Learning Infinite-horizon Average-reward Markov Decision Process with Constraints

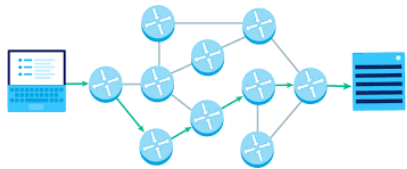
Liyu Chen, Rahul Jain, Haipeng Luo

University of Southern California

July 11, 2022

Motivation

In many real-world Reinforcement Learning (RL) applications, certain constraints need to be enforced along with reward maximization.



Motivation

In many real-world Reinforcement Learning (RL) applications, certain constraints need to be enforced along with reward maximization.

Constrained Markov Decision Process (CMDP): while the finite-horizon setting or discounted setting has received great attention, the infinite-horizon average-reward setting is much much less understood.

Our Contributions

We further extend our understanding of infinite-horizon average-reward CMDP.

	Assumption	Regret	Constraint Violation	Efficient
(Singh et al., 2020)	Ergodic	$T^{2/3}$	$T^{2/3}$	Yes
Ours	Ergodic	$\sqrt{T}$	Constant	Yes
	WC	$T^{2/3}$	$T^{2/3}$	Yes
	WC	$\sqrt{T}$	$\sqrt{T}$	No

WC: weakly-communicating.

Problem Formulation: Infinite-Horizon Average-Reward CMDP

The model is defined as tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, c, \tau, P)$. We assume only P is unknown.

Problem Formulation: Infinite-Horizon Average-Reward CMDP

The model is defined as tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, c, \tau, P)$. We assume only P is unknown.

learner starts in an arbitrary initial state $s_1 \in \mathcal{S}$.
for $t = 1, \ldots, T$ **do**
    learner observes state s_t, takes action $a_t \in \mathcal{A}$, and transits to the next state $s_{t+1} \sim P_{s_t, a_t}$.

Problem Formulation: Infinite-Horizon Average-Reward CMDP

- Average utility function: $J^{\pi,P,d}(s) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} d(s_t, a_t)|\pi, P, s_1 = s]$.

Problem Formulation: Infinite-Horizon Average-Reward CMDP

- Average utility function: $J^{\pi,P,d}(s) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} d(s_t, a_t)|\pi, P, s_1 = s]$.
- Optimal policy $\pi^\star$ is the solution of the following optimization problem:

$$\underset{\pi\in(\Delta_{\mathcal{A}})^S}{\text{argmax}}\ J^{\pi,P,r}(s), \quad \text{s.t. } J^{\pi,P,c}(s) \le \tau,$$

and also $J^{\pi^\star,P,r}(s) = J^\star$ for some constant $J^\star$.

Problem Formulation: Infinite-Horizon Average-Reward CMDP

- Average utility function: $J^{\pi,P,d}(s) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} d(s_t, a_t)|\pi, P, s_1 = s]$.
- Optimal policy $\pi^\star$ is the solution of the following optimization problem:

$$\operatorname*{argmax}_{\pi\in(\Delta_{\mathcal{A}})^S} J^{\pi,P,r}(s), \quad \text{s.t. } J^{\pi,P,c}(s) \le \tau,$$

and also $J^{\pi^\star,P,r}(s) = J^\star$ for some constant $J^\star$.

Learning Objective: ensure large reward while at the same time incurring small cost relative to the threshold τ.

$$\text{Regret: } R_T = \sum_{t=1}^{T}(J^\star - r(s_t, a_t)), \quad \text{Constraint Violation: } C_T = \sum_{t=1}^{T}(c(s_t, a_t) - \tau).$$

Results for Ergodic MDPs

Main Ideas: perform policy optimization to update policy incrementally.

Results for Ergodic MDPs

Main Ideas: perform policy optimization to update policy incrementally.

- Policy Evaluation: leverage ergodicity to estimate performance of learner's policy.

Results for Ergodic MDPs

Main Ideas: perform policy optimization to update policy incrementally.

- Policy Evaluation: leverage ergodicity to estimate performance of learner's policy.
- Policy Improvement: a novel value function estimator, and a new bonus term to deal with the error of transition estimation in the value function estimator.

Results for Ergodic MDPs

Main Ideas: perform policy optimization to update policy incrementally.

- Policy Evaluation: leverage ergodicity to estimate performance of learner's policy.
- Policy Improvement: a novel value function estimator, and a new bonus term to deal with the error of transition estimation in the value function estimator.
- Constraint Violation: primal-dual approach plus cost slack to achieve constant constraint violation.

Results for Ergodic MDPs

Main Ideas: perform policy optimization to update policy incrementally.

- Policy Evaluation: leverage ergodicity to estimate performance of learner's policy.
- Policy Improvement: a novel value function estimator, and a new bonus term to deal with the error of transition estimation in the value function estimator.
- Constraint Violation: primal-dual approach plus cost slack to achieve constant constraint violation.

Theorem

The described algorithm ensures $R_T = \tilde{\mathcal{O}}(\sqrt{T})$ *and* $C_T = \tilde{\mathcal{O}}(1)$.

This improves the $\tilde{\mathcal{O}}(T^{2/3})$ bounds of (Singh et al., 2020) in both metrics.

Results for Weakly Communicating MDPs

Main Idea: Finite-horizon approximation similar to (Wei et al., 2020).

- Divide T time steps into K episodes, each of length H.
- Treat each episode as interacting with a finite-horizon MDP, and solve it through the len of occupancy measure, where the expected reward and cost are both linear functions and thus easy to optimize.

Results for Weakly Communicating MDPs

Theorem

The described algorithm ensures $R_T = \tilde{\mathcal{O}}(T^{2/3})$ *and* $C_T = \tilde{\mathcal{O}}(T^{2/3})$.

Bottleneck of the Analysis: the span of value functions are bounded by H.

Results for Weakly Communicating MDPs

Theorem

The described algorithm ensures $R_T = \tilde{\mathcal{O}}(T^{2/3})$ *and* $C_T = \tilde{\mathcal{O}}(T^{2/3})$.

Bottleneck of the Analysis: the span of value functions are bounded by H.
Observation: the span of the value functions of optimal policy are properly bounded by smaller quantities.

Results for Weakly Communicating MDPs

Theorem

The described algorithm ensures $R_T = \tilde{\mathcal{O}}(T^{2/3})$ *and* $C_T = \tilde{\mathcal{O}}(T^{2/3})$.

Bottleneck of the Analysis: the span of value functions are bounded by H.
Observation: the span of the value functions of optimal policy are properly bounded by smaller quantities.

Theorem

Incorporating the constraints on the span of the value functions ensures $R_T = \tilde{\mathcal{O}}(\sqrt{T})$ *and* $C_T = \tilde{\mathcal{O}}(\sqrt{T})$.

Limitation: the algorithm is inefficient.

Conclusion

We further extend our understanding of infinite-horizon average-reward CMDP.

	Assumption	Regret	Constraint Violation	Efficient
(Singh et al., 2020)	Ergodic	$T^{2/3}$	$T^{2/3}$	Yes
Ours	Ergodic	$\sqrt{T}$	Constant	Yes
	WC	$T^{2/3}$	$T^{2/3}$	Yes
	WC	$\sqrt{T}$	$\sqrt{T}$	No

WC: weakly-communicating.