Challenging Common Assumptions in Convex Reinforcement Learning

POLITECNICO DI MILANO ⋆

UNIVERSITÀ DI BOLOGNA †

ETH ψ

Mirco Mutti⋆,†, Riccardo De Santiψ, Piersilvio De Bartolomeisψ, Marcello Restelli⋆

mirco.mutti@polimi.it rdesanti@ethz.ch pdebartol@ethz.ch marcello.restelli@polimi.it

Infinite Trials Setting

RL (dual)

$$\mathcal{J}_\infty(\pi) := r \cdot d^\pi$$

episodic with **horizon** T

reward vector r

state **distribution** d^π

convex/concave **objective** $\mathcal{F}$

Convex RL[1]

$$\zeta_\infty(\pi) := \mathcal{F}(d^\pi)$$

Finite Trials Setting

RL (dual)

$$\mathcal{J}_n(\pi) := \mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [r \cdot d_n]$$

episodic with **horizon** T

reward vector r

state visitation **frequency** with n episodes d_n

convex/concave **objective** $\mathcal{F}$

Convex RL

$$\zeta_n(\pi) := \mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [\mathcal{F}(d_n)]$$

Challenged Assumptions

Previous works in convex RL consider an infinite trials formulation to approximate a single trials one

	Finite Trials		Infinite Trials
RL	$\mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [r \cdot d_n]$	$=$	$r \cdot d^\pi$
Convex RL	$\mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [\mathcal{F}(d_n)]$	$\neq$	$\mathcal{F}(d^\pi)$

Challenged Assumption 1

The convex RL problem can be equivalently addressed with an infinite trials formulation

Finite Trials $\mathcal{J}_n(\pi) = \mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [r \cdot d_n] = r \cdot \mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [d_n] = r \cdot d^\pi = \mathcal{J}_\infty(\pi)$

Infinite Trials $\zeta_\infty(\pi) = \mathcal{F}(d^\pi) = \mathcal{F}(\mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [d_n]) \le \mathop{\mathbb{E}}_{d_n \sim p_n^\pi} [\mathcal{F}(d_n)] = \zeta_n(\pi)$

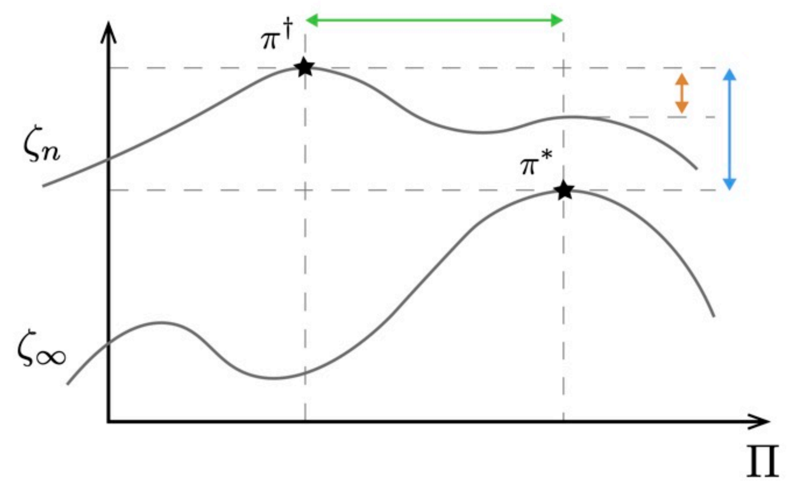
APPROXIMATION ERROR

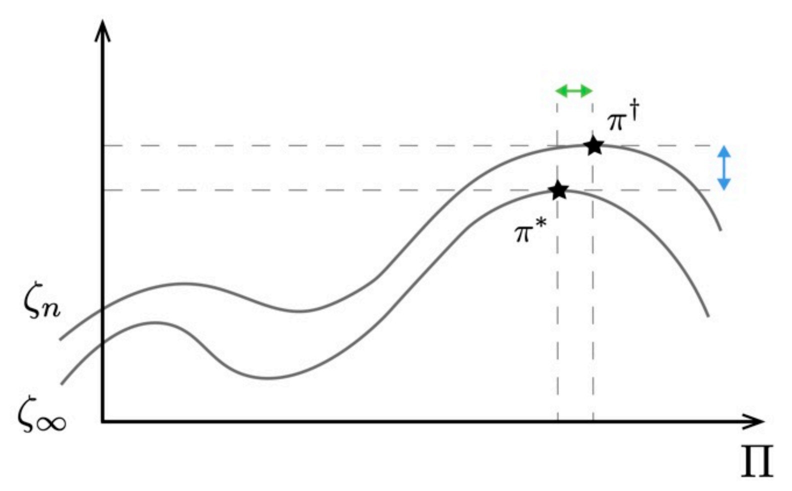
$$\left| \zeta_n(\pi^\dagger) - \zeta_n(\pi^\star) \right| \le 4LT \sqrt{\frac{2S \log(4T/\delta)}{n}}$$

$\pi^\dagger$ optimal finite trials policy

π^* optimal infinite trials policy

S, T, L are problem dependent





Challenged Assumption 2

The convex RL formulation is only slightly harder than the standard RL formulation

Challenged Assumption 3

The set of stationary randomized policies is sufficient for the convex RL formulation

Online Learning Setting

Single Trial Convex RL

$$\zeta_1(\pi) := \mathop{\mathbb{E}}_{d \sim p^\pi} [\mathcal{F}(d)]$$

unknown objective $\mathcal{F}$

approximate $\mathcal{F}$ from interactions (Bernstein polynomial)

N-episodes online **regret** $\mathcal{R}(N) := \sum_{t=1}^{N} V^* - V^{(t)}$

For any $\delta \in (0,1]$, unknown convex MDP, using OPE-UCBVI algorithm[2]

$$\mathcal{R}(N) \le \underbrace{N L_V L_{\mathcal{F}} (S/d_{\mathbf{w}})^{\frac{1}{2}}}_{\text{approximation term}} + \underbrace{\mathcal{R}_\delta(N)}_{\text{pure learning regret}^2}$$

V L_V-Lipschitz

$\mathcal{F}$ $L_{\mathcal{F}}$-Lipschitz

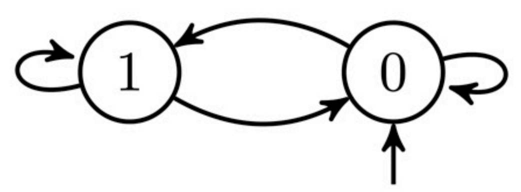
with probability $1-\delta$, where the regret is sub-linear in N

Empirical Validation

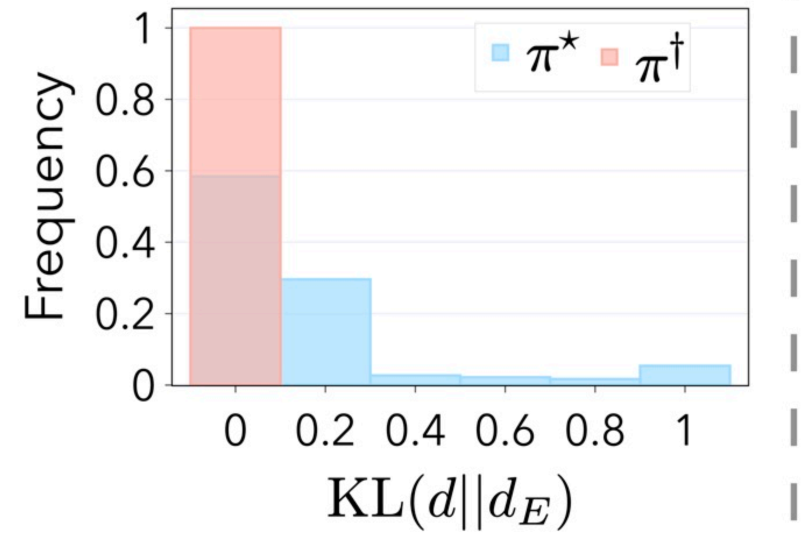
Objective $\mathcal{F}$		Application	Infinite Trials = Finite Trials
$r \cdot d$	$r \in \mathbb{R}^S, d \in \Delta(\mathcal{S})$	RL	✓
$\lVert d - d_E \rVert_p^p$ $KL(d \Vert d_E)$	$d, d_E \in \Delta(\mathcal{S})$	Imitation Learning	✗
$-d \cdot \log(d)$	$d \in \Delta(\mathcal{S})$	Pure Exploration	✗
$\mathrm{CVaR}_\alpha[r \cdot d]$ $r \cdot d - \mathbb{V}\mathrm{ar}[r \cdot d]$	$r \in \mathbb{R}^S, d \in \Delta(\mathcal{S})$	Risk-Averse RL	✗
$r \cdot d$, s.t. $\lambda \cdot d \le c$	$r, \lambda \in \mathbb{R}^S, c \in \mathbb{R}, d \in \Delta(\mathcal{S})$	Linearly Constrained RL	✓
$-\mathbb{E}_z\, d_{KL}(d_z \Vert \mathbb{E}_k\, d_k)$	$z \in \mathbb{R}^d, d_z, d_k \in \Delta(\mathcal{S})$	Diverse Skill Discovery	✗

Imitation Learning

$$\mathcal{F}(d) = \mathrm{KL}(d \Vert d_E)$$



d_E
$(1/3, 2/3)$



References

[1] Zahavy et al., *Reward is enough for convex mdps.* NeurIPS, 2021.

[2] Chatterji et al., *On the theory of reinforcement learning with once-per-episode feedback.* NeurIPS, 2021.