# INSTANCE-DEPENDENT REGRET FOR CMDPS WITH STEP-WISE CONSTRAINTS

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

## ABSTRACT

We study online learning in episodic tabular constrained Markov decision processes with *step-wise safety constraints*. In such a setting, the constraints induce a safe subgraph that shapes the variance of cumulative rewards under feasible policies and, consequently, the difficulty of learning. Exploiting this structure, however, requires learning which actions are safe while controlling constraint violations. We propose *Safe Variance-Adaptive Exploration* (SVAE), an efficient algorithm that learns candidate safe subgraphs and performs variance-adaptive optimistic planning within them. With high probability, SVAE achieves cumulative regret of order $\widetilde{\mathcal{O}}(\sqrt{SAH\min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}} + S\sqrt{AH^3\min\{K,\mathcal{C}\}} + S^2AH^2)$ over $K$ episodes, where $H$ is the horizon of a single episode, while $S$ and $A$ are the numbers of states and actions, respectively. Here, $\mathrm{Var}^\star$ is the maximum return variance among safe policies, $\mathbb{V}_\Sigma$ is the variance accumulated before the first unsafe action is encountered, and $\mathcal{C}$ captures the statistical complexity of eliminating actions incorrectly considered potentially safe. SVAE additionally attains $\widetilde{\mathcal{O}}(H\sqrt{SAK}+S^2AH^2)$ step-wise constraint violation and a gap-dependent violation bound that is polylogarithmic in $K$. Finally, we establish a lower bound showing that dependence on these instance-specific quantities is unavoidable.

## 1 INTRODUCTION

Constrained Markov decision processes (CMDPs) are widely used to model sequential decision making in robotics, autonomous driving, and wireless networks (Achiam et al., 2017; Liu et al., 2021; Bura et al., 2022). The objective is to maximize expected cumulative reward subject to safety or resource constraints, commonly imposed on expected cumulative costs (Altman, 1999; Miryoosefi & Jin, 2022; Stradi et al., 2025a). The potential harm from individual unsafe decisions motivates step-wise safety constraints, which impose a mean-cost threshold at each decision (Xiong et al., 2023; Shi et al., 2023). Violation is measured by summing positive excesses above this threshold without cancellation (Wei et al., 2024). In this work, we study online learning in episodic tabular constrained MDPs under this kind of *step-wise* constraint.

Instance-dependent regret bounds distinguish the statistical difficulty of individual problems—namely, instances—and provide sharper guarantees on easier ones (Zanette & Brunskill, 2019; Dann et al., 2021). In unconstrained reinforcement learning, variance-dependent analyses relate regret to the uncertainty induced by rewards and transitions (Talebi & Maillard, 2018; Zhou et al., 2023b; Zhao et al., 2023; Chen et al., 2025). Such bounds can retain worst-case guarantees while improving on low-variance instances, including nearly constant regret in deterministic environments (Zhou et al., 2023b; Zamir et al., 2026). In CMDPs, unknown costs and transitions determine which policies are feasible (Bura et al., 2022; Yu et al., 2025). Exploratory policies can be unsafe, so their returns need not share the low variance of safe policies. Whether the benefits of variance-dependent learning can be retained under step-wise constraints is still unclear.

Step-wise safety constraints can be represented by a safe subgraph that retains actions satisfying the cost threshold and leading only to states with feasible future decisions (Xiong et al., 2023; Shi et al., 2023). Our key observation is that even though learning under unknown constraints generally poses additional challenges, safety constraints can still simplify reward learning by removing stochastic branches or reducing differences in expected future rewards across possible next states under feasible policies. Even when the rewards and the transition kernel are fixed, changing costs can alter the safe



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.

subgraph and thereby reshape variances. However, small cost gaps and rare transitions to states with no feasible policy can still make safety identification difficult. Understanding how the safe subgraph relates to the variance and the statistical complexity is therefore central to instance-dependent regret analysis. These observations lead to the following questions:

*(i) Is it possible to design a computationally efficient algorithm that achieves variance-dependent regret bounds and sublinear cumulative step-wise violation?*
*(ii) How does the induced safe subgraph shape the statistical complexity of learning under step-wise constraints?*

To answer these questions, we first characterize how the induced safe subgraph shapes the variances and the safety identification complexity. We then propose Safe Variance-Adaptive Exploration (SVAE), which separates safe-subgraph learning from variance-adaptive planning. The algorithm learns candidate safe actions using variance-independent cost lower confidence bounds and observed successors, and plans optimistically over these sets using empirical reward and transition variances. We further evaluate SVAE and examine the effect of safe-subgraph properties on regret in synthetic experiments (Appendix H).

Our contributions can be summarized as follows.

- **Safe-subgraph structure and variance** In Section 3, we represent step-wise safety constraints through their induced safe subgraph and define two distinct variance quantities: $(i)$ the maximum safe-policy return variance $\mathrm{Var}^\star$ is the largest variance of the total episode return over all safe policies, and $(ii)$ the trajectory-based total variance $\mathbb{V}_\Sigma$ sums reward variances and transition variances before the first unsafe action in each episode. We establish a tight bound on transition variance in terms of stochastic branching depth and differences in expected cumulative rewards across successor states within the safe subgraph.
- **Safety identification complexity** In Section 3.3, we introduce a reward-independent instance quantity $\mathcal{C}$ to measure the difficulty of ruling out unsafe actions at states reachable under safe policies. Lower bounds for valid elimination establish its necessity.
- **Efficient variance-adaptive exploration** In Section 4, we develop SVAE, an efficient algorithm that exploits low return variance within the unknown safe subgraph without prior knowledge of a safe policy. SVAE maintains candidate action sets that retain all safe actions with high probability and uses the resulting candidate subgraph to adapt its exploration bonuses to empirical reward and next-state value variances.
- **Instance-dependent learning guarantees** In Section 5.1, we show that our algorithm attains, for a CMDP with $S$ states, $A$ actions, horizon $H$, a cumulative regret bound over $K$ episodes of order $\widetilde{\mathcal{O}}(\sqrt{SAH\min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}} + S\sqrt{AH^3\min\{K,\mathcal{C}\}} + S^2AH^2)$. Here $\widetilde{\mathcal{O}}$ hides polylogarithmic factors. Regret is polylogarithmic in $K$ on fixed instances with $\mathrm{Var}^\star = 0$, while both regret and cumulative step-wise violation retain $\widetilde{\mathcal{O}}(\sqrt{K})$ worst-case dependence. Step-wise violation is also polylogarithmic in $K$ for a fixed positive cost gap (Theorem 3).
- **Regret lower bound** In Section 5.2, we establish a regret lower bound, showing that both return variance and safety identification dependencies are unavoidable.

### 1.1 Related Work

**Safe reinforcement learning** Safe RL is a fundamental area of reinforcement learning, with diverse constraint formulations and measures of constraint violation. For episode-level constraints, regret and constraint-violation guarantees have been established in tabular settings (Efroni et al., 2020; Liu et al., 2021; Bura et al., 2022; Wei et al., 2022; Stradi et al., 2024; 2025b) and with linear function approximation (Ding et al., 2021; Ghosh et al., 2022; Kitamura et al., 2025). For instantaneous constraints, Amani et al. (2021) and Roknilamouki et al. (2025) ensure constraint satisfaction with high probability using known safe actions or policies. Without a known safe policy, Xiong et al. (2023) and Wei et al. (2024) bound cumulative positive step-wise violation. Xiong et al. (2023) study tabular MDPs with costs that depend only on the state. Wei et al. (2024) consider linear MDPs and assume that a feasible action exists at every state and step. We study tabular MDPs with unknown state–action costs and require only the existence of a safe policy from the initial state.

**Variance-dependent regret** Variance-dependent regret bounds can improve on worst-case rates in low-variance environments. Such guarantees have been established for episodic tabular MDPs (Zhou et al., 2023b), linear-mixture MDPs (Zhao et al., 2023), and infinite-horizon MDPs (Zamir et al., 2026), with recent work also combining variance and suboptimality gaps (Chen et al., 2025). In particular, Zhou et al. (2023b) derive bounds involving a maximum policy-return variance measure and a variance sum along the learner's trajectories. For CMDPs, Yu et al. (2025) use total-variance analysis to improve the horizon dependence of a worst-case regret bound under episode-level constraints. Jiang & Ye (2024) obtain a different form of instance dependence through an optimal-basis characterization of the CMDP linear program. We control step-wise violation and derive a regret bound that separates safe-policy return variance from the complexity of ruling out actions outside the true safe set.

**Graph structure and safe exploration** Graph structure has been used to guide safe exploration and refine regret bounds. Safe-region methods use reachability and returnability to expand the region that can be explored safely (Turchetta et al., 2016; Wachi & Sui, 2020). Shi et al. (2023) use safe subgraphs for learning under instantaneous constraints, assuming known transition supports and a known seed safe subgraph. For unconstrained MDPs, Bourel et al. (2020) combine variance-aware concentration with transition-support adaptation and obtain regret bounds involving local diameter and effective support. Our structural analysis examines how safety constraints change the graph available to safe policies and how this change affects their transition variance. We use this relation to bound the transition-variance contribution to regret in terms of stochastic branching and continuation-value differences within the safe subgraph.

Due to space constraints, additional related works are provided in Appendix A.

## 2 Problem Formulation

**Episodic CMDP** We consider an episodic constrained Markov decision process (CMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r, c, \tau)$ with a state space $\mathcal{S}$ of size $S$, an action space $\mathcal{A}$ of size $A$, and horizon $H$. For a positive integer $m$, let $[m] := \{1, \dots, m\}$. The transition kernel is $P = \{P_h : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}\}_{h=1}^{H}$, where $\Delta_{\mathcal{S}}$ denotes the set of probability distributions on $\mathcal{S}$, and $P_h(s' \mid s, a)$ denotes the transition probability from $(s, a)$ to $s'$ at step $h$. $r = \{r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]\}_{h=1}^{H}$ represents the reward function, and $r_h(s, a)$ is the reward of taking action $a$ in state $s$ at step $h$. A non-stationary policy $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^{H} \in \Pi$ specifies an action $\pi_h(s)$ at every state and step, and $\Pi$ denotes the set of all such policies. The learner interacts with the CMDP for $K$ episodes, each starting from a fixed initial state $s_1 \in \mathcal{S}$. For any episode $k$, the learner first chooses a policy $\pi^k$ based on previous observations. At step $h \in [H]$ of episode $k$, the learner observes a state $s_h^k$, chooses an action $a_h^k$, and transitions to a next state $s_{h+1}^k$ according to $P_h(\cdot \mid s_h^k, a_h^k)$. For a policy $\pi \in \Pi$, the state-value function $V_h^\pi(s)$ and the state-action value function $Q_h^\pi(s, a)$ are defined as:

$$V_h^\pi(s) := \mathbb{E}^\pi \left[ \sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s \right], \qquad Q_h^\pi(s, a) := \mathbb{E}^\pi \left[ \sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s, a_h = a \right].$$

The expectations start from $s$ at step $h$ and follow $\pi$, with the first action fixed to $a$ for $Q_h^\pi(s, a)$. We set $V_{H+1}^\pi \equiv 0$ and let $\mu_h^\pi(s) := \mathbb{P}^\pi(s_h = s \mid s_1)$ denote the step-$h$ state occupancy under $\pi$, for $h \in [H + 1]$.

Similarly, $c = \{c_h : \mathcal{S} \times \mathcal{A} \to [0, 1]\}_{h=1}^{H}$ denotes the cost function, and $c_h(s, a)$ is the cost of taking action $a$ in state $s$ at step $h$. Let $\tau \in [0, 1]$ denote the safety threshold. Under stochastic settings, the non-stationary reward and cost of state-action pair $(s, a)$ at step $h$ are $R_h(s, a)$ and $C_h(s, a)$ supported on $[0, 1]$, with means $r_h(s, a)$ and $c_h(s, a)$, respectively. Conditional on the selected state–action pairs, reward and cost observations are independent across visits, and each observation is conditionally independent of the next state. Write $R_h^k$ and $C_h^k$ for the samples observed at step $h$ of episode $k$. The threshold $\tau$ is known, whereas $P = \{P_h\}_{h=1}^{H}$, $r = \{r_h\}_{h=1}^{H}$, and $c = \{c_h\}_{h=1}^{H}$ are unknown.

**Safe subgraph** The CMDP induces a layered directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \mathcal{S} \times [H + 1]$ (Jin et al., 2020). For $h \in [H]$, the edge $((s, h), a, (s', h + 1))$ belongs to $\mathcal{E}$ exactly when $P_h(s' \mid s, a) > 0$. The safe subgraph $\mathcal{G}_{\text{safe}} = (\mathcal{V}_{\text{safe}}, \mathcal{E}_{\text{safe}})$ is defined through safe actions and

safe vertices by backward recursion. Set $\mathcal{V}_{\mathrm{safe},H+1} := \mathcal{S}$. For $h = H, \dots, 1$, define recursively

$$\mathcal{A}_{\mathrm{safe},h}(s) := \big\{a \in \mathcal{A} : c_h(s,a) \le \tau,\ \mathrm{supp}(P_h(\cdot \mid s,a)) \subseteq \mathcal{V}_{\mathrm{safe},h+1}\big\}, \tag{1}$$
$$\mathcal{V}_{\mathrm{safe},h} := \{s \in \mathcal{S} : \mathcal{A}_{\mathrm{safe},h}(s) \neq \varnothing\}.$$

An action $a$ is safe at $(s,h)$ if $c_h(s,a) \le \tau$ and all its successor vertices are safe. A vertex $(s,h)$ is safe when $s \in \mathcal{V}_{\mathrm{safe},h}$. The vertex and edge sets of $\mathcal{G}_{\mathrm{safe}}$ are $\mathcal{V}_{\mathrm{safe}} := \big\{(s,h) \in \mathcal{V} : s \in \mathcal{V}_{\mathrm{safe},h}\big\}$ and $\mathcal{E}_{\mathrm{safe}} := \big\{((s,h),a,(s',h+1)) \in \mathcal{E} : a \in \mathcal{A}_{\mathrm{safe},h}(s)\big\}$.

For a policy $\pi \in \Pi$, let $\mathcal{G}^\pi = (\mathcal{V}^\pi, \mathcal{E}^\pi)$ denote the subgraph generated by $\pi$ from $s_1$, containing the vertices and edges visited with positive probability (Shi et al., 2023). Its vertex and edge sets are $\mathcal{V}^\pi := \{(s,h) \in \mathcal{V} : \mu_h^\pi(s) > 0\}$ and $\mathcal{E}^\pi := \{((s,h),a,(s',h+1)) \in \mathcal{E} : (s,h) \in \mathcal{V}^\pi,\ a = \pi_h(s)\}$, respectively. Thus, $\mathcal{G}^\pi$ includes all positive-probability successors of the action selected at each reachable vertex. We write $\mathcal{G}^\pi \subseteq \mathcal{G}_{\mathrm{safe}}$ when $\mathcal{V}^\pi \subseteq \mathcal{V}_{\mathrm{safe}}$ and $\mathcal{E}^\pi \subseteq \mathcal{E}_{\mathrm{safe}}$. Therefore, the safe policy class is $\Pi_{\mathrm{safe}} := \{\pi \in \Pi : \mathcal{G}^\pi \subseteq \mathcal{G}_{\mathrm{safe}}\}$. Let $\pi^\star \in \arg\max_{\pi \in \Pi_{\mathrm{safe}}} V_1^\pi(s_1)$ be an optimal safe policy, and write $V_h^\star(s) = V_h^{\pi^\star}(s)$.

**Performance metrics** For the policies $\pi^1, \dots, \pi^K$ executed by the learner, we define the cumulative regret and cumulative step-wise violation as:

$$\mathcal{R}(K) := \sum_{k=1}^{K} \big(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\big), \tag{2}$$

$$\mathrm{Viol}(K) := \sum_{k=1}^{K} \sum_{h=1}^{H} \big[c_h(s_h^k, a_h^k) - \tau\big]_+, \tag{3}$$

where $[\cdot]_+ := \max\{\cdot, 0\}$. The regret $\mathcal{R}(K)$ compares the learner's cumulative expected return with that of the optimal safe policy, while $\mathrm{Viol}(K)$ accumulates positive mean-cost excesses along the learner's trajectories. Xiong et al. (2023) study step-wise violation with state-only costs $c(s)$, whereas our costs $c_h(s,a)$ depend on both the state and action and may vary across steps. Unlike constraints on expected cumulative costs, Equation (3) does not allow excess costs at one step to be offset by lower costs at another. We allow violations during exploration and control their cumulative magnitude, rather than require zero violation throughout learning as in Shi et al. (2023). Our goal is to obtain high-probability sublinear upper bounds in $K$ on regret and violation, with regret adapting to the variance of safe policies and the difficulty of identifying actions outside the safe action sets.

Our results generally require only that the initial state admit an unknown safe policy, which is a mild condition in the literature (Amani et al., 2021; Xiong et al., 2023; Shi et al., 2023; Wei et al., 2024; Roknilamouki et al., 2025).

**Assumption 1** (Feasibility)**.** *The initial state satisfies* $s_1 \in \mathcal{V}_{\mathrm{safe},1}$.

## 3 Safety-Induced Variance and Identification Complexity

Safety constraints shape the statistical difficulty of reinforcement learning, not only the set of feasible policies. They can impede exploration, but by restricting behavior they can also remove sources of reward uncertainty and simplify learning inside the feasible region. Under step-wise safety, these two effects are organized by the hidden safe subgraph. We show that its stochastic branching and successor-value differences control the transition variance available to reward learning, while actions outside the graph must be excluded based on cost or successor-viability evidence. This yields tight structural variance bounds together with a complexity for the separate task of safety identification.

### 3.1 Variance under Step-Wise Safety

We first introduce two variance quantities for reward uncertainty under step-wise safety.

**Maximum safe-policy return variance** We measure the return variability available to safe policies by the *maximum safe-policy return variance*. Write $\mathbb{V}(X)$ for the variance of $X$ and $\mathbb{V}(q,f) :=$

$\sum_x q(x)f(x)^2 - (\sum_x q(x)f(x))^2$ for the variance of $f$ under a distribution $q$. We define:

$$\mathrm{Var}^\star := \sup_{\pi \in \Pi_{\mathrm{safe}}} \mathbb{E}^\pi \left[ \sum_{h=1}^{H} \left\{ \mathbb{V}(R_h(s_h, a_h)) + \mathbb{V}\big(P_h(\cdot \mid s_h, a_h), V_{h+1}^\pi\big) \right\} \,\middle|\, s_1 \right]. \tag{4}$$

Equation (4) captures two sources of variance. The reward term $\mathbb{V}(R_h(s_h, a_h))$ measures noise in the reward observation, while $\mathbb{V}(P_h(\cdot \mid s_h, a_h), V_{h+1}^\pi)$ measures the transition-induced variance in continuation value. Their expected sum equals the variance of the total return under $\pi$, so $\mathrm{Var}^\star$ is the largest total-return variance over safe policies. It is a fixed property of the CMDP instance.

**Trajectory-based total variance** The preceding quantity is uniform over the safe policy class, but it can be conservative for a particular interaction sequence. A complementary view therefore uses the conditional variances actually encountered during interaction. For episode $k$, let $\tau_k$ be the first step at which the selected action lies outside the true safe set, with $\tau_k = H + 1$ if no such action is selected. We define the *trajectory-based total variance* as:

$$\mathbb{V}_\Sigma := \sum_{k=1}^{K} \sum_{h<\tau_k} \left\{ \mathbb{V}(R_h(s_h^k, a_h^k)) + \mathbb{V}\big(P_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^\star\big) \right\}. \tag{5}$$

This quantity depends on the learner's realized trajectories and is therefore algorithm dependent. It captures the local variance at state-action pairs visited before the first unsafe action is selected and measures transition uncertainty using the optimal safe value function. This truncation separates the interaction within the safe subgraph from the additional difficulty of identifying actions outside it.

**Remark 1.** *The two quantities capture different aspects of reward-learning uncertainty and neither generally subsumes the other. Similar policy-level and trajectory-level variance characterizations underlie variance-adaptive guarantees in unconstrained RL (Azar et al., 2017; Zanette & Brunskill, 2019; Zhou et al., 2023b; Zhao et al., 2023), and the corresponding quantities can be incomparable (Zhou et al., 2023b). The additional challenge under step-wise safety is that the region supporting these variance quantities is itself unknown, which leads to the separate safety identification complexity studied in Section 3.3.*

### 3.2 Structure of the Safe Subgraph and Transition Variance

Both variance measures introduced above are closely tied to the safe subgraph. We next study how the structure of the safe subgraph affects these quantities. In fact, it acts most directly on transition variance, through how often stochastic transitions can occur along a safe path and how large the value differences across their successors can be. We capture these two effects by the branching depth of the safe subgraph and successor-value spans.

Let $\gamma = (s_1, a_1, \ldots, s_H, a_H, s_{H+1})$ range over directed paths in $\mathcal{G}_{\mathrm{safe}}$ from $(s_1, 1)$ to step $H + 1$. The *branching depth* is the largest number of stochastic transitions on one such path, i.e., $B(\mathcal{G}_{\mathrm{safe}}) := \max_\gamma \sum_{h=1}^{H} \mathbf{1}\{|\mathrm{supp}(P_h(\cdot \mid s_h, a_h))| > 1\}$. For a nonempty finite set $X$ and a real-valued function $f$ on $X$, write $\mathrm{span}_X(f) := \max_{x \in X} f(x) - \min_{x \in X} f(x)$. For a policy $\pi \in \Pi_{\mathrm{safe}}$, the *successor-value span* $S^\pi$ measures the largest difference in continuation values among possible successors of an action selected by $\pi$ at any reachable vertex. Formally, $S^\pi := \max_{(s,h) \in \mathcal{V}^\pi,\, h \in [H]} \mathrm{span}_{\mathrm{supp}(P_h(\cdot \mid s, \pi_h(s)))}(V_{h+1}^\pi)$. Over the same safe paths, define the *optimal successor-value span* by $S^\star := \max_{\gamma,\, h \in [H]} \mathrm{span}_{\mathrm{supp}(P_h(\cdot \mid s_h, a_h))}(V_{h+1}^\star)$. This quantity measures the largest span of optimal safe continuation values across successors of actions on safe paths. The following result shows how these spans and the branching depth control transition variance.

**Theorem 1** (Safe-subgraph variance bound)**.** *Under Assumption 1, every $\pi \in \Pi_{\mathrm{safe}}$ satisfies:*

$$\mathbb{E}^\pi \left[ \sum_{h=1}^{H} \mathbb{V}\big(P_h(\cdot \mid s_h, a_h), V_{h+1}^\pi\big) \,\middle|\, s_1 \right] \le \frac{1}{4} (S^\pi)^2 B(\mathcal{G}_{\mathrm{safe}}). \tag{6}$$

*Moreover, every realized interaction sequence satisfies $\sum_{k=1}^{K} \sum_{h=1}^{\tau_k - 1} \mathbb{V}\big(P_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^\star\big) \le \frac{K}{4} (S^\star)^2 B(\mathcal{G}_{\mathrm{safe}})$.*

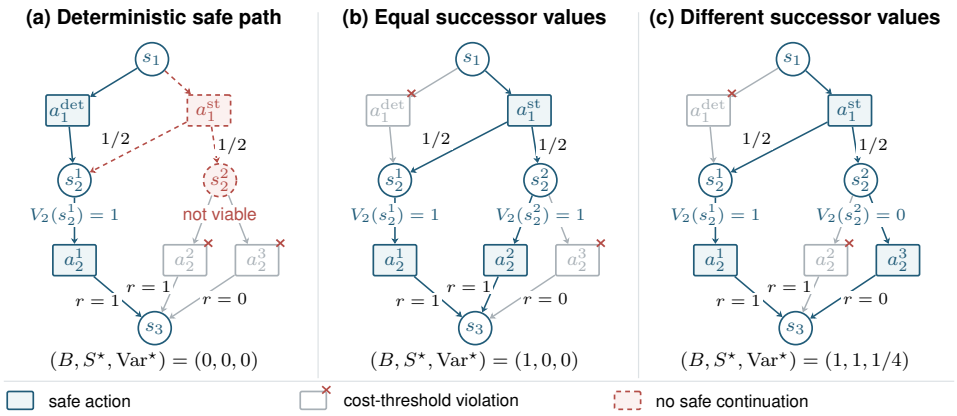



Figure 1: **The structure of the safe subgraph reshapes transition variance.** All panels share the same reward function and transition kernel and differ only in their costs. Changing only the costs removes the stochastic branch in panel (a) and retains it in panels (b)–(c). Its successors have equal continuation values and zero transition variance in panel (b), but unequal values and positive transition variance in panel (c). With $\tau = 1/2$, crossed actions have cost $1 > \tau$, while all other actions have cost 0. Blue solid elements denote the safe subgraph, gray solid elements denote actions excluded by the cost threshold and their transitions, and red dashed elements denote exclusion because a successor has no safe continuation.

The theorem bounds each safe policy's transition variance using $S^\pi$ and the trajectory-based transition variance using $S^\star$. Small branching depth limits how often transition randomness can affect continuation value, while a small successor-value span limits the contribution of each stochastic transition. The full proof is given in Appendix F.1.

Figure 1 shows how changing only the costs can change the safe subgraph, thereby removing or retaining stochastic branches and changing the continuation-value differences across their successors. The following proposition shows that the bound in Equation (6) is tight. In Appendix F.3, we provide the complete analysis and proof.

**Proposition 1.** *For every $n \geq 1$, there exists a family of CMDPs with $H = 2n$ that share the same transition kernel and deterministic rewards and differ only in the cost function $c$. For every $b \in \{0, \ldots, n\}$, the family contains an instance such that, for every $\pi \in \Pi_{\mathrm{safe}}$, $B(\mathcal{G}_{\mathrm{safe}}) = b$, $S^\pi = \mathbf{1}\{b > 0\}$, $\mathbb{E}^\pi\left[\sum_{h=1}^H \mathbb{V}\big(P_h(\cdot \mid s_h, a_h), V_{h+1}^\pi\big) \,\middle|\, s_1\right] = \frac{b}{4}$. Hence, Equation (6) holds with equality and $\mathrm{Var}^\star = b/4$.*

### 3.3 Safety Identification Complexity

The previous results reveal a potential benefit of safety: removing random branches or successor-value variation can reduce reward uncertainty. However, even a low-variance safe subgraph can still be difficult to identify. Intuitively, a rare non-viable branch can contribute little transition variance and yet require many executions to discover. This creates a distinct safety identification problem.

For an unsafe action, we first define a local complexity that captures the number of executions needed to identify it as unsafe. The action can be excluded either directly, when its mean cost exceeds the threshold, or recursively, through a successor state at which every action is unsafe. Formally, for every step $h \in [H]$, state $s \in \mathcal{S}$, and action $a \notin \mathcal{A}_{\mathrm{safe},h}(s)$, we define:

$$\kappa_h(s,a) := \begin{cases} \{c_h(s,a) - \tau\}^{-2}, & c_h(s,a) > \tau, \\ \min\limits_{\substack{s' \in \mathrm{supp}(P_h(\cdot|s,a)) \\ s' \notin \mathcal{V}_{\mathrm{safe},h+1}}} \dfrac{\kappa_{h+1}(s')}{P_h(s' \mid s,a)}, & c_h(s,a) \leq \tau, \end{cases} \qquad \kappa_h(s) := \sum_{a \notin \mathcal{A}_{\mathrm{safe},h}(s)} \kappa_h(s,a). \tag{7}$$

This definition mirrors the recursive structure of safety. A positive cost gap yields an inverse-squared term. Otherwise, one non-viable successor suffices to exclude the parent action, giving the minimum

**Algorithm 1** Safe Variance-Adaptive Exploration (SVAE)

**Input:** $\mathcal{S}$, $\mathcal{A}$, $H$, $K$, $s_1$, $\tau$, and $\delta$

Set $\iota \leftarrow \log(32S^2AHK/\delta)$ and initialize all empirical statistics
**for** $k = 1, 2, \ldots, K$ **do**
  Form $n_h^k$, $\widehat{r}_h^k$, $\widehat{c}_h^k$, $\widehat{P}_h^k$, $\widehat{\sigma}_{R,h}^{2,k}$, and $\widehat{\mathrm{supp}}_h^k$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$
  Set $\underline{c}_h^k$ by (9) for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$
  $s_1^k \leftarrow s_1$, $\hat{\mathcal{V}}_{\mathrm{safe},H+1}^k \leftarrow \mathcal{S}$, and $V_{H+1}^k(\cdot) \leftarrow 0$
  **for** $h = H, H-1, \ldots, 1$ **do**
    $\hat{\mathcal{A}}_{\mathrm{safe},h}^k(s) \leftarrow \{a \in \mathcal{A} : \underline{c}_h^k(s,a) \le \tau,\ \widehat{\mathrm{supp}}_h^k(s,a) \subseteq \hat{\mathcal{V}}_{\mathrm{safe},h+1}^k\}$ for all $s \in \mathcal{S}$
    $\hat{\mathcal{V}}_{\mathrm{safe},h}^k \leftarrow \{s \in \mathcal{S} : \hat{\mathcal{A}}_{\mathrm{safe},h}^k(s) \neq \varnothing\}$ ▷ *Safe-subgraph learning*
    Compute $b_h^k(s,a)$ by (10) for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ ▷ *Variance-adaptive bonus*
    $Q_h^k(s,a) \leftarrow \min\{H-h+1, \widehat{r}_h^k(s,a) + \widehat{P}_h^k(\cdot \mid s,a)^\top V_{h+1}^k + b_h^k(s,a)\}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
    **for** $s \in \mathcal{S}$ **do**
      $\pi_h^k(s) \in \begin{cases} \arg\max_{a \in \hat{\mathcal{A}}_{\mathrm{safe},h}^k(s)} Q_h^k(s,a), & \hat{\mathcal{A}}_{\mathrm{safe},h}^k(s) \neq \varnothing, \\ \arg\max_{a \in \mathcal{A}} Q_h^k(s,a), & \text{otherwise}, \end{cases}$
      $V_h^k(s) \leftarrow Q_h^k(s, \pi_h^k(s))$
    **end for**
  **end for**
  **for** $h = 1, 2, \ldots, H$ **do** ▷ *Policy execution*
    Play $a_h^k \leftarrow \pi_h^k(s_h^k)$, observe $(R_h^k, C_h^k, s_{h+1}^k)$, and update the empirical statistics
  **end for**
**end for**

over such successors, while reaching that successor incurs the factor $1/P_h(s' \mid s,a)$. At each state $s$, $\kappa_h(s)$ sums complexities of its unsafe actions.

**Instance-dependent identification complexity** The first unsafe action can only be selected after a safe path. Hence its vertex is reachable under some safe policy. This motivates the *safety identification complexity*, which we define as follows:

$$\mathcal{C}(\mathcal{M}) := \sum_{(s,h) \in \bigcup_{\pi \in \Pi_{\mathrm{safe}}} \mathcal{V}^\pi} \kappa_h(s), \tag{8}$$

where we set $\kappa_{H+1}(s) = 0$ for any state $s$. We write $\mathcal{C}$ for $\mathcal{C}(\mathcal{M})$ when the instance is clear from context. Thus $\mathcal{C}$ quantifies the difficulty of ruling out unsafe actions at the reachable boundary of the safe subgraph. Unlike the variance quantities, it is a reward-independent instance quantity determined by $(P, c, \tau, s_1)$.

The safety identification complexity also captures an unavoidable statistical cost of eliminating unsafe actions. Even on a simple tree-structured family, any $\delta$-valid elimination rule requires $\Omega(\kappa_h(s,a)\log(1/\delta))$ expected executions to eliminate an unsafe root action, while for pairwise-disjoint trees these requirements add to $\Omega(\mathcal{C}\log(1/\delta))$ expected executions. Thus, the cost-gap and rare-successor effects in $\kappa$, together with their accumulation in $\mathcal{C}$, capture genuine safety identification difficulty. The construction and proof are given in Appendix G.2.

## 4 Safe Variance-Adaptive Exploration

In this section, we provide our main algorithm, *Safe Variance-Adaptive Exploration* (SVAE). Section 3 shows that the safe subgraph can reduce reward uncertainty, while identifying it remains difficult even when the return variance is small. On the reward side, empirical variance can sharpen optimistic exploration, whereas on the cost side, empirical variance alone cannot reliably rule out unsafe actions. An unsafe action can exhibit zero empirical cost variance until a rare high-cost outcome is observed, as shown in Appendix E. SVAE therefore adopts an asymmetric design, combining variance-independent cost lower confidence bounds for safe-subgraph learning with variance-adaptive bonuses for reward exploration. We present SVAE in Algorithm 1.

**Safe-subgraph learning** SVAE maintains an optimistic outer approximation of the unknown safe subgraph. Fix a confidence parameter $\delta \in (0,1)$ and set $\iota := \log(32S^2AHK/\delta)$. Let $N_h^k(s,a)$

and $N_h^k(s,a,s')$ denote, respectively, the numbers of visits to $(s,a)$ and transitions from $(s,a)$ to $s'$ at step $h$ before episode $k$. Write $n_h^k(s,a) := \max\{N_h^k(s,a),1\}$. Let $\widehat{r}_h^k$, $\widehat{c}_h^k$, and $\widehat{P}_h^k$ denote the corresponding empirical estimates of $r_h$, $c_h$, and $P_h$, respectively, with detailed definitions given in Appendix B.1. The observed successor set is $\widehat{\mathrm{supp}}_h^k(s,a) := \{s' \in \mathcal{S} : N_h^k(s,a,s') > 0\}$.

We then compute the optimistic cost estimates for step $h \in [H]$, state $s \in \mathcal{S}$, and action $a \in \mathcal{A}$ as:

$$\underline{c}_h^k(s,a) := \widehat{c}_h^k(s,a) - \sqrt{2\iota / n_h^k(s,a)}. \tag{9}$$

SVAE constructs the candidate action sets $\hat{\mathcal{A}}_{\mathrm{safe},h}^k(s)$ and state sets $\hat{\mathcal{V}}_{\mathrm{safe},h}^k$ by applying the backward recursion in Equation (1), with optimistic cost estimates in place of true costs and observed successors in place of the full successor sets. The following lemma shows that this recursion retains all safe actions and states with high probability. Its proof is given in Appendix B.4.

**Lemma 1** (Optimism of the candidate sets)**.** *With probability at least $1-\delta/8$, it holds $\mathcal{A}_{\mathrm{safe},h}(s) \subseteq \hat{\mathcal{A}}_{\mathrm{safe},h}^k(s)$ and $\mathcal{V}_{\mathrm{safe},h} \subseteq \hat{\mathcal{V}}_{\mathrm{safe},h}^k$ for every episode $k \in [K]$, step $h \in [H]$, and state $s \in \mathcal{S}$.*

**Variance-adaptive planning** SVAE then performs optimistic value iteration on the estimated safe subgraph using variance-adaptive bonuses. For each step $h$ and state-action pair $(s,a)$, define the empirical reward variance $\widehat{\sigma}_{R,h}^{2,k}(s,a) := \frac{1}{n_h^k(s,a)} \sum_{\ell<k} (R_h^\ell - \widehat{r}_h^k(s,a))^2 \mathbf{1}\{(s_h^\ell, a_h^\ell) = (s,a)\}$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. For any function $V : \mathcal{S} \to \mathbb{R}$, write $\widehat{P}_h^k V(s,a) := \sum_{s' \in \mathcal{S}} \widehat{P}_h^k(s' \mid s,a) V(s')$. Let $Q_h^k$ and $V_h^k$ denote the state-action value and state-value estimates computed in Algorithm 1. SVAE uses the following empirical-Bernstein exploration bonus:

$$b_h^k(s,a) = 4\sqrt{\frac{\mathbb{V}(\widehat{P}_h^k(\cdot \mid s,a), V_{h+1}^k)\,\iota}{n_h^k(s,a)}} + 2\sqrt{\frac{\widehat{\sigma}_{R,h}^{2,k}(s,a)\,\iota}{n_h^k(s,a)}} + \frac{\{16(H-h)+5\}\iota}{n_h^k(s,a)}. \tag{10}$$

The first two terms account for transition and reward uncertainty through the empirical variances of estimated next-state values and reward observations, respectively. The last term provides a finite-sample correction even when the empirical variances vanish.

Using these bonuses, SVAE computes the state-action value estimates $Q_h^k$ by optimistic Bellman updates in the same backward pass that constructs the estimated safe subgraph. At each state $s$, it chooses $\pi_h^k(s)$ to maximize $Q_h^k(s,\cdot)$ over $\hat{\mathcal{A}}_{\mathrm{safe},h}^k(s)$ when this set is nonempty, and over $\mathcal{A}$ otherwise, with $V_h^k(s) = Q_h^k(s, \pi_h^k(s))$. SVAE then executes the resulting policy $\pi^k$ for one episode and uses the observed rewards, costs, and successor states to update the empirical statistics for the next episode.

**Remark 2.** *Variance-adaptive RL uses empirical next-state value variance, and in some cases reward variance, to sharpen optimism in unconstrained MDPs (Azar et al., 2017; Zanette & Brunskill, 2019; Zhou et al., 2023b). In SVAE, the continuation values entering the transition-variance term are obtained by planning over candidate safe action sets learned online. Hence, the bonus adapts to the return uncertainty under the evolving safety-induced graph structure.*

## 5 Learning Guarantees

In this section, we establish an instance-dependent regret bound and cumulative step-wise violation bounds for SVAE. We show how the variance and identification complexity introduced in Section 3 govern its learning guarantees. Throughout, we use $\Omega$ and $\Theta$ to hide universal constants. The notation $\widetilde{\mathcal{O}}$ additionally hides polylogarithmic factors in $S, A, H, K$, and $1/\delta$.

### 5.1 Regret and Constraint Violation

We first present the cumulative regret attained by Algorithm 1 over $K$ episodes.

**Theorem 2** (Regret bound)**.** *Suppose Assumption 1 holds. For any integer $K \geq 1$ and confidence parameter $\delta \in (0,1)$, SVAE (Algorithm 1) attains:*

$$\mathcal{R}(K) \leq \widetilde{\mathcal{O}}\Big(\sqrt{SAH \min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}} + S\sqrt{AH^3 \min\{K, \mathcal{C}\}} + S^2AH^2\Big),$$

*with probability at least $1-\delta$.*

Theorem 2 separates the variance-dependent contribution to regret from the cost of safety identification, and gives polylogarithmic regret on fixed instances with $\mathrm{Var}^\star = 0$. By taking the minimum of $\mathbb{V}_\Sigma$ and $K\mathrm{Var}^\star$, the bound refines the purely instance-dependent guarantee and can yield a smaller variance-dependent term when the learner encounters low-variance trajectories. When $\tau = 1$, our setting degenerates to unconstrained RL, and the bound recovers the variance-dependent term of MVP-V (Zhou et al., 2023b) in the time-inhomogeneous setting. Theorem 2 also yields $\widetilde{\mathcal{O}}(S\sqrt{AH^3K})$ regret in the worst case, matching the leading regret of OMDPD (Zhu et al., 2025). For uniformly bounded $\mathcal{C}$ and sufficiently large $K$, the leading regret term becomes $\widetilde{\mathcal{O}}(\sqrt{SAH^3K})$, which matches the corresponding minimax rate (Zhang et al., 2025). The proof is deferred to Appendix C.

For deterministic rewards, combining Theorem 2 with Theorem 1 and the universal bound $\mathrm{Var}^\star \le H^2/4$ yields the following graph-dependent corollary.

**Corollary 1** (Graph-dependent regret)**.** *Under the conditions of Theorem 2, if the rewards are deterministic, then with probability at least $1-\delta$, the reward regret of SVAE satisfies* $\mathcal{R}(K) \le \widetilde{\mathcal{O}}(\sqrt{SAHK\min\{H^2, (S^\star)^2 B(\mathcal{G}_{\mathrm{safe}})\}} + S\sqrt{AH^3\min\{K,\mathcal{C}\}} + S^2AH^2)$.

Corollary 1 bounds the variance-dependent regret term through the branching depth of the safe subgraph and the successor-value span of the optimal safe value function. In particular, with deterministic rewards, this term vanishes when $S^\star = 0$, even if the transitions are stochastic.

**Constraint violation** We then bound the cumulative step-wise constraint violation of SVAE. In addition to a general bound, we can obtain a gap-dependent bound under the following assumption. The gap in the assumption controls how easily costs exceeding the threshold can be identified. Similar gap-dependent assumptions have been studied in safe RL and bandits (Chen et al., 2022; Gangrade et al., 2024).

**Assumption 2** (Positive cost gap)**.** *There exists $\Delta_c > 0$ such that $c_h(s,a) - \tau \ge \Delta_c$ for every $h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$ with $c_h(s,a) > \tau$.*

We now give the cumulative step-wise constraint violation bounds of our algorithm.

**Theorem 3** (Cumulative constraint violation)**.** *Under Assumption 1, with probability at least $1-\delta$, SVAE satisfies* $\mathrm{Viol}(K) \le \widetilde{\mathcal{O}}(H\sqrt{SAK} + S^2AH^2)$. *If Assumption 2 also holds, then on the same event,* $\mathrm{Viol}(K) \le \widetilde{\mathcal{O}}(SAH/\Delta_c + S^2AH^2)$.

Theorem 3 gives a $\widetilde{\mathcal{O}}(\sqrt{K})$ cumulative constraint violation in general and a violation bound that is polylogarithmic in $K$ for a fixed positive cost gap. For state-only costs, pooling cost observations across actions and steps gives $\widetilde{\mathcal{O}}(\sqrt{SKH} + S^2AH^2)$ violation or $\widetilde{\mathcal{O}}(S/\Delta_c + S^2AH^2)$, which recovers the rates of SUCBVI (Xiong et al., 2023). If every state admits a safe action at every step, the bound becomes $\widetilde{\mathcal{O}}(H\sqrt{SAK})$, matching the bound of LSVI-AE (Wei et al., 2024) in the tabular setting. Under Assumption 2, if every unsafe action at a vertex reachable under a safe policy has cost exceeding $\tau$, we then have $\mathcal{C} \le SAH/\Delta_c^2$, as proved in Appendix C.7.

### 5.2 Regret Lower Bounds

We next establish a regret lower bound that captures the dependence on both reward variance and safety identification complexity.

**Theorem 4** (Lower bound)**.** *Let $S \ge 8$, $A \ge 2$, $H \ge 64$, and $K \ge SA$ be integers. Fix $v \in [0, H/128]$ and $\Gamma \in [16(A-1), SAH/64]$ satisfying $\min\{K,\Gamma\} \ge 16SA/H$. For any learning algorithm, there exists a CMDP $\mathcal{M}$ with $S$ states, $A$ actions, and horizon $H$ satisfying Assumption 1, such that $\mathrm{Var}^\star(\mathcal{M}) \in [3v/4, v]$, $\mathcal{C}(\mathcal{M}) = \Theta(\Gamma)$, and the algorithm's expected regret after $K$ episodes satisfies $\mathbb{E}_{\mathcal{M}}[\mathcal{R}(K)] = \Omega(\sqrt{SAHKv} + \sqrt{SAH\min\{K,\Gamma\}})$. Moreover, $\mathbb{V}_\Sigma \in [3Kv/4, Kv]$ for every realized interaction sequence on $\mathcal{M}$.*

Theorem 4 demonstrates that both maximum safe-policy return variance and safety identification complexity contribute to the difficulty of learning. In particular, zero safe-policy return variance does not eliminate the regret incurred by identifying safe actions. The proof uses two families of CMDPs with the prescribed variance and identification complexity. One requires distinguishing actions with similar expected rewards, while the other requires identifying a hidden safe action. Taking the larger of the two lower bounds yields the stated result. The full proof is provided in Appendix G.1.

## AI USE STATEMENT

In this work, we used generative AI tools to polish the writing for clarity and readability, assist with identifying relevant literature, help with experimental code and plotting, and, in a limited number of cases, help turn author-developed proof sketches into clearer lemma statements. All AI-assisted text, code, and mathematical statements were subsequently reviewed and verified by the authors, and the cited references were checked against the original sources.

## REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):2669–2678, 2018.

Eitan Altman. *Constrained Markov Decision Processes*, volume 7 of *Stochastic Modeling*. Chapman & Hall/CRC, Boca Raton, 1999.

Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 243–253. PMLR, 2021.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 2017.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, volume 30, pp. 908–918, 2017.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1056–1066. PMLR, 2020.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-François Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1047–1059, 2022.

Shulun Chen, Runlong Zhou, Zihan Zhang, Maryam Fazel, and Simon S. Du. Sharp gap-dependent variance-aware regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 124328–124364, 2025.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pp. 3123–3148. PMLR, 2022.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1–12, 2021.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R. Jovanović. Provably efficient safe exploration via primal-dual policy optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3304–3312. PMLR, 2021.

Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R. Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *Journal of Machine Learning Research*, 26(256):1–76, 2025.

Yihan Du and Wei Chen. Branching reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5494–5530. PMLR, 2022.

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.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. Iterative reachability estimation for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69764–69797, 2023.

Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. Safe linear bandits over unknown polytopes. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1755–1795. PMLR, 2024.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained RL with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13303–13315, 2022.

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

Ruiquan Huang, Jing Yang, and Yingbin Liang. Safe exploration incurs nearly no additional sample complexity for reward-free RL. In *International Conference on Learning Representations*, 2023.

Jiashuo Jiang and Yinyu Ye. Achieving instance-dependent sample complexity for constrained Markov decision process. *arXiv preprint arXiv:2402.16324*, 2024. URL https://arxiv.org/abs/2402.16324.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4860–4869. PMLR, 2020.

Toshinori Kitamura, Arnob Ghosh, Tadashi Kozuno, Wataru Kumagai, Kazumi Kasaura, Kenta Hoshino, Yohei Hosoe, and Yutaka Matsuo. Provably efficient RL under episode-wise safety in constrained MDPs with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 52632–52681, 2025.

Donghao Li, Ruiquan Huang, Cong Shen, and Jing Yang. Near-optimal conservative exploration in reinforcement learning under episode-wise constraints. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19527–19564. PMLR, 2023.

Mingyi Li, Taira Tsuchiya, and Kenji Yamanishi. Data- and variance-dependent regret bounds for online tabular MDPs. In *International Conference on Machine Learning*, 2026.

Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17183–17193, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15666–15698. PMLR, 2022.

Subhojyoti Mukherjee, Josiah P. Hanna, and Robert D. Nowak. ReVar: Strengthening policy evaluation via reduced variance sampling. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1413–1422. PMLR, 2022.

Subhojyoti Mukherjee, Josiah P. Hanna, and Robert D. Nowak. SaVeR: Optimal data collection strategy for safe policy evaluation in tabular MDP. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36531–36576. PMLR, 2024.

Mohamad Qadri, Gokul Swamy, Jonathan Francis, Michael Kaess, and Andrea Bajcsy. Your learned constraint is secretly a backward reachable tube. *Reinforcement Learning Journal*, 6:478–492, 2025.

Amirhossein Roknilamouki, Arnob Ghosh, Ming Shi, Fatemeh Nourzad, Eylem Ekici, and Ness Shroff. Provably efficient RL for linear MDPs under instantaneous safety constraints in non-convex feature spaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51957–51995. PMLR, 2025.

Ming Shi, Yingbin Liang, and Ness Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31243–31268. PMLR, 2023.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs: Handling stochastic and adversarial constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46692–46721. PMLR, 2024.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning adversarial MDPs with stochastic hard constraints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56920–56951. PMLR, 2025a.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained MDPs via policy optimization. In *International Conference on Learning Representations*, 2025b.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Policy optimization for CMDPs with bandit feedback: Learning stochastic and adversarial constraints. In *Forty-Second International Conference on Machine Learning*, 2025c.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pp. 770–805. PMLR, 2018.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite Markov decision processes with Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4312–4320, 2016.

Sharan Vaswani, Lin F. Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained MDPs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3110–3122, 2022.

Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained Markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9797–9806. PMLR, 2020.

Honghao Wei, Xin Liu, and Lei Ying. Triple-Q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3274–3307. PMLR, 2022.

Honghao Wei, Xin Liu, and Lei Ying. Safe reinforcement learning with instantaneous constraints: The role of aggressive exploration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21708–21716, 2024.

Yukuan Wei, Xudong Li, and Lin F. Yang. Near-optimal sample complexity bounds for constrained average-reward MDPs. In *International Conference on Learning Representations*, 2026.

Nuoya Xiong, Yihan Du, and Longbo Huang. Provably safe reinforcement learning with step-wise violation constraints. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54341–54353, 2023.

Kihyun Yu, Duksang Lee, William Overman, and Dabeen Lee. Improved regret bound for safe reinforcement learning via tighter cost pessimism and reward optimism. *Reinforcement Learning Journal*, 6:493–546, 2025.

Guy Zamir, Matthew Zurek, and Yudong Chen. Optimal variance-dependent regret bounds for infinite-horizon MDPs. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 7016–7061. PMLR, 2026.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7304–7312. PMLR, 2019.

Zihan Zhang, Yuxin Chen, Jason Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning. *Journal of the ACM*, 72(3):1–63, 2025.

Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4977–5020. PMLR, 2023.

Runlong Zhou, Ruosong Wang, and Simon Shaolei Du. Horizon-free and variance-dependent reinforcement learning for latent Markov decision processes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42698–42723. PMLR, 2023a.

Runlong Zhou, Zihan Zhang, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42878–42914. PMLR, 2023b.

Jiahui Zhu, Kihyun Yu, Dabeen Lee, Xin Liu, and Honghao Wei. An optimistic algorithm for online CMDPS with anytime adversarial constraints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80347–80372. PMLR, 2025.

## APPENDIX CONTENTS

**A Additional Related Work** **15**

**B Algorithm Details and Preliminaries** **16**

B.1 Empirical Estimates and Implementation . . . . . 16

B.2 Optimal Safe Policy . . . . . 17

B.3 Confidence Events . . . . . 17

B.4 Proof of Lemma 1 . . . . . 19

B.5 Optimism . . . . . 19

**C Regret Analysis** **19**

C.1 Martingale Bounds . . . . . 19

C.2 Safety Identification . . . . . 20

C.3 Bellman Residual Bounds . . . . . 22

C.4 Regret Decomposition . . . . . 23

C.5 Variance Bounds . . . . . 24

C.6 Proof of Theorem 2 . . . . . 28

C.7 A Special Case with Direct Cost Violations . . . . . 30

**D Constraint Violation** **30**

**E Cost Confidence Bounds at Zero Empirical Variance** **31**

**F Safe-Subgraph Variance** **32**

F.1 Proof of Theorem 1 . . . . . 33

F.2 Proof of Corollary 1 . . . . . 33

F.3 Proof of Proposition 1 . . . . . 33

**G Lower Bounds** **35**

G.1 Proof of Theorem 4 . . . . . 35

G.2 Lower Bounds for Action Elimination . . . . . 39

**H Experiments** **41**

H.1 Experimental Setup . . . . . 41

H.2 Stochastic Navigation . . . . . 42

H.3 Variance Adaptation . . . . . 43

H.4 Safety Identification Complexity . . . . . 44

H.5 Recursive Elimination . . . . . 45

H.6 Additional Experimental Details . . . . . 46

# A Additional Related Work

**Constrained policy optimization** Policy optimization methods approach constrained learning through different policy updates and penalty constructions. Achiam et al. (2017) use trust-region updates in Constrained Policy Optimization, while Tessler et al. (2019) use an adaptive penalty in Reward Constrained Policy Optimization to guide learning toward constraint satisfaction. For discounted CMDPs, Ding et al. (2025) analyze the convergence and sample complexity of natural policy gradient primal–dual methods with tabular and function-approximation parametrizations. In online CMDPs with adversarial losses and constraints, Stradi et al. (2025c) provide the first policy optimization primal-dual method to attain best-of-both-worlds results.

**Safe exploration and policy evaluation** Safety during data collection is distinct from feasibility of the policy returned after exploration. Miryoosefi & Jin (2022) reduce constrained RL and approachability to reward-free RL oracles without requiring safe exploration. Huang et al. (2023) instead use a known safe baseline to satisfy an expected cumulative-cost budget in every exploration episode with high probability, while Xiong et al. (2023) study reward-free exploration with cumulative step-wise violation control. For conservative exploration, Li et al. (2023) maintain an episode-wise performance threshold with high probability using a known baseline policy or one learned from suitable offline data. For policy evaluation, Mukherjee et al. (2022) use reward-variance estimates to guide sampling in tree-structured MDPs. Mukherjee et al. (2024) impose a cumulative safety constraint relative to a known default policy and, under their tractability assumptions, bound the mean squared error of SaVeR. In these evaluation problems, the goal is to reduce estimation error for a fixed target policy, rather than learn which safe policy to execute.

**Safe control** Control-theoretic and formal methods reason about safety through the future consequences of an action. Berkenkamp et al. (2017) combine learned dynamics with Lyapunov analysis to expand a region of attraction while maintaining stability. Alshiekh et al. (2018) use a shield to restrict or correct actions according to a temporal-logic specification. Reachability-based methods characterize states from which a violation can still be avoided. Ganai et al. (2023) combine reachability estimation with policy optimization, while Qadri et al. (2025) show that inverse constraint learning can recover a dynamics-dependent backward reachable tube rather than the underlying failure set. These approaches study safety through stability, specifications, or reachability, rather than through the variance-dependent regret of an online learner. They provide a related view of why local cost feasibility alone need not ensure safe continuation.

**Variance-dependent reinforcement learning** Variance has been used both to sharpen worst-case regret bounds and to obtain guarantees that adapt to a particular instance. Azar et al. (2017) use empirical next-state value variance in finite-horizon exploration bonuses, while Talebi & Maillard (2018) establish variance-aware regret bounds for undiscounted MDPs. Zhou et al. (2023a) study variance-dependent learning in latent MDPs whose contexts are revealed in hindsight. With known transitions, Li et al. (2026) obtain data-dependent regret bounds for adversarial losses and variance-dependent bounds for stochastic losses. A different trajectory model appears in branching RL, where an episode generates a tree rather than a single path and the analysis requires a corresponding law of total variance (Du & Chen, 2022). That model differs from the stochastic transitions along a single trajectory considered here. Thus, both the variance quantity and its role in a guarantee depend on the underlying MDP and observation model.

**Instance-dependent sample complexity** Sample-complexity bounds for CMDPs capture several sources of difficulty beyond return variance. With generative-model access, Vaswani et al. (2022) distinguish relaxed and strict feasibility in discounted CMDPs and quantify the role of the Slater constant in learning a near-optimal policy with no constraint violation. Wei et al. (2026) study constrained average-reward MDPs under the same sampling model, with bounds involving bias span, transient time, and a feasibility margin. Jiang & Ye (2024) instead identify nonzero basic variables and supporting constraints that determine an optimal occupancy measure of the CMDP linear program. These analyses concern feasibility margins, long-run dynamics, or the variables and constraints that characterize an optimal occupancy measure. Our safety identification complexity measures the difficulty of ruling out actions outside the true safe set using local cost observations and evidence that a successor admits no safe continuation.

# B Algorithm Details and Preliminaries

We first specify the empirical estimates and implementation of SVAE. We then fix an optimal safe policy, define the confidence events, and prove the candidate-set inclusions and optimism used in the analysis.

Fix $K \geq 1$ and $\delta \in (0,1)$. Recall that $\iota = \log(32S^2AHK/\delta)$, and set $L_K := 2+\log K$. Throughout the appendix, $\mathcal{O}(\cdot)$ hides universal multiplicative constants. We write $P_h f(s,a) := \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) f(s')$. A state $s$ is viable at step $h$ if $s \in \mathcal{V}_{\mathrm{safe},h}$.

## B.1 Empirical Estimates and Implementation

The estimates used in episode $k \in [K]$ of Algorithm 1 use only observations from episodes $\ell < k$. For $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in [H]$, define

$$N_h^k(s,a) := \sum_{\ell<k} \mathbf{1}\{(s_h^\ell, a_h^\ell) = (s,a)\},$$
$$N_h^k(s,a,s') := \sum_{\ell<k} \mathbf{1}\{(s_h^\ell, a_h^\ell, s_{h+1}^\ell) = (s,a,s')\}.$$

Write $n_h^k(s,a) := \max\{N_h^k(s,a), 1\}$. The empirical reward and cost estimates are

$$\begin{aligned} \widehat{r}_h^k(s,a) &:= \frac{1}{n_h^k(s,a)} \sum_{\ell<k} R_h^\ell \mathbf{1}\{(s_h^\ell, a_h^\ell) = (s,a)\}, \\ \widehat{c}_h^k(s,a) &:= \frac{1}{n_h^k(s,a)} \sum_{\ell<k} C_h^\ell \mathbf{1}\{(s_h^\ell, a_h^\ell) = (s,a)\}. \end{aligned} \tag{B.1}$$

For $N_h^k(s,a) \geq 1$, the empirical transition probabilities are

$$\widehat{P}_h^k(s' \mid s,a) := \frac{N_h^k(s,a,s')}{N_h^k(s,a)}. \tag{B.2}$$

At zero count, let $\widehat{P}_h^k(\cdot \mid s,a)$ be an arbitrary fixed probability distribution on $\mathcal{S}$. The empirical reward variance and observed successor set from Section 4 are

$$\widehat{\sigma}_{R,h}^{2,k}(s,a) := \frac{1}{n_h^k(s,a)} \sum_{\ell<k} \big(R_h^\ell - \widehat{r}_h^k(s,a)\big)^2 \mathbf{1}\{(s_h^\ell, a_h^\ell) = (s,a)\},$$
$$\widehat{\mathrm{supp}}_h^k(s,a) := \{s' \in \mathcal{S} : N_h^k(s,a,s') > 0\}.$$

Thus, at zero count, $\widehat{r}_h^k = \widehat{c}_h^k = \widehat{\sigma}_{R,h}^{2,k} = 0$ and the observed successor set is empty. The initialization of $\widehat{P}_h^k$ does not add elements to the observed successor set. The linear term in (10) then gives $Q_h^k(s,a) = H - h + 1$.

Before episode $k$, SVAE computes its candidate sets, bonuses, values, and policy from these estimates. Fix a deterministic rule for breaking ties in each maximization. At step $h$ of the dynamic program, $\hat{\mathcal{V}}_{\mathrm{safe},h+1}^k$ and $V_{h+1}^k$ are already available, so the candidate sets and Bellman updates can be computed in a single backward pass. The candidate sets, $\pi^k$, $Q_h^k$, and $V_h^k$ remain fixed during execution. New observations update the statistics for episode $k+1$.

The estimates can be maintained using visit counts, transition counts, and sums of rewards, costs, and squared rewards. In particular,

$$\widehat{\sigma}_{R,h}^{2,k}(s,a) = \frac{1}{n_h^k(s,a)} \sum_{\ell<k} (R_h^\ell)^2 \mathbf{1}\{(s_h^\ell, a_h^\ell) = (s,a)\} - \big(\widehat{r}_h^k(s,a)\big)^2.$$

Each observation updates these stored quantities in constant time. With dense transition arrays, checking observed successors and computing transition means and variances take $\mathcal{O}(S)$ operations for each $(s,a,h)$, while all action maximizations take $\mathcal{O}(SAH)$ operations. Each episode therefore requires

$\mathcal{O}(S^2AH)$ arithmetic operations, including its $H$ observation updates. The total is $\mathcal{O}(KS^2AH)$ operations with $\mathcal{O}(S^2AH)$ memory.

Let $\mathcal{F}_h^k$ be the $\sigma$-field generated by the history just before the reward, cost, and next state at step $h$ of episode $k$ are observed. Thus $s_h^k$ and $a_h^k$ are $\mathcal{F}_h^k$-measurable, while $\pi^k$, $Q_h^k$, and $V_h^k$ are fixed given $\mathcal{F}_1^k$. Conditional on $\mathcal{F}_h^k$, $R_h^k$ and $C_h^k$ have means $r_h(s_h^k, a_h^k)$ and $c_h(s_h^k, a_h^k)$, respectively, and the next state has distribution $P_h(\cdot \mid s_h^k, a_h^k)$. We order these histories by episode and step, and let $\mathcal{F}_{H+1}^k$ include the final observations of episode $k$.

### B.2 Optimal Safe Policy

The Bellman comparisons below use a policy that is optimal at every viable state. The next lemma supplies such a policy. We take it as $\pi^\star$ in Section 2 and retain this choice in $V^\star$, $\mathbb{V}_\Sigma$, and $S^\star$.

**Lemma B.1.** *Under Assumption 1, there exists a deterministic policy $\pi^\star \in \Pi_{\mathrm{safe}}$ that selects a safe action at every viable state. With $V_{H+1}^\star \equiv 0$, it satisfies*

$$V_h^\star(s) = \max_{a \in \mathcal{A}_{\mathrm{safe},h}(s)} \left\{ r_h(s,a) + P_h V_{h+1}^\star(s,a) \right\} \tag{B.3}$$

*for every $h \in [H]$ and $s \in \mathcal{V}_{\mathrm{safe},h}$. For every policy $\pi \in \Pi$ that selects safe actions at all viable states, $V_h^\pi(s) \le V_h^\star(s)$ for all $h \in [H]$ and $s \in \mathcal{V}_{\mathrm{safe},h}$. In particular, $\pi^\star$ maximizes $V_1^\pi(s_1)$ over $\pi \in \Pi_{\mathrm{safe}}$.*

*Proof.* Set $V_{H+1}^\star = 0$ and define $V_h^\star$ recursively by the right-hand side of (B.3). Because the maximizing set is finite and nonempty at every viable state, choose one maximizer as $\pi_h^\star(s)$, and define the policy arbitrarily at states outside $\mathcal{V}_{\mathrm{safe},h}$. Every selected safe action has all successors in $\mathcal{V}_{\mathrm{safe},h+1}$. Thus the recursion determines the value of $\pi^\star$ at every viable state, independently of its actions elsewhere. Assumption 1 therefore gives $\pi^\star \in \Pi_{\mathrm{safe}}$.

Let $\pi$ select a safe action at every viable state. If $V_{h+1}^\pi \le V_{h+1}^\star$ on $\mathcal{V}_{\mathrm{safe},h+1}$, then

$$\begin{aligned} V_h^\pi(s) &= r_h(s, \pi_h(s)) + P_h V_{h+1}^\pi(s, \pi_h(s)) \\ &\le r_h(s, \pi_h(s)) + P_h V_{h+1}^\star(s, \pi_h(s)) \le V_h^\star(s). \end{aligned}$$

Backward induction proves the comparison. For any $\pi \in \Pi_{\mathrm{safe}}$, choose an arbitrary safe action at each viable state that $\pi$ does not reach from $s_1$. This completes $\pi$ to a policy selecting safe actions at all viable states without changing its trajectory distribution or value from $s_1$. It follows that $V_1^\pi(s_1) \le V_1^\star(s_1)$ for every $\pi \in \Pi_{\mathrm{safe}}$. □

### B.3 Confidence Events

The cost and support bounds below preserve the safe action sets. The reward and transition bounds control optimism and Bellman residuals. All intersections below range over $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$, and $k \in [K]$. Define

$$\mathcal{E}_c := \bigcap_{s,a,h,k} \left\{ |\widehat{c}_h^k(s,a) - c_h(s,a)| \le \sqrt{\frac{2\iota}{n_h^k(s,a)}} \right\},$$

$$\mathcal{E}_P := \bigcap_{s,s',a,h,k} \left\{ \begin{array}{l} \mathbf{1}\{N_h^k(s,a) \ge 1\} |\widehat{P}_h^k(s' \mid s,a) - P_h(s' \mid s,a)| \\ \qquad \le \sqrt{\dfrac{2P_h(s' \mid s,a)\iota}{n_h^k(s,a)}} + \dfrac{\mathbf{1}\{P_h(s' \mid s,a) > 0\}\iota}{n_h^k(s,a)} \end{array} \right\},$$

$$\mathcal{E}_{\mathrm{supp}} := \bigcap_{s,a,h,k} \left\{ \widehat{\mathrm{supp}}_h^k(s,a) \subseteq \mathrm{supp}(P_h(\cdot \mid s,a)) \right\}.$$

For $N := N_h^k(s,a) \geq 2$, consider the three empirical Bernstein bounds

$$|\widehat{r}_h^k(s,a) - r_h(s,a)| \leq 2\sqrt{\frac{\widehat{\sigma}_{R,h}^{2,k}(s,a)\iota}{N}} + \frac{5\iota}{N}, \tag{B.4}$$

$$|(\widehat{P}_h^k - P_h)V_{h+1}^\star(s,a)| \leq 2\sqrt{\frac{\mathbb{V}(\widehat{P}_h^k(\cdot \mid s,a), V_{h+1}^\star)\iota}{N}} + \frac{5(H-h)\iota}{N}, \tag{B.5}$$

$$\widehat{\sigma}_{R,h}^{2,k}(s,a) \leq 2\mathbb{V}(R_h(s,a)) + \frac{2\iota}{N}. \tag{B.6}$$

Define $\mathcal{E}_{\mathrm{EB}}$ as the event that (B.4)–(B.6) hold for every $s, a, h, k$ with $N_h^k(s,a) \geq 2$. Set $\mathcal{E}_{\mathrm{cs}} := \mathcal{E}_c \cap \mathcal{E}_{\mathrm{supp}}$ and $\mathcal{E}_{\mathrm{est}} := \mathcal{E}_{\mathrm{cs}} \cap \mathcal{E}_P \cap \mathcal{E}_{\mathrm{EB}}$.

**Lemma B.2.** *The events satisfy* $\mathbb{P}(\mathcal{E}_{\mathrm{cs}}) \geq 1 - \delta/8$ *and* $\mathbb{P}(\mathcal{E}_{\mathrm{est}}) \geq 1 - 9\delta/16$.

*Proof.* For each $(s,a,h)$, generate an infinite sequence of independent copies of its reward, cost, and next-state observations. Revealing the next unused copy at each visit gives the same interaction distribution as the observation model. We apply concentration bounds to the first $n$ copies for every deterministic $n \in [K]$, then take a union bound over these counts before substituting the realized count $N_h^k(s,a)$.

For the cost and transition estimates, fix $(s,a,h)$ and $s' \in \mathcal{S}$. Write $C_i$ for the $i$th cost sample and let $Z_i$ indicate that the $i$th next state is $s'$. With $p = P_h(s' \mid s,a)$, Hoeffding's and Bernstein's inequalities give, for each $n \in [K]$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n C_i - c_h(s,a)\right| > \sqrt{\frac{2\iota}{n}}\right) \leq 2e^{-\iota},$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - p\right| > \sqrt{\frac{2p\iota}{n}} + \frac{2\iota}{3n}\right) \leq 2e^{-\iota}.$$

If $p = 0$, all $Z_i$ are zero almost surely. At zero count, the cost bound holds because $c_h(s,a) \leq 1 \leq \sqrt{2\iota}$. Every observed successor belongs to the true support almost surely, so $\mathbb{P}(\mathcal{E}_{\mathrm{supp}}) = 1$.

For the reward and value estimates, consider independent, identically distributed $X_1, \ldots, X_n \in [0,B]$. Let $\overline{X}_n$ and $\widehat{v}_n = n^{-1}\sum_i (X_i - \overline{X}_n)^2$ be the sample mean and variance. Applying Maurer & Pontil (2009, Theorem 4) to $X_i/B$ and $1 - X_i/B$ gives, for $B > 0$ and $n \geq 2$,

$$|\overline{X}_n - \mathbb{E}[X_1]| \leq \sqrt{\frac{2\widehat{v}_n\iota}{n-1}} + \frac{7B\iota}{3(n-1)} \leq 2\sqrt{\frac{\widehat{v}_n\iota}{n}} + \frac{5B\iota}{n}$$

with failure probability at most $4e^{-\iota}$. The case $B = 0$ is immediate. Apply this bound to the reward samples with $B = 1$ and to $V_{h+1}^\star$ evaluated at the sampled next states with $B = H - h$. The latter function is fixed by Lemma B.1. For the reward variance, write $R_i$ for the samples. The sample variance is at most the average squared deviation from $r_h(s,a)$. Bernstein's inequality therefore gives, with failure probability at most $e^{-\iota}$,

$$\begin{aligned}\widehat{v}_n &\leq \frac{1}{n}\sum_{i=1}^n (R_i - r_h(s,a))^2 \\ &\leq \mathbb{V}(R_h(s,a)) + \sqrt{\frac{2\mathbb{V}(R_h(s,a))\iota}{n}} + \frac{2\iota}{3n} \leq 2\mathbb{V}(R_h(s,a)) + \frac{2\iota}{n}.\end{aligned}$$

Here $(R_i - r_h(s,a))^2 \in [0,1]$ has mean $\mathbb{V}(R_h(s,a))$ and variance at most that mean. For each fixed triple and sample count, the three Bernstein bounds thus have total failure probability at most $(4+4+1)e^{-\iota} = 9e^{-\iota}$.

Taking a union bound over the deterministic sample counts, states, actions, and steps gives

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{\mathrm{cs}}^c) &\leq 2SAHKe^{-\iota} \leq \delta/8, \\ \mathbb{P}(\mathcal{E}_{\mathrm{est}}^c) &\leq \big[2(S+1)+9\big]SAHKe^{-\iota} \leq 13\delta/32 \leq 9\delta/16.\end{aligned}$$

□

### B.4 Proof of Lemma 1

*Proof.* We establish both inclusions on $\mathcal{E}_{\mathrm{cs}}$, including the state-set inclusion at step $H+1$. On $\mathcal{E}_{\mathrm{cs}}$, the cost bound gives $\underline{c}_h^k(s,a) \le c_h(s,a)$ for every $s,a,h,k$. Fix $k$ and argue backward from $\mathcal{V}_{\mathrm{safe},H+1} = \hat{\mathcal{V}}^k_{\mathrm{safe},H+1} = \mathcal{S}$. If $a \in \mathcal{A}_{\mathrm{safe},h}(s)$, then $\underline{c}_h^k(s,a) \le \tau$ and the induction hypothesis gives

$$\widehat{\mathrm{supp}}_h^k(s,a) \subseteq \mathrm{supp}(P_h(\cdot \mid s,a)) \subseteq \mathcal{V}_{\mathrm{safe},h+1} \subseteq \hat{\mathcal{V}}^k_{\mathrm{safe},h+1}.$$

Thus $a \in \hat{\mathcal{A}}^k_{\mathrm{safe},h}(s)$. A nonempty true safe action set therefore implies a nonempty candidate set, completing the induction. By Lemma B.2, $\mathbb{P}(\mathcal{E}_{\mathrm{cs}}) \ge 1 - \delta/8$, which proves the claim. □

### B.5 Optimism

The following consequence of Zhou et al. (2023b, Lemma 22) handles the dependence of the transition bonus on the optimistic value vector. For real vectors $p, z$ on the same finite index set, write $pz := \sum_i p_i z_i$.

**Lemma B.3.** *Let $p$ be a distribution on $[S]$, $B \ge 0$, $x > 0$, and $v, w \in [0,B]^S$ with $v \le w$ on $\mathrm{supp}(p)$. Then*

$$pw + \max\{4\sqrt{\mathbb{V}(p,w)x}, 16Bx\} \ge pv + 2\sqrt{\mathbb{V}(p,v)x} + 8Bx. \tag{B.7}$$

*Proof.* For $B > 0$, apply Zhou et al. (2023b, Lemma 22) on $\mathrm{supp}(p)$ with $c_1 = 4$, $c_2 = 16$, and $\iota/n = x$. Its monotonicity property applies because $c_1^2 = c_2$. Combining it with $\max\{4a, 16b\} \ge 2a + 8b$ for $a, b \ge 0$ proves the claim. For $B = 0$, both sides are zero. □

Combining candidate-set preservation with the monotone backup gives optimism on the viable states.

**Lemma B.4** (Optimism). *On $\mathcal{E}_{\mathrm{est}}$, $V_h^k(s) \ge V_h^\star(s)$ for every $k \in [K]$, $h \in [H+1]$, and $s \in \mathcal{V}_{\mathrm{safe},h}$.*

*Proof.* Fix $k$ and use backward induction from $h = H+1$. For a viable state $s$, let $a^\star = \pi_h^\star(s)$. By the inclusions proved in Appendix B.4, $a^\star$ is a candidate. Write $N := N_h^k(s,a^\star)$ and suppress $(s,a^\star,h,k)$ from the remaining empirical quantities. If $N \le 1$, the linear bonus gives $Q_h^k(s,a^\star) = H - h + 1 \ge V_h^\star(s)$.

If $N \ge 2$, set $x = \iota/N$ and $q = \widehat{P}_h^k(\cdot \mid s,a^\star)$. Since $\mathrm{supp}(q) \subseteq \mathcal{V}_{\mathrm{safe},h+1}$, the induction hypothesis holds there. Lemma B.3 and (B.5) yield

$$\begin{aligned} qV_{h+1}^k + 4\sqrt{\mathbb{V}(q,V_{h+1}^k)x} + 16(H-h)x &\ge qV_{h+1}^\star + 2\sqrt{\mathbb{V}(q,V_{h+1}^\star)x} + 8(H-h)x \\ &\ge P_h V_{h+1}^\star. \end{aligned}$$

Also, (B.4) gives $\hat{r}_h^k + 2\sqrt{\hat{\sigma}_{R,h}^{2,k}x} + 5x \ge r_h$. The unclipped backup is therefore at least $r_h + P_h V_{h+1}^\star = V_h^\star(s)$. Clipping at $H-h+1$ and maximizing over candidates preserve this inequality. □

## C Regret Analysis

We prove Theorem 2 under Assumption 1. The analysis separates each episode at its first action outside the true safe set. Before that action, an auxiliary safe policy relates the Bellman residuals to the safe-policy variance. We bound the number of episodes containing such an action in terms of $\mathcal{C}$ and use a general residual bound for their remaining steps.

### C.1 Martingale Bounds

We use the following consequence of Freedman's inequality (Freedman, 1975; Howard et al., 2020).

**Lemma C.1** (Martingale concentration). *Let $T \geq 1$ be an integer, let $(\mathcal{H}_t)_{t=0}^T$ be a filtration, and let $(\xi_t)_{t=1}^T$ be martingale differences with respect to this filtration. Suppose $|\xi_t| \leq b$ almost surely for every $t \in [T]$, where $b \geq 0$ is fixed. Then, for every $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$\sum_{t=1}^{T} \xi_t \leq \mathcal{O}\left(\sqrt{\left(\sum_{t=1}^{T} \mathbb{E}[\xi_t^2 \mid \mathcal{H}_{t-1}]\right) \log \frac{T+1}{\eta}} + b \log \frac{T+1}{\eta}\right).$$

*For the following consequences, let $B \geq 0$ be fixed and, for each $t \in [T]$, let $Y_t$ be $\mathcal{H}_t$-measurable with $0 \leq Y_t \leq B$ almost surely. Set $\mu_t := \mathbb{E}[Y_t \mid \mathcal{H}_{t-1}]$. For every $\eta \in (0,1)$, each of the following inequalities holds with probability at least $1-\eta$:*

$$\sum_{t=1}^{T} Y_t \leq 2 \sum_{t=1}^{T} \mu_t + \mathcal{O}\left(B \log \frac{T+1}{\eta}\right), \tag{C.1}$$

$$\sum_{t=1}^{T} \mu_t \leq 2 \sum_{t=1}^{T} Y_t + \mathcal{O}\left(B \log \frac{T+1}{\eta}\right). \tag{C.2}$$

*Proof.* The case $b = 0$ is immediate. Otherwise, apply Freedman's inequality at variance thresholds $b^2, 2b^2, \ldots, 2^{\lceil \log_2 T \rceil} b^2$, with failure probability $\eta/(T+1)$ at each threshold. The conditional variance sum is at most $Tb^2$, so these at most $T+1$ thresholds cover all its values. A union bound yields the first inequality, with the lowest interval absorbed by the term $b \log((T+1)/\eta)$. For the first consequence, apply it to $\xi_t = Y_t - \mu_t$ and use $\mathbb{E}[\xi_t^2 \mid \mathcal{H}_{t-1}] \leq \mathbb{E}[Y_t^2 \mid \mathcal{H}_{t-1}] \leq B\mu_t$. Young's inequality absorbs the resulting term $\mathcal{O}(\sqrt{B(\sum_t \mu_t) \log((T+1)/\eta)})$. Applying the same argument to $\xi_t = \mu_t - Y_t$ proves the second consequence. □

For the applications below, $t \leq KH$ and $\eta = \delta/16$ imply

$$\log \frac{t+1}{\eta} \leq \log \frac{32KH}{\delta} \leq 2\iota. \tag{C.3}$$

### C.2 Safety Identification

We use $\kappa_h(s,a)$ and $\mathcal{C}$ from Section 3.3 to bound the number of selections of unsafe actions in the candidate sets and, in turn, the number of episodes containing an unsafe action. The recursion for $\kappa_h$ ends at an action with a positive cost gap. Otherwise, it selects a non-viable successor and sums the sample requirements for ruling out all actions there. The next lemma expands this recursion into paths ending at positive-gap actions. Each leaf contributes its sample requirement $\Delta_\ell^{-2}$ divided by its path probability $q_\ell$.

**Lemma C.2** (Witness-tree representation). *Fix $h \in [H]$, $s \in \mathcal{S}$, and $a \notin \mathcal{A}_{\mathrm{safe},h}(s)$. There is a finite witness tree rooted at $(s,a,h)$ with a nonempty leaf set $\mathcal{L}$. For each leaf $\ell$, let $q_\ell \in (0,1]$ be the product of the transition probabilities on its path and $\Delta_\ell \in (0,1]$ the cost gap of its terminal action. Then*

$$\kappa_h(s,a) = \sum_{\ell \in \mathcal{L}} \frac{1}{q_\ell \Delta_\ell^2}. \tag{C.4}$$

*For SVAE, each execution of $(s,a,h)$ as a candidate action is assigned to the leaf determined by its episode policy, before the transitions on that path are observed. Across the $K$ episodes, let $N_\ell$ count the executions assigned to leaf $\ell$ for which every transition on that path occurs. On $\mathcal{E}_c$, the following bound holds for all $\ell \in \mathcal{L}$:*

$$N_\ell \leq 2 + \frac{8\iota}{\Delta_\ell^2} \leq \frac{10\iota}{\Delta_\ell^2}. \tag{C.5}$$

*Proof.* We prove (C.4) by backward induction. For the direct-cost branch, take one leaf with $q_\ell = 1$ and $\Delta_\ell = c_h(s,a) - \tau$. Then $\kappa_h(s,a) = \frac{1}{q_\ell \Delta_\ell^2}$.

Suppose now that the recursive branch applies. The minimizing set in (7) is nonempty because $a \notin \mathcal{A}_{\mathrm{safe},h}(s)$ and $c_h(s,a) \leq \tau$. Choose a minimizing successor $y$. Since $y \notin \mathcal{V}_{\mathrm{safe},h+1}$, $\mathcal{A}_{\mathrm{safe},h+1}(y) =$

$\varnothing$. By the induction hypothesis, each $\kappa_{h+1}(y, b)$, $b \in \mathcal{A}$, has a finite witness representation. Taking the union of these representations and multiplying every path probability by $P_h(y \mid s, a)$ gives

$$\kappa_h(s,a) = \frac{\kappa_{h+1}(y)}{P_h(y \mid s,a)} = \sum_{\ell \in \mathcal{L}} \frac{1}{q_\ell \Delta_\ell^2}.$$

The construction terminates after at most $H - h + 1$ layers. Each internal action has the selected successor $y$, whose children are indexed by all actions at $y$. A fixed episode policy selects the child $\pi_{h+1}^k(y)$ at each such successor and hence determines a unique leaf before any transition on the path is observed. The product of these transition probabilities is $q_\ell$.

Fix $\ell \in \mathcal{L}$ and write $c = \tau + \Delta_\ell$ for the mean cost of its terminal action. After $n \geq 1$ visits, its lower confidence bound $\underline{c}$ satisfies $\underline{c} \geq c - 2\sqrt{2\iota/n}$ on $\mathcal{E}_c$. Thus the terminal action is permanently removed by the cost test once $n > 8\iota/\Delta_\ell^2$. Before it is removed, it can be executed at most $1 + 8\iota/\Delta_\ell^2$ times.

It remains to account for paths whose transitions have not all been observed. After the terminal action is permanently rejected, a selected path to it can be completed at most once more. The first such completion records every transition on the path. In any later episode selecting the same leaf, suppose the root action is a candidate. Its recorded successor must then have a nonempty candidate set, so the algorithm selects a candidate action there. Repeating this argument along the recorded path would make the permanently rejected terminal action a candidate, a contradiction. Recorded supports only grow, and the terminal cost bound holds at all later sample counts. Adding this possible completion to the visits before removal proves (C.5). □

We now convert completed leaf paths into a bound on root executions. Weighting a completed path by $1/q_\ell$ makes its conditional mean equal to the indicator that the root is executed as a candidate.

**Lemma C.3.** *For SVAE, there exists an event $\mathcal{E}_{\text{cert}}$ with $\mathbb{P}(\mathcal{E}_{\text{cert}}) \geq 1 - \delta/16$ such that, on $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{cert}}$, the following bound holds simultaneously for all $h \in [H]$, $s \in \mathcal{S}$, and $a \notin \mathcal{A}_{\text{safe},h}(s)$:*

$$\sum_{k=1}^{K} \mathbf{1}\left\{ s_h^k = s,\ a_h^k = a,\ a \in \hat{\mathcal{A}}_{\text{safe},h}^k(s) \right\} \leq \mathcal{O}(\iota \kappa_h(s,a)). \tag{C.6}$$

*Proof.* Fix $(s, a, h)$ and the representation in Lemma C.2. Let $I_k$ denote the indicator in (C.6). If $I_k = 1$, let $\ell_k$ be the leaf determined by $\pi^k$, and let $Z_k$ indicate that all transitions along its path occur. Set $X_k = Z_k/q_{\ell_k}$ when $I_k = 1$ and $X_k = 0$ otherwise. Since $\pi^k$ is $\mathcal{F}_1^k$-measurable, $\ell_k$ and $q_{\ell_k}$ are $\mathcal{F}_h^k$-measurable. The Markov property and (C.4) give $\mathbb{E}[X_k \mid \mathcal{F}_h^k] = I_k$, and $0 \leq X_k \leq \kappa_h(s,a)$, where the second bound also uses $\Delta_{\ell_k} \leq 1$. On $\mathcal{E}_c$, (C.5) gives

$$\sum_{k=1}^{K} X_k = \sum_{\ell \in \mathcal{L}} \frac{N_\ell}{q_\ell} \leq 10\iota \sum_{\ell \in \mathcal{L}} \frac{1}{q_\ell \Delta_\ell^2} = 10\iota \kappa_h(s,a).$$

Apply (C.2) in episode order, conditioning the $k$th increment on $\mathcal{F}_h^k$. The variable $X_k$ is known by the end of episode $k$, before $\mathcal{F}_h^{k+1}$ for $k < K$. For the final increment, use the terminal history $\mathcal{F}_{H+1}^K$. Take $B = \kappa_h(s,a)$ and failure probability $\delta/(16SAH)$. Let $\mathcal{E}_{\text{cert}}$ be the intersection of the events supplied by these applications over all unsafe $(s, a, h)$. A union bound gives $\mathbb{P}(\mathcal{E}_{\text{cert}}) \geq 1 - \delta/16$. On $\mathcal{E}_{\text{cert}}$,

$$\sum_{k=1}^{K} I_k \leq 2 \sum_{k=1}^{K} X_k + \mathcal{O}\left( \kappa_h(s,a) \log \frac{16SAH(K+1)}{\delta} \right)$$

for every such triple. Since $\log(16SAH(K+1)/\delta) \leq 2\iota$, the bound on $\sum_k X_k$ above yields, on $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{cert}}$, $\sum_k I_k \leq \mathcal{O}(\iota \kappa_h(s,a))$, as required. □

We now apply this count to the first unsafe action in each episode. For episode $k$, define

$$\tau_k := \min\left\{ h \in [H] : a_h^k \notin \mathcal{A}_{\text{safe},h}(s_h^k) \right\}, \qquad \min \varnothing := H + 1.$$

The event $\{\tau_k \leq h\}$ belongs to $\mathcal{F}_h^k$, so $\tau_k$ is a stopping time for the within-episode history. All actions before $\tau_k$ are safe, so the state at step $\tau_k$ is reachable under a safe policy. The bound on unsafe candidate selections therefore controls the number of episodes with $\tau_k \leq H$.

**Proposition C.1.** *On* $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{cert}}$,

$$\sum_{k=1}^{K} \mathbf{1}\{\tau_k \le H\} \le \mathcal{O}(\iota \min\{K, \mathcal{C}\}). \tag{C.7}$$

*Proof.* If $\tau_k \le H$, the preceding safe actions reach $(s^k_{\tau_k}, \tau_k)$ with positive probability. Complete those actions with arbitrary safe actions at other viable states. The resulting policy is safe, so this vertex is included in the definition of $\mathcal{C}$. Its true safe action set is nonempty and, by the inclusions proved in Appendix B.4, so is its candidate set. Thus $a^k_{\tau_k}$ is an unsafe candidate action.

Apply Lemma C.3 to each unsafe action at a vertex reachable under a safe policy and sum the bounds. Each episode with $\tau_k \le H$ contributes one such execution, giving

$$\sum_{k=1}^{K} \mathbf{1}\{\tau_k \le H\} \le \mathcal{O}\left(\iota \sum_{(s,h)\in\bigcup_{\pi\in\Pi_{\text{safe}}} \mathcal{V}^\pi} \kappa_h(s)\right) = \mathcal{O}(\iota\mathcal{C}). \tag{C.8}$$

The count is also at most $K$. Combining these bounds and using $\iota \ge 1$ proves the claim. □

### C.3 Bellman Residual Bounds

We first control the transition estimation error, then derive the two Bellman residual bounds used before and after the first unsafe action. The next lemma converts coordinatewise transition bounds into mean and variance bounds for every bounded value vector, including the data-dependent value estimates used below.

**Lemma C.4.** *Let* $B \ge 0$, $x > 0$, *and let* $p, q$ *be distributions on* $[S]$ *satisfying*

$$|q_i - p_i| \le \sqrt{2p_i x} + \mathbf{1}\{p_i > 0\}x \qquad (i \in [S]). \tag{C.9}$$

*Then, for every* $z \in [0, B]^S$,

$$|(q-p)z| \le 2\sqrt{S\mathbb{V}(p,z)x} + 3SBx, \tag{C.10}$$

$$\mathbb{V}(q,z) \le 2\mathbb{V}(p,z) + 4SB^2x. \tag{C.11}$$

*Proof.* Let $d_i = z_i - pz$. Since $(q-p)\mathbf{1} = 0$ and $|d_i| \le B$,

$$\begin{aligned}|(q-p)z| &\le \sqrt{2x}\sum_i \sqrt{p_i}|d_i| + x\sum_{i:p_i>0}|d_i| \\ &\le \sqrt{2S\mathbb{V}(p,z)x} + SBx,\end{aligned}$$

which implies (C.10). Also,

$$\begin{aligned}\mathbb{V}(q,z) &\le \sum_{i=1}^{S} q_i d_i^2 \\ &\le \mathbb{V}(p,z) + \sqrt{2x}\sum_i \sqrt{p_i}d_i^2 + x\sum_{i:p_i>0} d_i^2 \\ &\le \mathbb{V}(p,z) + B\sqrt{2S\mathbb{V}(p,z)x} + SB^2x \le 2\mathbb{V}(p,z) + 2SB^2x.\end{aligned}$$

The last two inequalities use Cauchy–Schwarz and Young's inequality. This proves (C.11). □

Before the first unsafe action, we bound the Bellman residual using $V^\star$. For the remaining steps, we use a bound that applies to every action. For $h \in [H+1]$, define $D^k_h(s) := V^k_h(s) - V^\star_h(s)$ for $s \in \mathcal{V}_{\text{safe},h}$ and $D^k_h(s) := 0$ otherwise. On $\mathcal{E}_{\text{est}}$, Lemma B.4 gives $0 \le D^k_h \le H - h + 1$.

Fix a sufficiently large universal constant $C_0$. Define

$$\begin{aligned}\beta^k_h(s,a) :=& C_0\left[\sqrt{\frac{\{\mathbb{V}(R_h(s,a)) + \mathbb{V}(P_h(\cdot \mid s,a), V^\star_{h+1})\}\iota}{n^k_h(s,a)}}\right. \\ &\left. + \sqrt{\frac{S\mathbb{V}(P_h(\cdot \mid s,a), D^k_{h+1})\iota}{n^k_h(s,a)}} + \frac{SH\iota}{n^k_h(s,a)}\right],\end{aligned} \tag{C.12}$$

and

$$\widetilde{\beta}_h^k(s,a) := C_0 \left[ \sqrt{\frac{\{\mathbb{V}(R_h(s,a)) + S\mathbb{V}(P_h(\cdot \mid s,a), V_{h+1}^k)\}\iota}{n_h^k(s,a)}} + \frac{SH\iota}{n_h^k(s,a)} \right]. \tag{C.13}$$

The same $C_0$ is used in both definitions. The three terms in $\beta_h^k$ account for the reward and optimal-value variances, the next-state variance of $D_{h+1}^k$, and the linear concentration terms. The bound $\widetilde{\beta}_h^k$ uses $V_{h+1}^k$ directly.

**Lemma C.5** (Bellman residual bounds)**.** *On $\mathcal{E}_{\text{est}}$, for every $k \in [K]$, $h \in [H]$, and $s \in \mathcal{S}$, the following two bounds hold. For every $a \in \mathcal{A}_{\text{safe},h}(s)$,*

$$Q_h^k(s,a) - r_h(s,a) - P_h V_{h+1}^k(s,a) \le \beta_h^k(s,a). \tag{C.14}$$

*For every $a \in \mathcal{A}$,*

$$Q_h^k(s,a) - r_h(s,a) - P_h V_{h+1}^k(s,a) \le \widetilde{\beta}_h^k(s,a). \tag{C.15}$$

*Proof.* Fix $(s,a,h,k)$ and suppress these indices when unambiguous. Let $N = N_h^k(s,a)$ and $B = H - h$. If $N \le 1$, the residual is at most $H$ by clipping and nonnegative rewards. The linear term in either bound covers this case. Suppose $N \ge 2$, and set $x = \iota/N$, $p = P_h(\cdot \mid s,a)$, and $q = \widehat{P}_h^k(\cdot \mid s,a)$. The reward terms satisfy

$$\begin{aligned} 2\sqrt{\widehat{\sigma}_{R,h}^{2,k} x} + 5x + |\widehat{r}_h^k - r_h| &\le 4\sqrt{\widehat{\sigma}_{R,h}^{2,k} x} + 10x \\ &\le \mathcal{O}\Big(\sqrt{\mathbb{V}(R_h(s,a))x} + x\Big) \end{aligned} \tag{C.16}$$

by (B.4) and (B.6).

For a safe action $a \in \mathcal{A}_{\text{safe},h}(s)$, write $V = V_{h+1}^\star$ and $D = D_{h+1}^k$. On the true support, $V_{h+1}^k = V + D$ and $V, D \in [0,B]$. The empirical support is contained in the true support. Clipping gives

$$Q_h^k - r_h - pV_{h+1}^k \le b_h^k + (\widehat{r}_h^k - r_h) + (q-p)V + (q-p)D. \tag{C.17}$$

By (C.11), for $z = V, D$, $\sqrt{\mathbb{V}(q,z)x} \le \sqrt{2\mathbb{V}(p,z)x} + 2\sqrt{S}\,Bx$. Using $\mathbb{V}(q, V+D) \le 2\mathbb{V}(q,V) + 2\mathbb{V}(q,D)$, we obtain

$$\begin{aligned} &4\sqrt{\mathbb{V}(q,V+D)x} + 16Bx + |(q-p)V| + |(q-p)D| \\ &\quad \le (4\sqrt{2}+2)\sqrt{\mathbb{V}(q,V)x} + 4\sqrt{2}\sqrt{\mathbb{V}(q,D)x} + 2\sqrt{S\mathbb{V}(p,D)x} + (21+3S)Bx \\ &\quad \le \mathcal{O}\Big(\sqrt{\mathbb{V}(p,V)x} + \sqrt{S\mathbb{V}(p,D)x} + SBx\Big). \end{aligned} \tag{C.18}$$

The first inequality uses (B.5) for the fixed function $V$ and (C.10) for $D$. The second uses the two variance comparisons above. Combining (C.17), (C.16), and (C.18) proves (C.14).

Now let $a$ be arbitrary and write $V = V_{h+1}^k$. Lemma C.4 applies directly to this data-dependent function and gives

$$4\sqrt{\mathbb{V}(q,V)x} + 16Bx + |(q-p)V| \le \mathcal{O}\Big(\sqrt{S\mathbb{V}(p,V)x} + SBx\Big).$$

Together with (C.16) and the clipped backup, this proves (C.15). A sufficiently large $C_0$ makes both residual bounds hold. □

### C.4 Regret Decomposition

At viable states, we replace each unsafe action of $\pi^k$ by the corresponding action of $\pi^\star$. This gives a safe policy that agrees with $\pi^k$ before the first unsafe action and allows us to use the safe-policy variance. For $h \in [H]$ and $s \in \mathcal{V}_{\text{safe},h}$, define

$$\bar{\pi}_h^k(s) := \begin{cases} \pi_h^k(s), & \pi_h^k(s) \in \mathcal{A}_{\text{safe},h}(s), \\ \pi_h^\star(s), & \pi_h^k(s) \notin \mathcal{A}_{\text{safe},h}(s), \end{cases} \tag{C.19}$$

with arbitrary actions at states outside $\mathcal{V}_{\mathrm{safe},h}$. For $h \in [H+1]$, define $E_h^k(s) := V_h^k(s) - V_h^{\bar{\pi}^k}(s)$ for $s \in \mathcal{V}_{\mathrm{safe},h}$ and $E_h^k(s) := 0$ otherwise. Thus $D_h^k$ compares with the optimal safe value, whereas $E_h^k$ compares with the value of $\bar{\pi}^k$. For every $h \in [H]$, define the Bellman residual at the executed pair

$$f_h^k := V_h^k(s_h^k) - r_h(s_h^k, a_h^k) - P_h V_{h+1}^k(s_h^k, a_h^k).$$

**Lemma C.6.** *For every $k \in [K]$, the policy $\bar{\pi}^k$ is $\mathcal{F}_1^k$-measurable and belongs to $\Pi_{\mathrm{safe}}$. Couple $\pi^k$ and $\bar{\pi}^k$ using the same transition randomness. Their states then agree through step $\tau_k$, and their actions agree for all $h \in [H]$ with $h < \tau_k$.*

*For every $k \in [K]$ and $h \in [H]$ with $h < \tau_k$,*

$$\begin{aligned} D_h^k(s_h^k) - P_h D_{h+1}^k(s_h^k, a_h^k) &\le f_h^k, \\ E_h^k(s_h^k) - P_h E_{h+1}^k(s_h^k, a_h^k) &= f_h^k. \end{aligned} \tag{C.20}$$

*On $\mathcal{E}_{\mathrm{est}}$, $0 \le D_h^k(s) \le E_h^k(s) \le H - h + 1$ for every $k \in [K]$, $h \in [H+1]$, and $s \in \mathcal{V}_{\mathrm{safe},h}$.*

*Proof.* At every viable state, both cases in (C.19) select an action in $\mathcal{A}_{\mathrm{safe},h}(s)$. The support condition therefore makes $\bar{\pi}^k$ safe, and its definition makes it $\mathcal{F}_1^k$-measurable. For $h < \tau_k$, its action agrees with the executed action, so the coupled trajectories agree through the state at step $\tau_k$. Applying (B.3) to the safe action $a_h^k$ gives the inequality for $D_h^k$ in (C.20). The Bellman equation for $\bar{\pi}^k$ gives the equality for $E_h^k$. On $\mathcal{E}_{\mathrm{est}}$, Lemmas B.4 and B.1 give $V_h^k \ge V_h^\star \ge V_h^{\bar{\pi}^k}$ on $\mathcal{V}_{\mathrm{safe},h}$. Clipping and nonnegative rewards give the upper bound. □

Define

$$B_K := \sum_{k=1}^{K} \sum_{h<\tau_k} \min\{\beta_h^k(s_h^k, a_h^k), H\},$$

$$\widetilde{B}_K := \sum_{k=1}^{K} \sum_{h=\tau_k}^{H} \min\{\widetilde{\beta}_h^k(s_h^k, a_h^k), H\}.$$

The second inner sum is empty when $\tau_k = H + 1$. To pass from expected residuals to these realized sums, define the episode contribution

$$Y_k := \sum_{h<\tau_k} \min\{\beta_h^k(s_h^k, a_h^k), H\} + \sum_{h=\tau_k}^{H} \min\{\widetilde{\beta}_h^k(s_h^k, a_h^k), H\}, \qquad 0 \le Y_k \le H^2.$$

**Proposition C.2.** *On $\mathcal{E}_{\mathrm{est}}$, the reward regret satisfies $\mathcal{R}(K) \le \sum_{k=1}^{K} \mathbb{E}[Y_k \mid \mathcal{F}_1^k]$.*

*Proof.* Fix the history at the start of episode $k$. The Bellman equations for $\pi^k$, with $V_{H+1}^k = V_{H+1}^{\pi^k} = 0$, give

$$V_1^k(s_1) - V_1^{\pi^k}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H} f_h^k \,\middle|\, \mathcal{F}_1^k\right]. \tag{C.21}$$

If the episode's estimates satisfy the inequalities defining $\mathcal{E}_{\mathrm{est}}$, Lemma C.5 bounds the residual at every state and action. Hence, for every continuation trajectory, $\sum_h f_h^k \le Y_k$: use $\beta_h^k$ before $\tau_k$, $\widetilde{\beta}_h^k$ from $\tau_k$ onward, and $f_h^k \le H$ throughout. On $\mathcal{E}_{\mathrm{est}}$, optimism therefore yields

$$V_1^\star(s_1) - V_1^{\pi^k}(s_1) \le V_1^k(s_1) - V_1^{\pi^k}(s_1) \le \mathbb{E}[Y_k \mid \mathcal{F}_1^k].$$

Summing over $k$ proves the result. □

### C.5 Variance Bounds

Proposition C.2 reduces regret to conditional expectations of residual sums. We first control the visit-count and variance terms in these sums, then collect the concentration bounds that relate them to their conditional expectations. We write $\sum_{k,h}$ for $\sum_{k=1}^{K} \sum_{h=1}^{H}$; a restriction such as $h < \tau_k$ is applied within each episode.

The reward and optimal-value variances are fixed for each state, action, and step, whereas the value-error variances can change between episodes. The next lemma gives visit-count bounds for both types of weights.

**Lemma C.7** (Visit-count sums)**.** *Fix $(s, a, h)$ and enumerate its visits during the $K$ episodes. Set $n_j := \max\{j-1, 1\}$, the visit count before visit $j$. For any subset $J$ of these visit indices, write $m := |J|$. Then $\sum_{j\in J} n_j^{-1/2} \le 2\sqrt{m}$ and $\sum_{j\in J} n_j^{-1} \le L_K$.*

*Consequently, for any nonnegative weights $v_h(s, a)$ fixed across episodes and any nonnegative weights $z_h^k(s, a)$, the following bounds hold for every realized interaction sequence:*

$$\sum_{k,h<\tau_k} \sqrt{\frac{v_h(s_h^k, a_h^k)}{n_h^k(s_h^k, a_h^k)}} \le 2\sqrt{SAH \sum_{k,h<\tau_k} v_h(s_h^k, a_h^k)},$$

$$\sum_{k,h<\tau_k} \sqrt{\frac{z_h^k(s_h^k, a_h^k)}{n_h^k(s_h^k, a_h^k)}} \le \sqrt{SAHL_K \sum_{k,h<\tau_k} z_h^k(s_h^k, a_h^k)},$$

$$\sum_{k,h<\tau_k} \frac{1}{n_h^k(s_h^k, a_h^k)} \le SAHL_K.$$

*The same bounds hold with $h \ge \tau_k$.*

*Proof.* Both reciprocal sequences are nonincreasing, so the sum over a subset of size $m$ is maximized by the first $m$ visits. For $m \ge 2$,

$$2 + \sum_{q=2}^{m-1} q^{-1/2} \le 2 + \int_1^{m-1} x^{-1/2}\, dx = 2\sqrt{m-1} \le 2\sqrt{m}.$$

Similarly, $m \le K$ gives $2 + \sum_{q=2}^{m-1} q^{-1} \le 2 + \log K = L_K$. The case $m \le 1$ is immediate. For the fixed weights $v_h(s, a)$, first apply the bound $2\sqrt{m}$ at each $(s, a, h)$, and then apply Cauchy–Schwarz over the $SAH$ choices. For the varying weights, apply Cauchy–Schwarz to all visited terms and use $\sum_{k,h<\tau_k} 1/n_h^k(s_h^k, a_h^k) \le SAHL_K$. This proves all three bounds and also applies to $h \ge \tau_k$. □

After summing the visit-count factors, we need bounds on the next-state variances of $D^k$ and $E^k$. The next lemma applies to bounded nonnegative functions. Telescoping their squares before $\tau_k$ leaves a drift term and a boundary term at $\tau_k$. We will apply it to the positive parts of $D^k$ and $E^k$.

**Lemma C.8.** *For each $k \in [K]$, let $g_h^k : \mathcal{S} \to [0, H]$, $h \in [H+1]$, be $\mathcal{F}_1^k$-measurable functions. Assume $g_{H+1}^k \equiv 0$ and $g_h^k(s) = 0$ outside $\mathcal{V}_{\mathrm{safe},h}$. Let $u_h^k$, $k \in [K]$ and $h \in [H]$, be random variables. Suppose that, almost surely, for every $k \in [K]$ and $h \in [H]$ with $h < \tau_k$, $0 \le u_h^k \le H$ and*

$$g_h^k(s_h^k) - P_h g_{h+1}^k(s_h^k, a_h^k) \le u_h^k.$$

*Then, for every $\eta \in (0, 1)$, with probability at least $1 - \eta$,*

$$\sum_{k,h<\tau_k} \mathbb{V}(P_h(\cdot \mid s_h^k, a_h^k), g_{h+1}^k) \le \mathcal{O}\left( H \sum_{k,h<\tau_k} u_h^k + H^2 \sum_{k=1}^{K} \mathbf{1}\{\tau_k \le H\} + H^2 \log \frac{KH+1}{\eta} \right). \tag{C.22}$$

*Proof.* Set $I_h^k = \mathbf{1}\{h < \tau_k\}$ and $q_h^k = P_h g_{h+1}^k(s_h^k, a_h^k)$. The indicator $I_h^k$ is measurable after the step-$h$ action is chosen and before the successor is generated. In episode and step order, define the martingale differences

$$\xi_h^k := I_h^k \left\{ P_h[(g_{h+1}^k)^2](s_h^k, a_h^k) - (g_{h+1}^k(s_{h+1}^k))^2 \right\}.$$

Within episode $k$, telescoping gives

$$\begin{aligned}&\sum_{h<\tau_k}\left\{(g_{h+1}^k(s_{h+1}^k))^2-(q_h^k)^2\right\}\\&=\sum_{h<\tau_k}\left\{(g_h^k(s_h^k))^2-(q_h^k)^2\right\}-(g_1^k(s_1))^2+(g_{\tau_k}^k(s_{\tau_k}^k))^2\\&\le 2H\sum_{h<\tau_k}u_h^k+H^2\mathbf{1}\{\tau_k\le H\}.\end{aligned}$$

Indeed, $(g_h^k)^2-(q_h^k)^2\le 2H[g_h^k-q_h^k]_+\le 2Hu_h^k$. The terminal square is zero when $\tau_k=H+1$ and at most $H^2$ otherwise. When $\tau_k=1$, both sums are empty and the two boundary squares cancel.

Let

$$Z_g:=\sum_{k,h<\tau_k}\mathbb{V}(P_h(\cdot\mid s_h^k,a_h^k),g_{h+1}^k).$$

Adding the martingale differences to the deterministic sum above gives

$$Z_g\le\sum_{k,h}\xi_h^k+2H\sum_{k,h<\tau_k}u_h^k+H^2\sum_k\mathbf{1}\{\tau_k\le H\}.$$

The increments have absolute value at most $H^2$. For any $G\in[0,H]$ and an independent copy $G'$, the variance identity gives

$$\mathbb{V}(G^2)=\frac{1}{2}\mathbb{E}[(G^2-(G')^2)^2]\le 2H^2\mathbb{E}[(G-G')^2]=4H^2\mathbb{V}(G).$$

Apply this conditionally to $G=g_{h+1}^k(s_{h+1}^k)$. The predictable quadratic variation of the martingale is therefore at most $4H^2Z_g$. Lemma C.1 yields

$$\sum_{k,h}\xi_h^k\le\mathcal{O}\left(H\sqrt{Z_g\log\frac{KH+1}{\eta}}+H^2\log\frac{KH+1}{\eta}\right).$$

Young's inequality absorbs the square-root term into $Z_g/2$ and proves (C.22). □

For the steps from $\tau_k$ onward, we telescope the squares of $V^k$ up to step $H+1$. The terminal value is zero, and the square at $\tau_k$ enters with a negative sign, so no positive boundary term remains.

**Lemma C.9.** *There exists an event $\mathcal{E}_{\mathrm{suf}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{suf}})\ge 1-\delta/16$. On $\mathcal{E}_{\mathrm{est}}\cap\mathcal{E}_{\mathrm{suf}}$, the accumulated variance satisfies*

$$\sum_{k=1}^K\sum_{h=\tau_k}^H\mathbb{V}(P_h(\cdot\mid s_h^k,a_h^k),V_{h+1}^k)\le\mathcal{O}\left(H\widetilde{B}_K+H^2\sum_{k=1}^K\mathbf{1}\{\tau_k\le H\}+H^2\iota\right).\tag{C.23}$$

*Proof.* Set $J_h^k:=\mathbf{1}\{\tau_k\le h\}$, $q_h^k:=P_hV_{h+1}^k(s_h^k,a_h^k)$, and

$$u_h^k:=\min\Big\{H,1+\min\{\widetilde{\beta}_h^k(s_h^k,a_h^k),H\}\Big\}.$$

The indicators $J_h^k$ are predictable, and $V^k$ is fixed before episode $k$. In episode and step order, define the martingale differences $\xi_h^k:=J_h^k\left\{P_h[(V_{h+1}^k)^2](s_h^k,a_h^k)-(V_{h+1}^k(s_{h+1}^k))^2\right\}$. Let $Z$ denote the left-hand side of (C.23). The increments are bounded by $H^2$. As in the proof of Lemma C.8, their conditional variance sum is at most $4H^2Z$. Apply Lemma C.1 with failure probability $\delta/16$, and let $\mathcal{E}_{\mathrm{suf}}$ be the resulting event. By (C.3), on $\mathcal{E}_{\mathrm{suf}}$,

$$\sum_{k,h}\xi_h^k\le\mathcal{O}\Big(H\sqrt{Z\iota}+H^2\iota\Big).$$

We now bound the drift on $\mathcal{E}_{\mathrm{est}}$. The residual bound (C.15), $r_h\le 1$, and clipping give $V_h^k(s_h^k)-q_h^k\le u_h^k$. By the definition of $u_h^k$,

$$\sum_{k,h}J_h^ku_h^k\le\widetilde{B}_K+H\sum_{k=1}^K\mathbf{1}\{\tau_k\le H\}.\tag{C.24}$$

If $\tau_k \le H$, telescoping gives

$$\begin{aligned}
&\sum_{h=\tau_k}^{H} \left\{ (V_{h+1}^k(s_{h+1}^k))^2 - (q_h^k)^2 \right\} \\
&\quad = \sum_{h=\tau_k}^{H} \left\{ (V_h^k(s_h^k))^2 - (q_h^k)^2 \right\} - (V_{\tau_k}^k(s_{\tau_k}^k))^2 + (V_{H+1}^k(s_{H+1}^k))^2 \\
&\quad \le 2H \sum_{h=\tau_k}^{H} u_h^k.
\end{aligned}$$

The inequality follows from $(V_h^k)^2 - (q_h^k)^2 \le 2H[V_h^k - q_h^k]_+ \le 2Hu_h^k$ and $V_{H+1}^k \equiv 0$. For $\tau_k = H+1$, the sum is empty. Adding the martingale differences and summing over episodes gives $Z \le \sum_{k,h} \xi_h^k + 2H \sum_{k,h} J_h^k u_h^k$. On $\mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_{\mathrm{suf}}$, Young's inequality absorbs the term containing $\sqrt{Z}$ into $Z/2$ and gives

$$Z \le \mathcal{O}\left( H \sum_{k,h} J_h^k u_h^k + H^2 \iota \right).$$

Substituting (C.24) proves the claim. □

We next apply the preceding bounds to the residual sums and the auxiliary safe policy. Define the accumulated next-state variances of the two value differences by

$$\begin{aligned}
Z_D &:= \sum_{k,h<\tau_k} \mathbb{V}(P_h(\cdot \mid s_h^k, a_h^k), D_{h+1}^k), \\
Z_E &:= \sum_{k,h<\tau_k} \mathbb{V}(P_h(\cdot \mid s_h^k, a_h^k), E_{h+1}^k).
\end{aligned}$$

The auxiliary policy relates the variance accumulated before $\tau_k$ to a safe-policy return variance. For each episode, define

$$A_k := \sum_{h<\tau_k} \left\{ \mathbb{V}(R_h(s_h^k, a_h^k)) + \mathbb{V}(P_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^{\bar{\pi}^k}) \right\}. \tag{C.25}$$

The next lemma converts the expected residuals to realized sums and bounds the accumulated variances needed in the regret proof.

**Lemma C.10.** *There exists an event $\mathcal{E}_{\mathrm{mart}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{mart}}) \ge 1 - 5\delta/16$ such that both of the following bounds hold on $\mathcal{E}_{\mathrm{mart}}$:*

$$\sum_{k=1}^{K} \mathbb{E}[Y_k \mid \mathcal{F}_1^k] \le 2(B_K + \widetilde{B}_K) + \mathcal{O}(H^2\iota), \tag{C.26}$$

$$\sum_{k=1}^{K} A_k \le 2K\mathrm{Var}^\star + \mathcal{O}(H^3\iota). \tag{C.27}$$

*On $\mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_{\mathrm{mart}}$, the value-error variances also satisfy*

$$Z_D + Z_E \le \mathcal{O}\left( HB_K + H^2 \sum_{k=1}^{K} \mathbf{1}\{\tau_k \le H\} + H^2\iota \right), \tag{C.28}$$

*and (C.23) holds for the steps $h \ge \tau_k$.*

*Proof.* For the expected residual sum, apply (C.2) to $Y_k$ with $B = H^2$ and failure probability $\delta/16$, and let $\mathcal{E}_Y$ be the resulting event. Since $\sum_k Y_k = B_K + \widetilde{B}_K$, (C.26) holds on $\mathcal{E}_Y$, with $\mathbb{P}(\mathcal{E}_Y) \ge 1 - \delta/16$.

To control the safe-policy variance, condition on $\mathcal{F}_1^k$ and couple $\pi^k$ with $\bar{\pi}^k$ using the same transition randomness. Both policies are fixed given this history. Their states and actions agree for $h < \tau_k$. Since the summands in (C.25) are nonnegative, $A_k$ is at most the full variance sum along the trajectory of $\bar{\pi}^k$. Hence

$$\mathbb{E}[A_k \mid \mathcal{F}_1^k] \le \mathbb{E}^{\bar{\pi}^k}\left[\sum_{h=1}^{H}\left\{\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_h(\cdot \mid s_h,a_h), V_{h+1}^{\bar{\pi}^k})\right\} \,\middle|\, s_1\right] \le \mathrm{Var}^\star.$$

Iterating the Bellman variance decomposition identifies the middle expectation with the return variance of $\bar{\pi}^k$. The decomposition uses the conditional independence of the reward and successor state. Each reward variance is at most $1/4$, and each step-$h$ next-state value variance is at most $(H-h)^2/4$. Thus $0 \le A_k \le \frac{1}{4}(H + \sum_{j=0}^{H-1} j^2) \le H^3/3$. Apply (C.1) with $B = H^3/3$ and failure probability $\delta/16$, and let $\mathcal{E}_A$ be the resulting event. On $\mathcal{E}_A$,

$$\sum_{k=1}^{K} A_k \le 2\sum_{k=1}^{K} \mathbb{E}[A_k \mid \mathcal{F}_1^k] + \mathcal{O}(H^3\iota) \le 2K\mathrm{Var}^\star + \mathcal{O}(H^3\iota).$$

Thus $\mathbb{P}(\mathcal{E}_A) \ge 1 - \delta/16$.

For the value-error variances, first consider $[D_h^k]_+$ and $[E_h^k]_+$. These functions are $\mathcal{F}_1^k$-measurable, lie in $[0,H]$ on every sample path, and vanish at step $H+1$. The drift relations in Lemma C.6, together with $[x]_+ - [y]_+ \le [x-y]_+$ and $P_h[g_{h+1}]_+ \ge [P_h g_{h+1}]_+$, give, for $h < \tau_k$,

$$[D_h^k(s_h^k)]_+ - P_h[D_{h+1}^k]_+(s_h^k,a_h^k) \le [f_h^k]_+,$$
$$[E_h^k(s_h^k)]_+ - P_h[E_{h+1}^k]_+(s_h^k,a_h^k) \le [f_h^k]_+.$$

Since $0 \le [f_h^k]_+ \le H$, the assumptions of Lemma C.8 hold with $u_h^k = [f_h^k]_+$ on every sample path. Apply Lemma C.8 with $g_h^k = [D_h^k]_+$ and $g_h^k = [E_h^k]_+$, using this $u_h^k$ and $\eta = \delta/16$. Let $\mathcal{E}_D$ and $\mathcal{E}_E$ be the respective events. Each has probability at least $1 - \delta/16$. On $\mathcal{E}_{\mathrm{est}}$, Lemma C.6 gives $D^k, E^k \ge 0$, and Lemma C.5 gives $[f_h^k]_+ \le \min\{\beta_h^k(s_h^k,a_h^k), H\}$. Hence (C.28) holds on $\mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_D \cap \mathcal{E}_E$.

Set $\mathcal{E}_{\mathrm{mart}} := \mathcal{E}_Y \cap \mathcal{E}_A \cap \mathcal{E}_D \cap \mathcal{E}_E \cap \mathcal{E}_{\mathrm{suf}}$, where $\mathcal{E}_{\mathrm{suf}}$ is supplied by Lemma C.9. A union bound gives $\mathbb{P}(\mathcal{E}_{\mathrm{mart}}) \ge 1 - 5\delta/16$. □

### C.6 Proof of Theorem 2

The preceding bounds reduce the proof to controlling $B_K$ and $\widetilde{B}_K$. We bound $B_K$ using each of the two variance quantities and take the smaller bound. The number of episodes containing an unsafe action controls the remaining sum $\widetilde{B}_K$.

Define $\mathcal{E}_{\mathrm{main}} := \mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_{\mathrm{cert}} \cap \mathcal{E}_{\mathrm{mart}}$. Lemmas B.2, C.3, and C.10 give

$$\mathbb{P}(\mathcal{E}_{\mathrm{main}}) \ge 1 - \frac{9+1+5}{16}\delta \ge 1 - \delta. \tag{C.29}$$

We work on $\mathcal{E}_{\mathrm{main}}$ throughout this subsection.

Proposition C.2 and (C.26) give

$$\mathcal{R}(K) \le 2(B_K + \widetilde{B}_K) + \mathcal{O}(H^2\iota). \tag{C.30}$$

**Bounding $B_K$** Combining (C.28) with (C.7) yields

$$Z_D + Z_E \le \mathcal{O}\big(HB_K + H^2\iota\,[\min\{K,\mathcal{C}\}+1]\big). \tag{C.31}$$

For the steps $h < \tau_k$, $V_{h+1}^\star = V_{h+1}^{\bar{\pi}^k} + E_{h+1}^k - D_{h+1}^k$. The successor belongs to $\mathcal{V}_{\mathrm{safe},h+1}$, so this identity holds on the support of the executed transition. Applying $\mathbb{V}(X+Y+Z) \le 3\{\mathbb{V}(X) + \mathbb{V}(Y) + \mathbb{V}(Z)\}$ and including the nonnegative reward variances gives

$$\mathbb{V}_\Sigma \le 3\left(\sum_{k=1}^{K} A_k + Z_D + Z_E\right). \tag{C.32}$$

Lemma C.7 and (C.12) give

$$B_K \le \mathcal{O}\Big(\sqrt{SAH\iota\,\mathbb{V}_\Sigma} + \sqrt{S^2AH\iota L_K Z_D} + S^2AH^2\iota L_K\Big). \tag{C.33}$$

For the bound in terms of $\mathbb{V}_\Sigma$, substitute (C.31) into (C.33) to obtain

$$B_K \le \mathcal{O}\Big(\sqrt{SAH\iota\,\mathbb{V}_\Sigma} + S\sqrt{AH^2\iota L_K B_K} + S\sqrt{AH^3\min\{K,\mathcal{C}\}}\,\iota\sqrt{L_K} + S^2AH^2\iota L_K\Big).$$

Young's inequality bounds the term containing $\sqrt{B_K}$ by $B_K/2 + \mathcal{O}(S^2AH^2\iota L_K)$. Moving $B_K/2$ to the left yields

$$B_K \le \mathcal{O}\Big(\sqrt{SAH\iota\,\mathbb{V}_\Sigma} + S\sqrt{AH^3\min\{K,\mathcal{C}\}}\,\iota\sqrt{L_K} + S^2AH^2\iota L_K\Big). \tag{C.34}$$

For the bound in terms of $\mathrm{Var}^\star$, (C.32) and (C.27) give

$$\sqrt{SAH\iota\,\mathbb{V}_\Sigma} \le \mathcal{O}\Big(\sqrt{SAHK\mathrm{Var}^\star\,\iota} + \sqrt{SAH\iota(Z_D+Z_E)} + H^2\iota\sqrt{SA}\Big).$$

Substitute this inequality into (C.33) and use (C.31) for $Z_D + Z_E$. Since $S, L_K \ge 1$, the resulting terms involving $B_K$ and $\min\{K,\mathcal{C}\}$ are bounded by the corresponding terms above. Also, $H^2\iota\sqrt{SA} \le S^2AH^2\iota L_K$. Applying the same absorption step gives

$$B_K \le \mathcal{O}\Big(\sqrt{SAHK\mathrm{Var}^\star\,\iota} + S\sqrt{AH^3\min\{K,\mathcal{C}\}}\,\iota\sqrt{L_K} + S^2AH^2\iota L_K\Big). \tag{C.35}$$

Both bounds hold on $\mathcal{E}_{\mathrm{main}}$. Taking their minimum gives

$$B_K \le \mathcal{O}\Big(\sqrt{SAH\iota\,\min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}} + S\sqrt{AH^3\min\{K,\mathcal{C}\}}\,\iota\sqrt{L_K} + S^2AH^2\iota L_K\Big). \tag{C.36}$$

**Bounding $\widetilde{B}_K$** The remaining steps occur only in episodes with $\tau_k \le H$. Since each reward variance is at most $1/4$,

$$\sum_{k=1}^{K}\sum_{h=\tau_k}^{H}\mathbb{V}(R_h(s_h^k,a_h^k)) \le \frac{H}{4}\sum_{k=1}^{K}\mathbf{1}\{\tau_k \le H\}.$$

Apply Lemma C.7 to (C.13) over the visits with $h \ge \tau_k$:

$$\widetilde{B}_K \le \mathcal{O}\Bigg(\sqrt{SAH\iota\sum_{k=1}^{K}\sum_{h=\tau_k}^{H}\mathbb{V}(R_h(s_h^k,a_h^k))} + S\sqrt{AH\iota L_K\sum_{k=1}^{K}\sum_{h=\tau_k}^{H}\mathbb{V}(P_h(\cdot\mid s_h^k,a_h^k),V_{h+1}^k)} + S^2AH^2\iota L_K\Bigg).$$

The reward-variance bound above and (C.23) then give, using $S, A, H, L_K \ge 1$,

$$\widetilde{B}_K \le \mathcal{O}\Big(S\sqrt{AH^3\iota L_K\sum_{k=1}^{K}\mathbf{1}\{\tau_k \le H\}} + S\sqrt{AH^2\iota L_K\,\widetilde{B}_K} + S^2AH^2\iota L_K\Big).$$

Young's inequality bounds the term containing $\sqrt{\widetilde{B}_K}$ by $\widetilde{B}_K/2 + \mathcal{O}(S^2AH^2\iota L_K)$. Move $\widetilde{B}_K/2$ to the left and use (C.7) to obtain

$$\widetilde{B}_K \le \mathcal{O}\Big(S\sqrt{AH^3\min\{K,\mathcal{C}\}}\,\iota\sqrt{L_K} + S^2AH^2\iota L_K\Big). \tag{C.37}$$

*Proof of Theorem 2.* Substitute (C.36) and (C.37) into (C.30). Since $H^2\iota \le S^2AH^2\iota L_K$, this gives

$$\mathcal{R}(K) \le \mathcal{O}\Big(\sqrt{SAH\iota\,\min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}} + S\sqrt{AH^3\min\{K,\mathcal{C}\}}\,\iota\sqrt{L_K} + S^2AH^2\iota L_K\Big). \tag{C.38}$$

The probability statement follows from (C.29). Suppressing the logarithmic factors in $\iota$ and $L_K$ proves Theorem 2. □

### C.7 A Special Case with Direct Cost Violations

Suppose Assumptions 1 and 2 hold, and every unsafe action at a vertex reachable under a safe policy has cost exceeding $\tau$. Then $\mathcal{C} \le SAH/\Delta_c^2$, and SVAE satisfies, with probability at least $1-\delta$,

$$\mathcal{R}(K) \le \widetilde{\mathcal{O}}\Big(\sqrt{SAH\min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}} + S\sqrt{AH^3\min\{K, SAH/\Delta_c^2\}} + S^2AH^2\Big). \quad \text{(C.39)}$$

*Proof.* For every action contributing to $\mathcal{C}$, the direct-cost branch of (7) gives $\kappa_h(s,a) = (c_h(s,a)-\tau)^{-2} \le \Delta_c^{-2}$. There are at most $SAH$ such actions, so $\mathcal{C} \le SAH/\Delta_c^2$. Substitution into Theorem 2 yields (C.39) on the same event. □

## D Constraint Violation

We prove Theorem 3 on $\mathcal{E}_{\mathrm{cs}} = \mathcal{E}_c \cap \mathcal{E}_{\mathrm{supp}}$, which has probability at least $1-\delta/8$ by Lemma B.2. We separate steps with nonempty candidate sets from those with empty candidate sets. The cost bound controls the first contribution, and newly observed transitions control the second.

Let nemp and emp denote nonempty and empty candidate sets, respectively, and define

$$\mathrm{Viol}_{\mathrm{nemp}}(K) := \sum_{k=1}^{K}\sum_{h=1}^{H}\big[c_h(s_h^k,a_h^k)-\tau\big]_+ \mathbf{1}\Big\{\hat{\mathcal{A}}^k_{\mathrm{safe},h}(s_h^k) \neq \varnothing\Big\}, \quad \text{(D.1)}$$

$$\mathrm{Viol}_{\mathrm{emp}}(K) := \sum_{k=1}^{K}\sum_{h=1}^{H}\big[c_h(s_h^k,a_h^k)-\tau\big]_+ \mathbf{1}\Big\{\hat{\mathcal{A}}^k_{\mathrm{safe},h}(s_h^k) = \varnothing\Big\}.$$

Then, deterministically,

$$\mathrm{Viol}(K) = \mathrm{Viol}_{\mathrm{nemp}}(K) + \mathrm{Viol}_{\mathrm{emp}}(K). \quad \text{(D.2)}$$

**Nonempty candidate sets** We first bound $\mathrm{Viol}_{\mathrm{nemp}}(K)$. Every selected candidate action passes the cost test. This controls both its one-step violation and the number of times it can be selected with a positive cost gap.

**Lemma D.1.** *On $\mathcal{E}_c$, both $\mathrm{Viol}_{\mathrm{nemp}}(K) \le 4\sqrt{2}\,H\sqrt{SAK\iota}$ and the following gap-dependent bound hold:*

$$\mathrm{Viol}_{\mathrm{nemp}}(K) \le \sum_{c_h(s,a)>\tau}\left\{c_h(s,a)-\tau+\frac{8\iota}{c_h(s,a)-\tau}\right\}. \quad \text{(D.3)}$$

*The sum ranges over $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ with a positive cost gap.*

*Proof.* When the candidate set is nonempty, Algorithm 1 selects $a \in \hat{\mathcal{A}}^k_{\mathrm{safe},h}(s)$, so $\underline{c}_h^k(s,a) \le \tau$. On $\mathcal{E}_c$, this implies

$$c_h(s,a)-\tau \le c_h(s,a) - \widehat{c}_h^k(s,a) + \sqrt{\frac{2\iota}{n_h^k(s,a)}} \le 2\sqrt{\frac{2\iota}{n_h^k(s,a)}}.$$

Since $n_h^k = \max\{N_h^k, 1\}$, this also covers the zero-count case. Hence

$$[c_h(s,a)-\tau]_+ \le 2\sqrt{\frac{2\iota}{n_h^k(s,a)}}. \quad \text{(D.4)}$$

Let $M_h(s,a)$ count executions of $(s,a)$ at step $h$ with a nonempty candidate set. The reciprocal-count bound in Lemma C.7 applies to this subset of visits. Summing (D.4) and using $\sum_{s,a,h} M_h(s,a) \le KH$ gives

$$\begin{aligned}\mathrm{Viol}_{\mathrm{nemp}}(K) &\le 4\sqrt{2\iota}\sum_{s,a,h}\sqrt{M_h(s,a)} \\ &\le 4\sqrt{2SAH\iota\sum_{s,a,h}M_h(s,a)} \\ &\le 4\sqrt{2}\,H\sqrt{SAK\iota}.\end{aligned}$$

For the second bound, fix $(s, a, h)$ with $\Delta := c_h(s, a) - \tau > 0$. If $M_h(s, a) \geq 2$, the last counted execution has at least $M_h(s, a) - 1$ prior visits. Applying (D.4) there gives $\Delta \leq 2\sqrt{2\iota/(M_h(s, a) - 1)}$. Therefore, including the cases $M_h(s, a) \leq 1$,

$$M_h(s, a) \leq 1 + \frac{8\iota}{\Delta^2}. \tag{D.5}$$

Each counted execution contributes $\Delta$. Multiplying by $\Delta$ and summing over positive-gap triples proves (D.3). □

**Empty candidate sets** It remains to bound $\mathrm{Viol}_{\mathrm{emp}}(K)$. An empty candidate set can follow a candidate action only through a newly observed transition. We associate each block of empty candidate sets with this transition and bound its length by the remaining horizon.

**Lemma D.2.** *Suppose Assumption 1 holds. Then, on $\mathcal{E}_{\mathrm{cs}}$, $\mathrm{Viol}_{\mathrm{emp}}(K) \leq S^2AH(H-1)/2$.*

*Proof.* In each episode $k$, partition the steps with empty candidate sets into maximal contiguous blocks. A block $\{\ell, \ldots, r\}$ has $\ell \geq 2$: Assumption 1 and the inclusions proved in Appendix B.4 guarantee a nonempty candidate set at $s_1$. By maximality, the preceding action is a candidate, so

$$\widehat{\mathrm{supp}}^k_{\ell-1}(s^k_{\ell-1}, a^k_{\ell-1}) \subseteq \hat{\mathcal{V}}^k_{\mathrm{safe},\ell}. \tag{D.6}$$

The state $s^k_\ell$ has no candidate actions. Consequently,

$$s^k_\ell \notin \hat{\mathcal{V}}^k_{\mathrm{safe},\ell}, \qquad s^k_\ell \notin \widehat{\mathrm{supp}}^k_{\ell-1}(s^k_{\ell-1}, a^k_{\ell-1}). \tag{D.7}$$

Thus $(s^k_{\ell-1}, a^k_{\ell-1}, s^k_\ell, \ell - 1)$ is a newly observed transition. Associate the block with this transition.

Each transition is associated with at most one block. A given layer is visited once per episode. In later episodes, the successor remains in the observed support, so a candidate predecessor action forces it to have a nonempty candidate set by (D.6). It therefore cannot start another empty block through that action. There are at most $S^2A$ transitions at each parent layer $h$, and an associated block has length at most $H - h$. Since violation per step is at most one,

$$\mathrm{Viol}_{\mathrm{emp}}(K) \leq S^2A \sum_{h=1}^{H-1} (H - h) = \frac{S^2AH(H-1)}{2}. \qquad \square$$

*Proof of Theorem 3.* Substituting Lemmas D.1 and D.2 into (D.2) gives, on $\mathcal{E}_{\mathrm{cs}}$,

$$\mathrm{Viol}(K) \leq 4\sqrt{2}\, H\sqrt{SAK\iota} + \frac{S^2AH(H-1)}{2}.$$

If Assumption 2 also holds, every positive cost gap is at least $\Delta_c$. The second bound in Lemma D.1 then gives, on the same event,

$$\mathrm{Viol}(K) \leq SAH\left(1 + \frac{8\iota}{\Delta_c}\right) + \frac{S^2AH(H-1)}{2}.$$

The event has probability at least $1 - \delta/8$. Suppressing logarithmic factors proves both claims. □

## E Cost Confidence Bounds at Zero Empirical Variance

A sample with zero empirical variance may come from a cost distribution whose mean exceeds the safety threshold. We compare two distributions that can produce the same constant sample to obtain a necessary additive term in the confidence bound. We then apply this comparison to the lower confidence test used for action removal in SVAE.

Fix $\tau \in (0, 1)$ and $\delta \in (0, 1/2)$. A certification rule observes independent samples $X_1, X_2, \ldots$ from a common distribution $\nu$ on $[0, 1]$, may use independent auxiliary randomness, and returns a stopping time $T \in \mathbb{N} \cup \{\infty\}$ at which it declares that the mean is at most $\tau$. Here $\mathbb{N} = \{1, 2, \ldots\}$, and $\{T \leq n\}$ depends only on the first $n$ observations and the auxiliary randomness. Call the rule

$\delta$-safe if $\mathbb{P}_\nu(T < \infty) \le \delta$ for every such distribution with $\mathbb{E}_\nu[X] > \tau$. The probability includes both sources of randomness.

Set $\Delta_{\max} := \min\{\tau, 1-\tau\}$. For each $\Delta \in (0, \Delta_{\max})$, let $p_\Delta := 2\Delta/(1-\tau+\Delta)$. Under $\nu_0$, each sample equals $\tau - \Delta$ with probability one. Under $\nu_1$, it equals 1 with probability $p_\Delta$ and $\tau - \Delta$ otherwise. Both distributions are supported on $[0,1]$, with $\mathbb{E}_{\nu_0}[X] = \tau - \Delta$ and $\mathbb{E}_{\nu_1}[X] = \tau + \Delta$.

**Proposition E.1.** *Fix $\Delta \in (0, \Delta_{\max})$ and let $\nu_0, \nu_1$ be the distributions defined above. For every $\delta$-safe certification rule $T$ and every $n \in \mathbb{N}$, we have*

$$\mathbb{P}_{\nu_0}(T \le n) \le \frac{\delta}{(1-p_\Delta)^n}. \tag{E.1}$$

*Proof.* Since $p_\Delta \in (0,1)$, under $\nu_1$ the first $n$ observations all equal $\tau - \Delta$ with probability $(1-p_\Delta)^n$. On that event, the rule sees exactly the observations under $\nu_0$, and its auxiliary randomness has the same distribution. Thus

$$\delta \ge \mathbb{P}_{\nu_1}(T \le n) \ge (1-p_\Delta)^n \mathbb{P}_{\nu_0}(T \le n),$$

which proves (E.1). □

We now apply this comparison to confidence bounds based on empirical variance. For $t \in \mathbb{N}$, define $\overline{X}_t := t^{-1}\sum_{i=1}^t X_i$ and $\widehat{v}_t := t^{-1}\sum_{i=1}^t (X_i - \overline{X}_t)^2$. Let $(\rho_t)_{t\in\mathbb{N}}$ be deterministic functions from $[0,\infty) \times (0, 1/2)$ to $[0,\infty)$, and define the stopping rule

$$T := \inf\left\{t \in \mathbb{N} : \overline{X}_t + \rho_t(\widehat{v}_t, \delta) \le \tau\right\}, \tag{E.2}$$

with the convention $\inf \varnothing = +\infty$.

**Corollary E.1.** *Suppose the stopping rule in (E.2) is $\delta$-safe. Then, for every $n \in \mathbb{N}$,*

$$\rho_n(0,\delta) \ge \min\left\{\Delta_{\max}, \frac{(1-\tau)(1-\delta^{1/n})}{1+\delta^{1/n}}\right\}. \tag{E.3}$$

*In particular, for every $n \in \mathbb{N}$ satisfying $\log(1/\delta)/n \le \Delta_{\max}$, $\rho_n(0,\delta) \ge \frac{(1-\tau)\log(1/\delta)}{4n}$.*

*Proof.* Suppose (E.3) fails at some $n$, and choose

$$\rho_n(0,\delta) < \Delta < \min\left\{\Delta_{\max}, \frac{(1-\tau)(1-\delta^{1/n})}{1+\delta^{1/n}}\right\}.$$

Under $\nu_0$, all observations equal $\tau - \Delta$, so $\overline{X}_n = \tau - \Delta$, $\widehat{v}_n = 0$, and $T \le n$ almost surely. Proposition E.1 gives $(1-p_\Delta)^n \le \delta$. However, the chosen upper bound on $\Delta$ is equivalent to $p_\Delta < 1 - \delta^{1/n}$, a contradiction.

For the second claim, set $x = \log(1/\delta)/n \le \Delta_{\max} \le 1/2$. The inequalities $1 - e^{-x} \ge x/2$ and $1 + e^{-x} \le 2$ give $(1-\tau)(1-e^{-x})/(1+e^{-x}) \ge (1-\tau)x/4$. Also, $\Delta_{\max} \ge (1-\tau)x/4$. Substituting into (E.3) proves the claim. □

For the lower confidence test, consider $T = \inf\{t \in \mathbb{N} : \overline{X}_t - \rho_t(\widehat{v}_t, \delta) > \tau\}$ and suppose $\mathbb{P}_\nu(T < \infty) \le \delta$ for every distribution on $[0,1]$ with $\mathbb{E}_\nu[X] \le \tau$. Reflect the observations to $1 - X$ and the threshold to $1 - \tau$. The empirical variance is unchanged. The same constant-sample argument applies, since $\rho_n(0,\delta) < \Delta$ ensures that the strict test fires. It gives

$$\rho_n(0,\delta) \ge \min\left\{\Delta_{\max}, \frac{\tau(1-\delta^{1/n})}{1+\delta^{1/n}}\right\}. \tag{E.4}$$

In particular, $\rho_n(0,\delta) \ge \tau\log(1/\delta)/(4n)$ under the same conditions on $n$ as in the corollary. Thus uniformly valid action removal requires a positive radius at zero empirical variance.

## F Safe-Subgraph Variance

Throughout this section, Assumption 1 holds. We use the spans $S^\pi$ and $S^\star$ from Section 3.2 to bound the transition variances of safe-policy values and optimal safe values, respectively. The branching depth bounds the number of nonzero terms along a safe path. We then apply the resulting variance bounds to regret and provide matching cost constructions showing that the bounds are tight.

### F.1 Proof of Theorem 1

*Proof.* Fix $\pi \in \Pi_{\text{safe}}$. Its trajectory is almost surely a safe path in $\mathcal{G}_{\text{safe}}$, and every visited vertex belongs to $\mathcal{V}^\pi$. For a distribution $q$ and a bounded function $f$, let $m := \min_{x \in \text{supp}(q)} f(x)$ and $M := \max_{x \in \text{supp}(q)} f(x)$. Since $(f - m)(M - f) \geq 0$ on $\text{supp}(q)$,

$$\mathbb{V}(q, f) \leq (M - qf)(qf - m) \leq \frac{1}{4}(M - m)^2. \tag{F.1}$$

Apply (F.1) with $q = P_h(\cdot \mid s_h, a_h)$ and $f = V^\pi_{h+1}$. A deterministic transition has zero variance. Otherwise, the definition of $S^\pi$ bounds the successor-value span by $S^\pi$. Hence, almost surely,

$$\mathbb{V}(P_h(\cdot \mid s_h, a_h), V^\pi_{h+1}) \leq \frac{1}{4}(S^\pi)^2 \mathbf{1}\{|\text{supp}(P_h(\cdot \mid s_h, a_h))| > 1\}. \tag{F.2}$$

A safe path contains at most $B(\mathcal{G}_{\text{safe}})$ stochastic transitions. Summing (F.2) along the path and taking expectation gives

$$\mathbb{E}^\pi \left[ \sum_{h=1}^{H} \mathbb{V}(P_h(\cdot \mid s_h, a_h), V^\pi_{h+1}) \,\middle|\, s_1 \right] \leq \frac{1}{4}(S^\pi)^2 B(\mathcal{G}_{\text{safe}}). \tag{F.3}$$

For the second claim, fix a safe path $\gamma = (s_1, a_1, \ldots, s_H, a_H, s_{H+1})$ and a step $h \in [H]$. All successors of $a_h$ are viable, so $V^\star_{h+1}$ gives their optimal safe continuation values by Lemma B.1. Its span on $\text{supp}(P_h(\cdot \mid s_h, a_h))$ is at most $S^\star$ by definition. A deterministic transition has zero variance. Thus (F.1) gives

$$\mathbb{V}(P_h(\cdot \mid s_h, a_h), V^\star_{h+1}) \leq \frac{1}{4}(S^\star)^2 \mathbf{1}\{|\text{supp}(P_h(\cdot \mid s_h, a_h))| > 1\}. \tag{F.4}$$

The steps before $\tau_k$ can be completed to a full safe path because all successors of a safe action are viable. Each episode therefore contains at most $B(\mathcal{G}_{\text{safe}})$ stochastic transitions before $\tau_k$. Summing (F.4) gives

$$\sum_{k=1}^{K} \sum_{h < \tau_k} \mathbb{V}(P_h(\cdot \mid s^k_h, a^k_h), V^\star_{h+1}) \leq \frac{K}{4}(S^\star)^2 B(\mathcal{G}_{\text{safe}}), \tag{F.5}$$

which proves the second claim. □

### F.2 Proof of Corollary 1

*Proof.* For deterministic rewards, the reward-variance terms in (5) vanish, so (F.5) gives $\mathbb{V}_\Sigma \leq \frac{K}{4}(S^\star)^2 B(\mathcal{G}_{\text{safe}})$. The total reward of every safe policy lies in $[0, H]$, so its variance is at most $H^2/4$. Hence $\text{Var}^\star \leq H^2/4$ by (4). Combining these two bounds yields

$$\min\{\mathbb{V}_\Sigma, K\text{Var}^\star\} \leq \frac{K}{4} \min\{H^2, (S^\star)^2 B(\mathcal{G}_{\text{safe}})\}.$$

Substituting this inequality into Theorem 2 proves the corollary on the same event, with probability at least $1 - \delta$. □

### F.3 Proof of Proposition 1

To attain the variance bound, we keep the transition kernel and rewards fixed and vary the costs that determine which gadget transitions are safe. We first give the construction, then compute its branching depth, successor-value span, and return variance. Figure 2 shows the construction for $n = 3$ with zero, one, two, and three open gadgets.

**Example 1** (Gadget chain)**.** *Fix an integer $n \geq 1$ and set $H = 2n$. Gadget $i \in [n]$ starts at $s_i$ in layer $h_i := 2i - 1$, and $s_{n+1} := t$ is the terminal state. The transition kernel and deterministic rewards are fixed as follows.*

*At $s_i$, action $a_{\text{det},i}$ transitions deterministically to $s'_{A,i}$, while $a_{\text{st},i}$ transitions uniformly to $s'_{A,i}$ and $s'_{B,i}$. Both actions have reward zero. At the next layer, action $\alpha_i$ at $s'_{A,i}$ has reward one, and actions*

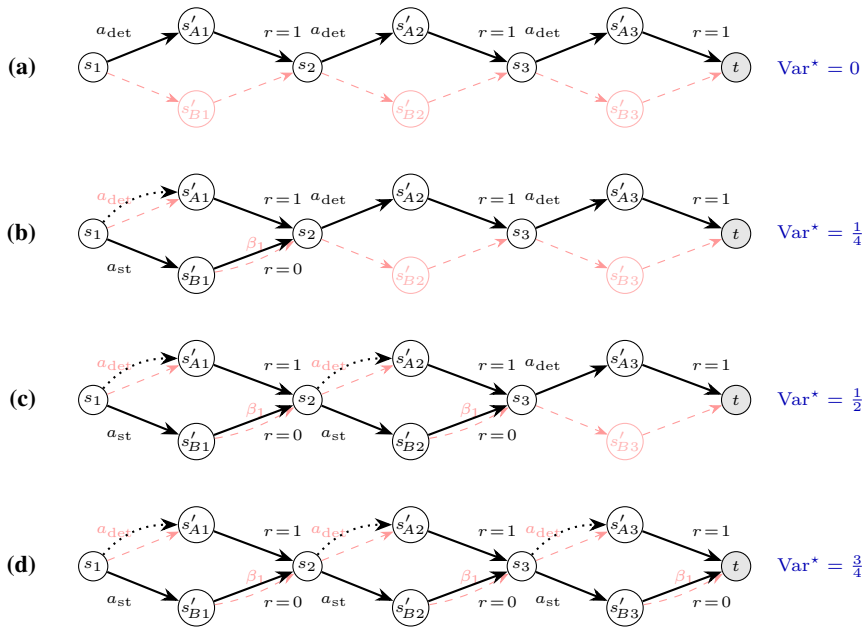



Figure 2: Four safe subgraph topologies on the same $H = 6$ gadget-chain MDP (Example 1). Each row shows the safe subgraph under a different cost function. Black solid edges are safe. Red dashed edges are unsafe. In a *closed* gadget, both actions at $s'_{B,i}$ are unsafe, so $a_{\mathrm{st},i}$ fails the successor-viability condition and only $a_{\mathrm{det},i}$ remains. In an *open* gadget, $a_{\mathrm{det},i}$ and $\beta_{1,i}$ are unsafe, leaving a successor-value gap of one by (F.10). Dotted black edges show the branch of $a_{\mathrm{st}}$ to $s'_{A,i}$. Gadget indices are omitted from the action labels.

$\beta_{1,i}, \beta_{2,i}$ *at* $s'_{B,i}$ *have rewards one and zero, respectively. All three actions transition deterministically to* $s_{i+1}$.

*Take* $\tau = 1/2$, *set* $c_{h_i+1}(s'_{A,i}, \alpha_i) = 0$, *and define*

$$\big(c_{h_i}(s_i, a_{\mathrm{det},i}), c_{h_i}(s_i, a_{\mathrm{st},i}), c_{h_i+1}(s'_{B,i}, \beta_{1,i}), c_{h_i+1}(s'_{B,i}, \beta_{2,i})\big) = \begin{cases} (0,0,1,1), & \text{if gadget } i \text{ is closed}, \\ (1,0,1,0), & \text{if gadget } i \text{ is open}. \end{cases} \tag{F.6}$$

*Every other state–action pair has cost one, reward zero, and a fixed deterministic transition to the next layer. For a cost configuration* $c$, *let* $b(c) \in \{0, \dots, n\}$ *denote the number of open gadgets.*

*Proof of Proposition 1.* A backward induction from $s_{n+1} = t$ using (F.6) gives, for every $i \in [n]$,

$$\mathcal{A}_{\mathrm{safe},h_i}(s_i) = \begin{cases} \{a_{\mathrm{st},i}\}, & \text{if gadget } i \text{ is open}, \\ \{a_{\mathrm{det},i}\}, & \text{if gadget } i \text{ is closed}, \end{cases} \tag{F.7}$$

and

$$\mathcal{A}_{\mathrm{safe},h_i+1}(s'_{A,i}) = \{\alpha_i\}, \qquad \mathcal{A}_{\mathrm{safe},h_i+1}(s'_{B,i}) = \begin{cases} \{\beta_{2,i}\}, & \text{if gadget } i \text{ is open}, \\ \varnothing, & \text{if gadget } i \text{ is closed}. \end{cases} \tag{F.8}$$

In a closed gadget, the state $s'_{B,i}$ is not viable. The action $a_{\mathrm{st},i}$ satisfies the immediate cost constraint but fails the successor-viability condition. In an open gadget, both successors of $a_{\mathrm{st},i}$ are viable, while $a_{\mathrm{det},i}$ and $\beta_{1,i}$ violate the cost threshold.

It follows from (F.7)–(F.8) that all policies in $\Pi_{\mathrm{safe}}$ select the same action at every state and step reachable from $s_1$ under a safe policy. Since both reward-layer branches merge at the next choice state,

$$\mathbb{P}^{\pi}(s_{h_i} = s_i \mid s_1) = 1, \qquad i \in [n]. \tag{F.9}$$

The stochastic choice $(s_i, a_{\mathrm{st},i}, h_i)$ is safe exactly when gadget $i$ is open. These are the only stochastic safe transitions, and every safe path visits all $n$ choice states. Hence $B(\mathcal{G}_{\mathrm{safe}}) = b(c)$.

We next compute the successor-value spans and variances. If gadget $i$ is closed, its safe choice action and its reward-layer transition are deterministic, so its next-state value variance is zero. If gadget $i$ is open, then

$$V^{\pi}_{h_i+1}(s'_{A,i}) = 1 + V^{\pi}_{h_i+2}(s_{i+1}), \qquad V^{\pi}_{h_i+1}(s'_{B,i}) = V^{\pi}_{h_i+2}(s_{i+1}). \tag{F.10}$$

The two successor values differ by one. All deterministic transitions have successor-value span zero, so $S^{\pi} = \mathbf{1}\{b(c) > 0\}$. At each open gadget,

$$\mathbb{V}\big(P_{h_i}(\cdot \mid s_i, a_{\mathrm{st},i}), V^{\pi}_{h_i+1}\big) = \frac{1}{4}\left(V^{\pi}_{h_i+1}(s'_{A,i}) - V^{\pi}_{h_i+1}(s'_{B,i})\right)^2 = \frac{1}{4}.$$

All rewards are deterministic. Using (F.9) and summing these next-state variances over the open gadgets gives

$$\mathbb{E}^{\pi}\left[\sum_{h=1}^{H} \mathbb{V}(P_h(\cdot \mid s_h, a_h), V^{\pi}_{h+1}) \,\middle|\, s_1\right] = \sum_{i:\ \text{gadget } i \text{ is open}} \frac{1}{4} = \frac{b(c)}{4}.$$

These identities hold for every $\pi \in \Pi_{\mathrm{safe}}$. Taking the supremum over this class gives $\mathrm{Var}^{\star} = b(c)/4$. The unique safe actions in (F.7)–(F.8) imply that these successor values coincide with $V^{\star}$. Thus $S^{\star} = \mathbf{1}\{b(c) > 0\}$ by definition. Substitution into (F.3) proves equality. For every $b \in \{0, \dots, n\}$, choosing exactly $b$ open gadgets gives the instance required by Proposition 1. □

# G Lower Bounds

Appendix G.1 proves the regret lower bound in Theorem 4. Appendix G.2 studies the sample requirements for action elimination under additional conditions on the transition structure and observations.

## G.1 Proof of Theorem 4

Assume the parameter conditions of Theorem 4. We use two families of CMDPs with the prescribed variance and safety identification complexity. The first requires learning which action has the better transition probability. The second requires identifying a safe action at each of several state–step pairs. We prove a regret lower bound for each family and then choose the instance giving the larger bound. Learning algorithms follow the protocol in Section 2, with their policies chosen before each episode. For a fixed instance $\mathcal{M}$, the expectation $\mathbb{E}_{\mathcal{M}}$ includes the algorithm's randomness and its observations.

**Common construction** Set $n := S - 3$ and use the states $\{x, g, b, c_1, \dots, c_n\}$, action set $[A]$, initial state $x$, and threshold $\tau = 1/2$. Use $d := \lfloor H/16 \rfloor$ three-step gadgets in the first $3d$ steps, each of which starts at $x$ and returns to $x$. The two families differ in these gadgets and in the costs and transitions at a set of state–step pairs after them. The last $R := \lfloor H/8 \rfloor$ steps provide the reward lost by an incorrect choice in the second family. Since $4d + R + 1 \le H$, these parts fit within the horizon.

Let $M := \lfloor \Gamma/[8(A-1)] \rfloor$. The parameter assumptions give $\Gamma/16 \le M(A-1) \le \Gamma/8$ and $2 \le M \le nd$, so we can choose $\mathcal{J} \subseteq \{3d+2, \dots, 4d+1\} \times [n]$ with $|\mathcal{J}| = M$. Here $(h, s) \in \mathcal{J}$ denotes context state $c_s$ at step $h$. We specify the transitions so that exactly one pair in $\mathcal{J}$ is reached in each episode, uniformly over the $M$ pairs. For $h \in \{3d+2, \dots, 4d+1\}$, let $M_h := |\{(t, s) \in \mathcal{J} : t \ge h\}|$ and set $M_{4d+2} := 0$. Whenever $M_h > 0$, every action at $x$ at step $h-1$ satisfies

$$P_{h-1}(c_s \mid x, a) = \frac{\mathbf{1}\{(h, s) \in \mathcal{J}\}}{M_h}, \qquad P_{h-1}(x \mid x, a) = \frac{M_{h+1}}{M_h}.$$

Every action at a pair in $\mathcal{J}$ moves to $g$ or $b$, as specified below, and the process then remains in that state until the episode ends. Since $\mu^{\pi}_{3d+1}(x) = 1 = M_{3d+2}/M$, induction gives

$$\mu^{\pi}_{h-1}(x) = \frac{M_h}{M}, \qquad \mu^{\pi}_h(c_s) = \frac{1}{M}, \quad (h, s) \in \mathcal{J}. \tag{G.1}$$

Thus the reached pair is uniform on $\mathcal{J}$, conditional on the past and independent of the chosen actions. All rewards and costs are deterministic. Unspecified rewards and costs are zero, and unspecified transitions are self-loops.

Write $D := nd(A-1)$ and $m := \min\{K, \Gamma\}$. We will use the bounds $n \geq 5S/8$, $d \geq H/20$, and $SAH/64 \leq D \leq SAH/16$, which follow from the parameter assumptions.

**Variance term** The first family places an unknown best action at each context state in each gadget. These actions determine the probabilities of receiving the gadget rewards; the costs at the pairs in $\mathcal{J}$ determine the safety identification complexity.

**Lemma G.1.** *Assume the parameter conditions of Theorem 4. For every learning algorithm following the protocol in Section 2, there exists a CMDP $\mathcal{M}_v$ with $S$ states, $A$ actions, and horizon $H$ satisfying Assumption 1, $\frac{3v}{4} \leq \mathrm{Var}^\star(\mathcal{M}_v) \leq v$, and $\frac{\Gamma}{4} \leq \mathcal{C}(\mathcal{M}_v) \leq \frac{\Gamma}{2}$. On this instance, every realized interaction sequence satisfies*

$$\frac{31Kv}{32} \leq \mathbb{V}_\Sigma \leq Kv. \tag{G.2}$$

*The algorithm's expected regret on $\mathcal{M}_v$ satisfies*

$$\mathbb{E}_{\mathcal{M}_v}[\mathcal{R}(K)] \geq \frac{3}{32\sqrt{2}}\sqrt{nAdKv}. \tag{G.3}$$

*Proof. Construction.* Set $\varepsilon := \frac{1}{8\sqrt{2}}\sqrt{nA/K}$ and $\rho := 2\sqrt{v/d}$. Since $K \geq nA$ and $v \leq d/4$, we have $0 < \varepsilon \leq 1/(8\sqrt{2})$ and $0 \leq \rho \leq 1$. Choose $(\omega_{i,s})_{(i,s)\in[d]\times[n]}$ independently and uniformly from $[A]$. In gadget $i$, at step $3i-2$, every action at $x$ transitions uniformly to one of $c_1, \ldots, c_n$. At step $3i-1$,

$$P_{3i-1}(g \mid c_s, a) = \frac{1}{2} + \varepsilon \mathbf{1}\{a = \omega_{i,s}\}, \qquad P_{3i-1}(b \mid c_s, a) = 1 - P_{3i-1}(g \mid c_s, a).$$

At step $3i$, every action yields reward $\rho$ at $g$ and zero at $b$, and both states return to $x$. All costs in these gadgets are zero.

At every pair in $\mathcal{J}$, action 1 has cost zero and all other actions have cost one. Every action there moves to $g$, and all subsequent rewards are zero. A policy choosing action 1 at every such pair is safe, so Assumption 1 holds. The transition probabilities and rewards lie in $[0,1]$ by the bounds on $\varepsilon$ and $\rho$.

*Instance properties.* Fix $\pi \in \Pi_{\mathrm{safe}}$ and gadget $i$, and let $\alpha_i := n^{-1}\sum_{s=1}^n \mathbf{1}\{\pi_{3i-1}(c_s) = \omega_{i,s}\}$. Each gadget returns to $x$, so its contribution to the return variance is the next-state value variance at $x$ plus the average over its context states. These two terms satisfy

$$\begin{aligned}
&\mathbb{V}\big(P_{3i-2}(\cdot \mid x, \pi_{3i-2}(x)), V^\pi_{3i-1}\big) + \frac{1}{n}\sum_{s=1}^n \mathbb{V}(P_{3i-1}(\cdot \mid c_s, \pi_{3i-1}(c_s)), V^\pi_{3i}) \\
&\quad = \varepsilon^2\rho^2\alpha_i(1-\alpha_i) + \frac{\rho^2}{4} - \varepsilon^2\rho^2\alpha_i \\
&\quad = \frac{\rho^2}{4} - \varepsilon^2\rho^2\alpha_i^2 \in \left[\frac{3\rho^2}{16}, \frac{\rho^2}{4}\right].
\end{aligned} \tag{G.4}$$

The first equality is the variance decomposition across the two transitions. The last inclusion uses $\varepsilon^2 \leq 1/128$. Summing (G.4) over $i$ and using $d\rho^2/4 = v$ gives $3v/4 \leq \mathrm{Var}^\star(\mathcal{M}_v) \leq v$. The later steps contribute no variance because their rewards are zero.

The only unsafe actions are the $M(A-1)$ actions with cost one at pairs in $\mathcal{J}$. All these pairs are reachable under a safe policy, and each such action has cost gap $1/2$. Therefore, $\mathcal{C}(\mathcal{M}_v) = 4M(A-1) \in [\Gamma/4, \Gamma/2]$.

We next compute $\mathbb{V}_\Sigma$. Under the optimal safe value function, $V^\star_{3i-1}(c_s) = (1/2+\varepsilon)\rho + V^\star_{3i+1}(x)$ for every $s \in [n]$, so the transition from $x$ to the context states has zero next-state value variance. In episode $k$, let $s_{i,k}$ be the visited context index and $a_{i,k}$ the selected action. Write $p_{i,k} := 1/2 + \varepsilon\mathbf{1}\{a_{i,k} = \omega_{i,s_{i,k}}\}$. The two successor values differ by $\rho$, so the context transition contributes $p_{i,k}(1-p_{i,k})\rho^2$. These gadgets are safe and occur before any pair in $\mathcal{J}$ is reached, so their contributions

are included in $\mathbb{V}_\Sigma$. All rewards are deterministic, and the remaining transitions before $\tau_k$ are deterministic or connect states with the same optimal safe value. Hence, pathwise,

$$\mathbb{V}_\Sigma = \sum_{k=1}^{K} \sum_{i=1}^{d} p_{i,k}(1-p_{i,k})\rho^2.$$

Since $p_{i,k} \in \{1/2, 1/2+\varepsilon\}$, $(1-4\varepsilon^2)Kv \le \mathbb{V}_\Sigma \le Kv$. The bound $\varepsilon^2 \le 1/128$ proves (G.2).

*Regret.* Each choice different from $\omega_{i,s}$ loses $\varepsilon\rho$ in expected reward. Since every gadget returns to $x$, these losses add over gadgets and episodes. We lower bound the number of such choices by averaging over the hidden actions.

Fix $(i,s)$ and fix all hidden actions except $\omega_{i,s}$. Let $\mathbb{P}_a$ be the interaction distribution when $\omega_{i,s} = a$, and let $\mathbb{P}_0$ be the reference distribution in which every action at this coordinate reaches $g$ with probability $1/2$. All other transition, reward, and cost distributions are unchanged. Write $\mathbb{E}_a$ and $\mathbb{E}_0$ for the corresponding expectations, and define

$$N_a(k) := \sum_{\ell=1}^{k} \mathbf{1}\{s^\ell_{3i-1} = c_s,\ a^\ell_{3i-1} = a\}.$$

Let $\mathbb{P}_a^{k-1}$ and $\mathbb{P}_0^{k-1}$ denote the distributions of the history through episode $k-1$, including the algorithm's random choices. Write KL for the Kullback–Leibler divergence. The two instances differ only in the transition following action $a$ at the fixed coordinate, and the KL divergence of one such observation is $-\frac{1}{2}\log(1-4\varepsilon^2) \le 4\varepsilon^2$. The adaptive chain rule gives

$$\mathrm{KL}(\mathbb{P}_0^{k-1} \| \mathbb{P}_a^{k-1}) \le 4\varepsilon^2 \mathbb{E}_0[N_a(k-1)].$$

Conditional on the history before episode $k$, the probability of choosing $a$ at $c_s$ is determined by the algorithm's policy distribution. The probability of reaching $c_s$ is $1/n$, independently of this choice. Applying Pinsker's inequality to the history distributions therefore gives

$$\mathbb{P}_a\{s^k_{3i-1} = c_s,\ a^k_{3i-1} = a\} \le \mathbb{P}_0\{s^k_{3i-1} = c_s,\ a^k_{3i-1} = a\} + \frac{\sqrt{2}\varepsilon}{n}\sqrt{\mathbb{E}_0[N_a(k-1)]}.$$

Summing over $k$ and averaging over $a$ yields

$$\begin{aligned}\frac{1}{A}\sum_{a=1}^{A} \mathbb{E}_a[N_a(K)] &\le \frac{K}{nA} + \frac{K\sqrt{2}\varepsilon}{nA}\sum_{a=1}^{A}\sqrt{\mathbb{E}_0[N_a(K)]} \\ &\le \frac{K}{nA} + \frac{K\sqrt{2}\varepsilon}{n}\sqrt{\frac{K}{nA}} \le \frac{5K}{8n}. \end{aligned} \tag{G.5}$$

The second inequality uses Cauchy–Schwarz and $\sum_a \mathbb{E}_0[N_a(K)] = K/n$. The last inequality uses the definition of $\varepsilon$ and $A \ge 2$.

By (G.5), averaging over all hidden actions and summing over the $nd$ coordinates gives at least $3dK/8$ incorrect choices in expectation. Multiplying by the loss $\varepsilon\rho$ gives

$$\mathbb{E}[\mathcal{R}(K)] \ge \frac{3dK}{8}\varepsilon\rho = \frac{3}{32\sqrt{2}}\sqrt{nAdKv}.$$

The expectation here also averages over the hidden actions. Some fixed choice therefore satisfies (G.3). The variance and safety identification complexity bounds hold for every such choice. □

**Safety identification term** The second family uses the same state–step pairs, but the unknown action at each pair now determines safety. The initial gadgets supply the target variance without contributing regret. An incorrect choice at a pair in $\mathcal{J}$ loses the reward from the last $R$ steps.

**Lemma G.2.** *Assume the parameter conditions of Theorem 4. For every learning algorithm following the protocol in Section 2, there exists a CMDP $\mathcal{M}_w$ with $S$ states, $A$ actions, and horizon $H$ satisfying Assumption 1, $\mathrm{Var}^\star(\mathcal{M}_w) = v$, and $\frac{\Gamma}{4} \le \mathcal{C}(\mathcal{M}_w) \le \frac{\Gamma}{2}$. On this instance, $\mathbb{V}_\Sigma = Kv$ for every realized interaction sequence. The algorithm's expected regret on $\mathcal{M}_w$ satisfies*

$$\mathbb{E}_{\mathcal{M}_w}[\mathcal{R}(K)] \ge \frac{1}{512}\sqrt{Dm}. \tag{G.6}$$

*Proof. Construction.* Use $d$ action-independent gadgets in the first $3d$ steps. In gadget $i$, every action at $x$ moves to $c_1$ at step $3i-2$, and every action at $c_1$ moves to $g$ or $b$, each with probability $1/2$, at step $3i-1$. At step $3i$, the rewards are $\rho := 2\sqrt{v/d}$ at $g$ and zero at $b$, and both states return to $x$. These rewards lie in $[0,1]$ because $v \le d/4$.

Choose $(\theta_j)_{j\in\mathcal{J}}$ independently and uniformly from $[A]$. At pair $j$, action $\theta_j$ has cost zero and moves to $g$. Every other action has cost one and moves to $b$. Both states have zero reward until the last $R$ steps. At each of those steps, the reward is $\Delta_r/R$ at $g$ and zero at $b$, where $\Delta_r := \sqrt{D/m}$. The bounds $R \ge H/10$, $m \ge 16SA/H$, and $D \le SAH/16$ imply $D \le R^2 m$, so $\Delta_r/R \le 1$. A policy choosing $\theta_j$ at every pair $j$ is safe, which proves Assumption 1.

*Instance properties.* The initial gadgets contribute $d\rho^2/4 = v$ to the return variance of every policy. Under every safe policy, the value is $\Delta_r$ at each pair in $\mathcal{J}$ and at $x$ whenever $x$ is reached after the gadgets. The action chosen at each pair moves deterministically to $g$. Thus the later steps contribute zero variance, and $\mathrm{Var}^\star(\mathcal{M}_w) = v$. The only unsafe actions are the $M(A-1)$ incorrect actions at these pairs. Every pair is reachable under a safe policy, and each incorrect action has cost gap $1/2$, so $\mathcal{C}(\mathcal{M}_w) = 4M(A-1) \in [\Gamma/4, \Gamma/2]$.

The initial gadgets lie in the true safe graph and occur before every pair in $\mathcal{J}$. Each contributes $\rho^2/4 = v/d$ to $\mathbb{V}_\Sigma$ in every episode. All rewards are deterministic. Under $V^\star$, every pair in $\mathcal{J}$ has continuation value $\Delta_r$, while safe actions at these pairs and all subsequent transitions are deterministic. Hence $\mathbb{V}_\Sigma = Kv$ for every realized interaction sequence.

*Regret.* An incorrect action at a pair in $\mathcal{J}$ loses $\Delta_r$ in reward. We bound the number of incorrect choices, first averaging over the hidden actions. Let $T_j(t)$ count visits to pair $j$ in the first $t$ episodes, set $T_j := T_j(K)$, and let $W_j$ count choices different from $\theta_j$ over the $K$ episodes.

Fix the sequence of reached pairs, all other hidden actions, the algorithm's random seed, and the transition randomness in the initial gadgets. Consider the actions tested at $j$ along the hypothetical history in which each test there has cost one and moves to $b$. List the distinct actions in the order in which they are first tested, stopping once all $A$ actions have been listed or the visits end. Append any untested actions in a fixed order. This ordering is independent of $\theta_j$, and the actual history agrees with the hypothetical one until $\theta_j$ is first tested. Also, $T_j$ depends only on the sequence of reached pairs. Conditional on $T_j = q$, the rank of $\theta_j$ is therefore uniform on $[A]$. If that rank is $r$, at least $\min\{q, r-1\}$ choices are incorrect; repeated tests cannot reduce this number. Averaging over the rank gives

$$\mathbb{E}[W_j \mid T_j = q] \ge \frac{1}{A}\sum_{r=1}^{A} \min\{q, r-1\} \ge \frac{1}{4}\min\{q, A-1\}. \tag{G.7}$$

It remains to bound the total number of visits before individual counts reach $A-1$. Let $J_k$ be the pair reached in episode $k$. Pathwise,

$$\sum_{j\in\mathcal{J}} \min\{T_j, A-1\} = \sum_{k=1}^{K} \mathbf{1}\{T_{J_k}(k-1) < A-1\}.$$

By (G.1), $J_k$ is uniform on $\mathcal{J}$ conditional on the past. Since $\sum_j T_j(k-1) = k-1$, we have $\mathbb{E}[T_{J_k}(k-1)] = (k-1)/M$. For $k \le \min\{K, \lfloor M(A-1)/2 \rfloor\}$, Markov's inequality therefore bounds the expectation of each corresponding indicator from below by $1/2$. Since $M(A-1) \ge 2$, $\lfloor M(A-1)/2 \rfloor \ge M(A-1)/4$, and hence

$$\mathbb{E}\left[\sum_{j\in\mathcal{J}} \min\{T_j, A-1\}\right] \ge \frac{1}{8}\min\{K, M(A-1)\} \ge \frac{m}{128}. \tag{G.8}$$

The last inequality uses $M(A-1) \ge \Gamma/16$ and $\min\{K, M(A-1)\} \ge m/16$. Combining (G.7) and (G.8) with the loss $\Delta_r$ per incorrect choice gives

$$\mathbb{E}[\mathcal{R}(K)] = \Delta_r \sum_{j\in\mathcal{J}} \mathbb{E}[W_j] \ge \frac{\Delta_r m}{512} = \frac{1}{512}\sqrt{Dm}.$$

The expectation also averages over the hidden actions, so some fixed choice satisfies (G.6). The variance and safety identification complexity bounds hold for every such choice. □

*Proof of Theorem 4.* By Lemmas G.1 and G.2, both instances lie in the required ranges of $\mathrm{Var}^\star$ and $\mathcal{C}$. The instance $\mathcal{M}_v$ satisfies $31Kv/32 \le \mathbb{V}_\Sigma \le Kv$ for every realized interaction sequence, while $\mathcal{M}_w$ satisfies $\mathbb{V}_\Sigma = Kv$. Thus either instance satisfies the stated range for $\mathbb{V}_\Sigma$. Choose $\mathcal{M} = \mathcal{M}_v$ when the right-hand side of (G.3) is at least that of (G.6), and choose $\mathcal{M} = \mathcal{M}_w$ otherwise. Then

$$\begin{aligned}\mathbb{E}_{\mathcal{M}}[\mathcal{R}(K)] &\ge \max\left\{\frac{3}{32\sqrt{2}}\sqrt{nAdKv}, \frac{1}{512}\sqrt{Dm}\right\} \\ &= \Omega\left(\sqrt{SAHKv} + \sqrt{SAH\min\{K,\Gamma\}}\right).\end{aligned}$$

The assumptions of Theorem 4 give $n \ge 5S/8$, $d \ge H/20$, and $D \ge SAH/64$. Substituting these bounds and using $\max\{x,y\} \ge (x+y)/2$ gives the final line. □

### G.2 Lower Bounds for Action Elimination

We next prove the elimination lower bound discussed in Section 3.3. The learner must distinguish an unsafe root from alternatives in which changing one leaf makes it safe. This gives a lower bound on the required leaf observations, which we convert into root executions.

**Instances and alternatives** Fix $\delta \in (0, 1/4]$ and $\tau = 1/2$, and consider an instance $\mathcal{M}$ with an unsafe root $u = (s, a, h)$. Use the witness-tree representation in Lemma C.2. Let $\mathcal{L}_u$ be its leaves and, for $\ell \in \mathcal{L}_u$, let $\Delta_\ell$ be its cost gap and $q_\ell > 0$ the product of the transition probabilities on its path. Assume the following conditions on the tree and its observations.

(i) Every nonterminal action satisfies the immediate cost constraint and has exactly one non-viable successor. All its other successors are viable. The root state has another action in $\mathcal{A}_{\mathrm{safe},h}(s)$.

(ii) Each nonroot state–step pair in the tree can be reached only through its designated parent action. Different branches use disjoint state–action pairs at each step.

(iii) At every leaf $\ell$, $0 < \Delta_\ell \le 1/4$, all successors are viable, and the cost has distribution $\mathrm{Bernoulli}(1/2 + \Delta_\ell)$. Define $\mathcal{M}^\ell$ by replacing this cost distribution with $\mathrm{Bernoulli}(1/2 - \Delta_\ell)$ and leaving all other observations unchanged.

(iv) In $\mathcal{M}$ and $\mathcal{M}^\ell$, conditional on the leaf state–action pair, the cost is independent of the pair consisting of the reward and next state. The reward and next state have the same joint distribution in both instances.

By Lemma C.2,

$$\kappa_h(s,a) = \sum_{\ell \in \mathcal{L}_u} \frac{1}{q_\ell \Delta_\ell^2}. \tag{G.9}$$

The leaf action is safe in $\mathcal{M}^\ell$. At its parent, this provides a safe action at the previously non-viable successor, while all other successors were already viable. Repeating this argument makes every ancestor action, including $u$, safe.

**Elimination rules** The learner chooses its policy before each episode, as in Section 2. An elimination rule removes $u$ at an episode boundary and never executes it afterward. Let $\sigma_u \in \{0, 1, \ldots\} \cup \{\infty\}$ be the number of completed episodes when it is removed, and let $N_u(t)$ count its executions through the end of episode $t$. Thus $\{\sigma_u \ge k\}$ is known before episode $k$ begins. Write $\mathrm{kl}(p,q) := p\log(p/q) + (1-p)\log((1-p)/(1-q))$. Terms with zero numerator are zero, and a positive numerator divided by zero contributes $+\infty$. The rule is $\delta$-valid at $u$ if

$$\begin{aligned}&\mathbb{P}_{\mathcal{M}}(\sigma_u < \infty) \ge 1 - \delta, \\ &\mathbb{P}_{\mathcal{M}^\ell}(\sigma_u < \infty) \le \delta \quad (\ell \in \mathcal{L}_u).\end{aligned} \tag{G.10}$$

**Lemma G.3** (Elimination lower bound)**.** *Consider the instance $\mathcal{M}$ and alternatives $\mathcal{M}^\ell$, $\ell \in \mathcal{L}_u$, satisfying conditions (i)–(iv) above. Every $\delta$-valid elimination rule at $u$, as defined in (G.10), satisfies*

$$\mathbb{E}_{\mathcal{M}}[N_u(\sigma_u)] \ge \frac{1}{32}\kappa_h(s,a)\log\frac{1}{2\delta}. \tag{G.11}$$

*Proof. Leaf observations.* Fix $\ell \in \mathcal{L}_u$, and let $N_\ell(t)$ count observations of its cost through episode $t$. By condition (iv), the KL divergence of the full observation at this leaf equals that of its Bernoulli cost. For $0 < \Delta_\ell \le 1/4$,

$$\mathrm{kl}(1/2 + \Delta_\ell, 1/2 - \Delta_\ell) = 2\Delta_\ell \log \frac{1 + 2\Delta_\ell}{1 - 2\Delta_\ell} \le 16\Delta_\ell^2,$$

where the inequality uses $\log((1+x)/(1-x)) \le 4x$ for $x \in [0, 1/2]$.

Let $E_n := \{\sigma_u \le n\}$, and stop the interaction history at the end of episode $\sigma_u \wedge n$. The stopped history, including the elimination decision, determines $E_n$. Since only the cost at leaf $\ell$ differs between the instances, the chain rule expresses the KL divergence of the stopped histories as the expected leaf count times the single-observation divergence. Data processing then gives

$$16\Delta_\ell^2 \mathbb{E}_{\mathcal{M}}[N_\ell(\sigma_u \wedge n)] \ge \mathrm{kl}(\mathbb{P}_{\mathcal{M}}(E_n), \mathbb{P}_{\mathcal{M}^\ell}(E_n)) .$$

Set $p := \mathbb{P}_{\mathcal{M}}(\sigma_u < \infty)$ and $q := \mathbb{P}_{\mathcal{M}^\ell}(\sigma_u < \infty)$. Letting $n$ tend to infinity, monotone convergence and lower semicontinuity give

$$16\Delta_\ell^2 \mathbb{E}_{\mathcal{M}}[N_\ell(\sigma_u)] \ge \mathrm{kl}(p, q) \ge \mathrm{kl}(1 - \delta, \delta), \tag{G.12}$$

where the last inequality uses $p \ge 1 - \delta > \delta \ge q$ and the monotonicity of binary KL.

*Root executions.* For each execution of $u$, the fixed episode policy determines a path through the witness tree before any transition along that path is observed. For $\ell \in \mathcal{L}_u$, let $T_\ell(t)$ count executions assigned to the path ending at $\ell$ through episode $t$. For every $n$,

$$\mathbb{E}_{\mathcal{M}}[N_\ell(\sigma_u \wedge n)] = q_\ell \mathbb{E}_{\mathcal{M}}[T_\ell(\sigma_u \wedge n)],$$
$$N_u(\sigma_u \wedge n) = \sum_{\ell \in \mathcal{L}_u} T_\ell(\sigma_u \wedge n).$$

To obtain the first identity, include episode $k$ exactly when $\sigma_u \ge k$, an event known before that episode. Conditional on an execution assigned to leaf $\ell$, all transitions on its path occur with probability $q_\ell$. Condition (ii) ensures that every observation at the leaf comes from this path. The second identity holds because each root execution is assigned to exactly one leaf. Letting $n$ tend to infinity and applying monotone convergence yields

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}}[N_u(\sigma_u)] &= \sum_{\ell \in \mathcal{L}_u} \frac{\mathbb{E}_{\mathcal{M}}[N_\ell(\sigma_u)]}{q_\ell} \\
&\ge \frac{1}{16} \mathrm{kl}(1 - \delta, \delta) \sum_{\ell \in \mathcal{L}_u} \frac{1}{q_\ell \Delta_\ell^2} \\
&= \frac{1}{16} \kappa_h(s, a) \, \mathrm{kl}(1 - \delta, \delta).
\end{aligned}$$

The inequality uses (G.12), and the last equality uses (G.9). Finally, since $\delta \le 1/4$,

$$\mathrm{kl}(1 - \delta, \delta) = (1 - 2\delta) \log \frac{1 - \delta}{\delta} \ge \frac{1}{2} \log \frac{1}{2\delta}.$$

Substitution proves (G.11). □

**Corollary G.1.** *Let $\mathcal{U}$ consist of all triples $(s, a, h)$ with $h \in [H]$, $a \notin \mathcal{A}_{\mathrm{safe},h}(s)$, and $(s, h)$ reachable under a safe policy, as indexed in (8). Suppose these triples are the roots of pairwise-disjoint witness trees satisfying conditions (i)–(iv) above. For $u \in \mathcal{U}$ and $\ell \in \mathcal{L}_u$, let $\mathcal{M}^{u,\ell}$ be the corresponding single-leaf alternative. For a learner using the elimination rules defined above at all roots, suppose $\mathbb{P}_{\mathcal{M}}\big(\bigcap_{u \in \mathcal{U}} \{\sigma_u < \infty\}\big) \ge 1 - \delta$, and $\mathbb{P}_{\mathcal{M}^{u,\ell}}(\sigma_u < \infty) \le \delta \quad (u \in \mathcal{U}, \ell \in \mathcal{L}_u)$. Then*

$$\mathbb{E}_{\mathcal{M}}\left[\sum_{u \in \mathcal{U}} N_u(\sigma_u)\right] \ge \frac{1}{32} \mathcal{C}(\mathcal{M}) \log \frac{1}{2\delta}.$$

*Proof.* The probability conditions imply the hypotheses of Lemma G.3 for every root. Applying (G.11) to every root and summing over $u \in \mathcal{U}$ proves the claim because $\sum_{u=(s,a,h) \in \mathcal{U}} \kappa_h(s, a) = \mathcal{C}(\mathcal{M})$. □

# H EXPERIMENTS

In this section, we conduct experiments to illustrate our theoretical findings and evaluate the performance of our algorithm. We focus on the following four questions:

**Q1:** *How does SVAE compare with baselines in regret and step-wise violation?*

**Q2:** *Can SVAE exploit low return variance induced by safety constraints?*

**Q3:** *Can safety identification remain difficult when safe-policy return variance is zero?*

**Q4:** *Does recursive elimination exclude unsafe branches while preserving safe continuations?*

Appendix H.1 describes the experimental setup and baselines. We evaluate regret and step-wise violation on stochastic navigation tasks in Appendix H.2 (Q1). We then study variance adaptation, safety identification complexity, and recursive elimination using controlled CMDPs in Appendices H.3–H.5 (Q2–Q4). Hyperparameter selection and implementation details are given in Appendix H.6.

## H.1 EXPERIMENTAL SETUP

**Learning setting** The reward means, mean costs, and transition probabilities are unknown. All algorithms learn from online interaction without a known safe policy. Rewards are Bernoulli in Q1 and deterministic in Q2–Q4, where we study the safe subgraph and safety identification without reward noise. The reward means remain fixed throughout learning. Costs are Bernoulli with mean $c_h(s,a)$, and next states follow the transition kernel. Estimates are maintained separately for each $(h,s,a)$. We set $\tau = 0.6$, $\delta = 0.05$, and run each algorithm for $K = 10{,}000$ episodes.

**Baselines** For navigation (Q1), we compare SVAE with SUCBVI (Xiong et al., 2023). Both methods use the same recursive safety filter, so the comparison focuses on their planning rules. Our SUCBVI implementation estimates unknown rewards and costs separately for each $(h,s,a)$. We also include MVP-V (Zhou et al., 2023b) and a variant with the recursive safety filter applied at every episode, denoted by *MVP-V + filter*.

**Ablations** We consider two variants of SVAE with Hoeffding-type bonuses. *SVAE (Hoeffding)* replaces the empirical reward and transition variances in (10) with $1/4$ and $(H-h)^2/4$, respectively. *SVAE (Hoeffding-P)* replaces only the transition variance, with P denoting the transition component. Both retain the finite-sample correction and recursive safety filter. We also consider *SVAE w/o propagation*, which retains variance-adaptive planning but eliminates actions only through their own cost lower confidence bounds. Q2 compares SVAE with both bonus variants, and Q4 compares it with the variant without propagation. Q3 evaluates SVAE as the environment parameters vary.

**Metrics** We evaluate regret using exact policy values and the optimal safe value. Step-wise violation is the cumulative positive excess of mean costs over the threshold along sampled trajectories, as defined in Section 2. The true model and safe sets are used only for offline tuning and evaluation. For navigation, we also report the strong regret (Stradi et al., 2025b) $\mathcal{R}_+(K) := \sum_{k=1}^{K}[V_1^\star(s_1) - V_1^{\pi^k}(s_1)]_+$, which does not allow cancellation across episodes. We track policy value and expected violation per episode using a moving average over up to 500 episodes.

We use the same observation streams and state/action relabelings to pair methods. For navigation (Q1), we average five paired seeds within each of eight independent environments, then average across environments. Each condition in Q2–Q4 uses 20 seeds. We report means with sample standard deviations (SDs) in parentheses. SDs are computed across environment means for Q1 and across seeds for Q2–Q4. Shading indicates one SD around the mean, with lower bounds clipped at zero for nonnegative metrics.

We tune all navigation methods with the same search budget on separate development and validation environments, then fix their hyperparameters before testing. SVAE and its ablations use the same planning and cost-confidence multipliers in the controlled experiments. Appendix H.6 gives the tuning procedure and selected values.

### H.2 Stochastic Navigation

We first compare the regret and step-wise violation of SVAE and the baselines in a $5 \times 5$ gridworld with four movement actions and horizon 20. The agent starts at the middle-left cell, and the middle-right cell is an absorbing goal. At each step, rewards are sampled independently from a Bernoulli distribution with mean $0.9$ at the goal and $0.1$ elsewhere. Reward noise is independent of costs and transitions. Each episode lasts $H$ steps, with the agent remaining at the goal after reaching it.

For each state–action pair, we sample the probability of the intended move uniformly from $[0.8, 0.95]$ when generating the environment. With the remaining probability, the agent moves in either perpendicular direction with equal probability. Attempts to move outside the grid leave the agent in the same cell. An action has mean cost in $[0.7, 0.85]$ if it can cross the boundary between columns 3 and 4 in rows 2–4. All other non-goal actions have mean costs in $[0.15, 0.3]$, and the cost at the goal is zero.

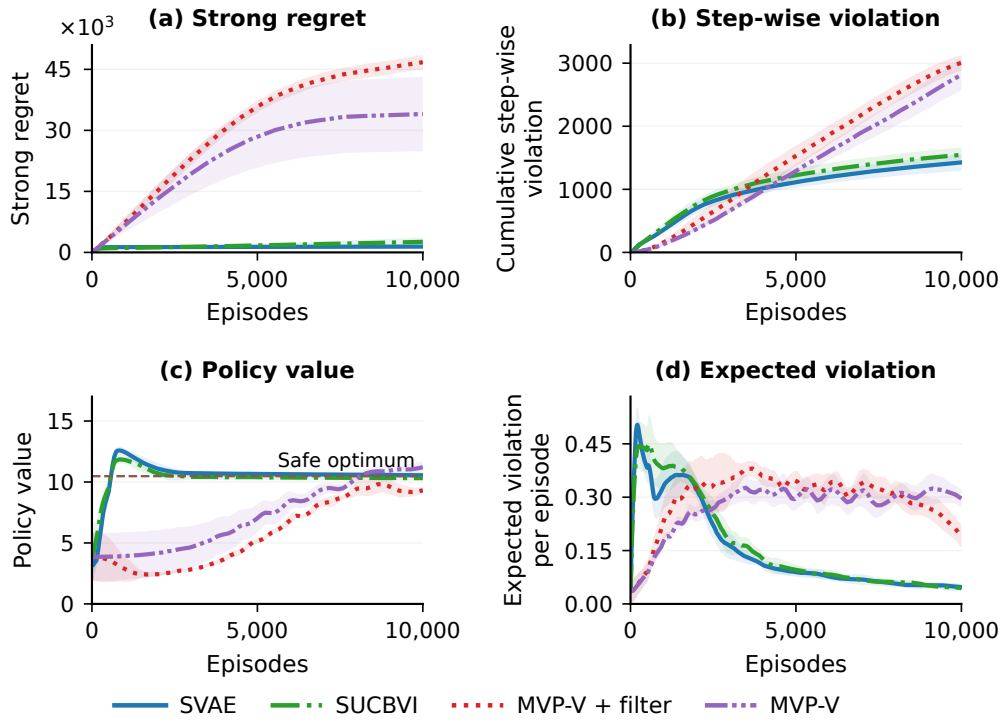



Figure 3: **Regret and safety in stochastic navigation.** Panels (a) and (b) show strong regret and cumulative step-wise violation. Panels (c) and (d) show policy value and expected violation per episode, using a moving average over up to 500 episodes. At $k = 0$, we evaluate the initial policy. Curves show means across eight environments after averaging five paired seeds per environment. Shaded areas show one SD across environment means. The horizontal line in (c) marks the mean optimal safe value.

**Results** Compared with SUCBVI, SVAE reduces final strong regret by $44.9\%$ and cumulative step-wise violation by $7.7\%$ (Table 1). The mean (SD) paired differences, computed as SVAE minus SUCBVI, are $-1{,}162.23$ $(947.46)$ for $\mathcal{R}_+$ and $-119.70$ $(45.10)$ for violation.

Table 1: **Regret and step-wise violation in stochastic navigation.** We report cumulative regret, strong regret, and step-wise violation at $K = 10{,}000$. Entries are means (SDs) across eight environments after averaging five paired seeds per environment. Bold numbers mark the lowest mean in each column, including ties at the displayed precision.

| Method | $\mathcal{R}(K)/10^3$ | $\mathcal{R}_+(K)/10^3$ | $\mathrm{Viol}(K)/10^3$ |
|---|---|---|---|
| SVAE | **-2.406 (0.947)** | **1.428 (0.112)** | **1.426 (0.132)** |
| SUCBVI | 0.150 (1.707) | 2.590 (1.037) | 1.546 (0.121) |
| MVP-V + filter | 45.989 (1.973) | 46.756 (1.776) | 3.005 (0.124) |
| MVP-V | 31.238 (9.735) | 34.020 (9.161) | 2.816 (0.246) |

SVAE's expected violation per episode decreases from $0.383$ in the first 500 episodes to $0.0474$ in the last 500 (Figure 3). Over the last 500 episodes, its mean policy value is $10.553$, above the average optimal safe value of $10.476$ while some constraint violation remains. Unconstrained MVP-V reaches a higher mean policy value of $11.206$ over the last 500 episodes, with greater expected violation of $0.2961$ per episode.

### H.3 Variance Adaptation

To study whether SVAE can exploit low return variance in an unknown safe subgraph, we use a two-step CMDP with eight states and four actions. The reward at the root is zero. Each of three root actions leads to two possible second-step states with probability $1/2$ each. The corresponding reward pairs are $(0.4, 0.4)$, $(0.25, 0.55)$, and $(0.1, 0.7)$. The fourth root action gives a deterministic return of $0.5$.

All three branching actions are available in every instance. We keep rewards and transitions fixed and vary costs to select the safe branch. The selected branching action and the deterministic action have mean cost $0.2$, while the other two root actions have mean cost $0.8$. All second-step mean costs are $0.2$. Thus, the optimal safe value is $0.5$, the expected-return gap between the two safe root actions is $0.1$, and $\mathcal{C} = 50$ in every instance.

Let $u \in \{0, 0.15, 0.30\}$ denote half the reward difference on the safe branch. Then $S^\star = 2u$, $\mathrm{Var}^\star = u^2$, and $B(\mathcal{G}_{\mathrm{safe}}) = 1$. The safe branching policy attains the variance bound in Theorem 1. When $u = 0$, both successor states have the same value, so the transition variance is zero despite the random next state. This is the zero-span case of Corollary 1.

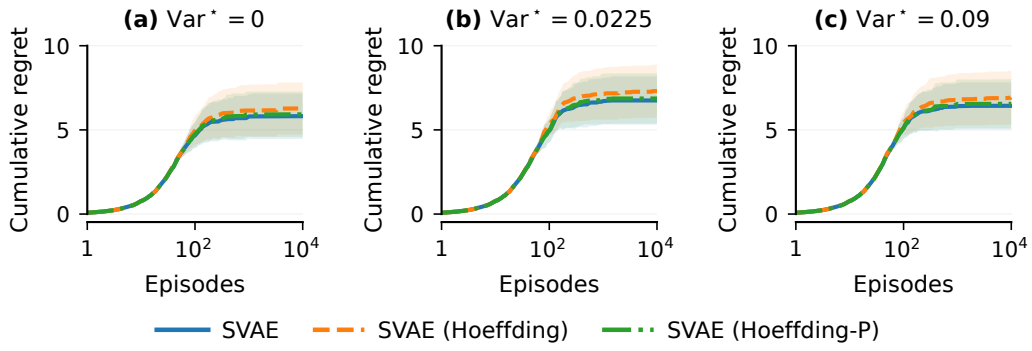



Figure 4: **Regret under different safe-policy return variances.** The three panels compare SVAE with its Hoeffding-type variants as $\mathrm{Var}^\star$ varies. Curves show means over 20 paired seeds, with shading indicating one SD. The episode axes use a logarithmic scale.

**Results** SVAE has lower final regret on average than both Hoeffding-type variants in all three conditions (Figure 4 and Table 2). Compared with SVAE (Hoeffding-P), SVAE selects suboptimal branching actions $1.10$–$1.20$ fewer times on average. This comparison isolates the effect of using empirical next-state value variance in the transition bonus, as both methods share the reward-variance estimates and safety filter.

Table 2: **Accumulated variance and regret differences.** We report SVAE's accumulated variance $\mathbb{V}_\Sigma$ and paired regret differences at $K = 10{,}000$ as means (SDs) over 20 seeds. Each difference $\Delta\mathcal{R}(K)$ is SVAE's regret minus the regret of the indicated Hoeffding-type variant.

| | | | $\Delta\mathcal{R}(K)$ | |
|---|---|---|---|---|
| $\mathrm{Var}^\star$ | $K\mathrm{Var}^\star$ | $\mathbb{V}_\Sigma$ | Hoeffding-P | Hoeffding |
| 0 | 0 | 0.000 (0.000) | −0.110 (0.031) | −0.460 (0.268) |
| 0.0225 | 225 | 0.739 (0.263) | −0.115 (0.198) | −0.560 (0.376) |
| 0.09 | 900 | 2.691 (1.078) | −0.120 (0.202) | −0.495 (0.386) |

The optimal safe action gives a deterministic return even when another safe action has positive return variance. As SVAE increasingly selects the optimal action, the variance it accumulates can therefore be much smaller than $K\mathrm{Var}^\star$. Let $a^\star$ and $a_b$ denote the deterministic optimal and safe branching root actions, respectively. Following Appendix B.1, $N_1^{K+1}(s_1, a)$ is the number of times action $a$ is taken at the root in the first $K$ episodes. Each branching action has an expected return $0.1$ below that of $a^\star$. Only $a_b$ contributes to $\mathbb{V}_\Sigma$, since this sum ends before the first unsafe action. Hence,

$$\mathcal{R}(K) = 0.1 \sum_{a \neq a^\star} N_1^{K+1}(s_1, a), \qquad \mathbb{V}_\Sigma = u^2 N_1^{K+1}(s_1, a_b).$$

SVAE's mean $\mathbb{V}_\Sigma$ is about $0.3\%$ of $K\mathrm{Var}^\star$ in both conditions with positive variance (Table 2). In these instances, the term $\min\{\mathbb{V}_\Sigma, K\mathrm{Var}^\star\}$ in Theorem 2 reflects the smaller variance along the learning trajectory.

### H.4 Safety Identification Complexity

We study safety identification in a two-step CMDP with four states and two actions. At the root, one action is safe and gives a return of $0.4$. The other gives a return of $0.5$ but reaches a state with no safe action with probability $p$. Both actions at this successor have mean cost $\tau + \Delta_c$, while all other mean costs are $0.2$. The root action with the higher return is therefore unsafe because it can reach a state with no safe continuation. Every safe policy has a deterministic return, so $\mathrm{Var}^\star = 0$. By (7), the identification complexity is $\mathcal{C} = 2/(p\Delta_c^2)$. We vary $p$ with $\Delta_c = 0.2$ fixed, then vary $\Delta_c$ with $p = 0.2$ fixed. We measure how often the unsafe root action is taken and how often its non-viable successor is visited.

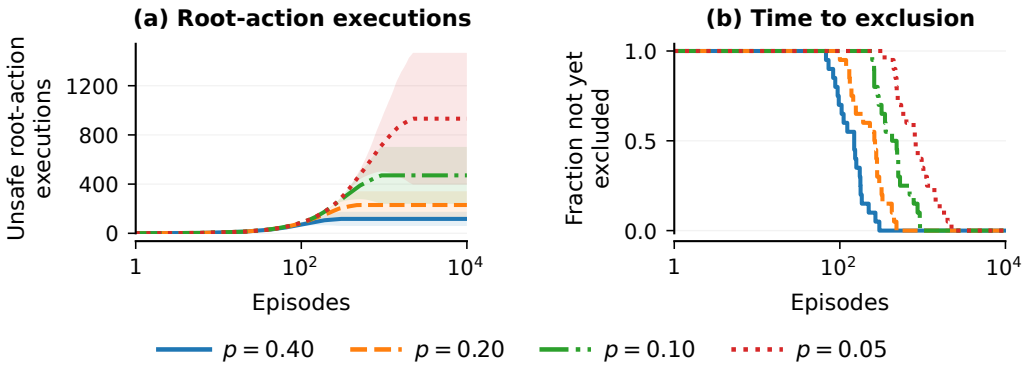



Figure 5: **Effect of transition probability on safety identification.** Panel (a) shows the mean cumulative number of times the unsafe root action is taken, with shading indicating one SD. Panel (b) shows the fraction of runs in which this action has not yet been excluded. For each $p$, we use 20 paired seeds and fix $\Delta_c = 0.2$. Both episode axes use a logarithmic scale.

**Results** As $p$ decreases from $0.40$ to $0.05$, the mean number of times the unsafe root action is taken increases from $117.1$ to $931.9$ (Figure 5). The mean number of visits to the successor remains $46.5$, and cumulative violation remains $9.29$ (Table 3). For each paired seed, all four conditions require

the same number of cost observations at the successor. A smaller transition probability increases the number of interactions needed to collect these observations, as reflected by the reachability factor in (7).

Table 3: **Effect of transition probability and cost gap on safety identification.** We report executions of the unsafe root action, visits to its non-viable successor, and cumulative step-wise violation at $K = 10{,}000$. Entries are means (SDs) over 20 seeds. The two parameter sweeps use separate seed sets. Bold rows mark the lowest identification complexity $\mathcal{C}$ in each sweep.

| $p$ | $\Delta_c$ | $\mathcal{C}$ | Root executions | Successor visits | Viol($K$) |
|---|---|---|---|---|---|
| *Varying the successor probability ($\Delta_c = 0.20$)* | | | | | |
| **0.40** | **0.20** | **125.0** | **117.1 (57.8)** | **46.5 (21.3)** | **9.29 (4.27)** |
| 0.20 | 0.20 | 250.0 | 230.1 (112.1) | 46.5 (21.3) | 9.29 (4.27) |
| 0.10 | 0.20 | 500.0 | 471.0 (231.5) | 46.5 (21.3) | 9.29 (4.27) |
| 0.05 | 0.20 | 1,000.0 | 931.9 (535.4) | 46.5 (21.3) | 9.29 (4.27) |
| *Varying the cost gap ($p = 0.20$)* | | | | | |
| 0.20 | 0.10 | 1,000.0 | 1,056.9 (599.7) | 214.1 (126.0) | 21.41 (12.60) |
| 0.20 | 0.15 | 444.4 | 591.7 (391.5) | 118.7 (82.5) | 17.81 (12.37) |
| 0.20 | 0.20 | 250.0 | 296.2 (123.0) | 59.3 (26.4) | 11.85 (5.28) |
| **0.20** | **0.30** | **111.1** | **126.7 (40.6)** | **26.0 (6.9)** | **7.79 (2.07)** |

Reducing the cost gap increases the number of observations needed at the successor. With $p = 0.2$ fixed, decreasing $\Delta_c$ from 0.30 to 0.10 increases the mean number of successor visits from 26.0 to 214.1. Root-action executions and cumulative violation also increase. Together, the two sweeps show how rare successors and small cost gaps increase $\mathcal{C}$ at zero safe-policy return variance, illustrating the separate safety identification term in Theorem 2. In every tested condition, SVAE excludes the unsafe root action within the episode budget and never takes it afterward. Cumulative violation then remains constant, consistent with the positive-gap bound in Theorem 3.

### H.5 Recursive Elimination

We study recursive elimination by comparing SVAE with SVAE w/o propagation in a five-step CMDP with three states, two actions, and deterministic transitions. The two branches give returns of 0.4 and 1. On the branch with the higher return, the state reached after $d = 3$ transitions has two actions with mean cost 0.8. All other costs are zero. An otherwise identical instance sets one of these two costs to zero, leaving a safe continuation. Both instances have $\mathrm{Var}^\star = 0$. We define detection as the first time the cost lower confidence bounds of both actions at the later state exceed $\tau$. We count entries into the branch with the higher return after detection.

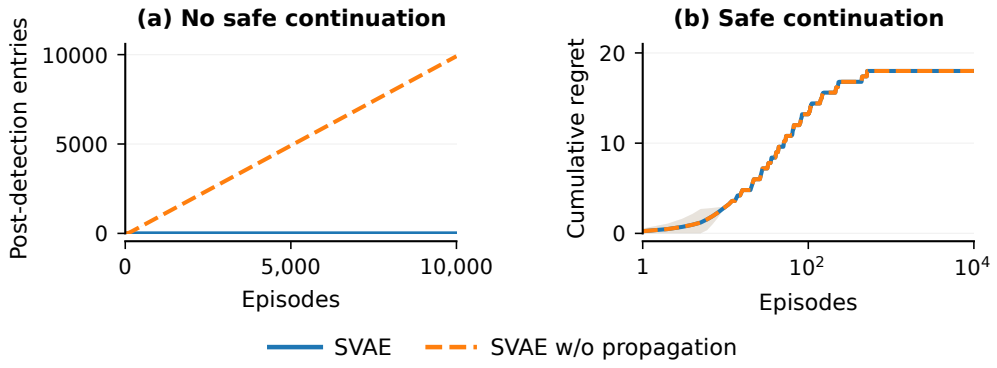



Figure 6: **Recursive elimination at propagation distance $d = 3$.** Panel (a) shows cumulative entries into the unsafe branch after detection. Panel (b) shows cumulative regret when the branch has a safe continuation, with a logarithmic episode axis. Curves show means over 20 paired seeds, with shading indicating one SD. The two methods have identical regret in (b) for every paired seed.

**Results** At $d = 3$, SVAE excludes the root action in the same backward pass that detects the absence of a safe continuation. It never enters the branch after detection in any of the 20 runs (Figure 6a). The variant without propagation continues to enter the branch, with $9{,}910.05$ entries after detection on average. SVAE incurs cumulative violation of $11.99$, compared with $1{,}994.00$ for the ablation. The $d = 1$ condition gives the same result: zero entries after detection for SVAE and $9{,}921.95$ for the ablation. Recursive elimination uses evidence from the later state to exclude the root action, illustrating the control of unsafe candidate selections in Lemma C.3.

When one action at the later state remains safe, SVAE retains the branch. SVAE and the ablation follow the same trajectories, with final regret $18.00$ and zero violation over the last 500 episodes (Figure 6b). SVAE's recursive update excludes the branch only when both downstream actions have been ruled out.

### H.6 Additional Experimental Details

**Hyperparameter selection** We scale the planning bonus by $\alpha$ and the cost-confidence radius by $\beta$. For navigation, each method is evaluated with 18 configurations. Methods with safety filtering use six planning multipliers combined with three cost-confidence multipliers, while unconstrained MVP-V uses 18 planning multipliers. To be eligible, configurations with safety filtering must have mean expected violation of at most $0.02$ per step. They must also retain all truly safe actions and keep a nonempty candidate set at every visited state that admits a safe continuation. Eligible configurations are ranked by increasing strong regret $\mathcal{R}_+$, and unconstrained MVP-V configurations by decreasing policy value. We first evaluate all configurations on two development environments with two seeds per environment. The best three are then evaluated on two separate validation environments, again with two seeds per environment. Both stages use $K = 10{,}000$ episodes. After validation, we select the best configuration and fix its hyperparameters before testing.

For navigation, $(\alpha, \beta)$ is $(0.0003, 0.30)$ for SVAE and $(0.001, 0.30)$ for SUCBVI. MVP-V + filter uses $(3 \times 10^{-6}, 0.30)$, and unconstrained MVP-V uses $\alpha \simeq 2.405 \times 10^{-6}$. In Q2–Q4, SVAE and its ablations use the same empirically chosen coefficients, $(\alpha, \beta) = (0.003, 0.17)$. The complete search grids, coefficients, and random seeds are included with the code.

**Implementation** For MVP-V, we treat each state–step pair as a separate state and normalize rewards by $H$. The empirical model is updated whenever a visit count doubles. We report results on the original reward scale. For paired comparisons, we compute the difference between the two methods' results for each paired seed. For navigation, we average these differences over the five seeds within each environment, then report the mean (SD) across the eight environments.