Learning Optimal Deterministic Policies with Stochastic Policy Gradients

POLITECNICO MILANO 1863

A. Montenegro, M. Mussi, A. M. Metelli, and M. Papini

{alessandro.montenegro, marco.mussi, albertomaria.metelli, matteo.papini}@polimi.it

RL³

Motivation

Why Policy Gradients? Real-world continuous control problems can be successfully tackled via **Stochastic Policy Gradients (PGs)**, by leveraging on **action** or **parameter-based** (**AB**/**PB**) exploration.

Why Deterministic Policies? Real-life artificial agents, especially in safety-critical scenarios, cannot accept stochastic policies, since they do not meet **reliability**, **safety**, and **traceability** standards.

Contributions

Focus Theoretical understanding of **learning** via **stochastic** PGs, then **deploy deterministic** policies:

- Framework for modelling the **AB** and **PB** **noise-injection** w.r.t. deterministic policies;
- **Last-iterate global convergence** of **AB** and **PB** PGs;
- **Exploration amount tuning**: how to optimize the trade-off between the sample complexity and the performance of the deployed deterministic policy;
- **AB** vs **PB**: assumptions and sample complexities.

General Last-Iterate Convergence

Assumptions

A Weak Gradient Domination

$$J_\dagger^* - J_\dagger(\boldsymbol{\theta}) \leqslant \alpha \left\| \nabla_{\boldsymbol{\theta}} J_\dagger(\boldsymbol{\theta}) \right\|_2 + \beta$$

B Smoothness

$$\left\| \nabla_{\boldsymbol{\theta}} J_\dagger(\boldsymbol{\theta}') - \nabla_{\boldsymbol{\theta}} J_\dagger(\boldsymbol{\theta}_2) \right\|_2 \leqslant L_{2,\dagger} \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} \right\|_2$$

C Bounded Variance

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{\theta}} J_\dagger(\boldsymbol{\theta}) \right] \leqslant V_\dagger / N$$

Last-Iterate Global Convergence

$$J_\dagger^* - \mathbb{E}\left[J_\dagger(\boldsymbol{\theta}_K) \right] \leqslant \epsilon + \beta$$

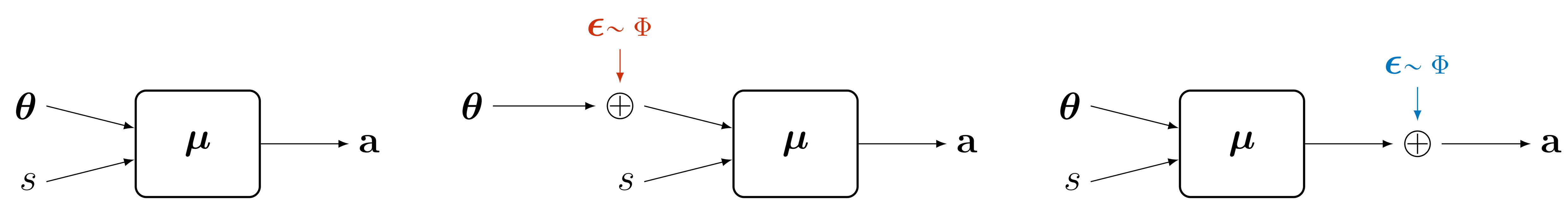
is ensured with a sample complexity

$$NK = \widetilde{\mathcal{O}}(\epsilon^{-3})$$

Setting

Noise-Injection Framework

Deterministic policy $\mu_{\boldsymbol{\theta}}$ — Stochastic hyperpolicy $\nu_{\boldsymbol{\theta}}$ — Stochastic policy $\pi_{\boldsymbol{\theta}}$



Performance Indices

$$J_{\mathrm{D}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{D}}(\cdot|\boldsymbol{\theta})}[R(\tau)] \qquad J_{\mathrm{P}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\boldsymbol{\theta}' \sim \nu_{\boldsymbol{\theta}}}\left[J_{\mathrm{D}}(\boldsymbol{\theta}') \right] \qquad J_{\mathrm{A}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}[R(\tau)]$$

Estimators

General Update $\quad \boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \zeta_t \widehat{\nabla}_{\boldsymbol{\theta}} J_\dagger(\boldsymbol{\theta}_t)$

$$\widehat{\nabla}_{\boldsymbol{\theta}} J_{\mathrm{P}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \log \nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) R(\tau_i)$$

$$\widehat{\nabla}_{\boldsymbol{\theta}} J_{\mathrm{A}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \left(\sum_{k=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{\tau_i,k} | \mathbf{s}_{\tau_i,k}) \right) \gamma^t r(\mathbf{s}_{\tau_i,t}, \mathbf{a}_{\tau_i,t})$$

Last-Iterate Convergence to Optimal Deterministic Policies

Assumptions | (1) Weak Gradient Domination (2) Lipschitz MDP (3) Lipschitz $\mu_{\boldsymbol{\theta}}$ (4) Bounded Noise Scores

PB Exploration

Generic σ_{P}

$$J_{\mathrm{D}}^* - \mathbb{E}\left[J_{\mathrm{D}}(\boldsymbol{\theta}_K) \right] \leqslant \epsilon + \beta + 3\sigma_{\mathrm{P}} L_P \sqrt{d_\Theta}$$

Asm. 1-4	$NK = \widetilde{\mathcal{O}}(\epsilon^{-3}\sigma_{\mathrm{P}}^{-4}(1-\gamma)^{-4}d_\Theta^2)$
+ Smooth MDP/$\mu_{\boldsymbol{\theta}}$	$NK = \widetilde{\mathcal{O}}(\epsilon^{-3}\sigma_{\mathrm{P}}^{-2}(1-\gamma)^{-5}d_\Theta)$

$\sigma_{\mathrm{P}} = \mathcal{O}(\epsilon(1-\gamma)^{-2}d_\Theta^{-1/2})$

$$J_{\mathrm{D}}^* - \mathbb{E}\left[J_{\mathrm{D}}(\boldsymbol{\theta}_K) \right] \leqslant \epsilon + \beta$$

Asm. 1-4	$NK = \widetilde{\mathcal{O}}(\epsilon^{-7}(1-\gamma)^{-12}d_\Theta^4)$
+ Smooth MDP/$\mu_{\boldsymbol{\theta}}$	$NK = \widetilde{\mathcal{O}}(\epsilon^{-5}(1-\gamma)^{-9}d_\Theta^2)$

AB Exploration

Generic σ_{A}

$$J_{\mathrm{D}}^* - \mathbb{E}\left[J_{\mathrm{D}}(\boldsymbol{\theta}_K) \right] \leqslant \epsilon + \beta + 3\sigma_{\mathrm{A}} L_A \sqrt{d_{\mathcal{A}}}$$

Asm. 1-4 + Smooth $\mu_{\boldsymbol{\theta}}$	$NK = \widetilde{\mathcal{O}}(\epsilon^{-3}\sigma_{\mathrm{A}}^{-4}(1-\gamma)^{-5}d_{\mathcal{A}}^2)$
+ Smooth MDP	$NK = \widetilde{\mathcal{O}}(\epsilon^{-3}\sigma_{\mathrm{A}}^{-2}(1-\gamma)^{-6}d_{\mathcal{A}})$

$\sigma_{\mathrm{A}} = \mathcal{O}(\epsilon(1-\gamma)^{-2}d_{\mathcal{A}}^{-1/2})$

$$J_{\mathrm{D}}^* - \mathbb{E}\left[J_{\mathrm{D}}(\boldsymbol{\theta}_K) \right] \leqslant \epsilon + \beta$$

Asm. 1-4 + Smooth $\mu_{\boldsymbol{\theta}}$	$NK = \widetilde{\mathcal{O}}(\epsilon^{-7}(1-\gamma)^{-13}d_{\mathcal{A}}^4)$
+ Smooth MDP	$NK = \widetilde{\mathcal{O}}(\epsilon^{-5}(1-\gamma)^{-10}d_{\mathcal{A}}^2)$

Deterministic Deploying Losses

PB and AB Solutions

$$\boldsymbol{\theta}_{\mathrm{P}}^* \in \arg\max_{\boldsymbol{\theta}\in\Theta} J_{\mathrm{P}}(\boldsymbol{\theta}) \qquad \boldsymbol{\theta}_{\mathrm{A}}^* \in \arg\max_{\boldsymbol{\theta}\in\Theta} J_{\mathrm{A}}(\boldsymbol{\theta})$$

Deterministic Deployment Bounds

Uniform Bound

$$|J_{\mathrm{D}}(\boldsymbol{\theta}) - J_{\mathrm{P}}(\boldsymbol{\theta})| \leqslant L_J \sqrt{d_\Theta}\sigma_{\mathrm{P}}$$
$$|J_{\mathrm{D}}(\boldsymbol{\theta}) - J_{\mathrm{A}}(\boldsymbol{\theta})| \leqslant L \sqrt{d_{\mathcal{A}}}\sigma_{\mathrm{A}}$$

Upper Bound

$$J_{\mathrm{D}}^* - J_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{P}}^*) \leqslant 2L_J \sqrt{d_\Theta}\sigma_{\mathrm{P}}$$
$$J_{\mathrm{D}}^* - J_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{A}}^*) \leqslant 2L \sqrt{d_{\mathcal{A}}}\sigma_{\mathrm{A}}$$

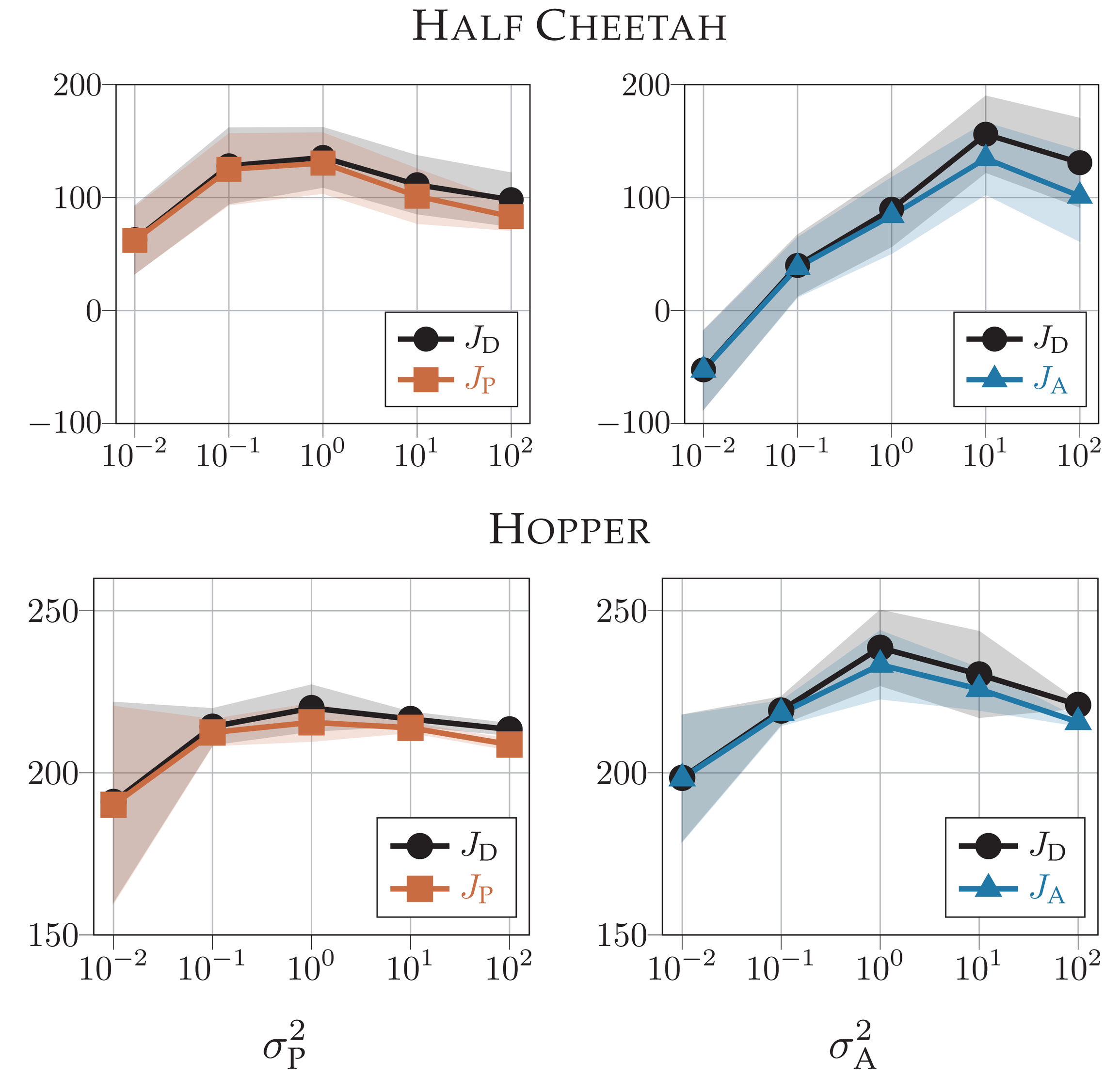
Lower Bound

$$J_{\mathrm{D}}^* - J_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{P}}^*) \geqslant 0.28 L_J \sqrt{d_\Theta}\sigma_{\mathrm{P}}$$
$$J_{\mathrm{D}}^* - J_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{A}}^*) \geqslant 0.28 L \sqrt{d_{\mathcal{A}}}\sigma_{\mathrm{A}}$$

Experimental Validation

Deterministic Deployment

Last-Iterate Performance: PB vs. AB



References

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *JAIR*, 2001.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 2010.

Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *AISTATS*, 2022.