Learning with Safety Constraints: Sample Complexity of Reinforcement Learning for Constrained MDPs

Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, Srinivas Shakkottai

T3: TEXAS A&M TRIADS FOR TRANSFORMATION
A President's Excellence Fund Initiative

Introduction

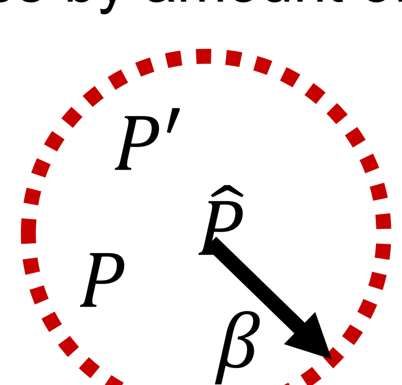
- Markov Decision Processes (MDPs) are useful to model real-world stochastic systems
  - Finding shortest path in a grid world
- However, there are physical limitations in many cases
  - Automated vehicles with no-collision constraint (safety)
  - A robot avoiding hitting walls while wandering around (safety)
  - Communication networks with link capacity constraints (transmitter safety constraints)
- Modeled by Constrained Markov Decision Processes

Constrained Markov Decision Process

- A finite-horizon CMDP is a tuple $M = < S, A, P, r, c, \bar{C}, s_0, H >$
  - S: state space. A: action space. P: transition kernel.
  - r: immediate reward matrix. c: immediate cost matrix.
  - $\bar{C}$: constraint bound with N constraints. s_0: initial state
  - H: horizon length
- Value function for CMDP M under a policy π:
  - $V_0^\pi(s_0) = \mathbb{E}[\sum_{h=0}^{H-1} r(s_h, a_h) | a_h \sim \pi(s_h, ., h)]$
- Constraint function i for CMDP M under a policy π:
  - $C_{i,0}^\pi(s_0) = \mathbb{E}[\sum_{h=0}^{H-1} c(i, s_h, a_h) | a_h \sim \pi(s_h, ., h)]$
- We solve
  - $\max_\pi V_0^\pi(s_0) \quad s.t. \quad C_{i,0}^\pi(s_0) \le \bar{C}_i \quad \forall i = \{1, \dots, N\}$
- Assumption: Problem is feasible
  - Solution to this problem may not be a deterministic policy [1]
  - Also depends on initial state distribution [1]

Constrained Reinforcement Learning

- Constrained-RL problem formulation is identical to CMDP problem, but without knowing system parameters
- A naïve way is to sample each state-action and obtain $\hat{P}$
- This approach works for unconstrained MDPs
- A CMDP with estimated model might not necessarily be feasible
- Need to expand the transition kernel space by amount of β and solve "Optimistic Planning" problem



- Thus, the problem would become feasible with high probability

CRL Solution Overview

- Here, we present two model-based algorithms
  - Offline: Optimistic Generative Model Based Learning, Optimistic-GMBL
  - Online: Online Constrained Reinforcement Learning, Online-CRL
- Both algorithms solve "Optimistic Planning" problem below
  - $\max_{M', \pi} V'^\pi_0(s_0) \quad s.t. \quad C'^\pi_{i,0}(s_0) \le \bar{C}_i \quad \forall i = \{1, \dots, N\}$
  - V' and C'_i are defined with respect to any P' inside the expanded transition kernel space

Optimistic-GMBL

- Input ϵ and δ
- Set visitation frequencies to 0
- **for** each (s, a)**:**
  - Sample that (s, a), $\frac{256\,|S|H^3}{\epsilon^2} \log \frac{12(N+2)|S||A|H}{\delta}$
  - Construct estimated transition kernel $\hat{P}$
- Construct class of CMDPs using $\hat{P}$ and inputs of algorithm
- Solve Optimistic Planning problem

Optimistic- GMBL satisfies the PAC result with sampling budget of

$$O\left(\frac{|S|^2 |A| H^3}{\epsilon^2} \log \frac{N}{\delta}\right)$$

Online-CRL

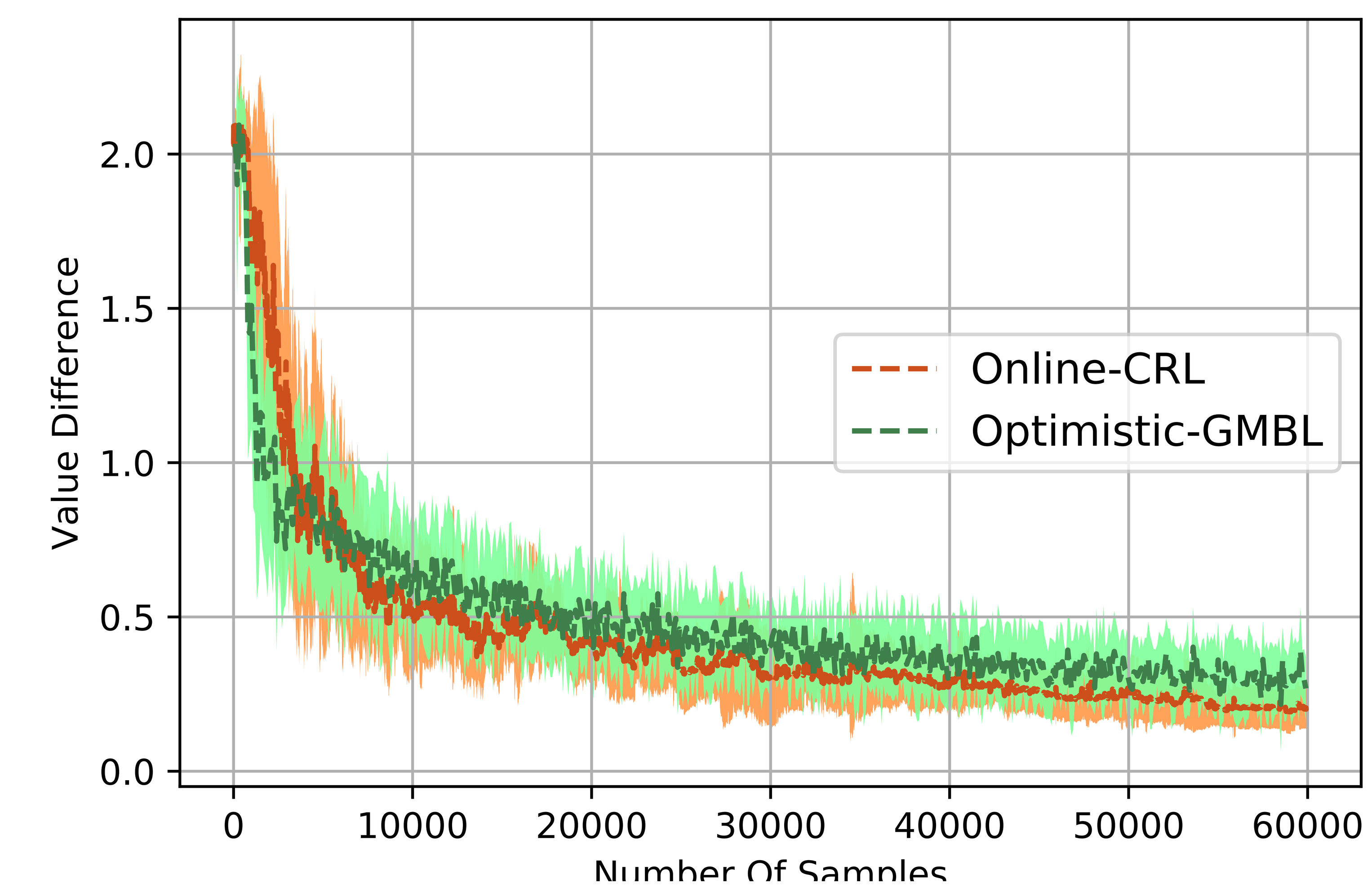
- Input ϵ and δ
- Set visitation frequencies to 0
- **while** there is (s, a) with less visitation frequency:
  - Construct estimated transition kernel $\hat{P}$
  - Construct class of CMDPs using $\hat{P}$ and inputs of algorithm
  - Solve Optimistic Planning problem
  - Employ the optimistic policy and collect data to update visitation frequencies

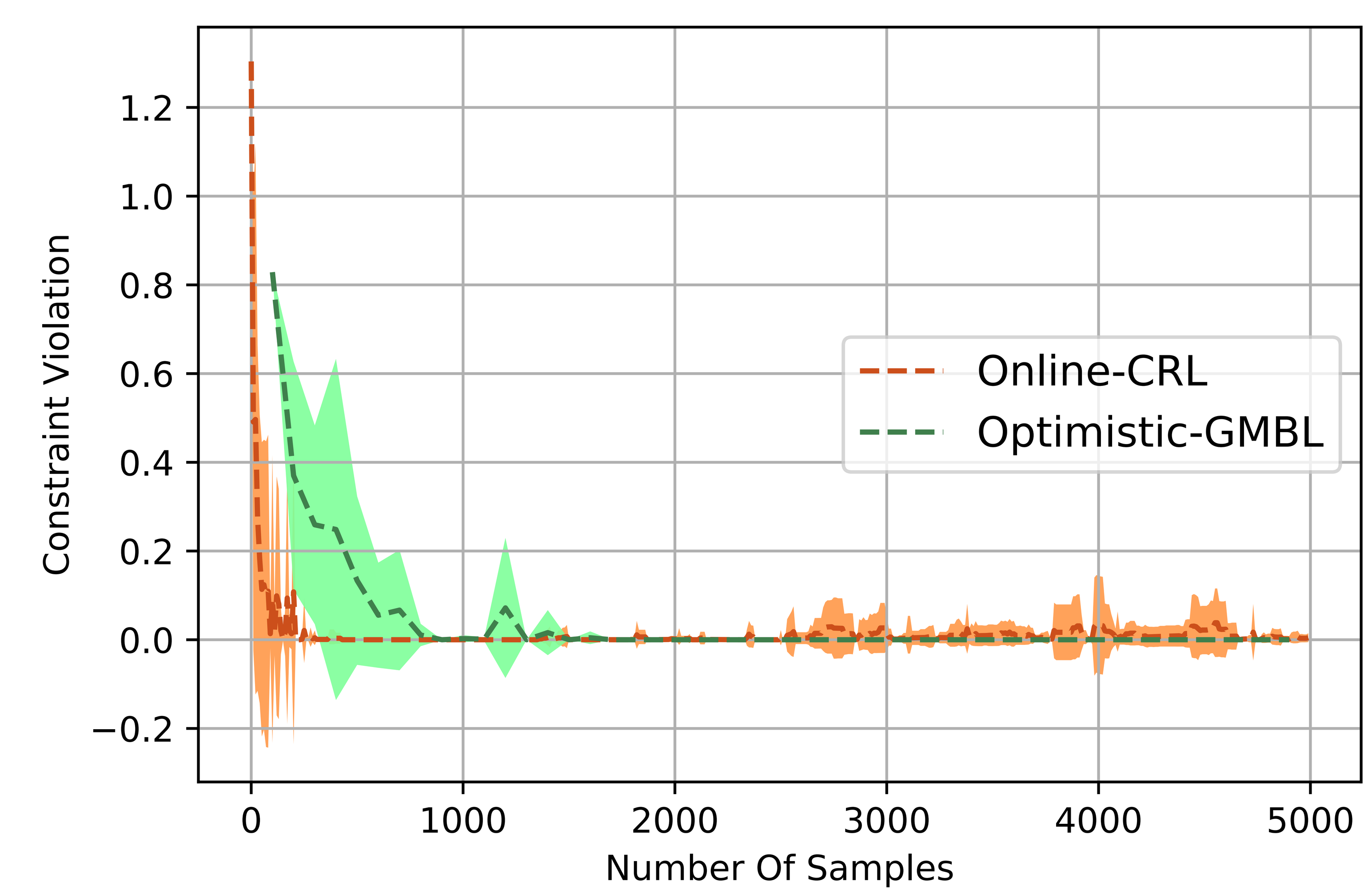
Online-CRL satisfies the PAC result with sampling budget of

$$O\left(\frac{|S|^2 |A| H^3}{\epsilon^2} \log \frac{N}{\delta}\right)$$

Experimental Result

- 5 × 5 Grid Network
- Horizon length of 10
- Use of action "Right" is limited by 2
- Online-CRL and Optimistic-GMBL have equal performance in terms of Value function
- Online-CRL is requires less sampling budget compared to Optimistic-GMBL in terms of Constraint violation





References

[1] Altman, Eitan. Constrained Markov decision processes. Vol. 7. CRC Press, 1999.