Regret Bounds for Stochastic Shortest Path Problems with Linear Function Approximation

Daniel Vial[*†], Advait Parulekar[*], Sanjay Shakkottai[*], R. Srikant[†]

[*]UT Austin, [†]UIUC

ICML 2022

Episodic stochastic shortest path (SSP)

SSP instance is defined by $(\mathcal{S}, \mathcal{A}, P, c, s_{init}, s_{goal})$, where

- $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces
- P and c are transition kernel and cost vector
- s_{init} and s_{goal} are initial and goal states

Episode k begins at state s_{init}, and for each $h = 1, 2, \ldots$

- agent chooses action $a_h^k \in \mathcal{A}$
- environment reveals cost $c(s_h^k, a_h^k)$ and next state $s_{h+1}^k \sim P(\cdot | s_h^k, a_h^k)$
- if $s_{h+1}^k = s_{goal}$, episode terminates

Applications: goal oriented tasks like games, navigation, etc.

Objective

Learner aims to minimize the K-episode regret

$$R(K) = \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H_k} c(s_h^k, a_h^k)}_{\text{total cost for learner}} - \underbrace{K J^\star(s_{init})}_{\substack{\mathbb{E}[\text{total cost}] \text{ for} \\ K \text{ episodes of opt}}}$$

Rosenberg et al. 2020 prove $R(K) \geq \Omega(\sqrt{SAK})$ for any algorithm

- plus near-optimal algorithms (see also Tarbouriech et al. 2020; Cohen et al. 2021; Tarbouriech et al. 2021; Chen et al. 2021; Jafarnia-Jahromi et al. 2021)

$poly(S, A)$ prohibitive in practice, so use function approximation

Linear function approximation

We consider the simplest case of linear MDPs, where

$$c(s,a) = \langle \phi(s,a), \theta \rangle, \quad P(\cdot|s,a) = \langle \phi(s,a), \mu(\cdot) \rangle$$

- $\phi(s,a) \in \mathbb{R}^d$ are known features, θ and $\mu(\cdot)$ are unknown
- see Jin et al. 2020 and (extensive) subsequent work

Under this assumption (and mild regularity conditions), optimal policy is

$$\pi^\star(s) = \arg\min_{a \in \mathcal{A}} \langle \phi(s,a), w^\star \rangle,$$

where $w^\star$ = fixed point of feature space Bellman operator $G : \mathbb{R}^d \to \mathbb{R}^d$

Main result

Basic idea of our algorithm:

- least-squares + value iteration to compute an optimistic estimate $\hat{w}$ of $w^\star$ (analogous to Jin et al. 2020 for finite horizons)
- update policy $\hat{\pi}(s) \leftarrow \arg\min_{a \in \mathcal{A}} \langle \phi(s,a), \hat{w} \rangle$

Theorem

If the above assumptions hold, our algorithm satisfies

$$R(K) \leq \tilde{O}\left(\sqrt{B_\star^3 d^3 K / c_{min}} + B_\star^2 d^3 / c_{min}\right),$$

where $B_\star = \max_{s \in \mathcal{S}} J^\star(s)$, $d =$ *feature dimension*, $K = \#$ *episodes, and* $c_{min} =$ *minimal cost of non-goal states (assumed* > 0*)*

Similar bounds in concurrent works Min et al. 2022; Chen et al. 2022

Thanks!

References:

Chen, Liyu et al. (2021). "Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path". In: *NeurIPS*.

Chen, Liyu, Rahul Jain, Haipeng Luo (2022). "Improved No-Regret Algorithms for Stochastic Shortest Path with Linear MDP". In: *ICML*.

Cohen, Alon et al. (2021). "Minimax regret for stochastic shortest path". In: *NeurIPS*.

Jafarnia-Jahromi, Mehdi et al. (2021). "Online Learning for Stochastic Shortest Path Model via Posterior Sampling". In: *arXiv preprint arXiv:2106.05335*.

Jin, Chi et al. (2020). "Provably efficient reinforcement learning with linear function approximation". In: *COLT*.

Min, Yifei et al. (2022). "Learning stochastic shortest path with linear function approximation". In: *ICML*.

Rosenberg, Aviv et al. (2020). "Near-optimal regret bounds for stochastic shortest path". In: *ICML*.

Tarbouriech, Jean et al. (2020). "No-regret exploration in goal-oriented reinforcement learning". In: *ICML*.

Tarbouriech, Jean et al. (2021). "Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret". In: *NeurIPS*.