



POLITECNICO
MILANO 1863

ADAPTIVE BATCH SIZE FOR SAFE POLICY GRADIENTS

M. PAPINI, M. PIROTTA AND M. RESTELLI

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



PROBLEM

- **Monotonically** improve a parametric **gaussian** policy π_θ in a **continuous** MDP, avoiding unsafe **oscillations** in the expected performance $J(\theta)$.
- Episodic Policy Gradient:
 - estimate $\widehat{\nabla}_\theta J(\theta)$ from a **batch** of N sample trajectories.
 - $\theta' \leftarrow \theta + \Lambda \widehat{\nabla}_\theta J(\theta)$
- 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 $J(\theta') - J(\theta)$ 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: [Pirota et al., 2013]

$$J(\theta') - J(\theta) \geq \underbrace{\frac{1}{1-\gamma} \int_{\mathcal{S}} d_{\mu}^{\pi_\theta}(s) \int_{\mathcal{A}} (\pi_\theta(a|s) - \pi_{\theta'}(a|s)) Q^{\pi_\theta}(s, a) da ds}_{\text{weighted advantage}} - \underbrace{\frac{\gamma}{2(1-\gamma)^2} \|\pi_{\theta'} - \pi_\theta\|_\infty^2 \|Q^{\pi_\theta}\|_\infty}_{\text{state distribution error}} = B_L(\theta', \theta)$$

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

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

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

SOLUTION: coordinate ascent

EXACT FRAMEWORK

Optimal step size:

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

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

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

APPROXIMATE FRAMEWORK

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

Optimal step size:

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

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

ADAPTIVE BATCH SIZE

- IDEA:**
- There are evidences that it is possible to adapt the batch size instead of the step length [Pirota 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

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

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

CHEBYSHEV-LIKE BOUNDS

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

Optimal meta-parameters:

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

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

BERNSTEIN-LIKE BOUNDS

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

Optimal meta-parameters:

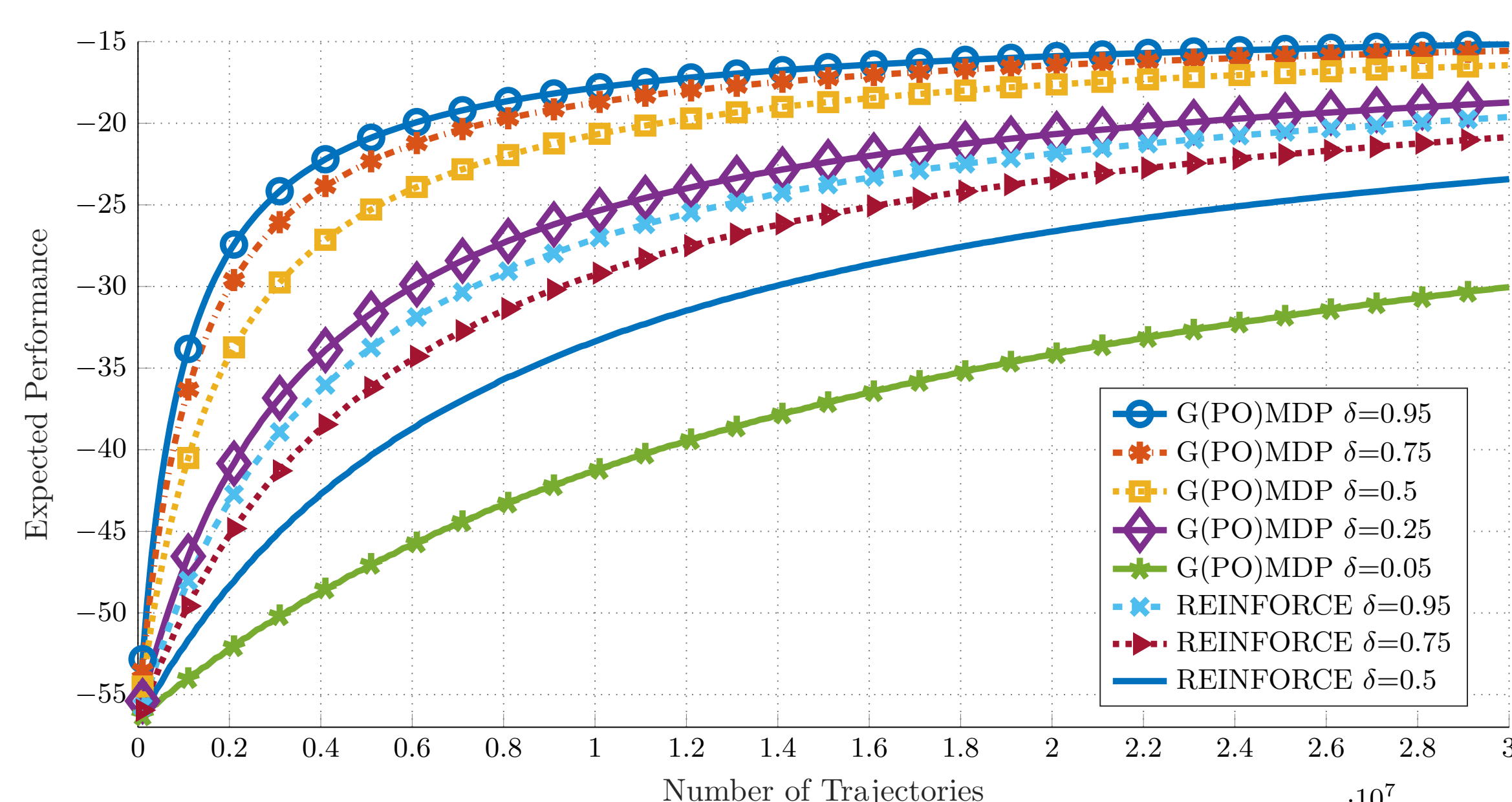
N^* has **no practical closed-form solution**: we **suggest** to find it with a linear search, knowing that:

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

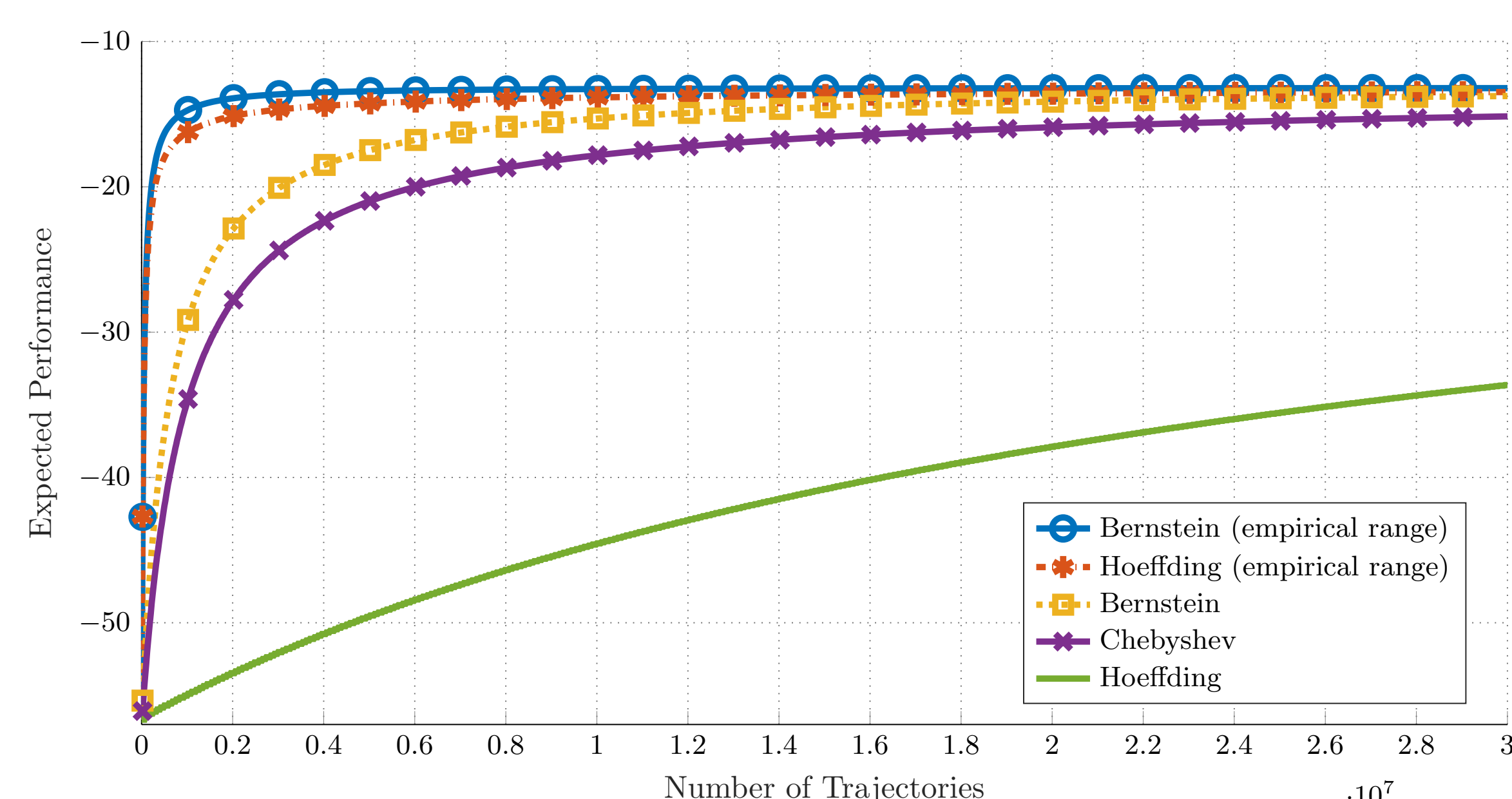
then compute Λ^* from $\epsilon(N^*)$

EMPIRICAL RESULTS (ONE-DIMENSIONAL LQG)

Comparing gradient estimation algorithms and values of δ



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



REFERENCES

- R. Bollapragada, R. Byrd, and J. Nocedal. Adaptive Sampling Strategies for Stochastic Optimization. *ArXiv e-prints*, 2017.
- Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In *International Conference on Machine Learning* 25, volume 307 of *ACM International Conference Proceeding Series*, pages 672–679. ACM, 2008.
- Matteo Pirota and Marcello Restelli. Cost-sensitive approach for batch size optimization. In *Optimizing the optimizers, NIPS 2016 Workshop*, 2016.
- Matteo Pirota, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *Advances in Neural Information Processing Systems* 26, pages 1394–1402, 2013.
- S. L. Smith, P.-J. Kindermans, and Q. V. Le. Don't Decay the Learning Rate, Increase the Batch Size. *ArXiv e-prints*, 2017.