Provably Efficient Risk-Sensitive Reinforcement Learning with Human Feedback

1st Xinyi Ni
Electrical and Computer Engineering
University of California, Davis
Davis, USA
xni@ucdavis.edu

2nd Lifeng Lai
Electrical and Computer Engineering
University of California, Davis
Davis, USA
lflai@ucdavis.edu

***Abstract*—Unlike standard Reinforcement Learning (RL) that relies directly on reward signals, reinforcement learning with human feedback (RLHF) integrates human feedback into the decision-making process. This paper introduces an online risk-sensitive RLHF framework that integrates static Conditional Value-at-Risk (CVaR), ensuring policy alignment with human-defined risk preferences. We propose a black-box reward estimation method for tabular Markov Decision Processes (MDPs) that integrates seamlessly with existing CVaR RL algorithms. While preserving the sample complexity of the base solver, our method incurs only a human feedback query complexity of $\tilde{\mathcal{O}}\left(\frac{d}{\beta^2\epsilon^2\alpha^2}\right)$ for trajectory-based comparisons, where d denotes the Eluder dimension of the reward function class, and $\tilde{\mathcal{O}}\left(\frac{SAH^2}{\beta^2\epsilon^2\alpha^2}\right)$ for state-action-based comparisons. We extend our framework to low-rank MDPs and develop a maximum likelihood estimator (MLE)-based reward inference method. Regret analysis is provided to guarantee the computational efficiency.**

***Index Terms*—reinforcement learning, risk-sensitive, human feedback, sample complexity, regret analysis**

I. Introduction

Reinforcement Learning (RL) is a control-theoretic framework for sequential decision-making, where agents interact with an unknown environment by taking actions, receiving reward feedback, and learning policies to maximize the expected cumulative reward [1]. Reward signals can range from objective measures of success, such as winning a game of Go [2], to hand-designed proxies for progress, like gaining gold in DOTA2. RL has achieved empirical success across diverse domains, including robotics, video games, and autonomous driving [3]. However, the availability and quality of reward signals are critical to this success, posing challenges in applications where designing a suitable reward function is difficult or where the goal depends on complex, human-centered evaluations.

To address this limitation, integrating human feedback into RL has emerged as a promising approach [4]–[10]. RL with human feedback (RLHF) introduces auxiliary reward signals provided by humans, enhancing the learning process by aligning the agent's behavior with nuanced objectives, including ethical considerations, safety, and robustness. This approach has demonstrated significant success in applications such as training Large Language Models (LLMs) [6], robotic control [5], and preference alignment tasks [4], highlighting its potential to tackle complex, human-centered decision-making challenges. While RLHF provides a mechanism for aligning RL agents with human preferences, existing approaches typically focus on maximizing the expected total reward without explicitly accounting for risk. However, in many real word applications, risk-sensitive decision-making is crucial. For instance, in AI safety, healthcare and autonomous systems, the consequences of rare but catastrophic events must be carefully managed, and human feedback should reflect not just preferences but also risk-aware objectives [7], [11]–[13].

Risk-sensitive RL extends the standard RL paradigm by optimizing for policies that account for risk measures rather than directly maximizing expected rewards [7], [14]–[21]. Integrating risk-sensitive criteria into RLHF aims to ensures that learned policies are not only aligned with human preferences but also robust to uncertain or extreme outcomes, bridging the gap between risk-aware decision-making and human-aligned learning. Like traditional RL, risk-sensitive RL depends on reward information to drive policy learning, and similar challenges arise when designing reward mechanisms. Specifically, the challenges inherent to RLHF, such as the reliability of human-provided feedback [4], scalability [5], and integration with algorithmic frameworks [6], persist in the risk-sensitive domain, further motivating a systematic study of this intersection.

In this work, we bridge RLHF and risk-sensitive RL by integrating static CVaR with *online* human feedback. Unlike conventional RLHF, which optimizes a linear expected return, the CVaR objective is nonlinear in the reward, so existing preference-based reward inference techniques do not directly extend to this setting. We first study tabular Markov Decision Processes (MDPs), where most existing CVaR RL results are established, and propose a black-box interface, **CVaR-HF-P2R**, that estimates rewards from online human feedback and

Accepted for presentation at the 2026 IEEE International Symposium on Information Theory (ISIT 2026). This work was supported by National Science Foundation under grant CCF-2232907 and ECCS-2448268.

can be seamlessly integrated with any CVaR RL algorithm. This interface preserves the sample complexity of the base solver and incurs only an additional query complexity of $\tilde{\mathcal{O}}\left(\frac{SAH^2}{\beta^2\epsilon^2\alpha^2}\right)$ for pairwise state–action comparisons, where S and A are the state and action space sizes, H is the horizon, β characterizes the link function for human feedback, α is the CVaR risk level, and ϵ is the accuracy parameter. To move beyond the inherent scalability limits of tabular MDPs, we then extend our framework to low-rank MDPs, which capture a broad class of structured environments including tabular, linear, and block MDPs [20], [22]–[24]. In this setting, we introduce **CVaR-HF-ELA**, which incorporates human feedback via Maximum Likelihood Estimation (MLE) and leverages ELA [24], a provably sample-efficient function-approximation method for CVaR RL, yielding effective reward inference and CVaR-based policy optimization with regret guarantees. Finally, we present **CVaR-HF-FAG**, a general interface that abstracts this reduction and can be instantiated with other provably efficient function-approximation-based solvers for CVaR RL.

Related Work: Recent studies have explored incorporating risk preferences into RLHF, with [7], [12], [13] being most relevant to ours but differing in both risk formulation and algorithmic assumptions. [7] focuses on online RLHF and uses *iterated* (dynamic) CVaR, which evaluates risk at each decision step. This differs fundamentally from the static CVaR used in our work, which assesses risk over the total trajectory reward. Consequently, their algorithm design and analysis, centered on dynamic Bellman updates and per-step preference feedback, do not directly apply to our setting. [12] considers *offline* distributional RLHF with pre-collected preferences, whereas we study an *online* setting with static CVaR, updating reward estimates from sequential comparisons via confidence sets and risk-aware Bellman operators rather than distributional estimation from a fixed dataset. [13] proposes a general online PbRL framework that supports a broad class of risk-sensitive objectives, but their analysis is restricted to tabular MDPs with trajectory-level rewards. In contrast, we assume decomposable rewards and further cover low-rank and function-approximation settings, using CVaR-specific confidence sets, tail perturbation bounds, and complexity measures (eluder and bracketing) that differ from their trajectory-embedding and nested-risk approach.

The remainder of this paper is organized as follows. Section II reviews MDPs, CVaR, and function approximation. Section III introduces the human-feedback model and presents a black-box CVaR-RLHF method with bounded additional query complexity. Section IV extends the framework to low-rank MDPs, using MLE-based reward inference and providing a regret analysis for infinite state spaces. Section V concludes.

II. Preliminaries

A. Episodic Markov Decision Processes (MDPs)

We consider an episodic MDP $(\mathcal{S}, \mathcal{A}, H, K, \mathbb{P}^\star, r^\star)$, where $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, H is the episode length, K is the number of episodes, and $\mathbb{P}_h^\star(\cdot \mid s, a)$ is the transition kernel at step h. The reward is deterministic, $r_h^\star : \mathcal{S}\times\mathcal{A} \to [0,1]$, following [7], [10], [25], [26]. Let $\Pi_{\mathcal{H}}$ be the class of history-dependent policies, where each $\pi \in \Pi_{\mathcal{H}}$ consists of $\pi_h : \mathcal{S} \times \mathcal{H}_h \to \Delta\mathcal{A}_{h=1}^H$, with $\mathcal{H}_h$ the history up to step h and $\Delta\mathcal{A}$ the simplex over $\mathcal{A}$. Each episode starts from an initial state s_1; at step h, the agent observes s_h, samples $a_h \sim \pi_h(s_h; \mathcal{H}_h)$, receives $r_h^\star(s_h, a_h)$, and transitions to $sh+1 \sim \mathbb{P}_h^\star(\cdot \mid s_h, a_h)$ until reaching s_{H+1}.

B. CVaR Optimization

We adopt Conditional Value-at-Risk (CVaR) as the risk measure. CVaR is a coherent and widely used risk measure in risk-sensitive RL [16]–[18], [25], [27]–[31]. For a random variable X and risk level $\alpha \in (0,1]$,

$$\text{CVaR}_\alpha(X) := \sup_{b\in\mathbb{R}} \left(b - \alpha^{-1}\mathbb{E}[(b-X)^+]\right),$$

where $x^+ = \max(x, 0)$. CVaR is preferred in risk-sensitive decision-making due to its coherency, which satisfies four key properties: (1) monotonicity, (2) translation invariance, (3) subadditivity, and (4) positive homogeneity, ensuring rational and consistent risk evaluation [32].

Since rewards are assumed deterministic and bounded in $[0,1]$, the total return lies in $[0,H]$, and the optimal $b^\star = \text{VaR}\alpha(X)$ also satisfies $b^\star \in [0,H]$ [29]. Thus we may restrict b to $[0,H]$ and rewrite:

$$\text{CVaR}_\alpha(X) := \sup_{b\in[0,H]} \left(b - \alpha^{-1}\mathbb{E}[(b-X)^+]\right). \tag{1}$$

Following [33], CVaR RL can be formulated on an augmented MDP with state space $\mathcal{S}^{\text{Aug}} = \mathcal{S} \times [0,H]$. There exists an optimal deterministic Markov policy $\rho^\star : \mathcal{S}^{\text{Aug}} \to \mathcal{A}$. Starting from (s_1, b_1), at each step h the agent chooses $a_h = \rho^\star(s_h, b_h)$, receives $r_h^\star(s_h, a_h)$, transitions via $\mathbb{P}h^\star(\cdot \mid s_h, a_h)$, and updates the budget $bh+1 = b_h - r_h^\star(s_h, a_h)$. The CVaR optimization problem can then be written as

$$\text{CVaR}_\alpha^\star(s_1) = \max_{b_1\in[0,H]} \{b_1 - \alpha^{-1} \min_{\rho\in\Pi^{\text{Aug}}} V_{1,\mathbb{P}^\star}^\rho(s_1, b_1)\}, \tag{2}$$

where Π^{Aug} is the class of deterministic Markov policies and

$$V_{h,\mathbb{P}^\star}^\rho(s_h, b_h) = \mathbb{E}_{\rho,\mathbb{P}^\star}\left[\left(b_h - \sum_{h'=h}^H r_{h'}^\star(s_{h'}, a_{h'})\right)^+ \middle| s_h, b_h\right].$$

C. Function Approximation

As tabular MDPs are restricted to finite state spaces, function approximation is needed in high-dimensional settings. We use the Eluder dimension to quantify its complexity [34]–[36].

Definition II.1 (Eluder Dimension)**.** For any function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, its Eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ is defined as the length of the longest sequence $\{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}$ such that there exists $\epsilon' \geq \epsilon$ for which, for all $i \in [n]$, x_i is ϵ-independent of its preceding sequence $\{x_1, \ldots, x_{i-1}\}$.

Specifically, this means there exist functions $f_i, g_i \in \mathcal{F}$ such that:

$$\sqrt{\sum_{j=1}^{i-1}((f_i - g_i)(x_j))^2} \leq \epsilon' \quad \text{and} \quad |(f_i - g_i)(x_i)| \geq \epsilon'.$$

Low-rank MDPs generalize tabular, linear and block MDPs, and have been widely studied in recent RL literature [24], [37]–[40].

Definition II.2 (Low-Rank MDPs)**.** A transition kernel $\mathcal{P} = \{\mathbb{P}\}_{h\in[H]}$ is said to admit a low-rank decomposition with rank d_{rank} if there exist embedding functions

$$\phi := \{\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\text{rank}}}\}_{h\in[H]}$$

and $\psi := \{\psi_h : \mathcal{S} \to \mathbb{R}^{d_{\text{rank}}}\}_{h\in[H]}$ such that:

$$\mathbb{P}_h(s'|s,a) = \langle \psi_h(s'), \phi_h(s,a)\rangle,$$

where $\|\phi_h(s,a)\|_2 \leq 1$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Furthermore, for any function $g : \mathcal{S} \to [0,1]$ and $h \in [H]$, it holds that

$$\left\| \int_{s\in\mathcal{S}} \psi_h(s) g(s) ds \right\|_2 \leq \sqrt{d_{\text{rank}}}.$$

A low-rank episodic MDP is one that admits such a decomposition.

The realizability assumption is crucial for deriving performance guarantees independent of the state-space size [41].

Assumption II.3. There exists a model class $\mathcal{F} = (\Psi, \Phi)$ such that $\psi_h \in \Psi$, $\phi_h \in \Phi$ for all $h \in [H]$.

D. Online RLHF

RLHF methods can be broadly categorized into *offline* and *online* [7], [10], [12], [13], [26], [42]. In offline RLHF, human preferences are pre-collected into a static dataset and used for supervised preference modeling without interaction during training [12], [26]. In contrast, online RLHF interleaves learning and data collection, allowing the agent to query a human oracle during policy execution [7], [9], [10], [13], [42], which is particularly suitable for dynamic or high-stakes settings. To incorporate CVaR into the learning objective, we adopt the online setting, where feedback takes two forms:

- **Trajectory-based preference:** the comparison oracle is defined over complete trajectories τ.
- (s,a)**-preference:** the comparison oracle is queried on specific state–action pairs (s,a).

Since (s,a)-preference can be viewed as a special case of trajectory preference with $\tau := \tau(s,a)$, we use τ to denote both cases. Throughout learning, the true reward function $r^\star$ is unknown: the agent does not observe numerical rewards and only receives comparisons from the human oracle.

Definition II.4 (Comparison oracle)**.** A comparison oracle takes in two trajectories τ_1, τ_2 and returns

$$o \sim \text{Ber}(\sigma(r^\star(\tau_1) - r^\star(\tau_2))),$$

where $\sigma(\cdot)$ is a link function, and $r^\star(\cdot)$ is the underlying reward function.

Here o is the human preference over τ_1, τ_2. The output $o = 1$ indicates $\tau_1 \succ \tau_2$, and $o = 0$ indicates $\tau_2 \succ \tau_1$. Moreover, we have the following assumption regarding the link function.

Assumption II.5 (Link function)**.** The link function σ satisfies the following properties:

- Completeness: $\sigma(0) = \frac{1}{2}$, and for any $x \in (-H, H)$, we have $\sigma(x) + \sigma(-x) = 1$.
- Regularity: For any $x \in (-H, H)$, we have $\sigma'(x) \geq \beta$ for some constant $\beta > 0$.

The completeness assumption enforces order-invariant comparisons between trajectories [7], while the regularity assumption is standard in bandit models [43], [44] and is crucial for the existence of an optimal policy [7], [10].

III. Risk-Sensitive RLHF with Tabular MDPs

Building on the episodic MDP framework, we incorporate human feedback into the risk-sensitive criteria. First, we assume that there is an underlying reward function that guides the human feedback.

Assumption III.1 (Underlying reward)**.** There is an unknown underlying reward $r^\star \in \mathcal{R}$ for some known function set $\mathcal{R} = \{r^\star : \mathcal{T} \to [0,H]\}$. Every reward $r^\star$ consists of H reward functions, i.e., $r^\star = \{r_h^\star : \mathcal{S} \times \mathcal{A} \to [0,1]\}_{h=1}^H$, and satisfies that for every trajectory $\tau = \{(s_1,a_1), \ldots, (s_H,a_H)\} \in \mathcal{T}$, we have $r^\star(\tau) := \sum_{h=1}^H r_h^\star(s_h,a_h)$. For a fixed trajectory $\tau_0 = \{(s_{0,1},a_{0,1}), \ldots, (s_{0,H},a_{0,H})\}$, we define a regularized reward

$$r_{\tau_0}^\star(\tau) := \sum_{h=1}^H r_h^\star(s_h,a_h) - r_h^\star(s_{0,h},a_{0,h}).$$

This assumption is commonly used for comparison feedback [4], [7], [9], [10], [26], [45]. Following [7], [10], we assume that human preference follows a Bernoulli distribution with a parameter determined by a general link function σ.

We consider a black-box setting where rewards are not directly observed but inferred from human feedback, and the resulting estimates are plugged into existing CVaR RL solvers. Our design is inspired by the P2R framework [10], which maintains a confidence set of plausible rewards $\mathcal{B}_r$ and, for each trajectory τ, checks whether all $r \in \mathcal{B}_r$ agree within a tolerance; only when they disagree does it query a comparison oracle against a fixed reference τ_0 and then tighten $\mathcal{B}_r$.

We extend this idea to the risk-sensitive setting and propose *CVaR-HF-P2R*. An overview is given in Fig.1, with details in Algorithm1. The key difference from P2R is that both the disagreement threshold and the confidence-set update rule are now explicitly α-dependent, reflecting the decision maker's risk attitude: for larger α (more risk-neutral), the interface tolerates larger reward estimation error, while for smaller α (more risk-averse), it enforces stricter agreement before skipping queries. In this way, CVaR-HF-P2R adapts the oracle

query strategy and reward filtering to the underlying risk sensitivity, rather than treating all errors uniformly as in the risk-neutral case.

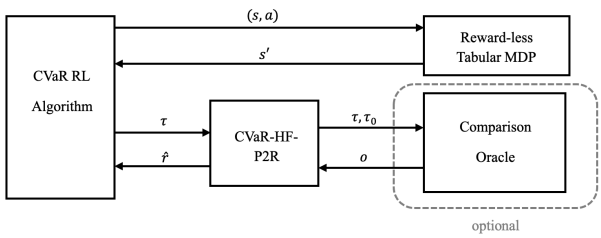

Fig. 1: Interaction protocol with CVaR-HF-P2R Interface

Algorithm 1 CVaR-HF-P2R

Given: risk tolerance $\alpha \in (0, 1]$.
Initialization: $\mathcal{B}_r \leftarrow \mathcal{R}$, $\mathcal{D} \leftarrow \{\}$, $\mathcal{D}_{\text{hist}} \leftarrow \{\}$.
Execute an arbitrary policy to collect a trajectory τ_0 with an arbitrary initial budget $b_0 \in [0, H]$.
Upon querying trajectory τ**:**
if $(\hat{r}, \tau) \in \mathcal{D}_{\text{hist}}$ **then**
  Return $\hat{r}$.
end if
if $\max_{r', r \in \mathcal{B}_r} \{r_{\tau_0}(\tau) - r'_{\tau_0}(\tau)\} < 2\epsilon_0\alpha$ **then**
  $\hat{r} \leftarrow r_{\tau_0}(\tau)$.
else
  Query the comparison oracle n times on τ and τ_0, compute the average comparison result $\bar{o}$, and set

$$\hat{r} \leftarrow \text{argmin}_{x \in [-H, H]} |\sigma(x) - \bar{o}|.$$

  Update $\mathcal{D} \leftarrow \mathcal{D} \cup (\hat{r}, \tau)$, $\mathcal{D}_{\text{hist}} \leftarrow \mathcal{D}_{\text{hist}} \cup (\hat{r}, \tau)$.
  Update $\mathcal{B}_r$ by

$$\mathcal{B}_r \leftarrow \left\{ r \in \mathcal{B}_r : \sum_{(\hat{r}, \tau) \in \mathcal{D}} (r_{\tau_0}(\tau) - \hat{r})^2 \le \frac{\epsilon_0^2 \alpha^2}{4} \right\}.$$

end if

We now provide the theoretical guarantees of Algorithm 1. Within the context of risk-sensitive RL with CVaR, we have the following lemma and its corresponding proposition, which demonstrate the robustness of tabular CVaR RL algorithms while facing some specific perturbed rewards.

Lemma III.2. *Any tabular CVaR RL algorithm* ***Alg*** *that outputs an ϵ-optimal policy ρ following an initial budget b on the true reward $r^\star$, i.e., $\left|CVaR_\alpha^\star(s_1|r^\star) - CVaR_\alpha^{\rho,b}(s_1|r^\star)\right| \le \epsilon$, is still able to output an ϵ-optimal policy even if the reward of each trajectory is perturbed by $||\varepsilon(\tau)||_\infty \le \mathcal{O}(\epsilon\alpha)$, with the same sample complexity.*

Proposition III.3. *Lemma III.2 remains valid for (s, a)-preference with $|r(s, a) - \hat{r}(s, a)| \le \mathcal{O}(\epsilon\alpha/H)$.*

Theorem III.4. *Suppose Assumptions II.5 and III.1 hold. Let $d := dim_E(\mathcal{R}, \epsilon)$ and $n = \Theta\left(\frac{\ln(d/\delta)}{\beta^2\epsilon^2\alpha^2}\right)$. By running any tabular CVaR RL algorithm* ***Alg*** *with Algorithm 1, we can learn an ϵ-optimal policy using $C(\epsilon, \alpha, \delta)$ samples (from* ***Alg****) and $\tilde{\mathcal{O}}\left(\frac{d}{\beta^2\epsilon^2\alpha^2}\right)$ queries to the comparison oracle with probability $1 - 2\delta$.*

Proposition III.5. *For (s, a)-preference, Theorem III.4 remains valid with a query complexity of $\tilde{\mathcal{O}}\left(\frac{SAH^2}{\beta^2\epsilon^2\alpha^2}\right)$, since in the tabular setup we have $d = SA$, and the perturbation bound on rewards changes from $\epsilon\alpha$ to $\epsilon\alpha/H$.*

To concretely illustrate how sample and query complexity scale when integrating CVaR-P2R with existing algorithms, we provide the following example:

Example III.6. *Running the CVaR-VI-DISC algorithm [29] within Algorithm 1 under (s, a)-preference yields an ϵ-optimal policy with $\tilde{\mathcal{O}}\left(\frac{S^2AH^3}{\epsilon^2\alpha^2}\right)$ episodes and $\tilde{\mathcal{O}}\left(\frac{SAH^2}{\beta^2\epsilon^2\alpha^2}\right)$ queries to the comparison oracle, which is sample-efficient.*

These results reinforce that, in the tabular setting, ***CVaR RLHF is no more difficult than standard CVaR RL.***

IV. Risk-Sensitive RLHF with Low-Rank MDPs

While the CVaR-P2R interface offers a sample-efficient solution for tabular CVaR-RLHF, it is fundamentally limited to episodic MDPs with small, finite state-action spaces. To overcome this limitation and scale to environments with infinite or high-dimensional state spaces, we extend our CVaR-RLHF framework with function approximation. Specifically, we focus on the class of low-rank MDPs, which subsume tabular MDPs, linear MDPs [22], and block MDPs [20], [23].

Low-rank MDPs model transition dynamics through low-dimensional latent representations. To solve such MDPs efficiently under CVaR objectives, we adopt the ELA algorithm [24] as our base RL solver. ELA achieves provable sample efficiency by iteratively estimating the latent structure through a maximum likelihood oracle and applying CVaR value iteration with optimism-based bonuses over an estimated model. It guarantees that the learned policy is near-optimal under the CVaR criterion with sample complexity scaling only with the latent dimension.

Building on this foundation and inspired by [7], [10], [24], [25], we introduce CVaR-HF-ELA, a function-approximation-based interface for CVaR-RLHF in low-rank MDPs. Our method replaces access to the true reward with a comparison oracle, enabling reward estimation from human feedback. In each episode k (Algorithm 2), the algorithm selects a reward estimate $\hat{r}^k$ based on the value function computed by ELA, uses ELA's model estimation module (ELA-ME) to obtain $\hat{\mathbb{P}}^k$ and the bonus term $\widehat{\texttt{BONS}}^k$, and then applies the ELA value-iteration-with-bonus routine (ELA-VI-BONS) with $(\hat{r}^k, \hat{\mathbb{P}}^k)$ to solve the CVaR problem, yielding a near-optimal policy $\hat{\rho}^k$ and initial budget $\hat{b}^k$. Human feedback is then queried to refine the reward, and the confidence set $\hat{\mathcal{B}}_r$ is updated via MLE

with radius $\hat{\beta}_R$, using a standard log-likelihood formulation as in [7], [9], [25], [26].

$$\mathcal{L}_k(r) := \sum_{i \le k} \log(\tilde{\sigma}(o_i, r_{\tau_0}(\tau_i))),$$

where

$$\tilde{\sigma}(o_i, r_{\tau_0}(\tau_i)) := o_i \cdot \sigma(r_{\tau_0}(\tau_i)) + (1 - o_i) \cdot \sigma(-r_{\tau_0}(\tau_i)).$$

Algorithm 2 CVaR-HF-ELA

Given: risk tolerance $\alpha \in (0, 1]$, number of iterations K, models $\mathcal{F} = \{\Psi, \Phi\}$.
Initialization: $\mathcal{B}_r \leftarrow \mathcal{R}$.
Execute an arbitrary policy to collect trajectory $\tau_0 = (s_{0,1}, a_{0,1}, \ldots, s_{0,H}, a_{0,H})$ with an arbitrary initial budget b.
for $k = 1, \ldots, K$ **do**
  Set $\hat{r}_k \leftarrow \text{argmin}_{r \in \hat{\mathcal{R}}_k} V^{\hat{\rho}^k}_{1, \hat{\mathbb{P}}_k, \text{BON}}(s_{k,1}, b; r_{\tau_0})$.
  for $h = 1, \ldots, H-1$ **do**
    $(\hat{\mathbb{P}}^k_h, \widehat{\text{BONS}}^k_h) \leftarrow \texttt{ELA-ME}(\hat{\psi}^k_h, \hat{\phi}^k_h)$.
  end for
  $(\hat{\rho}^k, \hat{b}^k) \leftarrow \texttt{ELA-VI-BONS}(\hat{r}^k, \hat{\mathbb{P}}^k, \hat{\psi}^k, \hat{\phi}^k)$.
  Execute the policy $\hat{\rho}^k$ with initial budget $\hat{b}^k$ to collect the trajectory $\tau_k = (s_{k,1}, a_{k,1}, \ldots, s_{k,H}, a_{k,H})$.
  Compare two trajectories τ_k, τ_0 and collect comparison oracle o_k from human feedback.
  Update the reward confidence set

$$\mathcal{B}_r \leftarrow \left\{ r \in \mathcal{B}_r : \mathcal{L}_k(r) > \max_{r' \in \mathcal{R}} \mathcal{L}_k(r') - \hat{\beta}_R \right\}.$$

end for

We first define the regret of Algorithm 2 as $\text{Regret}(K) = \sum_{k=1}^{K} \text{CVaR}_\alpha^{\pi^{\rho^\star, b^\star}}(s_{k,1}; r^\star) - \text{CVaR}_\alpha^{\pi^{\hat{\rho}^k, \hat{b}^k}}(s_{k,1}; r^\star)$, where $\pi^{\rho^\star, b^\star}$ is the optimal policy of CVaR RL running on true $r^\star$ and $\mathbb{P}^\star$ and $\pi^{\hat{\rho}^k, \hat{b}^k}$ is the output of CVaR-HF-ELA at k.

To facilitate the analysis, we relate this regret to a centered reward $r^\star_{\tau_0}$ defined using a fixed reference trajectory τ_0:

Lemma IV.1. *For the centered reward $r^\star_{\tau_0}$, we have*

$$\begin{aligned} &\textit{Regret}(K) \\ &\le \underbrace{\sum_{k=1}^{K} \textit{CVaR}_\alpha^{\pi^{\rho^\star, b^\star}}(s_{k,1}; r^\star_{\tau_0}) - \textit{CVaR}_\alpha^{\pi^{\hat{\rho}^k, \hat{b}^k}}(s_{k,1}; r^\star_{\tau_0})}_{\textit{Regret}(K | r^\star_{\tau_0})} \\ &\quad + \alpha^{-1} H. \end{aligned} \tag{3}$$

This lemma is key to our analysis: it shows that controlling the regret under the comparison-based reward $r^\star_{\tau_0}$ already bounds the original regret under $r^\star$ up to an additive constant. Hence, we can focus on the term $\text{Regret}(K | r^\star_{\tau_0})$.

Next, we note that Algorithm 2 adopts the MLE-based confidence set construction from [7], which maintains a likelihood-based confidence set over reward functions. This choice is crucial for handling infinite reward classes, whereas alternatives such as [10] assume a finite $\mathcal{R}$. The following lemma guarantees that the true reward function remains in the confidence set throughout learning.

Lemma IV.2. *(Lemma 11 in [7]) With properly choosing*

$$\hat{\beta}_{\mathcal{R}} = c \log(K) \cdot N_B(R, ||\cdot||_\infty, 1/K)/\delta)$$

and positive constant $\delta \in (0, 1]$, we have $r^\star \in \hat{\mathcal{R}}_k$ holds for every $k \in [K]$ with probability at least $1 - \delta$. Here $N_B(\mathcal{R}, ||\cdot||_\infty, 1/K)$ is the $1/K$-bracketing number of $\mathcal{R}$ under norm $||\cdot||_\infty$.

Building on these key lemmas, we now establish the following theorem, which provides upper bounds on the regret and sample complexity of CVaR-HF-ELA.

Theorem IV.3. *Suppose Assumptions II.5 and III.1 hold. For some positive constant $\delta > 0$, we set the estimation radius $\hat{\beta}_{\mathcal{R}} = c \log(K \cdot N_B(\mathcal{R}, ||\cdot||_\infty, 1/K)/\delta)$ for some constant c. Then, with probability at least $1 - 4\delta$, the regret of Algorithm 2 satisfies:*

$$\begin{aligned} &\textit{Regret}(K) \\ &\le \tilde{\mathcal{O}} \left(\alpha^{-1} d_E(\mathcal{R}) \cdot \frac{\hat{\beta}_{\mathcal{R}}}{\beta} + \alpha^{-1} H^3 A d^2_{\textit{rank}} \sqrt{K} \cdot \sqrt{\log \frac{|\mathcal{F}|}{\delta}} \right), \end{aligned}$$

where $N_B(\mathcal{R}, ||\cdot||_\infty, 1/K)$ is the $1/K$-bracketing number of $\mathcal{R}$ under norm $||\cdot||_\infty$ and $d_E(R) := \textit{dim}_E(\mathcal{R}, 1/\sqrt{K})$ is the eluder dimension of $\mathcal{R}$.

Remark IV.4. *Our CVaR-RLHF framework is **not** restricted to low-rank settings such as ELA; its modular design allows integration with* any provably efficient *function-approximation algorithm for CVaR-based RL, making it applicable to large-scale or continuous environments where tabular methods are infeasible. We introduce a general interface,* CVaR-HF-FAG *(**F**unction-**A**pproximation **G**eneralized), which cleanly decouples reward estimation from the underlying risk-sensitive RL solver. In each round k, the learner maintains a confidence set $\mathcal{B}_r$, selects a reward estimate $\hat{r}_k$ that is pessimistic for the CVaR objective under the current model, updates the dynamics $\hat{\mathbb{P}}_k$ via a function-approximation–based model estimation subroutine (`CVaR-ME-Alg`), and calls a downstream `CVaR-Solver` to obtain a near-optimal policy. The resulting policy is executed to generate a trajectory, and human comparisons are then used to refine $\mathcal{B}_r$ through a comparison oracle. Note that, when the function approximation algorithm in `CVaR-ME-Alg` (e.g. [24]) is provably sample-efficient and the `CVaR-Solver` outputs a near-optimal CVaR policy (e.g. [24], [25], [29]), the regret of our interface can be bounded using the same techniques as Theorem IV.3.*

V. Conclusion

In this work, we have proposed a provably efficient online risk-sensitive RLHF framework using static CVaR. We have proposed CVaR-P2R, a black-box reward estimation interface

for tabular MDPs that integrates seamlessly with existing CVaR RL algorithms and achieves provable sample efficiency with minimal additional complexity. To address scalability, we have extended our framework to infinite state spaces by leveraging low-rank structure and function approximation. We have proposed CVaR-HF-ELA algorithm and have demonstrated efficient learning with regret guarantees. Finally, we have presented a general interface CVaR-HF-FAG that abstracts our methodology and can accommodate any provably efficient function approximation solver for CVaR RL.

REFERENCES

[1] R. S. Sutton, "Reinforcement Learning: An Introduction," *A Bradford Book*, 2018.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the Game of Go Without Human Knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[3] V. Mnih, "Playing Atari with Deep Reinforcement Learning," *arXiv preprint arXiv:1312.5602*, 2013.

[4] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward Learning from Human Preferences and Demonstrations in Atari," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[7] Y. Chen, Y. Du, P. Hu, S. Wang, D. Wu, and L. Huang, "Provably Efficient Iterated CVaR Reinforcement Learning with Function Approximation and Human Feedback," in *Proc. International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] B. Zhu, M. Jordan, and J. Jiao, "Principled Reinforcement Learning with Human Feedback from Pairwise or K-Wise Comparisons," in *Proc. International Conference on Machine Learning*, Honolulu, Hawaii, 2023, pp. 43 037–43 067.

[10] Y. Wang, Q. Liu, and C. Jin, "Is RLHF More Difficult than Standard RL? A Theoretical Perspective," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] Y. Sun, N. Salami Pargoo, P. Jin, and J. Ortiz, "Optimizing autonomous driving for safety: A human-centric approach with llm-enhanced rlhf," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 76–80.

[12] S. Xu, B. Yue, H. Zha, and G. Liu, "Uncertainty-aware preference alignment in reinforcement learning from human feedback," in *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

[13] Y. Zhao, J. Aguilar Escamilla, W. Lu, and H. Wang, "Ra-pbrl: Provably efficient risk-aware preference-based reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 60 835–60 871, 2025.

[14] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of Conditional Value-at-Risk," *Journal of Risk*, vol. 2, pp. 21–42, 2000.

[15] C. Acerbi and D. Tasche, "On the Coherence of Expected Shortfall," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1487–1503, 2002.

[16] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the CVaR via Sampling," in *Proc. The AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Austin, TX, 2015.

[17] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy Gradient for Coherent Risk Measures," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[18] Y. Chow and M. Ghavamzadeh, "Algorithms for CVaR Optimization in MDPs," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[19] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-Sensitive and Robust Decision-Making: A CVaR Optimization Approach," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[20] S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudik, and J. Langford, "Provably Efficient RL with Rich Observations via Latent State Decoding," in *International Conference on Machine Learning*, Long Beach, CA, 2019, pp. 1665–1674.

[21] X. Ni and L. Lai, "Risk-Sensitive Reinforcement Learning with ϕ-Divergence-Risk," *IEEE Transactions on Information Theory*, 2025.

[22] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably Efficient Reinforcement Learning with Linear Function Approximation," in *Proc. Conference on Learning Theory*, Graz, Austria, 2020, pp. 2137–2143.

[23] Y. Efroni, D. Misra, A. Krishnamurthy, A. Agarwal, and J. Langford, "Provable RL with Exogenous Distractors via Multistep Inverse Dynamics," *arXiv preprint arXiv:2110.08847*, 2021.

[24] Y. Zhao, W. Zhan, X. Hu, H. Leung, F. Farnia, W. Sun, and J. D. Lee, "Provably Efficient CVaR RL in Low-Rank MDPs," *arXiv preprint arXiv:2311.11965*, 2023.

[25] K. Wang, N. Kallus, and W. Sun, "Near-Minimax-Optimal Risk-Sensitive Reinforcement Learning with CVaR," in *Proc. International Conference on Machine Learning*, Honolulu, Hawaii, 2023, pp. 35 864–35 907.

[26] W. Zhan, M. Uehara, N. Kallus, J. D. Lee, and W. Sun, "Provable Offline Reinforcement Learning with Human Feedback," *arXiv preprint arXiv:2305.14816*, 2023.

[27] L. A. Prashanth, "Policy Gradients for CVaR-Constrained MDPs," in *Proc. International Conference on Algorithmic Learning Theory*, Bled, Slovenia, 2014, pp. 155–169.

[28] Y. Du, S. Wang, and L. Huang, "Provably Efficient Risk-Sensitive Reinforcement Learning: Iterated CVaR and Worst Path," in *Proc. International Conference on Learning Representations*, 2022.

[29] X. Ni, G. Liu, and L. Lai, "Risk-Sensitive Reward-Free Reinforcement Learning with CVaR," in *Proc. International Conference on Machine Learning*, Vienna, Austria, 2024.

[30] X. Ni and L. Lai, "Robust Risk-Sensitive Reinforcement Learning with Conditional Value-at-Risk," in *Proc. IEEE Information Theory Workshop*, Shenzhen, China, 2024.

[31] Q. Zhang, S. Leng, X. Ma, Q. Liu, X. Wang, B. Liang, Y. Liu, and J. Yang, "CVaR-Constrained Policy Optimization for Safe Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[32] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath, "Coherent Measures of Risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[33] N. Bäuerle and J. Ott, "Markov Decision Processes with Average-Value-at-Risk Criteria," *Mathematical Methods of Operations Research*, vol. 74, pp. 361–379, 2011.

[34] R. Wang, R. R. Salakhutdinov, and L. Yang, "Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6123–6135, 2020.

[35] C. Jin, Q. Liu, and S. Miryoosefi, "Bellman Eluder Dimension: New Rich Classes of RL problems, and Sample-Efficient Algorithms," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 406–13 418, 2021.

[36] G. Li, P. Kamath, D. J. Foster, and N. Srebro, "Understanding the Eluder Dimension," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 737–23 750, 2022.

[37] A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup, "Invariant Causal Prediction for Block MDPs," in *Proc. International Conference on Machine Learning*, 2020, pp. 11 214–11 224.

[38] S. Sodhani, A. Zhang, and J. Pineau, "Multi-Task Reinforcement Learning with Context-Based Representations," in *Proc. International Conference on Machine Learning*, 2021, pp. 9767–9779.

[39] S. Sodhani, F. Meier, J. Pineau, and A. Zhang, "Block Contextual MDPs for Ccontinual Learning," in *Proc. Learning for Dynamics and Control Conference*, 2022, pp. 608–623.

[40] A. Modi, J. Chen, A. Krishnamurthy, N. Jiang, and A. Agarwal, "Model-Free Representation Learning and Exploration in Low-Rank MDPs," *Journal of Machine Learning Research*, vol. 25, no. 6, pp. 1–76, 2024.

[41] A. Agarwal, S. Kakade, A. Krishnamurthy, and W. Sun, "Flambe: Structural Complexity and Representation Learning of Low Rank MDPs," *Ad-*

vances in Neural Information Processing Systems, vol. 33, pp. 20 095–20 107, 2020.

[42] C. Ye, W. Xiong, Y. Zhang, H. Dong, N. Jiang, and T. Zhang, "Online iterative reinforcement learning from human feedback with general preference model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 81 773–81 807, 2024.

[43] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric Bandits: The Generalized Linear Case," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[44] L. Li, Y. Lu, and D. Zhou, "Provably Optimal Algorithms for Generalized Linear Contextual Bandits," in *Proc. International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 2071–2080.

[45] W. Zhan, M. Uehara, W. Sun, and J. D. Lee, "How to Query Human Feedback Efficiently in RL?" *arXiv preprint arXiv:2305.18505*, 2023.