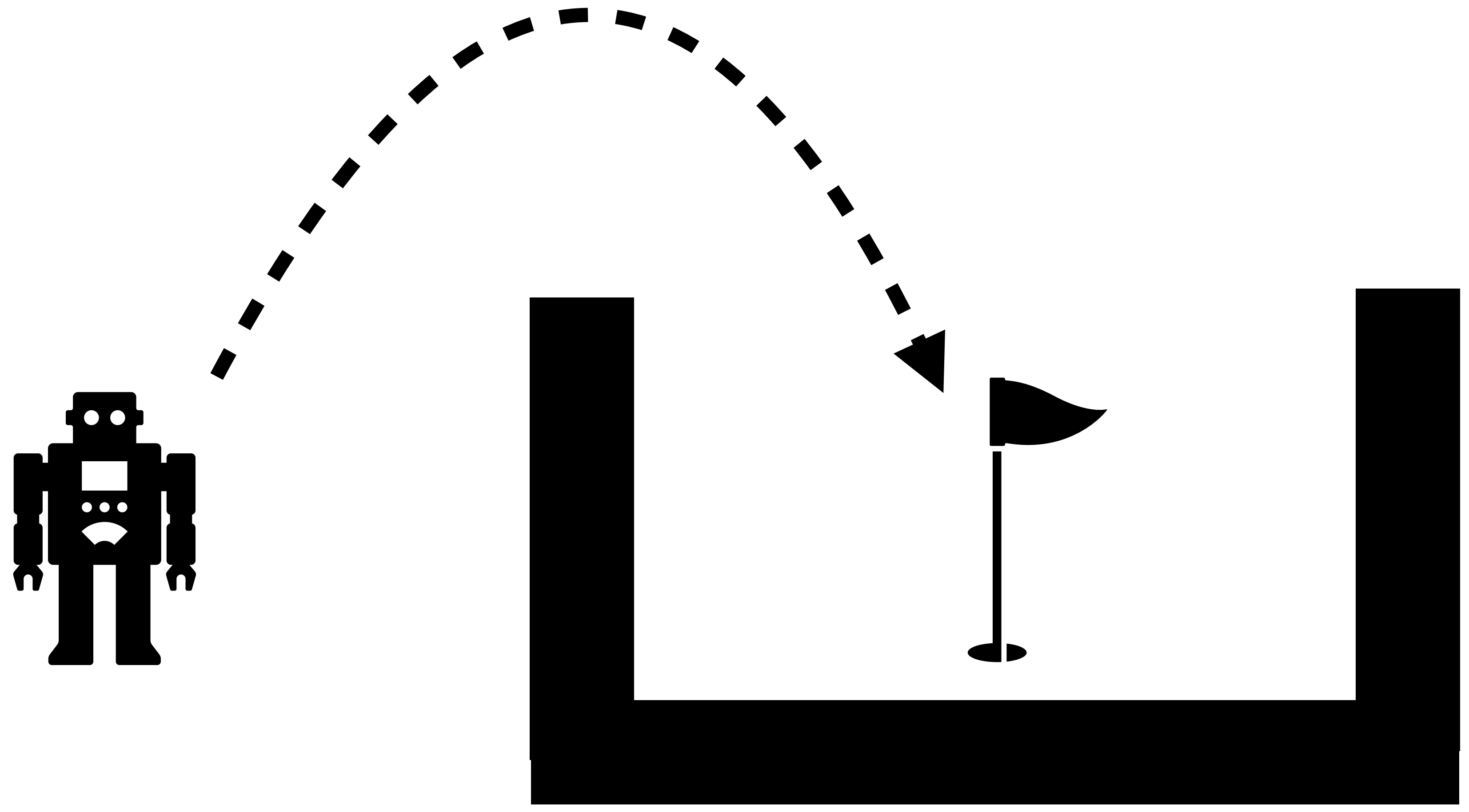
Last-Iterate Global Convergence of Policy Gradients for Constrained Reinforcement Learning

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

38th Conference on Neural Information Processing Systems (NeurIPS 2024)

Constrained Reinforcement Learning (CRL)
Introduction

- Real-world scenarios: reach a goal + meet structural/utility-based constraints
- Constrained RL: extension of RL with the possibility to account for constraints



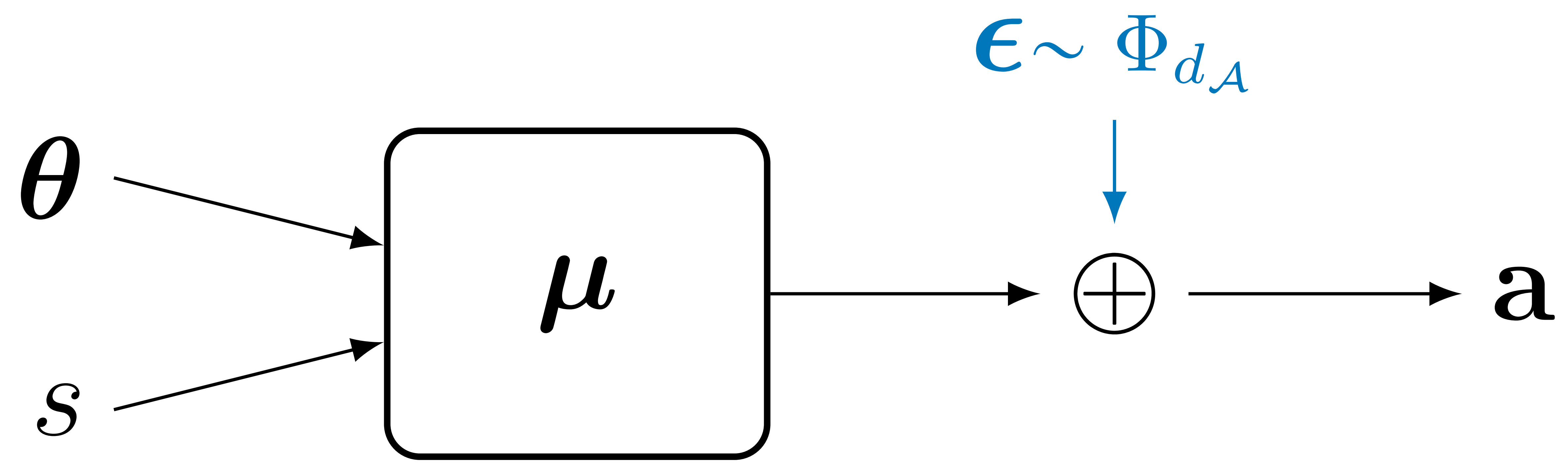
Policy Gradients (PGs) for CRL

Introduction

- Continuous State and Action Spaces
- Robustness to Actuators and Sensors Noise
- Robustness to Partial Observability
- Possibility to incorporate expert-knowledge in the Policy-design Phase

Action-based (AB) Exploration

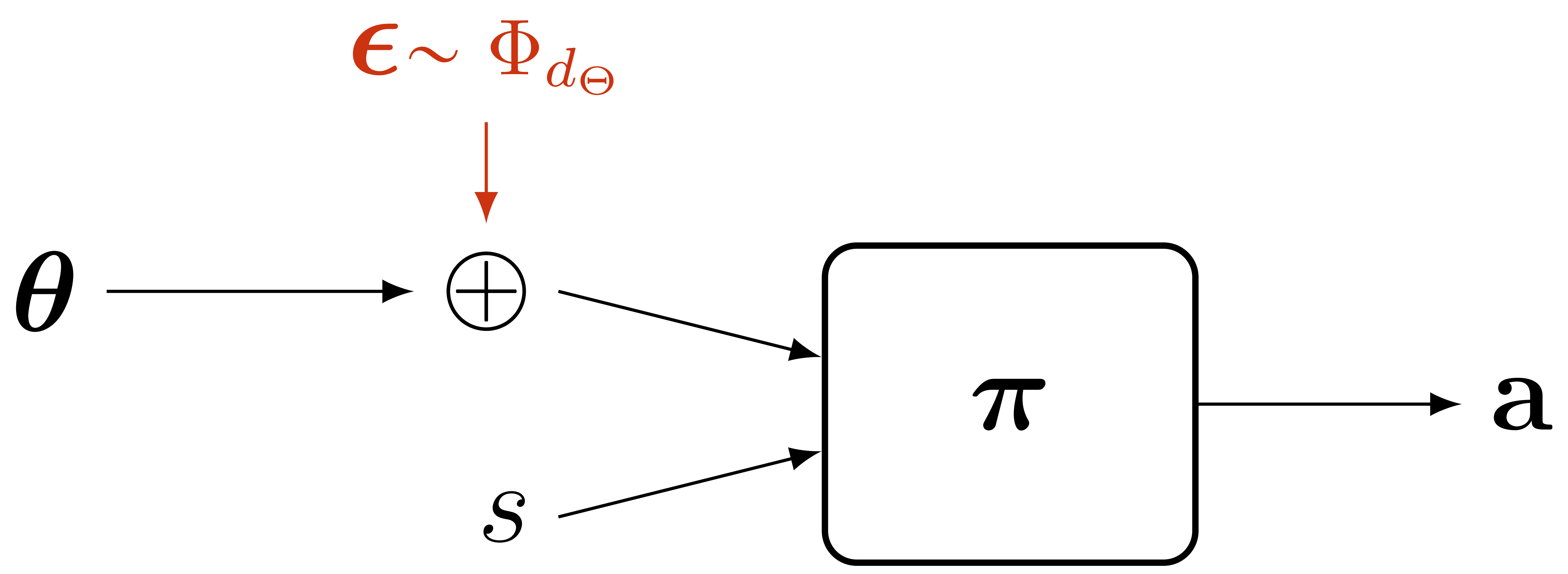
PGs Exploration Approaches



$$J_{\mathrm{A}}(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[R(\tau)\right]$$

Parameter-based (PB) Exploration

PGs Exploration Approaches



$$J_{\mathrm{P}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta} \sim \nu_{\boldsymbol{\rho}}} \left[\mathbb{E}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})} \left[R(\tau) \right] \right]$$

Constrained Optimization Problem

Setting

- Continuous State and Action spaces
- Multiple constraints on cost functions c_i
- Both exploration paradigms are supported
- Inexact Gradients

Constrained Optimization Problem

Setting

$$\min_{\boldsymbol{v}\in\mathcal{V}} J_{\dagger,0}(\boldsymbol{v}) \quad \text{s.t.} \quad J_{\dagger,i}(\boldsymbol{v}) \leqslant b_i, \quad \forall i \in [\![U]\!]$$

Constrained Optimization Problem

Setting

$$\min_{\boldsymbol{v}\in\mathcal{V}} \boxed{J_{\dagger,0}(\boldsymbol{v})} \quad \text{s.t.} \quad \boxed{J_{\dagger,i}(\boldsymbol{v})} \leqslant b_i, \quad \forall i \in [\![U]\!]$$

AB or PB approaches on costs c_i with $i \in \{0,1,\ldots,U\}$

Constrained Optimization Problem

Setting

$$\min_{\boldsymbol{v}\in\mathcal{V}} J_{\dagger,0}(\boldsymbol{v}) \quad \text{s.t.} \quad J_{\dagger,i}(\boldsymbol{v}) \leqslant \boxed{b_i,} \quad \forall i \in [\![U]\!]$$

i-th threshold

C-PG

Exploration-Agnostic Algorithm

Algorithm

Projected Alternate Ascent Descent on the ω-Regularized Lagrangian w.r.t. the Dual Variable

$$\downarrow \widehat{\nabla}_v \mathscr{L}_\omega(v, \lambda) \qquad \uparrow \widehat{\nabla}_\lambda \mathscr{L}_\omega(v, \lambda)$$

C-PG: Convergence

Exploration-Agnostic Algorithm

Assumptions:

1. ψ-Gradient Domination on parameterization ($\psi \in [1,2]$)
2. Regularity of $\mathscr{L}_\omega$
3. Existence of a saddle point

C-PG: Convergence

Exploration-Agnostic Algorithm

Theorem

$$\mathbb{E}[J_0(\boldsymbol{v}_k) - J_0(\boldsymbol{v}_0^*)] \leqslant \epsilon + \frac{\beta_1}{\alpha_1} + \frac{\omega}{2}\|\boldsymbol{\lambda}_0^*\|_2^2 \quad \text{and} \quad \mathbb{E}[(J_i(\boldsymbol{v}_k) - b_i)^+] \leqslant 4\epsilon + 4\frac{\beta_1}{\alpha_1} + \omega\|\boldsymbol{\lambda}_0^*\|_2 \ , \ \forall i \in [\![U]\!]$$

Holds for both exploration approaches

C-PG: Convergence

Exploration-Agnostic Algorithm

	$\boxed{\psi = 1}$	$\boxed{\psi = 2}$
Exact Gradients	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1}\log(\epsilon^{-1}))$
Estimated Gradients	$\mathcal{O}(\epsilon^{-6}\log(\epsilon^{-1}))$	$\mathcal{O}(\epsilon^{-4}\log(\epsilon^{-1}))$

Enforcing Constraints on Risks

Risk and Exploration Agnostic Algorithms

- AB and PB explorations have a **semantic difference** when enforcing constraints
- In order to induce safer behaviors, we can enforce **constraints on risk** measures
- We introduce a **unified risk measure** that extends the framework previously described
- **Additional parameter** to learn required

Conclusions

Our Contribution

- **Framework** to handle CRL with PGs (both AB and PB) in continuous spaces and with multiple constraints
- Both approaches exhibit **last-iterate global convergence** to a feasible (hyper)policy guarantees
- We extend the framework to handle **risk-based constraints**
- We **numerically** validate our results

Thank you!