Adaptive Batch Size for Safe Policy Gradients

M. Papini, M. Pirotta and M. Restelli

{matteo.papini, marcello.restelli}@polimi.it
{matteo.pirotta}@inria.fr

POLITECNICO MILANO 1863

Inria informatics mathematics

Problem

- **Monotonically** improve a parametric **gaussian** policy $\pi_{\boldsymbol{\theta}}$ in a **continuous** MDP, avoiding unsafe **oscillations** in the expected performance $J(\boldsymbol{\theta})$.
- Episodic Policy Gradient:
  - estimate $\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ from a **batch** of N sample trajectories.
  - $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- Tune step size α and batch size N to limit oscillations. **Not trivial**:
  - α: **trade-off** with speed of convergence $\leftarrow$ adaptive methods.
  - N: **trade-off** with total learning time $\leftarrow$ typically tuned by hand.
- **Lack** of cost sensitive solutions.

Contributions

1. We propose a per-component adaptive step size Λ which results in a greedy **coordinate descent** algorithm, improving over existing adaptive step-size methods.
2. We show a **duality** in the role played by Λ and N in maximizing the performance improvement $J(\boldsymbol{\theta}') - J(\boldsymbol{\theta})$ and how a **joint optimization** of the two meta-parameters can guarantee monotonic improvement with high probability.
3. We make a first step in the development of **practical methods** to jointly optimize the step size and the batch size.
4. We offer a preliminary **empirical evaluation** of the proposed methods on a simple control problem.

Non-scalar Adaptive Step Size

Formulation: $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \Lambda \nabla_{\boldsymbol{\theta}} J_{\mu}(\boldsymbol{\theta})$, $\Lambda = \mathrm{diag}(\alpha_1, \alpha_2, \dots, \alpha_m) \geq 0$

Goal: $\Lambda^* = \max_{\Lambda} J(\boldsymbol{\theta}') - J(\boldsymbol{\theta})$

Exact Framework

Optimal step size:

$$\alpha_k^* = \begin{cases} \frac{1}{2c} & \text{if } k = \min\left\{\arg\max_i |\nabla_{\theta_i} J_{\mu}(\boldsymbol{\theta})|\right\}, \\ 0 & \text{otherwise} \end{cases}$$

Improvement guarantee:

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq \frac{\|\nabla_{\boldsymbol{\theta}} J_{\mu}(\boldsymbol{\theta})\|_{\infty}^2}{4c}$$

$$c = \frac{RM_{\phi}^2}{(1-\gamma)^2\sigma^2}\left(\frac{|\mathcal{A}|}{\sqrt{2\pi}\sigma} + \frac{\gamma}{2(1-\gamma)}\right)$$

Approximate Framework

Optimal step size:

$$\alpha_k^* = \begin{cases} \dfrac{\left(\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty} - \epsilon\right)^2}{2c\left(\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty} + \epsilon\right)^2} & \text{if } k = \min\left\{\arg\max_i |\hat{\nabla}_{\theta_i} J(\boldsymbol{\theta})|\right\}, \\ 0 & \text{otherwise} \end{cases}$$

Improvement guarantee:

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq \frac{\left(\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty} - \epsilon\right)^4}{4c\left(\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty} + \epsilon\right)^2}.$$

Adaptive Batch Size

Goal: **Cost-sensitive** joint optimization: $\Lambda^*, N^* = \arg\max_{\Lambda, N} \frac{J(\boldsymbol{\theta}') - J(\boldsymbol{\theta})}{N}$

Chebyshev-like Bounds

Error bound: $\epsilon \leq \frac{d_\delta}{\sqrt{N}}$ with probability $(1-\delta)$

Optimal meta-parameters:

$$\alpha_k^* = \begin{cases} \frac{(13 - 3\sqrt{17})}{4c} & \text{if } k = \min\left\{\arg\max_i |\hat{\nabla}_{\theta_i} J(\boldsymbol{\theta})|\right\} \\ 0 & \text{otherwise} \end{cases} \qquad N^* = \left\lceil \frac{(13+3\sqrt{17})d_\delta^2}{2\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty}^2} \right\rceil,$$

	Chebyshev	Hoeddfing	Empirical Bernstein (Mnih et al., 2008)
d_δ	$\sqrt{\frac{Var[\hat{\nabla}_{\theta_i} J(\boldsymbol{\theta})]}{\delta}}$	$R\sqrt{\frac{\log 2/\delta}{2}}$	$\sqrt{2S_N \ln 3/\delta}$
f_δ	$\times$	$\times$	$3\mathbf{R} \ln 3/\delta$

Bernstein-like Bounds

Error bound: $\epsilon \leq \frac{d_\delta}{\sqrt{N}} + \frac{f_\delta}{N}$ with probability $1-\delta$

Optimal meta-parameters:

N^* has no practical closed-form solution: we suggest to find it with a linear search, knowing that:

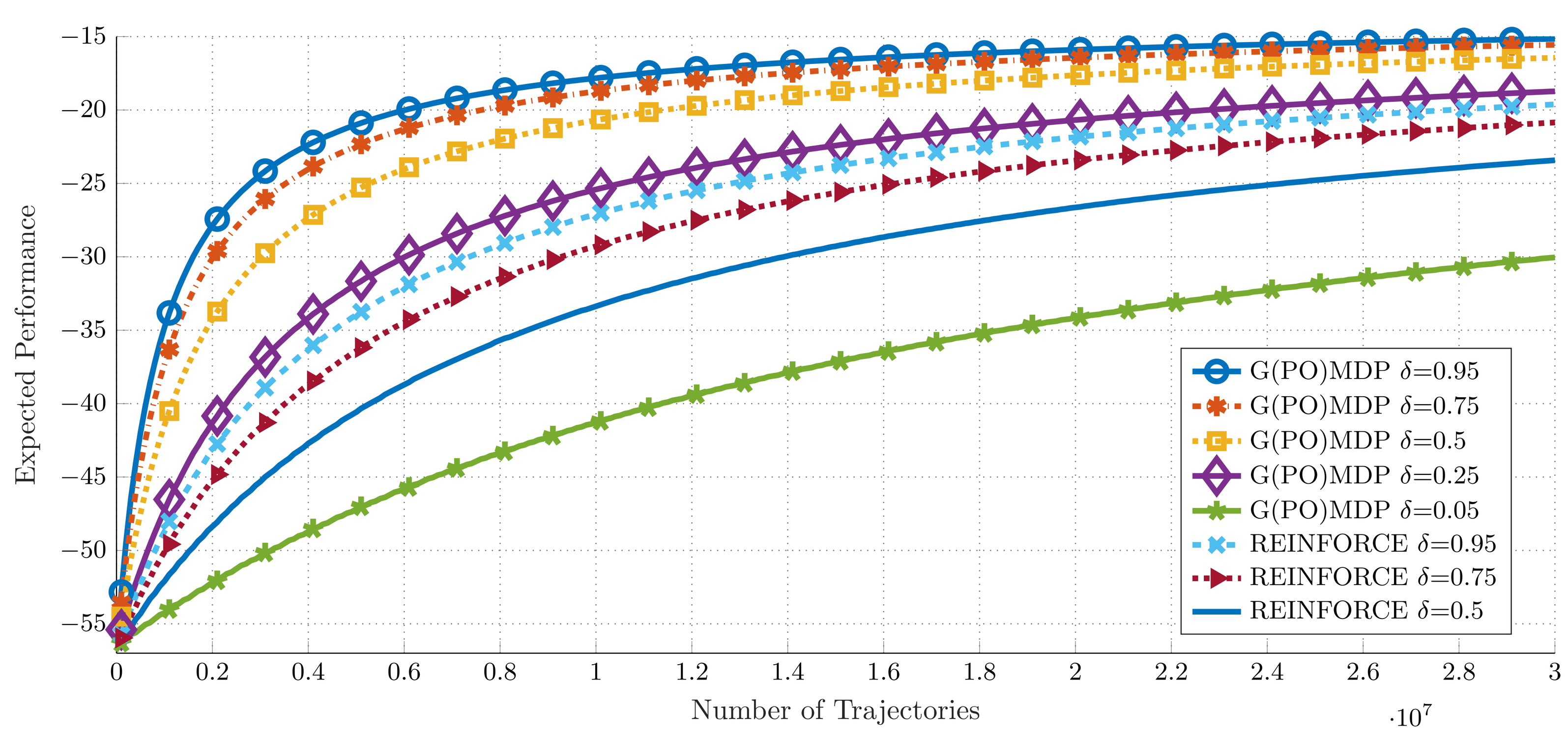
- $N^* \geq N_0 \triangleq \left(\dfrac{d_\delta + \sqrt{d_\delta^2 + 4f_\delta\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty}}}{2\left\|\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty}}\right)^2$
- the cost-sensitive objective is concave above N_0,

then compute Λ^* from $\epsilon(N^*)$

Empirical Results

One-Dimensional LQG

Comparing gradient estimation algorithms and values of δ



Comparing statistical bounds (using G(PO)MDP and $\delta = 0.95$)