Constrained episodic reinforcement learning in concave-convex and knapsack settings

Kianté Brantley, Miroslav Dudik,
Thodoris Lykouris, Sobhan Miryoosefi,
Max Simchowitz, Alex Slivkins, Wen Sun

Microsoft Research

Motivation

Traditional RL maximizes a single scalar objective function, which is not enough in many real world applications.

- Self-driving car needs to get to a destination quickly while satisfying gas budget
- Video game AI agents need to win the game while playing like human beings
- Robots need to achieve the task while avoiding applying large torques
- Constraints are easier to specify than a scalar reward or cost function

Main Ideas and Contribution

Efficient Exploration
in constrained episodic reinforcement learning setting (cMDP)
with a focus on

- Concave reward and Convex constraints (Concave-Convex setting)
- Hard constraints (Knapsack constraints)
- Empirical improvement over previous approaches

Our approach

- Optimism under Uncertainty
- Modular Analysis
- Novel application of mean-value theorem (circumventing challenges in Concave-Convex setting)

Preliminaries

States $\mathcal{S}$, Actions $\mathcal{A}$, Horizon H, no. episodes K

set $\mathcal{D}$ of d resources, episodic resource capacity $\xi(i)$

cMDP $\mathcal{M} = (p, r, \boldsymbol{c})$, True Model $\mathcal{M}^* = (p^*, r^*, \boldsymbol{c}^*)$

Transition probability p, reward r, resource consumption $\boldsymbol{c}$

Objective: minimize regret with respect to the following benchmark π^*

$$\max_{\pi} \mathbb{E}^{\pi,p^*}\Big[\sum_{h=1}^{H} r^*(s_h,a_h)\Big] \quad \text{s.t} \quad \forall i \in \mathcal{D} \;:\; \mathbb{E}^{\pi,p^*}\Big[\sum_{h=1}^{H} c^*(s_h,a_h,i)\Big] \le \xi(i)$$

Reward Regret: $\mathrm{RewReg}(k) = \mathbb{E}^{\pi^*,p^*}\Big[\sum_{h=1}^{H} r^*(s_h,a_h)\Big] - \frac{1}{k}\sum_{t=1}^{k}\mathbb{E}^{\pi_t,p^*}\Big[\sum_{h=1}^{H} r^*(s_h,a_h)\Big]$

Constraint Regret: $\mathrm{ConsReg}(k) = max_{i\in\mathcal{D}}\Big(\frac{1}{k}\sum_{t=1}^{k}\mathbb{E}^{\pi_t,p^*}\Big[\sum_{h=1}^{H} c^*(s_h,a_h,i)\Big] - \xi(i)\Big)$

Warm up: the Basic setting

ConRL

- For $k = 1,2,\dots,K$
  1. Let $\hat{p}_k, \hat{r}_k, \hat{\boldsymbol{c}}_k$ be the empirical estimates of the true model and define the bonus enhanced model $\mathcal{M}^{(k)} = (p^{(k)}, r^{(k)}, \boldsymbol{c}^{(k)})$

     $r^{(k)}(s,a) = \hat{r}_k(s,a) + \hat{b}_k(s,a) \quad \mathbf{c}^{(k)}(s,a) = \hat{\mathbf{c}}_k(s,a) - \hat{b}_k(s,a)\mathbf{1}_d$
  2. Let π_k be the solution to the following planning problem

     $$\max_{\pi} \mathbb{E}^{\pi,p^{(k)}}\Big[\sum_{h=1}^{H} r^{(k)}(s_h,a_h)\Big] \quad \text{s.t} \quad \forall i \in \mathcal{D} \;:\; \mathbb{E}^{\pi,p^{(k)}}\Big[\sum_{h=1}^{H} c^{(k)}(s_h,a_h,i)\Big] \le \xi(i)$$
  3. Run π_k for an episode and collect samples.

Bonus: $\hat{b}_k(s,a) = \tilde{O}(H\sqrt{1/N_k(s,a)})$ where $N_k(s,a) =$ # (s,a) visited

Main Result: with probability at least $1-\delta$ we have

$$\mathrm{RewReg}(k) \le \tilde{O}\Big(S\sqrt{AH^3}\cdot\frac{1}{\sqrt{(k)}}\Big) \qquad \mathrm{ConsReg}(k) \le \tilde{O}\Big(S\sqrt{AH^3}\cdot\frac{1}{\sqrt{(k)}}\Big)$$

Concave-Convex Setting

Setting and Objective: suppose $f:\mathbb{R}\to\mathbb{R}$ is concave and $g:\mathbb{R}^d\to\mathbb{R}$ is convex. Additionally assume that both are L-Lipschitz with respect to ℓ_1 norm. We want to be competitive against the following benchmark

$$\max_{\pi}\; f\Big(\mathbb{E}^{\pi,p^*}\Big[\sum_{h=1}^{H} r^*(s_h,a_h)\Big]\Big) \quad \text{s.t} \quad g\Big(\mathbb{E}^{\pi,p^*}\Big[\sum_{h=1}^{H} c^*(s_h,a_h,i)\Big]\Big) \le 0$$

$$\mathrm{ConvexRewReg}(k) = f\Big(\mathbb{E}^{\pi^*,p^*}\Big[\sum_{h=1}^{H} r^*(s_h,a_h)\Big]\Big) - f\Big(\frac{1}{k}\sum_{t=1}^{k}\mathbb{E}^{\pi_t,p^*}\Big[\sum_{h=1}^{H} r^*(s_h,a_h)\Big]\Big)$$

$$\mathrm{ConvexConsReg}(k) = g\Big(\frac{1}{k}\sum_{t=1}^{k}\mathbb{E}^{\pi_t,p^*}\Big[\sum_{h=1}^{H} c^*(s_h,a_h,i)\Big] - \xi(i)\Big)$$

Algorithm: We can no longer create a bonus enhanced model asme as basic setting by picking the extreme points of confidence interval; instead we merge steps (1.) and (2.) Into a convex program which finds the right bonus and also solves the planning problem simultaneously.

Main Result: with probability at least $1-\delta$, reward regret and constraint regret are upper bounded by $L\cdot\tilde{O}\big(S\sqrt{AH^3}\cdot K^{1/2}\big)$ and $Ld\cdot\tilde{O}\big(S\sqrt{AH^3}\cdot K^{1/2}\big)$ respectively.

Knapsack Setting

Setting and Objective: Fixed total episodes K; Each resource i has total budget B_i; This is a hard threshold. The goal is to maximizes the total reward across K episodes while not exceeding the hard threshold.

Algorithm: uses the basic ConRL algorithm with a smaller episodic resource capacity:

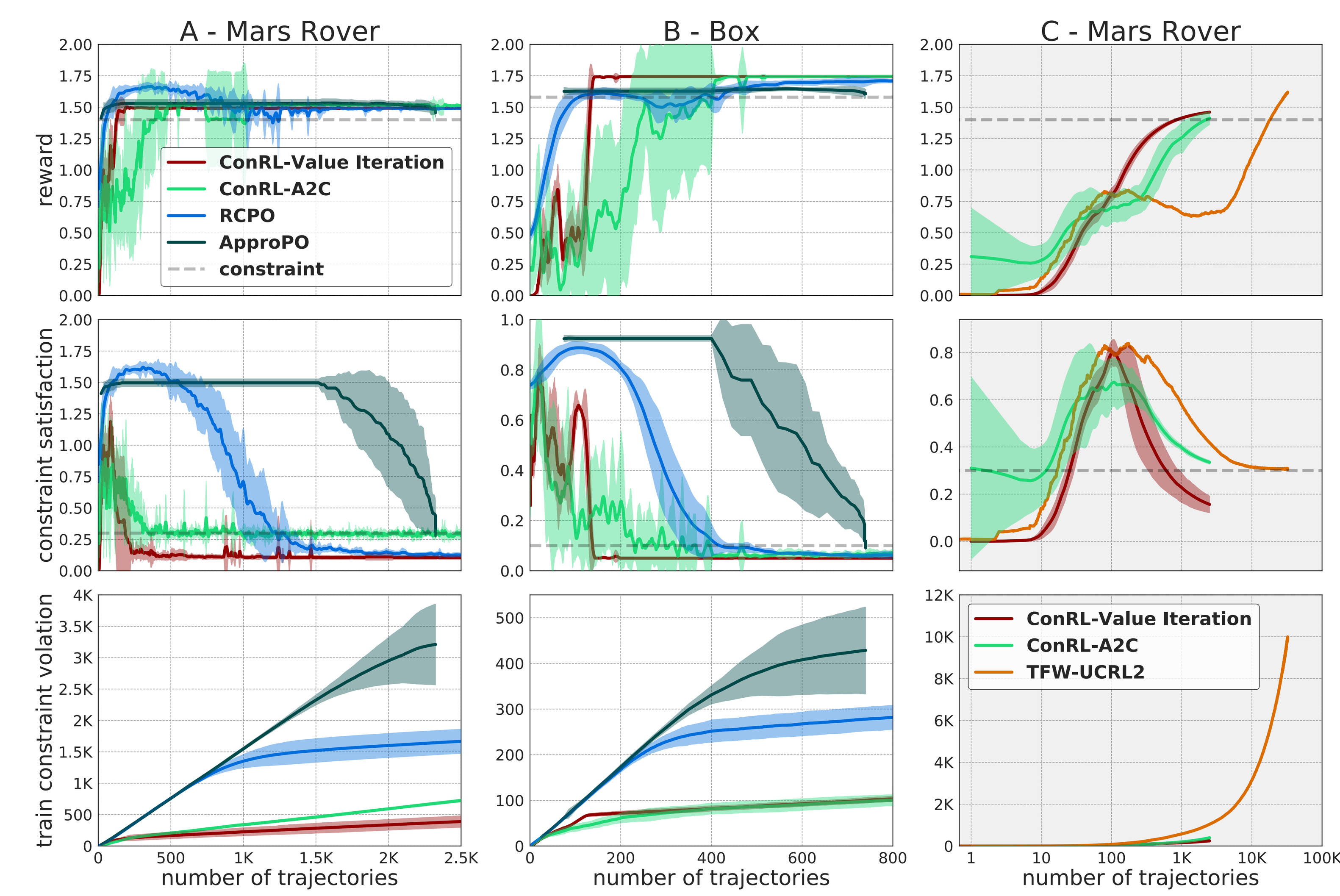
$$\max_{\pi} \mathbb{E}^{\pi,p^{(k)}}\Big[\sum_{h=1}^{H} r^{(k)}(s_h,a_h)\Big] \quad \text{s.t} \quad \forall i \in \mathcal{D} \;:\; \mathbb{E}^{\pi,p^{(k)}}\Big[\sum_{h=1}^{H} c^{(k)}(s_h,a_h,i)\Big] \le \frac{(1-\epsilon)B_i}{K}$$

Benchmark: A dynamic policy (could be history dependent) that maximizes total reward in K episodes while satisfying the hard constraints.

Main Result: Set ϵ properly, with high probability, the reward regret over K episodes is at most $O\Big(\frac{HS\sqrt{HAK}}{\min_i B_i}\Big)$, and the hard constraints are not violated.

(Compare to prior work, our budget can be as small as $\min_i B_i = \Omega(\sqrt{K})$)

Experiments



The performance of the algorithms as function of number of sample trajectories (trajectory = 30 samples); showing average and standard deviation over 10 runs. First two columns shows the comparison to the episodic approaches and the third column shows the comparison with the single-episodic approach.

References

[RCPO] Tessler, C., Mankowitz, D. J., and Mannor, S. (2019). Reward constrained policy optimization, *ICLR* 2019

[ApproPO] Miryoosefi, S., Brantley, K., Daume III, H., Dudik, M., & Schapire, R. E. (2019). Reinforcement learning with convex constraints. *NeurIPS* 2019.

[TFW-UCRL2] Cheung, W. C. (2019). Regret Minimization for Reinforcement Learning with Vectorial Feedback and Complex Objectives. *NeurIPS 2019.*