Safe Exploration in Gaussian Policy Gradient

Matteo Papini, Andrea Battistello and Marcello Restelli
matteo.papini@polimi.it, andrea.battistello1994@gmail.com, marcello.restelli@polimi.it
POLITECNICO MILANO 1863

Problem

- **Reinforcement Learning** for **continuous** control [Deisenroth et al., 2013]
- **Policy Gradient (PG)**: iteratively update **parametric** policy $\pi_{\boldsymbol{\theta}}$ via **gradient ascent** on performance $J(\boldsymbol{\theta})$ (expected cumulative reward):

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t)$$

- Convergence to local optimum guaranteed
- Intermediate policies may be arbitrarily bad!
- **Safe Exploration** [Amodei et al., 2016]: limit risks/costs of exploratory behavior

Motivation

- A working controller is provided as initial policy
- **Fine tuning**: improve it *online* via policy gradient
- Intermediate policies should never be (too much) worse than the initial one

State of the Art

Existing **safe PG** approaches [Pirotta et al., 2013, Papini et al., 2017, 2019]:

- Apply only to **Gaussian** policies with fixed variance:

$$\pi_{\boldsymbol{\theta}}(a|s) \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}(s), \sigma^2)$$

⚠ The variance parameter regulates exploration and has a big impact on convergence speed

- Focus on **monotonic improvement** guarantees:

$$J(\boldsymbol{\theta}_{t+1}) - J(\boldsymbol{\theta}_t) \geq 0$$

⚠ Very strict: exploration is totally sacrificed due to its immediate costs

Contributions

- We adopt a more general definition of safety that leaves room for exploration
- We propose a surrogate objective for the policy variance that takes the long-term benefits of exploration into account
- We extend the existing improvement guarantees for Gaussian policies to the adaptive-variance case

Setting

- *Shallow* Gaussian policy parametrization:

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sqrt{2\pi}\sigma_\omega} \exp\left\{-\frac{1}{2}\left(\frac{a-\mu_{\boldsymbol{v}}(s)}{\sigma_\omega}\right)^2\right\} \qquad \mu_{\boldsymbol{v}}(s) = \boldsymbol{v}^T\boldsymbol{\phi}(s), \quad \sigma_\omega = e^\omega \qquad \boldsymbol{\theta} = \begin{cases} \boldsymbol{v} & \text{mean parameter} \\ \omega & \text{variance parameter} \end{cases}$$

- **Safety requirement** (similar to Thomas et al. [2015]):

$$J(\boldsymbol{\theta}_{t+1}) - J(\boldsymbol{\theta}_t) \geq C_t \quad \text{with probability at least } 1-\delta \qquad \begin{array}{ll} C_t \geq 0: & \textit{required improvement} \\ C_t < 0: & \textit{bounded worsening} \end{array}$$

- Base algorithm: (normalized) REINFORCE with separate mean and variance updates

$$\begin{cases} \boldsymbol{v}_{t+1} \leftarrow \boldsymbol{v}_t + \alpha_t \nabla_{\boldsymbol{v}} J(\boldsymbol{v}_t, \omega_t) / \|\nabla_{\boldsymbol{v}} J(\boldsymbol{v}_t, \omega_t)\| \\ \omega_{t+1} \leftarrow \omega_t + \eta_t \nabla_\omega J(\boldsymbol{v}_{t+1}, \omega_t) / \|\nabla_\omega J(\boldsymbol{v}_{t+1}, \omega_t)\| \end{cases} \quad \text{naive update: too greedy!}$$

- **Adaptive PG**: we look for the *largest* step sizes satisfying the requirement at each iteration

Adaptive Exploration (Heuristic)

- We make α_t proportional to σ^2 to exploit its **smoothing effect** [Ahmed et al., 2019]:

$$\alpha_t = \alpha\sigma^2_{\omega_t}$$

- We introduce a **surrogate objective** for ω that accounts for the long-term advantages of exploration:

$$\mathcal{L}(\boldsymbol{v}_t, \omega_t) = \underbrace{J(\boldsymbol{v}_{t+1})}_{\text{performance after next mean update}} \simeq J(\boldsymbol{v}_t, \omega_t) + \alpha\sigma^2_{\omega_t}\|\nabla_{\boldsymbol{v}} J(\boldsymbol{v}_t, \omega_t)\|$$

Meta-Exploring Policy Gradient (MEPG):

$$\boldsymbol{v}_{t+1} \leftarrow \boldsymbol{v}_t + \alpha\sigma^2_{\omega_t}\nabla_{\boldsymbol{v}} J(\boldsymbol{v}_t, \omega_t) / \|\nabla_{\boldsymbol{v}} J(\boldsymbol{v}_t, \omega_t)\|, \qquad \omega_{t+1} \leftarrow \omega_t + \eta \underbrace{\nabla_\omega \mathcal{L}(\boldsymbol{v}_{t+1}, \omega_t)}_{\text{meta-gradient}} / \|\nabla_\omega \mathcal{L}(\boldsymbol{v}_{t+1}, \omega_t)\|$$

- Learning behavior still depends on **hyperparameters** α, η (step sizes)

Safe Exploration

- We extend improvement guarantees [Papini et al., 2019] for Gaussian policies to the adaptive-variance setting to obtain adaptive **safe step sizes** for **MEPG**:

$$\begin{cases} \alpha_t = \dfrac{\sigma^2_{\omega_t}}{F}\left(1 + \sqrt{1 - C_t/C_t^*}\right) \\ \eta_t = \dfrac{|\lambda_t|}{G}\left(sign(\lambda_t) + \sqrt{1 - C_t/C_t^*}\right) \end{cases}$$

- $\lambda_t = \dfrac{\nabla_\omega J(\boldsymbol{v}_t, \omega_t)^T \nabla_\omega \mathcal{L}(\boldsymbol{v}_t, \omega_t)}{\|\nabla_\omega \mathcal{L}(\boldsymbol{v}_t, \omega_t)\|}$ (scalar projection)
- F, G: smoothing constants $\mathcal{O}\left(R_{\max}\phi^2_{\max}(1-\gamma)^{-3}\right)$
- C_t^*: maximum ensurable improvement ($C_t \leq C_t^*$)
- η_t can be negative when exploration constrasts with immediate improvement (depending on C_t)

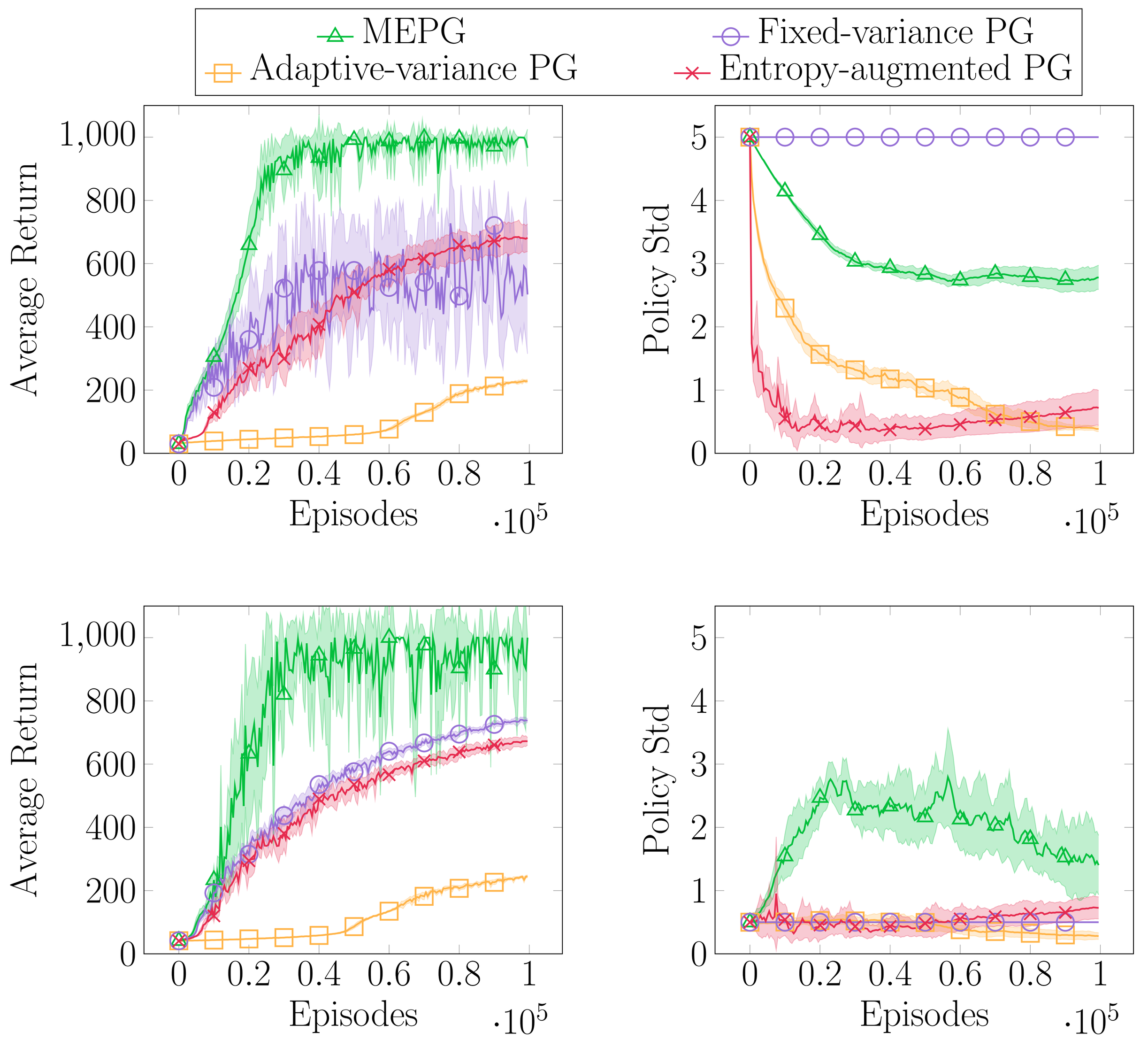
- The resulting **SEPG algorithm (Safely Exploring Policy Gradient)** guarantees $J(\boldsymbol{v}_{t+1}, \omega_{t+1}) - J(\boldsymbol{v}_t, \omega_t) \geq C_t$
- We devise high-probability variants for the approximate setting

Special Requirements

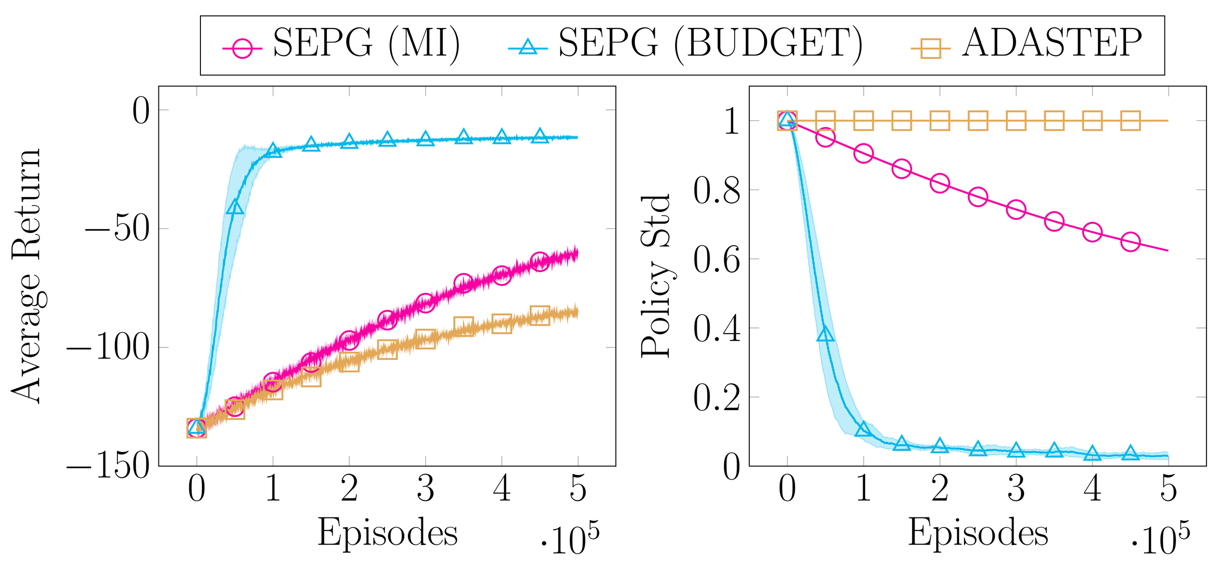
- **Monotonic improvement (MI)**: $C_t \equiv 0$
- **Fixed improvement**: constant C_t (possibly < 0)
- **Fixed threshold**: $C_t = J_{\min} - J(\boldsymbol{\theta}_t)$
- **Fine tuning (BUDGET)**: $C_t = J(\boldsymbol{\theta}_0) - J(\boldsymbol{\theta}_t)$

Experiments

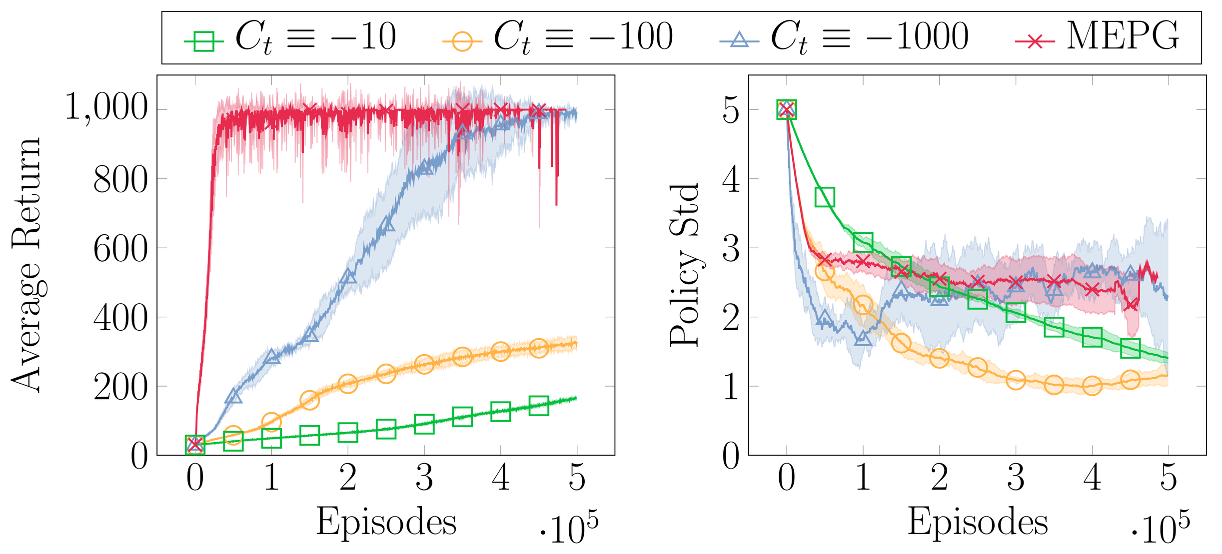
MEPG on Cart-Pole



SEPG on LQG



SEPG on Cart-Pole



References

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *ICML*, 2019.
Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. 2016.
Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.
Matteo Papini, Matteo Pirotta, and Marcello Restelli. Adaptive batch size for safe policy gradients. In *NeurIPS*, 2017.
Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *arXiv preprint arXiv:1905.03231*, 2019.
Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *NeurIPS*, pages 1394–1402, 2013.
Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *ICML*, 2015.