Provably Efficient Iterated CVaR Reinforcement Learning with Function Approximation and Human Feedback

Yu Chen[1] Yihan Du[2] Pihe Hu[1] Siwei Wang[3] Desheng Wu[4] Longbo Huang [1]

[1]IIIS, Tsinghua University [2]ECE, University of Illinois at Urbana-Champaign [3]Microsoft Research Asia [4]SEM, University of Chinese Academy of Sciences

Overview

This works presents a novel risk-sensitive RL framework that employs an Iterated Conditional Value-at-Risk (ICVaR) objective under both linear and general function approximations, and also integrates human feedback setting. We presents provably sample-efficient algorithms and provide rigorous theoretical analysis.

Motivation

Previous work [2] considering the ICVaR-RL only establishes regret guarantees for *tabular* MDPs, which is inapplicable to large state space. Moreover, many real-world applications of RL such as LLM [3, 4] learning from human feedbacks, underscoring the crucial role of infusing human feedback into risk-sensitive RL.

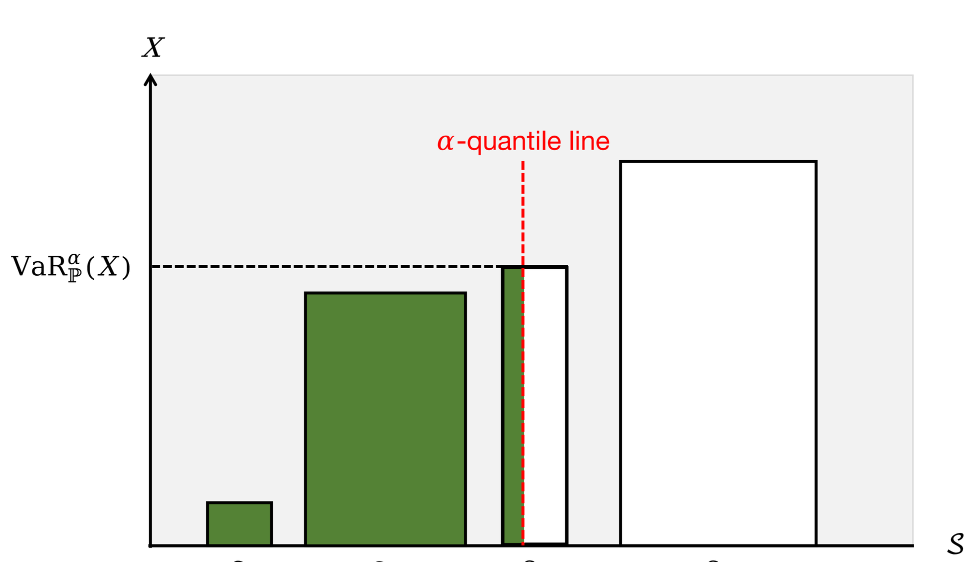
Formulation

- The Markov Decision Processes (MDPs) for traditional RL models

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, K, H, \{\mathbb{P}_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$$

- The Conditional Value-at-Risk (CVaR) operator

$$\mathrm{CVaR}_{\mathbb{P}}^{\alpha}(X) := \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{\alpha} \mathbb{E}\left[(x - X)^+\right] \right\}$$



- The Iterated CVaR objective:

$$J(\pi) = r_1(s_1, a_1) + \mathrm{CVaR}^{\alpha}_{s_2 \sim \mathbb{P}_1(\cdot|s_1,a_1)} \Big(r_2(s_2, a_2) + \mathrm{CVaR}^{\alpha}_{s_3 \sim \mathbb{P}_2(\cdot|s_2,a_2)} \Big(r_3(s_3, a_3) + \Big(\cdots \mathrm{CVaR}^{\alpha}_{s_H \sim \mathbb{P}_{H-1}(\cdot|s_{H-1},a_{H-1})} (r_H(s_H, a_H)) \Big) \Big) \Big)$$

ICVaR RL

Value and Q functions:

$$\begin{cases} Q_h^{\pi}(s,a) = r_h(s,a) + \mathrm{CVaR}^{\alpha}_{s' \sim \mathbb{P}_h(\cdot|s,a)}(V_{h+1}^{\pi}(s')) \\ V_h^{\pi}(s) = Q_h^{\pi}(s, \pi_h(s)) \\ V_{H+1}^{\pi}(s) = 0, \forall s \in \mathcal{S} \end{cases}$$

Optimal Policy:

$$V_h^{\pi^*}(s) = \max_{\pi} V_h^{\pi}(s)$$

Regret Metric:

$$\mathrm{Regret}(K) := \sum_{k=1}^{K} \left(V_1^{\pi^*}(s_{k,1}) - V_1^{\pi^k}(s_{k,1}) \right),$$

Function Approximation

Linear Function Approximation: For any step $h \in [H]$, there exists a vector $\theta_h \in \mathbb{R}^d$ with $\|\theta_h\|_2 \leq \sqrt{d}$ such that

$$\mathbb{P}_h(s' \mid s, a) = \langle \theta_h, \phi(s', s, a) \rangle$$

holds for any $(s', s, a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$. Moreover, the agent has access to the feature basis ϕ.

General Function Approximation: The transition kernels $\{\mathbb{P}_h\}_{h=1}^H \subset \mathcal{P}$ where $\mathcal{P}$ is a function class of transition kernels with the form $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. In addition, the agent has access to such function class $\mathcal{P}$.

Algorithm ICVaR-L

We develop a provably efficient (both *computationally* and *statistically*) algorithm ICVaR-L for ICVaR-RL with linear function approximation,

Algorithm 1 ICVaR-L

Require: risk level $\alpha \in (0, 1]$, approximation accuracy $\varepsilon > 0$, regularization parameter $\lambda > 0$, bonus multiplier $\widehat{\beta}$.

1: Initialize $\widehat{\mathbf{\Lambda}}_{1,h} \leftarrow \lambda \mathbf{I}$, $\widehat{\theta}_{1,h} \leftarrow \mathbf{0}$, $\widehat{V}_{k,H+1}(\cdot) \leftarrow 0$ for any $k \in [K]$ and $h \in [H]$.
2: **for** episode $k = 1, ..., K$ **do**
3: **for** step $h = H, ..., 1$ **do**
4: *// Optimistic value iteration*
5: $B_{k,h}(\cdot,\cdot) = \frac{\widehat{\beta}}{\alpha} \sup_{x \in \mathcal{N}_\varepsilon} \|\psi_{(x - \widehat{V}_{k,h+1})^+}(\cdot,\cdot)\|_{\widehat{\Lambda}_{k,h}^{-1}}$
6: $\widehat{Q}_{k,h}(\cdot,\cdot) = r_h(\cdot,\cdot) + [\mathbb{C}^{\alpha,\mathcal{N}_\varepsilon}_{\widehat{\theta}_{k,h}}(\widehat{V}_{k,h+1})](\cdot,\cdot) + 2\varepsilon + B_{k,h}(\cdot,\cdot)$
7: $\widehat{V}_{k,h}(\cdot) \leftarrow \min\left\{ \max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(\cdot, a), H \right\}$
8: $\pi_h^k(\cdot) \leftarrow \arg\max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(\cdot, a)$
9: **end for**
10: **for** step $h = 1, \cdots, H$ **do**
11: Observe the current state $s_{k,h}$, and take the action $a_{k,h} = \pi_h^k(s_{k,h})$
12: Calculate $x_{k,h} \leftarrow \arg\max_{x \in \mathcal{N}_\varepsilon} \|\psi_{(x - \widehat{V}_{k,h+1})^+}(s_{k,h}, a_{k,h})\|_{\widehat{\Lambda}_{k,h}^{-1}}$
13: $\widehat{\Lambda}_{k+1,h} \leftarrow \widehat{\Lambda}_{k,h} + \psi_{(x_{k,h} - \widehat{V}_{k,h+1})^+}(s_{k,h}, a_{k,h}) \psi_{(x_{k,h} - \widehat{V}_{k,h+1})^+}(s_{k,h}, a_{k,h})^\top$
14: $\widehat{\theta}_{k+1,h} \leftarrow \widehat{\Lambda}_{k,h}^{-1} \sum_{i=1}^{k} \psi_{(x_{i,h} - \widehat{V}_{i,h+1})^+}(s_{i,h}, a_{i,h})(x_{i,h} - \widehat{V}_{i,h+1})^+(s_{i,h+1})$ *// Solution to ridge regression*
15: **end for**
16: **end for**

Main Results of Linear FA

- **Key Components:**
  - **CVaR Operator Approximation(Line 6)** We take a supremum over the discrete finite set $\mathcal{N}_\varepsilon$ instead of the interval $[0, H]$ while guaranteeing that the error between the approximated CVaR operator and the true CVaR operator is at most 2ε.
  - **CVaR-Adaptive Ridge Regression(Lines 12-14)** We consider $\{\psi_{(x_{i,h} - \widehat{V}_{i,h+1})^+}\}_{i=1}^k$ as the regression features, which are different from $\{\psi_{\widehat{V}_{i,h+1}}\}_{i=1}^k$ used in previous risk-neutral linear mixture MDP works [7, 8].
- **Regret Upper Bound**

Theorem 1. *Suppose Assumption 1 holds, and for given $\delta \in (0, 1]$, set $\lambda = H^2$, $\varepsilon = dH\sqrt{\alpha^{H-3}/K}$, and the bonus multiplier $\widehat{\beta} = H\sqrt{d \log\left(\frac{H+KH^3}{\delta}\right)} + \sqrt{\lambda}$. Then, with probability at least $1 - 2\delta$, the regret of ICVaR-L (Algorithm 1) satisfies*

$$\mathrm{Regret}(K) \leqslant 4dH^2\sqrt{\frac{K}{\alpha^{H+1}}} + 2\widehat{\beta}\sqrt{\frac{KH}{\alpha^{H+1}}}\sqrt{8dH\log(K) + 4H^3\log\frac{4\log_2 K + 8}{\delta}}. \quad (10)$$

- **Regret Lower Bound**

Theorem 4. *Let $H \geqslant 2$, $d \geqslant 2$, and an interger $n \in [H-1]$. Then, for any algorithm, there exists an instance of Iterated CVaR RL under Assumption 1, such that the expected regret is lower bounded as follows:*

$$\mathbb{E}[\mathrm{Regret}(K)] \geqslant \Omega\left(d(H-n)\sqrt{\frac{K}{\alpha^n}} \right). \quad (47)$$

Algorithm ICVaR-L enjoys the regret uppper bound $\widetilde{O}(\sqrt{\alpha^{-(H+1)}(d^2H^4 + dH^6)K})$ and the regret lower bound $\Omega(d\sqrt{\alpha^{-(H-1)}K})$. We can see that ICVaR-L achieves a nearly minimax optimal with respect to factors d and K, and the factor $\sqrt{\alpha^{-H}}$ in our regret upper bound is unavoidable in general.

Human Feedback Setting

We consider the classic RLHF model [5, 6].

- **Human Feedback:** The agent cannot observe numerical reward signals, but only receives human feedback that describes human preferences for two different trajectories.
- **Underlying Ground Truth Reward:** There is a unknown underlying reward $\boldsymbol{r}^*$ in a known infinite function set $\mathcal{R}$.
- **Comparison Oracle:** A comparison oracle takes in two trajectories τ_1, τ_2 and returns

$$o \sim \mathrm{Ber}(\sigma(\boldsymbol{r}^*(\tau_1) - \boldsymbol{r}^*(\tau_2))),$$

where $\sigma(\cdot)$ is a known link function, e.g., sigmoid function (a.k.a. the BTV model [1]).

Algorithm ICVaR-HF

We firstly consider the risk-sensitive RL with human feedback and general FA and present the provably efficient algorithm ICVaR-HF.

Algorithm 2 ICVaR-HF

Execute an arbitrary policy to collect trajectory $\tau_0 = (s_{0,1}, a_{0,1}, \cdots, s_{0,H}, a_{0,H})$.
for $k = 1 \cdots K$ **do**
Receive the initial state $s_{k,1}$
Choose the estimated reward $\widehat{\boldsymbol{r}}^k \leftarrow \arg\max_{\boldsymbol{r} \in \widehat{\mathcal{R}}_k} \widetilde{V}_1^{\widehat{\mathcal{P}}_k}(s_{k,1}; \boldsymbol{r}_{\tau_0})$. *// Choose the estimated reward $\widehat{\boldsymbol{r}}^k$*
for $h = H, \cdots, 1$ **do**
$\widehat{Q}_{k,h}(\cdot,\cdot) \leftarrow \widehat{r}_h^k(\cdot,\cdot) - \widehat{r}_h^k(s_{0,h}, a_{0,h}) + \sup_{\mathbb{P}' \in \mathcal{P}_h} [\mathbb{C}^{\alpha}_{\mathbb{P}'}(\widehat{V}_{h+1})](\cdot,\cdot)$
$\widehat{V}_h(\cdot) \leftarrow \max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(\cdot, a)$, $\pi_h^k(\cdot) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(\cdot, a)$
end for
Execute the policy $\pi^k := \{\pi_h^k\}_{h=1}^H$. In every step h, receive state $s_{k,h}$ and execute action $a_{k,h} = \pi_{k,h}(s_{k,h})$. Then collect the trajectory $\tau_k = (s_{k,1}, a_{k,1}, s_{k,2}, a_{k,2}, \cdots, s_{k,H}, a_{k,H})$.
Compare two trajectories τ_k, τ_0 and collect observation o_k from human feedback.
Update the reward confidence set $\widehat{\mathcal{R}}_{k+1} \leftarrow \{\boldsymbol{r} \in \mathcal{R} : \mathcal{L}_k(\boldsymbol{r}) > \max_{\boldsymbol{r}' \in \mathcal{R}} \mathcal{L}_k(\boldsymbol{r}') - \widehat{\beta}_R\}$.
for $h = 1, \cdots, H$ **do**
$\widehat{\mathbb{P}}_{k+1,h} \leftarrow \arg\min_{\mathbb{P}' \in \mathcal{P}} \sum_{i=1}^{k} \mathrm{Dist}_{i,h}(\mathbb{P}', \delta_{k,h})$ *// Estimate the transition kernel $\mathbb{P}_h$*
$\widehat{\mathcal{P}}_{k+1,h} = \left\{ \mathbb{P}' \in \mathcal{P} : \sum_{i=1}^{k} \mathrm{Dist}_{i,h}(\mathbb{P}', \widehat{\mathbb{P}}_{i,h}) \leqslant \widehat{\gamma}^2 \right\}$ *// Construct the confidence set*
end for
end for

ICVaR-HF satisfies $\widetilde{O}(\sqrt{K})$ regret upper bound, which is stated below.

Theorem 3. *For some positive constant $\delta \in (0, 1]$, we set the estimation radius $\widehat{\beta}_R = c\log(K \cdot N_B(\mathcal{R}, \|\cdot\|_\infty, 1/K)/\delta)$ and $\widehat{\gamma} = 4H^2\left(2\log\left(\frac{2H \cdot N_C(\mathcal{P}, \|\cdot\|_{\infty,1}, 1/K)}{\delta}\right) + 1 + \sqrt{\log(5K^2/\delta)}\right)$ for some constant c. Denote Then with probability at least $1 - 4\delta$, the regret of Algorithm 2 satisfies*

$$\mathrm{Regret}(K) \leqslant \tilde{O}\left(\sqrt{KH^3\alpha^{-H-1}} \left(\sqrt{HD_P} + \sqrt{m^{-1}D_R} \right) \right), \quad (16)$$

where the dimension parameters $D_p := d_E(\mathcal{Z})\log(N_C(\mathcal{P}, \|\cdot\|_{\infty,1}, 1/K)$ detailed in Theorem 2, and $D_R := d_E(\mathcal{R})\log(N_B(\mathcal{R}, \|\cdot\|_\infty, 1/K))$. Here $d_E(\mathcal{R}) := \dim_E(\mathcal{R}, 1/\sqrt{K})$ is the eluder dimension of $\mathcal{R}$, and $N_B(\mathcal{R}, \|\cdot\|_\infty, 1/K)$ is the $1/K$-bracketing number of $\mathcal{R}$ under norm $\|\cdot\|_\infty$. [4]

References

[1] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[2] Yihan Du, Siwei Wang, and Longbo Huang. Provably efficient risk-sensitive reinforcement learning: Iterated CVar and worst path. In *The Eleventh International Conference on Learning Representations*, 2023.

[3] Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, September 2022.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[5] Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *arXiv preprint arXiv:2302.03201*, 2023.

[6] Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023.

[7] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[8] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.