# PAC SAMPLE COMPLEXITY OF CVAR-CONSTRAINED REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

High-stakes and safety-critical applications, such as financial investment and robotics, require sequential decision making that controls tail risk while optimizing average performance. Motivated by this fact, this paper studies Conditional Value-at-Risk (CVaR)-constrained reinforcement learning (RL) in the PAC setting. The goal is to minimize the expected cumulative performance cost subject to a CVaR constraint on the cumulative risk cost. We first prove the existence of an optimal policy that is Markovian on the augmented state space. Based on this property, we propose an efficient three-phase algorithm CVaR-PDS. This algorithm first solves linearly-constrained RL subproblems indexed by a grid of Value-at-Risk (VaR) candidates, then estimates a global model for subproblem solution evaluation, and finally performs post-hoc feasibility filtering and policy selection. We establish an $\widetilde{\mathcal{O}}(\frac{S^2AH^4\tau}{\alpha^2\varepsilon^3} + \frac{SAH^5\tau}{\alpha^2\varepsilon^4} + \frac{H^7\tau}{\zeta_\xi^2\alpha^2\varepsilon^3})$ sample complexity guarantee for algorithm CVaR-PDS, where $\tau$ is the constraint threshold, $\alpha$ is the risk level, and $\zeta_\xi$ is the Slater margin, together with an $\Omega(\frac{SAH^2}{\alpha\varepsilon^2})$ sample complexity lower bound. To the best of our knowledge, our work provides the first non-asymptotic (finite-sample) results for CVaR-constrained RL.

## 1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 2018) is a fundamental framework for sequential decision making, where an agent interacts with an unknown environment and learns to optimize long-term performance. Classical RL algorithms focus on the risk-neutral criterion of minimizing the expected cumulative cost. However, merely optimizing the expected cost fails to handle the risk control requirement in safety-critical applications, such as autonomous driving (Kiran et al., 2022), robotics (Brunke et al., 2022), and finance (Bai et al., 2025).

The above facts motivate the study of risk-sensitive RL (Mihatsch & Neuneier, 2002; García & Fernández, 2015; Gu et al., 2024), which goes beyond expected cost optimization by controlling the tail of the cumulative cost distribution. A standard tool for quantifying tail risk is the Conditional Value-at-Risk (CVaR), which admits the Rockafellar–Uryasev variational formulation (Rockafellar & Uryasev, 2000; 2002)

$$\mathrm{CVaR}_\alpha(X) \;=\; \min_{\nu\in\mathbb{R}}\big\{\nu + \tfrac{1}{\alpha}\,\mathbb{E}[(X-\nu)^+]\big\}, \tag{1}$$

where $(x)^+ := \max\{x, 0\}$ and $\alpha \in (0, 1]$ is the risk level. A $\nu$ that achieves the minimum can be chosen as the Value-at-Risk (VaR), i.e., the $1-\alpha$ quantile of $X$. Intuitively, $\mathrm{CVaR}_\alpha$ represents the expected value of the upper $\alpha$-tail of the cost distribution. Existing risk-sensitive RL studies (Bastani et al., 2022; Wang et al., 2023; Ni et al., 2024) have considered CVaR as the objective in RL.

However, in many real-world applications such as financial investment (Benati & Conde, 2022) and robotics (Hakobyan et al., 2019), the goal is often to optimize average performance while controlling rare catastrophic outcomes. This motivates using CVaR as a constraint on a separate cumulative risk cost. Therefore, we study the following CVaR-constrained RL problem:

$$\min_\pi\ \mathbb{E}_\pi\left[\sum_{h=1}^{H} z(s_h, a_h)\right] \quad \text{s.t.} \quad \mathrm{CVaR}_\alpha(C^\pi) \le \tau,$$



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

where $z$ is the performance cost function and $C^\pi$ is the cumulative risk cost, formulating the performance objective and risk control requirement, respectively. While there have been several prior works studying CVaR-constrained RL, they either show only asymptotic convergence to a locally optimal policy (Chow & Ghavamzadeh, 2014; Chow et al., 2018) or do not provide guarantees on the output policies (Prashanth, 2014; Kim & Oh, 2022; Zhang et al., 2025). None of the existing works establishes a non-asymptotic (finite-sample) guarantee relative to a globally optimal policy.

In this work, we investigate CVaR-constrained RL in the probably approximately correct (PAC) framework (Strehl et al., 2009; Dann et al., 2017) in episodic tabular MDPs with *unknown* performance costs, risk costs, and transition dynamics. The agent aims to output an $\varepsilon$-suboptimal and $\varepsilon$-feasible policy with probability at least $1-\delta$, using as few episodes as possible. For this problem, we first prove that there exists an optimal policy that is stochastic and Markovian on the augmented state space incorporating the cumulative risk cost so far. This allows us to restrict attention to this policy class rather than the entire class of history-dependent policies, and serves as a structural basis of our algorithm design and analysis.

Using the Rockafellar–Uryasev representation, we can formulate CVaR-constraint RL problem as a linearly-constrained RL subproblem with a fixed VaR value. This suggests applying primal-dual methods for CMDPs (Efroni et al., 2020), but two challenges remain. First, the VaR value can vary continuously, yielding infinitely many subproblems and making it computationally infeasible to solve them all. Second, Slater's condition may fail for some subproblems because no policy strictly satisfies the constraint.

To resolve these difficulties, we propose a computationally and sample efficient algorithm CVaR-PDS. First, we discretize risk costs to obtain finitely many VaR candidates and transfer the resulting guarantees to the original MDP with a controlled discretization error. CVaR-PDS then solves the subproblems via primal-dual learning, learns a global empirical model employing reward-free exploration (RFE) (Jin et al., 2020; Ménard et al., 2021), and finally evaluates the subproblem solutions and selects a policy under the learned model. Moreover, we establish a PAC and sample complexity analysis for CVaR-PDS, which requires Slater's condition only for the subproblem indexed by the VaR of the optimal policy on the discretized MDP, rather than all subproblems. Our fine-grained analysis of the post-hoc policy selection scheme regarding feasibility and optimality removes the need for Slater's condition for every subproblem, and overcomes the non-convexity difficulty of this problem. All proofs in this paper are deferred to Appendix.

Our contributions are summarized as follows.

1. We study CVaR-constrained RL under the PAC framework in episodic tabular MDPs. As a basis for our algorithm and analysis, we prove the existence of an optimal policy that is stochastic and Markovian on the augmented state space.
2. We design an efficient three-phase algorithm CVaR-PDS. It first performs optimistic primal-dual mirror descent on the linearly-constrained RL subproblems indexed by candidate VaRs, and estimates a global model using RFE for subproblem solution evaluation. Then, CVaR-PDS adopts a post-hoc policy selection scheme to filter out infeasible policies and pick a near-optimal policy.
3. We establish an $\widetilde{\mathcal{O}}(\frac{S^2AH^4\tau}{\alpha^2\varepsilon^3}+\frac{SAH^5\tau}{\alpha^2\varepsilon^4}+\frac{H^7\tau}{\zeta_\xi^2\alpha^2\varepsilon^3})$ sample complexity upper bound for algorithm CVaR-PDS, where $\zeta_\xi$ is the Slater margin, as well as an $\Omega(\frac{SAH^2}{\alpha\varepsilon^2})$ sample complexity lower bound. To the best of our knowledge, these are the first non-asymptotic (finite-sample) upper and lower bounds for CVaR-constrained RL.
4. Our analysis involves several novel ideas. For the upper bound, we couple the primal-dual and RFE-based evaluation errors to ensure that a near-optimal policy is retained in the feasibility set and derive the PAC guarantee. For the lower bound, we construct hard instances with informative states lying in the worst $\alpha$-fraction of trajectories, and embed the simultaneous feasibility and optimality requirements into local optimal policy identification problems over stochastic policies.

## 2 Related Work

**CVaR-Constrained RL.** Early approaches for CVaR-constrained RL use iterative planning and stochastic approximation methods (Borkar & Jain, 2014), and policy gradient and actor-critic updates

with asymptotic convergence to local optima (Chow & Ghavamzadeh, 2014; Prashanth, 2014; Chow et al., 2018). Deep approaches address CVaR-based safety constraints (Kim & Oh, 2022; Ying et al., 2022; Yang et al., 2023; Zhang et al., 2025). Kim et al. (2024) establish asymptotic optimality for spectral risk constraints, including CVaR, under assumptions including finite augmented state and outer parameter spaces. Lee et al. (2025) study optimized certainty equivalent (OCE) constraints, including CVaR, on per-stage rewards under discounted occupancy measures, and derive iteration bounds for approximate stationarity of a minimax relaxation. Ghosh & Moharrami (2025) prove sublinear regret and constraint violation for an entropic risk constraint, while the complete regret and violation results for their outlined CVaR extension remain open. None of the above works provides a rigorous non-asymptotic guarantee relative to a globally optimal policy for CVaR-constrained RL.

**RL with CVaR Objectives.** A separate line of works treats CVaR or broader risk measures as the objective rather than a constraint. Following the state augmentation approach in (Bäuerle & Ott, 2011), Keramati et al. (2020); Bastani et al. (2022); Wang et al. (2023) develop optimistic methods for tabular CVaR objectives. Du et al. (2023) study iterated CVaR, Ni et al. (2024) study reward-free exploration with the CVaR objective, and Kim & Min (2026) develop predictive CVaR Q-learning. CVaR-objective methods have also been analyzed with function approximation (Chen et al., 2024a; Zhao et al., 2024; Chen et al., 2024b). Broader risk measures considered as objectives in RL include Lipschitz dynamic risk measures (Liang & Luo, 2024) and OCEs (Wang et al., 2025).

**Reward-Free Exploration.** Reward-free exploration (RFE) (Jin et al., 2020; Kaufmann et al., 2021; Ménard et al., 2021; Zhang et al., 2021) explores an unknown environment without reward signals to learn a transition model that enables near-optimal planning for arbitrary reward functions. Our algorithm employs RFE to learn a global empirical model to evaluate the objective and constraint values of subproblem solutions.

## 3 Problem Formulation

We consider an episodic tabular Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, z, C)$, where $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, and $H$ is the horizon length of each episode. $P$ is the unknown transition kernel. $z : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the unknown performance cost function, so that $z(s, a)$ represents the expected performance cost if action $a$ is taken in state $s$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $C(\cdot \mid s, a)$ is the unknown risk cost distribution supported on $[0, 1]$. The interaction between the environment and agent is as follows. In each episode $k$, the agent chooses a policy $\pi^k$ and starts from a fixed initial state $s_1^k \equiv s_1$. At each step $h \in [H]$, the agent observes the current state $s_h^k$ and selects an action $a_h^k$ according to $\pi_h^k$. Then, it observes a random performance cost $z_h^k \in [0, 1]$ with mean $z(s_h^k, a_h^k)$, a random risk cost $c_h^k \sim C(\cdot \mid s_h^k, a_h^k)$, and transitions to a next state $s_{h+1}^k \sim P(\cdot \mid s_h^k, a_h^k)$. Conditional on the current state-action pair, the performance cost, risk cost, and next state are independent.

Let $\Pi_{\text{HS}}$ denote the class of history-dependent stochastic policies. Each policy in $\Pi_{\text{HS}}$ specifies a distribution on $\mathcal{A}$ given the complete history before each action, including the past trajectory, performance costs, and risk costs. For any $\pi \in \Pi_{\text{HS}}$, define the expected cumulative performance cost and the cumulative random risk cost as

$$f(\pi) \coloneqq \mathbb{E}_\pi \left[ \sum_{h=1}^{H} z(s_h, a_h) \right], \qquad C^\pi \coloneqq \sum_{h=1}^{H} c_h,$$

respectively, where $(s_h, a_h) \sim \pi$ and $c_h \sim C(\cdot \mid s_h, a_h)$ for any $h \in [H]$.

**Definition 3.1** ($(\delta, \varepsilon)$-PAC Guarantee)**.** Consider the *CVaR-constrained RL* problem

$$\min_{\pi \in \Pi_{\text{HS}}} f(\pi) \qquad \text{s.t. } \mathrm{CVaR}_\alpha(C^\pi) \le \tau, \tag{2}$$

where $\alpha \in (0, 1]$ is the risk level and $\tau \in [0, H]$ is the constraint threshold. We assume that Eq. (2) is feasible throughout the paper. Let $\pi^*$ denote an optimal policy of Eq. (2). Given an accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta \in (0, 1)$, we say an algorithm achieves the $(\delta, \varepsilon)$-*PAC guarantee* if, with probability at least $1 - \delta$, its output policy $\pi^{\text{out}}$ satisfies

$$f(\pi^{\text{out}}) - f(\pi^*) \le \varepsilon \quad \text{and} \quad \mathrm{CVaR}_\alpha(C^{\pi^{\text{out}}}) - \tau \le \varepsilon. \tag{3}$$

Equivalently, we say this algorithm is $(\delta, \varepsilon)$-*correct*. The sample complexity of a $(\delta, \varepsilon)$-correct algorithm is the number of episodes it uses to output a policy that satisfies Eq. (3). For notational convenience, define $g(\pi) \coloneqq \mathrm{CVaR}_\alpha(C^\pi) - \tau$. Let $\mathrm{VaR}_\alpha(X)$ denote the $(1 - \alpha)$-quantile of $X$.

# 4 State Augmentation and Markovian Recovery

Since CVaR depends on the cumulative risk cost, the original state alone may not contain enough information for an optimal decision. We therefore augment the state with the cumulative risk cost so far. Then, we show that it suffices to consider the class of stochastic policies that are Markovian on the augmented state space, instead of the entire history-dependent policy class $\Pi_{\text{HS}}$.

The resulting augmented MDP has state space $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}$. An augmented state $\widetilde{s}_h = (s_h, y_h)$ records the original state $s_h$ and the cumulative risk cost before the current step $y_h = \sum_{t=1}^{h-1} c_t$, where $\widetilde{s}_1 = (s_1, 0)$. Given the current augmented state $(s, y)$ and action $a$, the augmented next state is $(s', y + c)$, where $s' \sim P(\cdot \mid s, a)$ and $c \sim C(\cdot \mid s, a)$. After step $H$, define the terminal cumulative risk cost $y_{H+1} = \sum_{h=1}^{H} c_h$, which records the total risk cost received in this episode.

**Theorem 4.1.** *For every $\pi \in \Pi_{\text{HS}}$, there exists a stochastic and Markovian policy on the augmented state space $\widetilde{\pi} = \{\widetilde{\pi}_h : \mathcal{S} \times \mathbb{R} \to \triangle_{\mathcal{A}}\}_{h \in [H]}$ that satisfies $f(\widetilde{\pi}) = f(\pi)$ and $C^{\widetilde{\pi}} \overset{d}{=} C^{\pi}$.*

The proof demonstrates that, conditional on $h$ and $(s_h, y_h)$, the action distributions induced by $\pi$ and $\widetilde{\pi}$ coincide. The recovered Markovian policy $\widetilde{\pi}$ does not depend on the full trajectory history or individual risk costs at past steps. Since the first equality preserves the objective value and the second equality preserves the CVaR constraint value, for problem Eq. (2), *there exists an optimal policy that is stochastic and Markovian on the augmented state space $\widetilde{\mathcal{S}}$.*

# 5 A Computationally and Sample Efficient Algorithm CVaR-PDS

We propose a computationally and sample efficient algorithm CVaR-PDS, which runs on an MDP with discretized risk costs and consists of three phases: primal-dual learning, reward-free exploration (RFE), and post-hoc policy selection. Specifically, CVaR-PDS addresses the CVaR-constrained RL problem by solving linearly-constrained RL subproblems indexed by candidate VaRs, and carefully picks a near-optimal and near-feasible policy from subproblem solutions using a global empirical model learned by RFE.

## 5.1 Discretized MDP and Linearly-Constrained RL Subproblems

Since the cumulative risk cost space is continuous, to enable computational efficiency, we first introduce an MDP with discretized risk costs $\text{disc}(\mathcal{M})$ and then describe how to transform the CVaR-constrained RL problem into linearly-constrained RL subproblems.

Let $\xi \in (0, 1)$ be a discretization parameter, and define the floor-discretization map $\phi_\xi(c) \coloneqq \xi \lfloor c/\xi \rfloor$. Then, the discretized risk cost for each step lies in $\mathcal{C}_\xi \coloneqq \{\phi_\xi(c) : c \in [0, 1]\} = \{0, \xi, \ldots, \xi \lfloor 1/\xi \rfloor\}$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $C_\xi(\cdot \mid s, a)$ denote the discretized risk cost distribution such that $C_\xi(c' \mid s, a) \coloneqq \Pr_{c \sim C(\cdot \mid s,a)}(\phi_\xi(c) = c')$ for any $c' \in \mathcal{C}_\xi$. We define the MDP $\text{disc}(\mathcal{M}) = (\mathcal{S}, \mathcal{A}, H, P, z, C_\xi)$ where the risk cost is floor-discretized according to $\phi_\xi$. The cumulative discretized risk cost lies in $\mathcal{Y}_\xi \coloneqq \{0, \xi, 2\xi, \ldots, H \lfloor 1/\xi \rfloor \xi\}$.

For algorithm CVaR-PDS, we first transform $\mathcal{M}$ into $\text{disc}(\mathcal{M})$, perform learning on $\text{disc}(\mathcal{M})$ to obtain a policy $\pi$, and finally construct the adapted policy $\text{adapted}(\pi)$ for execution on the original MDP $\mathcal{M}$. Here $\text{adapted}(\pi)$ applies the decision rule of $\pi$ by replacing the original continuous risk cost on $\mathcal{M}$ with its floor-discretized counterpart. That is, at each step $h \in [H]$, $\text{adapted}(\pi)$ samples an action from $\pi_h(\cdot \mid s_h, y_h)$, observes the original risk cost $c_h$, and updates $y_{h+1} = y_h + \phi_\xi(c_h)$, where $y_1 = 0$. Note that this guarantees $y_h \in \mathcal{Y}_\xi$ and thus $\pi_h(\cdot \mid s_h, y_h)$ is always well-defined.

The following theorem converts the suboptimality and constraint violation guarantees on $\text{disc}(\mathcal{M})$ to those on $\mathcal{M}$. Let $\pi^*_{\text{disc}}$ denote an optimal policy of the CVaR-constrained RL problem on $\text{disc}(\mathcal{M})$.

**Theorem 5.1** (Suboptimality and Constraint Violation Transfer)**.** *Suppose that the policy $\pi$ learned on $\text{disc}(\mathcal{M})$ satisfies the following guarantees for some accuracy parameter $\varepsilon_0 > 0$:*

$$f(\pi) - f(\pi^*_{\text{disc}}) \le \varepsilon_0, \qquad \text{CVaR}_\alpha(C^\pi_{\text{disc}(\mathcal{M})}) - \tau \le \varepsilon_0.$$

*Then the adapted policy $\text{adapted}(\pi)$ run on the original MDP $\mathcal{M}$ satisfies*

$$f(\text{adapted}(\pi)) - f(\pi^*) \le \varepsilon_0, \qquad g(\text{adapted}(\pi)) \le \varepsilon_0 + H\xi. \tag{4}$$

With Theorem 5.1, now we focus on the CVaR-constrained RL problem on disc($\mathcal{M}$) and perform algorithm CVaR-PDS on it. Due to the feasibility of the optimal policy $\pi^*_{\text{disc}}$, it satisfies $\text{VaR}_\alpha(C^{\pi^*_{\text{disc}}}) \leq \text{CVaR}_\alpha(C^{\pi^*_{\text{disc}}}) \leq \tau$ by the definitions of VaR and CVaR. Therefore, it suffices for algorithm CVaR-PDS to consider the streamlined candidate VaR grid:

$$\mathcal{G}_\xi := \{0, \xi, 2\xi, \ldots, \lfloor \tau/\xi \rfloor \xi\},$$

rather than the whole cumulative risk cost space $\mathcal{Y}_\xi$. Let $N_\xi := |\mathcal{G}_\xi| = \lfloor \tau/\xi \rfloor + 1$. When $\tau \ll H$, e.g., in the long-horizon or strict threshold regime, the streamlined candidate VaR grid *significantly reduces the number of subproblems* CVaR-PDS needs to run on *from* $\lfloor H/\xi \rfloor + 1$ *to* $\lfloor \tau/\xi \rfloor + 1$.

Using the Rockafellar–Uryasev representation of CVaR (Eq. (1)), for each fixed candidate VaR value $\nu_i \in \mathcal{G}_\xi$, we define the linearly-constrained RL subproblem on disc($\mathcal{M}$):

$$\begin{aligned} \min_{\pi} \quad & f(\pi) := \mathbb{E}_\pi \left[ \sum_{h=1}^{H} z(s_h, a_h) \right], \\ \text{s.t.} \quad & g_i(\pi) := \nu_i + \frac{1}{\alpha} \mathbb{E}_\pi \left[ (C^\pi - \nu_i)^+ \right] - \tau \leq 0. \end{aligned} \tag{5}$$

Therefore, each subproblem Eq. (5) is convex and then primal-dual learning is applicable. Let $\nu_{j^*}$ denote the VaR of $C^{\pi^*_{\text{disc}}}$, and then it holds that $\text{CVaR}_\alpha(C^{\pi^*_{\text{disc}}}) = \nu_{j^*} + \frac{1}{\alpha}\mathbb{E}_{\pi^*_{\text{disc}}}\left[(C^{\pi^*_{\text{disc}}} - \nu_{j^*})^+\right]$. In other words, the linearly-constrained RL subproblem indexed by $\nu_{j^*}$ recovers the CVaR-constrained RL problem on disc($\mathcal{M}$). For the primal-dual analysis of subproblem $j^*$, we make the following assumption:

**Assumption 5.2** (Slater's Condition)**.** There exists a policy $\pi'$ and a constant $\zeta_\xi > 0$ such that

$$g_{j^*}(\pi') = \nu_{j^*} + \frac{1}{\alpha}\mathbb{E}_{\pi'}\Big[(C^{\pi'} - \nu_{j^*})^+\Big] - \tau \leq -\zeta_\xi.$$

Note that this assumption requires the existence of a strictly feasible solution *only for* the subproblem $j^*$ on disc($\mathcal{M}$), rather than all subproblems. This assumption holds, for example, if the optimal policy $\pi^*_{\text{disc}}$ on disc($\mathcal{M}$) is strictly feasible. Below we describe the three phases of algorithm CVaR-PDS.

### 5.2 Phase I: Optimistic Primal-Dual Learning

Algorithm 1 presents the psudeo-code of CVaR-PDS. For each candidate VaR value $\nu_i \in \mathcal{G}_\xi$, CVaR-PDS runs $K_{\text{PD}}$ episodes of optimistic primal-dual learning for the corresponding subproblem Eq. (5) (Lines 5–13). In Line 7, $\widehat{P}^{i,k}$, $\widehat{z}^{i,k}$, and $\widehat{C}^{i,k}_\xi$ are the estimated transition kernel, performance cost function, and risk cost distribution, respectively, and $N^{i,k}(s,a)$ counts the visits to $(s,a)$ before episode $k$.

For each subproblem, CVaR-PDS constructs the estimated state-action and state value functions $\widehat{W}^{i,k}_h, \widehat{U}^{i,k}_h$ for performance costs and $\widehat{Q}^{i,k}_h, \widehat{V}^{i,k}_h$ for risk costs via backward recursion with $\widehat{U}^{i,k}_{H+1} = 0$ and $\widehat{V}^{i,k}_{H+1} = (y - \nu_i)^+$ (Line 8). Specifically, here $\widehat{V}^{i,k}_1(s_1, 0)$ is an optimistic estimate of the expected positive part of the cumulative risk cost minus $\nu_i$, i.e., $\mathbb{E}_{(s_h,a_h)\sim\pi, c_h\sim C_\xi(\cdot|s_h,a_h)}[(\sum_{h=1}^{H} c_h - \nu_i)^+]$. In Line 8, we define the exploration bonuses as

$$b^{i,k}_Q(s,a) := \frac{\sqrt{2}H}{\sqrt{N^{i,k}(s,a)}} \left( \sqrt{S \log\left(\frac{16SAN_\xi K_{\text{PD}} H}{\delta}\right)} + \sqrt{\log\left(\frac{2^{\lfloor 1/\xi \rfloor + 4} SAN_\xi K_{\text{PD}} H}{\delta}\right)} \right),$$

$$b^{i,k}_W(s,a) := b^{i,k}_Q(s,a) + \sqrt{\frac{2\log(16SAN_\xi K_{\text{PD}} H/\delta)}{N^{i,k}(s,a)}}.$$

After $K_{\text{PD}}$ episodes, CVaR-PDS learns a solution $\bar{\pi}_i$ for subproblem $i$ (Line 13). In Phases II and III, it will evaluate candidate solutions $\{\bar{\pi}_i\}_{i\in[N_\xi]}$ to pick a near-optimal policy.

### 5.3 Phase II: Reward-Free Exploration

As illustrated in Algorithm 2, the Phase II of CVaR-PDS employs RFE to learn a global empirical model $(\overline{P}, \overline{z}, \overline{C}_\xi)$ for all subproblems to evaluate the objective and constraint values of candidate

**Algorithm 1** CVaR-PDS: Primal-Dual Learning with Post-Hoc Selection for CVaR-Constrained RL

1: **Input:** Target accuracy $\varepsilon$ for the original MDP $\mathcal{M}$, confidence parameter $\delta$, episode budget $K_{\text{PD}}$ for Phase I, Slater margin $\zeta_\xi$, initial Lagrange multiplier $\lambda^{\text{init}} \in [0, \frac{2H}{\zeta_\xi}]$, $\tau$, and $\alpha$.

2: **Initialization:** Set $\xi = \varepsilon/(2H)$, $\varepsilon_{\text{est}} = \alpha\varepsilon/8$, $\varepsilon_{\text{vio}} = \varepsilon/4$, $\Lambda = \frac{2H}{\zeta_\xi}$, $\eta = \sqrt{\frac{2\log A}{H^2(1+\Lambda/\alpha)^2 K_{\text{PD}}}}$, and $\eta_\lambda = \frac{\Lambda}{H(1+1/\alpha)\sqrt{K_{\text{PD}}}}$. Define $\phi_\xi(c) \coloneqq \xi\lfloor c/\xi \rfloor$ for any $c \in [0,1]$.

3: *# Phase I: Optimistic Primal-Dual Learning*

4: **for** each $\nu_i \in \mathcal{G}_\xi$ **do**

5: Initialize $\pi^{i,1}$ as the uniform policy and $\lambda^{i,1} \leftarrow \lambda^{\text{init}}$.

6: **for** episode $k = 1, \ldots, K_{\text{PD}}$ **do**

7: Update $\widehat{P}^{i,k}, \widehat{z}^{i,k}, \widehat{C}_\xi^{i,k}, N^{i,k}$. Execute $\pi^{i,k}$, apply $\phi_\xi$ to the observed risk cost on $\mathcal{M}$, and collect the trajectory.

8: Backward value iteration from $\widehat{U}_{H+1}^{i,k} = 0$ and $\widehat{V}_{H+1}^{i,k} = (y - \nu_i)^+$: For $h = H, \ldots, 1$,

$$\widehat{W}_h^{i,k}(s,y,a) \leftarrow \Big(\widehat{z}^{i,k}(s,a) + \widehat{P}^{i,k}(\cdot \mid s,a)^\top \mathbb{E}_{c\sim\widehat{C}_\xi^{i,k}(\cdot|s,a)}\Big[\widehat{U}_{h+1}^{i,k}(\cdot, y+c)\Big] - b_W^{i,k}(s,a)\Big)^+,$$

$$\widehat{Q}_h^{i,k}(s,y,a) \leftarrow \Big(\widehat{P}^{i,k}(\cdot \mid s,a)^\top \mathbb{E}_{c\sim\widehat{C}_\xi^{i,k}(\cdot|s,a)}\Big[\widehat{V}_{h+1}^{i,k}(\cdot, y+c)\Big] - b_Q^{i,k}(s,a)\Big)^+.$$

$$\widehat{U}_h^{i,k}(s,y) \leftarrow \langle \pi_h^{i,k}(\cdot \mid s,y), \widehat{W}_h^{i,k}(s,y,\cdot)\rangle, \quad \widehat{V}_h^{i,k}(s,y) \leftarrow \langle \pi_h^{i,k}(\cdot \mid s,y), \widehat{Q}_h^{i,k}(s,y,\cdot)\rangle.$$

9: For each $h \in [H]$, form $\widehat{L}_h^{i,k} \leftarrow \widehat{W}_h^{i,k} + \frac{\lambda^{i,k}}{\alpha}\widehat{Q}_h^{i,k}$.

10: For each $h \in [H]$, $\pi_h^{i,k+1}(\cdot \mid s,y) \propto \pi_h^{i,k}(\cdot \mid s,y)\exp\Big(-\eta\,\widehat{L}_h^{i,k}(s,y,\cdot)\Big)$.

11: Update: $\lambda^{i,k+1} \leftarrow \text{Proj}_{[0,\Lambda]}\Big(\lambda^{i,k} + \eta_\lambda\Big(\nu_i + \frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0) - \tau\Big)\Big)$.

12: **end for**

13: Set candidate solution $\bar{\pi}_i$ for subproblem $i$ as the uniform mixture of $\pi^{i,1}, \ldots, \pi^{i,K_{\text{PD}}}$.

14: **end for**

15: *# Phase II: Reward-Free Exploration (RFE)*

16: Run RFE (Algorithm 2) with evaluation accuracy $\varepsilon_{\text{est}}$ and confidence $\delta$ to compute $(\overline{P}, \overline{z}, \overline{C}_\xi)$.

17: *# Phase III: Post-Hoc Policy Evaluation and Selection*

18: **for** each $\nu_i \in \mathcal{G}_\xi$ **do**

19: Evaluate the candidate solution $\bar{\pi}_i$ under the globally learned model $(\overline{P}, \overline{z}, \overline{C}_\xi)$ and estimate $\bar{f}(\bar{\pi}_i)$ and $\bar{g}_i(\bar{\pi}_i)$ via Eq. (6).

20: **end for**

21: Construct the feasible solution index set $\mathcal{I}_{\text{valid}} = \big\{i \in [N_\xi] \mid \bar{g}_i(\bar{\pi}_i) \le \varepsilon_{\text{vio}} + \frac{1}{\alpha}\varepsilon_{\text{est}}\big\}$.

22: **return** $\pi^{\text{out}} = \text{adapted}(\bar{\pi}_{i^{\text{out}}})$, where $i^{\text{out}} = \arg\min_{i\in\mathcal{I}_{\text{valid}}} \bar{f}(\bar{\pi}_i)$.

solutions $\{\bar{\pi}_i\}_{i\in[N_\xi]}$ (in Phase III). Line 5 computes the uncertainty function $J_h^k(s,y,a)$ using the aggregate exploration bonus $b_{\text{RFE}}^k$ defined in Appendix D.2.1 to bound the value function estimation error. The greedy policy in Line 6 guides exploration and data collection. RFE stops and returns the learned global model after $K_{\text{stop}} - 1$ episodes (Lines 7–9). With high probability, the estimation error of the value functions for performance costs and risk costs at $(s_1, 0)$ under this model is controlled within $\varepsilon_{\text{est}}$ (we will bound $K_{\text{stop}}$ in the analysis).

### 5.4 Phase III: Post-Hoc Evaluation and Selection

Since there exists a subproblem that recovers the CVaR-constrained RL problem on $\text{disc}(\mathcal{M})$, in Phase III, CVaR-PDS selects a near-optimal and near-feasible policy from the candidate solutions $\{\bar{\pi}_i\}_{i\in[N_\xi]}$. Specifically, we re-evaluate every candidate solution in Line 19 using the global empirical model $(\overline{P}, \overline{z}, \overline{C}_\xi)$ learned in Phase II. Recall that for each $i \in [N_\xi]$, the candidate solution $\bar{\pi}_i$ is the uniform mixture of policies $\pi^{i,1}, \ldots, \pi^{i,K_{\text{PD}}}$. Let $\overline{U}_1^{i,k}(s_1,0)$ and $\overline{V}_1^{i,k}(s_1,0)$ denote the state value functions of policy $\pi^{i,k}$ for performance costs and risk costs under the global empirical model

**Algorithm 2** Reward-Free Exploration (RFE)

1: **Input:** RFE evaluation accuracy $\varepsilon_{\text{est}}$ and confidence parameter $\delta$.
2: **Initialization:** $N^1 \equiv 0$, and set $\pi^{\text{RFE},1}$ as the uniform policy.
3: **for** episode $k = 1, 2, \ldots$ **do**
4: Execute $\pi^{\text{RFE},k}$. Update $N^{k+1}$, $\widehat{P}^{k+1}$, $\widehat{z}^{k+1}$, $\widehat{C}_\xi^{k+1}$, and $b_{\text{RFE}}^{k+1}$.
5: Backward value iteration from $J_{H+1}^{k+1} \equiv 0$: For $h = H, \ldots, 1$,

$$J_h^{k+1}(s,y,a) \leftarrow \left(b_{\text{RFE}}^{k+1}(s,a) + \widehat{P}^{k+1}(\cdot \mid s,a)^\top \mathbb{E}_{c\sim\widehat{C}_\xi^{k+1}(\cdot|s,a)}\left(\max_{a'} J_{h+1}^{k+1}(\cdot, y+c, a')\right)\right) \wedge H.$$

6: Set $\pi_h^{\text{RFE},k+1}(s,y) \leftarrow \arg\max_{a'} J_h^{k+1}(s,y,a')$.
7: **if** $\max_{a'} J_1^{k+1}(s_1, 0, a') \le \varepsilon_{\text{est}}$, **then** set $K_{\text{stop}} = k+1$ and **break**.
8: **end for**
9: **Output:** $(\overline{P}, \overline{z}, \overline{C}_\xi) = (\widehat{P}^{K_{\text{stop}}}, \widehat{z}^{K_{\text{stop}}}, \widehat{C}_\xi^{K_{\text{stop}}})$.

$(\overline{P}, \overline{z}, \overline{C}_\xi)$. Then, we can estimate the objective and constraint values of $\bar{\pi}_i$ as follows:

$$\bar{f}(\bar{\pi}_i) = \frac{1}{K_{\text{PD}}} \sum_{k=1}^{K_{\text{PD}}} \overline{U}_1^{i,k}(s_1, 0), \qquad \bar{g}_i(\bar{\pi}_i) = \nu_i + \frac{1}{\alpha K_{\text{PD}}} \sum_{k=1}^{K_{\text{PD}}} \overline{V}_1^{i,k}(s_1, 0) - \tau. \tag{6}$$

The exploration strategy and stopping rule in Phase II (RFE) guarantee that with high probability, $|\bar{f}(\bar{\pi}_i) - f(\bar{\pi}_i)| \le \varepsilon_{\text{est}}$ and $|\bar{g}_i(\bar{\pi}_i) - g_i(\bar{\pi}_i)| \le \frac{1}{\alpha}\varepsilon_{\text{est}}$ for all $i \in [N_\xi]$.

Line 21 constructs a feasible solution index set $\mathcal{I}_{\text{valid}}$ using the estimated constraint values of candidate solutions in order to include $\bar{\pi}_{j^*}$. Its selection threshold $\varepsilon_{\text{vio}} + \frac{1}{\alpha}\varepsilon_{\text{est}}$ has two terms. The first term accounts for the constraint violation from Phase I. The second term covers the estimation error from Phase II with the $1/\alpha$ scaling in the Rockafellar–Uryasev formulation (Eq. (5)). This construction ensures that with high probability, $\bar{\pi}_{j^*}$ is retained in $\mathcal{I}_{\text{valid}}$ and every candidate solution contained in it has controlled true constraint violation. Line 22 then selects the candidate solution with the lowest estimated performance cost from $\mathcal{I}_{\text{valid}}$ and returns its adapted policy for execution on the original MDP $\mathcal{M}$.

## 6 Theoretical Guarantees of Algorithm CVaR-PDS

**Theorem 6.1** (Sample Complexity Upper Bound)**.** *Let* $K_{PD} = \widetilde{\mathcal{O}}\left(\frac{S(S+\lfloor \frac{2H}{\varepsilon} \rfloor + 1)AH^3 + H^4}{\alpha^2\varepsilon^2} + \frac{H^6}{\zeta_\xi^2\alpha^2\varepsilon^2}\right)$. *For any* $\delta \in (0,1)$ *and* $\varepsilon < \tau$*, with probability at least* $1-\delta$*, Algorithm 1 terminates, its output policy* adapted$(\bar{\pi}_{i^{\text{out}}})$ *is* $\varepsilon$*-optimal and* $\varepsilon$*-feasible on the original MDP* $\mathcal{M}$*, and the total number of episodes used is*

$$K_{\text{total}} = \widetilde{\mathcal{O}}\left(\frac{S^2AH^4\tau}{\alpha^2\varepsilon^3} + \frac{SAH^5\tau}{\alpha^2\varepsilon^4} + \frac{H^7\tau}{\zeta_\xi^2\alpha^2\varepsilon^3}\right).$$

Theorem 6.1 shows that algorithm CVaR-PDS effectively handles the CVaR-constrained RL problem and achieves the PAC guarantee on the original MDP $\mathcal{M}$. Specifically, the first term arises from the transition kernel and performance cost estimation. The second term is a price for estimating the discretized risk cost distribution. The third term captures the primal-dual learning cost for finding a near-optimal solution to the linearly-constrained RL subproblem in Phase I. To the best of our knowledge, this is the first non-asymptotic (finite-sample) result for CVaR-constrained RL.

*Proof sketch.* First, we analyze the primal-dual learning in Phase I. Note that the primal-dual analysis requires Slater's condition to derive the constraint violation guarantee, but the linearly-constrained RL subproblems other than that indexed by $\nu_{j^*}$ may not satisfy this condition. That is, it is only guaranteed that one output policy in these subproblems is good. Thanks to our carefully-designed post-hoc policy selection scheme, we can retain this policy (or a comparable one), and thus only need to perform the primal-dual analysis for subproblem $j^*$ and establish guarantees for $\bar{\pi}_{j^*}$ (see Corollary D.14), rather than analyze all subproblems.

For Phase II, according to the design of the uncertainty function $J_h^{k+1}$ and stopping rule, the value function estimation error under the learned global model $(\overline{P}, \overline{z}, \overline{C}_\xi)$ for any policy is controlled within $\varepsilon_{\text{est}}$. Then, we can establish the guarantees for the estimation error of the objective and constraint values for all candidate solutions $\{\bar{\pi}_i\}_{i\in[N_\xi]}$, and bound $K_{\text{stop}}$ by the progressive shrinkage of $J_h^{k+1}$ (see Theorem D.17).

Then we show how our post-hoc policy selection scheme in Phase III ensures the suboptimality and constraint violation guarantees of the output policy. With high probability, the primal-dual analysis in Phase I provides $f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) \le \varepsilon/4$ and $(g_{j^*}(\bar{\pi}_{j^*}))^+ \le \varepsilon_{\text{vio}}$. The RFE guarantees in Phase II give $\bar{g}_{j^*}(\bar{\pi}_{j^*}) \le g_{j^*}(\bar{\pi}_{j^*}) + \varepsilon_{\text{est}}/\alpha \le \varepsilon_{\text{vio}} + \varepsilon_{\text{est}}/\alpha$. Thus, by the construction of $\mathcal{I}_{\text{valid}}$, we have that $j^*$ is contained in $\mathcal{I}_{\text{valid}}$.

For suboptimality, the selection rule in Line 22 of Algorithm 1 gives $\bar{f}(\bar{\pi}_{i^{\text{out}}}) \le \bar{f}(\bar{\pi}_{j^*})$. Combining this with the estimation error $2\varepsilon_{\text{est}}$ incurred on $\bar{f}(\bar{\pi}_{i^{\text{out}}})$ and $\bar{f}(\bar{\pi}_{j^*})$, we can obtain the suboptimality guarantee. For feasibility, the filtering rule of $\mathcal{I}_{\text{valid}}$ ensures $\bar{g}_{i^{\text{out}}}(\bar{\pi}_{i^{\text{out}}}) \le \varepsilon_{\text{vio}} + \varepsilon_{\text{est}}/\alpha$. Taking into account the estimation error on $\bar{g}_{i^{\text{out}}}(\bar{\pi}_{i^{\text{out}}})$, we have $g_{i^{\text{out}}}(\bar{\pi}_{i^{\text{out}}}) \le \varepsilon_{\text{vio}} + 2\varepsilon_{\text{est}}/\alpha$. Using the definition of CVaR, we obtain $g(\bar{\pi}_{i^{\text{out}}}) \le g_{i^{\text{out}}}(\bar{\pi}_{i^{\text{out}}}) \le \varepsilon_{\text{vio}} + 2\varepsilon_{\text{est}}/\alpha$. Therefore, $\bar{\pi}_{i^{\text{out}}}$ is $\varepsilon/2$-suboptimal and $\varepsilon/2$-feasible on $\text{disc}(\mathcal{M})$.

Finally, by Theorem 5.1 and the choice of $\xi = \varepsilon/(2H)$, we obtain the PAC guarantee of CVaR-PDS on the original MDP $\mathcal{M}$. Summing up $N_\xi K_{\text{PD}}$ and $K_{\text{stop}}$ gives the sample complexity bound. □

## 7 Lower Bound for CVaR-Constrained RL

**Theorem 7.1** (Sample Complexity Lower Bound). *Consider $H > 2$, $A > 1$, $S > 5$, and $\alpha \in [\frac{1}{100}, 1]$. For any $\delta \in (0, \frac{1}{330e^4})$ and $\varepsilon \in (0, \frac{1}{160000e^4})$, there exists a family of CVaR-constrained RL instances where the expected number of episodes required by any $(\delta, \varepsilon)$-correct algorithm is $\Omega(\frac{SAH^2}{\alpha\varepsilon^2})$.*

Theorem 7.1 shows that the sample complexity for identifying a near-optimal and near-feasible policy for CVaR-constrained RL is at least linear in $S$, $A$, and $1/\alpha$, and quadratic in $H$ and $1/\varepsilon$. While there exists a gap between our upper and lower bounds, to the best of our knowledge, our work provides the first non-asymptotic upper and lower bounds for CVaR-constrained RL and takes an important step toward characterizing its optimal sample complexity.

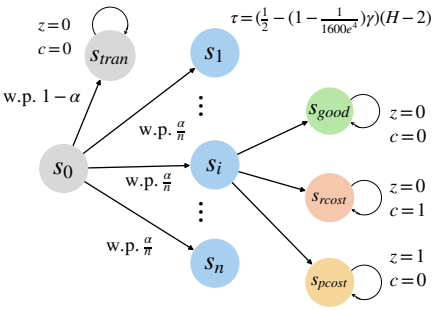



Figure 1: Hard instance for the lower bound.

*Proof sketch.* The construction of our hard instance is illustrated in Figure 1. In the first step, the agent reaches each of the $n = S-5$ bandit states $s_1, \ldots, s_n$ with probability $\frac{\alpha}{n}$, independent of the chosen action. In these bandit states, an action then leads to an absorbing state, where costs accumulate for $H-2$ steps. For each bandit state, there are two types of instances. The first one is called a reference instance, where $a_0$ has a lower risk cost but incurs a performance cost, while the other $m = A-1$ actions have higher risk but incur no performance cost. The second one is called a changed instance, which is generated via reducing the probability that one action $a_j$ reaches the risk state $s_{\text{rcost}}$ by $\gamma$ to match the risk cost of $a_0$.

With $\gamma = \Theta(\frac{\varepsilon}{H-2})$, $\varepsilon$-feasibility limits higher-risk actions on average, and $\varepsilon$-suboptimality then restricts $a_0$ at changed states. Roughly speaking, the optimal feasible action in a reference instance is $a_0$, while the optimal feasible action in a changed instance is the special $a_j$. Hence, with probability at least $1-\delta$, the algorithm needs to output a policy that assigns high probability (higher than $\frac{3}{4}$) to $a_0$ in a reference instance, while assigning low probability (lower than $\frac{1}{4}$) to $a_0$ in a changed instance for almost every bandit state $s_i$. That is, in expectation, the algorithm needs to distinguish between the two types of instances in at least half of the bandit states. However, since the two types of instances

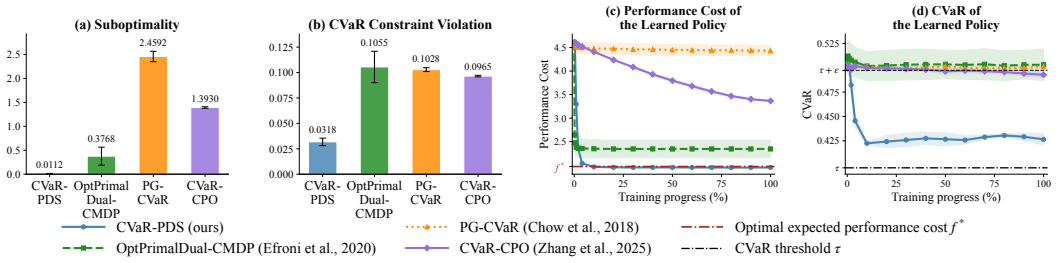



Figure 2: Subfigures (a) and (b) report the suboptimality and CVaR constraint violation of the output policies of the compared algorithms using the same total number of episodes. Subfigures (c) and (d) plot the expected cumulative performance cost and CVaR constraint violation of the learned policies during algorithm execution to illustrate the learning process. The X-axis represents the current episode as a percentage of the number of episodes used to learn the policy.

are very similar (differing only by $\gamma$ in one transition probability), we can obtain that the algorithm needs $\Omega(\frac{1}{\gamma^2})$ expected visits per state-action pair on average. Summing these bounds gives $\Omega(\frac{nm}{\gamma^2})$ expected visits on bandit states. Furthermore, recall that each episode can reach only one bandit state with probability $\alpha$, yielding $\Omega(\frac{SAH^2}{\alpha\varepsilon^2})$ expected episodes. □

## 8 Empirical Evaluations

In this section, we present empirical evaluations. We use an MDP with continuous risk costs, where $S = 9$, $A = 8$, $H = 12$, $\alpha = 0.05$, $\tau = 0.397323$, $\varepsilon = 0.1$, and $\delta = 0.005$. The performance costs are Bernoulli and risk costs follow mixtures of uniform distributions. Each evaluated algorithm uses the same total number of episodes and is independently run 10 times. We report means with 95% confidence intervals across runs. The details of the used instance and algorithm implementation are deferred to Appendix A.

We compare our algorithm CVaR-PDS with a classic constrained RL algorithm OptPrimalDual-CMDP (Efroni et al., 2020) and two CVaR-constrained RL algorithms PG-CVaR (Chow et al., 2018) and CVaR-CPO (Zhang et al., 2025). Figures 2 (a) and (b) show that CVaR-PDS achieves the lowest suboptimality in the objective and CVaR constraint violation among the compared algorithms. From the curves in Figures 2 (c) and (d), we see that CVaR-PDS successfully reduces the CVaR constraint violation while approaching the optimal performance cost during policy learning. These empirical results validate the effectiveness of CVaR-PDS in controlling tail risk while achieving near-optimal performance.

## 9 Conclusion and Future Work

This work investigates CVaR-constrained RL in the PAC setting. We design an efficient algorithm CVaR-PDS, which solves this problem through linearly-constrained RL subproblems indexed by candidate VaRs, employs RFE to learn a global empirical model, and performs post-hoc policy evaluation and selection based on the learned model. Both upper and lower sample complexity bounds are established.

There are several interesting future directions. One direction is to close the gap between the current upper and lower bounds, which may require sharper concentration of the discretized risk cost distributions and tighter control of the Lagrange multiplier range. Another direction is to extend our algorithm and analysis to function approximation with large state and action spaces. We believe that our first non-asymptotic results for CVaR-constrained RL in tabular MDPs provide an important foundation for further investigation of more general MDP models.

### AI USE STATEMENT

In this work, we used generative AI tools for language editing, checking the manuscript and references, and assisting with research code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work primarily studies the theory of CVaR-constrained reinforcement learning. We believe that this work does not involve negative societal impacts.

### REPRODUCIBILITY STATEMENT

We state all required assumptions in the paper and present complete proofs in Appendices B–H. Appendix A describes the detailed setup and MDP instance in our experiments. We provide the code for reproducing the experiments in supplementary materials.

## REFERENCES

Yahui Bai, Yuhe Gao, Runzhe Wan, Sheng Zhang, and Rui Song. A review of reinforcement learning in financial applications. *Annual Review Of Statistics And Its Application*, 12(1):209–232, 2025.

Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36259–36269, 2022.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods Of Operations Research*, 74(3):361–379, 2011. doi: 10.1007/s00186-011-0367-0.

Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017.

S. Benati and E. Conde. A relative robust approach on expected returns with bounded CVaR for portfolio selection. *European Journal Of Operational Research*, 296(1):332–352, 2022. doi: 10.1016/j.ejor.2021.04.038.

Vivek Borkar and Rahul Jain. Risk-constrained Markov decision processes. *IEEE Transactions On Automatic Control*, 59(9):2574–2579, 2014. doi: 10.1109/TAC.2014.2309262.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review Of Control, Robotics, And Autonomous Systems*, 5(1):411–444, 2022. ISSN 2573-5144. doi: 10.1146/annurev-control-042920-020211.

Yu Chen, Yihan Du, Pihe Hu, Siwei Wang, Desheng Wu, and Longbo Huang. Provably efficient iterated CVaR reinforcement learning with function approximation and human feedback. In *International Conference on Learning Representations*, 2024a.

Yu Chen, Xiangcheng Zhang, Siwei Wang, and Longbo Huang. Provable risk-sensitive distributional reinforcement learning with general function approximation. In *International Conference on Machine Learning*, 2024b.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal Of Machine Learning Research*, 18 (167):1–51, 2018.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv Preprint arXiv:2003.02189*, 2020. URL https://arxiv.org/abs/2003.02189.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal Of Machine Learning Research*, 16(42):1437–1480, 2015.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics Of Operations Research*, 44(2):377–399, 2019.

Arnob Ghosh and Mehrdad Moharrami. Online learning in risk sensitive constrained MDP. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19406–19425. PMLR, 2025. URL https://proceedings.mlr.press/v267/ghosh25c.html.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 46(12):11216–11235, 2024.

Astghik Hakobyan, Gyeong Chan Kim, and Insoon Yang. Risk-aware motion planning and control using CVaR-constrained optimization. *IEEE Robotics And Automation Letters*, 4(4):3924–3931, 2019. doi: 10.1109/LRA.2019.2929980.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Olav Kallenberg. *Foundations of Modern Probability*. Probability Theory and Stochastic Modelling. Springer International Publishing, 3rd edition, 2021. ISBN 9783030618711. doi: 10.1007/978-3-030-61871-1.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *International Conference On Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

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, pp. 4436–4443, 2020.

Dohyeong Kim and Songhwai Oh. TRC: Trust region conditional value at risk for safe reinforcement learning. *IEEE Robotics And Automation Letters*, 7(2):2621–2628, 2022. doi: 10.1109/LRA.2022.3141829.

Dohyeong Kim, Taehyun Cho, Seungyub Han, Hojun Chung, Kyungjae Lee, and Songhwai Oh. Spectral-risk safe reinforcement learning with convergence guarantees. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92465–92498, 2024. doi: 10.52202/079017-2936.

Ju-Hyun Kim and Seungki Min. Predictive CVaR Q-learning. In *International Conference on Learning Representations*, 2026.

B. Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions On Intelligent Transportation Systems*, 23(6):4909–4926, 2022. doi: 10.1109/TITS.2021.3054625.

Jane H. Lee, Baturay Saglam, Spyridon Pougkakiotis, Amin Karbasi, and Dionysis Kalogerias. Risk-averse constrained reinforcement learning with optimized certainty equivalents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41798–41833, 2025. doi: 10.52202/085713-1250.

Hao Liang and Zhiquan Luo. Regret bounds for risk-sensitive reinforcement learning with Lipschitz dynamic risk measures. In *International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1774–1782. PMLR, 2024. URL https://proceedings.mlr.press/v238/liang24a.html.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49 (2–3):267–290, 2002.

Xinyi Ni, Guanlin Liu, and Lifeng Lai. Risk-sensitive reward-free reinforcement learning with CVaR. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37999–38017. PMLR, 2024.

L. A. Prashanth. Policy gradients for CVaR-constrained MDPs. In *International Conference On Algorithmic Learning Theory*, pp. 155–169. Springer International Publishing, 2014. doi: 10.1007/978-3-319-11662-4_12.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal Of Risk*, 2:21–41, 2000.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal Of Banking & Finance*, 26(7):1443–1471, 2002. doi: 10.1016/S0378-4266(02)00271-6.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal Of Machine Learning Research*, 10(84):2413–2444, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *International Conference on Machine Learning*, pp. 35864–35907. PMLR, 2023.

Kaiwen Wang, Dawen Liang, Nathan Kallus, and Wen Sun. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63636–63661. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25bl.html.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the $L_1$ deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003.

Qisong Yang, Thiago D. Simão, Simon H. Tindemans, and Matthijs T. J. Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3):859–887, 2023. doi: 10.1007/s10994-022-06187-8.

ChengYang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3673–3680, July 2022. doi: 10.24963/ijcai.2022/510.

Andrea Zanette and Emma Brunskill. 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.

Qiyuan Zhang, Shu Leng, Xiaoteng Ma, Qihan Liu, Xueqian Wang, Bin Liang, Yu Liu, and Jun Yang. CVaR-constrained policy optimization for safe reinforcement learning. *IEEE Transactions On Neural Networks And Learning Systems*, 36(1):830–841, January 2025. doi: 10.1109/TNNLS.2023.3331304.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pp. 12402–12412. PMLR, 2021.

Yulai Zhao, Wenhao Zhan, Xiaoyan Hu, Ho-fung Leung, Farzan Farnia, Wen Sun, and Jason D. Lee. Provably efficient CVaR RL in low-rank MDPs. In *International Conference on Learning Representations*, 2024.

# Appendix

## NOTATION

The following table summarizes the notation used throughout the paper.

| Symbol | Meaning |
|---|---|
| $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, z, C)$ | Original finite-horizon MDP with stochastic performance cost observations and risk costs. |
| $\mathcal{S}, \mathcal{A}, S, A, H$ | State and action spaces, their cardinalities, and the horizon. |
| $P(\cdot \mid s, a)$ | Unknown controlled transition kernel. |
| $z_h^k$ | Observed performance cost. |
| $c_h, C(\cdot \mid s, a)$ | Observed risk cost and its unknown stationary kernel supported on $[0, 1]$. |
| $z(s, a)$ | Unknown conditional mean performance cost. |
| $C^\pi$ | Cumulative risk cost under policy $\pi$. |
| $\alpha, \tau$ | CVaR level and constraint threshold. |
| $f(\pi), g(\pi)$ | Expected performance objective and CVaR constraint function. |
| $\Pi_{\text{HS}}, \Pi_{\text{aug}}$ | History-dependent stochastic policies and augmented-Markov policies. |
| $y_h, \tilde{s}_h = (s_h, y_h)$ | Cumulative risk coordinate and continuous augmented state. |
| $\xi, \phi_\xi$ | Discretization parameter and floor-discretization map $\phi_\xi(c) = \xi\lfloor c/\xi \rfloor$. |
| $\text{disc}(\mathcal{M})$ | Discretized finite-alphabet MDP used by the algorithm and statistical analysis. |
| $\pi^*, \pi^*_{\text{disc}}$ | Selected optimal policies for the constrained problems on $\mathcal{M}$ and $\text{disc}(\mathcal{M})$, respectively. |
| $C_\xi(\cdot \mid s, a)$ | Discretized risk cost kernel induced by $\phi_\xi$. |
| $\mathcal{Y}_\xi$ | Cumulative-risk state grid used in the augmented state space. |
| $\mathcal{G}_\xi, N_\xi$ | Truncated VaR-candidate grid and its number of grid points. |
| $\mathcal{C}_\xi$ | One-step discretized risk cost alphabet. |
| $c_{\xi,h}$ | Discretized one-step risk cost $\phi_\xi(c_h)$ observed in $\text{disc}(\mathcal{M})$. |
| $i, \nu_i, j^*$ | Subproblem index, its VaR-grid value, and the index whose grid value equals the VaR of the selected $\pi^*_{\text{disc}}$. |
| $g_i(\pi)$ | Fixed-VaR constraint function for grid point $\nu_i$. |
| $\bar{\pi}_i$ | Uniform episode-level mixture of the policies learned for subproblem $i$ in Phase I. |
| $(i, k)$ | Lexicographically ordered subproblem and local-episode pair in Phase I. |
| $K_{\text{PD}}, K_{\text{stop}}$ | Episodes per subproblem in Phase I and the stopping model index in Phase II. |
| $\widehat{P}, \widehat{z}, \widehat{C}_\xi, N$ | Empirical transition kernel, performance cost mean, discretized risk cost distribution, and visit count. |
| $(\overline{P}, \overline{z}, \overline{C}_\xi)$ | Common empirical model learned in Phase II for post-hoc evaluation. |
| $q^{i,k}, \bar{n}^{i,k}; q_h^k, q^k, \bar{n}^k$ | Conditional occupancies and cumulative pseudo-counts under the two phase-specific clocks. |
| $R_{\text{PD}}, \Delta_{\text{PD}}(\lambda)$ | Explicit inverse-visitation bound and upper bound on the accumulated primal-dual gap in Phase I. |
| $\Gamma_{\text{PD}}(K), \Gamma_{\text{RFE}}(K)$ | Logarithmic envelopes used in the explicit inversions for Phases I and II. |
| $\mathcal{F}^{\text{PD}}_{i,k}, \mathcal{F}^{\text{RFE}}_{k-1}, \mathcal{F}^{\text{LB}}_{k-1}$ | Pre-episode filtrations for Phase I, Phase II, and the lower-bound experiment. |
| $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4, \mathcal{E}$ | Model and count concentration events, and their intersection. |
| $\varepsilon, \varepsilon_0$ | Target accuracies for $\mathcal{M}$ and $\text{disc}(\mathcal{M})$, respectively. |
| $\varepsilon_{\text{est}}, \varepsilon_{\text{vio}}$ | RFE evaluation accuracy and constraint tolerance in Phase I. |
| $\delta, \delta'$ | Overall confidence parameter and per-statement allocation $\delta/8$. |
| $\Lambda, \zeta_\xi$ | Lagrange multiplier bound and known Slater margin for fixed-VaR subproblem $j^*$. |
| $\text{adapted}(\pi)$ | Policy $\pi$ adapted for execution in $\mathcal{M}$. |
| $\pi^{\text{out}}$ | Adapted policy returned by CVaR-PDS in the original MDP $\mathcal{M}$. |

Appendix A provides the experimental details. The mathematical appendices then follow the proof narrative. Appendix B establishes the augmented-state policy reduction, and Appendix C constructs the common high-probability event. Appendix D analyzes Phases I–III of CVaR-PDS on $\text{disc}(\mathcal{M})$,

Table 1: Continuous risk cost distributions at every recurrent state.

| Action | Distribution | Action | Distribution |
|---|---|---|---|
| $a_0$ | $\mathrm{Unif}[19/1000, 1/50]$ | $a_4$ | $\mathrm{Unif}[2/1000, 3/1000]$ |
| $a_1$ | $\mathrm{Unif}[17/1000, 18/1000]$ | $a_5$ | $\mathrm{Unif}[18/1000, 19/1000]$ |
| $a_2$ | $\mathrm{Unif}[0, 1/1000]$ | $a_6$ | $\mathrm{Unif}[16/1000, 17/1000]$ |
| $a_3$ | $\mathrm{Unif}[15/1000, 16/1000]$ | $a_7$ | $\mathrm{Unif}[7/2000, 9/2000]$ |

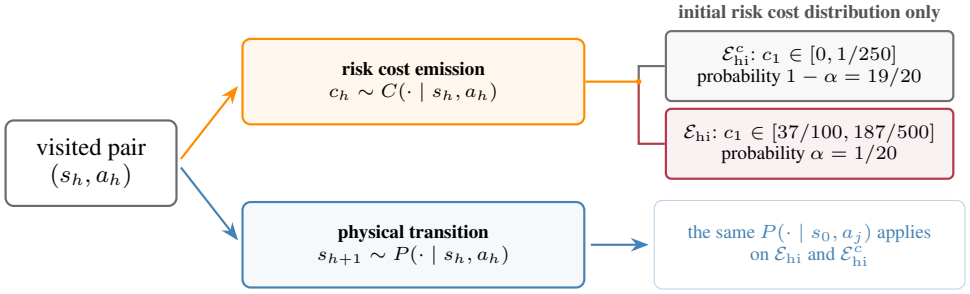



Figure 3: One-step environment generator. Conditional on $(s_h, a_h)$, the risk cost and next physical state are drawn independently. At $h = 1$, the event $\mathcal{E}_{\mathrm{hi}}$ identifies a risk cost component rather than a physical transition branch.

including the unknown-cost components in each phase. Appendix E transfers the resulting guarantee to the original MDP $\mathcal{M}$. Appendix F completes the main PAC proof, Appendix G proves the lower bound, and Appendix H collects the supporting inequalities.

# A Experimental Details

We first describe the MDP instance and report the exact LP reference values used in Figure 2. We then give the implementation and evaluation protocol under the same total number of training episodes.

## A.1 MDP Instance

We use a time-homogeneous MDP with horizon $H = 12$, initial state $s_0$, recurrent states $s_1, \ldots, s_8$, and actions $a_0, \ldots, a_7$. At each step, the learner observes a Bernoulli performance cost with unknown mean $z(s, a)$, a continuous risk cost, and the next state. Conditional on the visited state–action pair, these three draws are independent, and each step uses fresh draws.

The initial risk cost distribution is the same for every action:

$$C(\cdot \mid s_0, a) = \frac{19}{20}\mathrm{Unif}\left(0, \frac{1}{250}\right) + \frac{1}{20}\mathrm{Unif}\left(\frac{37}{100}, \frac{187}{500}\right).$$

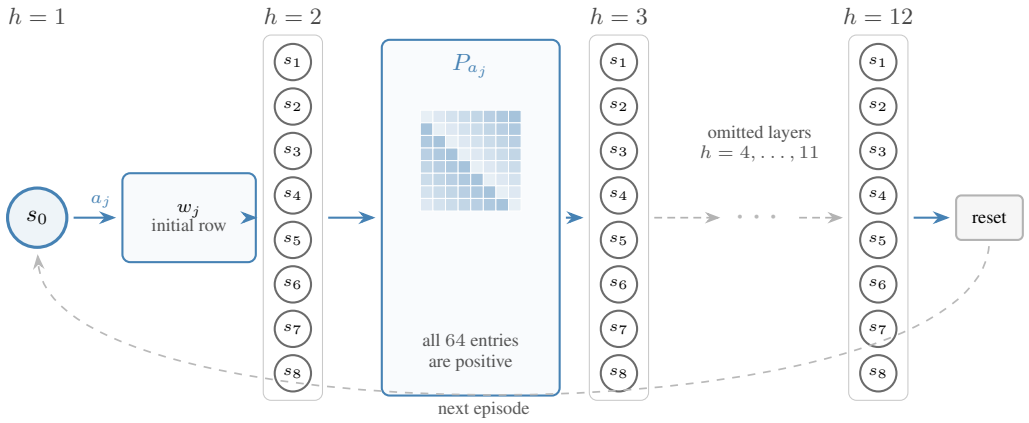
The second component has probability $\alpha = 1/20$, and the value of $c_1$ identifies which component was sampled. At recurrent states, the risk cost distribution depends only on the action and is given in Table 1.

Figure 3 depicts the risk cost emission and physical transition as separate draws. Observing the high-cost component does not add a physical state.

The physical dynamics are fully interconnected. Let $w_j(q) = b_{(q-j) \bmod 8}$, where $b = (3/8, 3/16, 1/8, 1/16, 1/16, 1/16, 1/16, 1/16)$. For $i, r, j \in \{0, \ldots, 7\}$,

$$P(s_{r+1} \mid s_0, a_j) = w_j(r), \qquad P(s_{r+1} \mid s_{i+1}, a_j) = b_{(r-i-j) \bmod 8}.$$

Thus every recurrent transition has positive probability, and both the current state and action affect subsequent occupancies. There is no return to $s_0$ within an episode. Figure 4 shows the resulting layered interaction.




| Offset | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $b_q$ | 3/8 | 3/16 | 1/8 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 |

At $s_0$, offsets refer to $s_1, \ldots, s_8$ in absolute order. At $s_{i+1}$, the current state and action rotate the same vector according to $b_{(r-i-j) \bmod 8}$.

Figure 4: Physical transitions in one episode. The initial row maps $s_0$ to all eight recurrent states. The representative block denotes the $8 \times 8$ recurrent transition matrix under action $a_j$. The omitted layers repeat the same cyclic rule.

Table 2: Bernoulli performance cost mean $z(s, a)$.

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ |
|---|---|---|---|---|---|---|---|---|
| $s_0$ | 1/3 | 1/2 | 3/4 | 7/12 | 2/3 | 4/5 | 9/10 | 19/20 |
| $s_1$ | 7/50 | 1/2 | 13/25 | 43/100 | 21/100 | 13/25 | 53/100 | 49/100 |
| $s_2$ | 12/25 | 1/10 | 27/100 | 3/10 | 1/4 | 7/20 | 29/100 | 11/50 |
| $s_3$ | 49/100 | 27/50 | 1/10 | 7/20 | 41/100 | 9/25 | 17/50 | 19/50 |
| $s_4$ | 57/100 | 51/100 | 23/50 | 4/25 | 12/25 | 29/100 | 59/100 | 6/25 |
| $s_5$ | 1/4 | 47/100 | 17/50 | 23/100 | 4/25 | 21/100 | 11/50 | 57/100 |
| $s_6$ | 1/5 | 41/100 | 51/100 | 11/20 | 11/20 | 9/50 | 11/20 | 27/100 |
| $s_7$ | 11/50 | 9/25 | 23/50 | 47/100 | 53/100 | 27/50 | 9/50 | 2/5 |
| $s_8$ | 1/5 | 19/100 | 9/25 | 8/25 | 39/100 | 29/100 | 19/100 | 9/50 |

Table 2 gives the unknown Bernoulli performance cost means used by the simulator.

For this finite-horizon instance, an occupancy-measure linear program computes an optimal policy $\pi^*$ for the original MDP $\mathcal{M}$ exactly. The resulting policy has $f(\pi^*) = 1.972065$ and $g(\pi^*) = \mathrm{CVaR}_{0.05}(C^{\pi^*}) - \tau = 0$, with $\mathrm{CVaR}_{0.05}(C^{\pi^*}) = \tau = 0.397323$, matching the horizontal reference lines in Figure 2(c,d).

### A.2 Algorithm Implementation

We train all four methods under a common episode budget and evaluate their final policies in the original MDP $\mathcal{M}$. We set $\varepsilon = 0.1$, $\delta = 0.005$, and $\xi = \varepsilon/(2H) = 1/240$. The cumulative risk state grid $\mathcal{Y}_\xi$ contains 2881 points. Without truncation, the VaR-candidate grid would also contain $\lfloor H/\xi \rfloor + 1 = 2881$ points, whereas truncation at $\tau$ leaves only $N_\xi = \lfloor \tau/\xi \rfloor + 1 = 96$ candidates. More generally, the number of fixed-VaR subproblems solved in Phase I scales with $\tau/\xi$ rather than $H/\xi$, and is therefore much smaller when $\tau \ll H$ in long-horizon or strict-constraint regimes. As explained in Section 5, candidates above $\tau$ are infeasible and are not part of $\mathcal{G}_\xi$. Phase I uses $K_{\mathrm{PD}} = 300$ episodes per VaR candidate, and Phase II runs a fixed segment of $K_{\mathrm{RFE}}^{\max} = 200$ episodes. Each CVaR-PDS trial therefore uses $96 \cdot 300 + 200 = 29{,}000$ training episodes. This fixed segment in Phase II is an engineering budget rather than the theoretical stopping time $K_{\mathrm{stop}}$.

We compare four methods at the same 29,000-episode budget. CVaR-PDS reports the episode-level mixture of iterates from Phase I at the selected VaR candidate. OptPrimalDual-CMDP enforces

Table 3: Frozen settings for comparison under the same total number of training episodes.

| Method | Settings |
|---|---|
| CVaR-PDS | $K_{\mathrm{PD}} = 300$, $K_{\mathrm{RFE}}^{\max} = 200$, $\eta_{\mathrm{OMD}} = 3$, $\eta_\lambda = 1$, $\Lambda = 24$, initial multiplier 0, and bonus scale $3 \times 10^{-6}$ |
| OptPrimalDual-CMDP | One update per episode, $\eta = \eta_\lambda = 1$, $\Lambda = 24$, and initial multiplier 0 |
| PG-CVaR | One trajectory per update. Policy, VaR, and multiplier scales are $(1, 0.5, 0.5)$ with decay exponents $(0.75, 0.60, 0.90)$. VaR and multiplier initialized at 0 |
| CVaR-CPO | Batch size 20, 64 critic quantiles, critic window 512, risk-index stride 8, temperature 0.2, KL radius 0.001, VaR step 0.01, multiplier step 0.1, and initial multiplier 20 |

expected risk and reports the corresponding mixture of training iterates. PG-CVaR and CVaR-CPO are finite-horizon tabular adaptations of the CVaR policy-gradient and trust-region methods, respectively. Each reports its final current augmented policy. These adaptations do not reproduce the original continuous-control implementations or guarantees.

The displayed $\Lambda$, learning rates, $K_{\mathrm{PD}}$, and bonus scale are frozen empirical choices. They are not the theorem's parameter schedule. The frozen empirical bonus profile also differs from the bonuses in Algorithms 1 and 2. For Phase I, let $n = N(s,a) \vee 1$ and

$$L_{\mathrm{PD}}^{\mathrm{emp}} = \log\left(\frac{2SA|\mathcal{Y}_\xi|K_{\mathrm{PD}}H}{\delta/4}\right).$$

Before applying the scale $3 \times 10^{-6}$, the transition, performance, and risk bonus components are

$$\left(H\sqrt{\frac{2SL_{\mathrm{PD}}^{\mathrm{emp}}}{n}}, \sqrt{\frac{L_{\mathrm{PD}}^{\mathrm{emp}}}{2n}}, H\sqrt{\frac{2|\mathcal{C}_\xi|L_{\mathrm{PD}}^{\mathrm{emp}}}{n}}\right).$$

Here $|\mathcal{Y}_\xi| = 2881$ and $|\mathcal{C}_\xi| = 241$. Thus, the first logarithm uses the full cumulative-risk grid size, not the 96 retained VaR candidates. For Phase II, the same profile uses

$$L_{\mathrm{RFE}}^{\mathrm{emp}}(n) = \log\left(\frac{3SAH}{\delta/4}\right) + S\log(8e(n+1))$$

and, before the same scale, the three components are

$$\left(H\sqrt{\frac{2L_{\mathrm{RFE}}^{\mathrm{emp}}(n)}{n}} + \frac{2HL_{\mathrm{RFE}}^{\mathrm{emp}}(n)}{3n}, \sqrt{\frac{L_{\mathrm{RFE}}^{\mathrm{emp}}(n)}{2n}}, H\sqrt{\frac{2|\mathcal{C}_\xi|L_{\mathrm{RFE}}^{\mathrm{emp}}(n)}{n}}\right).$$

The empirical uncertainty test applies the recursion in Algorithm 2 to these scaled bonuses for Phase II and the warm-start model.

The learner observes only sampled transitions, performance costs, and risk costs. The exact model is used only for the oracle and offline evaluation. For each final or checkpoint policy, we compute $f(\widehat{\pi})$ exactly under the physical transition model and estimate continuous CVaR from 100,000 independent trajectories. Checkpoint curves evaluate the current policies in the original MDP $\mathcal{M}$. The endpoint metrics are $f(\widehat{\pi}) - f(\pi^*)$ and $\mathrm{CVaR}_\alpha(C^{\widehat{\pi}}) - \tau$. Intervals are two-sided 95% Student-$t$ intervals across the ten independent trials. For CVaR-PDS, these curves retrospectively follow the local trajectory in Phase I at the VaR candidate selected after all three phases.

The CVaR-PDS configuration was calibrated on disjoint seeds 48–57. The two policy-gradient profiles were fixed on a predecessor instance. None of their screening or validation seeds appears among the reported seeds 9901–9910. The complete four-method run used five concurrent CPU workers on a 12-logical-core ARM processor with 24 GB memory. It required approximately two hours of wall-clock time and used no GPU.

## B Markovian Reduction via State Augmentation (Proof of Theorem 4.1)

*Proof of Theorem 4.1.* Related state-augmentation arguments for AVaR/CVaR objectives and constraints appear in Bäuerle & Ott (2011), Chow et al. (2018), and Bastani et al. (2022). Our direct

reduction preserves both the expected performance objective and the distribution of the cumulative risk cost.

Fix $\pi \in \Pi_{\text{HS}}$. Recall that $\widetilde{s}_h = (s_h, y_h)$ is the augmented state with cumulative risk cost $y_h = \sum_{t=1}^{h-1} c_t$, and let $\Pi_{\text{aug}}$ denote the class of time-dependent stochastic Markov policies on $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}$. For each step $h$, let $\mu_h$ be the distribution of $\widetilde{s}_h$ under $\pi$. Distinct pre-action histories can share the same augmented state, while $\pi$ may prescribe different action distributions on them. Hence, $\pi$ does not directly define an augmented-Markov rule.

The augmented state space $\mathcal{S} \times \mathbb{R}$ is standard Borel because $\mathcal{S}$ is finite and $\mathbb{R}$ is Polish, and the finite action space $\mathcal{A}$ is Borel. The disintegration theorem for conditional distributions (Kallenberg, 2021, Theorem 8.5) therefore gives a stochastic kernel $\widetilde{\pi}_h$ on $\mathcal{A}$ that is measurable in $(s, y)$ and factors the original step-$h$ joint distribution as

$$\Pr_{\pi}(s_h \in ds,\ y_h \in dy,\ a_h \in da) = \mu_h(ds, dy)\, \widetilde{\pi}_h(da \mid s, y).$$

The first factor records where $\pi$ arrives, and the second records how it acts there. Their product reconstructs the original joint distribution rather than imposing independence. More concretely, $\widetilde{\pi}_h$ averages the action choices of $\pi$ over all histories consistent with $(s, y)$. Equivalently, $\widetilde{\pi}_h(a \mid s, y) = \Pr_{\pi}(a_h = a \mid s_h = s,\ y_h = y)$ for $\mu_h$-almost every $(s, y)$. Its values on a $\mu_h$-null set may be chosen arbitrarily. Set $\widetilde{\pi} = (\widetilde{\pi}_h)_{h \in [H]} \in \Pi_{\text{aug}}$.

It remains to verify that executing $\widetilde{\pi}$ reproduces the augmented-state distribution $\mu_h$ at each step. We prove this claim by induction on $h$.

*Base case.* Both policies start at $\widetilde{s}_1 = (s_1, 0)$. Hence, their initial augmented-state distributions agree.

*Induction step.* Suppose both policies give $\widetilde{s}_h$ the same distribution $\mu_h$. Under $\pi$, the disintegration identity gives the joint distribution $\mu_h(ds, dy)\widetilde{\pi}_h(da \mid s, y)$ of $(\widetilde{s}_h, a_h)$. Under $\widetilde{\pi}$, the induction hypothesis supplies the same augmented-state distribution $\mu_h$, and the definition of $\widetilde{\pi}_h$ supplies the same conditional action kernel. The two policies therefore induce the same joint distribution of $(\widetilde{s}_h, a_h)$.

Under either policy, conditional on the complete history available before the action and on the chosen action $a_h$, the environment samples the next state and risk cost from the product kernel $P(\cdot \mid s_h, a_h)C(\cdot \mid s_h, a_h)$ by Section 3. This kernel depends on the history only through $(s_h, a_h)$. Averaging it over the matched distribution of $(\widetilde{s}_h, a_h)$ therefore gives the same distribution of $(s_{h+1}, c_h, y_h)$. Since $\widetilde{s}_{h+1} = (s_{h+1}, y_h + c_h)$, its distribution also agrees under both policies. This completes the induction.

At each $h \in [H]$, ignoring $y_h$ in the matched distribution of $(\widetilde{s}_h, a_h)$ gives the same distribution of $(s_h, a_h)$. Thus, the two policies have the same state–action occupancy measures. The induction also matches the distribution of the terminal coordinate $y_{H+1} = \sum_{h=1}^{H} c_h$.

By the definition of $f$ and the matched state–action occupancy measures, we have

$$f(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[z(s_h, a_h)] = \sum_{h=1}^{H} \mathbb{E}_{\widetilde{\pi}}[z(s_h, a_h)] = f(\widetilde{\pi}).$$

Under either policy, the cumulative risk cost equals its terminal coordinate: $C^{\pi} = y_{H+1}$ under $\pi$, and $C^{\widetilde{\pi}} = y_{H+1}$ under $\widetilde{\pi}$. The matched terminal-coordinate distributions therefore give $C^{\pi} \overset{d}{=} C^{\widetilde{\pi}}$ and hence $g(\pi) = g(\widetilde{\pi})$. Thus, every policy in $\Pi_{\text{HS}}$ can be replaced by a policy in $\Pi_{\text{aug}}$ with the same expected performance cost and cumulative risk distribution. □

## C Concentration Events and High-Probability Bounds

Our analysis uses eight concentration statements, grouped into four events and parameterized by $\delta' \in (0, 1)$. Events $\mathcal{E}_1$ and $\mathcal{E}_3$ each contain the transition, performance cost mean, and discretized risk cost distribution statements for one phase. Events $\mathcal{E}_2$ and $\mathcal{E}_4$ control visitation counts. This appendix defines all four events and proves their probability bounds before they are used in the algorithm analysis.

Order the episode pairs $(i,k) \in [N_\xi] \times [K_{\text{PD}}]$ lexicographically. Let $\mathcal{F}_{i,k}^{\text{PD}}$ be the sigma-field generated by all observations and algorithmic randomization in Phase I before episode $(i,k)$. Define

$$q^{i,k}(s,a) \coloneqq \sum_{h=1}^{H} \Pr\Big((s_h^{i,k}, a_h^{i,k}) = (s,a) \mid \mathcal{F}_{i,k}^{\text{PD}}\Big),$$

$$N^{i,k}(s,a) \coloneqq \sum_{(i',k')<(i,k)} \sum_{h=1}^{H} \mathbf{1}\{(s_h^{i',k'}, a_h^{i',k'}) = (s,a)\},$$

$$\bar{n}^{i,k}(s,a) \coloneqq \sum_{(i',k')<(i,k)} q^{i',k'}(s,a).$$

Both counts accumulate across subproblems, with $N^{1,1} = \bar{n}^{1,1} = 0$.

Phase II uses its own episode clock. Let $\mathcal{F}_{k-1}^{\text{RFE}}$ be the sigma-field containing all information available before its RFE episode $k$. Define $q_h^k(s,a) \coloneqq \Pr\big((s_h^k, a_h^k) = (s,a) \mid \mathcal{F}_{k-1}^{\text{RFE}}\big)$ for the stage-$h$ occupancy induced by $\pi^{\text{RFE},k}$ on $\text{disc}(\mathcal{M})$. Let $q^k(s,a) \coloneqq \sum_{h=1}^{H} q_h^k(s,a)$ and $\bar{n}^k(s,a) \coloneqq \sum_{j=1}^{k} q^j(s,a)$. Thus, $\bar{n}^0(s,a) = 0$ and $\bar{n}^k(s,a) = \bar{n}^{k-1}(s,a) + q^k(s,a)$ for $k \geq 1$. The empirical count $N^k(s,a)$ is measured before episode $k$. Event $\mathcal{E}_4$ therefore compares it with $\bar{n}^{k-1}(s,a)$.

In Phase I, index the transition, performance cost, and discretized risk cost observations by successive visits to each $(s,a)$ across all subproblems. Phase II uses its own observation sequences. The performance cost observations form an independent sequence with mean $z(s,a)$, and their centered differences are 1-sub-Gaussian. Conditional on $(s,a)$, the next state and discretized risk cost are independent, with distributions $P(\cdot \mid s,a)$ and $C_\xi(\cdot \mid s,a)$, respectively. Each empirical quantity at an adaptive visit count is the corresponding prefix statistic of these sequences. We therefore prove the concentration bounds simultaneously over all deterministic prefix lengths and then evaluate them at the realized counts. We do not condition on a random count. Discretization changes neither the transition dimension $S$ nor the visitation analysis because the transition kernel remains indexed by the original state-action pairs.

For $(i,k) \in [N_\xi] \times [K_{\text{PD}}]$, let $\widehat{P}^{i,k}(\cdot \mid s,a)$, $\widehat{z}^{i,k}(s,a)$, and $\widehat{C}_\xi^{i,k}(\cdot \mid s,a)$ denote the empirical transition kernel, performance cost mean, and discretized risk cost distribution, respectively. All three are computed from the $N^{i,k}(s,a)$ visits in all preceding episodes of Phase I. At positive counts, define the component bonuses

$$\begin{aligned} b_P^{i,k}(s,a) &\coloneqq H\sqrt{\frac{2S\log(2SAN_\xi K_{\text{PD}} H/\delta')}{N^{i,k}(s,a)}}, \\ b_z^{i,k}(s,a) &\coloneqq \sqrt{\frac{2\log(2SAN_\xi K_{\text{PD}} H/\delta')}{N^{i,k}(s,a)}}, \\ b_C^{i,k}(s,a) &\coloneqq H\sqrt{\frac{2((\lfloor 1/\xi \rfloor + 1)\log 2 + \log(SAN_\xi K_{\text{PD}} H/\delta'))}{N^{i,k}(s,a)}}. \end{aligned} \tag{7}$$

At zero count, set $(b_P^{i,k}, b_z^{i,k}, b_C^{i,k}) = (2H, 1, 2H)$.

In Phase II, for $N^k(s,a) > 0$, define $b_{P,\text{RFE}}^k(s,a) \coloneqq H\sqrt{\frac{2\beta(N^k(s,a),\delta')}{N^k(s,a)}} + \frac{2H}{3}\frac{\beta(N^k(s,a),\delta')}{N^k(s,a)}$, where $\beta(n,\delta') \coloneqq \log(3SAH/\delta') + S\log(8e(n+1))$. At zero count, define $b_{P,\text{RFE}}^k(s,a) \coloneqq 2H$.

At positive counts, the cost bonuses are

$$\begin{aligned} b_{z,\text{RFE}}^k(s,a) &\coloneqq \sqrt{\frac{2\log(2SA(N^k(s,a)+1)^2/\delta')}{N^k(s,a)}}, \\ b_{C,\text{RFE}}^k(s,a) &\coloneqq H\sqrt{\frac{2((\lfloor 1/\xi \rfloor + 1)\log 2 + \log(2SA(N^k(s,a)+1)^2/\delta'))}{N^k(s,a)}}. \end{aligned}$$

At zero count, set $b^k_{z,\mathrm{RFE}}(s,a) \coloneqq 1$ and $b^k_{C,\mathrm{RFE}}(s,a) \coloneqq 2H$. The subscripts follow the estimated objects: $b_z$ controls the scalar performance mean $z$, whereas $b_C$ controls the discretized risk cost distribution $C_\xi$. We use this convention in both phases.

- **Event $\mathcal{E}_1$ (Model Concentration in Phase I):** For all $(i,k) \in [N_\xi] \times [K_{\mathrm{PD}}]$ and all $(s,a)$, the following bounds hold

$$\|\widehat{P}^{i,k}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \le \begin{cases} \sqrt{\frac{2S\log(2SAN_\xi K_{\mathrm{PD}}H/\delta')}{N^{i,k}(s,a)}}, & N^{i,k}(s,a) > 0, \\ 2, & N^{i,k}(s,a) = 0, \end{cases}$$
$$|\widehat{z}^{i,k}(s,a) - z(s,a)| \le b^{i,k}_z(s,a), \qquad H\|\widehat{C}^{i,k}_\xi(\cdot \mid s,a) - C_\xi(\cdot \mid s,a)\|_1 \le b^{i,k}_C(s,a). \tag{8}$$

- **Event $\mathcal{E}_2$ (Visitation Count Concentration in Phase I):** For all $(i,k) \in [N_\xi] \times [K_{\mathrm{PD}}]$ and all $(s,a)$,

$$N^{i,k}(s,a) \ge \tfrac{1}{2}\bar{n}^{i,k}(s,a) - H\log\tfrac{SA}{\delta'}.$$

- **Event $\mathcal{E}_3$ (Time-Uniform Model Concentration in Phase II):** For all $k \in \mathbb{N}$, all $(s,a)$, and any $V : \mathcal{S} \to [0,H]$, the following bounds hold

$$\left|(\widehat{P}^k(\cdot \mid s,a) - P(\cdot \mid s,a))^\top V\right| \le b^k_{P,\mathrm{RFE}}(s,a)$$
$$|\widehat{z}^k(s,a) - z(s,a)| \le b^k_{z,\mathrm{RFE}}(s,a), \qquad H\|\widehat{C}^k_\xi(\cdot \mid s,a) - C_\xi(\cdot \mid s,a)\|_1 \le b^k_{C,\mathrm{RFE}}(s,a). \tag{9}$$

- **Event $\mathcal{E}_4$ (Count Concentration in Phase II):** For all $k \in \mathbb{N}$ and all $(s,a)$,

$$N^k(s,a) \ge \tfrac{1}{2}\,\bar{n}^{k-1}(s,a) - H\log\tfrac{SA}{\delta'}.$$

At zero count, the transition, performance cost mean, and discretized risk cost distribution bounds explicitly cover their maximal errors $2H$, $1$, and $2H$, respectively.

We define the overall good event as

$$\mathcal{E} \coloneqq \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4.$$

The lemmas below prove that each of the three statements in $\mathcal{E}_1$ and $\mathcal{E}_3$ fails with probability at most $\delta'$, while $\mathcal{E}_2$ and $\mathcal{E}_4$ each fail with probability at most $\delta'$. Their union bound will give $\Pr(\mathcal{E}) \ge 1 - 8\delta'$. We set $\delta' = \delta/8$. All lemmas and theorems that assume $\mathcal{E}$ condition on all four events.

### C.1 Model and Count Concentration Bounds

**Lemma C.1** (Transition Component of $\mathcal{E}_1$)**.** *The transition statement in event $\mathcal{E}_1$ holds with probability at least $1 - \delta'$.*

*Proof.* For any fixed state-action pair $(s,a)$ and a deterministic sample count $n \ge 1$, let $\widehat{P}_n(\cdot \mid s,a)$ be the empirical transition estimated from exactly $n$ samples. By Weissman's inequality (Weissman et al., 2003), for any $\varepsilon > 0$

$$\Pr\Big(\|\widehat{P}_n(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \ge \varepsilon\Big) \le 2^S \exp\left(-\frac{n\varepsilon^2}{2}\right).$$

Because Phase I reuses samples across all candidates, $N^{i,k}(s,a) \le N_\xi K_{\mathrm{PD}} H$. We apply a union bound over $(s,a)$ and $1 \le n \le N_\xi K_{\mathrm{PD}} H$. Substituting $\varepsilon = \sqrt{2S\log(2SAN_\xi K_{\mathrm{PD}}H/\delta')/n}$, we have

$$\Pr(\text{transition statement fails}) \le \sum_{s,a} \sum_{n=1}^{N_\xi K_{\mathrm{PD}} H} \Pr\left(\|\widehat{P}_n - P\|_1 \ge \sqrt{\frac{2S\log(2SAN_\xi K_{\mathrm{PD}}H/\delta')}{n}}\right)$$
$$\le SAN_\xi K_{\mathrm{PD}} H\, 2^S \exp\left(-S\log\frac{2SAN_\xi K_{\mathrm{PD}}H}{\delta'}\right) \le \delta',$$

where the last inequality holds for $S \geq 2$. If $S = 1$, the transition kernel is deterministic.

When $N^{i,k}(s,a) = 0$, we define the transition radius to be 2, the maximum possible categorical $L_1$ error. This explicit zero-count convention does not rely on enlarging the logarithm. □

**Lemma C.2** (Cost Components of $\mathcal{E}_1$)**.** *Each statement in Eq.* (8) *fails with probability at most* $\delta'$.

*Proof.* Fix $(s,a)$ and a deterministic positive count $n$. The sub-Gaussian tail bound for the independent centered performance cost observations gives

$$\Pr(|\widehat{z}_n(s,a) - z(s,a)| \geq r) \leq 2e^{-nr^2/2}.$$

With $M = N_\xi K_{\text{PD}} H$ and $r_n = \sqrt{2\log(2SAM/\delta')/n}$, the preceding tail is $\delta'/(SAM)$. A union bound over the $SA$ pairs and $1 \leq n \leq M$ therefore proves the performance mean statement. For the discretized risk cost distribution, Weissman's inequality gives

$$\Pr\Big(\|\widehat{C}_{\xi,n}(\cdot \mid s,a) - C_\xi(\cdot \mid s,a)\|_1 \geq r\Big) \leq 2^{\lfloor 1/\xi \rfloor + 1} e^{-nr^2/2}.$$

Substituting

$$r_n = \sqrt{\frac{2((\lfloor 1/\xi \rfloor + 1)\log 2 + \log(SAM/\delta'))}{n}}$$

makes this tail at most $\delta'/(SAM)$. The same union bound proves the discretized risk cost statement. The zero-count conventions following Eq. (7) cover the remaining cases directly. □

**Lemma C.3** (Visitation Count Concentration in Phase I)**.** *Event* $\mathcal{E}_2$ *holds with probability at least* $1 - \delta'$.

*Proof.* Fix $(s,a)$ and apply Lemma H.1 to the episodes $(i,k)$ in lexicographic order, with

$$X^{i,k} = \sum_{h=1}^{H} \mathbb{1}\{(s_h^{i,k}, a_h^{i,k}) = (s,a)\}, \qquad \mathbb{E}[X^{i,k} \mid \mathcal{F}_{i,k}^{\text{PD}}] = q^{i,k}(s,a).$$

Set $\delta_0 = \delta'/(SA)$. For each pair $(i,k)$, take the prefix consisting of the pairs preceding $(i,k)$. This gives the defining inequality of $\mathcal{E}_2$ simultaneously for all $(i,k) \in [N_\xi] \times [K_{\text{PD}}]$. A union bound over all $(s,a)$ gives $\Pr(\mathcal{E}_2^c) \leq \delta'$. □

**Lemma C.4** (Time-Uniform Count Concentration in Phase II)**.** *Event* $\mathcal{E}_4$ *holds with probability at least* $1 - \delta'$.

*Proof.* Fix $(s,a)$ and apply Lemma H.1 separately to the process generated in Phase II,

$$X_j = \sum_{h=1}^{H} \mathbb{1}\{(s_h^j, a_h^j) = (s,a)\}, \qquad \mathbb{E}[X_j \mid \mathcal{F}_{j-1}^{\text{RFE}}] = q^j(s,a),$$

with $\delta_0 = \delta'/(SA)$. Taking $n = k - 1$ gives the defining inequality of $\mathcal{E}_4$ simultaneously for all $k \in \mathbb{N}$. A union bound over all $(s,a)$ gives $\Pr(\mathcal{E}_4^c) \leq \delta'$. □

### C.2 Time-Uniform Model Concentration in Phase II

The transition bonus is evaluated at the random pre-episode count $N^k(s,a)$. We control the corresponding transition error by applying a summable fixed-count confidence allocation to the visit-indexed samples of each state-action pair.

**Lemma C.5** (Transition Component of $\mathcal{E}_3$)**.** *The transition statement in event* $\mathcal{E}_3$ *holds with probability at least* $1 - \delta'$.

*Proof.* The claim is immediate when $S = 1$, since the transition kernel is deterministic. Suppose $S \geq 2$. Fix $(s,a)$, let $X_1, X_2, \ldots$ be its visit-indexed sequence of next-state observations, and let $p_n$ be the empirical distribution of the first $n$ observations. By the successive-visit sampling assumption

above, this sequence is i.i.d. with distribution $P(\cdot \mid s, a)$. By Weissman's inequality (Weissman et al., 2003), for every deterministic $n \geq 1$

$$\Pr\left( \|p_n - P(\cdot \mid s, a)\|_1 > \sqrt{\frac{2(S \log 2 + \log(2SA(n+1)^2/\delta'))}{n}} \right) \leq \frac{\delta'}{2SA(n+1)^2}.$$

Taking a union bound over all $(s, a)$ and $n \geq 1$, and using $\sum_{n \geq 1}(n+1)^{-2} < 1$, shows that the displayed bound holds simultaneously for every state-action pair and every positive count with probability at least $1-\delta'$. We may therefore evaluate it pathwise at $n = N^k(s, a)$ without conditioning on this random count.

For $S \geq 2$, the logarithmic term satisfies

$$S \log 2 + \log \frac{2SA(n+1)^2}{\delta'} \leq \log \frac{3SAH}{\delta'} + S \log(8e(n+1)) = \beta(n, \delta').$$

Hence, for every $V : \mathcal{S} \to [0, H]$ and every positive count $n = N^k(s, a)$

$$\left| (\widehat{P}^k(\cdot \mid s, a) - P(\cdot \mid s, a))^\top V \right| \leq H \|\widehat{P}^k(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1 \leq H \sqrt{\frac{2\beta(n, \delta')}{n}} \leq b^k_{P,\mathrm{RFE}}(s, a).$$

At zero count, the categorical $L_1$ error is at most 2. The deterministic $L_1$-to-functional bound therefore gives error at most $2H = b^k_{P,\mathrm{RFE}}(s, a)$. Thus the transition statement holds simultaneously for every $k$, $(s, a)$, and bounded $V$. □

**Lemma C.6** (Cost Components of $\mathcal{E}_3$)**.** *Each statement in Eq.* (9) *fails with probability at most* $\delta'$.

*Proof.* Fix $(s, a)$ and $n \geq 1$. Hoeffding's inequality with

$$r_n = \sqrt{\frac{2 \log(2SA(n+1)^2/\delta')}{n}}$$

gives failure probability at most $\delta'/(SA(n+1)^2)$. For the discretized risk cost alphabet, whose size is $\lfloor 1/\xi \rfloor + 1$, Weissman's inequality with

$$r_n = \sqrt{\frac{2((\lfloor 1/\xi \rfloor + 1) \log 2 + \log(2SA(n+1)^2/\delta'))}{n}}$$

gives failure probability at most $\delta'/(2SA(n+1)^2)$. Summing either allocation over all $SA$ pairs and $n \geq 1$, and using $\sum_{n \geq 1}(n+1)^{-2} < 1$, proves the claimed $\delta'$ bound for each statement. The explicit zero-count values cover the maximal errors directly. Both bounds therefore hold at every realized model index without conditioning on its random count. □

Lemma C.5 also bounds the transition error for the random value function selected at $K_{\text{stop}}$.

# D PAC Analysis on disc($\mathcal{M}$)

For this analysis, fix $\xi \in (0, 1)$ and an accuracy target $\varepsilon_0 \in (0, 1]$ on disc($\mathcal{M}$). We analyze Phases I–III of Algorithm 1 with $\varepsilon_{\text{est}} = \alpha\varepsilon_0/4$ and $\varepsilon_{\text{vio}} = \varepsilon_0/2$. The choices in the main text correspond to $\varepsilon_0 = \varepsilon/2$ and $\xi = \varepsilon/(2H)$. Phases I and II control the transition, performance cost mean, and discretized risk cost distribution errors on their respective concentration events. Phase III combines the resulting candidate and evaluation guarantees.

## D.1 Phase I: Optimistic Primal-Dual Learning

The analysis follows the primal-dual structure of Efroni et al. (2020), but applies it to the stated recursion with all three unknown model components. All value functions are evaluated on $\mathcal{S} \times \mathcal{Y}_\xi$, and $C_\xi$ is the discretized risk cost distribution of disc($\mathcal{M}$).

### D.1.1 Transition Estimation Error and the Cumulative Exploration Bonus Bound

**Lemma D.1** (Transition Estimation Error in Phase I)**.** *Assume that the transition component of event $\mathcal{E}_1$ holds. Then, for any bounded value function $V : \mathcal{S} \to [0, H]$, the transition estimation error is bounded by the exploration bonus*

$$|(\widehat{P}^{i,k}(\cdot \mid s,a) - P(\cdot \mid s,a))^\top V| \le b_P^{i,k}(s,a), \quad \forall k \in [K_{PD}],\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

*Proof.* By Hölder's inequality,

$$|(\widehat{P}^{i,k}(\cdot \mid s,a) - P(\cdot \mid s,a))^\top V| \le \|\widehat{P}^{i,k}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \|V\|_\infty \le b_P^{i,k}(s,a).$$

The last inequality uses the transition component of event $\mathcal{E}_1$, which ensures $\|\widehat{P}^{i,k} - P\|_1 \le \sqrt{2S\log(2SAN_\xi K_{\mathrm{PD}} H/\delta')/(N^{i,k}(s,a) \vee 1)}$ whenever $N^{i,k}(s,a) > 0$. The zero-count case is covered by the radius 2. Therefore, $\|\widehat{P}^{i,k} - P\|_1 H \le b_P^{i,k}(s,a)$. This concludes the proof. □

**Lemma D.2** (Bound on the Expected Cumulative Exploration Bonus)**.** *Assume that event $\mathcal{E}_2$ holds. Then, the cumulative expected exploration bonus over $K_{PD}$ episodes is bounded by*

$$\begin{aligned}
&\sum_{k=1}^{K_{PD}} \sum_{h=1}^{H} \mathbb{E}_{\pi^{i,k},P,C_\xi}[b_P^{i,k}(s_h,a_h) \mid s_1, 0] \\
&\le H\sqrt{2S\log\tfrac{2SAN_\xi K_{PD} H}{\delta'}}\Big(2\sqrt{5}\sqrt{SAHK_{PD}} + 4SAH\log\tfrac{SA}{\delta'} + SAH\Big) + 2SAH^2 \\
&= \widetilde{\mathcal{O}}\Big(S\sqrt{AH^3K_{PD}} + SAH^2\sqrt{S}\Big).
\end{aligned}$$

*Proof.* Both $N^{i,k}(s,a)$ and $\pi^{i,k}$ are $\mathcal{F}_{i,k}^{\mathrm{PD}}$-measurable. Moreover, $\delta' = \delta/8$, $SA \ge 1$, and $\delta < 1$ imply $\log(SA/\delta') \ge \log 8 > 1$. Since $\bar{n}^{i,k}(s,a) \ge \sum_{k'=1}^{k-1} q^{i,k'}(s,a)$, event $\mathcal{E}_2$ supplies the count condition for the episodes of subproblem $i$. Applying Lemma H.5 with confidence parameter $\delta'$ yields

$$\begin{aligned}
&\sum_{k=1}^{K_{\mathrm{PD}}} \sum_{h=1}^{H} \mathbb{E}_{\pi^{i,k},P,C_\xi}[b_P^{i,k}(s_h,a_h) \mid s_1, 0] = \sum_{k=1}^{K_{\mathrm{PD}}} \sum_{h=1}^{H} \mathbb{E}\Big(b_P^{i,k}(s_h^{i,k}, a_h^{i,k}) \;\Big|\; \mathcal{F}_{i,k}^{\mathrm{PD}}\Big) \\
&\le H\sqrt{2S\log\frac{2SAN_\xi K_{\mathrm{PD}} H}{\delta'}} \sum_{k=1}^{K_{\mathrm{PD}}} \sum_{h=1}^{H} \mathbb{E}\left(\frac{1}{\sqrt{N^{i,k}(s_h^{i,k}, a_h^{i,k}) \vee 1}} \;\middle|\; \mathcal{F}_{i,k}^{\mathrm{PD}}\right) + 2SAH^2 \\
&\le H\sqrt{2S\log\frac{2SAN_\xi K_{\mathrm{PD}} H}{\delta'}}\Big(2\sqrt{5}\sqrt{SAHK_{\mathrm{PD}}} + 4SAH\log\tfrac{SA}{\delta'} + SAH\Big) + 2SAH^2.
\end{aligned}$$

At zero count, $b_P^{i,k}(s,a) = 2H$ and $\sqrt{N^{i,k}(s,a) \vee 1} = 1$. Since $\delta' = \delta/8$, $\delta < 1$, and $S, A, N_\xi, K_{\mathrm{PD}}, H \ge 1$, the square-root coefficient is at least $H\sqrt{2\log 16} > 2H$. Thus, the same pointwise bound covers every count before taking conditional expectations and summing. The additional $2SAH^2$ is a nonnegative relaxation. The square-root term is $\widetilde{\mathcal{O}}(S\sqrt{AH^3K_{\mathrm{PD}}})$, while the remaining count-independent terms are $\widetilde{\mathcal{O}}(SAH^2\sqrt{S})$. This gives the stated two-term bound without imposing an additional lower bound on $K_{\mathrm{PD}}$. □

### D.1.2 Optimistic Value Bounds and Extended Value Difference Decomposition

We first fix the notation for the subproblems analyzed in Phase I. For each grid point $\nu_i \in \mathcal{G}_\xi$ with $i \in [N_\xi]$, subproblem $i$ is the linearly-constrained problem in Eq. (5). Corollary E.4 and Lemma E.1 show that feasibility of the original problem implies feasibility of the constrained problem on $\mathrm{disc}(\mathcal{M})$. On the finite augmented MDP, the occupancy-measure polytope is compact and $f$ is affine. The CVaR constraint is continuous because its Rockafellar–Uryasev minimization ranges over the finite cumulative-risk state grid $\mathcal{Y}_\xi$. Hence an optimal policy for $\mathrm{disc}(\mathcal{M})$ exists. Recall from Section 5 that $\pi^*_{\mathrm{disc}}$ is such an optimal policy and $\nu_{j^*} = \mathrm{VaR}_\alpha(C^{\pi^*_{\mathrm{disc}}})$. The lemmas in this subsection and in

Appendix D.1.3 hold for any fixed subproblem $i \in [N_\xi]$ and any fixed reference policy. We introduce an optimizer only for subproblem $j^*$ before Lemma D.8. No optimizer is assumed to exist for an arbitrary grid point.

Recall the optimistic Lagrangian action-value function $\widehat{L}_h^{i,k}(s,y,a) = \widehat{W}_h^{i,k}(s,y,a) + \frac{\lambda^{i,k}}{\alpha}\widehat{Q}_h^{i,k}(s,y,a)$, which we call the *state-action Lagrangian value function*. With a slight abuse of notation, we define the corresponding *state Lagrangian value function* under the current policy $\pi^{i,k}$ as

$$\widehat{L}_h^{i,k}(s,y) \coloneqq \left\langle \widehat{L}_h^{i,k}(s,y,\cdot), \pi_h^{i,k}(\cdot \mid s,y) \right\rangle = \widehat{U}_h^{i,k}(s,y) + \frac{\lambda^{i,k}}{\alpha}\widehat{V}_h^{i,k}(s,y).$$

For episode $k$, the corresponding *true Lagrangian value function* under a policy $\pi$ is defined recursively as $L_h^{i,k,\pi}(s,y,a) \coloneqq z(s,a) + P(\cdot \mid s,a)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a)}[L_{h+1}^{i,k,\pi}(\cdot, y+c)]$, with terminal condition $L_{H+1}^{i,k,\pi}(s,y) = \frac{\lambda^{i,k}}{\alpha}(y-\nu_i)^+$. Its state analogue is $L_h^{i,k,\pi}(s,y) \coloneqq \left\langle L_h^{i,k,\pi}(s,y,\cdot), \pi_h(\cdot \mid s,y) \right\rangle = U_h^\pi(s,y) + \frac{\lambda^{i,k}}{\alpha}V_h^\pi(s,y)$. Here $V_h^\pi$ is the expected terminal cost $(y_{H+1}-\nu_i)^+$ under $\pi$ and the true model, with the fixed threshold suppressed in the notation.

**Lemma D.3** (Extended Value Difference Lemma)**.** *For any policy $\pi$ and any sequence of optimistic Lagrangian functions $\{\widehat{L}_h^{i,k}\}_{h=1}^H$, given the shared terminal condition $\widehat{L}_{H+1}^{i,k}(s,y) = L_{H+1}^{i,k,\pi}(s,y)$, the difference between the optimistic Lagrangian value and the true Lagrangian value under policy $\pi$ from the initial augmented state $(s_1,0)$ can be decomposed as follows. All expectations in this display start from $(s_1,0)$.*

$$\begin{aligned}\widehat{L}_1^{i,k}(s_1,0) - L_1^{i,k,\pi}(s_1,0) = &\sum_{h=1}^H \mathbb{E}_{\pi,P,C_\xi}\left[\left\langle \widehat{L}_h^{i,k}(\tilde{s}_h,\cdot), \pi_h^{i,k}(\cdot \mid \tilde{s}_h) - \pi_h(\cdot \mid \tilde{s}_h)\right\rangle\right] \\ &+ \sum_{h=1}^H \mathbb{E}_{\pi,P,C_\xi}\left[\widehat{L}_h^{i,k}(\tilde{s}_h,a_h) - z(s_h,a_h) - \mathbb{E}[\widehat{L}_{h+1}^{i,k}(\tilde{s}_{h+1}) \mid \tilde{s}_h, a_h]\right].\end{aligned}$$

*Proof.* By the definition of the value functions, we expand the difference at step $h$ recursively. For any augmented state $(s,y)$, adding and subtracting the term $\left\langle \widehat{L}_h^{i,k}(s,y,\cdot), \pi_h(\cdot \mid s,y)\right\rangle$ yields

$$\begin{aligned}\widehat{L}_h^{i,k}(s,y) - L_h^{i,k,\pi}(s,y) = &\left\langle \widehat{L}_h^{i,k}(s,y,\cdot), \pi_h^{i,k}(\cdot \mid s,y) - \pi_h(\cdot \mid s,y)\right\rangle \\ &+ \left\langle \widehat{L}_h^{i,k}(s,y,\cdot) - L_h^{i,k,\pi}(s,y,\cdot), \pi_h(\cdot \mid s,y)\right\rangle.\end{aligned}$$

Taking the expectation over $a_h \sim \pi_h(\cdot \mid s,y)$ in the second term, expand $L_h^{i,k,\pi}$ and add and subtract the expected next-step optimistic value. This gives

$$\begin{aligned}&\mathbb{E}_{a_h\sim\pi_h}\left[\widehat{L}_h^{i,k}(s,y,a_h) - L_h^{i,k,\pi}(s,y,a_h)\right] \\ &= \mathbb{E}_{a_h\sim\pi_h}\left[\widehat{L}_h^{i,k}(s,y,a_h) - z(s,a_h) - P(\cdot \mid s,a_h)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a_h)}[L_{h+1}^{i,k,\pi}(\cdot, y+c)]\right] \\ &= \mathbb{E}_{a_h\sim\pi_h}\left[\widehat{L}_h^{i,k}(s,y,a_h) - z(s,a_h) - P(\cdot \mid s,a_h)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a_h)}[\widehat{L}_{h+1}^{i,k}(\cdot, y+c)]\right] \\ &\quad + \mathbb{E}_{\substack{a_h\sim\pi_h(\cdot|s,y),\ s'\sim P(\cdot|s,a_h),\\ c\sim C_\xi(\cdot|s,a_h)}}\left[\widehat{L}_{h+1}^{i,k}(s',y+c) - L_{h+1}^{i,k,\pi}(s',y+c)\right].\end{aligned}$$

Unrolling this recursion from step $h=1$ down to $H$ over the trajectory generated by the reference policy $\pi$ and the true dynamics $P$, we accumulate the policy difference terms and the optimistic Bellman error terms. The unrolling leaves a final residual term at step $H+1$

$$\mathbb{E}_{\pi,P,C_\xi}\left[\widehat{L}_{H+1}^{i,k}(s_{H+1},y_{H+1}) - L_{H+1}^{i,k,\pi}(s_{H+1},y_{H+1})\right].$$

The shared terminal condition $\widehat{L}_{H+1}^{i,k}(s,y) = L_{H+1}^{i,k,\pi}(s,y) = \frac{\lambda^{i,k}}{\alpha}(y-\nu_i)^+$ makes this residual zero. □

Since $\widehat{W}_h^{i,k}, \widehat{Q}_h^{i,k} \in [0, H]$ and $\lambda^{i,k} \in [0, \Lambda]$, the optimistic Lagrangian action-value function is uniformly bounded by $M_L := H(1 + \frac{\Lambda}{\alpha})$, i.e., $\widehat{L}_h^{i,k}(s, y, a) \le M_L$ for all $(s, y, a)$, $h$, and $k$.

**Lemma D.4** (Online Mirror Descent Bound)**.** *For any fixed comparator policy $\pi$, we have*

$$\sum_{k=1}^{K_{PD}} \sum_{h=1}^{H} \mathbb{E}_{\pi,P,C_\xi}\Big[\Big\langle \widehat{L}_h^{i,k}(s_h, y_h, \cdot), \pi_h^{i,k}(\cdot \mid s_h, y_h) - \pi_h(\cdot \mid s_h, y_h)\Big\rangle \mid s_1, 0\Big] \le \frac{H \log A}{\eta} + \frac{\eta H M_L^2 K_{PD}}{2}.$$

*Proof.* Recall that the policy is updated via exponential weights

$$\pi_h^{i,k+1}(a \mid s, y) = \frac{\pi_h^{i,k}(a \mid s, y) \exp(-\eta \widehat{L}_h^{i,k}(s, y, a))}{\sum_{a'} \pi_h^{i,k}(a' \mid s, y) \exp(-\eta \widehat{L}_h^{i,k}(s, y, a'))}.$$

For any fixed state $(s, y)$ and step $h$, we omit the state dependence in the policy and value function notation for brevity (writing $\pi^{i,k}(a)$ and $\widehat{L}_h^{i,k}(a)$ instead of $\pi_h^{i,k}(a \mid s, y)$ and $\widehat{L}_h^{i,k}(s, y, a)$). Consider the KL divergence $D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,k}) = \sum_a \pi(a) \log \frac{\pi(a)}{\pi^{i,k}(a)}$. Using the update rule, the difference in KL divergence between two consecutive steps is

$$\begin{aligned} D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,k}) - D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,k+1}) &= \sum_{a \in \mathcal{A}} \pi(a) \log \frac{\pi^{i,k+1}(a)}{\pi^{i,k}(a)} \\ &= -\eta \Big\langle \widehat{L}_h^{i,k}(\cdot), \pi(\cdot) \Big\rangle - \log \sum_{a' \in \mathcal{A}} \pi^{i,k}(a') \exp(-\eta \widehat{L}_h^{i,k}(a')). \end{aligned}$$

By applying the inequalities $\exp(-x) \le 1 - x + \frac{x^2}{2}$ (valid for $x \ge 0$) and $\log(1 + z) \le z$, we can bound the log-partition function

$$\begin{aligned} \log \sum_{a' \in \mathcal{A}} \pi^{i,k}(a') \exp(-\eta \widehat{L}_h^{i,k}(a')) &\le \log \sum_{a' \in \mathcal{A}} \pi^{i,k}(a') \Big(1 - \eta \widehat{L}_h^{i,k}(a') + \frac{\eta^2}{2} (\widehat{L}_h^{i,k}(a'))^2\Big) \\ &\le -\eta \Big\langle \widehat{L}_h^{i,k}(\cdot), \pi^{i,k}(\cdot) \Big\rangle + \frac{\eta^2}{2} \sum_{a' \in \mathcal{A}} \pi^{i,k}(a') (\widehat{L}_h^{i,k}(a'))^2. \end{aligned}$$

Substituting this upper bound back into the KL difference equation and rearranging the terms, we obtain

$$\Big\langle \widehat{L}_h^{i,k}(\cdot), \pi^{i,k}(\cdot) - \pi(\cdot) \Big\rangle \le \frac{1}{\eta} \big(D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,k}) - D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,k+1})\big) + \frac{\eta}{2} \sum_{a' \in \mathcal{A}} \pi^{i,k}(a') (\widehat{L}_h^{i,k}(a'))^2.$$

Summing this inequality over $k = 1, \ldots, K_{\mathrm{PD}}$ creates a telescoping sum for the KL divergence

$$\sum_{k=1}^{K_{\mathrm{PD}}} \Big\langle \widehat{L}_h^{i,k}(\cdot), \pi^{i,k}(\cdot) - \pi(\cdot) \Big\rangle \le \frac{1}{\eta} D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,1}) - \frac{1}{\eta} D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,K_{\mathrm{PD}}+1}) + \frac{\eta M_L^2 K_{\mathrm{PD}}}{2}.$$

Since $\pi^{i,1}$ is initialized uniformly, $D_{\mathrm{KL}}(\pi \,\|\, \pi^{i,1}) \le \log A$, and the remaining KL divergence is non-negative ($D_{\mathrm{KL}} \ge 0$). Taking the expectation $\mathbb{E}_{\pi,P,C_\xi}$ over the state trajectory and summing over $h = 1, \ldots, H$ yields the final stated bound. □

**Lemma D.5** (On-Policy Optimistic Error)**.** *Fix a reference policy $\pi$. Suppose that the optimistic Lagrangian functions share the terminal condition above and have non-positive Bellman errors, namely*

$$\widehat{L}_h^{i,k}(s, y, a) - z(s, a) - P(\cdot \mid s, a)^\top \mathbb{E}_{c \sim C_\xi(\cdot \mid s,a)} \Big[\widehat{L}_{h+1}^{i,k}(\cdot, y + c)\Big] \le 0$$

*for every $k \in [K_{PD}]$, $h \in [H]$, and reachable $(s, y, a)$. Then*

$$\sum_{k=1}^{K_{PD}} \Big(\widehat{L}_1^{i,k}(s_1, 0) - L_1^{i,k,\pi}(s_1, 0)\Big) \le \frac{H \log A}{\eta} + \frac{\eta H M_L^2 K_{PD}}{2}.$$

*Proof.* Lemma D.3 splits each difference into the policy-update term and the Bellman-error term. The latter is non-positive by assumption. Summing over $k$ and applying Lemma D.4 to the remaining term gives

$$\sum_{k=1}^{K_{\text{PD}}}\Big(\widehat{L}_1^{i,k}(s_1,0)-L_1^{i,k,\pi}(s_1,0)\Big)\le\sum_{k=1}^{K_{\text{PD}}}\sum_{h=1}^{H}\mathbb{E}_{\pi,P,C_\xi}\Big[\Big\langle\widehat{L}_h^{i,k}(\tilde{s}_h,\cdot),\pi_h^{i,k}(\cdot\mid\tilde{s}_h)-\pi_h(\cdot\mid\tilde{s}_h)\Big\rangle\Big]$$
$$\le\frac{H\log A}{\eta}+\frac{\eta H M_L^2 K_{\text{PD}}}{2}.$$

□

### D.1.3 Primal-Dual Regret and Optimal Multipliers

For a fixed subproblem $i\in[N_\xi]$, we derive a primal-dual regret bound. We then bound the optimal Lagrange multiplier for subproblem $j^*$.

We first recall the objective and constraint functionals used throughout the analysis. The performance objective is $f(\pi)=U_1^\pi(s_1,0)=\mathbb{E}_\pi[\sum_{h=1}^H z(s_h,a_h)]$. The CVaR constraint violation on $\text{disc}(\mathcal{M})$ is $g(\pi)=\text{CVaR}_\alpha(C^\pi)-\tau$, and its surrogate for subproblem $i$ is $g_i(\pi)=\nu_i+\frac{1}{\alpha}\mathbb{E}_\pi[(C^\pi-\nu_i)^+]-\tau$ as defined in Eq. (5), which satisfies $g(\pi)\le g_i(\pi)$ for every $i\in[N_\xi]$ by the variational representation of CVaR.

Recall that the Lagrange multiplier $\lambda^{i,k}$ is updated via Online Gradient Descent (OGD) on the optimistic constraint violation $\widehat{g}^{i,k}=\nu_i+\frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)-\tau$, the empirical counterpart of $g_i(\pi^{i,k})$, with projection onto $[0,\Lambda]$. Since $\widehat{V}_1^{i,k},\nu_i\in[0,H]$, the optimistic violation is bounded by a constant $M_g\coloneqq H\big(1+\frac{1}{\alpha}\big)$.

**Lemma D.6** (Lagrange Multiplier OGD Bound)**.** *For any constant $\lambda\in[0,\Lambda]$, the regret of the Lagrange multiplier update rule satisfies*

$$\sum_{k=1}^{K_{PD}}\big(\lambda\widehat{g}^{i,k}-\lambda^{i,k}\widehat{g}^{i,k}\big)\le\frac{(\lambda^{i,1}-\lambda)^2}{2\eta_\lambda}+\frac{\eta_\lambda M_g^2 K_{PD}}{2}.$$

*Proof.* From the update $\lambda^{i,k+1}=\text{Proj}_{[0,\Lambda]}[\lambda^{i,k}+\eta_\lambda\widehat{g}^{i,k}]$ and the non-expansiveness of projection, for any $\lambda\in[0,\Lambda]$

$$\begin{aligned}(\lambda^{i,k+1}-\lambda)^2&=\Big(\text{Proj}_{[0,\Lambda]}[\lambda^{i,k}+\eta_\lambda\widehat{g}^{i,k}]-\text{Proj}_{[0,\Lambda]}(\lambda)\Big)^2\\&\le(\lambda^{i,k}+\eta_\lambda\widehat{g}^{i,k}-\lambda)^2\\&=(\lambda^{i,k}-\lambda)^2+2\eta_\lambda\widehat{g}^{i,k}(\lambda^{i,k}-\lambda)+\eta_\lambda^2(\widehat{g}^{i,k})^2.\end{aligned}$$

Rearranging and summing over $k$ gives

$$0\le(\lambda^{i,K_{\text{PD}}+1}-\lambda)^2\le(\lambda^{i,1}-\lambda)^2+2\eta_\lambda\sum_{k=1}^{K_{\text{PD}}}\widehat{g}^{i,k}(\lambda^{i,k}-\lambda)+\eta_\lambda^2\sum_{k=1}^{K_{\text{PD}}}(\widehat{g}^{i,k})^2,$$

which directly implies the stated bound. □

**Lemma D.7** (Primal-Dual Regret)**.** *Suppose the non-positive Bellman-error condition in Lemma D.5 holds. Then, for any fixed reference policy $\pi$, any input $\lambda^{\text{init}}\in[0,\Lambda]$, and any $\lambda\in[0,\Lambda]$, the optimistic Lagrangian gap against $\pi$ is bounded by*

$$\sum_{k=1}^{K_{PD}}\left(\widehat{U}_1^{i,k}(s_1,0)+\lambda\left(\nu_i-\tau+\frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)\right)-U_1^\pi(s_1,0)-\lambda^{i,k}\left(\nu_i-\tau+\frac{1}{\alpha}V_1^\pi(s_1,0)\right)\right)$$
$$\le\frac{H\log A}{\eta}+\frac{\eta H M_L^2 K_{PD}}{2}+\frac{(\lambda^{i,1}-\lambda)^2}{2\eta_\lambda}+\frac{\eta_\lambda M_g^2 K_{PD}}{2}.$$

*Proof.* By adding and subtracting the term $\lambda^{i,k}\Big(\nu_i - \tau + \frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)\Big)$ inside the summation, we can decouple the overall gap into a primal optimality gap and a dual regret gap

$$\begin{aligned}
&\sum_{k=1}^{K_{\text{PD}}}\left(\widehat{U}_1^{i,k}(s_1,0) + \lambda\left(\nu_i - \tau + \frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)\right) - U_1^{\pi}(s_1,0) - \lambda^{i,k}\left(\nu_i - \tau + \frac{1}{\alpha}V_1^{\pi}(s_1,0)\right)\right)\\
&= \sum_{k=1}^{K_{\text{PD}}}\left(\widehat{U}_1^{i,k}(s_1,0) + \lambda^{i,k}\left(\nu_i - \tau + \frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)\right) - U_1^{\pi}(s_1,0) - \lambda^{i,k}\left(\nu_i - \tau + \frac{1}{\alpha}V_1^{\pi}(s_1,0)\right)\right)\\
&\quad + \sum_{k=1}^{K_{\text{PD}}}\left(\lambda\left(\nu_i - \tau + \frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)\right) - \lambda^{i,k}\left(\nu_i - \tau + \frac{1}{\alpha}\widehat{V}_1^{i,k}(s_1,0)\right)\right)\\
&= \sum_{k=1}^{K_{\text{PD}}}\Big(\widehat{L}_1^{i,k}(s_1,0) - L_1^{i,k,\pi}(s_1,0)\Big) + \sum_{k=1}^{K_{\text{PD}}}\big(\lambda\widehat{g}^{i,k} - \lambda^{i,k}\widehat{g}^{i,k}\big),
\end{aligned}$$

Applying Lemmas D.5 and D.6 gives the result. □

We now specialize the analysis to subproblem $j^*$. The policy $\pi^*_{\text{disc}}$ is feasible for this subproblem because $\nu_{j^*} = \text{VaR}_\alpha(C^{\pi^*_{\text{disc}}})$ and hence $g_{j^*}(\pi^*_{\text{disc}}) = g(\pi^*_{\text{disc}}) \le 0$.

The augmented occupancy measures form a nonempty compact polytope, and every point in this polytope is induced by a time-dependent augmented-Markov policy. Both $f$ and $g_{j^*}$ are affine on the polytope. Consequently, the constrained minimum and the minimum of $f + \lambda g_{j^*}$ for every fixed $\lambda \ge 0$ are attained. Let $\pi^*_{j^*}$ denote a constrained optimizer.

Assumption 5.2 supplies a point with $g_{j^*} < 0$. Mixing it with any point in the relative interior of the occupancy polytope preserves strict feasibility for a sufficiently small positive mixing weight and produces a strictly feasible relative-interior point. The strong-duality theorem of Beck (2017, Theorem A.1) therefore applies and gives a supporting multiplier and dual attainment. We let

$$\lambda^*_{j^*} \in \arg\max_{\lambda\ge 0}\min_{\pi}\{f(\pi) + \lambda g_{j^*}(\pi)\}$$

denote an optimal Lagrange multiplier of subproblem $j^*$.

**Lemma D.8** (Bound on the Optimal Lagrange Multiplier of Subproblem $j^*$)**.** *Suppose Assumption 5.2 holds with margin $\zeta_\xi > 0$. Then $\lambda^*_{j^*} \le \frac{H}{\zeta_\xi}$.*

*Proof.* The compactness, affinity, and strict-feasibility conditions verified before the lemma give strong duality and dual attainment. Let $d_{j^*}(\lambda) \coloneqq \min_\pi\{f(\pi) + \lambda g_{j^*}(\pi)\}$ be the dual function of subproblem $j^*$. Strong duality gives

$$f(\pi^*_{j^*}) = \max_{\lambda\ge 0} d_{j^*}(\lambda) = d_{j^*}(\lambda^*_{j^*}).$$

By the definition of the dual function as an infimum, evaluating at the Slater policy $\pi'$ yields

$$d_{j^*}(\lambda^*_{j^*}) \le f(\pi') + \lambda^*_{j^*} g_{j^*}(\pi').$$

Combining and rearranging,

$$\lambda^*_{j^*}(-g_{j^*}(\pi')) \le f(\pi') - f(\pi^*_{j^*}).$$

By Assumption 5.2, $-g_{j^*}(\pi') \ge \zeta_\xi$. The performance cost lies in $[0, H]$, which gives $f(\pi') - f(\pi^*_{j^*}) \le H$. Therefore $\lambda^*_{j^*} \le H/\zeta_\xi$. □

**Lemma D.9** (Optimality Equivalence at Index $j^*$)**.** *The optimal policy of subproblem $j^*$ attains the optimal value on* disc($\mathcal{M}$)*, that is, $f(\pi^*_{j^*}) = f(\pi^*_{\text{disc}})$.*

*Proof.* Since $\nu_{j^*} = \text{VaR}_\alpha(C^{\pi^*_{\text{disc}}})$ minimizes the Rockafellar–Uryasev formula for $\pi^*_{\text{disc}}$, we have $g_{j^*}(\pi^*_{\text{disc}}) = g(\pi^*_{\text{disc}}) \le 0$, hence $\pi^*_{\text{disc}}$ is feasible for subproblem $j^*$. As $\pi^*_{j^*}$ is optimal for subproblem $j^*$, this gives $f(\pi^*_{j^*}) \le f(\pi^*_{\text{disc}})$. Conversely, by the variational representation of CVaR, $g(\pi) \le g_i(\pi)$ for every index $i$. Hence every policy feasible for subproblem $j^*$ is also feasible for the constrained problem on disc($\mathcal{M}$). Since $\pi^*_{\text{disc}}$ is optimal among policies feasible on disc($\mathcal{M}$), $f(\pi^*_{\text{disc}}) \le f(\pi^*_{j^*})$. Therefore $f(\pi^*_{j^*}) = f(\pi^*_{\text{disc}})$. □

### D.1.4 Estimation Error and Subproblem Guarantees

For the analysis, decompose the Phase I bonuses as $b_Q^{i,k} = b_P^{i,k} + b_C^{i,k}$ and $b_W^{i,k} = b_Q^{i,k} + b_z^{i,k}$, using the component bonuses in Eq. (7).

**Lemma D.10** (Value Function Optimism with Unknown Costs)**.** *On event $\mathcal{E}_1$, the Bellman errors of the two empirical recursions in Algorithm 1 are non-positive. In particular, for every reachable $(s, y, a)$ and $h \in [H]$,*

$$\widehat{W}_h^{i,k}(s,y,a) \le z(s,a) + P(\cdot \mid s,a)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(\cdot, y+c),$$
$$\widehat{Q}_h^{i,k}(s,y,a) \le P(\cdot \mid s,a)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} \widehat{V}_{h+1}^{i,k}(\cdot, y+c).$$

*Proof.* Fix a subproblem $i$ and episode $k$. For the performance continuation, add and subtract the term that uses the true transition kernel and the empirical discretized risk cost distribution. This gives

$$\begin{aligned}
&\left|\widehat{P}^{i,k}(\cdot \mid s,a)^\top \mathbb{E}_{c\sim \widehat{C}_\xi^{i,k}(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(\cdot, y+c) - P(\cdot \mid s,a)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(\cdot, y+c)\right| \\
&\quad \le \|\widehat{P}^{i,k}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \|\mathbb{E}_{c\sim \widehat{C}_\xi^{i,k}(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(\cdot, y+c)\|_\infty \\
&\qquad + \sum_{s'} P(s' \mid s,a) \left|\mathbb{E}_{c\sim \widehat{C}_\xi^{i,k}(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(s', y+c) - \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(s', y+c)\right| \\
&\quad \le b_P^{i,k}(s,a) + b_C^{i,k}(s,a).
\end{aligned}$$

The first term is Lemma D.1 applied to the empirical-risk average of the continuation, which also lies in $[0, H]$. The bounds $0 \le \widehat{U}_{h+1}^{i,k}, \widehat{V}_{h+1}^{i,k} \le H$ follow by backward induction from their terminal values, the stochastic empirical kernel, the assumption $z \in [0,1]$, and the clipping in Algorithm 1. For each fixed $s'$, Hölder's inequality on the finite alphabet $\mathcal{C}_\xi$ gives

$$\begin{aligned}
&\left|\mathbb{E}_{c\sim \widehat{C}_\xi^{i,k}(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(s', y+c) - \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(s', y+c)\right| \\
&\quad = \left|\sum_{c\in\mathcal{C}_\xi} \Big(\widehat{C}_\xi^{i,k}(c \mid s,a) - C_\xi(c \mid s,a)\Big) \widehat{U}_{h+1}^{i,k}(s', y+c)\right| \\
&\quad \le \|\widehat{C}_\xi^{i,k}(\cdot \mid s,a) - C_\xi(\cdot \mid s,a)\|_1 \|\widehat{U}_{h+1}^{i,k}(s', y+\cdot)\|_\infty \\
&\quad \le H \|\widehat{C}_\xi^{i,k}(\cdot \mid s,a) - C_\xi(\cdot \mid s,a)\|_1 \le b_C^{i,k}(s,a).
\end{aligned}$$

The last inequality is the risk cost component of event $\mathcal{E}_1$. Lemma C.2 establishes this component using Weissman's $L_1$ inequality. The same calculation with $\widehat{V}_{h+1}^{i,k}$ gives the identical bound for the risk continuation. The immediate performance-mean error is at most $b_z^{i,k}(s,a)$. Therefore, subtracting $b_Q^{i,k}$ and $b_W^{i,k}$ from the respective empirical backups gives the claimed lower confidence estimates. Clipping at zero preserves the inequalities because the true cost-to-go functions are non-negative. Multiplying the risk inequality by $\lambda^{i,k}/\alpha \ge 0$ and adding the performance inequality gives the non-positive Lagrangian Bellman error in Lemma D.5. □

Lemma D.2 controls the transition component. We use the shared inverse-visitation envelope

$$R_{\text{PD}} := 2\sqrt{5}\sqrt{SAHK_{\text{PD}}} + 4SAH \log \frac{SA}{\delta'} + SAH$$

in the additional cost-estimation bounds below.

**Lemma D.11** (Cumulative Cost-Estimation Bonuses)**.** *On $\mathcal{E}_2$, for any fixed subproblem $i \in [N_\xi]$,*

$$\sum_{k=1}^{K_{PD}} \sum_{h=1}^{H} \mathbb{E}_{\pi^{i,k},P,C_\xi}[b_C^{i,k}(s_h,a_h)] \le H\sqrt{2\log \frac{2^{\lfloor 1/\xi \rfloor + 1} SAN_\xi K_{PD} H}{\delta'}}\, R_{PD} + 2SAH^2,$$
$$\sum_{k=1}^{K_{PD}} \sum_{h=1}^{H} \mathbb{E}_{\pi^{i,k},P,C_\xi}[b_z^{i,k}(s_h,a_h)] \le \sqrt{2\log \frac{2SAN_\xi K_{PD} H}{\delta'}}\, R_{PD} + SAH.$$

*Proof.* At positive counts, both bonuses are their coefficients in Eq. (7) times $1/\sqrt{N^{i,k}(s,a)}$. The proof of Lemma D.2 bounds the expected sum of the corresponding inverse-count factor by $R_{\mathrm{PD}}$. At zero count, $\sqrt{N^{i,k}(s,a) \vee 1} = 1$, and the square-root coefficients exceed the respective bonuses $2H$ and 1 because $\delta' = \delta/8$, $\delta < 1$, and $S, A, N_\xi, K_{\mathrm{PD}}, H \geq 1$. The same pointwise bounds therefore hold at every count and remain valid after taking conditional expectations and summing. The additional $2SAH^2$ and $SAH$ are nonnegative relaxations. □

**Lemma D.12** (Lagrangian Estimation Error with Unknown Costs)**.** *On $\mathcal{E}_1$, for every $\lambda \in [0, \Lambda]$,*

$$\sum_{k=1}^{K_{PD}} \Big( \widehat{U}_1^{i,k}(s_1, 0) + \lambda \widehat{g}^{i,k} \Big) \geq \sum_{k=1}^{K_{PD}} \Big( U_1^{\pi^{i,k}}(s_1, 0) + \lambda g_i(\pi^{i,k}) \Big) - 2 \sum_{k=1}^{K_{PD}} \sum_{h=1}^{H} \mathbb{E}_{\pi^{i,k}, P, C_\xi} \left( b_W^{i,k}(s_h, a_h) + \frac{\lambda}{\alpha} b_Q^{i,k}(s_h, a_h) \right).$$

*Proof.* Fix a subproblem $i$ and episode $k$. Apply the two-term decomposition from the proof of Lemma D.10 to the empirical performance continuation $\widehat{U}_{h+1}^{i,k}$. This is the continuation that appears in the empirical Bellman recursion and gives

$$\left| \widehat{P}^{i,k}(\cdot \mid s,a)^\top \mathbb{E}_{c \sim \widehat{C}_\xi^{i,k}(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(\cdot, y+c) - P(\cdot \mid s,a)^\top \mathbb{E}_{c \sim C_\xi(\cdot|s,a)} \widehat{U}_{h+1}^{i,k}(\cdot, y+c) \right| \leq b_P^{i,k}(s,a) + b_C^{i,k}(s,a).$$

The same argument applies to $\widehat{V}_{h+1}^{i,k}$. Both empirical continuations lie in $[0, H]$, as established in the proof of Lemma D.10. In addition, $|\widehat{z}^{i,k}(s,a) - z(s,a)| \leq b_z^{i,k}(s,a)$ on $\mathcal{E}_1$. Clipping at zero can only increase the empirical optimistic value because the true values are non-negative. After averaging over $a \sim \pi_h^{i,k}(\cdot \mid s, y)$, the two one-step comparisons are therefore

$$\begin{aligned} U_h^{\pi^{i,k}}(s,y) - \widehat{U}_h^{i,k}(s,y) &\leq 2\mathbb{E}_{\pi_h^{i,k}} \Big( b_W^{i,k}(s,a) \Big) \\ &\quad + \mathbb{E}_{\substack{a \sim \pi_h^{i,k}(\cdot|s,y),\ s' \sim P(\cdot|s,a), \\ c \sim C_\xi(\cdot|s,a)}} \Big( U_{h+1}^{\pi^{i,k}}(s', y+c) - \widehat{U}_{h+1}^{i,k}(s', y+c) \Big), \\ V_h^{\pi^{i,k}}(s,y) - \widehat{V}_h^{i,k}(s,y) &\leq 2\mathbb{E}_{\pi_h^{i,k}} \Big( b_Q^{i,k}(s,a) \Big) \\ &\quad + \mathbb{E}_{\substack{a \sim \pi_h^{i,k}(\cdot|s,y),\ s' \sim P(\cdot|s,a), \\ c \sim C_\xi(\cdot|s,a)}} \Big( V_{h+1}^{\pi^{i,k}}(s', y+c) - \widehat{V}_{h+1}^{i,k}(s', y+c) \Big). \end{aligned}$$

The first copy of each bonus controls the empirical-to-true transition and discretized risk cost distribution errors. The second is the bonus subtracted in the empirical recursion. The empirical and true terminal values agree, and hence their terminal differences vanish. Iterating the two inequalities under the true trajectory distribution, multiplying the risk inequality by $\lambda/\alpha$, and summing over $k$ proves the claim. □

For $\lambda \in [0, \Lambda]$, define the following explicit upper bound on the accumulated primal-dual gap:

$$\begin{aligned} \Delta_{\mathrm{PD}}(\lambda) \coloneqq\ & \frac{H \log A}{\eta} + \frac{\eta H M_L^2 K_{\mathrm{PD}}}{2} + \frac{\Lambda^2}{2\eta_\lambda} + \frac{\eta_\lambda M_g^2 K_{\mathrm{PD}}}{2} \\ & + 2\left(1 + \frac{\lambda}{\alpha}\right) H \sqrt{2S \log \frac{2SAN_\xi K_{\mathrm{PD}} H}{\delta'}}\, R_{\mathrm{PD}} \\ & + 2\left(1 + \frac{\lambda}{\alpha}\right) H \sqrt{2\left( (\lfloor 1/\xi \rfloor + 1) \log 2 + \log \frac{SAN_\xi K_{\mathrm{PD}} H}{\delta'} \right)}\, R_{\mathrm{PD}} \\ & + 8\left(1 + \frac{\lambda}{\alpha}\right) SAH^2 + 2\sqrt{2 \log \frac{2SAN_\xi K_{\mathrm{PD}} H}{\delta'}}\, R_{\mathrm{PD}} + 2SAH. \end{aligned} \tag{10}$$

The first line is the explicit online-optimization contribution from Lemma D.7. The remaining lines follow from Lemmas D.2, D.11, and D.12.

**Theorem D.13** (Subproblem Error Bounds with Unknown Costs)**.** *Under Assumption 5.2 and the parameter and update choices of Algorithm 1, on $\mathcal{E}_1 \cap \mathcal{E}_2$, the candidate for subproblem $j^*$ satisfies*

$$\begin{aligned} f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) &\le \frac{\Delta_{PD}(0)}{K_{PD}}, \\ (g_{j^*}(\bar{\pi}_{j^*}))^+ &\le \frac{2\Delta_{PD}(\Lambda)}{\Lambda K_{PD}}. \end{aligned} \tag{11}$$

*Proof.* Apply Lemma D.7 with reference policy $\pi^*_{\text{disc}}$. Lemma D.10 verifies its Bellman-error premise, and $g_{j^*}(\pi^*_{\text{disc}}) \le 0$ makes the reference-policy multiplier term nonnegative after it is moved to the left. Lemma D.12 converts the empirical on-policy values to true values. Finally, Lemmas D.2 and D.11 control the accumulated bonuses. Thus, for every $\lambda \in [0, \Lambda]$,

$$\sum_{k=1}^{K_{\text{PD}}} \Big( f(\pi^{j^*,k}) - f(\pi^*_{\text{disc}}) + \lambda g_{j^*}(\pi^{j^*,k}) \Big) \le \Delta_{\text{PD}}(\lambda). \tag{12}$$

Linearity of the episode-level mixture in the occupancy measure turns the average of the left-hand side into $f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) + \lambda g_{j^*}(\bar{\pi}_{j^*})$. Setting $\lambda = 0$ gives the first inequality. For the second, dual optimality, Lemma D.9, and Lemma D.8 give

$$f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) \ge -\lambda^*_{j^*} g_{j^*}(\bar{\pi}_{j^*}), \qquad \Lambda - \lambda^*_{j^*} \ge \frac{\Lambda}{2}.$$

Substitution into Eq. (12) at $\lambda = \Lambda$ gives

$$(\Lambda - \lambda^*_{j^*}) g_{j^*}(\bar{\pi}_{j^*}) \le \frac{\Delta_{\text{PD}}(\Lambda)}{K_{\text{PD}}}.$$

If the constraint surrogate is non-positive, its positive part is zero. Otherwise, $\Lambda - \lambda^*_{j^*} \ge \Lambda/2$ gives the second inequality. □

**Remark.** Assumption 5.2 enters only through Lemma D.8 for subproblem $j^*$. The analysis does not require a feasibility margin for any other grid point.

For $K \ge 1$, write

$$\Gamma_{\text{PD}}(K) := \log \frac{2eSAN_\xi HK}{\delta'}.$$

**Corollary D.14** (Sample Complexity of Phase I with Unknown Costs)**.** *Under the conditions of Theorem D.13, there exists a deterministic integer $K_{PD}$ satisfying*

$$K_{PD} = \widetilde{\mathcal{O}}\left( \frac{S(S + \lfloor 1/\xi \rfloor + 1)AH^3 + H^4}{\alpha^2 \varepsilon_0^2} + \frac{H^6}{\zeta_\xi^2 \alpha^2 \varepsilon_0^2} \right), \tag{13}$$

*for which $\bar{\pi}_{j^*}$ satisfies*

$$f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) \le \frac{\varepsilon_0}{2}, \qquad (g_{j^*}(\bar{\pi}_{j^*}))^+ \le \frac{\varepsilon_0}{2}.$$

*Proof.* We retain the dependence on $K_{\text{PD}}$ until the inversion is complete. Substituting the learning rates in Algorithm 1, the definition of $R_{\text{PD}}$, and the two logarithmic bounds

$$\begin{aligned} \log \frac{2SAN_\xi K_{\text{PD}} H}{\delta'} &\le \Gamma_{\text{PD}}(K_{\text{PD}}), \\ (\lfloor 1/\xi \rfloor + 1) \log 2 + \log \frac{SAN_\xi K_{\text{PD}} H}{\delta'} &\le (\lfloor 1/\xi \rfloor + 2) \Gamma_{\text{PD}}(K_{\text{PD}}) \end{aligned}$$

into Eq. (10) gives, for $\lambda \in \{0, \Lambda\}$,

$$\begin{aligned}\Delta_{\mathrm{PD}}(\lambda) \le{}& \left( HM_L\sqrt{2\log A} + \Lambda M_g + 4\sqrt{10}\left(\left(1+\frac{\lambda}{\alpha}\right)H\left(\sqrt{S}+\sqrt{\lfloor 1/\xi\rfloor+2}\right)+1\right)\sqrt{SAH}\right) \\ &\times \sqrt{K_{\mathrm{PD}}(\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})+\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})^2)} \\ &+ 2\sqrt{2}\left(\left(1+\frac{\lambda}{\alpha}\right)H\left(\sqrt{S}+\sqrt{\lfloor 1/\xi\rfloor+2}\right)+1\right)SAH\left(4\log\frac{SA}{\delta'}+1\right) \\ &\times \left(\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})+\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})^2\right) \\ &+ \left(8\left(1+\frac{\lambda}{\alpha}\right)SAH^2+2SAH\right)\left(\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})+\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})^2\right).\end{aligned}$$

Let $C_{\mathrm{PD}}(\lambda)$ denote the bracketed coefficient of the square-root term in the preceding display, and let $D_{\mathrm{PD}}(\lambda)$ be the sum of the two coefficients multiplying $\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})+\Gamma_{\mathrm{PD}}(K_{\mathrm{PD}})^2$. If the objective error in Eq. (11) exceeds $\varepsilon_0/2$, then $K_{\mathrm{PD}} \le 2\Delta_{\mathrm{PD}}(0)/\varepsilon_0$. This is Lemma H.6 with

$$m = K_{\mathrm{PD}}, \quad \mathscr{C} = \frac{2C_{\mathrm{PD}}(0)}{\varepsilon_0}, \quad \mathscr{D} = \frac{2D_{\mathrm{PD}}(0)}{\varepsilon_0}, \quad \mathscr{A} = \mathscr{B} = \mathscr{E} = 1, \quad \kappa = \frac{2eSAN_\xi H}{\delta'}.$$

Under this substitution, the conclusion of Lemma H.6 gives a deterministic upper endpoint for $K_{\mathrm{PD}}$. If the constraint error exceeds $\varepsilon_0/2$, then $K_{\mathrm{PD}} \le 4\Delta_{\mathrm{PD}}(\Lambda)/(\Lambda\varepsilon_0)$. The same lemma applies with

$$m = K_{\mathrm{PD}}, \quad \mathscr{C} = \frac{4C_{\mathrm{PD}}(\Lambda)}{\Lambda\varepsilon_0}, \quad \mathscr{D} = \frac{4D_{\mathrm{PD}}(\Lambda)}{\Lambda\varepsilon_0}, \quad \mathscr{A} = \mathscr{B} = \mathscr{E} = 1, \quad \kappa = \frac{2eSAN_\xi H}{\delta'}.$$

This substitution gives a second deterministic upper endpoint. Choose $K_{\mathrm{PD}}$ to be one plus the ceiling of the maximum of 1 and these two endpoints. This integer is strictly larger than both endpoints. Therefore, neither target error can exceed $\varepsilon_0/2$.

The grid point $\nu_{j^*}$ and the positive-part expectation in Assumption 5.2 are nonnegative. Hence $g_{j^*}(\pi') \ge -\tau$, and the Slater inequality implies $\zeta_\xi \le \tau \le H$.

Finally, substitute $M_L = H(1+\Lambda/\alpha)$, $M_g = H(1+1/\alpha)$, and $\Lambda = 2H/\zeta_\xi$ into these two endpoints. Squaring the square-root coefficients produces the $S(S+\lfloor 1/\xi\rfloor+1)AH^3$, $H^4$, and $H^6/\zeta_\xi^2$ groups in Eq. (13). The linear and cross terms in each endpoint are no larger in polynomial order because $\varepsilon_0, \alpha \le 1$, $\zeta_\xi \le H$, and $S, A, H \ge 1$. The factor $8/3$ in Lemma H.6 changes only the absolute constant multiplying the squared logarithm. This proves that the deterministic choice above has the displayed rate and completes the proof. □

### D.2 Phase II: Reward-Free Exploration and Uniform Model Accuracy

Phase II learns the common empirical model used to evaluate candidates in Phase III. Its analysis follows the count argument of Ménard et al. (2021), but the uncertainty recursion directly includes the transition, discretized risk cost distribution, and performance cost mean bonuses. The use of concentration radii and the universal value bound $H$ gives the stated $H^3$ dependence.

For any policy $\pi$ and fixed grid point $\nu_i$, let $\widehat{W}_h^{k,\pi}$ and $\widehat{Q}_h^{k,\pi}$ be its performance and risk action-value functions under $(\widehat{P}^k, \widehat{z}^k, \widehat{C}_\xi^k)$. They evaluate the expected remaining performance cost and the expected terminal cost $(y_{H+1}-\nu_i)^+$, respectively. Let $\widehat{U}_h^{k,\pi}$ and $\widehat{V}_h^{k,\pi}$ be their averages under $\pi_h$. We suppress the dependence of these risk values on the fixed threshold $\nu_i$. The corresponding true values under $(P, z, C_\xi)$ omit the hat and the model index $k$. Appendix D.3 gives these policy-evaluation recursions at the stopped model.

#### D.2.1 Uniform Evaluation and Stopping

For the analysis, define the aggregate bonus by

$$b_{\mathrm{RFE}}^k = b_{P,\mathrm{RFE}}^k + b_{C,\mathrm{RFE}}^k + b_{z,\mathrm{RFE}}^k.$$

Its uncertainty recursion uses this combined bonus and the empirical discretized risk cost distribution:

$$J_h^k(s,y,a) := \min\left\{ b_{\mathrm{RFE}}^k(s,a) + \widehat{P}^k(\cdot \mid s,a)^\top \mathbb{E}_{c\sim\widehat{C}_\xi^k(\cdot|s,a)}\left[\max_{a'} J_{h+1}^k(\cdot, y+c, a')\right],\ H\right\}. \quad (14)$$

**Lemma D.15** (Uniform Evaluation with Unknown Costs)**.** *On $\mathcal{E}_3$, for every policy $\pi$, model index $k$, and reachable augmented state-action pair,*

$$|\widehat{W}_h^{k,\pi}(s,y,a) - W_h^\pi(s,y,a)| \le J_h^k(s,y,a),$$
$$|\widehat{Q}_h^{k,\pi}(s,y,a) - Q_h^\pi(s,y,a)| \le J_h^k(s,y,a).$$

*Proof.* We prove both bounds by backward induction on $h$. At $h = H+1$, the empirical and true terminal values agree, and $J_{H+1}^k = 0$. For the induction step, first consider the true performance continuation $U_{h+1}^\pi$. Adding and subtracting the term that uses the true transition kernel and the empirical discretized risk cost distribution gives

$$\begin{aligned}
&\left|\widehat{P}^k(\cdot \mid s,a)^\top \mathbb{E}_{c\sim\widehat{C}_\xi^k(\cdot|s,a)} U_{h+1}^\pi(\cdot, y+c) - P(\cdot \mid s,a)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} U_{h+1}^\pi(\cdot, y+c)\right| \\
&\quad\le \left|(\widehat{P}^k(\cdot \mid s,a) - P(\cdot \mid s,a))^\top \mathbb{E}_{c\sim\widehat{C}_\xi^k(\cdot|s,a)} U_{h+1}^\pi(\cdot, y+c)\right| \\
&\qquad + \sum_{s'} P(s' \mid s,a) \left|\mathbb{E}_{c\sim\widehat{C}_\xi^k(\cdot|s,a)} U_{h+1}^\pi(s', y+c) - \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} U_{h+1}^\pi(s', y+c)\right| \\
&\quad\le b_{P,\mathrm{RFE}}^k(s,a) + b_{C,\mathrm{RFE}}^k(s,a).
\end{aligned}$$

The empirical-risk average in the first term is a function from $\mathcal{S}$ to $[0,H]$. The transition statement in $\mathcal{E}_3$ therefore bounds this term directly by $b_{P,\mathrm{RFE}}^k(s,a)$. For each fixed $s'$, the finite-alphabet Hölder calculation displayed in the proof of Lemma D.10 gives

$$\begin{aligned}
&\left|\mathbb{E}_{c\sim\widehat{C}_\xi^k(\cdot|s,a)} U_{h+1}^\pi(s', y+c) - \mathbb{E}_{c\sim C_\xi(\cdot|s,a)} U_{h+1}^\pi(s', y+c)\right| \\
&\qquad \le H\|\widehat{C}_\xi^k(\cdot \mid s,a) - C_\xi(\cdot \mid s,a)\|_1 \le b_{C,\mathrm{RFE}}^k(s,a).
\end{aligned}$$

The last inequality is the risk cost component of event $\mathcal{E}_3$. Lemma C.6 establishes this component using Weissman's $L_1$ inequality. The same calculation with the true risk continuation $V_{h+1}^\pi$ gives the identical bound. The performance backup also contains the immediate error $|\widehat{z}^k(s,a) - z(s,a)| \le b_{z,\mathrm{RFE}}^k(s,a)$. After separating these model errors, the remaining difference between the empirical and true continuation values is averaged with respect to $s' \sim \widehat{P}^k(\cdot \mid s,a)$ and $c \sim \widehat{C}_\xi^k(\cdot \mid s,a)$. Applying the induction hypothesis gives the continuation term in Eq. (14). Clipping at $H$ is valid because both true and empirical cost-to-go functions lie in $[0,H]$. This proves the performance bound. The risk recursion has no immediate performance term, and the same $J^k$ remains a valid upper bound because it includes the nonnegative bonus $b_{z,\mathrm{RFE}}^k$.

For each fixed $k$, the event bounds are uniform over all functions from $\mathcal{S}$ to $[0,H]$, and the induction uses the common empirical model. The conclusion therefore holds simultaneously for every constituent policy. At $K_{\mathrm{stop}}$, average the initial action-value bounds within each constituent and then across the constituents of $\bar{\pi}_i$. Both averages preserve the same error bound. □

For every model index $K \ge 2$, let

$$\Gamma_{\mathrm{RFE}}(K) := \log\frac{3SAH}{\delta'} + S\log(8e(KH+1)) + (\lfloor 1/\xi \rfloor + 1)\log 2 + \log\frac{2SA(KH+1)^2}{\delta'}.$$

This logarithmic envelope controls all three bonuses in the stopping analysis.

**Lemma D.16** (Explicit Bound for a Non-Stopping Prefix)**.** *On $\mathcal{E}_3 \cap \mathcal{E}_4$, fix $T \ge 1$. If the stopping test fails at model indices $1,\dots,T$, then*

$$\begin{aligned}
T\varepsilon_{\mathrm{est}} < {}& 12H\sqrt{10\Gamma_{RFE}(T+1)SAHT} + 6SAH^2\sqrt{2\Gamma_{RFE}(T+1)}\big(4\log\tfrac{SA}{\delta'} + 1\big) \\
&+ 32SAH^2\left(\frac{2}{3}\Gamma_{RFE}(T+1) + 5\right)\log(T+1) \\
&+ 2SAH^2\left(\frac{2}{3}\Gamma_{RFE}(T+1) + 5\right)\big(4\log\tfrac{SA}{\delta'} + 1\big).
\end{aligned} \tag{15}$$

*Proof.* Condition on $\mathcal{E}_3 \cap \mathcal{E}_4$. The component statements are time-uniform. Hence, the argument applies to the fixed deterministic prefix without conditioning on a random stopping index. In Eq. (14), first replace the empirical transition kernel by $P$, and then replace the empirical discretized risk cost distribution by $C_\xi$. The corresponding components of $\mathcal{E}_3$ give the one-step inequality

$$\begin{aligned} J_h^k(s,y,a) &\le b_{\text{RFE}}^k(s,a) + \widehat{P}^k(\cdot \mid s,a)^\top \mathbb{E}_{c\sim \widehat{C}_\xi^k(\cdot|s,a)}\Big[\max_{a'} J_{h+1}^k(\cdot, y+c, a')\Big] \\ &\le 2b_{\text{RFE}}^k(s,a) + P(\cdot \mid s,a)^\top \mathbb{E}_{c\sim C_\xi(\cdot|s,a)}\Big[\max_{a'} J_{h+1}^k(\cdot, y+c, a')\Big]. \end{aligned}$$

The first inequality uses $\min\{x, H\} \le x$. The second uses $0 \le J_{h+1}^k \le H$ and $b_{P,\text{RFE}}^k + b_{C,\text{RFE}}^k \le b_{\text{RFE}}^k$. The exploratory policy selects an action attaining the maximum of $J_h^k$. Iterating the display from $J_{H+1}^k = 0$ under its true trajectory distribution gives

$$\max_{a'} J_1^k(s_1, 0, a') \le 2\sum_{h=1}^{H}\sum_{s,a} q_h^k(s,a) b_{\text{RFE}}^k(s,a) = 2\sum_{s,a} q^k(s,a) b_{\text{RFE}}^k(s,a). \tag{16}$$

The equality uses the fact that $b_{\text{RFE}}^k$ does not depend on $h$ and the definition $q^k = \sum_h q_h^k$. For $k \le T$, the definitions of the three bonuses, including their zero-count values, imply

$$b_{\text{RFE}}^k(s,a) \le 3H\sqrt{\frac{2\Gamma_{\text{RFE}}(T+1)}{N^k(s,a)\vee 1}} + H\frac{\frac{2}{3}\Gamma_{\text{RFE}}(T+1)+5}{N^k(s,a)\vee 1}. \tag{17}$$

Indeed, the three square-root terms are bounded by the first term on the right. The linear transition term is bounded by the second. At zero count, the left-hand side is at most $4H + 1 \le 5H$, which is also covered by the second term.

Sum Eq. (16) over $k = 1, \dots, T$ and use Eq. (17). Apply Lemma H.3 and split each state-action sequence when its pseudo-count first exceeds $4H\log(SA/\delta')$. Before that crossing, its total pseudo-count mass is at most $H[4\log(SA/\delta') + 1]$.

For the remaining indices, fix $(s,a)$ and apply Lemma H.4 to the normalized increments $q^k(s,a)/H$, whose cumulative sums are $\bar{n}^k(s,a)/H$. Each increment lies in $[0,1]$. On the late indices $\bar{n}^{k-1}(s,a)/H > 1$, the ratio of a normalized increment to the preceding normalized cumulative sum is exactly $q^k(s,a)/\bar{n}^{k-1}(s,a)$. Moreover, $\bar{n}^T(s,a)/H \le T$. The lemma and the factor four in Eq. (36) therefore give at most $16\log(T+1)$ for each late state-action sequence. Summing over $(s,a)$ gives the stronger late bound $16SA\log(T+1)$. Using $H \ge 1$ and adding the early mass yields

$$\sum_{k,s,a}\frac{q^k(s,a)}{N^k(s,a)\vee 1} \le 16SAH\log(T+1) + SAH\big(4\log\tfrac{SA}{\delta'} + 1\big).$$

For the inverse-square-root sum, event $\mathcal{E}_4$ gives the count condition in Lemma H.5. Since $\delta' = \frac{\delta}{8}$ implies $\log\frac{SA}{\delta'} \ge 1$, applying that lemma with $K = T$ gives

$$\begin{aligned} \sum_{k,s,a} q^k(s,a)\sqrt{\frac{2\Gamma_{\text{RFE}}(T+1)}{N^k(s,a)\vee 1}} &\le 2\sqrt{10\Gamma_{\text{RFE}}(T+1)SAHT} \\ &\quad + SAH\big(4\log\tfrac{SA}{\delta'} + 1\big)\sqrt{2\Gamma_{\text{RFE}}(T+1)}. \end{aligned}$$

Since the stopping condition fails through model index $T$, substitution gives Eq. (15). The constants are conservative. The second and fourth terms are the early-count contributions. □

**Theorem D.17** (Stopping Time with Unknown Costs)**.** *With $\varepsilon_0 \in (0,1]$ and $\varepsilon_{\text{est}} = \alpha\varepsilon_0/4$, on $\mathcal{E}_3 \cap \mathcal{E}_4$, Phase II reaches a finite stopping model index satisfying*

$$K_{\text{stop}} = \widetilde{\mathcal{O}}\Big(\frac{S(S + \lfloor 1/\xi\rfloor + 1)AH^3}{\varepsilon_{\text{est}}^2}\Big). \tag{18}$$

*The number of executed episodes is $K_{\text{stop}} - 1$ and obeys the same bound. At the stopped model, the post-hoc estimates in Eq.* (6) *satisfy, simultaneously for every $i \in [N_\xi]$,*

$$|\bar{f}(\bar{\pi}_i) - f(\bar{\pi}_i)| \le \varepsilon_{\text{est}}, \qquad |\bar{g}_i(\bar{\pi}_i) - g_i(\bar{\pi}_i)| \le \frac{\varepsilon_{\text{est}}}{\alpha}.$$

*Moreover,* $\Pr(\mathcal{E}_3 \cap \mathcal{E}_4) \ge 1 - 4\delta'$.

*Proof.* Condition on $\mathcal{E}_3 \cap \mathcal{E}_4$. Lemmas C.5 and C.6 give the three components of $\mathcal{E}_3$, and Lemma C.4 gives $\mathcal{E}_4$. Their simultaneous probability is at least $1 - 4\delta'$.

Algorithm 2 first tests model index 2. For the analysis, define the virtual uncertainty function $J^1$ by the same recursion as Eq. (14) at the initialized zero counts. The zero-count bonuses give $b^1_{\text{RFE}} = 4H + 1$. Backward induction then gives $J^1_1(s_1, 0, a) = H$ for every action. Since $\varepsilon_{\text{est}} = \alpha\varepsilon_0/4 \le 1/4 < H$, this virtual test fails.

Thus, if the first stopping index is $K_{\text{stop}}$, the virtual index 1 together with the actually tested indices $2, \dots, K_{\text{stop}} - 1$ forms a valid failed prefix. More generally, for any $T \ge 1$ whose tests fail at model indices $1, \dots, T$, Lemma D.16 gives Eq. (15).

Since $T \ge 1$, the definition of $\Gamma_{\text{RFE}}(T+1)$ gives

$$\Gamma_{\text{RFE}}(T+1) \le (S + \lfloor 1/\xi \rfloor + 5) \log \frac{24eSAHT}{\delta'}, \qquad \log(T+1) \le \log \frac{24eSAHT}{\delta'}.$$

Indeed, the four terms in $\Gamma_{\text{RFE}}(T+1)$ are bounded by one, $S$, $\lfloor 1/\xi \rfloor + 1$, and three copies of the common logarithm, respectively. After dividing Eq. (15) by $\varepsilon_{\text{est}}$, compare its terms directly with Lemma H.6 under the common logarithm $\log(24eSAHT/\delta')$. The coefficient of its square-root term is

$$\frac{12H\sqrt{10SAH(S + \lfloor 1/\xi \rfloor + 5)}}{\varepsilon_{\text{est}}}.$$

The coefficient of all remaining logarithmic terms is bounded by the larger of

$$\frac{64SAH^2(S + \lfloor 1/\xi \rfloor + 5)}{3\varepsilon_{\text{est}}}$$

and

$$\frac{6SAH^2}{\varepsilon_{\text{est}}}\left(4\log\frac{SA}{\delta'} + 1\right)\sqrt{2(S + \lfloor 1/\xi \rfloor + 5)} + \frac{160SAH^2}{\varepsilon_{\text{est}}} + \frac{4SAH^2(S + \lfloor 1/\xi \rfloor + 5)}{3\varepsilon_{\text{est}}}\left(4\log\frac{SA}{\delta'} + 1\right) + \frac{10SAH^2}{\varepsilon_{\text{est}}}\left(4\log\frac{SA}{\delta'} + 1\right).$$

Denote the displayed square-root coefficient by $C_{\text{RFE}}$ and the larger of the two remaining coefficients by $D_{\text{RFE}}$. Every failed prefix therefore satisfies Lemma H.6 with

$$m = T, \quad \mathscr{C} = C_{\text{RFE}}, \quad \mathscr{D} = D_{\text{RFE}}, \quad \mathscr{A} = \mathscr{B} = \mathscr{E} = 1, \quad \kappa = \frac{24eSAH}{\delta'}.$$

Under this substitution, the right-hand side in the conclusion of Lemma H.6 is a deterministic bound that depends only on the problem parameters. Choose $T$ to be one plus its ceiling. If all tests through model index $T$ failed, the same lemma would bound $T$ by the strictly smaller deterministic endpoint, a contradiction. Hence the actual first stopping index satisfies $K_{\text{stop}} \le T$ and is finite.

The algorithm records $K_{\text{stop}} = k + 1$ after an executed episode $k \ge 1$, which implies $K_{\text{stop}} \ge 2$. The failed prefix before the first stop has length $K_{\text{stop}} - 1$, and the same deterministic endpoint therefore controls the executed episode count.

The square of the displayed square-root coefficient is

$$\frac{1440SAH^3(S + \lfloor 1/\xi \rfloor + 5)}{\varepsilon^2_{\text{est}}}.$$

The leading endpoint term is $C^2_{\text{RFE}}$, displayed above. The remaining endpoint terms are $D_{\text{RFE}}$ and $C_{\text{RFE}}\sqrt{D_{\text{RFE}}}$ up to logarithms. Their polynomial orders are at most $SAH^2(S + \lfloor 1/\xi \rfloor + 5)/\varepsilon_{\text{est}}$ and $SAH^{5/2}(S + \lfloor 1/\xi \rfloor + 5)/\varepsilon^{3/2}_{\text{est}}$, respectively. Both are dominated by the displayed $C^2_{\text{RFE}}$ rate because $\varepsilon_{\text{est}} \le 1$ and $S, A, H \ge 1$. Applying asymptotic notation only after this explicit inversion proves Eq. (18). The corrected factor $8/3$ in Lemma H.6 affects only the absolute constant multiplying the squared logarithm.

At the stopping index, the stopping criterion gives $\max_{a'} J_1^{K_{\text{stop}}}(s_1, 0, a') \le \varepsilon_{\text{est}}$. Lemma D.15 bounds the initial value error for each constituent policy $\pi^{i,k}$ by this maximum. Averaging these bounds as in Eq. (6) gives $|\bar{f}(\bar{\pi}_i) - f(\bar{\pi}_i)| \le \varepsilon_{\text{est}}$ and the same bound for the mixture's terminal

positive-part risk functional. The latter functional is multiplied by $1/\alpha$ in $g_i$, while the fixed terms $\nu_i - \tau$ cancel. Hence $|\bar{g}_i(\bar{\pi}_i) - g_i(\bar{\pi}_i)| \le \varepsilon_{\text{est}}/\alpha$. The outer sums range over the original $SA$ pairs. Therefore, no augmented-state cardinality enters. □

The theorem separates what Phase II must accomplish from how the candidates were learned. A single stopped model evaluates every candidate learned in Phase I, and its uncertainty is controlled using counts indexed only by the original state-action pairs.

### D.3 Phase III: Candidate Selection and the Fixed-Grid Guarantee

At the stopped model, write the constituents of $\bar{\pi}_i$ as $\pi^{i,1}, \dots, \pi^{i,K_{\text{PD}}}$. To define the estimates in Eq. (6), initialize $\overline{U}^{i,k}_{H+1} = 0$ and $\overline{V}^{i,k}_{H+1}(s,y) = (y - \nu_i)^+$. For $h = H, \dots, 1$, compute

$$\begin{aligned}
\overline{W}^{i,k}_h(s,y,a) &= \overline{z}(s,a) + \overline{P}(\cdot \mid s,a)^\top \mathbb{E}_{c \sim \overline{C}_\xi(\cdot|s,a)}\Big(\overline{U}^{i,k}_{h+1}(\cdot, y+c)\Big),\\
\overline{Q}^{i,k}_h(s,y,a) &= \overline{P}(\cdot \mid s,a)^\top \mathbb{E}_{c \sim \overline{C}_\xi(\cdot|s,a)}\Big(\overline{V}^{i,k}_{h+1}(\cdot, y+c)\Big),\\
\overline{U}^{i,k}_h(s,y) &= \langle \pi^{i,k}_h(\cdot \mid s,y), \overline{W}^{i,k}_h(s,y,\cdot)\rangle,\\
\overline{V}^{i,k}_h(s,y) &= \langle \pi^{i,k}_h(\cdot \mid s,y), \overline{Q}^{i,k}_h(s,y,\cdot)\rangle.
\end{aligned}$$

These are policy-evaluation recursions under the common empirical model, with no exploration bonuses or clipping.

Phase I executes $N_\xi K_{\text{PD}}$ episodes, and Phase II executes $K_{\text{stop}} - 1$ episodes because $K_{\text{stop}}$ is the stopped model index. Thus, the exact number of executed episodes is $N_\xi K_{\text{PD}} + (K_{\text{stop}} - 1)$.

**Theorem D.18** (PAC Guarantee on $\text{disc}(\mathcal{M})$)**.** *Suppose Assumption 5.2 holds for the fixed-VaR subproblem $j^*$ on $\text{disc}(\mathcal{M})$ with known margin $\zeta_\xi > 0$, and choose $K_{PD}$ as in Corollary D.14. For every $\delta \in (0,1)$ and $\varepsilon_0 \in (0,1]$, with probability at least $1-\delta$, Phases I–III of Algorithm 1 terminate and select a policy $\bar{\pi}_{i^{\text{out}}}$ satisfying*

$$f(\bar{\pi}_{i^{\text{out}}}) - f(\pi^*_{\text{disc}}) \le \varepsilon_0, \qquad g(\bar{\pi}_{i^{\text{out}}}) \le \varepsilon_0$$

*on $\text{disc}(\mathcal{M})$. On the same event, the exact number of executed episodes is $N_\xi K_{PD} + (K_{\text{stop}} - 1)$. This count is bounded by $K_{\text{total}}$, where*

$$\begin{aligned}
K_{\text{total}} &= N_\xi K_{PD} + K_{\text{stop}},\\
N_\xi K_{PD} &= \widetilde{\mathcal{O}}\left(N_\xi \frac{S(S + \lfloor 1/\xi \rfloor + 1)AH^3 + H^4}{\alpha^2 \varepsilon_0^2}\right) + \widetilde{\mathcal{O}}\left(N_\xi \frac{H^6}{\zeta_\xi^2 \alpha^2 \varepsilon_0^2}\right),\\
K_{\text{stop}} &= \widetilde{\mathcal{O}}\left(\frac{S(S + \lfloor 1/\xi \rfloor + 1)AH^3}{\alpha^2 \varepsilon_0^2}\right).
\end{aligned} \tag{19}$$

*Proof.* Lemmas C.1 and C.2 give the three statements in $\mathcal{E}_1$, and Lemma C.3 gives $\mathcal{E}_2$. Lemmas C.5 and C.6 give the three statements in $\mathcal{E}_3$, and Lemma C.4 gives $\mathcal{E}_4$. A union bound over these eight statements gives total failure probability at most $8\delta' = \delta$. Condition on their intersection. Corollary D.14, with the episode choice in Eq. (13), gives the per-candidate cost of Phase I, which is incurred for all $N_\xi$ candidates. For subproblem $j^*$, it gives

$$f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) \le \frac{\varepsilon_0}{2}, \qquad (g_{j^*}(\bar{\pi}_{j^*}))^+ \le \frac{\varepsilon_0}{2}.$$

Theorem D.17 gives the episode cost of Phase II, incurred once. Its candidate-level conclusion gives, simultaneously for every candidate,

$$|\bar{f}(\bar{\pi}_i) - f(\bar{\pi}_i)| \le \varepsilon_{\text{est}}, \qquad |\bar{g}_i(\bar{\pi}_i) - g_i(\bar{\pi}_i)| \le \frac{\varepsilon_{\text{est}}}{\alpha}.$$

The feasibility filter retains $j^*$ because

$$\bar{g}_{j^*}(\bar{\pi}_{j^*}) \le g_{j^*}(\bar{\pi}_{j^*}) + \frac{\varepsilon_{\text{est}}}{\alpha} \le \varepsilon_{\text{vio}} + \frac{\varepsilon_{\text{est}}}{\alpha}.$$

By the definition of $i^{\text{out}}$ in Algorithm 1, $j^*$ is valid and the selection rule minimizes the estimated performance cost over the valid set. Hence,

$$f(\bar{\pi}_{i^{\text{out}}}) - f(\pi^*_{\text{disc}}) \le \bar{f}(\bar{\pi}_{i^{\text{out}}}) + \varepsilon_{\text{est}} - f(\pi^*_{\text{disc}}) \le \bar{f}(\bar{\pi}_{j^*}) + \varepsilon_{\text{est}} - f(\pi^*_{\text{disc}})$$
$$\le f(\bar{\pi}_{j^*}) - f(\pi^*_{\text{disc}}) + 2\varepsilon_{\text{est}} \le \frac{\varepsilon_0}{2} + 2\varepsilon_{\text{est}} \le \varepsilon_0.$$

Moreover, the variational representation gives $g(\pi) \le g_i(\pi)$ for every $i$. Since $i^{\text{out}}$ is valid,

$$g(\bar{\pi}_{i^{\text{out}}}) \le g_{i^{\text{out}}}(\bar{\pi}_{i^{\text{out}}}) \le \bar{g}_{i^{\text{out}}}(\bar{\pi}_{i^{\text{out}}}) + \frac{\varepsilon_{\text{est}}}{\alpha} \le \varepsilon_{\text{vio}} + \frac{2\varepsilon_{\text{est}}}{\alpha} = \varepsilon_0.$$

Here the last steps use $\varepsilon_{\text{est}} = \alpha\varepsilon_0/4$, $\varepsilon_{\text{vio}} = \varepsilon_0/2$, and $\alpha \le 1$. The exact count $N_\xi K_{\text{PD}} + (K_{\text{stop}} - 1)$ is at most $K_{\text{total}}$, which gives the bound in Eq. (19). This proves the stated statistical guarantee. □

This theorem is the fixed-grid statistical guarantee used by the transfer argument. Its episode count includes all $N_\xi$ primal-dual runs and the single reward-free exploration run, while the selection step requires no additional environment interaction.

# E TRANSFER TO THE ORIGINAL MDP

We transfer the guarantee on $\text{disc}(\mathcal{M})$ in Theorem D.18 to the original MDP $\mathcal{M}$, which has arbitrary risk cost distributions supported on $[0, 1]$. The argument uses pathwise couplings and does not require the kernels $C(\cdot \mid s, a)$ to be known. The MDPs $\mathcal{M}$ and $\text{disc}(\mathcal{M})$ share the performance cost function $z$. The couplings preserve the expected performance cost and control the change in the CVaR constraint by $H\xi$.

The construction builds on ideas from Bastani et al. (2022, Theorem 3.2) and Wang et al. (2023, Section 6). Those works study a CVaR objective. Here we adapt the arguments to a CVaR constraint.

## E.1 ADAPTATION AND COUPLING OF POLICIES

Recall from Section 5 the floor-discretization map $\phi_\xi$, the MDP $\text{disc}(\mathcal{M})$, the cumulative-risk state grid $\mathcal{Y}_\xi$, and the VaR-candidate grid $\mathcal{G}_\xi$. Write $c_{\xi,h} := \phi_\xi(c_h)$ for the discretized one-step risk cost. The pathwise inequality $0 \le c - \phi_\xi(c) < \xi$ is the only property of the discretization map used below.

**Adaptation of Policies.** To connect policies across $\mathcal{M}$ and $\text{disc}(\mathcal{M})$, we introduce two auxiliary constructions.

Given $\pi \in \Pi_{\text{aug}}$ designed for $\text{disc}(\mathcal{M})$, its *adapted policy* $\text{adapted}(\pi)$ runs in $\mathcal{M}$. The original environment continues to generate the risk costs $c_h$. Internally, $\text{adapted}(\pi)$ applies $\phi_\xi$ to every observation and maintains the discretized cumulative coordinate in $\mathcal{Y}_\xi$. Its internal augmented-state dynamics therefore match those of $\text{disc}(\mathcal{M})$:

$$\text{adapted}(\pi)_h(s_h,\, c_{1:h-1}) = \pi_h\left(s_h, \sum_{t=1}^{h-1} \phi_\xi(c_t)\right).$$

For a finite episode-level mixture, adaptation first samples the mixture component and then applies this construction to the sampled policy for the entire episode.

Conversely, let $\pi \in \Pi_{\text{aug}}$ be designed for $\mathcal{M}$. After observing the discretized risk cost $c_{\xi,h}$, the policy $\text{disc}(\pi)$ samples a latent risk cost $c_h$ conditional on its discretized value. For every measurable $B \subseteq [0, 1]$ and every $c' \in \mathcal{C}_\xi$ with $\Pr_{c\sim C(\cdot|s,a)}(\phi_\xi(c) = c') > 0$, the conditional distribution is

$$\Pr(c_h \in B \mid s_h = s, a_h = a, c_{\xi,h} = c') = \frac{\Pr_{c\sim C(\cdot|s,a)}(c \in B,\ \phi_\xi(c) = c')}{\Pr_{c\sim C(\cdot|s,a)}(\phi_\xi(c) = c')}.$$

At zero-probability grid points, the conditional distribution may be chosen arbitrarily. This reverse construction is used only to compare optimal values in the transfer proof. The learning algorithm never evaluates or samples from this conditional distribution. The policy stores the latent costs as internal memory and acts according to

$$\text{disc}(\pi)_h(a \mid s_h, c_{\xi,1:h-1}, c_{1:h-1}) = \pi_h\left(a \mid s_h, \sum_{t=1}^{h-1} c_t\right).$$

Integrating the conditional distribution over the discretized risk cost gives $c_h \sim C(\cdot \mid s_h, a_h)$ for every realized state-action history. An induction over $h$ then shows that $\mathrm{disc}(\pi)$ in $\mathrm{disc}(\mathcal{M})$ and $\pi$ in $\mathcal{M}$ induce the same joint distribution of $(s_h, a_h, c_h)_{h=1}^{H}$. Here $c_h$ is latent in $\mathrm{disc}(\mathcal{M})$ and observed in $\mathcal{M}$. The two MDPs also share the performance cost function $z$. The matched state-action distributions therefore imply

$$\mathbb{E}_{\mathrm{disc}(\pi),\,\mathrm{disc}(\mathcal{M})}\left[\sum_{h=1}^{H} z(s_h, a_h)\right] = \mathbb{E}_{\pi,\,\mathcal{M}}\left[\sum_{h=1}^{H} z(s_h, a_h)\right]. \tag{20}$$

The latent-cost construction makes $\mathrm{disc}(\pi)$ a history-dependent policy with internal memory rather than a member of $\Pi_{\mathrm{aug}}$. The following reduction is the finite-grid specialization of the argument in Appendix B. We state it locally because the reconstructed policy uses latent internal memory. It replaces that policy by an augmented-Markov policy for optimization and Slater comparisons over $\Pi_{\mathrm{aug}}$.

For any policy $\pi$ on $\mathrm{disc}(\mathcal{M})$, define its Markovization on every reachable augmented state by

$$\widetilde{\pi}_h(a \mid s, y) \coloneqq \Pr_{\pi}(a_h = a \mid s_h = s, y_h = y), \qquad y_h = \sum_{t<h} c_t^{\mathrm{disc}}.$$

The action distribution may be chosen arbitrarily at unreachable augmented states.

**Lemma E.1** (Augmented-Markov Reduction of Memory Policies)**.** *For any possibly history-dependent policy $\pi$ on* $\mathrm{disc}(\mathcal{M})$*, the Markovization defined above belongs to* $\Pi_{\mathrm{aug}}$ *and satisfies*

$$C^{\widetilde{\pi}}_{\mathrm{disc}(\mathcal{M})} \overset{\mathrm{d}}{=} C^{\pi}_{\mathrm{disc}(\mathcal{M})} \qquad \textit{and} \qquad f(\widetilde{\pi}) = f(\pi). \tag{21}$$

*In particular, the conclusion applies to the reconstructed policy* $\mathrm{disc}(\pi)$.

*Proof.* The definition of $\widetilde{\pi}$ matches the conditional action distribution of $\pi$ at every reachable augmented state. We first prove by induction that the two policies have the same distribution of $(s_h, y_h)$. The claim holds at $h = 1$ because they share the initial-state distribution and $y_1 = 0$. If it holds at step $h$, the definition of $\widetilde{\pi}_h$ gives the same joint distribution of $(s_h, y_h, a_h)$. The common transition kernel $P$ and discretized risk cost kernel $C_\xi(\cdot \mid s_h, a_h)$ then give the same distribution of $(s_{h+1}, y_{h+1})$. This proves the induction step.

The terminal identity $C^{\pi}_{\mathrm{disc}(\mathcal{M})} = y_{H+1}$ shows that the cumulative risk costs have the same distribution. Moreover, $f$ is the expectation of $\sum_{h=1}^{H} z(s_h, a_h)$ and is therefore determined by the matched state-action occupancy measures. Both conclusions in Eq. (21) follow. □

Thus, the auxiliary memory is used only to establish a coupling. It does not enlarge the set of risk distributions or expected performance costs attainable by augmented-Markov policies.

### E.2 Stochastic Dominance

**Lemma E.2** (From $\mathrm{disc}(\mathcal{M})$ to $\mathcal{M}$)**.** *For any $\pi \in \Pi_{\mathrm{aug}}$ designed for* $\mathrm{disc}(\mathcal{M})$*, or any finite episode-level mixture of such policies, almost surely*

$$C^{\pi}_{\mathrm{disc}(\mathcal{M})} \le C^{\mathrm{adapted}(\pi)}_{\mathcal{M}} \le C^{\pi}_{\mathrm{disc}(\mathcal{M})} + H\xi. \tag{22}$$

*Consequently,*

$$\Pr\Big(C^{\mathrm{adapted}(\pi)}_{\mathcal{M}} \le t\Big) \le \Pr\Big(C^{\pi}_{\mathrm{disc}(\mathcal{M})} \le t\Big) \le \Pr\Big(C^{\mathrm{adapted}(\pi)}_{\mathcal{M}} \le t + H\xi\Big), \qquad \forall\, t \in \mathbb{R}. \tag{23}$$

*Proof.* For an episode-level mixture, first sample the same component in both MDPs and use that component throughout the coupled episode. Conditional on this draw, the following construction applies unchanged. We construct the trajectories on a common probability space. Sample one initial state and use it in both MDPs. The induction invariant is that the trajectories have the same state $s_h$ and the same discretized accumulated risk cost $y_h = \sum_{t<h} \phi_\xi(c_t)$. Conditional on this aligned

augmented state, adapted($\pi$) and $\pi$ have the same action distribution. A common uniform random variable therefore produces the same action in both trajectories.

Given the shared pair $(s_h, a_h)$, draw one risk cost $c_h \sim C(\cdot \mid s_h, a_h)$. Assign $c_h$ to the trajectory in $\mathcal{M}$ and $c_{\xi,h} = \phi_\xi(c_h)$ to the trajectory in disc($\mathcal{M}$). Use another common uniform random variable to sample the next state from $P(\cdot \mid s_h, a_h)$. This construction preserves the induction invariant at step $h+1$, including when the policies are randomized. It follows that the state-action paths and discretized accumulated risk costs agree at every step. Since $0 \le c_h - \phi_\xi(c_h) < \xi$ for all $h$,

$$C_{\mathcal{M}}^{\text{adapted}(\pi)} - C_{\text{disc}(\mathcal{M})}^{\pi} = \sum_{h=1}^{H} (c_h - \phi_\xi(c_h)) \in [0,\, H\xi],$$

which gives Eq. (22). The pathwise inequalities imply the event inclusions

$$\{C_{\mathcal{M}}^{\text{adapted}(\pi)} \le t\} \subseteq \{C_{\text{disc}(\mathcal{M})}^{\pi} \le t\} \subseteq \{C_{\mathcal{M}}^{\text{adapted}(\pi)} \le t + H\xi\}.$$

Taking probabilities gives Eq. (23). □

**Lemma E.3** (From $\mathcal{M}$ to disc($\mathcal{M}$))**.** *For any $\pi \in \Pi_{\text{aug}}$ designed for $\mathcal{M}$, almost surely*

$$C_{\mathcal{M}}^{\pi} - H\xi \le C_{\text{disc}(\mathcal{M})}^{\text{disc}(\pi)} \le C_{\mathcal{M}}^{\pi}, \tag{24}$$

*and consequently* $\Pr(C_{\mathcal{M}}^{\pi} \le t) \le \Pr\Big(C_{\text{disc}(\mathcal{M})}^{\text{disc}(\pi)} \le t\Big) \le \Pr(C_{\mathcal{M}}^{\pi} \le t + H\xi)$ *for all* $t \in \mathbb{R}$.

*Proof.* Use the conditional distribution above to generate the latent costs of disc($\pi$). Conditional on $(s_h, a_h, c_{\xi,h})$, use the same latent cost $c_h$ as the true cost of the coupled trajectory in $\mathcal{M}$. Integrating over $c_{\xi,h}$ shows that, conditional on every preceding history and $(s_h, a_h)$, this cost has distribution $C(\cdot \mid s_h, a_h)$. Use common transition randomness to generate the next state in both trajectories. Induction over $h$ then produces a valid trajectory of $\pi$ in $\mathcal{M}$. The two policies use the same latent cumulative cost history and therefore select the same actions. The common transition randomness keeps their states aligned. Since $c_{\xi,h} = \phi_\xi(c_h)$ satisfies $0 \le c_h - \phi_\xi(c_h) < \xi$,

$$C_{\mathcal{M}}^{\pi} - H\xi = \sum_{h=1}^{H} (c_h - \xi) \le \sum_{h=1}^{H} c_{\xi,h} = C_{\text{disc}(\mathcal{M})}^{\text{disc}(\pi)} \le \sum_{h=1}^{H} c_h = C_{\mathcal{M}}^{\pi}.$$

This proves Eq. (24). The same event-inclusion argument as in Lemma E.2 gives the CDF sandwich. □

The two stochastic-dominance lemmas are pathwise. They therefore apply without any smoothness or density assumption on the risk cost distributions. The forward coupling also matches the state-action path for a finite episode-level mixture, so it preserves that mixture's expected performance objective.

**CVaR Transfer Inequalities.** Combining the pathwise couplings with the Rockafellar–Uryasev formula in Eq. (1) yields quantitative CVaR transfers.

**Corollary E.4** (CVaR Sandwich)**.** *Every $\pi \in \Pi_{\text{aug}}$ designed for* disc($\mathcal{M}$)*, and every finite episode-level mixture of such policies, satisfies*

$$\text{CVaR}_\alpha\Big(C_{\text{disc}(\mathcal{M})}^{\pi}\Big) \le \text{CVaR}_\alpha\Big(C_{\mathcal{M}}^{\text{adapted}(\pi)}\Big) \le \text{CVaR}_\alpha\Big(C_{\text{disc}(\mathcal{M})}^{\pi}\Big) + H\xi, \tag{25}$$

*whereas every $\pi \in \Pi_{\text{aug}}$ designed for $\mathcal{M}$ satisfies*

$$\text{CVaR}_\alpha\Big(C_{\text{disc}(\mathcal{M})}^{\text{disc}(\pi)}\Big) \le \text{CVaR}_\alpha(C_{\mathcal{M}}^{\pi}). \tag{26}$$

*Proof.* For every $\nu \in \mathbb{R}$, Eq. (22) and monotonicity of the positive-part function give

$$(C_{\text{disc}(\mathcal{M})}^{\pi} - \nu)^+ \le (C_{\mathcal{M}}^{\text{adapted}(\pi)} - \nu)^+ \le (C_{\text{disc}(\mathcal{M})}^{\pi} + H\xi - \nu)^+.$$

Taking expectations and minimizing the Rockafellar–Uryasev expression over $\nu$ yields

$$\text{CVaR}_\alpha\Big(C_{\text{disc}(\mathcal{M})}^{\pi}\Big) \le \text{CVaR}_\alpha\Big(C_{\mathcal{M}}^{\text{adapted}(\pi)}\Big) \le \text{CVaR}_\alpha\Big(C_{\text{disc}(\mathcal{M})}^{\pi} + H\xi\Big).$$

For the last term, set $\nu' = \nu - H\xi$. Then

$$\begin{aligned}\mathrm{CVaR}_\alpha\Big(C^\pi_{\mathrm{disc}(\mathcal{M})} + H\xi\Big) &= \min_{\nu\in\mathbb{R}}\left\{\nu + \frac{1}{\alpha}\mathbb{E}\Big((C^\pi_{\mathrm{disc}(\mathcal{M})} + H\xi - \nu)^+\Big)\right\} \\ &= H\xi + \min_{\nu'\in\mathbb{R}}\left\{\nu' + \frac{1}{\alpha}\mathbb{E}\Big((C^\pi_{\mathrm{disc}(\mathcal{M})} - \nu')^+\Big)\right\} \\ &= H\xi + \mathrm{CVaR}_\alpha\Big(C^\pi_{\mathrm{disc}(\mathcal{M})}\Big).\end{aligned}$$

This proves Eq. (25). The pathwise upper bound in Eq. (24) and monotonicity give Eq. (26) for every policy designed for $\mathcal{M}$. □

Floor discretization can only lower CVaR, while adapting a policy back to the original MDP can increase it by at most $H\xi$. This is the discretization error used in Theorem 5.1.

### E.3 Suboptimality and Constraint Violation Transfer

Recall that $\pi^*$ is optimal for $\mathcal{M}$ and that $\pi^*_{\mathrm{disc}}$ is optimal for $\mathrm{disc}(\mathcal{M})$. By Theorem 4.1, we choose $\pi^*$ to belong to $\Pi_{\mathrm{aug}}$.

*Proof of Theorem 5.1. Constraint Violation.* Apply the right inequality of the CVaR sandwich in Eq. (25) to $\pi$. This gives

$$\begin{aligned}g(\mathrm{adapted}(\pi)) &= \mathrm{CVaR}_\alpha\Big(C^{\mathrm{adapted}(\pi)}_{\mathcal{M}}\Big) - \tau \le \mathrm{CVaR}_\alpha\Big(C^\pi_{\mathrm{disc}(\mathcal{M})}\Big) + H\xi - \tau \\ &= \mathrm{CVaR}_\alpha\Big(C^\pi_{\mathrm{disc}(\mathcal{M})}\Big) - \tau + H\xi \le \varepsilon_0 + H\xi.\end{aligned}$$

*Suboptimality.* We first construct a feasible policy for $\mathrm{disc}(\mathcal{M})$ from $\pi^*$. By Lemma E.3, $\mathrm{CVaR}_\alpha(C^{\mathrm{disc}(\pi^*)}_{\mathrm{disc}(\mathcal{M})}) \le \mathrm{CVaR}_\alpha(C^{\pi^*}_{\mathcal{M}}) \le \tau$. The reconstructed policy $\mathrm{disc}(\pi^*)$ may depend on its latent-cost history and therefore need not belong to $\Pi_{\mathrm{aug}}$. Lemma E.1 provides an augmented-Markov representative $\widetilde{\pi}^* \in \Pi_{\mathrm{aug}}$ such that $C^{\widetilde{\pi}^*}_{\mathrm{disc}(\mathcal{M})} \overset{\mathrm{d}}{=} C^{\mathrm{disc}(\pi^*)}_{\mathrm{disc}(\mathcal{M})}$ and $f(\widetilde{\pi}^*) = f(\mathrm{disc}(\pi^*))$. The distributional identity makes $\widetilde{\pi}^*$ feasible for $\mathrm{disc}(\mathcal{M})$.

Eq. (20) also gives $f(\mathrm{disc}(\pi^*)) = f(\pi^*)$. Hence, the optimal feasible performance value on $\mathrm{disc}(\mathcal{M})$ is at most $f(\widetilde{\pi}^*) = f(\pi^*)$. By optimality of $\pi^*_{\mathrm{disc}}$,

$$f(\pi^*_{\mathrm{disc}}) \le f(\widetilde{\pi}^*) = f(\pi^*),$$

which establishes the required comparison between the optimal values.

Finally, the coupling in the proof of Lemma E.2 matches the state-action trajectory of $\mathrm{adapted}(\pi)$ in $\mathcal{M}$ with that of $\pi$ in $\mathrm{disc}(\mathcal{M})$. Since the performance cost function $z$ is shared by the two MDPs and the coupling matches their state-action paths, $f(\mathrm{adapted}(\pi)) = f(\pi)$. Hence,

$$f(\mathrm{adapted}(\pi)) - f(\pi^*) = f(\pi) - f(\pi^*) \le f(\pi) - f(\pi^*_{\mathrm{disc}}) \le \varepsilon_0.$$ □

**Remark.** The transfer in Eq. (4) has a one-sided error of size $H\xi$. Floor discretization makes $\mathrm{disc}(\mathcal{M})$ a relaxation of $\mathcal{M}$: every feasible policy in $\mathcal{M}$ has a feasible representative in $\mathrm{disc}(\mathcal{M})$ with the same performance cost. Suboptimality therefore transfers without a discretization term, while the discretization error appears in feasibility. Ceiling discretization would reverse this asymmetry and would require a separate perturbation argument for the optimal performance value.

## F Global PAC Guarantee and Total Sample Complexity (Proof of Theorem 6.1)

*Proof.* Apply Theorem D.18 with accuracy $\varepsilon_0 = \frac{\varepsilon}{2}$ and confidence parameter $\delta$. With probability at least $1-\delta$, the selected candidate $\bar{\pi}_{i^{\mathrm{out}}}$ is $\frac{\varepsilon}{2}$-suboptimal and $\frac{\varepsilon}{2}$-feasible on $\mathrm{disc}(\mathcal{M})$. With $\xi = \frac{\varepsilon}{2H}$, on every path where this guarantee holds, Theorem 5.1 shows that executing the adapted policy in $\mathcal{M}$

introduces no additional performance error and at most $H\xi = \frac{\varepsilon}{2}$ additional CVaR violation. Thus, the algorithm is $(\delta, \varepsilon)$-correct on the original MDP $\mathcal{M}$.

On the same event, Theorem D.18 also guarantees that the number of episodes used is bounded by $K_{\text{total}}$. Since $\varepsilon < \tau$ and $H \geq 1$, $1 \leq N_\xi = \lfloor \frac{\tau}{\xi} \rfloor + 1 \leq 1 + \frac{2H\tau}{\varepsilon} \leq \frac{3H\tau}{\varepsilon}$, and the one-step discretized risk cost alphabet $\mathcal{C}_\xi$ has at most $\frac{1}{\xi} + 1$ elements. Direct substitution into Eq. (19) gives

$$\begin{aligned} K_{\text{total}} &= N_\xi K_{\text{PD}} + K_{\text{stop}} \\ &= \widetilde{\mathcal{O}}\left( \frac{(N_\xi + 1)SAH^3\big(S + \frac{H}{\varepsilon}\big) + N_\xi H^4}{\alpha^2\varepsilon^2} + \frac{N_\xi H^6}{\zeta_\xi^2 \alpha^2 \varepsilon^2} \right) \\ &= \widetilde{\mathcal{O}}\left( \frac{N_\xi SAH^3\big(S + \frac{H}{\varepsilon}\big)}{\alpha^2\varepsilon^2} + \frac{N_\xi H^6}{\zeta_\xi^2 \alpha^2 \varepsilon^2} \right) = \widetilde{\mathcal{O}}\left( \frac{SAH^4\tau\big(S + \frac{H}{\varepsilon}\big)}{\alpha^2\varepsilon^3} + \frac{H^7\tau}{\zeta_\xi^2 \alpha^2 \varepsilon^3} \right) \\ &= \widetilde{\mathcal{O}}\left( \frac{S^2AH^4\tau}{\alpha^2\varepsilon^3} + \frac{SAH^5\tau}{\alpha^2\varepsilon^4} + \frac{H^7\tau}{\zeta_\xi^2 \alpha^2 \varepsilon^3} \right). \end{aligned}$$

The two terms contributed by Phase II have orders $\frac{S^2AH^3}{\alpha^2\varepsilon^2}$ and $\frac{SAH^4}{\alpha^2\varepsilon^3}$. They are absorbed by the corresponding estimation terms in Phase I because $N_\xi \geq 1$. The $N_\xi H^4$ term is absorbed using $H^4 \leq \frac{H^6}{\zeta_\xi^2}$, since Definition 3.1 and Assumption 5.2 imply $\zeta_\xi \leq \tau \leq H$. This proves the episode bound in Theorem 6.1. □

## G Proofs for Section 7: Information-Theoretic Lower Bound

This section contains the full proofs of the lower bound results stated in Section 7. The central idea is to embed many local output-separation problems into the MDP. In the auxiliary problem, an algorithm must commit a stochastic action distribution that places large mass on a designated action. The MDP construction later turns these designated actions into the actions required by the PAC guarantee. Since each local separation is sample-expensive and the states hosting these local problems are reached only through an $\alpha$-probability branch, solving the MDP is sample-expensive. We first prove the auxiliary separation lemma and then invoke it inside the MDP construction.

**Auxiliary Output-Separation Problem.** Consider an auxiliary bandit with $m + 1$ actions $a_0, a_1, \ldots, a_m$, where each action produces either a safe outcome or a high risk-cost outcome. An agent sequentially plays actions, possibly using internal randomization, stops at an almost surely finite stopping time, and then commits a stochastic policy $\pi^{\text{out}} \in \triangle_{\mathcal{A}}$. Let $T_i$ be the number of times $a_i$ is played.

Fix a bias parameter $\gamma \in (0, \frac{1}{4})$, and consider instances $\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_m$, where $p_\ell$ denotes the probability that action $a_\ell$ produces the safe outcome. Under the reference instance $\mathcal{I}_0$,

$$p_0 = \tfrac{1}{2} + \gamma, \qquad p_i = \tfrac{1}{2} \ \ \forall i \in [m].$$

Thus $a_0$ is the only action whose safe-outcome probability exceeds $\frac{1}{2}$. Under the changed instance $\mathcal{I}_j$ for each $j \in [m]$,

$$p_0 = \tfrac{1}{2} + \gamma, \qquad p_j = \tfrac{1}{2} + \gamma, \qquad p_i = \tfrac{1}{2} \ \ \forall i \in [m] \setminus \{j\}.$$

Thus $\mathcal{I}_0$ and $\mathcal{I}_j$ differ only in the safe-outcome probability of $a_j$, which rises from $\frac{1}{2}$ to $\frac{1}{2} + \gamma$. For use in the MDP reduction, define the designated target regions

$$R_{a_\ell} \coloneqq \big\{\pi \in \triangle_{\mathcal{A}} : \ \pi(a_\ell) > \tfrac{3}{4}\big\}, \quad \ell \in \{0, 1, \ldots, m\},$$

which are pairwise disjoint since no two actions can each carry more than $\frac{3}{4}$ of the mass. These regions are not optimality regions of the auxiliary bandit. Indeed, under $\mathcal{I}_j$, both $a_0$ and $a_j$ maximize the safe-outcome probability. The lemma below only lower-bounds the sample cost of forcing different output regions under $\mathcal{I}_0$ and $\mathcal{I}_j$.

For $p, q \in [0, 1]$, we write $\mathrm{kl}(p, q) \coloneqq p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$ for the Kullback–Leibler divergence between Bernoulli distributions with parameters $p$ and $q$. The following lemma lower-bounds the

number of plays of the changed action $a_j$ needed to separate these target regions. The two failure tolerances play asymmetric roles: a constant tolerance suffices on the reference instance $\mathcal{I}_0$, while the tolerance $\vartheta$ on the changed instance $\mathcal{I}_j$ enters the bound through a $\log(1/\vartheta)$ factor. Therefore, a smaller tolerance forces more plays.

**Lemma G.1.** *Consider the auxiliary output-separation problem, the target regions $R_{a_0}, \dots, R_{a_m}$, and the instances $\mathcal{I}_0, \mathcal{I}_1, \dots, \mathcal{I}_m$ described above, with $\gamma \in (0, \frac{1}{4})$. Fix any $\vartheta \in (0, \frac{1}{4})$ and any index $j \in [m]$. Then any algorithm that stops almost surely and whose committed policy satisfies $\Pr_{\mathcal{I}_0}[\pi^{\text{out}} \notin R_{a_0}] \le \frac{1}{4}$ and $\Pr_{\mathcal{I}_j}[\pi^{\text{out}} \notin R_{a_j}] \le \vartheta$ must satisfy*

$$\mathbb{E}_{\mathcal{I}_0}[T_j] \ge \frac{3}{16\gamma^2} \log\left(\frac{1}{4\vartheta}\right).$$

*Proof.* The proof is a change of measure through the divergence decomposition of Lemma H.2. It uses only the separation of target regions, not uniqueness of an optimal action in the auxiliary bandit.

The committed policy $\pi^{\text{out}}$ is a deterministic measurable function of the play history, observed transition outcomes, and the algorithm's internal randomization. The distribution of the algorithm's internal randomization is instance-independent. Hence, the indicator $\mathbb{1}\{\pi^{\text{out}} \in R_{a_0}\}$ qualifies as the random variable $\chi$ in Lemma H.2. Its expectation under each instance is the probability of the event $\{\pi^{\text{out}} \in R_{a_0}\}$.

The transition outcome of action $a_j$ follows $\mathrm{Ber}(\frac{1}{2})$ under $\mathcal{I}_0$ and $\mathrm{Ber}(\frac{1}{2}+\gamma)$ under $\mathcal{I}_j$, and every other action has identical outcome distributions under the two instances. Hence, Lemma H.2 applied with $\chi = \mathbb{1}\{\pi^{\text{out}} \in R_{a_0}\}$ gives

$$\mathbb{E}_{\mathcal{I}_0}[T_j] \cdot \mathrm{kl}\big(\tfrac{1}{2}, \tfrac{1}{2}+\gamma\big) \ge \mathrm{kl}\left(\Pr_{\mathcal{I}_0}[\pi^{\text{out}} \in R_{a_0}], \Pr_{\mathcal{I}_j}[\pi^{\text{out}} \in R_{a_0}]\right). \tag{27}$$

Here the left-hand side is the total information that the plays of $a_j$ can supply—the number of plays times the per-play divergence—while the right-hand side is the information the algorithm must have gathered to separate its output distributions on the two instances. Rearranging therefore lower-bounds the number of plays. We bound the two sides of Eq. (27) separately. For the left-hand side, a direct computation gives

$$\mathrm{kl}\big(\tfrac{1}{2}, \tfrac{1}{2}+\gamma\big) = \frac{1}{2}\log\left(\frac{1}{1-4\gamma^2}\right) \le \frac{1}{2} \cdot \frac{4\gamma^2}{1-4\gamma^2} \le 4\gamma^2,$$

where the first inequality uses $\log \frac{1}{1-x} \le \frac{x}{1-x}$ for $x \in [0,1)$, and the second uses $4\gamma^2 < \frac{1}{4}$, which holds since $\gamma < \frac{1}{4}$. Each play of the action whose safe-outcome probability changes from $\frac{1}{2}$ to $\frac{1}{2}+\gamma$ supplies little information, which is what forces roughly $\gamma^{-2}$ plays.

For the right-hand side, write $p \coloneqq \Pr_{\mathcal{I}_0}[\pi^{\text{out}} \in R_{a_0}]$ and $q \coloneqq \Pr_{\mathcal{I}_j}[\pi^{\text{out}} \in R_{a_0}]$. The assumption $\Pr_{\mathcal{I}_0}[\pi^{\text{out}} \notin R_{a_0}] \le \frac{1}{4}$ gives $p \ge \frac{3}{4}$. Since $R_{a_0} \cap R_{a_j} = \emptyset$, the event $\{\pi^{\text{out}} \in R_{a_0}\}$ implies $\{\pi^{\text{out}} \notin R_{a_j}\}$. Therefore, the assumption $\Pr_{\mathcal{I}_j}[\pi^{\text{out}} \notin R_{a_j}] \le \vartheta$ gives $q \le \Pr_{\mathcal{I}_j}[\pi^{\text{out}} \notin R_{a_j}] \le \vartheta$. For any $p \in [0,1]$ and $q \in (0,1)$, with the convention $0 \log 0 = 0$, we have

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \ge p\log\frac{1}{q} - \log 2,$$

since $p\log p + (1-p)\log(1-p) \ge -\log 2$ and $(1-p)\log\frac{1}{1-q} \ge 0$. If $q = 0$, then $\mathrm{kl}(p,q) = \infty$ for $p > 0$ and the bound holds trivially. Substituting $p \ge \frac{3}{4}$ and $q \le \vartheta$, we obtain

$$\mathrm{kl}(p,q) \ge \frac{3}{4}\log\frac{1}{\vartheta} - \log 2 = \frac{3}{4}\log\left(\frac{1}{4\vartheta}\right) + \frac{1}{2}\log 2 \ge \frac{3}{4}\log\left(\frac{1}{4\vartheta}\right).$$

Substituting the two bounds into Eq. (27) yields $\mathbb{E}_{\mathcal{I}_0}[T_j] \cdot 4\gamma^2 \ge \frac{3}{4}\log\big(\frac{1}{4\vartheta}\big)$, i.e., $\mathbb{E}_{\mathcal{I}_0}[T_j] \ge \frac{3}{16\gamma^2}\log\big(\frac{1}{4\vartheta}\big)$, as claimed. □

With Lemma G.1 established, we now prove the main sample complexity lower bound by realizing the local output-separation problems inside the MDP through the probability of avoiding $s_{\text{rcost}}$.

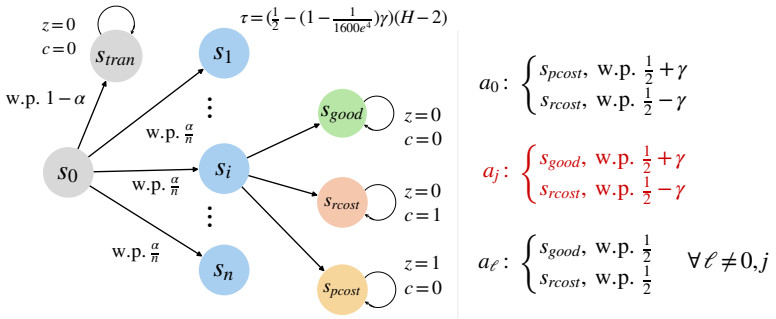



Figure 5: Hard instance for the lower bound.

*Proof of Theorem 7.1.* This proof follows the parallel-bandit lower-bound strategy of Dann & Brunskill (2015), while using a coordinate-freezing change-of-measure step to localize each comparison to one embedded bandit state.

Before specifying the instance, we isolate the three roles of the construction. First, PAC correctness must force the output policy to place large mass on a designated action on average across the embedded local problems. Second, feasibility and optimality must enforce this requirement in complementary ways: feasibility favors lower-risk actions, while optimality selects, among them, the one with lower performance cost. Third, these local problems must lie behind an $\alpha$-probability gate. Therefore, local samples translate into a factor $1/\alpha$ more episodes.

Consider the instance in Figure 5. There are states $s_0, s_{\text{tran}}, s_{\text{good}}, s_{\text{rcost}}, s_{\text{pcost}}, s_1, \ldots, s_n$ with $n = S - 5$, and actions $a_0, a_1, \ldots, a_m$ with $m = A - 1$. The two costs are deterministic and separated across states: the risk cost satisfies $c(s_{\text{rcost}}, \cdot) = 1$ and $c(s, \cdot) = 0$ elsewhere, while the performance cost satisfies $z(s_{\text{pcost}}, \cdot) = 1$ and $z(s, \cdot) = 0$ elsewhere. The transition from $s_0$ is action-independent: for every $a \in \mathcal{A}$, $P(s_i \mid s_0, a) = \frac{\alpha}{n}$ for all $i \in [n]$ and $P(s_{\text{tran}} \mid s_0, a) = 1 - \alpha$. The states $s_{\text{tran}}, s_{\text{good}}, s_{\text{rcost}}, s_{\text{pcost}}$ are absorbing. Hence, the agent either stays in $s_{\text{tran}}$ from step 2 onward or reaches one of the bandit states at step 2.

At each state $s_i$, $i \in [n]$, which we call a bandit state, we instantiate a local copy of the auxiliary separation problem through the probability of avoiding the risk cost state $s_{\text{rcost}}$. The bias parameter is $\gamma \in (0, \frac{1}{4})$, whose value is fixed later. Under the reference instance $\mathcal{I}_0^i$, the fallback action $a_0$ is lower-risk but performance costly:

$$P(s_{\text{pcost}} \mid s_i, a_0) = \tfrac{1}{2} + \gamma, \qquad P(s_{\text{rcost}} \mid s_i, a_0) = \tfrac{1}{2} - \gamma.$$

Thus $a_0$ lowers the probability of $s_{\text{rcost}}$ relative to the non-fallback actions, but pays for this reduction by entering the performance cost state $s_{\text{pcost}}$. Under $\mathcal{I}_0^i$, every non-fallback action is performance-free and transitions to $s_{\text{good}}$ and $s_{\text{rcost}}$ with equal probability:

$$P(s_{\text{good}} \mid s_i, a_\ell) = P(s_{\text{rcost}} \mid s_i, a_\ell) = \tfrac{1}{2}, \quad \forall \ell \in [m].$$

Consequently, $a_0$ is the only lower-risk action in $\mathcal{I}_0^i$. Under the changed instance $\mathcal{I}_j^i$ for each $j \in [m]$, only action $a_j$ is modified. It becomes lower-risk without adding performance cost:

$$P(s_{\text{good}} \mid s_i, a_j) = \tfrac{1}{2} + \gamma, \qquad P(s_{\text{rcost}} \mid s_i, a_j) = \tfrac{1}{2} - \gamma.$$

All other non-fallback actions keep the reference transitions. Thus $\mathcal{I}_0^i$ and $\mathcal{I}_j^i$ differ only in the transition kernel of $(s_i, a_j)$. After this bandit transition, the trajectory enters one of $s_{\text{good}}, s_{\text{rcost}}, s_{\text{pcost}}$ at step 3 and remains there for the remaining $H - 2$ steps. Set

$$\tau = \left(\tfrac{1}{2} - \left(1 - \frac{1}{1600e^4}\right)\gamma\right)(H - 2).$$

Since the cumulative risk cost accrued before the agent reaches $s_1, \ldots, s_n$ is 0, the augmented risk-budget coordinate stays 0 at every bandit state. Thus the objectives do not depend on any additional budget coordinate at these states. For any possibly history-dependent output policy, let $\pi(a|s_i)$ denote

the conditional probability of choosing $a$ at step 2 given $s_2 = s_i$. This conditional distribution is well-defined because every $s_i$ is reached with probability $\alpha/n > 0$, independently of the preceding action and history. The performance and risk objectives therefore depend only on these conditional marginals.

We index the instances by assignments. For an assignment $(j_1, \dots, j_n) \in \{0, 1, \dots, m\}^n$, let $\mathcal{M}_{j_1,\dots,j_n}$ denote the composite instance in which state $s_i$ realizes the bandit instance $\mathcal{I}^i_{j_i}$ for each $i \in [n]$. Fix an arbitrary assignment and work on the corresponding composite instance. Set $a^{*,i} = a_{j_i}$ as the designated action at $s_i$, and let $B = \{i \in [n] : j_i \neq 0\}$ collect the states whose designated action is non-fallback. Thus $a^{*,i} = a_0$ for $i \notin B$, while for $i \in B$, $a^{*,i}$ is the modified non-fallback action at $s_i$. Because the assignment is arbitrary, the per-instance bound derived below applies to every composite instance in the family.

Fix a $(\delta, \varepsilon)$-correct algorithm $\mathbb{A}$ for the CVaR-constrained RL problem. Let $\pi^{\text{out}}$ denote its output policy and $K$ the number of episodes it uses. Since the fixed composite instance belongs to the problem class, $\mathbb{A}$ must satisfy the PAC guarantee on it. If $\mathbb{A}$ fails to stop almost surely, or if $\mathbb{E}[K] = \infty$ on any composite instance in this family, the claimed lower bound holds trivially. We therefore assume throughout that $K$ is an almost surely finite stopping time with $\mathbb{E}[K] < \infty$ on every composite instance.

The first step translates PAC correctness into a high-mass requirement for the designated actions.

For the CVaR constraint, the risk cost is nonzero only if the trajectory enters $s_{\text{rcost}}$. Let $p_{\pi,i}$ denote the probability of avoiding $s_{\text{rcost}}$ after reaching $s_i$ under policy $\pi$. By construction,

$$\begin{aligned} p_{\pi,i} &= \sum_{a \in \mathcal{A}} \pi(a|s_i)(1 - P(s_{\text{rcost}} \mid s_i, a)) \\ &= \tfrac{1}{2} \sum_{a \in \mathcal{A}} \pi(a|s_i) + \gamma\, \pi(a_0|s_i) + \gamma\, \mathbb{1}\{i \in B\} \pi\big(a^{*,i}|s_i\big) \\ &= \tfrac{1}{2} + \gamma\big(\pi(a_0|s_i) + \mathbb{1}\{i \in B\} \pi\big(a^{*,i}|s_i\big)\big). \end{aligned}$$

Every action reaches $s_{\text{rcost}}$ with probability at most $\frac{1}{2}$ after a bandit state is reached. Hence $\Pr(C^\pi > 0) \leq \alpha/2 \leq \alpha$. Since $C^\pi = (H-2)\mathbb{1}\{\text{the trajectory enters } s_{\text{rcost}}\}$, the variational formula in Eq. (1) is minimized at 0. Therefore $\text{CVaR}_\alpha(C^\pi) = \mathbb{E}[C^\pi]/\alpha$. Combining this identity with the probability $\alpha/n$ of reaching each $s_i$ gives

$$\text{CVaR}_\alpha(C^\pi) = \frac{H-2}{n} \sum_{i=1}^{n} (1 - p_{\pi,i}). \tag{28}$$

Define the local avoidance deficit

$$d_i(\pi) := \tfrac{1}{2} + \gamma - p_{\pi,i} = \gamma\big(1 - \pi\big(a^{*,i}|s_i\big)\big) - \gamma\, \mathbb{1}\{i \in B\} \pi(a_0|s_i).$$

The statewise upper bound on $p_{\pi,i}$ gives $d_i(\pi) \geq 0$.

The policy that chooses $a^{*,i}$ at every $s_i$ avoids $s_{\text{rcost}}$ with probability $\frac{1}{2} + \gamma$ at every bandit state. Hence Eq. (28) gives its CVaR value as $(\frac{1}{2} - \gamma)(H-2)$. Since every policy in this instance satisfies $\Pr(C^\pi > 0) \leq \alpha/2$, the VaR value 0 minimizes the Rockafellar–Uryasev formula for every policy. In particular, the aligned grid point satisfies $\nu_{j^*} = 0$. At this grid point, the Rockafellar–Uryasev constraint value of the witness policy is

$$\big(\tfrac{1}{2} - \gamma\big)(H-2) - \tau = -\frac{\gamma(H-2)}{1600e^4}.$$

Thus every instance in the family satisfies Slater's condition with margin $\frac{\gamma(H-2)}{1600e^4}$.

On the $\varepsilon$-feasibility event $\text{CVaR}_\alpha(C^{\pi^{\text{out}}}) - \tau \leq \varepsilon$, Eq. (28) gives

$$\frac{1}{n} \sum_{i=1}^{n} d_i(\pi^{\text{out}}) \leq \frac{\gamma}{1600e^4} + \frac{\varepsilon}{H-2}. \tag{29}$$

For the performance objective, cost is incurred only in the bandit branch after entering $s_{\text{pcost}}$. Only the fallback action $a_0$ can lead there. Hence

$$f(\pi) = \frac{\alpha(H-2)}{n} \sum_{i=1}^{n} \big(\tfrac{1}{2} + \gamma\big)\pi(a_0|s_i). \tag{30}$$

To use $\varepsilon$-optimality, we identify $f^*$ under the fixed threshold $\tau$. Set $x_i = \pi(a_0 \mid s_i)$ for $i \notin B$. In a changed state, choosing $a^{*,i}$ with probability one has zero performance cost and zero local deficit. Thus, an optimal feasible policy needs no fallback mass there. We may therefore impose this choice when computing the optimum. In a reference state, $d_i(\pi) = \gamma(1 - x_i)$, and exact feasibility is equivalent to

$$0 \le x_i \le 1, \qquad \sum_{i \notin B} (1 - x_i) \le \frac{n}{1600e^4}.$$

After removing the common positive factor $\alpha(H-2)(\frac{1}{2} + \gamma)/n$ from Eq. (30), the remaining optimization is the linear program

$$\min_{(x_i)_{i \notin B}} \sum_{i \notin B} x_i \quad \text{s.t.} \quad \sum_{i \notin B} (1 - x_i) \le \frac{n}{1600e^4}, \quad 0 \le x_i \le 1.$$

Its optimum is $\max\{|\{i \notin B\}| - n/(1600e^4), 0\}$: each unit decrease in fallback mass consumes one unit of the deficit budget, and all units have the same objective coefficient. Consequently,

$$f^* = \frac{\alpha(H-2)}{n}\big(\tfrac{1}{2} + \gamma\big) \max\Big\{|\{i \notin B\}| - \frac{n}{1600e^4}, 0\Big\}.$$

Equivalently, if $|\{i \notin B\}|/n \ge \frac{1}{1600e^4}$, the budget removes $\frac{n}{1600e^4}$ units of fallback mass from the reference states. Otherwise, all reference fallback mass can be removed. Then Eq. (30) and the expression for $f^*$ imply that, on the $\varepsilon$-optimality event $f(\pi^{\text{out}}) - f^* \le \varepsilon$,

$$\frac{1}{n}\sum_{i \in B} \pi^{\text{out}}(a_0|s_i) \le \frac{\varepsilon}{\alpha(H-2)(\frac{1}{2}+\gamma)} + \frac{1}{n}\sum_{i \notin B}\big(1 - \pi^{\text{out}}(a_0|s_i)\big) - \min\Big\{\frac{1}{1600e^4}, \frac{|\{i \notin B\}|}{n}\Big\}. \tag{31}$$

On the intersection of these two PAC events, the feasibility and optimality controls combine as follows. Dividing Eq. (29) by $\gamma$ and using the expression for $d_i$ gives

$$\frac{1}{n}\sum_{i=1}^{n}\big(1 - \pi^{\text{out}}\big(a^{*,i}|s_i\big)\big) - \frac{1}{n}\sum_{i \in B}\pi^{\text{out}}(a_0|s_i) \le \frac{1}{1600e^4} + \frac{\varepsilon}{\gamma(H-2)}.$$

Also, since $d_i(\pi^{\text{out}}) \ge 0$, Eq. (29) restricted to $i \notin B$ yields

$$\frac{1}{n}\sum_{i \notin B}\big(1 - \pi^{\text{out}}(a_0|s_i)\big) \le \min\Big\{\frac{|\{i \notin B\}|}{n}, \frac{1}{1600e^4} + \frac{\varepsilon}{\gamma(H-2)}\Big\}.$$

Since

$$\min\Big\{\frac{|\{i \notin B\}|}{n}, \frac{1}{1600e^4} + \frac{\varepsilon}{\gamma(H-2)}\Big\} - \min\Big\{\frac{1}{1600e^4}, \frac{|\{i \notin B\}|}{n}\Big\} \le \frac{\varepsilon}{\gamma(H-2)},$$

Eq. (31) implies

$$\frac{1}{n}\sum_{i \in B}\pi^{\text{out}}(a_0|s_i) \le \frac{\varepsilon}{\alpha(H-2)(\frac{1}{2}+\gamma)} + \frac{\varepsilon}{\gamma(H-2)}.$$

Bounding the fallback mass over $i \in B$ by this inequality gives

$$\frac{1}{n}\sum_{i=1}^{n}\pi^{\text{out}}\big(a^{*,i}|s_i\big) \ge 1 - \frac{1}{1600e^4} - \frac{2\varepsilon}{\gamma(H-2)} - \frac{\varepsilon}{\alpha(H-2)(\frac{1}{2}+\gamma)}. \tag{32}$$

Thus the CVaR slack rules out cheap higher-risk actions through Eq. (29), while the optimality slack rules out the costly fallback in changed states through Eq. (31). Together they force high average

mass on the designated actions $a^{*,i}$, which is the condition needed to invoke the auxiliary target regions.

Fix $\gamma := \frac{1600e^4\varepsilon}{H-2}$. Choose the universal accuracy range in Theorem 7.1 such that $\varepsilon < \frac{1}{100\times 1600\, e^4}$. Together with $H \geq 3$ and $\alpha \geq \frac{1}{100}$, this implies $\gamma < \frac{1}{100} < \frac{1}{4}$ and $\gamma/\alpha < 1$. Since $\varepsilon/(H-2) = \gamma/(1600e^4)$ and $\frac{1}{2} + \gamma \geq \frac{1}{2}$, we obtain

$$\frac{1}{1600e^4} + \frac{2\varepsilon}{\gamma(H-2)} + \frac{\varepsilon}{\alpha(H-2)(\frac{1}{2}+\gamma)} \leq \frac{1}{1600e^4} + \frac{2}{1600e^4} + \frac{2}{1600e^4} = \frac{1}{320e^4}.$$

The Slater margin in the construction is therefore $\frac{\gamma(H-2)}{1600e^4} = \varepsilon$. Thus Eq. (32) gives $\frac{1}{n}\sum_{i=1}^{n} \pi^{\text{out}}\big(a^{*,i}|s_i\big) \geq 1 - \frac{1}{320e^4}$.

The second step lower-bounds the samples needed for these designated-action events. We first connect the target-mass requirement to the target regions $R_{a_\ell} = \{\pi : \pi(a_\ell) > \frac{3}{4}\}$ of the auxiliary separation problem. Let $G_i := \{\pi^{\text{out}}(\cdot|s_i) \in R_{a^{*,i}}\} = \{\pi^{\text{out}}(a^{*,i}|s_i) > \frac{3}{4}\}$ be the event that $\pi^{\text{out}}$ places more than $\frac{3}{4}$ mass on the designated action at $s_i$. On the complement $G_i^c$, $\pi^{\text{out}}(a^{*,i}|s_i) \leq \frac{3}{4}$, which gives $1 - \pi^{\text{out}}(a^{*,i}|s_i) \geq \frac{1}{4}$ and the pointwise bridge inequality

$$\mathbb{1}_{G_i^c} \leq 4\big(1 - \pi^{\text{out}}(a^{*,i}|s_i)\big). \tag{33}$$

By the $(\delta, \varepsilon)$-correctness of $\mathbb{A}$ on the composite instance under consideration, the guarantee Eq. (3) holds with probability at least $1-\delta$, and on this event, combining Eq. (32) with the bridge Eq. (33) averaged over $i \in [n]$ gives

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{G_i^c} \leq \frac{4}{n}\sum_{i=1}^{n}\big(1 - \pi^{\text{out}}(a^{*,i}|s_i)\big) \leq \frac{4}{320e^4} = \frac{1}{80e^4}.$$

Outside this event, the left-hand side is at most 1. The fixed assignment was arbitrary. Taking expectations under $\Pr_{\mathcal{M}_{j_1,\ldots,j_n}}$ and writing $c_0 := \frac{1}{80e^4} + \delta$ gives

$$\frac{1}{n}\sum_{i=1}^{n} \Pr_{\mathcal{M}_{j_1,\ldots,j_n}}[G_i^c] \leq c_0, \qquad \forall (j_1,\ldots,j_n) \in \{0,1,\ldots,m\}^n. \tag{34}$$

Choose the universal failure-probability range in Theorem 7.1 such that $\delta < \frac{1}{330e^4}$. We then have $c_0 \leq \frac{1}{80e^4} + \frac{1}{330e^4} < \frac{1}{32}$.

Eq. (34) bounds the average failure probability on every fixed instance, but the states carrying the failure mass may vary from one instance to another. We next introduce a prior that converts this fixed-instance average bound into guarantees after all coordinates except one are frozen. This averaging step plays the same role as the per-bandit failure-budget argument in Dann & Brunskill (2015). The random assignment then lets us freeze all other coordinates, which ensures that each local change-of-measure comparison differs only at one state-action pair.

To set up the prior, draw a random assignment $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_n)$ with independent coordinates, where each $\mathcal{J}_i$ equals 0 with probability $\frac{1}{2}$ and equals each $j \in [m]$ with probability $\frac{1}{2m}$. The reference index 0 thus receives half the prior mass, and the $m$ changed indices split the rest. This choice matches the reference-versus-changed-instance comparison in Lemma G.1.

Let $\bar{\Pr}$ denote the joint distribution of $\mathcal{J}$ and of a run of $\mathbb{A}$ on the composite instance selected by $\mathcal{J}$, with expectation $\bar{\mathbb{E}}$. Conditioned on $\{\mathcal{J} = (j_1,\ldots,j_n)\}$, the conditional distribution of the run coincides with $\Pr_{\mathcal{M}_{j_1,\ldots,j_n}}$.

For $i \in [n]$ and $\ell \in \{0,1,\ldots,m\}$, define the conditional failure probability

$$\bar{\delta}_{i,\ell} := \bar{\Pr}\big[\pi^{\text{out}}(\cdot|s_i) \notin R_{a_\ell} \mid \mathcal{J}_i = \ell\big].$$

Under $\bar{\Pr}$, the failure event at state $s_i$ is $G_i^c = \{\pi^{\text{out}}(\cdot|s_i) \notin R_{a_{\mathcal{J}_i}}\}$. Averaging Eq. (34) over $\mathcal{J}$ and decomposing the probability of each $G_i^c$ over the value of $\mathcal{J}_i$ gives

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\bar{\delta}_{i,0} + \frac{1}{2m}\sum_{j=1}^{m}\bar{\delta}_{i,j}\right) \leq c_0, \;\text{ hence }\; \sum_{i=1}^{n}\bar{\delta}_{i,0} \leq 2nc_0 \;\text{ and }\; \sum_{i=1}^{n}\sum_{j=1}^{m}\bar{\delta}_{i,j} \leq 2mnc_0. \tag{35}$$

Each of the two sums follows by keeping only its own bracketed term and dropping the other, which is nonnegative.

Set $\vartheta \coloneqq 8c_0$, which makes the term $\frac{2c_0}{\vartheta}$ appearing below equal to the clean constant $\frac{1}{4}$. Since $c_0 < \frac{1}{32}$, we have $\vartheta < \frac{1}{4}$, which keeps $\log \frac{1}{4\vartheta} > 0$ and hence the resulting per-state bound nontrivial.

To apply Lemma G.1 at a state $s_i$, we need a fixed bandit instance there. We therefore freeze the assignment of the other states and compare $\{\mathcal{J}_i = 0\}$ against $\{\mathcal{J}_i = j\}$. Fix $i \in [n]$ and $j \in [m]$, and write $\mathcal{J}_{-i}$ for the coordinates of $\mathcal{J}$ other than $\mathcal{J}_i$. Given $\mathcal{J}_{-i}$ and $\{\mathcal{J}_i = \ell\}$, the run of $\mathbb{A}$ takes place on one fixed composite instance. Hence

$$\delta_{i,\ell}(\mathcal{J}_{-i}) \coloneqq \bar{\Pr}\big[\pi^{\text{out}}(\cdot|s_i) \notin R_{a_\ell} \mid \mathcal{J}_i = \ell, \mathcal{J}_{-i}\big]$$

is the failure probability at $s_i$ for one fixed assignment of the other states. Since $\mathcal{J}_i$ and $\mathcal{J}_{-i}$ are independent, averaging over $\mathcal{J}_{-i}$ recovers $\mathbb{E}_{\mathcal{J}_{-i}}[\delta_{i,\ell}(\mathcal{J}_{-i})] = \bar{\delta}_{i,\ell}$.

At $(s_i, a_j)$, the probability of avoiding $s_{\text{rcost}}$ is $\frac{1}{2}$ when $\mathcal{J}_i = 0$ and $\frac{1}{2} + \gamma$ when $\mathcal{J}_i = j$. Given the same interaction history and selected state–action pair, all other observation distributions agree between the two instances. The output policy at $s_i$ is $\pi^{\text{out}}(\cdot \mid s_i)$. Let $T_j^i$ denote the number of times action $a_j$ is played at $s_i$.

Apply Lemma H.2 to the full interaction history, treating each state–action pair as an arm. Only $(s_i, a_j)$ contributes to the KL divergence. For every fixed $\mathcal{J}_{-i}$ satisfying $\delta_{i,0}(\mathcal{J}_{-i}) \le \frac{1}{4}$ and $\delta_{i,j}(\mathcal{J}_{-i}) \le \vartheta$, the proof of Lemma G.1 therefore gives

$$\bar{\mathbb{E}}\big[T_j^i \mid \mathcal{J}_i = 0, \mathcal{J}_{-i}\big] \ge \frac{3}{16\gamma^2} \log\!\left(\frac{1}{4\vartheta}\right).$$

It remains to average over the frozen coordinates. The two conditional failure probabilities satisfy $\mathbb{E}_{\mathcal{J}_{-i}}[\delta_{i,0}(\mathcal{J}_{-i})] = \bar{\delta}_{i,0}$ and $\mathbb{E}_{\mathcal{J}_{-i}}[\delta_{i,j}(\mathcal{J}_{-i})] = \bar{\delta}_{i,j}$. Markov's inequality therefore gives

$$\Pr_{\mathcal{J}_{-i}}\left[\delta_{i,0}(\mathcal{J}_{-i}) > \frac{1}{4}\right] \le 4\bar{\delta}_{i,0}, \qquad \Pr_{\mathcal{J}_{-i}}[\delta_{i,j}(\mathcal{J}_{-i}) > \vartheta] \le \frac{\bar{\delta}_{i,j}}{\vartheta}.$$

Hence the two conditions needed for the auxiliary lemma hold simultaneously with probability at least $1 - 4\bar{\delta}_{i,0} - \frac{\bar{\delta}_{i,j}}{\vartheta}$ over the draw of $\mathcal{J}_{-i}$. Since $T_j^i \ge 0$, averaging the previous conditional lower bound over $\mathcal{J}_{-i}$ and dropping the realizations where either condition fails gives

$$\begin{aligned}\bar{\mathbb{E}}\big[T_j^i \mid \mathcal{J}_i = 0\big] &= \mathbb{E}_{\mathcal{J}_{-i}}\big[\bar{\mathbb{E}}\big[T_j^i \mid \mathcal{J}_i = 0, \mathcal{J}_{-i}\big]\big] \\ &\ge \frac{3}{16\gamma^2} \log\!\left(\frac{1}{4\vartheta}\right)\left(1 - 4\bar{\delta}_{i,0} - \frac{\bar{\delta}_{i,j}}{\vartheta}\right).\end{aligned}$$

If the right-hand side is negative, the bound holds trivially. Summing over $i \in [n]$ and $j \in [m]$ makes the constants explicit

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} \bar{\mathbb{E}}\big[T_j^i \mid \mathcal{J}_i = 0\big] &\ge \frac{3}{16\gamma^2} \log\!\left(\frac{1}{4\vartheta}\right)\left(nm - 4m\sum_{i=1}^{n}\bar{\delta}_{i,0} - \frac{1}{\vartheta}\sum_{i=1}^{n}\sum_{j=1}^{m}\bar{\delta}_{i,j}\right) \\ &\ge \frac{3nm}{16\gamma^2} \log\!\left(\frac{1}{4\vartheta}\right)\left(1 - 8c_0 - \frac{2c_0}{\vartheta}\right) = \frac{3nm}{16\gamma^2} \log\!\left(\frac{1}{32c_0}\right)\left(\frac{3}{4} - 8c_0\right) \\ &> \frac{3nm}{32\gamma^2} \log\!\left(\frac{1}{32c_0}\right),\end{aligned}$$

where the second inequality applies Eq. (35), the equality uses $\vartheta = 8c_0$, and the last step uses $\frac{3}{4} - 8c_0 > \frac{1}{2}$, which holds since $c_0 < \frac{1}{32}$.

The last step converts play counts into episodes. Let $N_i$ denote the number of times algorithm $\mathbb{A}$ visits $s_i$, which gives $N_i \geq \sum_{j=1}^{m} T_j^i$. Since $\bar{\Pr}[\mathcal{J}_i = 0] = \frac{1}{2}$ and $N_i \geq 0$,

$$\begin{aligned}\bar{\mathbb{E}}\left[\sum_{i=1}^{n} N_i\right] &\geq \sum_{i=1}^{n} \bar{\Pr}[\mathcal{J}_i = 0]\, \bar{\mathbb{E}}[N_i \mid \mathcal{J}_i = 0] \\ &\geq \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \bar{\mathbb{E}}\big[T_j^i \mid \mathcal{J}_i = 0\big] \\ &> \frac{3nm}{64\gamma^2}\log\left(\frac{1}{32c_0}\right).\end{aligned}$$

To relate total visits to the episode count $K$, we track when the agent enters the bandit branch. For each episode $k$, let $E^k$ be the event that the agent reaches a bandit state in $\{s_1, \ldots, s_n\}$ during episode $k$, and let $\mathcal{F}_{k-1}^{\mathrm{LB}}$ denote the history before episode $k$ in the lower-bound experiment. From $s_0$ the $n$ destinations $s_1, \ldots, s_n$ each carry probability $\frac{\alpha}{n}$ and sum to $\alpha$. This transition is shared by all composite instances and is independent of the past. Therefore $\bar{\Pr}[E^k \mid \mathcal{F}_{k-1}^{\mathrm{LB}}] = \alpha$. Since each episode visits at most one bandit state, the visit counts satisfy $\sum_{i=1}^{n} N_i = \sum_{k=1}^{K} \mathbb{1}_{E^k}$.

Extend the probability space with unused episode randomness after the algorithm stops. Then $E^k$ is defined for every $k$, $\{K \geq k\} \in \mathcal{F}_{k-1}^{\mathrm{LB}}$, and $\bar{\Pr}(E^k \mid \mathcal{F}_{k-1}^{\mathrm{LB}}) = \alpha$. Since all summands are nonnegative, Tonelli's theorem and the tower property give

$$\begin{aligned}\bar{\mathbb{E}}\left[\sum_{i=1}^{n} N_i\right] &= \sum_{k\geq 1} \bar{\mathbb{E}}\big[\mathbb{1}_{\{K\geq k\}}\mathbb{1}_{E^k}\big] \\ &= \sum_{k\geq 1} \bar{\mathbb{E}}\big[\mathbb{1}_{\{K\geq k\}}\bar{\mathbb{E}}[\mathbb{1}_{E^k} \mid \mathcal{F}_{k-1}^{\mathrm{LB}}]\big] \\ &= \alpha \sum_{k\geq 1} \bar{\Pr}(K \geq k) = \alpha\bar{\mathbb{E}}[K].\end{aligned}$$

The final equality is the tail-sum identity for the nonnegative integer-valued $K$. It is finite by the standing assumption $\bar{\mathbb{E}}[K] < \infty$. The bandit branch is reached only an $\alpha$ fraction of the time. Therefore, recovering the required state visits costs a factor of $1/\alpha$ in episodes. This is where the $1/\alpha$ in the bound originates. Combining this identity with the visit-count lower bound above and dividing by $\alpha$ lower-bounds $\bar{\mathbb{E}}[K]$. Since $\bar{\mathbb{E}}[K]$ is the average of $\mathbb{E}_{\mathcal{M}_{j_1,\ldots,j_n}}[K]$ over the random assignment $\mathcal{J}$, some assignment $(j_1^*, \ldots, j_n^*)$ attains at least this average, and its composite instance $\mathcal{M}^* := \mathcal{M}_{j_1^*,\ldots,j_n^*}$ satisfies $\mathbb{E}_{\mathcal{M}^*}[K] > \frac{3nm}{64\alpha\gamma^2}\log\left(\frac{1}{32c_0}\right)$. Over the admissible range of $\delta$, we have $c_0 < (\frac{1}{80} + \frac{1}{330})e^{-4}$. Therefore, $\log\frac{1}{32c_0}$ and $\log\frac{1}{c_0}$ differ only by fixed positive constant factors. Recalling $n = S - 5$, $m = A - 1$, and $\gamma = \frac{1600e^4\varepsilon}{H-2}$, and using $c_0 = \frac{1}{80e^4} + \delta$, this is

$$\mathbb{E}_{\mathcal{M}^*}[K] = \Omega\left(\frac{SAH^2}{\alpha\varepsilon^2}\log\left(\frac{1}{\frac{1}{80e^4} + \delta}\right)\right).$$

The logarithmic factor is bounded below by a positive universal constant over the stated range of $\delta$. This proves the rate in Theorem 7.1. □

## H Technical Lemmas and Supporting Inequalities

This appendix collects the general probabilistic and algebraic tools invoked by the preceding proofs. Keeping them together separates reusable inequalities from the model-specific argument while preserving explicit references at every use.

**Lemma H.1** (Time-Uniform Bounded-Count Lower Bound)**.** *Let $H \geq 1$, let $(\mathcal{F}_j)_{j\geq 0}$ be a filtration, and let $X_j \in [0, H]$ be $\mathcal{F}_j$-measurable with conditional mean $q_j = \mathbb{E}[X_j \mid \mathcal{F}_{j-1}]$. For every $\delta_0 \in (0, 1)$, with probability at least $1 - \delta_0$, simultaneously for all $n \geq 0$,*

$$\sum_{j=1}^{n} X_j \geq \frac{1}{2}\sum_{j=1}^{n} q_j - H\log\frac{1}{\delta_0}.$$

*Proof.* Set $W_j = X_j/H$ and $P_j = q_j/H$. The chord inequality for $w \mapsto e^{-w}$ on $[0,1]$ gives

$$\mathbb{E}[e^{-W_j} \mid \mathcal{F}_{j-1}] \le 1 - P_j(1 - e^{-1}) \le e^{-P_j/2}.$$

The second inequality follows from $1 - e^{-1} \ge 1/2$ and $1 - x \le e^{-x}$. Define $M_n := \exp\{\sum_{j=1}^{n}(P_j/2 - W_j)\}$. This process is $\mathcal{F}_n$-measurable and integrable, satisfies $M_0 = 1$, and obeys

$$\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}e^{P_n/2}\mathbb{E}[e^{-W_n} \mid \mathcal{F}_{n-1}] \le M_{n-1}.$$

Ville's inequality for nonnegative supermartingales (Howard et al., 2020, Lemma 1) therefore gives $\Pr\{\sup_{n\ge 0} M_n \ge 1/\delta_0\} \le \delta_0$. On the complementary event, taking logarithms and multiplying by $H$ yields the stated simultaneous lower bound. □

**Lemma H.2** (Divergence Decomposition, Eq. (6) and Remark 2 in Garivier et al. (2019))**.** *Consider a bandit problem with actions $a_0, \dots, a_m$ and two instances $\omega = (\omega_0, \dots, \omega_m)$ and $\omega' = (\omega'_0, \dots, \omega'_m)$, where $\omega_i$ and $\omega'_i$ denote the outcome distributions of action $a_i$ under the two instances. An algorithm sequentially plays actions, possibly using internal randomization whose distribution does not depend on the instance, and stops at an almost surely finite stopping time. Let $T_i$ denote the number of plays of action $a_i$ upon stopping. Then, for every random variable $\chi \in [0,1]$ that is a measurable function of the history of plays, observed outcomes, and the internal randomization,*

$$\sum_{i=0}^{m} \mathbb{E}_\omega[T_i] \cdot D_{\mathrm{KL}}(\omega_i, \omega'_i) \ge \mathrm{kl}(\mathbb{E}_\omega[\chi], \mathbb{E}_{\omega'}[\chi]).$$

**Lemma H.3** (Counts to Pseudo-Counts, adapted from Lemma 8 in Ménard et al. (2021))**.** *On event $\mathcal{E}_4$, if $\bar{n}^{k-1}(s,a) > 4H\log(SA/\delta')$, then*

$$\frac{1}{N^k(s,a) \vee 1} \le \frac{4}{\bar{n}^{k-1}(s,a) \vee 1}, \qquad \frac{1}{\sqrt{N^k(s,a) \vee 1}} \le \frac{2}{\sqrt{\bar{n}^{k-1}(s,a) \vee 1}}. \tag{36}$$

*For every $K \ge 2$, the total visitation mass before this threshold is crossed satisfies*

$$\sum_{\substack{1\le k<K:\\ \bar{n}^{k-1}(s,a)\le 4H\log(SA/\delta')}} q^k(s,a) \le H\big(4\log\tfrac{SA}{\delta'} + 1\big). \tag{37}$$

*Proof.* This proof follows the analysis of Lemma 8 in Ménard et al. (2021), but adapts it to our aggregate-count setting where $N^k(s,a) = \sum_{j=1}^{k-1}\sum_{h=1}^{H} \mathbb{1}\{(s_h^j, a_h^j) = (s,a)\}$ and $\bar{n}^{k-1}(s,a) = \sum_{j=1}^{k-1} q^j(s,a)$ (summed over all steps $h$), rather than per-step counts. On $\mathcal{E}_4$, we have $N^k(s,a) \ge \frac{1}{2}\bar{n}^{k-1}(s,a) - H\log(SA/\delta')$. Above the stated threshold, this gives $N^k(s,a) > \frac{1}{4}\bar{n}^{k-1}(s,a) > 0$. Since $N^k(s,a)$ is integral, Eq. (36) follows. The sum in Eq. (37) consists of consecutive increments of $\bar{n}^k(s,a)$. Their cumulative mass is at most the threshold plus the last increment, which is at most $H$. □

**Lemma H.4** (Algebraic Sequence Bound, Lemma 9 in Ménard et al. (2021))**.** *For any $T \in \mathbb{N}$ and non-negative $(u_t)$ with $u_t \in [0,1]$, let $U_t = \sum_{\ell=1}^{t} u_\ell$. Then,*

$$\sum_{t=0}^{T} \frac{u_{t+1}}{U_t \vee 1} \le 4\log(U_{T+1} + 1).$$

The following lemma is a standard concentration result in tabular reinforcement learning. It bounds the expected sum of inverse-square-root empirical visitation counts under the count condition below, where empirical counts track expected counts up to a logarithmic offset. A related stage-indexed statement appears as Lemma 36 in Efroni et al. (2020). We give the aggregated-count proof with explicit constants below. Related count arguments also appear in Dann et al. (2017) and Zanette & Brunskill (2019).

**Lemma H.5** (Expected Inverse-Visitation Bound)**.** *Let $\mathcal{F}_{k-1}$ be the filtration containing all history strictly prior to episode $k$, and suppose $\log(SA/\delta') \geq 1$. Assume that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $k \in [K]$, the empirical visitation counts satisfy the following lower bound condition*

$$N^k(s,a) \geq \frac{1}{2}\sum_{j=1}^{k-1} q^j(s,a) - H\log\frac{SA}{\delta'},$$

*where $q_h^j(s,a)$ is the stage-$h$ conditional visitation probability and $q^j(s,a) = \sum_{h=1}^H q_h^j(s,a)$ is the expected visitation count of $(s,a)$ under policy $\pi^j$ and true dynamics $P$. Then,*

$$\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}\left(\frac{1}{\sqrt{N^k(s_h^k,a_h^k)\vee 1}}\,\middle|\,\mathcal{F}_{k-1}\right) \leq 2\sqrt{5}\sqrt{SAHK} + 4SAH\log\frac{SA}{\delta'} + SAH = \widetilde{\mathcal{O}}\Big(\sqrt{SAHK} + SAH\Big).$$

*Proof.* By the lower bound condition on the empirical counts, we have that for any episode $k$,

$$N^k(s,a)\vee 1 \geq \max\left\{\frac{1}{2}\sum_{j=1}^{k-1} q^j(s,a) - H\log\frac{SA}{\delta'}, 1\right\}.$$

Let $\bar{n}^{k-1}(s,a) = \sum_{j=1}^{k-1} q^j(s,a)$ denote the cumulative expected visitations. Whenever $\bar{n}^{k-1}(s,a) > 4H\log(SA/\delta')$, the count condition gives $N^k(s,a) \geq \frac{1}{4}\bar{n}^{k-1}(s,a)$.

We split the sum over episodes according to whether $\bar{n}^{k-1}(s,a)$ is at most $4H\log(SA/\delta')$. Using the conditional expectation $\mathbb{E}[\mathbb{1}\{(s_h^k,a_h^k)=(s,a)\} \mid \mathcal{F}_{k-1}] = q_h^k(s,a)$ and taking the sum over $h$, we can explicitly bound the expected sums for a fixed state-action pair. Since $q^k(s,a) \leq H$, the total expected visitations accumulated before the threshold is crossed cannot exceed $4H\log(SA/\delta') + H$.

$$\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}\left[\frac{\mathbb{1}\{(s_h^k,a_h^k)=(s,a)\}}{\sqrt{N^k(s,a)\vee 1}}\,\middle|\,\mathcal{F}_{k-1}\right] \\
&\leq \sum_{k:\bar{n}^{k-1}\leq 4H\log(SA/\delta')} q^k(s,a) + \sum_{k:\bar{n}^{k-1}>4H\log(SA/\delta')} \frac{q^k(s,a)}{\sqrt{\frac{1}{4}\bar{n}^{k-1}(s,a)}} \\
&\leq 4H\log\frac{SA}{\delta'} + H + 2\sum_{k:\bar{n}^{k-1}>4H\log(SA/\delta')} \frac{q^k(s,a)}{\sqrt{\bar{n}^{k-1}(s,a)}}.
\end{aligned}$$

In the late phase, $\bar{n}^{k-1}(s,a) > 4H\log(SA/\delta') \geq 4H \geq 4q^k(s,a)$ by the premise $\log(SA/\delta') \geq 1$ and the bound $q^k(s,a) \leq H$. This implies $\bar{n}^k(s,a) = \bar{n}^{k-1}(s,a) + q^k(s,a) \leq \frac{5}{4}\bar{n}^{k-1}(s,a)$. To extend the late-index sum to all episodes, we set $q^k(s,a)/\sqrt{\bar{n}^k(s,a)} = 0$ whenever $\bar{n}^k(s,a) = 0$. This is consistent because nonnegativity and $q^k(s,a) \leq \bar{n}^k(s,a)$ then imply $q^k(s,a) = 0$. We can therefore swap the denominator to the current episode index and write

$$\sum_{k:\bar{n}^{k-1}>4H\log(SA/\delta')} \frac{q^k(s,a)}{\sqrt{\bar{n}^{k-1}(s,a)}} \leq \sqrt{\frac{5}{4}}\sum_{k=1}^K \frac{q^k(s,a)}{\sqrt{\bar{n}^k(s,a)}} \leq \sqrt{5}\sqrt{\sum_{k=1}^K q^k(s,a)}.$$

For completeness, write $S_i = \sum_{j=1}^i z_j$ and use the same zero convention when $S_i = 0$. Such an index has $z_i = 0$. For every index with $S_i > 0$, since $z_i = S_i - S_{i-1}$ and $S_i \geq S_{i-1} \geq 0$,

$$\frac{z_i}{\sqrt{S_i}} = \frac{(\sqrt{S_i} - \sqrt{S_{i-1}})(\sqrt{S_i} + \sqrt{S_{i-1}})}{\sqrt{S_i}} \leq 2(\sqrt{S_i} - \sqrt{S_{i-1}}).$$

Summing this inequality telescopes and gives $\sum_{i=1}^{n} z_i/\sqrt{S_i} \le 2\sqrt{S_n}$, which proves the final inequality above. Summing this bound over all $S$ states and $A$ actions, and applying the Cauchy-Schwarz inequality over the state-action space, we obtain

$$\sum_{s,a}\sum_{k=1}^{K}\frac{q^k(s,a)}{\sqrt{N^k(s,a)\vee 1}} \le 2\sqrt{5}\sqrt{SA\sum_{s,a}\sum_{k=1}^{K}q^k(s,a)} + 4SAH\log\frac{SA}{\delta'} + SAH.$$

Since $\sum_{s,a}\sum_k q^k(s,a) = KH$, this explicitly gives the stated bound. □

**Lemma H.6** (Inversion of Functional Inequality, adapted from Lemma 13 in Ménard et al. (2021))**.** *Let $\mathscr{C}, \mathscr{D}, \mathscr{A}, \mathscr{B}, \mathscr{E}, \kappa > 0$ with $1 \le \mathscr{B} \le \mathscr{E}$ and $\kappa \ge e$. If $m \ge 1$ satisfies*

$$m \le \mathscr{C}\sqrt{m(\mathscr{A}\log(\kappa m) + \mathscr{B}\log(\kappa m)^2)} + \mathscr{D}\big(\mathscr{A}\log(\kappa m) + \mathscr{E}\log(\kappa m)^2\big),$$

*then, with $L = \frac{8}{3}\log\big(11\kappa^2(\mathscr{A}+\mathscr{E})(\mathscr{C}+\mathscr{D})\big)$, we have*

$$m \le \mathscr{C}^2(\mathscr{A}+\mathscr{B})\,L^2 + (\mathscr{D} + 2\mathscr{C}\sqrt{\mathscr{D}})(\mathscr{A}+\mathscr{E})\,L^2 + 1.$$

*Proof.* The proof follows the argument of Lemma 13 in Ménard et al. (2021), with the exponent in its crude inversion corrected below. Let $\ell := \log(\kappa m)$ and $Q := 11\kappa(\mathscr{A}+\mathscr{E})(\mathscr{C}+\mathscr{D})$.

Use $\log x \le x^\beta/\beta$ for $x > 0$ and $\beta > 0$. The choices $\beta = 1/4$ and $\beta = 1/8$ give $\mathscr{A}\ell + \mathscr{B}\ell^2 \le 64(\mathscr{A}+\mathscr{B})(\kappa m)^{1/4}$. Since $\kappa \ge 1$, the square-root term is at most $8\mathscr{C}\kappa\sqrt{\mathscr{A}+\mathscr{B}}\,m^{5/8}$.

The choices $\beta = 5/8$ and $\beta = 5/16$ similarly give $\mathscr{A}\ell + \mathscr{E}\ell^2 \le 11(\mathscr{A}+\mathscr{E})(\kappa m)^{5/8}$. Hence the second term is at most $11\mathscr{D}\kappa(\mathscr{A}+\mathscr{E})m^{5/8}$.

Because $1 \le \mathscr{B} \le \mathscr{E}$, we have $\sqrt{\mathscr{A}+\mathscr{B}} \le \mathscr{A}+\mathscr{E}$. The premise therefore implies $m \le Qm^{5/8}$. Dividing by $m^{5/8}$ gives $m^{3/8} \le Q$, and hence $m \le Q^{8/3}$. In particular, $Q \ge 1$. Since $\kappa \ge e$, we obtain $\ell \le \log\kappa + \frac{8}{3}\log Q \le \frac{8}{3}\log(\kappa Q) = L$, and $L \ge 1$. Consequently,

$$\mathscr{A}\ell + \mathscr{B}\ell^2 \le (\mathscr{A}+\mathscr{B})L^2, \qquad \mathscr{A}\ell + \mathscr{E}\ell^2 \le (\mathscr{A}+\mathscr{E})L^2.$$

Set $x := \sqrt{m}$, $a := \mathscr{C}\sqrt{\mathscr{A}+\mathscr{B}}L$, and $b := \mathscr{D}(\mathscr{A}+\mathscr{E})L^2$. The original inequality becomes $x^2 \le ax + b$. Since $a, b \ge 0$, the positive-root formula gives $x \le (a + \sqrt{a^2+4b})/2 \le a + \sqrt{b}$. Squaring and using $\sqrt{(\mathscr{A}+\mathscr{B})(\mathscr{A}+\mathscr{E})} \le \mathscr{A}+\mathscr{E}$ yields

$$m \le \mathscr{C}^2(\mathscr{A}+\mathscr{B})L^2 + (\mathscr{D} + 2\mathscr{C}\sqrt{\mathscr{D}})(\mathscr{A}+\mathscr{E})L^2.$$

□