
- Numerical Simulations

Contributions

- Regret Lower Bounds with Mediator feedback
- Importance Sampling for Mediator feedback
- New Randomized Algorithm: RANDOMIST and Regret Analysis
- Numerical Simulations

Contributions

- Regret Lower Bounds with Mediator feedback
- Importance Sampling for Mediator feedback
- New Randomized Algorithm: RANDOMIST and Regret Analysis
- Numerical Simulations

Regret Lower Bounds with Mediator Feedback

$\Theta = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_B\}$ with $\Delta = J(\boldsymbol{\theta}_A) - J(\boldsymbol{\theta}_B)$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) < \infty$ and $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) < \infty$ $\implies$ **constant regret**

$$\mathbb{E}\,\text{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\right)$$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) = \infty$ or $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) = \infty$ $\implies$ **logarithmic regret**

$$\mathbb{E}\,\text{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\log\left(\Delta^2 n\right)\right)$$

Regret Lower Bounds with Mediator Feedback

$\Theta = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_B\}$ with $\Delta = J(\boldsymbol{\theta}_A) - J(\boldsymbol{\theta}_B)$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) < \infty$ and $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) < \infty$ $\implies$ **constant regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\right)$$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) = \infty$ or $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) = \infty$ $\implies$ **logarithmic regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\log\left(\Delta^2 n\right)\right)$$

Regret Lower Bounds with Mediator Feedback

$\Theta = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_B\}$ with $\Delta = J(\boldsymbol{\theta}_A) - J(\boldsymbol{\theta}_B)$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) < \infty$ and $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) < \infty$ $\implies$ **constant regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\right)$$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) = \infty$ or $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) = \infty$ $\implies$ **logarithmic regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\log\left(\Delta^2 n\right)\right)$$

Regret Lower Bounds with Mediator Feedback

$\Theta = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_B\}$ with $\Delta = J(\boldsymbol{\theta}_A) - J(\boldsymbol{\theta}_B)$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) < \infty$ and $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) < \infty$ $\implies$ **constant regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\right)$$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) = \infty$ or $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) = \infty$ $\implies$ **logarithmic regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\log\left(\Delta^2 n\right)\right)$$

Regret Lower Bounds with Mediator Feedback

$\Theta = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_B\}$ with $\Delta = J(\boldsymbol{\theta}_A) - J(\boldsymbol{\theta}_B)$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) < \infty$ and $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) < \infty \implies$ **constant regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\right)$$

- If $D_{KL}(p_{\boldsymbol{\theta}_A} \| p_{\boldsymbol{\theta}_B}) = \infty$ or $D_{KL}(p_{\boldsymbol{\theta}_B} \| p_{\boldsymbol{\theta}_A}) = \infty \implies$ **logarithmic regret**

$$\mathbb{E}\,\mathrm{Regret}(n) \geqslant \mathcal{O}\left(\frac{1}{\Delta}\log\left(\Delta^2 n\right)\right)$$

Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1}\sum_{i=1}^{t-1} \underbrace{\omega_{\boldsymbol{\theta},t}(\tau_i)}_{\text{multiple importance sampling (Veach and Guibas, 1995)}} \mathcal{R}(\tau_i)$$

multiple importance sampling
(Veach and Guibas, 1995)

- Heavy-tail behavior, only **polynomial concentration** (Metelli et al., 2018, 2020):

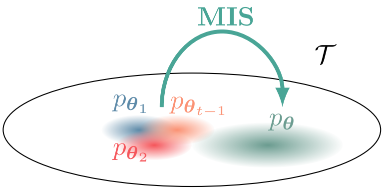
$$\widehat{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant \sqrt{\frac{1-\delta}{\delta(t-1)} \underbrace{d_2\left(p_{\boldsymbol{\theta}} \| \Phi_t\right)}_{\text{Rényi divergence}}}$$

Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1} \sum_{i=1}^{t-1} \underbrace{\frac{p_{\boldsymbol{\theta}}(\tau_i)}{\frac{1}{t-1}\sum_{j=1}^{t-1} p_{\boldsymbol{\theta}_j}(\tau_i)}} \mathcal{R}(\tau_i)$$

multiple importance sampling with **balance heuristic** (Owen, 2013)



- Heavy-tail behavior, only **polynomial concentration** (Metelli et al., 2018, 2020):

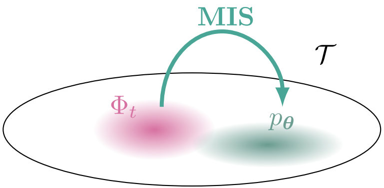
$$\widehat{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant \sqrt{\frac{1-\delta}{\delta(t-1)} \underbrace{d_2\left(p_{\boldsymbol{\theta}} \| \Phi_t\right)}_{\text{Renyi divergence}}}$$

Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1} \sum_{i=1}^{t-1} \underbrace{\frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}} \mathcal{R}(\tau_i)$$

multiple importance sampling with **balance heuristic** (Owen, 2013)



- Heavy-tail behavior, only **polynomial concentration** (Metelli et al., 2018, 2020):

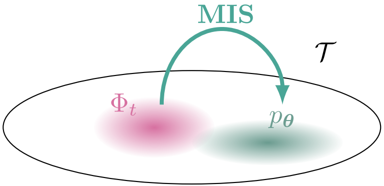
$$\widehat{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant \sqrt{\frac{1-\delta}{\delta(t-1)} \underbrace{d_2\left(p_{\boldsymbol{\theta}} \| \Phi_t\right)}_{\text{Renyi divergence}}}$$

Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1}\sum_{i=1}^{t-1} \underbrace{\frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}} \mathcal{R}(\tau_i)$$

multiple importance sampling with **balance heuristic** (Owen, 2013)



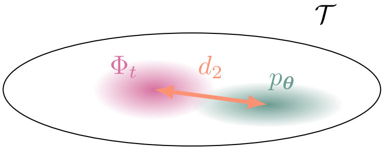
- Heavy-tail behavior, only **polynomial concentration** (Metelli et al., 2018, 2020):

$$\widehat{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant \sqrt{\frac{1-\delta}{\delta(t-1)} \underbrace{d_2\left(p_{\boldsymbol{\theta}} \| \Phi_t\right)}_{\text{Renyi divergence}}}$$

Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1}\sum_{i=1}^{t-1} \underbrace{\frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}} \mathcal{R}(\tau_i)$$



multiple importance sampling
with **balance heuristic**
(Owen, 2013)

- Heavy-tail behavior, only **polynomial concentration** (Metelli et al., 2018, 2020):

$$\widehat{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant \sqrt{\frac{1-\delta}{\delta(t-1)} \underbrace{d_2\left(p_{\boldsymbol{\theta}} \| \Phi_t\right)}_{\text{Renyi divergence}}}$$

Truncated Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1} \sum_{i=1}^{t-1} \underbrace{\check{\omega}_{\boldsymbol{\theta},t}(\tau_i)}_{\text{truncated multiple importance sampling (Ionides, 2008)}} \mathcal{R}(\tau_i)$$

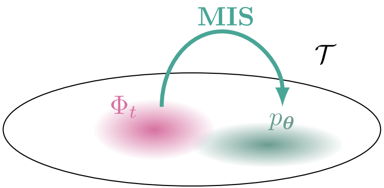
- If $M_t(\boldsymbol{\theta}) = \sqrt{\frac{(t-1) d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}{\log \frac{1}{\delta}}}$, we get exponential concentration (Papini et al., 2019; Metelli et al., 2020):

$$\check{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant 2.75 \sqrt{\frac{\log \frac{1}{\delta}}{t-1} \underbrace{d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}_{\text{Renyi divergence}}}$$

Truncated Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1} \sum_{i=1}^{t-1} \underbrace{\min\left\{ \frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}, M_t(\boldsymbol{\theta}) \right\}}_{\text{truncated multiple importance sampling (Ionides, 2008)}} \mathcal{R}(\tau_i)$$

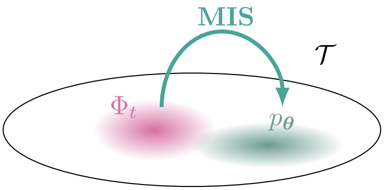


- If $M_t(\boldsymbol{\theta}) = \sqrt{\frac{(t-1) d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}{\log \frac{1}{\delta}}}$, we get exponential concentration (Papini et al., 2019; Metelli et al., 2020):

$$\check{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant 2.75 \sqrt{\frac{\log \frac{1}{\delta}}{t-1} \underbrace{d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}_{\text{Renyi divergence}}}$$

Truncated Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy



$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1} \sum_{i=1}^{t-1} \underbrace{\min\left\{ \frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}, M_t(\boldsymbol{\theta}) \right\}}_{\text{truncated multiple importance sampling (Ionides, 2008)}} \mathcal{R}(\tau_i)$$

- If $M_t(\boldsymbol{\theta}) = \sqrt{\frac{(t-1) d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}{\log \frac{1}{\delta}}}$, we get **exponential concentration** (Papini et al., 2019; Metelli et al., 2020):

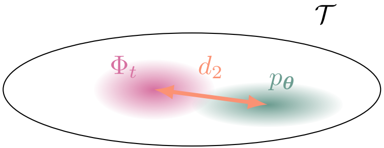
$$\check{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant 2.75 \sqrt{\frac{\log \frac{1}{\delta}}{t-1} \underbrace{d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}_{\text{Renyi divergence}}}$$

Truncated Importance Sampling for Mediator Feedback

- **Idea**: use **all** the samples to estimate the expected return of **any** policy

$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1}\sum_{i=1}^{t-1} \underbrace{\min\left\{\frac{p_{\boldsymbol{\theta}}(\tau_i)}{\Phi_t(\tau_i)}, M_t(\boldsymbol{\theta})\right\}}_{} \mathcal{R}(\tau_i)$$

truncated multiple
importance sampling
(Ionides, 2008)



- If $M_t(\boldsymbol{\theta}) = \sqrt{\frac{(t-1)d_2(p_{\boldsymbol{\theta}}\|\Phi_t)}{\log\frac{1}{\delta}}}$,
  we get exponential concentration (Papini et al., 2019; Metelli et al., 2020):

$$\check{J}_t(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) \leqslant 2.75\sqrt{\frac{\log\frac{1}{\delta}}{t-1}\underbrace{d_2\left(p_{\boldsymbol{\theta}}\|\Phi_t\right)}_{\text{Renyi divergence}}}$$

- **Previous Work**: OPTIMIST employs a UCB-like approach (Papini et al., 2019)
- **Idea**: perturb the estimate $\check{J}_t(\boldsymbol{\theta})$ (Kveton et al., 2019)
- For **finite** policy spaces:
  1. Compute expected return $\check{J}_t(\boldsymbol{\theta})$
  2. Generate perturbation $U_t(\boldsymbol{\theta})$
  3. Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta}\in\Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

- **Previous Work**: OPTIMIST employs a UCB-like approach (Papini et al., 2019)
- **Idea**: perturb the estimate $\check{J}_t(\boldsymbol{\theta})$ (Kveton et al., 2019)
- For **finite** policy spaces:
  1. Compute expected return $\check{J}_t(\boldsymbol{\theta})$
  2. Generate perturbation $U_t(\boldsymbol{\theta})$
  3. Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta}\in\Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

- **Previous Work**: OPTIMIST employs a UCB-like approach (Papini et al., 2019)
- **Idea**: perturb the estimate $\check{J}_t(\boldsymbol{\theta})$ (Kveton et al., 2019)
- For **finite** policy spaces:
  1. Compute expected return $\check{J}_t(\boldsymbol{\theta})$
  2. Generate perturbation $U_t(\boldsymbol{\theta})$
  3. Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta}\in\Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

- **Previous Work**: OPTIMIST employs a UCB-like approach (Papini et al., 2019)
- **Idea**: perturb the estimate $\check{J}_t(\boldsymbol{\theta})$ (Kveton et al., 2019)
- For **finite** policy spaces:
  1. Compute expected return $\check{J}_t(\boldsymbol{\theta})$
  2. Generate perturbation $U_t(\boldsymbol{\theta})$
  3. Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta} \in \Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

- **Previous Work**: OPTIMIST employs a UCB-like approach (Papini et al., 2019)
- **Idea**: perturb the estimate $\check{J}_t(\boldsymbol{\theta})$ (Kveton et al., 2019)
- For **finite** policy spaces:
  1. Compute expected return $\check{J}_t(\boldsymbol{\theta})$
  2. Generate perturbation $U_t(\boldsymbol{\theta})$
  3. Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta} \in \Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

- **Previous Work**: OPTIMIST employs a UCB-like approach (Papini et al., 2019)
- **Idea**: perturb the estimate $\check{J}_t(\boldsymbol{\theta})$ (Kveton et al., 2019)
- For **finite** policy spaces:
  1. Compute expected return $\check{J}_t(\boldsymbol{\theta})$
  2. Generate perturbation $U_t(\boldsymbol{\theta})$
  3. Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta} \in \Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$

- For **compact** policy spaces, the $\arg\max$ cannot be computed
- **Sample** from the distribution of being the max (D'Eramo et al., 2017) with MCMC (Beskos and Stuart, 2009):

$$\boldsymbol{\theta}_t \sim \Pr\left(\check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \Theta} \check{J}_t(\boldsymbol{\theta}') + U_t(\boldsymbol{\theta}')\right)$$

- For **compact** policy spaces, the $\arg\max$ cannot be computed
- **Sample** from the distribution of being the max (D'Eramo et al., 2017) with MCMC (Beskos and Stuart, 2009):

$$\boldsymbol{\theta}_t \sim \Pr\left(\check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \Theta} \check{J}_t(\boldsymbol{\theta}') + U_t(\boldsymbol{\theta}') \right)$$

$v = \max_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ and $\Delta = \min_{\boldsymbol{\theta} \neq \boldsymbol{\theta}^*} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta})$

Algorithm	Exploration	$\mathbb{E}\,\mathrm{Regret}(n)$	
		$v = \infty$	$v < \infty$
Greedy	-	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
UCB1 (Auer et al., 2002)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$
OPTIMIST (Papini et al., 2019)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
RANDOMIST	randomized	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
Lower Bound	-	$\mathcal{O}\left(\frac{1}{\Delta} \log(\Delta^2 n)\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\right)$

Finite Policy Space

$v = \max_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ and $\Delta = \min_{\boldsymbol{\theta} \neq \boldsymbol{\theta}^*} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta})$

Algorithm	Exploration	$\mathbb{E}\,\mathrm{Regret}(n)$	
		$v = \infty$	$v < \infty$
Greedy	-	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{v}{\Delta}\log\frac{v}{\Delta^2}\right)$
UCB1 (Auer et al., 2002)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$
OPTIMIST (Papini et al., 2019)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta}\log\frac{v}{\Delta^2}\right)$
RANDOMIST	randomized	$\mathcal{O}\left(\frac{1}{\Delta}\log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta}\log\frac{v}{\Delta^2}\right)$
Lower Bound	-	$\mathcal{O}\left(\frac{1}{\Delta}\log(\Delta^2 n)\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\right)$

Finite Policy Space

$v = \max_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ and $\Delta = \min_{\boldsymbol{\theta} \neq \boldsymbol{\theta}^*} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta})$

Algorithm	Exploration	$\mathbb{E}\,\mathrm{Regret}(n)$ $v = \infty$	$\mathbb{E}\,\mathrm{Regret}(n)$ $v < \infty$
Greedy	-	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
UCB1 (Auer et al., 2002)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$
OPTIMIST (Papini et al., 2019)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
RANDOMIST	randomized	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
Lower Bound	-	$\mathcal{O}\left(\frac{1}{\Delta} \log(\Delta^2 n)\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\right)$

$v = \max_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ and $\Delta = \min_{\boldsymbol{\theta} \neq \boldsymbol{\theta}^*} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta})$

Algorithm	Exploration	$\mathbb{E}\,\mathrm{Regret}(n)$	
		$v = \infty$	$v < \infty$
Greedy	-	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
UCB1 (Auer et al., 2002)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$
OPTIMIST (Papini et al., 2019)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
RANDOMIST	randomized	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
Lower Bound	-	$\mathcal{O}\left(\frac{1}{\Delta} \log(\Delta^2 n)\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\right)$

Finite Policy Space

$v = \max_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ and $\Delta = \min_{\boldsymbol{\theta} \neq \boldsymbol{\theta}^*} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta})$

Algorithm	Exploration	$\mathbb{E}\,\mathrm{Regret}(n)$	
		$v = \infty$	$v < \infty$
Greedy	-	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
UCB1 (Auer et al., 2002)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$
OPTIMIST (Papini et al., 2019)	deterministic	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
RANDOMIST	randomized	$\mathcal{O}\left(\frac{1}{\Delta} \log n\right)$	$\mathcal{O}\left(\frac{v}{\Delta} \log \frac{v}{\Delta^2}\right)$
Lower Bound	-	$\mathcal{O}\left(\frac{1}{\Delta} \log(\Delta^2 n)\right)$	$\mathcal{O}\left(\frac{1}{\Delta}\right)$

Compact Policy Space

$\Theta = [-D, D]^d$ and $v = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$

Algorithm	Approach	Complexity	$\mathbb{E}\,\mathrm{Regret}(n)$
OPTIMIST (Papini et al., 2019)	discretization	$t^{1+\frac{d}{2}}$	$\mathcal{O}\left(\sqrt{vdn}\right)$
RANDOMIST	MCMC sampling	dt^2	?

Compact Policy Space

$\Theta = [-D, D]^d$ and $v = \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$

Algorithm	Approach	Complexity	$\mathbb{E}\,\mathrm{Regret}(n)$
OPTIMIST (Papini et al., 2019)	discretization	$t^{1+\frac{d}{2}}$	$\mathcal{O}\left(\sqrt{vdn}\right)$
RANDOMIST	MCMC sampling	dt^2	?

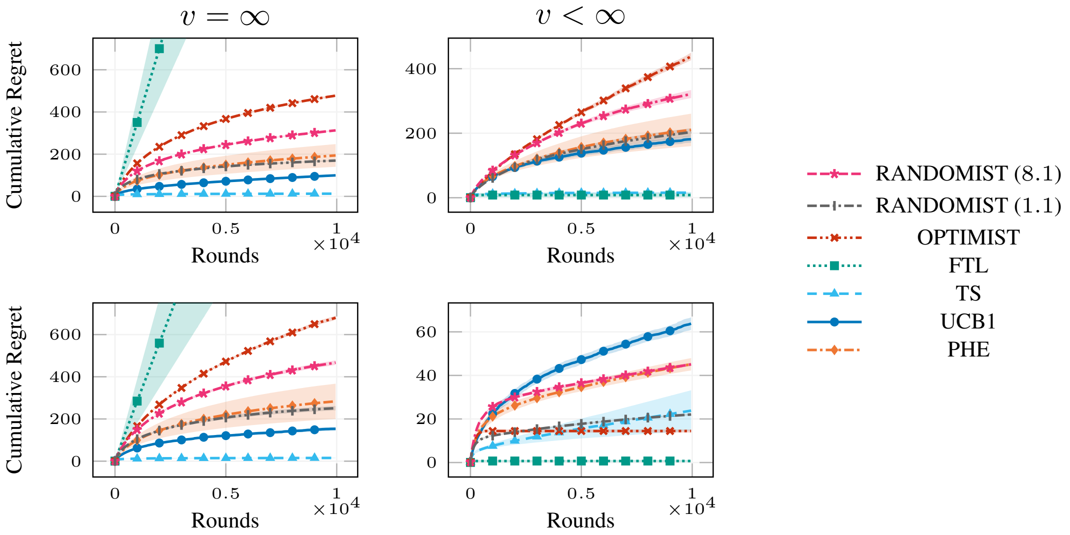
Compact Policy Space

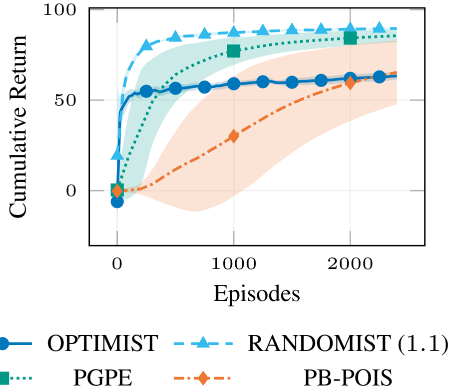
$\Theta = [-D, D]^d$ and $v = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$

Algorithm	Approach	Complexity	$\mathbb{E}\,\mathrm{Regret}(n)$
OPTIMIST (Papini et al., 2019)	discretization	$t^{1+\frac{d}{2}}$	$\mathcal{O}\left(\sqrt{vdn}\right)$
RANDOMIST	MCMC sampling	dt^2	?

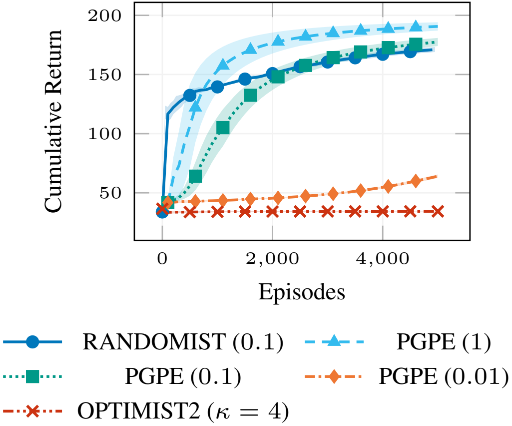
Numerical Simulations

Illustrative Examples ($\mathcal{T} = \mathbb{R}$)





- **Compact** policy space
- $d = 2$
- Parameter-based exploration (Sehnke et al., 2008)
- Gaussian hyperpolicy
- MCMC with Metropolis-Hastings



- **Compact** policy space
- $d = 4$
- OPTIMIST suffers from the discretization

Discussion and Conclusions

Contributions

- Formalization of **mediator feedback** and regret lower bounds
- Novel regret minimization algorithm **RANDOMIST**, its analysis and numerical simulations

Future works

- Improve/Match the lower bound
- Other perturbations for RANDOMIST
- Other applications of **mediator feedback** (e.g., variational inference, Bayesian networks)

Discussion and Conclusions

Contributions

- Formalization of **mediator feedback** and regret lower bounds
- Novel regret minimization algorithm **RANDOMIST**, its analysis and numerical simulations

Future works

- Improve/Match the lower bound
- Other perturbations for RANDOMIST
- Other applications of **mediator feedback** (e.g., variational inference, Bayesian networks)

Thank You for Your Attention!

Paper: `arxiv.org/pdf/2012.08225.pdf`

Code: `github.com/proceduralia/randomist`

Contact: `albertomaria.metelli@polimi.it`

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

References

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

A. Beskos and A. Stuart. Computational complexity of metropolis-hastings methods in high dimensions. *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 61–71, 2009.

M. P. Deisenroth, G. Neumann, and J. Peters. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2013.

C. D'Eramo, A. Nuara, M. Pirotta, and M. Restelli. Estimating the maximum expected value in continuous reinforcement learning problems. In *AAAI*, 2017.

E. L. Ionides. Truncated importance sampling. *JCGS*, 17(2):295–311, 2008.

B. Kveton, C. Szepesvári, M. Ghavamzadeh, and C. Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *IJCAI*, 2019.

A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *NeurIPS*, 2018.

A. M. Metelli, M. Papini, N. Montali, and M. Restelli. Importance sampling techniques for policy optimization. *JMLR*, 21(141):1–75, 2020.

A. B. Owen. Monte carlo theory, methods and examples. *Monte Carlo Theory, Methods and Examples*, 2013.

M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In *ICML*, 2019.

References (cont.)

F. Sehnke, C. Osendorfer, T. Rückstieß, A. Graves, J. Peters, and J. Schmidhuber. Policy gradients with parameter-based exploration for control. In *ICANN*, 2008.

E. Veach and L. J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In S. G. Mair and R. Cook, editors, *SIGGRAPH*, pages 419–428. ACM, 1995.