ADAPTIVE BATCH SIZE FOR SAFE POLICY GRADIENTS

POLITECNICO MILANO 1863

M. PAPINI, M. PIROTTA AND M. RESTELLI

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

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 $\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ from a **batch** of N sample trajectories.
  - $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \Lambda \widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- Tune step size α and batch size N to limit oscillations. **Not trivial**:
  - Λ: **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 *safe* 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

LOWER BOUND TO POLICY PERFORMANCE: [Pirotta et al., 2013]

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq \underbrace{\frac{1}{1-\gamma} \int_{\mathcal{S}} d_{\mu}^{\pi_{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \left(\pi_{\boldsymbol{\theta}}(a|s) - \pi_{\boldsymbol{\theta}'}(a|s)\right) Q^{\pi_{\boldsymbol{\theta}}}(s,a) \mathrm{d}a \mathrm{d}s}_{\textit{weighted advantage}}$$

$$- \underbrace{\frac{\gamma}{2(1-\gamma)^2} \left\|\pi_{\boldsymbol{\theta}'} - \pi_{\boldsymbol{\theta}}\right\|_{\infty}^2 \left\|Q^{\pi_{\boldsymbol{\theta}}}\right\|_{\infty}}_{\textit{state distribution error}} = B_L(\boldsymbol{\theta}', \boldsymbol{\theta})$$

Goal

$$\boxed{\Delta\boldsymbol{\theta}^* \in \arg\max_{\Delta\boldsymbol{\theta}} B_L(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}, \boldsymbol{\theta})}$$

- **Gaussian policy:** $\pi_{\boldsymbol{\theta}} \sim \mathcal{N}(\phi(s)^{\mathsf{T}}\boldsymbol{\theta}, \sigma^2)$
- **Gradient update:** $\Delta\boldsymbol{\theta} = \Lambda \nabla_{\boldsymbol{\theta}} J_{\mu}(\boldsymbol{\theta})$ with $\Lambda = \mathrm{diag}(\alpha_1, \alpha_2, \dots, \alpha_m) \geq 0$

If $\Lambda = \lambda I \rightarrow$ [Pirotta et al. 2013]

SOLUTION: *coordinate ascent*

EXACT FRAMEWORK

Optimal step size:

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

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

Improvement guarantee: $J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq (4c)^{-1} \left\|\nabla_{\boldsymbol{\theta}} J_{\mu}(\boldsymbol{\theta})\right\|_{\infty}^2$

APPROXIMATE FRAMEWORK

Given a policy **gradient estimate** $\widehat{\nabla}_{\boldsymbol{\theta}} J_{\mu}(\boldsymbol{\theta})$ s.t. $P\left(\left|\nabla_{\theta_i} J_{\mu}(\boldsymbol{\theta}) - \widehat{\nabla}_{\theta_i} J_{\mu}(\boldsymbol{\theta})\right| \geq \epsilon_i(N)\right) \leq \delta$

Optimal step size:

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

Improvement guarantee: $J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq \left(\left\|\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty} - \epsilon\right)^4 (4c)^{-1} \left(\left\|\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty} + \epsilon\right)^{-2}$

ADAPTIVE BATCH SIZE

IDEA:
- There are evidences that it is possible to adapt the batch size instead of the step length [Pirotta and Restelli, 2016, Bollapragada et al., 2017, Smith et al. 2017]
- In particular, in RL the cost of collecting new samples may be huge
- Small step size $\rightarrow$ lot of parameter update $\rightarrow$ high costs

Cost-sensitive joint optimization

$$\boxed{\{\Lambda^*, N^*\} \in \arg\max_{\Lambda, N} \frac{B_{\delta}(\Lambda, N)}{N}}$$

e.g., bound on the estimation error with N samples (see approx. framework)

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

CHEBYSHEV-LIKE BOUNDS

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

Optimal meta-parameters:

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

BERNSTEIN-LIKE BOUNDS

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

Optimal meta-parameters:

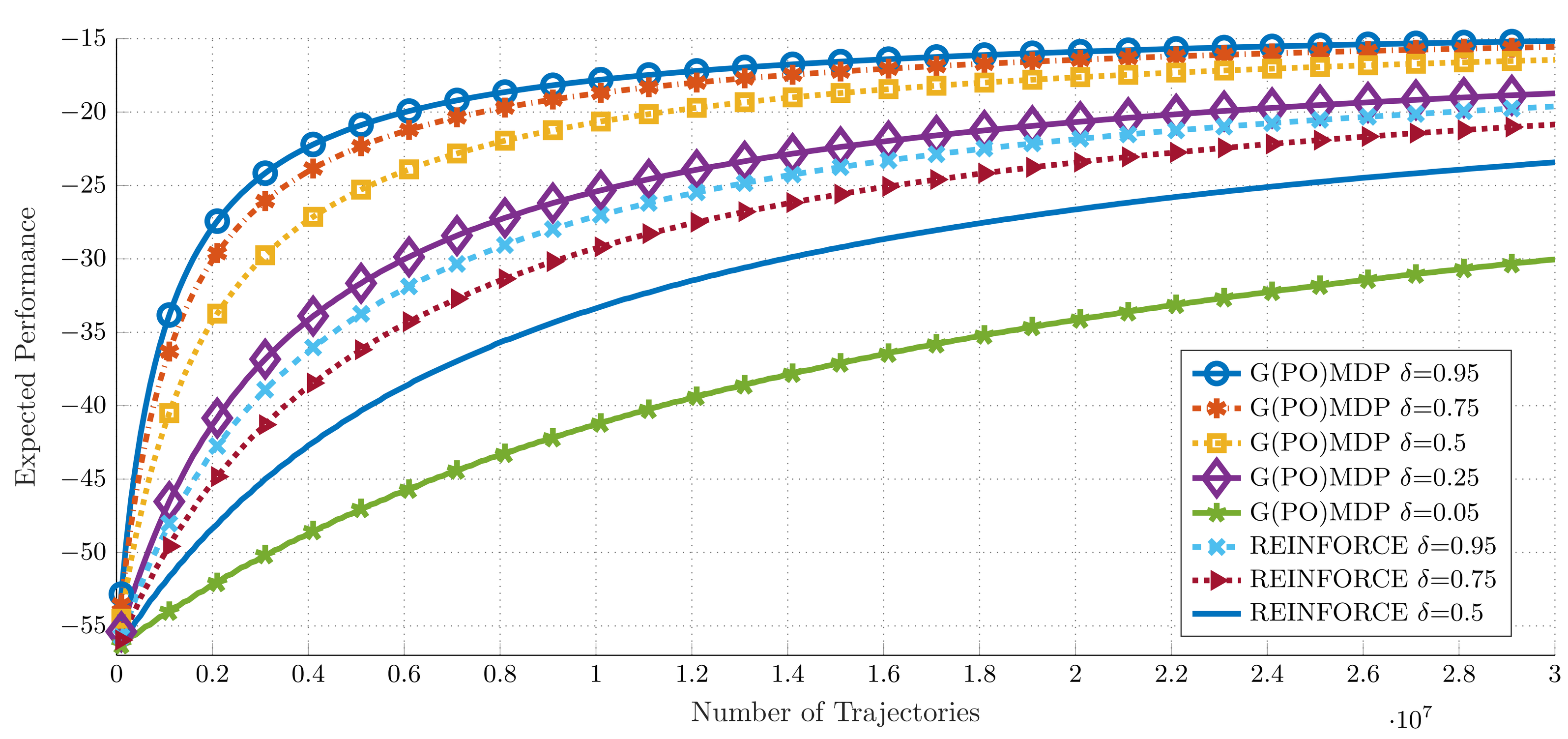
N^* has **no practical closed-form solution**: we **suggest** to find it with a linear search, knowing that:

- $N^* \geq N_0 \triangleq \left(d_{\delta} + \sqrt{d_{\delta}^2 + 4 f_{\delta} \left\|\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty}}\right)^2 2^{-2} \left\|\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right\|_{\infty}^{-2}$
- 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 $\delta = 0.95$)