Near-Minimax-Optimal Risk-Sensitive Reinforcement Learning with CVaR

Kaiwen Wang [1 2] **Nathan Kallus** [2] **Wen Sun** [1]

Abstract

In this paper, we study risk-sensitive Reinforcement Learning (RL), focusing on the objective of Conditional Value at Risk (CVaR) with risk tolerance τ. Starting with multi-arm bandits (MABs), we show the minimax CVaR regret rate is $\Omega(\sqrt{\tau^{-1}AK})$, where A is the number of actions and K is the number of episodes, and that it is achieved by an Upper Confidence Bound algorithm with a novel Bernstein bonus. For online RL in tabular Markov Decision Processes (MDPs), we show a minimax regret lower bound of $\Omega(\sqrt{\tau^{-1}SAK})$ (with normalized cumulative rewards), where S is the number of states, and we propose a novel bonus-driven Value Iteration procedure. We show that our algorithm achieves the optimal regret of $\widetilde{\mathcal{O}}(\sqrt{\tau^{-1}SAK})$ under a continuity assumption and in general attains a near-optimal regret of $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{SAK})$, which is minimax-optimal for constant τ. This improves on the best available bounds. By discretizing rewards appropriately, our algorithms are computationally efficient.

1. Introduction

Reinforcement Learning (RL) (Sutton & Barto, 2018) is the canonical framework for sequential decision making under uncertainty, with applications in personalizing recommendations (Bottou et al., 2013), robotics (Rajeswaran et al., 2017), healthcare (Murphy, 2003) and education (Singla et al., 2021). In vanilla RL, the objective is to maximize the *average* of returns, the cumulative rewards collected by the policy. As RL is increasingly applied in consequential settings, it is often necessary to account for risk beyond solely optimizing for the average.

[1]Computer Science, Cornell University [2]Operations Research and Information Engineering, Cornell Tech. Correspondence to: Kaiwen Wang <https://kaiwenw.github.io>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Conditional Value at Risk (CVaR) is a popular coherent measure of risk (Rockafellar & Uryasev, 2000; Filippi et al., 2020). For a random return X (where higher is better), the CVaR with risk tolerance $\tau \in (0, 1]$ is defined as

$$\mathrm{CVaR}_\tau(X) := \sup_{b\in\mathbb{R}}(b - \tau^{-1}\mathbb{E}[(b - X)^+]), \quad (1)$$

where $x^+ = \max(x, 0)$. $\mathrm{CVaR}_\tau(X)$ is the average outcome among the *worst* τ-percent of cases, and when X is continuous this exactly corresponds to those less than or equal to the τ-th quantile (Acerbi & Tasche, 2002), *i.e.*,

$$\mathrm{CVaR}_\tau(X) = \mathbb{E}[X \mid X \le F_X^\dagger(\tau)], \quad (2)$$

where $F_X^\dagger(\tau) = \inf\{x : F_X(x) \ge \tau\}$ is the τ-th quantile of X, a.k.a. the Value at Risk (VaR). A high risk tolerance $\tau = 1$ recovers the risk-neutral expectation, *i.e.*, $\mathrm{CVaR}_1(X) = \mathbb{E}X$. As τ decreases, CVaR_τ models the worst-case outcome, *i.e.*, $\lim_{\tau\to 0}\mathrm{CVaR}_\tau(X) = \operatorname{ess\,inf} X$. In the CVaR RL model we consider, X is the return of a policy, so our objective captures the tail-risk of the returns distribution. Another motivating perspective is that CVaR RL is equivalent to the robust MDP model, *i.e.*, expected value under worst-case perturbation of the transition kernel (Chow et al., 2015). Thus, CVaR RL is an attractive alterantive to vanilla RL in safety-critical applications.

In this paper, we provide algorithms with state-of-the-art regret guarantees for tabular, online decision making with the CVaR objective. To start, we prove tight lower bounds on the expected CVaR regret (formalized in Section 2) for both multi-arm bandit (MAB) and RL problems. We then propose BERNSTEIN-UCB, an Upper Confidence Bound (UCB) algorithm with a novel bonus constructed using Bernstein's inequality, and we prove it is minimax-optimal[1]. Compared to Brown-UCB (Tamkin et al., 2019), BERNSTEIN-UCB is minimax-optimal in general, without requiring reward distributions to be continuous.

We then turn to tabular RL with the CVaR objective. We propose CVaR-UCBVI, a novel bonus-driven Value Iteration (VI) algorithm in an augmented MDP. The augmented

[1]Following Azar et al. (2017), we say an algorithm is *minimax optimal* if its regret matches (up to log terms) our novel minimax lower bound, in all problem parameters. Sometimes, this is also referred to as "nearly-minimax-optimal" (Zhou et al., 2021a).

MDP framework of Bäuerle & Ott (2011) reveals Bellman equations for CVaR RL that served as the initial catalyst for the VI approach. We provide two choices of bonuses for CVaR-UCBVI, based on Hoeffding's and Bernstein's inequalities. With the Bernstein bonus, we guarantee a CVaR regret of $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{SAK})$, where K is the number of episodes, S, A are the sizes of the state and action spaces, respectively. This improves over the previous bound of $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{S^3AHK})$ (Bastani et al., 2022) in S and H, the horizon length. Note that we work under the normalized returns model, so there should be no H scaling in the bound (Jiang & Agarwal, 2018). If τ is a constant, our result is already minimax-optimal. Surprisingly, however, our lower bound actually scales as $\tau^{-1/2}$. Under an assumption that the returns of any policies are continuously distributed with lower-bounded density, we improve the upper bound on the regret of CVaR-UCBVI with the Bernstein bonus to $\widetilde{\mathcal{O}}(\sqrt{\tau^{-1}SAK})$. This establishes CVaR-UCBVI as the first algorithm with minimax-optimal regret for risk-sensitive CVaR RL, under the continuity assumption. Our key technical innovation is decomposing the regret using a novel simulation lemma for CVaR RL and precisely bounding the sum of variance bonuses with the Law of Total Variance (Azar et al., 2017).

1.1. Related Literature

CVaR MAB: Kagrecha et al. (2019) proposed a successive rejects algorithm for best CVaR arm identification, but it does not have regret guarantees. Tamkin et al. (2019) proposed two algorithms for CVaR MAB and analyze upper bounds on their regret, but not lower bounds. Their "CVaR-UCB" builds a confidence band for the reward distribution of each arm via Dvoretzky-Kiefer-Wolfowitz inequality, resulting in an optimistic estimate of CVaR. This leads to a suboptimal τ^{-1} dependence in the regret but may empirically work better if τ is not approaching 0. Their "Brown-UCB" is structurally similar to our BERNSTEIN-UCB, but they use a Hoeffding bonus that ensures optimism only if all arms have continuously distributed rewards (Brown, 2007, Theorem 4.2). We propose a Bernstein bonus that attains the minimax-optimal rate $\sqrt{\tau^{-1}AK}$ without any assumptions on the reward distribution.

Regret bounds for CVaR RL: To the best of our knowledge, Bastani et al. (2022) is the first and only work with regret bounds for CVaR RL (formalized in Section 2). Their algorithm iteratively constructs optimistic MDPs by routing unexplored states to a sink state with the maximum reward. This approach leads to a CVaR regret bound of $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{S^3AHK})$ (Bastani et al., 2022, Theorem 4.1), which is sub-optimal. The authors conjectured that bonus-based optimism could improve the bound by a $S\sqrt{H}$ factor. Our proposed CVaR-UCBVI indeed enjoys these improvements, leading to a $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{SAK})$ regret guarantee in Theorem 5.3. If returns are continuously distributed, we further improve the τ dependence, leading to the minimax-optimal result in Theorem 5.5.

CVaR RL without regret guarantees: Keramati et al. (2020) proposed a distributional RL approach (Bellemare et al., 2017) for RL with the CVaR objective. A key difference is that Keramati et al. (2020) focuses on the easier task of identifying a policy with high CVaR. On the other hand, Bastani et al. (2022) and our work focuses on algorithms with low CVaR regret, which guarantees safe exploration. Note that low-regret methods can be converted into probably approximately correct (PAC) CVaR RL, by taking the uniform mixture of policies from the low-regret algorithm.

Tamar et al. (2015) derived the policy gradient for the CVaR RL objective and showed asymptotic convergence to a local optimum. Chow & Ghavamzadeh (2014) developed actor-critic algorithms for the mean-CVaR objective, *i.e.*, maximizing expected returns subject to a CVaR constraint. Another motivating perspective for CVaR RL is its close ties to robust MDPs (Wiesemann et al., 2013). Specifically, Chow et al. (2015, Proposition 1) showed that the CVaR of returns is equivalent to the expected returns under the worst-case perturbation of the transition kernel in some uncertainty set. While the uncertainty set is not rectangular, Chow et al. (2015) derived tractable robust Bellman equations and proved convergence to a globally optimal CVaR policy. However, these methods for CVaR RL do not lead to low-regret algorithms, which is our focus.

Risk-sensitive RL with different risk measures: Prior works have also proved risk-sensitive RL regret bounds in the context of other risk measures that are not directly comparable to the CVaR RL setting we consider. Fei et al. (2020; 2021); Liang & Luo (2022) showed Bellman equations and regret guarantees with the entropic risk measure based on an exponential utility function. Du et al. (2022); Lam et al. (2023) studied the more conservative *Iterated* CVaR objective, which considers the risk of the reward-to-go at every step along the trajectory. In contrast, our setup aims to holistically maximize the CVaR of the *total* returns.

Risk-sensitive regret lower bounds: Fei et al. (2020); Liang & Luo (2022) showed regret lower bounds for risk-sensitive RL with the entropic risk measure. We show *tight* lower bounds for risk-sensitive MAB and RL with the CVaR objective, which to the best of our knowledge are the first lower bounds for this problem.

Safety in offline RL: While our focus is online RL, risk-aversion has also been studied in offline RL. Some past works include offline learning with risk measures (Urpí et al., 2021) and distributional robustness (Panaganti et al., 2022; Si et al., 2020; Kallus et al., 2022; Zhou et al., 2021b).

2. Problem Setup

As warmup, we consider CVaR_τ regret in a multi-arm bandit (MAB) problem with K episodes. At each episode $k \in [K]$, the learner selects an arm $a_k \in \mathcal{A}$ and observes reward $r_k \sim \nu(a_k)$, where $\nu(a)$ is the reward distribution of arm a. The learner's goal is to compete with the arm with the highest CVaR_τ value, *i.e.*, $a^\star = \max_{a\in\mathcal{A}} \mathrm{CVaR}_\tau(\nu(a))$, and minimize the regret, defined as $\mathrm{Regret}_\tau^{\mathrm{MAB}}(K) = \sum_{k=1}^K \mathrm{CVaR}_\tau(\nu(a^\star)) - \mathrm{CVaR}_\tau(\nu(a_k))$.

The focus of this paper is online RL, which generalizes the MAB (where $S = H = 1$). The statistical model is tabular Markov Decision Processes (MDPs) (Agarwal et al., 2021), with finite state space $\mathcal{S}$ of size S, finite action space $\mathcal{A}$ of size A, and horizon H. Let $\Pi_{\mathcal{H}}$ be the set of history-dependent policies, so each policy is $\pi = (\pi_h : \mathcal{S}\times\mathcal{H}_h \to \Delta(\mathcal{A}))_{h\in[H]}$ where $\mathcal{H}_h = \{(s_i, a_i, r_i)\}_{i\in[h-1]}$ is the history up to time h. At each episode $k \in [K]$, the learner plays a history-dependent policy $\pi^k \in \Pi_{\mathcal{H}}$, which induces a distribution over trajectories as follows. First, start at the fixed initial state $s_1 \in \mathcal{S}$. Then, for each time $h = 1, 2, \ldots, H$, sample an action $a_h \sim \pi_h^k(s_h, \mathcal{H}_h)$, which leads to a reward $r_h \sim R(s_h, a_h)$ and the next state $s_{h+1} \sim P^\star(s_h, a_h)$. Here, $P^\star : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the *unknown* Markov transition kernel and $R : \mathcal{S}\times\mathcal{A} \to \Delta([0,1])$ is the *known* reward distribution. The return is the sum of rewards from this process, $R(\pi) = \sum_{h=1}^H r_h$, which is a random variable. We posit the return is almost surely *normalized* in that $R(\pi) \in [0,1]$ w.p. 1 (as in Jiang & Agarwal, 2018, Section 2.1). We note normalized returns allows for sparsity in the rewards, and thus is strictly more general for regret upper bounds. Many prior works do not normalize, so their returns may scale with H. When comparing to their bounds, we make the scaling consistent by dividing them by H.

We focus on the setting we call *CVaR RL*, in which the learner's goal is to compete with a CVaR_τ-optimal policy, *i.e.*, $\pi^\star \in \arg\max_{\pi\in\Pi_{\mathcal{H}}} \mathrm{CVaR}_\tau(R(\pi))$. Toward this end, we define the regret as $\mathrm{Regret}_\tau^{\mathrm{RL}}(K) = \sum_{k=1}^K \mathrm{CVaR}_\tau^\star - \mathrm{CVaR}_\tau(R(\pi^k))$, where $\mathrm{CVaR}_\tau^\star = \mathrm{CVaR}_\tau(R(\pi^\star))$. CVaR RL captures vanilla risk-neutral RL when $\tau = 1$ and Worst Path RL (Du et al., 2022) when $\tau \to 0$. We prove lower bounds in expected regret. For upper bounds, we give high probability regret bounds, which implies upper bounds (with the same dependencies on problem parameters) in expected regret by integrating over the failure probability $\delta \in (0,1)$.

Notation: $[i : j] = \{i, i+1, \ldots, j\}$, $[n] = [1 : n]$ and $\Delta(\mathcal{S})$ is the set of distributions on $\mathcal{S}$. We set $L = \log(HSAK/\delta)$ (for MAB, $L = \log(AK/\delta)$), where δ is the desired failure probability provided to the algorithm. Please see Table 1 for a comprehensive list of notations.

3. Lower Bounds

We start with the minimax lower bound for CVaR_τ MAB.

Theorem 3.1. *Fix any $\tau \in (0, 1/2)$, $A \in \mathbb{N}$. For any algorithm, there is a MAB problem with Bernoulli rewards s.t. if $K \geq \sqrt{\frac{A-1}{8\tau}}$, then $\mathbb{E}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K)\right] \geq \frac{1}{24e}\sqrt{\frac{(A-1)K}{\tau}}$.*

Proof Sketch Our proof is inspired by the lower bound construction for the vanilla MAB (Lattimore & Szepesvári, 2020, Theorem 15.2). The key idea is to fix any learner, and construct two MAB instances that appear similar to the learner but in reality have very different CVaR value. Specifically, for any $\varepsilon \in (0,1)$, we need two reward distributions such that their KL-divergence is $\mathcal{O}(\varepsilon^2\tau^{-1})$ while their CVaRs differ by $\Omega(\tau^{-1}\varepsilon)$. We show that $\mathrm{Ber}(1-\tau)$ and $\mathrm{Ber}(1-\tau+\varepsilon)$ satisfy this. ■

Compared to the vanilla MAB minimax lower bound of $\Omega(\sqrt{AK})$, our result for CVaR_τ MAB has an extra $\sqrt{\tau^{-1}}$ factor. This proves that it is information-theoretically harder to be more risk-averse with CVaR_τ. While Brown-UCB of Tamkin et al. (2019) appears to match this lower bound, their proof hinges on the continuity of reward distributions, which is invalid for Bernoulli rewards. In Theorem 4.1, we show that BERNSTEIN-UCB is minimax-optimal over all reward distributions.

We next extend the above result to the RL setting.

Corollary 3.2. *Fix any $\tau \in (0, 1/2)$, $A, H \in \mathbb{N}$. For any algorithm, there is an MDP (with $S = \Theta(A^{H-1})$) s.t. if $K \geq \sqrt{\frac{S(A-1)}{8\tau}}$, then $\mathbb{E}\left[\mathrm{Regret}_\tau^{\mathrm{RL}}(K)\right] \geq \frac{1}{24e}\sqrt{\frac{S(A-1)K}{\tau}}$.*

The argument is to show that a class of MDPs with rewards only at the last layer essentially reduces to a MAB with exponentially many actions. Thus, the hardest CVaR RL problems are actually very big CVaR MAB problems. The bound does not scale with H as we've assumed returns to be normalized in $[0,1]$ (Jiang & Agarwal, 2018).

4. Risk Sensitive MAB

In this section, we propose a simple modification to the classic Upper Confidence Bound (UCB) with a novel Bernstein bonus that enjoys minimax-optimal regret. In the classic UCB algorithm (Auer et al., 2002), the bonus quantifies the confidence band from Hoeffding's inequality. Instead, we propose to build a confidence band of $\mu(b,a) = \mathbb{E}_{R\sim\nu(a)}[(b-R)^+]$ by using a Bernstein-based bonus (Eq. (3)). The standard deviation $\sqrt{\tau}$ in our bonus is crucial for obtaining the minimax-optimal regret.

Theorem 4.1. *For any $\delta \in (0,1)$, w.p. at least $1-\delta$,* BERNSTEIN-UCB *with $\varepsilon \leq \sqrt{A/2\tau K}$ enjoys*

$$\mathrm{Regret}_\tau^{\mathrm{MAB}}(K) \leq 4\sqrt{\tau^{-1}AKL} + 16\tau^{-1}AL^2.$$

Algorithm 1 BERNSTEIN-UCB

1: **Input:** risk tolerance τ, number of episodes K, failure probability δ, approximation parameter ε.
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3: Compute counts $N_k(a) = 1 \vee \sum_{i=1}^{k-1} \mathbb{I}[a_i = a]$.
4: Define pessimistic estimate of $\mu(b,a) = \mathbb{E}_{R\sim\nu(a)}[(b-R)^+]$, *i.e.*, for all b, a,

$$\widehat{\mu}_k(b,a) = \frac{1}{N_k(a)} \sum_{i=1}^{k-1} (b - r_i)^+ \mathbb{I}[a_i = a], \qquad \textsc{Bon}_k(a) = \sqrt{\frac{2\tau \log(AK/\delta)}{N_k(a)}} + \frac{\log(AK/\delta)}{N_k(a)}. \tag{3}$$

5: Compute ε-optimal solutions $\widehat{b}_{a,k}$, *i.e.*, for all a,

$$\widehat{f}_k(\widehat{b}_{a,k}, a) \geq \max_{b\in[0,1]} \widehat{f}_k(b,a) - \varepsilon, \quad \text{where,} \quad \widehat{f}_k(b,a) = b - \tau^{-1}(\widehat{\mu}_k(b,a) - \textsc{Bon}_k(a)).$$

6: Compute and pull the action for this episode, $a_k = \arg\max_{a\in\mathcal{A}} \widehat{f}_k(\widehat{b}_{a,k}, a)$. Receive reward $r_k \sim \nu(a_k)$.
7: **end for**

Proof Sketch First, we use Bernstein's inequality to build a confidence band of $\mu(b,a)$ at $b_a^\star = \arg\max_{b\in[0,1]}\{b - \tau^{-1}\mu(b,a)\}$. Conveniently, $b_a^\star$ is the τ-th quantile of $\nu(a)$, hence $\mathrm{Var}_{R\sim\nu(a)}((b_a^\star - R)^+) \leq \tau$. This proves pessimism with our Bernstein-based bonus, *i.e.*, $\widehat{\mu}_k(b_a^\star, a) - \textsc{Bon}_k(a) \leq \mu(b_a^\star, a)$. Pessimism in turn implies *optimism* in CVaR, *i.e.*, $\mathrm{CVaR}_\tau^\star \leq \widehat{\mathrm{CVaR}}_\tau^k := \max_{b\in[0,1]}\{b - \tau^{-1}(\widehat{\mu}_k(b,a_k) - \textsc{Bon}_k(a_k))\}$. This allows us to decompose regret into (1) the sum of bonuses plus (2) the difference between empirical and true CVaR of $\nu(a_k)$. (1) is handled using a standard pigeonhole argument. To bound (2), we prove a new concentration inequality for CVaR that holds for any bounded random variable (Theorem C.6), which may be of independent interest. ■

Up to log terms, the resulting bound matches our lower bound in Theorem 3.1, proving that BERNSTEIN-UCB is minimax-optimal. As noted earlier, under the assumption that rewards are continuous, Brown-UCB (Tamkin et al., 2019) also matches our novel lower bound. When working with continuous distributions, CVaR takes the convenient form in Eq. (2), which roughly suggests that Hoeffding's inequality on the lower τN data points suffices for a CVaR concentration bound (Brown, 2007, Theorem 4.2). This is why the bonus of Brown-UCB, which is $\sqrt{\frac{5\tau\log(3/\delta)}{N_k(a)}}$, does not yield optimism when continuity fails. In general, the $\frac{1}{N_k(a)}$ term in our bonus from Bernstein's inequality is needed for proving optimism in general.

Computational Efficiency: In Line 5, the objective function $\widehat{f}_k(\cdot, a)$ is concave and unimodal. So, its optimal value can be efficiently approximated, *e.g.*, by golden-section search (Kiefer, 1953) or gradient ascent in $1/\varepsilon^2 = \mathcal{O}(\tau K/A)$ steps (Boyd et al., 2004). Thus, BERNSTEIN-UCB is *both* minimax-optimal for regret and computationally efficient.

5. Risk Sensitive RL

We now shift gears to CVaR RL. First, we review the augmented MDP framework due to Bäuerle & Ott (2011) and derive Bellman equations for our problem (Bellemare et al., 2023). Using this perspective, we propose CVaR-UCBVI, a bonus-driven Value Iteration algorithm in the augmented MDP, which we show enjoys strong regret guarantees.

5.1. Augmented MDP and Bellman Equations

For any history-dependent $\pi \in \Pi_{\mathcal{H}}$, timestep $h \in [H]$, state $s_h \in \mathcal{S}$, budget $b_h \in [0,1]$, and history $\mathcal{H}_h$, define

$$V_h^\pi(s_h, b_h; \mathcal{H}_h) = \mathbb{E}_\pi\left[\left(b_h - \textstyle\sum_{t=h}^H r_t\right)^+ \mid s_h, \mathcal{H}_h\right].$$

Then, the CVaR RL objective can be formulated as,

$$\begin{aligned}\mathrm{CVaR}_\tau^\star &= \max_{\pi\in\Pi_{\mathcal{H}}} \max_{b\in[0,1]}\{b - \tau^{-1} V_1^\pi(s_1, b)\} \\ &= \max_{b\in[0,1]}\{b - \tau^{-1} \min_{\pi\in\Pi_{\mathcal{H}}} V_1^\pi(s_1, b)\}. \end{aligned} \tag{4}$$

Bäuerle & Ott (2011) showed a remarkable fact about $\min_{\pi\in\Pi_{\mathcal{H}}} V_1^\pi(s_1, b_1)$: there exists an optimal policy $\rho^\star = \{\rho_h^\star : \mathcal{S}^{\mathrm{Aug}} \to \mathcal{A}\}_{h\in[H]}$ that is deterministic and Markov in an augmented MDP, which we now describe. The augmented state is $(s,b) \in \mathcal{S}^{\mathrm{Aug}} := \mathcal{S} \times [0,1]$. Given any $b_1 \in [0,1]$, the initial state is (s_1, b_1). Then, for each $h = 1, 2, ..., H$, $a_h = \rho_h^\star(s_h, b_h), r_h \sim R(s_h, a_h), s_{h+1} \sim P^\star(s_h, a_h), b_{h+1} = b_h - r_h$. Intuitively, the extra state b_h is the amount of budget left from the initial b_1, and is a sufficient statistic of the history for the CVaR RL problem. Let Π^{Aug} denote the set of deterministic, Markov policies in the augmented MDP. Then, we may optimize over this simpler policy class without losing optimality!

Before we formalize the optimality result, we first derive Bellman equations (as in Bellemare et al., 2023, Chapter

7.8). For any $\rho \in \Pi^{\mathrm{Aug}}$, overload notation and define

$$V_h^\rho(s_h, b_h) = \mathbb{E}_\rho\left[\left(b_h - \textstyle\sum_{t=h}^H r_t\right)^+ \mid s_h, b_h\right],$$

where $r_h, \dots, r_H$ are the rewards generated by executing ρ in the augmented MDP starting from s_h, b_h at time step h. Observe that V_h^ρ satisfies the Bellman equations,

$$V_h^\rho(s_h, b_h) = \mathbb{E}_{a_h \sim \rho_h(s_h, b_h)}[U_h^\rho(s_h, b_h, a_h)],$$
$$U_h^\rho(s_h, b_h, a_h) = \mathbb{E}_{s_{h+1}, r_h}\left[V_{h+1}^\rho(s_{h+1}, b_{h+1})\right],$$

where $s_{h+1} \sim P^\star(s_h, a_h), r_h \sim R(s_h, a_h), b_{h+1} = b_h - r_h$ and $V_{H+1}^\rho(s, b) = b^+$. Analogously, define $V_h^\star$ and $\rho^\star$ inductively with the Bellman optimality equations,

$$V_h^\star(s_h, b_h) = \min_{a \in \mathcal{A}} U_h^\star(s_h, b_h, a_h),$$
$$\rho_h^\star(s_h, b_h) = \arg\min_{a \in \mathcal{A}} U_h^\star(s_h, b_h, a_h),$$
$$U_h^\star(s_h, b_h, a_h) = \mathbb{E}_{s_{h+1}, r_h}\left[V_{h+1}^\star(s_{h+1}, b_{h+1})\right],$$

where $s_{h+1} \sim P^\star(s_h, a_h), r_h \sim R(s_h, a_h), b_{h+1} = b_h - r_h$ and $V_{H+1}^\star(s, b) = b^+$. Armed with these definitions, we formalize the optimality result in the following theorem.

Theorem 5.1 (Optimality of Π^{Aug})**.** *For any $b \in [0, 1]$,*

$$V_1^\star(s_1, b) = V_1^{\rho^\star}(s_1, b) = \inf_{\pi \in \Pi_{\mathcal{H}}} V_1^\pi(s_1, b).$$

This is a known result in the infinite-horizon, discounted setting (Bäuerle & Ott, 2011; Bellemare et al., 2023). We provide a proof from first principles for the finite-horizon setting in Appendix F, by inductively unravelling the Bellman optimality equations. As a technical remark, we show optimality over history-dependent policies in the augmented MDP with memory, larger than the history-dependent class defined here.

These facts imply that we could compute $V_1^\star$ and $\rho^\star$ using dynamic programming (DP) if we knew the true transitions $P^\star$, and the procedure is similar to the classic Value Iteration procedure in vanilla RL. Based on Theorem 5.1 and Eq. (4), by executing $\rho^\star$ starting from $(s_1, b^\star)$ with $b^\star := \arg\max_{b \in [0,1]}\{b - \tau^{-1} V_1^\star(s_1, b)\}$, we achieve the maximum CVaR value in the original MDP. Below we leverage this DP perspective on the augmented MDP to design exploration algorithms to solve CVaR RL.

5.2. CVaR-UCBVI

In this section, we introduce our algorithm CVaR-UCBVI (Algorithm 2), an extension of the classic UCBVI algorithm of Azar et al. (2017) which attained the minimax-optimal regret for vanilla RL. Our contribution is showing that bonus-driven pessimism, which guarantees that the learned $\widehat{V}_{h,k}^\downarrow$ is a high probability lower confidence bound (LCB) on the optimal $V_h^\star$, is sufficient and in fact optimal for CVaR RL. This remarkably shows that the bonus-driven exploration paradigm from vanilla RL, *i.e.*, "optimism/pessimism under uncertainty," can be used to optimally conduct safe exploration for CVaR RL.

CVaR-UCBVI iterates over K episodes, where the k-th episode proceeds as follows. First, in Line 3, we compute an empirical estimate $\widehat{P}_k$ of the transition dynamics $P^\star$ using the previous episodes' data. Then, in Line 6, we inductively compute $\widehat{U}_{h,k}^\downarrow$ from $h = H$ to $h = 1$ by subtracting a bonus that accounts for the error from using $\widehat{P}_k$ instead of $P^\star$. Next, $\widehat{V}_{h,k}^\downarrow$ and $\widehat{\rho}^k$ are computed greedily w.r.t. $\widehat{U}_{h,k}^\downarrow$ to mimic the Bellman optimality equations (Section 5.1). Subtracting the bonus is key to showing $\widehat{U}_{h,k}^\downarrow$ (resp. $\widehat{V}_{h,k}^\downarrow$) is a high probability lower bound of $U_h^\star$ (resp. $V_h^\star$). Next, in Line 9, we compute $\widehat{b}_k$ using the pessimistic $\widehat{V}_{1,k}^\downarrow$. Similar to our MAB algorithm, this guarantees that $\widehat{\mathrm{CVaR}}_\tau^k := \widehat{b}_k - \tau^{-1}\widehat{V}_{1,k}^\downarrow(s_1, \widehat{b}_k)$ is an *optimistic* estimate of $\mathrm{CVaR}_\tau^\star$. Finally, in Line 10, we roll in with the learned, augmented policy $\widehat{\rho}^k$ starting from $\widehat{b}_k$ in the augmented MDP to collect data for the next iterate. We highlight that in Line 10, the algorithm is still only interacting with the original MDP described in Section 2. To roll in with an augmented policy, the algorithm can imagine this augmented MDP by keeping track of the b_h via the update $b_{h+1} = b_h - r_h$. There is virtually no overhead as it is only a scalar with known transitions.

5.3. The Hoeffding Bonus

Two types of bonuses may be used in CVaR-UCBVI: Hoeffding (Eq. (5)) and Bernstein (Eq. (6)). We now show that a simple Hoeffding bonus, defined below, can already provide the best CVaR regret bounds in the current literature:

$$\mathrm{BON}_{h,k}^{\mathrm{HOEFF}}(s, a) = \sqrt{\frac{L}{N_k(s, a)}}, \tag{5}$$

where $L = \log(HSAK/\delta)$.

Theorem 5.2. *For any $\delta \in (0, 1)$, w.p. at least $1 - \delta$,* CVaR-UCBVI *with the Hoeffding bonus (Eq. (5)) enjoys*

$$\mathrm{Regret}_\tau^{\mathrm{RL}}(K) \le 4e\tau^{-1}\sqrt{SAHK}L + 10e\tau^{-1}S^2AHL^2.$$

Proof Sketch The first step is to establish pessimism, *i.e.*, $\widehat{V}_{1,k}^\downarrow \le V_1^\star$, which implies optimism of $\widehat{\mathrm{CVaR}}_\tau^k \ge \mathrm{CVaR}_\tau^\star$. At this point, we cannot apply CVaR concentration as we did for MAB, since $\widehat{V}_{1,k}^\downarrow$ is not the empirical CVaR. Instead, we show that the simulation lemma (Lemma G.4) extends to the augmented MDP, which gives $V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k) - \widehat{V}_{1,k}^\downarrow(s_1, \widehat{b}_k) \le$

Algorithm 2 CVaR-UCBVI

Input: risk tolerance τ, number of episodes K, failure probability δ, bonus function $\textsc{Bon}_{h,k}(s,b,a)$.
for episode $k = 1, 2, \dots, K$ **do**
  Compute counts and empirical transition estimate,

$$N_k(s,a,s') = \sum_{h=1}^{H}\sum_{i=1}^{k-1} \mathbb{I}\left[(s_{h,i}, a_{h,i}, s_{h+1,i}) = (s,a,s')\right],$$
$$N_k(s,a) = 1 \vee \sum_{s'\in\mathcal{S}} N_k(s,a,s'),\ \widehat{P}_k(s' \mid s,a) = \frac{N_k(s,a,s')}{N_k(s,a)},$$

  For all $s \in \mathcal{S}, b \in [0,1]$, set $\widehat{V}^{\uparrow}_{H+1,k}(s,b) = \widehat{V}^{\downarrow}_{H+1,k}(s,b) = b^+$.
  for $h = H, H-1, \dots, 1$ **do**
    Define pessimistic estimates of $V^\star$ and $\widehat{\rho}^k$, *i.e.*, for all s, b, a,

$$\widehat{U}^{\downarrow}_{h,k}(s,b,a) = \widehat{P}_k(s,a)^\top \mathbb{E}_{r_h\sim R(s,a)}\left[\widehat{V}^{\downarrow}_{h+1,k}(\cdot, b-r_h)\right] - \textsc{Bon}_{h,k}(s,b,a),$$
$$\widehat{\rho}^k_h(s,b) = \arg\min_a \widehat{U}^{\downarrow}_{h,k}(s,b,a), \qquad \widehat{V}^{\downarrow}_{h,k}(s,b) = \max\left\{\widehat{U}^{\downarrow}_{h,k}(s,b,\widehat{\rho}^k_h(s,b)), 0\right\}.$$

    If using Bernstein bonus (Section 5.4), also define optimistic estimates for $V^\star$, *i.e.*, for all s, b, a,

$$\widehat{U}^{\uparrow}_{h,k}(s,b,a) = \widehat{P}_k(s,a)^\top \mathbb{E}_{r_h\sim R(s,a)}\left[\widehat{V}^{\uparrow}_{h+1,k}(\cdot, b-r_h)\right] + \textsc{Bon}_{h,k}(s,b,a),$$
$$\widehat{V}^{\uparrow}_{h,k}(s,b) = \min\left\{\widehat{U}^{\uparrow}_{h,k}(s,b,\widehat{\rho}^k_h(s,b)), 1\right\}.$$

  end for
  Calculate $\widehat{b}_k = \arg\max_{b\in[0,1]}\left\{b - \tau^{-1}\widehat{V}^{\downarrow}_{1,k}(s_1,b)\right\}$.
  Collect $\{(s_{h,k}, a_{h,k}, r_{h,k})\}_{h\in[H]}$ by rolling in $\widehat{\rho}^k$ starting from $(s_1, \widehat{b}_k)$ in the augmented MDP.
end for

$\mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\sum_{h=1}^H 2\textsc{Bon}^{\textsc{Hoeff}}_{h,k}(s_h,a_h) + \xi_{h,k}(s_h,a_h)\right]$. The expectation is over the distribution of rolling in $\widehat{\rho}^k$ from $\widehat{b}_k$, which is exactly how we explore and collect $s_{h,k}, a_{h,k}$. Thus, we can apply Azuma and elliptical potential to conclude the proof, as in the usual UCBVI analysis. ■

The leading term of the Hoeffding bound is $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{SAHK})$, which is optimal in S, A, K. Notably, it has a S factor improvement over the current best bound $\tau^{-1}\sqrt{S^3AHKL}$ from Bastani et al. (2022) (we've divided their bound by H to make returns scaling consistent). While Theorem 5.2 is already the tightest in the literature, our lower bound suggests the possibility of removing another $\sqrt{\tau^{-1}H}$.

5.4. Improved Bounds with the Bernstein Bonus

Precise design of the exploration bonus is critical to enabling tighter performance bounds, even in vanilla RL (Azar et al., 2017; Zanette & Brunskill, 2019). In this subsection, we propose the Bernstein bonus and prove two tighter regret bounds. The bonus depends on the sample variance, which recall, for any function f, is defined as $\mathrm{Var}_{s'\sim\widehat{P}_k(s,a)}(f(s')) = \widehat{P}_k(s,a)^\top\left(f(\cdot) - \bar{f}_N\right)^2$ with $\bar{f}_N = \widehat{P}_k(s,a)^\top f$ being the sample mean (Maurer & Pontil, 2009). We define the Bernstein bonus as follows,

$$\textsc{Bon}^{\textsc{Bern}}_{h,k}(s,b,a) = \sqrt{\frac{2\,\mathrm{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[\widehat{V}^{\downarrow}_{h+1,k}(s',b')\right]\right)L}{N_k(s,a)}}$$
$$+\sqrt{\frac{2\mathbb{E}_{s'\sim\widehat{P}_k(s,a),r_h}\left[\left(\widehat{V}^{\uparrow}_{h+1,k}(s',b') - \widehat{V}^{\downarrow}_{h+1,k}(s',b')\right)^2\right]L}{N_k(s,a)}}$$
$$+\frac{L}{N_k(s,a)}, \text{ where } b' = b - r_h, \text{ and } r_h \sim R(s,a). \tag{6}$$

When using the Bernstein bonus, Line 7 should be activated to compute optimistic estimates $\widehat{V}^{\uparrow}_{h,k}$ by adding the bonus. Together, $(\widehat{V}^{\downarrow}_{h,k}, \widehat{V}^{\uparrow}_{h,k})$ forms a tight confidence band around $V^\star_h$, which we will use to inductively prove pessimism and optimism at all $h \in [H]$ (as in Zanette & Brunskill, 2019). Compared to Zanette & Brunskill (2019), our Bernstein bonus also depends on the state augmentation b, since our UCBVI procedure is running in the

augmented MDP. We now prove the first Bernstein bound, which tightens the Hoeffding bound by a $\sqrt{H}$ factor.

Theorem 5.3. *For any $\delta \in (0,1)$, w.p. at least $1-\delta$,* CVaR-UCBVI *with the Bernstein bonus (Eq. (6)) enjoys a regret guarantee of*

$$\mathrm{Regret}_\tau^{\mathrm{RL}}(K) \le 10e\tau^{-1}\sqrt{SAKL} + \tau^{-1}\xi,$$

where $\xi \in \widetilde{\mathcal{O}}(SAHK^{1/4} + S^2AH)$ is a lower order term.

Proof Sketch We first establish pessimism and optimism of $\widehat{V}^{\downarrow}_{h,k}$ and $\widehat{V}^{\uparrow}_{h,k}$, similar to Zanette & Brunskill (2019). Then, apply Simulation lemma as in Theorem 5.2. The key observation is that the H sample variances from the bonus, *i.e.*, $\sum_{h=1}^H \mathrm{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}[\widehat{V}^{\downarrow}_{h+1,k}(s', b-r_h)]\right)$, can be reduced to a *single* variance $\mathrm{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left((\widehat{b}_k - \sum_{h=1}^H r_h)^+\right)$, which we bound by 1. To do so, we show that Azar et al. (2017)'s Law of Total Variance technique also applies to our $V^{\widehat{\rho}^k}$, which, unlike the value function of vanilla RL, depends on the state augmentation b. ■

The leading term of Theorem 5.3 is $\tau^{-1}\sqrt{SAK}$, and improves over Bastani et al. (2022) by a $S\sqrt{H}$ factor. Up to log terms, this matches our lower bound in all parameters except τ, which implies CVaR-UCBVI is minimax-optimal for a constant τ. In particular, $\tau = 1$ recovers the risk-neutral vanilla RL setting, where CVaR-UCBVI matches the minimax result (Azar et al., 2017). To get the optimal $\tau^{-1/2}$ scaling (Corollary 3.2), we cannot loosely bound each variance term by 1, as they should scale as τ if $\widehat{b}_k$ approximates the τ-th quantile of $R(\widehat{\rho}^k, \widehat{b}_k)$. We show this is indeed the case under a continuity assumption.

Assumption 5.4. For all $\rho \in \Pi^{\mathrm{Aug}}$ and $b_1 \in [0,1]$, the returns of rolling in ρ from b_1, *i.e.*, $R(\rho, b_1)$, is continuously distributed with a density lower bounded by $p_{\min}$.

Theorem 5.5. *Under Assumption 5.4, the bound in Theorem 5.3 can be refined to,*

$$\mathrm{Regret}_\tau^{\mathrm{RL}}(K) \le 12e\sqrt{\tau^{-1}SAKL} + \tau^{-1}p_{\min}^{-1/2}\xi.$$

Proof Sketch The only divergence from Theorem 5.3 is how we bound $\mathrm{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left((\widehat{b}_k - \sum_{h=1}^H r_h)^+\right)$. Since the density of $R(\widehat{\rho}^k, \widehat{b}_k)$ is lower bounded, the CVaR objective $f(b) = b - \tau^{-1}\mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[(b - \sum_{h=1}^H r_h)^+\right]$ is *strongly concave*. This implies that $\widehat{b}_k$ approximates the true τ-th quantile $b_k^\star = \arg\max_{b\in[0,1]} f(b)$, *i.e.*, $\frac{p_{\min}}{2}(\widehat{b}_k - b_k^\star)^2 \le \tau(f(b_k^\star) - f(\widehat{b}_k)) \le V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k) - \widehat{V}^{\downarrow}_{1,k}(s_1, \widehat{b}_k)$. Leveraging this fact, we show $\mathrm{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left((\widehat{b}_k - \sum_{h=1}^H r_h)^+\right) \le 2\tau + 4p_{\min}^{-1}(V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k) - \widehat{V}^{\downarrow}_{1,k}(s_1,\widehat{b}_k))$, which notably scales with τ. We conclude the proof by showing the error term, *i.e.*, $\sum_{k=1}^K (V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k) - \widehat{V}^{\downarrow}_{1,k}(s_1,\widehat{b}_k)) \in \widetilde{\mathcal{O}}(\sqrt{SAHK})$, is lower order. ■

The leading term of Theorem 5.5 is $\sqrt{\tau^{-1}SAK}$, which matches our lower bound in Corollary 3.2 and establishes the *optimality* of CVaR-UCBVI for return distributions satisfying Assumption 5.4. Notably, $p_{\min}$ only multiplies with the lower order term ξ. This result highlights the importance of the Bernstein bonus for CVaR RL – it improves the regret bound of the Hoeffding bonus by $\sqrt{\tau^{-1}H}$, whereas in vanilla RL, the improvement is $\sqrt{H}$.

With regards to Assumption 5.4, lower bounded densities, *i.e.*, strong monotonicity of the CDF, is standard for identifying the quantile (Ma et al., 2021). In fact, PAC results for quantile identification is not possible without some control of the mass at the quantile. As a simple example, consider estimating the 0.5-th quantile using N i.i.d. datapoints sampled from $\mathrm{Ber}(0.5)$. The correct answer is 0, but by symmetry, the sample median is always distributed as $\mathrm{Ber}(0.5)$ for any N. So we always have a 0.5-probability of being incorrect. We provide an information theoretic lower bound to rule out all estimators – not just the sample median – in Theorem I.1.

It nonetheless remains an open question whether Assumption 5.4 can be removed by eschewing identifying the quantile. In MABs Theorem 4.1, we circumvented the need to identify quantiles by decomposing the regret into (1) the sum of bonuses, plus, (2) the difference between the empirical and true CVaRs, both of which can be shown to have the correct $\tau^{-1/2}$ scaling. An analogous approach for RL is to decompose the regret into (1) $\sum_{h,k} \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k,\widehat{P}_k}\left[\mathrm{BON}^{\mathrm{BERN}}_{h,k}(s_{h,k}, b_{h,k}, a_{h,k})\right]$, plus, (2) $\sum_k \mathrm{CVaR}_\tau(\widehat{\rho}^k, \widehat{b}_k; \widehat{P}_k) - \mathrm{CVaR}_\tau(\widehat{\rho}^k, \widehat{b}_k)$. However, it is unclear if both terms can be unconditionally bounded by $\widetilde{\mathcal{O}}(\sqrt{\tau^{-1}SAK})$.

Remark on b-dependence: Although CVaR-UCBVI operates in the augmented MDP, our Hoeffding bonus has no dependence on the budget state b and matches the Hoeffding bonus of UCBVI (from vanilla RL; Azar et al., 2017). Intuitively, this is possible since the dynamics of b are known (we assume known reward distribution), so there is no need to explore in the b-dimension. In contrast to the Bernstein bonus of UCBVI, our Bernstein bonus depends on b and captures the variance of $\left(b - R(\widehat{\rho}^k, \widehat{b}_k)\right)^+$. This is crucial for obtaining the correct τ rate.

6. Computational Efficiency via Discretization

Previously, we assumed each line of Algorithm 2 was computed exactly. This is not computationally feasible since the dynamic programming (DP) step (Lines 6 and 7) needs to be done over all $b \in [0,1]$ and the calculation for $\widehat{b}_k$

(Line 9) involves maximizing a non-concave function. Following Bastani et al. (2022), we propose to discretize the rewards so the aforementioned steps need only be performed over a finite grid. Thus, we gain computational efficiency while maintaining the same statistical guarantees.

Fix a precision $\eta \in (0,1)$, define $\phi(r) = \eta\lceil r/\eta \rceil \wedge 1$, which *rounds-up* $r \in [0,1]$ to an η-net of $[0,1]$, henceforth referred to as "the grid". The discretized MDP $\mathrm{disc}(\mathcal{M})$ is an exact replica of the true MDP $\mathcal{M}$ with one exception: its rewards are post-processed with ϕ, *i.e.*, $R(s,a;\mathrm{disc}(\mathcal{M})) = R(s,a;\mathcal{M}) \circ \phi^{-1}$, where $\circ$ denotes composition.

In $\mathrm{disc}(\mathcal{M})$, the τ-th quantile of the returns distribution (the argmax of the CVaR objective) will be a multiple of η, so it suffices to compute $\widehat{V}_1^{\downarrow}(s_1, b)$ and maximize Line 9 over the grid. Since b transitions by subtracting rewards, which are multiples of η, b_h will always stay on the grid. Hence, the entire DP procedure (Lines 6 and 7) only needs to occur on the grid. In Appendix H.3, we show CVaR-UCBVI has a runtime of $\mathcal{O}(S^2\eta^{-2}AHK)$ in the discretized MDP. It's worth clarifying that CVaR-UCBVI is still interacting with $\mathcal{M}$, except that it internally discretizes the received rewards to simulate running in the $\mathrm{disc}(\mathcal{M})$ for computation's sake. Thus, we still want to compete with the strongest CVaR policy in the true MDP; we've just made our algorithm weaker by restricting it to run in an imagined $\mathrm{disc}(\mathcal{M})$.

Now, we show that the true regret only increases by $\mathcal{O}(K\eta)$, which can be made lower order by setting $\eta = K^{-1/2}$. Theorems 5.2 and 5.3 made no assumptions on the reward distribution, so they immediately apply to bound the $\mathrm{disc}(\mathcal{M})$ regret, *i.e.*, $\mathrm{Regret}_\tau^{\mathrm{RL}}(K;\mathrm{disc}(\mathcal{M})) = \sum_{k=1}^K \mathrm{CVaR}_\tau^\star(\mathrm{disc}(\mathcal{M})) - \mathrm{CVaR}_\tau(\widehat{\rho}^k, \widehat{b}_k; \mathrm{disc}(\mathcal{M}))$. We translate regret in $\mathrm{disc}(\mathcal{M})$ to regret in $\mathcal{M}$ via a coupling argument, inspired by Bastani et al. (2022). Let $Z_{\pi,\mathcal{M}}$ denote the returns from running π in $\mathcal{M}$. For random variables X, Y, we say Y stochastically dominates X, denoted $X \preceq Y$, if $\forall t \in \mathbb{R} : \Pr(Y \le t) \le \Pr(X \le t)$. Then, for any $\rho \in \Pi^{\mathrm{Aug}}, b_1 \in [0,1]$, we show two facts:

F1 Running ρ, b_1 in the imagined $\mathrm{disc}(\mathcal{M})$ is equivalent to running a reward-history-dependent policy, $\mathrm{adapted}(\rho, b_1)_h(s_h, r_{1:h-1}) = \rho_h(s_h, b_1 - \phi(r_1) - \ldots - \phi(r_{h-1}))$. Also, $Z_{\rho,b_1,\mathrm{disc}(\mathcal{M})} - H\eta \preceq Z_{\mathrm{adapted}(\rho,b_1),\mathcal{M}} \preceq Z_{\rho,b_1,\mathrm{disc}(\mathcal{M})}$.

F2 There exists a memory-history-dependent[2] policy $\mathrm{disc}(\rho, b)$ such that $Z_{\rho,b,\mathcal{M}} \preceq Z_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}$. Intuitively, when running in $\mathrm{disc}(\mathcal{M})$, once the discretized reward r_h is seen, a memory m_h is generated from the conditional reward distribution of rewards that get rounded-up to r_h. Thus, m_h is essentially sampled from the unconditional reward distribution. The memory-dependent policy $\mathrm{disc}(\rho, b)$ makes use of these samples to mimic running ρ, b in $\mathcal{M}$.

[2]The memory-MDP model is novel and key to our coupling argument. In Appendix F, we define this model and show that Π^{Aug} still contains the optimal policy over this seemingly larger class of memory-history-dependent policies, *i.e.*, Theorem 5.1 holds.

F1 implies $\mathrm{CVaR}_\tau(\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k); \mathcal{M}) \ge \mathrm{CVaR}_\tau(\widehat{\rho}^k, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) - \tau^{-1}H\eta$. **F2** implies $\mathrm{CVaR}_\tau^\star(\mathcal{M}) \le \mathrm{CVaR}_\tau^\star(\mathrm{disc}(\mathcal{M}))$. Combining these two facts, we have $\mathrm{Regret}_\tau^{\mathrm{RL}}(K;\mathcal{M}) \le \mathrm{Regret}_\tau^{\mathrm{RL}}(K;\mathrm{disc}(\mathcal{M})) + K\tau^{-1}H\eta$.

Translating Theorem 5.5 requires more care, as its proof relied on continuously distributed returns (Assumption 5.4) which is untrue in $\mathrm{disc}(\mathcal{M})$. We show that we only need the *true* returns distribution to be continuous.

Assumption 6.1. For all $\rho \in \Pi^{\mathrm{Aug}}$ and $b_1 \in [0,1]$, the returns distribution of $\mathrm{adapted}(\rho, b_1)$ in $\mathcal{M}$ is continuous, with a density lower bounded by $p_{\min}$.

With this premise, we can prove Theorem 5.5 for $\mathrm{disc}(\mathcal{M})$, with an extra term of $\widetilde{\mathcal{O}}(\tau^{-1}\sqrt{p_{\min}^{-1}SAHK\eta})$. In sum, setting $\eta = 1/\sqrt{K}$ ensures that CVaR-UCBVI is *both* near-minimax-optimal for regret and computationally efficient, with a runtime of $\mathcal{O}(S^2AHK^2)$. We note the superlinear-in-K runtime from discretization is not even avoidable in Lipschitz bandits (Wang et al., 2020), and we leave developing more scalable methods for future work.

7. Concluding Remarks

In this paper, we presented a more complete picture of risk-sensitive MAB and RL with CVaR by providing not only novel lower bounds but also procedures and analyses that both improve on the state of the art and match our lower bounds. One exception where a gap remains is CVaR RL with discontinuous returns and a risk tolerance that is not constant (or, not lower bounded); in this case, our lower and upper bounds differ by a factor of $\sqrt{\tau}$. We discuss the feasibility of closing this gap in Section 5.4.

A direction for future work is to develop algorithms with optimal regret guarantees for more general risk measures, *e.g.*, optimized certainty equivalent (OCE) (Ben-Tal & Teboulle, 2007). Another orthogonal direction is to extend our results beyond tabular MDPs. We believe that our techniques in this work are already enough for linear MDPs (Jin et al., 2020) where the transition kernel is linear in some *known* feature space. However, extending the results beyond linear models, such as to low-rank MDPs (Agarwal et al., 2020; Uehara et al., 2022) and block MDPs (Misra et al., 2020; Zhang et al., 2022) remains a challenge due to the fact that achieving point-wise optimism is harder when nonlinear function approximation is used.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-1846210 and IIS-2154711.

References

Acerbi, C. and Tasche, D. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. *Reinforcement learning: Theory and algorithms*. 2021. https://rltheorybook.github.io/.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Bastani, O., Ma, Y. J., Shen, E., and Xu, W. Regret bounds for risk-sensitive reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yJEUDfzsTX7.

Bäuerle, N. and Ott, J. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Bellemare, M. G., Dabney, W., and Rowland, M. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Ben-Tal, A. and Teboulle, M. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Bibaut, A., Kallus, N., Dimakopoulou, M., Chambaz, A., and van der Laan, M. Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. *Advances in Neural Information Processing Systems*, 34:19261–19273, 2021.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brown, D. B. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35 (6):722–730, 2007.

Chen, H. Chapter 2. order statistics. http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/notes/noteorder.pdf.

Chow, Y. and Ghavamzadeh, M. Algorithms for cvar optimization in mdps. *Advances in neural information processing systems*, 27, 2014.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

Du, Y., Wang, S., and Huang, L. Risk-sensitive reinforcement learning: Iterated cvar and the worst path. *arXiv preprint arXiv:2206.02678*, 2022.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446, 2021.

Filippi, C., Guastaroba, G., and Speranza, M. G. Conditional value-at-risk beyond finance: a survey. *International Transactions in Operational Research*, 27(3): 1277–1319, 2020.

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pp. 3395–3398. PMLR, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Kagrecha, A., Nair, J., and Jagannathan, K. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. *arXiv preprint arXiv:1906.00569*, 2019.

Kallus, N., Mao, X., Wang, K., and Zhou, Z. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, 2022.

Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4436–4443, 2020.

Kiefer, J. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3): 502–506, 1953.

Kisiala, J. Conditional value-at-risk: Theory and applications. *arXiv preprint arXiv:1511.00140*, 2015.

Lam, T., Verma, A., Low, B. K. H., and Jaillet, P. Risk-aware reinforcement learning with coherent risk measures and non-linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-RwZOVybbj.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Liang, H. and Luo, Z.-Q. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *arXiv preprint arXiv:2210.14051*, 2022.

Ma, Y., Jayaraman, D., and Bastani, O. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19235–19247, 2021.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. In *Advances in neural information processing systems*, 2022.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rajeswaran, A., Ghotra, S., Ravindran, B., and Levine, S. EPOpt: Learning robust neural network policies using model ensembles. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SyWvgP5el.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributional robust batch contextual bandits. *arXiv preprint arXiv:2006.05630*, 2020.

Singla, A., Rafferty, A. N., Radanovic, G., and Heffernan, N. T. Reinforcement learning for education: Opportunities and challenges. *arXiv preprint arXiv:2107.08828*, 2021.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Tamkin, A., Keramati, R., Dann, C., and Brunskill, E. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline RL in low-rank MDPs. In *ICLR*, 2022. URL https://openreview.net/forum?id=J4iSIR9fhY0.

Urpí, N. A., Curi, S., and Krause, A. Risk-averse offline reinforcement learning. 2021.

Van de Geer, S. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Wang, T., Ye, W., Geng, D., and Rudin, C. Towards practical lipschitz bandits. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pp. 129–138, 2020.

Wang, Y. and Gao, F. Deviation inequalities for an estimator of the conditional value-at-risk. *Operations Research Letters*, 38(3):236–239, 2010.

Wiesemann, W., Kuhn, D., and Rustem, B. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zhang, T. *Mathematical Analysis of Machine Learning Algorithms*. 2023. http://www.tongzhang-ml.org/lt-book.html.

Zhang, X., Song, Y., Uehara, M., Wang, M., Agarwal, A., and Sun, W. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pp. 26517–26547. PMLR, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021b.

Appendices

A. Notations

Table 1. List of Notations

x^+	$\max(x, 0)$, *i.e.*, the ReLU function.
$\mathcal{S}, \mathcal{A}, S, A$	State and action spaces, with sizes $S = \lvert\mathcal{S}\rvert$ and $A = \lvert\mathcal{A}\rvert$. In MAB, $S = 1$.
$\mathcal{S}^{\text{Aug}}$	Augmented state space $\mathcal{S} \times [0, 1]$ for CVaR RL.
$H \in \mathbb{N}$	Horizon of the RL problem. In MAB, $H = 1$.
$K \in \mathbb{N}$	Number of episodes.
$\delta \in (0, 1)$	Failure probability.
L	$\log(SAHK/\delta)$.
$\Delta(\mathcal{S})$	The set of distributions supported by $\mathcal{S}$.
$R(a) \in \Delta([0, 1])$	Reward distribution of arm a (for MAB).
$P^\star(s, a) \in \Delta(\mathcal{S})$	Ground truth transition kernel (for RL).
$R(s, a) \in \Delta([0, 1])$	Known reward distribution (for RL).
$R(\pi)$	Returns distribution of history-dependent policy π (for RL).
$R(\rho, b)$	Returns distribution of augmented policy $\rho \in \Pi^{\text{Aug}}$ starting from b (for RL).
$F^\dagger(t)$ for $t \in [0, 1]$	The t-th quantile function of X with CDF F, *i.e.*, $\inf\{x : F(x) \geq t\}$.
$\mathcal{I}_{h,k}(s, a)$	Indices of prior visits of s, a at h, *i.e.*, $\{i \in [k-1] : (s_{h,i}, a_{h,i}) = (s, a)\}$.
$N_{h,k}(s, a)$	Number of prior visits of s, a at h, *i.e.*, $\lvert\mathcal{I}_{h,k}(s, a)\rvert$.
$\xi_{h,k}(s, a)$	$\min\left\{2, \frac{2HSL}{N_{h,k}(s,a)}\right\}$.
$\mathcal{E}_k$	Trajectories from episodes $1, 2, ..., k-1$.
$\mathcal{H}_h (\mathcal{H}_{h,k})$	History up to and not including time h (in episode k).
$\Pi_{\mathcal{H}}$	Set of history-dependent policies.
Π^{Aug}	Set of Markov, deterministic policies in the augmented MDP.
(ρ, b)	The policy obtained from rolling in ρ starting from (s_1, b) in the augmented MDP.
$\text{disc}(\mathcal{M})$	The discretized MDP obtained by discretizing rewards, Section 6.
$\text{adapted}(\rho, b_1)$	The policy from adapting (ρ, b_1) in $\text{disc}(\mathcal{M})$ to $\mathcal{M}$.
$\text{disc}(\rho, b_1)$	The policy from discretizing (ρ, b_1) in $\mathcal{M}$ to $\text{disc}(\mathcal{M})$.

B. Concentration Lemmas

B.1. Uniform Hoeffding and Bernstein via Lipschitzness

Recall the classic Hoeffding and Bernstein inequalities (Theorems 2.8 and 2.10 in Boucheron et al., 2013). Let $X_{1:N}$ be i.i.d. random variables in $[0, 1]$, with mean μ and variance σ^2. Then, for any δ, w.p. at least $1 - \delta$, we have

$$\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \leq \frac{1}{2}\sqrt{\frac{\log(4/\delta)}{N}}, \tag{Hoeffding}$$

$$\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \leq \sqrt{\frac{2\sigma^2 \log(4/\delta)}{N}} + \frac{\log(4/\delta)}{N}. \tag{Bernstein}$$

Now we consider uniform inequalities for a function class. Specifically, let $X_{1:N}$ be i.i.d. copies of $X \in \mathcal{X}$ and $\mathcal{F}$ is a (potentially infinite) set of functions $f : \mathcal{X} \to [0, 1]$. Suppose $\mathcal{G}_\varepsilon \subset \mathcal{F}$ is a finite ℓ_∞-cover, a.k.a. ε-net, of $\mathcal{F}$ in the sense that: for any $f \in \mathcal{F}$, there exists $g \in \mathcal{G}_\varepsilon$ such that $\sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \varepsilon$.

Lemma B.1. *Let $\delta \in (0,1)$. We have w.p. at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f(X)\right| \le \sqrt{\frac{\log(4|\mathcal{G}_{1/N}|/\delta)}{N}}. \tag{Uniform Hoeffding}$$

If $N \ge 2\log(4|\mathcal{G}_{1/N}|/\delta)$, we also have

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f(X)\right| \le \sqrt{\frac{2\operatorname{Var}(f(X))\log(4|\mathcal{G}_{1/N}|/\delta)}{N}} + \frac{3\log(4|\mathcal{G}_{1/N}|/\delta)}{N}. \tag{Uniform Bernstein}$$

Proof. Apply a union bound over the elements of $\mathcal{G}_\varepsilon$. Then for any $f \in \mathcal{F}$,

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f(X)\right| &\le 2\varepsilon + \left|\frac{1}{N}\sum_{i=1}^{N} g(X_i) - \mathbb{E}g(X)\right| \\ &\le 2\varepsilon + \frac{1}{2}\sqrt{\frac{\log(4|\mathcal{G}_\varepsilon|/\delta)}{N}}.\end{aligned}$$

Setting $\varepsilon = 1/N$ gives the Uniform Hoeffding result. We also have

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f(X)\right| &\le 2\varepsilon + \sqrt{\frac{2\operatorname{Var}(g(X))\log(4|\mathcal{G}_\varepsilon|/\delta)}{N}} + \frac{\log(4|\mathcal{G}_\varepsilon|/\delta)}{N} \\ &\le 2\varepsilon + \sqrt{\frac{2\operatorname{Var}(f(X))\log(4|\mathcal{G}_\varepsilon|/\delta)}{N}} + \frac{\log(4|\mathcal{G}_\varepsilon|/\delta)}{N} + \varepsilon\sqrt{\frac{2\log(4|\mathcal{G}_\varepsilon|/\delta)}{N}},\end{aligned}$$

since $\sqrt{\operatorname{Var}(g(X))} - \sqrt{\operatorname{Var}(f(X))} \le \sqrt{\operatorname{Var}(f(X)-g(X))} \le \varepsilon$. By assumption, $\sqrt{\frac{2\log(4|\mathcal{G}_\varepsilon|/\delta)}{N}} \le 1$, so the total error is at most 3ε. Thus, setting $\varepsilon = 1/N$ gives the Uniform Bernstein result. □

A particularly important application of this for us is that $\mathcal{F}$ will be an finite set of functions f_b parameterized by a continuous parameter $b \in [0,1]$. These functions are C-Lipschitz in the b parameter, so to construct $\mathcal{G}_\varepsilon$, it suffices to take a grid over $[0,1]$ such that any element is ε/C close to the grid. This grid requires $\lceil C/\varepsilon \rceil$ atoms, and so $\log(|\mathcal{G}_{1/N}|) \le \log(CN)$.

Empirical Bernstein: By Theorems 4 and 6 of Maurer & Pontil (2009), we also have an empirical version of the uniform Bernstein, where we may replace $\operatorname{Var}(f(X))$ with $\frac{1}{N(N-1)}\sum_{i,j=1}^{N}(f(X_i)-f(X_j))^2$, *i.e.*, the empirical variance. Another useful result of Maurer & Pontil (2009) is their Theorem 10, which proves a fast convergence of empirical variance to the true variance: w.p. $1-\delta$,

$$\left|\widehat{\operatorname{Var}} f(X) - \operatorname{Var} f(X)\right| \le \sqrt{\frac{2\log(2/\delta)}{N-1}},$$

where $\widehat{\operatorname{Var}} f(X) = \frac{1}{N(N-1)}\sum_{i,j=1}^{N}(f(X_i)-f(X_j))^2$ is the sample variance of N datapoints. Note that $\widehat{\operatorname{Var}} f(X)$ is the variance under the empirical distribution of these N datapoints, and hence behaves like a variance. Since $\sqrt{\operatorname{Var}(X+Y)} \le \sqrt{\operatorname{Var}(X)} + \sqrt{\operatorname{Var}(Y)}$ by Cauchy-Schwartz, this can also be extended to be uniform by the above argument, *i.e.*,

$$\sup_{f\in\mathcal{F}}\left|\sqrt{\widehat{\operatorname{Var}} f(X)} - \sqrt{\operatorname{Var} f(X)}\right| \le 2\sqrt{\frac{\log(2|\mathcal{G}_{1/N}|/\delta)}{N-1}}.$$

Proof. For any f, let g be its neighbor in the net. Using the triangle inequality of variance, $\sqrt{\operatorname{Var}(X+Y)} \le \sqrt{\operatorname{Var}(X)} + \sqrt{\operatorname{Var}(Y)}$, we have

$$\begin{aligned}\left|\sqrt{\widehat{\operatorname{Var}} f(X)} - \sqrt{\operatorname{Var} f(X)}\right| &\le \left|\sqrt{\widehat{\operatorname{Var}} g(X)} - \sqrt{\operatorname{Var} g(X)}\right| + \sqrt{\widehat{\operatorname{Var}}((f-g)(X))} + \sqrt{\operatorname{Var}((f-g)(X))} \\ &\le 2\sqrt{\frac{\log(2|\mathcal{G}_{1/N}|/\delta)}{N-1}} + 2\varepsilon.\end{aligned}$$

Setting $\varepsilon = 1/N$ completes the proof. □

B.2. Tape Method for Tabular MAB and RL

In this section, we describe how we are able to prove claims about MAB and RL using uniform concentration inequalities over i.i.d. data, *i.e.*, without needing to use complicated uniform martingale inequalities, *e.g.*, Bibaut et al. (2021); Van de Geer (2000). We construct a probability space using a "tape" method inspired by Slivkins et al. (2019, Section 1.3.1). Compared to using black-box uniform martingale inequalities, our approach is potentially loose in log terms. However, our approach is much cleaner as we only need uniform concentrations for i.i.d. data. Thus, we prove everything from first principles, so that concentration inequalities do not distract from the main ideas.

First, consider the MAB problem. Before the protocol starts, nature constructs an (one-indexed) array with AK cells. For each $a \in \mathcal{A}, k \in [K]$, nature fills the index $[a, k]$ with an independent sample from $R(a)$. Whenever the learner pulls arm a on the k-th episode, it receives the contents of $(a, N_k(a) + 1)$ where that $N_k(a)$ is the number of times that a has been pulled up until now.

Notice that this formulation will never run out of rewards since we've seeded each arm with K cells. Also, this is equivalent to drawing a sample whenever the learner pulls arm a. Crucially, all the rewards are independent and so we can obtain concentration inequalities for $N_k(a)$ for any learner, even before it is executed.

In particular, for any function $f : [0, 1] \to [0, 1]$ of the rewards, we can union bound Hoeffding/Bernstein over the cells $[a, 1 : k]$ for all a, k to get: for any δ, w.p. at least $1 - \delta$, we have for all a, k,

$$\left|\frac{1}{N_k(a)} \sum_{i \in \mathcal{I}_k(a)} f(r_i) - \mathbb{E} f(R(a))\right| \le \frac{1}{2}\sqrt{\frac{\log(4AK/\delta)}{N_k(a)}},$$
$$\left|\frac{1}{N_k(a)} \sum_{i \in \mathcal{I}_k(a)} f(r_i) - \mathbb{E} f(R(a))\right| \le \sqrt{\frac{2 \operatorname{Var}(f(R(a))) \log(4AK/\delta))}{N_k(a)}} + \frac{\log(4AK/\delta)}{N_k(a)}.$$

Here $\mathcal{I}_k(a)$ is the indices where the learner has pulled arm a.

We now do something similar for tabular RL. Before the RL algorithm starts, nature constructs an (one-indexed) array with $SAHK$ cells. For each $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]$, nature fills the index $[s, a, h, k]$ with an independent sample from $P^\star(s, a)$. Whenever, the learner takes action a at state s and step h on episode k, it receives the next state via the content of $(s, a, h, N_{h,k}(s, a) + 1)$ where recall $N_{h,k}(s, a)$ is the number of times the learner has taken action a at state s and step h before the current episode. Then, for any function $f : \mathcal{S} \to [0, 1]$ of the states, we can union bound Hoeffding/Bernstein over the cells $[s, a, 1 : H, 1 : k]$ for all s, a, k to get: for any δ, w.p. at least $1 - \delta$, we have for all s, a, h, k,

$$\left|\frac{1}{N_k(s,a)} \sum_{h,i \in \mathcal{I}_k(s,a)} f(s_{h+1,i}) - \mathbb{E}_{s' \sim P^\star(s,a)} f(s')\right| \le \frac{1}{2}\sqrt{\frac{\log(4SAHK/\delta)}{N_k(s,a)}},$$
$$\left|\frac{1}{N_k(s,a)} \sum_{h,i \in \mathcal{I}_k(s,a)} f(s_{h+1,i}) - \mathbb{E}_{s' \sim P^\star(s,a)} f(s')\right| \le \sqrt{\frac{2 \operatorname{Var}_{s' \sim P^\star(s,a)}(f(s')) \log(4SAHK/\delta))}{N_k(s,a)}} + \frac{\log(4SAHK/\delta)}{N_k(s,a)},$$

where $\mathcal{I}_k(s, a)$ are the (h, i) pairs where the learner has visited (s, a), and $N_k(s, a)$ is the size of $\mathcal{I}_k(s, a)$.

Since these are standard Hoeffding/Bernstein results over i.i.d. data, the uniform concentration results from the previous section applies.

C. Concentration of CVaR

In this section, we derive general concentration results for the empirical CVaR, which may be of independent interest. The significance of our result is that it applies to any bounded random variable X, which may be continuous, discrete or neither. Prior concentration results from Brown (2007) require X to be continuous. Some later works (Wang & Gao, 2010; Kagrecha et al., 2019) did not explicitly mention their dependence on the continuity of X, but their proof appears to require it as well and is complicated by casework. We provide a simple new proof of this concentration based on the Acerbi integral formula for CVaR, Lemma C.3.

For any random variable X in $[0,1]$ with CDF F, the quantile function is defined as,

$$F^\dagger(t) = \inf\{x \in [0,1] : F(x) \geq t\} = \sup\{x \in [0,1] : F(x) < t\}.$$

The quantile has many useful properties (Chen, Lemma 1), which we recall here.

Lemma C.1. *For $t \in (0,1)$, $F^\dagger(t)$ is nondecreasing and left-continuous, and satisfies*

1. *For all $x \in \mathbb{R}$, $F^\dagger(F(x)) \leq x$.*
2. *For all $t \in (0,1)$, $F(F^\dagger(t)) \geq t$.*
3. *$F(x) \geq t \iff x \geq F^\dagger(t)$.*

The quantile is always a maximizer of the CVaR objective in Eq. (1). This is true for any random variable, discrete, continuous or neither.

Lemma C.2. *For any random variable X in $[0,1]$ with CDF F, we have*

$$F^\dagger(\tau) \in \arg\max_{b \in [0,1]} \left\{ b - \tau^{-1}\mathbb{E}\left[(b-X)^+\right] \right\}.$$

Proof. Recall the objective in Eq. (1), $f(b) = -b + \tau^{-1}\mathbb{E}[(b-X)^+]$. It has a subgradient of

$$\partial f(b) = -1 + \tau^{-1}(\Pr(X < b) + [0,1]\Pr(X = b)).$$

We want to show that $0 \in \partial f(F^\dagger(\tau))$, which is equivalent to showing

$$0 \overset{(a)}{\leq} \tau - \Pr(X < F^\dagger(\tau)) \overset{(b)}{\leq} \Pr(X = F^\dagger(\tau)).$$

For (b), observe that $\Pr(X < F^\dagger(\tau)) + \Pr(X = F^\dagger(\tau)) = F(F^\dagger(\tau)) \geq \tau$ (Lemma C.1). Hence, $\tau - \Pr(X < F^\dagger(\tau)) \leq \Pr(X = F^\dagger(\tau))$.

For (a), recall that $\Pr(X < F^\dagger(\tau)) = \lim_{n\to\infty} \Pr(X \leq F^\dagger(\tau) - n^{-1})$, since $\left\{X \leq F^\dagger(\tau) - 1\right\} \subset \left\{X \leq F^\dagger(\tau) - 1/2\right\} \subset \cdots \subset \bigcup_{n\in\mathbb{N}} \left\{X \leq F^\dagger(\tau) - n^{-1}\right\} = \left\{X < F^\dagger(\tau)\right\}$ and continuity of probability measures. If for any $n \in \mathbb{N}$, we had $\Pr(X \leq F^\dagger(\tau) - n^{-1}) \geq \tau$, *i.e.*, $F(F^\dagger(\tau) - n^{-1}) \geq \tau$, then by definition of $F^\dagger(\tau) = \inf\{x \in [0,1] : F(x) \geq t\}$, we have $F^\dagger(\tau) \leq F^\dagger(\tau) - n^{-1}$, which is a contradiction. Therefore, it must be that for all $n \in \mathbb{N}$, we have $\Pr(X \leq F^\dagger(\tau) - n^{-1}) < \tau$, so $\Pr(X < F^\dagger(\tau)) = \lim_{n\to\infty} \Pr(X \leq F^\dagger(\tau) - n^{-1}) \leq \tau$. □

The following interpretation of CVaR_τ due to Acerbi & Tasche (2002) will be very useful. An alternative proof was given in Kisiala (2015, Proposition 2.2).

Lemma C.3 (Acerbi's Integral Formula)**.** *For any non-negative random variable X with CDF F, we have*

$$\mathrm{CVaR}_\tau(X) = \tau^{-1}\int_0^\tau F^\dagger(y)\mathrm{d}y = \mathbb{E}\left[F^\dagger(U) \mid U \leq \tau\right],$$

where $U \sim \mathrm{Unif}([0,1])$.

Now suppose $X_{1:N}$ are i.i.d. copies of $X \in [0,1]$. Define the empirical CVaR as the CVaR of the empirical distribution $\widehat{F}_N(x) = \frac{1}{N}\sum_{i=1}^N \mathbb{I}[X_i \le x]$.

$$\widehat{\mathrm{CVaR}}_\tau(X_{1:N}) = \max_{b\in[0,1]}\left\{b - \frac{1}{N\tau}\sum_{i=1}^N (b - X_i)^+\right\}.$$

Let $X_{(i)}$ denote the i-th increasing order statistic.

Lemma C.4. *The maximum for the empirical CVaR is attained at the empirical quantile $X_{\lceil N\tau\rceil}$. Hence,*

$$\widehat{\mathrm{CVaR}}_\tau(X_{1:N}) = \left(1 - \frac{\lceil N\tau\rceil}{N\tau}\right)X_{\lceil N\tau\rceil} + \frac{1}{N\tau}\sum_{i=1}^{\lceil N\tau\rceil} X_{(i)}.$$

Proof. By Lemma C.2, the maximum is attained at the τ-th quantile of the empirical distribution, *i.e.*, $\widehat{F}_N^\dagger(\tau) = \inf\left\{x : \widehat{F}_N(x) \ge \tau\right\}$. Let $k \in [N]$ be the largest $X_{(k)}$ such that such that $\widehat{F}_N(X_{(k)}) = \frac{k}{N} < \tau \le \frac{k+1}{N} = \widehat{F}_N(X_{(k+1)})$. This implies that $\widehat{F}_N^\dagger(\tau) = X_{(k+1)}$. Note that $k < N\tau \le k+1$, so $k+1 = \lceil N\tau\rceil$. Thus,

$$\begin{aligned}
\widehat{\mathrm{CVaR}}_\tau(X_{1:N}) &= X_{(\lceil N\tau\rceil)} - \frac{1}{N\tau}\sum_{i=1}^N \left(X_{(\lceil N\tau\rceil)} - X_i\right)^+ \\
&= X_{(\lceil N\tau\rceil)} - \frac{1}{N\tau}\sum_{i=1}^{\lceil N\tau\rceil}\left(X_{(\lceil N\tau\rceil)} - X_{(i)}\right) \\
&= \left(1 - \frac{\lceil N\tau\rceil}{N\tau}\right)X_{(\lceil N\tau\rceil)} + \frac{1}{N\tau}\sum_{i=1}^{\lceil N\tau\rceil} X_{(i)}.
\end{aligned}$$

□

Lemma C.5. *Let $U_{1:N}$ be i.i.d. copies of* $\mathrm{Unif}([0,1])$. *Let $p \in (0,1)$. For any $\delta \in (0,1)$, w.p. at least $1-\delta$, we have*

$$\left|U_{\lceil Np\rceil} - p\right| \le \sqrt{\frac{3p(1-p)\log(2/\delta)}{N}} + \frac{5\log(2/\delta)}{N},$$

provided that $N \ge 25\log(2/\delta)$.

Proof. Let F be the distribution function of $\mathrm{Unif}([0,1])$, and $\widehat{F}_N$ the empirical distribution of $U_{1:N}$. Note that $U_{(\lceil Np\rceil)} = \widehat{F}_N^\dagger(p)$, by reasoning in the proof of Lemma C.4. So the left hand side is $\left|p - \widehat{F}_N^\dagger(p)\right|$.

Now consider any error $\varepsilon \in (0,1)$. We have

$$\widehat{F}_N^\dagger(p) \le p+\varepsilon \iff p \le \widehat{F}_N(p+\varepsilon) \iff p+\varepsilon - \widehat{F}_N(p+\varepsilon) \le \varepsilon,$$

and

$$\widehat{F}_N^\dagger(p) > p-\varepsilon \iff p > \widehat{F}_N(p-\varepsilon) \iff \widehat{F}_N(p-\varepsilon) - (p-\varepsilon) < \varepsilon.$$

In both cases, we can use Bernstein on $\mathbb{I}[U \le p-\varepsilon]$ or $\mathbb{I}[U \le p+\varepsilon]$ to obtain a bound depending on the variance. In the first case, $\sqrt{\mathrm{Var}(\mathbb{I}[U \le p-\varepsilon])} \le \sqrt{\mathrm{Var}(\mathbb{I}[U \le p])} + \sqrt{\mathrm{Var}(\mathbb{I}[U \le p-\varepsilon] - \mathbb{I}[U \le p])} \le p(1-p) + \varepsilon$. Similarly, $\sqrt{\mathrm{Var}(\mathbb{I}[U \le p+\varepsilon])} \le \sqrt{\mathrm{Var}(\mathbb{I}[U \le p])} + \sqrt{\mathrm{Var}(\mathbb{I}[U \le p+\varepsilon] - \mathbb{I}[U \le p])} \le p(1-p) + \varepsilon$. Thus, we have w.p. at least $1-\delta$,

$$(p+\varepsilon) - \widehat{F}_N(p+\varepsilon) \le \sqrt{\frac{2p(1-p)\log(2/\delta)}{N}} + \frac{\log(2/\delta)}{N} + \sqrt{\frac{2\varepsilon^2\log(2/\delta)}{N}},$$

and

$$\widehat{F}_N(p-\varepsilon)-(p-\varepsilon) < \sqrt{\frac{3p(1-p)\log(2/\delta)}{N}} + \frac{\log(2/\delta)}{N} + \sqrt{\frac{3\varepsilon^2\log(2/\delta)}{N}}.$$

So we can set $\varepsilon = \sqrt{\frac{3p(1-p)\log(2/\delta)}{N}} + \frac{5\log(2/\delta)}{N}$, as the third error term with this setting of ε is at most $\frac{3\log(2/\delta)}{N} + \frac{5\log(2/\delta)^{1.5}}{N^{1.5}} \le \frac{4\log(2/\delta)}{N}$ when $N \ge 25\log(2/\delta)$. Thus w.p. at least $1-\delta$, we have $\left|\widehat{F}_N^{\dagger}(p) - p\right| \le \varepsilon$. □

Theorem C.6. *Let $X_{1:N}$ be N i.i.d. copies of a random variable $X \in [0,1]$. Then for any $\delta \in (0,1)$, w.p. at least $1-\delta$, if $N \ge 25\log(2/\delta)$, then we have*

$$\left|\widehat{\mathrm{CVaR}}_\tau(X_{1:N}) - \mathrm{CVaR}_\tau(X)\right| \le \sqrt{\frac{3\log(2/\delta)}{N\tau}} + \frac{15\log(2/\delta)}{N\tau}.$$

Proof. We use the interpretation of the empirical CVaR in Lemma C.4. The first term is lower order since

$$\left|1 - \frac{\lceil N\tau\rceil}{N\tau}\right| \le \frac{1}{N\tau}.$$

Now recall that by the inverse CDF trick, we have $X_i = F^\dagger(U_i)$ where U_i are i.i.d. copies of $\mathrm{Unif}([0,1])$. Since $F^\dagger$ is non-decreasing, we have $X_{(i)} = F^\dagger(U_{(i)})$. Thus, the second term of Lemma C.4 is

$$\frac{1}{N\tau}\sum_{i=1}^{\lceil N\tau\rceil} X_{(i)} = \frac{1}{N\tau}\sum_{i=1}^{\lceil N\tau\rceil} F^\dagger(U_{(i)}) = \frac{1}{N\tau}\sum_{i=1}^{N} F^\dagger(U_i)\mathbb{I}\left[U_i \le U_{(\lceil N\tau\rceil)}\right],$$

which we want to show is close to $\mathrm{CVaR}_\tau(X) = \tau^{-1}\mathbb{E}\left[F^\dagger(U)\mathbb{I}\left[U\le\tau\right]\right]$ by Lemma C.3. If $U_{(\lceil N\tau\rceil)}$ were replaced by τ, we can simply invoke Bernstein and note that $\mathrm{Var}\left(F^\dagger(U)\mathbb{I}\left[U\le\tau\right]\right) \le \Pr(U\le\tau) = \tau$, which gives

$$\left|\frac{1}{N\tau}\sum_{i=1}^{N} F^\dagger(U_i)\mathbb{I}\left[U_i\le\tau\right] - \tau^{-1}\mathbb{E}\left[F^\dagger(U)\mathbb{I}\left[U\le\tau\right]\right]\right| \le \tau^{-1}\left(\sqrt{\frac{2\tau\log(2/\delta)}{N}} + \frac{\log(2/\delta)}{N}\right).$$

Thus, we just need to bound the difference term,

$$\left|\frac{1}{N\tau}\sum_{i=1}^{N} F^\dagger(U_i)\left(\mathbb{I}\left[U_i \le U_{(\lceil N\tau\rceil)}\right] - \mathbb{I}\left[U_i\le\tau\right]\right)\right|.$$

By Lemma C.5, we have $\left|U_{\lceil N\tau\rceil} - \tau\right| \le \varepsilon$ w.p. $1-\delta$, where $\varepsilon = \sqrt{\frac{3\tau(1-\tau)\log(2/\delta)}{N}} + \frac{5\log(2/\delta)}{N}$. So, for any U_i we have $\mathbb{I}\left[U_i \le U_{(\lceil N\tau\rceil)}\right] - \mathbb{I}\left[U_i\le\tau\right] \le \mathbb{I}\left[\tau\le U_i\le\tau+\varepsilon\right]$ and $\mathbb{I}\left[U_i\le\tau\right] \le \mathbb{I}\left[\tau-\varepsilon\le U_i\le\tau\right]$, so the difference term is at most,

$$\le \max\left\{\frac{1}{N\tau}\sum_{i=1}^{N} F^\dagger(U_i)\mathbb{I}\left[\tau-\varepsilon\le U_i\le\tau\right], \frac{1}{N\tau}\sum_{i=1}^{N} F^\dagger(U_i)\mathbb{I}\left[\tau\le U_i\le\tau+\varepsilon\right]\right\}.$$

By applying another Bernstein, and noting that $\sqrt{\mathrm{Var}(F^\dagger(U)\mathbb{I}\left[\tau-\varepsilon\le U\le\tau\right])} \le \varepsilon$, $\sqrt{\mathrm{Var}(F^\dagger(U)\mathbb{I}\left[\tau\le U\le\tau+\varepsilon\right])} \le \varepsilon$, we have this is at most

$$\begin{aligned}
&\tau^{-1}\left(\max\left\{\mathbb{E}\left[F^\dagger(U)\mathbb{I}\left[\tau-\varepsilon\le U\le\tau\right]\right], \mathbb{E}\left[F^\dagger(U)\mathbb{I}\left[\tau\le U\le\tau+\varepsilon\right]\right]\right\} + \sqrt{\frac{2\varepsilon^2\log(2/\delta)}{N}} + \frac{\log(2/\delta)}{N}\right)\\
&\le \tau^{-1}\left(\varepsilon + \sqrt{\frac{2\varepsilon^2\log(2/\delta)}{N}} + \frac{\log(2/\delta)}{N}\right)\\
&\le \sqrt{\frac{3\log(2/\delta)}{N\tau}} + \frac{5\log(2/\delta)}{N\tau} + \frac{3\log(2/\delta)}{N\sqrt{\tau}} + \frac{4\log^{1.5}(2/\delta)}{N^{1.5}\tau} + \frac{\log(2/\delta)}{N\tau}\\
&\le \sqrt{\frac{3\log(2/\delta)}{N\tau}} + \frac{15\log(2/\delta)}{N\tau}, \qquad \text{(when } N \ge 16\log(2/\delta)\text{)}
\end{aligned}$$

where the bound on ε occurs when $N \ge 25\log(2/\delta)$, which also implies the last inequality. □

D. Proofs for Lower Bounds

D.1. CVaR MAB Lower Bound

Let us first define some MAB notations that make explicit the dependence on the current MAB problem instance ν and the learner Alg. Recall that ν is a vector of A reward distributions, and in the k-th episode, Alg picks an action a_k based on the historical actions and rewards. Let $\Delta_a(\nu) = \mathrm{CVaR}_\tau^\star(\nu) - \mathrm{CVaR}_\tau(\nu(a))$ where $\mathrm{CVaR}_\tau^\star(\nu) = \max_{a\in\mathcal{A}} \mathrm{CVaR}_\tau(\nu(a))$. Let $\mathrm{Regret}_\tau^{\mathrm{MAB}}(K, \nu, \mathrm{Alg})$ denote the regret of running Alg in MAB ν for K episodes. Let $T_k(a)$ denote the number of times an arm a has been pulled up to time K.

For two distributions P, Q, recall the KL-divergence is defined as

$$D_{\mathrm{KL}}(P,Q) = \begin{cases} \int \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega)\mathrm{d}P(\omega)\Big), \text{ if } P \ll Q, \\ \infty, \text{ otherwise.} \end{cases}$$

A key inequality for lower bounds is the Bretagnolle-Huber inequality, cf. (Lattimore & Szepesvári, 2020, Theorem 14.2),

Lemma D.1 (Bretagnolle-Huber)**.** *Let P, Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and $A \in \mathcal{F}$ be any event. Then*

$$P(A) + Q(A^C) \geq \frac{1}{2}\exp(-D_{\mathrm{KL}}(P,Q)).$$

Lemma D.2 (Regret Decomposition)**.** *For any MAB instance ν and learner* Alg*, we have*

$$\mathbb{E}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K,\nu,\mathrm{Alg})\right] = \sum_{a\in\mathcal{A}} \Delta_a(\nu)\mathbb{E}[T_a(K)],$$

where the expectations are with respect to the trajectory of running Alg *in ν.*

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K,\nu,\mathrm{Alg})\right] \\
&= \sum_{k=1}^K \mathrm{CVaR}_\tau^\star(\nu) - \mathbb{E}[\mathrm{CVaR}_\tau(\nu(a_k))] \\
&= \sum_{k=1}^K \mathbb{E}\left[(\mathrm{CVaR}_\tau^\star(\nu) - \mathrm{CVaR}_\tau(\nu(a_k)))) \sum_{a\in\mathcal{A}} \mathbb{I}\left[a_k = a\right]\right] \\
&= \sum_{a\in\mathcal{A}}\sum_{k=1}^K \mathbb{E}[(\mathrm{CVaR}_\tau^\star(\nu) - \mathrm{CVaR}_\tau(\nu(a_k)))\mathbb{I}\left[a_k = a\right]].
\end{aligned}$$

Notice that if once we condition on a_k, if $a_k = a$, the difference $\mathrm{CVaR}_\tau^\star(\nu) - \mathrm{CVaR}_\tau(\nu(a_k))$ is simply $\Delta_a(\nu)$. If $a_k \neq a$, then we get 0. So, by the tower rule,

$$\mathbb{E}[(\mathrm{CVaR}_\tau^\star(\nu) - \mathrm{CVaR}_\tau(\nu(a_k)))\mathbb{I}\left[a_k = a\right]] = \mathbb{E}[\mathbb{I}\left[a_k = a\right]\Delta_a(\nu)].$$

Therefore, continuing from before,

$$\begin{aligned}
&= \sum_{a\in\mathcal{A}}\sum_{k=1}^K \mathbb{E}[\mathbb{I}\left[a_k = a\right]\Delta_a(\nu)] \\
&= \sum_{a\in\mathcal{A}} \Delta_a(\nu)\sum_{k=1}^K \mathbb{E}[\mathbb{I}\left[a_k = a\right]] \\
&= \sum_{a\in\mathcal{A}} \Delta_a(\nu)\mathbb{E}[T_a(K)].
\end{aligned}$$

□

Theorem 3.1. *Fix any $\tau \in (0, 1/2), A \in \mathbb{N}$. For any algorithm, there is a MAB problem with Bernoulli rewards s.t. if $K \geq \sqrt{\frac{A-1}{8\tau}}$, then $\mathbb{E}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K)\right] \geq \frac{1}{24e}\sqrt{\frac{(A-1)K}{\tau}}$.*

Proof of Theorem 3.1. Fix any $\tau \in (0, 1/2)$ and any MAB algorithm Alg. WLOG suppose $\mathcal{A} = [A]$. Define the shorthand, $\beta_c = \mathrm{Ber}(1 - \tau + c\varepsilon)$, *i.e.*, larger c implies ε more likelihood of pulling 1. Construct two MAB instances as follows,

$$\nu = (\beta_1, \beta_0, \ldots, \beta_0)$$
$$\nu' = (\beta_1, \beta_0, \ldots, \beta_0, \underbrace{\beta_2}_{\text{index } i}, \beta_0, \ldots, \beta_0), \text{ where } i = \arg\min_{a>1} \mathbb{E}_{\nu,\mathrm{Alg}}[T_a(K)].$$

For the first MAB instance ν, the optimal action is $a^\star(\nu) = 1$, and $\Delta_a(\nu) = \tau^{-1}\varepsilon$ for all $a > 1$. By Lemma D.2,

$$\begin{aligned}
\mathbb{E}_{\nu,\mathrm{Alg}}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K, \nu, \mathrm{Alg})\right] &= \sum_{a\in\mathcal{A}} \Delta_a(\nu)\mathbb{E}_{\nu,\mathrm{Alg}}[T_a(K)] \\
&= \tau^{-1}\varepsilon(K - \mathbb{E}_{\nu,\mathrm{Alg}}[T_1(K)]) \\
&\geq \tau^{-1}\varepsilon \Pr_{\nu,\mathrm{Alg}}\left(K - T_1(K) \geq \frac{K}{2}\right)\frac{K}{2} && \text{(Markov's inequality)} \\
&=\geq \tau^{-1}\varepsilon \Pr_{\nu,\mathrm{Alg}}\left(T_1(K) \leq \frac{K}{2}\right)\frac{K}{2}.
\end{aligned}$$

For the second MAB instance ν', the optimal action is $a^\star(\nu')$, and $\Delta_1(\nu) = \tau^{-1}\varepsilon$. By Lemma D.2,

$$\begin{aligned}
\mathbb{E}_{\nu',\mathrm{Alg}}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K, \nu', \mathrm{Alg})\right] &= \sum_{a\in\mathcal{A}} \Delta_a(\nu')\mathbb{E}_{\nu',\mathrm{Alg}}[T_a(K)] \\
&\geq \nu_1(\nu')\mathbb{E}_{\nu',\mathrm{Alg}}[T_1(K)] \\
&> \tau^{-1}\varepsilon \Pr_{\nu',\mathrm{Alg}}\left(T_1(K) > \frac{K}{2}\right)\frac{K}{2}. && \text{(Markov's inequality)}
\end{aligned}$$

Let $\mathbb{P}_{\nu,\mathrm{Alg}}$ denote the trajectory distribution from running Alg in MAB ν. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{Alg}}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K, \nu, \mathrm{Alg})\right] + \mathbb{E}_{\nu',\mathrm{Alg}}\left[\mathrm{Regret}_\tau^{\mathrm{MAB}}(K, \nu', \mathrm{Alg})\right] \\
&> \frac{K\varepsilon}{2\tau}\left(\Pr_{\nu,\mathrm{Alg}}\left(T_1(K) \leq \frac{K}{2}\right) + \Pr_{\nu',\mathrm{Alg}}\left(T_1(K) > \frac{K}{2}\right)\right) \\
&\geq \frac{K\varepsilon}{4\tau}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{\nu,\mathrm{Alg}}, \mathbb{P}_{\nu',\mathrm{Alg}})) && \text{(Bretagnolle-Huber Lemma D.1)} \\
&= \frac{K\varepsilon}{4\tau}\exp\left(-\sum_{a\in\mathcal{A}} \mathbb{E}_{\nu,\mathrm{Alg}}[T_a(K)]D_{\mathrm{KL}}(\nu(a), \nu'(a))\right) && \text{(Lattimore \& Szepesvári, 2020, Lemma 15.1)} \\
&= \frac{K\varepsilon}{4\tau}\exp(-\mathbb{E}_{\nu,\mathrm{Alg}}[T_i(K)]D_{\mathrm{KL}}(\nu(i), \nu'(i))) && \text{(other arms are the same for } \nu, \nu') \\
&= \frac{K\varepsilon}{4\tau}\exp(-\mathbb{E}_{\nu,\mathrm{Alg}}[T_i(K)]D_{\mathrm{KL}}(\nu(i), \nu'(i))) \\
&\geq \frac{K\varepsilon}{4\tau}\exp\left(-\frac{8K\varepsilon^2}{(A-1)\tau}\right).
\end{aligned}$$

The last inequality uses two facts. By definition of $i = \arg\min_{a>1} \mathbb{E}_{\nu,\mathrm{Alg}}[T_a(K)]$, $\mathbb{E}_{\nu,\mathrm{Alg}}[T_i(K)] \leq \frac{K}{A-1}$. Also, by Lemma D.4, $D_{\mathrm{KL}}(\nu(i), \nu'(i)) \leq 8\varepsilon^2\tau^{-1}$. Setting $\varepsilon^2 = \frac{(A-1)\tau}{8K}$ and noting $2\max\{a, b\} \geq a + b$ gives the desired lower bound. □

Lemma D.3. *For any $\tau \in (0, 1/2)$ and $\varepsilon \in [0, \tau]$, we have*

$$\mathrm{CVaR}_\tau(\mathrm{Ber}(1 - \tau + \varepsilon)) = \tau^{-1}\varepsilon.$$

Proof. The CDF of $X \sim \mathrm{Ber}(1-\tau+\varepsilon)$ is as follows,

$$F(x) = \begin{cases} 0, \text{ if } x < 0, \\ \tau - \varepsilon, \text{ if } x \in [0,1), \\ 1, \text{ if } x \geq 1. \end{cases}$$

Therefore, $F^\dagger(\tau) = \inf\{x : F(x) \geq \tau\} = 1$ for any $\varepsilon > 0$, and it is 0 when $\varepsilon = 0$. By Lemma C.2, we have

$$\mathrm{CVaR}_\tau(\mathrm{Ber}(1-\tau)) = 0 - \tau^{-1}\mathbb{E}\big[(0-X)^+\big] = 0,$$

and

$$\mathrm{CVaR}_\tau(\mathrm{Ber}(1-\tau+\varepsilon)) = 1 - \tau^{-1}\mathbb{E}\big[(1-X)^+\big] = 1 - \tau^{-1}(\tau-\varepsilon) = \tau^{-1}\varepsilon.$$

□

Lemma D.4. *For any $\tau \in (0, 1/2)$ and $\varepsilon \in [0, \tau]$, we have*

$$D_{\mathrm{KL}}(\mathrm{Ber}(1-\tau), \mathrm{Ber}(1-\tau+\varepsilon)) \leq 2\varepsilon^2\tau^{-1}.$$

Proof. By explicit computation, we have

$$\begin{aligned} & D_{\mathrm{KL}}(\mathrm{Ber}(1-\tau), \mathrm{Ber}(1-\tau+\varepsilon)) \\ &= \tau\log\left(\frac{\tau}{\tau-\varepsilon}\right) + (1-\tau)\log\left(\frac{1-\tau}{1-\tau+\varepsilon}\right) \\ &\leq \tau\log\left(\frac{\tau}{\tau-\varepsilon}\right) + \tau\log\left(\frac{\tau}{\tau+\varepsilon}\right) \\ &= -\tau\log\left(1-\frac{\varepsilon^2}{\tau^2}\right) \\ &\leq 2\varepsilon^2\tau^{-1}. \end{aligned}$$

The first inequality is because $f(x) = x\log\left(\frac{x}{x+\varepsilon}\right)$ is a decreasing function and $1-\tau \geq \tau$. The second inequality is because $-\log(1-x) \leq 2x$ for $x \in [0,1]$. □

D.2. Lower bound for CVaR RL

Corollary 3.2. *Fix any $\tau \in (0, 1/2), A, H \in \mathbb{N}$. For any algorithm, there is an MDP (with $S = \Theta(A^{H-1})$) s.t. if $K \geq \sqrt{\frac{S(A-1)}{8\tau}}$, then $\mathbb{E}\big[\mathrm{Regret}_\tau^{\mathrm{RL}}(K)\big] \geq \frac{1}{24e}\sqrt{\frac{S(A-1)K}{\tau}}$.*

Proof of Corollary 3.2. Fix any τ, A, H. Consider an MDP where the states are represented by an A-balanced tree with depth H (each node of the tree is a state). The initial state s_1 is the root, and based on the action a_1, transits to the a_1-th node in the next layer. The process repeats until we've reached one of the A^{H-1} leaves, where a reward is given (which also depends on the action taken at the leaf). There are no rewards until the last step. The number of states is $S = 1 + A + ... + A^{H-1}$, since the h-th layer of the tree has A^{h-1} states.

Since there are no rewards until the last step, running in this MDP reduces to a MAB with A^H "arms" where the "arms" are the sequences of actions $a_{1:H}$. So, by Theorem 3.1, for any RL algorithm, there is an MDP constructed this way (with Bernoulli rewards at the end) such that if $K \geq \sqrt{\frac{A^H-1}{8\tau}}$, then $\mathbb{E}\big[\mathrm{Regret}_\tau^{\mathrm{RL}}(K)\big] \geq \frac{1}{24e}\sqrt{\frac{(A^H-1)K}{\tau}}$. The key observation is that $A^H - 1 = (A-1)\big(A^{H-1} + A^{H-2} + \cdots + A + 1\big) = (A-1)S$. This concludes the proof. □

E. Proofs for BERNSTEIN-UCB

For any arm $a \in \mathcal{A}$, let $b_a^\star$ denote the τ-th quantile of $R(a)$, so

$$b_a^\star = \arg\max_{b\in[0,1]} \left\{ b - \tau^{-1}\mathbb{E}_{R\sim\nu(a)}\left[(b-R)^+\right]\right\}$$
$$\mathrm{CVaR}_\tau(R(a)) = b_a^\star - \tau^{-1}\mathbb{E}_{R\sim\nu(a)}\left[(b_a^\star - R)^+\right].$$

Let us denote

$$\mu(b,a) = \mathbb{E}_{R\sim\nu(a)}\left[(b-R)^+\right],$$
$$\widehat{\mu}_k(b,a) = \frac{1}{N_k(a)}\sum_{i=1}^{k-1}(b-r_i)^+\mathbb{I}\left[a_i = a\right].$$

Recall that $a^\star$ is the arm with the highest CVaR_τ.

For any $\delta \in (0,1)$, w.p. at least $1-\delta$, uniform Bernstein implies that for all b, a,

$$|\widehat{\mu}_k(b,a) - \mu(b,a)| \le \sqrt{\frac{2\tau\log(2AK/\delta)}{N_k(a)}} + \frac{\log(2AK/\delta)}{N_k(a)}. \tag{7}$$

Note that our bonus Eq. (3) is constructed to match the upper bound. This implies that $\widehat{\mu}_k - \textsc{Bon}_k$ is a pessimistic estimate of μ.

Lemma E.1 (Pessimism)**.** *For all $k \in [K]$*

$$\min_{a\in\mathcal{A}}\{\widehat{\mu}_k(b_a^\star, a) - \textsc{Bon}_k(a)\} \le \mu(b_{a^\star}^\star, a^\star).$$

Proof. Fix any $k \in [K]$. By Eq. (7), for all $a \in \mathcal{A}$,

$$\widehat{\mu}_k(b_a^\star, a) - \textsc{Bon}_k(a) \le \mu(b_a^\star, a).$$

Hence,

$$\begin{aligned}\min_{a\in\mathcal{A}}\{\widehat{\mu}_k(b_a^\star, a) - \textsc{Bon}_k(a)\} &\le \widehat{\mu}_k(b_{a^\star}^\star, a^\star) - \textsc{Bon}_k(a^\star)\\ &\le \mu(b_{a^\star}^\star, a^\star).\end{aligned}$$

□

Theorem 4.1. *For any $\delta \in (0,1)$, w.p. at least $1-\delta$,* BERNSTEIN-UCB *with $\varepsilon \le \sqrt{A/2\tau K}$ enjoys*

$$\mathrm{Regret}_\tau^{\mathrm{MAB}}(K) \le 4\sqrt{\tau^{-1}AK}L + 16\tau^{-1}AL^2.$$

Proof of Theorem 4.1.

$$
\begin{aligned}
&\mathrm{Regret}_\tau^{\mathrm{MAB}}(K) \\
&= \sum_{k=1}^{K} \mathrm{CVaR}_\tau^\star - \mathrm{CVaR}_\tau(R(a_k)) \\
&= \sum_{k=1}^{K} \left\{ b^\star_{a^\star} - \tau^{-1}\mu(b^\star_{a^\star}, a^\star) \right\} - \mathrm{CVaR}_\tau(\nu(a_k)) \\
&\le \sum_{k=1}^{K} \left\{ b^\star_{a^\star} - \tau^{-1} \min_{a\in\mathcal{A}} (\widehat{\mu}_k(b^\star_{a^\star}, a) - \textsc{Bon}_k(a)) \right\} - \mathrm{CVaR}_\tau(\nu(a_k)) && \text{(pessimism Lemma E.1)} \\
&= \sum_{k=1}^{K} \max_{a\in\mathcal{A}} \left\{ b^\star_{a^\star} - \tau^{-1} (\widehat{\mu}_k(b^\star_{a^\star}, a) - \textsc{Bon}_k(a)) \right\} - \mathrm{CVaR}_\tau(\nu(a_k)) \\
&\le K\varepsilon + \sum_{k=1}^{K} \max_{a\in\mathcal{A}} \left\{ \widehat{b}_{a,k} - \tau^{-1} \left(\widehat{\mu}_k(\widehat{b}_{a,k}, a) - \textsc{Bon}_k(a) \right) \right\} - \mathrm{CVaR}_\tau(\nu(a_k)) && (\widehat{b}_{a,k} \text{ is } \varepsilon\text{-optimal}) \\
&= K\varepsilon + \sum_{k=1}^{K} \left\{ \widehat{b}_{a_k,k} - \tau^{-1} \left(\widehat{\mu}_k(\widehat{b}_{a_k,k}, a_k) - \textsc{Bon}_k(a_k) \right) \right\} - \mathrm{CVaR}_\tau(\nu(a_k)) && \text{(defn. of } a_k) \\
&\le K\varepsilon + \sum_{k=1}^{K} \tau^{-1}\textsc{Bon}_k(a_k) + \max_{b\in[0,1]} \left\{ b - \tau^{-1}\widehat{\mu}_k(b, a_k) \right\} - \mathrm{CVaR}_\tau(\nu(a_k)) \\
&= K\varepsilon + \sum_{k=1}^{K} \tau^{-1}\textsc{Bon}_k(a_k) + \widehat{\mathrm{CVaR}}_\tau(\{r_i\}_{i\in\mathcal{I}_k(a_k)}) - \mathrm{CVaR}_\tau(\nu(a_k)) \\
&\le K\varepsilon + \sum_{k=1}^{K} \sqrt{\frac{2L}{N_k(a)\tau}} + \frac{L}{N_k(a)\tau} + \sqrt{\frac{3L}{N_k(a)\tau}} + \frac{15L}{N_k(a)\tau} && (\mathrm{CVaR}_\tau \text{ concentration Theorem C.6}) \\
&\le K\varepsilon + \sum_{k=1}^{K} \sqrt{\frac{10L}{N_k(a)\tau}} + \frac{16L}{N_k(a)\tau} \\
&\le K\varepsilon + \sqrt{10L\tau^{-1}} \cdot \sqrt{AKL} + 16L\tau^{-1} \cdot A\log(K). && \text{(elliptical potential Lemma G.12)}
\end{aligned}
$$

A technical detail is that CVaR_τ concentration only applies when $N_k(a) \ge 25L$. We can trivially bound the total regret of the episodes when $N_k(a) < 25L$ by $25AL$. Also, we remark the concentration step applies since $\{r_i\}_{i\in\mathcal{I}_k(a)}$ are i.i.d. via the tape framework, so we do not need to generalize Theorem C.6 to martingale sequences. Finally, setting $\varepsilon = \sqrt{\tau^{-1}A/2K}$ renders it lower order. □

F. Proofs for Augmented MDP

We first define the memory-MDP model, where the MDP is also equipped with a memory generator M_h, which generates $m_h \sim M_h(s_h, a_h, r_h, \mathcal{H}_h)$. These memories are stored into the history $\mathcal{H}_h = (s_t, a_t, r_t, m_t)_{t\in[h-1]}$ and may be used by history dependent policies in future time steps. Concretely, rolling out π proceeds as follows: for any $h \in [H]$, $a_h \sim \pi_h(s_h, \mathcal{H}_h)$, $s_{h+1} \sim P^\star(s_h, a_h)$, $r_h \sim R(s_h, a_h)$ and $m_h \sim M_h(s_h, a_h, r_h, \mathcal{H}_h)$.

We can also extend the above formulation to the augmented MDP, where the state is augmented with b as in Section 5.1. Here, the history is $\mathcal{H}_h^{\mathrm{Aug}} = (s_t, b_t, a_t, r_t, m_t)_{t\in[h-1]}$. Let $\Pi_{\mathcal{H}}^{\mathrm{Aug}}$ represent the set of history dependent policies in this augmented MDP with memory. Also, recall that Π^{Aug} is the set of Markov, deterministic policies in the augmented MDP.

The V function is defined for these multiple types of policies:

$$\pi \in \Pi_{\mathcal{H}} : V_h^\pi(s_h, b_h; \mathcal{H}_h) = \mathbb{E}_\pi\left[\left(b_h - \sum_{t=h}^H r_t\right)^+ \mid s_h, b_h, \mathcal{H}_h\right]$$

$$\rho \in \Pi^{\mathrm{Aug}} : V_h^\rho(s_h, b_h) = \mathbb{E}_\rho\left[\left(b_h - \sum_{t=h}^H r_t\right)^+ \mid s_h, b_h\right]$$

$$\rho \in \Pi_{\mathcal{H}}^{\mathrm{Aug}} : V_h^\rho(s_h, b_h; \mathcal{H}_h^{\mathrm{Aug}}) = \mathbb{E}_\rho\left[\left(b_h - \sum_{t=h}^H r_t\right)^+ \mid s_h, b_h, \mathcal{H}_h^{\mathrm{Aug}}\right]$$

Notice that rolling out ρ, b in the augmented MDP is equivalent to rolling out $\pi^{\rho,b}$ in the original MDP, where

$$\pi_h^{\rho,b}(s_h, \mathcal{H}_h) = \rho_h(s_h, b - r_1 - \ldots - r_{h-1}).$$

Thus, it's evident that their V functions should match.

Lemma F.1. *Fix any $\rho \in \Pi^{\mathrm{Aug}}, h \in [H]$, augmented state (s_h, b_h) and history $\mathcal{H}_h$. Then, we have $V_h^\rho(s_h, b_h) = V_h^{\pi^{\rho,b}}(s_h, b_h; \mathcal{H}_h)$ for $b = b_h + r_1 + \ldots + r_{h-1}$. In particular, we have $V_1^\rho(s_1, \cdot) = V_1^{\pi^{\rho,b}}(s_1, \cdot)$.*

Proof. The setting of b in the lemma satisfies $b_h = b - r_1 - \ldots - r_{h-1}$. Therefore, the trajectories of (ρ, b) and $\pi^{\rho,b}$ are exactly coupled. □

We now show the key result of this section. The theorem shows that the $V^\star, U^\star$ functions defined via the Bellman optimality equations (from Section 5.1) correspond to the V, U functions of $\rho^\star$. Furthermore, the Markov (in augmented state) and deterministic $\rho^\star$ is in fact an optimal policy amongst all history-dependent policies in the augmented MDP with memory! This result and our proof is analogous to the "Markov optimality theorem" of vanilla RL, *e.g.*, (Puterman, 2014), (Agarwal et al., 2021, Theorem 1.7).

Theorem F.2. *For all h we have $U_h^\star = U_h^{\rho^\star}$ and $V_h^\star = V_h^{\rho^\star}$. Furthermore, for all $s_h, b_h, \mathcal{H}_h^{\mathrm{Aug}}$, we have*

$$V_h^\star(s_h, b_h) = \inf_{\rho\in\Pi_{\mathcal{H}}^{\mathrm{Aug}}} V_h^\rho(s_h, b_h; \mathcal{H}_h^{\mathrm{Aug}}).$$

In particular, $V_1^\star(s_1, b) = \inf_{\rho\in\Pi_{\mathcal{H}}^{\mathrm{Aug}}} V_1^\rho(s_1, b)$ for all b.

Proof. We first prove the claim that $U_h^\star = U_h^{\rho^\star}$ and $V_h^\star = V_h^{\rho^\star}$. The base case of $H+1$ is trivial since $V_{H+1}(s, b) = b^+$ everywhere. For the inductive step, fix any $h \in [H]$ and suppose the claim is true for $h+1$. Then,

$$\begin{aligned}
U_h^{\rho^\star}(s_h, b_h, a_h) &= \mathbb{E}_{s_{h+1}\sim P^\star(s_h,a_h), r_h\sim R(s_h,a_h)}\left[V_{h+1}^{\rho^\star}(s_{h+1}, b_h - r_h)\right] && \text{(Bellman Eqns)}\\
&= \mathbb{E}_{s_{h+1}\sim P^\star(s_h,a_h), r_h\sim R(s_h,a_h)}\left[V_{h+1}^\star(s_{h+1}, b_h - r_h)\right] && \text{(IH)}\\
&= U_h^\star(s_h, b_h, a_h). && \text{(Bellman Opt. Eqns)}
\end{aligned}$$

This proves that $U_h^\star = U_h^{\rho^\star}$. For V,

$$\begin{aligned}
V_h^{\rho^\star}(s_h, b_h) &= \mathbb{E}_{a_h \sim \rho_h^\star(s_h,b_h)}\left[U_h^{\rho^\star}(s_h, b_h, a_h)\right] && \text{(Bellman Eqns)}\\
&= \mathbb{E}_{a_h \sim \rho_h^\star(s_h,b_h)}[U_h^\star(s_h, b_h, a_h)] && \text{(above claim)}\\
&= \min_{a_h \in \mathcal{A}} U_h^\star(s_h, b_h, a_h) && \text{(defn. of } \rho_h^\star)\\
&= V_h^\star(s_h, b_h). && \text{(Bellman Opt. Eqns)}
\end{aligned}$$

Therefore, we've shown that $V_h^\star = V_h^{\rho^\star}$.

We also prove the second claim inductively. The base case is again trivial since $V_{H+1}(s, b) = b^+$ everywhere. For the inductive step, fix any $h \in [H]$ and suppose the claim is true for $h+1$. Now fix any s_h, b_h and $\mathcal{H}_h^{\text{Aug}}$,

$$\begin{aligned}
&\inf_{\rho \in \Pi_{\mathcal{H}}^{\text{Aug}}} V_h^\rho(s_h, b_h; \mathcal{H}_h^{\text{Aug}})\\
&= \inf_{\rho \in \Pi_{\mathcal{H}}^{\text{Aug}}} \mathbb{E}_\rho\left[\left(b_h - \sum_{t=h}^H r_t\right)^+ \mid s_h, b_h, \mathcal{H}_h^{\text{Aug}}\right]\\
&\geq \inf_{\rho \in \Pi_{\mathcal{H}}^{\text{Aug}}} \mathbb{E}_{a_h, s_{h+1}, r_h, m_h}\left[\inf_{\rho' \in \Pi_{\mathcal{H}}^{\text{Aug}}} \mathbb{E}_{\rho'}\left[\left(b_h - \sum_{t=h}^H r_t\right)^+ \mid s_{h+1}, b_{h+1}, \mathcal{H}_{h+1}^{\text{Aug}}\right]\right]\\
&= \inf_{\rho \in \Pi_{\mathcal{H}}^{\text{Aug}}} \mathbb{E}_{a_h \sim \rho_h(s_h, b_h, \mathcal{H}_h^{\text{Aug}})}\left[\mathbb{E}_{s_{h+1} \sim P^\star(s_h,a_h), r_h \sim R(s_h,a_h)}\left[V_{h+1}^\star(s_{h+1}, b_h - r_h)\right]\right] && \text{(IH)}\\
&= \min_{a \in \mathcal{A}} \mathbb{E}_{s_{h+1} \sim P^\star(s_h,a_h), r_h \sim R(s_h,a_h)}\left[V_{h+1}^\star(s_{h+1}, b_h - r_h)\right] && (\star)\\
&= V_h^\star(s_h, b_h). && \text{(by defn.)}
\end{aligned}$$

There are three key steps. First, the inequality is due to expanding out one step, where $a_h \sim \rho_h(s_h, b_h, \mathcal{H}_h^{\text{Aug}}), s_{h+1} \sim P^\star(s_h, a_h), r_h \sim R(s_h, a_h), m_h \sim M_h(s_h, a_h, r_h, \mathcal{H}_h)$, then push the inf for future steps inside the expectation. Second, the IH invocation is significant as it essentially removes dependence of the memory hallucinations m_h. Third, the step marked with $\star$ is significant since, regardless of the history, the current best action is just to minimize the inner function (which is independent of the history). We also have $V_h^\star(s_h, b_h) \leq \inf_{\rho \in \Pi_{\mathcal{H}}^{\text{Aug}}} V_h^\rho(s_h, b_h; \mathcal{H}_h^{\text{Aug}})$ since by the first part of the claim, $V_h^\star$ is the value of $\rho^\star \in \Pi_{\mathcal{H}}^{\text{Aug}}$. Thus, we've shown $V_h^\star(s_h, b_h) = \inf_{\rho \in \Pi_{\mathcal{H}}^{\text{Aug}}} V_h^\rho(s_h, b_h; \mathcal{H}_h^{\text{Aug}})$. □

As a corollary of the above theorem, we can restrict the policy class to history-dependent policies on the non-augmented MDP (and without history).

Theorem 5.1 (Optimality of Π^{Aug})**.** *For any* $b \in [0, 1]$,

$$V_1^\star(s_1, b) = V_1^{\rho^\star}(s_1, b) = \inf_{\pi \in \Pi_{\mathcal{H}}} V_1^\pi(s_1, b).$$

Proof of Theorem 5.1. The first equality is directly from Theorem F.2. We now prove the second equality. For any b,

$$\begin{aligned}
&\min_{\rho \in \Pi^{\text{Aug}}} V_1^{\pi^{\rho,b}}(s_1, b)\\
&= \min_{\rho \in \Pi^{\text{Aug}}} V_1^\rho(s_1, b) && \text{(Lemma F.1)}\\
&= \min_{\pi \in \Pi_{\mathcal{H}}^{\text{Aug}}} V_1^\pi(s_1, b) && \text{(Theorem F.2)}\\
&\leq \min_{\pi \in \Pi_{\mathcal{H}}} V_1^\pi(s_1, b)\\
&\leq \min_{\rho \in \Pi^{\text{Aug}}} V_1^{\pi^{\rho,b}}(s_1, b).
\end{aligned}$$

The last two inequalities is due to considering strictly smaller sets of policies. Therefore, we have equality throughout, which proves the claim. □

G. Proofs for CVaR-UCBVI

G.1. The high probability good event

In this section, we derive all the high probability results needed in the remainder of the proof. Fix any failure probability $\delta \in (0,1)$. Then, w.p. at least $1-\delta$, for all $h \in [H], k \in [K], s \in \mathcal{S}, a \in \mathcal{A}$, we have, for all $b \in [0,1], s' \in \mathcal{S}$,

$$\left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top \mathbb{E}_{r_h}\left[V^\star_{h+1}(\cdot, b - r_h)\right]\right| \le \sqrt{\frac{L}{N_k(s,a)}}, \tag{8}$$

$$\left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top \mathbb{E}_{r_h}\left[V^\star_{h+1}(\cdot, b - r_h)\right]\right| \le \sqrt{\frac{2\operatorname{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[V^\star_{h+1}(s', b - r_h)\right]\right)L}{N_k(s,a)}} + \frac{L}{N_k(s,a)}, \tag{9}$$

$$\left|\widehat{P}_k(s' \mid s,a) - P^\star(s' \mid s,a)\right| \le \sqrt{\frac{2P^\star(s' \mid s,a)L}{N_k(s,a)}} + \frac{L}{N_k(s,a)}. \tag{10}$$

where $r_h \sim R(s,a)$ in the expectations.

Proof. Eq. (8) is due to uniform Hoeffding applied to $\mathbb{E}_{s_{h+1},r_h}\left[V^\star_{h+1}(s_{h+1}, b - r_h)\right]$, which is 1-Lipschitz in b by Lemma G.1. Eq. (10) is due to standard Bernstein's inequality on the indicator random variable on (s,a,s'), *i.e.*, $\mathbb{I}\left[(s_{h,k}, a_{h,k}, s_{h+1,k}) = (s,a,s')\right]$. Eq. (9) is due to uniform empirical Bernstein applied to $\mathbb{E}_{s_{h+1},r_h}\left[V^\star_{h+1}(s_{h+1}, b - r_h)\right]$. In Appendix B, we derive and review these uniform results. □

Lemma G.1. *For any $h \in [H]$ and $s \in \mathcal{S}$, $V^\star_h(s,\cdot)$ is 1-Lipschitz.*

Proof. We proceed by induction. Let $b, b' \in [0,1]$ be arbitrary. At $h = H+1$, $\left|V^\star_{H+1}(s,b) - V^\star_{H+1}(s,b')\right| = |b^+ - (b')^+| \le |b - b'|$ since the ReLU is 1-Lipschitz. For the inductive step, fix any h and suppose the claim is true at $h+1$. Then $|V^\star_h(s,b) - V^\star_h(s,b')| = \left|\min_a \mathbb{E}_{s_{h+1},r_h}\left[V^\star_{h+1}(s_{h+1}, b - r_h)\right] - \min_a \mathbb{E}_{s_{h+1},r_h}\left[V^\star_{h+1}(s_{h+1}, b' - r_h)\right]\right| \le \max_a \left|\mathbb{E}_{s_{h+1},r_h}\left[V^\star_{h+1}(s_{h+1}, b - r_h) - V^\star_{h+1}(s_{h+1}, b' - r_h)\right]\right| \le |b - b'|$, by the IH. The expectations are over $s_{h+1} \sim P^\star(s,a)$ and $r_h \sim R(s,a)$. □

We now show that the projected error between $\widehat{P}_k(s,a)$ and $P^\star$ can be bounded in two ways.

Lemma G.2. *For any $\delta \in (0,1)$, w.p. at least $1-\delta$, we have for all $f : \mathcal{S} \to [0,1]$,*

$$\left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top f\right| \le \min\left\{8\sqrt{\frac{SL}{N_k(s,a)}}, \frac{\mathbb{E}_{s'\sim P^\star(s,a)}[f(s')]}{H} + \xi_k(s,a)\right\},$$

$$\text{where } \xi_k(s,a) := \min\left\{1, \frac{2HSL}{N_k(s,a)}\right\}.$$

Proof. Fix any $f : \mathcal{S} \to [0,1]$. The first bound of $\sqrt{\frac{SL}{N_k(s,a)}}$ follows from applying Hoeffding on an ε-net of the space of f's, *i.e.*, for each g in the net, we have $\left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top g\right| \le \sqrt{\frac{L}{N_k(s,a)}}$. This ε-net has ℓ_2 bounded by $\sqrt{S}$. This gives the metric entropy $\log(1 + 2\sqrt{S}/\varepsilon)^S \approx S\log(S/\varepsilon)$. Setting $\varepsilon = \frac{1}{HK}$ makes the error lower order, *i.e.*, $\frac{1}{HK} \le \frac{1}{N_k(s,a)}$, which gives the uniform result over all f's. The detailed proof is in (Agarwal et al., 2021, Lemma 7.2).

The second bound also appears in Agarwal et al. (2021) as Lemma 7.8. We prove its proof for completeness:

$$\begin{aligned}
\left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top f\right| &\le \sum_{s'}\left|\widehat{P}_k(s' \mid s,a) - P^\star(s' \mid s,a)\right| f(s') \\
&\le \sum_{s'} f(s')\sqrt{\frac{2P^\star(s' \mid s,a)L}{N_k(s,a)}} + \frac{f(s')L}{N_k(s,a)} && \text{(Eq. (10))} \\
&\le \sqrt{S\frac{\sum_{s'} 2P^\star(s' \mid s,a)f^2(s')L}{N_k(s,a)}} + \frac{SL}{N_k(s,a)} && \text{(C-S)} \\
&\le \frac{SHL}{N_k(s,a)} + \frac{\sum_{s'} P^\star(s' \mid s,a)f(s')}{H} + \frac{SL}{N_k(s,a)}. && \text{(AM-GM)}
\end{aligned}$$

Finally, since $\widehat{P}_k(s,a)^\top f$ and $P^\star(s,a)^\top f$ are both in $[0,1]$, a trivial bound is 1, which is why $\xi_{h,k}$ can be truncated. □

Finally, we also have consequences of Azuma's inequality Lemma G.11. W.p. at least $1-\delta$, for all $h \in [H]$,

$$\sum_{k=1}^K \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}[2\text{B}\textsc{on}_{h,k}(s_h,b_h,a_h) + \xi_{h,k}(s_h,a_h) \mid \mathcal{E}_k] \le 6L + 2\sum_{k=1}^K 2\text{B}\textsc{on}_{h,k}(s_{h,k},b_{h,k},a_{h,k}) + \xi_{h,k}(s_{h,k},a_{h,k}), \tag{Azuma 1}$$

where we used the fact that WLOG we truncated the bonus to be at most 1 (by sentence below Eq. (B$\textsc{on}\bigstar$)), so $\|2\text{B}\textsc{on}_{h,k} + \xi_{h,k}\|_\infty \le 3$. $\mathcal{E}_k$ denotes the complete trajectories from episodes $1, 2, ..., k-1$

For the Bernstein bonus proofs, we'll also need,

$$\begin{aligned}
&\sum_{h=1}^H\sum_{k=1}^K \mathbb{E}_{s'\sim P^\star(s_{h,k},a_{h,k}),r\sim R(s_{h,k},a_{h,k})}\left[\left(\widehat{V}^\uparrow_{h+1,k}(s',b_{h,k}-r) - \widehat{V}^\downarrow_{h+1,k}(s',b_{h,k}-r)\right)^2 \mid \mathcal{E}_k, \mathcal{H}_{h,k}\right] \\
&\le \sqrt{HKL} + \sum_{h=1}^H\sum_{k=1}^K\left(\widehat{V}^\uparrow_{h+1,k}(s_{h+1,k},b_{h+1,k}) - \widehat{V}^\downarrow_{h+1,k}(s_{h+1,k},b_{h+1,k})\right)^2, && \text{(Azuma 2)}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{h=1}^H\sum_{k=1}^K \mathbb{E}_{s'\sim P^\star(s_{h,k},a_{h,k}),r\sim R(s_{h,k},a_{h,k})}\left[\left(V^{\widehat{\rho}^k}_{h+1}(s',b_{h,k}-r) - \widehat{V}^\downarrow_{h+1,k}(s',b_{h,k}-r)\right)^2 \mid \mathcal{E}_k, \mathcal{H}_{h,k}\right] \\
&\le \sqrt{HKL} + \sum_{h=1}^H\sum_{k=1}^K\left(V^{\widehat{\rho}^k}_{h+1}(s_{h+1,k},b_{h+1,k}) - \widehat{V}^\downarrow_{h+1,k}(s_{h+1,k},b_{h+1,k})\right)^2, && \text{(Azuma 3)}
\end{aligned}$$

where we've used that the envelope is at most 1 and $b_{h+1,k} = b_{h,k} - r_{h,k}$. Here, $\mathcal{H}_{h,k} = (s_{t,k}, a_{t,k}, r_{t,k})_{t\in[h-1]}$ denotes the history before h for the k-th episode. Also, for all $h \in [H]$,

$$\begin{aligned}
&\sum_{k=1}^K \text{Var}_{s'\sim P^\star(s_{h,k},a_{h,k})}\left(\mathbb{E}_{r\sim R(s_{h,k},a_{h,k})}\left[V^{\widehat{\rho}^k}_{h+1}(s',b_{h,k}-r)\right]\right) \\
&\le 2L + 2\sum_{k=1}^K \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\text{Var}_{s'\sim P^\star(s_h,a_h)}\left(\mathbb{E}_{r\sim R(s_h,a_h)}\left[V^{\widehat{\rho}^k}_{h+1}(s',b_{h,k}-r)\right]\right) \mid \mathcal{E}_k\right]. && \text{(Azuma 4)}
\end{aligned}$$

Also, for all $h, t \in [H]$ where $t \ge h$,

$$\sum_{k=1}^K \mathbb{E}_{\widehat{\rho}^k, s_h=s_{h,k}, b_h=b_{h,k}}\left[2\text{B}\textsc{on}^{\text{B}\textsc{ern}}_{t,k}(s_t,b_t,a_t) + \xi_{t,k}(s_t,a_t) \mid \mathcal{E}_k\right] \le 6L + 2\sum_{k=1}^K 2\text{B}\textsc{on}^{\text{B}\textsc{ern}}_{t,k}(s_{t,k},b_{t,k},a_{t,k}) + \xi_{t,k}(s_{t,k},a_{t,k}). \tag{Azuma 5}$$

Finally a standard Azuma also gives, for all $h \in [H]$,

$$\sum_{k=1}^{K} \mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\left[2\text{BON}_{h,k}^{\text{BERN}}(s_h, b_h, a_h) + \xi_{h,k}(s_h, a_h) \mid \mathcal{E}_k\right] \le 3\sqrt{KL} + \sum_{k=1}^{K} 2\text{BON}_{h,k}^{\text{BERN}}(s_{h,k}, b_{h,k}, a_{h,k}) + \xi_{h,k}(s_{h,k}, a_{h,k}) \tag{Azuma 6}$$

Henceforth, we always condition on the union of these high probability statements to be true.

G.2. Key lemmas for CVaR-UCBVI

In general, the bonus should be designed to satisfy for all $h \in [H], k \in [K]$,

$$\forall s, b, a : \left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top \mathbb{E}_{r_h \sim R(s,a)}\left[V_{h+1}^\star(\cdot, b - r_h)\right]\right| \le \text{BON}_{h,k}(s,b,a). \tag{BON★}$$

The bonus only needs to satisfy this inequality for our proofs to work. WLOG, since the left hand side is the difference between two numbers in $[0,1]$, we can always assume bonus to be truncated by 1, i.e. has envelope 1.

We say that pessimism is satisfied at $h \in [H], k \in [K]$ if

$$\forall s, b : \widehat{V}_{h,k}^{\downarrow}(s,b) \le V_h^\star(s,b). \tag{Pessimism ($V^\downarrow$)}$$

Lemma G.3. *For any $k \in [K], h \in [H]$, suppose Pessimism ($V^\downarrow$) holds at $(h+1,k)$ and BON★ holds at (h,k). Then Pessimism ($V^\downarrow$) holds at (h,k).*

Proof. First, we prove pessimism for $\widehat{U}_{h,k}^{\downarrow}$. For any s, b, a, we have

$$\begin{aligned}
&\widehat{U}_{h,k}^{\downarrow}(s,b,a) - U_h^\star(s,b,a) \\
&= \widehat{P}_k(s,a)^\top \mathbb{E}_{r_h \sim R(s,a)}\left[\widehat{V}_{h+1,k}^{\downarrow}(\cdot, b - r_h)\right] - \text{BON}_{h,k}(s,b,a) - P^\star(s,a)^\top \mathbb{E}_{r_h \sim R(s,a)}\left[V_{h+1}^\star(\cdot, b - r_h)\right] \\
&\le \left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top \mathbb{E}_{r_h \sim R(s,a)}\left[V_{h+1}^\star(\cdot, b - r_h)\right] - \text{BON}_{h,k}(s,b,a) && \text{(IH)} \\
&\le 0. && \text{by BON★}
\end{aligned}$$

To complete the proof, if $\widehat{V}_{h,k}^{\downarrow}(s,b) = 0$, it is trivially pessimistic, and if not,

$$\begin{aligned}
\widehat{V}_{h,k}^{\downarrow}(s,b) - V_h^\star(s,b) &= \min_a\left\{\widehat{U}_{h,k}^{\downarrow}(s,b,a)\right\} - \min_a\{U_h^\star(s,b,a)\} \\
&\le \max_a\left\{\widehat{U}_{h,k}^{\downarrow}(s,b,a) - U_h^\star(s,b,a)\right\} \\
&\le 0.
\end{aligned}$$

□

Remarkably, we show Simulation lemma also holds for CVaR-UCBVI. Here, it is also required that the bonus satisfies BON★. As for notation, recall $\mathcal{E}_k$ represents the episodes before and not including k.

Lemma G.4 (Simulation Lemma)**.** *Fix any $k \in [K], t \in [H]$. Then, for all s_t, b_t, we have*

$$\begin{aligned}
&V_t^{\widehat{\rho}^k}(s_t, b_t) - \widehat{V}_{t,k}^{\downarrow}(s_t, b_t) \\
&\le \sum_{h=t}^{H} \mathbb{E}_{\widehat{\rho}^k, s_t, b_t}\left[\text{BON}_{h,k}(s_h, b_h, a_h) + \left(P^\star(s_h, a_h) - \widehat{P}_k(s_h, a_h)\right)^\top \widehat{V}_{h+1,k}^{\downarrow}(\cdot, b_{h+1}) \mid \mathcal{E}_k\right].
\end{aligned} \tag{11}$$

Furthermore, if we assume that BON★ holds, then for all s_t, b_t,

$$V_t^{\widehat{\rho}^k}(s_t, b_t) - \widehat{V}_{t,k}^{\downarrow}(s_t, b_t) \le \sum_{h=t}^{H} (1 + 1/H)^{h-t} \mathbb{E}_{\widehat{\rho}^k, s_t, b_t}[2\text{BON}_{h,k}(s_h, b_h, a_h) + \xi_{h,k}(s_h, a_h) \mid \mathcal{E}_k]. \tag{12}$$

In particular,

$$V_1^{\widehat{\rho}^k}(s_1,b) - \widehat{V}^{\downarrow}_{1,k}(s_1,b) \le e\sum_{h=1}^{H} \mathbb{E}_{\widehat{\rho}^k}[2\textsc{Bon}_{h,k}(s_h,b_h,a_h) + \xi_{h,k}(s_h,a_h) \mid \mathcal{E}_k].$$

Proof. Fix any k and t. All expectations in the proof will condition on $\mathcal{E}_k$; this way, the randomness is only from rolling in the policy $\widehat{\rho}^k$ and not over any of the prior episodes.

First claim: Let's first show Eq. (11) by induction. The base case is $t = H+1$, we have $V_{H+1}^{\widehat{\rho}^k}(s,b) = \widehat{V}^{\downarrow}_{H+1,k}(s,b) = b^+$, so $V_{H+1}^{\widehat{\rho}^k} - \widehat{V}_{H+1,k} = 0$.

For the inductive step, fix any $t \le H$ and suppose Eq. (11) is true for $t+1$. Let us denote $a_t = \widehat{\rho}_t^k(s_t,b_t) = \arg\min_a \widehat{U}_{t,k}(s_t,b_t,a)$, so $\widehat{V}^{\downarrow}_{t,k}(s_t,b_t) = \max\left\{\widehat{U}^{\downarrow}_{t,k}(s_t,b_t,a_t),0\right\} \ge \widehat{U}^{\downarrow}_{t,k}(s_t,b_t,a_t) \ge \widehat{P}_{t,k}(s_t,a_t)^\top \mathbb{E}_{r_t}\left[\widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] - \textsc{Bon}_{t,k}(s_t,b_t,a_t)$, where $b_{t+1} = b_t - r_t$ is the random next budget. So, we have

$$\begin{aligned}
&V_t^{\widehat{\rho}^k}(s_t,b_t) - \widehat{V}^{\downarrow}_{t,k}(s_t,b_t)\\
&\le U_t^{\widehat{\rho}^k}(s_t,b_t,a_t) - \widehat{U}^{\downarrow}_{t,k}(s_t,b_t,a_t)\\
&= \textsc{Bon}_{t,k}(s_t,b_t,a_t) - \widehat{P}_{t,k}(s_t,a_t)^\top \mathbb{E}_{r_t}\left[\widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] + P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[V_{t+1,k}^{\widehat{\rho}^k}(\cdot,b_{t+1})\right]\\
&= \textsc{Bon}_{t,k}(s_t,b_t,a_t) + \left(P_t^\star(s_t,a_t) - \widehat{P}_{t,k}(s_t,a_t)\right)^\top \mathbb{E}_{r_t}\left[\widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right]\\
&+ P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[V_{t+1}^{\widehat{\rho}^k}(\cdot,b_{t+1}) - \widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right]\\
&\le \textsc{Bon}_{t,k}(s_t,b_t,a_t) + \left(P_t^\star(s_t,a_t) - \widehat{P}_{t,k}(s_t,a_t)\right)^\top \mathbb{E}_{r_t}\left[\widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right]\\
&+ \mathbb{E}_{s_{t+1}\sim P_t^\star(s_t,a_t)}\left[\sum_{h=t+1}^{H} \mathbb{E}_{\widehat{\rho}^k,s_{t+1},b_{t+1}}\left[\textsc{Bon}_{h,k}(s_h,b_h,a_h) + \left(P^\star(s_h,a_h) - \widehat{P}_k(s_h,a_h)\right)^\top \widehat{V}^{\downarrow}_{h+1,k}(\cdot,b_{h+1})\right]\right] \qquad \text{(IH)}\\
&= \sum_{h=t}^{H} \mathbb{E}_{\widehat{\rho}^k,s_t,b_t}\left[\textsc{Bon}_{h,k}(s_h,b_h,a_h) + \left(P^\star(s_h,a_h) - \widehat{P}_k(s_h,a_h)\right)^\top \widehat{V}^{\downarrow}_{h+1,k}(\cdot,b_{h+1})\right].
\end{aligned}$$

This concludes the proof for the first claim.

Second claim: Now let us show Eq. (12) by induction. The base case at $t = H+1$ is same as the first claim. For the inductive step, fix any $t \le H$ and suppose Eq. (12) is true for $t+1$. Then, continuing from the line before invoking the IH

of the first claim, we have

$$
\begin{aligned}
&V_t^{\widehat{\rho}^k}(s_t,b_t) - \widehat{V}^{\downarrow}_{t,k}(s_t,b_t) \\
&\le \textsc{Bon}_{t,k}(s_t,b_t,a_t) + \left(P_t^\star(s_t,a_t) - \widehat{P}_{t,k}(s_t,a_t)\right)^\top \mathbb{E}_{r_t}\left[\widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] \\
&+ P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[\left(V_{t+1}^{\widehat{\rho}^k}(\cdot,b_{t+1}) - \widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right)\right] \\
&= \textsc{Bon}_{t,k}(s_t,b_t,a_t) + \left(P_t^\star(s_t,a_t) - \widehat{P}_{t,k}(s_t,a_t)\right)^\top \mathbb{E}_{r_t}\left[\widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1}) - V^\star_{t+1}(\cdot,b_{t+1})\right] \\
&+ \left(P_t^\star(s_t,a_t) - \widehat{P}_{t,k}(s_t,a_t)\right)^\top \mathbb{E}_{r_t}\left[V^\star_{t+1}(\cdot,b_{t+1})\right] + P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[V_{t+1}^{\widehat{\rho}^k}(\cdot,b_{t+1}) - \widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] \\
&\le \textsc{Bon}_{t,k}(s_t,b_t,a_t) + \xi_{t,k}(s_t,a_t) + \frac{1}{H} P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[V^\star_{t+1}(\cdot,b_{t+1}) - \widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] && \text{(Lemma G.2)} \\
&+ \textsc{Bon}_{t,k}(s_t,b_t,a_t) + P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[V_{t+1}^{\widehat{\rho}^k}(\cdot,b_{t+1}) - \widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] && \text{(premise (}\textsc{Bon}\bigstar\text{))} \\
&\le 2\textsc{Bon}_{t,k}(s_t,b_t,a_t) + \xi_{t,k}(s_t,a_t) + (1+1/H) P_t^\star(s_t,a_t)^\top \mathbb{E}_{r_t}\left[V_{t+1}^{\widehat{\rho}^k}(\cdot,b_{t+1}) - \widehat{V}^{\downarrow}_{t+1,k}(\cdot,b_{t+1})\right] && (V^\star \le V^{\widehat{\rho}^k}) \\
&\le 2\textsc{Bon}_{t,k}(s_t,b_t,a_t) + \xi_{t,k}(s_t,a_t) \\
&+ (1+1/H)\sum_{h=t+1}^{H} (1+1/H)^{h-t-1} \mathbb{E}_{\widehat{\rho}^k,s_t,b_t}[2\textsc{Bon}_{h,k}(s_h,b_h,a_h) + \xi_{h,k}(s_h,a_h)]. && \text{(IH)}
\end{aligned}
$$

This completes the inductive proof. Observing that $(1+1/H)^H \le \exp(1/H)^H = e$ gives the corollary. □

G.3. CVaR-UCBVI with Hoeffding Bonus

The Hoeffding bonus $\textsc{Bon}^{\textsc{Hoeff}}_{h,k}(s,a)$ defined in 5 satisfies the crucial bonus requirement $\textsc{Bon}\bigstar$ by the uniform Hoeffding's inequality result of Eq. (8). Thus, we have pessimism for all k, h with the Hoeffding bonus.

Theorem 5.2. *For any $\delta \in (0,1)$, w.p. at least $1-\delta$,* CVaR-UCBVI *with the Hoeffding bonus (Eq. (5)) enjoys*

$$
\mathrm{Regret}^{\mathrm{RL}}_\tau(K) \le 4e\tau^{-1}\sqrt{SAHKL} + 10e\tau^{-1}S^2AHL^2.
$$

Proof of Theorem 5.2. Let $R(\rho^k, \widehat{b}_k)$ denote the distribution of returns from rolling in $\widehat{\rho}^k$ starting at $\widehat{b}_k$. For any k, we have

$$
\begin{aligned}
\mathrm{CVaR}_\tau(R(\widehat{\rho}^k,\widehat{b}_k)) &= \max_{b\in[0,1]}\left\{b - \tau^{-1}\mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\left(b - \sum_{t\in[H]} r_t\right)^+\right]\right\} \\
&\ge \widehat{b}_k - \tau^{-1}\mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\left(\widehat{b}_k - \sum_{t\in[H]} r_t\right)^+\right] \\
&= \widehat{b}_k - \tau^{-1} V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k). && (13)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathrm{Regret}_\tau^{\mathrm{RL}}(K) \\
&= \sum_{k=1}^{K} \mathrm{CVaR}_\tau^\star - \mathrm{CVaR}_\tau(R(\widehat{\rho}^k, \widehat{b}_k)) \\
&= \sum_{k=1}^{K} \left\{ b^\star - \tau^{-1} V_1^\star(s_1, b^\star) \right\} - \mathrm{CVaR}_\tau(R(\widehat{\rho}^k, \widehat{b}_k)) \\
&\le \sum_{k=1}^{K} \left\{ b^\star - \tau^{-1} \widehat{V}_{1,k}^{\downarrow}(s_1, b^\star) \right\} - \mathrm{CVaR}_\tau(R(\widehat{\rho}^k, \widehat{b}_k)) && \text{(Pessimism } (V^\downarrow)\text{)} \\
&\le \sum_{k=1}^{K} \left\{ \widehat{b}_k - \tau^{-1} \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k) \right\} - \left\{ \widehat{b}_k - \tau^{-1} V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k) \right\} && \text{(defn. of } \widehat{b}_k \text{ and Eq. (13))} \\
&= \tau^{-1} \sum_{k=1}^{K} \left(V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k) - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k) \right) \\
&\le e\tau^{-1} \sum_{(h,k) \in [H] \times [K]} \mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k} \left[2\mathrm{B}\mathrm{ON}_{h,k}^{\mathrm{HOEFF}}(s_h, a_h) + \xi_{h,k}(s_h, a_h) \mid \mathcal{E}_k \right] && \text{(Simulation Lemma G.4)} \\
&\le 6e\tau^{-1} HL + 2e\tau^{-1} \sum_{(h,k) \in [H] \times [K]} 2\mathrm{B}\mathrm{ON}_{h,k}^{\mathrm{HOEFF}}(s_{h,k}, a_{h,k}) + \xi_{h,k}(s_{h,k}, a_{h,k}) && \text{(Eq. (Azuma 1))} \\
&\le 6e\tau^{-1} HL + 2e\tau^{-1} \left(2\sqrt{L} \cdot \sqrt{SAHKL} + 2HSL \cdot SA \log(K) \right) && \text{(elliptical potential Lemma G.12)} \\
&\le 4e\tau^{-1} \sqrt{SAHK} L + 10e\tau^{-1} S^2 AHL^2,
\end{aligned}
$$

which concludes the proof. □

G.4. CVaR-UCBVI with Bernstein Bonus

When running Algorithm 2 with the Bernstein bonus $\mathrm{B}\mathrm{ON}_{h,k}^{\mathrm{BERN}}(s, b, a)$ (Eq. (6)), we need to also show that $\widehat{V}_{h,k}^{\uparrow}$ are optimistic for $V_h^\star$. We say that optimism is satisfied at $(h, k) \in [H] \times [K]$ if

$$
\forall s, b : V_h^\star(s, b) \le \widehat{V}_{h,k}^{\uparrow}(s, b). \qquad \text{(Optimism } (V^\uparrow)\text{)}
$$

Lemma G.5. *For any $k \in [K], h \in [H]$, suppose Optimism $(V^\uparrow)$ holds at $(h+1, k)$ and* BON★ *holds at (h, k). Then Optimism $(V^\uparrow)$ holds at (h, k).*

Proof. First, we prove optimism for $\widehat{U}_{h,k}^{\uparrow}$. For any s, b, a we have

$$
\begin{aligned}
&U_h^\star(s, b, a) - \widehat{U}_{h,k}^{\uparrow}(s, b, a) \\
&= P^\star(s, a)^\top \mathbb{E}_{r_h} \left[V_{h+1}^\star(\cdot, b - r_h) \right] - \widehat{P}_k(s, a)^\top \mathbb{E}_{r_h} \left[\widehat{V}_{h+1,k}^{\uparrow}(\cdot, b - r_h) \right] - \mathrm{B}\mathrm{ON}_{h,k}(s, b, a) \\
&\le \left(P^\star(s, a) - \widehat{P}_k(s, a) \right)^\top \mathbb{E}_{r_h} \left[V_{h+1}^\star(\cdot, b - r_h) \right] - \mathrm{B}\mathrm{ON}_{h,k}(s, b, a) && \text{(IH)} \\
&\le 0. && \text{by Eq. (B}\mathrm{ON}\text{★)}
\end{aligned}
$$

To complete the proof, if $\widehat{V}^\uparrow(s, b) = 1$, then it is trivially optimistic, and if not

$$
\begin{aligned}
V_h^\star(s, b) - \widehat{V}_{h,k}^{\uparrow}(s, b) &= \min_a U_h^\star(s, b, a) - \widehat{U}_{h,k}^{\uparrow}(s, b, \widehat{\rho}^k(s, b)) \\
&\le U_h^\star(s, b, \widehat{\rho}^k(s, b)) - \widehat{U}_{h,k}^{\uparrow}(s, b, \widehat{\rho}^k(s, b)) \\
&\le 0.
\end{aligned}
$$

□

Lemma G.6. *For any $(h,k) \in [H] \times [K]$, if Pessimism ($V^{\downarrow}$) and Optimism ($V^{\uparrow}$) both hold at $(h+1,k)$, then* BON★ *holds at (h,k) for the Bernstein bonus* BON^{BERN}.

Proof. Recall that uniform empirical Bernstein Eq. (9) gave us the following inequality: for all s, a and b,

$$\left|\left(\widehat{P}_k(s,a) - P^\star(s,a)\right)^\top \mathbb{E}_{r_h}\left[V^\star_{h+1}(\cdot, b - r_h)\right]\right| \le \sqrt{\frac{2\operatorname{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[V^\star_{h+1}(s', b-r_h)\right]\right)L}{N_k(s,a)}} + \frac{L}{N_k(s,a)}.$$

Eq. (9) revisited.

Now apply the useful triangle inequality of variances $\sqrt{\operatorname{Var}(X)} \le \sqrt{\operatorname{Var}(Y)} + \sqrt{\operatorname{Var}(X-Y)}$ (Zanette & Brunskill, 2019, Eqn. 51),

$$\begin{aligned}\sqrt{\operatorname{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[V^\star_{h+1}(s', b-r_h)\right]\right)} &\le \sqrt{\operatorname{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[\widehat{V}^{\downarrow}_{h+1}(s', b-r_h)\right]\right)} \\ &\quad + \sqrt{\operatorname{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[V^\star_{h+1}(s', b-r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b-r_h)\right]\right)}.\end{aligned}$$

The first term is in the bonus. The second term is bounded by the correction term of the bonus as follows,

$$\begin{aligned}&\operatorname{Var}_{s'\sim\widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h}\left[V^\star_{h+1}(s', b-r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b-r_h)\right]\right) \\ &\le \mathbb{E}_{s'\sim\widehat{P}_k(s,a)}\left[\left(\mathbb{E}_{r_h}\left[V^\star_{h+1}(s', b-r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b-r_h)\right]\right)^2\right] \\ &\le \mathbb{E}_{s'\sim\widehat{P}_k(s,a), r\sim R(s,a)}\left[\left(V^\star_{h+1}(s', b-r) - \widehat{V}^{\downarrow}_{h+1,k}(s', b-r)\right)^2\right] && \text{(Jensen)} \\ &\le \mathbb{E}_{s'\sim\widehat{P}_k(s,a), r\sim R(s,a)}\left[\left(\widehat{V}^{\uparrow}_{h+1,k}(s', b-r) - \widehat{V}^{\downarrow}_{h+1,k}(s', b-r)\right)^2\right]. && \text{(premise: } \widehat{V}^{\downarrow}_{h+1,k} \le V^\star_{h+1} \le \widehat{V}^{\uparrow}_{h+1,k}\text{)}\end{aligned}$$

This upper bound is a part of the Bernstein bonus. Thus, we've shown that BON^{BERN} dominates the error, and so BON★ is satisfied at (h,k). □

A key corollary of Lemmas G.3, G.5 and G.6 is that we have Pessimism ($V^{\downarrow}$) and Optimism ($V^{\uparrow}$) *for all* $(h,k) \in [H]\times[K]$ with the Bernstein bonus. Indeed, for any k, we first apply Lemma G.6 to get that BON★ is satisfied at (H,k) (as optimism/pessimism trivially holds at $H+1$). Then, apply Lemmas G.3 and G.5 to get Pessimism ($V^{\downarrow}$) and Optimism ($V^{\uparrow}$) at (H,k). Then, apply Lemma G.6 to get that BON★ is satisfied at $(H-1,k)$. Continue in this fashion until we've shown BON★, Pessimism ($V^{\downarrow}$) and Optimism ($V^{\uparrow}$) for all $h \in [H]$.

Theorem G.7. *The Bernstein bonus satisfies* BON★, *Pessimism ($V^{\downarrow}$) and Optimism ($V^{\uparrow}$) at all $(h,k) \in [H]\times[K]$.*

We now prove the regret guarantee for the Bernstein bonus. The main body of the proof for Theorem 5.3 and Theorem 5.5 will be the same. The proofs will only diverge at the end when bounding the sum of variances, where we invoke Assumption 5.4.

Theorem 5.3. *For any $\delta \in (0,1)$, w.p. at least $1-\delta$,* CVaR-UCBVI *with the Bernstein bonus (Eq. (6)) enjoys a regret guarantee of*

$$\text{Regret}^{\text{RL}}_\tau(K) \le 10e\tau^{-1}\sqrt{SAKL} + \tau^{-1}\xi,$$

where $\xi \in \widetilde{\mathcal{O}}(SAHK^{1/4} + S^2AH)$ is a lower order term.

Theorem 5.5. *Under Assumption 5.4, the bound in Theorem 5.3 can be refined to,*

$$\text{Regret}^{\text{RL}}_\tau(K) \le 12e\sqrt{\tau^{-1}SAKL} + \tau^{-1}p_{\min}^{-1/2}\xi.$$

Proof of Theorem 5.3 and Theorem 5.5. Following the same initial steps as Theorem 5.2, we have

$$\text{Regret}^{\text{RL}}_\tau(K) \le 6e\tau^{-1}HL + 2e\tau^{-1}\sum_{(h,k)\in[H]\times[K]} 2\text{BON}^{\text{BERN}}_{h,k}(s_{h,k}, b_{h,k}, a_{h,k}) + \xi_{h,k}(s_{h,k}, a_{h,k}).$$

The proof boils down to bounding the sum.

Logarithmic-in-K terms: First, note that any $\mathcal{O}(1/N_k(s,a))$ term will scale logarithmically in K. This includes the $\xi_{h,k}(s,a)$ term, as well as a $\frac{2L}{N_k(s,a)}$ from the bonus. Thus,

$$\sum_{(h,k)\in[H]\times[K]} \frac{2L+2HSL}{N_k(s_{h,k},a_{h,k})} \le 4HSL\cdot SA\log(K) = 4S^2AHL^2. \qquad \text{(Lemma G.12)}$$

The variance correction term of the bonus:

$$\sum_{(h,k)\in[H]\times[K]} \sqrt{\frac{\mathbb{E}_{s'\sim\widehat{P}_k(s_{h,k},a_{h,k}),r\sim R(s_{h,k},a_{h,k})}\left[\left(\widehat{V}^{\uparrow}_{h+1,k}(s',b_{h,k}-r)-\widehat{V}^{\downarrow}_{h+1,k}(s',b_{h,k}-r)\right)^2\right]}{N_k(s_{h,k},a_{h,k})}}. \qquad (14)$$

First, apply a Cauchy Schwarz. The $\sum_{h,k}\frac{1}{N_k(s_{h,k},a_{h,k})}$ term is at most SAL by elliptical potential Lemma G.12. Then, translate $\widehat{P}$ to $P^\star$ via Lemma G.2 to get

$$\le \sqrt{SAL\left(\sum_{(h,k)\in[H]\times[K]} 8\sqrt{\frac{SL}{N_k(s_{h,k},a_{h,k})}} + \sum_{(h,k)\in[H]\times[K]} \mathbb{E}_{s'\sim P^\star(s_{h,k},a_{h,k}),r\sim R(s_{h,k},a_{h,k})}\left[\left(\widehat{V}^{\uparrow}_{h+1,k}(s',b_{h,k}-r)-\widehat{V}^{\downarrow}_{h+1,k}(s',b_{h,k}-r)\mid\mathcal{E}_k,\mathcal{H}_{h,k}\right)^2\right]\right)}$$

Then, switch to the empirical s,b using Eq. (Azuma 2),

$$\le \sqrt{SAL\left(8\sqrt{SL}\sqrt{SAHKL}+\sqrt{HKL}+\sum_{(h,k)\in[H]\times[K]}\left(\widehat{V}^{\uparrow}_{h+1,k}(s_{h+1,k},b_{h+1,k})-\widehat{V}^{\downarrow}_{h+1,k}(s_{h+1,k},b_{h+1,k})\right)^2\right)}.$$

For the sum term, since each term is at most 1, we have $\left(\widehat{V}^{\uparrow}_{h,k}(s_{h,k},b_{h,k})-\widehat{V}^{\downarrow}_{h,k}(s_{h,k},b_{h,k})\right)^2 \le \left(\widehat{V}^{\uparrow}_{h,k}(s_{h,k},b_{h,k})-\widehat{V}^{\downarrow}_{h,k}(s_{h,k},b_{h,k})\right)$. Then, applying a simulation-like Lemma G.8 to each summand bounds the sum by,

$$\begin{aligned}
&\sum_{h=2}^{H}\sum_{k=1}^{K}\sum_{t=h}^{H}\mathbb{E}_{\widehat{\rho}^k,s_h=s_{h,k},b_h=b_{h,k}}\left[2\text{B}\text{ON}^{\text{BERN}}_{t,k}(s_t,b_t,a_t)+\xi_{t,k}(s_t,a_t)\mid\mathcal{E}_k\right] && \text{(simulation-like Lemma G.8)}\\
&\le 6H^2L+2\sum_{h=2}^{H}\sum_{t=h}^{H}\sum_{k=1}^{K}2\text{B}\text{ON}^{\text{BERN}}_{t,k}(s_{t,k},b_{t,k},a_{t,k})+\xi_{t,k}(s_t,a_t) && \text{(Eq. (Azuma 5))}\\
&\le 6H^2L+2H\sum_{t=1}^{H}\sum_{k=1}^{K}2\text{B}\text{ON}^{\text{BERN}}_{t,k}(s_{t,k},b_{t,k},a_{t,k})+\xi_{t,k}(s_t,a_t).
\end{aligned}$$

Now, we can loosely bound each Bernstein bonus by $2\sqrt{\frac{2L}{N_k(s,a)}}+\frac{L}{N_k(s,a)}$, so by elliptical potential Lemma G.12,

$$\begin{aligned}
\sum_{t=1}^{H}\sum_{k=1}^{K}2\text{B}\text{ON}^{\text{BERN}}_{t,k}(s_{t,k},b_{t,k},a_{t,k})+\xi_{t,k}(s_t,a_t) &\le 4\sqrt{2L}\cdot\sqrt{SAHKL}+(L+2HSL)\cdot SA\log(K)\\
&\le 6\sqrt{SAHK}L+3S^2AHL^2. \qquad (15)
\end{aligned}$$

Therefore, we've shown that

$$\sum_{h=2}^{H}\sum_{k=1}^{K}\sum_{t=h}^{H}\mathbb{E}_{\widehat{\rho}^k,s_h=s_{h,k},b_h=b_{h,k}}\left[2\text{B}\text{ON}^{\text{BERN}}_{t,k}(s_t,b_t,a_t)+\xi_{t,k}(s_t,a_t)\mid\mathcal{E}_k\right]\le 12\sqrt{SAH^3K}L+12S^2AH^2L^2. \qquad (16)$$

Combining everything together, we have

$$\begin{aligned}\text{Eq. (14)} &\le \sqrt{SAL\Big(9S\sqrt{AHK}L + 12\sqrt{SAH^3K}L + 12S^2AH^2L^2\Big)} \\ &\le 5SAHK^{1/4}L + 4S^{3/2}AHL^2,\end{aligned}$$

which is lower order in K (the dominant term should scale as $K^{1/2}$).

Bounding the empirical variance term: We now shift our focus to the variance term of the bonus,

$$\sum_{(h,k)\in[H]\times[K]} \sqrt{\frac{\mathrm{Var}_{s'\sim\widehat{P}_k(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r\sim R(s_{h,k},a_{h,k})}\Big[\widehat{V}^{\downarrow}_{h+1,k}(s', b_{h,k}-r)\Big]\Big)}{N_k(s_{h,k},a_{h,k})}}. \tag{17}$$

The key idea here is to apply a sequential Law of Total Variance Lemma G.13, but to do so, we must first convert $\sqrt{\mathrm{Var}_{s'\sim\widehat{P}_k(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[\widehat{V}^{\downarrow}_{h+1,k}(s', b_{h,k}-r_h)\Big]\Big)}$ to $\sqrt{\mathrm{Var}_{s'\sim P^\star_h(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[V^{\widehat{\rho}^k}_{h+1}(s', b_{h,k}-r_h)\Big] \mid \mathcal{E}_k\Big)}$. So we need to bound the difference term, *i.e.*,

$$\sum_{(h,k)\in[H-1]\times[K]} \sqrt{\frac{\mathrm{Var}_{s'\sim\widehat{P}_k(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[\widehat{V}^{\downarrow}_{h+1,k}(s', b_{h,k}-r_h)\Big]\Big)}{N_k(s_{h,k},a_{h,k})}} - \sqrt{\frac{\mathrm{Var}_{s'\sim P^\star_h(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[V^{\widehat{\rho}^k}_{h+1}(s', b_{h,k}-r_h)\Big] \mid \mathcal{E}_k\Big)}{N_k(s_{h,k},a_{h,k})}} \tag{18}$$

Switch the empirical variance to the (conditional) population one, which incurs a $\sum_{h,k} 2\sqrt{\frac{L}{N_k(s_{h,k},a_{h,k})}}$ term (Appendix B). Then, use $\sqrt{\mathrm{Var}(X+Y)} \le \sqrt{\mathrm{Var}(X)} + \sqrt{\mathrm{Var}(Y)}$ (Eqn 51 of (Zanette & Brunskill, 2019)) to get,

$$\begin{aligned}&\le \sum_{(h,k)\in[H-1]\times[K]} \frac{2\sqrt{L}}{N_k(s_{h,k},a_{h,k})} + \sum_{(h,k)\in[H-1]\times[K]} \sqrt{\frac{\mathrm{Var}_{s'\sim P^\star(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[\widehat{V}^{\downarrow}_{h+1,k}(s', b_{h+1,k}) - V^{\widehat{\rho}^k}_{h+1}(s', b_{h+1,k})\Big] \mid \mathcal{E}_k\Big)}{N_k(s_{h,k},a_{h,k})}} \\ &\le 2\sqrt{L}\cdot SA\log(K) + \sqrt{SAL \sum_{(h,k)\in[H-1]\times[K]} \mathrm{Var}_{s'\sim P^\star(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[V^{\widehat{\rho}^k}_{h+1}(s', b_{h,k}-r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b_{h,k}-r_h)\Big] \mid \mathcal{E}_k\Big)},\end{aligned}$$

where the second inequality is due to elliptical potential Lemma G.12 and Cauchy-Schwarz. Focusing on the sum inside the square root,

$$\begin{aligned}&\sum_{(h,k)\in[H-1]\times[K]} \mathrm{Var}_{s'\sim P^\star(s_{h,k},a_{h,k})}\Big(\mathbb{E}_{r_h}\Big[V^{\widehat{\rho}^k}_{h+1}(s', b_{h,k}-r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b_{h,k}-r_h)\Big] \mid \mathcal{E}_k\Big) \\ &\le \sum_{(h,k)\in[H-1]\times[K]} \mathbb{E}_{s',r_h\sim(P^\star\circ R)(s_{h,k},a_{h,k})}\Big[\Big(V^{\widehat{\rho}^k}_{h+1}(s', b_{h,k}-r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b_{h,k}-r_h)\Big)^2 \mid \mathcal{E}_k\Big] && \text{(Jensen)} \\ &\le \sqrt{HKL} + \sum_{(h,k)\in[H-1]\times[K]} \Big(V^{\widehat{\rho}^k}_{h+1}(s_{h+1,k}, b_{h+1,k}) - \widehat{V}^{\downarrow}_{h+1,k}(s_{h+1,k}, b_{h+1,k})\Big)^2 && \text{(Eq. (Azuma 3))} \\ &\le \sqrt{HKL} + \sum_{(h,k)\in[H-1]\times[K]} \Big(V^{\widehat{\rho}^k}_{h+1}(s_{h+1,k}, b_{h+1,k}) - \widehat{V}^{\downarrow}_{h+1,k}(s_{h+1,k}, b_{h+1,k})\Big) && \text{(r.v. is in } [0,1]) \\ &\le \sqrt{HKL} + \sum_{(h,k)\in[H-1]\times[K]} \sum_{t=h}^{H} \mathbb{E}_{\widehat{\rho}^k, s_h=s_{h,k}, b_h=b_{h,k}}\big[2\mathrm{Bon}^{\mathrm{Bern}}_{t,k}(s_t,b_t,a_t) + \xi_{t,k}(s_t,a_t)\big] \\ &&& \text{(simulation lemma Lemma G.4)} \\ &\le \sqrt{HKL} + 12\sqrt{SAH^3K}L + 12S^2AH^2L^2. && \text{(Eq. (16))}\end{aligned}$$

Combining the steps, we have shown that the switching cost to the population variance is at most

$$\text{Eq. (18)} \le 2SAL^2 + \sqrt{SAL\Big(13\sqrt{SAH^3KL} + 12S^2AH^2L^2\Big)} \le 4SAHK^{1/4}L + 6S^{3/2}AHL^2$$

which is again lower order.

(Key step) Bounding the dominant (population) variance term:

$$\sum_{(h,k)\in[H]\times[K]} \sqrt{\frac{\mathrm{Var}_{s'\sim P_h^\star(s_{h,k},a_{h,k})}\left(\mathbb{E}_{r\sim R(s_{h,k},a_{h,k})}\left[V_{h+1}^{\widehat{\rho}^k}(s',b_{h,k}-r)\right] \mid \mathcal{E}_k\right)}{N_k(s_{h,k},a_{h,k})}}, \tag{19}$$

First apply a Cauchy-Schwarz (as usual) and then the law of total variance,

$$\leq \sqrt{SAL \sum_{(h,k)\in[H]\times[K]} \mathrm{Var}_{s'\sim P_h^\star(s_{h,k},a_{h,k})}\left(\mathbb{E}_{r\sim R(s_{h,k},a_{h,k})}\left[V_{h+1}^{\widehat{\rho}^k}(s',b_{h,k}-r)\right]\right)}$$

$$\leq \sqrt{SAL\left(2HL + 2\sum_{(h,k)\in[H]\times[K]} \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\mathrm{Var}_{s'\sim P_h^\star(s_h,a_h)}\left(\mathbb{E}_{r\sim R(s_h,a_h)}\left[V_{h+1}^{\widehat{\rho}^k}(s',b_h-r)\right]\right) \mid \mathcal{E}_k\right]\right)}$$

(Eq. (Azuma 4))

$$\leq \sqrt{SAL\left(2HL + 2\sum_{(h,k)\in[H]\times[K]} \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\mathrm{Var}_{s'\sim P_h^\star(s_h,a_h),r\sim R(s_h,a_h)}\left(V_{h+1}^{\widehat{\rho}^k}(s',b_h-r)\right)\right] \mid \mathcal{E}_k\right)}$$

(joint variance is larger)

$$= \sqrt{SAL\left(2HL + 2\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right)\right)}, \qquad \text{(Law of Total Variance Lemma G.9)}$$

Below, we give two ways to bound the sum,

$$\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right). \tag{20}$$

The first way is to simply bound it by a probability, which is trivially at most 1. This results in Theorem 5.3. To prove Theorem 5.5, we show a second more refined approach, which uses Assumption 5.4 to bound each variance by τ, plus a lower order term.

Before doing so, we first prepare to conclude the proof by recapping all the terms in the regret decomposition. First, we collected $4S^2AHL^2$ from the $1/N_k(s,a)$ terms. Eq. (14) is the correction term inside the Bernstein bonus. Eq. (18) is the switching cost from empirical variance (in the Bernstein bonus) to the population variance that we want to bound now, which is Eq. (19). We also multiply back the $2\sqrt{2L} \leq 3\sqrt{L}$ factor we omitted from above. So,

$$\begin{aligned}
&4S^2AHL^2 + 3\sqrt{L}(\text{Eq. (14)} + \text{Eq. (18)} + \text{Eq. (19)}) \\
&\leq 4S^2AHL^2 + 3\sqrt{L}\Big(\left(5SAHK^{1/4}L + 4S^{3/2}AHL^2\right) + \left(4SAHK^{1/4}L + 6S^{3/2}AHL^2\right) + \text{Eq. (19)}\Big) \\
&= 27SAHK^{1/4}L^2 + 34S^2AHL^3 + 3\sqrt{L}\cdot\text{Eq. (19)}.
\end{aligned}$$

Thus, the regret is at most

$$\begin{aligned}
\mathrm{Regret}_\tau^{\mathrm{RL}}(K) &\leq 6e\tau^{-1}HL + 2e\tau^{-1}\Big(27SAHK^{1/4}L^2 + 34S^2AHL^3 + 3\sqrt{L}\cdot\text{Eq. (19)}\Big) \\
&\leq 6e\tau^{-1}\sqrt{L}\cdot\text{Eq. (19)} + 54e\tau^{-1}SAHK^{1/4}L^2 + 70e\tau^{-1}S^2AHL^3.
\end{aligned}$$

First bound for Eq. (19) (for Theorem 5.3): Since $x^+ = x\mathbb{I}[x \geq 0]$, for any random variable $X \in [0,1]$, we have $\mathrm{Var}(X^+) \leq \mathbb{E}\left[X^2\mathbb{I}[X \geq 0]\right] \leq \Pr(X \geq 0)$. Therefore,

$$\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right) \leq \sum_{k=1}^{K} \Pr_{\widehat{\rho}^k,\widehat{b}_k}\left(\sum_{h=1}^{H} r_h \leq \widehat{b}_k \mid \mathcal{E}_k\right) \leq K$$

Certainly, each probability is bounded by 1. Hence,

$$\text{Eq. (19)} \le \sqrt{SAL(2HL+2K)} \le \sqrt{2SAKL} + \sqrt{2SAH}L.$$

Combining everything together, we get

$$\operatorname{Regret}_\tau^{\mathrm{RL}}(K) \le 6\sqrt{2}e\tau^{-1}\sqrt{SAK}L + 54e\tau^{-1}SAHK^{1/4}L^2 + 82e\tau^{-1}S^2AHL^3,$$

which concludes the proof for Theorem 5.3.

Second bound for Eq. (19) (for Theorem 5.5): Define

$$f(b) = b - \tau^{-1}\mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\left(b - \sum_{t=1}^H r_t\right)^+ \mid \mathcal{E}_k\right], \qquad b_k^\star = \arg\max_{b\in[0,1]} f(b),$$
$$\widehat{f}(b) = b - \tau^{-1}\widehat{V}_{1,k}^{\downarrow}(s_1,b), \qquad \widehat{b}_k = \arg\max_{b\in[0,1]} \widehat{f}(b).$$

So, $b_k^\star$ as the τ-th quantile of $R(\widehat{\rho}^k,\widehat{b}_k)$. By pessimism, we have $\widehat{V}_{1,k}^{\downarrow}(s_1,b) \le V_1^\star(s_1,b) \le \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[\left(b-\sum_{t=1}^H r_t\right)^+\right]$, since $V_1^\star$ is the minimum amongst all history-dependent policies, including $(\widehat{\rho}^k,\widehat{b}_k)$. Thus, we have $\widehat{f}(b) \ge f(b)$ for all $b\in[0,1]$. In particular, we have

$$\begin{aligned}
f(b_k^\star) - f(\widehat{b}_k) &= f(b_k^\star) - \widehat{f}(\widehat{b}_k) + \widehat{f}(\widehat{b}_k) - f(\widehat{b}_k) \\
&\le f(b_k^\star) - \widehat{f}(b_k^\star) + \widehat{f}(\widehat{b}_k) - f(\widehat{b}_k) && (\widehat{b}_k \text{ is argmax of } \widehat{f}) \\
&\le \widehat{f}(\widehat{b}_k) - f(\widehat{b}_k) && (f(b_k^\star) \le \widehat{f}(b_k^\star) \text{ by pessimism}) \\
&\le \tau^{-1}\left(V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k) - \widehat{V}_{1,k}^{\downarrow}(s_1,\widehat{b}_k)\right).
\end{aligned}$$

Now we invoke Assumption 5.4 with Lemma G.10 which implies

$$\begin{aligned}
&\frac{p_{\min}}{2\tau}\left(b_k^\star - \widehat{b}_k\right)^2 \le f(b_k^\star) - f(\widehat{b}_k) \le \tau^{-1}\left(V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k) - \widehat{V}_{1,k}^{\downarrow}(s_1,\widehat{b}_k)\right) \\
&\implies \left(b_k^\star - \widehat{b}_k\right)^2 \le 2p_{\min}^{-1}\left(V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k) - \widehat{V}_{1,k}^{\downarrow}(s_1,\widehat{b}_k)\right).
\end{aligned}$$

Using $\operatorname{Var}(X) \le 2(\operatorname{Var}(Y) + \operatorname{Var}(X-Y))$,

$$\begin{aligned}
\text{Eq. (20)} &= \sum_{k=1}^K \operatorname{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{t=1}^H r_t\right)^+ \mid \mathcal{E}_k\right) \\
&\le 2\sum_{k=1}^K \operatorname{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left(\left(b_k^\star - \sum_{t=1}^H r_t\right)^+ \mid \mathcal{E}_k\right) + 2\sum_{k=1}^K \operatorname{Var}_{\widehat{\rho}^k,\widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{t=1}^H r_t\right)^+ - \left(b_k^\star - \sum_{t=1}^H r_t\right)^+ \mid \mathcal{E}_k\right) \\
&\le 2\sum_{k=1}^K \Pr_{\widehat{\rho}^k,\widehat{b}_k}\left(\sum_{t=1}^H r_t \le b_k^\star \mid \mathcal{E}_k\right) + 2\sum_{k=1}^K\left(\widehat{b}_k - b_k^\star\right)^2 && (\text{ReLU is 1-Lipschitz}) \\
&\le 2K\tau + 4p_{\min}^{-1}\sum_{k=1}^K\left(V_1^{\widehat{\rho}^k}(s_1,\widehat{b}_k) - \widehat{V}_{1,k}^{\downarrow}(s_1,\widehat{b}_k)\right) \\
&\le 2K\tau + 4p_{\min}^{-1}\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\widehat{\rho}^k,\widehat{b}_k}\left[2\textsc{Bon}_{h,k}^{\textsc{Bern}}(s_h,b_h,a_h) + \xi_{h,k}(s_h,a_h) \mid \mathcal{E}_k\right] && (\text{simulation Lemma G.4}) \\
&\le 2K\tau + 4p_{\min}^{-1}\left(3\sqrt{KL} + 6\sqrt{SAHKL} + 3S^2AHL^2\right). && (\text{Eq. (Azuma 6) and the loose bound in Eq. (15)})
\end{aligned}$$

Therefore,

$$\text{Eq. (19)} \leq \sqrt{SAL\Big(2HL + 4K\tau + 72p_{\min}^{-1}\sqrt{SAHKL}L + 24p_{\min}^{-1}S^2AHL^2\Big)}$$
$$\leq 2\sqrt{\tau SAKL} + 9p_{\min}^{-1/2}SAHK^{1/4}L + 5p_{\min}^{-1/2}S^{3/2}AHL^2.$$

Combining everything together, we get

$$\text{Regret}_\tau^{\text{RL}}(K) \leq 12e\sqrt{\tau^{-1}SAKL} + \Big(54 + 9p_{\min}^{-1/2}\Big)e\tau^{-1}SAHK^{1/4}L^2 + \Big(70 + 5p_{\min}^{-1/2}\Big)e\tau^{-1}S^2AHL^3.$$

This concludes the proof for Theorem 5.5.

□

Lemma G.8. *Fix any $k \in [K]$. Then for all $t \in [H]$, for all s_t, b_t, we have*

$$\widehat{V}_{t,k}^{\uparrow}(s_t, b_t) - \widehat{V}_{t,k}^{\downarrow}(s_t, b_t) \leq \sum_{h=t}^{H}(1 - 1/H)^{t-h}\mathbb{E}_{\widehat{\rho}^k, s_t, b_t}[2\text{Bon}_{h,k}(s_h, b_h, a_h) + \xi_{h,k}(s_h, a_h) \mid \mathcal{E}_k].$$

Proof. In this proof, all expectations are conditioned on $\mathcal{E}_k$. We proceed by induction. The base case at $t = H+1$ trivially holds since $\widehat{V}_{H+1,k}^{\uparrow}(s, b) - \widehat{V}_{H+1,k}^{\downarrow}(s, b) = b^+ - b^+ = 0$. Now suppose $t \leq H$ and suppose the claim holds for $t+1$. Then setting $a_t = \widehat{\rho}^k(s_t, b_t)$, we have

$$\begin{aligned}
&\widehat{V}_{t,k}^{\uparrow}(s_t, b_t) - \widehat{V}_{t,k}^{\downarrow}(s_t, b_t) \\
&\leq \widehat{U}_{t,k}^{\uparrow}(s_t, b_t, a_t) - \widehat{U}_{t,k}^{\downarrow}(s_t, b_t, a_t) \\
&= \widehat{P}_{t,k}(s_t, a_t)^\top \mathbb{E}_{r_t}\Big[\widehat{V}_{t+1,k}^{\uparrow}(\cdot, b_t - r_t) - \widehat{V}_{t+1,k}^{\downarrow}(\cdot, b_t - r_t)\Big] + 2\text{Bon}_{t,k}(s_t, b_t, a_t) \\
&= \Big(\widehat{P}_{t,k}(s_t, a_t) - P_t^{\star}(s_t, a_t)\Big)^\top \mathbb{E}_{r_t}\Big[\widehat{V}_{t+1,k}^{\uparrow}(\cdot, b_t - r_t) - \widehat{V}_{t+1,k}^{\downarrow}(\cdot, b_t - r_t)\Big] + 2\text{Bon}_{t,k}(s_t, b_t, a_t) \\
&+ P_t^{\star}(s_t, a_t)^\top \mathbb{E}_{r_t}\Big[\widehat{V}_{t+1,k}^{\uparrow}(\cdot, b_t - r_t) - \widehat{V}_{t+1,k}^{\downarrow}(\cdot, b_t - r_t)\Big] \\
&\leq 2\text{Bon}_{t,k}(s_t, b_t, a_t) + \xi_{t,k}(s_t, a_t) + (1 - 1/H)P_t^{\star}(s_t, a_t)^\top \mathbb{E}_{r_t}\Big[\widehat{V}_{t+1,k}^{\uparrow}(\cdot, b_t - r_t) - \widehat{V}_{t+1,k}^{\downarrow}(\cdot, b_t - r_t)\Big] \\
&\hspace{30em} \text{(Lemma G.2)} \\
&\leq 2\text{Bon}_{t,k}(s_t, b_t, a_t) + \xi_{t,k}(s_t, a_t) + \sum_{h=t+1}^{H}(1 - 1/H)^{1+t-(h+1)}\mathbb{E}_{\widehat{\rho}^k, s_t, b_t}[2\text{Bon}_{h,k}(s_h, b_h, a_h) + \xi_{h,k}(s_h, a_h)],
\end{aligned}$$

where $\mathbb{E}_{r_t}$ is short for $\mathbb{E}_{r_t \sim R_t(s_t, a_t)}$. □

Lemma G.9. *For any $k \in [K]$, we have*

$$\text{Var}_{\widehat{\rho}^k, \widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{h=1}^{H} r_h\right)^+ \mid \mathcal{E}_k\right) = \sum_{h=1}^{H-1}\mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\Big[\text{Var}_{s' \sim P^\star(s_h, a_h), r_h \sim R(s,a)}\Big(V_{h+1}^{\widehat{\rho}^k}(s', b_h - r_h)\Big) \mid \mathcal{E}_k\Big].$$

Proof. Apply Law of Total Variance Lemma G.13 with $Y = \Big(\widehat{b}_k - \sum_{h=1}^{H} r_h\Big)^+$, $X_h = (s_h, a_h, r_{h-1})$ for $h \in [H]$ (when $h = 1$, r_0 is omitted), and $\mathcal{H} = \mathcal{E}_k$ being the trajectories from the past episodes $1, 2, ..., k-1$. Here, s_h, a_h, r_h are collected from rolling in with $\widehat{\rho}^k$ starting from $\widehat{b}_k$, and note that $\widehat{b}_k$ is a constant conditioned on $\mathcal{E}_k$. Lemma G.13 gives

$$\begin{aligned}
\text{Var}_{\widehat{\rho}^k, \widehat{b}_k}\left(\left(\widehat{b}_k - \sum_{h=1}^{H} r_h\right)^+ \mid \mathcal{E}_k\right) &= \mathbb{E}[\text{Var}(Y \mid X_{1:H}, \mathcal{E}_k) \mid \mathcal{E}_k] + \sum_{h=1}^{H}\mathbb{E}[\text{Var}(\mathbb{E}[Y \mid X_{1:h}, \mathcal{E}_k] \mid X_{1:h-1}, \mathcal{E}_k) \mid \mathcal{E}_k] \\
&= \sum_{h=1}^{H}\mathbb{E}[\text{Var}(\mathbb{E}[Y \mid X_{1:h}, \mathcal{E}_k] \mid X_{1:h-1}, \mathcal{E}_k) \mid \mathcal{E}_k].
\end{aligned}$$

The first term is zero because once we condition on $X_{1:H}, \mathcal{E}_k$, the term $\left(\widehat{b}_k - \sum_t r_t\right)^+$ is a constant, and variance of constants is zero. Now consider each summand. The outer expectation is taken over $s_{1:h-1}, a_{1:h-1}, r_{1:h-2}$ from rolling in $\widehat{\rho}^k$. The variance is taken over $s_h \sim P^\star(s_{h-1}, a_{h-1}), r_{h-1} \sim R_{h-1}(s_{h-1}, a_{h-1})$, and deterministically picking $a_h = \widehat{\rho}^k_h(s_h, b_h)$ where $b_h = \widehat{b}_k - \sum_{t=1}^{h-1} r_t$. The inner expectation is over the remainder of the trajectory, which is $s_{h+1:H}, a_{h+1:H}, r_{h:H}$. Therefore,

$$
\begin{aligned}
&\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X_{1:h}, \mathcal{E}_k] \mid X_{1:h-1}, \mathcal{E}_k) \mid \mathcal{E}_k] \\
&= \mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\left[\mathrm{Var}\left(\mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\left[\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid X_{1:h}, \mathcal{E}_k\right] \mid X_{1:h-1}, \mathcal{E}_k\right) \mid \mathcal{E}_k\right] \\
&= \mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\left[\mathrm{Var}\left(U_h^{\widehat{\rho}^k}(s_h, b_h, a_h) \mid X_{1:h-1}, \mathcal{E}_k\right) \mid \mathcal{E}_k\right] && (b_{h-1} = \widehat{b}_k - r_1 - \ldots - r_{h-1}) \\
&= \mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\left[\mathrm{Var}_{s_h \sim P^\star_{h-1}(s_{h-1}, a_{h-1}), r_{h-1} \sim R_{h-1}(s_{h-1}, a_{h-1})}\left(U_h^{\widehat{\rho}^k}(s_h, b_h, a_h)\right) \mid \mathcal{E}_k\right] \\
&= \mathbb{E}_{\widehat{\rho}^k, \widehat{b}_k}\left[\mathrm{Var}_{s_h \sim P^\star_{h-1}(s_{h-1}, a_{h-1}), r_{h-1} \sim R_{h-1}(s_{h-1}, a_{h-1})}\left(V_h^{\widehat{\rho}^k}(s_h, b_h)\right) \mid \mathcal{E}_k\right]. && (a_h = \widehat{\rho}^k(s_h, b_h))
\end{aligned}
$$

Note that in the special case of $h = 1$, we have s_1 is not random and r_0 is omitted. So, the variance is taken over a constant, which is zero. □

Lemma G.10. *Let π be a history-dependent policy such that its return distribution $R(\pi)$ is continuously distributed and has a density p lower bounded by $p_{\min}$. Then, the function*

$$
f(b) = \tau^{-1}\mathbb{E}_\pi\left[\left(b - \sum_h r_h\right)^+\right] - b
$$

is $\tau^{-1}p_{\min}$ strongly convex.

Proof. Observe that

$$
f'(b) = \tau^{-1}\Pr_\pi\left(\sum_h r_h \le b\right) - 1.
$$

Since we've assumed that $R(\pi)$ is continuously distributed with density p, so

$$
f''(b) = \tau^{-1}p(b).
$$

Since $f''(b) \ge \tau^{-1}p_{\min}$ for all b, we have f'' is strongly convex with that parameter. □

G.5. Auxiliary Lemmas

Lemma G.11 (Azuma)**.** *Let $\{X_i\}_{i\in[N]}$ be a sequence of random variables supported on $[0,1]$, adapted to filtration $\{\mathcal{F}_i\}_{i\in[N]}$. For any $\delta \in (0,1)$, we have w.p. at least $1-\delta$,*

$$
\sum_{t=1}^{N}\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \le \sum_{t=1}^{N} X_t + \sqrt{N\log(2/\delta)}, \qquad \text{(Standard Azuma)}
$$

$$
\sum_{t=1}^{N}\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \le 2\sum_{t=1}^{N} X_t + 2\log(1/\delta). \qquad \text{(Multiplicative Azuma)}
$$

Proof. For standard Azuma, see Zhang (2023, Theorem 13.4). For multiplicative Azuma, apply (Zhang, 2023, Theorem 13.5) with $\lambda = 1$. The claim follows, since $\frac{1}{1-\exp(-\lambda)} \le 2$. □

Below we recall the standard elliptical potential lemma (Lattimore & Szepesvári, 2020; Agarwal et al., 2021). Regarding terminology, we remark that this lemma is also known as the "pigeonhole argument" in the tabular RL literature (Azar et al., 2017; Zanette & Brunskill, 2019). The term "elliptical potential" is more commonly used in the linear MDP setting (Jin et al., 2020), of which tabular RL is a special case.

Lemma G.12 (Elliptical Potential)**.** *For any sequence of states and actions* $\{s_{h,k}, a_{h,k}\}_{h\in[H],k\in[K]}$*, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_k(s_{h,k},a_{h,k})} \le SA\log(K),$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{N_k(s_{h,k},a_{h,k})}} \le \sqrt{HSAK\log(K)}.$$

Proof. For the first claim, observe that in the sum, $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \dots$ can appear at most SA times. And since we run for K episodes, the maximum denominator is K. Therefore, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_k(s_{h,k},a_{h,k})} \le SA\sum_{k=1}^{K}\frac{1}{k} \le SA\log(K).$$

For the second claim,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{N_k(s_{h,k},a_{h,k})}} &\le \sqrt{KH\left(\sum_{k,h=1}^{K,H}\frac{1}{N_k(s_{h,k},a_{h,k})}\right)}\\ &\le \sqrt{SAHK\log(K)}.\end{aligned}$$

□

Lemma G.13 (Sequential Law of Total Conditional Variance)**.** *For any random variables* $Y, X_1, X_2, ..., X_N, \mathcal{H}$*, we have*

$$\mathrm{Var}(Y \mid \mathcal{H}) = \mathbb{E}[\mathrm{Var}(Y \mid X_{1:N}, \mathcal{H}) \mid \mathcal{H}] + \sum_{t=1}^{N}\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X_{1:t}, \mathcal{H}] \mid X_{1:t-1}, \mathcal{H}) \mid \mathcal{H}].$$

Notice, for each summand, the inner expectation is taken over Y*, the variance is taken over* X_t*, and outer expectation is taken over* $X_{1:t-1}$*.*

Proof. Recall the Law of Total Conditional Variance (LTCV): for any random variables Y, Z_1, Z_2,

$$\mathrm{Var}(Y \mid Z_1) = \mathbb{E}[\mathrm{Var}(Y \mid Z_1, Z_2) \mid Z_1] + \mathrm{Var}(\mathbb{E}[Y \mid Z_1, Z_2] \mid Z_1).$$

We now inductively prove the desired claim by recursively applying the (LTCV). The base case is $N = 1$, which follows immediately from LTCV applied to $Z_1 = \mathcal{H}, Z_2 = X_1$. For the inductive case, fix any N and suppose the claim is true for N. Now consider $N + 1$, where we have $Y, X_1, X_2, ..., X_{N+1}, \mathcal{H}$. By the IH, we have

$$\mathrm{Var}(Y \mid \mathcal{H}) = \mathbb{E}[\mathrm{Var}(Y \mid X_{1:N}, \mathcal{H}) \mid \mathcal{H}] + \sum_{t=1}^{N}\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X_{1:t}, \mathcal{H}] \mid X_{1:t-1}, \mathcal{H}) \mid \mathcal{H}].$$

Now applying LTCV on the first term with $Z_1 = (X_{1:N}, \mathcal{H}), Z_2 = X_{N+1}$, we have

$$\mathrm{Var}(Y \mid X_{1:N}, \mathcal{H}) = \mathbb{E}[\mathrm{Var}(Y \mid X_{1:N+1}, \mathcal{H}) \mid X_{1:N}, \mathcal{H}] + \mathrm{Var}(\mathbb{E}[Y \mid X_{1:N+1}, \mathcal{H}] \mid X_{1:N}, \mathcal{H}),$$

and therefore,

$$\mathbb{E}[\mathrm{Var}(Y \mid X_{1:N}, \mathcal{H}) \mid \mathcal{H}] = \mathbb{E}[\mathrm{Var}(Y \mid X_{1:N+1}, \mathcal{H}) \mid \mathcal{H}] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X_{1:N+1}, \mathcal{H}] \mid X_{1:N}, \mathcal{H}) \mid \mathcal{H}],$$

which concludes the proof. □

H. Proofs for CVaR-UCBVI with discretized rewards

Recall the discretized MDP $\mathrm{disc}(\mathcal{M})$, as introduced in Section 6. It is a copy of the true MDP $\mathcal{M}$ except its rewards are rounded to an ε-net. *I.e.*, let $\phi(r) = \eta\lceil r/\eta \rceil$ be the rounding up operator of r onto the net, so that $0 \le \phi(r) - r \le \eta$. Concretely, the reward distribution is $R(s,a;\mathrm{disc}(\mathcal{M})) = R(s,a;\mathcal{M}) \circ \phi^{-1}$. Our proofs are inspired by Bastani et al. (2022, Lemma B.1,Lemma B.5).

From $\mathrm{disc}(\mathcal{M})$ to $\mathcal{M}$: Fix any $\rho \in \Pi^{\mathrm{Aug}}$ and $b \in [0,1]$ (which we'll run in $\mathrm{disc}(\mathcal{M})$). Then define an adapted policy, which is a history-dependent policy in $\mathcal{M}$, as follows,

$$\mathrm{adapted}(\rho,b)_h(s_h, r_{1:h-1}) = \rho_h(s_h, b_1 - \phi(r_1) - \ldots - \phi(r_{h-1})).$$

Intuitively, as $\mathrm{adapted}(\rho,b)$ runs $\mathcal{M}$, it uses the history to emulate the evolution of b in $\mathrm{disc}(\mathcal{M})$. Let $Z_{\rho,b,\mathrm{disc}(\mathcal{M})}$ be the returns from running ρ, b in $\mathrm{disc}(\mathcal{M})$. Let $Z_{\mathrm{adapted}(\rho,b),\mathcal{M}}$ be the returns from running $\mathrm{adapted}(\rho,b)$ in $\mathcal{M}$. We show that $Z_{\rho,b,\mathrm{disc}(\mathcal{M})} - H\eta \le Z_{\mathrm{adapted}(\rho,b),\mathcal{M}} \le Z_{\rho,b,\mathrm{disc}(\mathcal{M})}$ w.p. 1 via a coupling argument.

Lemma H.1. *We almost surely have $Z_{\rho,b,\mathrm{disc}(\mathcal{M})} - H\eta \le Z_{\mathrm{adapted}(\rho,b),\mathcal{M}} \le Z_{\rho,b,\mathrm{disc}(\mathcal{M})}$. Therefore, if $F_{\rho,b,\mathrm{disc}(\mathcal{M})}$ is the CDF of $Z_{\rho,b,\mathrm{disc}(\mathcal{M})}$ and $F_{\mathrm{adapted}(\rho,b),\mathcal{M}}$ is the CDF of $Z_{\mathrm{adapted}(\rho,b),\mathcal{M}}$, we have*

$$\forall x \in \mathbb{R} : F_{\rho,b,\mathrm{disc}(\mathcal{M})}(x) \le F_{\mathrm{adapted}(\rho,b),\mathcal{M}}(x) \le F_{\rho,b,\mathrm{disc}(\mathcal{M})}(x + H\eta).$$

Proof. Let $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ be the trajectory of running $\mathrm{adapted}(\rho,b)$ in $\mathcal{M}$. Let $\widehat{s}_1, \widehat{a}_1, \widehat{r}_1, \widehat{s}_2, \widehat{a}_2, \widehat{r}_2, \ldots$ be the trajectory of running ρ, b in $\mathrm{disc}(\mathcal{M})$. We couple these two trajectories by making $\mathrm{adapted}(\rho,b)$ in $\mathcal{M}$ follow ρ, b in $\mathrm{disc}(\mathcal{M})$. Set $\widehat{s}_1 = s_1$. By definition of $\mathrm{adapted}(\rho,b)$, $\widehat{a}_1 = a_1$. By definition of $\mathrm{disc}(\mathcal{M})$, $\widehat{r}_1 = \phi(r_1)$. Continuing in this fashion, we have $\widehat{s}_t = s_t, \widehat{a}_t = a_t, \widehat{r}_t = \phi(r_t)$ for all $t \in [H]$. This is a valid coupling since the actions are sampled from the exact same distribution, *i.e.*, $a_h \sim \rho_h(b - \phi(r_1) - \ldots - \phi(r_{h-1}))$, and by the transitions of $\widehat{b}_h$ in $\mathrm{disc}(\mathcal{M})$, we have $\widehat{b}_h = b - \widehat{r}_1 - \ldots - \widehat{r}_{h-1}$ which is exactly what was inputted into ρ_h by $\mathrm{adapted}(\rho,b)$.

Since $r \le \phi(r)$ for all r, we have

$$Z_{\mathrm{adapted}(\rho,b),\mathcal{M}} = \sum_{t=1}^{H} r_t \le \sum_{t=1}^{H} \phi(r_t) = \sum_{t=1}^{H} \widehat{r}_t = Z_{\rho,b,\mathrm{disc}(\mathcal{M})}.$$

Since $\phi(r) - \eta \le r$ for all r, we have

$$Z_{\rho,b,\mathrm{disc}(\mathcal{M})} = \sum_{t=1}^{H} \phi(r_t) \le -H\eta + \sum_{t=1}^{H} r_t = Z_{\mathrm{adapted}(\rho,b),\mathcal{M}} - H\eta.$$

To conclude the proof, recall a basic fact about couplings and stochastic comparisons. For two random variables X, Y in the same probability space and a constant c, if $\mathbb{P}(X \le Y + c) = 1$, we have $F_Y(t) \le F_X(t + c)$ for all x. This is because $F_Y(x) - F_X(x + c) = \mathbb{P}(Y \le t \cap X > t + c) \le \mathbb{P}(X - Y > c) = 0$. □

From $\mathcal{M}$ to $\mathrm{disc}(\mathcal{M})$: Fix any $\rho \in \Pi^{\mathrm{Aug}}$ and $b \in [0,1]$ (which we'll run in $\mathcal{M}$). Then define a discretized policy, which is a history-dependent policy in the discretized MDP $\mathrm{disc}(\mathcal{M})$ *with memory* (as in Appendix F), as follows,

$$\mathrm{disc}(\rho,b)_h(s_h, m_{1:h-1}) = \rho_h(s_h, b - m_1 - \cdots - m_{h-1}),$$
$$m_h \sim R(s_h, a_h) \mid \phi(R(s_h, a_h)) = r_h.$$

With this definition, although we receive reward $\widehat{r}_h$ (on the discrete grid) when running in $\mathrm{disc}(\mathcal{M})$, the memory element m_h exactly imitates a random reward that would have been received in the true MDP $\mathcal{M}$. Then, the discretized policy $\mathrm{disc}(\rho,b)$ will instead follow these exact rewards m_h rather than what has been received.

Let $Z_{\rho,b,\mathcal{M}}$ be the returns from running ρ, b in $\mathcal{M}$. Let $Z_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}$ be the returns from running $\mathrm{disc}(\rho,b)$ in $\mathrm{disc}(\mathcal{M})$ (which memory as described above). We show that $Z_{\rho,b,\mathcal{M}} \le Z_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}$ w.p. 1 via a coupling argument.

Lemma H.2. *We almost surely have $Z_{\rho,b,\mathcal{M}} \le Z_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}$. Therefore, if $F_{\rho,b,\mathcal{M}}$ is the CDF of $Z_{\rho,b,\mathcal{M}}$ and $F_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}$ is the CDF of $Z_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}$, we have*

$$\forall x \in \mathbb{R} : F_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}(x) \le F_{\rho,b,\mathcal{M}}.$$

Proof. Let $s_1, a_1, r_1, s_2, a_2, r_2, ...$ be the trajectory of running ρ, b in $\mathcal{M}$. Let $\widehat{s}_1, \widehat{a}_1, \widehat{r}_1, \widehat{m}_1, \widehat{s}_2, \widehat{a}_2, \widehat{r}_2, \widehat{m}_2, ...$ be the trajectory of running $\mathrm{disc}(\rho, b)$ in $\mathrm{disc}(\mathcal{M})$ with memory. We couple these two trajectories by making ρ, b in $\mathcal{M}$ follow $\mathrm{disc}(\rho, b)$ in $\mathrm{disc}(\mathcal{M})$. Set $s_1 = \widehat{s}_1$. By definition of $\mathrm{disc}(\rho, b)$, $a_1 = \widehat{a}_1$. Then, set $r_1 = \widehat{m}_1$. Note that r_1 is sampled by first sampling a discrete $\widehat{r}_1$, then sampling $\widehat{m}_1$ from the conditional reward distribution of the interval that rounds to $\widehat{r}_1$. By law of total probability, this is indeed equivalent to sampling directly from the unconditional reward distribution. Continuing in this fashion, we have $\widehat{s}_t = s_t, \widehat{a}_t = a_t, \widehat{r}_t = \phi(r_t), \widehat{m}_t = m_t$ for all $t \in [H]$. Importantly, the policies actions match because $b - \widehat{m}_1 - ... - \widehat{m}_{h-1} = b - r_1 - ... - r_{h-1}$. Therefore, we always have

$$Z_{\rho,b,\mathcal{M}} = \sum_{t=1}^{H} r_t \leq \sum_{t=1}^{H} \phi(r_t) = \sum_{t=1}^{H} \widehat{r}_t = Z_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}.$$

As with the previous proof, stochastic dominance implies the claim on CDFs. □

Now we show two useful consequences of the above coupling results.

Theorem H.3. *Fix any* $\rho \in \Pi^{\mathrm{Aug}}$ *and* $b_1 \in [0, 1]$. *Then,*

$$\forall b : 0 \leq \mathbb{E}_{\mathrm{adapted}(\rho,b_1),\mathcal{M}}\left[\left(b - \sum_{h=1}^{H} r_h\right)^+\right] - \mathbb{E}_{\rho,b_1,\mathrm{disc}(\mathcal{M})}\left[\left(b - \sum_{h=1}^{H} r_h\right)^+\right] \leq H\eta,$$

and

$$-\tau^{-1} H\eta \leq \mathrm{CVaR}_\tau(\mathrm{adapted}(\rho, b_1), \mathcal{M}) - \mathrm{CVaR}_\tau(\rho, b, \mathrm{disc}(\mathcal{M})) \leq 0.$$

Proof. Let $f(b) = \mathbb{E}_{\mathrm{adapted}(\rho,b_1),\mathcal{M}}\left[\left(b - \sum_{h=1}^{H} r_h\right)^+\right]$ and let $F = F_{\mathrm{adapted}(\rho,b_1),\mathcal{M}}$. Similarly, let $\mathrm{disc}(f)(b) = \mathbb{E}_{\rho,b_1,\mathrm{disc}(\mathcal{M})}\left[\left(b - \sum_{h=1}^{H} r_h\right)^+\right]$ and $\mathrm{disc}(F) = F_{\rho,b_1,\mathrm{disc}(\mathcal{M})}$. Both $f(0) = \mathrm{disc}(f)(0) = 0$. Also, their derivatives are F and $\mathrm{disc}(F)$ respectively. By Lemma H.1, $\mathrm{disc}(F)(t) \leq F(t) \leq \mathrm{disc}(F)(t + H\eta)$.

First, we show $\mathrm{disc}(f)(b) - f(b) \leq 0$. By the fundamental theorem of Calculus,

$$\mathrm{disc}(f)(b) - f(b) = \int_0^b \mathrm{disc}(F)(t) - F(t)\mathrm{d}t \leq 0.$$

Next, we show $f(b) - \mathrm{disc}(f)(b) \leq H\eta$. By the fundamental theorem of Calculus (FTC),

$$\begin{aligned}
f(b) - \mathrm{disc}(f)(b) &\leq \int_0^b F(t)\mathrm{d}t - \int_0^b \mathrm{disc}(F)(t)\mathrm{d}t \\
&\leq \int_{H\eta}^{b+H\eta} \mathrm{disc}(F)(t)\mathrm{d}t - \int_0^b \mathrm{disc}(F)(t)\mathrm{d}t && \text{(Lemma H.1)} \\
&\leq \int_{b}^{b+H\eta} \mathrm{disc}(F)(t)\mathrm{d}t \\
&= \mathrm{disc}(f)(b + H\eta) - \mathrm{disc}(f)(b) && \text{(FTC)} \\
&\leq H\eta, && \text{(ReLU is 1-Lipschitz)}
\end{aligned}$$

Altogether, we've shown $0 \leq f(b) - \mathrm{disc}(f)(b) \leq H\eta$ for all b. This immediately implies the claims about CVaR:

$$\begin{aligned}
&\mathrm{CVaR}_\tau(\mathrm{adapted}(\rho, b_1), \mathcal{M}) - \mathrm{CVaR}_\tau(\rho, b_1, \mathrm{disc}(\mathcal{M})) \\
&= \max_b \left\{b - \tau^{-1} f(b)\right\} - \max_b \left\{b - \tau^{-1}\,\mathrm{disc}(f)(b)\right\} \\
&\leq \tau^{-1} \max_b (\mathrm{disc}(f)(b) - f(b)) \\
&\leq 0,
\end{aligned}$$

and similarly,

$$\begin{aligned}&\mathrm{CVaR}_\tau(\rho, b_1, \mathrm{disc}(\mathcal{M})) - \mathrm{CVaR}_\tau(\mathrm{adapted}(\rho, b_1), \mathcal{M})\\ &\le \tau^{-1} \max_b (f(b) - \mathrm{disc}(f)(b))\\ &\le \tau^{-1} H\eta.\end{aligned}$$

□

Theorem H.4. *We have,*

$$\forall b \in [0,1] : V_1^\star(s_1, b; \mathrm{disc}(\mathcal{M})) \le V_1^\star(s_1, b; \mathcal{M}),$$

which implies

$$\mathrm{CVaR}_\tau^\star(\mathrm{disc}(\mathcal{M})) \ge \mathrm{CVaR}_\tau^\star(\mathcal{M}).$$

Proof. Fix any $\rho \in \Pi^{\mathrm{Aug}}, b \in [0,1]$. By Lemma H.2, we have $F_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}(x) \le F_{\rho,b,\mathcal{M}}$. Applying the same FTC-style arguments in Theorem H.3, we have

$$\forall b' : \mathbb{E}_{\mathrm{disc}(\rho,b),\mathrm{disc}(\mathcal{M})}\left[\left(b' - \sum_{h=1}^H r_h\right)^+\right] \le \mathbb{E}_{\rho,b,\mathcal{M}}\left[\left(b' - \sum_{h=1}^H r_h\right)^+\right].$$

Setting $b' = b$ and using the definition of V functions, we have $V^{\mathrm{disc}(\rho,b)}(s_1, b; \mathrm{disc}(\mathcal{M})) \le V_1^\rho(s_1, b; \mathcal{M})$. Then, since $V_1^\star$ is the minimum history-dependent policy in this memory-MDP (Theorem F.2) implies that

$$V_1^\star(s_1, b; \mathrm{disc}(\mathcal{M})) \le V_1^{\mathrm{disc}(\rho,b)}(s_1, b; \mathrm{disc}(\mathcal{M})) \le V_1^\rho(s_1, b; \mathcal{M}).$$

Since $\rho \in \Pi^{\mathrm{Aug}}$ was arbitrary and the minimum is attained by $\rho^\star \in \Pi^{\mathrm{Aug}}$ (Theorem F.2) this implies that $V_1^\star(s_1, b; \mathrm{disc}(\mathcal{M})) \le V_1^\star(s_1, b; \mathcal{M})$, as needed. For the CVaR claim,

$$\begin{aligned}&\mathrm{CVaR}_\tau^\star(\mathcal{M}) - \mathrm{CVaR}_\tau^\star(\mathrm{disc}(\mathcal{M}))\\ &= \max_b\left\{b - \tau^{-1} V_1^\star(s_1, b; \mathcal{M})\right\} - \max_b\left\{b - \tau^{-1} V_1^\star(s_1, b; \mathrm{disc}(\mathcal{M}))\right\}\\ &\le \tau^{-1} \max_b (V_1^\star(s_1, b; \mathrm{disc}(\mathcal{M})) - V_1^\star(s_1, b; \mathcal{M})) \le 0.\end{aligned}$$

□

In the proof above, we highlight that $\mathrm{disc}(\mathcal{M})$ is the MDP with memory. In the proof of Bastani et al. (2022, Lemma B.6), this detail was glossed over as their "history-dependent policy" in Bastani et al. (2022, Lemma B.5) does not exactly fit into the vanilla history-dependent policy framework (as in Section 2); their policies are coupled through time with the α parameter, which is disallowed *a priori* by the history-dependent policy framework. Our formalism with the memory-MDP resolves this ambiguity.

H.1. Amendment of Bernstein proof in the discretized MDP

In this section, we amend the proof of the "Second bound for Eq. (19)" in the Bernstein regret bound.

Theorem H.5. *Suppose we're running* CVaR-UCBVI *in the discretized MDP and assume Assumption 6.1 holds. For any $\delta \in (0,1)$, w.p. at least δ, we have the regret of Theorem 5.5 plus an additional term,*

$$18e\tau^{-1}\sqrt{p_{\min}^{-1} SAHK\eta L}.$$

Thus, setting $\eta = 1/\sqrt{K}$ makes this an lower order term.

Proof of Theorem H.5. Recall that

$$\text{Eq. (19)} \le \sqrt{SAL\left(2HL + 2\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k, \widehat{b}_k, \mathrm{disc}(\mathcal{M})}\left(\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right)\right)}$$

where note the randomness is taken over trajectories from $\mathrm{disc}(\mathcal{M})$, as that's the MDP we're working in. Define

$$f(b) = b - \tau^{-1}\mathbb{E}_{\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k), \mathcal{M}}\left[\left(b - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right], \qquad b_k^\star = \arg\max_{b \in [0,1]} f(b),$$
$$\widehat{f}(b) = b - \tau^{-1}\widehat{V}_{1,k}^{\downarrow}(s_1, b), \qquad \widehat{b}_k = \arg\max_{b \in [0,1]} \widehat{f}(b).$$

A priori, the pessimism argument a priori only applies to policies in the discretized MDP. But thanks to Theorem H.4, it also holds here,

$$\widehat{V}_{1,k}^{\downarrow}(s_1, b) \le V_1^\star(s_1, b; \mathrm{disc}(\mathcal{M})) \le V_1^\star(s_1, b; \mathcal{M}) \le \mathbb{E}_{\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k), \mathcal{M}}\left[\left(b - \sum_{t=1}^{H} r_t\right)^+\right].$$

Thus, we also have $\widehat{f}(b) \ge f(b)$ for all b, and the same argument as before gives

$$\begin{aligned} f(b_k^\star) - f(\widehat{b}_k) &\le \tau^{-1}\left(\mathbb{E}_{\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k), \mathcal{M}}\left[\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right] - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k)\right) \\ &\le \tau^{-1}\left(V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k)\right) + \tau^{-1}H\eta. \qquad \text{(Theorem H.3)} \end{aligned}$$

By Assumption 6.1 (which applies to $\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k)$), Lemma G.10 applies and we have

$$\left(b_k^\star - \widehat{b}_k\right)^2 \le 2p_{\min}^{-1}\left(V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k) + H\eta\right).$$

Using $\mathrm{Var}(X) \le 2(\mathrm{Var}(Y) + \mathrm{Var}(X - Y))$,

$$\begin{aligned} &\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k, \widehat{b}_k, \mathrm{disc}(\mathcal{M})}\left(\left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right) \\ &\le 2\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k, \widehat{b}_k, \mathrm{disc}(\mathcal{M})}\left(\left(b_k^\star - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right) + 2\sum_{k=1}^{K} \mathrm{Var}_{\widehat{\rho}^k, \widehat{b}_k, \mathrm{disc}(\mathcal{M})}\left(\left(b_k^\star - \sum_{t=1}^{H} r_t\right)^+ - \left(\widehat{b}_k - \sum_{t=1}^{H} r_t\right)^+ \mid \mathcal{E}_k\right) \\ &\le 2\sum_{k=1}^{K} \Pr_{\widehat{\rho}^k, \widehat{b}_k, \mathrm{disc}(\mathcal{M})}\left(\sum_{t=1}^{H} r_t \le b_k^\star\right) + 2\sum_{k=1}^{K}\left(\widehat{b}_k - b_k^\star\right)^2 \qquad \text{(ReLU is 1-Lipschitz)} \\ &\le 2\sum_{k=1}^{K} \Pr_{\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k), \mathcal{M}}\left(\sum_{t=1}^{H} r_t \le b_k^\star\right) + 4p_{\min}^{-1}\sum_{k=1}^{K}\left(V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k) + H\eta\right) \qquad \text{(Lemma H.1)} \\ &\le 2K\tau + 4p_{\min}^{-1}HK\eta + 4p_{\min}^{-1}\sum_{k=1}^{K}\left(V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k)\right). \end{aligned}$$

Previously in Theorem 5.5, we also had the $2K\tau + 4p_{\min}^{-1}\sum_{k=1}^{K}\left(V_1^{\widehat{\rho}^k}(s_1, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) - \widehat{V}_{1,k}^{\downarrow}(s_1, \widehat{b}_k)\right)$ term. This can be bounded as before. We've only incurred an extra $4p_{\min}^{-1}HK\eta$ term. So, Eq. (19) incurs an extra

$$\sqrt{SAL(2 \cdot 4p_{\min}^{-1}HK\eta)}.$$

Finally, Eq. (19) gets multiplied by $6e\tau^{-1}\sqrt{L}$, which gives a final extra regret of $18e\tau^{-1}\sqrt{p_{\min}^{-1}SAHK\eta L}$. □

H.2. Translating regret from discretized MDP to true MDP

In this section, we prove that running our algorithms in the imagined discretized MDP also has low regret in the true MDP. Let us first recall the setup. When running CVaR-UCBVI, we will provide the discretization parameter η. The algorithm will discretize the rewards received from the environment when updating b – in this way, it is running in its own hallucinated discretized MDP, while the regret we care about is in the true MDP. Since the algorithm is essentially running in the hallucinated discretized MDP, our regret bound applies in the discretized MDP, *i.e.*, roll-outs in expectations are with respect to $\mathrm{disc}(\mathcal{M})$,

$$\sum_{k=1}^{K} \mathrm{CVaR}_\tau^\star(\mathrm{disc}(\mathcal{M})) - \mathrm{CVaR}_\tau(\widehat{\rho}^k, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) \le C.$$

When rolling out $\widehat{\rho}^k$ from $\widehat{b}_k$ in the hallucinated discretized MDP, we are essentially running $\mathrm{adapted}(\widehat{\rho}^k)$ as described in Theorem H.3. So the *true* regret in the real MDP is,

$$\begin{aligned}
&\sum_{k=1}^{K} \mathrm{CVaR}_\tau^\star(\mathcal{M}) - \mathrm{CVaR}_\tau(\mathrm{adapted}(\widehat{\rho}^k, \widehat{b}_k); \mathcal{M}) \\
&\le \sum_{k=1}^{K} \mathrm{CVaR}_\tau^\star(\mathrm{disc}(\mathcal{M})) - \mathrm{CVaR}_\tau(\widehat{\rho}^k, \widehat{b}_k; \mathrm{disc}(\mathcal{M})) + \tau^{-1} H\eta \qquad \text{(Theorems H.3 and H.4)} \\
&\le C + K\tau^{-1} H\eta.
\end{aligned}$$

In other words, when discretizing our algorithm, we pay an extra regret of at most $K\tau^{-1}H\eta$, where η is the discretization parameter. Setting $\eta = 1/K$ renders this term lower order.

H.3. Computational Complexity

In this section, we compute the running time complexity of CVaR-UCBVI under discretization of η. There are two places where discretization comes in,

1. At each h, we only compute $\widehat{U}^{\downarrow}_{h,k}(s, b, a)$ for all s, a and b in the grid. So assuming each step takes T_{step}, the total run time of DP is $\mathcal{O}(SAH\eta^{-1}T_{step})$.
2. When computing $\widehat{b}_k$, we only need to search over $\mathrm{grid}_\eta([0,1])$, since we know that the returns distribution is supported on the $\mathrm{grid}_\eta([0,1])$. Thus, the optimal solution, which is the τ-th quantile, lives on the grid. This computation costs $\mathcal{O}(\eta^{-1})$, which is lower order.

So the total runtime is $\mathcal{O}(K \cdot SAH\eta^{-1}T_{step})$. For running with the Hoeffding bonus, each step is dominated by computing the expectation $\widehat{P}_k(s,a)^\top \mathbb{E}_{r_h \sim R(s,a)}\left[\widehat{V}^{\downarrow}_{h+1,k}(\cdot, b - r_h)\right]$, as the bonus term is a constant. In the discretized MDP, this expectation can be computed using only grid elements, so $T_{step} = \mathcal{O}(S\eta^{-1})$.

When running with the Bernstein bonus, we also need to consider the complexity of computing the bonus term. In the bonus term (Eq. (6)), the expectation term $\mathbb{E}_{s' \sim \widehat{P}_k(s,a), r_h \sim R(s,a)}\left[\left(\widehat{V}^{\uparrow}_{h+1,k}(s', b - r_h) - \widehat{V}^{\downarrow}_{h+1,k}(s', b - r_h)\right)^2\right]$ can be computed in $\mathcal{O}(S\eta^{-1})$. Notably, the variance term $\mathrm{Var}_{s' \sim \widehat{P}_k(s,a)}\left(\mathbb{E}_{r_h \sim R(s,a)}\left[\widehat{V}^{\downarrow}_{h+1,k}(s', b')\right]\right)$ can also be computed in $\mathcal{O}(S\eta^{-1})$ by first computing the empirical mean (which takes $\mathcal{O}(S\eta^{-1})$). So for the Berstein bonus, we also have $T_{step} = \mathcal{O}(S\eta^{-1})$.

So the total running time of CVaR-UCBVI with discretized rewards is $\mathcal{O}(S^2AHK\eta^{-2})$. As remarked by (Auer et al., 2008; Azar et al., 2017), we can also reduce the computational cost by selectively recomputing the DP after sufficiently many observations have passed.

I. Minimax Lower Bounds for Quantile Estimation

Theorem I.1. *Let $m(p) = \mathbb{I}[p > 1/2]$ be the median of $\mathrm{Ber}(p)$. For any n,*

$$\inf_{f:\{0,1\}^n \to [0,1]} \sup_{p \in [0,1]} \mathbb{E}_{Y_1,\dots,Y_n \sim_{iid} \mathrm{Ber}(p),\, \hat{m}_n \sim \mathrm{Ber}(f(Y_1,\dots,Y_n))}[(\hat{m}_n - m(p))^2] \geq \frac{1}{2}.$$

That is, the minimax mean-squared error of any (potentially randomized) estimator for the median of a Bernoulli based on n observations thereon is bounded away from 0 for all n.

Proof. Let $g(f,p)$ denote the value of the game (the objective of the above inf-sup). Let $\mathbb{P}_p$ denote the measure of $(Y_1,\dots,Y_n) \sim \mathrm{Ber}(p)^n$. Fix any $\epsilon \in (0, \frac{e-1}{e+1}]$. Then

$$\begin{aligned}
\inf_{f:\{0,1\}^n \to [0,1]} \sup_{p \in [0,1]} g(f,p) &\geq \inf_{f:\{0,1\}^n \to [0,1]} \left(\frac{1}{2} g(f,(1+\epsilon)/2) + \frac{1}{2} g(f,(1-\epsilon)/2) \right) \\
&\geq \inf_{f:\{0,1\}^n \to \{0,1\}} \left(\frac{1}{2} g(f,(1+\epsilon)/2) + \frac{1}{2} g(f,(1-\epsilon)/2) \right) \\
&= \inf_{f:\{0,1\}^n \to \{0,1\}} \left(\frac{1}{2} \mathbb{P}_{(1+\epsilon)/2}(f(Y_1,\dots,Y_n) \neq 1) + \frac{1}{2} \mathbb{P}_{(1-\epsilon)/2}(f(Y_1,\dots,Y_n) \neq 0) \right) \\
&= \inf_{f:\{0,1\}^n \to \{0,1\}} \frac{1}{2} - \left(\frac{1}{2} \mathbb{P}_{(1+\epsilon)/2}(f(Y_1,\dots,Y_n) = 1) - \frac{1}{2} \mathbb{P}_{(1-\epsilon)/2}(f(Y_1,\dots,Y_n) = 1) \right) \\
&\geq \frac{1}{2} - \frac{1}{2} \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}_{(1+\epsilon)/2}, \mathbb{P}_{(1-\epsilon)/2})} \\
&= \frac{1}{2} - \frac{1}{2} \sqrt{\frac{1}{2} n\epsilon \log((1+\epsilon)/(1-\epsilon))} \\
&\geq \frac{1}{2} - \sqrt{\frac{n}{8}} \sqrt{\frac{e+1}{e-1}} \epsilon.
\end{aligned}$$

The first line is because worst-case risk upper bounds any Bayesian risk. The second because Bayesian risk is optimized by non-randomized estimators. The third by writing the expectation of a 0-1 variable as a probability. The fourth by total probability. The fifth by Pinsker's inequality. The sixth by evaluating the divergence. And the last by convexity of $\log((1+\epsilon)/(1-\epsilon))$.

Since $\epsilon \in (0, \frac{e-1}{e+1}]$ was arbitrary (for fixed n), the conclusion is reached. □