Last-Iterate Global Convergence of Policy Gradients for Constrained Reinforcement Learning

A. Montenegro, M. Mussi, M. Papini, and A. M. Metelli

{alessandro.montenegro, marco.mussi, matteo.papini, albertomaria.metelli}@polimi.it

Motivations

Policy Gradients for Constrained RL? PGs handle **continuous state-action spaces**, making them suitable for real-world constrained control problems, learning via **AB** or **PB** exploration paradigms.

Constraints on Risks? They enforce **safer behaviors** by imposing constraints on risk measures instead of on expected costs.

Contributions

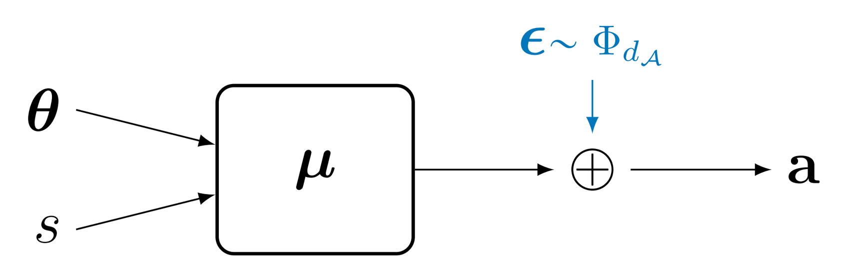
Exploration-Agnostic Framework for solving **risk-constrained continuous** control problems via policy-based primal-dual methods under **general** policy **parameterization** and with either **AB** or **PB** exploration approaches.

Last-Iterate Global Convergence to a **feasible policy** for the exploration-agnostic method `C-PG` with:

- **regularized Lagrangian** w.r.t. the dual variable
- **weak gradient domination** w.r.t. the parameters
- multiple **constraints on expected costs**

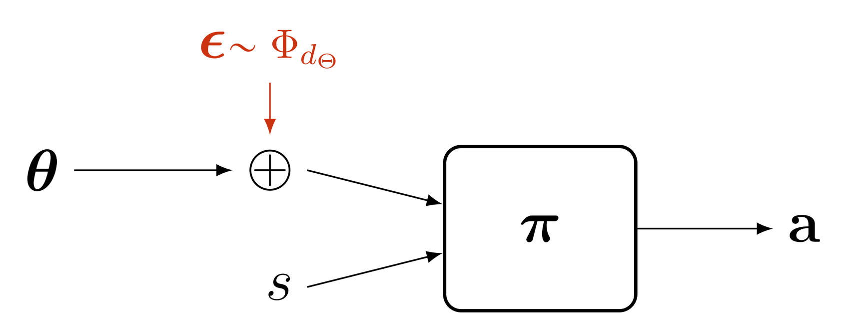
AB and PB Exploration

Action-Based Exploration



$$J_A(\theta) = \mathop{\mathbb{E}}_{\tau \sim p_A(\cdot|\theta)}[R(\tau)]$$

Parameter-Based Exploration



$$J_P(\theta) = \mathop{\mathbb{E}}_{\theta \sim \nu_\rho}\left[\mathop{\mathbb{E}}_{\tau \sim p_A(\cdot|\theta)}[R(\tau)]\right]$$

Solving Risk Constrained Optimization Problem via PGs

Exploration-Agnostic Constrained Optimization Problem

$$\min_{\upsilon \in \mathcal{V}} J_0(\upsilon) \quad \text{s.t.} \quad J_i(\upsilon) \leqslant b_i, \quad \forall i \in [\![U]\!]$$

which is equivalent to

$$\min_{\upsilon \in \mathcal{V}} \max_{\lambda \geq 0_U} \mathcal{L}_0(\upsilon, \lambda)$$

Algorithm

Projected Alternate Ascent-Descent on the Ridge-Regularized Lagrangian $\mathcal{L}_\omega(\upsilon, \lambda)$ w.r.t. λ

$$\mathcal{L}_\omega(\upsilon, \lambda) := J_0(\upsilon) + \langle \lambda, \mathbf{J}(\upsilon) - \mathbf{b} \rangle - \frac{\omega}{2}\|\lambda\|_2^2$$

Unified Risk Measure

$$\min_{\eta \in \mathbb{R}} \mathcal{J}(\theta, \eta)$$

$\rightarrow$

$$\mathcal{J}(\theta, \eta) := \mathop{\mathbb{E}}_{\tau \sim p(\cdot|\theta)}[f(C(\tau), \eta)] + g(\eta)$$

Risk Mapping

	f	g
Cost	$C(\tau)$	0
MV	$C(\tau)(1-2\kappa\eta+\kappa C(\tau))$	$\kappa\eta^2$
CVaR$_\alpha$	$(1-\alpha)^{-1}(C(\tau)-\eta)^+$	η
Chance	$\mathbb{1}\{C(\tau) \geqslant n\}$	0

Last-Iterate Global Convergence

Assumptions

1. ψ-Gradient Domination
2. $\mathcal{L}_\omega$ Regularity
3. $\widehat{\mathbb{V}\mathrm{ar}}[\widehat{\nabla}\mathcal{L}_\omega]$ Bounded
4. Saddle Point Existence

ψ-Gradient Domination $\psi \in [1, 2]$ $\rightarrow$

$$\|\nabla_\upsilon \mathcal{L}_0(\upsilon, \lambda)\|_2^\psi \geqslant \alpha_1 \left(\mathcal{L}_0(\upsilon, \lambda) - \min_{\upsilon' \in \mathcal{V}} \mathcal{L}_0(\upsilon', \lambda)\right) - \beta_1$$

Under 1-4 and with constraints on expected costs

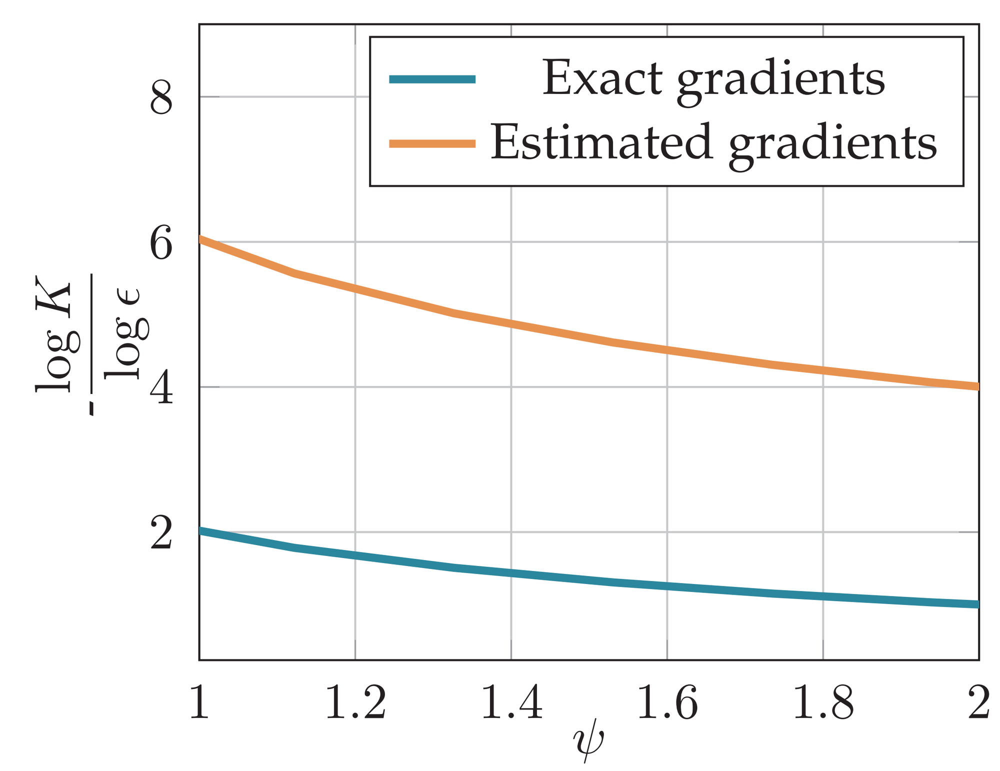
$$\mathbb{E}[J_0(\upsilon_k) - J_0(\upsilon_0^*)] \leqslant \epsilon + \frac{\beta_1}{\alpha_1} + \frac{\omega}{2}\|\lambda_0^*\|_2^2 \quad \text{and} \quad \mathbb{E}[(J_i(\upsilon_k) - b_i)^+] \leqslant 4\epsilon + 4\frac{\beta_1}{\alpha_1} + \omega\|\lambda_0^*\|_2, \ \forall i \in [\![U]\!]$$

Exact Gradients

	$\psi = 1$ (GD)	$\psi = 2$ (PL)
Fixed ω	$\omega^{-1}\epsilon^{-1}$	$\omega^{-1}\log(\epsilon^{-1})$
$\omega = \mathcal{O}(\epsilon)$	ϵ^{-2}	$\epsilon^{-1}\log(\epsilon^{-1})$

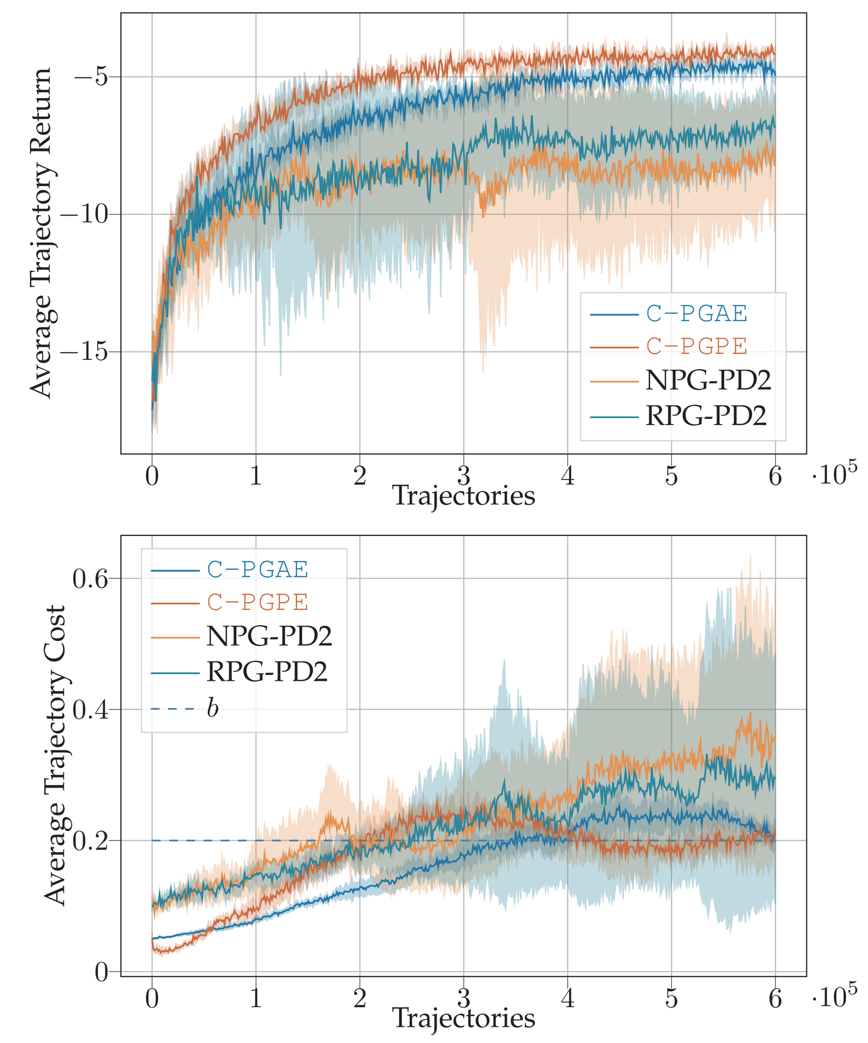
Estimated Gradients

	$\psi = 1$ (GD)	$\psi = 2$ (PL)
Fixed ω	$\omega^{-3}\epsilon^{-3}\log(\epsilon^{-1})$	$\omega^{-3}\epsilon^{-1}\log(\epsilon^{-1})$
$\omega = \mathcal{O}(\epsilon)$	$\epsilon^{-6}\log(\epsilon^{-1})$	$\epsilon^{-4}\log(\epsilon^{-1})$



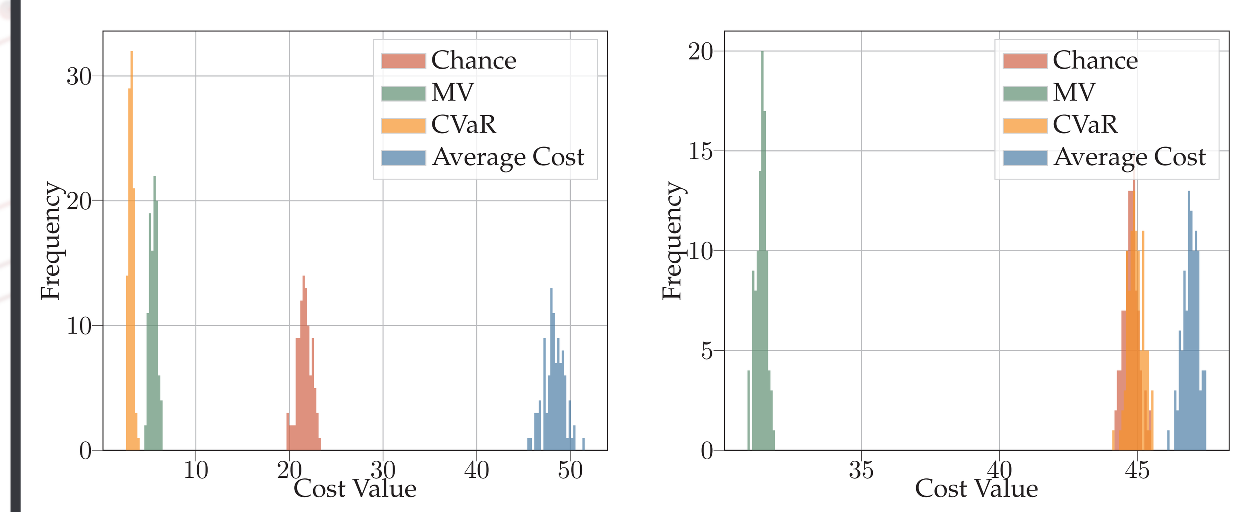
Experiments

Number of Trajectories Study



Risk Minimization Study

Last-Iterate Cost Distribution



Last-Iterate Performance

Cost	CVaR$_\alpha$	MV	Chance
26.91 ± 0.09	23.07 ± 0.26	24.23 ± 0.23	26.34 ± 0.16
25.80 ± 0.14	23.08 ± 0.24	0.75 ± 0.26	23.07 ± 0.26

References

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *JAIR*, 2001.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *NeurIPS*, 36, 2024.

Alessandro Montenegro, Marco Mussi, Alberto Maria Metelli, and Matteo Papini. Learning optimal deterministic policies with stochastic policy gradients. *ICML*, 2024.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 2010.