Motivation

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

- Self-driving cars need 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

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

Preliminaries

- States $S$, Actions $A$, Horizon $H$, no. episodes $K$
- set $\delta$ of $H$ resources, episodic resource capacity $\xi(t)$
- True Model $\mathcal{M} = (P, r, c, \gamma)$
- Transition probability $P$, reward $r$, resource consumption $c$

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

$\max x E^{x^*}[\sum_k r_s(s_k, a_k)] s.t. \forall i \in \mathcal{S} : E^{x^*}[\sum_k r_s(s_k, a_k, i)] \leq \xi(i)$

Reward Regret: $\text{RewReg}(k) = E^{x^*}[\sum_k r_s(s_k, a_k)] - 1/k \sum_k E^{x^*}[\sum_k r_s(s_k, a_k)]$

Constraint Regret: $\text{ConsReg}(k) = \max_{x, a, i} \left(1 - \frac{1}{k} \sum_k E^{x^*}[\sum_k r_s(s_k, a_k, i)] - \xi(i) \right)$

Constrained Episodic RL

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

$\text{RewReg}(k) \leq \tilde{O}\left(\sqrt{AH^2 \cdot \frac{1}{\sqrt{k}}}\right)$

$\text{ConsReg}(k) \leq \tilde{O}\left(\sqrt{AH^2 \cdot \frac{1}{\sqrt{k}}}\right)$

Concave-Convex Setting

Setting and Objective: suppose $f : R \rightarrow R$ is concave and $g : R^d \rightarrow R$ is convex. Additionally assume that both are $L$-Lipschitz with respect to $\ell_1$ norm. We want to be competitive against the following benchmark

$\max x f(E^{x^*}[\sum_k r_s(s_k, a_k)]) s.t. g(E^{x^*}[\sum_k r_s(s_k, a_k, i)]) \leq 0$

Algorithm: We can no longer create a bonus enhanced model as in 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 $\ell_1(\sqrt{AH^2 \cdot \kappa^2})$ and $\ell_2(\sqrt{AH^2 \cdot \kappa^2})$ 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 maximize the total reward across K episodes while not exceeding the hard thresholds.

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

$\max x E^{x^*}[\sum_k r_s(s_k, a_k)] s.t. \forall i \in \mathcal{S} : E^{x^*}[\sum_k r_s(s_k, a_k, i)] \leq (1 - c)B_i$

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

Main Result: Set $\epsilon$ properly, with high probability, the reward regret over $K$ episodes is at most $O\left(HS^2/HAK \cdot \min_{B_i} B_i\right)$ and the hard constraints are not violated.

(Compare to prior work, our budget can be as small as $\min_{B_i} B_i = \tilde{O}(\epsilon^{1/3})$)

Experiments

The performance of the algorithms as function of number of sample trajectories (trajectory * 50 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-episode approach.

References