Reinforcement Learning with Convex Constraints

Sobhan Miryoosefi[1], Kianté Brantley[2], Hal Daumé III[2,3], Miroslav Dudík[3], Robert E. Schapire[3]

[1]*Princeton University*, [2]*University of Maryland*, [3]*Microsoft Research*

Main ideas

find a policy satisfying some *(convex) constraints* on the observed average *"measurement vector"*

Constraint-based RL:
- more natural in many applications

Examples of constraints we allow:
- [previously studied] *orthant constraints*: bounds on total wear, probability of bad events (safety), . . .
- [new] bound on distance to expert behavior, distance to uniform distribution (diversity), . . .
- bound on reward can be incorporated (as a constraint)

Our game-theoretic approach:
- relies on ability to approximately solve standard RL
- uses any off-the-shelf RL algorithm
- satisfies rigorous theoretical guarantees

Convex feasibility problem

Model: $M = (\mathcal{S}, \mathcal{A}, \beta, P_s, P_z)$ vector-valued MDP

$\mathcal{S}$ states, $\mathcal{A}$ actions, β initial distribution

transitions: $s_{i+1} \sim P_s(\cdot|s_i, a_i)$, $s_0 \sim \beta$

$\pi \in \Pi$ stationary policy, $a_i \sim \pi(s_i)$

vector measurements: $\mathbf{z}_i \sim P_z(\cdot|s_i, a_i)$, $\mathbf{z}_i \in \mathbb{R}^d$

> Find a policy $\pi \in \Pi$ such that *long-term measurement* $\overline{\mathbf{z}}(\pi)$ lies in a convex target set $\mathcal{C}$

where (for some discount factor $\gamma \in [0, 1)$)

$$\overline{\mathbf{z}}(\pi) \triangleq \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i \mathbf{z}_i \;\middle|\; \pi\right]$$

We consider mixed policies $\mu \in \Delta(\Pi)$
(distributions over finitely many policies in Π):

$$\overline{\mathbf{z}}(\mu) \triangleq \mathbb{E}_{\pi\sim\mu}\left[\overline{\mathbf{z}}(\pi)\right]$$

Our approach (similar to Abernethy et al., 2011)

- **Solve a harder problem**

$$\min_{\mu\in\Delta(\Pi)} \operatorname{dist}(\overline{\mathbf{z}}(\mu), \mathcal{C})$$

where dist is Euclidean distance

- **Form the game.** If $\mathcal{C}$ is a cone[†], then $\operatorname{dist}(\mathbf{x}, \mathcal{C}) = \max_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}\cdot\mathbf{x}$, where Λ is a suitable convex set. So:

$$\min_{\mu\in\Delta(\Pi)} \max_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}\cdot\overline{\mathbf{z}}(\mu)$$

- **Solve two-person *zero-sum game***

$$\underbrace{\max_{\boldsymbol{\lambda}\in\Lambda}}_{\text{no-regret learner}} \quad \underbrace{\min_{\mu\in\Delta(\Pi)}}_{\text{best-response}} \quad \boldsymbol{\lambda}\cdot\overline{\mathbf{z}}(\mu)$$

- **Key insight:** best response (i.e., μ-player) is standard reinforcement learning!

RL task with scalar reward of $r_i = -\boldsymbol{\lambda}\cdot z_i$ at each step i

† For an arbitrary convex $\mathcal{C}$, we apply a reduction to obtain a similar algorithm and guarantees.

Our algorithm: ApproPO

(approachability-based policy optimization)

Initialize $\boldsymbol{\lambda}_1 \in \Lambda$ arbitrarily

For $t = 1, \dots, T$
- compute approximately optimal policy π_t for standard RL with scalar reward $r = -\boldsymbol{\lambda}_t\cdot\mathbf{z}$
- $\hat{\mathbf{z}}_t \leftarrow$ approximate long-term measurement for π_t
- update $\boldsymbol{\lambda}_{t+1}$ from $\boldsymbol{\lambda}_t$ based on gradient $-\hat{\mathbf{z}}_t$
(following online gradient descent algorithm)

Return $\mu = \frac{1}{T}\sum_{t=1}^{T} \pi_t$

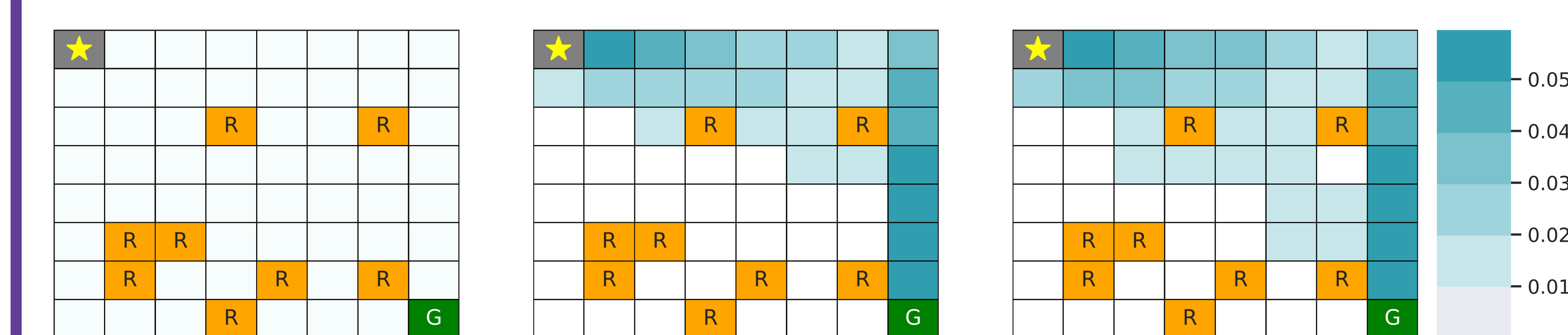
Guarantees and further details

Convergence guarantee:

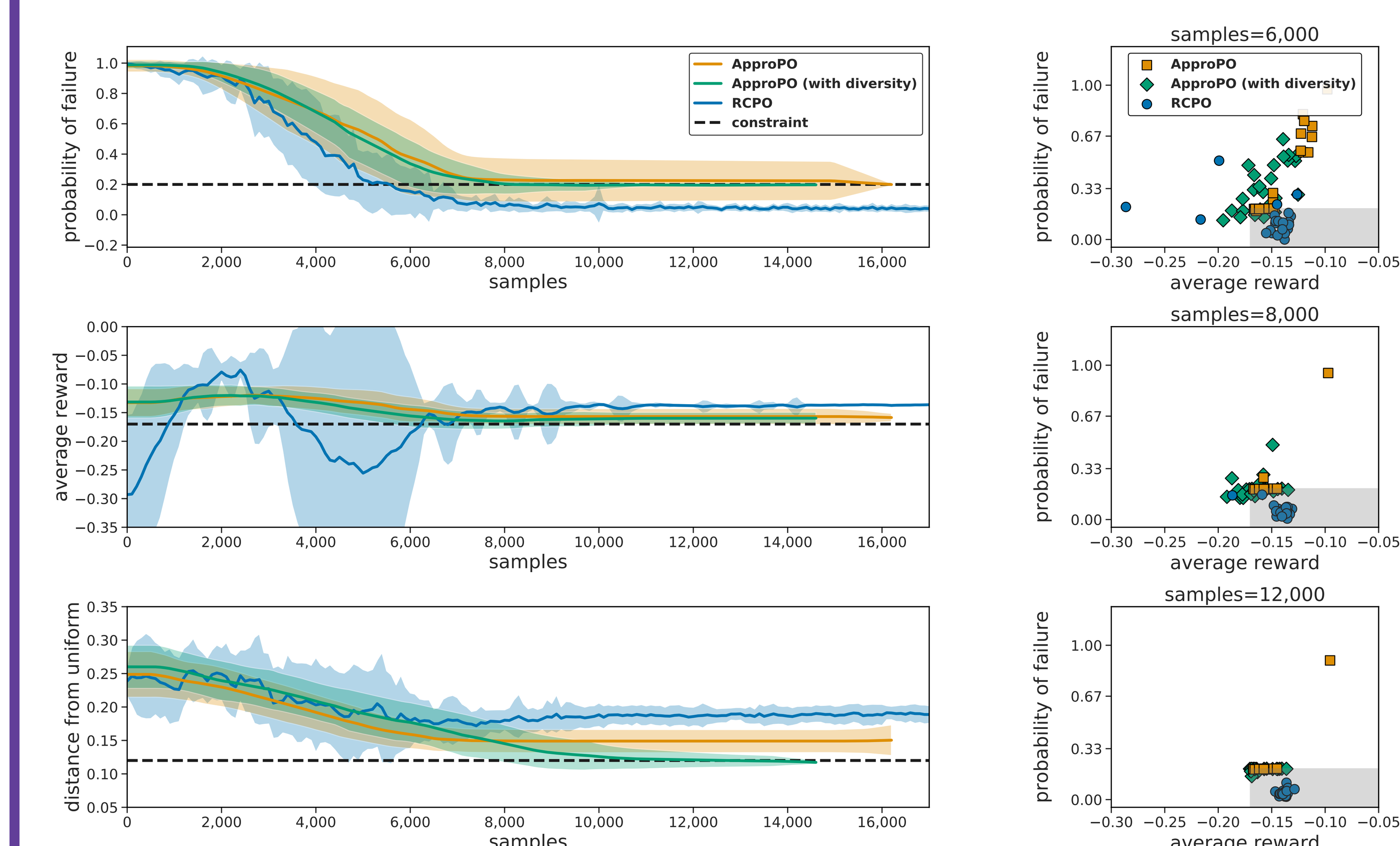
$$\operatorname{dist}(\overline{\mathbf{z}}(\mu), \mathcal{C}) \le \min_{\mu\in\Delta(\Pi)} \operatorname{dist}(\overline{\mathbf{z}}(\mu), \mathcal{C}) + O(T^{-1/2})$$

For feasibility problem: *positive (not best) response* is enough
- *positive response*: given $\boldsymbol{\lambda}$, find policy π that achieves a positive long-term reward in a standard MDP with scalar reward $r_i = -\boldsymbol{\lambda}\cdot\mathbf{z}_i$.
- modified algorithm either reports infeasibility or returns μ that converges to $\mathcal{C}$ with rate of $O(T^{-1/2})$.
- substantial efficiency improvements:
  - only run off-the-shelf RL until the reward positive;
  - cache previous results and check if they give positive reward.

Experiments



Left: The Mars rover domain. ***Middle, Right***: Visitation probabilities of ApproPO without and with diversity constraints, respectively.



Left: Convergence of constraint values (mean and standard deviation across 25 replicates). ***Right***: Each symbol corresponds to a separate replicate at a given number of samples.

[RCPO baseline] C. Tessler, D. Mankowitz, S. Mannor. Reward Constrained Policy Optimization. *ICLR* 2019.