FINDING THE STOCHASTIC SHORTEST PATH WITH LOW REGRET: THE ADVERSARIAL COST AND UNKNOWN TRANSITION CASE

Liyu Chen & Haipeng Luo

Liyu Chen

October 8, 2021

University of Southern California

Problem Formulation

SSP with adversarial costs: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + costs $\{c_k\}_{k=1}^{K}$

Problem Formulation

SSP with adversarial costs: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + costs $\{c_k\}_{k=1}^K$

Learning Protocol (for $k = 1, \ldots, K$)

1. environment chooses c_k possibly based on learner's algorithm and history;

Problem Formulation

SSP with adversarial costs: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + costs $\{c_k\}_{k=1}^{K}$

Learning Protocol (for $k = 1, \ldots, K$)

1. environment chooses c_k possibly based on learner's algorithm and history;
2. learner starts in state $s_k^1 = s_0, i \leftarrow 1$;
3. learner sequentially takes action a_k^i, observes states s_k^{i+1}, and increases counter $i \leftarrow i + 1$ until $s_k^i = g$;

Problem Formulation

SSP with adversarial costs: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + costs $\{c_k\}_{k=1}^K$

Learning Protocol (for $k = 1, \ldots, K$)

1. environment chooses c_k possibly based on learner's algorithm and history;
2. learner starts in state $s_k^1 = s_0, i \leftarrow 1$;
3. learner sequentially takes action a_k^i, observes states s_k^{i+1}, and increases counter $i \leftarrow i + 1$ until $s_k^i = g$;
4. learner observes c_k (full information) or $\{c(s_k^i, a_k^i)\}_{i=1}^{I_k}$ (bandit feedback) and suffer cost $\sum_{i=1}^{I_k} c(s_k^i, a_k^i)$.

Problem Formulation

SSP with adversarial costs: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + costs $\{c_k\}_{k=1}^K$

Learning Protocol (for $k = 1, \ldots, K$)

1. environment chooses c_k possibly based on learner's algorithm and history;
2. learner starts in state $s_k^1 = s_0$, $i \leftarrow 1$;
3. learner sequentially takes action a_k^i, observes states s_k^{i+1}, and increases counter $i \leftarrow i+1$ until $s_k^i = g$;
4. learner observes c_k (full information) or $\{c(s_k^i, a_k^i)\}_{i=1}^{I_k}$ (bandit feedback) and suffer cost $\sum_{i=1}^{I_k} c(s_k^i, a_k^i)$.

Objective: minimize regret w.r.t. the best stationary proper policy in hindsight ($\pi^\star = \mathrm{argmin}_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^K J_k^{\pi^\star}(s_0)$)

$$R_K = \sum_{k=1}^K \left(\sum_{i=1}^{I_k} c_k(s_k^i, a_k^i) - J_k^{\pi^\star}(s_0) \right).$$

Existing Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes

$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

$B_\star$: upper bound on the expected cost of optimal policy

SSP with stochastic costs (Tarbouriech et al., 2020; Cohen et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021): $\Theta\left(B_\star\sqrt{SAK}\right)$.

Existing Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes

$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

$B_\star$: upper bound on the expected cost of optimal policy

SSP with stochastic costs (Tarbouriech et al., 2020; Cohen et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021): $\Theta\left(B_\star\sqrt{SAK}\right)$.

SSP with adversarial costs:

- (Rosenberg and Mansour, 2020): $\tilde{\mathcal{O}}\left(\frac{D}{c_{\min}}\sqrt{K}\right)$ or $\tilde{\mathcal{O}}\left(T_\star K^{3/4}\right)$ with known transition; $\tilde{\mathcal{O}}\left(\frac{DS}{c_{\min}}\sqrt{AK}\right)$ or $\tilde{\mathcal{O}}\left(T_\star S\sqrt{A}K^{3/4}\right)$ with unknown transition.

Existing Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes

$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

$B_\star$: upper bound on the expected cost of optimal policy

SSP with stochastic costs (Tarbouriech et al., 2020; Cohen et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021): $\Theta\left(B_\star\sqrt{SAK}\right)$.

SSP with adversarial costs:

- (Rosenberg and Mansour, 2020): $\tilde{\mathcal{O}}\left(\frac{D}{c_{\min}}\sqrt{K}\right)$ or $\tilde{\mathcal{O}}\left(T_\star K^{3/4}\right)$ with known transition; $\tilde{\mathcal{O}}\left(\frac{DS}{c_{\min}}\sqrt{AK}\right)$ or $\tilde{\mathcal{O}}\left(T_\star S\sqrt{A}K^{3/4}\right)$ with unknown transition.
- (Chen et al., 2021): With known transition, $\Theta\left(\sqrt{DT_\star K}\right)$ in the full information setting, and $\Theta\left(\sqrt{DT_\star SAK}\right)$ in the bandit feedback setting.

Our Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes

$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

	Full information	Bandit feedback
Adaptive adversary	$\tilde{\mathcal{O}}\left(\sqrt{S^2ADT_\star K}\right)$	
Stochastic adversary		
Lower Bounds	$\Omega\left(\sqrt{DT_\star K} + D\sqrt{SAK}\right)$	$\Omega\left(\sqrt{SADT_\star K} + D\sqrt{SAK}\right)$

Our contributions:

1. strictly improve (Rosenberg and Mansour, 2020) in the full information setting;

Our Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes

$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

	Full information	Bandit feedback
Adaptive adversary	$\tilde{\mathcal{O}}\left(\sqrt{S^2ADT_\star K}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{S^3A^2DT_\star K}\right)$
Stochastic adversary		
Lower Bounds	$\Omega\left(\sqrt{DT_\star K} + D\sqrt{SAK}\right)$	$\Omega\left(\sqrt{SADT_\star K} + D\sqrt{SAK}\right)$

Our contributions:

1. strictly improve (Rosenberg and Mansour, 2020) in the full information setting;
2. the first result in the most challenging bandit feedback, unknown transition setting;

Our Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes

$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

	Full information	Bandit feedback
Adaptive adversary	$\tilde{\mathcal{O}}\left(\sqrt{S^2ADT_\star K}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{S^3A^2DT_\star K}\right)$
Stochastic adversary	$\tilde{\mathcal{O}}\left(\sqrt{DT_\star K} + DS\sqrt{AK}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{SADT_\star K} + DS\sqrt{AK}\right)$
Lower Bounds	$\Omega\left(\sqrt{DT_\star K} + D\sqrt{SAK}\right)$	$\Omega\left(\sqrt{SADT_\star K} + D\sqrt{SAK}\right)$

Our contributions:

1. strictly improve (Rosenberg and Mansour, 2020) in the full information setting;
2. the first result in the most challenging bandit feedback, unknown transition setting;
3. achieve near optimal regret under the weaker stochastic adversary.

Techniques

1. Extend the loop-free reduction of (Chen et al., 2020) to the unknown transition setting.

Techniques

1. Extend the loop-free reduction of (Chen et al., 2020) to the unknown transition setting.
2. Introduce a data dependent bound on the transition estimation error, which can be controlled by the skewed occupancy measure introduced in (Chen et al., 2020).

Techniques

1. Extend the loop-free reduction of (Chen et al., 2020) to the unknown transition setting.
2. Introduce a data dependent bound on the transition estimation error, which can be controlled by the skewed occupancy measure introduced in (Chen et al., 2020).
3. For the bandit feedback setting, we further propose and utilize two optimistic cost estimators inspired by the idea of upper occupancy bounds from (Jin et al., 2020) for loop-free SSP.

Techniques

1. Extend the loop-free reduction of (Chen et al., 2020) to the unknown transition setting.
2. Introduce a data dependent bound on the transition estimation error, which can be controlled by the skewed occupancy measure introduced in (Chen et al., 2020).
3. For the bandit feedback setting, we further propose and utilize two optimistic cost estimators inspired by the idea of upper occupancy bounds from (Jin et al., 2020) for loop-free SSP.
4. For the weaker stochastic adversaries, we augment the loop-free reduction to allow the learner to switch to a fast policy (to reach goal in shortest time) at any time if necessary.

SEE YOU IN THE POSTER SESSION!