A SONG OF FIRE AND ICE: POLICY OPTIMIZATION IN STOCHASTIC AND ADVERSARIAL CMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing state-of-the-art Policy Optimization (PO) methods for Constrained Markov Decision Processes (CMDPs) assume Slater's condition and give vacuous regret-violation bounds for small slackness ρ. When Slater's condition fails, some primal-dual PO methods artificially shrink their dual spaces to achieve sublinear bounds. We feed Lyapunov drift-plus-penalty surrogate losses to a PO subroutine for adversarial unconstrained MDPs. Since the surrogates are unbounded, we keep the feasibility state outside the optimizer and normalize the surrogates with an epoch-based geometric-doubling trick. Our main algorithm, Feasible Iterative Restarts in Expectation (FIRE), is Slater-free, projection-free, has $\widetilde{\mathcal{O}}(T^{\frac{1}{2}})$ regret and strong violation in the stochastic case, and a Pareto lower bound shows that these exponents cannot be jointly improved without Slater's condition. Being an adversarial extension, In-Hindsight Constrained Exploration (ICE) has $\widetilde{\mathcal{O}}(T^{\frac{3}{4}})$ regret and weak violation for oblivious adversaries, which improve to $\widetilde{\mathcal{O}}(T^{\frac{1}{2}})$ if the prefix-regret is non-negative. We prove that ICE's reflected queue cannot certify sublinear strong violation, identifying the structure a stronger adversarial bound needs. Synthetic experiments illustrate our mechanisms and tradeoffs.

1 INTRODUCTION

Online learning in Constrained Markov Decision Processes (CMDPs) (Efroni et al., 2020; Müller et al., 2024; Wei et al., 2018) lies at the intersection of Reinforcement Learning (RL) (Sutton & Barto, 2018) and online learning with constraints (Cesa-Bianchi & Lugosi, 2006). The learner aims to sublinearly bound the *regret* and the *violation* in T, i.e., the total number of episodes. Many works dealt with *weak violation* (Efroni et al., 2020; Qiu et al., 2020), in which the positive violations are offset by negative ones (Ghosh et al., 2022; Wei et al., 2023). In contrast, *strong violation* (Stradi et al., 2025b) focuses only on positive violations. The CMDP losses and constraint violation costs are either *stochastic* or *adversarial*. Numerous methods focus on stochastic losses and constraints (Zheng & Ratliff, 2020; Efroni et al., 2020) or adversarial losses with stochastic constraints (Wei et al., 2018; Qiu et al., 2020), and handling adversarial constraints is less common. Germano et al. (2023) and Stradi et al. (2024b) first provided sublinear regret and violation bounds for adversarial constraints, but with Slater's condition.

In *full feedback*–the loss and costs for all state-action pairs are revealed when an episode ends (Wei et al., 2018; Qiu et al., 2020)–rather than *bandit feedback* (Müller et al., 2023; Müller et al., 2024), where the loss and costs for only the visited state-action pairs are revealed. Among approaches that sublinearly bound regret and violation (e.g., linear programming (LP) (Efroni et al., 2020) and upper confidence (Zheng & Ratliff, 2020)), *primal-dual* methods (Stradi et al., 2024a;b; 2025a; Müller et al., 2024) are popular. However, these methods rely on Slater's condition, which assumes a policy π° exists that satisfies all constraints with at least $\rho > 0$ slack (Stradi et al., 2025b; Germano et al., 2023). The bounds of such algorithms scale with $1/\rho$, leading to vacuous and huge suboptimal results if ρ is small. Moreover, this condition can fail even when the CMDP is feasible.

Utilizing *Occupancy Measures (OM)* (Altman, 1999; Rosenberg & Mansour, 2019b) to tackle CMDPs requires solving a global optimization problem (Stradi et al., 2025c; 2026; Zhu et al., 2025), and OM techniques are as computationally expensive as directly solving the CMDP via LP (Müller et al., 2024). Although *Policy Optimization (PO)* methods (Stradi et al., 2025a;d;e) directly optimize over the policy space (i.e., locally update the policy at each state), few works focus on PO. To date, every

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.

Table 1: Our results (boldfaced entries) versus five **S**tate-**O**f-**T**he-**A**rt (SOTA) methods. A method is *projection-free* if no episode projects onto, or optimizes over, a set of occupancy measures or policies.

SOTA Work	Approach	Regime	Regret/Violation	Violation Type	Benchmark	Slater-free	Projection-free
Stradi et al. (2025c)	OM	Stochastic	$\widetilde{\mathcal{O}}(T^{1/2})/\widetilde{\mathcal{O}}(T^{1/2})$	Strong	Not Weakened	Yes	No
Stradi et al. (2025c)	OM	Adversarial	$\widetilde{\mathcal{O}}(T^{1/2})/\widetilde{\mathcal{O}}(T^{1/2})$	Weak	Weakened	Yes	No
Roy & Das (2026)	OM	Adversarial	$\widetilde{\mathcal{O}}(T^{1/2})/\widetilde{\mathcal{O}}(T^{1/2})$	Strong	Not Weakened	Yes	No
Stradi et al. (2025a)	PO	Stochastic	$\widetilde{\mathcal{O}}(T^{1/2})/\widetilde{\mathcal{O}}(T^{1/2})$	Strong	Not Weakened	No	Yes
Stradi et al. (2025d)	PO	Stochastic	$\widetilde{\mathcal{O}}(T^{3/4})/\widetilde{\mathcal{O}}(T^{3/4})$	Weak	Not Weakened	No	Yes
Stradi et al. (2025d)	PO	Adversarial	$\widetilde{\mathcal{O}}(T^{3/4})/\widetilde{\mathcal{O}}(T^{3/4})$	Weak	Not Weakened	No	Yes
Stradi et al. (2025e)	PO	Stochastic	$\widetilde{\mathcal{O}}(T^{3/4})/\widetilde{\mathcal{O}}(T^{3/4})$	Weak	Not Weakened	No	Yes
Stradi et al. (2025e)	PO	Adversarial	$\widetilde{\mathcal{O}}(T^{3/4})/\widetilde{\mathcal{O}}(T^{3/4})$	Weak	Not Weakened	No	Yes
FIRE (ours)	**PO**	**Stochastic**	$\boldsymbol{\widetilde{\mathcal{O}}(T^{1/2})/\widetilde{\mathcal{O}}(T^{1/2})}$	**Strong**	**Not Weakened**	**Yes**	**Yes**
General ICE (ours)	**PO**	**Adversarial**	$\boldsymbol{\widetilde{\mathcal{O}}(T^{3/4})/\widetilde{\mathcal{O}}(T^{3/4})}$	**Weak**	**Not Weakened**	**Yes**	**Yes**
Conditional ICE (ours)	**PO**	**Adversarial**	$\boldsymbol{\widetilde{\mathcal{O}}(T^{1/2})/\widetilde{\mathcal{O}}(T^{1/2})}$	**Weak**	**Not Weakened**	**Yes**	**Yes**

PO-based work assumes Slater's condition and involves a $\rho > 0$ slack. Stradi et al. (2025c) provides sublinear bounds without Slater's condition in the stochastic case by employing OM, but significantly weakens the regret benchmark by $\rho/1+\rho$ and considers weak violation in the adversarial case. Roy & Das (2026) sublinearly bounded the regret and strong violation in the adversarial case without Slater's condition by reducing the CMDP to a Constrained Online Convex Optimization (COCO) (Guo et al., 2022) problem. However, their approach was OM-based, computed a Euclidean projection in every episode, and excluded the stochastic setting entirely.

Our Contributions: On viewing the limitations of prior work, we see a gap: *"With bandit feedback and unknown transitions, can we have Slater-free, projection-free PO algorithms for stochastic and adversarial CMDPs?"*. We answer this question positively and list our contributions below.

- We map violations to a persistent virtual queue with a Lyapunov drift-plus-penalty approach and use the `PO-DB` algorithm (Luo et al., 2021) as the regret minimizer. The surrogates are unbounded, whereas `PO-DB` requires bounded losses. Therefore, we keep the feasibility state outside the optimizer, normalize the surrogates on geometrically doubling scales, and restart only the optimizer (Section 3). Unlike the OM-based Stradi et al. (2025c) and Roy & Das (2026), we never project onto occupancy measures.
- Unlike Stradi et al. (2025a), **F**easible **I**terative **R**estarts in **E**xpectation (FIRE) has a Slater-free $\widetilde{\mathcal{O}}(T^{1/2})$ regret and strong violation (Section 4.2). Its regret order is minimax-optimal via the unconstrained bandit-MDP subclass, and its $T^{1/2}$ regret and strong-violation orders are jointly Pareto optimal (Section 4.3).
- **I**n-Hindsight **C**onstrained **E**xploration (ICE) has $\widetilde{\mathcal{O}}(T^{3/4})$ regret and weak violation for oblivious adversaries on the unweakened benchmark, unlike Stradi et al. (2025c). If the prefix non-negative regret holds, ICE delivers $\widetilde{\mathcal{O}}(T^{1/2})$ bounds (Section 5.1). The PO methods of Stradi et al. (2025d;e) show $\widetilde{\mathcal{O}}(T^{3/4})$ bounds in both regimes if $\rho \geq T^{-1/8}H\sqrt{112}$ fails and cap their dual spaces to $[0, T^{1/4}]$. FIRE and ICE have ρ-free bounds without caps.
- We prove that a uniformly bounded reflected queue can coexist with linear strong violation. Hence, ICE's weak-violation guarantee reflects the mechanism, not the analysis. We also show that the prefix non-negative regret criterion precisely separates ICE's $T^{3/4}$ and $T^{1/2}$ rates (Section 5.2).

Table 1 compares FIRE and ICE with SOTA methods. We have more related work in Appendix A, revise `PO-DB` in Appendix B, and give all proofs in Appendices C, D, and E. Unlike the SOTA (Stradi et al., 2025a;d;e), Section 6 demonstrates executable implementations and loss-feasibility tradeoffs. Appendix F contains more experimental details and results.

2 Problem Setting

Throughout the paper, for any $n \in \mathbb{N}$ and $z \in \mathbb{R}$, we define $[n] \equiv \{1, \ldots, n\}$, $[n]^{-1} \equiv \{0, \ldots, n-1\}$, and $[z]^+$ or $[z]_+ \equiv \max(0, z)$. Let $\Delta(\mathcal{B})$ be the probability simplex on $\mathcal{B}$, and let $d = |X||A|H$.

Constrained Markov Decision Process: It is defined as $\mathcal{M} = (T, H, X, A, P, \{\boldsymbol{\ell}_t, \boldsymbol{c}_t\}_{t=1}^T)$ (Altman, 1999), where T is the total number of episodes; $H > 0$ is the finite horizon, i.e., the length of each episode; X and A are finite state and action spaces, respectively; $P : X \times A \times X \to [0, 1]$ is an unknown transition function; $\boldsymbol{\ell}_t \in [0, 1]^d$ is a loss vector, whose component $\ell_{t,h}(x, a)$ encodes the learner's loss when taking $a \in A$ in $x \in X$; $\boldsymbol{c}_t \in [-1, 1]^d$ is a cost vector, whose component $c_{t,h}(x, a)$ denotes the learner's constraint violation cost on taking $a \in A$ in $x \in X$. When playing action $a \in A$, $P(x' \mid x, a)$ is the probability of transitioning from $x \in X$ to the next state $x' \in X$. Without loss of generality, we assume that X is partitioned into layers $X_0, \ldots, X_H$, such that $X_0 = \{x_0\}$ and $X_H = \{x_H\}$. Also, $P(x' \mid x, a) > 0$ only if $x' \in X_{h+1}$ and $x \in X_h$ for some $h \in [H]^{-1}$.

Value Functions: Let the policy $\pi = \{\pi_t\}_{t=1}^T$ be a collection with $\pi_t(\cdot \mid x) \in \Delta(A)$ for $x \in X$. The layer-wise occupancy measure (Altman, 1999; Rosenberg & Mansour, 2019b) induced by π is given as $q_{t,h}^{\pi,P}(x, a) \coloneqq \Pr(x_h = x, a_h = a)$ with $\sum_{(x,a,h)} q_{t,h}^{\pi,P}(x, a) = H$. Given any vector $\boldsymbol{y}$, we define a V-function $\forall h \in [H]^{-1}$ and $\forall x \in X_h$ as $V^{\pi,P}(\boldsymbol{y}) \coloneqq \mathbb{E}_{\pi,P}[\sum_{h=0}^{H-1} y_h(x_h, a_h)] = \langle q^{\pi,P}, \boldsymbol{y} \rangle$, where the $y_h(x_h, a_h)$'s are the components of $\boldsymbol{y}$. Thus, we have $V^{\pi,P}(\boldsymbol{\ell}_t) \in [0, H]$ and $V^{\pi,P}(\boldsymbol{c}_t) \in [-H, H]$. When clear from the context, we omit writing π and P in $V^{\pi,P}(\boldsymbol{y})$ and $q^{\pi,P}$.

Learning Protocol: The learning procedure is repeated over T episodes, where the learner chooses a policy π_t for every $t \in [T]$. Then, depending on the CMDP type, we choose the vectors $\boldsymbol{\ell}_t$ and $\boldsymbol{c}_t$ appropriately. We always have bandit feedback and break ties by choosing the lexicographically smallest optimal occupancy under a fixed ordering.

2.1 Stochastic Regime

Let $\boldsymbol{\ell}_t$ and $\boldsymbol{c}_t$ be sampled from stationary distributions. The pairs $(\boldsymbol{\ell}_t, \boldsymbol{c}_t)$ are i.i.d. across episodes with means $\boldsymbol{\ell}$ and $\boldsymbol{c}$, and independent of the learner's internal randomness and of the transition draws in every episode, including episode t. Hence, for the pre-vector filtration $\mathcal{G}_t$ (given in Equation 28), we have $\mathbb{E}[\boldsymbol{\ell}_t \mid \mathcal{G}_t] = \boldsymbol{\ell}$ and $\mathbb{E}[\boldsymbol{c}_t \mid \mathcal{G}_t] = \boldsymbol{c}$. Given $\mathcal{G}_t$ and $(\boldsymbol{\ell}_t, \boldsymbol{c}_t)$, the episode-$t$ trajectory is generated by π_t and P. We define the non-empty feasible set $\Pi_{\text{stoc}}^\star \coloneqq \{\pi \mid V^{\pi,P}(\boldsymbol{c}) \leq 0\}$ as consisting of all policies that satisfy the constraints in expectation. Using deterministic tie-breaking, we fix $\pi_{\text{stoc}}^\star \in \arg\min_{\pi \in \Pi_{\text{stoc}}^\star} V^{\pi,P}(\boldsymbol{\ell})$. Thus, the regret and strong violation over T episodes are

$$\mathfrak{R}_T^{\text{stoc}} \coloneqq \sum_{t=1}^T V^{\pi_t,P}(\boldsymbol{\ell}) - \sum_{t=1}^T V^{\pi_{\text{stoc}}^\star,P}(\boldsymbol{\ell}), \text{ and } \mathfrak{V}_T^{\text{stoc}} \coloneqq \sum_{t=1}^T \left[V^{\pi_t,P}(\boldsymbol{c})\right]^+. \tag{1}$$

2.2 Adversarial Regime

With bandit feedback in adversarial CMDPs, no algorithm can attain sublinear regret and violation against a best-in-hindsight policy that satisfies the constraints on average (Mannor et al., 2009). Thus, it is standard (Stradi et al., 2025d;e) to let the non-empty feasible set have all fixed policies in hindsight that satisfy the constraints in every episode: $\Pi_{\text{adv}}^\star \coloneqq \{\pi \in \Pi \mid V^{\pi,P}(\boldsymbol{c}_t) \leq 0,\ \forall t \in [T]\}$. Before the learner's interaction with the CMDP, the adversary independently fixes the entire sequence $\{\boldsymbol{\ell}_t, \boldsymbol{c}_t\}_{t=1}^T$. Again with deterministic tie-breaking, we fix $\pi_{\text{adv}}^\star = \arg\min_{\pi \in \Pi_{\text{adv}}^\star} \sum_{t=1}^T V^{\pi,P}(\boldsymbol{\ell}_t)$, and the regret and weak violation over T episodes are

$$\mathfrak{R}_T^{\text{adv}} \coloneqq \sum_{t=1}^T V^{\pi_t,P}(\boldsymbol{\ell}_t) - \sum_{t=1}^T V^{\pi_{\text{adv}}^\star,P}(\boldsymbol{\ell}_t), \text{ and } \mathfrak{V}_T^{\text{adv}} \coloneqq \left[\sum_{t=1}^T V^{\pi_t,P}(\boldsymbol{c}_t)\right]^+. \tag{2}$$

3 Proposed Methodology

Figure 1 presents our solution workflow. We treat the CMDP instance as an unconstrained adversarial MDP problem using the components below. Also, we maintain a feasibility queue and restart the policy optimizer only when the surrogate range exceeds its current scale.

Regime-Specific Cost Evaluation: We utilize g_t as a computable proxy for the expected episodic constraint cost of π_t. In the stochastic case, g_t is a predictable lower confidence value for $V^{\pi_t,P}(\boldsymbol{c})$. In the adversarial case, g_t is the realized trajectory cost and satisfies $\mathbb{E}[g_t \mid \mathcal{G}_t^{\text{I}}] = V^{\pi_t,P}(\boldsymbol{c}_t)$ (the filtration $\mathcal{G}_t^{\text{I}}$ is defined in Equation 8).

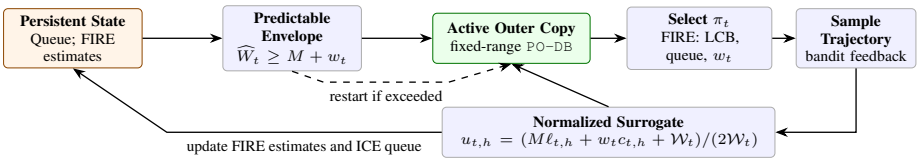


Figure 1: Depiction of our approach. The common wrapper retains the queue and FIRE estimates when an outer optimizer copy is restarted.

Virtual Queue Dynamics: Inspired by Sinha & Vaze (2024), we maintain a persistent virtual queue ζ_t to track the cumulative constraint violations across the episodes, with $\zeta_0 = 0$. ζ_t evolves as $\zeta_t \coloneqq \zeta_{t-1} + [g_t]^+$ and $\zeta_t \coloneqq [\zeta_{t-1} + g_t]^+$ in the stochastic and adversarial cases, respectively.

Surrogate Loss and Predictable Envelope: Let $\varphi(\cdot)$ be a non-decreasing and convex Lyapunov function. At episode t, let $M > 0$ scale $\boldsymbol{\ell}_t$ and let $w_t \geq 0$ be a queue multiplier fixed before the current trajectory is sampled. From the drift-plus-penalty framework of Neely (2010), we build a surrogate loss $\boldsymbol{s}_t$ by adding $M\boldsymbol{\ell}_t$ with the cost vector $\boldsymbol{c}_t$, weighted by the multiplier $w_t = \varphi'(\cdot)$. Also, w_t uses ζ_t with an indicator in the stochastic case, and ζ_{t-1} in the adversarial case. Thus, we have

$$\boldsymbol{s}_t \coloneqq M\boldsymbol{\ell}_t + w_t\boldsymbol{c}_t;\ V^{\pi_t,P}(\boldsymbol{s}_t) \coloneqq MV^{\pi_t,P}(\boldsymbol{\ell}_t) + w_t V^{\pi_t,P}(\boldsymbol{c}_t);\ \langle q^{\pi_t,P}, \boldsymbol{s}_t\rangle = V^{\pi_t,P}(\boldsymbol{s}_t). \quad (3)$$

FIRE uses $w_t = \varphi'(\zeta_t)\mathbf{1}\{g_t > 0\}$ and ICE uses $w_t = \varphi'(\zeta_{t-1})$. Every coordinate of $\boldsymbol{s}_t$ lies in the range $[-w_t, M + w_t]$. Before requesting π_t, the wrapper computes a predictable envelope $\widehat{W}_t \geq M + w_t$. FIRE uses a one-step-ahead queue bound because it finalizes its current multiplier after selecting π_t, and ICE uses $\widehat{W}_t = M + w_t$ directly.

Scale Interval and Outer Copy: We maintain levels by doubling the guesses $W^{[k+1]} = 2W^{[k]}$. If $\widehat{W}_t > W^{[k]}$, we increase the level until it covers the envelope and start a fresh PO-DB instance. We define a *scale interval* as the maximal consecutive block at one active level, and its optimizer instance as an *outer copy*. We reset that copy's policy weights, bonuses, and counts, retaining the global queue and FIRE estimates. These restarts differ from PO-DB's internal count epochs, which refresh its transition confidence sets. We write k_t for the active level, $\mathcal{W}_t = W^{[k_t]}$, ϱ_t for the nonempty-copy ordinal, and ω_r for copy r's scale. Since a jump can skip levels, ω_r need not equal $W^{[r]}$.

Normalization and Local Prefix: Since $|\boldsymbol{s}_t| \leq M + w_t \leq \mathcal{W}_t$ coordinatewise, we supply the visited coordinates of $\boldsymbol{u}_t = (\boldsymbol{s}_t + \mathcal{W}_t \cdot \mathbf{1})/(2\mathcal{W}_t) \in [0,1]^d$ to the active copy, where $\mathbf{1}$ is the vector with all ones. Since every occupancy has mass H, we have $V^{\pi,P}(\boldsymbol{u}_t) - V^{\pi',P}(\boldsymbol{u}_t) = {}^{V^{\pi,P}(\boldsymbol{s}_t) - V^{\pi',P}(\boldsymbol{s}_t)}/_{2\mathcal{W}_t}$, for any two policies π and π'. A *local prefix* of length m consists of the first m episodes of one outer copy. We retain the global T in every copy's learning rate and confidence bounds.

Proposition 3.1 (Local-Prefix Guarantee)**.** *We suppose that $S = |X|$, $D = S|A|$, $L = \log({}^{TD}/_{\varepsilon})$, and $K = \log|A| + \log({}^{1}/_{\varepsilon}) + \frac{1}{2}\log H$. Exact tabular PO-DB (Luo et al., 2021) is run with the learning rate $\eta_{\mathrm{PO}} = \min\{{}^{1}/_{(24H^3)}, {}^{1}/_{\sqrt{DHT}}\}$. We define $C_{\mathrm{opt}} = H\sqrt{DH}\max\{K, 18\}$, $C_{\mathrm{exp}} = 480HS\sqrt{SLDH}$, and $C_{\mathrm{add}} = 24H^4K + 240HSD\sqrt{SL}\log(D/\varepsilon)$. Conditional on a copy's starting history, for any fixed comparator π and deterministic $m \leq T$, with probability at least $1 - 8\varepsilon$:*

$$\sum_{i=1}^{m}\left[V^{\pi_i,P}(\boldsymbol{u}_i) - V^{\pi,P}(\boldsymbol{u}_i)\right] \leq C_{\mathrm{opt}}(\sqrt{T} + m/\sqrt{T}) + C_{\mathrm{exp}}\sqrt{m} + C_{\mathrm{add}}. \quad (4)$$

The current loss vector must be fixed before its trajectory (Appendix C.1). Appendices C.2 and C.2.1 prove Proposition 3.1 with exact numerical coefficients, which require no unknown model parameters and grow only poly-logarithmically with T. However, it depends conservatively on H, S, and $|A|$ (Section 6 measures the consequence). The wrapper's confidence allocation uses the 9ε failure bound.

Lemma 3.2. *We fix a comparator π before interaction and define $B_{\mathrm{PO}} \coloneqq 12C_{\mathrm{opt}} + \frac{8}{\sqrt{3}}C_{\mathrm{exp}} + {}^{8C_{\mathrm{add}}}/_{\sqrt{T}}$. For any $\delta \in (0,1)$, we initialize each outer copy with $\varepsilon_{\mathrm{PO}} = {}^{\delta}/_{27T^2}$ and let $W_{j',\max} \coloneqq \max\{W^{[1]}, \max_{t \leq j'} \widehat{W}_t\}$. With probability at least $1 - {}^{\delta}/_{3}$, we have simultaneously $\forall j' \leq T$,*

$$\sum_{t=1}^{j'}\left[V^{\pi_t,P}(\boldsymbol{s}_t) - V^{\pi,P}(\boldsymbol{s}_t)\right] \leq B_{\mathrm{PO}}W_{j',\max}\sqrt{T}. \quad (5)$$

We prove the above lemma in Appendix C.4 by union-bounding over deterministic copy indices and local prefix lengths before inserting the realized restart times. Then, we undo normalization and sum the increasing scales (Appendices C.3 and C.4).

Why not a standard drift-plus-penalty reduction? Neely (2010) and Sinha & Vaze (2024) feed the surrogate to a scale-adaptive learner, which absorbs a growing multiplier. Roy & Das (2026) does the same at the price of a projection per episode. Bandit `PO-DB` is not scale-free as it needs losses in $[0, 1]$. The growing surrogate range is unknown in advance, and per-episode rescaling would control only a regret with time-varying weights. We bridge this gap by (i) keeping the feasibility state outside the optimizer, so the drift telescopes over the whole horizon while `PO-DB` is certified copy by copy; (ii) normalizing on geometric levels that stay fixed within each outer copy, so the restarted copies' certificates sum to $B_{\mathrm{PO}} W_{j',\max}\sqrt{T}$ (Lemma 3.2). Thus, the unknown surrogate range costs only a constant factor, regardless of the number of copies, for any ratio $\chi > 1$ (Theorem E.10).

4 FIRE: FOR THE STOCHASTIC REGIME

Throughout this section, we consider the stochastic setup of Section 2.1. We write $q_t = q^{\pi_t, P}$ and $q^\star = q^{\pi^\star_{\mathrm{stoc}}, P}$. The deterministic tie-breaking makes $q^\star$ fixed by the population instance.

Confidence Allocation: We fix the desired final failure probability $\delta \in (0, 1)$. Before the learner's interaction with the CMDP, we deterministically set $\delta_{\mathrm{stat}} = \delta/9$, $\varepsilon_{\mathrm{PO}} = \delta/27T^2$, and $\delta_{\mathrm{pop}} = \delta/3T^2$. The statistical event is the intersection of simultaneous cost-mean confidence, transition confidence, and the predictable inverse-count bound. Since each event has failure probability at most δ_{stat}, their total budget is $\delta/3$.

4.1 EMPIRICAL ESTIMATES AND LOWER CONFIDENCE VALUES

Let $N_t(x, a)$ and $N_t(x, a, x')$ respectively be the visitation counts for (x, a) and (x, a, x') through episode $t \in [T]$. We define the estimated cost $\widehat{c}_t(x, a) = \sum_{\tau=1}^{t} c_\tau(x,a)\mathbf{1}_\tau(x,a)/N_t(x,a)$ and the estimated transition $\widehat{P}_t(x' \mid x, a) = N_t(x,a,x')/N_t(x,a)$, for $N_t(x, a) > 0$. When $N_t(x, a) = 0$, we set $\widehat{c}_t(x, a) = 0$ and set the entire row $\widehat{P}_t(\cdot \mid x, a)$ to an arbitrary fixed probability distribution supported on X_{h+1} when $x \in X_h$. Let $\widehat{\boldsymbol{c}}_t$ be the vector whose components are the $\widehat{c}_t(x, a)$'s.

We denote two constants $d_{\mathrm{sa}} = |X||A|$ and $d_{\mathrm{tr}} = |X|^2|A|$ for our convenience. Given any confidence input, $\eta \in (0, 1)$, we define $L_c(\eta) = \log 2d_{\mathrm{sa}}T/\eta$ and $L_P(\eta) = \log 2d_{\mathrm{tr}}T/\eta$. Since the costs lie in $[-1, 1]$, with probability at least $1 - \eta$, we have $|\widehat{c}_t(x, a) - c(x, a)| \le \xi_t(x, a)$, where $\xi_t(x, a) = \min\{2, \sqrt{2L_c(\eta)/1 \vee N_t(x,a)}\}$ (proof in Appendix D.3). Additionally, we use the completely explicit coordinate-wise Hoeffding radius as $\epsilon_t(x, a, x') = \min\{1, \sqrt{L_P(\eta)/2N_t(x,a)}\}$ when $N_t(x, a) \ge 1$, and $\epsilon_t(x, a, x') = 1$ if $N_t(x, a) = 0$. Hence, we have $|\widehat{P}_t(x' \mid x, a) - P(x' \mid x, a)| \le \epsilon_t(x, a, x')$ (proof in Appendix D.4), with probability at least $1 - \eta$. For any structurally known transition row, including a singleton successor layer, we set $\widehat{P}_t = P$ and $\epsilon_t = 0$ on that row. We apply the same convention to all rows in the known-transition experimental control. The visit-count estimates $\widehat{\boldsymbol{c}}_t$, $\widehat{P}_t$, and radii are classic confidence-based RL tools that follow standard optimistic MDP constructions (Rosenberg & Mansour, 2019b; Jin et al., 2020).

We define a purely empirical, step-wise confidence bonus as $b_t(x, a) \coloneqq \min\{2 + 2H, \xi_t(x, a) + H \sum_{x'} \epsilon_t(x, a, x')\}$. Let $\boldsymbol{b}_t$ be the vector whose components are the $b_t(x, a)$'s. For any policy π, we define the exactly computable confidence width $\beta_t(\pi)$ as

$$\beta_t(\pi) \coloneqq V^{\pi, \widehat{P}_t}(\boldsymbol{b}_t) = \mathbb{E}_{\pi, \widehat{P}_t}\left[\sum_{h=0}^{H-1} b_t(x_h, a_h)\right]. \tag{6}$$

For our purposes, we adapt the classic Performance Difference Lemma (Cai et al., 2020; Kakade & Langford, 2002) to compare value functions across different transition functions. We state and prove this formally in Lemma D.1 (in Appendix D.2).

For all $t \ge 2$, we let $\widehat{V}_t \coloneqq V^{\pi_t, \widehat{P}_{t-1}}(\widehat{\boldsymbol{c}}_{t-1})$, and define the predictable *Lower Confidence Bound (LCB)*, g_t, of the expected constraint cost for π_t as

$$g_t \coloneqq \widehat{V}_t - \beta_{t-1}(\pi_t), \tag{7}$$

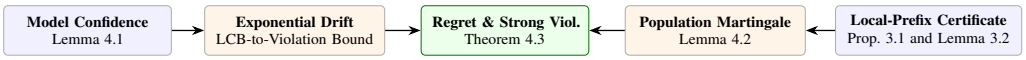


Figure 2: Proof roadmap of Theorem 4.3. The LCB queue controls strong violation. The certified optimizer and a population martingale bound the surrogate regret.

We initialize $g_1 = 0$, and it absolutely need not be an LCB. For $t \geq 2$, $\beta_{t-1}(\pi_t)$ bounds the model error and ensures $g_t \leq V^{\pi_t,P}(\boldsymbol{c})$ on the simultaneous confidence event. We formalize this in Lemma 4.1 (proof in Appendix D.7). We derive the helper results, i.e., the occupancy simulation inequality and the predictable inverse counts, in Appendices D.5 and D.6.

Lemma 4.1 (Value Confidence and Width Sum)**.** *We suppose $\eta = \delta_{\mathrm{stat}}$. There exists a deterministic $\Gamma_\beta = \widetilde{\mathcal{O}}(\sqrt{T})$ (in Equation 34). With probability at least $1-3\delta_{\mathrm{stat}}$, for every $t \geq 2$, we simultaneously have $|\widehat{V}_t - V^{\pi_t,P}(\boldsymbol{c})| \leq \beta_{t-1}(\pi_t)$, $g_t \leq V^{\pi_t,P}(\boldsymbol{c}) \leq g_t + 2\beta_{t-1}(\pi_t)$, and $\sum_{t=2}^{T} \beta_{t-1}(\pi_t) \leq \Gamma_\beta$.*

4.2 Bounding the Regret and Strong Violation

We describe our method in Algorithm 2, ***F**easible **I**terative **R**estarts in **E**xpectation (FIRE)*. Let $\varphi(z) = e^{\lambda z} - 1$, $I_t = \mathbf{1}\{g_t > 0\}$, $\zeta_t = \zeta_{t-1} + [g_t]^+$, and $w_t = \varphi'(\zeta_t) I_t = \lambda e^{\lambda \zeta_t} I_t$. Appendix D.1 defines the filtration and proves that q_t and w_t are fixed before the fresh stochastic vectors appear. The *realized surrogates* $\boldsymbol{s}_t$ are in Equation 3 and let the *population surrogates* be $\bar{\boldsymbol{s}}_t = M\boldsymbol{\ell} + w_t \boldsymbol{c}$. Therefater, we write $S_t(\pi) = \langle q^{\pi,P}, \boldsymbol{s}_t \rangle$ and $\bar{S}_t(\pi) = \langle q^{\pi,P}, \bar{\boldsymbol{s}}_t \rangle$.

Let C_{opt}, C_{exp}, and C_{add} be the coefficients in Proposition 3.1, evaluated at $\varepsilon_{\mathrm{PO}}$. We define the variables $L_Z = \log {6T^2}/{\delta}$, $B_Z = 16\sqrt{2} \cdot H\sqrt{L_Z} + {32HL_Z}/{3\sqrt{T}}$, and $B_F = B_{\mathrm{PO}} + B_Z$, which is deterministic and known before interaction. Other important constants are $M = 1$, $\lambda = {1}/{H+2eB_F\sqrt{T}}$, and $W^{[1]} = M + \lambda e^{\lambda H}$. For $t \geq 2$, we know that $\widehat{V}_t \leq H$ and $\beta_{t-1}(\pi_t) \geq 0$, so we have $g_t \leq H$. Hence, we see $\zeta_t \leq \zeta_{t-1} + H$ and $M + w_t \leq M + \lambda e^{\lambda(\zeta_{t-1}+H)} = \widehat{W}_t \leq W^{[k_t]}$. Because the components $s_{t,h} \in [-w_t, M + w_t]$, we have $u_{t,h} \in [0,1]$ under the normalization in Section 3.

Lemma 4.2 (Population Surrogate Regret)**.** *With probability at least $1 - {2\delta}/{3}$ and $\overline{W}_T = M + \lambda e^{\lambda(\zeta_T+H)}$, we have $\mathcal{J}_T^{\mathrm{pop}} := \sum_{t=1}^{T} \left(\bar{S}_t(\pi_t) - \bar{S}_t(\pi^\star_{\mathrm{stoc}})\right) \leq B_F \overline{W}_T \sqrt{T}$.*

Lemma 4.2 is proved in Appendix D.8. A regret decomposition inequality connects $\mathcal{J}_T^{\mathrm{pop}}$ and $\mathfrak{R}_T^{\mathrm{stoc}}$ and yields the theorem below (proof in Appendix D.9).

Theorem 4.3 (FIRE Bounds)**.** *On the event of Lemma 4.1, we have $M\mathfrak{R}_T^{\mathrm{stoc}} + e^{\lambda\zeta_T} - 1 \leq \mathcal{J}_T^{\mathrm{pop}}$. With the coefficients of Proposition 3.1, with probability at least $1-\delta$, we get $\mathfrak{R}_T^{\mathrm{stoc}} \leq B_F\sqrt{T} + 1 = \widetilde{\mathcal{O}}(\sqrt{T})$, $\zeta_T \leq {\log\left(2[B_F\sqrt{T}+1+HT]\right)}/{\lambda}$, and $\mathfrak{V}_T^{\mathrm{stoc}} \leq H + \zeta_T + 2\Gamma_\beta = \widetilde{\mathcal{O}}(\sqrt{T})$.*

A visual proof roadmap of Theorem 4.3 is in Figure 2. Explicitly, $B_F\sqrt{T} = \widetilde{\mathcal{O}}(H^{1.5}|X|^2\sqrt{|A|T} + H^4 + H|X|^{2.5}|A|)$, which is `PO-DB`'s order (Theorem B.1) up to a factor ${|X|}/{\sqrt{H}}$ in the leading term and an additive term. Also, we have $\Gamma_\beta = \widetilde{\mathcal{O}}(H^{2.5}|X|^{1.5}\sqrt{|A|T} + H^2|X|^2|A|)$ (Appendix D.9).

4.3 Optimality: A Pareto Lower Bound without Slater's

FIRE's $T^{1/2}$ regret exponent is minimax-optimal via the unconstrained bandit MDP subclass, and our result implies that its regret and strong-violation exponents are jointly Pareto-optimal.

Theorem 4.4 (Pareto Lower Bound)**.** *For every integer $T \geq 1$, there exist two stochastic CMDP instances ν_0 and ν_1, which depend only on T and satisfy: 1) Horizon $H = 1$ with one non-terminal state, two actions–a_0 (safe action) and a_1 (risky action), and known and deterministic transitions. Both instances have a non-empty feasible set but no strictly feasible policy; 2) The losses and costs are i.i.d. across episodes and lie in $[0,1]$ and $[-1,1]$, respectively; 3) Every possibly randomized online algorithm, even under full feedback, has $\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}] + 2\sqrt{T}\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] \geq (1 - {1}/{2\sqrt{2}})T$; and 4) every such algorithm satisfies $\mathbb{P}_{\nu_0}(\mathfrak{R}_T^{\mathrm{stoc}} > {T}/{2}) + \mathbb{P}_{\nu_1}(\mathfrak{V}_T^{\mathrm{stoc}} \geq {\sqrt{T}}/{4}) \geq 1 - {1}/{2\sqrt{2}}$.*

Consequently, any uniform $o(T)$ expected-regret bound forces $\Omega(\sqrt{T})$ expected violation on one of the constructed instances. With its expectation and probability corollaries, we prove Theorem 4.4 in Appendix D.10, where the risky action has identical losses but costs with means 0 and ${1}/{(2\sqrt{T})}$.



Figure 3: Proof roadmap of Theorems 5.4 and 5.6, sharing the surrogate and martingale bounds.

5 ICE: AN EXTENSION TO THE ADVERSARIAL REGIME

Given in Algorithm 3, ***I**n-Hindsight **C**onstrained **E**xploration (ICE)* ("in hindsight" refers to its benchmark) extends the wrapper of Section 3 to oblivious adversaries. Only tIt changes onlyost signal and the queue change roadmap in Figure 3). We use the setup of Section 2.2, write $q^\star = q^{\pi^\star_{\mathrm{adv}},P}$, and let $\mathcal{F}_{t-1}$ contain all observations and internal randomness through episode $t-1$. With the active scale level k_t, the pre-trajectory filtration adds the selected policy, the scale, and the previous queue:

$$\mathcal{G}^{\mathrm{I}}_t \coloneqq \mathcal{F}_{t-1} \vee \sigma(\pi_t, k_t, W^{[k_t]}, \zeta_{t-1}). \tag{8}$$

Unlike Section 4.1, $g_t \coloneqq \sum_{h=0}^{H-1} c_{t,h}(x_{t,h}, a_{t,h})$ is the *realized episodic cost*, with $\mathbb{E}[g_t \mid \mathcal{G}^{\mathrm{I}}_t] = V^{\pi_t,P}(\boldsymbol{c}_t) = \langle q_t, \boldsymbol{c}_t\rangle$ and $|g_t| \le H$. The reflected signed queue is $\zeta_t \coloneqq [\zeta_{t-1} + g_t]^+$ with $\zeta_0 = 0$, and the multiplier is $w_t \coloneqq \varphi'(\zeta_{t-1})$. Oblivious base vectors and the previous-queue multiplier make $\boldsymbol{s}_t$ predictable, so $\widehat{W}_t = M + w_t$ needs no look-ahead. The multiplier w_t and the fully normalized vector $\boldsymbol{u}_t = \frac{\boldsymbol{s}_t + W^{[k_t]}\cdot\mathbf{1}}{2W^{[k_t]}} \in [0,1]^d$ are $\mathcal{G}^{\mathrm{I}}_t$-measurable. With the coefficients of Proposition 3.1, with probability at least $1 - \delta/3$, simultaneously for every global prefix $u \le T$, we have the surrogate regret as

$$\mathcal{J}^{\mathrm{I}}_u \coloneqq \sum_{t=1}^{u} \Big[V^{\pi_t,P}(\boldsymbol{s}_t) - V^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{s}_t)\Big] = \sum_{t=1}^{u} \langle q_t - q^\star, \boldsymbol{s}_t\rangle \le B_{\mathrm{PO}} W_{\max}\sqrt{T}, \tag{9}$$

where $W_{\max} \coloneqq \max_{t\le T}(M + w_t)$, by the fixed comparator, the predictable-surrogate protocol (Appendix C), and Lemma 3.2. The queue-noise martingale difference is $D_t \coloneqq w_t(g_t - V^{\pi_t,P}(\boldsymbol{c}_t))$.

Lemma 5.1 (Uniform Queue-Noise Concentration)**.** *With probability at least $1 - \delta/3$, simultaneously for every $u \le T$, we have $\sum_{t=1}^{u} D_t \le B_{\mathrm{MD}} W_{\max}\sqrt{T}$, where $B_{\mathrm{MD}} \coloneqq 8H\sqrt{2\log\left(3T^2/\delta\right)}$.*

Lemma 5.1 is proved in Appendix E.1, and we set $B \coloneqq B_{\mathrm{PO}} + B_{\mathrm{MD}}$. On every sample path, we have $0 \le \zeta_t \le Ht$, and the number K_Q of nonempty outer copies under the quadratic potential satisfies $K_Q \le 1 + \lceil \log_2(1 + HT/M)\rceil$ (Lemma E.2). Lemma 5.2 connects the realized and the expected weak violation (proof in Appendix E.3).

Lemma 5.2 (Weak-Violation Bridge)**.** *Independent of the potential, with probability at least $1 - \delta/3$, we have $\mathfrak{V}^{\mathrm{adv}}_T \le \zeta_T + 2H\sqrt{2T\log(3/\delta)}$.*

5.1 BOUNDING THE REGRET AND WEAK VIOLATION

Conditional ICE: We assume that, almost surely under the law of ICE on the fixed instance, the actual occupancies and the same full-horizon feasible comparator satisfy, for every $u \in [T]$,

$$\mathfrak{R}^{\mathrm{adv}}_u(\pi^\star_{\mathrm{adv}}) \coloneqq \sum_{t=1}^{u} \Big[V^{\pi_t,P}(\boldsymbol{\ell}_t) - V^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{\ell}_t)\Big] = \sum_{t=1}^{u} \langle q^{\pi_t,P} - q^{\pi^\star_{\mathrm{adv}},P}, \boldsymbol{\ell}_t\rangle \ge 0. \tag{10}$$

We call Equation 10 the "Prefix Non-negative Regret" (PNR) condition. With $\varphi(z) = e^{\lambda z} - 1$, "Conditional ICE" satisfies Lemma 5.3 and Theorem 5.4 (proofs in Appendices E.4 and E.5).

Lemma 5.3 (Exponential Drift-plus-Penalty)**.** *With $0 < \lambda H \le 1$, we have $\forall t \in [T]$: $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \varphi(\zeta_t) - \varphi(\zeta_{t-1}) \le V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{eH^2\lambda^2}{2} e^{\lambda\zeta_{t-1}} + 1$.*

Theorem 5.4 (Conditional ICE Bounds)**.** *We set $M = \sqrt{T}$, $\alpha = 4B + 2eH^2$, and $\lambda = 1/\alpha\sqrt{T}$. With the coefficients of Proposition 3.1, assuming Equation 10 holds almost surely for all prefixes, with probability at least $1 - \delta$, we have: $\mathfrak{R}^{\mathrm{adv}}_T \le \widetilde{\mathcal{O}}(\sqrt{T})$ and $\mathfrak{V}^{\mathrm{adv}}_T = \widetilde{\mathcal{O}}(\sqrt{T})$.*

General ICE: Without Equation 10, we have the method "General ICE". On setting $\varphi(z) = z^2/2$ and $w_t = \zeta_{t-1}$, we propose Lemma 5.5 and Theorem 5.6 with their proofs respectively in Appendix E.6 and Appendix E.7.

Lemma 5.5 (Quadratic Drift-plus-Penalty). *For all $t \in [T]$, we have $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \frac{1}{2}\zeta_t^2 - \frac{1}{2}\zeta_{t-1}^2 \leq V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{1}{2}H^2$.*

Theorem 5.6 (General ICE Bounds). *Having $M = \sqrt{T}$ and with the coefficients of Proposition 3.1, with probability at least $1-\delta$, we have: $\mathfrak{R}_T^{\mathrm{adv}} \leq \widetilde{\mathcal{O}}(T^{3/4})$ and $\mathfrak{V}_T^{\mathrm{adv}} = \widetilde{\mathcal{O}}(T^{3/4})$.*

5.2 What Can a Reflected Queue Certify?

ICE's limitations are structural findings about the reflected-queue mechanism, not gaps in its analysis, and they identify what a stronger adversarial guarantee needs: *Strong violation needs a predictable positive-part signal*. By Proposition 5.7 (proof in Appendix E.8), a uniformly bounded reflected queue can coexist with linear strong violation.

Proposition 5.7. *We fix $Q > 0$ and $T \geq 1$. Let $g_1, \ldots, g_T \in [-Q, Q]$, set $z_0 = 0$, and recursively define $z_t = [z_{t-1} + g_t]^+$. If $z_t \leq Q$ for every $t \in [T]$, then $\sum_{t=1}^T [g_t]^+ \leq Q\lceil T/2 \rceil$. The bound is attained by $g_t = Q$ for an odd t and $g_t = -Q$ for an even t. In particular, for every even T, this attaining sequence satisfies $z_T = 0$, $[\sum_{t=1}^T g_t]^+ = 0$, and $\sum_{t=1}^T [g_t]^+ = QT/2$.*

Note that ICE need not generate the attaining sequence. For $Q \leq 1$, the attaining sequence embeds in a one-step CMDP with a zero-cost feasible action and another action of cost g_t. Always choosing the latter gives zero weak violation and $QT/2$ strong violation for even T. Hence, no analysis of the reflected queue alone can certify sublinear strong violation. Within our reduction, that would need queue increments dominating $[V^{\pi_t,P}(\boldsymbol{c}_t)]^+$ while the surrogate fed to `PO-DB` stays fixed before the trajectory (as in Appendix C.1). FIRE's predictable LCB gate provides exactly this, because a stochastic constraint is stationary and learnable across episodes. With adversarial constraints and bandit feedback, the only cost signal, g_t, arrives after the trajectory. OM-based methods use it in biased subgradients (Roy & Das, 2026), paying a projection per episode. Thereby, Roy & Das (2026) get a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and strong violation, stronger than ICE's. However, ICE's advantage is computational (Appendix F.10). Per episode, FIRE and ICE run 4–8 times faster than the OM-based WC-OPS (Stradi et al., 2025c) and BAG-U (Roy & Das, 2026) (Table 9), and FIRE's saving grows with the horizon, from about 1.1 ms at $H = 1$ to 2.2–3.8 ms at $H = 8$ (Table 10).

PNR is what separates $T^{3/4}$ from $T^{1/2}$: In the proof of Theorem 5.6, the $T^{3/4}$ rate comes only from the negative prefix regret $\mathfrak{R}_u^{\mathrm{adv}}(\pi_{\mathrm{adv}}^\star) \geq -Hu$, which ICE can incur by violating the constraints, and the PNR removes this term. Under PNR, even General ICE, unchanged and unaware of PNR, attains $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds (given in Corollary E.11). The exponential potential of Conditional ICE makes the regret linear rather than quadratic in B (shown in Corollary E.11). PNR plays the role of the non-negative-regret condition under which Sinha & Vaze (2024) sharpen their violation bound. It holds for every occupancy sequence exactly when the fixed comparator minimizes every episode's loss (described in Proposition E.9 in Appendix E.9), a simple sufficient class. Note that Theorem 5.4 requires PNR only on histories that ICE can generate.

6 Synthetic Experiments

This work is primarily theoretical, and real-world datasets or simulated environments do not exist in the relevant literature. However, we ran FIRE and ICE on small CMDPs to check that the complete wrappers execute as specified and to observe their finite-budget tradeoffs, not to test the asymptotic rates. We term the indicator I_t (in Section 4 and Algorithm 2) as FIRE's *gate*, which is open in episode t when $I_t = 1$ and closed when $I_t = 0$. Appendix F provides full details, tables, ablations, and experimental suites. Our code is zipped in the supplementary material.

Setup: In the *stochastic boundary* task, two risky actions have lower losses than the safe action but mean costs of 0.45 and 0.16. Thus, the always-safe policy is feasible, and Slater's condition fails. In the *oblivious switching* task, a fixed sequence of eight phases alternates two loss-cost profiles, each favoring a different risky action. Both tasks have $H = 2$, state-layers of sizes $(1, 2, 1)$, three actions, unknown transitions, and bandit feedback. The comparator is always safe, so a negative regret reflects violation. We run FIRE with the practical λ (see the next paragraph) and General ICE with its prescribed $M = \sqrt{T}$, at $T = 20{,}000$ on ten seeds (95% Student-t intervals). We compare our methods with the Slater-based PO method PDB-PS (Stradi et al., 2025d), the OM-based Slater-free

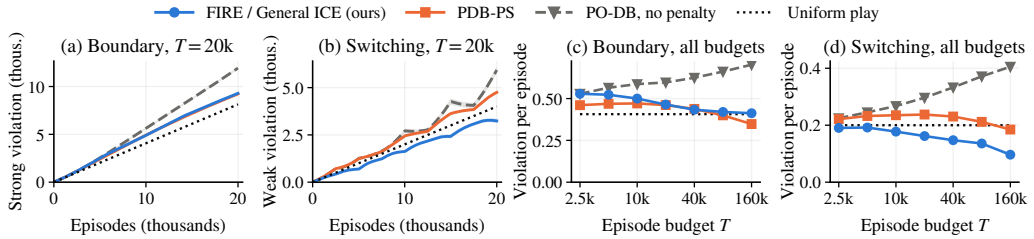


Figure 4: *(a)-(b)* Violation over the episodes at $T = 20{,}000$. *(c)-(d)* Violation per episode when all parameters are set from the budget T. Mean curves with 95% confidence over ten seeds.

WC-OPS (Stradi et al., 2025c) and BAG-U (Roy & Das, 2026), *uniform play* (i.e., picks each action with probability $1/3$), and the *matched no-penalty* PO−DB (or *control*, i.e., PO−DB on the losses alone with the initial normalization and no restarts).

Certified vs Practical parameters: Theorem 4.3 (which is a statement about rates) prescribes $\lambda = 1/(H+2eB_F\sqrt{T})$. On the boundary task at $T = 20{,}000$, the certificate gives $B_F \approx 1.54\times10^6$: 70% of which is the exploration term $8/\sqrt{3}C_{\exp}$ and 30% the $8C_{\text{add}}/\sqrt{T}$ term. Hence, $\lambda \approx 8.4\times10^{-10}$, and the multiplier $\lambda e^{\lambda\zeta_t}$ would double only after the queue grew by $\log 2/\lambda \approx 8.2\times10^8 \gg HT$. Thus, the certified penalty is inert, and the certified bounds are vacuous ($B_F\sqrt{T}+1 \approx 2.2\times10^8 \gg HT$) until $T \gtrsim 7\times10^{11}$ (regret) and 4×10^{16} (violation). Our main runs use $\lambda = \min\{1/H, 4/(H\sqrt{T})\} \approx 0.014$, which keeps the $T^{-1/2}$ scaling but replaces $2eB_F$ by $H/4$ (a factor of 1.7×10^7). It carries no theorem, since the proof of Theorem 4.3 would need $B_F \le He^{-\lambda H}/8 \approx 0.24$.

Findings: (1) Mechanisms (Figure 4a-b, Tables 2-3): FIRE's gate stays closed, and FIRE coincides with the control until the confidence width falls below the estimated cost (median 5,538 episodes). Then, its queue bends the violation downward. On the switching task, the control and PDB-PS drift toward the risky action with the lower average loss, whereas ICE's queue tilts it toward the risky action with a negative current cost. FIRE and General ICE lower the control's violation by 22% and 45% at a higher regret. FIRE and practical PDB-PS are statistically indistinguishable (paired violation difference 63 ± 75), and General ICE has $1{,}510 \pm 71$ lower weak violation than PDB-PS at $1{,}374 \pm 67$ higher regret; **(2) Budgets (Figures 4c-d, 6, Table 5):** With all parameters set from $T \in [2{,}500, 160{,}000]$, ICE's weak violation grows with exponent 0.76 ± 0.03 over $T \ge 20{,}000$ (PDB-PS: 0.88 ± 0.01), whereas its strong violation stays nearly linear (0.97 ± 0.02; Proposition 5.7). Once FIRE's gate opens, its restarts (224 at $T = 160{,}000$) keep resetting PO−DB, so its per-episode violation approaches uniform play; **(3) Penalty strength (Figure 5, Table 4)**: All methods trade regret for violation along nearly one line, and stronger penalties move FIRE and ICE only a short way, as the normalization caps the loss gaps that PO−DB sees, and certified FIRE's penalty is inert (Appendix F.4); **(4) Components (Figure 7, Table 6)**: Without its width, FIRE's gate opens at once, and FIRE restarts 161 times and stays near uniform play. The gate and the potential change its violation by less than 1%, and known transitions lower it by 9%; **(5) PNR and Horizons (Figures 8, 9, Tables 7-8)**: Under PNR, the certified Conditional ICE matches the control. For $H \in \{1, 4, 8\}$, ICE stays below the control, whereas FIRE's gate never opens for $H \ge 4$; **(6) OM baselines (Figure 10, Tables 9-10):** WC-OPS and BAG-U run 4–8 times slower than FIRE and ICE, and BAG-U's prescribed penalty is inert. WC-OPS keeps its iterates in an estimated feasible set and has 31% less strong violation than FIRE on the boundary task, but twice ICE's weak violation on the switching task. WC-OPS's regret against $\rho/1+\rho$ times the unconstrained optimum is vacuous.

7 Conclusion

Our main result, the Slater-free and projection-free FIRE, has $\widetilde{\mathcal{O}}(T^{1/2})$ regret and strong violation with jointly Pareto-optimal exponents. Its adversarial extension, ICE, has $\widetilde{\mathcal{O}}(T^{3/4})$ regret and weak violation, or $\widetilde{\mathcal{O}}(T^{1/2})$ under PNR, which exactly separates the two rates. A reflected queue cannot certify sublinear strong violation in the adversarial case, and $T^{1/2}$ rates without PNR remain open. Tighter constants, restarts that retain progress, and function approximation are potential future works.

AI USE STATEMENT

We used generative AI tools only to polish the writing and format tables. We did not use generative AI tools for any other tasks. We reviewed all AI-polished text and formatting, and we take responsibility for the final content of this work.

ETHICS STATEMENT

This work advances online learning in CMDPs. Our work may have potential societal consequences, none of which we feel must be highlighted here. Hence, we believe that this work raises no questions or concerns regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

We have made every effort to make our results reproducible. All assumptions and complete proofs are in Appendices C, D, and E, and Appendix C.2.1 gives every numerical coefficient of the optimizer certificate. Appendix F lists every instance, method, parameter, ablation, and protocol of Section 6. The zipped supplementary contains the complete code and detailed setup and execution instructions.

REFERENCES

E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Papini, Alberto Maria Metelli, and Nicola Gatti. Online learning with off-policy feedback in adversarial mdps. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/409. URL https://doi.org/10.24963/ijcai.2024/409.

Qinbo Bai, Vaneet Aggarwal, and Ather Gattami. Provably sample-efficient model-free algorithm for mdps with peak constraints. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Martino Bernasconi, Federico Cacciamani, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Safe learning in tree-form sequential decision making: Handling hard and soft constraints. In *International Conference on Machine Learning*, pp. 1854–1873. PMLR, 2022.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2767–2783. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/castiglioni22a.html.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022b. Curran Associates Inc. ISBN 9781713871088.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006. ISBN 0521841089.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pp. 3246–3270. PMLR, 2022.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, pp. 3304–3312. PMLR, 2021.

Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25900. URL https://doi.org/10.1609/aaai.v37i6.25900.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Jacopo Germano, Francesco Emanuele Stradi, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. A best-of-both-worlds algorithm for constrained mdps with long-term constraints. *arXiv e-prints*, pp. arXiv–2304, 2023.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Hengquan Guo, Honghao Wei, Xin Liu, and Lei Ying. Online convex optimization with hard constraints: towards the best of two worlds and beyond. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Shinji Ito, Kevin Jamieson, Haipeng Luo, Arnab Maiti, and Taira Tsuchiya. Adapting to stochastic and adversarial losses in episodic mdps with aggregate bandit feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010. ISSN 1532-4435.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:208624190.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Tal Lancewicki and Yishay Mansour. Near-optimal regret using policy optimization in online mdps with aggregate bandit feedback. *arXiv preprint arXiv:2502.04004*, 2025.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jordan Lekeufack and Michael I Jordan. An optimistic algorithm for online convex optimization with adversarial constraints. *arXiv preprint arXiv:2412.08060*, 2024.

Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3944–3952. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/liakopoulos19a.html.

Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

Shie Mannor, John N. Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *J. Mach. Learn. Res.*, 10:569–590, June 2009. ISSN 1532-4435.

Davide Maran, Pierriccardo Olivieri, Francesco Emanuele Stradi, Giuseppe Urso, Nicola Gatti, and Marcello Restelli. Online markov decision processes configuration with continuous decision space. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14315–14322, Mar. 2024. doi: 10.1609/aaai.v38i13.29344. URL https://ojs.aaai.org/index.php/AAAI/article/view/29344.

Adrian Müller, Pragnya Alatur, Giorgia Ramponi, and Niao He. Cancellation-free regret bounds for lagrangian approaches in constrained markov decision processes. In *Sixteenth European Workshop on Reinforcement Learning*. OpenReview, 2023.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Michael J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan and Claypool Publishers, 2010. ISBN 160845455X.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pp. 1804–1812, Red Hook, NY, USA, 2010. Curran Associates Inc.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pp. 2827–2835. PMLR, 2021.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a0872cc5b5ca4cc25076f3d868e1bdf8-Paper.pdf.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019b.

Srinjoy Roy and Swagatam Das. Optimal regret and hard violation for constrained markov decision processes with adversarial losses and constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=EsInBaX0ko.

Abhishek Sinha and Rahul Vaze. Optimal algorithms for online convex optimization with adversarial constraints. *Advances in Neural Information Processing Systems*, 37:41274–41302, 2024.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs with adversarial losses and stochastic hard constraints. In *Seventeenth European Workshop on Reinforcement Learning*, 2024a. URL https://openreview.net/forum?id=1Jl0PErLDe.

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*, ICML'24. JMLR.org, 2024b.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning constrained markov decision processes with non-stationary rewards and constraints. *CoRR*, 2024c.

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 (ICLR)*, 2025a.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, et al. Learning adversarial mdps with stochastic hard constraints. In *Forty-Second International Conference on Machine Learning*, pp. 1–8, 2025b.

Francesco Emanuele Stradi, Eleonora Fidelia Chiefari, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Beyond slater's condition in online cmdps with stochastic and adversarial constraints. *arXiv preprint arXiv:2509.20114*, 2025c.

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 *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025d.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Best-of-both-worlds policy optimization for cmdps with bandit feedback, 2025e. URL https://arxiv.org/abs/2410.02269.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Taming adversarial constraints in CMDPs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=sT9nd1WQ76.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and Xingyu Zhou. Provably efficient model-free algorithms for non-stationary cmdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 6527–6570. PMLR, 2023.

Xiaohan Wei, Hao Yu, and Michael J. Neely. Online learning in weakly coupled markov decision processes: A convergence time study. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(1), April 2018. doi: 10.1145/3179415. URL https://doi.org/10.1145/3179415.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pp. 620–629. PMLR, 2020.

Jiahui Zhu, Kihyun Yu, Dabeen Lee, Xin Liu, and Honghao Wei. An optimistic algorithm for online cmdps with anytime adversarial constraints. In *Forty-second International Conference on Machine Learning*, 2025.

A More Related Works

We divide the prior literature into three divisions and dedicate one subsection to each. First, we provide a comprehensive review of advances in online learning for CMDPs. Next, we survey notable works applying online learning to traditional MDPs. Finally, we briefly discuss relevant works on the classical problem of online learning under constraints (Cesa-Bianchi & Lugosi, 2006).

A.1 Online Learning in CMDPs

Most early literature focuses on stochastic losses and constraints. For episodic CMDPs with unknown transitions and bandit feedback, Efroni et al. (2020) applied linear programming and primal-dual methods to achieve sublinear regret and constraint violations. Assuming known transitions, Zheng & Ratliff (2020) attained $\widetilde{\mathcal{O}}(T^{3/4})$ regret and high-probability violation bounds in fully stochastic environments. Furthermore, Bai et al. (2023) established sublinear regret under deterministic rewards, unknown transitions, and peak stochastic constraints.

Restricting attention to stochastic settings, several studies (Liu et al., 2021; Müller et al., 2024; Stradi et al., 2025a) derived sublinear bounds for the strong violation of constraints. Meanwhile, weak violation guarantees have been explored through both model-free (Ghosh et al., 2022; Wei et al., 2023) and model-based (Ding et al., 2021; Chen et al., 2022) approaches. Extending beyond stochasticity, Stradi et al. (2024b) bounded the weak violation under adversarial losses. To bound the strong violation in fully stochastic scenarios, Stradi et al. (2025a) relied heavily on Slater's condition, whereas Müller et al. (2023) utilized an augmented Lagrangian approach requiring access to a strictly feasible policy.

Bridging these regimes, Germano et al. (2023) proposed a best-of-both-worlds framework that accommodates both stochastic and adversarial settings. While not fully adversarial, Ding & Lavaei (2023); Wei et al. (2023) and Stradi et al. (2024c) established guarantees for non-stationary environments, provided the variances of the losses and constraints are bounded. Very recently, the `OMDPD` algorithm (Zhu et al., 2025) successfully tackled adversarial constraints, achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation bounds without relying on Slater's condition. However, two important limitations of `OMDPD` were: first, losses were stochastic across all episodes; second, full feedback was used instead of the challenging bandit feedback under adversarial constraints.

A.2 Online Learning in MDPs

In the realm of traditional MDPs, the seminal `UCRL2` algorithm (Jaksch et al., 2010) popularized the use of confidence sets for transition estimation, achieving an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound in undiscounted settings. Transitioning to adversarial environments, Neu et al. (2010) established an $\widetilde{\mathcal{O}}(T^{2/3})$ regret bound for undiscounted MDPs against oblivious adversaries. Later, Rosenberg & Mansour (2019b) applied entropic regularization to episodic MDPs with unknown transitions and full feedback, securing an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound for adversarial losses.

When restricted to bandit feedback, early efforts by Rosenberg & Mansour (2019a) attained $\widetilde{\mathcal{O}}(T^{3/4})$ regret in episodic adversarial MDPs with unknown transitions. This was subsequently improved by the elegant `UOB-REPS` algorithm (Jin et al., 2020), which was the first to achieve the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in this challenging setup. Further advancements include the work of Lee et al. (2020), who provided data-dependent, high-probability $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds against adaptive adversaries using simple learning-rate schedules and standard unbiased estimators. Additional specialized settings have also been explored, including off-policy regret bounds for adversarial MDPs (Bacchiocchi et al., 2024) and online configurations for MDPs with continuous decision spaces and stochastic losses (Maran et al., 2024).

Recent literature has also explored *aggregate bandit feedback*, a setting in which the learner observes only the total episodic loss, rather than individual step-wise penalties. Under this feedback model, Lancewicki & Mansour (2025) pioneered policy optimization algorithms for finite-horizon adversarial MDPs, successfully addressing both known and unknown transition dynamics, thereby improving upon prior results. Expanding on this, Ito et al. (2025) introduced the first best-of-both-worlds algorithm for aggregate feedback. For known transitions, their method achieves $\mathcal{O}(\log T)$ regret in

stochastic environments and $\mathcal{O}(\sqrt{T})$ regret against adversaries. When transitions are unknown, Ito et al. (2025) leveraged confidence-based techniques to maintain rigorous $\widetilde{\mathcal{O}}(\sqrt{T})$ regret guarantees.

A.3 Online Learning with Constraints

Liakopoulos et al. (2019) investigated long-term budget constraints chosen by an adversary; however, their regret formulation is measured against a comparator that only satisfies the budget over a fixed time window. Subsequent advancements by Castiglioni et al. (2022a) and Castiglioni et al. (2022b) introduced the first best-of-both-worlds frameworks for managing long-term constraints. Additionally, hard constraint violations have been extensively analyzed across various domains, including simple stochastic environments (Pacchiano et al., 2021; Bernasconi et al., 2022), standard Online Convex Optimization (OCO) (Guo et al., 2022), and Constrained OCO (COCO) (Sinha & Vaze, 2024). Notably, Sinha & Vaze (2024) were the first to demonstrate that an online policy in COCO can guarantee $\mathcal{O}(\sqrt{T})$ regret and $\widetilde{\mathcal{O}}(\sqrt{T})$ violation without relying on restrictive supplementary assumptions. More recently, Lekeufack & Jordan (2024) explored a scenario in which both loss predictions and constraints are available prior to decision-making. By leveraging the analytical tools developed by Sinha & Vaze (2024), Lekeufack & Jordan (2024) were able to slightly tighten the existing $\mathcal{O}(\sqrt{T})$ regret and $\widetilde{\mathcal{O}}(\sqrt{T})$ violation bounds.

B Revisiting the PO-DB Base Optimizer

Algorithm 1 Policy Optimization with Dilated Bonuses (PO-DB)

Require: $\varepsilon \in (0,1)$, $\eta_{\mathrm{PO}} = \min\{1/(24H^3), 1/\sqrt{|X||A|HT}\}$, $\gamma = 2\eta_{\mathrm{PO}}H$.

Set internal epoch $j = 1$ and $\widetilde{\mathcal{P}}_1$ to all layered transition kernels.
Initialize running counts $C(x,a)$, $C(x,a,x')$ and snapshots $C_1(x,a)$, $C_1(x,a,x')$ to zero.
for $i = 1, \dots, T$ (or *until the outer wrapper restarts this copy*) **do**
  For each state, set $\pi_i(a \mid x) \propto \exp[-\eta_{\mathrm{PO}} \sum_{\tau<i}(\widetilde{Q}_\tau(x,a) - B_\tau(x,a))]$.
  Compute $\overline{q}_i$, $\underline{q}_i$, b_i^{PO}, and B_i using π_i and $\widetilde{\mathcal{P}}_j$.
  Execute π_i and observe the visited normalized losses $l_i(x_h, a_h)$.
  Set $L_{i,h} = \sum_{v=h}^{H-1} l_i(x_v, a_v)$ and $\widetilde{Q}_i(x,a) = L_{i,h}\mathbf{1}_i(x,a)/(\overline{q}_i(x,a) + \gamma)$ for $x \in X_h$.
  Increment the running counts along the observed trajectory.
  if some visited pair satisfies $C(x_h, a_h) \geq \max\{1, 2C_j(x_h, a_h)\}$ **then**
    $j \leftarrow j + 1$; Copy all running counts into $C_j(x,a)$, $C_j(x,a,x')$.
    Form P'_j and update $\widetilde{\mathcal{P}}_j$ using Equation 11.
  end if
end for

For the reader's convenience, we briefly revisit the PO-DB algorithm from Luo et al. (2021) in Algorithm 1. Let $C_j(x,a)$ and $C_j(x,a,x')$ be respectively the cumulative visitation-count snapshots for (x,a) and (x,a,x') at the start of the j-th internal epoch. Based on the counters, the empirical transition

$$P'_j(x' \mid x, a) = \frac{C_j(x,a,x')}{\max\{1, C_j(x,a)\}}$$

is computed for positive counts. At zero counts, we use a fixed probability row supported on the next layer. Also, for epoch j, let $\widetilde{\mathcal{P}}_j$ be the confidence set of the transition P, maintained by PO-DB. Restricting all candidates $\widetilde{P}$ to layered probability kernels, we define

$$\widetilde{\mathcal{P}}_j \coloneqq \left\{ \widetilde{P} : \left| \mathcal{U}_j^{(x,a,x')} \right| \leq \kappa_j(x' \mid x, a), \forall (x,a,x') \in X_h \times A \times X_{h+1}, h = 0, \dots, H-1 \right\}, \quad (11)$$

where $\mathcal{U}_j^{(x,a,x')} = \widetilde{P}(x' \mid x, a) - P'_j(x' \mid x, a)$, and $\kappa_j(x' \mid x, a)$ is set as

$$4\sqrt{\frac{P'_j(x' \mid x, a) \ln\left(\frac{T|X||A|}{\varepsilon}\right)}{\max\{1, C_j(x,a)\}}} + \frac{28 \ln\left(\frac{T|X||A|}{\varepsilon}\right)}{3 \max\{1, C_j(x,a)\}}$$

If a transition row is known, we intersect each confidence set with that exact row. The singleton successor layer already enforces this restriction. This also specifies the known-transition experimental control. The true transition function remains in the set, and every occupancy width can only decrease, and the same certificate applies. The confidence set $\widetilde{\mathcal{P}}_j$ is updated when the visitation count of a certain (x, a)-pair is doubled, and the period between two model updates is called an *epoch*. Inside the main loop of PO−DB, bonus functions are constructed for all nonterminal state–action pairs as

$$b_t^{\mathrm{PO}}(x) = \mathbb{E}_{a\sim\pi_t(\cdot|x)}\left[\frac{3\gamma H + H(\overline{q}_t(x,a) - \underline{q}_t(x,a))}{\overline{q}_t(x,a)+\gamma}\right], \text{ and}$$

$$B_t(x,a) = b_t^{\mathrm{PO}}(x) + \left(1+\frac{1}{H}\right)\max_{\widetilde{P}\in\widetilde{\mathcal{P}}_j}\mathbb{E}_{x'\sim\widetilde{P}(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right],$$

with $\overline{q}_t(x,a) = \max_{\widetilde{P}\in\widetilde{\mathcal{P}}_j} q^{\pi_t,\widetilde{P}}(x,a)$, $\underline{q}_t(x,a) = \min_{\widetilde{P}\in\widetilde{\mathcal{P}}_j} q^{\pi_t,\widetilde{P}}(x,a)$, and $B_t(x_H,a) = 0$ for all a. In the theorem below, we state the regret bound attained by PO−DB.

Theorem B.1 (Luo et al. (2021))**.** *With probability at least $1-\mathcal{O}(\varepsilon)$, for any oblivious sequence of losses $l_t \in [0,1]^d$, any fixed policy $\pi^\star$, and any valid occupancy measure, PO−DB satisfies*

$$\sum_{t=1}^{T}\left[V^{\pi_t,P}(l_t) - V^{\pi^\star,P}(l_t)\right] \le \widetilde{\mathcal{O}}\left(H^2|X|\sqrt{|A|T} + H^4\right).$$

C Predictable Local-Prefix Interface for PO−DB

We prove a numerical local-prefix bound for exact tabular PO−DB and show how it applies at adaptive restarts. First, we condition on a copy's starting history and fix a deterministic local prefix. Then, we union-bound over copy indices and prefix lengths before inserting the realized lengths. Our explicit transition-width estimate avoids importing hidden constants from an asymptotic result.

C.1 Local Filtration and Loss Protocol

We fix one restarted copy and condition on the complete outer history at its beginning. Then, we relabel its local episodes by $i = 1, 2, \ldots$. Let $\mathscr{P}_i$ contain the conditioned starting history, all local trajectories and feedback before local episode i, and all internal random variables used by the copy before it produces π_i. After the policy is produced and the environment's normalized vector $\boldsymbol{u}_i \in [0,1]^d$ is fixed (before the trajectory is sampled), we define $\mathscr{Q}_i = \mathscr{P}_i \vee \sigma(\pi_i, \boldsymbol{u}_i)$.

The admissible protocol only requires that $\boldsymbol{u}_i$ be $\mathscr{Q}_i$-measurable and that, conditional on $\mathscr{Q}_i$, the trajectory be generated by the fixed transition function P and policy π_i. Thus, $\boldsymbol{u}_i$ may depend on the entire past and π_i, but not on the current trajectory. Both ICE and FIRE satisfy this protocol.

We suppose that $q_i = q^{\pi_i,P}$. Also, if we suppose that $(x_{i,h}, a_{i,h})$ is the visited state-action pair at layer h and $L_{i,h} = \sum_{j=h}^{H-1} u_{i,j}(x_{i,j}, a_{i,j})$, then we have

$$\mathbb{E}[\mathbf{1}\{x_{i,h} = x, a_{i,h} = a\} \mid \mathscr{Q}_i] = q_{i,h}(x,a), \text{and}$$

$$\mathbb{E}[L_{i,h} \mid x_{i,h} = x, a_{i,h} = a, \mathscr{Q}_i] = Q_h^{\pi_i,P}(\boldsymbol{u}_i; x, a).$$

Both identities follow directly from the conditional trajectory law. These are exactly the conditional-expectation identities used to analyze PO−DB's importance-weighted return estimator. Neither identity requires fixing the entire loss sequence in advance.

C.2 Prefix-Validity of the Optimizer

We retain the global T in the optimizer's learning rate and confidence parameters. The performance-difference identity and exponential-weights potential argument apply to every realized vector sequence. The stochastic steps also use only the two conditional identities above, bounded suffix returns, and adaptive visitation counts. Thus, the analysis permits current losses that depend on the selected policy, provided the trajectory has not yet been sampled. In the next subsection, we rigorously prove these statements quantitatively for each deterministic prefix. Its 8ε failure bound implies the 9ε bound used in the restart ledger.

C.2.1 Explicit Numerical Coefficients

We give conservative numerical constants for the tabular PO-DB optimizer. Importantly, we retain its published algorithm and prove a coarse transition-width estimate, rather than importing an unspecified constant from an asymptotic bound. The resulting constants preserve the dependence on the episode count up to logarithms, but have worse dependence on state and horizon dimensions than the sharp published PO-DB bound. This point is essential when using the constants to instantiate FIRE or ICE.

Dimensions and Local Protocol: We use layers $X_0, \dots, X_H$, with a fixed initial state and one terminal state. We let $S = |X|$, let $|A| \geq 1$ be the number of available actions, and use $D = S|A|$ as an upper bound on the number of non-terminal state–action pairs. All unqualified state, state–action pair, loss, count, and bonus sums exclude the terminal state. In particular, every occupancy has total mass H, and we keep $d = \sum_{h=0}^{H-1} |X_h||A|$ for the vector dimension defined in Section 2. Thus, $d \leq D$ and D is only a convenient upper bound and does not add terminal coordinates to the loss vectors. We fix integers $H, T \geq 1$, a confidence parameter $\varepsilon \in (0,1)$, and write

$$L = \log(TD/\varepsilon), \qquad K = \log|A| + \log(1/\varepsilon) + \tfrac{1}{2}\log H, \tag{12}$$

$$\eta_{\mathrm{PO}} = \min\{1/(24H^3), 1/\sqrt{DHT}\}, \qquad \gamma_{\mathrm{PO}} = 2\eta_{\mathrm{PO}} H. \tag{13}$$

The normalized loss vector $\boldsymbol{u}_t \in [0,1]^d$ is fixed before the episode's trajectory. Conditional on the past, selected policy, and this vector, the trajectory is generated using the true transition P and policy π_t. The vector may depend on the past and on π_t. We fix a comparator policy π before the copy begins, so it is measurable with respect to the conditioned starting history. We abbreviate $q_t = q^{\pi_t,P}$ and $q^\pi = q^{\pi,P}$, and write $q_t(x) = \sum_a q_t(x,a)$ and $q^{\pi(x)} = \sum_a q^\pi(x,a)$. We define the local realized-vector regret by

$$R_m^{\mathrm{PO}} = \sum_{t=1}^{m} \left[V^{\pi_t,P}(\boldsymbol{u}_t) - V^{\pi,P}(\boldsymbol{u}_t)\right] = \sum_{t=1}^{m} \langle q_t - q^\pi, \boldsymbol{u}_t \rangle.$$

All statements below concern a deterministic local prefix $m \leq T$. The algorithm retains the global T in Equation 12 and Equation 13.

Theorem C.1. *We define $C_{\mathrm{opt}} = H\sqrt{DH}\max\{K, 18\}$, $C_{\mathrm{exp}} = 480HS\sqrt{SLDH}$, and $C_{\mathrm{add}} = 24H^4K + 240HSD\sqrt{SL}\log(D/\varepsilon)$. We run exact tabular PO-DB with the parameters in Equation 13. For any fixed comparator π and deterministic $m \leq T$, with failure probability at most 8ε, we have*

$$\sum_{t=1}^{m} \left[V^{\pi_t,P}(\boldsymbol{u}_t) - V^{\pi,P}(\boldsymbol{u}_t)\right] \leq C_{\mathrm{opt}}(\sqrt{T} + m/\sqrt{T}) + C_{\mathrm{exp}}\sqrt{m} + C_{\mathrm{add}}.$$

The coefficients are deterministic and require no knowledge of P. For fixed $H, S, |A|$, they grow only poly-logarithmically with T whenever ε^{-1} is bounded by a polynomial in T.

Proof. We use the exact tabular PO-DB algorithm of Luo et al. (2021). In this subsection, t indexes local episodes within one optimizer copy, and its counts reset at an outer restart and are distinct from FIRE's global counts. At the start of local episode t, let $n_t^{\mathrm{PO}}(x,a)$ be its internal-epoch count snapshot and let $N_t^{\mathrm{PO}}(x,a)$ be its actual count from local episodes before t. Every count trigger copies all running counts into the new snapshot. The transition confidence set $\mathcal{P}_t^{\mathrm{PO}}$ is row rectangular, consists of layered stochastic transition functions, and uses the coordinate radius

$$\kappa_t(x' \mid x,a) = 4\sqrt{\frac{\widehat{P}_t^{\mathrm{PO}}(x' \mid x,a)L}{n_t^{\mathrm{PO}}(x,a) \vee 1}} + \frac{28L}{3(n_t^{\mathrm{PO}}(x,a) \vee 1)}. \tag{14}$$

For a known row, we restrict every candidate transition function to its exact value, reducing occupancy widths. For other rows, we take any fixed probability row when the count is zero. The initial confidence set may equivalently be all layered transitions. We write $\bar{q}_t(x,a) = \max_{P' \in \mathcal{P}_t^{\mathrm{PO}}} q^{\pi_t,P'}(x,a)$, $\underline{q}_t(x,a) = \min_{P' \in \mathcal{P}_t^{\mathrm{PO}}} q^{\pi_t,P'}(x,a)$, and $\Delta_t^{\mathrm{PO}}(x,a) = \bar{q}_t(x,a) - \underline{q}_t(x,a)$.

We write $\mathbf{1}_t(x,a)$ for the indicator that (x,a) is visited. For a visited pair at layer h, let $L_{t,h}$ be its observed suffix loss, which lies in $[0,H]$. The return estimate, base bonus, and dilated bonus are

$$\begin{aligned}\widehat{Q}_t(x,a) &= \frac{\mathbf{1}_t(x,a)L_{t,h}}{\bar{q}_t(x,a)+\gamma_{\mathrm{PO}}},\\ b_t^{\mathrm{PO}}(x) &= \sum_a \pi_t(a\mid x)\frac{3\gamma_{\mathrm{PO}}H + H\Delta_t^{\mathrm{PO}}(x,a)}{\bar{q}_t(x,a)+\gamma_{\mathrm{PO}}},\\ B_t^{\mathrm{PO}}(x,a) &= b_t^{\mathrm{PO}}(x) + (1+1/H)\max_{P'\in\mathcal{P}_t^{\mathrm{PO}}} \mathbb{E}_{x'\sim P'(\cdot|x,a),\,a'\sim\pi_t(\cdot|x')} B_t^{\mathrm{PO}}(x',a'),\end{aligned} \tag{15}$$

with terminal bonus zero. The state-wise exponential-weights loss is $\widehat{Q}_t(x,a) - B_t^{\mathrm{PO}}(x,a)$. These are the same optimizer quantities reviewed in Appendix B: if j is the active internal epoch, then $n_t^{\mathrm{PO}} = C_j$, $\widehat{P}_t^{\mathrm{PO}} = P'_j$, and $\mathcal{P}_t^{\mathrm{PO}} = \widetilde{\mathcal{P}}_j$. Our γ_{PO}, $\widehat{Q}_t$, and B_t^{PO} are denoted there by γ, $\widetilde{Q}_t$, and B_t, respectively. The only probability statement we import is the explicit transition-confidence event from Jin et al. (2020, Lemma 2), which states

$$\mathcal{E}_P = \{P \in \mathcal{P}_t^{\mathrm{PO}} \text{ for every local } t \le T\},\ \mathbb{P}(\mathcal{E}_P^{\mathrm{c}}) \le 4\varepsilon.$$

The radius in Equation 14 dominates the radius used there: the latter has coefficients $2, 14/3$ and denominator $(n-1)\vee 1$. For $n \ge 2$, $n/(n-1) \le 2$ bounds its square-root and linear terms by those in Equation 14. For $n \le 1$, the claim follows directly. Unvisited rows are unrestricted (or have radii larger than one). The cited proof uses empirical Bernstein concentration and applies to adaptive policies. We do not use the asymptotic occupancy-width conclusion from that source.

Optimizer Analysis: Throughout this proof, $\mathbb{E}_t$ conditions on the past, the policy, and the current normalized loss vector. In particular, q_t, $\bar{q}_t$, b_t^{PO}, and B_t^{PO} are measurable before the trajectory. We define $Q_t(x,a) = Q_h^{\pi_t,P}(\boldsymbol{u}_t;x,a)$ for $x \in X_h$. Conditional on visiting (x,a), the suffix loss has mean $Q_t(x,a)$.

A concentration inequality with its constant: We consider one layer. Its possible visited pairs form a categorical variable with probabilities $q_t(y)$. Let $p_t(y) \ge q_t(y)$ be predictable, let $c_t(y) \in [0,1]$ be predictable, and let $Z_t \in [0,1]$ be observed after the selected y, with conditional mean $z_t(y)$. Whenever $p_t(y) = 0$, we also have $q_t(y) = 0$ and define the ratio $q_t(y)/p_t(y)$ as zero. The sampled indicator then vanishes almost surely. We claim that, with failure probability at most α, for a deterministic m,

$$\sum_{t=1}^{m}\sum_y \left[\frac{c_t(y)\mathbf{1}_t(y)Z_t}{p_t(y)+\gamma_{\mathrm{PO}}} - \frac{q_t(y)c_t(y)z_t(y)}{p_t(y)}\right] \le \frac{\log(1/\alpha)}{2\gamma_{\mathrm{PO}}}. \tag{16}$$

To verify the claim, for $x \ge 0$ we differentiate to obtain

$$\frac{d}{dx}\left[\log(1+2x) - \frac{2x}{1+x}\right] = \frac{2x^2}{(1+2x)(1+x)^2} \ge 0.$$

The bracket vanishes at zero, so $\exp(2x/(1+x)) \le 1+2x$. For $r \in [0,1]$, convexity then gives

$$\exp\left(\frac{2\gamma_{\mathrm{PO}}r}{p+\gamma_{\mathrm{PO}}}\right) \le 1 + r\left[\exp\left(\frac{2\gamma_{\mathrm{PO}}}{p+\gamma_{\mathrm{PO}}}\right) - 1\right] \le 1 + \frac{2\gamma_{\mathrm{PO}}r}{p} \qquad (\text{for } p>0).$$

We apply this with $r = c_t(y)Z_t$ and average over the visited pair and suffix trajectory. Since precisely one pair is visited in the layer,

$$\begin{aligned}\mathbb{E}_t \exp\left(2\gamma_{\mathrm{PO}}\sum_y \frac{c_t(y)\mathbf{1}_t(y)Z_t}{p_t(y)+\gamma_{\mathrm{PO}}}\right) &\le 1 + 2\gamma_{\mathrm{PO}}\sum_y \frac{q_t(y)c_t(y)z_t(y)}{p_t(y)}\\ &\le \exp\left(2\gamma_{\mathrm{PO}}\sum_y \frac{q_t(y)c_t(y)z_t(y)}{p_t(y)}\right).\end{aligned}$$

Iteration of conditional expectation shows that the exponential of the centered sum has expectation at most one. Markov's inequality proves Equation 16. Applying it to each of the H layers with $\alpha = \varepsilon/H$ and summing gives a bound $H\log(H/\varepsilon)/(2\gamma_{\mathrm{PO}})$ with failure at most ε. For the applications below,

we may use $p_t(y) = \max\{\bar{q}_t(y), q_t(y)\}$ in the proof. On $\mathcal{E}_P$, this equals the actual denominator parameter $\bar{q}_t(y)$, so the concentration event is defined without conditioning on a future event.

The learner-bias term: We suppose that $\mathsf{Bias}_1 = \sum_{t=1}^m \sum_{x,a} q^\pi(x)\pi_t(a \mid x)[Q_t(x,a) - \widehat{Q}_t(x,a)]$ and $Y_t = \sum_{x,a} q^\pi(x)\pi_t(a \mid x)\widehat{Q}_t(x,a)$. We see that $0 \le Y_t \le H^2/\gamma_{\mathrm{PO}}$ because $\sum_{x,a} q^\pi(x)\pi_t(a \mid x) = H$ and $\widehat{Q}_t \le H/\gamma_{\mathrm{PO}}$. Set $D = H^2/\gamma_{\mathrm{PO}}$ and $X_t = \mathbb{E}_t[Y_t] - Y_t$. Then, $\mathbb{E}_t[X_t] = 0$, $|X_t| \le D$. For $z \in [-1,1]$, we know that $e^z \le 1 + z + z^2$. Consequently, we have

$$\mathbb{E}_t[e^{X_t/D}] \le 1 + \mathbb{E}_t[X_t^2]/D^2 \le \exp(\mathbb{E}_t[X_t^2]/D^2).$$

The same argument gives, with failure at most ε,

$$\sum_{t=1}^m X_t \le D^{-1} \sum_{t=1}^m \mathrm{Var}_t(Y_t) + D\log(1/\varepsilon). \tag{17}$$

On $\mathcal{E}_P$, the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\mathrm{Var}_t(Y_t) &\le H \sum_{x,a} q^\pi(x)\pi_t(a \mid x)\mathbb{E}_t[\widehat{Q}_t(x,a)^2] \\
&\le H^3 \sum_{x,a} q^\pi(x)\pi_t(a \mid x)\frac{q_t(x,a)}{(\bar{q}_t(x,a) + \gamma_{\mathrm{PO}})^2} \\
&\le H^3 \sum_{x,a} \frac{q^\pi(x)\pi_t(a \mid x)}{\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}}.
\end{aligned}$$

Also, since $\underline{q}_t \le q_t \le \bar{q}_t$ on this event,

$$\sum_{x,a} q^\pi(x)\pi_t(a \mid x)Q_t(x,a) - \mathbb{E}_t[Y_t] \le \sum_{x,a} q^\pi(x)\pi_t(a \mid x)\frac{H[\Delta_t^{\mathrm{PO}}(x,a) + \gamma_{\mathrm{PO}}]}{\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}}.$$

We insert these two inequalities into Equation 17, and use $D^{-1} = \gamma_{\mathrm{PO}}/H^2$ and $\gamma_{\mathrm{PO}} = 2\eta_{\mathrm{PO}}H$, to obtain

$$\mathsf{Bias}_1 \le \frac{H}{2\eta_{\mathrm{PO}}}\log(1/\varepsilon) + \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x)\frac{H\Delta_t^{\mathrm{PO}}(x,a) + 2\gamma_{\mathrm{PO}}H}{\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}}. \tag{18}$$

Unless otherwise indicated, time sums in this proof run over $t \le m$.

The comparator-bias term: In Equation 16, we take $c_t(x,a) = q^\pi(x)\pi(a \mid x) \le 1$ and $Z_t = L_{t,h}/H$. Summing its layerwise conclusions and multiplying by H gives, on $\mathcal{E}_P$ and an event with failure at most ε,

$$\sum_{t,x,a} q^\pi(x)\pi(a \mid x)\left[\widehat{Q}_t(x,a) - \frac{q_t(x,a)Q_t(x,a)}{\bar{q}_t(x,a)}\right] \le \frac{H^2}{2\gamma_{\mathrm{PO}}}\log(H/\varepsilon).$$

The subtracted population term is at most $q^\pi(x)\pi(a \mid x)Q_t(x,a)$, since $q_t \le \bar{q}_t$. Thus, we see

$$\mathsf{Bias}_2 \coloneqq \sum_{t,x,a} q^\pi(x)\pi(a \mid x)[\widehat{Q}_t(x,a) - Q_t(x,a)] \le \frac{H}{4\eta_{\mathrm{PO}}}\log(H/\varepsilon). \tag{19}$$

The exponential-weights term: First, we check the bounded-loss condition. The bonus satisfies $b_t^{\mathrm{PO}}(x) \le 4H$, since $\Delta_t^{\mathrm{PO}} \le \bar{q}_t$ and $\gamma_{\mathrm{PO}}/(\bar{q}_t + \gamma_{\mathrm{PO}}) \le 1$. Its recursion implies $B_t^{\mathrm{PO}}(x,a) \le 4H^2(1+1/H)^H < 12H^2$. Therefore, we have

$$\eta_{\mathrm{PO}}B_t^{\mathrm{PO}}(x,a) \le 1/(2H) \text{ and } \eta_{\mathrm{PO}}\widehat{Q}_t(x,a) \le \eta_{\mathrm{PO}}H/\gamma_{\mathrm{PO}} = 1/2.$$

We give the exponential-weights argument explicitly. For a fixed state x, we abbreviate $z_t(a) = \widehat{Q}_t(x,a) - B_t^{\mathrm{PO}}(x,a)$ and define $\mathcal{Z}_t = \sum_a \exp(-\eta_{\mathrm{PO}}\sum_{i<t} z_i(a))$. The preceding bounds imply $|\eta_{\mathrm{PO}}z_t(a)| \le 1/2$. Because $e^v \le 1 + v + v^2$ for $|v| \le 1$, and $\log(1+r) \le r$ for $r > -1$, we obtain

$$\begin{aligned}
\log\frac{\mathcal{Z}_{t+1}}{\mathcal{Z}_t} &= \log\sum_a \pi_t(a \mid x)e^{-\eta_{\mathrm{PO}}z_t(a)} \\
&\le -\eta_{\mathrm{PO}}\sum_a \pi_t(a \mid x)z_t(a) + \eta_{\mathrm{PO}}^2\sum_a \pi_t(a \mid x)z_t(a)^2.
\end{aligned}$$

For each action, its exponential is a summand in $\mathcal{Z}_{m+1}$. Averaging the corresponding logarithmic lower bounds against $\pi(\cdot \mid x)$ gives

$$\log \frac{\mathcal{Z}_{m+1}}{\mathcal{Z}_1} \geq -\eta_{\mathrm{PO}} \sum_{t=1}^{m} \sum_{a} \pi(a \mid x) z_t(a) - \log|A|,$$

$$\sum_{t=1}^{m} \sum_{a} [\pi_t(a \mid x) - \pi(a \mid x)] z_t(a) \leq \frac{\log|A|}{\eta_{\mathrm{PO}}} + \eta_{\mathrm{PO}} \sum_{t=1}^{m} \sum_{a} \pi_t(a \mid x) z_t(a)^2.$$

We multiply by $q^\pi(x)$, sum over states, use $\sum_x q^\pi(x) = H$, and apply $(r-s)^2 \leq 2r^2 + 2s^2$ to obtain

$$\begin{aligned}\mathsf{EW} &\coloneqq \sum_{t,x,a} q^\pi(x)[\pi_t(a \mid x) - \pi(a \mid x)][\widehat{Q}_t(x,a) - B_t^{\mathrm{PO}}(x,a)] \\ &\leq \frac{H \log|A|}{\eta_{\mathrm{PO}}} + 2\eta_{\mathrm{PO}} \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x)\widehat{Q}_t(x,a)^2 + 2\eta_{\mathrm{PO}} \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x) B_t^{\mathrm{PO}}(x,a)^2.\end{aligned}$$

In Equation 16, we now take $Z_t = 1$ and $c_t(x,a) = \gamma_{\mathrm{PO}} q^\pi(x)\pi_t(a \mid x)/(\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}) \leq 1$. For a globally defined concentration event, we use $p_t = \max\{\bar{q}_t, q_t\}$ and replace $\bar{q}_t$ by p_t also in this coefficient. The quantities coincide on $\mathcal{E}_P$. After summing layers and multiplying by $2\eta_{\mathrm{PO}} H^2/\gamma_{\mathrm{PO}}$, we obtain

$$\begin{aligned}2\eta_{\mathrm{PO}} \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x)\widehat{Q}_t(x,a)^2 &\leq 2\eta_{\mathrm{PO}} H^2 \sum_{t,x,a} \frac{q^\pi(x)\pi_t(a \mid x)}{\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}} + \frac{\eta_{\mathrm{PO}} H^3}{\gamma_{\mathrm{PO}}^2} \log(H/\varepsilon) \\ &= \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x) \frac{\gamma_{\mathrm{PO}} H}{\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}} + \frac{H}{4\eta_{\mathrm{PO}}} \log(H/\varepsilon).\end{aligned}$$

Here, we used $\widehat{Q}_t^2 \leq H^2 \mathbf{1}_t/(\bar{q}_t + \gamma_{\mathrm{PO}})^2$ and $q_t/\bar{q}_t \leq 1$ on $\mathcal{E}_P$. This concentration event fails with probability at most ε. Finally, since $2\eta_{\mathrm{PO}}(B_t^{\mathrm{PO}})^2 \leq B_t^{\mathrm{PO}}/H$, so we have

$$\begin{aligned}\mathsf{EW} \leq \frac{H \log|A|}{\eta_{\mathrm{PO}}} &+ \frac{H}{4\eta_{\mathrm{PO}}} \log(H/\varepsilon) \\ &+ \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x) \left[\frac{\gamma_{\mathrm{PO}} H}{\bar{q}_t(x,a) + \gamma_{\mathrm{PO}}} + \frac{B_t^{\mathrm{PO}}(x,a)}{H} \right].\end{aligned} \tag{20}$$

Eliminating the comparator-weighted bonus: On adding Equation 18, Equation 19, and Equation 20 we obtain

$$\begin{aligned}&\sum_{t,x,a} q^\pi(x)[\pi_t(a \mid x) - \pi(a \mid x)][Q_t(x,a) - B_t^{\mathrm{PO}}(x,a)] \\ &\quad \leq \frac{HK}{\eta_{\mathrm{PO}}} + \sum_{t,x} q^\pi(x) b_t^{\mathrm{PO}}(x) + \frac{1}{H} \sum_{t,x,a} q^\pi(x)\pi_t(a \mid x) B_t^{\mathrm{PO}}(x,a).\end{aligned} \tag{21}$$

The numerical logarithmic coefficient is $\log|A| + \frac{1}{2}\log(1/\varepsilon) + \frac{1}{2}\log(H/\varepsilon) = K$. For completeness, we detail the deterministic dilated-bonus step. Because the confidence set is row rectangular, we choose a transition function $\widetilde{P}_t$ whose rows attain the maxima in Equation 15. Compactness of the row sets ensures such choices exist. On $\mathcal{E}_P$, we have

$$\begin{aligned}B_t^{\mathrm{PO}}(x,a) &= b_t^{\mathrm{PO}}(x) + (1 + 1/H)\mathbb{E}_{\widetilde{P}_t, \pi_t}[B_t^{\mathrm{PO}}(x', a')] \\ &\geq b_t^{\mathrm{PO}}(x) + (1 + 1/H)\mathbb{E}_{P, \pi_t}[B_t^{\mathrm{PO}}(x', a')].\end{aligned}$$

Let $F_t(x) = \sum_a \pi_t(a \mid x) B_t^{\mathrm{PO}}(x,a)$ and let $F_{t,h} = \sum_{x \in X_h} q^\pi(x) F_t(x)$, with $F_{t,H} = 0$. The comparator flow equation and the preceding inequality imply

$$\sum_{x \in X_h, a} q^\pi(x)[\pi_t(a \mid x) - \pi(a \mid x)] B_t^{\mathrm{PO}}(x,a) \leq F_{t,h} - (1 + 1/H) F_{t,h+1} - \sum_{x \in X_h} q^\pi(x) b_t^{\mathrm{PO}}(x).$$

We sum this over layers and add the bonus terms on the right side of Equation 21. All intermediate $F_{t,h}$ terms cancel, leaving $(1 + 1/H)F_t(x_0)$ for episode t. Therefore, the performance-difference identity gives

$$R_m^{\mathrm{PO}} \le HK/\eta_{\mathrm{PO}} + (1 + 1/H) \sum_{t=1}^{m} F_t(x_0).$$

Expansion of the bonus recursion yields

$$\begin{aligned} F_t(x_0) &= \sum_{h=0}^{H-1} (1 + 1/H)^h \sum_{x \in X_h} q^{\pi_t, \widetilde{P}_t}(x) b_t^{\mathrm{PO}}(x) \\ &\le (1 + 1/H)^{H-1} V^{\pi_t, \widetilde{P}_t}(b_t^{\mathrm{PO}}). \end{aligned}$$

Since $(1 + 1/H)^H \le e < 3$ and $b_t^{\mathrm{PO}} \ge 0$, we conclude

$$R_m^{\mathrm{PO}} \le HK/\eta_{\mathrm{PO}} + 3 \sum_{t=1}^{m} V^{\pi_t, \widetilde{P}_t}(b_t^{\mathrm{PO}}).$$

For every pair (x, a), we have $q^{\pi_t, \widetilde{P}_t}(x, a) \le \bar{q}_t(x, a)$. Since b_t^{PO} averages over π_t and is independent of its action argument, we get

$$\begin{aligned} V^{\pi_t, \widetilde{P}_t}(b_t^{\mathrm{PO}}) &= \sum_{x,a} q^{\pi_t, \widetilde{P}_t}(x, a) \frac{3\gamma_{\mathrm{PO}} H + H\Delta_t^{\mathrm{PO}}(x, a)}{\bar{q}_t(x, a) + \gamma_{\mathrm{PO}}} \\ &\le 3\gamma_{\mathrm{PO}} HD + H \sum_{x,a} \Delta_t^{\mathrm{PO}}(x, a). \end{aligned}$$

Thus, on $\mathcal{E}_P$ and the three concentration events, we have

$$R_m^{\mathrm{PO}} \le HK/\eta_{\mathrm{PO}} + 18\eta_{\mathrm{PO}} H^2 Dm + 3H \sum_{t=1}^{m} \sum_{x,a} \Delta_t^{\mathrm{PO}}(x, a). \tag{22}$$

Quite clearly, up to this point, the total failure probability is at most 7ε.

From snapshot counts to row errors: The internal epoch rule gives $n_t^{\mathrm{PO}}(x, a) \vee 1 \ge (N_t^{\mathrm{PO}}(x, a) \vee 1)1/2$. Indeed, a zero snapshot and a positive running count trigger an update after the first visit. For a positive snapshot, the running count stays below twice that snapshot until the next trigger, and the trigger copies all counts. Hence, the relation holds at the beginning of each episode.

On $\mathcal{E}_P$, both P and any $P' \in \mathcal{P}_t^{\mathrm{PO}}$ satisfy the same coordinate constraints. The empirical row sums to one when its count is positive. Therefore, the Cauchy-Schwarz inequality, the triangle inequality, and Equation 14 together imply

$$\begin{aligned} D_t^{\mathrm{PO}}(x, a) &\coloneqq \sup_{P' \in \mathcal{P}_t^{\mathrm{PO}}} \|P'(\cdot \mid x, a) - P(\cdot \mid x, a)\|_1 \\ &\le \min\left\{2, 8\sqrt{\frac{SL}{n_t^{\mathrm{PO}}(x, a) \vee 1}} + \frac{56SL}{3(n_t^{\mathrm{PO}}(x, a) \vee 1)}\right\} \\ &\le \min\left\{2, 8\sqrt{\frac{2SL}{N_t^{\mathrm{PO}}(x, a) \vee 1}} + \frac{112SL}{3(N_t^{\mathrm{PO}}(x, a) \vee 1)}\right\}. \end{aligned}$$

Also, the bound holds when the count is zero: the L_1 diameter is at most two and the second argument exceeds two ($S \ge 2$ and $L \ge \log 2$). For any $a, b, v \ge 0$, we have

$$\min\{2, av + bv^2\} \le av + \min\{2, bv^2\} \le (a + \sqrt{2b})v.$$

Putting $v = \sqrt{SL/(N_t^{\mathrm{PO}}(x, a) \vee 1)}$, $a = 8\sqrt{2}$, $b = 112/3$, and since $8\sqrt{2} + \sqrt{224/3} < 20$, we obtain the upper bound

$$D_t^{\mathrm{PO}}(x, a) \le \frac{20\sqrt{SL}}{\sqrt{N_t^{\mathrm{PO}}(x, a) \vee 1}}. \tag{23}$$

A simulation identity for every target state: We fix an episode, a candidate P', and a target state $x \in X_h$. Let $f_v(y)$ be the probability of reaching x from $y \in X_v$ under P' and π_t. Then, we have $0 \le f_v \le 1$ and $f_h(y) = \mathbf{1}\{y = x\}$. The telescoping identity for replacing P' by P (one transition at a time) gives

$$q^{\pi_t,P'}(x) - q^{\pi_t,P}(x) = \sum_{v=0}^{h-1} \sum_{y \in X_v, b} q_t(y,b) \sum_z [P'(z \mid y,b) - P(z \mid y,b)] f_{v+1}(z).$$

Also, this follows by telescoping the conditional expectations of $f_v(X_v)$ along the trajectory generated by P and π_t. Taking absolute values yields

$$|q^{\pi_t,P'}(x) - q_t(x)| \le \sum_{v<h} \sum_{y \in X_v, b} q_t(y,b) D_t^{\mathrm{PO}}(y,b).$$

For a fixed x, its action probabilities are the same for every candidate transition function. Therefore,

$$\begin{aligned} \sum_a \Delta_t^{\mathrm{PO}}(x,a) &= \max_{P' \in \mathcal{P}_t^{\mathrm{PO}}} q^{\pi_t,P'}(x) - \min_{P' \in \mathcal{P}_t^{\mathrm{PO}}} q^{\pi_t,P'}(x) \\ &\le 2 \sum_{v<h} \sum_{y \in X_v, b} q_t(y,b) D_t^{\mathrm{PO}}(y,b). \end{aligned}$$

Summing over at most S non-terminal states gives the deliberately coarse but uniform estimate as

$$\sum_{x,a} \Delta_t^{\mathrm{PO}}(x,a) \le 2S \sum_{x,a} q_t(x,a) D_t^{\mathrm{PO}}(x,a). \tag{24}$$

The maxima for different coordinates may be attained by different transition functions, and the preceding bound allows this.

A predictable weighted-count bound: We fix one non-terminal pair $y = (x,a)$ and write $J_t = \mathbf{1}_t(y)$, $r_t = q_t(y)$, and $v_t = (N_t^{\mathrm{PO}}(y) \vee 1)^{-1/2} \in [0,1]$. All these local proof variables are distinct from FIRE's gate and queue multiplier. We have $\mathbb{E}_t[J_t] = r_t$. With $c = 1 - e^{-1} > 1/2$, concavity of $1 - e^{-w}$ on $[0,1]$ yields $1 - e^{-w} \ge cw$. Thus,

$$\mathbb{E}_t[e^{-v_t J_t}] = 1 - r_t(1 - e^{-v_t}) \le e^{-c r_t v_t}.$$

Hence, the term $\exp(\sum_{t \le m} [c r_t v_t - v_t J_t])$ has expectation at most one. Applying Markov's inequality with error probability ε / D gives

$$\sum_{t=1}^m r_t v_t \le c^{-1} \left[\sum_{t=1}^m J_t v_t + \log(D/\varepsilon) \right] \le 2 \sum_{t=1}^m J_t v_t + 2 \log(D/\varepsilon).$$

We take a union bound over at most D pairs. If a pair is visited n times, its realized weighted count is

$$\sum_{k=0}^{n-1} (k \vee 1)^{-1/2} \le 2\sqrt{n}.$$

For $n = 0, 1$ the above result immediately follows. For $n \ge 2$, the left side is $2 + \sum_{k=2}^{n-1} k^{-1/2} \le 2 + \int_1^{n-1} x^{-1/2}\, dx = 2\sqrt{n-1} \le 2\sqrt{n}$. If n_y is the number of visits to pair y through episode m, then $\sum_y n_y = Hm$. Using the Cauchy-Schwarz inequality now gives, with failure at most ε,

$$\begin{aligned} \sum_{t=1}^m \sum_{x,a} \frac{q_t(x,a)}{\sqrt{N_t^{\mathrm{PO}}(x,a) \vee 1}} &\le 4 \sum_y \sqrt{n_y} + 2D \log(D/\varepsilon) \\ &\le 4\sqrt{DHm} + 2D \log(D/\varepsilon). \end{aligned} \tag{25}$$

Combining Equation 23, Equation 24, and Equation 25 proves

$$\sum_{t=1}^m \sum_{x,a} \Delta_t^{\mathrm{PO}}(x,a) \le 160 S \sqrt{SLDH} \sqrt{m} + 80 SD \sqrt{SL} \log(D/\varepsilon). \tag{26}$$

Completing the certificate: On the intersection of the transition event and four concentration events, we substitute Equation 26 into Equation 22 to get

$$R_m^{\mathrm{PO}} \le HK/\eta_{\mathrm{PO}} + 18\eta_{\mathrm{PO}}H^2Dm + 480HS\sqrt{SLDH}\sqrt{m} + 240HSD\sqrt{SL}\log(D/\varepsilon).$$

The total failure probability is at most $4\varepsilon + \varepsilon(4) = 8\varepsilon$. The global learning rate satisfies

$$1/\eta_{\mathrm{PO}} \le 24H^3 + \sqrt{DHT} \text{ and } \eta_{\mathrm{PO}} \le 1/\sqrt{DHT}.$$

Therefore, we have

$$HK/\eta_{\mathrm{PO}} + 18\eta_{\mathrm{PO}}H^2Dm \le 24H^4K + H\sqrt{DH}K\sqrt{T} + 18H\sqrt{DH}\cdot m/\sqrt{T}.$$

The exact numerical coefficients C_{opt}, C_{exp}, and C_{add} prove Theorem C.1. □

Using the coefficients in the wrapper: We evaluate these expressions at $\varepsilon = \varepsilon_{\mathrm{PO}}$ and may use them directly as the coefficients in Proposition 3.1. In particular, their 8ε failure bound is stronger than the conservative 9ε interface used by the outer confidence allocation. Appendix C.3 supplies the conditional continuation and simultaneous prefix event needed at random restart endpoints. The coefficients are proven upper bounds, and we do not need the smallest admissible coefficients. The conservative width bound can make the prescribed exponential parameter small at moderate episode budgets. It preserves the stated T exponents, while its dependence on H, $|X|$, and $|A|$ is looser than the sharp asymptotic result of Luo et al. (2021). Finally, this certificate is for exact occupancy extrema and exact dilated-bonus calculations. Additional approximation errors require a separate analysis.

C.3 Random Starts and Random Endpoints

Let τ_r be the random global start time of the outer copy r. For $r, m \in [T]$, we condition on the outer history at τ_r whenever that copy is started. Then, we construct an auxiliary continuation on the same probability space. Until the outer stopping trigger activates, we use the real copy's policies, vectors, and trajectories. After the trigger activates, we continue from that copy's last state, supply the zero loss vector on every remaining local episode, sample fresh trajectories from P under its own policies, and update its state normally through local episode T. This continuation is used only in the proof. The auxiliary sequence satisfies the protocol of the preceding subsections. Hence, for every deterministic prefix m, its certificate fails with conditional probability at most $9\varepsilon_{\mathrm{PO}}$.

We say the event $E_{r,m}$ holds if copy r is never started, if its real duration is less than m, or if Equation 4 holds for its first m real local episodes. If the real duration is at least m, its first m episodes match those of the auxiliary continuation. Thus, $\bar{E}_{r,m}$ is contained in the auxiliary certificate's failure event. Hence, conditional on the history at the start of copy r, the failure probability is at most $9\varepsilon_{\mathrm{PO}}$. Taking expectations on both sides gives $\Pr(\bar{E}_{r,m}) \le 9\varepsilon_{\mathrm{PO}}$. On putting $\varepsilon_{\mathrm{PO}} = \delta/27T^2$, we have

$$\Pr\left(\bigcap_{r=1}^{T}\bigcap_{m=1}^{T} E_{r,m}\right) \ge 1 - \left(T^2\cdot\frac{\delta}{3T^2}\right) = 1 - \frac{\delta}{3}. \tag{27}$$

Only after constructing this simultaneous event do we insert the random realized duration of each copy. Being an ordinary deterministic union bound, no independence between restarted copies is required.

C.4 Proof of Lemma 3.2

Lemma C.2. *We fix a comparator π before interaction and define $B_{\mathrm{PO}} \coloneqq 12C_{\mathrm{opt}} + \frac{8}{\sqrt{3}}C_{\mathrm{exp}} + 8C_{\mathrm{add}}/\sqrt{T}$. For any $\delta \in (0,1)$, we initialize each outer copy with $\varepsilon_{\mathrm{PO}} = \delta/27T^2$ and let $W_{j',\max} \coloneqq \max\{W^{[1]}, \max_{t\le j'}\widehat{W}_t\}$. With probability at least $1-\delta/3$, we have simultaneously $\forall j' \le T$,*

$$\sum_{t=1}^{j'}\left[V^{\pi_t,P}(\boldsymbol{s}_t) - V^{\pi,P}(\boldsymbol{s}_t)\right] \le B_{\mathrm{PO}}W_{j',\max}\sqrt{T}.$$

Proof. We consider the simultaneous local-prefix event constructed in Appendix C.3, and fix a global prefix $j' \le T$. Now, we enumerate the nonempty outer copies that intersect this prefix by

Algorithm 2 Feasible Iterative Restarts in Expectation (FIRE)

Require: T, H, $|X|$, $|A|$, δ.
Set $M = 1$ and the confidence inputs from Section 4.
Compute C_{opt}, C_{exp}, and C_{add} at $\varepsilon_{\mathrm{PO}}$, then B_F.
Set $\lambda = 1/(H+2eB_F\sqrt{T})$ and $W^{[1]} = M + \lambda e^{\lambda H}$.
Initialize $\zeta_0 = 0$, $k = 1$, all global counts to zero, and a fresh `PO-DB` copy with T and $\varepsilon_{\mathrm{PO}}$.
for $t = 1, \ldots, T$ **do**
  Set $\widehat{W}_t = M + \lambda e^{\lambda(\zeta_{t-1}+H)}$ and $k_{\mathrm{old}} = k$.
  while $\widehat{W}_t > W^{[k]}$ **do**
    $W^{[k+1]} = 2W^{[k]}$; $k \leftarrow k + 1$.
  end while
  if $k > k_{\mathrm{old}}$ **then**
    Start one fresh `PO-DB` copy and retain the global queue and estimators.
  end if
  Set $k_t = k$, $\mathcal{W}_t = W^{[k_t]}$, and obtain π_t from the active copy.
  Set $g_1 = 0$ and for $t \geq 2$, compute g_t using Equation 7.
  Set $I_t = \mathbf{1}\{g_t > 0\}$, $\zeta_t = \zeta_{t-1} + [g_t]^+$, and $w_t = \lambda e^{\lambda\zeta_t} I_t$.
  Execute π_t and observe its bandit trajectory.
  Form $s_{t,h} = M\ell_{t,h} + w_t c_{t,h}$ on visited pairs; Feed $u_{t,h} = (s_{t,h}+\mathcal{W}_t)/(2\mathcal{W}_t)$ to the active copy.
  Update the global counts, $\widehat{P}_t$, $\widehat{\boldsymbol{c}}_t$, and $\boldsymbol{b}_t$.
end for

$r = 1, \ldots, R$. Let m_r be the number of their episodes in the prefix, and let ω_r be the scale used by copy r. These scales are random, but each is fixed at the start of its corresponding copy. Since a new copy is created only after a factor-two increase, even a jump that skips unused scale levels will satisfy $\omega_{r+1} \geq 2\omega_r$.

The trigger chooses the smallest scale not below the predictable envelope $\widehat{W}_t$. Consequently, we have $\omega_R \leq 2W_{j',\max}$. Thus, the elementary geometric and Cauchy-Schwarz bounds are

$$\sum_{r=1}^{R} \omega_r < 2\omega_R \leq 4W_{j',\max}, \ \sum_{r=1}^{R} \omega_r m_r \leq \omega_R \sum_{r=1}^{R} m_r \leq 2W_{j',\max}T, \text{ and}$$

$$\sum_{r=1}^{R} \omega_r \sqrt{m_r} \leq \sqrt{\sum_{r=1}^{R} \omega_r^2} \sqrt{\sum_{r=1}^{R} m_r} < \frac{2}{\sqrt{3}} \omega_R \sqrt{T} \leq \frac{4}{\sqrt{3}} W_{j',\max}\sqrt{T}.$$

For every episode in the r-th copy, the affine normalization is exact because all policy occupancies have mass H. Thus, we have

$$V^{\pi_t,P}(\boldsymbol{s}_t) - V^{\pi,P}(\boldsymbol{s}_t) = 2\omega_r \left[V^{\pi_t,P}(\boldsymbol{u}_t) - V^{\pi,P}(\boldsymbol{u}_t)\right].$$

We apply Proposition 3.1 at the realized local prefix m_r, multiply it by $2\omega_r$, and sum over r. Using the preceding three bounds gives us

$$\sum_{t=1}^{j'} \left[V^{\pi_t,P}(\boldsymbol{s}_t) - V^{\pi,P}(\boldsymbol{s}_t)\right] \leq \left(12C_{\mathrm{opt}} + \frac{8}{\sqrt{3}} C_{\mathrm{exp}}\right) W_{j',\max}\sqrt{T} + 8C_{\mathrm{add}} W_{j',\max}$$
$$= B_{\mathrm{PO}} W_{j',\max}\sqrt{T}.$$

Since the prefix event has probability at least $1 - \delta/3$, the conclusion is simultaneous for every $j' \leq T$. □

D Omitted Details for the Stochastic Regime

D.1 Filtrations and Predictability

Let $\mathcal{B}_{t-1}$ contain the learner's internal random variables, trajectories, and observed feedback through episode $t-1$, together with $(\boldsymbol{\ell}_\tau, \boldsymbol{c}_\tau)$ for $\tau < t$. This history determines the active outer copy and

scale. After the active copy has produced π_t, FIRE computes g_t, I_t, and ζ_t, all before the fresh episode-t vectors are sampled. We define the pre-vector filtration $\mathcal{G}_t$ as

$$\mathcal{G}_t \coloneqq \mathcal{B}_{t-1} \vee \sigma\left(\pi_t, g_t, I_t, \zeta_t, k_t, \mathcal{W}_t\right). \tag{28}$$

Then, $q_t = q^{\pi_t, P}$ and $w_t = \lambda e^{\lambda \zeta_t} I_t$ are $\mathcal{G}_t$-measurable. The stochastic model in Section 2.1 implies $\mathbb{E}[\boldsymbol{\ell}_t \mid \mathcal{G}_t] = \boldsymbol{\ell}$ and $\mathbb{E}[\boldsymbol{c}_t \mid \mathcal{G}_t] = \boldsymbol{c}$. After the full current vectors are fixed, we define another filtration

$$\mathcal{H}_t \coloneqq \mathcal{G}_t \vee \sigma(\boldsymbol{\ell}_t, \boldsymbol{c}_t).$$

Conditional on $\mathcal{H}_t$, the trajectory is generated by the fixed transition function P and policy π_t. Thus, for every layered coordinate (x, a, h), we have

$$\mathbb{E}[\mathbf{1}\{x_{t,h} = x, a_{t,h} = a\} \mid \mathcal{H}_t] = q_{t,h}(x, a).$$

This identity supplies the conditional unbiasedness required by the importance-weighted PO-DB estimators. It also shows why the current surrogate may be stochastic. The policy and multiplier are fixed before the current vector, and the vector is fixed before the current trajectory.

D.2 Proof of Lemma D.1

Lemma D.1 (Model-Shift Performance Difference)**.** *First, we fix a policy π, transition functions P, $\widehat{P}$, and cost vectors $\boldsymbol{c}$, $\widehat{\boldsymbol{c}}$. Second, for $h \in [H+1]^{-1}$ and $x \in X_h$, we define $D_h(x) \coloneqq V_h^{\pi,\widehat{P}}(x \mid \widehat{\boldsymbol{c}}) - V_h^{\pi,P}(x \mid \boldsymbol{c})$, $\mathfrak{C}_j \coloneqq \widehat{c}_j(x_j, a_j) - c_j(x_j, a_j)$, and $\mathfrak{P}_j(x') = \widehat{P}(x' \mid x_j, a_j) - P(x' \mid x_j, a_j)$. Then, we have the following*

$$D_h(x) = \mathbb{E}_{\pi,\widehat{P}}\left[\sum_{j=h}^{H-1}\left(\mathfrak{C}_j + \sum_{x'}\mathfrak{P}_j(x') \cdot V_{j+1}^{\pi,P}(x' \mid \boldsymbol{c})\right) \,\middle|\, x_h = x\right]. \tag{29}$$

Proof. Since both value functions equal zero at layer H, we have $D_H(x) = 0$. The sum on the right side of Equation 29 is empty at $h = H$ and is also zero. Hence, the identity holds at the terminal layer. We fix an $h < H$ and suppose the identity holds at layer $h+1$. Then, we subtract the two Bellman equations to get

$$\begin{aligned} D_h(x) = \mathbb{E}_{a\sim\pi_h(\cdot|x)}\Bigg[&\widehat{c}_h(x,a) - c_h(x,a) \\ &+ \sum_{x'}\widehat{P}(x' \mid x,a) \cdot V_{h+1}^{\pi,\widehat{P}}(x' \mid \widehat{\boldsymbol{c}}) - \sum_{x'}P(x' \mid x,a) \cdot V_{h+1}^{\pi,P}(x' \mid \boldsymbol{c})\Bigg]. \end{aligned}$$

We add and subtract $\sum_{x'}\widehat{P}(x' \mid x,a) \cdot V_{h+1}^{\pi,P}(x' \mid \boldsymbol{c})$ inside the brackets. After collecting the terms, we observe

$$\begin{aligned} D_h(x) = \mathbb{E}_{a\sim\pi_h(\cdot|x)}\Bigg[&\widehat{c}_h(x,a) - c_h(x,a) \\ &+ \sum_{x'}\Big[\widehat{P}(x' \mid x,a) - P(x' \mid x,a)\Big] \cdot V_{h+1}^{\pi,P}(x' \mid \boldsymbol{c}) \\ &+ \sum_{x'}\widehat{P}(x' \mid x,a) \cdot D_{h+1}(x')\Bigg]. \end{aligned}$$

The first two terms form the summand of Equation 29 for $j = h$. Quite clearly, conditional on (x_h, a_h), the last term is $\mathbb{E}_{x_{h+1}\sim\widehat{P}(\cdot|x_h,a_h)}[D_{h+1}(x_{h+1})]$. We substitute the induction hypothesis and apply the law of iterated expectations. This concatenates the $j = h$ summand with the sum from $j = h+1$ to $H-1$, proving the identity at layer h. Backward induction completes the proof. □

D.3 Proof of the Cost-Confidence event

Proposition D.2 (Cost Confidence)**.** *For any confidence input $\eta \in (0,1)$, with probability at least $1-\eta$, simultaneously for all $t \le T$ and (x,a), we have $|\widehat{c}_t(x,a) - c(x,a)| \le \xi_t(x,a)$, where $\xi_t(x,a) = \min\{2, \sqrt{2L_c(\eta)/1 \vee N_t(x,a)}\}$.*

Proof. We fix a state-action pair (x,a) and enumerate its visits. Although we draw the fresh cost vector before the trajectory, our constructed stochastic model makes it conditionally independent of the policy and transition randomness that determine whether (x,a) is visited. Consequently, if $\mathcal{C}_i$ is the cost observed on the i-th visit, then $\mathcal{C}_i \in [-1,1]$ and its conditional mean (given the visit-indexed past) is $c(x,a)$. Thus, $\mathcal{C}_i - c(x,a)$ is a bounded martingale difference. For every deterministic sample size $n \ge 1$, Hoeffding's martingale inequality gives

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i - c(x,a)\right| > \epsilon\right) \le 2e^{-n\epsilon^2/2}.$$

On setting $\epsilon = \sqrt{2L_c(\eta)/n}$, the right side of the above inequality is $2e^{-L_c(\eta)} = \eta/(d_{\mathrm{sa}}T)$. We union-bound over the d_{sa} pairs and the T possible positive sample sizes (at most). If the square-root radius exceeds 2, capping at 2 remains valid because it is the largest possible difference between any two numbers in $[-1,1]$. For $n = 0$, we have $\widehat{c}_t(x,a) = 0$. Since $L_c(\eta) \ge \log 2$, we have $\xi_t(x,a) = \min\{2, \sqrt{2L_c(\eta)}\} \ge 1$, so $|c(x,a)| \le \xi_t(x,a)$. □

D.4 Proof of the Transition-Confidence event

Proposition D.3 (Transition Confidence)**.** *For any confidence input $\eta \in (0,1)$, with probability at least $1-\eta$, the inequality $|\widehat{P}_t(x' \mid x,a) - P(x' \mid x,a)| \le \epsilon_t(x,a,x')$ holds simultaneously for every $t \le T$ and every layered transition coordinate (x,a,x'). Here, $\epsilon_t(x,a,x')$ is the coordinate-wise Hoeffding radius defined as $\epsilon_t(x,a,x') = \min\{1, \sqrt{L_P(\eta)/2N_t(x,a)}\}$ when $N_t(x,a) \ge 1$, and $\epsilon_t(x,a,x') = 1$ if $N_t(x,a) = 0$, on unknown rows. On known rows, we use the exact estimate and zero radius.*

Proof. We fix a triple (x,a,x') and enumerate visits to (x,a). Let Y_i indicate that the successor at the i-th visit is x'. Conditional on the history immediately before that transition, we have $Y_i \in \{0,1\}$ and $\mathbb{E}[Y_i] = P(x' \mid x,a)$. Therefore, for every deterministic sample size $n \ge 1$, Hoeffding's martingale inequality gives

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}Y_i - P(x' \mid x,a)\right| > \epsilon\right) \le 2e^{-2n\epsilon^2}.$$

On setting $\epsilon = \sqrt{L_P(\eta)/(2n)}$, the right side of the above inequality becomes $2e^{-L_P(\eta)} = \eta/d_{\mathrm{tr}}T$. We union-bound over at most d_{tr} layered transition coordinates and T positive sample sizes. Also, we cap the radius at 1 because the two probabilities differ by at most 1. When the sample size is zero, i.e., $n = 0$, the explicitly assigned unit radius also dominates the difference. For structurally known rows, the estimate equals the true row, and the zero radius makes the claim deterministic. Thereafter, reducing these radii preserves all subsequent upper bounds. □

D.5 Occupancy Simulation

Proposition D.4 (Occupancy Simulation Inequality)**.** *For every policy π and two layered transition functions P and $\widehat{P}$, we have*

$$\sum_{(x,a,h)} |q_h^{\pi,\widehat{P}}(x,a) - q_h^{\pi,P}(x,a)| \le H \sum_{(x,a,h)} q_h^{\pi,P}(x,a) \sum_{x'} |\widehat{P}(x' \mid x,a) - P(x' \mid x,a)|. \tag{30}$$

Proof. Suppose μ_h^P and $\mu_h^{\widehat{P}}$ be the two state distributions and $\mathfrak{d}_h = \left\|\mu_h^{\widehat{P}} - \mu_h^P\right\|_1$. They share the same deterministic initial state, so $\mathfrak{d}_0 = 0$. Applying the same policy to two state distributions

does not change their L_1-distance. Therefore, adding and subtracting the distribution obtained by propagating $\mu_h^P \pi_h$ through $\widehat{P}$ gives

$$\mathfrak{d}_{h+1} \le \mathfrak{d}_h + \sum_{x,a} q_h^{\pi,P}(x,a) \left\| \widehat{P}(\cdot \mid x,a) - P(\cdot \mid x,a) \right\|_1 .$$

Performing the induction yields

$$\mathfrak{d}_h \le \sum_{j=0}^{h-1} \sum_{x,a} q_j^{\pi,P}(x,a) \left\| \widehat{P}(\cdot \mid x,a) - P(\cdot \mid x,a) \right\|_1 .$$

Moreover, we have $\sum_{x,a} |q_h^{\pi,\widehat{P}}(x,a) - q_h^{\pi,P}(x,a)| = \mathfrak{d}_h$. This is because both occupancies use the same conditional action probabilities. Then, we sum the preceding inequality over $h = 0, \dots, H-1$. Each layer-j row-error term occurs at most H times, which proves Equation 30. □

D.6 Predictable Inverse-Count

Proposition D.5 (Predictable Inverse-Count Sums)**.** *Let $Y_{t,j} \in \{0,1\}$ indicate that the pair $j = (x,a)$ is visited in episode $t \in [T]$. Also, we set $p_{t,j} = \mathbb{E}[Y_{t,j} \mid \mathcal{G}_t] = q_t(x,a)$, $N_{t-1,j} = \sum_{\tau<t} Y_{\tau,j}$, and $L_N = \log 2d_{\mathrm{sa}}/\eta$. With probability at least $1-\eta$, simultaneously for every j,*

$$\sum_{t=1}^{T} \frac{p_{t,j}}{\sqrt{1 \vee N_{t-1,j}}} \le 4\sqrt{N_{T,j}} + 2 + 3L_N, \text{ and} \tag{31}$$

$$\sum_{t=1}^{T} \frac{p_{t,j}}{1 \vee N_{t-1,j}} \le 4 + 2\log(1 + N_{T,j}) + 3L_N. \tag{32}$$

Proof. We fix the index-pair j and let f_t be either $(1 \vee N_{t-1,j})^{-1/2}$ or $(1 \vee N_{t-1,j})^{-1}$. Note that the weight is $\mathcal{G}_t$-measurable and lies in $[0,1]$. We define the following quantities

$$S = \sum_{t=1}^{T} p_{t,j} f_t, \; A = \sum_{t=1}^{T} Y_{t,j} f_t, \; D = S - A = \sum_{t=1}^{T} (p_{t,j} - Y_{t,j}) f_t.$$

For $u \in [0,1]$, we have $1 - e^{-u} \ge u/2$. Thus, conditional on $\mathcal{G}_t$, we have

$$\begin{aligned} \mathbb{E}\left[\exp\left(\frac{1}{2} p_{t,j} f_t - Y_{t,j} f_t \right) \middle| \mathcal{G}_t \right] &= e^{p_{t,j} f_t / 2} \left(1 - p_{t,j} + p_{t,j} e^{-f_t} \right) \\ &\le e^{p_{t,j} f_t / 2 - p_{t,j}(1 - e^{-f_t})} \le 1. \end{aligned}$$

Let $\mathcal{K}_t$ be $\mathcal{G}_t$ enlarged by the full episode-t vectors, trajectory, and feedback. Then, we get $\mathcal{G}_t \subseteq \mathcal{K}_t \subseteq \mathcal{G}_{t+1}$. The product of these conditional factors is a non-negative supermartingale with respect to the interlaced filtration $\mathcal{G}_1, \mathcal{K}_1, \mathcal{G}_2, \mathcal{K}_2, \dots$. Hence, by Markov's inequality, we observe

$$\Pr\left(\frac{1}{2} S - A > L_N \right) \le e^{-L_N}.$$

A union bound over the d_{sa} pairs and the two choices of f_t simultaneously yields, $S \le 2A + 2L_N$. We use the looser $3L_N$ in the statement. If the pair is visited $n = N_{T,j}$ times, the preceding counts at its visits are $0, 1, \dots, n-1$. Therefore, we have

$$A_{1/2} \le 1 + \sum_{r=1}^{n-1} r^{-1/2} \le 1 + \int_0^n u^{-1/2} \, du = 1 + 2\sqrt{n},$$

$$A_1 \le 1 + \sum_{r=1}^{n-1} r^{-1} \le 2 + \log(1+n).$$

To cover $n = 0$, we substitute the above inequalities into $S \le 2A + 3L_N$ and loosen the additive constants as in Equation 31 and Equation 32. □

D.7 Proof of Lemma 4.1

Lemma D.6 (Value Confidence and Width Sum)**.** *We suppose $\eta = \delta_{\mathrm{stat}}$. There exists a deterministic $\Gamma_\beta = \widetilde{\mathcal{O}}(\sqrt{T})$ (in Equation 34). With probability at least $1-3\delta_{\mathrm{stat}}$, for every $t \geq 2$, we simultaneously have $|\widehat{V}_t - V^{\pi_t,P}(\boldsymbol{c})| \leq \beta_{t-1}(\pi_t)$, $g_t \leq V^{\pi_t,P}(\boldsymbol{c}) \leq g_t + 2\beta_{t-1}(\pi_t)$, and $\sum_{t=2}^{T} \beta_{t-1}(\pi_t) \leq \Gamma_\beta$.*

Proof. Having the confidence $\eta = \delta_{\mathrm{stat}}$, we intersect Propositions D.2, D.3, and D.5. Thus, the total failure probability of these intersected events is at most $3\delta_{\mathrm{stat}}$. Now, we fix $t \geq 2$ and apply Lemma D.1 with $\widehat{P} = \widehat{P}_{t-1}$ and $\widehat{\boldsymbol{c}} = \widehat{\boldsymbol{c}}_{t-1}$. For each pair, we define its local absolute-error bound as

$$e_{t-1}(x,a) = |\widehat{c}_{t-1}(x,a) - c(x,a)| + H \sum_{x'} |\widehat{P}_{t-1}(x' \mid x,a) - P(x' \mid x,a)|.$$

The confidence events give $e_{t-1} \leq \xi_{t-1} + H \sum_{x'} \epsilon_{t-1}$. Independently, costs in $[-1,1]$ and probability rows give $e_{t-1} \leq 2 + 2H$. Hence, $e_{t-1}(x,a) \leq b_{t-1}(x,a)$, including where the bonus is capped. Since $|V_{h+1}^{\pi_t,P}(x' \mid \boldsymbol{c})| \leq H$, the model-shift identity implies

$$\begin{aligned} |\widehat{V}_t - V^{\pi_t,P}(\boldsymbol{c})| &\leq \mathbb{E}_{\pi_t,\widehat{P}_{t-1}} \left[\sum_{h=0}^{H-1} e_{t-1}(x_h,a_h) \right] \\ &\leq \mathbb{E}_{\pi_t,\widehat{P}_{t-1}} \left[\sum_{h=0}^{H-1} b_{t-1}(x_h,a_h) \right] = \beta_{t-1}(\pi_t). \end{aligned}$$

This proves $|\widehat{V}_t - V^{\pi_t,P}(\boldsymbol{c})| \leq \beta_{t-1}(\pi_t)$. Substituting $g_t = \widehat{V}_t - \beta_{t-1}(\pi_t)$ proves $g_t \leq V^{\pi_t,P}(\boldsymbol{c}) \leq g_t + 2\beta_{t-1}(\pi_t)$. Now, we sum the widths and evaluate all logarithms at δ_{stat} and put

$$a_P \coloneqq \max\left\{1, \sqrt{\frac{L_P}{2}}\right\},\ A_\beta \coloneqq \sqrt{2L_c} + H|X|a_P.$$

The definitions of the two radii imply the following, including at zero visits,

$$\begin{aligned} \xi_{t-1}(x,a) &\leq \frac{\sqrt{2L_c}}{\sqrt{1 \vee N_{t-1}(x,a)}}, \\ \sum_{x'} \epsilon_{t-1}(x,a,x') &\leq \frac{|X|a_P}{\sqrt{1 \vee N_{t-1}(x,a)}}. \end{aligned}$$

The cap in the definition of $b_t(x,a)$ (in Section 4.1) can only decrease the bonus. Thus, we observe

$$b_{t-1}(x,a) \leq \frac{A_\beta}{\sqrt{1 \vee N_{t-1}(x,a)}}.$$

On adding and subtracting the true occupancy, we get

$$\begin{aligned} \sum_{t=2}^{T} \beta_{t-1}(\pi_t) = &\sum_{t=2}^{T} \sum_{x,a} q^{\pi_t,P}(x,a) b_{t-1}(x,a) \\ &+ \sum_{t=2}^{T} \sum_{x,a} [q^{\pi_t,\widehat{P}_{t-1}}(x,a) - q^{\pi_t,P}(x,a)] b_{t-1}(x,a). \end{aligned}$$

In the first term of the above equation, we have $q^{\pi_t,P}(x,a) = p_{t,(x,a)}$. Then, we apply Proposition D.5 to the first term. Since $\sum_{x,a} N_T(x,a) = HT$, using the Cauchy-Schwarz inequality gives $\sum_{x,a} \sqrt{N_T(x,a)} \leq \sqrt{d_{\mathrm{sa}} HT}$. Thus, the first term is at most

$$A_\beta \left[4\sqrt{d_{\mathrm{sa}} HT} + d_{\mathrm{sa}}(2 + 3L_N) \right]. \tag{33}$$

For the second term, we apply Proposition D.4 episode by episode. Then, we use the transition-confidence event and Proposition D.5 to have

$$\begin{aligned}\sum_{t=2}^{T}\sum_{x,a}&|q^{\pi_t,\widehat{P}_{t-1}}(x,a)-q^{\pi_t,P}(x,a)|\\&\le H|X|a_P\sum_{t=2}^{T}\sum_{x,a}\frac{q^{\pi_t,P}(x,a)}{\sqrt{1\vee N_{t-1}(x,a)}}\\&\le H|X|a_P\left[4\sqrt{d_{\mathrm{sa}}HT}+d_{\mathrm{sa}}(2+3L_N)\right]=:\Gamma_P.\end{aligned}$$

Since we know $b_{t-1}(x,a)\le 2+2H\le 4H$, the absolute value of the second term in the occupancy decomposition is at most $4H\Gamma_P$. Combining this with Equation 33 gives us

$$\Gamma_\beta = A_\beta\left[4\sqrt{d_{\mathrm{sa}}HT}+d_{\mathrm{sa}}(2+3L_N)\right]+4H\Gamma_P. \tag{34}$$

This proves $\sum_{t=2}^{T}\beta_{t-1}(\pi_t)\le\Gamma_\beta$. Also, for fixed problem dimensions, we have $\Gamma_\beta=\widetilde{\mathcal{O}}(\sqrt{T})$. □

D.8 Realized-to-Population Bridge and Proof of Lemma 4.2

Lemma D.7 (Population Surrogate Regret)**.** *With probability at least* $1-2\delta/3$ *and* $\overline{W}_T = M+\lambda e^{\lambda(\zeta_T+H)}$*, we have* $\mathcal{J}_T^{\mathrm{pop}} \coloneqq \sum_{t=1}^{T}\left(\bar{S}_t(\pi_t)-\bar{S}_t(\pi^\star_{\mathrm{stoc}})\right)\le B_F\overline{W}_T\sqrt{T}$.

Proof. First, we define the learner-plus-comparator difference as

$$Z_t \coloneqq \langle q_t-q^\star, \boldsymbol{s}_t-\bar{\boldsymbol{s}}_t\rangle.$$

Recall from Appendix D.1 that q_t, w_t, and $q^\star$ are all $\mathcal{G}_t$-measurable. Thus, we have the martingale term as

$$\mathbb{E}[Z_t\mid\mathcal{G}_t]=\langle q_t-q^\star, M\cdot\mathbb{E}[\boldsymbol{\ell}_t-\boldsymbol{\ell}\mid\mathcal{G}_t]+w_t\cdot\mathbb{E}[\boldsymbol{c}_t-\boldsymbol{c}\mid\mathcal{G}_t]\rangle=0.$$

Note that the martingale term contains both the learner and the fixed comparator. Let $\mathcal{W}_t=W^{[k_t]}$ be the active scale in episode t. Since occupancies have total mass H,

$$\|q_t-q^\star\|_1\le 2H.$$

The loss differences have magnitude at most one, the cost differences have magnitude at most two, and so $M+w_t\le\mathcal{W}_t$. Consequently, we have $\|\boldsymbol{s}_t-\bar{\boldsymbol{s}}_t\|_\infty\le M+2w_t\le 2\mathcal{W}_t$ and

$$|Z_t|\le 4H\mathcal{W}_t,\ \mathbb{E}[Z_t^2\mid\mathcal{G}_t]\le 16H^2(\mathcal{W}_t)^2. \tag{35}$$

Let $\varrho_t\in[T]$ be the ordinal of the nonempty copy receiving episode t. Both ϱ_t and $\mathcal{W}_t$ are $\mathcal{G}_t$-measurable. For each deterministic $r\in[T]$, we set

$$U_t^{(r)}=\mathbf{1}\{\varrho_t=r\}\frac{Z_t}{\mathcal{W}_t}.$$

The predictable normalization, together with Equation 35, gives

$$\mathbb{E}[U_t^{(r)}\mid\mathcal{G}_t]=0,\ |U_t^{(r)}|\le 4H, \mathbb{E}[(U_t^{(r)})^2\mid\mathcal{G}_t]\le 16H^2.$$

With $L_Z=\log 6T^2/\delta$ and failure probability at most δ_{pop}, applying the two-sided Freedman's inequality at the deterministic global horizon T gives

$$\left|\sum_{t=1}^{T}U_t^{(r)}\right|\le 4H\sqrt{2TL_Z}+\frac{8}{3}HL_Z. \tag{36}$$

A union bound over the T possible copy ordinals costs at most $T\delta_{\mathrm{pop}}=\delta/3T\le\delta/3$. Since the inactive episodes contribute exactly zero, the deterministic-horizon sum in Equation 36 is already the complete sum of copy r. One should note that we have not substituted a random endpoint and have used no optional stopping criterion.

Let T_r and ω_r be the realized length and scale of nonempty copy r, and let W_{last} be the largest active scale. The scales ω_r may be random, but are constant within each copy and satisfy $\omega_{r+1} \geq 2\omega_r$. Therefore, we have the following pathwise

$$\sum_r \omega_r \leq 2W_{\text{last}}, \ \sum_r (\omega_r)^2 \leq \frac{4}{3} W_{\text{last}}^2.$$

When the trigger is activated, the chosen scale is the smallest one that covers the predictable envelope. Hence, $W_{\text{last}} \leq 2 \max_t \widehat{W}_t \leq 2\overline{W}_T$, and thus

$$\sum_r \omega_r \leq 4\overline{W}_T, \tag{37}$$

$$\sum_r \omega_r \sqrt{T_r} \leq \sqrt{\sum_r (\omega_r)^2} \sqrt{\sum_r T_r} \leq \frac{4}{\sqrt{3}} \overline{W}_T \sqrt{T}. \tag{38}$$

We multiply Equation 36 by ω_r, sum over the nonempty copies, and use Equation 37. This gives

$$\left| \sum_{t=1}^{T} Z_t \right| \leq 16\sqrt{2} H \overline{W}_T \sqrt{T L_Z} + \frac{32}{3} H \overline{W}_T L_Z = B_Z \overline{W}_T \sqrt{T}. \tag{39}$$

For an episode in scale ω_r, $\boldsymbol{u}_t = {}^{\boldsymbol{s}_t + \omega_r \mathbf{1}}/_{2\omega_r}$. Since both q_t and $q^\star$ have mass H, the affine shift cancels, and we get

$$V^{\pi_t, P}(\bar{\boldsymbol{s}}_t) - V^{\pi^\star_{\text{stoc}}, P}(\bar{\boldsymbol{s}}_t) = \langle q_t - q^\star, \bar{\boldsymbol{s}}_t \rangle = 2\omega_r \langle q_t - q^\star, \boldsymbol{u}_t \rangle - Z_t. \tag{40}$$

On the wrapper's deterministic-prefix PO event, Proposition 3.1 can be evaluated at every realized T_r. Equation 37, Equation 38, and $\max_r \omega_r \leq 2\overline{W}_T$ yield

$$\begin{aligned}
&2 \sum_r \omega_r \left[C_{\text{opt}} \left(\sqrt{T} + \frac{T_r}{\sqrt{T}} \right) + C_{\text{exp}} \sqrt{T_r} + C_{\text{add}} \right] \\
&\leq 8 C_{\text{opt}} \overline{W}_T \sqrt{T} + 4 C_{\text{opt}} \overline{W}_T \sqrt{T} + \frac{8}{\sqrt{3}} C_{\text{exp}} \overline{W}_T \sqrt{T} + 8 C_{\text{add}} \overline{W}_T \\
&= B_{\text{PO}} \overline{W}_T \sqrt{T},
\end{aligned}$$

by the definition of B_{PO} in Lemma 3.2. On summing Equation 40 and using Equation 39 we get

$$\mathcal{J}_T^{\text{pop}} \leq (B_{\text{PO}} + B_Z) \overline{W}_T \sqrt{T} = B_F \overline{W}_T \sqrt{T}.$$

The PO-event and the population-martingale event each fail with probability at most ${}^{\delta}/_{3}$. With the success probability at least $1 - {}^{2\delta}/_{3}$, we have $\mathcal{J}_T^{\text{pop}} \leq B_F \overline{W}_T \sqrt{T}$. □

D.9 Proof of Theorem 4.3

Theorem D.8 (FIRE Bounds)**.** *On the event of Lemma 4.1, we have* $M\mathfrak{R}_T^{\text{stoc}} + e^{\lambda \zeta_T} - 1 \leq \mathcal{J}_T^{\text{pop}}$. *With the coefficients of Proposition 3.1, with probability at least* $1 - \delta$*, we get* $\mathfrak{R}_T^{\text{stoc}} \leq B_F \sqrt{T} + 1 = \widetilde{\mathcal{O}}(\sqrt{T})$*,* $\zeta_T \leq {}^{\log\left(2[B_F \sqrt{T} + 1 + HT]\right)}/_{\lambda}$*, and* $\mathfrak{V}_T^{\text{stoc}} \leq H + \zeta_T + 2\Gamma_\beta = \widetilde{\mathcal{O}}(\sqrt{T})$.

Proof. **Stochastic Regret Decomposition:** For any differentiable, convex function $\varphi(\cdot)$ and $y \geq x$, convexity yields $\varphi(x) \geq \varphi(y) + \varphi'(y)(x - y) \implies \varphi(y) - \varphi(x) \leq \varphi'(y)(y - x)$. Since $\zeta_t \geq \zeta_{t-1}$, we can easily write

$$\varphi(\zeta_t) - \varphi(\zeta_{t-1}) \leq \varphi'(\zeta_t)[g_t]^+ = \varphi'(\zeta_t) I_t g_t = w_t g_t.$$

For all $t \geq 2$, we have from $g_t \leq V^{\pi_t, P}(\boldsymbol{c}) \leq g_t + 2\beta_{t-1}(\pi_t)$ that $I_t g_t \leq I_t V^{\pi_t, P}(\boldsymbol{c})$. Multiplying by the non-negative $\varphi'(\zeta_t)$ preserves the inequality. For $t = 1$, we see that $I_1 = w_1 = 0$, and thus, the same conclusion holds. Hence, we get

$$M V^{\pi_t, P}(\boldsymbol{\ell}) + \varphi(\zeta_t) - \varphi(\zeta_{t-1}) \leq M V^{\pi_t, P}(\boldsymbol{\ell}) + w_t V^{\pi_t, P}(\boldsymbol{c}) = \bar{S}_t(\pi_t). \tag{41}$$

It is implied by the feasibility of $\pi^\star_{\text{stoc}}$ and $w_t \geq 0$ that

$$\bar{S}_t(\pi^\star_{\text{stoc}}) = MV^{\pi^\star_{\text{stoc}},P}(\boldsymbol{\ell}) + w_t V^{\pi^\star_{\text{stoc}},P}(\boldsymbol{c}) \leq MV^{\pi^\star_{\text{stoc}},P}(\boldsymbol{\ell}).$$

We subtract $\bar{S}_t(\pi^\star_{\text{stoc}})$ from both sides of Equation 41, sum over t, and telescope the drift from $\zeta_0 = 0$. Thus, the terms become $M\mathfrak{R}_T^{\text{stoc}}$ and $\varphi(\zeta_T) = e^{\lambda\zeta_T} - 1$, proving $M\mathfrak{R}_T^{\text{stoc}} + e^{\lambda\zeta_T} - 1 \leq \mathcal{J}_T^{\text{pop}}$.

A deterministic cap: The restart analysis is non-circular even before the good event is invoked. Indeed, $\widehat{V}_t \leq H$ and $\beta_{t-1}(\pi_t) \geq 0$, so $g_t \leq H$ for $t \geq 2$, while $g_1 = 0$. On every sample path, induction gives $\zeta_t \leq Ht$. Consequently, before interaction, the scale index is bounded by

$$K_F^0 = 1 + \left\lceil \log_2 \frac{M + \lambda e^{\lambda HT}}{W^{[1]}} \right\rceil^+.$$

$\lambda \leq 1/(2eB_F\sqrt{T})$ and $B_F \geq B_Z > 0$ show that the above quantity is $\mathcal{O}(\sqrt{T})$ for fixed problem dimensions. Note that the confidence was allocated using the simpler deterministic T^2 copy-prefix cap, not K_F^0 or the eventual realized count.

Bounding $\mathfrak{R}_T^{\text{stoc}}$ and $\mathfrak{V}_T^{\text{stoc}}$: We intersect the statistical event of Lemma 4.1, the PO−DB prefix event, and the population-martingale event. The intersection has a success probability of at least $1 - \delta$ by the confidence allocation in Section 4. We put $Z = e^{\lambda\zeta_T} \geq 1$, and the drift inequality just proved and Lemma 4.2 give

$$\mathfrak{R}_T^{\text{stoc}} + Z - 1 \leq B_F\sqrt{T}\left(1 + \lambda e^{\lambda H} Z\right).$$

The deterministic choices of the FIRE's parameters satisfy

$$\lambda H \leq 1,\ B_F\sqrt{T}\lambda e^{\lambda H} \leq eB_F\sqrt{T}\lambda = \frac{eB_F\sqrt{T}}{H + 2eB_F\sqrt{T}} \leq \frac{1}{2}.$$

On moving the resulting exponential term to the left, we get

$$\mathfrak{R}_T^{\text{stoc}} + \frac{1}{2}Z \leq B_F\sqrt{T} + 1. \tag{42}$$

Safely dropping $Z/2 \geq 0$ proves $\mathfrak{R}_T^{\text{stoc}} \leq B_F\sqrt{T} + 1$. As each episodic loss lies in $[0, H]$, we clearly have $\mathfrak{R}_T^{\text{stoc}} \geq -HT$. On inserting this lower bound in Equation 42, we have

$$\frac{1}{2}Z \leq B_F\sqrt{T} + 1 + HT$$
$$e^{\lambda\zeta_T} = Z \leq 2(B_F\sqrt{T} + 1 + HT).$$

Taking the logarithm and dividing both sides by $\lambda > 0$ proves $\zeta_T \leq \log\big(2[B_F\sqrt{T}+1+HT]\big)/\lambda$. For every $t \geq 2$, using $g_t \leq V^{\pi_t,P}(\boldsymbol{c}) \leq g_t + 2\beta_{t-1}(\pi_t)$ yields

$$V^{\pi_t,P}(\boldsymbol{c}) \leq g_t + 2\beta_{t-1}(\pi_t) \leq [g_t]^+ + 2\beta_{t-1}(\pi_t).$$

The right side of the above relation is non-negative, so we can write

$$[V^{\pi_t,P}(\boldsymbol{c})]^+ \leq [g_t]^+ + 2\beta_{t-1}(\pi_t).$$

Note that the first episode contributes at most H. On summing, using $\sum_t [g_t]^+ = \zeta_T$, and using $\sum_{t=2}^T \beta_{t-1}(\pi_t) \leq \Gamma_\beta$ gives

$$\mathfrak{V}_T^{\text{stoc}} \leq H + \sum_{t=2}^T [g_t]^+ + 2\sum_{t=2}^T \beta_{t-1}(\pi_t) \leq H + \zeta_T + 2\Gamma_\beta,$$

Recall that B_F and Γ_β contain only problem-dependent and logarithmic factors, and $\lambda^{-1} = H + 2eB_F\sqrt{T}$. Thus, the regret $\mathfrak{R}_T^{\text{stoc}}$ and strong violation $\mathfrak{V}_T^{\text{stoc}}$ are both $\widetilde{\mathcal{O}}(\sqrt{T})$.

We sharpen the restart count only aftproving er the preceding good evved. From the queue bound, we have

$$\overline{W}_T \leq M + 2\lambda e^{\lambda H}(B_F\sqrt{T} + 1 + HT) \leq M + \frac{B_F\sqrt{T} + 1 + HT}{B_F\sqrt{T}},$$

where $\lambda e^{\lambda H} \leq e\lambda \leq (2B_F\sqrt{T})^{-1}$ was used. Hence, $\overline{W}_T$ is polynomial in T, and the realized scale index obeys $1 + \left\lceil \log_2 \overline{W}_T/W^{[1]} \right\rceil^+ = \mathcal{O}(\log T)$. This final sharpening plays no role in confidence allocation. □

Explicit Dimension Dependence and Admissible λ: Let $S = |X|$ and $D = S|A|$ as in Proposition 3.1. Its coefficients give

1. $C_{\mathrm{opt}}\sqrt{T} = H^{1.5}\sqrt{S|A|T}\max\{K, 18\}$,
2. $C_{\mathrm{exp}}\sqrt{T} = 480H^{1.5}S^2\sqrt{|A|LT}$, and
3. $8C_{\mathrm{add}} = 192H^4K + 1920HS^{2.5}|A|\sqrt{L}\log(D/\varepsilon_{\mathrm{PO}})$.

We have $B_Z\sqrt{T} = 16\sqrt{2}H\sqrt{L_Z T} + 32HL_Z/3$. Since $\varepsilon_{\mathrm{PO}} = \delta/27T^2$ and $H \le S$, the factors K, L, $\log(D/\varepsilon_{\mathrm{PO}})$, and L_Z are all $\mathcal{O}(\log(S|A|T/\delta))$. Hence, we get $B_F\sqrt{T} = (12C_{\mathrm{opt}} + 8/\sqrt{3}C_{\mathrm{exp}} + B_Z)\sqrt{T} + 8C_{\mathrm{add}} = \widetilde{\mathcal{O}}(H^{1.5}S^2\sqrt{|A|T} + H^4 + HS^{2.5}|A|)$. In Equation 34, we have $A_\beta = \widetilde{\mathcal{O}}(HS)$ and $\Gamma_P = \widetilde{\mathcal{O}}(HS[\sqrt{S|A|HT} + S|A|])$, so $\Gamma_\beta = \widetilde{\mathcal{O}}(H^{2.5}S^{1.5}\sqrt{|A|T} + H^2S^2|A|)$, and $\zeta_T \le \lambda^{-1}\log(2[B_F\sqrt{T} + 1 + HT]) = \widetilde{\mathcal{O}}(H + B_F\sqrt{T})$. Finally, the proof above uses λ only through $\lambda H \le 1$ and $B_F\sqrt{T}\lambda e^{\lambda H} \le 1/2$, which persist for every smaller $\lambda > 0$. Hence, Theorem 4.3 holds for every $\lambda \in (0, 1/(H+2eB_F\sqrt{T})]$, with its bound on ζ_T evaluated at that λ.

D.10 More on Section 4.3

As already seen in Section 4.2, FIRE gives high-probability upper bounds on the stochastic regret and strong violation. In this subsection, we derive a result that clarifies exactly what is optimal about their common $\sqrt{T}$ dependence, i.e., it is a *joint* lower bound. To put it differently, in the construction below, always choosing a_0 gives zero violation but linear regret under ν_0. Thus, we assert that FIRE's $\sqrt{T}$ regret and strong violation exponents are Pareto-optimal. We do not claim anything about the individual minimax optimality of FIRE's strong violation. To connect the high-probability upper bounds to expectations, we choose $\delta = T^{-2}$ for $T \ge 2$. Both quantities are at most HT on every path, so the failure event adds at most H/T to either upper bound. Therefore, the tradeoff below also applies to FIRE's $\widetilde{\mathcal{O}}(\sqrt{T})$ expected-regret guarantee.

Let $\mathbb{P} \ll \mathbb{Q}$ denote that the measure $\mathbb{P}$ is absolutely continuous with respect to the measure $\mathbb{Q}$, and let $\mathrm{TV}(\mathbb{P}, \mathbb{Q})$ denote the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$. We use the following lemma to show how much a bounded statistic's expectation can differ between two probability measures. One should note that a small KL divergence keeps these expectations close, thereby quantifying the consequences of statistical indistinguishability. In our construction, it connects regret on one instance to strong violation on the other.

Lemma D.9 (Bounding the Change-of-Measure)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space, let X be measurable with $0 \le X \le T$, and suppose $\mathbb{P} \ll \mathbb{Q}$. Then, we can show*

$$|\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]| \le T \cdot \mathrm{TV}(\mathbb{P}, \mathbb{Q}) \le T\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}\|\mathbb{Q})}. \tag{43}$$

Proof. For every outcome, the layer-cake representation gives $X = \int_0^T \mathbf{1}\{X > s\}\, ds$. Therefore, by Tonelli's theorem, we have

$$\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X] = \int_0^T \big(\mathbb{P}(X > s) - \mathbb{Q}(X > s)\big)\, ds.$$

By the definition $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|$, the absolute value of every integrand is at most $\mathrm{TV}(\mathbb{P}, \mathbb{Q})$. Taking absolute values and integrating proves the first inequality in Equation 43. The application of Pinsker's inequality, i.e., $\mathrm{TV}(\mathbb{P}, \mathbb{Q})^2 \le \mathrm{KL}(\mathbb{P}\|\mathbb{Q})/2$, proves the second inequality in Equation 43.

For the sake of completeness, the scalar inequality behind the application of Pinsker is

$$\mathrm{kl}(p\|q) \ge 2(p-q)^2,$$

where kl denotes the binary relative entropy. Indeed, for fixed $q \in (0,1)$, the second derivative with respect to p of $\mathrm{kl}(p\|q) - 2(p-q)^2$ equals $1/[p(1-p)] - 4 \ge 0$. Also, the function and its first derivative vanish at $p = q$. For $p \in \{0,1\}$ the inequality follows by continuity, and for $q \in \{0,1\}$ it follows directly using extended relative entropy. The data-processing inequality gives $\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \ge \mathrm{kl}(\mathbb{P}(A)\|\mathbb{Q}(A))$ for every measurable event A. Taking the supremum over A yields Pinsker's inequality. □

We state and prove in Theorem 4.4 that, without Slater's condition, any online algorithm guaranteeing $\mathcal{O}(\sqrt{T})$ expected regret uniformly must incur $\Omega(\sqrt{T})$ expected strong violation on some instance. This limitation holds even with known transition and full feedback. Additionally, we also establish a corresponding probability tradeoff.

Theorem D.10 (Pareto Lower Bound)**.** *For every integer $T \geq 1$, there exist two stochastic CMDP instances ν_0 and ν_1, which depend only on T and satisfy: 1) Horizon $H = 1$ with one non-terminal state, two actions–a_0 (safe action) and a_1 (risky action), and known and deterministic transitions. Both instances have a non-empty feasible set but no strictly feasible policy; 2) The losses and costs are i.i.d. across episodes and lie in $[0, 1]$ and $[-1, 1]$, respectively; 3) Every possibly randomized online algorithm, even under full feedback, has $\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\text{stoc}}] + 2\sqrt{T}\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\text{stoc}}] \geq (1 - 1/2\sqrt{2})T$; and 4) every such algorithm satisfies $\mathbb{P}_{\nu_0}(\mathfrak{R}_T^{\text{stoc}} > T/2) + \mathbb{P}_{\nu_1}(\mathfrak{V}_T^{\text{stoc}} \geq \sqrt{T}/4) \geq 1 - 1/2\sqrt{2}$.*

Consequently, no algorithm can guarantee $\mathcal{O}(\sqrt{T})$ expected stochastic regret and $o(\sqrt{T})$ expected strong violation simultaneously and uniformly without Slater's condition. More precisely, if an algorithm satisfies $\mathbb{E}[\mathfrak{R}_T^{\text{stoc}}] \leq C\sqrt{T}$ uniformly, then, whenever $\sqrt{T} \geq 2C/1-1/(2\sqrt{2})$, the instance ν_1 satisfies $\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\text{stoc}}] \geq 1/4(1 - 1/2\sqrt{2})\sqrt{T}$.

Without imposing a regret guarantee, the tradeoff still implies $\max\{\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\text{stoc}}], \mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\text{stoc}}]\} \geq 1/3(1 - 1/2\sqrt{2})\sqrt{T}$. If for some $C > 0$ and $\delta \in (0, 1)$ an algorithm uniformly guarantees $\mathbb{P}(\mathfrak{R}_T^{\text{stoc}} \leq C\sqrt{T}) \geq 1 - \delta$ then, whenever $\sqrt{T} > 2C$, we have $\mathbb{P}_{\nu_1}(\mathfrak{V}_T^{\text{stoc}} \geq \sqrt{T}/4) \geq 1 - \frac{1}{2\sqrt{2}} - \delta$.

Proof. We fix a $T \geq 1$ and set $\epsilon := 1/2\sqrt{T}$. Since $T \geq 1$, we have $0 < \epsilon \leq 1/2$. We denote the two actions by a_0 (*safe* action) and a_1 (*risky* action). In both instances ν_0 and ν_1, one action is chosen at the only non-terminal state, and then moves deterministically to the terminal state. Thus, the transition is known and contains no statistical uncertainty.

In every episode and in both instances, we use the deterministic losses as

$$\ell_t(a_0) = 1,\ \ell_t(a_1) = 0. \tag{44}$$

The costs of a_0 are identically zero and the cost of a_1 is a random variable $Y_t \in \{-1, +1\}$, independent across episodes. Under the instance ν_0 let

$$\mathbb{P}_{\nu_0}(Y_t = +1) = \mathbb{P}_{\nu_0}(Y_t = -1) = \frac{1}{2}.$$

On the other hand, under the instance ν_1 let

$$\mathbb{P}_{\nu_1}(Y_t = +1) = \frac{1+\epsilon}{2},\ \mathbb{P}_{\nu_1}(Y_t = -1) = \frac{1-\epsilon}{2}.$$

Recall that all losses belong to $[0, 1]$ and all costs belong to $[-1, 1]$. Let a stationary policy choose a_1 with probability $p \in [0, 1]$. Its mean cost under ν_0 is zero, because $\mathbb{E}_{\nu_0}[Y_t] = 0$. Therefore, every policy is feasible under ν_0, but no policy is strictly feasible. By Equation 44, the unique loss-minimizing feasible policy under ν_0 is $p = 1$ (namely, always choose a_1). Under the instance ν_1, we have

$$\mathbb{E}_{\nu_1}[Y_t] = (+1)\frac{1+\epsilon}{2} + (-1)\frac{1-\epsilon}{2} = \epsilon.$$

Hence, the mean cost of the policy is $p\epsilon$. The feasibility implies $p\epsilon \leq 0$, together with $p \geq 0$ and $\epsilon > 0$, forces $p = 0$. Thus, a_0 is the unique feasible policy under ν_1, and its mean cost is exactly zero. The second instance also has no strictly feasible policy.

Let $p_t \in [0, 1]$ be the conditional probability (determined by the algorithm's history and internal randomness) of choosing a_1 in episode t, and we put $\mathcal{N}_T := \sum_{t=1}^{T} p_t$. The random variable $\mathcal{N}_T$ satisfies $0 \leq \mathcal{N}_T \leq T$. Under the instance ν_0, the comparator always chooses a_1 and has loss zero. The learner's expected loss in episode t, conditional on its selected policy, is $1 - p_t$. Consequently, pathwise as a function of the algorithm's selected policies, we have

$$\mathfrak{R}_T^{\text{stoc}} = T - \mathcal{N}_T \text{ under } \nu_0,\ \mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\text{stoc}}] = T - \mathbb{E}_{\nu_0}[\mathcal{N}_T]. \tag{45}$$

Under the instance ν_1, the population cost of the episode-t policy is $p_t\epsilon \geq 0$. By the definition of positive violation, we have

$$\mathfrak{V}_T^{\mathrm{stoc}} = \sum_{t=1}^{T} [p_t\epsilon]^+ = \epsilon\mathcal{N}_T \text{ under } \nu_1,\ \mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] = \epsilon\mathbb{E}_{\nu_1}[\mathcal{N}_T]. \tag{46}$$

Now, we need to show that the two expectations of $\mathcal{N}_T$ cannot be far apart. We give the algorithm the entire vector (ℓ_t, c_t) after episode t, which is strictly at least as much information as the bandit feedback.

Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the distributions of the resulting full transcript, including the algorithm's internal randomization, under ν_0 and ν_1. The algorithm uses the same conditional action-selection function after any common transcript. Note that the losses and the safe-action costs are identical in the two instances. Thus, the chain rule for relative entropy leaves only the T independent observations of Y_t and gives

$$\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) = T \cdot \mathrm{kl}\left(\frac{1}{2}\middle\|\frac{1+\epsilon}{2}\right). \tag{47}$$

The Bernoulli divergence in Equation 47 is

$$\mathrm{kl}\left(\frac{1}{2}\middle\|\frac{1+\epsilon}{2}\right) = \frac{1}{2}\log\frac{1/2}{(1+\epsilon)/2} + \frac{1}{2}\log\frac{1/2}{(1-\epsilon)/2} = -\frac{1}{2}\log(1-\epsilon^2).$$

For every $x \in [0,1)$, $-\log(1-x) = \int_0^x (1-u)^{-1}\,du \leq x/(1-x)$. Taking $x = \epsilon^2$ and using $\epsilon^2 \leq 1/4$ yields

$$-\frac{1}{2}\log(1-\epsilon^2) \leq \frac{\epsilon^2}{2(1-\epsilon^2)} \leq \frac{2}{3}\epsilon^2 \leq \epsilon^2. \tag{48}$$

Combining Equation 47 and Equation 48, and substituting $\epsilon^2 = 1/4T$, gives

$$\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \leq T\epsilon^2 = \frac{1}{4}. \tag{49}$$

Pinsker's inequality and Equation 49 also give

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1)} \leq \frac{1}{2\sqrt{2}}. \tag{50}$$

We apply Lemma D.9 to $\mathcal{N}_T \in [0,T]$ and use Equation 49, to obtain

$$|\mathbb{E}_{\nu_0}[\mathcal{N}_T] - \mathbb{E}_{\nu_1}[\mathcal{N}_T]| \leq \frac{T}{2\sqrt{2}}.$$

Equation 45 and Equation 46, together with $1/\epsilon = 2\sqrt{T}$, now imply

$$\begin{aligned}\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}] + 2\sqrt{T}\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] &= T - \mathbb{E}_{\nu_0}[\mathcal{N}_T] + \mathbb{E}_{\nu_1}[\mathcal{N}_T] \\ &\geq (T - |\mathbb{E}_{\nu_0}[\mathcal{N}_T] - \mathbb{E}_{\nu_1}[\mathcal{N}_T]|) \\ &\geq T\,(1 - 1/2\sqrt{2})\,.\end{aligned}$$

This proves $\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}] + 2\sqrt{T}\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] \geq (1 - 1/2\sqrt{2})T$. Now, suppose $\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}] \leq C\sqrt{T}$ and rearranging the identity gives

$$\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] \geq \frac{1}{2}\left(1 - \frac{1}{2\sqrt{2}}\right)\sqrt{T} - \frac{C}{2}.$$

The condition $\sqrt{T} \geq 2C/1-1/(2\sqrt{2})$ implies $C/2 \leq [1-1/(2\sqrt{2})]\sqrt{T}/4$. Substituting this inequality into the above relation proves $\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] \geq 1/4(1 - 1/2\sqrt{2})\sqrt{T}$.

We suppose that $M_T \coloneqq \max\{\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}], \mathbb{E}_{\nu_1}\mathfrak{V}_T^{\mathrm{stoc}}\}$. As both terms are nonnegative, the left-hand side of $\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}] + 2\sqrt{T}\mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}] \geq (1 - 1/2\sqrt{2})T$ is at most $(1+2\sqrt{T})M_T$. Since $T \geq 1$ implies $1 + 2\sqrt{T} \leq 3\sqrt{T}$, we obtain

$$M_T \geq \frac{[1 - 1/2\sqrt{2}]T}{1 + 2\sqrt{T}} \geq \frac{1}{3}\left(1 - \frac{1}{2\sqrt{2}}\right)\sqrt{T},$$

Algorithm 3 In-Hindsight Constrained Exploration (ICE)

Require: T, H, $|X|$, $|A|$, δ.

Set $\varepsilon_{\mathrm{PO}} = \delta/(27T^2)$ and compute C_{opt}, $C_{\exp}$, C_{add}, B_{PO}, B_{MD}, and $B = B_{\mathrm{PO}} + B_{\mathrm{MD}}$.
Set $M = \sqrt{T}$. For General ICE, use $\varphi(z) = z^2/2$. For Conditional ICE, use $\lambda = 1/[(4B + 2eH^2)\sqrt{T}]$ and $\varphi(z) = e^{\lambda z} - 1$.
Set $\zeta_0 = 0$, $k = 1$, and $W^{[1]} = M + \varphi'(0)$.
Initialize a fresh `PO-DB` copy with $(T, \eta_{\mathrm{PO}}, \varepsilon_{\mathrm{PO}})$.
for $t = 1, \dots, T$ **do**
  Set $w_t = \varphi'(\zeta_{t-1})$ and $\widehat{W}_t = M + w_t$.
  if $\widehat{W}_t > W^{[k]}$ **then**
    while $\widehat{W}_t > W^{[k]}$ **do**
      Set $W^{[k+1]} = 2W^{[k]}$ and $k \leftarrow k + 1$.
    end while
    Discard the old optimizer state and initialize one fresh copy with $(T, \eta_{\mathrm{PO}}, \varepsilon_{\mathrm{PO}})$.
  end if
  Obtain π_t from the active copy and execute it for one episode.
  Observe the loss and cost on every visited state-action pair.
  Feed the visited normalized losses $u_{t,h}(x_{t,h}, a_{t,h}) = [M\ell_{t,h}(x_{t,h}, a_{t,h}) + w_t c_{t,h}(x_{t,h}, a_{t,h}) + W^{[k]}]/(2W^{[k]})$ to the active copy.
  Compute $g_t = \sum_{h=0}^{H-1} c_{t,h}(x_{t,h}, a_{t,h})$ and update $\zeta_t = [\zeta_{t-1} + g_t]^+$.
end for

which proves $\max\{\mathbb{E}_{\nu_0}[\mathfrak{R}_T^{\mathrm{stoc}}], \mathbb{E}_{\nu_1}[\mathfrak{V}_T^{\mathrm{stoc}}]\} \geq 1/3(1 - 1/2\sqrt{2})\sqrt{T}$. It remains to prove the probability statements. We define the transcript event as $A \coloneqq \{\mathcal{N}_T < T/2\}$. Since $\mathfrak{R}_T^{\mathrm{stoc}} = T - \mathcal{N}_T$ under the instance ν_0, we have $A \subseteq \{\mathfrak{R}_T^{\mathrm{stoc}} > T/2\}$. Since $\mathfrak{V}_T^{\mathrm{stoc}} = \epsilon\mathcal{N}_T$ under ν_1 and $\epsilon = 1/2\sqrt{T}$, we again have $A^{\mathsf{c}} \subseteq \{\mathfrak{V}_T^{\mathrm{stoc}} \geq \sqrt{T}/4\}$. Therefore, by the definition of total variation and Equation 50,

$$\begin{aligned}
\mathbb{P}_{\nu_0}\left(\mathfrak{R}_T^{\mathrm{stoc}} > T/2\right) + \mathbb{P}_{\nu_1}\left(\mathfrak{V}_T^{\mathrm{stoc}} \geq \sqrt{T}/4\right) &\geq \mathbb{P}_0(A) + \mathbb{P}_1(A^{\mathsf{c}}) \\
&= 1 + \mathbb{P}_0(A) - \mathbb{P}_1(A) \\
&\geq 1 - |\mathbb{P}_0(A) - \mathbb{P}_1(A)| \\
&\geq 1 - \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \geq 1 - \frac{1}{2\sqrt{2}}.
\end{aligned}$$

This proves $\mathbb{P}_{\nu_0}(\mathfrak{R}_T^{\mathrm{stoc}} > T/2) + \mathbb{P}_{\nu_1}(\mathfrak{V}_T^{\mathrm{stoc}} \geq \sqrt{T}/4) \geq 1 - 1/2\sqrt{2}$. Finally, suppose $\mathbb{P}(\mathfrak{R}_T^{\mathrm{stoc}} \leq C\sqrt{T}) \geq 1 - \delta$ holds and $\sqrt{T} > 2C$. Then, $T/2 > C\sqrt{T}$, and hence

$$\mathbb{P}_{\nu_0}(\mathfrak{R}_T^{\mathrm{stoc}} > T/2) \leq \mathbb{P}_{\nu_0}(\mathfrak{R}_T^{\mathrm{stoc}} > C\sqrt{T}) \leq \delta.$$

On subtracting the above bound from $\mathbb{P}_{\nu_0}(\mathfrak{R}_T^{\mathrm{stoc}} > T/2) + \mathbb{P}_{\nu_1}(\mathfrak{V}_T^{\mathrm{stoc}} \geq \sqrt{T}/4) \geq 1 - 1/2\sqrt{2}$ proves $\mathbb{P}_{\nu_1}(\mathfrak{V}_T^{\mathrm{stoc}} \geq \sqrt{T}/4) \geq 1 - \frac{1}{2\sqrt{2}} - \delta$. Thus, our algorithmic construction was full feedback and known deterministic transitions. Any bandit feedback algorithm with unknown transitions is a special case of an algorithm facing this more informative protocol. Hence, the lower bound also applies. □

Remark D.11. *Theorem 4.4 does not prove that every algorithm must incur $\Omega(\sqrt{T})$ strong violation on every instance, or even on some instance if regret is ignored. In other words, in the displayed two-instance construction, always choosing a_0 gives zero violation but incurs linear regret under ν_0.*

E Omitted Details for the Adversarial Regime

E.1 Proof of Lemma 5.1

Lemma E.1 (Uniform Queue-Noise Concentration)**.** *With probability at least $1 - \delta/3$, simultaneously for every $u \leq T$, we have $\sum_{t=1}^{u} D_t \leq B_{\mathrm{MD}} W_{\max}\sqrt{T}$, where $B_{\mathrm{MD}} \coloneqq 8H\sqrt{2\log\left(3T^2/\delta\right)}$.*

Proof. By the predictability of w_t, the fact that $\mathbb{E}[g_t \mid \mathcal{G}_t^{\mathrm{I}}] = V^{\pi_t,P}(\boldsymbol{c}_t)$, and the definition $D_t \coloneqq w_t\left(g_t - V^{\pi_t,P}(\boldsymbol{c}_t)\right)$, we obtain

$$\mathbb{E}[D_t \mid \mathcal{G}_t^{\mathrm{I}}] = w_t\left(\mathbb{E}[g_t \mid \mathcal{G}_t^{\mathrm{I}}] - V^{\pi_t,P}(\boldsymbol{c}_t)\right) = 0.$$

Thus, (D_t) is a martingale-difference sequence. Let $\mathcal{W}_t = W^{[k_t]}$ be the active scale in episode t. $\mathcal{W}_t$ is $\mathcal{G}_t^{\mathrm{I}}$-measurable, and the range test yields $M + w_t \le \mathcal{W}_t$. Since $|g_t| \le H$ and $|V^{\pi_t,P}(\boldsymbol{c}_t)| \le H$, we have

$$|D_t| \le 2Hw_t \le 2H\mathcal{W}_t.$$

Let $\varrho_t \in [T]$ be the ordinal of the nonempty outer copy receiving episode t. As the trigger is evaluated before the episode-t trajectory, ϱ_t is $\mathcal{G}_t^{\mathrm{I}}$-measurable. For each deterministic copy ordinal r, we define

$$U_t^{(r)} \coloneqq \mathbf{1}\{\varrho_t = r\}\frac{D_t}{\mathcal{W}_t}.$$

Because the selector and denominator are predictable, $(U_t^{(r)})_{t\le T}$ is a martingale-difference sequence and $|U_t^{(r)}| \le 2H$. For every deterministic pair $(r,u) \in [T]^2$, the Azuma-Hoeffding inequality gives

$$\Pr\left(\sum_{t=1}^{u} U_t^{(r)} > 2H\sqrt{2u\log\left(\frac{3T^2}{\delta}\right)}\right) \le \frac{\delta}{3T^2}. \tag{51}$$

A deterministic union bound over all (r,u)-pairs has cost at most $\delta/3$. Now, suppose we fix a global prefix u and let ω_r be the scale of the r-th nonempty copy. Consecutive nonempty scales satisfy $\omega_{r+1} \ge 2\omega_r$, even if a range jump skips unused levels. The last active scale is at most twice the largest required range. Thus, we see that $\sum_{r:\ r \text{ started by } u} \omega_r < 2\omega_{\varrho_u} \le 4W_{\max}$.

On the simultaneous event from Equation 51, we have

$$\begin{aligned}
\sum_{t=1}^{u} D_t &= \sum_r \omega_r \sum_{t=1}^{u} U_t^{(r)} \\
&\le 2H\sqrt{2u\log\left(\frac{3T^2}{\delta}\right)} \sum_{r:\ r \text{ started by } u} \omega_r \\
&\le 8H\sqrt{2\log\left(\frac{3T^2}{\delta}\right)} W_{\max}\sqrt{T} \\
&= B_{\mathrm{MD}} W_{\max}\sqrt{T}.
\end{aligned}$$

Since the union bound covers every deterministic global prefix, the conclusion holds simultaneously for all $u \le T$. □

E.2 Restart Bounds

Let K_Q and K_E count the nonempty outer copies of ICE, including the initial copy, for the quadratic and exponential potentials, respectively.

Lemma E.2 (Restart Bounds)**.** *On every sample path, we have $0 \le \zeta_t \le Ht$. The quadratic potential $\varphi(z) = {}^{z^2}/_2$ satisfies $K_Q \le 1 + \lceil \log_2 ({}^{M+HT}/_M) \rceil$, whereas the exponential potential $\varphi(z) = e^{\lambda z} - 1$ satisfies the crude deterministic bound $K_E \le 1 + \left\lceil \log_2 \left({}^{M+\lambda e^{\lambda HT}}/_{M+\lambda}\right)\right\rceil$.*

Proof. The queue ζ_t is non-negative by definition. Because $g_t \le H$, we can write $\zeta_t = [\zeta_{t-1}+g_t]^+ \le [\zeta_{t-1} + H]^+ = \zeta_{t-1} + H$, where the last equality uses $\zeta_{t-1} \ge 0$. Starting from $\zeta_0 = 0$, induction gives $0 \le \zeta_t \le Ht$ for every t.

For the quadratic potential, $\varphi'(z) = z$. Thus, every required range is at most $M + HT$, while the initial range is M. Since each restart increases the active guess by at least a factor of two, the result $K_Q \le 1 + \lceil \log_2 ({}^{M+HT}/_M) \rceil$ holds.

For the exponential potential, $\varphi'(z) = \lambda e^{\lambda z}$. Every required range is at most $M + \lambda e^{\lambda HT}$, while the initial range is $M + \lambda$. By geometric doubling, we get $K_E \le 1 + \left\lceil \log_2 \left({}^{M+\lambda e^{\lambda HT}}/_{M+\lambda}\right)\right\rceil$. Both of these statements hold on every sample path before any confidence event is invoked. □

E.3 Proof of Lemma 5.2

Lemma E.3 (Weak-Violation Bridge). *Independent of the potential, with probability at least* $1 - \delta/3$, *we have* $\mathfrak{V}_T^{\mathrm{adv}} \le \zeta_T + 2H\sqrt{2T\log(3/\delta)}$.

Proof. We define $\Delta_t := V^{\pi_t,P}(\boldsymbol{c}_t) - g_t$. By $\mathbb{E}[g_t \mid \mathcal{G}_t^{\mathrm{I}}] = V^{\pi_t,P}(\boldsymbol{c}_t)$, it is clearly implied that $\mathbb{E}[\Delta_t \mid \mathcal{G}_t^{\mathrm{I}}] = 0$, and $|\Delta_t| \le 2H$. Thus, the Azuma-Hoeffding inequality with failure $\delta/3$ gives

$$\sum_{t=1}^{T} \Delta_t \le 2H\sqrt{2T\log\left(\frac{3}{\delta}\right)}. \tag{52}$$

Moreover, the queue recursion implies $\zeta_t = [\zeta_{t-1} + g_t]^+ \ge \zeta_{t-1} + g_t$. Summing and using $\zeta_0 = 0$ yields $\sum_{t=1}^{T} g_t \le \zeta_T$. On the event of Equation 52, we have

$$\begin{aligned}\sum_{t=1}^{T} V^{\pi_t,P}(\boldsymbol{c}_t) &= \sum_{t=1}^{T} g_t + \sum_{t=1}^{T} \Delta_t \\ &\le \zeta_T + 2H\sqrt{2T\log\left(\frac{3}{\delta}\right)}.\end{aligned}$$

The right-hand side of the above relation is nonnegative, so taking the positive part on the left proves $\mathfrak{V}_T^{\mathrm{adv}} \le \zeta_T + 2H\sqrt{2T\log(3/\delta)}$. □

E.4 Proof of Lemma 5.3

Lemma E.4 (Exponential Drift-plus-Penalty). *With* $0 < \lambda H \le 1$, *we have* $\forall t \in [T]$: $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \varphi(\zeta_t) - \varphi(\zeta_{t-1}) \le V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{eH^2\lambda^2}{2}e^{\lambda\zeta_{t-1}} + 1$.

Proof. Since we have chosen $\varphi(z) = e^{\lambda z} - 1$, we observe

$$\varphi(\zeta_t) - \varphi(\zeta_{t-1}) = e^{\lambda\zeta_t} - e^{\lambda\zeta_{t-1}}. \tag{53}$$

For every real x, we have $e^{[x]^+} \le e^x + 1$. When $x \ge 0$, the left side of the inequality is e^x, while for $x < 0$ it is $1 < 1 + e^x$. Applying this with $x = \lambda(\zeta_{t-1} + g_t)$ gives

$$e^{\lambda\zeta_t} \le e^{\lambda\zeta_{t-1}}e^{\lambda g_t} + 1. \tag{54}$$

On substituting Equation 54 into Equation 53, we get

$$\varphi(\zeta_t) - \varphi(\zeta_{t-1}) \le e^{\lambda\zeta_{t-1}}(e^{\lambda g_t} - 1) + 1. \tag{55}$$

Taylor's theorem states that, for some ξ between 0 and y, we have $e^y - 1 = y + e^{\xi}y^2/2$. When $y \le 1$, this is at most $y + ey^2/2$. Since $|g_t| \le H$ and $\lambda H \le 1$, we have

$$e^{\lambda g_t} - 1 \le \lambda g_t + \frac{e}{2}\lambda^2 H^2. \tag{56}$$

On combining Equation 55 and Equation 56, and using $w_t = \lambda e^{\lambda\zeta_{t-1}}$, we have

$$\varphi(\zeta_t) - \varphi(\zeta_{t-1}) \le w_t g_t + \frac{eH^2\lambda^2}{2}e^{\lambda\zeta_{t-1}} + 1.$$

We add $MV^{\pi_t,P}(\boldsymbol{\ell}_t)$ to both sides, and add and subtract $w_t V^{\pi_t,P}(\boldsymbol{c}_t)$ on the right side of the above inequality. Thus, the definition of $\boldsymbol{s}_t$ and of D_t proves $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \varphi(\zeta_t) - \varphi(\zeta_{t-1}) \le V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{eH^2\lambda^2}{2}e^{\lambda\zeta_{t-1}} + 1$. □

E.5 Proof of Theorem 5.4

Theorem E.5 (Conditional ICE Bounds)**.** *We set $M = \sqrt{T}$, $\alpha = 4B + 2eH^2$, and $\lambda = 1/\alpha\sqrt{T}$. With the coefficients of Proposition 3.1, assuming Equation 10 holds almost surely for all prefixes, with probability at least $1 - \delta$, we have: $\mathfrak{R}_T^{\mathrm{adv}} \le \widetilde{\mathcal{O}}(\sqrt{T})$ and $\mathfrak{V}_T^{\mathrm{adv}} = \widetilde{\mathcal{O}}(\sqrt{T})$.*

Proof. Let $t^\star \in \arg\max_{t\in[T]} \zeta_t$, we define $\zeta_{\max} \coloneqq \zeta_{t^\star}$, and $Y \coloneqq e^{\lambda\zeta_{\max}}$. Episode-wise feasibility of $\pi^\star_{\mathrm{adv}}$ and $w_t \ge 0$ imply

$$V^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{s}_t) = MV^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{\ell}_t) + w_t V^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{c}_t) \le MV^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{\ell}_t). \tag{57}$$

We subtract $MV^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{\ell}_t)$ from the two sides of Lemma 5.3. By Equation 57, replacing the resulting term $-MV^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{\ell}_t)$ on the right by $-V^{\pi^\star_{\mathrm{adv}},P}(\boldsymbol{s}_t)$ can only enlarge that side. Summing through an arbitrary global prefix u and telescoping the potential gives

$$M\mathfrak{R}_u^{\mathrm{adv}}(\pi^\star_{\mathrm{adv}}) + e^{\lambda\zeta_u} - 1 \le \mathcal{J}_u^{\mathrm{I}} + \sum_{t=1}^{u} D_t + \frac{eH^2\lambda^2}{2}\sum_{t=1}^{u} e^{\lambda\zeta_{t-1}} + u. \tag{58}$$

Now, we intersect the events of Equation 9 and Lemma 5.1. Both are uniform in u, so they can be applied at the random prefix $t^\star$. Moreover, for any $u \le T$, we have

$$W_{\max} \le M + \lambda Y,\ \sum_{t=1}^{u} e^{\lambda\zeta_{t-1}} \le TY.$$

Therefore, by Equation 58, we get

$$M\mathfrak{R}_u^{\mathrm{adv}}(\pi^\star_{\mathrm{adv}}) + e^{\lambda\zeta_u} - 1 \le B\sqrt{T}(M + \lambda Y) + \frac{eH^2\lambda^2}{2}TY + T.$$

On inserting $M = \sqrt{T}$ and $\lambda = 1/\alpha\sqrt{T}$, we obtain

$$\sqrt{T}\mathfrak{R}_u^{\mathrm{adv}}(\pi^\star_{\mathrm{adv}}) + e^{\lambda\zeta_u} - 1 \le (B+1)T + \left(\frac{B}{\alpha} + \frac{eH^2}{2\alpha^2}\right)Y.$$

Since $\alpha = 4B + 2eH^2$, we have $B/\alpha \le 1/4$. Moreover, $\alpha \ge 2eH^2 \ge 2$, and hence

$$\frac{eH^2}{2\alpha^2} \le \frac{1}{4\alpha} \le \frac{1}{8}.$$

The sum of the two coefficients is at most $3/8 < 1/2$. Thus, for every $u \le T$, we have

$$\sqrt{T}\cdot\mathfrak{R}_u^{\mathrm{adv}}(\pi^\star_{\mathrm{adv}}) + e^{\lambda\zeta_u} - 1 \le (B+1)T + \frac{1}{2}Y. \tag{59}$$

At $u = t^\star$, the condition in Equation 10 makes the regret term non-negative and $e^{\lambda\zeta_{t^\star}} = Y$. Hence, we observe

$$\begin{aligned} Y - 1 &\le (B+1)T + \frac{1}{2}Y \\ \implies Y &\le 2(B+1)T + 2. \end{aligned} \tag{60}$$

On taking logarithms and using $\lambda^{-1} = \alpha\sqrt{T}$, we get

$$\zeta_{\max} \le \alpha\sqrt{T}\log(2(B+1)T + 2) = \widetilde{\mathcal{O}}(\sqrt{T}).$$

It also follows that

$$W_{\max} \le \sqrt{T} + \frac{2(B+1)T + 2}{\alpha\sqrt{T}} = \widetilde{\mathcal{O}}(\sqrt{T}).$$

Now, we sharpen the deterministic exponential restart cap. Substitution of $W_{\max}$ into the scale count gives $K_E = \mathcal{O}(\log T)$. We did not use this sharper conclusion to set confidence. At $u = T$, we drop the non-negative term $e^{\lambda\zeta_T} - 1$ from Equation 59 and use Equation 60 to get

$$\begin{aligned} \sqrt{T}\cdot\mathfrak{R}_T^{\mathrm{adv}} &\le (B+1)T + \frac{1}{2}Y \\ &\le 2(B+1)T + 1. \end{aligned}$$

Dividing by $\sqrt{T}$ proves the regret bound in $\mathfrak{R}_T^{\mathrm{adv}} \le \widetilde{\mathcal{O}}(\sqrt{T})$. On the event of Lemma 5.2, we have

$$\begin{aligned}\mathfrak{V}_T^{\mathrm{adv}} &\le \zeta_T + 2H\sqrt{2T\log(3/\delta)} \\ &\le \zeta_{\max} + 2H\sqrt{2T\log(3/\delta)} = \widetilde{\mathcal{O}}(\sqrt{T}).\end{aligned}$$

The wrapper event, weighted-martingale event, and weak-violation event each fail with probability at most $\delta/3$. A union bound proves the claimed probability $1-\delta$. □

E.6 Proof of Lemma 5.5

Lemma E.6 (Quadratic Drift-plus-Penalty)**.** *For all $t \in [T]$, we have $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \frac{1}{2}\zeta_t^2 - \frac{1}{2}\zeta_{t-1}^2 \le V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{1}{2}H^2$.*

Proof. We recall the notation $[x]_+$ to mean $\max(0,x)$ for any $x \in \mathbb{R}$. Since $[x]_+^2 \le x^2$ for every real x, we have

$$\zeta_t^2 = [\zeta_{t-1} + g_t]_+^2 \le (\zeta_{t-1} + g_t)^2 = \zeta_{t-1}^2 + 2\zeta_{t-1}g_t + g_t^2.$$

We subtract ζ_{t-1}^2, divide by two, and use $g_t^2 \le H^2$ in the above relation to get

$$\frac{1}{2}\zeta_t^2 - \frac{1}{2}\zeta_{t-1}^2 \le \zeta_{t-1}g_t + \frac{1}{2}H^2.$$

In the above relation, we first add $MV^{\pi_t,P}(\boldsymbol{\ell}_t)$. Then, we add and subtract $\zeta_{t-1}V^{\pi_t,P}(\boldsymbol{c}_t)$. As $w_t = \zeta_{t-1}$, the definitions of $\boldsymbol{s}_t$ and D_t yield $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \frac{1}{2}\zeta_t^2 - \frac{1}{2}\zeta_{t-1}^2 \le V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{1}{2}H^2$. □

E.7 Proof of Theorem 5.6

Theorem E.7 (General ICE Bounds)**.** *Having $M = \sqrt{T}$ and with the coefficients of Proposition 3.1, with probability at least $1-\delta$, we have: $\mathfrak{R}_T^{\mathrm{adv}} \le \widetilde{\mathcal{O}}(T^{3/4})$ and $\mathfrak{V}_T^{\mathrm{adv}} = \widetilde{\mathcal{O}}(T^{3/4})$.*

Proof. Let $t^\star \in \arg\max_{t\in[T]} \zeta_t$ and $\zeta_{\max} \coloneqq \zeta_{t^\star}$. The feasibility of $\pi_{\mathrm{adv}}^\star$ and $\zeta_{t-1} \ge 0$ imply

$$V^{\pi_{\mathrm{adv}}^\star,P}(\boldsymbol{s}_t) \le MV^{\pi_{\mathrm{adv}}^\star,P}(\boldsymbol{\ell}_t). \tag{61}$$

We utilize Equation 61 in the same subtraction step as above. Then, we sum $MV^{\pi_t,P}(\boldsymbol{\ell}_t) + \frac{1}{2}\zeta_t^2 - \frac{1}{2}\zeta_{t-1}^2 \le V^{\pi_t,P}(\boldsymbol{s}_t) + D_t + \frac{1}{2}H^2$ through an arbitrary prefix u, and use $\zeta_0 = 0$. Thus, we have

$$M\mathfrak{R}_u^{\mathrm{adv}}(\pi_{\mathrm{adv}}^\star) + \frac{1}{2}\zeta_u^2 \le \mathcal{J}_u^{\mathrm{I}} + \sum_{t=1}^{u} D_t + \frac{1}{2}H^2 u.$$

On the two uniform events and for every $u \le T$, we see that

$$M\mathfrak{R}_u^{\mathrm{adv}}(\pi_{\mathrm{adv}}^\star) + \frac{1}{2}\zeta_u^2 \le B\sqrt{T}(M + \zeta_{\max}) + \frac{1}{2}H^2 u, \tag{62}$$

because $W_{\max} \le M + \zeta_{\max}$. Since every episodic loss value belongs to $[0,H]$, for every prefix,

$$\mathfrak{R}_u^{\mathrm{adv}}(\pi_{\mathrm{adv}}^\star) \ge -Hu \ge -HT. \tag{63}$$

We apply Equation 62 at the random prefix $u = t^\star$ and use Equation 63 to get

$$\frac{1}{2}\zeta_{\max}^2 - B\sqrt{T}\zeta_{\max} \le MHT + BM\sqrt{T} + \frac{1}{2}H^2 T.$$

Multiplying by two and completing the square

$$(\zeta_{\max} - B\sqrt{T})^2 \le B^2T + 2MHT + 2BM\sqrt{T} + H^2T.$$

Taking the square root gives the explicit one-sided bound as

$$\zeta_{\max} \le B\sqrt{T} + \sqrt{B^2T + 2MHT + 2BM\sqrt{T} + H^2T}.$$

On putting $M = \sqrt{T}$, we obtain

$$\zeta_{\max} \le B\sqrt{T} + \sqrt{B^2T + 2HT^{3/2} + 2BT + H^2T} = \widetilde{\mathcal{O}}(T^{3/4}).$$

We evaluate Equation 62 at $u = T$, drop the term $\zeta_T^2/2 \ge 0$, and divide by $M = \sqrt{T}$ to get

$$\mathfrak{R}_T^{\mathrm{adv}} \le B(M + \zeta_{\max}) + \frac{1}{2}H^2\sqrt{T} = \widetilde{\mathcal{O}}(T^{3/4}).$$

On the weak-bridge event, we have

$$\begin{aligned}\mathfrak{V}_T^{\mathrm{adv}} &\le \zeta_T + 2H\sqrt{2T\log(3/\delta)} \\ &\le \zeta_{\max} + 2H\sqrt{2T\log(3/\delta)} = \widetilde{\mathcal{O}}(T^{3/4}).\end{aligned}$$

The three events each fail with probability at most $\delta/3$. Consequently, the two bounds hold simultaneously with probability at least $1-\delta$. On combining $K_Q \le 1 + \lceil \log_2 (M+HT/M) \rceil$ and $M = \sqrt{T}$, we obtain

$$K_Q \le 1 + \left\lceil \log_2(1 + H\sqrt{T}) \right\rceil = \mathcal{O}(\log T).$$

Again, we do not use this random count to choose the confidence levels. □

E.8 Proof of Proposition 5.7

Proposition E.8. *We fix $Q > 0$ and $T \ge 1$. Let $g_1, \dots, g_T \in [-Q, Q]$, set $z_0 = 0$, and recursively define $z_t = [z_{t-1} + g_t]^+$. If $z_t \le Q$ for every $t \in [T]$, then $\sum_{t=1}^T [g_t]^+ \le Q\lceil T/2 \rceil$. The bound is attained by $g_t = Q$ for an odd t and $g_t = -Q$ for an even t. In particular, for every even T, this attaining sequence satisfies $z_T = 0$, $[\sum_{t=1}^T g_t]^+ = 0$, and $\sum_{t=1}^T [g_t]^+ = QT/2$.*

Proof. We call an interval $[u, v] \subseteq [T]$ a *positive block* if $g_t > 0$ for every $t \in [u, v]$, and if the interval is maximal with this property. Let the positive blocks be denoted as $[u_1, v_1], \dots, [u_B, v_B]$ in increasing order.

Within a positive block, the projection in $z_t = [z_{t-1} + g_t]^+$ is inactive because $z_{t-1} \ge 0$ and $g_t > 0$. Hence, for every block b, we have

$$z_{v_b} = z_{u_b - 1} + \sum_{t=u_b}^{v_b} g_t.$$

The assumed queue bound $z_{v_b} \le Q$ and nonnegativity $z_{u_b-1} \ge 0$ give

$$\sum_{t=u_b}^{v_b} g_t = z_{v_b} - z_{u_b-1} \le Q. \tag{64}$$

Every positive block contains at least one index. Moreover, two consecutive positive blocks are separated by at least one index t with $g_t \le 0$. Otherwise, they would form a single maximal block. Therefore, the B blocks and the $B-1$ required separating indices occupy at least $2B-1$ time indices. Since there are only T indices, we get

$$2B - 1 \le T,\ B \le \frac{T+1}{2},\ B \le \left\lceil \frac{T}{2} \right\rceil.$$

Only positive blocks contribute to $\sum_t [g_t]^+$. Then, summing Equation 64 over the B blocks proves

$$\sum_{t=1}^T [g_t]^+ = \sum_{b=1}^B \sum_{t=u_b}^{v_b} g_t \le BQ \le Q \left\lceil \frac{T}{2} \right\rceil.$$

For the proposed attaining sequence, the recursion $z_t = [z_{t-1} + g_t]^+$ gives $z_t = Q$ at every odd t and $z_t = 0$ at every even t. There are $\lceil T/2 \rceil$ positive increments, each equal to Q, proving equality in $\sum_{t=1}^T [g_t]^+ \le Q\lceil T/2 \rceil$. If T is even, the increments cancel pairwise, the terminal queue is zero, and the three identities, i.e., $z_T = 0$, $[\sum_{t=1}^T g_t]^+ = 0$, and $\sum_{t=1}^T [g_t]^+ = QT/2$, follow directly. □

Consequently, even a pathwise uniform bound $\max_{t \le T} z_t \le Q = \mathcal{O}(1)$ on the entire signed-queue trajectory does not imply sublinear strong violation.

E.9 Proof of Proposition E.9

The path-wise prefix nonnegative-regret condition used by Conditional ICE in Equation 10 is clearest when related to the loss geometry. Let $\mathcal{Q}(P)$ denote the occupancy-measure polytope induced by the fixed transition function P.

Proposition E.9. *We fix an occupancy $q^\star \in \mathcal{Q}(P)$ and an oblivious loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$. The following statements are equivalent: 1) for every sequence $q_1, \dots, q_T \in \mathcal{Q}(P)$ and every $u \in [T]$, we have $\sum_{t=1}^{u} \langle q_t - q^\star, \boldsymbol{\ell}_t \rangle \geq 0$; 2) $q^\star$ is a common per-episode loss minimizer, i.e., $q^\star \in \bigcap_{t=1}^{T} \arg\min_{q \in \mathcal{Q}(P)} \langle q, \boldsymbol{\ell}_t \rangle$.*

Proof. First, we assume that $q^\star \in \bigcap_{t=1}^{T} \arg\min_{q \in \mathcal{Q}(P)} \langle q, \boldsymbol{\ell}_t \rangle$ holds. Then, we fix any sequence $q_1, \dots, q_T \in \mathcal{Q}(P)$ and any prefix $u \in [T]$. For every $t \leq u$, the optimality of $q^\star$ for the episode-t linear objective implies $\langle q_t, \boldsymbol{\ell}_t \rangle \geq \langle q^\star, \boldsymbol{\ell}_t \rangle$. Subtracting the right side and summing these u nonnegative inequalities proves $\sum_{t=1}^{u} \langle q_t - q^\star, \boldsymbol{\ell}_t \rangle \geq 0$.

Conversely, we assume that $\sum_{t=1}^{u} \langle q_t - q^\star, \boldsymbol{\ell}_t \rangle \geq 0$ holds. Then, we fix an episode $s \in [T]$ and an arbitrary occupancy measure $q \in \mathcal{Q}(P)$. We choose the test sequence

$$q_t = q^\star \ (t < s), \ q_s = q,$$

and choose the later occupancies arbitrarily. We apply $\sum_{t=1}^{u} \langle q_t - q^\star, \boldsymbol{\ell}_t \rangle \geq 0$ at the prefix $u = s$. Since every summand with $t < s$ is zero, the resulting inequality is $\langle q - q^\star, \boldsymbol{\ell}_s \rangle \geq 0$. As $q \in \mathcal{Q}(P)$ was chosen arbitrarily, $q^\star$ minimizes the episode-s linear objective over $\mathcal{Q}(P)$. Since $s \in [T]$ was also arbitrary, $q^\star \in \bigcap_{t=1}^{T} \arg\min_{q \in \mathcal{Q}(P)} \langle q, \boldsymbol{\ell}_t \rangle$ follows.

For the sufficient class used by Conditional ICE, we also suppose that $q^\star$ satisfies every episode's constraint, so it is an admissible fixed comparator. Let $\bar{q}$ be any feasible comparator minimizing the total loss. Since $q^\star$ minimizes each episode's unconstrained loss, we have for any $t \in [T]$ that $\langle \bar{q} - q^\star, \boldsymbol{\ell}_t \rangle \geq 0$. The feasibility and total-loss optimality of $\bar{q}$, together with feasibility of $q^\star$, imply that the sum of these nonnegative gaps is zero. Hence, every gap is zero and $\bar{q}$ is itself a common per-episode minimizer. This applies to the tie-broken comparator. The already proved implication from $q^\star \in \bigcap_{t=1}^{T} \arg\min_{q \in \mathcal{Q}(P)} \langle q, \boldsymbol{\ell}_t \rangle$ to $\sum_{t=1}^{u} \langle q_t - q^\star, \boldsymbol{\ell}_t \rangle \geq 0$ then applies to the occupancy sequence generated on every possible history.

To conclude, if the oblivious losses have a common unconstrained minimizer $q^\star$ that also satisfies every episode's constraint, then every minimum-total-loss feasible comparator is a common per-episode minimizer. In particular, the tie-broken comparator used by ICE satisfies the prefix nonnegative-regret hypothesis of Theorem 5.4 for every policy sequence, and hence for every ICE-generated history. □

E.10 Proof of Theorem E.10

Theorem E.10. *Let $u \leq T$ be a fixed global prefix containing R nonempty outer copies of the wrapper with geometric ratio $\chi > 1$. Let ω_r and m_r be the active scale and the number of executed episodes in nonempty copy r, with $\sum_{r=1}^{R} m_r = u \leq T$. We set $W_{\max} := \max\{W^{[1]}, \max_{t \leq u} \widehat{W}_t\}$, where the wrapper always selects the smallest geometric level covering its predictable envelope. Suppose every copy's normalized local regret obeys $\mathrm{Reg}_r(m_r) \leq C_{\mathrm{opt}}(\sqrt{T} + m_r/\sqrt{T}) + C_{\exp}\sqrt{m_r} + C_{\mathrm{add}}$, with nonnegative certified coefficients. Undoing the affine normalization yields $\sum_{r=1}^{R} 2\omega_r \mathrm{Reg}_r(m_r) \leq B_{\mathrm{PO}}(\chi) W_{\max} \sqrt{T}$, where*

$$B_{\mathrm{PO}}(\chi) := 2\left[\left(C_{\mathrm{opt}} + \frac{C_{\mathrm{add}}}{\sqrt{T}}\right)\frac{\chi^2}{\chi - 1} + C_{\mathrm{opt}}\chi + C_{\exp}\frac{\chi^2}{\sqrt{\chi^2 - 1}}\right].$$

Proof. If no scale increase occurs on the prefix, the last active scale is $W^{[1]} \leq W_{\max}$. Otherwise, the immediately preceding geometric level (which need not have received an episode) is strictly below the predictable envelope that triggered the last scale. The smallest covering level is χ times that preceding level. Therefore, in either case, $\omega_R < \chi W_{\max}$.

If one trigger skips several scales without executing an episode, omitting those empty scales only decreases every sum below. Hence, we can bound the nonempty scales by the full reverse geometric sequence ending at ω_R. First,

$$\sum_{r=1}^{R} \omega_r \le \omega_R \sum_{j=0}^{R-1} \chi^{-j} < \omega_R \sum_{j=0}^{\infty} \chi^{-j} = \omega_R \frac{\chi}{\chi-1} < \frac{\chi^2}{\chi-1} W_{\max},$$

where the final inequality uses $\omega_R < \chi W_{\max}$. Second, monotonicity of the guesses and $\sum_r m_r \le T$ give

$$\sum_{r=1}^{R} \omega_r m_r \le \omega_R \sum_{r=1}^{R} m_r < \chi W_{\max} T.$$

Third, the Cauchy-Schwarz inequality gives

$$\begin{aligned}\sum_{r=1}^{R} \omega_r \sqrt{m_r} &\le \left(\sum_{r=1}^{R} (\omega_r)^2\right)^{1/2} \left(\sum_{r=1}^{R} m_r\right)^{1/2} \\ &< \omega_R \left(\sum_{j=0}^{\infty} \chi^{-2j}\right)^{1/2} \cdot \sqrt{T} \\ &= \omega_R \frac{\chi}{\sqrt{\chi^2-1}} \sqrt{T} < \frac{\chi^2}{\sqrt{\chi^2-1}} W_{\max} \sqrt{T}.\end{aligned}$$

Now, we multiply $\mathrm{Reg}_r(m_r) \le C_{\mathrm{opt}} \left(\sqrt{T} + {}^{m_r}/_{\sqrt{T}}\right) + C_{\mathrm{exp}} \sqrt{m_r} + C_{\mathrm{add}}$ by $2\omega_r$ and sum over r. The four resulting terms are bounded, respectively, by

$$\begin{aligned} 2C_{\mathrm{opt}} \sqrt{T} \sum_r \omega_r &\le 2C_{\mathrm{opt}} \frac{\chi^2}{\chi-1} W_{\max} \sqrt{T}, \\ \frac{2C_{\mathrm{opt}}}{\sqrt{T}} \sum_r \omega_r m_r &\le 2C_{\mathrm{opt}} \chi W_{\max} \sqrt{T}, \\ 2C_{\mathrm{exp}} \sum_r \omega_r \sqrt{m_r} &\le 2C_{\mathrm{exp}} \frac{\chi^2}{\sqrt{\chi^2-1}} W_{\max} \sqrt{T}, \\ 2C_{\mathrm{add}} \sum_r \omega_r &\le 2 \frac{C_{\mathrm{add}}}{\sqrt{T}} \frac{\chi^2}{\chi-1} W_{\max} \sqrt{T}. \end{aligned}$$

Adding these four inequalities proves $\sum_{r=1}^{R} 2\omega_r \,\mathrm{Reg}_r(m_r) \le B_{\mathrm{PO}}(\chi) W_{\max} \sqrt{T}$ and $B_{\mathrm{PO}}(\chi) \coloneqq 2\left[(C_{\mathrm{opt}} + {}^{C_{\mathrm{add}}}/_{\sqrt{T}}) {}^{\chi^2}/_{\chi-1} + C_{\mathrm{opt}} \chi + C_{\mathrm{exp}} {}^{\chi^2}/_{\sqrt{\chi^2-1}}\right]$. Putting $\chi = 2$ gives $\chi^2/(\chi-1) = 4$ and $\chi^2/\sqrt{\chi^2-1} = 4/\sqrt{3}$, which recovers B_{PO}. □

E.11 General ICE under PNR

Corollary E.11 (General ICE under PNR)**.** *We run General ICE with $M = \sqrt{T}$ and $\varphi(z) = {}^{z^2}/_2$, exactly as in Theorem 5.6. If, in addition, Equation 10 holds almost surely for every prefix under the law of General ICE, then, with probability at least $1-\delta$, we have*

$$\zeta_{\max} \le (2B + H + 1)\sqrt{T},\ \mathfrak{R}_T^{\mathrm{adv}} \le \left(2B^2 + (H+2)B + \tfrac{1}{2}H^2\right)\sqrt{T},\ \textit{and}$$
$$\mathfrak{V}_T^{\mathrm{adv}} \le (2B + H + 1)\sqrt{T} + 2H\sqrt{2T \log({}^3/_\delta)}.$$

Proof. We work on the intersection of the events of Equation 9, Lemma 5.1, and Lemma 5.2, which has probability at least $1-\delta$. The proof of Theorem 5.6 shows that Equation 62 holds on this event

simultaneously for every $u \le T$, so it also holds at the random prefix $u = t^\star$. There, PNR gives $\mathfrak{R}^{\mathrm{adv}}_{t^\star}(\pi^\star_{\mathrm{adv}}) \ge 0$, and $t^\star \le T$ yields

$$\frac{1}{2}\zeta^2_{\max} \le B\sqrt{T}(M + \zeta_{\max}) + \frac{1}{2}H^2 T.$$

With $M = \sqrt{T}$, this becomes $\zeta^2_{\max} - 2B\sqrt{T}\zeta_{\max} - (2B + H^2)T \le 0$. Since $B^2 + 2B + H^2 \le (B + H + 1)^2$, we get

$$\zeta_{\max} \le B\sqrt{T} + \sqrt{(B^2 + 2B + H^2)T} \le (2B + H + 1)\sqrt{T}.$$

We evaluate Equation 62 at $u = T$, drop $\varsigma^2_T/2 \ge 0$, and divide by $M = \sqrt{T}$ to obtain

$$\mathfrak{R}^{\mathrm{adv}}_T \le B(M + \zeta_{\max}) + \frac{1}{2}H^2\sqrt{T} \le \left(B(2B + H + 2) + \frac{1}{2}H^2\right)\sqrt{T}.$$

Finally, Lemma 5.2 and $\zeta_T \le \zeta_{\max}$ give the violation bound. We recall that since $B = B_{\mathrm{PO}} + B_{\mathrm{MD}}$ grows only poly-logarithmically with T, all three bounds are $\widetilde{\mathcal{O}}(\sqrt{T})$. By contrast, Theorem 5.4 gives $\mathfrak{R}^{\mathrm{adv}}_T \le 2(B + 1)\sqrt{T} + 1/\sqrt{T}$, which is linear in B, and a violation bound with the extra factor $\log(2(B + 1)T + 2)$. □

Thus, the negative prefix regret is the only term that separates the $T^{3/4}$ rates of Theorem 5.6 from the $T^{1/2}$ rates, and General ICE adapts to PNR without being told that it holds.

F Experimental Details and Additional Results

This appendix completes Section 6. Appendices F.1–F.3 describe the instances, the methods, and the protocol. Appendices F.4–F.10 report every suite in full, and Appendix F.11 states the scope of the conducted synthetic experiments.

F.1 Instances

Layered Dynamics: As stated earlier, every CMDP instance has $|A| = 3$ actions, the layers $X_0 = \{x_0\}$, $X_h = \{x_{h,1}, x_{h,2}\}$ for $h \in [H - 1]$, and $X_H = \{x_H\}$. Thus, $|X| = 2H$, and $H = 2$ gives the four states x_0, $x_{1,1}$, $x_{1,2}$, and x_2. We write $i(x) \in \{1, 2\}$ for the index of x within its layer, with $i(x_0) = 1$. For the physical actions a_0 (safe), a_1, and a_2, we set $P(x_{h+1,1} \mid x, a) = p_a - 0.08(i(x) - 1)$ with $(p_{a_0}, p_{a_1}, p_{a_2}) = (0.72, 0.27, 0.49)$, and the remaining mass goes to $x_{h+1,2}$. Every state of the last non-terminal layer moves to x_H. Except for the known-transition controls, the learners do not know P, and only the singleton terminal layer is structurally known (Section 4.1). For every seed, a seeded permutation (which all methods share) relabels the actions. Hence, the safe action has no fixed index.

Stochastic Boundary: At every state x, the mean losses of (a_0, a_1, a_2) are $(0.66, 0.06, 0.34) + 0.12(i(x) - 1)$, and the mean costs are $(0, 0.45, 0.16)$. The realized losses are Bernoulli with these means. The realized costs of a_1 and a_2 lie in $\{-1, +1\}$ with these means, and the cost of a_0 is identically zero. Since both risky actions have positive mean costs and every state is reachable under every policy, every policy that plays a risky action with positive probability has $V^{\pi,P}(\boldsymbol{c}) > 0$. Thus, the always-safe policy is the unique feasible policy, $V^{\pi^\star_{\mathrm{stoc}},P}(\boldsymbol{c}) = 0$, and no policy is strictly feasible, i.e., Slater's condition fails with $\rho = 0$. Also, $V^{\pi,P}(\boldsymbol{c}) \ge 0$ for every π, so the strong and the weak violations coincide on this task.

Oblivious Switching: We split the T episodes into eight phases of length $T/8$. In the even phases, the losses of (a_0, a_1, a_2) are $(0.55, 0.05, 0.45) + 0.12(i(x) - 1)$ and the costs are $(0, 0.6, -0.3)$. In the odd phases, the losses are $(0.55, 0.80, 0.15) + 0.12(i(x) - 1)$ and the costs are $(0, -0.4, 0.7)$. The whole sequence is fixed before the interaction. Let $\bar{q}(a) = \sum_x q(x, a)$ be the total occupancy of action a under a fixed policy. Since the costs do not depend on the state, the policy satisfies every episode's constraint only if $0.6\bar{q}(a_1) - 0.3\bar{q}(a_2) \le 0$ and $-0.4\bar{q}(a_1) + 0.7\bar{q}(a_2) \le 0$. The first inequality gives $\bar{q}(a_2) \ge 2\bar{q}(a_1)$, and the second gives $\bar{q}(a_1) \ge 7/4\bar{q}(a_2) \ge 7/2\bar{q}(a_1)$. Hence, $\bar{q}(a_1) = \bar{q}(a_2) = 0$, and $\pi^\star_{\mathrm{adv}}$ is the always-safe policy. In contrast, the per-phase loss minimizers (a_1 in the even and a_2 in the odd phases) have positive costs in their own phases.

PNR Task: This task keeps the costs of the switching task, but a_0 has zero loss at every state and in every episode, whereas (a_1, a_2) have the losses $(0.55, 0.75)$ in the even phases and $(0.80, 0.45)$ in the odd phases, plus $0.12(i(x) - 1)$. Thus, the always-safe policy minimizes every episode's loss and satisfies every constraint. Hence, the PNR (in Equation 10) condition holds for every occupancy sequence (Proposition E.9).

Uniform Play: Since the costs do not depend on the state, the uniform policy has the expected cost $0.61/3 \approx 0.2033$ per step on the boundary and 0.1 per step on both adversarial tasks. We compute its regret and violations exactly from its occupancy measure and utilize them as seed-independent references.

F.2 Methods and Parameters

FIRE and ICE: We implement Algorithms 2 and 3 around exact tabular `PO-DB` (Algorithm 1), with the LCB of Equation 7, the persistent queue and FIRE statistics, the predictable envelopes, the factor-two scale levels, and the outer restarts. Every copy of `PO-DB` uses the published η_{PO} and $\gamma = 2\eta_{\mathrm{PO}}H$ ($\eta_{\mathrm{PO}} \approx 1.44 \times 10^{-3}$ at $T = 20{,}000$), the confidence sets of Equation 11, and $\varepsilon_{\mathrm{PO}} = \delta/27T^2$ with $\delta = 0.05$. FIRE uses $M = 1$ and $W^{[1]} = M + \lambda e^{\lambda H}$, and General ICE uses $M = \sqrt{T}$, $\varphi(z) = z^2/2$, and $W^{[1]} = M$. At $T = 20{,}000$, $H = 2$, $|X| = 4$, and $|A| = 3$, the certified coefficients of Appendix C.2.1 give $B_F \approx 1.54 \times 10^6$, the FIRE parameter $\lambda \approx 8.42 \times 10^{-10}$, and the Conditional ICE parameter $\lambda \approx 1.14 \times 10^{-9}$. The certified regret bound $B_F\sqrt{T} + 1 \approx 2.2 \times 10^8$ of Theorem 4.3 then exceeds the trivial bound $HT = 4 \times 10^4$. Thus, our main runs use the practical $\lambda = \min\{1/H, c/H\sqrt{T}\}$ with $c = 4$, which carries no theorem, and the calibration controls use certified values. Conditional ICE uses either the same practical λ or a certified value. The saved calibration and PNR suites were run with the looser constant $12C_{\mathrm{opt}} + 8/\sqrt{3}C_{\mathrm{exp}} + 8C_{\mathrm{add}}$ in place of B_{PO} (Appendix C.4), which gives the smaller $\lambda \approx 1.95 \times 10^{-11}$ (FIRE) and $\lambda \approx 2.66 \times 10^{-11}$ (Conditional ICE). These values remain certified: Theorem 4.3 holds for every smaller λ (Appendix D.9), and the proof of Theorem 5.4 uses α only through $B/\alpha \le 1/4$ and $\alpha \ge 2eH^2$, which hold for every larger α.

PDB-PS: We implement the single-constraint PDB-PS with FS-PODB of Stradi et al. (2025d) (their Algorithms 2 and 3). We use the rewards $r = 1 - \ell$, which preserves the regret since every occupancy has mass H. The primal step is $1/(2HC\sqrt{T})$ with fixed share $1/T$. The dual variable is projected onto $[0, T^{1/4}]$ with the step $1/(D\log(|A||X|^2T^2/\delta_B)\sqrt{T})$ and $\delta_B = \delta/14$. The transition confidence sets are the L_1 balls of Rosenberg & Mansour (2019b), which our implementation represents exactly for successor layers with at most two states, i.e., for all of our instances. The source's text updates the range with the next multiplier, whereas its Algorithm 2 prints the current one. We use the former, since it covers the next episode's surrogate. The prescribed constants are $C = 252|X||A|H = 6{,}048$ and $D = 84{,}672H^2|X|^2|A| = 16{,}257{,}024$ on our main tasks. The practical choice $C = D = 1$ was fixed before any run and carries no theorem.

Controls and Ablations: The matched no-penalty control runs `PO-DB` with the method's initial normalization, i.e., on the losses $(M\ell_{t,h} + W^{[1]})/(2W^{[1]})$, and it never uses the costs, grows its scale, or restarts. Thus, it produces exactly FIRE's policies until FIRE's first restart. Each ablation changes one component: *Plug-in* drives FIRE's queue with $\widehat{V}_t$ instead of g_t; *No gate* sets $I_t \equiv 1$; *Quadratic* uses $\varphi(z) = z^2/2$, $w_t = \zeta_t I_t$, $\widehat{W}_t = M + \zeta_{t-1} + H$, and $M = \sqrt{T}$, as in General ICE; *Warm* keeps the optimizer's weights, bonuses, and counts at every scale increase; *Known P* removes the transition part of FIRE's width and replaces the `PO-DB` confidence sets with the true rows (Appendix B). None of these variants carries a theorem.

F.3 Protocol, Metrics, and Statistics

Protocol: We fixed all suites and parameters in two written protocols before the corresponding runs, one for the main comparison and one for the five extension suites, and we report every configuration and every seed. After an initial run, we tried to make the arithmetic platform-independent, re-executed every suite without changing any configuration, and reported the re-execution. We use the seeds 0 to 9. For each seed, the environment draws, the action relabeling, and the uniform variables that drive the action. Transition samples come from separate seeded streams, and all methods share them (common random numbers). The learners observe only the visited states, actions, successors,

Table 2: Endpoints at $T = 20{,}000$ for Figure 4a–Figure 4b and the calibration controls (mean $\pm$ 95% half-width over ten seeds). The two certified values of λ are described in Appendix F.2.

	Stochastic boundary		Oblivious switching	
Method	**Regret**	**Strong violation**	**Regret**	**Weak violation**
FIRE / General ICE (ours)	$-13{,}167 \pm 74$	$9{,}286 \pm 66$	$-3{,}885 \pm 106$	$3{,}240 \pm 127$
PDB-PS (practical $C = D = 1$)	$-13{,}273 \pm 104$	$9{,}223 \pm 96$	$-5{,}259 \pm 75$	$4{,}749 \pm 91$
`PO-DB`, no penalty	$-16{,}605 \pm 133$	$11{,}928 \pm 126$	$-6{,}229 \pm 163$	$5{,}933 \pm 166$
FIRE (looser certified λ)	$-16{,}672 \pm 130$	$11{,}983 \pm 128$	–	–
FIRE (sharper certified λ)	$-14{,}521 \pm 107$	$10{,}338 \pm 94$	–	–
PDB-PS (prescribed C, D)	$-11{,}725 \pm 0.043$	$8{,}135 \pm 0.032$	$-4{,}456 \pm 0.011$	$4{,}000 \pm 0.014$
Uniform play (exact)	$-11{,}723$	$8{,}133$	$-4{,}456$	$4{,}000$

Table 3: Both violation measures (i.e., strong and weak), restarts, and the fraction of episodes with an open gate at $T = 20{,}000$.

Task	**Method**	**Regret**	**Strong Violation**	**Weak Violation**	**Restarts**	**Gate open**
Boundary	FIRE (ours)	$-13{,}167 \pm 74$	$9{,}286 \pm 66$	$9{,}286 \pm 66$	4.6 ± 0.7	0.512
	PDB-PS	$-13{,}273 \pm 104$	$9{,}223 \pm 96$	$9{,}223 \pm 96$	0.0 ± 0.0	–
	`PO-DB`, no penalty	$-16{,}605 \pm 133$	$11{,}928 \pm 126$	$11{,}928 \pm 126$	0.0 ± 0.0	–
Switching	General ICE (ours)	$-3{,}885 \pm 106$	$3{,}303 \pm 101$	$3{,}240 \pm 127$	5.0 ± 0.0	–
	PDB-PS	$-5{,}259 \pm 75$	$4{,}750 \pm 91$	$4{,}749 \pm 91$	0.0 ± 0.0	–
	`PO-DB`, no penalty	$-6{,}229 \pm 163$	$6{,}192 \pm 232$	$5{,}933 \pm 166$	0.0 ± 0.0	–

losses, and costs, and they never draw random numbers themselves. We did not perform any kind of hyperparameter search.

Metrics: In every episode, we evaluate $V^{\pi_t,P}(\boldsymbol{\ell}_t)$ and $V^{\pi_t,P}(\boldsymbol{c}_t)$ exactly from the occupancy measure, with the population means on the stochastic task. The comparator comes from a lexicographic occupancy LP. We report the signed regret of Equation 1 and Equation 2, and we save both violation measures for every run.

Statistics: We report means with 95% Student-t half-widths over the ten seeds ($t_{0.975,9} \approx 2.262$). For two methods on the same task, we also report the mean and the half-width of the per-seed paired differences, which the common random numbers make informative. The curves use pointwise intervals, and we do not adjust any intervals for multiple comparisons.

F.4 Main Comparison and Certified Constants

Tables 2 and 3 report the endpoints behind Figure 4a–Figure 4b and the calibration controls.

Gate of FIRE and Restarts: FIRE's LCB first becomes positive between episodes $4{,}596$ and $6{,}116$ (median $5{,}538$). The first restart follows one episode later, since $W^{[1]}$ equals the envelope at $\zeta = 0$, and FIRE's policies coincide with those of the no-penalty control until then. After the gate first opens, it stays open in 66% to 75% of the episodes, and FIRE ends with 4.6 ± 0.7 restarts. At $T = 20{,}000$, FIRE plays the safe action at x_0 with probability 0.345 ± 0.015, against 0.087 ± 0.028 for the no-penalty control. General ICE penalizes from the first episodes (its first restart occurs by episode 4) and ends with five restarts.

Comparison with PDB-PS: On the stochastic boundary task, FIRE and practical PDB-PS have the paired differences 106 ± 93 in regret and 63 ± 75 in strong violation. On the oblivious switching task, General ICE has $1{,}374 \pm 67$ higher regret and $1{,}510 \pm 71$ lower weak violation. Its strong violation ($3{,}303 \pm 101$) is close to its weak violation, so ICE's reflected queue does not hide positive costs behind negative ones on this instance. We do not claim that this holds in general (Proposition 5.7).

Certified Constants: With the looser $\lambda \approx 1.95 \times 10^{-11}$ of the calibration suite (in Appendix F.2), FIRE's multiplier stays below 10^{-10}, and certified FIRE has 55 ± 17 more strong violations than the no-penalty control. This differs from it essentially through the initial normalization. With the sharper $\lambda \approx 8.42 \times 10^{-10}$, the multiplier stays below 10^{-9}. However, once the queue exceeds about 10^2, the envelope $M + \lambda e^{\lambda(\zeta_{t-1}+H)}$ exceeds $W^{[1]}$ in floating point, and the wrapper doubles the scale once in every seed. The coarser normalization slows `PO-DB`, so the strong violation is $1{,}591 \pm 126$

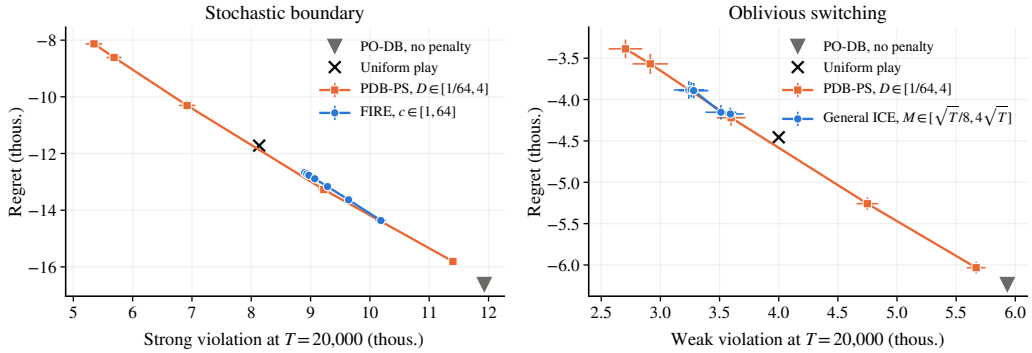


Figure 5: Endpoint regret against violation at $T = 20{,}000$ under the penalty sweeps (means and 95% intervals over ten seeds). Stronger penalties move the points up and to the left.

below the control's, at a $2{,}084 \pm 153$ higher regret. In exact calculations, the first positive queue value triggers this doubling for every $\lambda > 0$. The looser λ misses it because the envelope's increase stays below double precision. In both cases, the penalty itself is inert. With its prescribed constants, PDB-PS has the primal and dual steps 2.9×10^{-7} and 1.5×10^{-11}, and its endpoints stay within 0.03% of uniform play. Thus, the certified constants of FIRE and PDB-PS are far too conservative for moderate budgets.

F.5 Penalty-Strength Sweeps

Table 4: Every point of the penalty sweeps in Figure 5 (strong violation on the boundary, weak violation on the switching task).

Task	Setting	Regret	Violation	Restarts
Stochastic boundary	FIRE, $c = 1$	$-14{,}366 \pm 84$	$10{,}186 \pm 91$	1.0 ± 0.0
	FIRE, $c = 2$	$-13{,}634 \pm 79$	$9{,}640 \pm 74$	2.8 ± 0.3
	FIRE, $c = 4$	$-13{,}167 \pm 74$	$9{,}286 \pm 66$	4.6 ± 0.7
	FIRE, $c = 8$	$-12{,}892 \pm 91$	$9{,}071 \pm 75$	7.5 ± 2.3
	FIRE, $c = 16$	$-12{,}769 \pm 84$	$8{,}970 \pm 73$	16.3 ± 4.4
	FIRE, $c = 32$	$-12{,}724 \pm 71$	$8{,}937 \pm 67$	40.9 ± 4.5
	FIRE, $c = 64$	$-12{,}675 \pm 66$	$8{,}900 \pm 63$	82.7 ± 9.1
	PDB-PS, $D = 1/64$	$-8{,}130 \pm 155$	$5{,}348 \pm 141$	0.0 ± 0.0
	PDB-PS, $D = 1/16$	$-8{,}612 \pm 147$	$5{,}688 \pm 128$	0.0 ± 0.0
	PDB-PS, $D = 1/4$	$-10{,}305 \pm 172$	$6{,}918 \pm 141$	0.0 ± 0.0
	PDB-PS, $D = 1$	$-13{,}273 \pm 104$	$9{,}223 \pm 96$	0.0 ± 0.0
	PDB-PS, $D = 4$	$-15{,}809 \pm 67$	$11{,}399 \pm 72$	0.0 ± 0.0
Oblivious switching	General ICE, $M = \sqrt{T}/8$	$-3{,}891 \pm 93$	$3{,}281 \pm 126$	8.0 ± 0.0
	General ICE, $M = \sqrt{T}/4$	$-3{,}892 \pm 89$	$3{,}280 \pm 113$	7.0 ± 0.0
	General ICE, $M = \sqrt{T}/2$	$-3{,}882 \pm 92$	$3{,}260 \pm 109$	6.0 ± 0.0
	General ICE, $M = \sqrt{T}$	$-3{,}885 \pm 106$	$3{,}240 \pm 127$	5.0 ± 0.0
	General ICE, $M = 2\sqrt{T}$	$-4{,}154 \pm 88$	$3{,}511 \pm 131$	4.0 ± 0.0
	General ICE, $M = 4\sqrt{T}$	$-4{,}174 \pm 76$	$3{,}591 \pm 112$	3.0 ± 0.0
	PDB-PS, $D = 1/64$	$-3{,}387 \pm 113$	$2{,}705 \pm 142$	0.0 ± 0.0
	PDB-PS, $D = 1/16$	$-3{,}569 \pm 122$	$2{,}914 \pm 149$	0.0 ± 0.0
	PDB-PS, $D = 1/4$	$-4{,}220 \pm 97$	$3{,}597 \pm 121$	0.0 ± 0.0
	PDB-PS, $D = 1$	$-5{,}259 \pm 75$	$4{,}749 \pm 91$	0.0 ± 0.0
	PDB-PS, $D = 4$	$-6{,}032 \pm 72$	$5{,}669 \pm 80$	0.0 ± 0.0

Figure 5 and Table 4 show every point of the penalty sweeps. On the boundary task, raising FIRE's c from 1 to 64 lowers its strong violation from 10,186 to 8,900 and raises its restarts from 1.0 to 82.7. Also, lowering PDB-PS's D from 4 to $1/64$ lowers its strong violation from 11,399 to 5,348. On the switching task, lowering ICE's M from $4\sqrt{T}$ to $\sqrt{T}$ lowers its weak violation from 3,591 to 3,240, and smaller values of M change little, whereas PDB-PS spans the range from 5,669 to 2,705.

One Frontier: In both tasks, the points of all methods, the uniform policy, and the no-penalty control lie close to one line. This is because moving probability from risky actions to the safe action raises

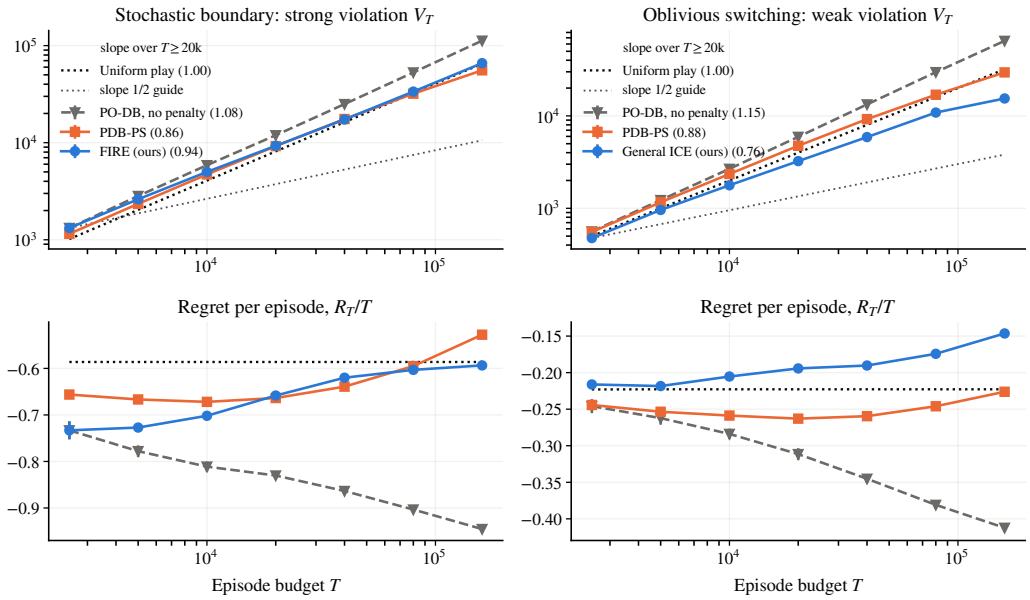


Figure 6: *(first row)* Violation on a log-log scale with the mean exponent over $T \geq 20{,}000$ in the legend. *(second row)* Regret per episode when all parameters are set from the budget T. Mean curves over ten seeds with 95% confidence intervals are plotted.

the loss and lowers the cost at nearly fixed rates. Hence, the methods differ in how far they move along this line within the budget.

Why FIRE and ICE Move Less? For two actions a and a' at the same state, let $\Delta\ell$ and Δc be their loss and cost gaps in episode t. Since $\mathcal{W}_t \geq M + w_t$, the normalized surrogate of Section 3 satisfies

$$|u_{t,h}(x,a) - u_{t,h}(x,a')| = \frac{|M\Delta\ell + w_t\Delta c|}{2\mathcal{W}_t} \leq \frac{(M+w_t)\max\{|\Delta\ell|,|\Delta c|\}}{2(M+w_t)}$$
$$= \frac{1}{2}\cdot\max\{|\Delta\ell|,|\Delta c|\},$$

for every multiplier w_t. Thus, a stronger multiplier can tilt these gaps toward the cost, but it cannot enlarge them beyond this cap, and it triggers more restarts. The same cap holds for PDB-PS, whose range $\Xi_t \geq 2(1+\lambda_t)$ also covers its surrogate $\ell + \lambda_t(1+c)$. The difference lies in the timing. On the boundary task, the PDB-PS's multiplier starts to grow after the first episode and exceeds one within 200 episodes for $D = 1/64$ (within 700 for $D = 1/16$). In contrast, FIRE's gate first opens after 5,482 to 6,294 episodes (medians over the seeds) for every c. This is because FIRE runs the unpenalized optimizer before its gate opens, and c affects that optimizer only through $W^{[1]}$.

F.6 Budget Scaling

We run the main comparison for $T \in \{2{,}500, 5{,}000, 10{,}000, 20{,}000, 40{,}000, 80{,}000, 160{,}000\}$ and set every parameter from T exactly as at $T = 20{,}000$, i.e., FIRE's $\lambda = \min\{1/H, 4/H\sqrt{T}\}$, ICE's $M = \sqrt{T}$, the rates of `PO-DB`, and the steps of PDB-PS (Figure 4c, Figure 4d, Figure 6, and Table 5). Hence, each budget is a separately configured run, not a prefix of a longer run. For each of the ten seeds, we fit the slope of $\log V_T$ against $\log T$ over $T \geq 20{,}000$, and we report the mean slope with its 95% interval. It should be noted that these local exponents describe the budgets we ran, and they do not establish asymptotic rates.

Switching Task: General ICE has the smallest weak violation at every budget, and its weak violation per episode falls from 0.19 to 0.096. Its weak violation grows with the exponent 0.76 ± 0.03, against 0.88 ± 0.01 for PDB-PS and 1.15 ± 0.01 for the no-penalty control, whose loss-greedy policy violates more and more. The strong violation of General ICE grows with exponent 0.97 ± 0.02 and reaches

Table 5: Endpoint violation for every budget (strong violation on the boundary task, weak violation on the switching task), the mean exponent over $T \geq 20{,}000$ with its 95% interval, and the mean number of restarts of our methods.

Task	Method	2.5k	5k	10k	20k	40k	80k	160k	Slope
Boundary	FIRE (ours)	1,322	2,614	5,000	9,286	17,337	33,596	65,978	0.944 ± 0.003
	PDB-PS	1,151	2,344	4,709	9,223	17,524	31,975	55,621	0.864 ± 0.008
	PO-DB, no penalty	1,322	2,824	5,865	11,928	24,976	52,907	112,395	1.079 ± 0.005
	Uniform play	1,017	2,033	4,067	8,133	16,267	32,533	65,067	1.00
	Restarts of FIRE	0.0	1.0	1.1	4.6	37.7	105.3	223.7	–
Switching	General ICE (ours)	476	959	1,775	3,240	5,897	10,867	15,435	0.764 ± 0.029
	PDB-PS	554	1,162	2,351	4,749	9,220	16,944	29,573	0.879 ± 0.013
	PO-DB, no penalty	561	1,226	2,673	5,933	13,314	29,721	64,857	1.151 ± 0.011
	Uniform play	500	1,000	2,000	4,000	8,000	16,000	32,000	1.00
	Restarts of General ICE	4.0	4.0	5.0	5.0	5.2	6.0	6.0	–

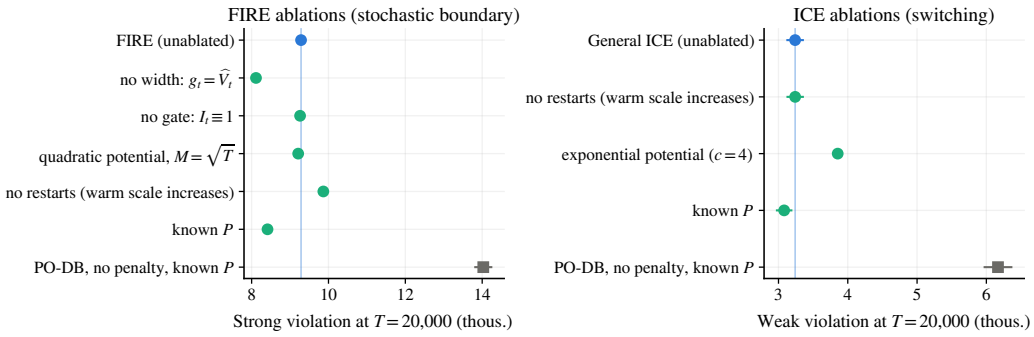


Figure 7: Endpoint violation of every ablation at $T = 20{,}000$ (mean curves over ten seeds with 95% confidence interval). The vertical blue line denotes the unablated method.

24,619 at $T = 160{,}000$, compared with a weak violation of 15,435. Thus, the reflected queue lets positive and negative episode costs cancel, as Proposition 5.7 permits. General ICE restarts between 4 and 6 times at every budget, in line with the logarithmic count K_Q of Lemma E.2.

Boundary Task: At $T = 2{,}500$, FIRE's gate never opens, and FIRE coincides with the no-penalty control. For larger budgets, the gate first opens later as T grows (median episode 4,060 at $T = 5{,}000$ and 8,519 at $T = 160{,}000$), since the logarithmic factors of the width grow with T. After that, the LCB of the restarted, nearly uniform policy stays positive. At $T = 160{,}000$, the gate is open in every episode after episode 80,000, and the queue grows by about 0.26 per episode. With $\lambda = 0.005$, the envelope doubles whenever the queue grows by $\log 2/\lambda \approx 139$, i.e., about every 530 episodes. Thus, FIRE restarts 37.7, 105.3, and 223.7 times at $T = 40{,}000$, 80,000, and 160,000, and each restart returns PO-DB to the uniform policy before it can learn the safe action. Consequently, FIRE's strong violation per episode decreases from 0.53 to 0.41 and then settles at the level of uniform play (0.407), and its local exponent rises back toward one (0.97 between 80,000 and 160,000). The no-penalty control converges to the risky actions (exponent 1.08 ± 0.01), whereas PDB-PS, which never restarts, lowers its violation per episode from 0.46 to 0.35 (exponent 0.86 ± 0.01).

Two Regimes of FIRE: This restart cycle arises because the practical λ exceeds the certified value by more than seven orders of magnitude. With a certified λ, the multiplier stays negligible, and the scale doubles at most once within our budgets, so FIRE's penalty has no effect beyond the normalization (Appendix F.4). Hence, at these budgets, neither parameterization gives FIRE a good finite-budget tradeoff on this task. Closing this gap could be an open problem that motivates tighter certificates and restarts that retain progress. The regret per episode (Figure 6, second row) mirrors the violation: it rises for the methods that lower their violation and falls for the no-penalty control.

F.7 Component Ablations

Figure 7 and Table 6 change one component at a time.

Table 6: Component ablations at $T = 20{,}000$. The paired differences are relative to the unablated method on the same seeds.

Task	Variant	Regret	Violation	Δ violation (paired)	Restarts
Boundary	FIRE (unablated)	$-13{,}167 \pm 74$	$9{,}286 \pm 66$	–	4.6
	no width: $g_t = \widehat{V}_t$	$-11{,}699 \pm 5$	$8{,}114 \pm 4$	$-1{,}173 \pm 64$	161.0
	no gate: $I_t \equiv 1$	$-13{,}139 \pm 76$	$9{,}261 \pm 68$	-25 ± 11	4.6
	quadratic potential, $M = \sqrt{T}$	$-13{,}081 \pm 111$	$9{,}212 \pm 95$	-75 ± 46	2.6
	no restarts (warm scale increases)	$-14{,}155 \pm 137$	$9{,}866 \pm 124$	580 ± 93	9.5
	known P	$-12{,}075 \pm 28$	$8{,}413 \pm 26$	-873 ± 58	49.0
	PO-DB, no penalty, known P	$-18{,}700 \pm 237$	$14{,}029 \pm 236$	$4{,}742 \pm 211$	0.0
Switching	General ICE (unablated)	$-3{,}885 \pm 106$	$3{,}240 \pm 127$	–	5.0
	no restarts (warm scale increases)	$-3{,}782 \pm 107$	$3{,}241 \pm 126$	1 ± 77	5.0
	exponential potential ($c = 4$)	$-4{,}376 \pm 22$	$3{,}855 \pm 41$	615 ± 113	65.6
	known P	$-3{,}757 \pm 98$	$3{,}082 \pm 119$	-158 ± 39	5.0
	PO-DB, no penalty, known P	$-6{,}448 \pm 190$	$6{,}168 \pm 208$	$2{,}928 \pm 169$	0.0

Confidence Width: With the plug-in $\widehat{V}_t$, FIRE's gate opens within the first 26 episodes and stays open in more than 99% of the episodes. Thus, the queue grows from the start, and FIRE restarts 161 times. Its policy then stays near uniform, and its endpoints are close to uniform play (regret $-11{,}699 \pm 5$ against $-11{,}723$, and strong violation $8{,}114 \pm 4$ against 8,133). The width delays the penalty until the model certifies a positive cost, and it is what makes g_t a lower confidence bound in Lemma 4.1.

Gate and Potential: Removing the gate changes FIRE's strong violation by -25 ± 11, and the quadratic potential (with $M = \sqrt{T}$) changes it by -75 ± 46, i.e., by less than 1% in both cases. On the switching task, where PNR fails, the exponential potential of Conditional ICE raises the weak violation by 615 ± 113 and restarts 65.6 times.

Outer Restarts: Keeping the optimizer's state at every scale increase raises FIRE's strong violation by 580 ± 93 and leaves ICE's weak violation unchanged (1 ± 77). A restart resets PO-DB to the uniform policy, which is safer than the loss-greedy policy that FIRE learns before its gate opens. Thus, the restarts that our analysis requires do not hurt constraint control on these tasks.

Transition Learning: Known transitions lower FIRE's strong violation by 873 ± 58 (9%) and ICE's weak violation by 158 ± 39 (around 5%). With known P, FIRE's width loses its transition part, so the gate is open in 84% of the episodes and FIRE restarts 49 times. The no-penalty control with known P exploits the risky actions faster and incurs more violations (14,029 and 6,168).

F.8 The PNR Regime

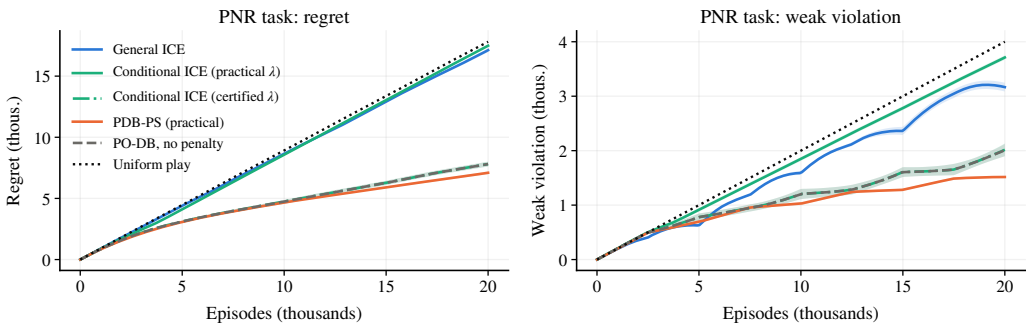


Figure 8: Regret and weak violation on the PNR task (mean and pointwise 95% intervals over ten seeds). The certified Conditional ICE (dash-dotted) coincides with the no-penalty PO-DB (dashed).

On the PNR task, the comparator has zero loss. Hence, every regret increment is nonnegative, and PNR holds on every history. Since the losses and the constraints agree, the unpenalized PO-DB already moves toward the safe action. Conditional ICE with a certified λ (2.66×10^{-11} or 1.14×10^{-9}, Appendix F.2) is indistinguishable from it (paired differences 1.4 ± 3.7 in regret and 6.4 ± 12.7 in

Table 7: Endpoints on the PNR task at $T = 20{,}000$.

Method	Regret	Weak Violation	Restarts	λ
General ICE	$17{,}122 \pm 148$	$3{,}167 \pm 81$	5.0	–
Conditional ICE (practical λ)	$17{,}486 \pm 65$	$3{,}711 \pm 39$	62.6	0.0141
Conditional ICE (looser certified λ)	$7{,}810 \pm 194$	$2{,}016 \pm 117$	0.0	2.66×10^{-11}
Conditional ICE (sharper certified λ)	$7{,}810 \pm 197$	$2{,}016 \pm 116$	0.0	1.14×10^{-9}
PDB-PS (practical)	$7{,}097 \pm 110$	$1{,}516 \pm 19$	0.0	–
PO-DB, no penalty	$7{,}808 \pm 197$	$2{,}010 \pm 114$	0.0	–
Uniform play (exact)	17,811	4,000	–	–

Table 8: Endpoint violation at $T = 20{,}000$ for every horizon (strong violation on the boundary task, weak violation on the switching task).

Task	Method	$H = 1$	$H = 2$	$H = 4$	$H = 8$
Boundary	FIRE (ours)	$4{,}161 \pm 13$	$9{,}286 \pm 66$	$19{,}915 \pm 137$	$33{,}730 \pm 104$
	PDB-PS	$6{,}538 \pm 57$	$9{,}223 \pm 96$	$15{,}047 \pm 112$	$28{,}401 \pm 176$
	PO-DB, no penalty	$8{,}059 \pm 62$	$11{,}928 \pm 126$	$19{,}915 \pm 137$	$33{,}730 \pm 104$
	Uniform play	4,067	8,133	16,267	32,533
Switching	General ICE (ours)	$1{,}418 \pm 88$	$3{,}240 \pm 127$	$7{,}413 \pm 100$	$15{,}847 \pm 30$
	PDB-PS	$3{,}072 \pm 58$	$4{,}749 \pm 91$	$7{,}985 \pm 153$	$15{,}137 \pm 179$
	PO-DB, no penalty	$3{,}805 \pm 117$	$5{,}933 \pm 166$	$9{,}787 \pm 261$	$16{,}448 \pm 66$
	Uniform play	2,000	4,000	8,000	16,000

weak violation for the looser value, which the sharper one changes by -0.2 ± 5.5 and -0.3 ± 6.3), since its multiplier is negligible at this budget. In contrast, General ICE and Conditional ICE with the practical λ grow their scales, restart 5 and 62.6 times, and stay close to uniform play in regret (17,122 and 17,486 against 17,811). PDB-PS attains the lowest regret and weak violation ($7{,}097 \pm 110$ and $1{,}516 \pm 19$). Thus, when the losses and the constraints agree, a growing penalty scale slows a fixed-range optimizer, and the certified Conditional ICE avoids this slowdown.

F.9 Other Horizons

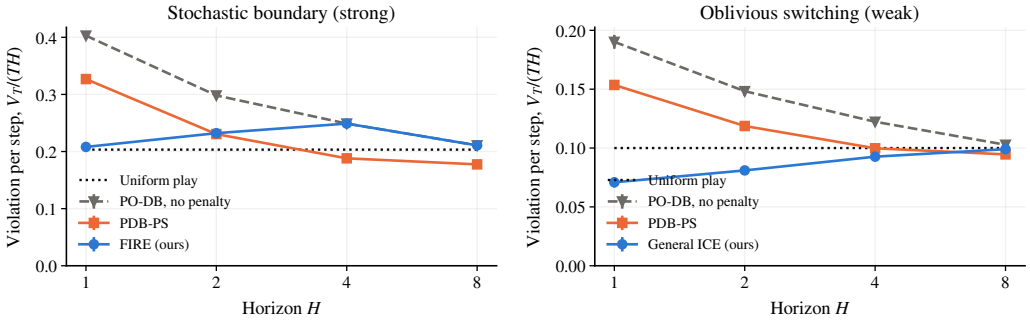


Figure 9: Violation per step, i.e., $V_T/(TH)$, at $T = 20{,}000$ for $H \in \{1, 2, 4, 8\}$ (mean curves over ten seeds with 95% confidence interval). For $H \geq 4$, FIRE's gate never opens, and FIRE coincides with the no-penalty control.

We repeat the main comparison for $H \in \{1, 4, 8\}$ at $T = 20{,}000$, where $|X| = 2H$ (Figure 9 and Table 8).

Switching Task: General ICE has lower weak violation than the no-penalty control at every horizon, specifically by $2{,}387 \pm 97$, $2{,}693 \pm 164$, $2{,}374 \pm 221$, and 600 ± 53 for $H = 1, 2, 4, 8$. It also has lower weak violation than PDB-PS for $H \leq 4$, whereas PDB-PS has 710 ± 161 less weak violation at $H = 8$.

Boundary Task: FIRE has the lowest strong violation at $H = 1$ ($4{,}161 \pm 13$, against $6{,}538 \pm 57$ for PDB-PS), where it restarts 64.7 times on average and stays close to uniform play. For $H \in \{4, 8\}$, the width $\beta_{t-1}(\pi_t)$ sums the bonuses of H steps over more states, and it stays above the estimated

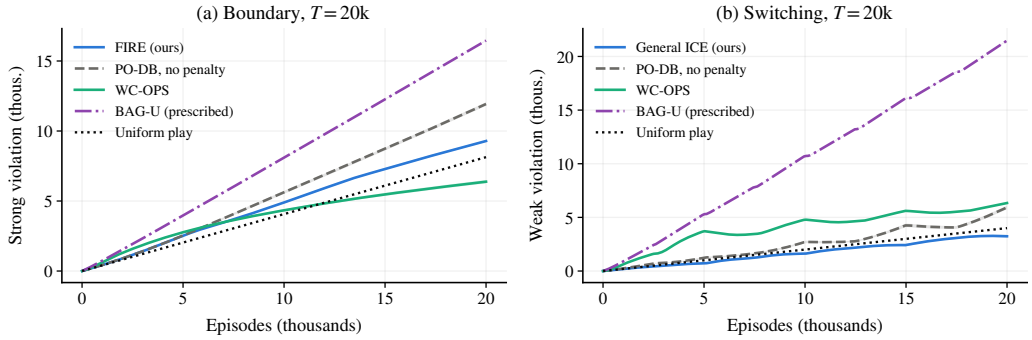


Figure 10: Violation over the episodes at $T = 20{,}000$ for our methods, the OM-based Slater-free baselines, and the references (mean curves with pointwise 95% intervals over ten seeds).

Table 9: Endpoints at $T = 20{,}000$ against WC-OPS and BAG-U (mean $\pm$ 95% half-width over ten seeds), with the learner time per episode (ms) and the episodic program that each learner solves.

	Stochastic boundary			Oblivious switching			
Method	**Regret**	**Strong viol.**	**ms/ep.**	**Regret**	**Weak viol.**	**ms/ep.**	**Program**
FIRE / General ICE (ours)	$-13{,}167 \pm 74$	$9{,}286 \pm 66$	0.39	$-3{,}885 \pm 106$	$3{,}240 \pm 127$	0.22	none
PDB-PS (practical $C = D = 1$)	$-13{,}273 \pm 104$	$9{,}223 \pm 96$	0.36	$-5{,}259 \pm 75$	$4{,}749 \pm 91$	0.35	none
PO-DB, no penalty	$-16{,}605 \pm 133$	$11{,}928 \pm 126$	0.22	$-6{,}229 \pm 163$	$5{,}933 \pm 166$	0.21	none
WC-OPS	$-10{,}260 \pm 215$	$6{,}381 \pm 123$	1.81	$-5{,}453 \pm 42$	$6{,}339 \pm 40$	1.75	KL proj.
BAG-U, prescribed θ	$-21{,}216 \pm 31$	$16{,}460 \pm 28$	1.56	$-14{,}639 \pm 21$	$21{,}482 \pm 37$	1.54	Eucl. proj.
Uniform play (exact)	$-11{,}723$	$8{,}133$	–	$-4{,}456$	$4{,}000$	–	–

cost in all 20,000 episodes (the final widths average 1.59 and 7.05). Thus, FIRE's gate never opens, and FIRE coincides with the no-penalty control, whereas PDB-PS lowers the strong violation by $4{,}868 \pm 148$ and $5{,}329 \pm 149$. Hence, at moderate budgets, the conservative width is the main obstacle for FIRE on longer horizons.

F.10 Comparison with OM-Based Slater-Free Methods

Methods: We implement the two OM-based Slater-free methods of Table 1 that handle bandit feedback and unknown transitions, WC-OPS (Stradi et al., 2025c, Algorithm 1) and BAG-U (Roy & Das, 2026, Algorithm 6). WC-OPS plays $\pi_t(a \mid x) \propto q_t(x, a)$ for an occupancy iterate q_t on (x, a, x') triples, forms implicit-exploration loss estimates, and takes a KL mirror-descent step over the occupancies whose induced transition lies in its confidence set. Also, $\langle \widehat{\boldsymbol{c}}_t - \boldsymbol{b}_t, q \rangle \le 0$ is satisfied for its cost estimate $\widehat{\boldsymbol{c}}_t$ and bonus $\boldsymbol{b}_t$. We use its prescribed $\eta = \gamma$ and radii with $\delta/30$, since its guarantees hold with probability $1 - 30\delta$. Its adaptive weights equal the running mean in all our runs, because the cumulative observed cost never exceeds its threshold C_t. BAG-U takes an AdaGrad step on the biased subgradient $\omega \widehat{\boldsymbol{\ell}}_t + \varphi'(\zeta_t)\omega \widehat{\boldsymbol{c}}_t \mathbf{1}\{C_t > 0\}$, where C_t is the observed episode cost, and projects onto the confidence occupancy polytope. We use its prescribed $\omega = 1/(2|X|H|A|)$, $\Lambda = \omega\sqrt{H}$, and $\theta = 1/(2m(T)) \approx 1.7 \times 10^{-4}$, with $\delta = 0.01$ as in its experiments. Its pseudocode increments the queue by $\omega[\langle \rho_t, \boldsymbol{c}_t \rangle]^+$, which is not observable under bandit feedback with unknown transitions. Hence, we use the observable $\omega[C_t]^+$, whose conditional mean upper-bounds it by Jensen's inequality. We time only the learners' own calls, which share the upper-occupancy-bound code, so the time difference isolates the per-episode programs.

Results: Table 9 and Figure 10 report the endpoints and curves at $T = 20{,}000$, and Table 10 reports the learner time per episode as H grows. **(i)** ***Computation:*** The OM baselines spend 1.3–4.2 ms per episode in their convex programs alone. At $H = 2$, their learner time is 4–8 times that of FIRE and ICE, and the absolute gap grows with H, from about 1.1 ms at $H = 1$ to 2.2–3.8 ms at $H = 8$. These timings reflect our Python implementations on tiny instances rather than tuned solvers, but the per-episode program is the structural cost that FIRE and ICE avoid; **(ii)** ***Constraint control:*** WC-OPS keeps its iterates in the estimated feasible set, and on the boundary task its strong

Table 10: Learner time per episode on the boundary task for $H \in \{1, 2, 4, 8\}$ ($T = 2{,}000$, three seeds, one process, mean $\pm$ 95% half-width), the part spent in the per-episode convex program, and the program's number of decision variables.

Learner time per episode (ms)	$\mathbf{H = 1}$	$\mathbf{H = 2}$	$\mathbf{H = 4}$	$\mathbf{H = 8}$
FIRE (ours): `PO-DB` + wrapper	0.30 ± 0.03	0.38 ± 0.02	0.48 ± 0.10	0.64 ± 0.06
WC-OPS	1.48 ± 0.22	1.78 ± 0.22	2.65 ± 0.23	4.47 ± 0.18
BAG-U	1.36 ± 0.05	1.54 ± 0.04	2.02 ± 0.03	2.82 ± 0.24
of which the convex program (WC-OPS)	1.34	1.60	2.41	4.15
of which the convex program (BAG-U)	1.27	1.41	1.84	2.58
Decision variables of the program	3	12	36	84

violation is $2{,}905 \pm 158$ below FIRE's at a $2{,}907 \pm 251$ higher regret. A normalization does not cap its occupancy steps, whereas the normalization caps the loss gaps that `PO-DB` sees at $\frac{1}{2}\max\{|\Delta\ell|, |\Delta c|\}$ (Appendix F.5), and FIRE's restarts reset its progress. On the switching task, WC-OPS's estimate averages the costs of both phases and its adaptive weights never activate, so its weak violation is $3{,}099 \pm 143$ above ICE's. Prescribed BAG-U keeps the cost weight $\theta e^{\theta\zeta_t}$ below 2.1×10^{-4}, so its penalty is inert, and its AdaGrad steps chase the risky actions faster than the control; **(iii)** ***Guarantees:*** In the adversarial regime, WC-OPS's guarantee is also weaker than ICE's. Its Theorem 4.6 (Stradi et al., 2025c) bounds an α-regret, the shortfall of the learner's cumulative reward below α times that of the best unconstrained policy in hindsight, with $\alpha = \rho/1+\rho$. Without a strictly feasible policy, as in our switching task, $\rho = 0$, and this guarantee is vacuous. WC-OPS's violation bound there concerns the weak violation. ICE's regret is against the feasible comparator and requires no margin.

Summary: At these budgets, WC-OPS buys faster constraint control on the stochastic boundary task with a convex program in every episode. On the adversarial switching task, ICE has less violation, a stronger guarantee, and a cheaper update. FIRE and ICE avoid the per-episode program, and their finite-budget weakness is the normalization-and-restart mechanism that makes a fixed-range PO optimizer usable.

F.11 Scope of the Experiments

Our experiments cover three tabular tasks with at most 15 non-terminal states, three actions, and budgets of at most $160{,}000$ episodes. They test executability, mechanisms, and finite-budget tradeoffs. They do not verify the asymptotic rates or cover function approximation, and our comparison with OM-based methods (Appendix F.10) is limited to these tasks. Our PDB-PS implementation represents the L_1 confidence sets exactly only for successor layers with at most two states, which holds for all of our instances. The practical λ, C, and D carry no theorem, and our intervals are pointwise and unadjusted for multiple comparisons.