Regret Bounds for Risk-Sensitive Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In safety-critical applications of reinforcement learning such as healthcare and robotics, it is often desirable to optimize risk-sensitive objectives that account for tail outcomes rather than expected reward. We prove the first regret bounds for reinforcement learning under a general class of risk-sensitive objectives including the popular CVaR objective. Our theory is based on a novel characterization of the CVaR objective as well as a novel optimistic MDP construction.

1 Introduction

There has been recent interest in *risk-sensitive reinforcement learning*, which replaces the usual expected reward objective with one that accounts for variation in possible outcomes. One of the most popular risk-sensitive objectives is the *conditional value-at-risk (CVaR)* objective [1, 2, 3, 4], which is the average risk at some tail of the distribution of returns (i.e., cumulative rewards) under a given policy [5, 6]. More generally, we consider a broad class of objectives in the form of a weighted integral of quantiles of the return distribution, of which CVaR is a special case.

A key question is providing regret bounds for risk-sensitive reinforcement learning. While there has been some work studying this question, it has focused on a specific objective called the entropic risk measure [7, 8], leaving open the question of bounds for more general risk-sensitive objectives. There has also been work on optimistic exploration for CVaR [9], but without any regret bounds.

We provide the first regret bounds for risk-sensitive reinforcement learning with objectives of form

$$\Phi(\pi) = \int_0^1 F^{\dagger}_{Z^{(\pi)}}(\tau) \cdot dG(\tau), \tag{1}$$

where $Z^{(\pi)}$ is the random variable encoding the return of policy π, $F_{Z^{(\pi)}}$ is its quantile function (roughly speaking, the inverse CDF), and G is a weighting function over the quantiles. This class captures a broad range of useful objectives, and has been studied in prior work [10, 4].

We focus on the episodic setting, where the agent interacts with the environment, modeled by a Markov decision process (MDP), over a fixed sequence of episodes. Its goal is to minimize the regret—i.e., the gap between the objective value it achieves compared to the optimal policy. Our approach is based on the upper confidence bound strategy [11, 12], which makes decisions according to an optimistic estimate of the MDP. We prove that this algorithm (denoted $\mathfrak{A}$) has regret

$$\text{regret}(\mathfrak{A}) = \tilde{O}\left(T^2 \cdot L_G \cdot |\mathcal{S}|^{3/2} \cdot |\mathcal{A}| \cdot \sqrt{K}\right),$$

where T is the length of a single episode, L_G is the Lipschitz constant for the weighting function G, $|\mathcal{S}|$ is the number of states in the MDP, $|\mathcal{A}|$ is the number of actions, and K is the number of episodes (Theorem 4.1). Importantly, it achieves the optimal rate $\sqrt{K}$ achievable for typical expected return

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

objectives (which is a lower bound in our setting since expected return is an objective in the class we consider, taking $G(\tau) = \tau$). For CVaR objectives, we have $L_G = 1/\alpha$, where α is the size of the tail considered—e.g., when α is small, it averages over outliers with particularly small return.

The main challenge behind proving our result is bounding the gap between the objective value for the estimated MDP and the true MDP. In particular, even if we have a uniform bound $\|F_{\hat{Z}^{(\pi)}} - F_{Z^{(\pi)}}\|_\infty$ on the CDFs of the estimated return $\hat{Z}^{(\pi)}$ and the true return $Z^{(\pi)}$, we need to translate this to a bound on the corresponding objective values. To do so, we prove that equivalently, we have

$$\Phi(\pi) = 2T - \int_{\mathbb{R}} G(F_{Z^{(\pi)}}(x)) \cdot dx.$$

This equivalent expression for Φ follows by variable substitution and integration by parts when $F_{Z^{(\pi)}}$ is invertible (so $F^{\dagger}_{Z^{(\pi)}}(\tau) = F^{-1}_{Z^{(\pi)}}(\tau)$), but the general case requires significantly more care. We show that it holds for an arbitrary CDF $F_{Z^{(\pi)}}$.

In addition to our regret bound, we provide several other useful results for MDPs with risk-sensitive objectives. In particular, optimal policies for risk-sensitive objectives may be non-Markov. For CVaR objectives, it is known that the optimal policy only needs to depend on the cumulative return accrued so far [13]. We prove that this holds in general for objectives of the form (1) (Theorem 3.1). Furthermore, the cumulative return so far is a continuous component; we prove that discretizing this component yields an arbitrarily close approximation of the true MDP (Theorem 3.2).

Related work. To the best of our knowledge, the only prior work on regret bounds for risk-sensitive reinforcement learning is specific to the entropic risk objective [7, 8]:

$$J(\pi) = \frac{1}{\beta} \log \mathbb{E}_{Z^{(\pi)}} \left[e^{\beta Z^{(\pi)}} \right],$$

where $\beta \in \mathbb{R}_{>0}$ is a hyperparameter. As $\beta \to 0$, this objective recovers the expected return objective; for $\beta < 0$, it encourages risk aversion by upweighting negative returns; and for $\beta > 0$, it encourages risk seeking behaviors by upweighting positive returns. This objective is amenable to theoretical analysis since the value function satisfies a variant of the Bellman equation called the *exponential Bellman equation*; however, it is a narrow family of risk measures and is not widely used in practice.

In contrast, we focus on a much broader class of risk measures including the popular CVaR objective [1], which is used to minimize tail losses. To the best of our knowledge, we provide the first regret bounds for the CVaR objective and for the wide range of objectives given by (1).

2 Problem Formulation

Markov decision process. We consider a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, D, P, \mathbb{P}, T)$, with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, initial state distribution $D(s)$, finite time horizon T, transition probabilities $P(s' \mid s, a)$, and reward measure $\mathbb{P}_{R(s,a)}$; without loss of generality, we assume $r \in [0, 1]$ with probability one. A *history* is a sequence

$$\xi \in \mathcal{Z} = \bigcup_{t=1}^{T} \mathcal{Z}_t \qquad \text{where} \qquad \mathcal{Z}_t = (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{t-1} \times \mathcal{S}$$

Intuitively, a history captures the interaction between an agent and $\mathcal{M}$ up to step t. We consider stochastic, time-varying, history-dependent policies $\pi_t(a_t \mid \xi_t)$, where t is the time step. Given π, the history $\Xi_t^{(\pi)}$ generated by π up to step t is a random variable with probability measure

$$\mathbb{P}_{\Xi_t^{(\pi)}}(\xi_t) = \begin{cases} D(s_1) & \text{if } t = 1 \\ \mathbb{P}_{\Xi_{t-1}^{(\pi)}}(\xi_{t-1}) \cdot \pi_t(a_t \mid \xi_{t-1}) \cdot \mathbb{P}_{R(s_t,a_t)}(r_t) \cdot P(s_{t+1} \mid s_t, a_t) & \text{otherwise,} \end{cases}$$

where for all $\tau \in [T]$ we use the notation

$$\xi_\tau = ((s_1, a_1, r_1), ..., (s_{\tau-1}, a_{\tau-1}, r_{\tau-1}), s_\tau).$$

Finally, an *episode* (or *rollout*) is a history $\xi \in \mathcal{Z}_T$ of length T generated by a given policy π.

Bellman equation. The *return* of π on step t is the random variable $(Z_t^{(\pi)}(\xi_t))(\xi_T) = \sum_{\tau=t}^{T} r_t$, where $\xi_T \sim \mathbb{P}_{\Xi_T^{(\pi)}}(\cdot \mid \Xi_t^{(\pi)} = \xi_t)$—i.e., it is the reward from step t given that the current history is ξ_t. Defining $Z_{T+1}^{(\pi)}(\xi, s) = 0$, the *distributional Bellman equation* [14, 9] is

$$F_{Z_t^{(\pi)}(\xi)}(x) = \sum_{a \in \mathcal{A}} \pi_t(a \mid \xi) \sum_{s' \in \mathcal{S}} P(s' \mid S(\xi), a) \int F_{Z_{t+1}^{(\pi)}(\xi \circ (a,r,s'))}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r),$$

where $S(\xi) = s$ for $\xi = (..., s)$ is the current state in history ξ, and F_X is the cumulative distribution function (CDF) of random variable X. Finally, the *cumulative return* of π is $Z^{(\pi)} = Z_1^{(\pi)}(\xi)$, where $\xi = (s) \in \mathcal{Z}_1$ for $s \sim D$ is the initial history; in particular, we have

$$F_{Z^{(\pi)}}(x) = \int F_{Z_1^{(\pi)}(\xi)}(x) \cdot dD(s).$$

Risk-sensitive objective. The *quantile function* of a random variable X is

$$F_X^{\dagger}(\tau) = \inf \{x \in \mathbb{R} \mid F_X(x) \geq \tau\}.$$

Note that if F_X is strictly monotone, then it is invertible and we have $F_X^{\dagger}(\tau) = F_X^{-1}(\tau)$. Now, our objective is given by the Riemann-Stieljes integral

$$\Phi_{\mathcal{M}}(\pi) = \int_0^1 F_{Z^{(\pi)}}^{\dagger}(\tau) \cdot dG(\tau),$$

where $G(\tau)$ is a given CDF over quantiles $\tau \in [0, 1]$. This objective was originally studied in [15] for the reinforcement learning setting. For example, choosing $G(\tau) = \min\{\tau/\alpha, 1\}$ (i.e., the CDF of the distribution Uniform($[0, \alpha]$)) for $\alpha \in [0, 1]$ yields the α-conditional value at risk (CVaR) objective; furthermore, taking $\alpha = 1$ yields the usual expected cumulative reward objective. In addition, choosing $G(\tau) = \mathbb{1}(\tau \leq \alpha)$ for $\alpha \in [0, 1]$ yields the α value at risk (VaR) objective. Other risk sensitive-objectives can also be captured in this form, for example the Wang measure [16], and the cumulative probability weighting (CPW) metric [17]. We call any policy

$$\pi_{\mathcal{M}}^* \in \arg\max_{\pi} \Phi_{\mathcal{M}}(\pi).$$

an *optimal policy*—i.e., it maximizes the given objective for $\mathcal{M}$.

Assumptions. First, we have the following assumption on the quantile function for $Z^{(\pi)}$:

Assumption 2.1. $F_{Z^{(\pi)}}^{\dagger}(1) = T$.

Since T is the maximum reward attainable in an episode, this assumption says that the maximum reward is attained with *some* nontrivial probability. This assumption is very minor; for any given MDP $\mathcal{M}$, we can modify $\mathcal{M}$ to include a path achieving reward T with arbitrarily low probability.

Assumption 2.2. G is L_G-Lipschitz continuous for some $L_G \in \mathbb{R}_{>0}$, and $G(0) = 0$.

For example, for the α-CVaR objective, we have $L_G = 1/\alpha$.

Assumption 2.3. We are given an algorithm for computing $\pi_{\mathcal{M}}^*$ for a given MDP $\mathcal{M}$.

For CVaR objectives, existing algorithms [13] can compute $\pi_{\mathcal{M}}^*$ with any desired approximation error. For completeness, we give a formal description of the procedure in Appendix D. When unambiguous, we drop the dependence on $\mathcal{M}$ and simply write π^*.

Finally, our goal is to learn while interacting with the MDP $\mathcal{M}$ across a fixed number of episodes K. In particular, at the beginning of each episode $k \in [K]$, our algorithm chooses a policy $\pi^{(k)} = \mathfrak{A}(H_k)$, where $H_k = \{\xi_{T,\kappa}\}_{\kappa=1}^{k-1}$ is the random set of episodes observed so far, to use for the duration of episode k. Then, our goal is to design an algorithm $\mathfrak{A}$ that aims to minimize *regret*, which measures the expected sub-optimality with respect to π^*:

$$\text{regret}(\mathfrak{A}) = \mathbb{E}\left[\sum_{k \in [K]} \Phi(\pi^*) - \Phi(\pi^{(k)})\right].$$

Finally, for simplicity, we assume that the initial state distribution D is known; in practice, we can remove this assumption using a standard strategy.

3 Optimal Risk-Sensitive Policies

In this section, we characterize properties of the optimal risk-sensitive policy $\pi^*_{\mathcal{M}}$. First, we show that it suffices to consider policies dependent on the current state and the cumulative rewards obtained so far, rather than the entire history. Second, the cumulative reward is a continuous quantity, making it difficult to compute the optimal policy; we prove that discretizing this component does not significantly reduce the objective value. For CVaR objectives, these results imply that existing algorithms can be used to compute the optimal risk-sensitive policy [13].

Augmented state space. We show there exists an optimal policy $\pi^*_t(a_t \mid y_t, s_t)$ that only depends on the current state s_t and cumulative reward $y_t = J(\xi_t) = \sum_{\tau=1}^{t-1} r_\tau$ obtained so far. To this end, let

$$\mathcal{Z}_t(y, s) = \{\xi \in \mathcal{Z}_t \mid J(\xi_t) \leq y \wedge s_t = s\}$$

be the set of length t histories ξ with cumulative reward at most y so far, and current state s. For any history-dependent policy π, define the alternative policy $\tilde{\pi}$ by

$$\tilde{\pi}_t(a_t \mid \xi_t) = \mathbb{E}_{\Xi_t^{(\pi)}}\left[\pi_t(a_t \mid \Xi_t^{(\pi)}) \;\middle|\; \Xi_t^{(\pi)} \in \mathcal{Z}_t(J(\xi_t), s_t)\right].$$

Note that $\tilde{\pi}$ only depends on ξ_t through $y_t = J(\xi_t)$ and s_t, we can define $\tilde{\pi}_t(a_t \mid \xi_t) = \tilde{\pi}_t(a_t \mid y_t, s_t)$.

Theorem 3.1. *For any policy π, we have $\Phi(\tilde{\pi}) = \Phi(\pi)$.*

We give a proof in Appendix A. In particular, given any optimal policy π^*, we have $\Phi(\tilde{\pi}^*) = \Phi(\pi^*)$; thus, we have $\tilde{\pi}^* \in \arg\max_\pi \Phi(\pi)$. Finally, we note that this result has already been shown for CVaR objectives [13]; our theorem generalizes the existing result to any risk-sensitive objective that can be expressed as a weighted integral of the quantile function.

Augmented MDP. As a consequence of Theorem 3.1, it suffices to consider the augmented MDP $\tilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{D}, \tilde{P}, \tilde{\mathbb{P}}, T)$. First, $\tilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}$ is the augmented state space; for a state $(s, y) \in \tilde{\mathcal{S}}$, the first component encodes the current state and the second encodes the cumulative rewards so far. The initial state distribution is a probability measure

$$\tilde{D}((s, y)) = D(s) \cdot \delta_0(y),$$

where δ_0 is the Dirac delta measure placing all probability mass on $y = 0$ (i.e., the cumulative reward so far is initially zero). The transitions are given by the product measure

$$\tilde{P}((s', y') \mid (s, y), a) = P(s' \mid s, a) \cdot \mathbb{P}_{R(s,a)}(y' - y),$$

i.e., the second component of the state space is incremented as $y' = y + r$, where r is the reward achieved in the original MDP. Finally, the rewards are now only provided on the final step:

$$\mathbb{P}_{R_t((s,y),a)}(r) = \begin{cases} \delta_y(r) & \text{if } t = T \\ 0 & \text{otherwise,} \end{cases}$$

i.e., the reward at the end of a rollout is simply the cumulative reward so far, as encoded by the second component of the state. By Theorem 3.1, it suffices to compute the optimal policy for $\tilde{\mathcal{M}}$ over history-independent policies $\pi_t(a_t \mid \tilde{s}_t)$:

$$\max_{\pi \in \Pi_{\text{ind}}} \Phi_{\tilde{\mathcal{M}}}(\pi) = \max_\pi \Phi_{\mathcal{M}}(\pi),$$

where Π_{ind} is the set of history-independent policies. Once we have $\pi^*_{\tilde{\mathcal{M}}}$, we can use it in $\mathcal{M}$ by defining $\pi_{\mathcal{M}}(a \mid \xi, s) = \pi^*_{\tilde{\mathcal{M}}}(a \mid J(\xi), s)$.

Discretized augmented MDP. Planning over $\tilde{\mathcal{M}}$ is complicated by the fact that the second component of its state space is continuous. Thus, we consider an η-discretization of $\tilde{\mathcal{M}}$, for some $\eta \in \mathbb{R}_{>0}$. To this end, we modify the reward function so that it only produces rewards in $\eta \cdot \mathbb{N} = \{\eta \cdot n \mid n \in \mathbb{N}\}$, by always rounding the reward up. Then, sums of these rewards are contained in $\eta \cdot \mathbb{N}$, so we can replace the second component of $\tilde{S}$ with $\eta \cdot \mathbb{N}$. In particular, we consider the discretized MDP $\hat{\mathcal{M}} = (\hat{\mathcal{S}}, \mathcal{A}, \tilde{D}, \hat{P}, \tilde{\mathbb{P}}, T)$, where $\hat{\mathcal{S}} = \mathcal{S} \times (\eta \cdot \mathbb{N})$, and transition probability measure

$$\hat{P}((s', y') \mid (s, y), a) = P(s' \mid s, a) \cdot (\mathbb{P}_{R(s,a)} \circ \phi^{-1})(y' - y)$$

where $\phi(r) = \eta \cdot \lceil r/\eta \rceil$. That is, $\mathbb{P}_{R(s,a)}$ is replaced with the pushforward measure $\mathbb{P}_{R(s,a)} \circ \phi^{-1}$, which gives reward $\eta \cdot i$ with probability $\mathbb{P}_{R(s,a)}[\eta \cdot (i-1) < r \leq \eta \cdot i]$.

Now, we prove that the optimal policy $\pi^*_{\hat{\mathcal{M}}}$ for the discretized augmented MDP $\hat{\mathcal{M}}$ achieves objective value close to the optimal policy $\pi^*_{\mathcal{M}}$ for the original MDP $\mathcal{M}$. Importantly, we want to consider measure performance of both policies based on the objective $\Phi_{\mathcal{M}}$ of the original MDP $\mathcal{M}$. To do so, we need a way to use $\pi^*_{\hat{\mathcal{M}}}$ in $\mathcal{M}$. Note that $\pi^*_{\hat{\mathcal{M}}}$ depends only on the state $\hat{s} = (s, y)$, where $s \in \mathcal{S}$ is a state of the original MDP $\mathcal{M}$, and $y \in \eta \cdot \mathbb{N}$ is a discretized version of the cumulative reward obtained so far. Thus, we can run $\pi^*_{\hat{\mathcal{M}}}$ in $\mathcal{M}$ by simply rounding the reward r_t at each step t up to the nearest value $\hat{r}_t \in \eta \cdot \mathbb{N}$ at each step—i.e., $\hat{r}_t = \phi(r_t)$; then, we increment the internal state as $y_t = y_{t-1} + \hat{r}_t$. We call the resulting policy $\pi_{\mathcal{M}}$ the version of $\pi^*_{\hat{\mathcal{M}}}$ *adapted* to $\mathcal{M}$. Then, our next result says that the performance of $\pi_{\mathcal{M}}$ is not too much worse than the performance of $\pi^*_{\mathcal{M}}$.

Theorem 3.2. *Let $\pi^*_{\hat{\mathcal{M}}}$ be the optimal policy for the discretized augmented MDP $\hat{\mathcal{M}}$, let $\pi_{\mathcal{M}}$ be the policy $\pi^*_{\hat{\mathcal{M}}}$ adapted to the original MDP $\mathcal{M}$, and let $\pi^*_{\mathcal{M}}$ be the optimal (history-dependent) policy for the original MDP $\mathcal{M}$. Then, we have*

$$\Phi_{\mathcal{M}}(\pi_{\mathcal{M}}) \geq \Phi_{\mathcal{M}}(\pi^*_{\mathcal{M}}) - \eta.$$

We give a proof in Appendix B. Note that we can set η to be sufficiently small to achieve any desired error level (i.e., choose ϵ/T, where ϵ is the desired error). The only cost is in computation time. Note that the number of states in $\hat{\mathcal{M}}$ is still infinite; however, since the cumulative return satisfies $y \in [0, H]$, it suffices to take $\hat{\mathcal{S}} = \mathcal{S} \times (\epsilon \cdot [\lceil H/\eta \rceil])$; then, $\hat{\mathcal{M}}$ has $|\hat{\mathcal{S}}| = |\mathcal{S}| \cdot \lceil H/\eta \rceil$ states.

4 Upper Confidence Bound Algorithm

Here, we present our upper confidence bound (UCB) algorithm (summarized in Algorithm 1). At a high level, for each episode, our algorithm constructs an estimate $\mathcal{M}^{(k)}$ of the underlying MDP $\mathcal{M}$ based on the prior episodes $i \in [k-1]$; to ensure exploration, it optimistically inflates the estimate of the reward probability measure $\mathbb{P}$. Then, it plans in $\mathcal{M}^{(k)}$ to obtain an optimistic policy $\pi^{(k)} = \pi^*_{\mathcal{M}^{(k)}}$, and uses this policy to act in the MDP for episode k.

Optimistic MDP. We define $\mathcal{M}^{(k)}$. Without loss of generality, we assume $\mathcal{S}$ includes a distinguished state s_∞ with rewards $F_{R(s_\infty,a)}(r) = \mathbb{1}(r \geq 1)$ (i.e., achieve the maximum reward $r = 1$ with probability one), and transitions $P(s_\infty \mid s, a) = \mathbb{1}(s = s_\infty)$ and $P(s' \mid s_\infty, a) = \mathbb{1}(s' = s_\infty)$ (i.e., inaccessible from other states and only transitions to itself). Our construction of $\hat{\mathcal{M}}^{(k)}$ uses s_∞ for optimism. Now, let $\tilde{\mathcal{M}}^{(k)}$ be the MDP using the empirical estimates of the transitions and rewards:

$$\tilde{P}^{(k)}(s' \mid s, a) = \frac{N_{k,t}(s, a, s')}{N_{k,t}(s, a)}$$

$$F_{\tilde{R}^{(k)}(s,a)}(r) = \frac{1}{N_{k,t}(s, a)} \sum_{i=1}^{k-1} \sum_{t=1}^{T} \mathbb{1}(r \leq r_{i,t}) \cdot \mathbb{1}\left(s_{i,t} = s \wedge a_{i,t} = a\right).$$

Then, let $\hat{\mathcal{M}}^{(k)}$ be the optimistic MDP; in particular, its transitions

$$\hat{P}^{(k)}(s' \mid s, a) = \begin{cases} \mathbb{1}(s' = s_\infty) & \text{if } s = s_\infty \\ 1 - \sum_{s' \in \mathcal{S} \setminus \{s_\infty\}} \hat{P}^{(k)}(s' \mid s, a) & \text{if } s' = s_\infty \\ \max\left\{\tilde{P}^{(k)}(s' \mid s, a) - \epsilon_P^{(k)}(s, a),\ 0\right\} & \text{otherwise} \end{cases}$$

transition to the optimistic state s_∞ when uncertain, and its rewards

$$F_{\hat{R}^{(k)}(s,a)}(r) = \begin{cases} \mathbb{1}(r \geq 1) & \text{if } s = s_\infty \\ 1 & \text{if } r \geq 1 \\ \max\left\{F_{\tilde{R}^{(k)}(s,a)}(r) - \epsilon_R^{(k)}(s, a),\ 0\right\} & \text{otherwise} \end{cases}$$

optimistically shift the reward CDF downwards. Here, $\epsilon_P^{(k)}(s, a)$ and $\epsilon_R^{(k)}(s, a)$ are defined in Section 5; intuitively, they are high-probability upper bounds on the errors of the empirical estimates $\tilde{P}^{(k)}(\cdot \mid s, a)$ and $F_{\tilde{R}^{(k)}(s,a)}$ of the transitions and rewards, respectively.

Algorithm 1 Upper Confidence Bound Algorithm

for $k \in [K]$ **do**
  Compute $\mathcal{M}^{(k)}$ and $\pi^{(k)} = \pi^*_{\mathcal{M}^{(k)}}$ using prior episodes $\{\xi^{(i)} \mid i \in [k-1]\}$
  Execute $\pi^{(k)}$ in the true MDP $\mathcal{M}$ and observe episode $\xi^{(k)} = [(s_{k,t}, a_{k,t}, r_{k,t})]_{t=1}^T \cup [s_{k,T+1}]$
end for

Theoretical guarantees. We have the following upper bound on the regret of Algorithm 1.

Theorem 4.1. *Denote Algorithm 1 by* $\mathfrak{A}$*. For any* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$*, we have*

$$\text{regret}(\mathfrak{A}) \leq 4T^2 \cdot L_G \cdot |\mathcal{S}| \cdot |\mathcal{A}| \cdot \sqrt{5|\mathcal{S}| \cdot K \cdot \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)} = \tilde{\mathcal{O}}(\sqrt{K}).$$

We briefly compare our bound to existing ones in the setting of expected return objectives. The dependence on the number of episodes K matches existing bounds [11, 12]; since this is optimal in the setting of expected return, and expected return is a special case of our setting (with $G(\tau) = \tau$), our bound is also optimal in K. In terms of the dependence on the number of states $|\mathcal{S}|$, our bound has an extra $\sqrt{|\mathcal{S}|}$ factor compared to the UCRL2 algorithm [11], and an extra $|\mathcal{S}|$ factor compared to the improved bound of the UCBVI algorithm [12]. In terms of the dependence on the number of actions $|\mathcal{A}|$, our bound has an extra $\sqrt{|\mathcal{S}|}$ factor compared to both UCRL2 and UCBVI. Our dependence on the horizon length T is T^2, compared to a $T^{3/2}$ in UCBVI and T in a variant of UCBVI [12] utilizing a carefully designed variance-based bonus.

5 Proof of Theorem 4.1

We prove Theorem 4.1; we defer proofs of several lemmas to Appendix C.

At a high level, the proof proceeds in three steps. First, we prove our key Lemma 5.1, which expresses the objective Φ in terms of an integral of the weighted CDF of the return. This lemma allows us to translate bounds on the difference between CDFs of the estimated return $\hat{Z}^{(\pi)}$ and the true return $Z^{(\pi)}$ into bounds on the difference between corresponding objective values. The proof of this lemma is divided into three parts that deal with different sets of points in the domain of the quantile function $F^\dagger_{Z^{(\pi)}}$: (i) discontinuous; (ii) continuous and strictly monotone; (iii) continuous and non-strictly monotone. This result is used throughout the remainder of the proof.

Second, we define $\mathcal{E}$ to be the event where the optimistic estimated MDP $\hat{\mathcal{M}}^{(k)}$ falls into a certain confidence set around the true MDP $\mathcal{M}$ for each $k \in [K]$; in Lemma 5.2, we prove that $\mathcal{E}$ holds with high probability. Then, in Lemma 5.6, we prove that under event $\mathcal{E}$, the objective values of $\hat{\mathcal{M}}^{(k)}$ and $\mathcal{M}$ are close. To prove this lemma, we separately show that (i) the objective values of the estimated MDP $\tilde{\mathcal{M}}^{(k)}$ (estimated without optimism) and $\mathcal{M}$ are close (Lemma 5.4), and (ii) the objective values of $\hat{\mathcal{M}}^{(k)}$ and $\tilde{\mathcal{M}}^{(k)}$ are close (Lemma 5.5).

Third, in Lemma 5.7, we prove that under event $\mathcal{E}$, the MDP $\hat{\mathcal{M}}^{(k)}$ is indeed optimistic. Together, these results imply the regret bound using the standard UCB proof strategy.

We proceed with the proof. First, we have our key result providing an equivalent expression for Φ:

Lemma 5.1. *We have*

$$\Phi(\pi) = T - \int_{\mathbb{R}} G(F_{Z^{(\pi)}}(x)) \cdot dx.$$

Proof. First, note that by integration by parts, we have

$$\begin{aligned}\Phi(\pi) = \int_0^1 F^\dagger_{Z^{(\pi)}}(\tau) \cdot dG(\tau) &= \left[F^\dagger_{Z^{(\pi)}}(\tau) \cdot G(\tau)\right]_0^1 - \int_0^1 G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau) \\ &= T - \int_0^1 G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau),\end{aligned}$$

where the last line follows by Assumptions 2.1 & 2.2. Thus, it suffices to show that

$$\int_0^1 G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau) = \int_{\mathbb{R}} G(F_{Z^{(\pi)}}(x)) \cdot dx.$$

The quantile function $F^\dagger_{Z^{(\pi)}}$ is monotonically increasing and left-continuous [18], so this integral is equivalently a Lebesgue-Stieltjes integral [19]. Dividing the unit interval $I = [0, 1]$ into disjoint sets

$$\begin{aligned}
I^{(1)} &= \{\tau \in \mathbb{I} \mid F^\dagger_{Z^{(\pi)}}(\tau) \text{ is discontinuous}\} \\
I^{(2)} &= \{\tau \in \mathbb{I} \mid F^\dagger_{Z^{(\pi)}}(\tau) \text{ is continuous and strictly monotone}\} \\
I^{(3)} &= \{\tau \in \mathbb{I} \mid F^\dagger_{Z^{(\pi)}}(\tau) \text{ is continuous and non-strictly monotone}\},
\end{aligned}$$

then we have

$$\int_0^1 G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau) = \int_{I^{(1)}} G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau) + \int_{I^{(2)}} G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau) + \int_{I^{(3)}} G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau).$$

We consider each of the three terms separately and then combine them to finish the proof.

First term. Note that $I^{(1)} = \{\tau_i^{(1)}\}_{i=1}^\infty$ is countable since monotone functions can have countably many discontinuities. Also, for each $i \in \mathbb{N}$, the measure assigned to $\tau_i^{(1)}$ by $dF^\dagger_{Z^{(\pi)}}$ is

$$dF^\dagger_{Z^{(\pi)}}(\{\tau_i^{(1)}\}) = \lim_{\tau \to \tau_i^{(1)}+} F^\dagger_{Z^{(\pi)}}(\tau) - F^\dagger_{Z^{(\pi)}}(\tau_i^{(1)}) =: x_i^{(1)+} - x_i^{(1)}.$$

Thus, we have

$$\begin{aligned}
\int_{I^{(1)}} G(\tau) \cdot dF^\dagger_{Z^{(\pi)}}(\tau) = \sum_{i=1}^\infty G(\tau_i^{(1)}) \cdot (x_i^{(1)+} - x_i^{(1)}) &= \sum_{i=1}^\infty G(\tau_i^{(1)}) \cdot \int_{x_i^{(1)}}^{x_i^{(1)+}} dx \\
&= \sum_{i=1}^\infty \int_{x_i^{(1)}}^{x_i^{(1)+}} G(F_{Z^{(\pi)}}(x)) \cdot dx \\
&= \sum_{i=1}^\infty \int_{F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\})} G(F_{Z^{(\pi)}}(x)) \cdot dx \\
&= \int_{F^{-1}_{Z^{(\pi)}}(I^{(1)})} G(F_{Z^{(\pi)}}(x)) \cdot dx.
\end{aligned}$$

On the second line, we have used the fact that $F_{Z^{(\pi)}}(x) = \tau_i^{(1)}$ for all $x \in [x_i^{(1)}, x_i^{(1)+})$. To see this fact, note that since $x \geq x_i^{(1)}$, by monotonicity of $F_{Z^{(\pi)}}$, we have $F_{Z^{(\pi)}}(x) \geq F_{Z^{(\pi)}}(x_i^{(1)}) = \tau_i^{(1)}$. Furthermore, if $F_{Z^{(\pi)}}(x) > \tau_i^{(1)}$, then we would have

$$\begin{aligned}
x_i^{(1)+} = \lim_{\tau \to \tau_i^{(1)}+} F^\dagger_{Z^{(\pi)}}(\tau) &= \lim_{\tau \to \tau_i^{(1)+}} \inf\{x' \in \mathbb{R} \mid F_{Z^{(\pi)}}(x') \geq \tau\} \\
&\leq \inf\{x' \in \mathbb{R} \mid F_{Z^{(\pi)}}(x') \geq F_{Z^{(\pi)}}(x)\} \\
&\leq x,
\end{aligned}$$

where the first inequality follows since $F_{Z^{(\pi)}}(x) \geq \tau$ for τ sufficiently close to $\tau_i^{(1)}$, and the second since $x \in \{x' \in \mathbb{R} \mid F_{Z^{(\pi)}}(x') \geq F_{Z^{(\pi)}}(x)\}$. Since we have assumed $x < x_i^{(1)+}$, we have a contradiction, so $F_{Z^{(\pi)}}(x) \leq \tau_i^{(1)}$. Thus, it follows that $F_{Z^{(\pi)}}(x) = \tau_i^{(1)}$, as claimed. The third line follows since

$$F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\}) = [x_i^{(1)}, x_i^{(1)+}) \qquad \text{or} \qquad F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\}) = [x_i^{(1)}, x_i^{(1)+}].$$

In particular, for any $x \in F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\})$, we have $F_{Z^{(\pi)}}(x) = \tau_i^{(1)}$, so

$$x \geq \inf\{x \in \mathbb{R} \mid F_{Z^{(\pi)}}(x) \geq \tau_i^{(1)}\} = x_i^{(1)}.$$

Conversely, we have

$$x_i^{(1)+} = \lim_{\tau \to \tau_i^{(1)}+} F^{\dagger}_{Z^{(\pi)}}(\tau) = \lim_{\tau \to \tau_i^{(1)+}} \inf\{x' \in \mathbb{R} \mid F_{Z^{(\pi)}}(x') \geq \tau\} \geq x$$

since $F_{Z^{(\pi)}}(x') \leq \tau_i^{(1)} < \tau$ for all $x' \leq x$ so the infimum must be $\geq x$. These two arguments show that $F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\}) \subseteq [x_i^{(1)}, x_i^{(1)+}]$. The fact that $[x_i^{(1)}, x_i^{(1)+}) \subseteq F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\})$ follows by the same argument as for the second line. The claim follows. Finally, the fourth line follows since the sets $F^{-1}_{Z^{(\pi)}}(\{\tau_i^{(1)}\})$ are disjoint.

Second term. For any $\tau \in I^{(2)}$, then $F^{-1}_{Z^{(\pi)}}$ exists at τ, and we have $F^{\dagger}_{Z^{(\pi)}}(\tau) = F^{-1}_{Z^{(\pi)}}(\tau)$. Thus, by a substitution $\tau = F_{Z^{(\pi)}}(x)$, we have

$$\int_{I^{(2)}} G(\tau) \cdot dF^{\dagger}_{Z^{(\pi)}}(\tau) = \int_{F^{-1}_{Z^{(\pi)}}(I^{(2)})} G(F_{Z^{(\pi)}}(x)) \cdot dx.$$

Third term. We can divide $I^{(3)}$ into a union of disjoint intervals $I^{(3)} = \bigcup_{i=1}^{\infty} I_i^{(3)}$, where

$$I_i^{(3)} = \{\tau \in [0,1] \mid F^{\dagger}_{Z^{(\pi)}}(\tau) = x_i^{(3)}\}$$

for some $x_i^{(3)} \in \mathbb{R}$; there are only be countably many such intervals (since each one contains a distinct rational number). Then, we have

$$\int_{I^{(3)}} G(\tau) \cdot dF^{\dagger}_{Z^{(\pi)}}(\tau) = 0 = \int_{F^{-1}_{Z^{(\pi)}}(I^{(3)})} G(F_{Z^{(\pi)}}(x)) \cdot dx,$$

since $F^{-1}_{Z^{(\pi)}}(I^{(3)}) = \{x_i^{(3)}\}_{i=1}^{\infty}$ has measure zero according to the Lebesgue measure dx.

Final proof. Finally, note that $F^{-1}_{Z^{(\pi)}}(I^{(1)})$, $F^{-1}_{Z^{(\pi)}}(I^{(2)})$, and $F^{-1}_{Z^{(\pi)}}(I^{(3)})$ cover $\mathbb{R}$ and are disjoint except possibly on a set of measure zero, so

$$\int_{F^{-1}_{Z^{(\pi)}}(I^{(1)})} G(F_{Z^{(\pi)}}(x)) \cdot dx + \int_{F^{-1}_{Z^{(\pi)}}(I^{(2)})} G(F_{Z^{(\pi)}}(x)) \cdot dx + \int_{F^{-1}_{Z^{(\pi)}}(I^{(3)})} G(F_{Z^{(\pi)}}(x)) \cdot dx$$
$$= \int_{\mathbb{R}} G(F_{Z^{(\pi)}}(x)) \cdot dx.$$

The claim follows. □

Next, given $\delta \in \mathbb{R}_{>0}$, define $\mathcal{E}$ to be the event where the following hold:

$$\|\tilde{P}^{(k)}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \leq \sqrt{\frac{2|\mathcal{S}|}{N^{(k)}(s,a)} \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)} =: \epsilon_P^{(k)}(s,a) \quad (\forall s \in \mathcal{S}, a \in \mathcal{A})$$

$$\|F_{\tilde{R}^{(k)}(s,a)} - F_{R(s,a)}\|_\infty \leq \sqrt{\frac{1}{2N^{(k)}(s,a)} \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)} =: \epsilon_R^{(k)}(s,a) \quad (\forall s \in \mathcal{S}, a \in \mathcal{A}).$$

Lemma 5.2. *We have* $\mathbb{P}[\mathcal{E} \mid \{N^{(k)}(s,a)\}_{k \in [K], s \in \mathcal{S}, a \in \mathcal{A}}] \geq 1 - \delta$.

Next, let $\tilde{Z}^{(k,\pi)}$ and $\hat{Z}^{(k,\pi)}$ be the returns for policy π for $\tilde{\mathcal{M}}^{(k)}$ and $\hat{\mathcal{M}}^{(k)}$, respectively, let $\Phi = \Phi_{\mathcal{M}}$, $\tilde{\Phi}^{(k)} = \Phi_{\tilde{\mathcal{M}}^{(k)}}$, and $\hat{\Phi}^{(k)} = \Phi_{\hat{\mathcal{M}}^{(k)}}$, and let $\pi^* = \pi^*_{\mathcal{M}}$, $\tilde{\pi}^{(k)} = \pi^*_{\tilde{\mathcal{M}}^{(k)}}$, and $\hat{\pi}^{(k)} = \pi^*_{\hat{\mathcal{M}}^{(k)}}$. Now, we prove two key results: (i) $\hat{\Phi}^{(k)}$ is close to Φ, and (ii) $\hat{\Phi}^{(k)}$ is optimistic compared to Φ. To this end, we have the following key lemma; its proof depends critically on Lemma 5.1.

Lemma 5.3. *Consider MDPs* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, D, P, \mathbb{P}, T)$ *and* $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, D, P', \mathbb{P}', T)$*, such that* $\|P'(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \leq \epsilon_P(s,a)$ *and* $\|F_{R'(s,a)} - F_{R(s,a)}\|_\infty \leq \epsilon_R(s,a)$*. Then, we have*

$$|\Phi'(\pi) - \Phi(\pi)| \leq T \cdot L_G \cdot B(\pi) \qquad (\forall k \in [K], \pi),$$

where

$$B(\pi) = \mathbb{E}_{\Xi_T^{(\pi)}}\left[\sum_{t=1}^{T} \epsilon_P(s_t,a_t) + \epsilon_R(s_t,a_t)\right].$$

Our next lemma characterizes the connection between $\tilde{\Phi}^{(k)}$ and Φ.

Lemma 5.4. *On event $\mathcal{E}$ and conditioned on $\{N^{(k)}(s,a)\}_{k\in[K],s\in\mathcal{S},a\in\mathcal{A}}$, we have*

$$|\tilde{\Phi}^{(k)}(\pi) - \Phi(\pi)| \leq T \cdot L_G \cdot B^{(k)}(\pi) \qquad (\forall k \in [K], \pi),$$

where

$$B^{(k)}(\pi) = \mathbb{E}_{\Xi_T^{(\pi)}}\left[\sum_{t=1}^{T} \epsilon_P^{(k)}(s_t,a_t) + \epsilon_R^{(k)}(s_t,a_t) \;\middle|\; \{N^{(k)}(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}\right].$$

Proof. The result follows since on event $\mathcal{E}$ and conditioned on $\{N^{(k)}(s,a)\}_{k\in[K],s\in\mathcal{S},a\in\mathcal{A}}$, $\tilde{\mathcal{M}}^{(k)}$ and $\mathcal{M}$ satisfy the conditions of Lemma 5.3 for all $k \in [K]$. □

Our next lemma characterizes the connection between $\hat{\Phi}^{(k)}$ and $\tilde{\Phi}^{(k)}$.

Lemma 5.5. *For each $k \in [K]$ and any policy π, we have*

$$|\hat{\Phi}^{(k)}(\pi) - \tilde{\Phi}(\pi)| \leq 2T \cdot |\mathcal{S}| \cdot B^{(k)}(\pi).$$

Proof. The result follows since by definition of $\hat{\mathcal{M}}^{(k)}$, $\hat{\mathcal{M}}^{(k)}$ and $\tilde{\mathcal{M}}^{(k)}$ satisfy the condition of Lemma 5.3 with $\epsilon_P(s,a) = 2|S| \cdot \epsilon_P^{(k)}(s,a)$ and $\epsilon_R(s,a) = \epsilon_R^{(k)}(s,a)$ for all $k \in [K]$. □

Now, we prove the first key claim—i.e., $\hat{\Phi}^{(k)}$ is close to Φ.

Lemma 5.6. *On event $\mathcal{E}$, for all $k \in [K]$ and any policy π, we have*

$$|\hat{\Phi}^{(k)}(\pi) - \Phi(\pi)| \leq 4T \cdot L_G \cdot |\mathcal{S}| \cdot B^{(k)}(\pi).$$

Proof. Note that

$$|\hat{\Phi}^{(k)}(\pi) - \Phi(\pi)| \leq |\hat{\Phi}^{(k)}(\pi) - \tilde{\Phi}^{(k)}(\pi)| + |\tilde{\Phi}^{(k)}(\pi) - \Phi(\pi)| \leq 4T \cdot L_G \cdot |\mathcal{S}| \cdot B^{(k)}(\pi),$$

where the second inequality follows by Lemmas 5.4 & 5.5. □

Now, we prove the second key claim—i.e., $\hat{\Phi}^{(k)}$ is optimistic compared to Φ.

Lemma 5.7. *On event $\mathcal{E}$, we have $\hat{\Phi}^{(k)}(\pi) \geq \Phi(\pi)$ for all $k \in [K]$ and all policies π.*

With these two key claims, the proof of Theorem 4.1 follows by a standard upper confidence bound argument; we give the proof in Appendix C.4.

6 Conclusion

We have proposed a novel regret bound for risk sensitive reinforcement learning that applies to a broad class of objective functions, including the popular conditional value-at-risk (CVaR) objective. Our results recover the usual $\sqrt{K}$ dependence on the number of episodes, and also highlights dependence on the Lipschitz constant L_G of the integral of the weighting function G used to define the objective. Future work includes extending these ideas to the setting of function approximation and understanding whether alternative exploration strategies such as Thompson sampling are applicable.

References

[1] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[2] Yichuan Charlie Tang, Jian Zhang, and Ruslan Salakhutdinov. Worst cases policy gradients. *arXiv preprint arXiv:1911.03618*, 2019.

[3] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[4] Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[6] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in neural information processing systems*, 27, 2014.

[7] Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pages 3198–3207. PMLR, 2021.

[9] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[10] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[11] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[12] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

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

[14] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[15] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[16] Shaun S Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of risk and insurance*, pages 15–36, 2000.

[17] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[18] Paul Embrechts and Marius Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432, 2013.

[19] Elias M Stein and Rami Shakarchi. Real analysis. In *Real Analysis*. Princeton University Press, 2009.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
   (c) Did you describe the limitations of your work? [Yes] **Assumptions are stated in Sec 2.**
   (d) Did you discuss any potential negative societal impacts of your work? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] **See Section 2.**
   (b) Did you include complete proofs of all theoretical results? [Yes] **See Section 5 and supplement.**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Theorem 3.1

In this section, we prove Theorem 3.1, which says that it suffices to the augmented state space (y, s) rather than the whole history ξ. First, we have the following lemma.

Lemma A.1. *For any $y \in \mathbb{R}$ and $s \in \mathcal{S}$, we have*

$$\mathbb{E}_{\Xi_t^{(\pi)}}\left[\pi_t(a \mid \Xi_t^{(\pi)}) \cdot \mathbb{1}\left(\Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right)\right] = \mathbb{E}_{\Xi_t^{(\pi)}}\left[\tilde{\pi}_t(a \mid \Xi_t^{(\pi)}) \cdot \mathbb{1}\left(\Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right)\right].$$

Proof. Note that

$$\begin{aligned}
&\mathbb{E}_{\Xi_t^{(\pi)}}\left[\tilde{\pi}_t(a \mid \Xi_t^{(\pi)}) \;\middle|\; \Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right] \\
&= \mathbb{E}_{\Xi_t^{(\pi)}}\left[\mathbb{E}_{\tilde{\Xi}_t^{(\pi)}}\left[\pi_t(a_t \mid \tilde{\Xi}_t^{(\pi)}) \;\middle|\; \tilde{\Xi}_t^{(\pi)} \in \mathcal{Z}_t(J(\Xi_t^{(\pi)}), s_t)\right] \;\middle|\; \Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right] \\
&= \mathbb{E}_{\Xi_t^{(\pi)}}\left[\mathbb{E}_{\tilde{\Xi}_t^{(\pi)}}\left[\pi_t(a_t \mid \tilde{\Xi}_t^{(\pi)}) \;\middle|\; \tilde{\Xi}_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right] \;\middle|\; \Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right] \\
&= \mathbb{E}_{\tilde{\Xi}_t^{(\pi)}}\left[\pi_t(a_t \mid \tilde{\Xi}_t^{(\pi)}) \;\middle|\; \Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)\right].
\end{aligned}$$

The claim follows by replacing $\tilde{\Xi}_t^{(\pi)}$ with $\Xi_t^{(\pi)}$ and multiplying by $\mathbb{P}_{\Xi_t^{(\pi)}}[\Xi_t^{(\pi)} \in \mathcal{Z}_t(y,s)]$. □

Next, let

$$D_t^{(\pi)}(y,s) = \mathbb{P}_{\Xi_t^{(\pi)}}\left[J(\Xi_t^{(\pi)}) \le y \wedge S(\Xi_t^{(\pi)}) = s\right]$$

be the probability of a history achieving current cumulative return at most y and ending in state s.

Lemma A.2. *We have $D_t^{(\pi)} = D_t^{(\tilde{\pi})}$.*

Proof. We prove by induction. The base case $t = 1$ follows trivially. For the inductive case, note that

$$\begin{aligned}
D_{t+1}^{(\pi)}(y,s'') &= \int \mathbb{1}(\xi' \in \mathcal{Z}_{t+1}(y,s'')) \cdot d\mathbb{P}_{\Xi_{t+1}^{(\pi)}}(\xi') \\
&= \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathbb{1}(\xi \circ (a,r,s') \in \mathcal{Z}_{t+1}(y,s'')) \cdot P(s' \mid S(\xi), a) \cdot d\mathbb{P}_{R(s,a)}(r) \cdot \pi(a \mid \xi) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi) \\
&= \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathbb{1}(J(\xi) + r \le y) \cdot \mathbb{1}(s' = s'') \cdot P(s' \mid S(\xi), a) \cdot d\mathbb{P}_{R(s,a)}(r) \cdot \pi(a \mid \xi) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi) \\
&= \int \sum_{a\in\mathcal{A}} \mathbb{1}(J(\xi) + r \le y) \cdot P(s'' \mid S(\xi), a) \cdot d\mathbb{P}_{R(s,a)}(r) \cdot \pi(a \mid \xi) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi),
\end{aligned}$$

where the first line follows by definition of $D_{t+1}^{(\pi)}$, the second by the inductive formula for $\mathbb{P}_{\Xi_{t+1}^{(\pi)}}$, the third since $J(\xi \circ (a,r,s')) = J(\xi) + r$, and the fourth by summing over s'. Continuing, we have

$$\begin{aligned}
D_{t+1}^{(\pi)}(y,s'') &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \pi(a \mid \xi) \cdot \mathbb{1}(J(\xi) + r \le y) \cdot \mathbb{1}(S(\xi) = s) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \pi(a \mid \xi) \cdot \mathbb{1}(\xi \in \mathcal{Z}_t(y-r,s)) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \tilde{\pi}(a \mid \xi) \cdot \mathbb{1}(\xi \in \mathcal{Z}_t(y-r,s)) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r),
\end{aligned}$$

where the first line follows by introducing $\mathbb{1}(S(\xi) = s)$ and rearranging, the second by definition of $\mathcal{Z}_t$, and the third by Lemma A.1. Continuing, we have

$$\begin{aligned}
D_{t+1}^{(\pi)}(y, s'') &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \mathbb{1}(\xi \in \mathcal{Z}_t(y-r, s)) \cdot d\mathbb{P}_{\Xi_t^{(\pi)}}(\xi) \cdot \tilde{\pi}(a \mid y-r, s) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int D_t^{(\pi)}(y-r, s) \cdot \tilde{\pi}(a \mid y-r, s) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int D_t^{(\tilde{\pi})}(y-r, s) \cdot \tilde{\pi}(a \mid y-r, s) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \mathbb{1}(\xi \in \mathcal{Z}_t(y-r, s)) \cdot d\mathbb{P}_{\Xi_t^{(\tilde{\pi})}}(\xi) \cdot \tilde{\pi}(a \mid y-r, s) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r),
\end{aligned}$$

where the first line follows since $\tilde{\pi}$ is independent of ξ and by rearranging, the second by definition of $D_t^{(\pi)}$, the third by induction, and the fourth by definition of $D_t^{(\tilde{\pi})}$. Continuing, we have

$$\begin{aligned}
D_{t+1}^{(\pi)}(y, s'') &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \tilde{\pi}(a \mid \xi) \cdot \mathbb{1}(\xi \in \mathcal{Z}_t(y-r, s)) \cdot d\mathbb{P}_{\Xi_t^{(\tilde{\pi})}}(\xi) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \int \tilde{\pi}(a \mid \xi) \cdot \mathbb{1}(J(\xi) + r \le y) \cdot \mathbb{1}(S(\xi) = s) \cdot d\mathbb{P}_{\Xi_t^{(\tilde{\pi})}}(\xi) \cdot P(s'' \mid s, a) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{a\in\mathcal{A}} \int \mathbb{1}(J(\xi) + r \le y) \cdot P(s'' \mid S(\xi), a) \cdot d\mathbb{P}_{R(s,a)}(r) \cdot \tilde{\pi}(a \mid \xi) \cdot d\mathbb{P}_{\Xi_t^{(\tilde{\pi})}}(\xi) \\
&= \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \int \mathbb{1}(J(\xi) + r \le y) \cdot \mathbb{1}(s' = s'') \cdot P(s' \mid S(\xi), a) \cdot d\mathbb{P}_{R(s,a)}(r) \cdot \tilde{\pi}(a \mid \xi) \cdot d\mathbb{P}_{\Xi_t^{(\tilde{\pi})}}(\xi) \\
&= \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \int \mathbb{1}(\xi \circ (a, r, s') \in \mathcal{Z}_{t+1}(y, s'')) \cdot P(s' \mid S(\xi), a) \cdot d\mathbb{P}_{R(s,a)}(r) \cdot \tilde{\pi}(a \mid \xi) \cdot d\mathbb{P}_{\Xi_t^{(\tilde{\pi})}}(\xi) \\
&= \int \mathbb{1}(\xi' \in \mathcal{Z}_{t+1}(y, s'')) \cdot d\mathbb{P}_{\Xi_{t+1}^{(\pi)}}(\xi') \\
&= D_{t+1}^{(\tilde{\pi})}(y, s''),
\end{aligned}$$

where the first line follows by definition of $\tilde{\pi}$, the second by definition of $\mathcal{Z}_t$, the third by summing over s and rearranging, the fourth by introducing $\mathbb{1}(s' = s'')$, the fifth by definition of $\mathcal{Z}_{t+1}$, the sixth by the inductive formula for $\mathbb{P}_{\Xi_{t+1}^{(\pi)}}$, and the seventh by the definition of $D_{t+1}^{(\tilde{\pi})}$. The claim follows. $\square$

Now, we prove Theorem 3.1. By Lemma A.2, we have

$$\begin{aligned}
F_{Z^{(\pi)}}(x) &= \int \mathbb{1}(J(\xi) \le x) \cdot d\mathbb{P}_{\Xi_T^{(\pi)}}(\xi) \\
&= \sum_{s\in\mathcal{S}} \int \mathbb{1}(J(\xi) \le x) \cdot \mathbb{1}(S(\xi) = s) \cdot d\mathbb{P}_{\Xi_T^{(\pi)}}(\xi) \\
&= \sum_{s\in\mathcal{S}} \int \mathbb{1}(\xi \in \mathcal{Z}_T(x, s)) \cdot d\mathbb{P}_{\Xi_T^{(\pi)}}(\xi) \\
&= \sum_{s\in\mathcal{S}} \int \mathbb{1}(\xi \in \mathcal{Z}_T(x, s)) \cdot d\mathbb{P}_{\Xi_T^{(\tilde{\pi})}}(\xi) \\
&= \int \mathbb{1}(J(\xi) \le x) \cdot d\mathbb{P}_{\Xi_T^{(\tilde{\pi})}}(\xi) \\
&= F_{Z^{(\tilde{\pi})}}(x).
\end{aligned}$$

Theorem 3.1 follows straightforwardly from this result. $\square$

B Proof of Theorem 3.2

We construct a sequence of MDPs $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_T$, such that $\mathcal{M}_0 = \tilde{\mathcal{M}}$ and $\mathcal{M}_T = \hat{\mathcal{M}}$, and where we can bound the incremental errors

$$\Phi_{\tilde{\mathcal{M}}}(\pi^*_{\mathcal{M}_\tau}) - \Phi_{\hat{M}}(\pi^*_{\mathcal{M}_{\tau-1}}),$$

noting a policy for one of the MDPs can be used in all the other MDPs. For each $\tau \in [T]$, the MDP $\mathcal{M}_\tau$ discretizes the reward assigned on the tth step of $\mathcal{M}_{\tau-1}$—more precisely, it discretizes the transitions since the rewards are only assigned on the last step based on the cumulative reward recorded in the second component of the state space. Formally, $\mathcal{M}_\tau$ is identical to $\tilde{\mathcal{M}}$, except it uses the (time-varying) transition probability measure $\hat{P}^{(\tau)}$ defined by

$$\hat{P}^{(\tau)}_t((s', y') \mid (s, y), a) = \begin{cases} P(s' \mid s, a) \cdot (\mathbb{P}_{R(s,a)} \circ \phi^{-1})(y' - y) & \text{if } t \leq \tau \\ P(s' \mid s, a) \cdot \mathbb{P}_{R(s,a)}(y' - y) & \text{otherwise.} \end{cases}$$

Then, $\mathcal{M}_\tau$ is identical to $\mathcal{M}_{\tau-1}$ except $\mathbb{P}_{R(s,a)}$ is replaced with $\mathbb{P}_{R(s,a)} \circ \phi^{-1}$ on step τ. We prove three lemmas showing a lower bound on the value of a policy π for $\mathcal{M}_\tau$ when adapted to $\mathcal{M}_{\tau-1}$.

Lemma B.1. *Given $\tau \in [T]$, let $\mathcal{M} = \mathcal{M}_{\tau-1}$ and $\hat{\mathcal{M}} = \mathcal{M}_\tau$ (so compared to $\mathcal{M}$, $\hat{\mathcal{M}}$ replaces $\mathbb{P}_{R(s,a)}$ with $\mathbb{P}_{R(s,a)} \circ \phi^{-1}$ on step τ in its transitions). Given any policy $\hat{\pi}$ for $\hat{\mathcal{M}}$, define the policy*

$$\pi_t(a \mid s, y, \alpha) = \hat{\pi}_t(a \mid s, y + \alpha)$$

for $\mathcal{M}$, where we initialize the (extra) policy internal state $\alpha_1 = 0$, and we update $\alpha_{\tau+1} = \phi(r_\tau) - r_\tau$ on step τ and $\alpha_{t+1} = \alpha_t$ otherwise. Then, for all $x, y, \alpha \in \mathbb{R}$, for $t > \tau$, we have

$$F_{Z^{(\pi)}_t(s,y,\alpha)}(x) = F_{\hat{Z}^{(\hat{\pi})}_t(s,y+\alpha)}(x),$$

and for $t \leq \tau$, we have

$$F_{Z^{(\pi)}_t(s,y,0)}(x) \leq F_{\hat{Z}^{(\hat{\pi})}_t(s,y)}(x + \eta),$$

where $Z^{(\pi)}_t$ (resp., $\hat{Z}^{(\hat{\pi})}_t$) is the return of $\mathcal{M}$ (resp., $\hat{\mathcal{M}}$) from step t for policy π (resp., $\hat{\pi}$).

Proof. We prove by backwards induction on t. The base case $t = T$ follows by definition (and since the reward measure does not change from $\mathcal{M}$ to $\hat{\mathcal{M}}$). For $t > \tau$, we have

$$\begin{aligned} F_{Z^{(\pi)}_t(s,y,\alpha)}(x) &= \sum_{a \in A} \sum_{s' \in S} \int \pi_t(a \mid s, y, \alpha) \cdot P(s' \mid s, a) \cdot F_{Z^{(\pi)}_{t+1}(s',y+r,\alpha)}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\ &= \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y + \alpha) \cdot P(s' \mid s, a) \cdot F_{\hat{Z}^{(\hat{\pi})}_{t+1}(s',y+\alpha+r)}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\ &= F_{\hat{Z}^{(\hat{\pi})}_t(s,y+\alpha)}(x), \end{aligned}$$

where the second line follows by induction and by the definition of π. Next, for $t = \tau$, we have

$$\begin{aligned} F_{Z^{(\pi)}_t(s,y,0)}(x) &= \sum_{a \in A} \sum_{s' \in S} \int \pi_t(a \mid s, y, 0) \cdot P(s' \mid s, a) \cdot F_{Z^{(\pi)}_{t+1}(s',y+r,\phi(r)-r)}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\ &= \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y) \cdot P(s' \mid s, a) \cdot F_{\hat{Z}^{(\hat{\pi})}_{t+1}(s',y+\phi(r))}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\ &= \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y) \cdot P(s' \mid s, a) \cdot F_{\hat{Z}^{(\hat{\pi})}_{t+1}(s',y+\phi(r))}(x - \phi(r) + \phi(r) - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\ &\leq \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y) \cdot P(s' \mid s, a) \cdot F_{\hat{Z}^{(\hat{\pi})}_{t+1}(s',y+\phi(r))}(x - \phi(r) + \eta) \cdot d\mathbb{P}_{R(s,a)}(r) \\ &= \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y) \cdot P(s' \mid s, a) \cdot F_{\hat{Z}^{(\hat{\pi})}_{t+1}(s',y+\rho)}(x - \rho + \eta) \cdot d\mathbb{P}_{R(s,a)} \circ \phi^{-1}(\rho) \\ &= F_{\hat{Z}^{(\hat{\pi})}_t(s,y)}(x + \eta), \end{aligned}$$

where the first line uses the update from $\alpha_t = 0$ to $\alpha_{t+1} = r - \phi(r)$ on this step, the second line follows by induction and by the definition of π, the fourth line follows by monotonicity of $F_{\hat{Z}_{t+1}^{(\hat{\pi})}(s',y+r)}$, and the fifth line follows by a change of variables $\rho = \phi(r)$. For $t < \tau$, we have

$$\begin{aligned}
F_{Z_t^{(\pi)}(s,y,0)}(x) &= \sum_{a \in A} \sum_{s' \in S} \int \pi_t(a \mid s, y, 0) \cdot P(s' \mid s, a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r,0)}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y) \cdot P(s' \mid s, a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r,0)}(x - r) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&\leq \sum_{a \in A} \sum_{s' \in S} \int \hat{\pi}_t(a \mid s, y) \cdot P(s' \mid s, a) \cdot F_{\hat{Z}_{t+1}^{(\hat{\pi})}(s',y+r,0)}(x - r + \eta) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= F_{\hat{Z}_t^{(\hat{\pi})}(s',0)}(x + \eta),
\end{aligned}$$

where the second line follows by the definition of π, and the third line follows by induction. The claim follows. □

Lemma B.2. *For any monotonically increasing F, we have $F^\dagger(F(x)) \leq x$, and $F(F^\dagger(\tau)) \geq \tau$.*

Proof. See Proposition 1 in [18]. □

Lemma B.3. *Let $F, G : \mathbb{R} \to \mathbb{R}$ be monotonically increasing. If $F(x) \leq G(x + \eta)$ for all $x \in \mathbb{R}$, then we have $F^\dagger(\tau) \geq G^\dagger(\tau) - \eta$ for all $\tau \in \mathbb{R}$.*

Proof. By assumption, $G(x) \geq F(x - \eta)$. Substituting $x = F^\dagger(\tau) + \eta$ into this formula, we obtain

$$G(F^\dagger(\tau) + \eta) \geq F(F^\dagger(\tau)) \geq \tau,$$

where the second inequality follows by Lemma B.2. Also by Lemma B.2, since G is monotonically increasing, so is $G^\dagger$, so we can apply $G^\dagger$ to each side of the inequality to obtain

$$G^\dagger(\tau) \leq G^\dagger(G(F^\dagger(\tau) + \eta)) \leq F^\dagger(\tau) + \eta,$$

where the second inequality follows by Lemma B.2. The claim follows. □

Lemma B.4. *Consider the same setup as in Lemma B.1. Let $\hat{\pi}$ be a policy for $\hat{\mathcal{M}}$, and let π be the policy defined in Lemma B.1 that adapts $\hat{\pi}$ to $\mathcal{M}$. Then, we have*

$$\Phi(\pi) \geq \hat{\Phi}(\hat{\pi}) - \eta,$$

where Φ is the objective for $\mathcal{M}$ and $\hat{\Phi}$ is the objective for $\hat{\mathcal{M}}$.

Proof. Let $\hat{Z} = \hat{Z}_1^{(\hat{\pi})}(s_1, 0)$ and $Z = Z_1^{(\pi)}(s_1, 0, 0)$. Applying Lemma B.3 to the inequality in Lemma B.1, we have

$$F_Z^\dagger(\tau) \geq F_{\hat{Z}}^\dagger(\tau) - \eta.$$

Integrating this inequality, we have

$$\Phi(\pi) = \int F_Z^\dagger(\tau) \cdot dG(\tau) \geq \int \left(F_{\hat{Z}}^\dagger(\tau) - \eta \right) \cdot dG(\tau) = \hat{\Phi}(\hat{\pi}) - \eta,$$

as claimed. □

Lemma B.5. *Consider the same setup as in Lemma B.1. Given any policy π for $\mathcal{M}$, define the policy*

$$\hat{\pi}_t(a \mid s, y, \alpha) = \pi_t(a \mid s, y + \alpha)$$

for $\hat{\mathcal{M}}$, where we initialize $\alpha_1 = 0$, and we update $\alpha_{\tau+1} = r$ on step τ, where r is a random variable with probability measure

$$\mathbb{P}_{R(s,a)}(r \mid \phi(r) = \rho),$$

and $\alpha_{t+1} = \alpha_t$ otherwise. Then, for all $x, y, \alpha \in \mathbb{R}$, for $t > \tau$, we have

$$F_{\hat{Z}_t^{(\hat{\pi})}(s,y,\alpha)}(x) = F_{Z_t^{(\pi)}(s,y+\alpha)}(x),$$

and for $t \leq \tau$, we have

$$F_{\hat{Z}_t^{(\hat{\pi})}(s,y,0)}(x) \leq F_{Z_t^{(\pi)}(s,y)}(x).$$

Proof. We prove by backwards induction on T. The base case $t = T$ follows by definition. For $t > \tau$, we have

$$\begin{aligned}
F_{\hat{Z}_t^{(\hat{\pi})}(s,y,\alpha)}(x) &= \sum_{a\in A}\sum_{s'\in S}\int \hat{\pi}_t(a \mid s,y,\alpha)\cdot P(s' \mid s,a)\cdot F_{\hat{Z}_{t+1}^{(\pi)}(s',y+\rho,\alpha)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(\rho)\\
&= \sum_{a\in A}\sum_{s'\in S}\int \pi_t(a \mid s,y+\alpha)\cdot P(s' \mid s,a)\cdot F_{Z_{t+1}^{(\pi)}(s',y+\alpha+\rho)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(\rho)\\
&= F_{Z_t^{(\pi)}(s,y+\alpha)}(x),
\end{aligned}$$

where the second line follows by induction and by the definition of π. Next, for $t = \tau$, we have

$$\begin{aligned}
&F_{\hat{Z}_t^{(\hat{\pi})}(s,c,0)}(x)\\
&= \sum_{a\in A}\sum_{s'\in S}\int \hat{\pi}_t(a \mid s,y,0)\cdot P(s' \mid s,a)\cdot F_{\hat{Z}_{t+1}^{(\hat{\pi})}(s',y+\rho,r-\rho)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(r \mid \phi(r)=\rho)\cdot d\mathbb{P}_{R(s,a)}\circ\phi^{-1}(\rho)\\
&= \sum_{a\in A}\sum_{s'\in S}\int \pi_t(a \mid s,y)\cdot P(s' \mid s,a)\cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(r \mid \phi(r)=\rho)\cdot d\mathbb{P}_{R(s,a)}\circ\phi^{-1}(\rho)\\
&\le \sum_{a\in A}\sum_{s'\in S}\int \pi_t(a \mid s,y)\cdot P(s' \mid s,a)\cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\cdot d\mathbb{P}_{R(s,a)}(r \mid \phi(r)=\rho)\cdot d\mathbb{P}_{R(s,a)}\circ\phi^{-1}(\rho)\\
&= \sum_{a\in A}\sum_{s'\in S}\int \pi_t(a \mid s,y)\cdot P(s' \mid s,a)\cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\cdot d\mathbb{P}_{R(s,a)}(r)\\
&= F_{Z_t^{(\pi)}(s,y)}(x),
\end{aligned}$$

where the second line uses the update from $\alpha_t = 0$ to $\alpha_{t+1} = r - \rho$ on this step, the third line follows by induction and by the definition of $\hat{\pi}$, the fourth line follows by monotonicity of $F_{Z_{t+1}^{(\pi)}(s',y+r)}$, and the fifth line follows by the definition of conditional probability—in particular,

$$\begin{aligned}
&\int F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\cdot d\mathbb{P}_{R(s,a)}(r \mid \phi(r)=\rho)\cdot d\mathbb{P}_{R(s,a)}\circ\phi^{-1}(\rho)\\
&= \int \frac{\int F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\cdot \mathbb{1}(r\in\phi^{-1}(\rho))\cdot d\mathbb{P}_{R(s,a)}(r)}{\int \mathbb{1}(r'\in\phi^{-1}(\rho))\cdot d\mathbb{P}_{R(s,a)}(r')}\cdot d\mathbb{P}_{R(s,a)}\circ\phi^{-1}(\rho)\\
&= \int F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\int \frac{\cdot\mathbb{1}(r\in\phi^{-1}(\rho))}{\int \mathbb{1}(r'\in\phi^{-1}(\rho))\cdot d\mathbb{P}_{R(s,a)}(r')}\cdot d\mathbb{P}_{R(s,a)}\circ\phi^{-1}(\rho)\cdot d\mathbb{P}_{R(s,a)}(r)\\
&= \int F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\sum_{i=1}^{\infty}\frac{\cdot\mathbb{1}(r\in B_i)}{\mathbb{P}(R(s,a)\in B_i)}\cdot\mathbb{P}(R(s,a)\in B_i)\cdot d\mathbb{P}_{R(s,a)}(r)\\
&= \int F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\cdot d\mathbb{R}_{R(s,a)}(r),
\end{aligned}$$

where in the third line, $B_i = (\eta\cdot(i-1), \eta\cdot i]$. For $t < \tau$, we have

$$\begin{aligned}
F_{\hat{Z}_t^{(\hat{\pi})}(s,y,0)}(x) &= \sum_{a\in A}\sum_{s'\in S}\int \hat{\pi}_t(a \mid s,y,0)\cdot P(s' \mid s,a)\cdot F_{\hat{Z}_{t+1}^{(\hat{\pi})}(s',y+\rho,0)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(\rho)\\
&= \sum_{a\in A}\sum_{s'\in S}\int \pi_t(a \mid s,y)\cdot P(s' \mid s,a)\cdot F_{\hat{Z}_{t+1}^{(\hat{\pi})}(s',y+\rho,0)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(\rho)\\
&\le \sum_{a\in A}\sum_{s'\in S}\int \pi_t(a \mid s,y)\cdot P(s' \mid s,a)\cdot F_{Z_{t+1}^{(\pi)}(s',y+\rho,0)}(x-\rho)\cdot d\mathbb{P}_{R(s,a)}(\rho)\\
&= F_{Z_t^{(\pi)}(s',y,0)}(x),
\end{aligned}$$

where the second line follows by the definition of π, and the third line follows by induction. The claim follows. □

Next, we prove two lemmas showing a converse—namely, a lower bound on the value of a policy π for $\mathcal{M}_{\tau-1}$ when adapted to $\mathcal{M}_\tau$.

Lemma B.6. *Consider the same setup as in Lemma B.1. Letting π be a policy for $\mathcal{M}$, and $\hat{\pi}$ be the policy defined in Lemma B.5 that adapts π to $\hat{\mathcal{M}}$. Then, we have*

$$\hat{\Phi}(\hat{\pi}) \geq \Phi(\pi),$$

where $\hat{\Phi}$ is the objective for $\hat{\mathcal{M}}$ and Φ is the objective for $\mathcal{M}$.

Proof. Let $Z = Z_1^{(\pi)}(s_1, 0)$ and $\hat{Z} = Z_1^{(\hat{\pi})}(s_1, 0, 0)$. Applying Lemma B.3 to the inequality in Lemma B.5, we have

$$F_{\hat{Z}}^{\dagger}(\tau) \geq F_Z^{\dagger}(\tau).$$

Integrating this inequality, we have

$$\hat{\Phi}(\hat{\pi}) = \int F_{\hat{Z}}^{\dagger}(\tau) \cdot dG(\tau) \geq \int F_Z^{\dagger}(\tau) \cdot dG(\tau) = \Phi(\pi),$$

as claimed. □

Finally, we prove Theorem 3.2. Let π_T^T be the optimal policy for $\mathcal{M}_T$, and let π_τ^T be the policy defined in Lemma B.4 adapting π_τ^T from $\mathcal{M}_\tau$ to $\mathcal{M}_{\tau-1}$ for each $\tau \in [T]$.

$$\Phi_0(\pi_0^T) \geq \Phi_1(\pi_1^T) - \eta \geq \Phi_2(\pi_2^T) \geq ... \geq \Phi_T(\pi_T^T) - T \cdot \eta,$$

where each inequality follows by Lemma B.4. Similarly, let π_0^0 be the optimal policy for $\mathcal{M}_0$, and let π_τ^0 be the policy defined in Lemma B.6 adapting $\pi_{\tau-1}^0$ from $\mathcal{M}_{\tau-1}$ to $\mathcal{M}_\tau$. Then, we have

$$\Phi_T(\pi_T^0) \geq \Phi_{T-1}(\pi_{T-1}^0) \geq ... \geq \Phi_0(\pi_0^0),$$

where each inequality follows by Lemma B.6. Furthermore, by optimality of π_T^T for Φ_T, we also have $\Phi_T(\pi_T^T) \geq \Phi_T(\pi_T^0)$; together, these three inequalities imply

$$\Phi_0(\pi_0^T) \geq \Phi_0(\pi_0^0) - T \cdot \eta.$$

Finally, note that $\pi_0^0 = \pi_{\tilde{\mathcal{M}}}^*$ is the optimal policy for $\tilde{\mathcal{M}} = \mathcal{M}_0$, and $\pi_T^0 = \pi_{\hat{\mathcal{M}}}$ is π_0^0 adapted to $\hat{\mathcal{M}}$; also, $\Phi_0 = \Phi_{\tilde{\mathcal{M}}}$ is the objective for $\tilde{\mathcal{M}}$. Thus, we have

$$\Phi_{\tilde{\mathcal{M}}}(\pi_{\hat{\mathcal{M}}}) \geq \Phi_{\tilde{\mathcal{M}}}(\pi_{\tilde{\mathcal{M}}}^*) - T \cdot \eta.$$

By Theorem 3.1, the optimal policy for $\tilde{\mathcal{M}}$ equals the optimal history-dependent policy for the original MDP $\mathcal{M}$, so the claim follows. □

C Proof of Lemmas for Section 5

C.1 Proof of Lemma 5.2

Proof. First, by the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality and a union bound, for each $k \in [K]$, conditioned on $\{N^{(k)}(s,a)\}_{s \in \mathcal{S}, a \in \mathcal{A}}$, with probability at least $1 - \delta/(2K)$, we have

$$\|F_{\tilde{R}^{(k)}(s,a)} - F_{R(s,a)}\|_\infty \leq \epsilon_R^{(k)}(s,a) \qquad (\forall s \in \mathcal{S}, a \in \mathcal{A}).$$

Similarly, by Hoeffding's inequality, an ℓ_1 concentration bound for multinomial distribution, and a union bound, for each $k \in [K]$, conditioned on $\{N^{(k)}(s,a)\}_{s \in \mathcal{S}, a \in \mathcal{A}}$, we have

$$\|\tilde{P}^{(k)}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \leq \epsilon_P^{(k)}(s,a) \quad (\forall s \in \mathcal{S}, a \in \mathcal{A}).$$

Thus, both of these bounds hold for all $k \in [K]$ with probability at least $1 - \delta$. The claim follows. □

C.2 Proof of Lemma 5.3

Proof. First, we prove that for all policies π, we have

$$\|F_{Z'^{(\pi)}} - F_{Z^{(\pi)}}\|_\infty \leq B(\pi).$$

To this end, let $G(r) = F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)$; note that $G(-\infty) = 1$ and $G(\infty) = 0$. Then, by integration by parts, we have

$$\begin{aligned}
F_{Z_t^{(\pi)}(s,y)}(x) &= \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot P(s' \mid s,a) \cdot G(r) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= -\int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot P(s' \mid s,a) \cdot F_{R(s,a)}(r) \cdot dG(r),
\end{aligned}$$

and similarly for $F_{Z_t'^{(\pi)}(s,y)}(x)$. Next, note that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} |F_{Z_t'^{(\pi)}(s,y)}(x) - F_{Z_t^{(\pi)}(s,y)}(x)| \\
&= \sup_{x\in\mathbb{R}} \Bigg| \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \left(P'(s' \mid s,a) - P(s' \mid s,a)\right) \int F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x-r) d\mathbb{P}'_{R'(s,a)}(r) \\
&\qquad + \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) P(s' \mid s,a) \int \left(F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x-r) - F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \right) d\mathbb{P}'_{R'(s,a)}(r) \Bigg| \\
&\qquad - \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) P(s' \mid s,a) \int \left(F_{R'(s,a)}(r) - F_{R(s,a)}(r)\right) dF_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \\
&\leq \sup_{x\in\mathbb{R}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot |P'(s' \mid s,a) - P(s' \mid s,a)| \\
&\qquad + \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \int \sup_{x'\in\mathbb{R}} |F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x') - F_{Z_{t+1}^{(\pi)}(s',y+r)}(x')| \cdot d\mathbb{P}'_{R'(s,a)}(r) \\
&\qquad + \sum_{a\in\mathcal{A}} \pi(a \mid s,y) \cdot \sup_{r'\in\mathbb{R}} |F_{R'(s,a)}(r') - F_{R(s,a)}(r')| \\
&\leq \mathbb{E}\left[\epsilon_P(s,a) + \epsilon_R(s,a) + \sup_{x'\in\mathbb{R}} |F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x') - F_{Z_{t+1}^{(\pi)}(s',y+r)}(x')| \right].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\epsilon_t^{(\pi)} &:= \mathbb{E}\left[\sup_{x\in\mathbb{R}} |F_{Z_t'^{(\pi)}(s,y)}(x) - F_{Z_t^{(\pi)}(s,y)}(x)| \right] \\
&= \mathbb{E}\left[\epsilon_{s,a}^{(k,P)} + \epsilon_{s,a}^{(k,R)} + \sup_{x'\in\mathbb{R}} |F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x') - F_{Z_{t+1}^{(\pi)}(s',y+r)}(x')| \right] \\
&\leq \mathbb{E}\left[\epsilon_P(s,a) + \epsilon_R(s,a)\right] + \epsilon_{t+1}^{(k,\pi)} \\
&= \mathbb{E}\left[\sum_{\tau=t}^{T} \epsilon_P(s_\tau, a_\tau) + \epsilon_R(s_\tau, a_\tau) \right],
\end{aligned}$$

where the last step follows by induction. Finally, we have

$$\begin{aligned}
|\Phi'(\pi) - \Phi(\pi)| &= \left| \int_0^T \left(G(F_{Z'^{(\pi)}}(x)) - G(F_{Z^{(\pi)}}(x))\right) \cdot dx \right| \\
&\leq L_G \int_0^T |F_{Z'^{(\pi)}}(x) - F_{Z^{(\pi)}}(x)| \cdot dx \\
&\leq T \cdot L_G \cdot \epsilon_1^{(\pi)},
\end{aligned}$$

where the first line follows by Lemma 5.1. The claim follows since $\epsilon_1^{(\pi)}$ equals the desired bound. □

C.3 Proof of Lemma 5.7

Proof. First, we prove that $F_{\hat{Z}_t^{(k,\pi)}(s,y)}(x) \leq F_{Z_t^{(\pi)}(s,y)}(x)$. The case $s = s_\infty$ is straightforward, since its transitions and rewards are equal in $\mathcal{M}$ and $\hat{\mathcal{M}}$, and it only transitions to itself. For $s \neq s_\infty$, we prove by induction on t. The base case $t = T$ follows by definition. Then, we have

$$\begin{aligned}
F_{\hat{Z}_t^{(k,\pi)}(s,y)}(x) &= \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \hat{P}^{(k)}(s' \mid s,a) \cdot F_{\hat{Z}_{t+1}^{(k,\pi)}(s',y+r)}(x-r) \cdot d\hat{\mathbb{P}}_{\hat{R}^{(k)}(s,a)}(r) \\
&\leq \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \hat{P}^{(k)}(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \cdot d\hat{\mathbb{P}}_{\hat{R}^{(k)}(s,a)}(r) \\
&= \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \hat{P}^{(k)}(s' \mid s,a) \cdot F_{\hat{R}^{(k)}(s,a)}(x'-x) \cdot dF_{Z_{t+1}^{(\pi)}(s',y+r)}(x') \\
&\leq \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \hat{P}^{(k)}(s' \mid s,a) \cdot F_{R(s,a)}(x'-x) \cdot dF_{Z_{t+1}^{(\pi)}(s',y+r)}(x') \\
&= \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \hat{P}^{(k)}(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \cdot d\mathbb{P}_{R(s,a)}(r),
\end{aligned} \tag{2}$$

where the second line follows by induction, the third by integration by parts and substituting $x' = x - r$, the fourth since $F_{R(s,a)}(r) = 1 = F_{\hat{R}^{(k)}(s,a)}(r)$ for $r \geq 1$, and for $r < 1$, on event $\mathcal{E}$, we have

$$F_{R(s,a)}(r) \geq \max\left\{F_{\tilde{R}^{(k)}(s,a)}(r) - \epsilon_R^{(k)}(s,a), 0\right\} = F_{\hat{R}^{(k)}(s,a)}(r),$$

and the fifth by integration by parts and substituting $r = x' - x$. Next, since $s \neq s_\infty$, we have

$$\hat{P}^{(k)}(s_\infty \mid s,a) = 1 - \sum_{s'\in\mathcal{S}\setminus\{s_\infty\}} \hat{P}^{(k)}(s' \mid s,a) = \sum_{s'\in\mathcal{S}\setminus\{s_\infty\}} P(s' \mid s,a) - \hat{P}^{(k)}(s' \mid s,a),$$

so we can decompose the summand $\hat{P}^{(k)}(s_\infty \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s_\infty,y+r)}(x-r)$ (i.e., $s' = s_\infty$) in (2) and distribute it across the other summands; in particular, the summands $s' \neq s_\infty$ become

$$\begin{aligned}
&\hat{P}^{(k)}(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) + \left(P(s' \mid s,a) - \hat{P}^{(k)}(s' \mid s,a)\right) \cdot F_{Z_{t+1}^{(\pi)}(s_\infty,y+r)}(x-r) \\
&\leq \hat{P}^{(k)}(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) + \left(P(s' \mid s,a) - \hat{P}^{(k)}(s' \mid s,a)\right) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \\
&= P(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r),
\end{aligned} \tag{3}$$

where the second line follows since $F_{Z_{t+1}^{(\pi)}(s_\infty,y+r)}(x-r) \leq F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)$ for all $s' \neq s_\infty$, and since $P(s' \mid s,a) - \hat{P}^{(k)}(s' \mid s,a) \geq 0$ on event $\mathcal{E}$. Continuing from (2), we have

$$\begin{aligned}
F_{\hat{Z}_t^{(k,\pi)}(s,y)}(x) &\leq \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s,y) \cdot \hat{P}^{(k)}(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&\leq \int \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}\setminus\{s_\infty\}} \pi(a \mid s,y) \cdot P(s' \mid s,a) \cdot F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r) \cdot d\mathbb{P}_{R(s,a)}(r) \\
&= F_{Z_t^{(\pi)}(s,y)}(x),
\end{aligned} \tag{4}$$

where the second line follows by distributing the summand $s' = s_\infty$ and applying (3). Since $\mathcal{M}$ and $\hat{\mathcal{M}}$ have the same initial state distribution, we have $F_{\hat{Z}^{(k,\pi)}}(x) \leq F_{Z^{(\pi)}}(x)$. By Lemma 5.1, we have

$$\hat{\Phi}^{(k)}(\pi) = T - \int_0^T G(F_{\hat{Z}^{(k,\pi)}}(x)) \cdot dx \geq T - \int_0^T G(F_{Z^{(\pi)}}(x)) \cdot dx = \Phi(\pi),$$

where the inequality follows from (4) and since G is monotone. The claim follows. □

C.4 Proof of Theorem 4.1

Proof. We prove Theorem 4.1. Note that on event $\mathcal{E}$, we have

$$
\begin{aligned}
&\text{regret}(\mathfrak{A}) \\
&= \sum_{k=1}^{K} \Phi(\pi^*) - \Phi(\hat{\pi}^{(k)}) \\
&\leq \sum_{k=1}^{K} \hat{\Phi}^{(k)}(\pi^*) - \Phi(\hat{\pi}^{(k)}) \\
&\leq \sum_{k=1}^{K} \hat{\Phi}^{(k)}(\hat{\pi}^{(k)}) - \Phi(\hat{\pi}^{(k)}) \\
&\leq \sum_{k=1}^{K} 2T \cdot L_G \cdot |\mathcal{S}| \cdot B^{(k)}(\hat{\pi}^{(k)}) \\
&= 2TL_G \sqrt{5|\mathcal{S}|^3 \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)} \cdot \mathbb{E}_{\Xi_T^{(\hat{\pi}^{(1:K)})}} \left[\sum_{k=1}^{K} \sum_{t=1}^{T} \frac{1}{\sqrt{N^{(k)}(s_t, a_t)}} \;\middle|\; \{N^{(k)}(s,a)\}_{k \in [K], s \in \mathcal{S}, a \in \mathcal{A}} \right],
\end{aligned}
$$

where the first inequality follows by Lemma 5.7, the second follows by optimality of $\hat{\pi}^{(k)}$ for $\hat{\Phi}^{(k)}$, and the third by Lemma 5.6. Furthermore, note that

$$
\begin{aligned}
\sum_{k=1}^{K} \sum_{t=1}^{T} \frac{1}{\sqrt{N^{(k)}(s_t, a_t)}} &\leq \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{\mathbb{1}((s,a) \in \mathcal{W}_k)}{\sqrt{N^{(k)}(s,a)}} \\
&= T \cdot \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \frac{\mathbb{1}((s,a) \in \mathcal{W}_k)}{\sqrt{N^{(k)}(s,a)}} \\
&\leq T \cdot \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \frac{1}{\sqrt{k}} \\
&\leq 2T \cdot |\mathcal{S}| \cdot |\mathcal{A}| \cdot \sqrt{K}.
\end{aligned}
$$

The claim follows. □

D The Optimal Policy for CVaR Objectives

In this section, we describe how to compute the optimal policy for the CVaR objective when the MDP is known; this approach is described in detail in [13]. In particular, consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, D, P, \mathbb{P}, T)$, and our goal is to compute a policy π that maximizes its CVaR objective.

Step 1: CVaR objective. We begin by rewriting the CVaR objective in a form that is more amenable to optimization. First, we have the following key result:

Lemma D.1. *For any random variable Z, we have*

$$
\mathit{CVaR}_\alpha(Z) = \min_{\rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \mathbf{E}_Z \left[(Z - \rho)^+ \right] \right\},
$$

where the minimum is achieved by $\rho^* = \mathit{VaR}(Z)$.

As a consequence of this lemma, we have

$$
\begin{aligned}
\max_{\pi \in \Pi} \text{CVaR}(Z^{(\pi)}) &= \max_{\pi \in \Pi} \inf_{\rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho)^+ \right] \right\} \\
&= \inf_{\rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \max_{\pi \in \Pi} \mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho)^+ \right] \right\}.
\end{aligned}
$$

Thus, we have

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho^*)^+ \right],$$

where

$$\rho^* = \arg\min_{\rho \in \mathbb{R}} J(\rho) \qquad \text{where} \qquad J(\rho) = \rho + \frac{1}{1-\alpha} \cdot \max_{\pi \in \Pi} \mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho)^+ \right].$$

The main challenge is evaluating the maximum over $\pi \in \Pi$ in $J(\rho)$. To do so, we construct another MDP whose objective is $\mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho)^+ \right]$ for the appropriate choice of initial state distribution.

Step 2: Construct alternative MDP. The MDP we construct is $\tilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{D}, \tilde{P}, \tilde{R}, T)$, where the states are $\tilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}$, the (time-varying, deterministic) rewards $\tilde{R} : \tilde{\mathcal{S}} \times [T] \to \mathbb{R}$ are

$$\tilde{R}((s,r),t) = \begin{cases} \max\{r, 0\} & \text{if } t = T \\ 0 & \text{otherwise,} \end{cases}$$

and the transitions are

$$\tilde{P}((s',r') \mid (s,r),a) = P(s' \mid s,a) \times \mathbb{P}_{R(s,a)}(r' - r),$$

noting that $\tilde{P}$ is a (conditional) probability measure since the state space $\tilde{\mathcal{S}}$ includes a continuous component; in practice, we discretize the continuous component of the state space.

Step 3: Value iteration. Letting S_1 be the random initial state of the original MDP $\mathcal{M}$ (with distribution D), we have

$$\mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho)^+ \right] = \mathbb{E}_{S_1} \left[\tilde{V}_1^{(\pi)}((S_1, -\rho)) \right],$$

where $\tilde{V}_1^{(\pi)}$ is the value function of policy π for MDP $\tilde{\mathcal{M}}$ on step $t = 1$. Thus, we have

$$\max_{\pi \in \Pi} \mathbb{E}_{Z^{(\pi)}} \left[(Z^{(\pi)} - \rho)^+ \right] = \mathbb{E}_{S_1} \left[\tilde{V}_1^*((S_1, -\rho)) \right],$$

where $\tilde{V}_1^*$ is the value function of the optimal policy for $\tilde{\mathcal{M}}$. Intuitively, this strategy works because the augmented component of the state space r captures the cumulative reward so far plus its initial value $-\rho$; then, by the definition of $\tilde{R}$, the reward is r^+, which implies that $\tilde{V}_1^{(\pi)}((s, -\rho))$ is the expectation of the random variable $(Z^{(\pi)} - \rho)^+$. Thus, we can compute $\max_{\pi \in \Pi} \mathbb{E}_{Z^{(\pi)}} [(Z^{(\pi)} - \rho)^+]$ by performing value iteration on $\tilde{\mathcal{M}}$ to compute $\tilde{V}_1^{(\pi)}$. In particular, we have

$$\tilde{V}_T^*((s,r)) = \max\{r, 0\},$$

and

$$\tilde{V}_t^*((s,r)) = \max_{a \in A} \int \tilde{V}_{t+1}^*((s',r')) \cdot d\tilde{P}((s',r') \mid (s,r),a)$$

for all $t \in \{1, ..., T-1\}$. Then, given an initial state s_1, we construct state $\tilde{s}_1 = (s_1, -\rho^*)$, where

$$\rho^* = \arg\min_{\rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \tilde{V}_1^{(\pi)}((s, -\rho)) \right\},$$

and then acting optimally in $\tilde{\mathcal{M}}$ according to $\tilde{V}_t^*$.

E Experiments

We consider a classic frozen lake problem with a finite horizon. The agent moves to a block next to its current state at each timestep t and has a slipping probability of 0.1 in its moving direction if the next state is an ice block. The objective is to maximize the cumulative reward without falling into holes. The agent needs to choose among paths which correspond to different levels of risk and rewards. In other words, the agent should account for the tradeoff between the cumulative reward and risk of slipping into holes. We use a map with four paths of the same lengths that have different rewards at

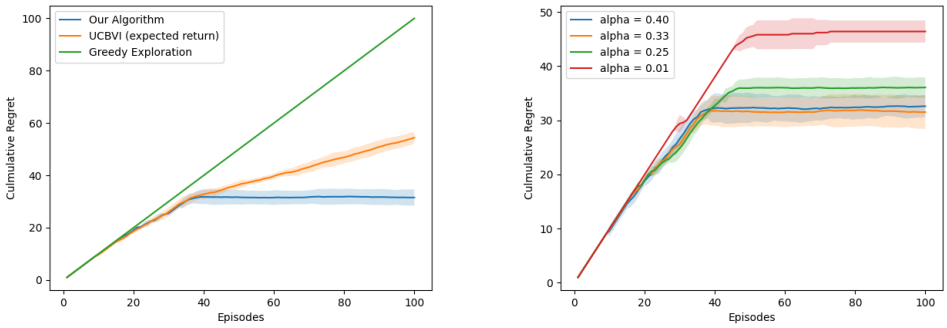

Figure 1: Results on the frozen lake environment. Left: Regret of our algorithm vs. UCBVI (with expected return) and a greedy exploration strategy. Right: Regret of our algorithm across different α values. We show mean and standard deviation across five random seeds.

the end and different levels of risk of falling into holes. We consider $\alpha \in \{0.40, 0.33, 0.25, 0.01\}$, which correspond to optimal policies of choosing paths with best possible returns of $\{6, 4, 2, 1\}$ and success probabilities of $\{0.729, 0.81, 0.9, 1\}$, respectively (failure corresponds to zero return).

Figure 1 (left) shows the comparison in cumulative regret between our algorithm, UCBVI (which maximizes expected returns, not our risk-sensitive objective), and the an algorithm that optimizes our risk-sensitive objective but explores in a greedy way (i.e., use the best policy for the current estimated MDP without any optimism), for $\alpha = 0.33$. The regret is measured in terms of the CVaR objective with respect to the optimal policy for the same CVaR objective. While UCBVI outperforms greedy, neither of them converge; in contrast, our algorithm converges within 40 episodes.

Figure 1 (right) compares the regret between our algorithm under different values of α using the CVaR objective. Note that smaller values of α tend to lead our algorithm to converge more slowly; this result matches our theory since smaller α corresponds to larger L_G. Intuitively, more samples are needed to get a good estimate of the objective as α becomes small since the CVaR objective is the average return over a tiny fraction of samples, causing high variance in our estimate of the objective.