Convergence Analysis of Policy Gradient Methods with Dynamic Stochasticity

A. Montenegro, M. Mussi, M. Papini, and A. M. Metelli
{alessandro.montenegro, marco.mussi, matteo.papini, albertomaria.metelli}@polimi.it

Motivation and Contribution

Why Policy Gradients?
They successfully tackle real-world **continuous control problems**.

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

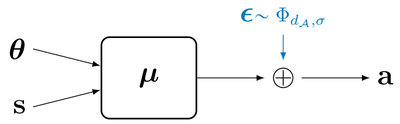
Why Dynamic Stochasticity?
Theorists consider it **fixed** to show convergence, but to deploy deterministic policies it has to be set small. **Practitioners learn it, not** caring of **convergence**.

PES: Decrease the Stochasticity.
Convergence to the optimal **deterministic** policy in $\tilde{\mathcal{O}}(\epsilon^{-5})$, as if keeping a small fixed stochasticity!

SL-PG: Learn the Stochasticity.
Learn a stochasticity mapping in $[\sigma_{\min}, \sigma_{\max}]$. Convergence to the optimal **stochastic** policy in $\tilde{\mathcal{O}}(\epsilon^{-3}\sigma_{\min}^{-2})$, but stronger assumptions are required!

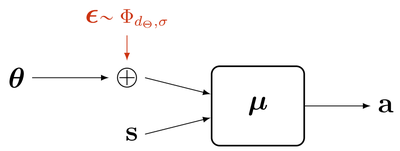
AB and PB Exploration

Action-Based Exploration



$$J_{\mathrm{A}}(\boldsymbol{\theta}, \sigma) = \mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta},\sigma)} [R(\tau)]$$

Parameter-Based Exploration



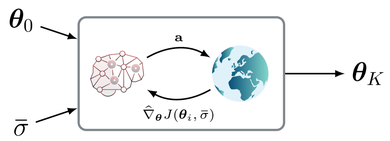
$$J_{\mathrm{P}}(\boldsymbol{\theta}, \sigma) = \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim \nu_{\boldsymbol{\rho},\sigma}} \left[\mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{D}}(\cdot|\boldsymbol{\theta})} [R(\tau)] \right]$$

Convergence: Fixed Stochasticity vs. Learning the Stochasticity

Conditions for Convergence

1. Weak Gradient Domination: $J(\boldsymbol{\theta}^*(\sigma), \sigma) - J(\boldsymbol{\theta}, \sigma) \leqslant \alpha \|\nabla J(\boldsymbol{\theta}, \sigma)\|_2 + \beta$
2. Objective Smoothness: $\|\nabla^2 J(\boldsymbol{\theta}, \sigma)\|_2 \leqslant L_2$
3. Bounded Variance: $\mathbb{V}\mathrm{ar}[\widehat{\nabla} J(\boldsymbol{\theta}, \sigma)] \leqslant V N^{-1} \sigma^{-2}$

Fixed Stochasticity



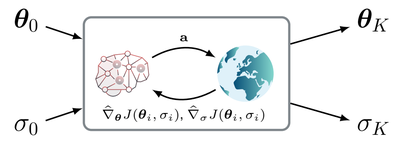
Last-Iterate Convergence

$$J(\boldsymbol{\theta}^*(\sigma), \sigma) - \mathbb{E}[J(\boldsymbol{\theta}_K, \sigma)] \leqslant \epsilon + \beta$$
with
$$NK = \tilde{\mathcal{O}}(\epsilon^{-3}\sigma^{-2})$$

✓ $\sigma = \epsilon \implies$ Deterministic policies in $NK = \tilde{\mathcal{O}}(\epsilon^{-5})$

✗ σ has to be selected, it affects $J(\cdot)$

Learning Stochasticity



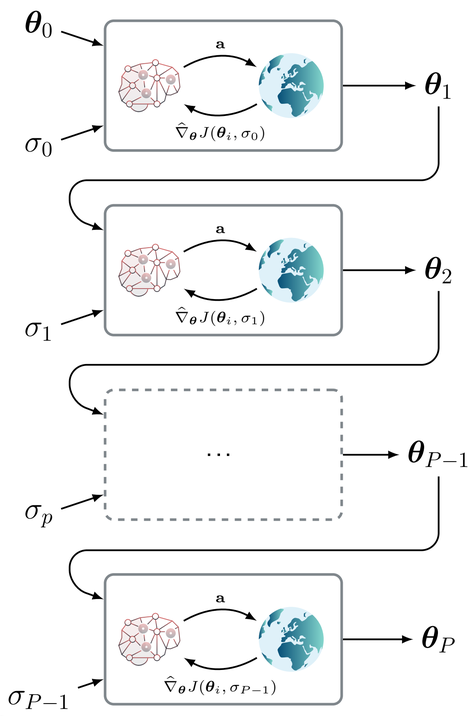
Last-Iterate Convergence

$$J^* - \mathbb{E}[J(\boldsymbol{\theta}_K, \sigma_K)] \leqslant \epsilon + \beta$$
with
$$NK = \tilde{\mathcal{O}}(\epsilon^{-3}\sigma_{\min}^{-2})$$

✓ Select $\bar{\sigma}$, then σ is learned to maximize $J(\cdot)$

✗ Final σ is not controllable, no guarantees on deterministic deployment

PES: Convergence Under a Scheduled Decreasing Stochasticity



Stochasticity Schedule

$$\sigma_p = \sigma_{\max} (p+1)^{-y}$$

Assumptions

1. $(\alpha_{\mathrm{D}}, \beta_{\mathrm{D}})$-WGD on J_{D}
2. J_{D} Smoothness
3. $\mu_{\boldsymbol{\theta}}$ and MDP Regularity
4. Bounded Noise Score

$\implies$ ✓ Phases' Convergence

Last-Iterate Convergence

Number of Phases $P = (\epsilon/\sigma_{\max})^{-1/y}$

$$J_{\mathrm{D}}^* - \mathbb{E}[J_{\mathrm{D}}(\boldsymbol{\theta}_{P-1})] \leqslant \mathcal{O}(\epsilon) + \beta_{\mathrm{D}}$$
with
$$NK = \tilde{\mathcal{O}}(\epsilon^{-5})$$

Switching Stochasticity

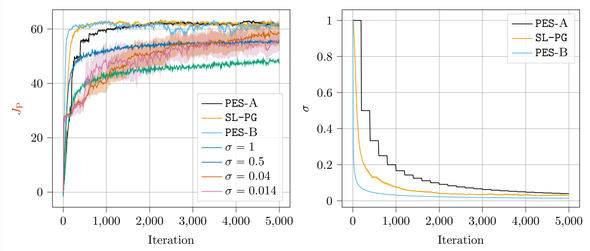
What happens when the stochasticity σ changes?

1. Stochasticity Change: $|J(\boldsymbol{\theta}, \sigma_1) - J(\boldsymbol{\theta}, \sigma_2)| \leqslant D|\sigma_1 - \sigma_2|$
2. Optimum Plugging Loss: $J^*(\sigma_1) - J(\boldsymbol{\theta}^*(\sigma_2), \sigma_1) \leqslant 2D|\sigma_1 - \sigma_2|$
3. Distance Between Optima: $|J^*(\sigma_1) - J^*(\sigma_2)| \leqslant D|\sigma_1 - \sigma_2|$

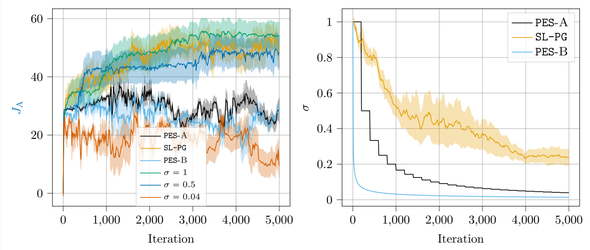
where $D_{\mathrm{A}} = \mathcal{O}(T\sqrt{d_{\mathcal{A}}})$ and $D_{\mathrm{P}} = \mathcal{O}(\sqrt{d_\Theta})$

Does it Work?

Parameter-based Exploration



Action-based Exploration



Deterministic Deployment

Method	PB	Δ_{P}	AB	Δ_{A}
PES-A	61.14 ± 1.1	-0.12	28.93 ± 9.89	0.44
SL-PG	61.43 ± 2.73	-0.21	49.83 ± 13.38	-1.19
PES-B	60.40 ± 2.31	-0.27	29.24 ± 8.74	0.33

References

J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *JAIR*, 2001.

A. Montenegro, M. Mussi, A. M. Metelli, and M. Papini. Learning optimal deterministic policies with stochastic policy gradients. *ICML*, 2024.