Learning Stochastic Shortest Path with Linear Function Approximation

Yale UCLA

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

[1]Dept. of Statistics and Data Science, Yale University [2]Dept. of Computer Science, UCLA

ICML

Problem Setting

- **Online Stochastic Shortest Path (SSP)**: $M(\mathcal{S}, \mathcal{A}, \mathbb{P}, c, s_{\text{init}}, g)$.
  - State space $\mathcal{S}$, action space $\mathcal{A}$
  - Initial state s_{init}, **goal state** g
  - Cost function $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$
    - the goal state incurs zero cost, i.e., $c(g, \cdot) \equiv 0$
  - Transition probability function $\mathbb{P}(s'|s, a)$
    - the goal state is an absorbing state, i.e., $\mathbb{P}(g|g, \cdot) \equiv 1$
  - *SSP is a generalization of episodic finite-horizon MDPs and discounted infinite-horizon MDPs: the horizon length varies across episodes, and can be random*
- **Policy**: A policy π is a map from $\mathcal{S}$ to $\mathcal{A}$
- **Goal**: Minimize the cumulative cost over all episodes
- **Value function**: Expected accumulative cost for policy π: $V^\pi(s) = \lim_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} c(s_t, \pi(s_t))|s_1 = s]$
- **Regret**: The sum of sub-optimality over K episodes:

$$R_K := \sum_{k=1}^{K}\sum_{i=1}^{I_k} c_{k,i} - K \cdot V^{\pi^*}(s_{\text{init}})$$

I_k: length of the k-th episode, π^*: the optimal policy

- **Linear mixture SSP**: There exists an *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle$, where ϕ is some known d-dimensional feature mapping

Our Contributions

- We develop 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
    - Introduce an auxiliary discount factor by perturbing the transition kernel
- A regret upper bound for LEVIS with Hoeffding-type bonus (a simple algorithm)
- A near-optimal regret upper bound for LEVIS with Bernstein-type bonus (a more complicate algorithm)

Main Results: Algorithm

- Notation:
  - β_t: confidence radius, q: transition bonus, ρ : cost perturbation
  - $\boldsymbol{\phi}_V(s, a) = \int_{\mathcal{S}} \boldsymbol{\phi}(s'|s, a)V(s')\mathrm{d}s'$, $c_\rho(s, a) = \max\{c(s, a), \rho\}$
- **Main algorithm:**

Algorithm 1 LEVIS

for episode $k = 1, 2, \dots, K$ **do**
  while $s_t \neq g$ **do**
    Take action a_t greedy w.r.t. the Q function
    Receive $c(s_t, a_t)$ and s_{t+1}
    $\Sigma_t \leftarrow \Sigma_{t-1} + \boldsymbol{\phi}_V(s_t, a_t)\boldsymbol{\phi}_V(s_t, a_t)^\top$
    $\mathrm{b}_t \leftarrow \mathrm{b}_{t-1} + \boldsymbol{\phi}_V(s_t, a_t)V(s_{t+1})$
    if $\det(\Sigma_t)$ or t doubles **then**
      Update model estimate $\widehat{\boldsymbol{\theta}} = \Sigma_t^{-1}\mathrm{b}_t$
      Update the confidence region of $\hat{\boldsymbol{\theta}}$
      $$\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_{\Sigma_t} \leq \beta_t\}$$
      $Q(\cdot,\cdot) \leftarrow \text{DEVI}(\mathcal{C}, 1/t, 1/t, \rho)$
      $V(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q(s, \cdot)$
    end if
    $t \leftarrow t + 1$
  end while
end for

- **Subroutine:**

Algorithm 2 DEVI

Input: Confidence set $\mathcal{C}$, error parameter ϵ, transition bonus q, cost perturbation ρ
Initialize: $i \leftarrow 0$, $Q^{(0)} \leftarrow 0$, $V^{(0)} \leftarrow 0$
while $\|V^{(i)} - V^{(i-1)}\|_\infty \geq \epsilon$ **do**
  $Q^{(i+1)}(\cdot,\cdot) \leftarrow c_\rho(\cdot,\cdot) + (1 - q)\min_{\boldsymbol{\theta}}\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot,\cdot)\rangle$
  $V^{(i+1)}(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q^{(i+1)}(\cdot, a)$
  $i \leftarrow i + 1$
end while
Output: $Q^{(i+1)}(\cdot,\cdot)$

Main Results: Theory

Theorem 1 (Hoeffding-type 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 expected cost of the optimal policy, $c_{\min} > 0$ is the lower bound of the per-step cost. The bound degrades to $\widetilde{O}(K^{2/3})$ for general cost functions, i.e., $c_{\min} = 0$.

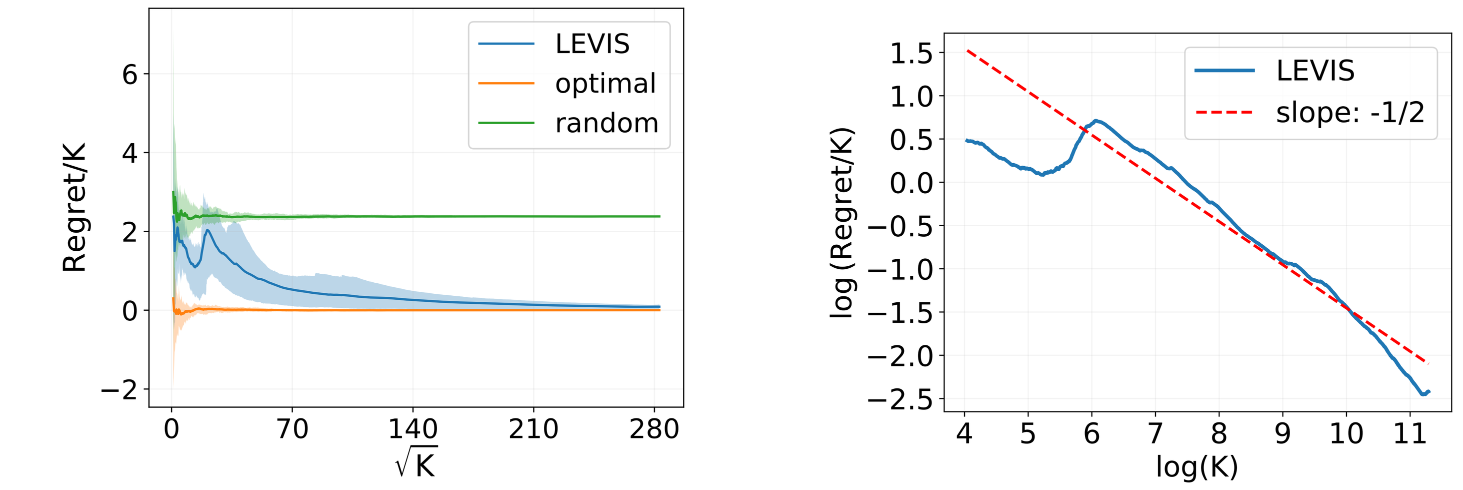
Theorem 2 (Lower bound)
Under technical assumptions, any algorithm for linear mixture SSP incurs an expected regret of at least $\Omega(dB_\star\sqrt{K})$.

Theorem 3 (Near-optimal upper 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.

Open problems:
- How to remove the dependence on $c_{\min}$?
- How to achieve $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bound when $c_{\min} = 0$?

Experiments

Numerical experiments corroborate our theory that LEVIS achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret:



(a) Plot of average regret (b) Log-log plot of regret