Provably Efficient CVaR RL in Low-rank MDPs

Yulai Zhao*, Wenhao Zhan*, Xiaoyan Hu*, Ho-fung Leung, Farzan Farnia, Wen Sun, Jason D. Lee

Princeton University, The Chinese University of Hong Kong, Cornell University

Motivation

- In risk-sensitive Reinforcement Learning (RL), the goal is to maximize the *conditional-value-at-risk* (CVaR)

$$\mathrm{CVaR}_\tau(R(\pi)) \coloneqq \sup_{c\in[0,H]} \left\{c - \tau^{-1}\cdot\mathbb{E}[c - R(\pi)^+]\right\}$$

where $R(\pi)$ is the random return of policy π and τ is the *risk tolerance*.

- Prior work [1] establishes regret guarantees for *tabular* MDPs, which is inapplicable to large state space

Contributions

- We study CVaR RL in *low-rank* MDPs [2], where the transition kernel admits *unknown* low-rank decomposition and the state space can be arbitrarily large
- We propose *REpresentation Learning for CVaR* (ELA), an online algorithm that learns a near-optimal policy with *polynomial sample complexity*
- Computational-wise, we propose *REpresentation Learning with LSVI for CVaR* (ELLA) as an *efficient oracle* to compute a near-optimal policy in the learned model, i.e., linear MDPs
- To our knowledge, this is the *first provably sample-efficient and computation-efficient* CVaR RL algorithm in low-rank MDPs.

Problem Statement

1. Risk-Sensitive RL and Augmented MDP

- Bäuerle and Ott [3] show that the optimal CVaR policy can be solved in the *augmented MDP*, where the state is augmented by the budget variable c. Define *augmented policy* $\pi_h: \mathcal{S}\times[0,H]\to\Delta(\mathcal{A})$.
- The Q-function is defined as ($c_{t+1} = c_t - r_t$)

$$Q^\pi_{h,\mathcal{P}^*}(s,c,a) \coloneqq \mathbb{E}_{\pi,\mathcal{P}^*}\left[\left(c_h - \sum_{t=h}^{H} r_t(s_t,a_t)\right)^+ \middle|\, s_h = s, c_h = c, a_h = a\right]$$

- The value function is defined as

$$V^\pi_{h,\mathcal{P}^*}(s,c) \coloneqq \mathbb{E}_{\pi,\mathcal{P}^*}\left[\left(c_h - \sum_{t=h}^{H} r_t(s_t,a_t)\right)^+ \middle|\, s_h = s, c_h = c\right]$$

- The goal is to learn the *optimal augmented policy* π^* and initial budget c^* such that

$$\mathrm{CVaR}_\tau(R(\pi^*,c^*)) = \mathrm{CVaR}^*_\tau \coloneqq \max_{c\in[0,H]}\left[c - \tau^{-1}\cdot\min_\pi V^\pi_{1,\mathcal{P}^*}(s_1,c)\right]$$

Problem Statement (Cont'd)

2. Low-rank MDPs

- We consider an episodic MDP M with episode length H, state space S, and a finite action space A. At each episode, a trajectory $\tau = (s_1, a_1, s_2, \cdots, s_H, a_H)$ is generated by an agent, where (a) $s_1 \in S$ is a fixed starting state,3 (b) at step h, the agent chooses action according to a history-dependent policy $a_h \sim \pi_h(\cdot\,|s_h, \tau_{h-1})$ (c) the model transits to the next state $s_{h+1}\sim\mathcal{P}^*_h(\cdot\,|s_h,a_h)$ and receive reward $r_h: S\times A \mapsto \Delta([0,1])$.
- To address large state space, we consider low-rank MDPs [2], where

$$\mathcal{P}^*_h(s'|s,a) = \langle \psi^*_h(s'), \phi^*_h(s,a)\rangle$$

  - where $\psi^*_h: \mathcal{S}\to\mathbb{R}^d$ and $\phi^*_h: \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ are *unknown* embedding functions.
  - Moreover, $\|\phi^*_h(s,a)\|_2 \le 1$ for all $(h,s,a)\in[H]\times S\times A$, and for any function
  - $g: S\mapsto[0,1], h\in[H]$, and $\|\int_{s\in S}\psi^*_h(s)g(s)ds\|_2 \le \sqrt{d}$.
- The learner has access to function classes Ψ and Φ that satisfy the *realizability* assumption, i.e., $\psi^*_h\in\Psi$ and $\phi^*_h\in\Phi$.

Main Results

Algorithm 1 *ELA*

Require: Risk tolerance $\tau\in(0,1]$, number of iterations K, parameters $\{\lambda^k\}_{k\in[K]}$ and $\{\alpha^k\}_{k\in[K]}$, models $\mathcal{F}=\{\Psi,\Phi\}$, failure probability $\delta\in(0,1)$.

1: Set datasets $\mathcal{D}_h, \widetilde{\mathcal{D}}_h \leftarrow \emptyset$ for each $h\in[H-1]$.
2: Initialize the exploration policy $\pi^0 \leftarrow \{\pi^0_h(s,c) = U(\mathcal{A}), \text{for any } (s,c)\in\mathcal{S}\times[0,H]\}_{h\in[H]}$.
3: Initialize the budget $c^0 \leftarrow 1$.
4: **for** iteration $k = 1,\ldots,K$ **do**
5: Collect a tuple $(\tilde{s}_1,\tilde{a}_1,s'_2)$ by taking $\tilde{a}_1\sim U(\mathcal{A})$, $s'_2\sim P^*_1(\cdot|\tilde{s}_1,\tilde{a}_1)$.
6: Update $\widetilde{\mathcal{D}}_1 \leftarrow \widetilde{\mathcal{D}}_1\cup\{(\tilde{s}_1,\tilde{a}_1,s'_2)\}$.
7: **for** $h = 1,\cdots,H-1$ **do**
8: Collect two transition tuples $(s_h,a_h,\tilde{s}_{h+1})$ and $(\tilde{s}_{h+1},\tilde{a}_{h+1},s'_{h+2})$ by first rolling out π^{k-1} starting from (s_1,c^{k-1}) into state s_h, taking $a_h\sim U(\mathcal{A})$, and receiving $\tilde{s}_{h+1}\sim P^*_h(\cdot|s_h,a_h)$, then taking $\tilde{a}_{h+1}\sim U(\mathcal{A})$ and receiving $s'_{h+2}\sim P^*_{h+1}(\cdot|\tilde{s}_{h+1},\tilde{a}_{h+1})$.
9: Update $\mathcal{D}_h\leftarrow\mathcal{D}_h\cup\{(s_h,a_h,\tilde{s}_{h+1})\}$.
10: Update $\widetilde{\mathcal{D}}_{h+1}\leftarrow\widetilde{\mathcal{D}}_{h+1}\cup\{(\tilde{s}_{h+1},\tilde{a}_{h+1},s'_{h+2})\}$ if $h\le H-2$.
11: Learn representations via MLE

$$\widehat{P}_h \coloneqq (\widehat{\psi}_h,\widehat{\phi}_h) \leftarrow \arg\max_{(\psi,\phi)\in\mathcal{F}} \sum_{(s_h,a_h,s_{h+1})\in\{\mathcal{D}_h+\widetilde{\mathcal{D}}_h\}} \log\langle\psi(s_{h+1}),\phi(s_h,a_h)\rangle$$

12: Update empirical covariance matrix $\widehat{\Sigma}_h = \sum_{(s,a)\in\mathcal{D}_h}\widehat{\phi}_h(s,a)\widehat{\phi}_h(s,a)^\top + \lambda^k I_d$.
13: Set the exploration bonus:

$$\widehat{b}_h(s,a)\leftarrow\begin{cases}\min\left(\alpha^k\sqrt{\widehat{\phi}_h(s,a)\widehat{\Sigma}_h^{-1}\widehat{\phi}_h(s,a)^\top},2\right) & h\le H-2\\ 0 & h = H-1\end{cases}$$

14: **end for**
15: Run Value-Iteration (VI) and obtain $c^k\leftarrow\arg\max_{c\in[0,H]}\left\{c-\tau^{-1}\min_\pi V^\pi_{1,\widehat{\mathcal{P}},\widehat{b}}(s_1,c)\right\}$.
16: Set $\pi^k\leftarrow\arg\min_\pi V^\pi_{1,\widehat{\mathcal{P}},\widehat{b}}(s_1,c^k)$.
17: **end for**

Ensure: Uniformly sample k from $[K]$, return $(\widehat{\pi},\widehat{c}) = (\pi^k,c^k)$.

Main Results (Cont'd)

1. ELA (REprensentation Learning for CVAR)

This algorithm has the following key components.

- **Data Collection (Lines 8-10).** We collect two (disjoint) sets of transition tuples to compute the bonus terms and estimate the transition kernels. These two datasets are different in their (marginalized) distributions and facilitate the regret analysis.
- **MLE oracle (Line 11).** Model transitions are estimated through the MLE oracle.
- **Value Iteration (Line 15).** Based on the learned model, the algorithm runs Value-Iteration (VI) with the exploration bonus term. Such a value function is used to perform VI and update policy on the learned model $\hat{P}$, since the learner has no prior knowledge of the real model transitions. Therefore, obtaining an accurate estimation of the model determines the quality of the output policy.

We remark that the exact VI in Line 15 is *not computationally efficient* due to the continuity of c and potentially large state space S. To overcome such a computational barrier, in the next algorithm, we provide a computationally efficient planning oracle that performs LSVI with discretized reward function (with sufficiently high precision).

Theoretical guarantees. This presents the first regret/sample complexity bounds for CVaR RL with function approximation, where exploring the unknown action/space spaces posits extra difficulty.

Theorem 4.1. *Fix $\delta\in(0,1)$. Set the parameters in Algorithm 1 as:*

$$\alpha^k = O\left(\sqrt{H^2(|\mathcal{A}|+d^2)}\log\left(\frac{|\mathcal{F}|Hk}{\delta}\right)\right),\quad \lambda^k = O\left(d\log\left(\frac{|\mathcal{F}|Hk}{\delta}\right)\right).$$

We have two equivalent interpretations of the theoretical results. In terms of PAC bound, with probability at least $1-\delta$, the regret is bounded by

$$\sum_{k=1}^{K}\mathrm{CVaR}^*_\tau - \mathrm{CVaR}_\tau(R(\pi^k,c^k)) = \tilde{O}\left(\tau^{-1}H^3Ad^2\sqrt{K}\cdot\sqrt{\log(|\mathcal{F}|/\delta)}\right).$$

Alternatively, we can interpret in terms of sample complexity: w.p. at least $1-\delta$, to present an ϵ-optimal policy and budget pair s.t. $\mathrm{CVaR}^*_\tau - \mathrm{CVaR}_\tau(R(\widehat{\pi},\widehat{c}))\le\epsilon$. The total number of trajectories required is upper bounded by

$$\tilde{O}\left(\frac{H^7A^2d^4\log(|\mathcal{F}|/\delta)}{\tau^2\epsilon^2}\right).$$

Main Results (Cont'd)

2. ELLA (REprensentation Learning with LSVI for CVAR).

- In Algorithm 1, the VI in line 15 is not computational efficient since objective $c - \tau^{-1}\min_\pi V^\pi_{1,\hat{P},\hat{b}}(s_1,c)$ is not concave, which brings significant computational overhead.
- We introduce a feasible planning oracle for this step. Moreover, we introduce a novel LSVI-UCB, stated in Algorithm 2. Particularly, the following theorem characterizes the computational cost for finding an ϵ-optimal policy

Theorem 5.1 (Informal). *Let the parameters in Algorithm 1 and 2 take appropriate values, then we have with probability at least $1-\delta$ that $\mathrm{CVaR}^*_\tau - \mathrm{CVaR}_\tau(R(\widehat{\pi},\widehat{c}))\le\epsilon$ where $(\widehat{\pi},\widehat{c})$ is the returned policy and initial budget by Algorithm 3. In total, the sample complexity is upper bounded by $\tilde{O}\left(\frac{H^7A^2d^4\log\frac{|\mathcal{F}|}{\delta}}{\tau^2\epsilon^2}\right)$. The MLE oracle is called $\tilde{O}\left(\frac{H^7A^2d^4\log\frac{|\mathcal{F}|}{\delta}}{\tau^2\epsilon^2}\right)$ times and the rest of the computation cost is $\tilde{O}\left(\frac{H^{19}A^3d^{12}\log\frac{|\mathcal{F}|}{\delta}}{\upsilon^{10}\tau^6\epsilon^6}\right)$.*

References

[1] Wang, Kaiwen, Nathan Kallus, and Wen Sun. "Near-Minimax-Optimal Risk-Sensitive Reinforcement Learning with CVaR." arXiv preprint arXiv:2302.03201 (2023).

[2] Agarwal, Alekh, et al. "Flambe: Structural complexity and representation learning of low rank MDPs." Advances in neural information processing systems 33 (2020): 20095-20107.

[3] Bäuerle, Nicole, and Jonathan Ott. "Markov decision processes with average-value-at-risk criteria." Mathematical Methods of Operations Research 74 (2011): 361-379.

[4] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory, pages 2137–2143. PMLR.

[5] Uehara, M., Zhang, X., and Sun, W. (2022). Representation learning for online and offline RL in low-rank MDPs. In International Conference on Learning Representations.

Acknowledgements

The work presented in this paper is partially supported by a research grant from the Research Grants Council, Hong Kong, China (RGC Ref. No. CUHK 14206820).