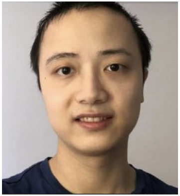
Learning Stochastic Shortest Path with Linear Function Approximation

Yifei Min[1] Jiafan He[2] Tianhao Wang[1] Quanquan Gu[2]

[1]Department of Statistics and Data Science, Yale
[2]Department of Computer Science, UCLA

ICML 2022

Stochastic Shortest Path (SSP)

- Online SSP: a type of goal-oriented RL problem
  - Episodic interaction: each episode starts from an initial state and ends when the agent reaches the goal state g
  - Cost: each state-action pair (s, a) incurs a cost $c(s, a)$
  - Goal: to minimize the cumulative cost over all episodes

Stochastic Shortest Path (SSP)

- Online SSP: a type of goal-oriented RL problem
  - Episodic interaction: each episode starts from an initial state and ends when the agent reaches the goal state g
  - Cost: each state-action pair (s, a) incurs a cost $c(s, a)$
  - Goal: to minimize the cumulative cost over all episodes
- SSP is a generalization of episodic finite-horizon MDPs and discounted infinite-horizon MDPs
  - *The horizon length varies across episodes, and can be random*

Stochastic Shortest Path (SSP)

- Online SSP: a type of goal-oriented RL problem
  - Episodic interaction: each episode starts from an initial state and ends when the agent reaches the goal state g
  - Cost: each state-action pair (s, a) incurs a cost $c(s, a)$
  - Goal: to minimize the cumulative cost over all episodes
- SSP is a generalization of episodic finite-horizon MDPs and discounted infinite-horizon MDPs
  - *The horizon length varies across episodes, and can be random*
- Beyond tabular SSP: linear function approximation
  - Existing works on tabular SSP (Rosenberg et al. 2020; Cohen et al. 2021; Tarbouriech et al. 2021, ...)
  - Linear mixture SSP: assume that there exists an *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle$
  - Linear mixture model is common in RL literature (Ayoub et al. 2020; Zhou et al. 2021b, ...)

Stochastic Shortest Path (SSP)

- Online SSP: a type of goal-oriented RL problem
  - Episodic interaction: each episode starts from an initial state and ends when the agent reaches the goal state g
  - Cost: each state-action pair (s, a) incurs a cost $c(s, a)$
  - Goal: to minimize the cumulative cost over all episodes
- SSP is a generalization of episodic finite-horizon MDPs and discounted infinite-horizon MDPs
  - *The horizon length varies across episodes, and can be random*
- Beyond tabular SSP: linear function approximation
  - Existing works on tabular SSP (Rosenberg et al. 2020; Cohen et al. 2021; Tarbouriech et al. 2021, ...)
  - Linear mixture SSP: assume that there exists an *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle$
  - Linear mixture model is common in RL literature (Ayoub et al. 2020; Zhou et al. 2021b, ...)

This work: efficiently learn linear mixture SSP

Linear Mixture SSP: Algorithmic Design

- Two approaches for SSP in existing literature:
  - By reduction to finite-horizon MDP (Cohen et al. 2021; Chen et al. 2021, ...)
  - By (implicitly) viewing SSP as an infinite-horizon problem (Tarbouriech et al. 2021; Vial et al. 2021, ...)

Linear Mixture SSP: Algorithmic Design

- Two approaches for SSP in existing literature:
  - By reduction to finite-horizon MDP (Cohen et al. 2021; Chen et al. 2021, ...)
  - By (implicitly) viewing SSP as an infinite-horizon problem (Tarbouriech et al. 2021; Vial et al. 2021, ...)
- `LEVIS`: a novel optimistic value-iteration algorithm for linear mixture SSP
  - Model estimate updating criteria: coupling features with time
    - Determinant-doubling + time-step-doubling
  - Optimistic planning: contraction via perturbation
    - There is no discount factor in SSP $\rightarrow$ no contraction for EVI
    - Introduce an auxiliary discount factor by perturbing the transition probability

Linear Mixture SSP: Algorithm

Algorithm 1 `LEVIS`

for episode $k = 1, 2, \ldots, K$ **do**
  while $s_t \neq g$ **do**
    Greedily take action a_t, and receive $c(s_t, a_t)$ and s_{t+1}
    $\boldsymbol{\Sigma}_t \leftarrow \boldsymbol{\Sigma}_{t-1} + \phi_V(s_t, a_t)\phi_V(s_t, a_t)^\top$
    if $\det(\boldsymbol{\Sigma}_t)$ or t doubles **then**
      Update model estimate $\widehat{\boldsymbol{\theta}}$ and its confidence region
      Call `DEVI` to update estimate of the value functions

Algorithm 2 `DEVI`

while $\|V^{(i)} - V^{(i-1)}\|_\infty \geq \epsilon$ **do**
  $Q^{(i+1)}(\cdot, \cdot) \leftarrow c_\rho(\cdot, \cdot) + (1 - q) \min\langle \boldsymbol{\theta}, \phi_{V^{(i)}}(\cdot, \cdot)\rangle$
  $V^{(i+1)}(\cdot) \leftarrow \min_a Q^{(i+1)}(\cdot, a)$

- Determinant-doubling + time-step-doubling
- Perturb the transition probability

Linear Mixture SSP: Theory

Theorem (Regret upper bound)

Under technical assumptions, the proposed algorithm `LEVIS` *achieves a* $\widetilde{\mathcal{O}}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ *regret, where* d *is the feature dimension,* $B_\star$ *is the cost of the optimal policy,* $c_{\min} > 0$ *is the lower bound of the per-step cost.*

Linear Mixture SSP: Theory

Theorem (Regret upper bound)

Under technical assumptions, the proposed algorithm `LEVIS` achieves a $\widetilde{\mathcal{O}}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ *regret, where* d *is the feature dimension,* $B_\star$ *is the cost of the optimal policy,* $c_{\min} > 0$ *is the lower bound of the per-step cost.*

Theorem (Regret lower bound)

Under technical assumptions, any algorithm for linear mixture SSP incurs at least an expected regret of $\Omega(dB_\star\sqrt{K})$.

Linear Mixture SSP: Theory

Theorem (Regret upper bound)

Under technical assumptions, the proposed algorithm `LEVIS` *achieves a* $\widetilde{\mathcal{O}}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ *regret, where* d *is the feature dimension,* $B_\star$ *is the cost of the optimal policy,* $c_{\min} > 0$ *is the lower bound of the per-step cost.*

Theorem (Regret lower bound)

Under technical assumptions, any algorithm for linear mixture SSP incurs at least an expected regret of $\Omega(dB_\star\sqrt{K})$.

- There is a $\sqrt{B_\star}$-gap between the upper and lower bound. How to do better?

Linear Mixture SSP: Near-optimal Regret

- Design Bernstein-type confidence region to reduce the dependence on $B_\star$
  - Similar technique has been used in online/offline RL (Zhou et al. 2021a; Zhang et al. 2021; Min et al. 2021, ...)

Linear Mixture SSP: Near-optimal Regret

- Design Bernstein-type confidence region to reduce the dependence on $B_\star$
  - Similar technique has been used in online/offline RL (Zhou et al. 2021a; Zhang et al. 2021; Min et al. 2021, ...)

Theorem (Near-optimal regret bound)

Under technical assumptions, by using a refined Bernstein-type confidence region in algorithm `LEVIS`*, it can achieve* $\widetilde{\mathcal{O}}(dB_\star\sqrt{K/c_{\min}})$ *regret.*

Linear Mixture SSP: Near-optimal Regret

- Design Bernstein-type confidence region to reduce the dependence on $B_\star$
  - Similar technique has been used in online/offline RL (Zhou et al. 2021a; Zhang et al. 2021; Min et al. 2021, ...)

Theorem (Near-optimal regret bound)

Under technical assumptions, by using a refined Bernstein-type confidence region in algorithm `LEVIS`*, it can achieve* $\widetilde{\mathcal{O}}(dB_\star\sqrt{K/c_{\min}})$ *regret.*

- There is still a remaining gap of $1/\sqrt{c_{\min}}$
- Future work: how to remove the dependence on $c_{\min}$?

THANK YOU!

Reference:

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M. and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

Chen, L., Jafarnia-Jahromi, M., Jain, R. and Luo, H. (2021). Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems* **34** 10849–10861.

Cohen, A., Efroni, Y., Mansour, Y. and Rosenberg, A. (2021). Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems* **34** 28350–28361.

Min, Y., Wang, T., Zhou, D. and Gu, Q. (2021). Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems* **34** 7598–7610.

Rosenberg, A., Cohen, A., Mansour, Y. and Kaplan, H. (2020). Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*. PMLR.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M. and Lazaric, A. (2021). Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems* **34** 6843–6855.

Vial, D., Parulekar, A., Shakkottai, S. and Srikant, R. (2021). Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593* .

ZHANG, Z., YANG, J., JI, X. and DU, S. S. (2021). Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems* **34**.

ZHOU, D., GU, Q. and SZEPESVARI, C. (2021a). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

ZHOU, D., HE, J. and GU, Q. (2021b). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.