BEST-OF-BOTH-WORLDS POLICY OPTIMIZATION FOR HEAVY-TAILED MARKOV DECISION PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study finite-horizon tabular Markov decision processes (MDPs) with heavy-tailed loss functions (*i.e.*, every one-step loss is assumed only to have a bounded ε-th raw moment for a known $\varepsilon \in (1, 2]$). Previously, *best-of-both-worlds* (BOBW) guarantee for this problem has only been achieved by Chen et al. (2026) using an *occupancy measure* (OM)-based algorithm. In this work, we develop a *policy optimization* (PO)-based approach, which is preferable to OM-based approaches in practice because PO updates the policy locally at each state and does not require solving a global optimization problem. In addition to the computational advantage, for known transitions, the heavy-tail terms in our regret bounds improve upon those of Chen et al. (2026) by factors of $\mathcal{O}(S^{2-2/\varepsilon}A^{2-3/\varepsilon+1/\varepsilon^2})$ and $\mathcal{O}(S^2A^{2-1/\varepsilon})$ in the adversarial and stochastic regimes, respectively, where S and A denote the sizes of the state and action spaces. For unknown transitions, our method does *not* require the truncated non-negativity assumption imposed by Chen et al. (2026); moreover, our heavy-tail terms are smaller by factors of $\mathcal{O}(S^{3-2/\varepsilon}A^{3-3/\varepsilon+1/\varepsilon^2})$ and $\mathcal{O}(HS^3A^{3-1/\varepsilon}\log T)$ in the adversarial and stochastic regimes, respectively, where T denotes the number of episodes.

1 INTRODUCTION

Heavy-tailed feedback is common in sequential decision-making problems where rare yet consequential observations cannot be ruled out. In such problems, the boundedness and sub-Gaussian assumptions may be restrictive: the feedback may have infinite variance even though a lower-order moment is finite. In this line of research, Bubeck et al. (2013) pioneer the study of stochastic heavy-tailed multi-armed bandits (MABs), where each arm a is associated with an unknown but fixed distribution ν_a and the i.i.d. losses $Y_{t,a} \sim \nu_a$ satisfy $\mathbb{E}[|Y_{t,a}|^\varepsilon] \le \sigma$ for some known $\varepsilon \in (1, 2]$ and $\sigma > 0$. This work initiates a broad line of research on stochastic heavy-tailed MABs (Agrawal et al., 2021), heavy-tailed linear bandits (Medina & Yang, 2016; Shao et al., 2018), and reinforcement learning (RL) with heavy-tailed losses (Zhuang & Sui, 2021).

Heavy-tailed feedback is not the only challenge a learner may face in practice: the nature of the underlying environment may also be unknown. This nature is commonly modeled through two canonical regimes. In the *adversarial regime*, the loss distributions may be selected arbitrarily and may vary across rounds. In the *stochastic regime*, by contrast, the loss distributions are unknown yet fixed, and the learner receives i.i.d. samples from them in each round. Although many algorithms achieve (nearly) optimal regret in one of these regimes, regime-specific algorithms typically require the regime to be known in advance. If the nature of the regime is unknown, an adversarially robust algorithm may therefore be overly conservative in a stochastic regime, whereas an algorithm tailored to stochastic regimes may incur linear regret under adversarial losses. This motivates the study of *best-of-both-worlds* (BOBW) algorithms, which, *without* knowing the regime in advance, simultaneously achieve (nearly) optimal worst-case regret in adversarial regimes and (nearly) optimal gap-dependent regret in stochastic regimes.

Beginning with the seminal work on MABs (Bubeck & Slivkins, 2012), BOBW learning has been studied for combinatorial semi-bandits (Zimmert et al., 2019), linear bandits (Lee et al., 2021), and episodic MDPs with known and unknown transitions (Jin & Luo, 2020; Jin et al., 2021). For MDPs with bounded feedback, Dann et al. (2023b) further show that BOBW guarantees can be achieved through *policy optimization* (PO) rather than global *occupancy-measure* (OM) updates. Meanwhile,

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: Qualitative comparisons with existing literature for BOBW MDPs.

Algorithm	Method	Loss Model	Transitions
Jin & Luo (2020)	OM-based	Nonnegative Bounded Losses	Known
Jin et al. (2021)	OM-based	Nonnegative Bounded Losses	Known
			Unknown
Dann et al. (2023b)	PO-based	Nonnegative Bounded Losses	Known
			Unknown
Chen et al. (2026)	OM-based	Heavy-tailed Losses	Known
		Heavy-tailed Losses w/ Truncated Non-negativity	Unknown
Algorithm 1 (**Ours**)	PO-based	Heavy-tailed Losses	Known
			Unknown

the BOBW property under heavy-tailed feedback is subsequently studied for MABs and linear bandits (Huang et al., 2022; Zhao et al., 2025), but its extension to MDPs remained unexplored until very recently, when Chen et al. (2026) develop OM-based algorithms with the first BOBW guarantees for heavy-tailed MDPs. Their approach, however, requires solving a global constrained optimization problem over the OM polytope instead of performing local policy updates. More importantly, in the known-transition setting with $\varepsilon = 2$, which includes bounded losses as a special case, the leading terms in their regret in adversarial and stochastic regimes incur additional factors of $SA^{3/4}$ and $S^2A^{3/2}$, respectively, compared with the best-known BOBW bounds for bounded-loss MDPs (Dann et al., 2023b). In addition, the unknown-transition guarantee of Chen et al. (2026) relies on a truncated non-negativity condition, which is generally difficult to verify in practice even for the special case of heavy-tailed MABs (Cheng et al., 2024), and Chen et al. (2026) explicitly identify removing this assumption as an open problem. These limitations motivate our central question: can a state-wise PO method achieve both heavy-tail robustness and BOBW adaptation, for known as well as unknown transitions, without imposing a truncated non-negativity assumption?

Challenges of BOBW PO for Heavy-tailed MDPs The key challenge in designing a BOBW algorithm for online learning problems with heavy-tailed losses is that heavy-tailed robustness and BOBW adaptation pull an online learner in opposite directions. In the setting of heavy-tailed MDPs, robustification asks the learner to clip or skip rare observations, but achieving a stochastic fast rate in the BOBW adaptation requires every clipping, exploration, and transition-estimation cost to shrink with the probability mass assigned to suboptimal decisions. The tension is sharper for PO: the learner does not optimize a single global OM. Instead, it updates its policy locally at every state and must strategically propagate all local regret guarantees through the transition dynamics without incurring an occupancy-distribution mismatch. The main difficulties and corresponding technical ideas are as follows.

(i) *Centered clipping compensation.* We construct an importance-weighted estimator of the (loss-based) advantage function $\bar{Q}_t(s,a) \coloneqq Q_t^{\pi_t}(s,a) - V_t^{\pi_t}(s)$ (Sutton & Barto, 2018), instead of the standard uncentered Q-value (for example, the Q-function estimator used by Luo et al. (2021); Dann et al. (2023b)). In the running of the follow-the-regularized-leader (FTRL), this allows a tail-compensation bonus $d_t(s,a)$ controlled by a comparator-independent proxy that vanishes at every deterministic policy. It can therefore be absorbed into a local bonus and propagated by the dilated Bellman recursion of Luo et al. (2021); Dann et al. (2023b), without requiring the algorithm to know the optimal action $\pi^\star(s)$.

(ii) *Signed visitation bias compensation.* In PO with bounded non-negative losses (Luo et al., 2021; Dann et al., 2023b), the visitation inflation and the transition estimation produce positive bias. In the heavy-tailed setting, the bias can take either sign, so the previous one-sided treatment no longer applies. To this end, we bound the bias by a statewise upper bound c_t and define a back-

Table 2: Comparisons with the regret of Chen et al. (2026) regarding the main terms with polynomial dependence on either T or Δ, where $\Delta \in \mathbb{R}_+^{SA}$ is defined in Definition 2.2. All the additive remainders of order $\mathcal{O}_\varepsilon(\mathrm{poly}(\sigma, H, S, A, \log T))$ with no dependence on Δ, which are dominated by the main terms for sufficiently large T in the adversarial regime and for sufficiently small $\Delta(s,a)$'s in the stochastic regime, are omitted. The common heavy-tail terms of Chen et al. (2026) and our work are $B_{\mathrm{adv}}^{\mathrm{ht}} := \sigma^{1/\varepsilon} H^{1+1/\varepsilon} (SA)^{1-1/\varepsilon} T^{1/\varepsilon}$ and $B_{\mathrm{stoc}}^{\mathrm{ht}} := H^{\varepsilon/(\varepsilon-1)} \sigma^{1/(\varepsilon-1)} \omega_{1/\varepsilon}(\Delta) \log T$, where $\omega_q(\Delta) := \sum_{s \neq s_H, a \neq \pi^\star(s)} \Delta(s,a)^{-q/(1-q)}$ for $q \in (0,1)$, with $\Delta_{\min} := \min_{s \neq s_H, a \neq \pi^\star(s)} \Delta(s,a) > 0$. The additional term of Chen et al. (2026) is $G := H^6 \sigma^{2/\varepsilon} S^2 (\omega_{1/2}(\Delta) + \Delta_{\min}^{-1}) \log T$ and the additional terms in our work are $B_{\mathrm{adv}}^{\mathrm{tr}} := \sigma^{1/\varepsilon} H^2 S \sqrt{AT}$, $B_{\mathrm{stoc}}^{\mathrm{frac}} := H^{\frac{2\varepsilon+1}{2\varepsilon-1}} \sigma^{\frac{2}{2\varepsilon-1}} S^{\frac{2}{2\varepsilon-1}} A^{\frac{1}{2\varepsilon-1}} \omega_{1/(2\varepsilon)}(\Delta) (\log T)^{\frac{2\varepsilon}{2\varepsilon-1}}$, and $B_{\mathrm{stoc}}^{\mathrm{tr}} := \sigma^{2/\varepsilon} H^4 S^2 A (\log T)^2 / \Delta_{\min}$. The final column reports the improvements in heavy-tail terms, which are the only main terms for known P. For unknown P, these terms dominate the other main terms contributed by learning the unknown P when $1 < \varepsilon < 2$ in adversarial regimes and when $1 < \varepsilon < 2$ and the gap $\Delta(s,a)$'s are sufficiently small in stochastic regimes; even when they do not dominate, our main terms still improve upon those of Chen et al. (2026) by an $\mathcal{O}_\varepsilon(S^{3/2} A^{7/4} / \sqrt{H})$ factor in adversarial regimes and an $\mathcal{O}_\varepsilon(SA^{2-1/\varepsilon}/H)$ factor in stochastic regimes.

Trans.	Regime	Chen et al. (2026)	Ours	Improvement in Heavy-tail Term
Known	**Adv.**	$\widetilde{\mathcal{O}}_\varepsilon(S^{2-2/\varepsilon} A^{2-3/\varepsilon+1/\varepsilon^2} B_{\mathrm{adv}}^{\mathrm{ht}})$	$\widetilde{\mathcal{O}}_\varepsilon(B_{\mathrm{adv}}^{\mathrm{ht}})$	$S^{2-2/\varepsilon} A^{2-3/\varepsilon+1/\varepsilon^2}$
	Stoc.	$\mathcal{O}_\varepsilon(S^2 A^{2-1/\varepsilon} B_{\mathrm{stoc}}^{\mathrm{ht}})$	$\mathcal{O}_\varepsilon(B_{\mathrm{stoc}}^{\mathrm{ht}})$	$S^2 A^{2-1/\varepsilon}$
Unknown	**Adv.**	$\widetilde{\mathcal{O}}_\varepsilon(S^{3-2/\varepsilon} A^{3-3/\varepsilon+1/\varepsilon^2} B_{\mathrm{adv}}^{\mathrm{ht}})$	$\widetilde{\mathcal{O}}_\varepsilon(B_{\mathrm{adv}}^{\mathrm{ht}} + B_{\mathrm{adv}}^{\mathrm{tr}})$	$S^{3-2/\varepsilon} A^{3-3/\varepsilon+1/\varepsilon^2}$
	Stoc.	$\mathcal{O}_\varepsilon(HS^3 A^{3-1/\varepsilon} \log T \cdot B_{\mathrm{stoc}}^{\mathrm{ht}} + G)$	$\mathcal{O}_\varepsilon(B_{\mathrm{stoc}}^{\mathrm{ht}} + B_{\mathrm{stoc}}^{\mathrm{frac}} + B_{\mathrm{stoc}}^{\mathrm{tr}})$	$HS^3 A^{3-1/\varepsilon} \log T$

ward compensation function C_t whose recursion transports twice the future c_t. The factor-two construction ensures that the induced auxiliary loss has a nonnegative Bellman residual. The performance-difference identity can then be used to control the entire signed-bias contribution.

(iii) *Comparator removal before the heavy-tail power.* In the bounded-loss case, to obtain the final BOBW guarantee, Dann et al. (2023b) establish a second-order bound of $\widetilde{\mathcal{O}}(\sum_{s,a}(\mathbb{E}[\sum_{t=1}^T \widetilde{\mu}_t(s)\pi_t(a \mid s)(1 - \pi_t(a \mid s))])^{\frac{1}{2}})$, where $\widetilde{\mu}_t(s)$ is an inflated state-occupancy proxy and π_t is the policy adopted by the learner in episode t. In the heavy-tailed case, the corresponding local geometry is fractional rather than quadratic. We instead prove, for every deterministic comparator policy π, a data-dependent bound of the form $\widetilde{\mathcal{O}}(\sum_s \sum_{a \neq \pi(s)} \widetilde{n}_T(s,a)^{1/\varepsilon})$, where $\widetilde{n}_T(s,a) := \mathbb{E}[\sum_{t=1}^T \widetilde{\mu}_t(s)\pi_t(a \mid s)]$. In this way, transition uncertainty can be converted into visitation counts before applying the fractional $1/\varepsilon$ power, and can still be absorbed by the self-bounding inequality, which thus permits a polylogarithmic stochastic regret guarantee even when transitions are unknown.

Main Results and Broader Implications We compare our results with those in the most related works in Tables 1 and 2. Moreover, when $\varepsilon = 2$, the main terms in our results across both transition settings and both regimes match those of Dann et al. (2023b) on all parameters. Algorithmically, we believe some developments in our work may also be of interest beyond heavy-tailed MDPs. The centered clipping-compensation mechanism and our new design of the dilated-bonus avoid the conditional "loss-shifting" operation used in the BOBW analysis for PO with a Tsallis regularizer by Dann et al. (2023b). Moreover, combining the clipping compensation bonus with a $\widetilde{\mu}_t$-adaptive learning-rate schedule removes the need for the action-level explicit exploration, which is required even in previous BOBW algorithms for heavy-tailed MABs (Zhao et al., 2025).

2 Preliminaries

We consider a finite-horizon episodic tabular MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, \{\ell_t\}_{t=1}^T)$ with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a horizon H, a transition kernel P, and loss functions $\{\ell_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{t=1}^T$. W.l.o.g., we assume $\mathcal{S}$ consists of disjoint layers $\mathcal{S} = \mathcal{S}_0 \cup \cdots \cup \mathcal{S}_H$, where $\mathcal{S}_0 = \{s_0\}$ and $\mathcal{S}_H = \{s_H\}$ are the initial and terminal layers. Let $h(s)$ denote the layer containing state s.

Transitions are only possible between consecutive layers; that is, $P(\cdot \mid s,a)$ is supported only on $\mathcal{S}_{h(s)+1}$. We denote the sizes of the state and action spaces by $S \coloneqq |\mathcal{S}|$ and $A \coloneqq |\mathcal{A}|$, respectively. A randomized Markov policy π assigns an action distribution $\pi(\cdot \mid s) \in \Delta(\mathcal{A})$ to each state s; for a deterministic policy, let $\pi(s)$ be its chosen action. For any transition kernel $\widetilde{P}$ and policy π, the state and state-action occupancy measures are $\mu^{\widetilde{P},\pi}(s) \coloneqq \mathbb{P}_{\widetilde{P},\pi}(s_{h(s)} = s)$ and $\mu^{\widetilde{P},\pi}(s,a) \coloneqq \mu^{\widetilde{P},\pi}(s)\pi(a \mid s)$, respectively; we abbreviate $\mu^{P,\pi}$ as μ^{π} when the context is clear.

Learning Protocol The learner interacts with the environment for T episodes, knowing $\mathcal{S}$, its layer structure, and $\mathcal{A}$. We consider both known and unknown transitions. Let $\mathcal{F}_{t-1}$ be the σ-algebra generated by all the observations before episode t. At the beginning of episode t, the learner uses $\mathcal{F}_{t-1}$ to specify a policy π_t, which remains fixed throughout the episode. Meanwhile, the environment chooses a conditional joint distribution for the loss table $Y_t = \{Y_t(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}$, possibly depending on $\mathcal{F}_{t-1}$, but the coordinates of Y_t may be arbitrarily dependent. We write $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and $\ell_t(s,a) \coloneqq \mathbb{E}_t[Y_t(s,a)]$. Conditional on $\mathcal{F}_{t-1}$, the loss table is sampled independently of the randomness in the learner's actions and the transitions. Starting from $s_{t,0} = s_0$, at each stage $h = 0,\dots,H-1$, the learner draws $a_{t,h} \sim \pi_t(\cdot \mid s_{t,h})$, incurs loss $Y_t(s_{t,h},a_{t,h})$, and observes the next state $s_{t,h+1} \sim P(\cdot \mid s_{t,h},a_{t,h})$. Upon reaching $s_{t,H} = s_H$, the learner receives the individual losses $\{Y_t(s_{t,h},a_{t,h})\}_{h=0}^{H-1}$, and then moves on to episode $t+1$. Note that the learner observes only the losses at visited state-action pairs; the remaining table entries and the loss distribution are not revealed.

In this work, we impose only the heavy-tailed condition on the losses (Bubeck et al., 2013).

Definition 2.1 (Heavy-tailed Loss Distributions)**.** *There exist some fixed $\varepsilon \in (1,2]$ and $\sigma > 0$ such that the ε-th (raw) moment of the loss satisfies $\mathbb{E}_t[|Y_t(s,a)|^{\varepsilon}] \le \sigma$ almost surely for every t,s,a.*

We write $L_\sigma \coloneqq \sigma^{1/\varepsilon}$ for the corresponding loss scale. In the previous work on BOBW MDPs with heavy-tailed losses by Chen et al. (2026), the truncated non-negativity assumption (Huang et al., 2022; Chen et al., 2025) is imposed in the unknown-transition case: the heavy-tailed loss distribution chosen by the environment satisfies $\mathbb{E}_t[\ell_t(s,a)\cdot\mathbb{1}\{|\ell_t(s,a)| > M\}] \ge 0$ for arbitrary $M \ge 0$. In this work, we do not need such an assumption. Nevertheless, we assume the learner has access to the knowledge of (ε,σ), which is also required by Chen et al. (2026). Moreover, Genalti et al. (2024) have shown that the worst-case lower bound for the setting with known (ε,σ) cannot be attained when either ε or σ is unknown, even in the special case of heavy-tailed MABs.

For a policy π, let $Q_t^{\pi}(s,a)$ and $V_t^{\pi}(s)$ denote its expected cumulative loss from (s,a) and s, respectively, under the true kernel P and the conditional mean loss function ℓ_t. With $V_t^{\pi}(s_H) \coloneqq 0$, they satisfy the Bellman equations

$$Q_t^{\pi}(s,a) \coloneqq \ell_t(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_t^{\pi}(s')]\,, \quad V_t^{\pi}(s) \coloneqq \sum_a \pi(a \mid s) Q_t^{\pi}(s,a)\,.$$

The learner's performance is measured by its *regret* against a fixed deterministic policy π:

$$\mathrm{Reg}_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T}\left(V_t^{\pi_t}(s_0) - V_t^{\pi}(s_0)\right)\right], \tag{1}$$

where the expectation is taken over the randomness of the learner, transitions, and losses. In the *adversarial regime*, the environment may determine the loss distributions arbitrarily. We also consider the *adversarial regime with a self-bounding constraint* (Zimmert & Seldin, 2021; Jin et al., 2021), defined as follows.

Definition 2.2 (Adversarial Regime with a Self-bounding Constraint)**.** *Fix a deterministic policy $\pi^\star$, a gap function $\Delta \in \mathbb{R}_{\ge 0}^{SA}$ satisfying $\Delta(s,\pi^\star(s)) = 0$ and $\Delta(s,a) > 0$, $\forall s \ne s_H$ and $a \ne \pi^\star(s)$, and a corruption level $\mathcal{C} \ge 0$. The environment satisfies a $(\pi^\star,\Delta,\mathcal{C})$ self-bounding condition if*

$$\mathrm{Reg}_T(\pi^\star) \ge \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s\ne s_H,a\ne\pi^\star(s)} \mu^{\pi_t}(s,a)\Delta(s,a)\right] - \mathcal{C}\,. \tag{2}$$

In the *stochastic regime*, the loss tables are drawn i.i.d. across episodes from a fixed distribution with mean ℓ. In this case, $\pi^\star \in \arg\max_\pi \mathrm{Reg}_T(\pi)$ is the deterministic optimal policy and $\Delta(s,a) =$

$Q^{\pi^\star}(s,a) - V^{\pi^\star}(s)$. The performance-difference identity then gives (2) with equality and $\mathcal{C} = 0$. More generally, $\mathcal{C}$ allows the cumulative violation of this gap identity due to adversarial corruption, which subsumes the *stochastic regime with adversarial corruptions* (Lykouris et al., 2018).

Additional Notation We use $\mathcal{O}_\varepsilon(\cdot)$ to suppress all constants that may depend on ε but not on any other parameters. $\widetilde{\mathcal{O}}_\varepsilon(\cdot)$ suppresses all constants (possibly depending on ε) and all logarithmic factors. For $f, g : \mathbb{N}_+ \to \mathbb{R}_+$, we sometimes write $f \lesssim g$ if $f = \mathcal{O}(g)$ and $f \lesssim_\varepsilon g$ if $f = \mathcal{O}_\varepsilon(g)$. For arbitrary $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we sometimes write $p_{t,s} \coloneqq \pi_t(\cdot \mid s)$ and $p_{t,s,a} \coloneqq \pi_t(a \mid s)$.

3 Algorithm

Our algorithm runs FTRL with an α-Tsallis regularizer locally at each state, using an advantage function estimator and three bonus terms to handle clipping bias, visitation uncertainty, and the propagation of local regret through the MDP. We describe our algorithm for unknown transitions with a transition confidence set $\mathcal{P}_t$, which reduces to $\{P\}$ when transitions are known. The complete procedure is given in Algorithm 1.

Transition Confidence Sets When transitions are unknown, as in previous work (Jin et al., 2020; Luo et al., 2021), we estimate them from previously observed trajectories. For $s \in \mathcal{S}_h$ and $s' \in \mathcal{S}_{h+1}$, let $N_t(s,a) \coloneqq \sum_{\tau=1}^{t-1} \mathbb{1}\{s_{\tau,h} = s,\ a_{\tau,h} = a\}$ denote the number of visits to (s,a) before episode t, and let $N_t(s,a,s') \coloneqq \sum_{\tau=1}^{t-1} \mathbb{1}\{s_{\tau,h} = s,\ a_{\tau,h} = a,\ s_{\tau,h+1} = s'\}$ denote the number of such visits followed by a transition to s'. In particular, $N_1(s,a) = N_1(s,a,s') = 0$. We use $N_t^+(s,a) \coloneqq \max\{1, N_t(s,a)\}$ and the empirical transition $\widehat{P}_t(s' \mid s,a) \coloneqq N_t(s,a,s')/N_t(s,a)$ whenever $N_t(s,a) > 0$; otherwise, $\widehat{P}_t(s' \mid s,a) = 1/S_{h(s')}$. Let $\overline{T} \coloneqq \max\{2,T\}$, $\delta \coloneqq \overline{T}^{-64}$, and $\iota \coloneqq \log(16HS^2A\overline{T}^2/\delta)$. For $s \in \mathcal{S}_h$, we construct a coordinatewise confidence radius and a confidence set of transition kernels as follows:

$$\mathrm{rad}_t(s,a,s') \coloneqq 2\sqrt{\frac{\widehat{P}_t(s' \mid s,a)\iota}{N_t^+(s,a)}} + \frac{14\iota}{3N_t^+(s,a)}\,, \qquad \mathcal{P}_t \coloneqq \prod_{s,a} \mathcal{P}_t(s,a)\,, \tag{3}$$

$$\text{where } \mathcal{P}_t(s,a) \coloneqq \left\{ q \in \Delta(\mathcal{S}_{h+1}) : |q(s') - \widehat{P}_t(s' \mid s,a)| \le \mathrm{rad}_t(s,a,s')\,, \forall s' \in \mathcal{S}_{h+1} \right\}. \tag{4}$$

The confidence set induces upper and lower bounds on state visitation. For any policy π, define $\overline{\mu}_t^\pi(s) \coloneqq \max_{\widetilde{P} \in \mathcal{P}_t} \mu^{\widetilde{P},\pi}(s)$ and $\underline{\mu}_t^\pi(s) \coloneqq \min_{\widetilde{P} \in \mathcal{P}_t} \mu^{\widetilde{P},\pi}(s)$. On the event $\mathcal{E}_t \coloneqq \{P \in \mathcal{P}_t\}$, we have $\underline{\mu}_t^\pi(s) \le \mu^\pi(s) \le \overline{\mu}_t^\pi(s)$. Let $\mathcal{E} \coloneqq \bigcap_{t=1}^T \mathcal{E}_t$.

Advantage Function Estimator $\widetilde{Q}_t$ and Clipping Compensation Bonus d_t In heavy-tailed online learning, the constructed loss estimator may have prohibitively large magnitude, which can make the FTRL update unstable. A typical way to address this is to clip the estimator when it becomes too large. However, this can introduce a prohibitively large bias for the optimal action, which cannot be directly controlled by FTRL and in turn requires the truncated non-negativity assumption (Huang et al., 2022; Chen et al., 2026). To eliminate this assumption, following the bonus-shifted FTRL principle of Zhao et al. (2025), we construct a bonus term $d_t(s,a)$ and subtract it from the clipped loss estimate used by FTRL to compensate for the clipping bias.

However, the standard importance-weighted estimator of the uncentered Q-value used by Luo et al. (2021); Dann et al. (2023b) would leave the clipping-compensation cost too large to be propagated through a comparator-independent dilated bonus B_t (introduced below). We instead construct an importance-weighted estimator of the (loss-based) *advantage function* $\bar{Q}_t(s,a) \coloneqq Q_t^{\pi_t}(s,a) - V_t^{\pi_t}(s)$ (Sutton & Barto, 2018). Subtracting the state-value baseline does not change the performance-difference identity, but it gives the relevant heavy-tail moment an additional policy-concentration factor. Consequently, the clipping bonus $d_t(s,a)$ can be further controlled by a local, $\pi^\star(s)$-independent upper bound that vanishes at every deterministic policy. As a result, $d_t(s,a)$ can be incorporated into a local bonus b_t and propagated by the dilated Bellman recursion B_t of Luo et al. (2021); Dann et al. (2023b), without requiring the algorithm to know $\pi^\star(s)$.

Algorithm 1 Best-of-both-worlds Policy Optimization for Heavy-tailed MDPs

Input: $T, H, \mathcal{S}, \mathcal{A}, \varepsilon, \sigma$.
Set $G_0(s) = 1$ and all transition counts to zero.
for $t = 1, \dots, T$ **do**
  Construct $\widehat{P}_t$ and $\mathcal{P}_t$ from (3)–(4).
  for $h = 0, \dots, H-1$ **do**
    Using π_t on layers $0, \dots, h-1$, compute $\overline{\mu}_t(s), \underline{\mu}_t(s)$ according to the definitions of $\overline{\mu}_t(s)$ and $\underline{\mu}_t(s)$ in Section 3 for all $s \in \mathcal{S}_h$.
    Compute $\widetilde{\mu}_t(s), G_t(s)$, set $\beta_t(s)$ via (13), and solve (15) for $\pi_t(\cdot \mid s)$.
  end for
  Compute c_t, then compute C_t backward via (8).
  Compute h_t, z_t via (9), ν_t, b_t via (11) and (10), and B_t backward via (12).
  Compute $\mathsf{D}_{\pi_t(\cdot|s)}(a)$ via (5), $M_t(s,a)$, and $d_t(s,a)$ via (7).
  Execute π_t in the true MDP and observe $\{Y_t(s_{t,h}, a_{t,h})\}_{h=0}^{H-1}$.
  Compute $R_{t,h}$, then $\widetilde{Q}_t$ via (6) and $\widehat{Q}_t$ via (7).
  Set g_t via (14) and update transition counts.
end for

In detail, for $p \in \mathrm{int}(\Delta(\mathcal{A}))$, we define a *centered* importance-weighting kernel and its ε-th moment

$$\mathsf{K}_p(a,b) \coloneqq \frac{\mathbb{1}\{a=b\}}{p_a} - 1\,, \qquad \mathsf{D}_p(a) \coloneqq \sum_b p_b |\mathsf{K}_p(a,b)|^\varepsilon\,. \tag{5}$$

Let $R_{t,h} \coloneqq \sum_{j=h}^{H-1} Y_t(s_{t,j}, a_{t,j})$. We construct the advantage function estimator

$$\widetilde{Q}_t(s,a) \coloneqq \frac{\mathbb{1}\{s_{t,h}=s\} R_{t,h}}{\widetilde{\mu}_t(s)} \mathsf{K}_{p_{t,s}}(a, a_{t,h})\,, \tag{6}$$

where $\widetilde{\mu}_t(s) \coloneqq \overline{\mu}_t(s) + \gamma_t$ is an inflated upper occupancy with $\gamma_t \coloneqq \min\{1, \Gamma/t\}$ (Γ is specified in the learning-rate schedule below). Via the identity $\sum_b p_b \mathsf{K}_p(a,b) q_b = q_a - \langle p, q\rangle$ for any $q \in \mathbb{R}^A$, we can show that $\mathbb{E}_t[\widetilde{Q}_t(s,a)] = \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)} \bar{Q}_t(s,a)$.

For $p \in \Delta(\mathcal{A})$, let $\rho(p) \coloneqq \min\{\|p\|_\infty, 1 - \|p\|_\infty\}$ and $\bar{p}_a \coloneqq \min\{p_a, \rho(p)\}$. After constructing $\widetilde{Q}_t(s,a)$, we truncate the estimator and compensate for the resulting bias. Define the local moment scale $\Sigma_t(s) \coloneqq H^\varepsilon \sigma \widetilde{\mu}_t(s)^{1-\varepsilon}$ and the threshold constant $k_\alpha \coloneqq (1-\alpha)/64$. On $\mathcal{E}_t$, the conditional ε-th moment of $\widetilde{Q}_t(s,a)$ is at most $\Sigma_t(s)\mathsf{D}_{p_{t,s}}(a)$. Therefore, we define the clipped advantage estimator $\widehat{Q}_t(s,a)$ and the clipping compensation bonus $d_t(s,a)$ by

$$\widehat{Q}_t(s,a) \coloneqq \widetilde{Q}_t(s,a)\mathbb{1}\{|\widetilde{Q}_t(s,a)| \le M_t(s,a)\}\,, \quad d_t(s,a) \coloneqq \Sigma_t(s)\mathsf{D}_{p_{t,s}}(a) M_t(s,a)^{1-\varepsilon}\,, \tag{7}$$

where $M_t(s,a) \coloneqq k_\alpha \beta_t(s) \bar{p}_{t,s,a}^{\alpha-1}$ is the clipping threshold and $\beta_t(s)$ is the (inverse) learning rate specified in (13) below. On $\mathcal{E}_t$, $d_t(s,a)$ upper bounds the clipping bias $\tau_t(s,a) \coloneqq \mathbb{E}_t[\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)]$ and is subtracted from the estimate in the FTRL update (15).

Signed Visitation-bias Compensation Bonus C_t The visitation denominator $\widetilde{\mu}_t(s)$ in $\widetilde{Q}_t(s,\cdot)$ creates a bias $r_t(s,\cdot) \coloneqq \bar{Q}_t(s,\cdot) - \mathbb{E}_t[\widetilde{Q}_t(s,\cdot)]$ even before truncation. On $\mathcal{E}_t$, the discrepancy between $\widetilde{\mu}_t(s)$ and the true occupancy $\mu_t^{\pi_t}(s)$ is bounded by $w_t(s) \coloneqq \overline{\mu}_t(s) - \underline{\mu}_t(s) + \gamma_t$. Let $c_t(s) \coloneqq 2HL_\sigma w_t(s)/\widetilde{\mu}_t(s)$ be the statewise upper bound on $|r_t(s,a)|$ for every action a. To control the regret contributed by $r_t(s,\cdot)$, we follow Dann et al. (2023b) to construct a Bellman equation $C_t(s,\cdot)$ for $c_t(s,\cdot)$ and include $C_t(s,\cdot)$ in the FTRL update (15). However, for the canonical bounded nonnegative losses considered by Dann et al. (2023b), the corresponding shrinkage bias admits a one-sided treatment using the standard Bellman equation for $c_t(s,\cdot)$, whereas in our case, the advantage function $\bar{Q}_t(s,a)$ can take either sign, so that argument no longer applies. To cope with this issue, we propagate this c_t backward through an amplified Bellman recursion:

$$C_t(s,a) \coloneqq \max_{\widetilde{P} \in \mathcal{P}_t} \mathbb{E}_{s' \sim \widetilde{P}(\cdot|s,a), a' \sim \pi_t(\cdot|s')} \left[2c_t(s') + C_t(s',a')\right]\,, \tag{8}$$

with $C_t(s_H, a) = 0$. We denote $P^C_{t,s,a} \in \arg\max_{\widetilde{P}\in\mathcal{P}_t} \mathbb{E}_{s'\sim\widetilde{P}(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[2c_t(s') + C_t(s', a')]$. Importantly, the recursion defining the backward compensation function C_t transports *twice* the future values of c_t. The factor-two construction ensures that the artificial loss $\mathcal{L}^{\text{bias}}_t(s, a) := r_t(s, a) + C_t(s, a) + c_t(s)$ has a nonnegative Bellman residual, which allows the performance-difference identity to control its contribution to regret.

Dilated Bonus B_t For $p \in \Delta(\mathcal{A})$, let $\psi_\alpha(p) := -\frac{1}{\alpha}\sum_a (p_a^\alpha - p_a)$ be the α-Tsallis entropy regularizer with $\alpha \in (0, 1)$. Define

$$h_\alpha(p) := -\psi_\alpha(p) = \frac{1}{\alpha}\left(\sum_a p_a^\alpha - 1\right), \quad z_\alpha(p) := \sum_a \bar{p}_a^\alpha, \tag{9}$$

which correspond to the penalty and stability terms in the FTRL regret bound, respectively. We denote $h_t(s) := h_\alpha(p_{t,s})$ and $z_t(s) := z_\alpha(p_{t,s})$, and set $K_{\text{ht}} := 2^{14}(1-\alpha)^{1-\varepsilon}$. To tackle the occupancy mismatch between π_t and the optimal policy, following Luo et al. (2021); Dann et al. (2023b), we construct a local bonus b_t; its dilated Bellman recursion defines B_t, which is included as a bonus in the FTRL update. Formally, in our problem, $b_t(s)$ is defined as

$$b_t(s) := \mathbb{1}\{t \geq 2\}\big(\beta_t(s) - \beta_{t-1}(s)\big)h_t(s) + K_{\text{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s) + \nu_t(s), \tag{10}$$

$$\nu_t(s) := \frac{16\rho(p_{t,s})^{1-\alpha}}{(1-\alpha)\beta_t(s)}\sum_a \pi_t(a \mid s)C_t(s, a)^2, \tag{11}$$

where (a) the first term in (10) accounts for the penalty term of FTRL, (b) the second term comes from the stability term of FTRL, the clipping bias τ_t, and the clipping bonus d_t, and (c) the third term accounts for the visitation-bias bonus C_t. We then construct a dilated Bellman equation for b_t:

$$B_t(s, a) := b_t(s) + \left(1 + \frac{1}{H}\right)\max_{\widetilde{P}\in\mathcal{P}_t} \mathbb{E}_{s'\sim\widetilde{P}(\cdot|s,a)a'\sim\pi_t(\cdot|s')}[B_t(s', a')], \tag{12}$$

with $B_t(s_H, a) = b_t(s_H)$. We denote $P^B_{t,s,a} \in \arg\max_{\widetilde{P}\in\mathcal{P}_t} \mathbb{E}_{s'\sim\widetilde{P}(\cdot|s,a)a'\sim\pi_t(\cdot|s')}[B_t(s', a')]$. The dilation factor $1 + 1/H$ enables the local regret bounds to be propagated backward and controlled in terms of the bonus value under the learner's occupancy μ^{π_t}.

Learning-rate Schedule and Overall FTRL Update To balance the penalty and stability terms of FTRL, we propose the following $\widetilde{\mu}_t(s)$-dependent (inverse) learning rate $\beta_t(s)$:

$$\beta_t(s) := \beta_0 + c_\beta H L_\sigma G_t(s)^{1/\varepsilon}, \tag{13}$$

where $G_t(s) := G_{t-1}(s) + \widetilde{\mu}_t(s)^{1-\varepsilon}$ and $G_0(s) := 1$. The initial learning rate is set to $\beta_0 := 2^{12}eK_{\text{saf}}H^3A^{1-\alpha}L_\sigma/(1-\alpha)$ to ensure a stable update of FTRL, where $\Gamma := \left\lceil\left(2^{12}eK_{\text{saf}}H^2A^{1-\alpha}/((1-\alpha)c_\beta)\right)^{\varepsilon}\right\rceil$, $c_\beta := 2^{12}/(1-\alpha)$, and $K_{\text{saf}} := c_\beta/\alpha + K_{\text{ht}}c_\beta^{1-\varepsilon} + 1$. Further, we define the loss function $g_t(s, \cdot)$ used in FTRL as

$$g_t(s, \cdot) := \widehat{Q}_t(s, \cdot) - d_t(s, \cdot) - B_t(s, \cdot) - C_t(s, \cdot). \tag{14}$$

At the beginning of episode t, Algorithm 1 solves the following FTRL problem at each state $s \in \mathcal{S}$:

$$\pi_t(\cdot \mid s) \in \arg\min_{p\in\Delta(\mathcal{A})}\left\{\left\langle p, \sum_{\tau=1}^{t-1} g_\tau(s, \cdot)\right\rangle + \beta_t(s)\psi_\alpha(p)\right\}. \tag{15}$$

4 Analysis

In this section, we present unified results for both known and unknown transitions. Define the expected actual and inflated local counts as $n_T(s, a) := \mathbb{E}[\sum_{t=1}^T \mu^{\pi_t}(s)\pi_t(a \mid s)]$ and $\widetilde{n}_T(s, a) := \mathbb{E}[\sum_{t=1}^T \widetilde{\mu}_t(s)\pi_t(a \mid s)]$. For a deterministic comparator π, let $\mathsf{D}_T(\pi) := \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a}[\mu^{\pi_t}(s, a) - \mu^\pi(s, a)]_+\right]$. We write $\mathbb{1}^U\{x\}$ for x when transitions are unknown and for 0 when they are known. We present only the main terms in the bounds below; the full statements, including additive terms of order $\mathcal{O}_\varepsilon(\text{poly}(\sigma, H, S, A, \log T))$ without dependence on Δ, are given in Appendices B and C.

We first present the following theorem.

Theorem 4.1 (Data-dependent Bound (Informal)). *Under Definition 2.1, Algorithm 1 with $\alpha = 1/\varepsilon$ satisfies, for every deterministic policy π,*

$$\mathrm{Reg}_T(\pi) \lesssim_\varepsilon H\sigma^{1/\varepsilon}(\log T)^{1-1/\varepsilon} \sum_{s\in\mathcal{S}\setminus\mathcal{S}_H} \sum_{a\neq\pi(s)} \widetilde{n}_T(s,a)^{1/\varepsilon} + \mathbb{1}^U\left\{\sigma^{1/\varepsilon} H^{3/2} S \log T \cdot \sqrt{A\mathsf{D}_T(\pi)}\right\}.$$

Remark 4.2. *In general, a BOBW analysis needs to charge data-dependent costs to suboptimal state-action visits. In the bounded-feedback analysis of Dann et al. (2023b), this is achieved through a second-order bound of $\widetilde{\mathcal{O}}(\sum_{s,a}\sqrt{\mathbb{E}[\sum_{t=1}^T \widetilde{\mu}_t(s)\pi_t(a\mid s)(1-\pi_t(a\mid s))]})$. In the heavy-tailed case, the corresponding local geometry is fractional rather than quadratic. The construction of our bonus B_t instead yields, for every deterministic comparator π, a data-dependent bound of $\widetilde{\mathcal{O}}_\varepsilon(\sum_s \sum_{a\neq\pi(s)} \widetilde{n}_T(s,a)^{1/\varepsilon})$ as shown above. For unknown transitions, Theorem C.2 further guarantees a coordinatewise conversion $\widetilde{n}_T(s,a) \lesssim n_T(s,a) + \sqrt{\mathsf{X}_T\, n_T(s,a)} + \mathsf{Z}_T$, where X_T, Z_T are confidence-complexity quantities. We apply this conversion before taking the $1/\varepsilon$ power. Transition uncertainty therefore produces terms involving $n_T(s,a)^{1/(2\varepsilon)}$ that can still be absorbed by the self-bounding inequality, thereby yielding $\mathcal{O}(\log T)$ regret even for unknown transitions.*

Before presenting the proof sketch of Theorem 4.1, we first state the following regret guarantees, which follow from Theorem 4.1.

Theorem 4.3 (Regret Guarantees (Informal)). *Under Definition 2.1, Algorithm 1 with $\alpha = 1/\varepsilon$ guarantees the following bound in the adversarial regime for every deterministic policy π,*

$$\mathrm{Reg}_T(\pi) = \mathcal{O}_\varepsilon\left(\sigma^{1/\varepsilon}\left[H^{1+1/\varepsilon}(SA)^{1-1/\varepsilon}T^{1/\varepsilon}(\log T)^{1-1/\varepsilon} + \mathbb{1}^U\left\{H^2 S\log T\sqrt{AT}\right\}\right]\right). \quad (16)$$

If the self-bounding condition in Definition 2.2 also holds, Algorithm 1 achieves

$$\mathrm{Reg}_T(\pi^\star) = \mathcal{O}_\varepsilon\left(\mathcal{U} + \mathcal{U}^{1-1/\varepsilon}\mathcal{C}^{1/\varepsilon} + \mathbb{1}^U\left\{\mathcal{V} + \mathcal{V}^{1-1/(2\varepsilon)}\mathcal{C}^{1/(2\varepsilon)} + \mathcal{Z} + \sqrt{\mathcal{Z}\mathcal{C}}\right\}\right), \quad (17)$$

where $\mathcal{U} := H^{\varepsilon/(\varepsilon-1)}\sigma^{1/(\varepsilon-1)}\omega_{1/\varepsilon}(\Delta)\log T$, $\mathcal{V} := H^{(2\varepsilon+1)/(2\varepsilon-1)}\sigma^{2/(2\varepsilon-1)}S^{2/(2\varepsilon-1)}A^{1/(2\varepsilon-1)} \times (\log T)^{(2\varepsilon)/(2\varepsilon-1)}\omega_{1/(2\varepsilon)}(\Delta)$, and $\mathcal{Z} := \sigma^{2/\varepsilon}H^4S^2A\log^2 T/\Delta_{\min}$.

Proof Sketch of Theorem 4.1. We restrict attention to the part of the regret bound conditioned on $\mathcal{E}$ because the contribution from $\mathcal{E}^c$ is only of order $\mathcal{O}_\varepsilon(\mathrm{poly}(\sigma,H,S,A,\log T))$. Fix an arbitrary deterministic comparator policy π. We slightly abuse notation by defining $B_t^\pi(s) := B_t(s,\pi(s))$, $\overline{B}_t(s) := \langle \pi_t(\cdot\mid s), B_t(s,\cdot)\rangle$, and $\kappa_H := 1 + 1/H$. We start with the regret decomposition:

$$\mathrm{Reg}_T(\pi) = \mathbb{E}\left[\sum_{t=1}^T V_t^{\pi_t}(s_0) - V_t^\pi(s_0)\right] = \mathbb{E}\left[\sum_{t=1}^T\sum_{s\in\mathcal{S}}\mu^\pi(s)\left\langle \pi_t(\cdot\mid s) - e_{\pi(s)}, \bar{Q}_t(s,\cdot)\right\rangle\right]$$

$$= \sum_{s\in\mathcal{S}}\mu^\pi(s)\left(\underbrace{\mathbb{E}\left[\sum_{t=1}^T\left\langle \pi_t(\cdot\mid s) - e_{\pi(s)}, \bar{Q}_t(s,\cdot) - B_t(s,\cdot)\right\rangle\right]}_{\mathsf{Loc}^\pi(s)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T\left\langle \pi_t(\cdot\mid s) - e_{\pi(s)}, B_t(s,\cdot)\right\rangle\right]}_{\mathsf{Bon}^\pi(s)}\right), \quad (18)$$

where the second equality uses the performance-difference identity and the fact that $\langle \pi^1(\cdot\mid s) - \pi^2(\cdot\mid s), \mathbf{1}\rangle = 0$ for any pair of policies π^1 and π^2. As in prior works on PO (Luo et al., 2021; Dann et al., 2023b), our goal is to show that the local learning term $\mathsf{Loc}^\pi(s) \lesssim \mathbb{E}[\sum_{t=1}^T (b_t(s) + \overline{B}_t(s)/H)]$. This will, in turn, guarantee that $\sum_s \mu^\pi(s)(\mathsf{Loc}^\pi(s) + \mathsf{Bon}^\pi(s)) \lesssim V^{\pi_t}(s_0;b_t)$, which can then be controlled. To this end, we decompose

$$\mathsf{Loc}^\pi(s) = \underbrace{\mathbb{E}\left[\sum_{t=1}^T\left\langle \pi_t(\cdot\mid s) - e_{\pi(s)}, g_t(s,\cdot)\right\rangle\right]}_{\mathsf{FTRL\text{-}Regret}^\pi(s)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T\left\langle \pi_t(\cdot\mid s) - e_{\pi(s)}, \tau_t(s,\cdot) + d_t(s,\cdot)\right\rangle\right]}_{\mathsf{Clipping\text{-}Bias}^\pi(s)}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^T\left\langle \pi_t(\cdot\mid s) - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle\right]}_{\mathsf{Estimator\text{-}Bias}^\pi(s)}. \quad (19)$$

Bounding $\mathsf{FTRL\text{-}Regret}^{\pi}(s)$ Lemma D.7, together with Lemma C.11, shows that

$$\mathsf{FTRL\text{-}Regret}^{\pi}(s) \lesssim \mathbb{E}\left[\sum_{t=1}^{T}\left(\beta_t(s)-\beta_{t-1}(s)\right)h_t(s)+\sum_{t=1}^{T}\left(\beta_t^{1-\varepsilon}\Sigma_t(s)z_t(s)+\nu_t(s)+\bar{B}_t(s)/H\right)\right]. \tag{20}$$

Bounding $\mathsf{Clipping\text{-}Bias}^{\pi}(s)$ Lemma C.8 shows that $|\tau_t(s,a)| \le d_t(s,a)$. Therefore,

$$\mathsf{Clipping\text{-}Bias}^{\pi}(s) \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq a^\star}\pi_t(a\mid s)d_t(s,a)\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\Sigma_t(s)\beta_t^{1-\varepsilon}(s)z_t(s)\right]. \tag{21}$$

Substituting (20) and (21) into (19) and using the definition of b_t in (10) shows that

$$\mathsf{Loc}^{\pi}(s) \lesssim \mathbb{E}\left[\sum_{t=1}^{T}\left(b_t(s)+\frac{1}{H}\bar{B}_t(s)\right)\right]+\mathsf{Estimator\text{-}Bias}^{\pi}(s). \tag{22}$$

Substituting this bound into (18) leads to

$$\begin{aligned}
\mathrm{Reg}_T(\pi) &\lesssim \mathbb{E}\left[\sum_{t,s}\mu^{\pi}(s)\left(b_t(s)+\kappa_H\bar{B}_t(s)-B_t(s,\pi(s))\right)\right]+\sum_{s\in\mathcal{S}}\mu^{\pi}(s)\mathsf{Estimator\text{-}Bias}^{\pi}(s)\\
&\le \kappa_H\mathbb{E}\left[\sum_t\left(\sum_{s\in\mathcal{S}}\mu^{\pi}(s)\bar{B}_t(s)-\sum_{s\in\mathcal{S}\setminus\mathcal{S}_0}\mu^{\pi}(s)\bar{B}_t(s)\right)\right]+\sum_{s\in\mathcal{S}}\mu^{\pi}(s)\mathsf{Estimator\text{-}Bias}^{\pi}(s)\\
&\le e\mathbb{E}\left[\sum_t V^{P_t^B,\pi_t}(s_0;b_t)\right]+\sum_{s\in\mathcal{S}}\mu^{\pi}(s)\mathsf{Estimator\text{-}Bias}^{\pi}(s).
\end{aligned} \tag{23}$$

Bounding $\mathsf{Estimator\text{-}Bias}^{\pi}(s)$ To bound $\mathsf{Estimator\text{-}Bias}^{\pi}(s)$, we introduce an auxiliary loss function $\mathcal{L}_t^{\mathrm{bias}}(s,a) \coloneqq r_t(s,a)+C_t(s,a)+c_t(s)$, whose Bellman residual can be shown to be nonnegative: $\zeta_t(s,a) \coloneqq \mathcal{L}_t^{\mathrm{bias}}(s,a)-\mathbb{E}_{P,\pi_t}[\mathcal{L}_t^{\mathrm{bias}}(s',a')\mid s,a]\ge 0$ on $\mathcal{E}$ (see Lemma C.14). Therefore,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}}\mu^{\pi}(s)\mathsf{Estimator\text{-}Bias}^{\pi}(s)=\sum_{s\in\mathcal{S}}\mu^{\pi}(s)\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\pi_t(\cdot\mid s)-e_{\pi(s)},\mathcal{L}_t^{\mathrm{bias}}(s,a)\right\rangle\right]\\
&=\sum_{t,s,a}\left(\mu^{\pi_t}(s,a)-\mu^{\pi^*}(s,a)\right)\zeta_t(s,a)\lesssim\sigma^{1/\varepsilon}H^{3/2}S\log T\cdot\sqrt{A\mathsf{D}_T(\pi)},
\end{aligned} \tag{24}$$

where the inequality follows from Lemma C.13.

Bounding $V^{P_t^B,\pi_t}(s_0;b_t)$ Lemma C.17 shows that $\mathbb{E}\left[\sum_t V^{P_t^B,\pi_t}(s_0;\nu_t)\right] \lesssim_{\varepsilon} \mathrm{poly}(\sigma,H,S,A,\log T)$. Meanwhile, Lemma C.16 shows that

$$\mathbb{E}\left[\sum_t V^{P_t^B,\pi_t}(s_0;b_t-\nu_t)\right]\lesssim H\sum_{s,a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha}(\log T)^{1-\alpha}. \tag{25}$$

Substituting (24) and (25) into (23) concludes the proof.

□

5 Conclusion

In this work, we develop a BOBW PO-based method for heavy-tailed MDPs, improving the results of previous OM-based methods by Chen et al. (2026) while eliminating their truncated nonnegativity assumption in the unknown-transition case and avoiding the need to solve a complicated global constrained optimization problem. One promising direction is to extend our algorithm and results to the function-approximation setting, which we leave for future work.

AI USE STATEMENT

The authors have developed the main ideas and guided the overall direction of this work. Generative AI tools have assisted in developing some proofs from ideas provided by the authors, helped write parts of the proofs, and improved the presentation of the paper. We have not used generative AI tools for any other tasks that require disclosure. We have checked and verified all proof derivations and text generated by LLMs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on developing and analyzing the BOBW algorithm for MDPs with heavy-tailed losses. The algorithm and the results are purely theoretical. We are not aware of any potential negative impacts or ethical concerns arising from our work.

REPRODUCIBILITY STATEMENT

We have stated all assumptions and notations explicitly, described every algorithmic step in detail, and provided complete pseudocode. We have stated and explained all theoretical results, together with their conditions and parameter dependencies, in the main text. The appendices provide detailed proofs and the supporting technical lemmas.

REFERENCES

Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pp. 26–62. PMLR, 2021.

Kumar Ashutosh, Jayakrishnan Nair, Anmol Kagrecha, and Krishna Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 622–630. PMLR, 2021.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 116–120. PMLR, 2016.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Cheng Chen, Canzhe Zhao, and Shuai Li. Simultaneously learning stochastic and adversarial bandits under the position-based model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 6, 2022.

Yu Chen, Jiatai Huang, Yan Dai, and Longbo Huang. uniinf: Best-of-both-worlds algorithm for parameter-free heavy-tailed mabs. In *International Conference on Learning Representations*, volume 2025, pp. 71762–71798, 2025.

Yu Chen, Yuhao Liu, Jiatai Huang, Yihan Du, and Longbo Huang. Best-of-both-worlds for heavy-tailed markov decision processes. *arXiv preprint arXiv:2602.01295*, 2026.

Duo Cheng, Xingyu Zhou, and Bo Ji. Taming heavy-tailed losses in adversarial bandits and the best-of-both-worlds setting. *Advances in Neural Information Processing Systems*, 37:48083–48129, 2024.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, Proceedings of Machine Learning Research. PMLR, 2023a.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. Best of both worlds policy optimization. In *International Conference on Machine Learning*, pp. 6968–7008. PMLR, 2023b.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 2021.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ε, u)-adaptive regret minimization in heavy-tailed bandits. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1882–1915. PMLR, 2024.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *international conference on machine learning*, pp. 9173–9200. PMLR, 2022.

Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. *Advances in Neural Information Processing Systems*, 36:56576–56588, 2023.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, Proceedings of Machine Learning Research. PMLR, 2021a.

Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021b.

Shinji Ito and Kei Takemura. Best-of-three-worlds linear bandit algorithm with variance-adaptive regret bounds. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, Proceedings of Machine Learning Research. PMLR, 2023a.

Shinji Ito and Kei Takemura. An exploration-by-optimization approach to best of both worlds in linear bandits. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2522–2563. PMLR, 2024a.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, 2024b.

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*, pp. 4860–4869. PMLR, 2020.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33:16557–16566, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34: 20491–20502, 2021.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: Ftrl with general regularizers and multiple optimal arms. *Advances in Neural Information Processing Systems*, 36:30918–30978, 2023.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna P. Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11269–11278, 2019.

Minhyun Kang and Gi-Soo Kim. Heavy-tailed linear bandit with huber regression. In *Uncertainty in Artificial Intelligence*, pp. 1027–1036. PMLR, 2023.

Fang Kong, Yichi Zhou, and Shuai Li. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research. PMLR, 2022.

Fang Kong, Canzhe Zhao, and Shuai Li. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, Proceedings of Machine Learning Research. PMLR, 2023.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pp. 6142–6151. PMLR, 2021.

Kyungjae Lee and Sungbin Lim. Minimax optimal bandits for heavy tail rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5280–5294, 2022.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.

Xiang Li and Qiang Sun. Variance-aware decision making with linear function approximation under heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *international conference on machine learning*, pp. 4154–4163. PMLR, 2019.

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.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pp. 114–122, 2018.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. *CoRR*, 2022.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pp. 1287–1295. PMLR, 2014.

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Artin Tajdini, Jonathan Scarlett, and Kevin Jamieson. Improved regret bounds for linear bandits with heavy-tailed rewards. *arXiv preprint arXiv:2506.04775*, 2025.

Taira Tsuchiya and Shinji Ito. A simple and adaptive learning rate for ftrl in online learning with minimax regret of $\Theta(T^{2/3})$ and its application to best-of-both-worlds. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, Proceedings of Machine Learning Research. PMLR, 2023.

Sattar Vakili, Keqin Liu, and Qing Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5): 759–767, 2013.

Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-tailed linear bandits: Huber regression with one-pass update. *arXiv preprint arXiv:2503.00419*, 2025.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2020.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2936–2942. ijcai.org, 2020.

Canzhe Zhao, Shinji Ito, and Shuai Li. Heavy-tailed linear bandits: Adversarial robustness, best-of-both-worlds, and beyond. *arXiv preprint arXiv:2508.13679*, 2025.

Han Zhong, Jiayi Huang, Lin Yang, and Liwei Wang. Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 34:15710–15720, 2021.

Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 3385–3393. PMLR, 2021.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pp. 7683–7692. PMLR, 2019.

CONTENTS

A Additional Literature Review **15**

B Proofs for the Known-transition Case **16**

B.1 Proof of the Main Results 16

B.2 Bounding FTRL-Regret$^{\pi}(s)$ 22

B.3 Bounding Clipping-Bias$^{\pi}(s)$ 30

B.4 Bounding Estimator-Bias$^{\pi}(s)$ 31

B.5 Bounding $V^{\pi_t}(s_0; b_t)$ 33

C Proofs for the Unknown-transition Case **39**

C.1 Proof of the Main Results 42

C.2 Bounding FTRL-Regret$^{\pi}(s)$ 54

C.3 Bounding Clipping-Bias$^{\pi}(s)$ 58

C.4 Bounding Estimator-Bias$^{\pi}(s)$ 59

C.5 Bounding $V^{P_t^B, \pi_t}(s_0; b_t)$ 65

D Auxiliary Lemmas **71**

D.1 Value-difference and Performance-difference Identities 71

D.2 Dilated-bonus Reduction 72

D.3 Tsallis Stability and FTRL on the Full Simplex 74

D.4 Transition Concentration and Occupancy Comparison 78

A Additional Literature Review

Best-of-both-worlds Online Learning The BOBW objective is to design a single algorithm that guarantees worst-case robustness in adversarial regimes while achieving fast gap-dependent regret in stochastic regimes. Early BOBW bandit algorithms follow a detect-and-switch paradigm: they use statistical tests to distinguish the two regimes and then invoke an algorithm specialized to the detected regime (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017). A second line avoids explicit regime detection and derives adaptive guarantees from a single FTRL procedure, often through self-bounding analyses (Wei & Luo, 2018; Zimmert & Seldin, 2021; Ito, 2021a; Jin et al., 2023). Such analyses can also provide refined guarantees in intermediate regimes, including stochastic regimes with adversarial corruptions (Lykouris et al., 2018; Zimmert & Seldin, 2021).

The BOBW principle also extends to structured feedback models, including combinatorial semi-bandits (Zimmert et al., 2019; Ito, 2021b; Tsuchiya et al., 2023), linear bandits (Lee et al., 2021; Kong et al., 2023; Ito & Takemura, 2023a; Dann et al., 2023a; Ito & Takemura, 2023b; Ito et al., 2024a), online learning to rank (Chen et al., 2022), and bandits with graph feedback (Erez & Koren, 2021; Ito et al., 2022; Rouyer et al., 2022; Kong et al., 2022; Dann et al., 2023a; Ito et al., 2024b; Tsuchiya & Ito, 2024). For episodic MDPs with bounded losses, Jin & Luo (2020) and Jin et al. (2021) establish BOBW guarantees for known and unknown transitions, respectively, using OM-based methods. Subsequently, Dann et al. (2023b) develop PO-based BOBW algorithms using the dilated bonuses proposed by Luo et al. (2021) to propagate state-wise regret bounds through the MDP.

Online Learning with Heavy-tailed Losses Bubeck et al. (2013) initiate the study of stochastic MABs under heavy-tailed conditions, introducing robust estimators and establishing worst-case and gap-dependent upper and lower bounds. Subsequent work sharpens minimax and instance-dependent guarantees and develops algorithms with different robustness and adaptivity properties (Vakili et al., 2013; Lee et al., 2020; Wei & Srivastava, 2020; Agrawal et al., 2021; Ashutosh et al., 2021; Lee & Lim, 2022; Genalti et al., 2024). Furthermore, related works explore risk-aware bandits with unbounded rewards (Kagrecha et al., 2019) and Lipschitz bandits with heavy-tailed feedback (Lu et al., 2019). For structured action spaces, Medina & Yang (2016) first study stochastic heavy-tailed linear bandits using truncation and median-of-means estimators. Later work improves the regret guarantees, considers finite arm sets and weaker moment conditions, and develops regression-based approaches (Shao et al., 2018; Xue et al., 2020; Zhong et al., 2021; Kang & Kim, 2023; Wang et al., 2025; Tajdini et al., 2025). Heavy-tailed reinforcement learning has also been studied in tabular MDPs (Zhuang & Sui, 2021) and under linear function approximation (Huang et al., 2023; Li & Sun, 2024).

The adversarial and BOBW line for heavy-tailed MABs begins with Huang et al. (2022), who obtain nearly optimal guarantees but require the truncated non-negativity assumption. Cheng et al. (2024) use FTRL with a log-barrier regularizer to develop adversarial and detect-and-switch-based BOBW algorithms that do not rely on the truncated non-negativity assumption. Besides, Chen et al. (2025) develop a parameter-free BOBW algorithm for heavy-tailed MABs without knowing the parameters of the heavy-tailed distribution (ε, σ), whereas the truncated non-negativity assumption is still required. More recently, Zhao et al. (2025) introduce bonus-shifted FTRL, obtain a purely FTRL-based BOBW heavy-tailed MAB algorithm without the truncated non-negativity assumption, and extend the framework to adversarial and BOBW heavy-tailed linear bandits. Closest to the present work, Chen et al. (2026) provide the first BOBW guarantees for heavy-tailed MDPs using OM-based FTRL, whereas the truncated non-negativity assumption is needed in the unknown-transition case.

B Proofs for the Known-transition Case

Let $c_{\mathrm{cnt},\varepsilon} \coloneqq 3(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}) = \mathcal{O}_\varepsilon(1)$. Recall from Section 2 that $L_\sigma = \sigma^{1/\varepsilon}$, and from Section 3 that $K_{\mathrm{saf}} = \frac{2^{12}\varepsilon^2}{\varepsilon-1} + 2^{26-12\varepsilon} + 1$, and $\Gamma = \lceil e^\varepsilon K_{\mathrm{saf}}^\varepsilon H^{2\varepsilon} A^{\varepsilon-1}\rceil$. For ease of presentation, we define the following shorthand notation. Let $\mathsf{G}_T \coloneqq \Gamma\big(1+\log(eT)\big)$ and $\Lambda_T \coloneqq 1+\varepsilon\log(eT)$. Then $\sum_{t=1}^T \gamma_t \le \mathsf{G}_T$. Let $\mathcal{L}_T \coloneqq \log\big(16HS^2A\overline{T}^{66}\big)(2+\log(eT))$, where we recall from Section 3 that $\overline{T} = \max\{2,T\}$. Define

$$\mathsf{R}_{\mathrm{low}} \coloneqq HA^{1-\alpha}(\beta_0 + 2^\alpha c_\beta HL_\sigma)/\alpha + 20SHL_\sigma\mathsf{G}_T = \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}\Big[H^4A^{2-\frac{2}{\varepsilon}} + SH^{2\varepsilon+1}A^{\varepsilon-1}\log T\Big]\Big)\,.$$

Theorem B.1 (Formal Version of Theorem 4.1 for Known Transitions)**.** *Under Definition 2.1, Algorithm 1 satisfies, for every deterministic policy π,*

$$\mathrm{Reg}_T(\pi) \le c_{\mathrm{cnt},\varepsilon}HL_\sigma\Lambda_T^{1-\alpha}\sum_{s\in\mathcal{S}\setminus\mathcal{S}_H}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^\alpha + \mathsf{R}_{\mathrm{low}}\,. \tag{26}$$

For the self-bounding result under Definition 2.2, recall from Section 1 that

$$\omega_{1/\varepsilon}(\Delta) = \sum_{s\in\mathcal{S}\setminus\mathcal{S}_H}\sum_{a\neq\pi^\star(s)}\Delta(s,a)^{-\frac{1}{\varepsilon-1}}\,. \tag{27}$$

Define

$$\mathsf{A}_T \coloneqq c_{\mathrm{cnt},\varepsilon}HL_\sigma\Lambda_T^{1-\alpha}\,, \tag{28}$$

$$\mathcal{U} \coloneqq \mathsf{A}_T^{\frac{1}{1-\alpha}}\omega_{1/\varepsilon}(\Delta) = \mathsf{A}_T^{\frac{\varepsilon}{\varepsilon-1}}\omega_{1/\varepsilon}(\Delta)\,, \tag{29}$$

$$\mathsf{R}_{\mathrm{infl}} \coloneqq \mathsf{A}_T(SA)^{1-\alpha}(S\mathsf{G}_T)^\alpha = \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}H^3SA^{2-\frac{2}{\varepsilon}}\log T\Big)\,, \tag{30}$$

$$\overline{\mathsf{R}}_{\mathrm{low}} \coloneqq \mathsf{R}_{\mathrm{low}} + \mathsf{R}_{\mathrm{infl}} = \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}H^3S\Big[A^{2-\frac{2}{\varepsilon}} + H^{2\varepsilon-2}A^{\varepsilon-1}\Big]\log T\Big)\,.$$

Theorem B.2 (Formal Version of Theorem 4.3 for Known Transitions)**.** *Under Definition 2.1 and known transitions, Algorithm 1 satisfies, for every adaptive adversarial loss sequence,*

$$\max_\pi \mathrm{Reg}_T(\pi) = \mathcal{O}_\varepsilon\left(\sigma^{\frac{1}{\varepsilon}}H^{1+\frac{1}{\varepsilon}}(SA)^{1-\frac{1}{\varepsilon}}T^{\frac{1}{\varepsilon}}\Lambda_T^{1-\frac{1}{\varepsilon}}\right) + \overline{\mathsf{R}}_{\mathrm{low}}\,. \tag{31}$$

Suppose the environment satisfies the $(\pi^\star,\Delta,\mathcal{C})$ self-bounding condition in Definition 2.2. Then

$$\mathrm{Reg}_T(\pi^\star) = \mathcal{O}_\varepsilon\left(\mathcal{U} + \mathcal{U}^{1-\frac{1}{\varepsilon}}\mathcal{C}^{\frac{1}{\varepsilon}} + \overline{\mathsf{R}}_{\mathrm{low}}\right)\,. \tag{32}$$

B.1 Proof of the Main Results

B.1.1 Proof of the Master Bound

Proof of Theorem B.1. Fix an arbitrary deterministic comparator policy π. All state sums below range over nonterminal states, and all occupancies use the true kernel P. Recall $p_{t,s} = \pi_t(\cdot \mid s)$ and write $B_t^\pi(s) \coloneqq B_t(s,\pi(s))$, $\overline{B}_t(s) \coloneqq \langle p_{t,s}, B_t(s,\cdot)\rangle$, and $\kappa_H \coloneqq 1+1/H$. We start with the regret decomposition:

$$\begin{aligned}\mathrm{Reg}_T(\pi) &= \mathbb{E}\left[\sum_{t=1}^T\big(V_t^{\pi_t}(s_0) - V_t^\pi(s_0)\big)\right] = \mathbb{E}\left[\sum_{t,s}\mu^\pi(s)\,\big\langle p_{t,s} - e_{\pi(s)}, \bar{Q}_t(s,\cdot)\big\rangle\right]\\ &= \sum_s \mu^\pi(s)\big(\mathsf{Loc}^\pi(s) + \mathsf{Bon}^\pi(s)\big)\,,\end{aligned} \tag{33}$$

where

$$\mathsf{Loc}^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^T\big\langle p_{t,s} - e_{\pi(s)}, \bar{Q}_t(s,\cdot) - B_t(s,\cdot)\big\rangle\right]\,,$$

$$\mathsf{Bon}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, B_t(s,\cdot)\right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\overline{B}_t(s) - B_t^{\pi}(s)\right)\right].$$

The second equality in (33) is the performance-difference identity: centering $Q_t^{\pi_t}(s,\cdot)$ by $V_t^{\pi_t}(s)\mathbf{1}$ leaves the inner product unchanged because $\left\langle p_{t,s} - e_{\pi(s)}, \mathbf{1}\right\rangle = 0$. Our goal is to bound $\mathsf{Loc}^{\pi}(s)$ by the local bonus $b_t(s) + \overline{B}_t(s)/H$, an initial regularizer term, and the estimator bias. The bonus recursion will then cancel the comparator-weighted contribution of $\mathsf{Bon}^{\pi}(s)$.

Recall the estimator-mean discrepancy and clipping-bias vector:

$$r_t(s,a) = \bar{Q}_t(s,a) - \mathbb{E}_t[\widetilde{Q}_t(s,a)] = \frac{\gamma_t}{\widetilde{\mu}_t(s)}\bar{Q}_t(s,a)\,, \tag{34}$$

where the last equality follows from (79), and $\tau_t(s,\cdot) = \mathbb{E}_t[\widetilde{Q}_t(s,\cdot) - \widehat{Q}_t(s,\cdot)]$. The definition (14) gives the exact vector identity

$$\bar{Q}_t(s,\cdot) - B_t(s,\cdot) = \mathbb{E}_t[g_t(s,\cdot)] + \tau_t(s,\cdot) + d_t(s,\cdot) + r_t(s,\cdot) + C_t(s,\cdot)\,. \tag{35}$$

Since $p_{t,s} - e_{\pi(s)}$ is $\mathcal{F}_{t-1}$-measurable, taking its inner product with this identity and applying the tower property yields

$$\mathsf{Loc}^{\pi}(s) = \mathsf{FTRL\text{-}Regret}^{\pi}(s) + \mathsf{Clipping\text{-}Bias}^{\pi}(s) + \mathsf{Estimator\text{-}Bias}^{\pi}(s)\,, \tag{36}$$

with the same three terms as in the proof sketch:

$$\mathsf{FTRL\text{-}Regret}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, g_t(s,\cdot)\right\rangle\right],$$

$$\mathsf{Clipping\text{-}Bias}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, \tau_t(s,\cdot) + d_t(s,\cdot)\right\rangle\right],$$

$$\mathsf{Estimator\text{-}Bias}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle\right].$$

Bounding $\mathsf{FTRL\text{-}Regret}^{\pi}(s)$ Fix s and let $\widetilde{p}_{t+1}$ be the lookahead iterate in (464), with the current inverse learning rate $\beta_t(s)$. Define its stability contribution by

$$\mathsf{S}_t(s) \coloneqq \langle p_{t,s} - \widetilde{p}_{t+1}, g_t(s,\cdot)\rangle - \beta_t(s) D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_{t,s})\,.$$

The clock (13) makes $\beta_t(s)$ nondecreasing. Thus Lemma D.7, specifically (466) with comparator $e_{\pi(s)}$, gives the pathwise inequality

$$\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, g_t(s,\cdot)\right\rangle \le \beta_1(s)h_1(s) + \sum_{t=2}^{T}(\beta_t(s) - \beta_{t-1}(s))h_t(s) + \sum_{t=1}^{T}\mathsf{S}_t(s)\,. \tag{37}$$

Taking expectations and applying the conditional stability bound in Lemma B.10 to each $\mathsf{S}_t(s)$ yields

$$\begin{aligned}\mathsf{FTRL\text{-}Regret}^{\pi}(s) \le{}& \mathbb{E}[\beta_1(s)h_1(s)] + \mathbb{E}\left[\sum_{t=2}^{T}(\beta_t(s) - \beta_{t-1}(s))h_t(s)\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^{T}\left[\frac{64}{1-\alpha}k_\alpha^{2-\varepsilon}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s) + \nu_t(s) + \frac{\overline{B}_t(s)}{H}\right]\right].\end{aligned} \tag{38}$$

Bounding $\mathsf{Clipping\text{-}Bias}^{\pi}(s)$ Lemma B.11 gives $|\tau_t(s,a)| \le d_t(s,a)$, hence $0 \le \tau_t(s,a) + d_t(s,a) \le 2d_t(s,a)$. For $a^\star = \pi(s)$, the comparator coordinate has coefficient $p_{t,s,a^\star} - 1 \le 0$, so

$$\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \le 2\sum_{a\neq a^\star} p_{t,s,a}d_t(s,a) \le 4k_\alpha^{1-\varepsilon}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s)\,.$$

The last inequality is (126), obtained by substituting the clipping thresholds and compensation d_t. Consequently,

$$\mathsf{Clipping\text{-}Bias}^{\pi}(s) \le 4k_\alpha^{1-\varepsilon}\mathbb{E}\left[\sum_{t=1}^{T}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s)\right]. \tag{39}$$

The heavy-tail coefficients in (38) and (39) obey

$$\frac{64}{1-\alpha}k_\alpha^{2-\varepsilon} + 4k_\alpha^{1-\varepsilon} = 5\cdot 64^{\varepsilon-1}(1-\alpha)^{1-\varepsilon} \le 320(1-\alpha)^{1-\varepsilon} \le K_{\mathrm{ht}}\,, \tag{40}$$

using $k_\alpha = (1-\alpha)/64$, $1 < \varepsilon \le 2$, and $K_{\mathrm{ht}} = 2^{14}(1-\alpha)^{1-\varepsilon}$. Thus the changing-rate term, both heavy-tail terms, and $\nu_t(s)$ are bounded by $b_t(s)$ as defined in (10). Substituting these bounds into (36) proves the one-state inequality

$$\mathsf{Loc}^\pi(s) \le \mathbb{E}[\beta_1(s)h_1(s)] + \mathbb{E}\left[\sum_{t=1}^{T}\left(b_t(s) + \frac{\overline{B}_t(s)}{H}\right)\right] + \mathsf{Estimator\text{-}Bias}^\pi(s)\,. \tag{41}$$

The same argument holds for every comparator action $a^\star$.

We now substitute (41) into (33) and combine its $\overline{B}_t(s)/H$ term with $\mathsf{Bon}^\pi(s)$:

$$\mathrm{Reg}_T(\pi) \le \mathcal{I}_0(\pi) + \mathbb{E}\left[\sum_{t=1}^{T}\mathcal{D}_t(\pi)\right] + \mathsf{S}(\pi)\,, \tag{42}$$

where $\mathcal{I}_0(\pi) := \mathbb{E}\left[\sum_s \mu^\pi(s)\beta_1(s)h_1(s)\right]$, $\mathsf{S}(\pi) := \sum_s \mu^\pi(s)\mathsf{Estimator\text{-}Bias}^\pi(s)$, and $\mathcal{D}_t(\pi) := \sum_s \mu^\pi(s)[b_t(s) + \kappa_H\overline{B}_t(s) - B_t^\pi(s)]$. The bonus recursion (12) and the comparator occupancy flow cancel all noninitial terms:

$$\mathcal{D}_t(\pi) = \kappa_H\left[\sum_s \mu^\pi(s)\overline{B}_t(s) - \sum_{s\neq s_0}\mu^\pi(s)\overline{B}_t(s)\right] = \kappa_H\overline{B}_t(s_0)\,.$$

Unrolling the same recursion under π_t, using $b_t \ge 0$ and $\kappa_H^H < e < 3$, gives $\mathcal{D}_t(\pi) \le 3V^{\pi_t}(s_0;b_t)$. These are precisely the cancellation and unrolling steps of Lemma D.3 in Section D.2. Therefore,

$$\mathrm{Reg}_T(\pi) \le \mathcal{I}_0(\pi) + 3\mathbb{E}\left[\sum_{t=1}^{T}V^{\pi_t}(s_0;b_t)\right] + \mathsf{S}(\pi)\,. \tag{43}$$

Bounding $\mathsf{Estimator\text{-}Bias}^\pi(s)$ As in the proof sketch, introduce the auxiliary loss $\mathcal{L}_t^{\mathrm{bias}}(s,a) = r_t(s,a)+C_t(s,a)+c_t(s)$ and its Bellman residual $\zeta_t(s,a) = \mathcal{L}_t^{\mathrm{bias}}(s,a) - \mathbb{E}_{P,\pi_t}[\mathcal{L}_t^{\mathrm{bias}}(s',a') \mid s,a]$. Lemma B.12 proves $\zeta_t \ge 0$ and identifies $\mathcal{L}_t^{\mathrm{bias}}$ as the Q-function of π_t for loss ζ_t. Adding the action-independent $c_t(s)$ to the inner product has no effect. The performance-difference identity and (136) therefore give

$$\begin{aligned}\mathsf{S}(\pi) &= \mathbb{E}\left[\sum_{t,s}\mu^\pi(s)\left\langle p_{t,s} - e_{\pi(s)}, \mathcal{L}_t^{\mathrm{bias}}(s,\cdot)\right\rangle\right] = \mathbb{E}\left[\sum_{t,s,a}\left(\mu^{\pi_t}(s,a) - \mu^\pi(s,a)\right)\zeta_t(s,a)\right]\\ &\le 4\mathbb{E}\left[\sum_{t,s}\mu^{\pi_t}(s)c_t(s)\right] \le 8SHL_\sigma\sum_{t=1}^{T}\gamma_t\,.\end{aligned} \tag{44}$$

Substitution into (43) gives

$$\mathrm{Reg}_T(\pi) \le \mathcal{I}_0(\pi) + 8SHL_\sigma\sum_{t=1}^{T}\gamma_t + 3\mathbb{E}\left[\sum_{t=1}^{T}V^{\pi_t}(s_0;b_t)\right]\,. \tag{45}$$

Bounding $V^{\pi_t}(s_0;b_t)$ The two components of this scalar value are controlled separately. Lemma B.17 gives $\mathbb{E}\left[\sum_t V^{\pi_t}(s_0;\nu_t)\right] \le 4SHL_\sigma\mathsf{G}_T$. For $b_t - \nu_t$, the comparator-removal and clock estimates in the proof of Proposition B.13, specifically (168), give

$$\mathbb{E}\left[\sum_{t=1}^{T}V^{\pi_t}(s_0;b_t-\nu_t)\right] \le \left(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\right)HL_\sigma\Lambda_T^{1-\alpha}\sum_s\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^\alpha\,.$$

Adding the two bounds is (157). Using $\sum_t\gamma_t \le \mathsf{G}_T$ and $c_{\mathrm{cnt},\varepsilon} = 3(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon})$ in (45) now yields

$$\mathrm{Reg}_T(\pi) \le c_{\mathrm{cnt},\varepsilon}HL_\sigma\Lambda_T^{1-\alpha}\sum_s\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^\alpha + \mathcal{I}_0(\pi) + 20SHL_\sigma\mathsf{G}_T\,. \tag{46}$$

It remains to bound the initial regularizer term retained above. Since $\gamma_1 = 1$, we have $1 \le \widetilde{\mu}_1(s) \le 2$ and $G_1(s) = 1 + \widetilde{\mu}_1(s)^{1-\varepsilon} \le 2$, so $\beta_1(s) \le \beta_0 + 2^\alpha c_\beta H L_\sigma$. Concavity gives $\sum_a p_{1,s,a}^\alpha \le A^{1-\alpha}$ and hence $h_1(s) \le A^{1-\alpha}/\alpha$. Using $\sum_s \mu^\pi(s) = H$, we obtain

$$\mathcal{I}_0(\pi) \le \frac{HA^{1-\alpha}}{\alpha}\big(\beta_0 + 2^\alpha c_\beta H L_\sigma\big)\,.$$

Inserting this into (46) and using the definition of $\mathsf{R}_{\mathrm{low}}$ proves (26). □

B.1.2 Adversarial Specialization

Proof of (31). Fix a deterministic comparator policy π. For every realization of the algorithm, the total inflated state-action mass is

$$\begin{aligned}\sum_{s,a}\sum_{t=1}^{T}\widetilde{\mu}_t(s)p_{t,s,a} &= \sum_{t=1}^{T}\sum_{s}\widetilde{\mu}_t(s)\sum_{a}p_{t,s,a}\\ &= \sum_{t=1}^{T}\sum_{s}\big(\mu^{\pi_t}(s)+\gamma_t\big)\\ &= \sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s) + S\sum_{t=1}^{T}\gamma_t = HT + S\sum_{t=1}^{T}\gamma_t\\ &\le HT + S\mathsf{G}_T\,. \end{aligned} \tag{47}$$

Here the equalities use $\sum_a p_{t,s,a} = 1$, the definition of $\widetilde{\mu}_t(s)$ in Section 3, and the occupancy identity $\sum_s \mu^{\pi_t}(s) = H$ for the layered model in Section 2; the inequality uses $\sum_t \gamma_t \le \mathsf{G}_T$ from Section 4.

According to the definition of $\widetilde{n}_T(s,a)$ in Section 4, taking expectations in (47) gives

$$\sum_{s,a}\widetilde{n}_T(s,a) = \mathbb{E}\left[\sum_{s,a}\sum_{t=1}^{T}\widetilde{\mu}_t(s)p_{t,s,a}\right] \le HT + S\mathsf{G}_T\,. \tag{48}$$

Let $\mathcal{I}_\pi := \{(s,a) : a \ne \pi(s)\}$. The set $\mathcal{I}_\pi$ has at most SA elements. Hölder's inequality with exponents $1/\alpha$ and $1/(1-\alpha)$ therefore yields

$$\begin{aligned}\sum_{s}\sum_{a\ne\pi(s)}\widetilde{n}_T(s,a)^\alpha &= \sum_{(s,a)\in\mathcal{I}_\pi}\widetilde{n}_T(s,a)^\alpha\cdot 1 \le \left(\sum_{(s,a)\in\mathcal{I}_\pi}\widetilde{n}_T(s,a)\right)^{\alpha}|\mathcal{I}_\pi|^{1-\alpha}\\ &\le (SA)^{1-\alpha}\left(\sum_{s,a}\widetilde{n}_T(s,a)\right)^{\alpha}\\ &\le (SA)^{1-\alpha}(HT+S\mathsf{G}_T)^\alpha\,. \end{aligned} \tag{49}$$

The penultimate inequality uses nonnegative counts and $|\mathcal{I}_\pi| \le SA$, and the last applies (48).

Substituting (49) into the complete master bound (26) gives

$$\mathrm{Reg}_T(\pi) \le c_{\mathrm{cnt},\varepsilon} H L_\sigma \Lambda_T^{1-\alpha}(SA)^{1-\alpha}(HT+S\mathsf{G}_T)^\alpha + \mathsf{R}_{\mathrm{low}}\,. \tag{50}$$

The right-hand side of (50) is independent of π. Taking the maximum over deterministic comparators gives the first inequality in (31).

To retain all additive terms, use $0 < \alpha < 1$ and subadditivity:

$$\begin{aligned}\mathsf{A}_T(SA)^{1-\alpha}(HT+S\mathsf{G}_T)^\alpha + \mathsf{R}_{\mathrm{low}} &\le \mathsf{A}_T(SA)^{1-\alpha}(HT)^\alpha + \mathsf{A}_T(SA)^{1-\alpha}(S\mathsf{G}_T)^\alpha + \mathsf{R}_{\mathrm{low}}\\ &= \mathsf{A}_T(SA)^{1-\alpha}(HT)^\alpha + \overline{\mathsf{R}}_{\mathrm{low}}\,.\end{aligned}$$

The second equality is exactly the definition of $\mathsf{R}_{\mathrm{infl}}$ and $\overline{\mathsf{R}}_{\mathrm{low}}$ in (30). Substituting $\mathsf{A}_T = c_{\mathrm{cnt},\varepsilon} H L_\sigma \Lambda_T^{1-\alpha}$ and $\alpha = 1/\varepsilon$ proves (31). In particular, the inflated-count contribution cannot simply be dropped on the grounds that it is logarithmic in T. □

B.1.3 SELF-BOUNDING SPECIALIZATION

We use two elementary inequalities.

Lemma B.3 (Gap-weighted Hölder inequality)**.** *Let $n_i \geq 0$ and $\Delta_i > 0$. Then*

$$\sum_i n_i^{\frac{1}{\varepsilon}} \leq \left(\sum_i \Delta_i n_i\right)^{\frac{1}{\varepsilon}} \left(\sum_i \Delta_i^{-\frac{1}{\varepsilon-1}}\right)^{\frac{\varepsilon-1}{\varepsilon}} . \tag{51}$$

Proof. For every index i, insert the positive gap Δ_i without applying an inequality:

$$n_i^{\frac{1}{\varepsilon}} = (\Delta_i n_i)^{\frac{1}{\varepsilon}} \Delta_i^{-\frac{1}{\varepsilon}} . \tag{52}$$

Apply Hölder's inequality to the two nonnegative sequences

$$a_i := (\Delta_i n_i)^{\frac{1}{\varepsilon}} , \qquad b_i := \Delta_i^{-\frac{1}{\varepsilon}} , \tag{53}$$

with conjugate exponents ε and $\varepsilon/(\varepsilon-1)$. Substituting (52)– (53) gives the full calculation

$$\begin{aligned}
\sum_i n_i^{\frac{1}{\varepsilon}} &= \sum_i (\Delta_i n_i)^{\frac{1}{\varepsilon}} \Delta_i^{-\frac{1}{\varepsilon}} \\
&\leq \left(\sum_i \left[(\Delta_i n_i)^{\frac{1}{\varepsilon}}\right]^{\varepsilon}\right)^{\frac{1}{\varepsilon}} \left(\sum_i \left[\Delta_i^{-\frac{1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}}\right)^{\frac{\varepsilon-1}{\varepsilon}} \\
&= \left(\sum_i \Delta_i n_i\right)^{\frac{1}{\varepsilon}} \left(\sum_i \Delta_i^{-\frac{1}{\varepsilon-1}}\right)^{\frac{\varepsilon-1}{\varepsilon}} .
\end{aligned}$$

The inequality is Hölder's inequality; simplifying the powers gives (51). □

Lemma B.4 (Solving the self-bounding inequality)**.** *Let $\alpha \in (0,1)$ and $x, D, \mathcal{C}, R \geq 0$. If*

$$x \leq D(x + \mathcal{C})^{\alpha} + R , \tag{54}$$

then

$$x \leq c_\alpha \left(D^{\frac{1}{1-\alpha}} + D\mathcal{C}^{\alpha} + R\right) \tag{55}$$

for a constant c_α depending only on α.

Proof. Because $0 < \alpha < 1$, the map $u \mapsto u^\alpha$ is subadditive on $\mathbb{R}_+$. Therefore

$$(x + \mathcal{C})^{\alpha} \leq x^{\alpha} + \mathcal{C}^{\alpha} . \tag{56}$$

We next derive the precise Young inequality used to absorb Dx^α. Let $p := \frac{1}{\alpha}$, $q := \frac{1}{1-\alpha}$, $\mathsf{A}_0 := \left(\frac{x}{2}\right)^{\alpha}$, $\mathsf{B}_0 := 2^\alpha D$. Then p and q are conjugate, $\mathsf{A}_0\mathsf{B}_0 = Dx^\alpha$, and Young's inequality $ab \leq a^p/p + b^q/q$ gives

$$Dx^{\alpha} = \mathsf{A}_0\mathsf{B}_0 \leq \frac{\mathsf{A}_0^p}{p} + \frac{\mathsf{B}_0^q}{q} = \frac{\alpha x}{2} + (1-\alpha)2^{\frac{\alpha}{1-\alpha}} D^{\frac{1}{1-\alpha}} \leq \frac{x}{2} + c_\alpha^{(0)} D^{\frac{1}{1-\alpha}} , \tag{57}$$

where $c_\alpha^{(0)} := (1-\alpha)2^{\alpha/(1-\alpha)}$ depends only on α.

Substitute (56) and then (57) into the complete assumed inequality (54):

$$x \leq D(x + \mathcal{C})^{\alpha} + R \leq Dx^{\alpha} + D\mathcal{C}^{\alpha} + R \leq \frac{x}{2} + c_\alpha^{(0)} D^{\frac{1}{1-\alpha}} + D\mathcal{C}^{\alpha} + R . \tag{58}$$

Move the term $x/2$ in (58) to the left and multiply by two:

$$x \leq 2c_\alpha^{(0)} D^{\frac{1}{1-\alpha}} + 2D\mathcal{C}^{\alpha} + 2R \leq c_\alpha \left(D^{\frac{1}{1-\alpha}} + D\mathcal{C}^{\alpha} + R\right) ,$$

where $c_\alpha := 2\max\{c_\alpha^{(0)}, 1\}$ depends only on α. This proves (55). □

Proof of (32). The right-hand side of (32) is nonnegative. If $\mathrm{Reg}_T(\pi^\star) < 0$, the claimed inequality holds immediately. We may therefore assume

$$\mathrm{Reg}_T(\pi^\star) \geq 0\,. \tag{59}$$

The self-bounding condition (2) and nonnegativity of the gap sum also imply $\mathrm{Reg}_T(\pi^\star) + \mathcal{C} \geq 0$, so every fractional power below is well defined.

For every suboptimal pair $a \neq \pi^\star(s)$, define the expected floor count $f_T(s,a) \coloneqq \mathbb{E}\left[\sum_{t=1}^T \gamma_t p_{t,s,a}\right]$. According to the definition of $\widetilde{n}_T(s,a)$ in Section 4, substituting $\widetilde{\mu}_t(s) = \mu^{\pi_t}(s) + \gamma_t$ gives the exact identity

$$\begin{aligned}\widetilde{n}_T(s,a) &= \mathbb{E}\left[\sum_{t=1}^T \big(\mu^{\pi_t}(s) + \gamma_t\big) p_{t,s,a}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \mu^{\pi_t}(s) p_{t,s,a}\right] + \mathbb{E}\left[\sum_{t=1}^T \gamma_t p_{t,s,a}\right] = n_T(s,a) + f_T(s,a)\,.\end{aligned}$$

Since $0 < \alpha < 1$, subadditivity of $x \mapsto x^\alpha$ gives

$$\widetilde{n}_T(s,a)^\alpha = \big(n_T(s,a) + f_T(s,a)\big)^\alpha \leq n_T(s,a)^\alpha + f_T(s,a)^\alpha\,. \tag{60}$$

We next bound the complete sum of floor counts. Using $\sum_a p_{t,s,a} = 1$ and then the bound $\sum_{t=1}^T \gamma_t \leq \mathsf{G}_T$ in Section 4,

$$\begin{aligned}\sum_s \sum_{a \neq \pi^\star(s)} f_T(s,a) &= \sum_s \sum_{a \neq \pi^\star(s)} \mathbb{E}\left[\sum_{t=1}^T \gamma_t p_{t,s,a}\right] \\ &\leq \sum_s \mathbb{E}\left[\sum_{t=1}^T \gamma_t \sum_{a \in \mathcal{A}} p_{t,s,a}\right] = S \sum_{t=1}^T \gamma_t \leq S\mathsf{G}_T\,.\end{aligned} \tag{61}$$

There are at most SA suboptimal state-action pairs. Hölder's inequality therefore gives

$$\sum_s \sum_{a \neq \pi^\star(s)} f_T(s,a)^\alpha \leq (SA)^{1-\alpha} \left(\sum_s \sum_{a \neq \pi^\star(s)} f_T(s,a)\right)^\alpha \tag{62}$$

$$\leq (SA)^{1-\alpha} (S\mathsf{G}_T)^\alpha\,. \tag{63}$$

Inequality (63) substitutes the total floor-count estimate (61) into (62).

Apply the master bound (26) with comparator $\pi^\star$, substitute (60) for every suboptimal pair, and then substitute (63) into the entire floor sum. Using the definitions (28) and (30), the complete calculation is

$$\begin{aligned}\mathrm{Reg}_T(\pi^\star) &\leq \mathsf{A}_T \sum_s \sum_{a \neq \pi^\star(s)} \widetilde{n}_T(s,a)^\alpha + \mathsf{R}_{\mathrm{low}} \\ &\leq \mathsf{A}_T \sum_s \sum_{a \neq \pi^\star(s)} n_T(s,a)^\alpha + \mathsf{A}_T \sum_s \sum_{a \neq \pi^\star(s)} f_T(s,a)^\alpha + \mathsf{R}_{\mathrm{low}} \\ &\leq \mathsf{A}_T \sum_s \sum_{a \neq \pi^\star(s)} n_T(s,a)^\alpha + \mathsf{A}_T (SA)^{1-\alpha} (S\mathsf{G}_T)^\alpha + \mathsf{R}_{\mathrm{low}} \\ &= \mathsf{A}_T \sum_s \sum_{a \neq \pi^\star(s)} n_T(s,a)^\alpha + \overline{\mathsf{R}}_{\mathrm{low}}\,.\end{aligned} \tag{64}$$

The first three lines apply the master bound, split each count using (60), and bound the floor contribution by (63); the last substitutes $\mathsf{R}_{\mathrm{infl}}$ and $\overline{\mathsf{R}}_{\mathrm{low}}$ from (30).

Apply Lemma B.3 to the index set $i = (s,a)$ with $a \neq \pi^\star(s)$. Since $\alpha = 1/\varepsilon$ and $\omega_{1/\varepsilon}(\Delta)$ is defined in (27), the result is

$$\sum_s \sum_{a \neq \pi^\star(s)} n_T(s,a)^\alpha \leq \left(\sum_s \sum_{a \neq \pi^\star(s)} \Delta(s,a) n_T(s,a)\right)^\alpha \left(\sum_s \sum_{a \neq \pi^\star(s)} \Delta(s,a)^{-\frac{1}{\varepsilon - 1}}\right)^{1-\alpha}$$

$$= \omega_{1/\varepsilon}(\Delta)^{1-\alpha} \left(\sum_s \sum_{a \neq \pi^\star(s)} \Delta(s,a) n_T(s,a) \right)^{\alpha} . \tag{65}$$

The gap-weighted actual count is exactly the expected gap sum in the self-bounding condition. Indeed, according to the definition of $n_T(s,a)$ in Section 4 and because $\Delta(s,\pi^\star(s)) = 0$,

$$\begin{aligned} \sum_s \sum_{a \neq \pi^\star(s)} \Delta(s,a) n_T(s,a) &= \sum_s \sum_{a \neq \pi^\star(s)} \Delta(s,a) \mathbb{E}\left[\sum_{t=1}^{T} \mu^{\pi_t}(s) p_{t,s,a} \right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{s,a} \mu^{\pi_t}(s,a) \Delta(s,a) \right] \leq \mathrm{Reg}_T(\pi^\star) + \mathcal{C} . \end{aligned} \tag{66}$$

Line (66) is the rearranged self-bounding condition (2).

Substitute (65) and (66) into the complete count term in (64):

$$\mathrm{Reg}_T(\pi^\star) \leq \mathsf{A}_T \omega_{1/\varepsilon}(\Delta)^{1-\alpha} \big(\mathrm{Reg}_T(\pi^\star) + \mathcal{C} \big)^{\alpha} + \overline{\mathsf{R}}_{\mathrm{low}} . \tag{67}$$

Define

$$D \coloneqq \mathsf{A}_T \omega_{1/\varepsilon}(\Delta)^{1-\alpha} . \tag{68}$$

Equations (59) and (67) verify every hypothesis of Lemma B.4 with $x = \mathrm{Reg}_T(\pi^\star)$ and $R = \overline{\mathsf{R}}_{\mathrm{low}}$. Hence

$$\mathrm{Reg}_T(\pi^\star) \leq c_\alpha \left(D^{\frac{1}{1-\alpha}} + D\mathcal{C}^{\alpha} + \overline{\mathsf{R}}_{\mathrm{low}} \right) . \tag{69}$$

The two powers of D can be identified directly from (68) and (29):

$$D^{\frac{1}{1-\alpha}} = \big(\mathsf{A}_T \omega_{1/\varepsilon}(\Delta)^{1-\alpha} \big)^{\frac{1}{1-\alpha}} = \mathsf{A}_T^{\frac{1}{1-\alpha}} \omega_{1/\varepsilon}(\Delta) = \mathcal{U} , \tag{70}$$

$$D\mathcal{C}^{\alpha} = \mathsf{A}_T \omega_{1/\varepsilon}(\Delta)^{1-\alpha} \mathcal{C}^{\alpha} = \mathcal{U}^{1-\alpha} \mathcal{C}^{\alpha} = \mathcal{U}^{1-\frac{1}{\varepsilon}} \mathcal{C}^{\frac{1}{\varepsilon}} . \tag{71}$$

Substituting (70)– (71) into (69) proves (32), after renaming the constant c_α as c_ε.

□

B.2 Bounding $\mathsf{FTRL\text{-}Regret}^{\pi}(s)$

The next four subsections follow the decomposition in (19). We begin with the moment and centered-kernel estimates shared by the FTRL stability and clipping arguments. We then verify pointwise safety and bound the conditional stability cost.

B.2.1 Heavy-tailed Estimator Properties

Moment of a Suffix Return Let $\mathcal{H}_t^{\mathrm{path}} \coloneqq \sigma((s_{t,h}, a_{t,h})_{h=0}^{H-1}, s_{t,H})$ be the sigma-field generated by the episode-t trajectory, excluding losses.

Lemma B.5 (Heavy-tailed suffix return)**.** *For every layer h,*

$$\mathbb{E}_t \left[|R_{t,h}|^{\varepsilon} \;\middle|\; \mathcal{H}_t^{\mathrm{path}} \right] \leq (H-h)^{\varepsilon} \sigma \leq H^{\varepsilon} \sigma \qquad \textit{almost surely} . \tag{72}$$

Equivalently, the same inequality holds after conditioning on any event in $\mathcal{H}_t^{\mathrm{path}}$ with positive conditional probability.

Proof. Fix an episode t and a layer h, and write $Z_{t,j} \coloneqq Y_t(s_{t,j}, a_{t,j})$ for $j \in \{h, \dots, H-1\}$. Conditional on $\mathcal{F}_{t-1} \vee \mathcal{H}_t^{\mathrm{path}}$, the trajectory indices are fixed and the loss table retains its pre-episode conditional distribution. Thus Definition 2.1 gives

$$\mathbb{E}_t \left[|Z_{t,j}|^{\varepsilon} \mid \mathcal{H}_t^{\mathrm{path}} \right] = \sum_{(u,b)} \mathbb{1}\{(s_{t,j}, a_{t,j}) = (u,b)\} \mathbb{E}_t \left[|Y_t(u,b)|^{\varepsilon} \mid \mathcal{H}_t^{\mathrm{path}} \right]$$

$$= \sum_{(u,b)} \mathbb{1}\{(s_{t,j}, a_{t,j}) = (u,b)\}\mathbb{E}_t[|Y_t(u,b)|^\varepsilon]$$
$$\leq \sigma\,. \tag{73}$$

The first equality partitions over the path-selected coordinate, the second uses conditional independence of the loss table and trajectory given $\mathcal{F}_{t-1}$, and the inequality applies Definition 2.1.

Since $\varepsilon \geq 1$, conditional Minkowski's inequality applies to the finite family $(Z_{t,j})_{j=h}^{H-1}$ and yields the single calculation

$$\left(\mathbb{E}_t\left[|R_{t,h}|^\varepsilon \middle| \mathcal{H}_t^{\text{path}}\right]\right)^{\frac{1}{\varepsilon}} = \left(\mathbb{E}_t\left[\left|\sum_{j=h}^{H-1} Z_{t,j}\right|^\varepsilon \middle| \mathcal{H}_t^{\text{path}}\right]\right)^{\frac{1}{\varepsilon}}$$
$$\leq \sum_{j=h}^{H-1} \left(\mathbb{E}_t\left[|Z_{t,j}|^\varepsilon \mid \mathcal{H}_t^{\text{path}}\right]\right)^{\frac{1}{\varepsilon}}$$
$$\leq \sum_{j=h}^{H-1} \sigma^{\frac{1}{\varepsilon}} = (H-h)\sigma^{\frac{1}{\varepsilon}}\,.$$

The first inequality is conditional Minkowski, which allows dependent loss coordinates, and the second applies (73).

Raising this nonnegative bound to the power ε gives $\mathbb{E}_t\left[|R_{t,h}|^\varepsilon \mid \mathcal{H}_t^{\text{path}}\right] \leq (H-h)^\varepsilon\sigma \leq H^\varepsilon\sigma$, which is (72). If $A \in \mathcal{H}_t^{\text{path}}$ has positive conditional probability, multiplying (72) by $\mathbb{1}\{A\}$ and applying the tower property gives the same bound conditional on A. □

Algebra of the Centered Action Kernel

Lemma B.6 (Kernel identities). *Fix $p \in \operatorname{int}(\Delta(\mathcal{A}))$ and $a \in \mathcal{A}$. Then*

$$\mathsf{D}_p(a) = p_a^{1-\varepsilon}(1-p_a)^\varepsilon + (1-p_a)\,, \tag{74}$$
$$\mathsf{D}_p(a)\bar{p}_a^{\varepsilon-1} \leq 2\,, \tag{75}$$
$$p_a\mathsf{D}_p(a)\bar{p}_a^{(\varepsilon-1)(1-\alpha)} \leq 2\bar{p}_a^\alpha\,. \tag{76}$$

Proof. The centered kernel in (5) satisfies

$$\mathsf{K}_p(a,a) = \frac{1}{p_a} - 1 = \frac{1-p_a}{p_a}\,, \tag{77}$$
$$\mathsf{K}_p(a,b) = -1, \qquad b \neq a\,. \tag{78}$$

Substituting (77)– (78) directly into the definition of $\mathsf{D}_p(a)$ gives

$$\mathsf{D}_p(a) = p_a\left|\frac{1-p_a}{p_a}\right|^\varepsilon + \sum_{b\neq a} p_b|-1|^\varepsilon = p_a^{1-\varepsilon}(1-p_a)^\varepsilon + (1-p_a)\,,$$

This proves (74).

Because $\bar{p}_a \leq p_a$ and $\varepsilon - 1 > 0$, the map $x \mapsto x^{\varepsilon-1}$ is nondecreasing. Hence

$$\mathsf{D}_p(a)\bar{p}_a^{\varepsilon-1} \leq \mathsf{D}_p(a)p_a^{\varepsilon-1} = (1-p_a)^\varepsilon + (1-p_a)p_a^{\varepsilon-1} \leq 1+1 = 2\,.$$

This proves (75).

To prove (76), put $\theta := (\varepsilon-1)(1-\alpha) = \frac{(\varepsilon-1)^2}{\varepsilon}$ and consider the two cases in the definition of $\bar{p}_a$.

Suppose first that $\bar{p}_a = p_a$. Direct substitution of (74) gives

$$p_a\mathsf{D}_p(a)\bar{p}_a^\theta = p_a\left[p_a^{1-\varepsilon}(1-p_a)^\varepsilon + (1-p_a)\right]p_a^\theta$$
$$= p_a^{2-\varepsilon+\theta}(1-p_a)^\varepsilon + p_a^{1+\theta}(1-p_a) = p_a^\alpha(1-p_a)^\varepsilon + p_a^{1+\theta}(1-p_a)$$

$$\leq p_a^\alpha + p_a^\alpha = 2\bar{p}_a^\alpha .$$

Suppose next that $\bar{p}_a < p_a$. By the definition $\bar{p}_a = \min\{p_a, \rho(p)\}$, this case can occur only when a is a maximizer of p, $p_a > 1/2$, and $\bar{p}_a = \rho(p) = 1 - p_a$. Substituting $\bar{p}_a = 1 - p_a$ and (74) gives

$$\begin{aligned} p_a \mathsf{D}_p(a)\bar{p}_a^\theta &= p_a \left[p_a^{1-\varepsilon}(1-p_a)^\varepsilon + (1-p_a)\right](1-p_a)^\theta \\ &= p_a^{2-\varepsilon}(1-p_a)^{\varepsilon+\theta} + p_a(1-p_a)^{1+\theta} \leq (1-p_a)^\alpha + (1-p_a)^\alpha = 2\bar{p}_a^\alpha . \end{aligned}$$

The two cases prove (76). □

Mean, Moment, and Clipping Tail of the Advantage Function Estimator

Lemma B.7. *Fix an episode t, a state $s \in \mathcal{S}_h$, and an action a. Conditional on $\mathcal{F}_{t-1}$,*

$$\mathbb{E}_t[\widetilde{Q}_t(s,a)] = \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)} \bar{Q}_t(s,a) , \tag{79}$$

$$\mathbb{E}_t[|\widetilde{Q}_t(s,a)|^\varepsilon] \leq \Sigma_t(s) \mathsf{D}_{p_{t,s}}(a) , \tag{80}$$

$$\mathbb{E}_t\left[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|\right] \leq d_t(s,a) . \tag{81}$$

Proof. Fix t, $s \in \mathcal{S}_h$, and $a \in \mathcal{A}$, and abbreviate $p := p_{t,s}$ and $\widetilde{\mu} := \widetilde{\mu}_t(s)$. We first identify the conditional mean of the suffix return when the state-action pair at layer h is fixed. For every $b \in \mathcal{A}$ with $\mathbb{P}(s_{t,h} = s, a_{t,h} = b \mid \mathcal{F}_{t-1}) > 0$, the tower property, the conditional independence of the loss table from the trajectory randomization, and the Bellman definitions of Q_t^π and V_t^π in Section 2 give

$$\begin{aligned} \mathbb{E}_t[R_{t,h} \mid s_{t,h} = s, a_{t,h} = b] &= \mathbb{E}_t\left[Y_t(s,b) + \sum_{j=h+1}^{H-1} Y_t(s_{t,j}, a_{t,j}) \,\middle|\, s_{t,h} = s, a_{t,h} = b \right] \\ &= \ell_t(s,b) + \mathbb{E}_{s' \sim P(\cdot|s,b)}\left[V_t^{\pi_t}(s')\right] \\ &= Q_t^{\pi_t}(s,b) . \end{aligned} \tag{82}$$

The equalities expand the return, take the conditional means of the current loss and continuation return, and use the Bellman definition of $Q_t^{\pi_t}$ in Section 2. A zero-probability conditioning event contributes zero below, regardless of the assigned conditional value.

Using the estimator definition (6), partitioning over the action taken at state s, and then applying (82), we obtain

$$\begin{aligned} \mathbb{E}_t[\widetilde{Q}_t(s,a)] &= \frac{1}{\widetilde{\mu}} \sum_{b \in \mathcal{A}} \mathbb{E}_t[I_t(s)\mathbb{1}\{a_{t,h} = b\} R_{t,h}]\, \mathsf{K}_p(a,b) \\ &= \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}} \sum_{b \in \mathcal{A}} p_b \mathsf{K}_p(a,b) Q_t^{\pi_t}(s,b) \qquad (83) \\ &= \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}} \left(Q_t^{\pi_t}(s,a) - \sum_{b \in \mathcal{A}} p_b Q_t^{\pi_t}(s,b) \right) \\ &= \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}} \bar{Q}_t(s,a) . \end{aligned}$$

Here (83) uses $\mathbb{P}(s_{t,h} = s, a_{t,h} = b \mid \mathcal{F}_{t-1}) = \mu^{\pi_t}(s) p_b$ and (82); the next two equalities apply the centered-kernel identity and the definition of $\bar{Q}_t$ in Section 3, proving (79).

We next prove the ε-moment bound. Since $\{s_{t,h} = s, a_{t,h} = b\} \in \mathcal{H}_t^{\text{path}}$, the tower property and Lemma B.5 imply

$$\mathbb{E}_t[|R_{t,h}|^\varepsilon | s_{t,h} = s, a_{t,h} = b] \leq H^\varepsilon \sigma \tag{84}$$

whenever the conditioning event has positive probability. Substituting (6) and then (84) gives

$$\mathbb{E}_t[|\widetilde{Q}_t(s,a)|^\varepsilon] = \frac{1}{\widetilde{\mu}^\varepsilon} \sum_{b \in \mathcal{A}} \mathbb{E}_t[I_t(s)\mathbb{1}\{a_{t,h} = b\} |R_{t,h}|^\varepsilon]\, |\mathsf{K}_p(a,b)|^\varepsilon$$

$$
\begin{aligned}
&= \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}^{\varepsilon}} \sum_{b\in\mathcal{A}} p_b \mathbb{E}_t[|R_{t,h}|^{\varepsilon} | s_{t,h}=s, a_{t,h}=b]\, |\mathsf{K}_p(a,b)|^{\varepsilon} \\
&\le \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}^{\varepsilon}} H^{\varepsilon}\sigma \sum_{b\in\mathcal{A}} p_b |\mathsf{K}_p(a,b)|^{\varepsilon} = \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}^{\varepsilon}} H^{\varepsilon}\sigma \mathsf{D}_p(a) \qquad (85)\\
&\le H^{\varepsilon}\sigma \widetilde{\mu}^{1-\varepsilon} \mathsf{D}_p(a) \\
&= \Sigma_t(s) \mathsf{D}_{p_{t,s}}(a)\,.
\end{aligned}
$$

The final three lines use the definition of $\mathsf{D}_p(a)$ in (5), $\mu^{\pi_t}(s) \le \widetilde{\mu}$, and the definition of $\Sigma_t(s)$ in Section 3, respectively. This proves (80).

Finally, by the clipped-estimator definition in (7),

$$
|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)| = |\widetilde{Q}_t(s,a)| \mathbb{1}\{|\widetilde{Q}_t(s,a)| > M_t(s,a)\} \le |\widetilde{Q}_t(s,a)|^{\varepsilon} M_t(s,a)^{1-\varepsilon}\,. \qquad (86)
$$

Line (86) uses $|x| \le |x|^{\varepsilon} M^{1-\varepsilon}$ on the event $|x| > M$, which is valid because $\varepsilon > 1$. Taking $\mathbb{E}_t$ in (86), applying (80), and then substituting the definition (7) yields

$$
\begin{aligned}
\mathbb{E}_t\Big[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|\Big] &\le \mathbb{E}_t[|\widetilde{Q}_t(s,a)|^{\varepsilon}] M_t(s,a)^{1-\varepsilon} \\
&\le \Sigma_t(s) \mathsf{D}_{p_{t,s}}(a) M_t(s,a)^{1-\varepsilon} = d_t(s,a)\,, \qquad (87)
\end{aligned}
$$

and equation (87) establishes (81). □

B.2.2 Pointwise Safety

The following two lemmas verify that the compensated local loss lies in the Tsallis stability region.

Lemma B.8 (Bounds on the action bonus)**.** *For every t, s, a,*

$$
0 \le d_t(s,a) \le M_t(s,a)\,. \qquad (88)
$$

Proof. Fix t, s, a and abbreviate $p = p_{t,s}$, $\Sigma = \Sigma_t(s)$, $M_a = M_t(s,a)$, $d_a = d_t(s,a)$, and $\beta = \beta_t(s)$. All factors in the definition of d_a are nonnegative, so $d_a \ge 0$. Directly substituting (7) and then the definition of $M_t(s,a)$ in Section 3 gives

$$
\frac{d_a}{M_a} = \Sigma \mathsf{D}_p(a) M_a^{-\varepsilon} = \frac{\Sigma}{(k_\alpha \beta)^{\varepsilon}} \mathsf{D}_p(a) \bar{p}_a^{\varepsilon-1} \le \frac{2\Sigma}{(k_\alpha\beta)^{\varepsilon}}\,, \qquad (89)
$$

where the inequality follows from (75).

The clock definition gives $G_t(s) \ge \widetilde{\mu}_t(s)^{1-\varepsilon}$. Since $\beta_t(s) = \beta_0 + c_\beta H L_\sigma G_t(s)^{1/\varepsilon}$ and $\beta_0 \ge 0$,

$$
\beta^{\varepsilon} \ge (c_\beta H L_\sigma)^{\varepsilon} G_t(s) \ge c_\beta^{\varepsilon} H^{\varepsilon} L_\sigma^{\varepsilon} \widetilde{\mu}_t(s)^{1-\varepsilon} = c_\beta^{\varepsilon} H^{\varepsilon} \sigma \widetilde{\mu}_t(s)^{1-\varepsilon} = c_\beta^{\varepsilon} \Sigma\,.
$$

Substituting this bound into (89) yields

$$
\frac{d_a}{M_a} \le \frac{2}{k_\alpha^{\varepsilon} c_\beta^{\varepsilon}} = \frac{2}{(k_\alpha c_\beta)^{\varepsilon}} = \frac{2}{64^{\varepsilon}} < 1\,.
$$

The last equality uses $k_\alpha c_\beta = 64$ from the definitions in Section 3. Multiplying by $M_a > 0$ proves (88). □

Lemma B.9 (Pointwise size of c_t, C_t, b_t, B_t)**.** *For every episode t,*

$$
0 \le c_t(s) \le 2HL_\sigma\,, \qquad (90)
$$

$$
0 \le C_t(s,a) \le 4H^2 L_\sigma\,, \qquad (91)
$$

$$
b_t(s) \le K_{\mathrm{saf}} H L_\sigma A^{1-\alpha} \widetilde{\mu}_t(s)^{-\frac{\varepsilon-1}{\varepsilon}}\,, \qquad (92)
$$

$$
0 \le B_t(s,a) \le \frac{k_\alpha \beta_t(s)}{32H} \le \frac{M_t(s,a)}{32H}\,, \qquad (93)
$$

$$
0 \le C_t(s,a) \le \frac{M_t(s,a)}{16}\,. \qquad (94)
$$

Moreover,

$$
|g_t(s,a)| \le \frac{1-\alpha}{16} \beta_t(s) \bar{p}_{t,s,a}^{\alpha-1}\,. \qquad (95)
$$

Proof. We prove the displayed bounds in the order in which they are stated. Since $\widetilde{\mu}_t(s) = \mu^{\pi_t}(s) + \gamma_t$ and both summands are nonnegative,

$$0 \le \frac{\gamma_t}{\widetilde{\mu}_t(s)} \le 1\,. \tag{96}$$

Substituting (96) directly into the definition of $c_t(s)$ in Section 3 gives

$$0 \le c_t(s) = 2HL_\sigma \frac{\gamma_t}{\widetilde{\mu}_t(s)} \le 2HL_\sigma\,, \tag{97}$$

and equation (97) establishes (90).

Define the state-averaged continuation envelope $\bar{C}_t(s) \coloneqq \sum_{a\in\mathcal{A}} p_{t,s,a} C_t(s,a)$. The recursion (8) can then be written as

$$C_t(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[2c_t(s') + \bar{C}_t(s')\right]\,. \tag{98}$$

Repeatedly substituting (98), using the terminal convention $C_t(s_H,\cdot) = c_t(s_H) = 0$, and then applying (90) yields

$$C_t(s,a) = 2\mathbb{E}_{P,\pi_t}\left[\sum_{j=h(s)+1}^{H-1} c_t(s_j)\,\middle|\, s_{h(s)} = s, a_{h(s)} = a\right] \tag{99}$$

$$\le 2\big(H - h(s) - 1\big)\max_x c_t(x) \tag{100}$$

$$\le 4H^2 L_\sigma\,. \tag{101}$$

Line (99) is the exact backward unrolling of (98). Line (100) bounds the number of future nonterminal layers by $H - h(s) - 1$. Line (101) uses (90) and $H - h(s) - 1 \le H$. Thus (91) holds.

For the local scalar bonus, put $x_t(s) \coloneqq \widetilde{\mu}_t(s)^{1-\varepsilon}$. The clock update is $G_t(s) = G_{t-1}(s) + x_t(s)$ and $\beta_t(s) = \beta_0 + c_\beta HL_\sigma G_t(s)^\alpha$. Since $0 < \alpha \le 1$ and $(u+v)^\alpha - u^\alpha \le v^\alpha$ for $u, v \ge 0$,

$$\beta_t(s) - \beta_{t-1}(s) = c_\beta HL_\sigma\left(G_t(s)^\alpha - G_{t-1}(s)^\alpha\right) \le c_\beta HL_\sigma x_t(s)^\alpha\,.$$

Furthermore, concavity of $u \mapsto u^\alpha$ and $\sum_a p_{t,s,a} = 1$ imply

$$\sum_a (p_{t,s,a})^\alpha \le A^{1-\alpha}\,,$$

$$h_t(s) = \frac{1}{\alpha}\left(\sum_a (p_{t,s,a})^\alpha - 1\right) \le \frac{A^{1-\alpha}}{\alpha}\,,$$

$$z_t(s) = \sum_a \bar{p}_{t,s,a}^{\,\alpha} \le \sum_a (p_{t,s,a})^\alpha \le A^{1-\alpha}\,.$$

Consequently, the changing-rate part of $b_t(s)$ satisfies

$$\mathbb{1}\{t \ge 2\}(\beta_t(s) - \beta_{t-1}(s))h_t(s) \le \frac{c_\beta}{\alpha} HL_\sigma A^{1-\alpha} x_t(s)^\alpha\,. \tag{102}$$

For the heavy-tail part, substitute $\Sigma_t(s) = H^\varepsilon L_\sigma^\varepsilon x_t(s)$ and use $\beta_t(s) \ge c_\beta HL_\sigma G_t(s)^\alpha$ together with $G_t(s) \ge x_t(s)$. Because $1 - \varepsilon < 0$, these lower bounds on $\beta_t(s)$ may be raised to the power $1-\varepsilon$ in the reverse direction:

$$\begin{aligned} K_{\mathrm{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon} z_t(s) &\le K_{\mathrm{ht}} H^\varepsilon L_\sigma^\varepsilon x_t(s)\left(c_\beta HL_\sigma x_t(s)^\alpha\right)^{1-\varepsilon} A^{1-\alpha} \\ &= K_{\mathrm{ht}} c_\beta^{1-\varepsilon} HL_\sigma A^{1-\alpha} x_t(s)^\alpha\,. \end{aligned} \tag{103}$$

Line (103) uses $\alpha(1-\varepsilon) = \alpha - 1$ and hence $1 + \alpha(1-\varepsilon) = \alpha$.

For the squared-bias correction, substitute (91) into (11), use $\rho(p_{t,s})^{1-\alpha} \le 1$ and $\beta_t(s) \ge \beta_0$, and obtain

$$\nu_t(s) \le \frac{16}{(1-\alpha)\beta_0} \sum_a p_{t,s,a}(4H^2 L_\sigma)^2 = \frac{256H^4 L_\sigma^2}{(1-\alpha)\beta_0}\,. \tag{104}$$

The definition of β_0 in Section 3 implies $\beta_0 \geq 512H^3L_\sigma/(1-\alpha)$. Moreover, $\widetilde{\mu}_t(s) \leq \mu^{\pi_t}(s)+\gamma_t \leq 2$, so

$$x_t(s)^\alpha = \widetilde{\mu}_t(s)^{-\frac{\varepsilon-1}{\varepsilon}} \geq 2^{-\frac{\varepsilon-1}{\varepsilon}} \geq \frac{1}{2}\,. \tag{105}$$

Using (105), $A^{1-\alpha} \geq 1$, and the lower bound on β_0 in (104) gives

$$\nu_t(s) \leq \frac{1}{2}HL_\sigma \leq HL_\sigma A^{1-\alpha}x_t(s)^\alpha\,. \tag{106}$$

Substituting (102), (103), and (106) into the definition (10) gives the complete chain

$$b_t(s) \leq \left(\frac{c_\beta}{\alpha} + K_{\mathrm{ht}}c_\beta^{1-\varepsilon} + 1\right) HL_\sigma A^{1-\alpha}x_t(s)^\alpha = K_{\mathrm{saf}}HL_\sigma A^{1-\alpha}\widetilde{\mu}_t(s)^{-\frac{\varepsilon-1}{\varepsilon}}\,, \tag{107}$$

and equation (107) establishes (92).

Let $b_t^{\max} := \max_x b_t(x)$. Unrolling the nonnegative recursion (12) over at most H nonterminal layers gives

$$0 \leq B_t(s,a) \leq \sum_{j=0}^{H-1}\left(1+\frac{1}{H}\right)^j b_t^{\max} \leq H\left(1+\frac{1}{H}\right)^{H-1} b_t^{\max} \leq eHb_t^{\max}\,.$$

Because $\widetilde{\mu}_t(x) \geq \gamma_t$ and the exponent in (92) is negative,

$$B_t(s,a) \leq eK_{\mathrm{saf}}H^2L_\sigma A^{1-\alpha}\gamma_t^{-\frac{\varepsilon-1}{\varepsilon}}\,. \tag{108}$$

If $t \leq \Gamma$, then $\gamma_t = 1$ and $\beta_t(s) \geq \beta_0$. Substituting the definition of β_0 in Section 3 and $k_\alpha = (1-\alpha)/64$ gives

$$\frac{B_t(s,a)}{k_\alpha\beta_t(s)} \leq \frac{eK_{\mathrm{saf}}H^2L_\sigma A^{1-\alpha}}{k_\alpha\beta_0} \leq \frac{eK_{\mathrm{saf}}H^2L_\sigma A^{1-\alpha}}{64eK_{\mathrm{saf}}H^3L_\sigma A^{1-\alpha}} = \frac{1}{64H}\,.$$

If $t > \Gamma$, then $\gamma_t = \Gamma/t$. For every $\tau \leq t$, $\widetilde{\mu}_\tau(s) \leq 2$, and because $1-\varepsilon < 0$, $\widetilde{\mu}_\tau(s)^{1-\varepsilon} \geq 2^{1-\varepsilon}$. Therefore

$$\begin{aligned} G_t(s) &\geq \sum_{\tau=1}^{t} 2^{1-\varepsilon} = 2^{1-\varepsilon}t\,,\\ \beta_t(s) &\geq c_\beta HL_\sigma 2^{\frac{1-\varepsilon}{\varepsilon}}t^{\frac{1}{\varepsilon}}\,. \end{aligned} \tag{109}$$

Substituting $\gamma_t = \Gamma/t$ into (108) and dividing by (109) yields

$$\begin{aligned} \frac{B_t(s,a)}{k_\alpha\beta_t(s)} &\leq \frac{2eK_{\mathrm{saf}}HA^{1-\alpha}}{k_\alpha c_\beta}\Gamma^{-\frac{\varepsilon-1}{\varepsilon}}t^{\frac{\varepsilon-2}{\varepsilon}}\\ &\leq \frac{2eK_{\mathrm{saf}}HA^{1-\alpha}}{k_\alpha c_\beta}\Gamma^{-\frac{1}{\varepsilon}}\\ &\leq \frac{1}{32H}\,. \end{aligned}$$

The first inequality uses $2^{(\varepsilon-1)/\varepsilon} \leq 2$, the second uses $t \geq \Gamma$ and $(\varepsilon-2)/\varepsilon \leq 0$, and the third uses the definition of Γ in Section 3 and $k_\alpha c_\beta = 64$. Combining both cases gives

$$0 \leq B_t(s,a) \leq \frac{k_\alpha\beta_t(s)}{32H}\,.$$

Since $\bar{p}_{t,s,a} \leq 1$ and $\alpha - 1 < 0$, the threshold definition of $M_t(s,a)$ in Section 3 gives

$$M_t(s,a) = k_\alpha\beta_t(s)\bar{p}_{t,s,a}^{\alpha-1} \geq k_\alpha\beta_t(s)\,, \tag{110}$$

and hence

$$0 \leq B_t(s,a) \leq \frac{k_\alpha\beta_t(s)}{32H} \leq \frac{M_t(s,a)}{32H}\,. \tag{111}$$

Equation (111) is exactly (93).

For C_t, use (91), $\beta_t(s) \geq \beta_0$, and the explicit choice of β_0:

$$\frac{C_t(s,a)}{k_\alpha \beta_t(s)} \leq \frac{4H^2 L_\sigma}{k_\alpha \beta_0} \leq \frac{1}{16}\,. \tag{112}$$

Substituting (110) into (112) gives the full chain

$$0 \leq C_t(s,a) \leq \frac{k_\alpha \beta_t(s)}{16} \leq \frac{M_t(s,a)}{16}\,,$$

which is (94).

Finally, substituting the definition (14) and then applying the clipping bound $|\widehat{Q}_t(s,a)| \leq M_t(s,a)$, Lemma B.8, (93), and (94) gives

$$\begin{aligned} |g_t(s,a)| &\leq |\widehat{Q}_t(s,a)| + d_t(s,a) + B_t(s,a) + C_t(s,a) && (113)\\ &\leq M_t(s,a) + M_t(s,a) + \frac{M_t(s,a)}{32H} + \frac{M_t(s,a)}{16} \\ &< 4M_t(s,a) = \frac{1-\alpha}{16}\beta_t(s)\bar{p}_{t,s,a}^{\alpha-1}\,. && (114) \end{aligned}$$

Line (114) uses $M_t(s,a) = k_\alpha \beta_t(s)\bar{p}_{t,s,a}^{\alpha-1}$ and $4k_\alpha = (1-\alpha)/16$. Equations (113)–(114) establish (95). □

B.2.3 Conditional Stability and the FTRL Bound

Fix a state s for this section and abbreviate $\beta_t := \beta_t(s)$, $\Sigma_t := \Sigma_t(s)$, $M_{t,a} := M_t(s,a)$. Let $a^\star$ be an arbitrary comparator action at this state. For each t, let $\widetilde{p}_{t+1}$ denote the same-inverse-learning-rate look-ahead point in (464) for this state's realized local-loss sequence. It is used only in the stability analysis and is not $\pi_{t+1}(\cdot \mid s)$.

Lemma B.10 (Conditional stability cost)**.** *For every t,*

$$\begin{aligned} &\mathbb{E}_t\left[\langle p_{t,s} - \widetilde{p}_{t+1}, g_t(s,\cdot)\rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_{t,s})\right] \\ &\leq \left(64(1-\alpha)^{-1}k_\alpha^{2-\varepsilon}\right)\Sigma_t \beta_t^{1-\varepsilon} z_t(s) + \nu_t(s) + \frac{1}{H}\left\langle p_{t,s}, B_t(s,\cdot)\right\rangle\,. \end{aligned} \tag{115}$$

Proof. For the fixed state s, write $M_{t,a} := M_t(s,a)$ and define the pathwise stability quantity

$$\mathsf{S}_t(s) := \langle p_{t,s} - \widetilde{p}_{t+1}, g_t(s,\cdot)\rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_{t,s})\,.$$

The pointwise estimate (95) verifies the hypothesis of Lemma D.6. Applying that lemma with $p = p_{t,s}$, $q = \widetilde{p}_{t+1}$, $g = g_t(s,\cdot)$, and $\beta = \beta_t$ gives

$$\mathsf{S}_t(s) \leq \frac{4}{(1-\alpha)\beta_t}\sum_{a\in\mathcal{A}} \bar{p}_{t,s,a}^{2-\alpha} g_t(s,a)^2\,.$$

Substituting the definition $g_t(s,a) = \widehat{Q}_t(s,a) - d_t(s,a) - B_t(s,a) - C_t(s,a)$ and applying $(x_1+x_2+x_3+x_4)^2 \leq 4\sum_{j=1}^4 x_j^2$ gives

$$\mathsf{S}_t(s) \leq \frac{16}{(1-\alpha)\beta_t}\sum_{a\in\mathcal{A}} \bar{p}_{t,s,a}^{2-\alpha}\Big(\widehat{Q}_t(s,a)^2 + d_t(s,a)^2 + B_t(s,a)^2 + C_t(s,a)^2\Big)\,. \tag{116}$$

All quantities in (116), except the clipped estimator $\widehat{Q}_t$, are $\mathcal{F}_{t-1}$-measurable. We bound the four displayed square terms explicitly.

On the clipping event in (7), the relation $\varepsilon \leq 2$ gives

$$\widehat{Q}_t(s,a)^2 = |\widetilde{Q}_t(s,a)|^2 \mathbb{1}\{|\widetilde{Q}_t(s,a)| \leq M_{t,a}\} \leq |\widetilde{Q}_t(s,a)|^\varepsilon M_{t,a}^{2-\varepsilon}\,. \tag{117}$$

Taking $\mathbb{E}_t$ in (117) and applying (80) yields

$$\mathbb{E}_t[\widehat{Q}_t(s,a)^2] \leq \Sigma_t \mathsf{D}_{p_{t,s}}(a) M_{t,a}^{2-\varepsilon}\,. \tag{118}$$

For the deterministic tail-compensation term, Lemma B.8 and the definition (7) give

$$d_t(s,a)^2 \le d_t(s,a)M_{t,a} = \Sigma_t \mathsf{D}_{p_{t,s}}(a) M_{t,a}^{2-\varepsilon}\,. \tag{119}$$

The common weighted factor in (118) and (119) is evaluated by substituting the threshold before applying (75):

$$\begin{aligned}
\bar{p}_{t,s,a}^{2-\alpha}\mathsf{D}_{p_{t,s}}(a)M_{t,a}^{2-\varepsilon} &= \bar{p}_{t,s,a}^{2-\alpha}\mathsf{D}_{p_{t,s}}(a)\left(k_\alpha\beta_t\bar{p}_{t,s,a}^{\alpha-1}\right)^{2-\varepsilon} \\
&= (k_\alpha\beta_t)^{2-\varepsilon}\mathsf{D}_{p_{t,s}}(a)\bar{p}_{t,s,a}^{2-\alpha+(\alpha-1)(2-\varepsilon)} \\
&= (k_\alpha\beta_t)^{2-\varepsilon}\mathsf{D}_{p_{t,s}}(a)\bar{p}_{t,s,a}^{\alpha+\varepsilon-1} \\
&= (k_\alpha\beta_t)^{2-\varepsilon}\left(\mathsf{D}_{p_{t,s}}(a)\bar{p}_{t,s,a}^{\varepsilon-1}\right)\bar{p}_{t,s,a}^{\alpha} \le 2(k_\alpha\beta_t)^{2-\varepsilon}\bar{p}_{t,s,a}^{\alpha}\,.
\end{aligned} \tag{120}$$

The exponent simplification uses $\alpha\varepsilon = 1$, and the last inequality applies (75).

Substituting (118), (119), and (120) into the complete $\widehat{Q}_t$- and d_t-contribution from (116) gives

$$\begin{aligned}
&\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}\left(\mathbb{E}_t[\widehat{Q}_t(s,a)^2] + d_t(s,a)^2\right) \\
&\le \frac{16}{(1-\alpha)\beta_t}\sum_a 2\Sigma_t\bar{p}_{t,s,a}^{2-\alpha}\mathsf{D}_{p_{t,s}}(a)M_{t,a}^{2-\varepsilon} \\
&\le \frac{64}{1-\alpha}k_\alpha^{2-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}\sum_a \bar{p}_{t,s,a}^{\alpha} \\
&= \frac{64}{1-\alpha}k_\alpha^{2-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s)\,.
\end{aligned} \tag{121}$$

The first inequality applies (118) and (119), producing a factor of two; the second uses (120), and the equality substitutes the definition of $z_t(s)$ in Section 3.

For the C_t-square contribution, the coordinatewise inequality

$$\bar{p}_{t,s,a}^{2-\alpha} \le p_{t,s,a}\rho(p_{t,s})^{1-\alpha} \tag{122}$$

holds for every action. If $\bar{p}_{t,s,a} = p_{t,s,a}$, then $p_{t,s,a} \le \rho(p_{t,s})$ and ${p_{t,s,a}}^{2-\alpha} \le p_{t,s,a}\rho(p_{t,s})^{1-\alpha}$; if $\bar{p}_{t,s,a} = \rho(p_{t,s})$, then $\rho(p_{t,s}) \le p_{t,s,a}$ and the same inequality follows. Substituting (122) into the C_t-part of (116) yields

$$\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}C_t(s,a)^2 \le \frac{16\rho(p_{t,s})^{1-\alpha}}{(1-\alpha)\beta_t}\sum_a p_{t,s,a}C_t(s,a)^2 = \nu_t(s)\,, \tag{123}$$

where the equality is the definition of $\nu_t(s)$ in (11).

For the B_t-square contribution, nonnegativity of B_t, together with (122) and (93), gives

$$\begin{aligned}
\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}B_t(s,a)^2 &\le \frac{16\rho(p_{t,s})^{1-\alpha}}{(1-\alpha)\beta_t}\left(\max_a B_t(s,a)\right)\sum_a p_{t,s,a}B_t(s,a) \\
&\le \frac{16\rho(p_{t,s})^{1-\alpha}}{(1-\alpha)\beta_t}\frac{k_\alpha\beta_t}{32H}\sum_a p_{t,s,a}B_t(s,a) \\
&= \frac{\rho(p_{t,s})^{1-\alpha}}{128H}\sum_a p_{t,s,a}B_t(s,a) \\
&\le \frac{1}{H}\left\langle p_{t,s}, B_t(s,\cdot)\right\rangle\,.
\end{aligned} \tag{124}$$

The equality substitutes $k_\alpha = (1-\alpha)/64$, and the final inequality uses $\rho(p_{t,s})^{1-\alpha} \le 1$ and $1/128 \le 1$.

Taking $\mathbb{E}_t$ in (116) and substituting the complete bounds (121), (123), and (124) gives

$$\mathbb{E}_t[\mathsf{S}_t(s)] \le \frac{64}{1-\alpha}k_\alpha^{2-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) + \nu_t(s) + \frac{1}{H}\left\langle p_{t,s}, B_t(s,\cdot)\right\rangle\,.$$

This proves (115). □

Applying the variable-rate FTRL inequality (466) and summing the conditional stability bound gives the cumulative FTRL estimate recorded in (38). The clipping term is controlled in Section B.3; the two estimates are combined in (41), within Section B.1.1.

B.3 Bounding Clipping-Bias$^{\pi}(s)$

Fix a state s and a comparator action $a^\star$, and abbreviate $\beta_t = \beta_t(s)$, $\Sigma_t = \Sigma_t(s)$, and $M_{t,a} = M_t(s,a)$. We use the same look-ahead point $\widetilde{p}_{t+1}$ as in Section B.2. The clipping-tail estimate above is compensated coordinatewise by d_t, giving the following comparator-sensitive bound.

Lemma B.11 (Clipping compensation)**.** *Recall from Section 3 that* $\tau_t(s,a) = \mathbb{E}_t[\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)]$. *Then* $|\tau_t(s,a)| \le d_t(s,a)$ *and*

$$\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \le 2 \sum_{a \ne a^\star} p_{t,s,a} d_t(s,a)\,. \tag{125}$$

Moreover,

$$2 \sum_{a \ne a^\star} p_{t,s,a} d_t(s,a) \le 4 k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} z_t(s)\,. \tag{126}$$

Proof. The vector $p_{t,s} - e_{a^\star}$ is $\mathcal{F}_{t-1}$-measurable. By the triangle inequality for conditional expectation and the clipping-tail estimate (81), each coordinate satisfies

$$|\tau_t(s,a)| = \Big|\mathbb{E}_t[\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)]\Big| \le \mathbb{E}_t\Big[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|\Big] \le d_t(s,a)\,.$$

Consequently,

$$0 \le \tau_t(s,a) + d_t(s,a) \le 2 d_t(s,a) \qquad \text{for every } a\,. \tag{127}$$

Expanding the policy-difference inner product and applying (127) gives

$$\begin{aligned} &\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \\ &= \sum_{a \ne a^\star} p_{t,s,a}\big(\tau_t(s,a) + d_t(s,a)\big) + (p_{t,s,a^\star} - 1)\big(\tau_t(s,a^\star) + d_t(s,a^\star)\big) \qquad (128)\\ &\le 2 \sum_{a \ne a^\star} p_{t,s,a} d_t(s,a)\,. \qquad (129)\end{aligned}$$

The comparator-coordinate summand in (128) is nonpositive because $p_{t,s,a^\star} - 1 \le 0$ and $\tau_t(s,a^\star) + d_t(s,a^\star) \ge 0$. Every non-comparator coordinate is bounded by the upper inequality in (127). Thus (129) proves (125).

For each action a, substitute $d_t(s,a)$ from (7) and $M_t(s,a)$ from Section 3 without applying an inequality:

$$\begin{aligned} p_{t,s,a} d_t(s,a) &= p_{t,s,a} \Sigma_t \mathsf{D}_{p_{t,s}}(a) \left(k_\alpha \beta_t \bar{p}_{t,s,a}^{\alpha-1}\right)^{1-\varepsilon} \\ &= k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} p_{t,s,a} \mathsf{D}_{p_{t,s}}(a) \bar{p}_{t,s,a}^{(\alpha-1)(1-\varepsilon)} \\ &= k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} p_{t,s,a} \mathsf{D}_{p_{t,s}}(a) \bar{p}_{t,s,a}^{(\varepsilon-1)(1-\alpha)} \qquad (130)\\ &\le 2 k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} \bar{p}_{t,s,a}^{\alpha}\,. \qquad (131)\end{aligned}$$

Line (130) uses $(\alpha-1)(1-\varepsilon) = (\varepsilon-1)(1-\alpha)$. Line (131) applies (76) with $p = p_{t,s}$.

Substituting (131) into the complete right-hand side of (125) gives

$$\begin{aligned} \langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \le 2 \sum_{a \ne a^\star} p_{t,s,a} d_t(s,a) &\le 4 k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} \sum_{a \ne a^\star} \bar{p}_{t,s,a}^{\alpha} \\ &\le 4 k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} \sum_{a \in \mathcal{A}} \bar{p}_{t,s,a}^{\alpha} = 4 k_\alpha^{1-\varepsilon} \Sigma_t \beta_t^{1-\varepsilon} z_t(s)\,. \end{aligned}$$

The last equality uses the definition of $z_t(s)$ in Section 3, proving (126). □

B.4 BOUNDING Estimator-Bias$^{\pi}(s)$

For known transitions, (79) gives the exact bias $r_t(s,a) = \gamma_t \bar{Q}_t(s,a)/\widetilde{\mu}_t(s)$ in (34). We bound its contribution together with C_t after weighting by the comparator occupancies. The signed Bellman-residual argument below gives the known-transition bound (136).

Lemma B.12 (Nonnegative Bellman residual for signed losses)**.** *Fix an episode t. Recall from Section 3 that*

$$\mathcal{L}_t^{\mathrm{bias}}(s,a) = r_t(s,a) + C_t(s,a) + c_t(s)\,. \tag{132}$$

Recall from Section 4 that

$$\zeta_t(s,a) = \mathcal{L}_t^{\mathrm{bias}}(s,a) - \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim p_{t,s'}}}\left[\mathcal{L}_t^{\mathrm{bias}}(s',a')\right]\,. \tag{133}$$

Then

$$\zeta_t(s,a) = r_t(s,a) + c_t(s) + \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim p_{t,s'}}}[c_t(s') - r_t(s',a')]\,, \tag{134}$$

and consequently

$$0 \le \zeta_t(s,a) \le 2c_t(s) + 2\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s')\right]\,. \tag{135}$$

For every deterministic policy π,

$$\sum_s \mu^{\pi}(s)\,\langle p_{t,s} - \pi(\cdot\mid s), r_t(s,\cdot) + C_t(s,\cdot)\rangle \le 4\sum_s \mu^{\pi_t}(s)c_t(s) \le 8SHL_\sigma\gamma_t\,. \tag{136}$$

Proof. We first control the signed local bias r_t. By (34), the raw-moment condition and Bellman definitions in Section 2, and the definition of $\bar{Q}_t(s,a)$ in Section 3,

$$\begin{aligned}|r_t(s,a)| &= \frac{\gamma_t}{\widetilde{\mu}_t(s)}|\bar{Q}_t(s,a)| = \frac{\gamma_t}{\widetilde{\mu}_t(s)}|Q_t^{\pi_t}(s,a) - V_t^{\pi_t}(s)|\\ &\le \frac{\gamma_t}{\widetilde{\mu}_t(s)}\left(|Q_t^{\pi_t}(s,a)| + |V_t^{\pi_t}(s)|\right) \le 2HL_\sigma\frac{\gamma_t}{\widetilde{\mu}_t(s)} = c_t(s)\,.\end{aligned} \tag{137}$$

The last inequality uses $|Q_t^{\pi_t}(s,a)|, |V_t^{\pi_t}(s)| \le HL_\sigma$, which follows from the moment condition and the Bellman definitions; the equality uses $c_t(s)$ from Section 3. In particular,

$$-c_t(s) \le r_t(s,a) \le c_t(s) \qquad \text{for every } s,a\,. \tag{138}$$

Define the policy averages

$$\bar{r}_t(s) \coloneqq \sum_a p_{t,s,a} r_t(s,a)\,, \qquad \bar{C}_t(s) \coloneqq \sum_a p_{t,s,a} C_t(s,a)\,. \tag{139}$$

The recursion (8) is equivalently

$$C_t(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[2c_t(s') + \bar{C}_t(s')\right]\,. \tag{140}$$

Starting from (133), directly substituting (132), (139), and (140) gives

$$\begin{aligned}\zeta_t(s,a) &= r_t(s,a) + C_t(s,a) + c_t(s) - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\bar{r}_t(s') + \bar{C}_t(s') + c_t(s')\right]\\ &= r_t(s,a) + c_t(s) + \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s') - \bar{r}_t(s')\right]\,.\end{aligned} \tag{141}$$

The last equality substitutes the recursion for C_t and cancels the continuation terms $\bar{C}_t(s')$, proving (134).

The coordinatewise bracket (138) implies

$$0 \le r_t(s,a) + c_t(s) \le 2c_t(s)\,, \tag{142}$$

$$0 \le c_t(s') - \bar{r}_t(s') \le 2c_t(s') \qquad \text{for every } s'\,. \tag{143}$$

Inequality (143) follows because $\bar{r}_t(s') = \sum_a p_{t,s',a} r_t(s',a)$ is a convex combination of coordinates satisfying (138). Substituting (142)– (143) into (141) gives the complete two-sided estimate

$$0 \le \zeta_t(s,a) \le 2c_t(s) + 2\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s')\right]\,, \tag{144}$$

and equation (144) establishes (135).

We now prove the global comparison in (136). The terminal convention gives $\mathcal{L}_t^{\mathrm{bias}}(s_H, \cdot) = 0$. Rearranging (133) yields the Bellman recursion

$$\mathcal{L}_t^{\mathrm{bias}}(s,a) = \zeta_t(s,a) + \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim p_{t,s'}}}\left[\mathcal{L}_t^{\mathrm{bias}}(s',a')\right]. \tag{145}$$

Backward induction over the layers in (145) shows that $\mathcal{L}_t^{\mathrm{bias}}$ is the Q-function of policy π_t for the artificial one-step loss ζ_t.

Since $c_t(s)$ is the same for every action at state s,

$$\sum_s \mu^{\pi}(s)\left\langle p_{t,s} - \pi(\cdot \mid s), r_t(s,\cdot) + C_t(s,\cdot)\right\rangle = \sum_s \mu^{\pi}(s)\left\langle p_{t,s} - \pi(\cdot \mid s), \mathcal{L}_t^{\mathrm{bias}}(s,\cdot)\right\rangle. \tag{146}$$

Applying the performance-difference identity to the artificial loss ζ_t and its Q-function $\mathcal{L}_t^{\mathrm{bias}}$ gives

$$\begin{aligned}
&\sum_s \mu^{\pi}(s)\left\langle p_{t,s} - \pi(\cdot \mid s), \mathcal{L}_t^{\mathrm{bias}}(s,\cdot)\right\rangle \\
&= V^{\pi_t}(s_0;\zeta_t) - V^{\pi}(s_0;\zeta_t) &(147)\\
&= \sum_{s,a}\big(\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)\big)\zeta_t(s,a) \le \sum_{s,a}\mu^{\pi_t}(s,a)\zeta_t(s,a). &(148)
\end{aligned}$$

Line (148) uses $\mu^{\pi}(s,a) \ge 0$ and $\zeta_t(s,a) \ge 0$, established in (144).

Substituting the upper bound from (135) into the complete right-hand side of (148) gives

$$\begin{aligned}
\sum_{s,a}\mu^{\pi_t}(s,a)\zeta_t(s,a) &\le 2\sum_{s,a}\mu^{\pi_t}(s,a)c_t(s) + 2\sum_{s,a}\mu^{\pi_t}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s')\right] \\
&= 2\sum_s \mu^{\pi_t}(s)c_t(s) + 2\sum_{h=0}^{H-2}\sum_{s'\in\mathcal{S}_{h+1}}\mu^{\pi_t}(s')c_t(s') &(149)\\
&\le 4\sum_s \mu^{\pi_t}(s)c_t(s). &(150)
\end{aligned}$$

The two transformations used in (149) are

$$2\sum_{s,a}\mu^{\pi_t}(s,a)c_t(s) = 2\sum_s \mu^{\pi_t}(s)c_t(s), \tag{151}$$

$$2\sum_{s,a}\mu^{\pi_t}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s')\right] = 2\sum_{h=0}^{H-2}\sum_{s'\in\mathcal{S}_{h+1}}\mu^{\pi_t}(s')c_t(s'). \tag{152}$$

Equality (151) uses $\sum_a \mu^{\pi_t}(s,a) = \mu^{\pi_t}(s)$. Equality (152) follows from the layerwise occupancy-flow identity

$$\sum_{s\in\mathcal{S}_h}\sum_a \mu^{\pi_t}(s,a)P(s' \mid s,a) = \mu^{\pi_t}(s'), \qquad s' \in \mathcal{S}_{h+1}.$$

The transported state sum satisfies

$$\sum_{h=0}^{H-2}\sum_{s'\in\mathcal{S}_{h+1}}\mu^{\pi_t}(s')c_t(s') \le \sum_s \mu^{\pi_t}(s)c_t(s), \tag{153}$$

because its layer range contains every nonterminal layer except the initial layer. Substituting (153) into (149) yields (150).

Finally, direct substitution of the definition of $c_t(s)$ in Section 3 and $\widetilde{\mu}_t(s) = \mu^{\pi_t}(s) + \gamma_t$ gives

$$\begin{aligned}
\sum_s \mu^{\pi_t}(s)c_t(s) &= 2HL_\sigma\gamma_t\sum_s \frac{\mu^{\pi_t}(s)}{\mu^{\pi_t}(s)+\gamma_t} &(154)\\
&\le 2HL_\sigma\gamma_t\sum_s 1 = 2SHL_\sigma\gamma_t. &(155)
\end{aligned}$$

Combining (146), (147)–(148), (150), and (155) gives

$$\sum_s \mu^{\pi}(s)\left\langle p_{t,s} - \pi(\cdot \mid s), r_t(s,\cdot) + C_t(s,\cdot)\right\rangle \le 4\sum_s \mu^{\pi_t}(s)c_t(s) \le 8SHL_\sigma\gamma_t,$$

which is (136). □

B.5 Bounding $V^{\pi_t}(s_0; b_t)$

The remaining term in (45) is $\mathbb{E}[\sum_{t=1}^{T} V^{\pi_t}(s_0; b_t)]$, since $V^{\pi_t}(s_0; b_t) = \sum_s \mu^{\pi_t}(s) b_t(s)$. The following proposition gives the complete count bound. Its proof uses the comparator-removal, clock, and squared-bias estimates established in the subsections below.

Proposition B.13 (Global bonus-count bound)**.** *Fix a deterministic comparator policy π. Then*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s)b_t(s)\right] \tag{156}$$

$$\leq \left(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\right) HL_\sigma \Lambda_T^{1-\alpha} \sum_{s}\sum_{a\neq\pi(s)} \widetilde{n}_T(s,a)^\alpha + 4SHL_\sigma \mathsf{G}_T\,. \tag{157}$$

Proof. Fix a state s and abbreviate $x_t \coloneqq \widetilde{\mu}_t(s)^{1-\varepsilon}$, $G_t \coloneqq G_t(s)$. We first bound the changing-rate component of $b_t(s)$. For $t \geq 2$, direct substitution of the clock definition (13) gives

$$\begin{aligned}
&\mu^{\pi_t}(s)(\beta_t(s) - \beta_{t-1}(s))h_t(s) \\
&= \mu^{\pi_t}(s)c_\beta HL_\sigma\big(G_t^\alpha - G_{t-1}^\alpha\big)h_t(s) \\
&\leq \mu^{\pi_t}(s)c_\beta HL_\sigma\left(\alpha 5^{1-\alpha}x_t G_t^{\alpha-1}\right)\left(\frac{1}{\alpha}\sum_{a\neq\pi(s)} {p_{t,s,a}}^\alpha\right) && (158)\\
&= 5^{1-\alpha}c_\beta HL_\sigma \mu^{\pi_t}(s) x_t G_t^{\alpha-1} \sum_{a\neq\pi(s)} {p_{t,s,a}}^\alpha \\
&\leq 5^{1-\alpha}c_\beta HL_\sigma \widetilde{\mu}_t(s)^{2-\varepsilon} G_t^{\alpha-1} \sum_{a\neq\pi(s)} {p_{t,s,a}}^\alpha\,. && (159)
\end{aligned}$$

Line (158) uses the clock increment (177) and the comparator-sensitive entropy bound (170). Line (159) uses $\mu^{\pi_t}(s) \leq \widetilde{\mu}_t(s)$ and then substitutes $x_t = \widetilde{\mu}_t(s)^{1-\varepsilon}$. When $t = 1$, the changing-rate component in (10) is zero, while the right-hand side of (159) is nonnegative. Hence the same upper bound holds after multiplying the left-hand side by $\mathbb{1}\{t \geq 2\}$ for all t.

We next bound the heavy-tail component. Substitute $\Sigma_t(s) = H^\varepsilon L_\sigma^\varepsilon x_t$ and $\beta_t(s) \geq c_\beta HL_\sigma G_t^\alpha$ before applying the comparator bound on $z_t(s)$:

$$\begin{aligned}
&\mu^{\pi_t}(s)K_{\mathrm{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s) \\
&= \mu^{\pi_t}(s)K_{\mathrm{ht}}H^\varepsilon L_\sigma^\varepsilon x_t \beta_t(s)^{1-\varepsilon} z_t(s) \\
&\leq \mu^{\pi_t}(s)K_{\mathrm{ht}}H^\varepsilon L_\sigma^\varepsilon x_t \left(c_\beta HL_\sigma G_t^\alpha\right)^{1-\varepsilon}\left(2\sum_{a\neq\pi(s)} {p_{t,s,a}}^\alpha\right) && (160)\\
&= 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}HL_\sigma \mu^{\pi_t}(s) x_t G_t^{\alpha-1} \sum_{a\neq\pi(s)} {p_{t,s,a}}^\alpha && (161)\\
&\leq 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}HL_\sigma \widetilde{\mu}_t(s)^{2-\varepsilon} G_t^{\alpha-1} \sum_{a\neq\pi(s)} {p_{t,s,a}}^\alpha\,. && (162)
\end{aligned}$$

The inequality direction in (160) is reversed when the lower bound on $\beta_t(s)$ is raised to the power $1-\varepsilon < 0$; the same line uses (171). Line (161) uses $H^\varepsilon (H)^{1-\varepsilon} = H$, $L_\sigma^\varepsilon L_\sigma^{1-\varepsilon} = L_\sigma$, and $\alpha(1-\varepsilon) = \alpha - 1$. Line (162) again uses $\mu^{\pi_t}(s)x_t \leq \widetilde{\mu}_t(s)x_t = \widetilde{\mu}_t(s)^{2-\varepsilon}$.

Define the coefficient

$$c_{\mathrm{bon},\varepsilon} \coloneqq 5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\,. \tag{163}$$

Adding the complete bounds (159) and (162), and then summing over episodes, gives

$$\sum_{t=1}^{T}\mu^{\pi_t}(s)\Big[\mathbb{1}\{t\geq 2\}(\beta_t(s)-\beta_{t-1}(s))h_t(s) + K_{\mathrm{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s)\Big]$$

$$\leq c_{\mathrm{bon},\varepsilon} H L_\sigma \sum_{a \neq \pi(s)} \sum_{t=1}^{T} \widetilde{\mu}_t(s)^{2-\varepsilon} {p_{t,s,a}}^{\alpha} G_t(s)^{\alpha-1} . \tag{164}$$

For every fixed action a, the sequence $q_t := p_{t,s,a}$ is nonnegative, bounded by one, and predictable before the episode-t trajectory is sampled. Applying (186) to each action separately gives the explicit family of inequalities

$$\sum_{t=1}^{T} \widetilde{\mu}_t(s)^{2-\varepsilon} {p_{t,s,a}}^{\alpha} G_t(s)^{\alpha-1} \leq \left(\sum_{t=1}^{T} \widetilde{\mu}_t(s) p_{t,s,a} \right)^{\alpha} \Lambda_T^{1-\alpha} , \qquad a \neq \pi(s) . \tag{165}$$

Substituting every inequality in (165) into the entire right-hand side of (164) yields the pathwise bound

$$\sum_{t=1}^{T} \mu^{\pi_t}(s) \Big[\mathbb{1}\{t \geq 2\} (\beta_t(s) - \beta_{t-1}(s)) h_t(s) + K_{\mathrm{ht}} \Sigma_t(s) \beta_t(s)^{1-\varepsilon} z_t(s) \Big]$$

$$\leq c_{\mathrm{bon},\varepsilon} H L_\sigma \Lambda_T^{1-\alpha} \sum_{a \neq \pi(s)} \left(\sum_{t=1}^{T} \widetilde{\mu}_t(s) p_{t,s,a} \right)^{\alpha} . \tag{166}$$

According to the definition of $\widetilde{n}_T(s,a)$ in Section 4, taking expectations in (166) and applying Jensen's inequality, valid because $x \mapsto x^\alpha$ is concave on $\mathbb{R}_+$, gives, for each pair (s,a),

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} \widetilde{\mu}_t(s) p_{t,s,a} \right)^{\alpha} \right] \leq \left(\mathbb{E}\left[\sum_{t=1}^{T} \widetilde{\mu}_t(s) p_{t,s,a} \right] \right)^{\alpha} = \widetilde{n}_T(s,a)^{\alpha} . \tag{167}$$

Substituting (167) into the expected version of (166) and summing over states gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) \Big(\mathbb{1}\{t \geq 2\} (\beta_t(s) - \beta_{t-1}(s)) h_t(s) + K_{\mathrm{ht}} \Sigma_t(s) \beta_t(s)^{1-\varepsilon} z_t(s) \Big) \right]$$

$$\leq c_{\mathrm{bon},\varepsilon} H L_\sigma \Lambda_T^{1-\alpha} \sum_{s} \sum_{a \neq \pi(s)} \widetilde{n}_T(s,a)^{\alpha} . \tag{168}$$

The bound (168) accounts for the changing-rate and heavy-tail summands of $b_t(s)$ in (10). We now bound the remaining squared-bias summand $\nu_t(s)$. Summing (190) over t, taking expectation, and applying the bound $\sum_{t=1}^{T} \gamma_t \leq \mathsf{G}_T$ in Section 4 gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) \nu_t(s) \right] \leq 4 S H L_\sigma \sum_{t=1}^{T} \gamma_t \leq 4 S H L_\sigma \mathsf{G}_T . \tag{169}$$

Finally, substitute the three-term definition (10) into the left-hand side of the proposition and apply (168) and (169) to the complete sum:

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) b_t(s) \right] \leq c_{\mathrm{bon},\varepsilon} H L_\sigma \Lambda_T^{1-\alpha} \sum_{s} \sum_{a \neq \pi(s)} \widetilde{n}_T(s,a)^{\alpha} + 4 S H L_\sigma \mathsf{G}_T$$

$$= \left(5^{1-\alpha} c_\beta + 2 K_{\mathrm{ht}} c_\beta^{1-\varepsilon} \right) H L_\sigma \Lambda_T^{1-\alpha} \sum_{s} \sum_{a \neq \pi(s)} \widetilde{n}_T(s,a)^{\alpha} + 4 S H L_\sigma \mathsf{G}_T .$$

The last equality substitutes $c_{\mathrm{bon},\varepsilon}$ from (163), proving (157). □

B.5.1 Comparator-sensitive Tsallis Quantities

Lemma B.14 (Removing an arbitrary comparator action)**.** *For every $p \in \Delta(\mathcal{A})$ and every action $a^\star$,*

$$h_\alpha(p) \leq \frac{1}{\alpha} \sum_{a \neq a^\star} p_a^{\alpha} , \tag{170}$$

$$z_\alpha(p) \leq 2 \sum_{a \neq a^\star} p_a^{\alpha} . \tag{171}$$

Proof. We first prove (170). Substituting the definition of h_α from (9) and separating the comparator coordinate gives

$$h_\alpha(p) = \frac{1}{\alpha}\left(\sum_a p_a^\alpha - 1\right) = \frac{1}{\alpha}\left(p_{a^\star}^\alpha - 1 + \sum_{a\neq a^\star} p_a^\alpha\right) \le \frac{1}{\alpha}\sum_{a\neq a^\star} p_a^\alpha .$$

The inequality uses $p_{a^\star}^\alpha \le 1$.

We next prove (171). Fix $j \in \arg\max_a p_a$ and consider separately the two cases in the definition of $\rho(p)$.

Suppose first that $p_j \le 1/2$. Then $\|p\|_\infty = p_j \le 1/2$, so $\rho(p) = \|p\|_\infty$ and $p_a \le \rho(p)$ for every a. Therefore $\bar{p}_a = p_a$ for every action. Moreover, $p_{a^\star} \le p_j \le 1/2$ implies $1 - p_{a^\star} \ge p_{a^\star}$. Since $0 < \alpha < 1$, the map $x \mapsto x^\alpha$ is subadditive on $\mathbb{R}_+$, and hence

$$\sum_{a\neq a^\star} p_a^\alpha \ge \left(\sum_{a\neq a^\star} p_a\right)^\alpha = (1 - p_{a^\star})^\alpha \ge p_{a^\star}^\alpha . \tag{172}$$

The first inequality is subadditivity, and the second follows from $1 - p_{a^\star} \ge p_{a^\star}$ and monotonicity. Consequently,

$$z_\alpha(p) = \sum_a \bar{p}_a^\alpha = \sum_a p_a^\alpha = p_{a^\star}^\alpha + \sum_{a\neq a^\star} p_a^\alpha \le 2\sum_{a\neq a^\star} p_a^\alpha .$$

The inequality applies (172).

Suppose now that $p_j > 1/2$. Then j is the unique maximizer, $\rho(p) = 1 - p_j$, $\bar{p}_j = \rho(p)$, and $\bar{p}_a = p_a$ for every $a \neq j$. Thus

$$z_\alpha(p) = \rho(p)^\alpha + \sum_{a\neq j} p_a^\alpha , \tag{173}$$

$$\sum_{a\neq j} p_a^\alpha \ge \left(\sum_{a\neq j} p_a\right)^\alpha = (1 - p_j)^\alpha = \rho(p)^\alpha . \tag{174}$$

These relations follow from the definition of z_α and subadditivity, respectively.

If $a^\star = j$, substituting (174) into (173) gives

$$z_\alpha(p) = \rho(p)^\alpha + \sum_{a\neq j} p_a^\alpha \le 2\sum_{a\neq j} p_a^\alpha = 2\sum_{a\neq a^\star} p_a^\alpha .$$

The remaining case is $a^\star \neq j$. Since $p_j \ge p_{a^\star}$, $p_j^\alpha \ge p_{a^\star}^\alpha$. Therefore

$$\sum_a p_a^\alpha = p_{a^\star}^\alpha + \sum_{a\neq a^\star} p_a^\alpha \le p_j^\alpha + \sum_{a\neq a^\star} p_a^\alpha \le 2\sum_{a\neq a^\star} p_a^\alpha , \tag{175}$$

where the last inequality uses $j \neq a^\star$. Finally, $\bar{p}_a \le p_a$ gives

$$z_\alpha(p) = \sum_a \bar{p}_a^\alpha \le \sum_a p_a^\alpha \le 2\sum_{a\neq a^\star} p_a^\alpha .$$

The last inequality is (175). Together with the preceding two cases, this proves (171). □

B.5.2 The Occupancy-adaptive Clock

Lemma B.15 (Clock increment and logarithmic potential)**.** *Fix a state s and put $x_t := \widetilde{\mu}_t(s)^{1-\varepsilon}$, $G_t := 1 + \sum_{\tau=1}^t x_\tau$. For every $t \ge 2$,*

$$x_t \le 4G_{t-1} , \tag{176}$$

and consequently

$$G_t^\alpha - G_{t-1}^\alpha \le \alpha 5^{1-\alpha} x_t G_t^{\alpha-1} . \tag{177}$$

Moreover,

$$\sum_{t=1}^T \frac{x_t}{G_t} \le \log G_T \le \Lambda_T . \tag{178}$$

Proof. Fix the state s throughout the proof and write $r \coloneqq \varepsilon - 1 \in (0, 1]$. Since $\mu^{\pi_t}(s) \le 1$ and $\gamma_t \le 1$, one has $\widetilde{\mu}_t(s) = \mu^{\pi_t}(s) + \gamma_t \le 2$. Because $1 - \varepsilon = -r < 0$, this upper bound on $\widetilde{\mu}_t(s)$ gives the lower bound

$$x_t = \widetilde{\mu}_t(s)^{1-\varepsilon} = \widetilde{\mu}_t(s)^{-r} \ge 2^{-r} = 2^{1-\varepsilon}. \tag{179}$$

Consequently,

$$G_{t-1} = 1 + \sum_{\tau=1}^{t-1} x_\tau \ge 1 + (t-1)2^{1-\varepsilon} \ge (t-1)2^{1-\varepsilon}. \tag{180}$$

The first inequality sums (179) over $\tau < t$.

We prove (176) by considering the two branches of the floor $\gamma_t = \min\{1, \Gamma/t\}$. If $t \le \Gamma$, then $\gamma_t = 1$ and $\widetilde{\mu}_t(s) = \mu^{\pi_t}(s) + 1 \ge 1$. Since the exponent $1 - \varepsilon$ is nonpositive,

$$x_t = \widetilde{\mu}_t(s)^{1-\varepsilon} \le 1^{1-\varepsilon} = 1 \le G_{t-1} \le 4G_{t-1}.$$

Here $G_{t-1} \ge G_0 = 1$.

Suppose next that $t > \Gamma$. Then $\gamma_t = \Gamma/t$, and $\widetilde{\mu}_t(s) \ge \gamma_t$. Since $1 - \varepsilon = -r < 0$, direct substitution gives

$$x_t = \widetilde{\mu}_t(s)^{-r} \le \gamma_t^{-r} = \left(\frac{t}{\Gamma}\right)^r. \tag{181}$$

Combining (181) with (180) yields the complete ratio calculation

$$\frac{x_t}{G_{t-1}} \le \frac{(t/\Gamma)^r}{(t-1)2^{-r}} = 2^r \Gamma^{-r} \frac{t^r}{t-1} = 2^r \Gamma^{-r} t^{r-1} \frac{t}{t-1} \le 2 \cdot 1 \cdot 1 \cdot 2 = 4.$$

The four factors are bounded by $2, 1, 1, 2$ because $0 < r \le 1$, $\Gamma \ge 1$, and $t \ge 2$. Together with the case $t \le \Gamma$, this proves (176).

Since $G_t = G_{t-1} + x_t$, equation (176) also gives

$$G_t = G_{t-1} + x_t \le G_{t-1} + 4G_{t-1} = 5G_{t-1},$$
$$G_{t-1} \ge \frac{G_t}{5}. \tag{182}$$

To prove (177), apply the supporting-line inequality for the concave function $u \mapsto u^\alpha$ at G_{t-1} and then use (182):

$$G_t^\alpha - G_{t-1}^\alpha \le \alpha G_{t-1}^{\alpha-1}(G_t - G_{t-1}) = \alpha x_t G_{t-1}^{\alpha-1} \le \alpha x_t \left(\frac{G_t}{5}\right)^{\alpha-1} = \alpha 5^{1-\alpha} x_t G_t^{\alpha-1}.$$

The first inequality uses concavity, and the second uses $G_{t-1} \ge G_t/5$ and $\alpha - 1 < 0$. This proves (177).

We finally establish (178). Set $u_t \coloneqq x_t/G_{t-1} \ge 0$. Since $G_t = G_{t-1}(1 + u_t)$ and $u/(1+u) \le \log(1+u)$ for every $u \ge 0$, one has

$$\log G_t - \log G_{t-1} = \log\left(\frac{G_t}{G_{t-1}}\right) = \log(1 + u_t) \ge \frac{u_t}{1 + u_t} = \frac{x_t/G_{t-1}}{1 + x_t/G_{t-1}} = \frac{x_t}{G_t}. \tag{183}$$

Summing (183) from $t = 1$ to T and using $G_0 = 1$ gives

$$\sum_{t=1}^{T} \frac{x_t}{G_t} \le \sum_{t=1}^{T} (\log G_t - \log G_{t-1}) = \log G_T - \log G_0 = \log G_T. \tag{184}$$

For the upper bound on $\log G_T$, use $\widetilde{\mu}_t(s) \ge \gamma_t$ and $1 - \varepsilon < 0$ to obtain $x_t \le \gamma_t^{1-\varepsilon}$. Since $\Gamma \ge 1$, $\gamma_t = \min\{1, \Gamma/t\} \ge 1/t$, and therefore

$$G_T = 1 + \sum_{t=1}^{T} x_t \le 1 + \sum_{t=1}^{T} \gamma_t^{1-\varepsilon} \le 1 + \sum_{t=1}^{T} t^{\varepsilon-1} \le 1 + T\, T^{\varepsilon-1} = 1 + T^\varepsilon \le (eT)^\varepsilon.$$

The last inequality holds for $T \ge 1$ and $\varepsilon > 1$. Taking logarithms and using the definition of Λ_T in Section 4 gives

$$\log G_T \le \varepsilon \log(eT) \le 1 + \varepsilon \log(eT) = \Lambda_T. \tag{185}$$

Equations (184) and (185) prove (178). □

Lemma B.16 (Weighted clock-to-count inequality)**.** *Fix a state s, an action a, and any nonnegative predictable sequence $q_t \in [0,1]$. Then*

$$\sum_{t=1}^{T} \widetilde{\mu}_t(s)^{2-\varepsilon} q_t^{\alpha} G_t(s)^{\alpha-1} \le \left(\sum_{t=1}^{T} \widetilde{\mu}_t(s) q_t \right)^{\alpha} \Lambda_T^{1-\alpha} . \tag{186}$$

Proof. Fix the state s and action a, and abbreviate $G_t \coloneqq G_t(s)$ and $x_t \coloneqq \widetilde{\mu}_t(s)^{1-\varepsilon}$. We first verify the factorization used in the Hölder step. Direct substitution gives

$$\big(\widetilde{\mu}_t(s) q_t\big)^{\alpha} \left(\frac{x_t}{G_t} \right)^{1-\alpha} = \widetilde{\mu}_t(s)^{\alpha} q_t^{\alpha} \left(\frac{\widetilde{\mu}_t(s)^{1-\varepsilon}}{G_t} \right)^{1-\alpha} = \widetilde{\mu}_t(s)^{\alpha+(1-\varepsilon)(1-\alpha)} q_t^{\alpha} G_t^{\alpha-1} . \tag{187}$$

The exponent $\alpha + (1-\varepsilon)(1-\alpha)$ of $\widetilde{\mu}_t(s)$ in (187) satisfies

$$\begin{aligned} \alpha + (1-\varepsilon)(1-\alpha) &= \frac{1}{\varepsilon} + (1-\varepsilon)\left(1 - \frac{1}{\varepsilon}\right) = \frac{1}{\varepsilon} - \frac{(\varepsilon-1)^2}{\varepsilon} \\ &= \frac{1-(\varepsilon-1)^2}{\varepsilon} = \frac{2\varepsilon - \varepsilon^2}{\varepsilon} = 2 - \varepsilon . \end{aligned} \tag{188}$$

Substituting (188) into (187) proves the pointwise identity

$$\widetilde{\mu}_t(s)^{2-\varepsilon} q_t^{\alpha} G_t^{\alpha-1} = \big(\widetilde{\mu}_t(s) q_t\big)^{\alpha} \left(\frac{x_t}{G_t} \right)^{1-\alpha} . \tag{189}$$

Apply Hölder's inequality to the two nonnegative sequences $a_t \coloneqq \big(\widetilde{\mu}_t(s) q_t\big)^{\alpha}$, $b_t \coloneqq \left(\frac{x_t}{G_t}\right)^{1-\alpha}$, with conjugate exponents $1/\alpha$ and $1/(1-\alpha)$. Using (189), the resulting calculation is

$$\begin{aligned} \sum_{t=1}^{T} \widetilde{\mu}_t(s)^{2-\varepsilon} q_t^{\alpha} G_t^{\alpha-1} &= \sum_{t=1}^{T} \big(\widetilde{\mu}_t(s) q_t\big)^{\alpha} \left(\frac{x_t}{G_t} \right)^{1-\alpha} \\ &\le \left(\sum_{t=1}^{T} \left[\big(\widetilde{\mu}_t(s) q_t\big)^{\alpha} \right]^{\frac{1}{\alpha}} \right)^{\alpha} \left(\sum_{t=1}^{T} \left[\left(\frac{x_t}{G_t} \right)^{1-\alpha} \right]^{\frac{1}{1-\alpha}} \right)^{1-\alpha} \\ &= \left(\sum_{t=1}^{T} \widetilde{\mu}_t(s) q_t \right)^{\alpha} \left(\sum_{t=1}^{T} \frac{x_t}{G_t} \right)^{1-\alpha} \le \left(\sum_{t=1}^{T} \widetilde{\mu}_t(s) q_t \right)^{\alpha} \Lambda_T^{1-\alpha} . \end{aligned}$$

The first inequality is Hölder with conjugate exponents $1/\alpha$ and $1/(1-\alpha)$, the next equality simplifies the powers, and the last inequality applies (178). This proves (186). □

B.5.3 The Squared-bias Correction

Lemma B.17 (Global cost of the signed-bias square correction)**.** *For every episode t,*

$$\sum_{s} \mu^{\pi_t}(s) \nu_t(s) \le 4SHL_{\sigma}\gamma_t . \tag{190}$$

Proof. Fix episode t. The pointwise envelope (91) gives $0 \le C_t(s,a) \le 4H^2 L_{\sigma}$, and therefore

$$C_t(s,a)^2 \le 4H^2 L_{\sigma}\, C_t(s,a) \qquad \text{for every } s, a . \tag{191}$$

Substituting (191) into the definition (11), and only then applying $\rho(p_{t,s})^{1-\alpha} \le 1$ and $\beta_t(s) \ge \beta_0$, gives

$$\begin{aligned} \nu_t(s) &= \frac{16\rho(p_{t,s})^{1-\alpha}}{(1-\alpha)\beta_t(s)} \sum_{a} p_{t,s,a} C_t(s,a)^2 \\ &\le \frac{64H^2 L_{\sigma} \rho(p_{t,s})^{1-\alpha}}{(1-\alpha)\beta_t(s)} \sum_{a} p_{t,s,a} C_t(s,a) \le \frac{64H^2 L_{\sigma}}{(1-\alpha)\beta_0} \sum_{a} p_{t,s,a} C_t(s,a) . \end{aligned}$$

The choice of β_0 in Section 3 implies

$$\frac{64H^2L_\sigma}{(1-\alpha)\beta_0} = \frac{64H^2L_\sigma}{2^{12}eK_{\text{saf}}H^3A^{1-\alpha}L_\sigma} = \frac{1}{64eK_{\text{saf}}HA^{1-\alpha}} \le \frac{1}{H}\,.$$

The equalities substitute β_0 and simplify, and the inequality uses $K_{\text{saf}} \ge 1$ and $A^{1-\alpha} \ge 1$. Consequently,

$$\nu_t(s) \le \frac{1}{H}\sum_a p_{t,s,a}C_t(s,a)\,. \tag{192}$$

We now evaluate the occupancy-weighted continuation term in (192). Direct backward unrolling of (8) gives, for $s \in \mathcal{S}_h$,

$$C_t(s,a) = 2\mathbb{E}_{P,\pi_t}\left[\sum_{j=h+1}^{H-1} c_t(s_j)\middle| s_h = s, a_h = a\right]\,. \tag{193}$$

Multiplying (193) by $\mu^{\pi_t}(s)p_{t,s,a}$, summing over state-action pairs, and then reordering the two layer sums gives

$$\begin{aligned}\sum_s \mu^{\pi_t}(s)\sum_a p_{t,s,a}C_t(s,a) &= 2\mathbb{E}_{P,\pi_t}\left[\sum_{h=0}^{H-2}\sum_{j=h+1}^{H-1} c_t(s_j)\right]\\ &= 2\mathbb{E}_{P,\pi_t}\left[\sum_{j=1}^{H-1} j\, c_t(s_j)\right]\\ &\le 2H\mathbb{E}_{P,\pi_t}\left[\sum_{j=1}^{H-1} c_t(s_j)\right] \le 2H\sum_s \mu^{\pi_t}(s)c_t(s)\,.\end{aligned} \tag{194}$$

The first equality takes the occupancy expectation of (193); the second counts the j predecessor layers of layer j. The inequalities use $j \le H$, convert back to occupancies, and include all nonterminal states.

Substituting (194) into the sum of (192) over states gives

$$\sum_s \mu^{\pi_t}(s)\nu_t(s) \le \frac{1}{H}\sum_s \mu^{\pi_t}(s)\sum_a p_{t,s,a}C_t(s,a) \le 2\sum_s \mu^{\pi_t}(s)c_t(s)\,. \tag{195}$$

Finally, the direct calculation (154)– (155) gives $\sum_s \mu^{\pi_t}(s)c_t(s) \le 2SHL_\sigma\gamma_t$. Substituting this bound into (195) yields

$$\sum_s \mu^{\pi_t}(s)\nu_t(s) \le 4SHL_\sigma\gamma_t\,,$$

which is (190). □

C PROOFS FOR THE UNKNOWN-TRANSITION CASE

The following formal statements specialize the unified results in Section 4 to unknown transitions and include all additive remainders. The exact coordinatewise conversion is stated here as a supporting theorem.

Notation and Explicit Remainder Budgets The order summaries below use $H \le S$, $1 < \varepsilon \le 2$, and $\mathcal{L}_T = \mathcal{O}(\log(eHSAT)\log(eT))$. Set

$$\mathsf{G}_T := \Gamma(1+\log(eT)), \qquad \Lambda_T := 1+\varepsilon\log(eT), \qquad \mathcal{L}_T := \iota(2+\log(eT)). \tag{196}$$

Let c_{occ} be a sufficiently large universal constant used in Lemma D.13 and in the quadratic conversion below, and define

$$\mathsf{X}_T := c_{\mathrm{occ}} HS^2 A\mathcal{L}_T, \qquad \mathsf{Y}_T := c_{\mathrm{occ}} HS^4 A\mathcal{L}_T, \qquad \mathsf{Z}_T := 2\big(\mathsf{Y}_T + \mathsf{G}_T + 1\big), \tag{197}$$
$$\mathsf{Z}_T = \mathcal{O}_\varepsilon\big(HS^4 A\log(eHSAT)\log(eT)\big).$$

For a deterministic policy π, recall the definition of $\mathsf{D}_T(\pi)$ from Section 4. Let

$$c_{\mathrm{cnt},\varepsilon} := 3\left(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\right), \qquad \mathsf{A}_T := c_{\mathrm{cnt},\varepsilon} HL_\sigma \Lambda_T^{1-\alpha}. \tag{198}$$

For the adversarial specialization, define the complete additive budget

$$\mathsf{R}^U_{\mathrm{adv}} := \mathsf{R}_0^{\mathrm{unk}} + \mathsf{A}_T(SA)^{1-\alpha}\left[(SA\mathsf{Z}_T)^\alpha + \frac{(SA\mathsf{X}_T)^\alpha}{4}\right]. \tag{199}$$

The definition of $\mathsf{R}_0^{\mathrm{unk}}$ is given in (209) below.

For the full-sequence analysis, define the deterministic inverse-learning-rate cap

$$\beta_{\max} := \beta_0 + c_\beta HL_\sigma\left(1+\sum_{t=1}^{T}\gamma_t^{1-\varepsilon}\right)^{\frac{1}{\varepsilon}}. \tag{200}$$

The deterministic envelopes below have two different scopes. The quantities $\mathsf{F}_{\mathrm{stab}}$, $\mathsf{F}_{\mathrm{clip}}$, and $\mathsf{F}_{\mathrm{bias}}$ are uniform one-state bounds: the first two dominate $\mathsf{e}_t^{\mathrm{stab}}(s)$ and $\mathsf{e}_t^{\mathrm{clip}}(s)$, while the third dominates the fixed-state signed-bias inner product. By contrast, $\mathsf{F}_{\mathrm{dil}}$ and F_ν directly dominate complete one-episode quantities. This distinction is reflected in the definition of $\mathsf{F}_{\mathrm{seq}}$:

$$\mathsf{F}_{\mathrm{stab}} := \frac{64}{1-\alpha}\beta_{\max}A^{1-\alpha}, \tag{201}$$
$$\mathsf{F}_{\mathrm{clip}} := 4k_\alpha^{1-\varepsilon}H^\varepsilon\sigma\beta_0^{1-\varepsilon}A^{1-\alpha}\overline{T}^\varepsilon + 2HL_\sigma\overline{T}, \tag{202}$$
$$\mathsf{F}_{\mathrm{bias}} := 4HL_\sigma(1+\overline{T}) + 8H^2L_\sigma, \tag{203}$$
$$\mathsf{F}_{\mathrm{dil}} := (1+3e)K_{\mathrm{saf}}H^3L_\sigma A^{1-\alpha}\overline{T}^{1-\alpha}, \tag{204}$$
$$\mathsf{F}_\nu := K_{\mathrm{saf}}H^2L_\sigma A^{1-\alpha}\overline{T}^{1-\alpha}, \tag{205}$$
$$\mathsf{F}_{\mathrm{seq}} := H\big(\mathsf{F}_{\mathrm{stab}} + \mathsf{F}_{\mathrm{clip}} + \mathsf{F}_{\mathrm{bias}}\big) + \mathsf{F}_{\mathrm{dil}} + 3\mathsf{F}_\nu \tag{206}$$
$$= \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}\left[H^4A^{2-\frac{2}{\varepsilon}} + H^2A^{1-\frac{1}{\varepsilon}}\overline{T} + H^{4-2\varepsilon}A^{(1-\frac{1}{\varepsilon})(2-\varepsilon)}\overline{T}^\varepsilon\right]\Big).$$

In (206), $\sum_s \mu^\pi(s) = H$ converts the first three statewise envelopes into episodewise bounds, while $\mathsf{F}_{\mathrm{dil}}$ and F_ν are already episodewise; the factor three is the coefficient of the squared-C_t correction in the global reduction.

There is a constant $c_{\nu,\varepsilon} > 0$, depending only on ε, such that

$$\mathsf{R}_\nu := c_{\nu,\varepsilon}H^3L_\sigma\left(\sqrt{H\mathsf{X}_T(1+\log(eT))} + \mathsf{Y}_T + S(\Gamma+1)\right) \tag{207}$$
$$= \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}H^4S^4A\log(eHSAT)\log T\Big).$$

To prevent distinct proof coefficients from being hidden under the same symbol, fix positive constants $c_{\varepsilon,0}, c_{\varepsilon,1}, c_{\varepsilon,2}, c_{\varepsilon,3}, c_{\varepsilon,4}$, all depending only on ε, with

$$c_{\varepsilon,2} \coloneqq 2c_{\varepsilon,1}^2 . \tag{208}$$

The roles of these constants are fixed as follows: $c_{\varepsilon,0}$ multiplies the aggregate additive remainder budget, $c_{\varepsilon,1}$ multiplies the transition term in the master bound, $c_{\varepsilon,2}$ defines $\mathcal{Z}$, $c_{\varepsilon,3}$ multiplies the coordinatewise conversion remainder, and $c_{\varepsilon,4}$ is the constant in the final absorption bound. Define

$$\begin{aligned} \mathsf{R}_0^{\mathrm{unk}} &\coloneqq c_{\varepsilon,0}\Big[HA^{1-\alpha}\beta_0 + SHL_\sigma \mathsf{G}_T + L_\sigma H^2 S^4 A^2 \mathcal{L}_T + \mathsf{R}_\nu + T\delta\,\mathsf{F}_{\mathrm{seq}} + 1\Big] \\ &= \mathcal{O}_\varepsilon\Big(1 + \sigma^{\frac{1}{\varepsilon}} H^2 S^4 A(H^2+A)\log(HSAT)\log T\Big) . \end{aligned} \tag{209}$$

Explicit Parameter Dependence of the Additive Remainder Budgets Recall from Section 3 that $\iota = \log\big(16HS^2A\overline{T}^{66}\big)$ and $\overline{T} = \max\{2, T\}$. Consequently,

$$\mathcal{L}_T = \iota(2 + \log(eT)) , \qquad \mathsf{G}_T = \Gamma(1 + \log(eT)) . \tag{210}$$

The recalled expression for ι follows by substituting $\delta = \overline{T}^{-64}$ into its definition in Section 3. Introduce the ε-only constants

$$b_{0,\varepsilon} \coloneqq \frac{2^{12} e K_{\mathrm{saf}}}{1-\alpha} , \qquad g_{0,\varepsilon} \coloneqq \left(\frac{2^{12} e K_{\mathrm{saf}}}{(1-\alpha)c_\beta}\right)^{\varepsilon} .$$

Substituting the definitions of β_0 and Γ in Section 3, $\alpha = 1/\varepsilon$, and $L_\sigma = \sigma^{1/\varepsilon}$ gives

$$\beta_0 = b_{0,\varepsilon} H^3 A^{1-\frac{1}{\varepsilon}} \sigma^{\frac{1}{\varepsilon}} , \tag{211}$$

$$\Gamma = \big\lceil g_{0,\varepsilon} H^{2\varepsilon} A^{\varepsilon-1} \big\rceil = \mathcal{O}_\varepsilon\big(1 + H^{2\varepsilon} A^{\varepsilon-1}\big) . \tag{212}$$

Moreover, the definition $\gamma_t = \min\{1, \Gamma/t\}$ yields

$$\begin{aligned} 1 + \sum_{t=1}^{T} \gamma_t^{1-\varepsilon} &= 1 + \sum_{t=1}^{T} \left(\max\left\{1, \frac{t}{\Gamma}\right\}\right)^{\varepsilon-1} \\ &\le 1 + \min\{T, \Gamma\} + \Gamma^{1-\varepsilon} \sum_{\Gamma < t \le T} t^{\varepsilon-1} \le 1 + \Gamma + \frac{\overline{T}^{\varepsilon}}{\Gamma^{\varepsilon-1}} . \end{aligned} \tag{213}$$

Substituting (211) and (213) into (200) gives

$$\beta_{\max} \le \sigma^{\frac{1}{\varepsilon}} \left[b_{0,\varepsilon} H^3 A^{1-\frac{1}{\varepsilon}} + c_\beta H \left(1 + \Gamma + \frac{\overline{T}^{\varepsilon}}{\Gamma^{\varepsilon-1}}\right)^{\frac{1}{\varepsilon}} \right] .$$

The subadditivity of $x \mapsto x^{1/\varepsilon}$, the lower bound $\Gamma \ge 1$, and (212) further imply the parameter-only form

$$\beta_{\max} = \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}} \Big[H^3 A^{1-\frac{1}{\varepsilon}} + H\overline{T}\Big]\Big) .$$

Indeed, $(1 + \Gamma + \overline{T}^{\varepsilon}/\Gamma^{\varepsilon-1})^{1/\varepsilon} \le 1 + \Gamma^{1/\varepsilon} + \overline{T}$ and $\Gamma^{1/\varepsilon} = \mathcal{O}_\varepsilon(H^2 A^{1-1/\varepsilon})$.

Using (197) and (207), one obtains

$$\mathsf{R}_\nu = \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}} \Big[H^4 S\sqrt{A\mathcal{L}_T(1+\log(eT))} + H^4 S^4 A \mathcal{L}_T + H^3 S\big(1 + H^{2\varepsilon} A^{\varepsilon-1}\big)\Big]\Big) . \tag{214}$$

The full-sequence confidence-failure budget has the following explicit parameter dependence. Direct substitution of (211) into the first term of (202) gives

$$H^\varepsilon \sigma \beta_0^{1-\varepsilon} A^{1-\alpha} = H^\varepsilon L_\sigma^\varepsilon \left(b_{0,\varepsilon} H^3 A^{1-\frac{1}{\varepsilon}} L_\sigma\right)^{1-\varepsilon} A^{1-\frac{1}{\varepsilon}} = b_{0,\varepsilon}^{1-\varepsilon} L_\sigma H^{3-2\varepsilon} A^{(1-\frac{1}{\varepsilon})(2-\varepsilon)} . \tag{215}$$

The first equality in (215) uses $\sigma = L_\sigma^\varepsilon$ and substitutes (211); the second equality collects the powers of H, A, and L_σ. Substituting (215) and the parameter-only bound on $\beta_{\max}$ into (201)–(206) yields

$$\mathsf{F}_{\mathrm{seq}} = \mathcal{O}_\varepsilon\Big(L_\sigma\big[H^4A^{2-\frac{2}{\varepsilon}} + H^2A^{1-\frac{1}{\varepsilon}}\overline{T} + H^{4-2\varepsilon}A^{(1-\frac{1}{\varepsilon})(2-\varepsilon)}\overline{T}^{\varepsilon} + H^2\overline{T} + H^3 + H^3A^{1-\frac{1}{\varepsilon}}\overline{T}^{1-\frac{1}{\varepsilon}}\big]\Big). \tag{216}$$

The shorter summary in (206) uses $\overline{T}^{1-1/\varepsilon} \le \sqrt{\overline{T}}$ and $2xy \le x^2 + y^2$ to absorb the last term into the first two terms. Since $T \le \overline{T}$ and $\delta = \overline{T}^{-64}$, multiplying (216) by $T\delta$ gives

$$\begin{aligned} T\delta\,\mathsf{F}_{\mathrm{seq}} &= \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}\big[H^4A^{2-\frac{2}{\varepsilon}}\overline{T}^{-63} + H^2A^{1-\frac{1}{\varepsilon}}\overline{T}^{-62} + H^{4-2\varepsilon}A^{(1-\frac{1}{\varepsilon})(2-\varepsilon)}\overline{T}^{\varepsilon-63} \\ &\qquad + H^2\overline{T}^{-62} + H^3\overline{T}^{-63} + H^3A^{1-\frac{1}{\varepsilon}}\overline{T}^{-62-\frac{1}{\varepsilon}}\big]\Big) \\ &= \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}H^4A^{2-\frac{2}{\varepsilon}}\overline{T}^{\varepsilon-63}\Big). \end{aligned} \tag{217}$$

Finally, substituting (211), (212), (214), and (217) into (209) gives

$$\begin{aligned} \mathsf{R}_0^{\mathrm{unk}} = \mathcal{O}_\varepsilon\Bigg(1 + \sigma^{\frac{1}{\varepsilon}}\Bigg[&H^4A^{2-\frac{2}{\varepsilon}} + SH\big(1 + H^{2\varepsilon}A^{\varepsilon-1}\big)(1+\log(eT)) + H^2S^4A^2\mathcal{L}_T \\ &+ H^4S\sqrt{A\mathcal{L}_T(1+\log T)} + H^4S^4A\mathcal{L}_T + H^3S\big(1 + H^{2\varepsilon}A^{\varepsilon-1}\big) + H^4A^{2-\frac{2}{\varepsilon}}\overline{T}^{-63} \\ &+ H^2A^{1-\frac{1}{\varepsilon}}\overline{T}^{-62} + H^{4-2\varepsilon}A^{(1-\frac{1}{\varepsilon})(2-\varepsilon)}\overline{T}^{\varepsilon-63} + H^2\overline{T}^{-62} + H^3\overline{T}^{-63} \\ &+ H^3A^{1-\frac{1}{\varepsilon}}\overline{T}^{-62-\frac{1}{\varepsilon}}\Bigg]\Bigg). \end{aligned} \tag{218}$$

The shorter summaries at the definitions follow because $\Gamma = \mathcal{O}_\varepsilon(H^{2\varepsilon}A^{\varepsilon-1})$ and $H \le S$ imply $\Gamma = \mathcal{O}_\varepsilon(HS^4A)$ and $H^3S\Gamma = \mathcal{O}_\varepsilon(H^4S^4A)$. The confidence-failure contribution is absorbed by $\sigma^{1/\varepsilon}H^4A^{2-2/\varepsilon}$ since $\overline{T}^{\varepsilon-63} \le 1$.

Theorem C.1 (Formal Version of Theorem 4.1 for Unknown Transitions)**.** *Under Definition 2.1, Algorithm 1 satisfies, for every deterministic policy π,*

$$\mathrm{Reg}_T(\pi) \le \mathsf{A}_T \sum_s \sum_{a \ne \pi(s)} \widetilde{n}_T(s,a)^\alpha + c_{\varepsilon,1}HL_\sigma\sqrt{\mathsf{X}_T\,\mathsf{D}_T(\pi)} + \mathsf{R}_0^{\mathrm{unk}}. \tag{219}$$

Theorem C.2 (Coordinatewise upper-occupancy conversion)**.** *For every state-action pair,*

$$\widetilde{n}_T(s,a) \le n_T(s,a) + \sqrt{\mathsf{X}_T\,n_T(s,a)} + \mathsf{Z}_T. \tag{220}$$

Consequently,

$$\widetilde{n}_T(s,a)^\alpha \le n_T(s,a)^\alpha + \mathsf{X}_T^{\frac{\alpha}{2}}n_T(s,a)^{\frac{\alpha}{2}} + \mathsf{Z}_T^\alpha. \tag{221}$$

For the fast-rate result under Definition 2.2, set

$$q_\varepsilon := \frac{\alpha}{2} = \frac{1}{2\varepsilon}. \tag{222}$$

Over all pairs $i = (s,a)$ with $a \ne \pi^\star(s)$, recall from Section 1 that

$$\omega_{1/\varepsilon}(\Delta) = \sum_i \Delta_i^{-\frac{\alpha}{1-\alpha}} = \sum_i \Delta_i^{-\frac{1}{\varepsilon-1}}, \tag{223}$$

$$\omega_{1/(2\varepsilon)}(\Delta) = \sum_i \Delta_i^{-\frac{q_\varepsilon}{1-q_\varepsilon}} = \sum_i \Delta_i^{-\frac{1}{2\varepsilon-1}}. \tag{224}$$

We also recall the definition of $\Delta_{\min}$ from Section 1. Let

$$\mathcal{U} := \mathsf{A}_T^{\frac{1}{1-\alpha}}\omega_{1/\varepsilon}(\Delta), \tag{225}$$

$$\mathsf{B}_T := \mathsf{A}_T\mathsf{X}_T^{\frac{\alpha}{2}}, \tag{226}$$

$$\mathcal{V} \coloneqq \mathsf{B}_T^{\frac{1}{1-q_\varepsilon}}\,\omega_{1/(2\varepsilon)}(\Delta)\,, \tag{227}$$

$$\mathcal{Z} \coloneqq \frac{c_{\varepsilon,2}L_\sigma^2 H^3 \mathsf{X}_T}{\Delta_{\min}}\,, \tag{228}$$

$$\begin{aligned}\mathsf{R}_{\mathrm{const}} &\coloneqq \mathsf{R}_0^{\mathrm{unk}} + c_{\varepsilon,3}\mathsf{A}_T S(A-1)\mathsf{Z}_T^{\alpha} \\ &= \mathsf{R}_0^{\mathrm{unk}} + \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}} H^{1+\frac{1}{\varepsilon}} S^{1+\frac{4}{\varepsilon}} A^{1+\frac{1}{\varepsilon}} \big(\log(eHSAT)\big)^{\frac{1}{\varepsilon}} \log T\Big)\,.\end{aligned} \tag{229}$$

Theorem C.3 (Formal Version of Theorem 4.3 for Unknown Transitions)**.** *Under Definition 2.1 and unknown transitions, Algorithm 1 satisfies the following regret guarantees. Let $m \coloneqq S(A-1) \le SA$. For every adaptive adversarial loss sequence,*

$$\begin{aligned}\max_\pi \mathrm{Reg}_T(\pi) \le \mathsf{A}_T m^{1-\alpha}\Big(HT + \sqrt{SA\,\mathsf{X}_T\,HT} + SA\,\mathsf{Z}_T\Big)^{\alpha} \\ + c_{\varepsilon,1}\sqrt{c_{\mathrm{occ}}}L_\sigma H^2 S\sqrt{AT\mathcal{L}_T} + \mathsf{R}_0^{\mathrm{unk}}\,.\end{aligned} \tag{230}$$

In particular,

$$\max_\pi \mathrm{Reg}_T(\pi) = \widetilde{\mathcal{O}}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}\Big[H^{1+\frac{1}{\varepsilon}}(SA)^{1-\frac{1}{\varepsilon}}T^{\frac{1}{\varepsilon}} + H^2 S\sqrt{AT}\Big]\Big) + \mathsf{R}_{\mathrm{adv}}^{U}\,. \tag{231}$$

If Definition 2.2 further holds, then

$$\mathrm{Reg}_T(\pi^\star) \le c_{\varepsilon,4}\Big(\mathcal{U} + \mathcal{U}^{1-\alpha}\mathcal{C}^{\alpha} + \mathcal{V} + \mathcal{V}^{1-q_\varepsilon}\mathcal{C}^{q_\varepsilon} + \mathcal{Z} + \sqrt{\mathcal{Z}\mathcal{C}} + \mathsf{R}_{\mathrm{const}}\Big)\,. \tag{232}$$

The four named complexity quantities in (232) have the explicit parameter dependence

$$\mathcal{U} = \mathcal{O}_\varepsilon\Big(H^{\frac{\varepsilon}{\varepsilon-1}}\sigma^{\frac{1}{\varepsilon-1}}\Lambda_T\omega_{1/\varepsilon}(\Delta)\Big)\,, \tag{233}$$

$$\mathcal{V} = \mathcal{O}_\varepsilon\left(H^{\frac{2\varepsilon+1}{2\varepsilon-1}}\sigma^{\frac{2}{2\varepsilon-1}}S^{\frac{2}{2\varepsilon-1}}A^{\frac{1}{2\varepsilon-1}}\Lambda_T^{\frac{2(\varepsilon-1)}{2\varepsilon-1}}\mathcal{L}_T^{\frac{1}{2\varepsilon-1}}\omega_{1/(2\varepsilon)}(\Delta)\right)\,, \tag{234}$$

$$\mathcal{Z} = \mathcal{O}_\varepsilon\left(\frac{\sigma^{\frac{2}{\varepsilon}}H^4S^2A\mathcal{L}_T}{\Delta_{\min}}\right)\,, \tag{235}$$

$$\mathsf{R}_{\mathrm{const}} = \mathsf{R}_0^{\mathrm{unk}} + \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}\Lambda_T^{1-\frac{1}{\varepsilon}}\big[H^{1+\frac{1}{\varepsilon}}S^{1+\frac{4}{\varepsilon}}A^{1+\frac{1}{\varepsilon}}\mathcal{L}_T^{\frac{1}{\varepsilon}} + H^3SA^{2-\frac{1}{\varepsilon}}(1+\log(eT))^{\frac{1}{\varepsilon}}\big]\Big)\,. \tag{236}$$

Remark C.4 (Known-transition specialization)**.** *If the transition kernel is known, take $\mathcal{P}_t = \{P\}$ and redo the confidence lemmas with zero occupancy width. Then $\mathsf{X}_T = \mathsf{Y}_T = 0$ in the analysis, and the transition-width terms $\mathcal{V}$ and $\mathcal{Z}$ disappear. Moreover, $\mathcal{E}_t$ and $\mathcal{E}$ hold surely, so every term multiplied by $\mathbb{1}\{\mathcal{E}_t^c\}$ or $\mathbb{1}\{\mathcal{E}^c\}$—and hence the entire $T\delta\mathsf{F}_{\mathrm{seq}}$ correction—vanishes exactly rather than merely being negligible. The remaining visitation-floor compensation is the one already present in the known-transition construction.*

C.1 Proof of the Main Results

C.1.1 Proof of the Master Bound

Proof of Theorem C.1. Fix an arbitrary deterministic comparator policy π. We expand the proof sketch of Theorem 4.1 in the same order, retaining the initial regularizer and confidence-failure terms. Write $\mathcal{E}_t = \{P \in \mathcal{P}_t\}$ and $\mathcal{E} = \bigcap_{t=1}^T \mathcal{E}_t$; (478) gives $\mathbb{P}(\mathcal{E}^c) \le \delta$. All state sums range over nonterminal states, and occupancies without an explicit kernel use P. Recall $p_{t,s} = \pi_t(\cdot \mid s)$, and write $B_t^\pi(s) \coloneqq B_t(s,\pi(s))$, $\overline{B}_t(s) \coloneqq \langle p_{t,s}, B_t(s,\cdot)\rangle$, and $\kappa_H \coloneqq 1 + 1/H$. We start with the regret decomposition:

$$\begin{aligned}\mathrm{Reg}_T(\pi) &= \mathbb{E}\left[\sum_{t=1}^T \big(V_t^{\pi_t}(s_0) - V_t^{\pi}(s_0)\big)\right] = \mathbb{E}\left[\sum_{t,s}\mu^\pi(s)\,\big\langle p_{t,s} - e_{\pi(s)}, \bar{Q}_t(s,\cdot)\big\rangle\right] \\ &= \sum_s \mu^\pi(s)\big(\mathsf{Loc}^\pi(s) + \mathsf{Bon}^\pi(s)\big)\,,\end{aligned} \tag{237}$$

where

$$\mathsf{Loc}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, \bar{Q}_t(s,\cdot) - B_t(s,\cdot)\right\rangle\right],$$
$$\mathsf{Bon}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, B_t(s,\cdot)\right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\overline{B}_t(s) - B_t^{\pi}(s)\right)\right].$$

The performance-difference identity proves the second equality in (237); subtracting $V_t^{\pi_t}(s)\mathbf{1}$ from the Q-vector has no effect since $\langle p_{t,s} - e_{\pi(s)}, \mathbf{1}\rangle = 0$. As in the sketch, we first control the local learning term and then use dilation to cancel the comparator-weighted bonus terms.

The estimator mean in Lemma C.8 gives, on every history,

$$r_t(s,a) = \bar{Q}_t(s,a) - \mathbb{E}_t[\widetilde{Q}_t(s,a)] = \frac{\widetilde{\mu}_t(s) - \mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)}\bar{Q}_t(s,a). \tag{238}$$

Recall $\tau_t(s,\cdot) = \mathbb{E}_t[\widetilde{Q}_t(s,\cdot) - \widehat{Q}_t(s,\cdot)]$. Substituting these definitions into (14) gives

$$\bar{Q}_t(s,\cdot) - B_t(s,\cdot) = \mathbb{E}_t[g_t(s,\cdot)] + \tau_t(s,\cdot) + d_t(s,\cdot) + r_t(s,\cdot) + C_t(s,\cdot). \tag{239}$$

Taking the inner product with the $\mathcal{F}_{t-1}$-measurable vector $p_{t,s} - e_{\pi(s)}$ and using the tower property therefore yields

$$\mathsf{Loc}^{\pi}(s) = \mathsf{FTRL\text{-}Regret}^{\pi}(s) + \mathsf{Clipping\text{-}Bias}^{\pi}(s) + \mathsf{Estimator\text{-}Bias}^{\pi}(s), \tag{240}$$

where

$$\mathsf{FTRL\text{-}Regret}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, g_t(s,\cdot)\right\rangle\right],$$
$$\mathsf{Clipping\text{-}Bias}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, \tau_t(s,\cdot) + d_t(s,\cdot)\right\rangle\right],$$
$$\mathsf{Estimator\text{-}Bias}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle\right].$$

Bounding $\mathsf{FTRL\text{-}Regret}^{\pi}(s)$ Apply Lemma D.7 to the entire realized loss sequence at state s. With $\widetilde{p}_{t+1}$ defined using $\beta_t(s)$ as in (464), put

$$\mathsf{S}_t(s) \coloneqq \langle p_{t,s} - \widetilde{p}_{t+1}, g_t(s,\cdot)\rangle - \beta_t(s) D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_{t,s}).$$

The inverse learning rates are nondecreasing by (13), so (466) gives

$$\sum_{t=1}^{T}\left\langle p_{t,s} - e_{\pi(s)}, g_t(s,\cdot)\right\rangle \le \beta_1(s)h_1(s) + \sum_{t=2}^{T}(\beta_t(s) - \beta_{t-1}(s))h_t(s) + \sum_{t=1}^{T}\mathsf{S}_t(s).$$

For each episode, Lemma C.11 bounds $\mathbb{E}_t[\mathsf{S}_t(s)]$ on the predictable event $\mathcal{E}_t$ and retains the correction $\mathbb{1}\{\mathcal{E}_t^c\}\mathsf{e}_t^{\mathrm{stab}}(s)$. Taking expectations gives

$$\mathsf{FTRL\text{-}Regret}^{\pi}(s) \le \mathbb{E}[\beta_1(s)h_1(s)] + \mathbb{E}\left[\sum_{t=2}^{T}(\beta_t(s) - \beta_{t-1}(s))h_t(s)\right]$$
$$+ \mathbb{E}\left[\sum_{t=1}^{T}\left[\frac{64k_\alpha^{2-\varepsilon}}{1-\alpha}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s) + \nu_t(s) + \frac{\overline{B}_t(s)}{H} + \mathbb{1}\{\mathcal{E}_t^c\}\mathsf{e}_t^{\mathrm{stab}}(s)\right]\right]. \tag{241}$$

This calculation takes expectations of the full pathwise FTRL inequality; it does not condition that inequality on the whole-horizon event $\mathcal{E}$.

Bounding $\mathsf{Clipping\text{-}Bias}^{\pi}(s)$ On $\mathcal{E}_t$, Lemma C.8 gives $|\tau_t(s,a)| \le d_t(s,a)$. Thus, for $a^\star = \pi(s)$, the comparator coordinate contributes a nonpositive term, and Lemma C.12 gives

$$\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \le 2\sum_{a\neq a^\star} p_{t,s,a} d_t(s,a) \le 4k_\alpha^{1-\varepsilon}\Sigma_t(s)\beta_t(s)^{1-\varepsilon} z_t(s) \quad \text{on } \mathcal{E}_t\,.$$

The same lemma controls the complementary event by $\mathsf{e}_t^{\mathrm{clip}}(s)$. Summing its full-sequence bound (349) yields

$$\mathsf{Clipping\text{-}Bias}^{\pi}(s) \le \mathbb{E}\left[\sum_{t=1}^{T}\left[4k_\alpha^{1-\varepsilon}\Sigma_t(s)\beta_t(s)^{1-\varepsilon} z_t(s) + \mathbb{1}\{\mathcal{E}_t^c\}\mathsf{e}_t^{\mathrm{clip}}(s)\right]\right]. \tag{242}$$

By (40), the sum of the stability and clipping coefficients is at most K_{ht}. The changing-rate term, both heavy-tail terms, and $\nu_t(s)$ therefore fit into $b_t(s)$ in (10). Substituting into (240) gives

$$\begin{aligned}\mathsf{Loc}^{\pi}(s) \le \mathbb{E}[\beta_1(s)h_1(s)] + \mathbb{E}\left[\sum_{t=1}^{T}\left(b_t(s) + \frac{\overline{B}_t(s)}{H}\right)\right] + \mathsf{Estimator\text{-}Bias}^{\pi}(s)\\ + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}_t^c\}\left(\mathsf{e}_t^{\mathrm{stab}}(s) + \mathsf{e}_t^{\mathrm{clip}}(s)\right)\right].\end{aligned} \tag{243}$$

The argument also holds with any comparator action $a^\star$.

Substituting (243) into (237) now gives

$$\mathrm{Reg}_T(\pi) \le \mathcal{I}_0(\pi) + \mathbb{E}\left[\sum_{t=1}^{T}\mathcal{D}_t(\pi)\right] + \mathsf{S}(\pi) + \mathcal{F}_{\mathrm{loc}}(\pi)\,, \tag{244}$$

where $\mathcal{I}_0(\pi) := \mathbb{E}\left[\sum_s \mu^\pi(s)\beta_1(s)h_1(s)\right]$, $\mathsf{S}(\pi) := \sum_s \mu^\pi(s)\mathsf{Estimator\text{-}Bias}^{\pi}(s)$, and $\mathcal{D}_t(\pi) := \sum_s \mu^\pi(s)[b_t(s) + \kappa_H\overline{B}_t(s) - B_t^\pi(s)]$. Here $\mathcal{F}_{\mathrm{loc}}(\pi)$ is exactly (250), the comparator-weighted sum of the last line of (243). Lemma C.5 gives $\mathcal{F}_{\mathrm{loc}}(\pi) \le HT\delta(\mathsf{F}_{\mathrm{stab}} + \mathsf{F}_{\mathrm{clip}})$.

On $\mathcal{E}_t$, the true kernel is feasible in (12). Consequently, $B_t^\pi(s) \ge b_t(s) + \kappa_H \mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[\overline{B}_t(s')\right]$. The comparator occupancy flow then cancels the noninitial terms, and unrolling under the maximizing kernel P_t^B gives

$$\mathcal{D}_t(\pi) \le \kappa_H\overline{B}_t(s_0) \le \kappa_H^H V^{P_t^B,\pi_t}(s_0; b_t) \le 3V^{P_t^B,\pi_t}(s_0; b_t) \qquad \text{on } \mathcal{E}_t\,.$$

This is Lemma D.4 in Section D.2. On $\mathcal{E}_t^c$, take $|\mathcal{D}_t(\pi)|$ and use (256). Since $b_t \ge 0$, the good-event indicator can be dropped from its value, yielding

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathcal{D}_t(\pi)\right] \le 3\mathbb{E}\left[\sum_{t=1}^{T}V^{P_t^B,\pi_t}(s_0; b_t)\right] + T\delta\mathsf{F}_{\mathrm{dil}}\,.$$

Thus (244) becomes

$$\mathrm{Reg}_T(\pi) \le \mathcal{I}_0(\pi) + 3\mathbb{E}\left[\sum_{t=1}^{T}V^{P_t^B,\pi_t}(s_0; b_t)\right] + \mathsf{S}(\pi) + T\delta\left[H(\mathsf{F}_{\mathrm{stab}} + \mathsf{F}_{\mathrm{clip}}) + \mathsf{F}_{\mathrm{dil}}\right]. \tag{245}$$

Bounding $\mathsf{Estimator\text{-}Bias}^{\pi}(s)$ Introduce $\mathcal{L}_t^{\mathrm{bias}}(s,a) = r_t(s,a) + C_t(s,a) + c_t(s)$ and its Bellman residual $\zeta_t(s,a) = \mathcal{L}_t^{\mathrm{bias}}(s,a) - \mathbb{E}_{P,\pi_t}[\mathcal{L}_t^{\mathrm{bias}}(s',a') \mid s,a]$. On $\mathcal{E}$, Lemma C.14 gives $\zeta_t \ge 0$ for every t, and $\mathcal{L}_t^{\mathrm{bias}}$ is the Q-function of π_t for loss ζ_t. Adding $c_t(s)\mathbf{1}$ has no effect on the local inner product. The performance-difference identity therefore gives, on $\mathcal{E}$,

$$\begin{aligned}\sum_s \mu^\pi(s)\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle &= \sum_{s,a}\left(\mu^{\pi_t}(s,a) - \mu^\pi(s,a)\right)\zeta_t(s,a)\\ &\le \sum_{s,a}[\mu^{\pi_t}(s,a) - \mu^\pi(s,a)]_+\,\zeta_t(s,a)\,.\end{aligned}$$

Keeping this positive occupancy discrepancy is what gives the comparator-sensitive bound in Lemma C.13. Apply that lemma on $\mathcal{E}$, and bound the absolute contribution of $\mathcal{E}^c$ by (255). We obtain

$$\begin{aligned}\mathsf{S}(\pi) \le c_{\varepsilon,1} H L_\sigma \sqrt{\mathsf{X}_T\, \mathsf{D}_T(\pi)} + c_\varepsilon L_\sigma H^2 S^4 A^2 \mathcal{L}_T \\ + c_\varepsilon H L_\sigma (\mathsf{Y}_T + S\mathsf{G}_T) + HT\delta \mathsf{F}_{\mathrm{bias}}\,.\end{aligned} \tag{246}$$

Substituting this into (245) yields

$$\begin{aligned}\mathrm{Reg}_T(\pi) \le \mathcal{I}_0(\pi) + c_{\varepsilon,1} H L_\sigma \sqrt{\mathsf{X}_T\, \mathsf{D}_T(\pi)} + 3\mathbb{E}\left[\sum_{t=1}^T V^{P_t^B,\pi_t}(s_0;b_t)\right] + c_\varepsilon L_\sigma H^2 S^4 A^2 \mathcal{L}_T \\ + c_\varepsilon H L_\sigma (\mathsf{Y}_T + S\mathsf{G}_T) + T\delta\left[H(\mathsf{F}_{\mathrm{stab}} + \mathsf{F}_{\mathrm{clip}} + \mathsf{F}_{\mathrm{bias}}) + \mathsf{F}_{\mathrm{dil}}\right]\,.\end{aligned} \tag{247}$$

Bounding $V^{P_t^B,\pi_t}(s_0;b_t)$ Lemma C.17 controls the squared-confidence part:

$$\mathbb{E}\left[\sum_{t=1}^T V^{P_t^B,\pi_t}(s_0;\nu_t)\right] \le \mathsf{R}_\nu + T\delta \mathsf{F}_\nu\,.$$

For the remaining changing-rate and heavy-tail parts, Lemma C.16 gives

$$\mathbb{E}\left[\sum_{t=1}^T V^{P_t^B,\pi_t}(s_0;b_t-\nu_t)\right] \le \left(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\right) H L_\sigma \Lambda_T^{1-\alpha} \sum_s \sum_{a\neq\pi(s)} \widetilde{n}_T(s,a)^\alpha\,.$$

Adding these two component bounds, as in Proposition C.15, and using $\mathsf{A}_T = 3(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon})HL_\sigma\Lambda_T^{1-\alpha}$, gives

$$3\mathbb{E}\left[\sum_{t=1}^T V^{P_t^B,\pi_t}(s_0;b_t)\right] \le \mathsf{A}_T \sum_s \sum_{a\neq\pi(s)} \widetilde{n}_T(s,a)^\alpha + 3\mathsf{R}_\nu + 3T\delta\mathsf{F}_\nu\,.$$

Substituting into (247) gives

$$\mathrm{Reg}_T(\pi) \le \mathsf{A}_T \sum_s \sum_{a\neq\pi(s)} \widetilde{n}_T(s,a)^\alpha + c_{\varepsilon,1} H L_\sigma \sqrt{\mathsf{X}_T\, \mathsf{D}_T(\pi)} + \mathcal{R}_{\mathrm{rem}}\,, \tag{248}$$

where, with $\mathsf{F}_{\mathrm{seq}}$ from (206),

$$\mathcal{R}_{\mathrm{rem}} := \mathcal{I}_0(\pi) + c_\varepsilon L_\sigma H^2 S^4 A^2 \mathcal{L}_T + c_\varepsilon H L_\sigma (\mathsf{Y}_T + S\mathsf{G}_T) + 3\mathsf{R}_\nu + T\delta\mathsf{F}_{\mathrm{seq}}\,. \tag{249}$$

To finish, $\gamma_1 = 1$ and $0 \le \overline{\mu}_1(s) \le 1$ give $1 \le \widetilde{\mu}_1(s) \le 2$, $G_1(s) \le 2$, and $\beta_1(s) \le \beta_0 + 2^\alpha c_\beta H L_\sigma \le 2\beta_0$. The last inequality follows from $2^\alpha c_\beta H L_\sigma/\beta_0 = 2^\alpha/(eK_{\mathrm{saf}}H^2A^{1-\alpha}) \le 1$. Using $h_1(s) \le A^{1-\alpha}/\alpha$ and $\sum_s \mu^\pi(s) = H$, we obtain

$$\mathcal{I}_0(\pi) \le \frac{2}{\alpha} H A^{1-\alpha}\beta_0\,, \qquad H L_\sigma \mathsf{Y}_T \le c_{\mathrm{occ}} L_\sigma H^2 S^4 A^2 \mathcal{L}_T\,.$$

Here the second bound uses (197) and $A \ge 1$. Thus, for the sufficiently large constant $c_{\varepsilon,0}$ in (209),

$$\mathcal{R}_{\mathrm{rem}} \le c_{\varepsilon,0}\left[HA^{1-\alpha}\beta_0 + L_\sigma H^2 S^4 A^2 \mathcal{L}_T + SHL_\sigma \mathsf{G}_T + \mathsf{R}_\nu + T\delta\mathsf{F}_{\mathrm{seq}} + 1\right] = \mathsf{R}_0^{\mathrm{unk}}\,.$$

Combining this with (248) proves (219). □

Confidence-failure Corrections We now bound the four confidence-failure contributions used in the global reduction and the bonus-count bound. The deterministic envelopes are established in Sections C.2 and C.3.

Lemma C.5 (Full-sequence confidence-failure corrections)**.** *For every deterministic comparator policy π, define*

$$\mathcal{F}_{\mathrm{loc}}(\pi) := \mathbb{E}\left[\sum_{t=1}^T \sum_s \mu^\pi(s)\mathbb{1}\{\mathcal{E}_t^c\}\left(\mathsf{e}_t^{\mathrm{clip}}(s) + \mathsf{e}_t^{\mathrm{stab}}(s)\right)\right]\,, \tag{250}$$

$$\mathcal{F}_{\text{bias}}(\pi) := \mathbb{E}\left[\mathbb{1}\{\mathcal{E}^c\}\left|\sum_{t=1}^{T}\sum_{s}\mu^{\pi}(s)\left\langle p_{t,s}-e_{\pi(s)}, r_t(s,\cdot)+C_t(s,\cdot)\right\rangle\right|\right], \tag{251}$$

$$\mathcal{F}_{\text{dil}}(\pi) := \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}_t^c\}|\mathcal{D}_t(\pi)|\right],$$

$$\mathcal{F}_{\nu} := \mathbb{E}\left[\mathbb{1}\{\mathcal{E}^c\}\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right], \tag{252}$$

where

$$\mathcal{D}_t(\pi) := \sum_{s}\mu^{\pi}(s)\left[b_t(s)+\left(1+\frac{1}{H}\right)\langle p_{t,s}, B_t(s,\cdot)\rangle - B_t(s,\pi(s))\right]. \tag{253}$$

Then

$$\mathcal{F}_{\text{loc}}(\pi) \le HT\delta\left(\mathsf{F}_{\text{clip}}+\mathsf{F}_{\text{stab}}\right), \tag{254}$$

$$\mathcal{F}_{\text{bias}}(\pi) \le HT\delta\,\mathsf{F}_{\text{bias}}, \tag{255}$$

$$\mathcal{F}_{\text{dil}}(\pi) \le T\delta\,\mathsf{F}_{\text{dil}}, \tag{256}$$

$$\mathcal{F}_{\nu} \le T\delta\,\mathsf{F}_{\nu}. \tag{257}$$

In particular, the confidence-failure corrections are controlled without a lower bound on any action probability.

Proof. Because $\mathcal{E} = \bigcap_{t=1}^{T}\mathcal{E}_t$, the indicator comparison

$$\mathbb{1}\{\mathcal{E}_t^c\} \le \mathbb{1}\{\mathcal{E}^c\}, \qquad \sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}_t^c\} \le T\mathbb{1}\{\mathcal{E}^c\} \tag{258}$$

holds pathwise. We bound the four quantities in (250)–(252) separately.

For the local stability and clipping corrections, substitute (342) into (250), use the occupancy-mass identity $\sum_s \mu^{\pi}(s) = H$, and then apply (258):

$$\begin{aligned}\mathcal{F}_{\text{loc}}(\pi) &\le \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}_t^c\}\sum_{s}\mu^{\pi}(s)\left(\mathsf{F}_{\text{clip}}+\mathsf{F}_{\text{stab}}\right)\right] = H\left(\mathsf{F}_{\text{clip}}+\mathsf{F}_{\text{stab}}\right)\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}_t^c\}\right]\\ &\le HT\left(\mathsf{F}_{\text{clip}}+\mathsf{F}_{\text{stab}}\right)\mathbb{P}(\mathcal{E}^c) \le HT\delta\left(\mathsf{F}_{\text{clip}}+\mathsf{F}_{\text{stab}}\right).\end{aligned}$$

Here, the equality uses $\sum_s \mu^{\pi}(s) = H$, and the last two inequalities use (258) and (478), respectively, proving (254).

For the signed occupancy-confidence term, the event-free bounds (323) and (336) imply, for every episode,

$$\begin{aligned}&\left|\sum_{s}\mu^{\pi}(s)\left\langle p_{t,s}-e_{\pi(s)}, r_t(s,\cdot)+C_t(s,\cdot)\right\rangle\right|\\ &\le \sum_{s}\mu^{\pi}(s)\left\|p_{t,s}-e_{\pi(s)}\right\|_1\max_{a}\left(|r_t(s,a)|+C_t(s,a)\right)\\ &\le \sum_{s}\mu^{\pi}(s)\,2\left[2HL_{\sigma}(1+\overline{T})+4H^2L_{\sigma}\right] = H\left[4HL_{\sigma}(1+\overline{T})+8H^2L_{\sigma}\right] = H\mathsf{F}_{\text{bias}}.\end{aligned}$$

Here, the first inequality is ℓ_1-ℓ_∞ Hölder, the second uses $\left\|\pi_t - e_{\pi(s)}\right\|_1 \le 2$, (336), and (323), and the equality uses (203). Substituting into (251) gives

$$\mathcal{F}_{\text{bias}}(\pi) \le HT\mathsf{F}_{\text{bias}}\mathbb{E}[\mathbb{1}\{\mathcal{E}^c\}] = HT\mathsf{F}_{\text{bias}}\mathbb{P}(\mathcal{E}^c) \le HT\delta\mathsf{F}_{\text{bias}},$$

which proves (255).

We next control the complete robust-dilation bracket. By nonnegativity of b_t and B_t, the triangle inequality, and $\sum_s \mu^{\pi}(s) = H$, direct substitution of (253) gives

$$|\mathcal{D}_t(\pi)| \le \sum_{s}\mu^{\pi}(s)b_t(s) + \left(1+\frac{1}{H}\right)\sum_{s}\mu^{\pi}(s)\langle p_{t,s}, B_t(s,\cdot)\rangle$$

$$+\sum_s \mu^\pi(s) B_t(s,\pi(s)) \le H b_t^{\max} + \left(2+\frac{1}{H}\right) H \max_{s,a} B_t(s,a)$$
$$\le H b_t^{\max} + 3eH^2 b_t^{\max} \le (1+3e) H^2 K_{\mathrm{saf}} H L_\sigma A^{1-\alpha} \overline{T}^{1-\alpha} = \mathsf{F}_{\mathrm{dil}} .$$

Here, the third inequality uses $2 + H^{-1} \le 3$ and $\max_{s,a} B_t(s,a) \le eHb_t^{\max}$ from (335), the fourth uses the bound on $b_t^{\max}$ there, and the equality uses (204). Hence

$$\mathcal{F}_{\mathrm{dil}}(\pi) \le \mathsf{F}_{\mathrm{dil}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}_t^c\}\right] \le T \mathsf{F}_{\mathrm{dil}} \mathbb{P}(\mathcal{E}^c) \le T\delta \mathsf{F}_{\mathrm{dil}} ,$$

which proves (256).

Finally, $0 \le \nu_t(s) \le b_t^{\max}$ by (336). The occupancy mass under every layered kernel is H, so

$$\sum_s \mu^{P_t^B, \pi_t}(s) \nu_t(s) \le b_t^{\max} \sum_s \mu^{P_t^B, \pi_t}(s) = H b_t^{\max} \le K_{\mathrm{saf}} H^2 L_\sigma A^{1-\alpha} \overline{T}^{1-\alpha} = \mathsf{F}_\nu . \tag{259}$$

Substituting (259) into (252) gives

$$\mathcal{F}_\nu \le T \mathsf{F}_\nu \mathbb{P}(\mathcal{E}^c) \le T\delta \mathsf{F}_\nu ,$$

which proves (257). □

C.1.2 Adversarial Specialization

Proof of (230). Fix an arbitrary deterministic comparator policy π and define $\mathcal{I}_\pi := \{(s,a) : a \ne \pi(s)\}$, $m_\pi := |\mathcal{I}_\pi| \le S(A-1) = m$. The count-summation argument in the proof of (31) in Section B.1 applies to any nonnegative vector $(x_i)_{i \in \mathcal{I}_\pi}$. Keeping the sharper cardinality bound $m_\pi \le m$ gives

$$\sum_{i \in \mathcal{I}_\pi} x_i^\alpha \le m_\pi^{1-\alpha} \left(\sum_{i \in \mathcal{I}_\pi} x_i\right)^\alpha \le m^{1-\alpha} \left(\sum_{i \in \mathcal{I}_\pi} x_i\right)^\alpha . \tag{260}$$

The repeated Hölder calculation is omitted. Applying (260) with $x_i = \widetilde{n}_T(i)$ yields the comparator-sensitive count bound needed below.

We next control the total upper count. Summing the coordinatewise inequality (220) over all SA nonterminal state-action pairs gives the direct substitution

$$\sum_{s,a} \widetilde{n}_T(s,a) \le \sum_{s,a} n_T(s,a) + \sqrt{\mathsf{X}_T} \sum_{s,a} \sqrt{n_T(s,a)} + SA\mathsf{Z}_T . \tag{261}$$

The occupancy-mass calculation in the proof of (31) in Section B.1 also gives

$$\sum_{s,a} n_T(s,a) = HT . \tag{262}$$

Cauchy–Schwarz and (262) imply

$$\sum_{s,a} \sqrt{n_T(s,a)} \le \sqrt{\left(\sum_{s,a} 1\right)\left(\sum_{s,a} n_T(s,a)\right)} = \sqrt{SA\,HT} . \tag{263}$$

Substituting (262) and (263) into (261) gives

$$\sum_{s,a} \widetilde{n}_T(s,a) \le HT + \sqrt{SA\mathsf{X}_T\,HT} + SA\mathsf{Z}_T .$$

Since $\mathcal{I}_\pi$ is a subset of all state-action pairs and all counts are nonnegative,

$$\sum_{i \in \mathcal{I}_\pi} \widetilde{n}_T(i) \le \sum_{s,a} \widetilde{n}_T(s,a) \le HT + \sqrt{SA\mathsf{X}_T\,HT} + SA\mathsf{Z}_T . \tag{264}$$

Combining (260) and (264) yields

$$\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha}\le m^{1-\alpha}\left(HT+\sqrt{SA\,\mathsf{X}_T\,HT}+SA\mathsf{Z}_T\right)^{\alpha}. \tag{265}$$

The comparator-deviation term is bounded episodewise by the learner's total occupancy mass:

$$\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\le\sum_{s,a}\mu^{\pi_t}(s,a)=H. \tag{266}$$

Taking expectations and summing (266) over t gives

$$\mathsf{D}_T(\pi)=\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\right]\le HT. \tag{267}$$

Substitute (265) and (267) into the master bound (219). The complete substituted inequality is

$$\begin{aligned}\mathrm{Reg}_T(\pi)\le{}&\mathsf{A}_T m^{1-\alpha}\left(HT+\sqrt{SA\,\mathsf{X}_T\,HT}+SA\mathsf{Z}_T\right)^{\alpha}\\&+c_{\varepsilon,1}HL_\sigma\sqrt{\mathsf{X}_T\,HT}+\mathsf{R}_0^{\mathrm{unk}}.\end{aligned} \tag{268}$$

Using the definition $\mathsf{X}_T=c_{\mathrm{occ}}HS^2A\mathcal{L}_T$, the transition term in (268) becomes

$$\begin{aligned}c_{\varepsilon,1}HL_\sigma\sqrt{\mathsf{X}_T\,HT}&=c_{\varepsilon,1}HL_\sigma\sqrt{c_{\mathrm{occ}}HS^2A\mathcal{L}_T\,HT}\\&=c_{\varepsilon,1}\sqrt{c_{\mathrm{occ}}}\,L_\sigma H^2S\sqrt{AT\mathcal{L}_T}=\mathcal{O}_\varepsilon\left(L_\sigma H^2S\sqrt{AT\mathcal{L}_T}\right).\end{aligned} \tag{269}$$

Substituting (269) into (268) proves (230). Since the bound holds for every deterministic comparator, taking the maximum over π preserves the same right-hand side.

We finally verify the displayed leading-order form. From (198),

$$\mathsf{A}_T=c_{\mathrm{cnt},\varepsilon}HL_\sigma\Lambda_T^{1-\alpha}=\widetilde{\mathcal{O}}_\varepsilon(HL_\sigma),\qquad \mathsf{X}_T=\widetilde{\mathcal{O}}(HS^2A),\qquad \mathsf{Z}_T=\widetilde{\mathcal{O}}(HS^4A+\Gamma).$$

Since $0<\alpha<1$, subadditivity gives

$$\left(HT+\sqrt{SA\,\mathsf{X}_T\,HT}+SA\mathsf{Z}_T\right)^{\alpha}\le(HT)^{\alpha}+(SA\,\mathsf{X}_T\,HT)^{\frac{\alpha}{2}}+(SA\mathsf{Z}_T)^{\alpha}. \tag{270}$$

The contribution $\mathsf{A}_T m^{1-\alpha}(HT)^{\alpha}$ generated by the summand $(HT)^{\alpha}$ in (270) satisfies

$$\mathsf{A}_T m^{1-\alpha}(HT)^{\alpha}\le c_{\mathrm{cnt},\varepsilon}HL_\sigma\Lambda_T^{1-\alpha}(SA)^{1-\alpha}H^{\alpha}T^{\alpha}=\widetilde{\mathcal{O}}_\varepsilon\left(L_\sigma H^{1+\alpha}(SA)^{1-\alpha}T^{\alpha}\right). \tag{271}$$

With $\alpha=1/\varepsilon$ and $L_\sigma=\sigma^{1/\varepsilon}$, the expression in (271) equals the heavy-tail leading contribution displayed in (231).

For the contribution generated by $(SA\mathsf{X}_THT)^{\alpha/2}$, use $m\le SA$ and $uv\le u^2+v^2/4$ with $u=(HT)^{\alpha/2}$ and $v=(SA\mathsf{X}_T)^{\alpha/2}$:

$$\mathsf{A}_T m^{1-\alpha}(SA\mathsf{X}_THT)^{\alpha/2}\le\mathsf{A}_T(SA)^{1-\alpha}\left[(HT)^{\alpha}+\frac{(SA\mathsf{X}_T)^{\alpha}}{4}\right]. \tag{272}$$

The remaining count contribution is at most $\mathsf{A}_T(SA)^{1-\alpha}(SA\mathsf{Z}_T)^{\alpha}$. Combining these bounds with (270) and (269) gives the complete inequality

$$\max_{\pi}\mathrm{Reg}_T(\pi)\le 2\mathsf{A}_T(SA)^{1-\alpha}(HT)^{\alpha}+c_{\varepsilon,1}\sqrt{c_{\mathrm{occ}}}L_\sigma H^2S\sqrt{AT\mathcal{L}_T}+\mathsf{R}_{\mathrm{adv}}^{U}. \tag{273}$$

This proves (231) with the explicit budget (199); its logarithmic dependence on T has not been discarded. It also gives the unknown-transition specialization of (16). □

C.1.3 SELF-BOUNDING SPECIALIZATION

Occupancy Variation under a Policy Change

Lemma C.6 (Policy-to-occupancy variation)**.** *For any two policies π, π' under the same transition kernel,*

$$\sum_{s,a} |\mu^{\pi}(s,a) - \mu^{\pi'}(s,a)| \le 2H \sum_{s} \mu^{\pi}(s) \sum_{a} |\pi(a \mid s) - \pi'(a \mid s)| \,. \tag{274}$$

Consequently, if $\pi^\star$ is deterministic,

$$\mathsf{D}_T(\pi^\star) \le 2H \sum_{s} \sum_{a \neq \pi^\star(s)} n_T(s,a) \,. \tag{275}$$

Proof. We construct an explicit coupling of the two trajectories generated by π and π' under the common transition kernel P. The two coupled trajectories start at the same initial state s_0. As long as their states and actions have agreed at all previous layers, they occupy the same current state s. At that state, draw their actions from a maximal coupling of $\pi(\cdot \mid s)$ and $\pi'(\cdot \mid s)$. Conditional on the common state s, the probability that the two actions disagree is

$$d_{\mathrm{TV}}(s) := \mathrm{TV}\big(\pi(\cdot \mid s), \pi'(\cdot \mid s)\big) = \frac{1}{2} \sum_{a} |\pi(a \mid s) - \pi'(a \mid s)| \,. \tag{276}$$

If the coupled actions agree, use the same transition draw from $P(\cdot \mid s, a)$. After the first action disagreement, complete the coupling arbitrarily.

Let D_h denote the event that the first action disagreement occurs at layer h, and let $D := \bigcup_{h=0}^{H-1} D_h$. Before the first disagreement, the common state at layer h has the same law as the layer-h state under π. Therefore

$$\mathbb{P}(D_h) = \sum_{s \in \mathcal{S}_h} \mathbb{P}(D_0^c \cap \cdots \cap D_{h-1}^c \,, s_h = s)\, d_{\mathrm{TV}}(s) \le \sum_{s \in \mathcal{S}_h} \mu^{\pi}(s) d_{\mathrm{TV}}(s) \,. \tag{277}$$

The inequality in (277) drops the event that no earlier disagreement occurred. Summing over the disjoint first-disagreement events and substituting (276) gives

$$\begin{aligned} \mathbb{P}(D) &= \sum_{h=0}^{H-1} \mathbb{P}(D_h) \le \sum_{h=0}^{H-1} \sum_{s \in \mathcal{S}_h} \mu^{\pi}(s) d_{\mathrm{TV}}(s) \\ &= \frac{1}{2} \sum_{s} \mu^{\pi}(s) \sum_{a} |\pi(a \mid s) - \pi'(a \mid s)| \,. \end{aligned} \tag{278}$$

For each layer h, let $Z_h = (s_h, a_h)$ and $Z'_h = (s'_h, a'_h)$ denote the two coupled state-action pairs. On D^c the pairs agree at every layer, so the coupling inequality gives

$$\mathrm{TV}\big(\mathcal{L}(Z_h), \mathcal{L}(Z'_h)\big) \le \mathbb{P}(Z_h \neq Z'_h) \le \mathbb{P}(D) \,.$$

The ℓ_1 distance between two probability mass functions equals twice their total-variation distance. Hence, for every layer h,

$$\sum_{s \in \mathcal{S}_h} \sum_{a} |\mu^{\pi}(s,a) - \mu^{\pi'}(s,a)| = 2\,\mathrm{TV}\big(\mathcal{L}(Z_h), \mathcal{L}(Z'_h)\big) \le 2\mathbb{P}(D) \,. \tag{279}$$

Summing (279) over the H nonterminal layers and then substituting (278) yields

$$\begin{aligned} \sum_{s,a} |\mu^{\pi}(s,a) - \mu^{\pi'}(s,a)| &= \sum_{h=0}^{H-1} \sum_{s \in \mathcal{S}_h, a} |\mu^{\pi}(s,a) - \mu^{\pi'}(s,a)| \le 2H\mathbb{P}(D) \\ &\le H \sum_{s} \mu^{\pi}(s) \sum_{a} |\pi(a \mid s) - \pi'(a \mid s)| \\ &\le 2H \sum_{s} \mu^{\pi}(s) \sum_{a} |\pi(a \mid s) - \pi'(a \mid s)| \,. \end{aligned}$$

Here, the second inequality uses (278); the last relaxes the constant to obtain (274).

We now prove the consequence for a deterministic $\pi^\star$. The vectors $\mu^{\pi_t} = (\mu^{\pi_t}(s,a))_{s,a}$ and $\mu^{\pi^\star} = (\mu^{\pi^\star}(s,a))_{s,a}$ are nonnegative and have the same total mass H. Therefore

$$\sum_{s,a}[\mu^{\pi_t}(s,a) - \mu^{\pi^\star}(s,a)]_+ = \frac{1}{2}\sum_{s,a}|\mu^{\pi_t}(s,a) - \mu^{\pi^\star}(s,a)|\,. \tag{280}$$

Because $\pi^\star$ is deterministic, its action distribution at state s is $e_{\pi^\star(s)}$, and hence

$$\begin{aligned}\sum_a |p_{t,s,a} - \pi^\star(a \mid s)| &= |\pi_t(\pi^\star(s) \mid s) - 1| + \sum_{a\neq\pi^\star(s)} p_{t,s,a} \\ &= 2\sum_{a\neq\pi^\star(s)} p_{t,s,a}\,. \end{aligned}\tag{281}$$

Substitute (274) and (281) into (280):

$$\begin{aligned}\sum_{s,a}[\mu^{\pi_t}(s,a) - \mu^{\pi^\star}(s,a)]_+ &\le \frac{1}{2}\left[2H\sum_s \mu^{\pi_t}(s)\sum_a |p_{t,s,a} - \pi^\star(a\mid s)|\right] \\ &= 2H\sum_s \mu^{\pi_t}(s)\sum_{a\neq\pi^\star(s)} p_{t,s,a} = 2H\sum_s\sum_{a\neq\pi^\star(s)}\mu^{\pi_t}(s,a)\,. \end{aligned}\tag{282}$$

According to the definitions of $\mathsf{D}_T(\pi)$ and $n_T(s,a)$ in Section 4, summing (282) over $t = 1,\dots,T$ and taking expectations gives

$$\begin{aligned}\mathsf{D}_T(\pi^\star) &= \mathbb{E}\left[\sum_{t=1}^T\sum_{s,a}[\mu^{\pi_t}(s,a) - \mu^{\pi^\star}(s,a)]_+\right] \\ &\le 2H\sum_s\sum_{a\neq\pi^\star(s)}\mathbb{E}\left[\sum_{t=1}^T \mu^{\pi_t}(s,a)\right] = 2H\sum_s\sum_{a\neq\pi^\star(s)} n_T(s,a)\,,\end{aligned}$$

which is (275). □

A Three-exponent Absorption Lemma

Lemma C.7 (Absorption)**.** *Let $p, q \in (0,1)$ and let $R, \mathcal{C}, U, V, Z, K \ge 0$. If*

$$R \le U^{1-p}(R+\mathcal{C})^p + V^{1-q}(R+\mathcal{C})^q + \sqrt{Z(R+\mathcal{C})} + K\,, \tag{283}$$

then

$$R \le c_{p,q}\Big(U + U^{1-p}\mathcal{C}^p + V + V^{1-q}\mathcal{C}^q + Z + \sqrt{Z\mathcal{C}} + K\Big)\,. \tag{284}$$

Proof. Since $p, q \in (0,1)$, the maps $x \mapsto x^p$ and $x \mapsto x^q$ are subadditive on $\mathbb{R}_+$, and the square-root map is also subadditive. Therefore

$$(R+\mathcal{C})^p \le R^p + \mathcal{C}^p\,, \qquad (R+\mathcal{C})^q \le R^q + \mathcal{C}^q\,, \qquad \sqrt{R+\mathcal{C}} \le \sqrt{R} + \sqrt{\mathcal{C}}\,. \tag{285}$$

Substituting (285) into (283) gives

$$R \le U^{1-p}R^p + U^{1-p}\mathcal{C}^p + V^{1-q}R^q + V^{1-q}\mathcal{C}^q + \sqrt{ZR} + \sqrt{Z\mathcal{C}} + K\,. \tag{286}$$

We next absorb the three R-dependent quantities $U^{1-p}R^p$, $V^{1-q}R^q$, and $\sqrt{ZR}$ appearing in (286). The weighted Young factorization is the same as in the proof of Lemma B.4 in Section B.1. With the scaling parameter $\lambda > 0$, it gives

$$U^{1-p}R^p \le p\lambda R + (1-p)\lambda^{-\frac{p}{1-p}}U\,. \tag{287}$$

To absorb each of the three terms using $R/6$, choose $\lambda_U := 1/(6p)$ in (287):

$$U^{1-p}R^p \le \frac{1}{6}R + (1-p)(6p)^{\frac{p}{1-p}}U\,. \tag{288}$$

Applying the same weighted Young inequality with (V, q) and choosing $\lambda_V := 1/(6q)$ gives, explicitly,

$$V^{1-q}R^q = (\lambda_V R)^q \left(\lambda_V^{-\frac{q}{1-q}} V\right)^{1-q} \le q\lambda_V R + (1-q)\lambda_V^{-\frac{q}{1-q}} V$$
$$= \frac{1}{6}R + (1-q)(6q)^{\frac{q}{1-q}} V\,. \tag{289}$$

For the square-root term, the elementary inequality $ab \le \lambda a^2 + b^2/(4\lambda)$, valid for every $\lambda > 0$, with $a = \sqrt{R}$ and $b = \sqrt{Z}$ gives

$$\sqrt{ZR} \le \lambda R + \frac{Z}{4\lambda}\,. \tag{290}$$

Taking $\lambda_Z := 1/6$ in (290) yields

$$\sqrt{ZR} \le \frac{1}{6}R + \frac{3}{2}Z\,. \tag{291}$$

Substituting (288), (289), and (291) into (286) gives

$$R \le \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) R + (1-p)(6p)^{\frac{p}{1-p}} U + U^{1-p}\mathcal{C}^p$$
$$+ (1-q)(6q)^{\frac{q}{1-q}} V + V^{1-q}\mathcal{C}^q + \frac{3}{2}Z + \sqrt{Z\mathcal{C}} + K$$
$$= \frac{1}{2}R + (1-p)(6p)^{\frac{p}{1-p}} U + U^{1-p}\mathcal{C}^p$$
$$+ (1-q)(6q)^{\frac{q}{1-q}} V + V^{1-q}\mathcal{C}^q + \frac{3}{2}Z + \sqrt{Z\mathcal{C}} + K\,. \tag{292}$$

Subtracting $R/2$ from both sides of (292) and multiplying by two gives

$$R \le 2(1-p)(6p)^{\frac{p}{1-p}} U + 2U^{1-p}\mathcal{C}^p$$
$$+ 2(1-q)(6q)^{\frac{q}{1-q}} V + 2V^{1-q}\mathcal{C}^q + 3Z + 2\sqrt{Z\mathcal{C}} + 2K\,. \tag{293}$$

Define

$$c_{p,q} := \max\left\{2(1-p)(6p)^{\frac{p}{1-p}}, 2(1-q)(6q)^{\frac{q}{1-q}}, 3, 2\right\}.$$

which bounds every coefficient in (293) and proves (284). □

Proof of (232). If $\mathrm{Reg}_T(\pi^\star) < 0$, (232) follows from its nonnegative right-hand side; hence assume $\mathrm{Reg}_T(\pi^\star) \ge 0$.

Let $\mathcal{I}^\star := \{i = (s,a) : a \ne \pi^\star(s)\}$, $n_i := n_T(s,a)$, $\widetilde{n}_i := \widetilde{n}_T(s,a)$, $\Delta_i := \Delta(s,a)$, and define the total expected gap charge $\mathsf{Gap}_T := \sum_{i\in\mathcal{I}^\star} \Delta_i n_i$. The gap-charge calculation in the proof of (32) in Section B.1, specifically (66), applies without any transition assumption. It gives

$$0 \le \mathsf{Gap}_T \le \mathrm{Reg}_T(\pi^\star) + \mathcal{C}\,. \tag{294}$$

We omit the repeated rearrangement of the self-bounding condition.

We first convert the upper counts in the leading term of the master bound. For each $i \in \mathcal{I}^\star$, the power inequality (221) gives

$$\widetilde{n}_i^\alpha \le n_i^\alpha + \mathsf{X}_T^{\frac{\alpha}{2}} n_i^{\frac{\alpha}{2}} + \mathsf{Z}_T^\alpha\,. \tag{295}$$

Multiply (295) by A_T and sum over $i \in \mathcal{I}^\star$. The direct substitution is

$$\mathsf{A}_T \sum_{i\in\mathcal{I}^\star} \widetilde{n}_i^\alpha \le \mathsf{A}_T \sum_{i\in\mathcal{I}^\star} n_i^\alpha + \mathsf{A}_T \mathsf{X}_T^{\frac{\alpha}{2}} \sum_{i\in\mathcal{I}^\star} n_i^{\frac{\alpha}{2}} + \mathsf{A}_T |\mathcal{I}^\star| \mathsf{Z}_T^\alpha$$
$$= \mathsf{A}_T \sum_{i\in\mathcal{I}^\star} n_i^\alpha + \mathsf{B}_T \sum_{i\in\mathcal{I}^\star} n_i^{q_\varepsilon} + \mathsf{A}_T |\mathcal{I}^\star| \mathsf{Z}_T^\alpha\,, \tag{296}$$

where the equality uses $q_\varepsilon = \alpha/2$ from (222) and $\mathsf{B}_T = \mathsf{A}_T \mathsf{X}_T^{\alpha/2}$ from (226).

The proof of Lemma B.3 in Section B.1 applies verbatim to any $r \in (0,1)$, using the conjugate exponents $1/r$ and $1/(1-r)$. Thus

$$\sum_{i\in\mathcal{I}^\star} n_i^r \le \mathsf{Gap}_T^r \left(\sum_{i\in\mathcal{I}^\star} \Delta_i^{-\frac{r}{1-r}} \right)^{1-r} . \tag{297}$$

We omit the repeated factorization and Hölder calculation, and retain the applications to the two different exponents below.

Apply (297) first with $r = \alpha$. By (223) and (225),

$$\mathsf{A}_T \sum_{i\in\mathcal{I}^\star} n_i^\alpha \le \mathsf{A}_T \mathsf{Gap}_T^\alpha \omega_{1/\varepsilon}(\Delta)^{1-\alpha} = \left[\mathsf{A}_T^{\frac{1}{1-\alpha}} \omega_{1/\varepsilon}(\Delta) \right]^{1-\alpha} \mathsf{Gap}_T^\alpha = \mathcal{U}^{1-\alpha} \mathsf{Gap}_T^\alpha . \tag{298}$$

Here, the last equality uses (225).

Apply (297) next with $r = q_\varepsilon$. By (224) and (227),

$$\mathsf{B}_T \sum_{i\in\mathcal{I}^\star} n_i^{q_\varepsilon} \le \mathsf{B}_T \mathsf{Gap}_T^{q_\varepsilon} \omega_{1/(2\varepsilon)}(\Delta)^{1-q_\varepsilon} = \left[\mathsf{B}_T^{\frac{1}{1-q_\varepsilon}} \omega_{1/(2\varepsilon)}(\Delta) \right]^{1-q_\varepsilon} \mathsf{Gap}_T^{q_\varepsilon} = \mathcal{V}^{1-q_\varepsilon} \mathsf{Gap}_T^{q_\varepsilon} . \tag{299}$$

Substitute (298) and (299) into (296). Since $|\mathcal{I}^\star| \le S(A-1)$, the entire upper-count contribution obeys

$$\mathsf{A}_T \sum_{i\in\mathcal{I}^\star} \widetilde{n}_i^\alpha \le \mathcal{U}^{1-\alpha} \mathsf{Gap}_T^\alpha + \mathcal{V}^{1-q_\varepsilon} \mathsf{Gap}_T^{q_\varepsilon} + \mathsf{A}_T S(A-1) \mathsf{Z}_T^\alpha . \tag{300}$$

The three terms respectively account for the actual count, the fractional transition correction, and the additive conversion remainder.

We now control the transition-sensitive comparator-deviation term. By (275),

$$\mathsf{D}_T(\pi^\star) \le 2H \sum_{i\in\mathcal{I}^\star} n_i = 2H \sum_{i\in\mathcal{I}^\star} \Delta_i n_i \, \Delta_i^{-1} \le \frac{2H}{\Delta_{\min}} \sum_{i\in\mathcal{I}^\star} \Delta_i n_i = \frac{2H}{\Delta_{\min}} \mathsf{Gap}_T . \tag{301}$$

Here, Definition 2.2 and finiteness of the state and action spaces give $\Delta_i \ge \Delta_{\min} > 0$ for every $i \in \mathcal{I}^\star$, which justifies the last inequality. Substituting this bound into the transition term and using $c_{\varepsilon,2} = 2c_{\varepsilon,1}^2$ gives

$$\begin{aligned} c_{\varepsilon,1} H L_\sigma \sqrt{\mathsf{X}_T \, \mathsf{D}_T(\pi^\star)} &\le c_{\varepsilon,1} H L_\sigma \sqrt{\mathsf{X}_T \frac{2H}{\Delta_{\min}} \mathsf{Gap}_T} = \sqrt{\frac{2c_{\varepsilon,1}^2 L_\sigma^2 H^3 \mathsf{X}_T}{\Delta_{\min}} \mathsf{Gap}_T} \\ &= \sqrt{\frac{c_{\varepsilon,2} L_\sigma^2 H^3 \mathsf{X}_T}{\Delta_{\min}} \mathsf{Gap}_T} = \sqrt{\mathcal{Z} \mathsf{Gap}_T} . \end{aligned} \tag{302}$$

Here, the inequality uses (301) and the final equality uses the definition of $\mathcal{Z}$ in (228).

Set $R \coloneqq \mathrm{Reg}_T(\pi^\star)$. Substituting (300) and (302) into (219) gives

$$\begin{aligned} R &\le \mathcal{U}^{1-\alpha} \mathsf{Gap}_T^\alpha + \mathcal{V}^{1-q_\varepsilon} \mathsf{Gap}_T^{q_\varepsilon} + \sqrt{\mathcal{Z} \mathsf{Gap}_T} + \mathsf{A}_T S(A-1) \mathsf{Z}_T^\alpha + \mathsf{R}_0^{\mathrm{unk}} \\ &\le \mathcal{U}^{1-\alpha} \mathsf{Gap}_T^\alpha + \mathcal{V}^{1-q_\varepsilon} \mathsf{Gap}_T^{q_\varepsilon} + \sqrt{\mathcal{Z} \mathsf{Gap}_T} + \mathsf{R}_{\mathrm{const}} . \end{aligned} \tag{303}$$

Here, the last inequality uses (229) with $c_{\varepsilon,3} \ge 1$, so that $\mathsf{A}_T S(A-1) \mathsf{Z}_T^\alpha + \mathsf{R}_0^{\mathrm{unk}} \le \mathsf{R}_{\mathrm{const}}$.

By (294), $0 \le \mathsf{Gap}_T \le R + \mathcal{C}$. Monotonicity of x^α, x^{q_ε}, and $\sqrt{x}$ on $\mathbb{R}_+$ therefore gives

$$\mathsf{Gap}_T^\alpha \le (R+\mathcal{C})^\alpha , \qquad \mathsf{Gap}_T^{q_\varepsilon} \le (R+\mathcal{C})^{q_\varepsilon} , \qquad \sqrt{\mathsf{Gap}_T} \le \sqrt{R+\mathcal{C}} . \tag{304}$$

Substituting all three inequalities from (304) into (303) yields the exact absorption premise

$$R \le \mathcal{U}^{1-\alpha} (R+\mathcal{C})^\alpha + \mathcal{V}^{1-q_\varepsilon} (R+\mathcal{C})^{q_\varepsilon} + \sqrt{\mathcal{Z}(R+\mathcal{C})} + \mathsf{R}_{\mathrm{const}} . \tag{305}$$

Apply Lemma C.7 to (305) with $p=\alpha$, $q=q_\varepsilon$, $U=\mathcal{U}$, $V=\mathcal{V}$, $Z=\mathcal{Z}$, $K=\mathsf{R}_{\mathrm{const}}$. The concrete conclusion (284) is

$$R \le c_{\varepsilon,4}\Big(\mathcal{U}+\mathcal{U}^{1-\alpha}\mathcal{C}^{\alpha}+\mathcal{V}+\mathcal{V}^{1-q_\varepsilon}\mathcal{C}^{q_\varepsilon}+\mathcal{Z}+\sqrt{\mathcal{Z}\mathcal{C}}+\mathsf{R}_{\mathrm{const}}\Big), \tag{306}$$

which is exactly (232).

For the parameter dependence of $\mathcal{U}$, substituting $\mathsf{A}_T = c_{\mathrm{cnt},\varepsilon} H L_\sigma \Lambda_T^{1-\alpha}$ from (198), $L_\sigma=\sigma^{1/\varepsilon}$, and $\alpha=1/\varepsilon$ gives

$$\mathsf{A}_T^{\frac{1}{1-\alpha}} = \mathcal{O}_\varepsilon\Big(H^{\frac{\varepsilon}{\varepsilon-1}}\sigma^{\frac{1}{\varepsilon-1}}\Lambda_T\Big),$$

and hence

$$\mathcal{U} = \mathcal{O}_\varepsilon\Big(H^{\frac{\varepsilon}{\varepsilon-1}}\sigma^{\frac{1}{\varepsilon-1}}\Lambda_T\omega_{1/\varepsilon}(\Delta)\Big). \tag{307}$$

This proves (233). The remaining named terms depend on transition uncertainty and are expanded below.

For $\mathcal{V}$, first substitute $\mathsf{B}_T=\mathsf{A}_T\mathsf{X}_T^{\alpha/2}$ from (226) into (227) without any intervening relaxation:

$$\mathcal{V} = \Big(\mathsf{A}_T\mathsf{X}_T^{\frac{\alpha}{2}}\Big)^{\frac{1}{1-q_\varepsilon}}\omega_{1/(2\varepsilon)}(\Delta) = \mathsf{A}_T^{\frac{2\varepsilon}{2\varepsilon-1}}\mathsf{X}_T^{\frac{1}{2\varepsilon-1}}\omega_{1/(2\varepsilon)}(\Delta). \tag{308}$$

Substituting $\mathsf{A}_T = c_{\mathrm{cnt},\varepsilon}H\sigma^{1/\varepsilon}\Lambda_T^{1-1/\varepsilon}$ and $\mathsf{X}_T = c_{\mathrm{occ}}HS^2A\mathcal{L}_T$ into (308) gives

$$\begin{aligned}\mathcal{V} &= \mathcal{O}_\varepsilon\Big(H^{\frac{2\varepsilon}{2\varepsilon-1}}\sigma^{\frac{2}{2\varepsilon-1}}\Lambda_T^{\frac{2(\varepsilon-1)}{2\varepsilon-1}}H^{\frac{1}{2\varepsilon-1}}S^{\frac{2}{2\varepsilon-1}}A^{\frac{1}{2\varepsilon-1}}\mathcal{L}_T^{\frac{1}{2\varepsilon-1}}\omega_{1/(2\varepsilon)}(\Delta)\Big)\\ &= \mathcal{O}_\varepsilon\Big(H^{\frac{2\varepsilon+1}{2\varepsilon-1}}\sigma^{\frac{2}{2\varepsilon-1}}S^{\frac{2}{2\varepsilon-1}}A^{\frac{1}{2\varepsilon-1}}\Lambda_T^{\frac{2(\varepsilon-1)}{2\varepsilon-1}}\mathcal{L}_T^{\frac{1}{2\varepsilon-1}}\omega_{1/(2\varepsilon)}(\Delta)\Big),\end{aligned} \tag{309}$$

which is (234).

For $\mathcal{Z}$, directly substitute $L_\sigma=\sigma^{1/\varepsilon}$ and $\mathsf{X}_T = c_{\mathrm{occ}}HS^2A\mathcal{L}_T$ into (228):

$$\mathcal{Z} = \frac{c_{\varepsilon,2}L_\sigma^2H^3\mathsf{X}_T}{\Delta_{\min}} = \frac{c_{\varepsilon,2}c_{\mathrm{occ}}\sigma^{\frac{2}{\varepsilon}}H^3(HS^2A\mathcal{L}_T)}{\Delta_{\min}} = \mathcal{O}_\varepsilon\left(\frac{\sigma^{\frac{2}{\varepsilon}}H^4S^2A\mathcal{L}_T}{\Delta_{\min}}\right), \tag{310}$$

which is (235).

For $\mathsf{R}_{\mathrm{const}}$, first substitute its definition and only then use subadditivity of $x^{1/\varepsilon}$:

$$\begin{aligned}&\mathsf{R}_{\mathrm{const}}\\ &= \mathsf{R}_0^{\mathrm{unk}} + c_{\varepsilon,3}\mathsf{A}_T S(A-1)\mathsf{Z}_T^{\frac{1}{\varepsilon}}\\ &\le \mathsf{R}_0^{\mathrm{unk}} + \mathcal{O}_\varepsilon\Big(H\sigma^{\frac{1}{\varepsilon}}\Lambda_T^{1-\frac{1}{\varepsilon}}SA\big(\mathsf{Y}_T^{\frac{1}{\varepsilon}}+\mathsf{G}_T^{\frac{1}{\varepsilon}}+1\big)\Big)\\ &= \mathsf{R}_0^{\mathrm{unk}} + \mathcal{O}_\varepsilon\Big(\sigma^{\frac{1}{\varepsilon}}\Lambda_T^{1-\frac{1}{\varepsilon}}\big[H^{1+\frac{1}{\varepsilon}}S^{1+\frac{4}{\varepsilon}}A^{1+\frac{1}{\varepsilon}}\mathcal{L}_T^{\frac{1}{\varepsilon}} + H^3SA^{2-\frac{1}{\varepsilon}}(1+\log(eT))^{\frac{1}{\varepsilon}}\big]\Big).\end{aligned} \tag{311}$$

The first inequality in (311) uses $A-1\le A$, $\mathsf{Z}_T = 2(\mathsf{Y}_T+\mathsf{G}_T+1)$, and $(x+y+z)^{1/\varepsilon}\le x^{1/\varepsilon}+y^{1/\varepsilon}+z^{1/\varepsilon}$. The final line substitutes $\mathsf{Y}_T = c_{\mathrm{occ}}HS^4A\mathcal{L}_T$ and $\mathsf{G}_T=\Gamma(1+\log(eT))$, and then uses $\Gamma^{1/\varepsilon}=\mathcal{O}_\varepsilon(H^2A^{1-1/\varepsilon})$. This proves (236), while $\mathsf{R}_0^{\mathrm{unk}}$ itself is expanded in (218).

The two corruption interactions follow by direct exponentiation of the definitions of $\mathcal{U}$ and $\mathcal{V}$:

$$\mathcal{U}^{1-\alpha}\mathcal{C}^{\alpha} = \mathsf{A}_T\omega_{1/\varepsilon}(\Delta)^{1-\alpha}\mathcal{C}^{\alpha} = \mathcal{O}_\varepsilon\Big(H\sigma^{\frac{1}{\varepsilon}}\Lambda_T^{1-\frac{1}{\varepsilon}}\omega_{1/\varepsilon}(\Delta)^{1-\frac{1}{\varepsilon}}\mathcal{C}^{\frac{1}{\varepsilon}}\Big), \tag{312}$$

$$\begin{aligned}\mathcal{V}^{1-q_\varepsilon}\mathcal{C}^{q_\varepsilon} &= \mathsf{B}_T\omega_{1/(2\varepsilon)}(\Delta)^{1-q_\varepsilon}\mathcal{C}^{q_\varepsilon} = \mathsf{A}_T\mathsf{X}_T^{\frac{1}{2\varepsilon}}\omega_{1/(2\varepsilon)}(\Delta)^{1-\frac{1}{2\varepsilon}}\mathcal{C}^{\frac{1}{2\varepsilon}}\\ &= \mathcal{O}_\varepsilon\Big(H^{1+\frac{1}{2\varepsilon}}\sigma^{\frac{1}{\varepsilon}}S^{\frac{1}{\varepsilon}}A^{\frac{1}{2\varepsilon}}\Lambda_T^{1-\frac{1}{\varepsilon}}\mathcal{L}_T^{\frac{1}{2\varepsilon}}\omega_{1/(2\varepsilon)}(\Delta)^{1-\frac{1}{2\varepsilon}}\mathcal{C}^{\frac{1}{2\varepsilon}}\Big).\end{aligned}$$

Finally, (310) gives

$$\sqrt{\mathcal{Z}\mathcal{C}} = \mathcal{O}_\varepsilon\left(\sigma^{\frac{1}{\varepsilon}}H^2S\sqrt{\frac{A\mathcal{L}_T\mathcal{C}}{\Delta_{\min}}}\right). \tag{313}$$

□

C.2 Bounding FTRL-Regret$^{\pi}(s)$

We follow the four-part proof structure of Section B. The FTRL decomposition is applied to the full realized loss sequence; the conditional moment and stability estimates distinguish the predictable events $\mathcal{E}_t$ and $\mathcal{E}_t^c$.

C.2.1 Heavy-tailed Estimator Properties

The suffix-return and centered-kernel estimates are shared prerequisites for the stability argument here and the clipping argument in Section C.3.

Moment of a Suffix Return The suffix-return moment bound (72) from Lemma B.5 in Appendix B.2 applies here: its proof uses only Definition 2.1, conditional independence of the loss table and trajectory randomization given $\mathcal{F}_{t-1}$, and conditional Minkowski's inequality.

Algebra of the Centered Action Kernel The centered-kernel identities (74)–(76) are supplied directly by Lemma B.6 in Appendix B.2, since the centered kernel and $\mathsf{D}_p(a)$ have the same definitions here.

Mean, Moment, and Clipping Tail of the Advantage Function Estimator

Lemma C.8. *Fix an episode t, a state $s \in \mathcal{S}_h$, and an action a. Conditional on $\mathcal{F}_{t-1}$,*

$$\mathbb{E}_t[\widetilde{Q}_t(s,a)] = \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)} \bar{Q}_t(s,a)\,, \tag{314}$$

$$\mathbb{E}_t[|\widetilde{Q}_t(s,a)|^{\varepsilon}] \le \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)} \Sigma_t(s) \mathsf{D}_{p_{t,s}}(a)\,. \tag{315}$$

Consequently, on $\mathcal{E}_t$,

$$\mathbb{E}_t[|\widetilde{Q}_t(s,a)|^{\varepsilon}] \le \Sigma_t(s) \mathsf{D}_{p_{t,s}}(a)\,, \tag{316}$$

$$\mathbb{E}_t\big[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|\big] \le d_t(s,a)\,, \tag{317}$$

whereas on every history, without assuming $\mathcal{E}_t$,

$$\mathbb{E}_t[|\widetilde{Q}_t(s,a)|^{\varepsilon}] \le \gamma_t^{-1} \Sigma_t(s) \mathsf{D}_{p_{t,s}}(a)\,, \tag{318}$$

$$\mathbb{E}_t\big[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|\big] \le \gamma_t^{-1} d_t(s,a)\,. \tag{319}$$

Proof. Fix t, $s \in \mathcal{S}_h$, and a, and write $p = p_{t,s}$. The mean and moment calculations through (85) in the proof of Lemma B.7, Appendix B.2, require only a positive $\mathcal{F}_{t-1}$-measurable denominator, not $\mu^{\pi_t}(s) \le \widetilde{\mu}_t(s)$. With the present $\widetilde{\mu}_t(s)$ they give (314) and (315), so we omit the repeated calculations.

On $\mathcal{E}_t$, the true kernel belongs to $\mathcal{P}_t$, and hence

$$\mu^{\pi_t}(s) \le \overline{\mu}_t(s) \le \underline{\mu}_t(s) + \gamma_t = \widetilde{\mu}_t(s)\,. \tag{320}$$

Substituting (320) into (315) gives (316). Without assuming $\mathcal{E}_t$, the deterministic inequalities

$$\frac{\mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)} \le \frac{1}{\widetilde{\mu}_t(s)} \le \frac{1}{\gamma_t} \tag{321}$$

follow from $\mu^{\pi_t}(s) \le 1$ and $\widetilde{\mu}_t(s) \ge \gamma_t$. Substituting (321) into (315) gives (318).

The deterministic clipping inequality and its application to the good-event moment bound are also the same as in the proof of Lemma B.7. Applying that argument to (316) proves (317) on $\mathcal{E}_t$. On every history, applying the same inequality instead to (318) gives

$$\mathbb{E}_t[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|] \le \gamma_t^{-1} \Sigma_t(s) \mathsf{D}_p(a) M_t(s,a)^{1-\varepsilon} = \gamma_t^{-1} d_t(s,a)\,,$$

which proves (319). □

C.2.2 POINTWISE SAFETY

We next verify the pointwise stability condition and the deterministic envelopes needed on confidence-failure histories.

The action-bonus bound (88) from Lemma B.8 in Appendix B.2 applies on every history: the threshold and tail-bonus definitions, (75), and $G_t(s) \geq \widetilde{\mu}_t(s)^{1-\varepsilon}$ also hold for the denominator $\widetilde{\mu}_t(s)$ defined in Section 3.

Lemma C.9 (Pointwise size of c_t, C_t, b_t, B_t)**.** *For every episode t,*

$$0 \leq c_t(s) \leq 2HL_\sigma\,, \tag{322}$$

$$0 \leq C_t(s,a) \leq 4H^2L_\sigma\,, \tag{323}$$

$$b_t(s) \leq K_{\mathrm{saf}} HL_\sigma A^{1-\alpha} \widetilde{\mu}_t(s)^{-\frac{\varepsilon-1}{\varepsilon}}\,, \tag{324}$$

$$0 \leq B_t(s,a) \leq \frac{k_\alpha \beta_t(s)}{32H} \leq \frac{M_t(s,a)}{32H}\,, \tag{325}$$

$$0 \leq C_t(s,a) \leq \frac{M_t(s,a)}{16}\,. \tag{326}$$

Moreover,

$$|g_t(s,a)| \leq \frac{1-\alpha}{16}\beta_t(s)\bar{p}_{t,s,a}^{\alpha-1}\,. \tag{327}$$

Proof. We prove the displayed bounds in the order in which they are stated.

First, $w_t(s) = \overline{\mu}_t(s) - \underline{\mu}_t(s) + \gamma_t$ is nonnegative, and $\underline{\mu}_t(s) \geq 0$ implies

$$0 \leq w_t(s) \leq \overline{\mu}_t(s) + \gamma_t = \widetilde{\mu}_t(s)\,. \tag{328}$$

Substituting (328) into the definition of $c_t(s)$ in Section 3 gives

$$0 \leq c_t(s) = 2HL_\sigma \frac{w_t(s)}{\widetilde{\mu}_t(s)} \leq 2HL_\sigma\,,$$

which proves (322).

Let $\bar{C}_t(s) := \sum_a p_{t,s,a} C_t(s,a)$. Rectangularity gives a single selector P_t^C for which

$$C_t(s,a) = \mathbb{E}_{s' \sim P_t^C(\cdot|s,a)}\left[2c_t(s') + \bar{C}_t(s')\right]\,. \tag{329}$$

Repeated substitution of (329), together with the terminal convention, yields

$$C_t(s,a) = 2\mathbb{E}_{P_t^C,\pi_t}\left[\sum_{j=h(s)+1}^{H-1} c_t(s_j) \middle| s_{h(s)} = s, a_{h(s)} = a\right] \tag{330}$$

$$\leq 2(H - h(s) - 1)\max_x c_t(x) \leq 4H^2L_\sigma\,. \tag{331}$$

Line (331) uses (322) and $H - h(s) - 1 \leq H$. Equations (330)–(331) establish (323).

The remaining scalar-bonus calculation is the same as in the proof of Lemma B.9 in Appendix B.2. Indeed, with $x_t(s) = \widetilde{\mu}_t(s)^{1-\varepsilon}$, the definitions of G_t, β_t, Σ_t, h_t, z_t, and ν_t are unchanged, while $\gamma_t \leq \widetilde{\mu}_t(s) = \overline{\mu}_t(s) + \gamma_t \leq 2$ and (323) provide the same bounds used there. Thus the changing-rate, heavy-tail, and squared-confidence estimates in that proof give (324); we omit their repeated algebra.

Let $b_t^{\max} := \max_x b_t(x)$. The optimistic recursion takes maxima of expectations of nonnegative quantities. Unrolling the nonnegative recursion (12) over at most H nonterminal layers gives

$$0 \leq B_t(s,a) \leq \sum_{j=0}^{H-1}\left(1 + \frac{1}{H}\right)^j b_t^{\max} \leq H\left(1 + \frac{1}{H}\right)^{H-1} b_t^{\max} \leq eHb_t^{\max}\,. \tag{332}$$

Because $\widetilde{\mu}_t(x) \geq \gamma_t$ and the exponent in (324) is negative,

$$B_t(s,a) \leq eK_{\mathrm{saf}} H^2 L_\sigma A^{1-\alpha} \gamma_t^{-\frac{\varepsilon-1}{\varepsilon}}\,. \tag{333}$$

Starting from (333), the cases $t \le \Gamma$ and $t > \Gamma$ are exactly the corresponding cases in the proof of Lemma B.9. Their only occupancy inputs are $\widetilde{\mu}_t(s) \ge \gamma_t$ and $\widetilde{\mu}_t(s) \le 2$, which hold here on every history. They therefore give (325) with the same constants. The final comparisons for C_t and g_t in that proof use only (323), Lemma B.8, and the threshold and constant definitions, so they also apply unchanged and establish (326) and (327). □

Lemma C.10 (Event-free deterministic envelopes for the full sequence)**.** *For every episode t, state s, and action a, the quantities used by the full-sequence FTRL analysis satisfy*

$$\beta_t(s) \le \beta_{\max}\,, \qquad \gamma_t^{-1} \le \overline{T}\,, \tag{334}$$

$$b_t^{\max} := \max_x b_t(x) \le K_{\mathrm{saf}} H L_\sigma A^{1-\alpha} \overline{T}^{1-\alpha}\,, \qquad \max_{x,b} B_t(x,b) \le e H b_t^{\max}\,, \tag{335}$$

$$|r_t(s,a)| \le 2HL_\sigma\big(1+\gamma_t^{-1}\big) \le 4HL_\sigma \overline{T}\,, \qquad 0 \le \nu_t(s) \le b_t(s) \le b_t^{\max}\,. \tag{336}$$

Consequently, all quantities in (201)–(205) are finite deterministic envelopes that do not use a lower bound on any action probability.

Proof. Because $1-\varepsilon < 0$ and $\widetilde{\mu}_\tau(s) \ge \gamma_\tau$, direct substitution into the clock gives

$$G_t(s) = 1 + \sum_{\tau=1}^{t} \widetilde{\mu}_\tau(s)^{1-\varepsilon} \le 1 + \sum_{\tau=1}^{t} \gamma_\tau^{1-\varepsilon} \le 1 + \sum_{\tau=1}^{T} \gamma_\tau^{1-\varepsilon}\,,$$

$$\beta_t(s) = \beta_0 + c_\beta H L_\sigma G_t(s)^{\frac{1}{\varepsilon}} \le \beta_0 + c_\beta H L_\sigma \left(1 + \sum_{\tau=1}^{T} \gamma_\tau^{1-\varepsilon}\right)^{\frac{1}{\varepsilon}} = \beta_{\max}\,.$$

Here, the last equality uses (200). Also, $t \le T \le \overline{T}$ and $\Gamma \ge 1$ imply

$$\gamma_t = \min\{1, \Gamma/t\} \ge \min\{1, 1/t\} = \frac{1}{t} \ge \frac{1}{\overline{T}}\,, \tag{337}$$

which proves both bounds in (334).

The pointwise local-bonus estimate (324) and $\widetilde{\mu}_t(x) \ge \gamma_t$ give

$$b_t^{\max} \le K_{\mathrm{saf}} H L_\sigma A^{1-\alpha} \gamma_t^{-\frac{\varepsilon-1}{\varepsilon}} = K_{\mathrm{saf}} H L_\sigma A^{1-\alpha} \gamma_t^{-(1-\alpha)} \le K_{\mathrm{saf}} H L_\sigma A^{1-\alpha} \overline{T}^{1-\alpha}\,. \tag{338}$$

Here, the equality uses $(\varepsilon-1)/\varepsilon = 1-\alpha$ and the last inequality uses (337). Bellman unrolling (332) then gives

$$\max_{x,b} B_t(x,b) \le e H b_t^{\max}\,, \tag{339}$$

so (338)–(339) establish (335).

For the visitation-denominator bias, substitute (238) before applying any upper bound:

$$|r_t(s,a)| = \frac{|\widetilde{\mu}_t(s) - \mu^{\pi_t}(s)|}{\widetilde{\mu}_t(s)} |\bar{Q}_t(s,a)| \le \frac{\widetilde{\mu}_t(s) + \mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)}\, 2HL_\sigma$$
$$\le 2HL_\sigma \left(1 + \frac{1}{\gamma_t}\right) \le 4HL_\sigma \overline{T}\,.$$

Here, the first inequality uses $|\bar{Q}_t(s,a)| \le 2HL_\sigma$, the second uses $\mu^{\pi_t}(s) \le 1$ and $\widetilde{\mu}_t(s) \ge \gamma_t$, and the last uses $1 + \gamma_t^{-1} \le 2\overline{T}$. Since $\nu_t(s)$ is a nonnegative summand of $b_t(s)$ in (10), $0 \le \nu_t(s) \le b_t(s) \le b_t^{\max}$, completing (336). □

C.2.3 Conditional Stability and the FTRL Bound

Fix a state s and abbreviate $\beta_t := \beta_t(s)$, $\Sigma_t := \Sigma_t(s)$, $M_{t,a} := M_t(s,a)$. Let $a^\star$ be an arbitrary comparator action at this state. For each t, let $\widetilde{p}_{t+1}$ denote the same-inverse-learning-rate look-ahead point in (464) for this state's cumulative loss sequence. Define the analysis-only confidence-failure corrections

$$\mathsf{e}_t^{\mathrm{clip}}(s) := 2\big(1+\gamma_t^{-1}\big) k_\alpha^{1-\varepsilon} \Sigma_t(s) \beta_t(s)^{1-\varepsilon} z_t(s) + \frac{2HL_\sigma}{\gamma_t}\,, \tag{340}$$

$$\mathsf{e}_t^{\mathrm{stab}}(s) \coloneqq \frac{32k_\alpha^2}{1-\alpha}\,\beta_t(s)z_t(s)\,. \tag{341}$$

The deterministic envelopes in (201)–(202) satisfy

$$\mathsf{e}_t^{\mathrm{clip}}(s) \le \mathsf{F}_{\mathrm{clip}}\,, \qquad \mathsf{e}_t^{\mathrm{stab}}(s) \le \mathsf{F}_{\mathrm{stab}} \qquad \text{for every } t, s\,. \tag{342}$$

Lemma C.11 (Conditional stability on the full FTRL sequence)**.** *For every t,*

$$\begin{aligned}&\mathbb{E}_t\left[\langle p_{t,s} - \widetilde{p}_{t+1}, g_t(s,\cdot)\rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_{t,s})\right]\\ &\le \frac{64}{1-\alpha}k_\alpha^{2-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) + \nu_t(s) + \frac{1}{H}\langle p_{t,s}, B_t(s,\cdot)\rangle + \mathbb{1}\{\mathcal{E}_t^c\}\mathsf{e}_t^{\mathrm{stab}}(s)\,.\end{aligned} \tag{343}$$

Moreover, $\mathsf{e}_t^{\mathrm{stab}}(s) \le \mathsf{F}_{\mathrm{stab}}$.

Proof. Define the pathwise stability quantity

$$\mathsf{S}_t(s) \coloneqq \langle p_{t,s} - \widetilde{p}_{t+1}, g_t(s,\cdot)\rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_{t,s})\,.$$

The pointwise bound (327) permits exactly the vector-Tsallis and square-splitting argument in the proof of Lemma B.10 in Appendix B.2. We therefore recall its pathwise conclusion:

$$\mathsf{S}_t(s) \le \frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}\Big(\widehat{Q}_t(s,a)^2 + d_t(s,a)^2 + B_t(s,a)^2 + C_t(s,a)^2\Big)\,. \tag{344}$$

On $\mathcal{E}_t$, (316), Lemma B.8, and (75) supply the same moment, action-bonus, and kernel bounds as in that proof. Thus its $\widehat{Q}_t$- and d_t-square calculation applies without change and gives

$$\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}\left(\mathbb{E}_t[\widehat{Q}_t(s,a)^2] + d_t(s,a)^2\right) \le \frac{64}{1-\alpha}k_\alpha^{2-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) \qquad \text{on } \mathcal{E}_t\,. \tag{345}$$

On $\mathcal{E}_t^c$, the event-free inequalities $|\widehat{Q}_t(s,a)| \le M_{t,a}$ and $d_t(s,a) \le M_{t,a}$ imply

$$\begin{aligned}&\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}\left(\widehat{Q}_t(s,a)^2 + d_t(s,a)^2\right)\\ &\le \frac{32}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}M_{t,a}^2 = \frac{32k_\alpha^2}{1-\alpha}\,\beta_t\sum_a \bar{p}_{t,s,a}^{\alpha} = \mathsf{e}_t^{\mathrm{stab}}(s)\,.\end{aligned} \tag{346}$$

Here, the equality uses $M_{t,a} = k_\alpha\beta_t\bar{p}_{t,s,a}^{\alpha-1}$.

The C_t- and B_t-square calculations in Lemma B.10 also apply on every history, using (11), nonnegativity of B_t, and (325):

$$\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}C_t(s,a)^2 \le \nu_t(s)\,, \tag{347}$$

$$\frac{16}{(1-\alpha)\beta_t}\sum_a \bar{p}_{t,s,a}^{2-\alpha}B_t(s,a)^2 \le \frac{1}{H}\langle p_{t,s}, B_t(s,\cdot)\rangle\,. \tag{348}$$

Taking $\mathbb{E}_t$ in (344), splitting over the predictable events $\mathcal{E}_t$ and $\mathcal{E}_t^c$, and substituting (345), (346), (347), and (348) proves (343). Finally, $\beta_t \le \beta_{\max}$ and $z_t(s) \le A^{1-\alpha}$ give

$$\mathsf{e}_t^{\mathrm{stab}}(s) \le \frac{32k_\alpha^2}{1-\alpha}\,\beta_{\max}A^{1-\alpha} \le \mathsf{F}_{\mathrm{stab}}\,,$$

where the last inequality uses (201). □

The variable-rate FTRL inequality (466), together with the preceding conditional stability bound, controls the cumulative FTRL term. We combine it with clipping compensation in Section C.1.1, retaining both confidence-failure corrections.

C.3 BOUNDING Clipping-Bias$^{\pi}(s)$

Fix a state s and comparator action $a^\star$, and use the abbreviations $\beta_t = \beta_t(s)$, $\Sigma_t = \Sigma_t(s)$, and $M_{t,a} = M_t(s,a)$. The look-ahead point $\widetilde{p}_{t+1}$ and the corrections $\mathsf{e}_t^{\mathrm{clip}}(s)$ and $\mathsf{e}_t^{\mathrm{stab}}(s)$ are those defined in Section C.2.

Lemma C.12 (Clipping compensation on the full FTRL sequence)**.** *Let* $\tau_t(s,a) := \mathbb{E}_t[\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)]$. *Then, for every comparator action* $a^\star$,

$$\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \le 4k_\alpha^{1-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) + \mathbb{1}\{\mathcal{E}_t^c\}\mathsf{e}_t^{\mathrm{clip}}(s)\,. \tag{349}$$

Moreover, the confidence-failure correction in (349) obeys the first inequality in (342).

Proof. On the predictable event $\mathcal{E}_t$, (317) gives $|\tau_t(s,a)| \le d_t(s,a)$. The proof for this part is the same as the proof of Lemma B.11 in Appendix B.3, so we omit the repeated calculation. The definitions of $M_{t,a}$, $d_t(s,a)$, and Σ_t, and the kernel bound (76), are exactly the inputs used there. For use also on $\mathcal{E}_t^c$, recall the event-free intermediate estimate

$$p_{t,s,a}d_t(s,a) \le 2k_\alpha^{1-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}\bar{p}_{t,s,a}^{\alpha}\,. \tag{350}$$

The resulting good-event bound is

$$\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle \le 4k_\alpha^{1-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) \qquad \text{on } \mathcal{E}_t\,. \tag{351}$$

We next derive a probability-floor-free bound on $\mathcal{E}_t^c$. The safe tail bound (319) gives $|\tau_t(s,a)| \le \gamma_t^{-1}d_t(s,a)$. In addition, the centered kernel has the exact first absolute moment

$$\begin{aligned}\sum_{b\in\mathcal{A}} p_{t,s,b}|\mathsf{K}_{p_{t,s}}(a,b)| &= p_{t,s,a}\left|\frac{1}{p_{t,s,a}} - 1\right| + \sum_{b\neq a} p_{t,s,b}\\ &= (1-p_{t,s,a}) + (1-p_{t,s,a}) = 2(1-p_{t,s,a}) \le 2\,.\end{aligned} \tag{352}$$

Conditional Jensen, (72), and (352) therefore imply

$$\begin{aligned}|\tau_t(s,a)| &\le \mathbb{E}_t\big[|\widetilde{Q}_t(s,a) - \widehat{Q}_t(s,a)|\big] \le \mathbb{E}_t[|\widetilde{Q}_t(s,a)|]\\ &\le \frac{\mu^{\pi_t}(s)}{\widetilde{\mu}_t(s)}HL_\sigma\sum_b p_{t,s,b}|\mathsf{K}_{p_{t,s}}(a,b)| \le \frac{2HL_\sigma}{\gamma_t}\,.\end{aligned} \tag{353}$$

The final inequality uses $\mu^{\pi_t}(s) \le 1$ and $\widetilde{\mu}_t(s) \ge \gamma_t$.

Expanding the policy-difference inner product and retaining the negative term $-d_t(s,a^\star)$ gives the complete bad-event calculation

$$\begin{aligned}\langle p_{t,s} - e_{a^\star}, \tau_t(s,\cdot) + d_t(s,\cdot)\rangle &= \sum_a p_{t,s,a}\big(\tau_t(s,a) + d_t(s,a)\big) - \tau_t(s,a^\star) - d_t(s,a^\star)\\ &\le \sum_a p_{t,s,a}\big(|\tau_t(s,a)| + d_t(s,a)\big) + |\tau_t(s,a^\star)| - d_t(s,a^\star)\\ &\le \big(1+\gamma_t^{-1}\big)\sum_a p_{t,s,a}d_t(s,a) + |\tau_t(s,a^\star)|\\ &\le 2\big(1+\gamma_t^{-1}\big)k_\alpha^{1-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) + \frac{2HL_\sigma}{\gamma_t}\\ &= \mathsf{e}_t^{\mathrm{clip}}(s)\,.\end{aligned} \tag{354}$$

Here, the second inequality uses $|\tau_t(s,a)| \le \gamma_t^{-1}d_t(s,a)$ and discards $-d_t(s,a^\star) \le 0$; the last inequality sums (350) over actions and uses (353).

Combining (351) and (354) over the two predictable events $\mathcal{E}_t$ and $\mathcal{E}_t^c$ proves (349). It remains to verify the deterministic envelope. Direct substitution of $\Sigma_t = H^\varepsilon\sigma\widetilde{\mu}_t^{1-\varepsilon}$, followed by $\widetilde{\mu}_t \ge \gamma_t$, $\beta_t \ge \beta_0$, $z_t(s) \le A^{1-\alpha}$, and $1+\gamma_t^{-1} \le 2\gamma_t^{-1}$, gives

$$2(1+\gamma_t^{-1})k_\alpha^{1-\varepsilon}\Sigma_t\beta_t^{1-\varepsilon}z_t(s) \le 4k_\alpha^{1-\varepsilon}H^\varepsilon\sigma\gamma_t^{-\varepsilon}\beta_0^{1-\varepsilon}A^{1-\alpha} \le 4k_\alpha^{1-\varepsilon}H^\varepsilon\sigma\beta_0^{1-\varepsilon}A^{1-\alpha}\overline{T}^{\varepsilon}\,.$$

The last line uses $\gamma_t^{-1} \le \overline{T}$. The remaining term in (340) satisfies $2HL_\sigma/\gamma_t \le 2HL_\sigma\overline{T}$. Substituting these two estimates into (340) gives $\mathsf{e}_t^{\mathrm{clip}}(s) \le \mathsf{F}_{\mathrm{clip}}$. □

C.4 BOUNDING Estimator-Bias$^{\pi}(s)$

We now control the comparator-occupancy-weighted signed estimator bias. We first state and prove the aggregate signed-bias bound, then establish the Bellman-residual estimate used in its proof. The robust dilated reduction is given in Section D.2, and the confidence-failure contributions are controlled in Lemma C.5 within Section C.1.1.

For a deterministic comparator π, abbreviate

$$d_t^{\pi}(s,a) \coloneqq [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ , \qquad d_t^{\pi}(s) \coloneqq \sum_a d_t^{\pi}(s,a) . \tag{355}$$

Notice that

$$d_t^{\pi}(s,a) \le \mu^{\pi_t}(s,a) , \qquad d_t^{\pi}(s) \le \mu^{\pi_t}(s) . \tag{356}$$

Lemma C.13 (Comparator-sensitive signed-bias bound)**.** *For every deterministic policy π,*

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_{s}\mu^{\pi}(s)\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle\right] \\ &\le c_{\varepsilon,1} H L_{\sigma}\sqrt{\mathsf{X}_T\,\mathsf{D}_T(\pi)} + c_{\varepsilon} L_{\sigma} H^2 S^4 A^2 \mathcal{L}_T + c_{\varepsilon} H L_{\sigma}\big(\mathsf{Y}_T + S\mathsf{G}_T\big) . \end{aligned} \tag{357}$$

Proof. On $\mathcal{E}_t$, Lemma C.14 identifies $\mathcal{L}_t^{\mathrm{bias}}$ as the Q-function of π_t for the nonnegative loss ζ_t. The scalar cancellation and performance-difference identity are identical to (146) and (147) in the proof of Lemma B.12 in Appendix B.4, so we omit those steps. Here we retain the positive occupancy discrepancy to obtain

$$\begin{aligned} &\sum_s \mu^{\pi}(s)\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle \\ &= \sum_{s,a}\big(\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)\big)\zeta_t(s,a) \le \sum_{s,a} d_t^{\pi}(s,a)\zeta_t(s,a) . \end{aligned} \tag{358}$$

Line (358) follows from $\zeta_t(s,a) \ge 0$ and the definition $d_t^{\pi}(s,a) = [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+$.

Substituting the three summands from (397) into (358), multiplying by $\mathbb{1}\{\mathcal{E}\}$, summing over t, and taking expectations gives

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s}\mu^{\pi}(s)\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle\right] \le 2\mathsf{C}_{\mathrm{cur}} + 2\mathsf{C}_{\mathrm{next}} + 2\mathsf{C}_{\mathrm{rad}} , \tag{359}$$

where

$$\mathsf{C}_{\mathrm{cur}} \coloneqq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s} d_t^{\pi}(s) c_t(s)\right] ,$$

$$\mathsf{C}_{\mathrm{next}} \coloneqq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^{\pi}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s')\right]\right] , \tag{360}$$

$$\mathsf{C}_{\mathrm{rad}} \coloneqq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^{\pi}(s,a)\sum_x |P_t^C(x \mid s,a) - P(x \mid s,a)| F_t(x)\right] , \tag{361}$$

$$F_t(x) \coloneqq \mathbb{E}_{P_t^C,\pi_t}\left[\sum_{j=h(x)}^{H-1} c_t(s_j)\,\middle|\, s_{h(x)} = x\right] .$$

We bound the three quantities in (359) separately and then substitute all three bounds back into that formula.

Bounding $\mathsf{C}_{\mathrm{cur}}$ For the current-state quantity, define the predictable weight

$$g_t(s) \coloneqq \min\left\{1, \frac{d_t^\pi(s)}{\widetilde{\mu}_t(s)}\right\}. \tag{362}$$

The occupancies μ^{π_t}, the policy π_t, and the comparator are fixed before the episode-t transition randomization, so the weight in (362) is $\mathcal{F}_{t-1}$-measurable. On $\mathcal{E}$, (356) and $\mu^{\pi_t}(s) \le \overline{\mu}_t(s) \le \widetilde{\mu}_t(s)$ imply $d_t^\pi(s) \le \widetilde{\mu}_t(s)$, and hence $g_t(s) = d_t^\pi(s)/\widetilde{\mu}_t(s)$. Direct substitution of $c_t(s) = 2HL_\sigma(\overline{\mu}_t(s) - \underline{\mu}_t(s) + \gamma_t)/\widetilde{\mu}_t(s)$ gives

$$\mathsf{C}_{\mathrm{cur}} = 2HL_\sigma\,\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s} g_t(s)\big(\overline{\mu}_t(s) - \underline{\mu}_t(s) + \gamma_t\big)\right]. \tag{363}$$

Let $P_t^{\mathrm{up},s}$ and $P_t^{\mathrm{low},s}$ be the measurable selectors attaining $\overline{\mu}_t(s)$ and $\underline{\mu}_t(s)$. On $\mathcal{E}$, $\underline{\mu}_t(s) \le \mu^{\pi_t}(s) \le \overline{\mu}_t(s)$, and therefore

$$\overline{\mu}_t(s) - \underline{\mu}_t(s) = \left|\mu^{P_t^{\mathrm{up},s},\pi_t}(s) - \mu^{\pi_t}(s)\right| + \left|\mu^{P_t^{\mathrm{low},s},\pi_t}(s) - \mu^{\pi_t}(s)\right|. \tag{364}$$

For either selector, the quadratic quantity required by Lemma D.13 satisfies

$$\begin{aligned} X_g &= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s}\mu^{\pi_t}(s)g_t(s)^2\right] \\ &= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s}\mu^{\pi_t}(s)\frac{d_t^\pi(s)^2}{\widetilde{\mu}_t(s)^2}\right] \le \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s} d_t^\pi(s)\right] && (365) \\ &\le \mathsf{D}_T(\pi). && (366) \end{aligned}$$

For (365), the positive discrepancy satisfies $d_t^\pi(s) \le \mu^{\pi_t}(s) \le \widetilde{\mu}_t(s)$, and hence

$$\mu^{\pi_t}(s)\frac{d_t^\pi(s)^2}{\widetilde{\mu}_t(s)^2} \le \widetilde{\mu}_t(s)\frac{d_t^\pi(s)^2}{\widetilde{\mu}_t(s)^2} = \frac{d_t^\pi(s)^2}{\widetilde{\mu}_t(s)} \le d_t^\pi(s).$$

According to the definition of $\mathsf{D}_T(\pi)$ in Section 4, line (366) follows after summing over actions.

Applying Lemma D.13 once to $P_t^{\mathrm{up},s}$ and once to $P_t^{\mathrm{low},s}$, and using (366), gives the two explicit bounds

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s} g_t(s)\left|\mu^{P_t^{\mathrm{up},s},\pi_t}(s) - \mu^{\pi_t}(s)\right|\right] \le \sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + \mathsf{Y}_T, \tag{367}$$

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s} g_t(s)\left|\mu^{P_t^{\mathrm{low},s},\pi_t}(s) - \mu^{\pi_t}(s)\right|\right] \le \sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + \mathsf{Y}_T. \tag{368}$$

The floor part satisfies

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s} g_t(s)\gamma_t\right] \le \sum_{t=1}^{T}\sum_{s}\gamma_t = S\sum_{t=1}^{T}\gamma_t \le S\mathsf{G}_T. \tag{369}$$

Substituting (364), (367), (368), and (369) into (363) yields

$$\mathsf{C}_{\mathrm{cur}} \le 2HL_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right]. \tag{370}$$

Bounding $\mathsf{C}_{\mathrm{next}}$ For the next-state quantity $\mathsf{C}_{\mathrm{next}}$, define $\widetilde{d}_t^\pi(x) \coloneqq \sum_{s,a} d_t^\pi(s,a)P(x \mid s,a)$, where only predecessors in the layer immediately before x contribute. Using (356) and the true occupancy-flow equation gives

$$\widetilde{d}_t^\pi(x) \le \sum_{s,a}\mu^{\pi_t}(s,a)P(x \mid s,a) = \mu^{\pi_t}(x), \tag{371}$$

$$\sum_x \widetilde{d}_t^{\pi}(x) \le \sum_{s,a} d_t^{\pi}(s,a)\,. \tag{372}$$

The definition (360) can now be rewritten exactly as

$$\mathsf{C}_{\text{next}} = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\widetilde{d}_t^{\pi}(x)c_t(x)\right]. \tag{373}$$

Set

$$\widetilde{g}_t(x) \coloneqq \min\left\{1, \frac{\widetilde{d}_t^{\pi}(x)}{\widetilde{\mu}_t(x)}\right\}. \tag{374}$$

On $\mathcal{E}$, (371) implies $\widetilde{d}_t^{\pi}(x) \le \mu^{\pi_t}(x) \le \widetilde{\mu}_t(x)$, so the minimum in (374) equals the ratio. Its quadratic quantity satisfies

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\mu^{\pi_t}(x)\widetilde{g}_t(x)^2\right] = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\mu^{\pi_t}(x)\frac{\widetilde{d}_t^{\pi}(x)^2}{\widetilde{\mu}_t(x)^2}\right] \le \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\widetilde{d}_t^{\pi}(x)\right] \tag{375}$$

$$\le \mathsf{D}_T(\pi)\,. \tag{376}$$

The bounds $\widetilde{d}_t^{\pi}(x) \le \mu^{\pi_t}(x) \le \widetilde{\mu}_t(x)$ give the pointwise calculation

$$\mu^{\pi_t}(x)\frac{\widetilde{d}_t^{\pi}(x)^2}{\widetilde{\mu}_t(x)^2} \le \widetilde{\mu}_t(x)\frac{\widetilde{d}_t^{\pi}(x)^2}{\widetilde{\mu}_t(x)^2} = \frac{\widetilde{d}_t^{\pi}(x)^2}{\widetilde{\mu}_t(x)} \le \widetilde{d}_t^{\pi}(x)\,.$$

This gives (375). According to the definition of $\mathsf{D}_T(\pi)$ in Section 4, (376) then follows from (372).

For the same upper and lower selectors introduced above, the weighted occupancy lemma applied with the concrete weight $\widetilde{g}_t(x)$ gives the two separate inequalities

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\widetilde{g}_t(x)\left|\mu^{P_t^{\text{up},x},\pi_t}(x) - \mu^{\pi_t}(x)\right|\right] \le \sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + \mathsf{Y}_T\,, \tag{377}$$

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\widetilde{g}_t(x)\left|\mu^{P_t^{\text{low},x},\pi_t}(x) - \mu^{\pi_t}(x)\right|\right] \le \sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + \mathsf{Y}_T\,. \tag{378}$$

The floor contribution for the weight $\widetilde{g}_t(x)$ is

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\widetilde{g}_t(x)\gamma_t\right] \le S\sum_{t=1}^{T}\gamma_t \le S\mathsf{G}_T\,. \tag{379}$$

Substituting the definition of $c_t(x)$ into (373) first gives the exact identity

$$\mathsf{C}_{\text{next}} = 2HL_\sigma\,\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,x}\widetilde{g}_t(x)\big(\overline{\mu}_t(x) - \underline{\mu}_t(x) + \gamma_t\big)\right].$$

Using (364), the selector bounds (377)– (378), and (379) gives

$$\begin{aligned}\mathsf{C}_{\text{next}} &\le 2HL_\sigma\left[\big(\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + \mathsf{Y}_T\big) + \big(\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + \mathsf{Y}_T\big) + S\mathsf{G}_T\right] \\ &= 2HL_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right].\end{aligned} \tag{380}$$

Bounding C_{rad} For the transition-radius quantity, (322) implies

$$0 \le F_t(x) \le H\max_y c_t(y) \le 2H^2L_\sigma\,. \tag{381}$$

We then bound C_{rad} in the following steps.

(i) *Selector-relative radius via Young's inequality.* We adapt the transition-bias argument of Dann et al. (2023b, Appendix F) to the positive discrepancy d_t^π and the present heavy-tail scale L_σ. Write $Q_t = P_t^C$ and $\rho_t(s,a) = \iota/N_t^+(s,a)$. Fix (t,s,a,x) on $\mathcal{E}$, and abbreviate $e = |Q_t(x \mid s,a) - P(x \mid s,a)|$ and $\rho = \rho_t(s,a)$. Then

$$e \le 8\sqrt{P(x \mid s,a)\rho} + 56\rho \le 8\sqrt{Q_t(x \mid s,a)\rho} + 8\sqrt{e\rho} + 56\rho \le 8\sqrt{Q_t(x \mid s,a)\rho} + \frac{e}{2} + 88\rho\,,$$

where the first inequality is by (479), which applies because $Q_t(\cdot \mid s,a) = P_{t,s,a}^C \in \mathcal{P}_t$ and we work on $\mathcal{E}$; the second inequality uses $P(x \mid s,a) \le Q_t(x \mid s,a) + e$, which is immediate from the definition of e, together with $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$; and the third inequality is the AM–GM bound $8\sqrt{e\rho} = 2\sqrt{e/2}\,\sqrt{32\rho} \le e/2 + 32\rho$. Subtracting $e/2$ from both sides and multiplying by 2 gives $e \le 16\sqrt{Q_t(x \mid s,a)\rho} + 176\rho$. Consequently, for every $0 < \lambda \le 1$,

$$\begin{aligned}
|Q_t(x \mid s,a) - P(x \mid s,a)| &\le 16\sqrt{Q_t(x \mid s,a)\rho_t(s,a)} + 176\rho_t(s,a) \\
&\le \lambda Q_t(x \mid s,a) + \left(\frac{64}{\lambda} + 176\right)\rho_t(s,a) \\
&\le \lambda Q_t(x \mid s,a) + \frac{240}{\lambda}\rho_t(s,a)\,, \qquad (382)
\end{aligned}$$

where the first inequality is the rearranged bound above; the second inequality is Young's inequality $16\sqrt{Q_t(x \mid s,a)\rho} = 2\sqrt{\lambda Q_t(x \mid s,a)}\sqrt{64\rho/\lambda} \le \lambda Q_t(x \mid s,a) + 64\rho/\lambda$; and the last inequality uses $\lambda \le 1$, so that $176 \le 176/\lambda$. Multiplying (382) by the nonnegative weight $d_t^\pi(s,a)F_t(x)$ (recall $c_t \ge 0$, hence $F_t \ge 0$), summing over (t,s,a,x), and taking expectations on $\mathcal{E}$, the definition (361) separates the radius cost into a propagated term and a reciprocal-count term:

$$\begin{aligned}
\mathsf{C}_{\mathrm{rad}} \le \lambda\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^\pi(s,a)\sum_x Q_t(x \mid s,a)F_t(x)\right] \\
+ \frac{240}{\lambda}\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^\pi(s,a)\rho_t(s,a)\sum_x F_t(x)\right]. \qquad (383)
\end{aligned}$$

All successor sums here are over $x \in \mathcal{S}_{h(s)+1}$, with $F_t = 0$ on the terminal layer.

(ii) *Propagation through a feasible suffix kernel.* For $h(y) > h(s)$, let $K_t(y \mid s,a)$ be the probability of visiting y in layer $h(y)$ when starting from (s,a) and then following Q_t, π_t. Set this quantity to zero when $h(y) \le h(s)$, and define, for nonterminal y,

$$\widehat{d}_t^\pi(y) := \sum_{s,a:\ h(s)<h(y)} d_t^\pi(s,a)K_t(y \mid s,a)\,. \qquad (384)$$

By the definition of F_t, $\sum_x Q_t(x \mid s,a)F_t(x)$ is the expected sum of c_t over the nonterminal states visited after (s,a) under (Q_t,π_t), which equals $\sum_{y\notin\mathcal{S}_H} K_t(y \mid s,a)c_t(y)$. Multiplying by $d_t^\pi(s,a)$ and interchanging the finite nonnegative sums over (s,a) and y gives, on every history,

$$\sum_{s,a} d_t^\pi(s,a)\sum_x Q_t(x \mid s,a)F_t(x) = \sum_{y\notin\mathcal{S}_H} \widehat{d}_t^\pi(y)c_t(y)\,. \qquad (385)$$

Next, fix a predecessor layer $h < h(y)$, and let $P_t^{(h)}$ be the hybrid kernel that uses P for transitions out of layers below h and Q_t from layer h onward. On $\mathcal{E}$,

$$\begin{aligned}
\sum_{s\in\mathcal{S}_h,a} d_t^\pi(s,a)K_t(y \mid s,a) &\le \sum_{s\in\mathcal{S}_h,a} \mu^{P,\pi_t}(s,a)K_t(y \mid s,a) \\
&= \mu^{P_t^{(h)},\pi_t}(y) \le \overline{\mu}_t(y)\,, \qquad (386)
\end{aligned}$$

where the first inequality uses $d_t^\pi(s,a) \le \mu^{\pi_t}(s,a) = \mu^{P,\pi_t}(s,a)$ from (356); the equality holds because $P_t^{(h)}$ coincides with P before layer h, so that $\mu^{P_t^{(h)},\pi_t}(s,a) = \mu^{P,\pi_t}(s,a)$ for $s \in \mathcal{S}_h$, and coincides with Q_t from layer h onward, so that the probability of reaching y from

(s,a) is exactly $K_t(y \mid s,a)$; and the last inequality holds because both P and Q_t belong to $\mathcal{P}_t$ on $\mathcal{E}$, so the rowwise rectangularity of $\mathcal{P}_t$ implies $P_t^{(h)} \in \mathcal{P}_t$, while $\overline{\mu}_t(y)$ is the maximal occupancy of y over $\mathcal{P}_t$. Summing (386) over the at most H predecessor layers $h < h(y)$ and using $\widetilde{\mu}_t(y) = \overline{\mu}_t(y) + \gamma_t \geq \overline{\mu}_t(y)$ yields

$$0 \leq \widehat{d}_t^{\pi}(y) \leq H\overline{\mu}_t(y) \leq H\widetilde{\mu}_t(y) \quad \text{on } \mathcal{E}\,. \tag{387}$$

Moreover, a suffix started from (s,a) visits at most one state in each of at most H nonterminal layers, so that $\sum_y K_t(y \mid s,a) \leq H$ and hence

$$\sum_{y \notin \mathcal{S}_H} \widehat{d}_t^{\pi}(y) \leq H \sum_{s,a} d_t^{\pi}(s,a)\,. \tag{388}$$

(iii) *Normalized weights and the sharp square-root term.* Define on every history, including confidence-failure histories,

$$g_t^{\mathrm{rad}}(y) := \min\left\{1, \frac{\widehat{d}_t^{\pi}(y)}{H\widetilde{\mu}_t(y)}\right\}. \tag{389}$$

This is a $[0,1]$-valued $\mathcal{F}_{t-1}$-measurable weight: Q_t, π_t, the true occupancies, and the fixed comparator are determined before the episode-t randomization. The definition does not condition the algorithm on the future-dependent event $\mathcal{E}$. Its quadratic quantity in Lemma D.13 satisfies

$$\begin{aligned} X_{g^{\mathrm{rad}}} &= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,y}\mu^{\pi_t}(y)\frac{\widehat{d}_t^{\pi}(y)^2}{H^2\widetilde{\mu}_t(y)^2}\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,y}\frac{\widehat{d}_t^{\pi}(y)^2}{H^2\widetilde{\mu}_t(y)}\right] \leq \frac{1}{H}\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,y}\widehat{d}_t^{\pi}(y)\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^{\pi}(s,a)\right] \leq \mathsf{D}_T(\pi)\,, \end{aligned} \tag{390}$$

where the equality combines the definition (495) of X_g with (387), which makes the minimum in (389) inactive on $\mathcal{E}$; the first inequality uses $\mu^{\pi_t}(y) \leq \overline{\mu}_t(y) \leq \widetilde{\mu}_t(y)$ on $\mathcal{E}$; the second inequality uses $\widehat{d}_t^{\pi}(y)/(H\widetilde{\mu}_t(y)) \leq 1$, again by (387); the third inequality is (388); and the last inequality follows from the definition of $\mathsf{D}_T(\pi)$, since $d_t^{\pi}(s,a) = [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+$ and $\mathbb{1}\{\mathcal{E}\} \leq 1$. Write $w_t(y) = \overline{\mu}_t(y) - \underline{\mu}_t(y) + \gamma_t$, so that $c_t(y) = 2HL_\sigma w_t(y)/\widetilde{\mu}_t(y)$. Then

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,y} g_t^{\mathrm{rad}}(y) w_t(y)\right] &\leq 2\left(\sqrt{\mathsf{X}_T X_{g^{\mathrm{rad}}}} + \mathsf{Y}_T\right) + S\sum_{t=1}^{T}\gamma_t \\ &\leq 2\sqrt{\mathsf{X}_T \mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\,, \end{aligned} \tag{391}$$

where the first inequality decomposes $w_t(y)$ by (364), applies Lemma D.13 with $g_{\max} = 1$ separately to the upper and lower occupancy selectors, and bounds the floor part by $\sum_y g_t^{\mathrm{rad}}(y)\gamma_t \leq S\gamma_t$ because $g_t^{\mathrm{rad}} \leq 1$; and the second inequality uses (390) and $\sum_t \gamma_t \leq \mathsf{G}_T$. Consequently, the first term of (383) is bounded by

$$\begin{aligned} &\lambda\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^{\pi}(s,a)\sum_x Q_t(x \mid s,a)F_t(x)\right] \\ &= \lambda\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,y}\widehat{d}_t^{\pi}(y)c_t(y)\right] = 2\lambda H^2 L_\sigma \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,y} g_t^{\mathrm{rad}}(y)w_t(y)\right] \\ &\leq 2\lambda H^2 L_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right], \end{aligned} \tag{392}$$

where the first equality is (385); the second equality substitutes $c_t(y) = 2HL_\sigma w_t(y)/\widetilde{\mu}_t(y)$ and uses $\widehat{d}_t^{\pi}(y)/(H\widetilde{\mu}_t(y)) = g_t^{\mathrm{rad}}(y)$ on $\mathcal{E}$; and the inequality is (391). Unlike the former absorption, this inequality holds uniformly in H.

(iv) *Reciprocal-count remainder and the choice of λ.* We use the envelope (381) *only* for the second term of (383):

$$\frac{240}{\lambda}\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s,a} d_t^{\pi}(s,a)\rho_t(s,a)\sum_x F_t(x)\right]$$
$$\le \frac{480H^2L_\sigma S\iota}{\lambda}\mathbb{E}\left[\sum_{t,s,a}\frac{\mu^{\pi_t}(s,a)}{N_t^+(s,a)}\right] \le \frac{480}{\lambda}L_\sigma H^2S^2A\mathcal{L}_T\,, \tag{393}$$

where the first inequality uses $0 \le F_t(x) \le 2H^2L_\sigma$ from (381), the fact that each (s,a) has at most S successors x, $\rho_t(s,a) = \iota/N_t^+(s,a)$, $d_t^{\pi}(s,a) \le \mu^{\pi_t}(s,a)$ from (356), and $\mathbb{1}\{\mathcal{E}\} \le 1$; and the second inequality applies Lemma D.10 to each of the at most SA nonterminal pairs and uses $\mathcal{L}_T = \iota(2+\log(eT))$. Substituting (392) and (393) into (383) gives

$$\begin{aligned}\mathsf{C}_{\mathrm{rad}} &\le 2\lambda H^2L_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right] + \frac{480}{\lambda}L_\sigma H^2S^2A\mathcal{L}_T\\ &= 2HL_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right] + 480L_\sigma H^3S^2A\mathcal{L}_T\\ &\le 2HL_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right] + 480L_\sigma H^2S^4A^2\mathcal{L}_T\,,\end{aligned} \tag{394}$$

where the equality chooses $\lambda = 1/H \in (0,1]$, and the last inequality uses $H \le S$, which holds because there is at least one nonterminal state per layer, so that $H^3S^2A \le H^2S^4A^2$. This step uses $H \le S$ only to bound the *additive reciprocal-count remainder*; it does not compare the two square-root coefficients. The extra terms are already covered by $\mathsf{R}_0^{\mathrm{unk}}$ in (209) after choosing its ε-only constant sufficiently large.

Putting Everything Together Finally, substitute (370), (380), and (394) into the complete decomposition (359):

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s}\mu^{\pi}(s)\left\langle p_{t,s} - e_{\pi(s)}, r_t(s,\cdot) + C_t(s,\cdot)\right\rangle\right]\\ &\le 12HL_\sigma\left[2\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + 2\mathsf{Y}_T + S\mathsf{G}_T\right] + 960L_\sigma H^2S^4A^2\mathcal{L}_T\\ &\le c_{\varepsilon,1}HL_\sigma\sqrt{\mathsf{X}_T\mathsf{D}_T(\pi)} + c_\varepsilon L_\sigma H^2S^4A^2\mathcal{L}_T + c_\varepsilon HL_\sigma(\mathsf{Y}_T + S\mathsf{G}_T)\,,\end{aligned}$$

which is (357). □

The following Bellman-residual estimate supplies the nonnegativity and pointwise upper bound used in Lemma C.13.

Lemma C.14 (Nonnegative Bellman residual for a signed bias)**.** *On $\mathcal{E}_t$, recall from Section 3 that*

$$\mathcal{L}_t^{\mathrm{bias}}(s,a) = r_t(s,a) + C_t(s,a) + c_t(s)\,, \tag{395}$$

and recall from Section 4 that

$$\zeta_t(s,a) = \mathcal{L}_t^{\mathrm{bias}}(s,a) - \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim p_{t,s'}}}\left[\mathcal{L}_t^{\mathrm{bias}}(s',a')\right]\,. \tag{396}$$

Then $\zeta_t(s,a) \ge 0$ for every state-action pair. Moreover,

$$\begin{aligned}\zeta_t(s,a) &\le 2c_t(s) + 2\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s')\right]\\ &\quad + 2\sum_{s'}|P_t^C(s'\mid s,a) - P(s'\mid s,a)|\mathbb{E}_{P_t^C,\pi_t}\left[\sum_{j=h(s')}^{H-1}c_t(s_j)\,\middle|\, s_{h(s')} = s'\right]\,.\end{aligned} \tag{397}$$

Proof. Work on $\mathcal{E}_t$. Feasibility of $P \in \mathcal{P}_t$ gives $\underline{\mu}_t(s) \le \mu^{\pi_t}(s) \le \overline{\mu}_t(s)$. Using $\widetilde{\mu}_t(s) = \overline{\mu}_t(s) + \gamma_t$ and $w_t(s) = \overline{\mu}_t(s) - \underline{\mu}_t(s) + \gamma_t$, the numerator in (238) satisfies

$$|\widetilde{\mu}_t(s) - \mu^{\pi_t}(s)| = |\overline{\mu}_t(s) + \gamma_t - \mu^{\pi_t}(s)| \le \overline{\mu}_t(s) - \underline{\mu}_t(s) + \gamma_t = w_t(s)\,. \tag{398}$$

The centered-Q estimate $|\bar{Q}_t(s,a)| \le 2HL_\sigma$ is the same as in the proof of Lemma B.12 in Appendix B.4; see (137). We omit its repeated derivation. Together with (398) and (238), it gives

$$|r_t(s,a)| \le 2HL_\sigma \frac{w_t(s)}{\widetilde{\mu}_t(s)} = c_t(s)\,. \tag{399}$$

Define the state-averaged confidence envelope

$$\bar{C}_t(s) := \sum_{a\in\mathcal{A}} p_{t,s,a} C_t(s,a)\,. \tag{400}$$

Because P_t^C is the rectangular selector attaining the maximum in (8),

$$C_t(s,a) = \mathbb{E}_{s'\sim P_t^C(\cdot|s,a)}\left[2c_t(s') + \bar{C}_t(s')\right]\,. \tag{401}$$

Starting from the definition (396), substitute (395), (400), and (401) without yet applying an inequality:

$$\begin{aligned}\zeta_t(s,a) &= r_t(s,a) + c_t(s) + C_t(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[c_t(s') + \bar{r}_t(s') + \bar{C}_t(s')\right]\\ &= r_t(s,a) + c_t(s) + \mathbb{E}_P\left[c_t(s') - \bar{r}_t(s')\right]\\ &\quad + \mathbb{E}_{P_t^C}\left[2c_t(s') + \bar{C}_t(s')\right] - \mathbb{E}_P\left[2c_t(s') + \bar{C}_t(s')\right]\,,\end{aligned} \tag{402}$$

where $\bar{r}_t(s') := \sum_{a'} p_{t,s',a'} r_t(s',a')$ and the symbols $\mathbb{E}_P, \mathbb{E}_{P_t^C}$ in (402) refer to the next state from the fixed pair (s,a).

Given (399), the bounds for the local term $r_t(s,a) + c_t(s)$ and the transported term $\mathbb{E}_P[c_t(s') - \bar{r}_t(s')]$ are exactly those proved in (142)–(143) of Lemma B.12 in Appendix B.4. Thus their sum lies between zero and $2c_t(s) + 2\mathbb{E}_P[c_t(s')]$, and we omit the identical bracket calculation. The additional transition-selection difference in (402) is nonnegative because $P_t^C(\cdot \mid s,a)$ maximizes the expectation of $2c_t + \bar{C}_t$ over the row set and $P(\cdot \mid s,a)$ is feasible on $\mathcal{E}_t$. Thus

$$\zeta_t(s,a) \ge 0\,. \tag{403}$$

For the upper bound, it remains to control the transition-selection difference:

$$\begin{aligned}&\mathbb{E}_{P_t^C}[2c_t(s') + \bar{C}_t(s')] - \mathbb{E}_P[2c_t(s') + \bar{C}_t(s')]\\ &\le \sum_{s'} |P_t^C(s' \mid s,a) - P(s' \mid s,a)|\big(2c_t(s') + \bar{C}_t(s')\big)\,.\end{aligned} \tag{404}$$

The remaining quantity in (404) is the explicit envelope $2c_t(s') + \bar{C}_t(s')$. Averaging (401) over $a' \sim p_{t,s'}$ and iterating backward from the terminal layer gives

$$\begin{aligned}\bar{C}_t(s') &= 2\mathbb{E}_{P_t^C,\pi_t}\left[\sum_{j=h(s')+1}^{H-1} c_t(s_j)\,\middle|\, s_{h(s')} = s'\right]\,,\\ 2c_t(s') + \bar{C}_t(s') &= 2\mathbb{E}_{P_t^C,\pi_t}\left[\sum_{j=h(s')}^{H-1} c_t(s_j)\,\middle|\, s_{h(s')} = s'\right]\,.\end{aligned} \tag{405}$$

Substituting the local-term bound recalled above, (404), and (405) into (402) gives exactly (397).

Finally, the terminal convention and (396) give the same Bellman identification as (145) in the proof of Lemma B.12 in Appendix B.4. The backward-induction argument is unchanged and is omitted: $\mathcal{L}_t^{\mathrm{bias}}$ is the Q-function of π_t for the artificial loss ζ_t, which is nonnegative by (403). □

C.5 Bounding $V^{P_t^B,\pi_t}(s_0;b_t)$

The remaining bonus term in (247) is $\mathbb{E}[\sum_{t=1}^T V^{P_t^B,\pi_t}(s_0;b_t)]$, where $V^{P_t^B,\pi_t}(s_0;b_t) = \sum_s \mu^{P_t^B,\pi_t}(s) b_t(s)$. The following proposition gives the complete count bound. Its proof uses the changing-rate, heavy-tail, and squared-confidence estimates established in the subsections below.

Proposition C.15 (Optimistic bonus-count bound)**.** *For every deterministic policy π,*

$$3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)b_t(s)\right] \le \mathsf{A}_T\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha} + c_\varepsilon SHL_\sigma \mathsf{G}_T + 3\mathsf{R}_\nu + 3T\delta\,\mathsf{F}_\nu\,. \tag{406}$$

Proof. Use the same three-component decomposition of $b_t(s)$ as in the final step of the proof of Proposition B.13 in Appendix B.5. Write $b_t^{\mathrm{rate}}(s) = \mathbb{1}\{t\ge 2\}(\beta_t(s)-\beta_{t-1}(s))h_t(s)$ and $b_t^{\mathrm{ht}}(s) = K_{\mathrm{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s)$ for the first two components. The optimistic occupancy weight and the squared-confidence component give

$$3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)b_t(s)\right] = 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\big(b_t^{\mathrm{rate}}(s)+b_t^{\mathrm{ht}}(s)\big)\right] + 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right]. \tag{407}$$

Apply (410) to the first expectation in (407), with $q_t(s,a)=\widetilde{\mu}_t(s)p_{t,s,a}$ and $a_s^\star=\pi(s)$:

$$\begin{aligned} &3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\big(b_t^{\mathrm{rate}}(s)+b_t^{\mathrm{ht}}(s)\big)\right] \\ &\le 3\left(5^{1-\alpha}c_\beta + 2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\right)HL_\sigma\Lambda_T^{1-\alpha}\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha} \\ &= \mathsf{A}_T\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha}\,. \end{aligned} \tag{408}$$

Here, the last equality uses the definitions of $c_{\mathrm{cnt},\varepsilon}$ and A_T in (198).

Apply Lemma C.17 to the second expectation in (407). Multiplying (413) by three gives

$$3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right] \le 3\mathsf{R}_\nu + 3T\delta\,\mathsf{F}_\nu\,. \tag{409}$$

Substituting both (408) and (409) into (407) produces the complete bound

$$\begin{aligned} 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)b_t(s)\right] &\le \mathsf{A}_T\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha} + 3\mathsf{R}_\nu + 3T\delta\,\mathsf{F}_\nu \\ &\le \mathsf{A}_T\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^{\alpha} + c_\varepsilon SHL_\sigma\mathsf{G}_T + 3\mathsf{R}_\nu + 3T\delta\,\mathsf{F}_\nu\,. \end{aligned}$$

Here, the last inequality adds $c_\varepsilon SHL_\sigma\mathsf{G}_T \ge 0$, giving (406). □

C.5.1 Comparator-sensitive Tsallis Quantities

We directly invoke Lemma B.14 in Appendix B.5, since h_α, z_α, and ρ have the same definitions here. Thus the comparator-removal bounds (170) and (171) are available for every comparator action.

C.5.2 The Occupancy-adaptive Clock

The clock estimates (176)–(178) from Lemma B.15 in Appendix B.5 apply on every history: they use only the definition of G_t, $\gamma_t = \min\{1,\Gamma/t\}$ with $\Gamma\ge 1$, and $\gamma_t \le \widetilde{\mu}_t(s) = \overline{\mu}_t(s)+\gamma_t \le 2$, which follows from $0\le\overline{\mu}_t(s)\le 1$ and $\gamma_t\le 1$.

We also use the weighted clock-to-count inequality (186) from Lemma B.16 in Appendix B.5, with the present $\widetilde{\mu}_t(s)$ and $G_t(s)$. Its factorization and Hölder argument require only the logarithmic-potential estimate (178) just established.

C.5.3 Changing-rate and Heavy-tail Parts

Lemma C.16 (Optimistic penalty and heavy-tail bonus)**.** *Fix a deterministic comparator policy π. Then*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\left(\mathbb{1}\{t\geq 2\}(\beta_t(s)-\beta_{t-1}(s))h_t(s)+K_{\mathrm{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s)\right)\right]$$
$$\leq\left(5^{1-\alpha}c_\beta+2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}\right)HL_\sigma\Lambda_T^{1-\alpha}\sum_{s}\sum_{a\neq\pi(s)}\widetilde{n}_T(s,a)^\alpha\,. \tag{410}$$

Proof. Fix a state s and use the abbreviations $\mu_t^B(s):=\mu^{P_t^B,\pi_t}(s)$, $x_t:=\widetilde{\mu}_t(s)^{1-\varepsilon}$, $G_t:=G_t(s)$. Because $P_t^B\in\mathcal{P}_t$ and $\overline{\mu}_t(s)$ is the maximum state occupancy over all kernels in $\mathcal{P}_t$, one has, pathwise,

$$\mu_t^B(s)=\mu^{P_t^B,\pi_t}(s)\leq\max_{\widetilde{P}\in\mathcal{P}_t}\mu^{\widetilde{P},\pi_t}(s)=\overline{\mu}_t(s)\leq\overline{\mu}_t(s)+\gamma_t=\widetilde{\mu}_t(s)\,. \tag{411}$$

Here, the middle equality follows from the definition of $\overline{\mu}_t(s)$ in Section 3, and the last equality follows from the definition of $\widetilde{\mu}_t(s)$ there.

The changing-rate and heavy-tail calculations in Proposition B.13, Appendix B.5, use only an occupancy weight bounded by $\widetilde{\mu}_t(s)$, so (411) permits the weight $\mu_t^B(s)$. Applying (159), (162), and (166) with the present clock gives

$$\sum_{t=1}^{T}\mu_t^B(s)\left[\mathbb{1}\{t\geq 2\}(\beta_t(s)-\beta_{t-1}(s))h_t(s)+K_{\mathrm{ht}}\Sigma_t(s)\beta_t(s)^{1-\varepsilon}z_t(s)\right]$$
$$\leq c_{\mathrm{ph}}HL_\sigma\Lambda_T^{1-\alpha}\sum_{a\neq\pi(s)}\left(\sum_{t=1}^{T}\widetilde{\mu}_t(s)p_{t,s,a}\right)^\alpha, \tag{412}$$

where $c_{\mathrm{ph}}=5^{1-\alpha}c_\beta+2K_{\mathrm{ht}}c_\beta^{1-\varepsilon}$.

The Jensen step (167) also applies to the present count, whose expectation is, according to the definition of $\widetilde{n}_T(s,a)$ in Section 4, $\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{\mu}_t(s)p_{t,s,a}\right]=\widetilde{n}_T(s,a)$. Taking expectations in (412), applying this same Jensen step, and summing over states gives (410). □

C.5.4 The Squared-confidence Correction

Lemma C.17 (Optimistic cost of ν_t)**.** *The squared-confidence correction satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right]\leq\mathsf{R}_\nu+T\delta\,\mathsf{F}_\nu\,. \tag{413}$$

Proof. We first derive an event-free pointwise upper bound for $\nu_t(s)$. The square-to-linear step in the proof of Lemma B.17 in Appendix B.5 applies unchanged, since (323) gives the same envelope $0\leq C_t(s,a)\leq 4H^2L_\sigma$ on every history. We therefore use (191) and omit its repeated derivation. The unknown-transition bound additionally needs decay in t. Since $\widetilde{\mu}_\tau(s)\leq 2$ for every τ, one has $\widetilde{\mu}_\tau(s)^{1-\varepsilon}\geq 2^{1-\varepsilon}$. Directly from the clock definition,

$$G_t(s)=1+\sum_{\tau=1}^{t}\widetilde{\mu}_\tau(s)^{1-\varepsilon}\geq 1+t2^{1-\varepsilon}\geq t2^{1-\varepsilon}\,,$$
$$\beta_t(s)=\beta_0+c_\beta HL_\sigma G_t(s)^\alpha\geq c_\beta HL_\sigma\left(t2^{1-\varepsilon}\right)^\alpha=c_\beta 2^{(1-\varepsilon)\alpha}HL_\sigma t^\alpha=:c'_{\beta,\varepsilon}HL_\sigma t^\alpha\,. \tag{414}$$

Here, the bound on $G_t(s)$ uses the clock-increment bound and that on $\beta_t(s)$ uses (13); $c'_{\beta,\varepsilon}>0$ depends only on ε.

After the common square-to-linear step and $\rho(p_{t,s})^{1-\alpha} \le 1$, retain $\beta_t(s)$ in the denominator and apply (414):

$$\begin{aligned}\nu_t(s) &\le \frac{64H^2L_\sigma}{(1-\alpha)\beta_t(s)} \sum_a p_{t,s,a} C_t(s,a) \\ &\le \frac{64H^2L_\sigma}{(1-\alpha)c'_{\beta,\varepsilon} H L_\sigma t^\alpha} \sum_a p_{t,s,a} C_t(s,a) \le c_\varepsilon H t^{-\alpha} \sum_a p_{t,s,a} C_t(s,a)\,. \end{aligned} \tag{415}$$

The second inequality in (415) substitutes the lower bound for $\beta_t(s)$; the last inequality absorbs only constants depending on ε.

We next convert the C_t-cost weighted by the occupancy of the robust-bonus selector P_t^B into an upper-occupancy width. For each predecessor layer $h \in \{0,\dots,H-2\}$, define the hybrid kernel $P_t^{B\to C,h}$ to use P_t^B on rows whose states lie in layers $0,\dots,h-1$ and to use P_t^C on rows whose states lie in layers $h,\dots,H-1$. Rectangularity implies $P_t^{B\to C,h} \in \mathcal{P}_t$. Iterating the recursion (8) from a state $s \in \mathcal{S}_h$ gives

$$\sum_a p_{t,s,a} C_t(s,a) = 2 \sum_{j=h+1}^{H-1} \mathbb{E}_{P_t^C,\pi_t}\left[c_t(s_{t,j}) \mid s_{t,h} = s\right]\,. \tag{416}$$

Here, the terminal continuation is zero and the factor two in (8) appears at each future state.

Multiply (416) by $\mu^{P_t^B,\pi_t}(s)$, sum first over states in layer h, and then over h. The resulting hybrid occupancies satisfy

$$\begin{aligned}&\sum_s \mu^{P_t^B,\pi_t}(s) \sum_a p_{t,s,a} C_t(s,a) \\ &= 2 \sum_{h=0}^{H-2} \sum_{j=h+1}^{H-1} \sum_{x\in\mathcal{S}_j} \mu^{P_t^{B\to C,h},\pi_t}(x) c_t(x) \\ &\le 2 \sum_{h=0}^{H-2} \sum_{j=h+1}^{H-1} \sum_{x\in\mathcal{S}_j} \overline{\mu}_t(x) c_t(x) = 2 \sum_{j=1}^{H-1} j \sum_{x\in\mathcal{S}_j} \overline{\mu}_t(x) c_t(x) \le 2H \sum_x \overline{\mu}_t(x) c_t(x)\,. \end{aligned} \tag{417}$$

Here, the first equality uses the hybrid-kernel construction; the first inequality uses rectangularity, which gives $P_t^{B\to C,h} \in \mathcal{P}_t$ and hence hybrid occupancy at most $\overline{\mu}_t(x)$ on every history; the next equality counts the j predecessor layers, and the last inequality uses $j \le H$.

The definition of c_t in Section 3 and $\overline{\mu}_t(x) \le \widetilde{\mu}_t(x) = \overline{\mu}_t(x) + \gamma_t$ give

$$\overline{\mu}_t(x) c_t(x) = 2HL_\sigma \overline{\mu}_t(x) \frac{\overline{\mu}_t(x) - \underline{\mu}_t(x) + \gamma_t}{\widetilde{\mu}_t(x)} \le 2HL_\sigma \big(\overline{\mu}_t(x) - \underline{\mu}_t(x) + \gamma_t\big)\,. \tag{418}$$

Here, the inequality uses $\overline{\mu}_t(x)/\widetilde{\mu}_t(x) \le 1$.

Multiplying (415) by $\mu^{P_t^B,\pi_t}(s)$, summing over s, and substituting (417) and (418) gives the complete episodewise bound

$$\begin{aligned}\sum_s \mu^{P_t^B,\pi_t}(s) \nu_t(s) &\le c_\varepsilon H t^{-\alpha} \sum_s \mu^{P_t^B,\pi_t}(s) \sum_a p_{t,s,a} C_t(s,a) \\ &\le c_\varepsilon H t^{-\alpha} \left(2H \sum_x \overline{\mu}_t(x) c_t(x)\right) \\ &\le c_\varepsilon H t^{-\alpha} \left(4H^2 L_\sigma \sum_x [\overline{\mu}_t(x) - \underline{\mu}_t(x) + \gamma_t]\right) \\ &\le c_\varepsilon H^3 L_\sigma t^{-\alpha} \sum_x \big(\overline{\mu}_t(x) - \underline{\mu}_t(x) + \gamma_t\big)\,. \end{aligned} \tag{419}$$

Here, the three inequalities use (415), (417), and (418), respectively.

We now bound the two occupancy-width terms in (419) by applying Lemma D.13 explicitly to the upper and lower selectors. For every state x, let $P_t^{\mathrm{up},x}$ and $P_t^{\mathrm{low},x}$ be the measurable selectors satisfying

$$\mu^{P_t^{\mathrm{up},x},\pi_t}(x) = \overline{\mu}_t(x)\,, \qquad \mu^{P_t^{\mathrm{low},x},\pi_t}(x) = \underline{\mu}_t(x)\,.$$

On $\mathcal{E}$, the true occupancy lies between these two extrema, so

$$\overline{\mu}_t(x) - \underline{\mu}_t(x) = \left|\mu^{P_t^{\mathrm{up},x},\pi_t}(x) - \mu^{\pi_t}(x)\right| + \left|\mu^{P_t^{\mathrm{low},x},\pi_t}(x) - \mu^{\pi_t}(x)\right|. \tag{420}$$

Take the predictable weight $g_t(x) := t^{-\alpha} \in [0,1]$. Its quadratic quantity in (495) satisfies

$$\begin{aligned} X_g &= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_x \mu^{\pi_t}(x)t^{-2\alpha}\right] \\ &\le \sum_{t=1}^{T} t^{-2\alpha}\sum_x \mu^{\pi_t}(x) = H\sum_{t=1}^{T} t^{-2\alpha} \le H(1+\log(eT))\,. \end{aligned}$$

Here, the equality uses $\sum_x \mu^{\pi_t}(x) = H$ and the last inequality uses $2\alpha = 2/\varepsilon \ge 1$ to bound the power sum by the harmonic sum.

Applying (496) to the concrete upper selector gives

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_x t^{-\alpha}\left|\mu^{P_t^{\mathrm{up},x},\pi_t}(x) - \mu^{\pi_t}(x)\right|\right] \\ &\le \sqrt{\mathsf{X}_T X_g} + \mathsf{Y}_T \le \sqrt{H\mathsf{X}_T(1+\log(eT))} + \mathsf{Y}_T\,. \end{aligned} \tag{421}$$

For the concrete lower selector, the weighted occupancy lemma yields the distinct inequality

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_x t^{-\alpha}\left|\mu^{P_t^{\mathrm{low},x},\pi_t}(x) - \mu^{\pi_t}(x)\right|\right] \\ &\le \sqrt{\mathsf{X}_T X_g} + \mathsf{Y}_T \le \sqrt{H\mathsf{X}_T(1+\log(eT))} + \mathsf{Y}_T\,. \end{aligned} \tag{422}$$

Thus, substituting both (421) and (422) into (420) yields

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_x t^{-\alpha}(\overline{\mu}_t(x) - \underline{\mu}_t(x))\right] \le 2\sqrt{H\mathsf{X}_T(1+\log(eT))} + 2\mathsf{Y}_T\,. \tag{423}$$

Both selectors contribute to this bound.

For the floor term, split the sum at Γ. Since $\gamma_t = 1$ for $t \le \Gamma$ and $\gamma_t = \Gamma/t$ for $t > \Gamma$, the integral comparison gives

$$\begin{aligned} \sum_{t=1}^{T} t^{-\alpha}\gamma_t &\le \sum_{t=1}^{\lceil\Gamma\rceil} t^{-\alpha} + \Gamma\sum_{t>\Gamma} t^{-1-\alpha} \le \lceil\Gamma\rceil + \Gamma\int_{\Gamma}^{\infty} x^{-1-\alpha}\,dx \\ &= \lceil\Gamma\rceil + \frac{\Gamma^{1-\alpha}}{\alpha} \le c_\varepsilon(\Gamma+1)\,. \end{aligned} \tag{424}$$

The last inequality in (424) uses $\Gamma \ge 1$ and α depending only on ε. Therefore

$$\sum_{t=1}^{T}\sum_x t^{-\alpha}\gamma_t = S\sum_{t=1}^{T} t^{-\alpha}\gamma_t \le c_\varepsilon S(\Gamma+1)\,. \tag{425}$$

Multiplying (419) by $\mathbb{1}\{\mathcal{E}\}$, summing over t, taking expectations, and using (423) and (425) gives

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_s \mu^{P_t^B,\pi_t}(s)\nu_t(s)\right] \\ &\le c_\varepsilon H^3 L_\sigma\left[2\sqrt{H\mathsf{X}_T(1+\log(eT))} + 2\mathsf{Y}_T + c_\varepsilon S(\Gamma+1)\right] \le \mathsf{R}_\nu\,, \end{aligned}$$

where the last line is the definition (207) after choosing $c_{\nu,\varepsilon}$ sufficiently large.

Finally, split the full expectation into $\mathcal{E}$ and $\mathcal{E}^c$. The bad-event ν_t term is the quantity bounded in (257); hence

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right] \\
&= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t,s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right] + \mathbb{E}\left[\mathbb{1}\{\mathcal{E}^c\}\sum_{t,s}\mu^{P_t^B,\pi_t}(s)\nu_t(s)\right] \le \mathsf{R}_\nu + T\delta\,\mathsf{F}_\nu\,. \qquad (426)
\end{aligned}$$

Equation (426) is exactly (413). □

D AUXILIARY LEMMAS

D.1 VALUE-DIFFERENCE AND PERFORMANCE-DIFFERENCE IDENTITIES

Lemma D.1 (Value-difference identity)**.** *Fix a layered kernel P, a policy π, and two deterministic one-step loss functions $g, g' : (\mathcal{S} \setminus \mathcal{S}_H) \times \mathcal{A} \to \mathbb{R}$. Let $V_g^{P,\pi}$ and $V_{g'}^{P,\pi}$ be their finite-horizon value functions. Then*

$$V_g^{P,\pi}(s_0) - V_{g'}^{P,\pi}(s_0) = \sum_{s,a} \mu^{P,\pi}(s,a)\big(g(s,a) - g'(s,a)\big)\,. \tag{427}$$

Proof. Let $(s_0, a_0, \ldots, s_H)$ be a trajectory generated by (P, π). By the definition of a finite-horizon value function,

$$V_g^{P,\pi}(s_0) = \mathbb{E}_{P,\pi}\left[\sum_{h=0}^{H-1} g(s_h, a_h)\right], \qquad V_{g'}^{P,\pi}(s_0) = \mathbb{E}_{P,\pi}\left[\sum_{h=0}^{H-1} g'(s_h, a_h)\right].$$

Subtracting, exchanging the finite sum and expectation, and grouping each state-action pair by its occupancy probability gives (427). □

D.1.1 PERFORMANCE DIFFERENCE

Lemma D.2 (Performance-difference identity)**.** *For any deterministic policy π,*

$$V_t^{\pi_t}(s_0) - V_t^{\pi}(s_0) = \sum_s \mu^{\pi}(s)\left\langle p_{t,s} - e_{\pi(s)}, Q_t^{\pi_t}(s,\cdot)\right\rangle \tag{428}$$

$$= \sum_s \mu^{\pi}(s)\left\langle p_{t,s} - e_{\pi(s)}, \bar{Q}_t(s,\cdot)\right\rangle. \tag{429}$$

Proof. Fix episode t and let $(s_0, a_0, s_1, a_1, \ldots, s_H)$ be a trajectory generated by the comparator policy π under the true transition kernel P. For each nonterminal layer h, the conditional expectation of the policy-difference term at the state s_h is

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left\langle p_{t,s_h} - e_{\pi(s_h)}, Q_t^{\pi_t}(s_h,\cdot)\right\rangle\right] \\
&= \mathbb{E}_\pi\left[\sum_{a\in\mathcal{A}} p_{t,s_h,a} Q_t^{\pi_t}(s_h,a) - Q_t^{\pi_t}(s_h,a_h)\right] \\
&= \mathbb{E}_\pi[V_t^{\pi_t}(s_h) - \ell_t(s_h,a_h) - \mathbb{E}[V_t^{\pi_t}(s_{h+1}) \mid s_h, a_h]] \\
&= \mathbb{E}_\pi[V_t^{\pi_t}(s_h) - \ell_t(s_h,a_h) - V_t^{\pi_t}(s_{h+1})]\,.
\end{aligned} \tag{430}$$

where the second equality uses the Bellman equations for $V_t^{\pi_t}$ and $Q_t^{\pi_t}$ in Section 2, and the third uses the tower property.

Summing (430) over $h = 0, \ldots, H-1$ gives the finite-horizon telescoping calculation

$$\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_\pi\left[\left\langle p_{t,s_h} - e_{\pi(s_h)}, Q_t^{\pi_t}(s_h,\cdot)\right\rangle\right] \\
&= \mathbb{E}_\pi\left[\sum_{h=0}^{H-1} (V_t^{\pi_t}(s_h) - V_t^{\pi_t}(s_{h+1})) - \sum_{h=0}^{H-1} \ell_t(s_h,a_h)\right] \\
&= V_t^{\pi_t}(s_0) - \mathbb{E}_\pi\left[\sum_{h=0}^{H-1} \ell_t(s_h,a_h)\right] - V_t^{\pi_t}(s_H) \\
&= V_t^{\pi_t}(s_0) - V_t^{\pi}(s_0)\,.
\end{aligned}$$

where the last equality uses $V_t^{\pi_t}(s_H) = 0$. Expressing the left-hand side through comparator occupancies gives (428).

For the centered identity, use $\bar{Q}_t(s,\cdot) = Q_t^{\pi_t}(s,\cdot) - V_t^{\pi_t}(s)\mathbf{1}$ and $\langle p_{t,s} - e_{\pi(s)}, \mathbf{1}\rangle = \sum_a p_{t,s,a} - 1 = 0$. Therefore

$$\langle p_{t,s} - e_{\pi(s)}, \bar{Q}_t(s,\cdot)\rangle = \langle p_{t,s} - e_{\pi(s)}, Q_t^{\pi_t}(s,\cdot)\rangle \,,$$

and substituting this equality into (428) proves (429). □

D.2 Dilated-bonus Reduction

This section presents the reduction from the global regret to local regret guarantees at each state, by leveraging the dilated bonus function proposed by Luo et al. (2021); Dann et al. (2023b). These reductions convert statewise policy-regret inequalities into bounds on the cumulative scalar bonus. We first treat a fixed known kernel, then give the one-episode version for the optimistic kernel used with unknown transitions. All occupancies without an explicit kernel are evaluated under the true transition kernel P.

Known Transitions

Lemma D.3 (Dilated local-to-global reduction)**.** *Fix a layered transition kernel P. Let $b_t(s) \geq 0$ and let B_t satisfy (12) with $\mathcal{P}_t = \{P\}$. Suppose that for every state s and deterministic action $a^\star$,*

$$\mathbb{E}\left[\sum_t \langle p_{t,s} - e_{a^\star}, Q_t^{\pi_t}(s,\cdot) - B_t(s,\cdot)\rangle\right] \leq X_{a^\star}(s) + \mathbb{E}\left[\sum_t b_t(s) + \frac{1}{H}\langle p_{t,s}, B_t(s,\cdot)\rangle\right]. \tag{431}$$

Then every deterministic policy π satisfies

$$\mathrm{Reg}_T(\pi) \leq \sum_s \mu^\pi(s) X_{\pi(s)}(s) + 3\mathbb{E}\left[\sum_{t=1}^T \sum_s \mu^{\pi_t}(s) b_t(s)\right]. \tag{432}$$

Proof. Fix a deterministic comparator policy π and recall from Section 4 that, for every episode t and state s, $\overline{B}_t(s) = \langle p_{t,s}, B_t(s,\cdot)\rangle$, $B_t^\pi(s) = \langle \pi(\cdot \mid s), B_t(s,\cdot)\rangle$. The performance-difference identity for the mean loss ℓ_t gives

$$\mathrm{Reg}_T(\pi) = \mathbb{E}\left[\sum_{t=1}^T \sum_s \mu^\pi(s)\,\langle p_{t,s} - \pi(\cdot \mid s), Q_t^{\pi_t}(s,\cdot)\rangle\right]. \tag{433}$$

Add and subtract $B_t(s,\cdot)$ inside every inner product in (433). Since π is deterministic, the hypothesis (431) can be applied at state s with $a^\star = \pi(s)$. Multiplying that statewise inequality by $\mu^\pi(s)$ and summing over states gives

$$\begin{aligned}\mathrm{Reg}_T(\pi) \leq \sum_s \mu^\pi(s) X_{\pi(s)}(s) &+ \mathbb{E}\left[\sum_{t=1}^T \sum_s \mu^\pi(s)\left(b_t(s) + \frac{1}{H}\overline{B}_t(s)\right)\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^T \sum_s \mu^\pi(s)\left(\overline{B}_t(s) - B_t^\pi(s)\right)\right]. \end{aligned} \tag{434}$$

The first bonus sum is the occupancy-weighted bound from (431); the second accounts for the B_t term added and subtracted in (433).

For a fixed episode t, collect the two bonus lines of (434) into

$$\begin{aligned}\mathcal{R}_t &:= \sum_s \mu^\pi(s) b_t(s) + \frac{1}{H}\sum_s \mu^\pi(s)\overline{B}_t(s) + \sum_s \mu^\pi(s)\left(\overline{B}_t(s) - B_t^\pi(s)\right) \\ &= \sum_s \mu^\pi(s) b_t(s) + \left(1 + \frac{1}{H}\right)\sum_s \mu^\pi(s)\overline{B}_t(s) - \sum_s \mu^\pi(s) B_t^\pi(s)\,. \end{aligned} \tag{435}$$

We now evaluate the comparator-policy bonus sum $\sum_s \mu^\pi(s) B_t^\pi(s)$ in (435) exactly. Averaging the recursion (12) over $a \sim \pi(\cdot \mid s)$ gives

$$B_t^\pi(s) = b_t(s) + \left(1 + \frac{1}{H}\right)\mathbb{E}_{\substack{a\sim\pi(\cdot|s)\\ s'\sim P(\cdot|s,a)}}\left[\overline{B}_t(s')\right]. \tag{436}$$

Multiplying (436) by $\mu^\pi(s)$, summing over nonterminal states, and using layerwise occupancy flow yields

$$\sum_s \mu^\pi(s) B_t^\pi(s) = \sum_s \mu^\pi(s) b_t(s) + \left(1 + \frac{1}{H}\right) \sum_{h=0}^{H-2} \sum_{s' \in \mathcal{S}_{h+1}} \mu^\pi(s') \overline{B}_t(s') \tag{437}$$

$$= \sum_s \mu^\pi(s) b_t(s) + \left(1 + \frac{1}{H}\right) \sum_{s \neq s_0} \mu^\pi(s) \overline{B}_t(s)\,. \tag{438}$$

The terminal state contributes zero by the terminal convention, and the initial state has no predecessor. These observations explain the range of the transported sum in (437) and the exclusion of s_0 in (438).

Substituting the complete expression (438) into (435) gives

$$\begin{aligned}\mathcal{R}_t &= \sum_s \mu^\pi(s) b_t(s) + \left(1 + \frac{1}{H}\right) \sum_s \mu^\pi(s) \overline{B}_t(s) \\ &\quad - \left[\sum_s \mu^\pi(s) b_t(s) + \left(1 + \frac{1}{H}\right) \sum_{s \neq s_0} \mu^\pi(s) \overline{B}_t(s)\right] \\ &= \left(1 + \frac{1}{H}\right) \mu^\pi(s_0) \overline{B}_t(s_0) = \left(1 + \frac{1}{H}\right) \overline{B}_t(s_0)\,,\end{aligned} \tag{439}$$

where the final line uses $\mu^\pi(s_0) = 1$.

Averaging (12) over the executed action $a \sim p_{t,s}$ gives

$$\overline{B}_t(s) = b_t(s) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{\substack{a \sim p_{t,s} \\ s' \sim P(\cdot|s,a)}} \left[\overline{B}_t(s')\right]\,. \tag{440}$$

Backward unrolling of (440) from the initial state gives

$$\overline{B}_t(s_0) = \mathbb{E}_{\pi_t, P}\left[\sum_{h=0}^{H-1} \left(1 + \frac{1}{H}\right)^h b_t(s_h)\right]\,. \tag{441}$$

Substituting (441) into (439) and using nonnegativity of b_t gives

$$\begin{aligned}\mathcal{R}_t &= \mathbb{E}_{\pi_t, P}\left[\sum_{h=0}^{H-1} \left(1 + \frac{1}{H}\right)^{h+1} b_t(s_h)\right] \leq \left(1 + \frac{1}{H}\right)^H \mathbb{E}_{\pi_t, P}\left[\sum_{h=0}^{H-1} b_t(s_h)\right] \\ &< e \sum_s \mu^{\pi_t}(s) b_t(s) < 3 \sum_s \mu^{\pi_t}(s) b_t(s)\,.\end{aligned} \tag{442}$$

The first inequality uses $h + 1 \leq H$, and the remaining bounds use $(1 + 1/H)^H < e < 3$.

Finally, substitute (442) for every $\mathcal{R}_t$ in (434) and sum over t:

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\leq \sum_s \mu^\pi(s) X_{\pi(s)}(s) + \mathbb{E}\left[\sum_{t=1}^T \mathcal{R}_t\right] \\ &\leq \sum_s \mu^\pi(s) X_{\pi(s)}(s) + 3\mathbb{E}\left[\sum_{t=1}^T \sum_s \mu^{\pi_t}(s) b_t(s)\right]\,.\end{aligned}$$

The final displayed inequality is (432). □

Optimistic Transitions

Lemma D.4 (One-episode robust dilated-bonus inequality)**.** *Let $b_t(s) \geq 0$ and let B_t satisfy (12) over the rectangular confidence set $\mathcal{P}_t$, with maximizing kernel P_t^B. On $\mathcal{E}_t = \{P \in \mathcal{P}_t\}$, every deterministic policy π satisfies*

$$\sum_s \mu^\pi(s) \left[b_t(s) + \left(1 + \frac{1}{H}\right) \langle p_{t,s}, B_t(s, \cdot)\rangle - B_t(s, \pi(s))\right]$$

$$\leq 3 \sum_s \mu^{P_t^B, \pi_t}(s) b_t(s) \,. \tag{443}$$

Proof. Fix an episode t for which $\mathcal{E}_t$ occurs, and recall from Section 4 that the statewise policy-averaged robust bonus is $\overline{B}_t(s) = \langle p_{t,s}, B_t(s, \cdot) \rangle$. Because $\mathcal{E}_t = \{P \in \mathcal{P}_t\}$, the true transition row $P(\cdot \mid s, \pi(s))$ is feasible in the maximization in (12). Therefore, for every nonterminal state s,

$$\begin{aligned} B_t(s, \pi(s)) &= b_t(s) + \left(1 + \frac{1}{H}\right) \max_{\widetilde{P} \in \mathcal{P}_t} \mathbb{E}_{\substack{s' \sim \widetilde{P}(\cdot \mid s, \pi(s)) \\ a' \sim p_{t,s'}}} \left[B_t(s', a')\right] \\ &\geq b_t(s) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{\substack{s' \sim P(\cdot \mid s, \pi(s)) \\ a' \sim p_{t,s'}}} \left[B_t(s', a')\right] \\ &= b_t(s) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{s' \sim P(\cdot \mid s, \pi(s))} \left[\overline{B}_t(s')\right] \,. \end{aligned} \tag{444}$$

The first line of (444) is the definition (12); the second line evaluates the maximum at the feasible true row $P(\cdot \mid s, \pi(s))$; and the third line uses the definition of $\overline{B}_t(s')$.

The comparator occupancy-flow and telescoping calculation is the same as in the proof of Lemma D.3 in the present subsection, specifically (437)–(439). Starting from the row inequality (444) gives the corresponding upper bound instead of the known-transition equality. We omit the repeated layer sums and record the resulting bound:

$$\sum_s \mu^{\pi}(s) \left[b_t(s) + \left(1 + \frac{1}{H}\right) \overline{B}_t(s) - B_t(s, \pi(s)) \right] \leq \left(1 + \frac{1}{H}\right) \overline{B}_t(s_0) \,. \tag{445}$$

The remaining quantity on the right-hand side of (445) is $(1 + H^{-1})\overline{B}_t(s_0)$, which is bounded next. Let $d \coloneqq 1 + H^{-1}$. By the definition of the rectangular maximizing selector P_t^B, the recursion (12) implies, for every state s,

$$\begin{aligned} \overline{B}_t(s) &= \sum_a p_{t,s,a} b_t(s) + d \sum_a p_{t,s,a} \mathbb{E}_{s' \sim P_t^B(\cdot \mid s,a)} \left[\overline{B}_t(s')\right] \\ &= b_t(s) + d \mathbb{E}_{\substack{a \sim p_{t,s} \\ s' \sim P_t^B(\cdot \mid s,a)}} \left[\overline{B}_t(s')\right] \,. \end{aligned} \tag{446}$$

For the selected kernel P_t^B, the Bellman unrolling and the estimate $d^{h+1} \leq d^H \leq e < 3$ are identical to (441)–(442) in the proof of Lemma D.3 in the present subsection, with P replaced by P_t^B. We omit that repeated calculation. Applied to (446), it gives

$$d\overline{B}_t(s_0) \leq 3 \sum_s \mu^{P_t^B, \pi_t}(s) b_t(s) \,.$$

Together with (445), this proves (443). □

D.3 Tsallis Stability and FTRL on the Full Simplex

D.3.1 A Self-contained Tsallis Stability Inequality

For the scalar function $\phi_\alpha(x) \coloneqq -\frac{x^\alpha - x}{\alpha}$, $x \geq 0$, write $d_\alpha(y, x) \coloneqq \phi_\alpha(y) - \phi_\alpha(x) - \phi'_\alpha(x)(y - x)$. Then $\phi''_\alpha(x) = (1 - \alpha)x^{\alpha - 2}$, $x > 0$.

Lemma D.5 (Scalar Tsallis conjugate bound)**.** *Let $x > 0$ and $u \in \mathbb{R}$ satisfy*

$$|u| \leq \frac{1 - \alpha}{8} x^{\alpha - 1} \,. \tag{447}$$

Then

$$\sup_{y \geq 0} \left\{ u(x - y) - d_\alpha(y, x) \right\} \leq \frac{2}{1 - \alpha} x^{2 - \alpha} u^2 \,. \tag{448}$$

Proof. For fixed $x > 0$ and $u \in \mathbb{R}$, define $F_{x,u}(y) := u(x-y) - d_\alpha(y,x)$, $y \geq 0$. Since $d_\alpha(\cdot, x)$ is strictly convex on $(0,\infty)$, $F_{x,u}$ is strictly concave there. Its derivative is

$$F'_{x,u}(y) = -u - \phi'_\alpha(y) + \phi'_\alpha(x) = -u + y^{\alpha-1} - x^{\alpha-1}\,, \tag{449}$$

where (449) uses $\phi'_\alpha(z) = \alpha^{-1} - z^{\alpha-1}$. Thus an interior maximizer must satisfy

$$y^{\alpha-1} = x^{\alpha-1} + u\,. \tag{450}$$

Let $v := u/x^{\alpha-1}$. Condition (447) implies

$$\begin{aligned} |v| &\leq \frac{1-\alpha}{8} < 1\,, \\ x^{\alpha-1} + u &= x^{\alpha-1}(1+v) > 0\,. \end{aligned} \tag{451}$$

Hence the stationary point in (450) is positive and equals $y_\star = x(1+v)^{-1/(1-\alpha)}$. Because $F_{x,u}$ is strictly concave and $F'_{x,u}(y) \to +\infty$ as $y \downarrow 0$ while $F'_{x,u}(y) \to -u - x^{\alpha-1} < 0$ as $y \to \infty$, the point $y_\star$ is the unique global maximizer over $y \geq 0$.

The maximizer remains in a constant-factor neighborhood of x. Indeed,

$$\begin{aligned} |\log(1+v)| &\leq \frac{|v|}{1-|v|} \leq \frac{(1-\alpha)/8}{1-1/8} = \frac{1-\alpha}{7}\,, \\ \left|\log \frac{y_\star}{x}\right| &= \frac{|\log(1+v)|}{1-\alpha} \leq \frac{1}{7}\,. \end{aligned}$$

where the first inequality is the logarithm bound and the second uses (451). Since $e^{-1/7} > 3/4$ and $e^{1/7} < 4/3$, the interval joining x and $y_\star$ lies in $[3x/4, 4x/3]$.

For $z \in [3x/4, 4x/3]$, the exponent $\alpha - 2 < 0$ and therefore

$$\phi''_\alpha(z) = (1-\alpha)z^{\alpha-2} \geq (1-\alpha)(4x/3)^{\alpha-2} \geq \frac{9}{16}(1-\alpha)x^{\alpha-2}\,. \tag{452}$$

Inequality (452) uses $(4/3)^{\alpha-2} \geq (4/3)^{-2} = 9/16$. The integral representation of the scalar Bregman divergence now yields

$$d_\alpha(y_\star, x) = (y_\star - x)^2 \int_0^1 (1-r)\phi''_\alpha(x + r(y_\star - x))\, dr \geq \frac{9}{32}(1-\alpha)x^{\alpha-2}(y_\star - x)^2\,. \tag{453}$$

where the inequality follows from (452).

Set $m := \frac{9}{16}(1-\alpha)x^{\alpha-2}$. Substituting (453) into $F_{x,u}(y_\star)$ and applying $ab - \frac{m}{2}b^2 \leq a^2/(2m)$ gives

$$\begin{aligned} F_{x,u}(y_\star) &= u(x - y_\star) - d_\alpha(y_\star, x) \\ &\leq |u|\,|x - y_\star| - \frac{m}{2}|x - y_\star|^2 \leq \frac{u^2}{2m} = \frac{8}{9(1-\alpha)}x^{2-\alpha}u^2 \leq \frac{2}{1-\alpha}x^{2-\alpha}u^2\,. \end{aligned} \tag{454}$$

Since $y_\star$ is the global maximizer, line (454) proves (448). □

Let $p \in \mathrm{int}(\Delta(\mathcal{A}))$. Recall that $\rho(p) := \min\{\|p\|_\infty, 1 - \|p\|_\infty\}$ and $\bar{p}_a := \min\{p_a, \rho(p)\}$.

Lemma D.6 (Vector Tsallis stability)**.** *If $g \in \mathbb{R}^A$ satisfies*

$$|g_a| \leq \frac{1-\alpha}{16}\,\beta \bar{p}_a^{\alpha-1} \qquad \textit{for every } a\,, \tag{455}$$

then, for every $q \in \Delta(\mathcal{A})$,

$$\langle p - q, g\rangle - \beta D_{\psi_\alpha}(q,p) \leq \frac{4}{(1-\alpha)\beta}\sum_a \bar{p}_a^{2-\alpha} g_a^2\,. \tag{456}$$

Proof. Because $\psi_\alpha(p) = \sum_a \phi_\alpha(p_a)$ on the simplex, its Bregman divergence is separable:

$$D_{\psi_\alpha}(q,p) = \sum_{a\in\mathcal{A}} d_\alpha(q_a,p_a)\,. \tag{457}$$

We distinguish whether p has a coordinate larger than $1/2$.

Suppose first that $\|p\|_\infty \le 1/2$. Then $\rho(p) = \|p\|_\infty$ and $\bar{p}_a = p_a$ for every a. Using (457) and enlarging the feasible set from the simplex to independent coordinates $y_a \ge 0$ gives

$$\begin{aligned}\langle p-q, g\rangle - \beta D_{\psi_\alpha}(q,p) &= \sum_a \left[g_a(p_a-q_a) - \beta d_\alpha(q_a,p_a)\right] \\ &\le \sum_a \sup_{y\ge 0}\left[g_a(p_a-y) - \beta d_\alpha(y,p_a)\right] \\ &= \beta\sum_a \sup_{y\ge 0}\left[\frac{g_a}{\beta}(p_a-y) - d_\alpha(y,p_a)\right]\,. \end{aligned}\tag{458}$$

For every coordinate, (455) implies

$$\left|\frac{g_a}{\beta}\right| \le \frac{1-\alpha}{16}p_a^{\alpha-1} \le \frac{1-\alpha}{8}p_a^{\alpha-1}\,.$$

Applying Lemma D.5 with $x = p_a$ and $u = g_a/\beta$ to each supremum in (458) yields

$$\begin{aligned}\langle p-q,g\rangle - \beta D_{\psi_\alpha}(q,p) &\le \beta\sum_a \frac{2}{1-\alpha}p_a^{2-\alpha}\frac{g_a^2}{\beta^2} \\ &= \frac{2}{(1-\alpha)\beta}\sum_a \bar{p}_a^{2-\alpha}g_a^2 \le \frac{4}{(1-\alpha)\beta}\sum_a \bar{p}_a^{2-\alpha}g_a^2\,.\end{aligned}$$

This proves (456) in the first case.

Suppose next that $p_j = \|p\|_\infty > 1/2$. Then

$$\rho(p) = 1-p_j = \sum_{a\ne j} p_a\,, \qquad \bar{p}_j = \rho(p)\,, \qquad \bar{p}_a = p_a \quad (a\ne j)\,. \tag{459}$$

Because both p and q have total mass one, $\langle p-q, \mathbf{1}\rangle = 0$. Hence, with $u_a := g_a - g_j$ for $a \ne j$ and $u_j := 0$,

$$\langle p-q,g\rangle = \langle p-q, g-g_j\mathbf{1}\rangle = \sum_{a\ne j}(p_a-q_a)u_a\,. \tag{460}$$

For $a \ne j$, one has $p_a \le \sum_{b\ne j} p_b = \rho(p)$. Since $\alpha - 1 < 0$, the inequality $p_a \le \rho(p)$ implies $\rho(p)^{\alpha-1} \le p_a^{\alpha-1}$. Using (455) for coordinates a and j therefore gives

$$|u_a| \le |g_a| + |g_j| \le \frac{1-\alpha}{16}\beta p_a^{\alpha-1} + \frac{1-\alpha}{16}\beta\rho(p)^{\alpha-1} \le \frac{1-\alpha}{8}\beta p_a^{\alpha-1}\,. \tag{461}$$

Using (457), the nonnegativity of every scalar Bregman divergence, and (460), we obtain

$$\begin{aligned}\langle p-q,g\rangle - \beta D_{\psi_\alpha}(q,p) &= \sum_{a\ne j}\left[u_a(p_a-q_a) - \beta d_\alpha(q_a,p_a)\right] - \beta d_\alpha(q_j,p_j) \\ &\le \sum_{a\ne j}\sup_{y\ge 0}\left[u_a(p_a-y) - \beta d_\alpha(y,p_a)\right] \\ &\le \frac{2}{(1-\alpha)\beta}\sum_{a\ne j} p_a^{2-\alpha}(g_a-g_j)^2 \\ &\le \frac{4}{(1-\alpha)\beta}\left(\sum_{a\ne j} p_a^{2-\alpha}g_a^2 + g_j^2\sum_{a\ne j}p_a^{2-\alpha}\right)\,.\end{aligned}\tag{462}$$

where the penultimate inequality applies Lemma D.5 using (461), and the last uses $(g_a - g_j)^2 \le 2g_a^2 + 2g_j^2$.

Since $2 - \alpha > 1$, nonnegativity gives

$$\sum_{a \neq j} p_a^{2-\alpha} \le \left(\sum_{a \neq j} p_a \right)^{2-\alpha} = \rho(p)^{2-\alpha} = \bar{p}_j^{2-\alpha} . \tag{463}$$

Substituting (463) and (459) into (462) gives exactly (456). □

D.3.2 FTRL Decomposition with Changing Inverse Learning Rate

Lemma D.7 (FTRL decomposition on the full simplex)**.** *Let $g_1, \dots, g_T \in \mathbb{R}^A$, let $0 < \beta_1 \le \cdots \le \beta_T$, and define*

$$p_t \in \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \left\langle p, \sum_{\tau < t} g_\tau \right\rangle + \beta_t \psi_\alpha(p) \right\} .$$

For every t, define the same-inverse-learning-rate look-ahead point

$$\widetilde{p}_{t+1} \in \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \left\langle p, \sum_{\tau \le t} g_\tau \right\rangle + \beta_t \psi_\alpha(p) \right\} . \tag{464}$$

Then, for every $u \in \Delta(\mathcal{A})$,

$$\sum_{t=1}^{T} \langle p_t - u, g_t \rangle \le \beta_1 \big(\psi_\alpha(u) - \psi_\alpha(p_1)\big) + \sum_{t=2}^{T} (\beta_t - \beta_{t-1}) \big(\psi_\alpha(u) - \psi_\alpha(p_t)\big) + \sum_{t=1}^{T} \left(\langle p_t - \widetilde{p}_{t+1}, g_t \rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_t) \right) . \tag{465}$$

In particular, if $u = e_{a^\star}$ is a simplex vertex, then $\psi_\alpha(u) = 0$ and

$$\sum_{t=1}^{T} \langle p_t - e_{a^\star}, g_t \rangle \le \beta_1 h_\alpha(p_1) + \sum_{t=2}^{T} (\beta_t - \beta_{t-1}) h_\alpha(p_t) + \sum_{t=1}^{T} \left(\langle p_t - \widetilde{p}_{t+1}, g_t \rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_t) \right) . \tag{466}$$

The point $\widetilde{p}_{t+1}$ is analysis-only. It is not the next executed policy, whose inverse learning rate is β_{t+1}.

Proof. Define the cumulative losses and the regularized value functions by

$$L_t := \sum_{\tau=1}^{t} g_\tau , \qquad \Phi_t(L) := \min_{p \in \Delta(\mathcal{A})} \big\{ \langle p, L \rangle + \beta_t \psi_\alpha(p) \big\} , \qquad L_0 := 0 .$$

The Tsallis objective is strictly convex. Moreover, its one-sided derivative tends to $-\infty$ when any zero coordinate is increased, so both p_t and $\widetilde{p}_{t+1}$ lie in the relative interior of $\Delta(\mathcal{A})$. By the definitions of these two minimizers,

$$\Phi_t(L_{t-1}) = \langle p_t, L_{t-1} \rangle + \beta_t \psi_\alpha(p_t) , \tag{467}$$

$$\Phi_t(L_t) = \langle \widetilde{p}_{t+1}, L_t \rangle + \beta_t \psi_\alpha(\widetilde{p}_{t+1}) . \tag{468}$$

The variational inequality for the minimizer p_t is

$$\langle L_{t-1} + \beta_t \nabla \psi_\alpha(p_t), q - p_t \rangle \ge 0 \qquad \text{for every } q \in \Delta(\mathcal{A}) . \tag{469}$$

The variational inequality is taken in the affine hull of the simplex.

Taking $q = \widetilde{p}_{t+1}$ in (469), substituting (467)–(468), and using $L_t = L_{t-1} + g_t$ gives

$$\begin{aligned}\Phi_t(L_t) - \Phi_t(L_{t-1}) &= \langle \widetilde{p}_{t+1} - p_t, L_{t-1}\rangle + \langle \widetilde{p}_{t+1}, g_t\rangle + \beta_t\big(\psi_\alpha(\widetilde{p}_{t+1}) - \psi_\alpha(p_t)\big)\\ &= \langle p_t, g_t\rangle - \langle p_t - \widetilde{p}_{t+1}, g_t\rangle + \langle \widetilde{p}_{t+1} - p_t, L_{t-1}\rangle + \beta_t\big(\psi_\alpha(\widetilde{p}_{t+1}) - \psi_\alpha(p_t)\big)\\ &\geq \langle p_t, g_t\rangle - \langle p_t - \widetilde{p}_{t+1}, g_t\rangle + \beta_t\,\langle \nabla\psi_\alpha(p_t), p_t - \widetilde{p}_{t+1}\rangle\\ &\quad + \beta_t\big(\psi_\alpha(\widetilde{p}_{t+1}) - \psi_\alpha(p_t)\big)\\ &= \langle p_t, g_t\rangle - \big(\langle p_t - \widetilde{p}_{t+1}, g_t\rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_t)\big)\ . \end{aligned} \tag{470}$$

where the inequality uses optimality of p_t, and the last equality uses the definition of the Bregman divergence.

Rearranging (470) and subtracting $\langle u, g_t\rangle$ gives the one-step inequality

$$\langle p_t - u, g_t\rangle \leq \Phi_t(L_t) - \Phi_t(L_{t-1}) - \langle u, g_t\rangle + \langle p_t - \widetilde{p}_{t+1}, g_t\rangle - \beta_t D_{\psi_\alpha}(\widetilde{p}_{t+1}, p_t)\,. \tag{471}$$

The value-function increments telescope as

$$\sum_{t=1}^{T}\big(\Phi_t(L_t) - \Phi_t(L_{t-1})\big) = \Phi_T(L_T) - \Phi_1(L_0) + \sum_{t=2}^{T}\big(\Phi_{t-1}(L_{t-1}) - \Phi_t(L_{t-1})\big)\,. \tag{472}$$

For $t \geq 2$, the vector p_t minimizes the objective defining $\Phi_t(L_{t-1})$. Using the same vector p_t as a feasible point in the objective defining $\Phi_{t-1}(L_{t-1})$ gives

$$\begin{aligned}\Phi_{t-1}(L_{t-1}) - \Phi_t(L_{t-1}) &\leq [\langle p_t, L_{t-1}\rangle + \beta_{t-1}\psi_\alpha(p_t)] - [\langle p_t, L_{t-1}\rangle + \beta_t\psi_\alpha(p_t)]\\ &= (\beta_{t-1} - \beta_t)\psi_\alpha(p_t)\,. \end{aligned} \tag{473}$$

At the two endpoints and for the comparator loss, one has

$$\Phi_1(L_0) = \Phi_1(0) = \beta_1\psi_\alpha(p_1)\,, \tag{474}$$

$$\begin{aligned}\Phi_T(L_T) &\leq \langle u, L_T\rangle + \beta_T\psi_\alpha(u)\,,\\ \sum_{t=1}^{T}\langle u, g_t\rangle &= \langle u, L_T\rangle\ . \end{aligned} \tag{475}$$

Substituting (473) and (474)– (475) into (472) gives the full penalty calculation

$$\begin{aligned}&\sum_{t=1}^{T}\left(\Phi_t(L_t) - \Phi_t(L_{t-1}) - \langle u, g_t\rangle\right)\\ &\leq \beta_T\psi_\alpha(u) - \beta_1\psi_\alpha(p_1) + \sum_{t=2}^{T}(\beta_{t-1} - \beta_t)\psi_\alpha(p_t)\\ &= \beta_1\big(\psi_\alpha(u) - \psi_\alpha(p_1)\big) + \sum_{t=2}^{T}(\beta_t - \beta_{t-1})\big(\psi_\alpha(u) - \psi_\alpha(p_t)\big)\,. \end{aligned} \tag{476}$$

Line (476) uses the identity $\beta_T = \beta_1 + \sum_{t=2}^{T}(\beta_t - \beta_{t-1})$.

Summing (471) over t and substituting (476) proves (465). For the vertex $u = e_{a^\star}$, the direct substitution into (9) gives $\psi_\alpha(e_{a^\star}) = 0$. Therefore

$$\psi_\alpha(e_{a^\star}) - \psi_\alpha(p_t) = -\psi_\alpha(p_t) = h_\alpha(p_t)\,, \tag{477}$$

and the same identity holds for $t = 1$. Substituting (477) into (465) proves (466). □

Remark D.8. *The proof above is the standard "penalty plus stability" decomposition. It is included to make the argument self-contained; no stochastic assumption is used.*

D.4 Transition Concentration and Occupancy Comparison

Lemma D.9 (Uniform empirical Bernstein confidence event)**.** *The event $\mathcal{E} = \cap_{t=1}^{T}\{P \in \mathcal{P}_t\}$ satisfies*

$$\mathbb{P}(\mathcal{E}) \geq 1 - \delta\,. \tag{478}$$

Moreover, on $\mathcal{E}$, every $\widetilde{P} \in \mathcal{P}_t$ satisfies

$$|\widetilde{P}(s' \mid s,a) - P(s' \mid s,a)| \le 8\sqrt{\frac{P(s' \mid s,a)\iota}{N_t^+(s,a)}} + \frac{56\iota}{N_t^+(s,a)} \tag{479}$$

for all t, s, a, s'.

Proof. Fix an admissible triple (s,a,s') and abbreviate $p := P(s' \mid s,a)$. Enumerate the visits to (s,a) in chronological order, and let X_j be the indicator that the next state at the j-th visit is s'. The strong Markov property of the controlled process and the fact that P is fixed imply

$$\mathbb{E}[X_j \mid X_1, \dots, X_{j-1}, \text{all information before visit } j] = p\,.$$

Thus $X_j - p$ is a bounded martingale-difference sequence, and the usual Bernstein–Freedman argument gives, for every $n \ge 1$ and $x > 0$,

$$\mathbb{P}\left(|\widehat{p}_n - p| > \sqrt{\frac{2px}{n}} + \frac{2x}{3n}\right) \le 2e^{-x}\,. \tag{480}$$

Here $\widehat{p}_n = n^{-1}\sum_{j=1}^n X_j$.

We convert (480) into the empirical radius used by the algorithm. Put $d := |\widehat{p}_n - p|$ and $a_x := x/n$. On the complementary event in (480),

$$\begin{aligned} d &\le \sqrt{2pa_x} + \frac{2a_x}{3} \\ &\le \sqrt{2\widehat{p}_n a_x} + \sqrt{2da_x} + \frac{2a_x}{3} \\ &\le \sqrt{2\widehat{p}_n a_x} + \frac{d}{4} + 2a_x + \frac{2a_x}{3}\,. \end{aligned} \tag{481}$$

where the second inequality uses $\sqrt{p} \le \sqrt{\widehat{p}_n} + \sqrt{d}$, and the third uses Young's inequality. Rearranging gives

$$d \le \frac{4\sqrt{2}}{3}\sqrt{\widehat{p}_n a_x} + \frac{32}{9}a_x \le 2\sqrt{\frac{\widehat{p}_n x}{n}} + \frac{14x}{3n}\,. \tag{482}$$

The last line uses $4\sqrt{2}/3 \le 2$ and $32/9 \le 14/3$.

Apply (482) with $x = \iota$ for every sample size $n \in \{1, \dots, T\}$ and every admissible triple. The number of such events is at most HS^2AT, and hence the union bound and the definition of ι give

$$\mathbb{P}(\text{at least one empirical constraint fails}) \le 2HS^2AT e^{-\iota} = 2HS^2AT\frac{\delta}{16HS^2A\overline{T}^2} \le \delta\,.$$

where the last inequality uses $T \le \overline{T}$ and $\overline{T} \ge 2$. If $N_t(s,a) = 0$, the confidence constraint holds automatically because its linear radius $14\iota/3$ exceeds one. Thus all constraints in (4) hold simultaneously with probability at least $1-\delta$, proving (478).

We now derive (479). Fix a history on the preceding event, write $n = N_t^+(s,a)$ and $a_\iota = \iota/n$, and let $\widehat{p} = \widehat{P}_t(s' \mid s,a)$. Because both P and $\widetilde{P} \in \mathcal{P}_t$ satisfy the empirical constraint,

$$|\widetilde{P}(s' \mid s,a) - p| \le |\widetilde{P}(s' \mid s,a) - \widehat{p}| + |\widehat{p} - p| \le 4\sqrt{\widehat{p}\,a_\iota} + \frac{28}{3}a_\iota\,. \tag{483}$$

To replace $\widehat{p}$ by p, use the true-variance bound (481):

$$\sqrt{\widehat{p}\,a_\iota} \le \sqrt{p\,a_\iota} + \sqrt{|\widehat{p} - p|a_\iota} \le \sqrt{p\,a_\iota} + \sqrt{\sqrt{2p\,a_\iota}\,a_\iota + \frac{2}{3}a_\iota^2} \le 2\sqrt{p\,a_\iota} + 2a_\iota\,.$$

where the last inequality follows from subadditivity of the square root and Young's inequality. Substituting into (483) yields

$$|\widetilde{P}(s' \mid s,a) - p| \le 8\sqrt{p\,a_\iota} + 8a_\iota + \frac{28}{3}a_\iota \le 8\sqrt{\frac{P(s' \mid s,a)\iota}{N_t^+(s,a)}} + \frac{56\iota}{N_t^+(s,a)}\,,$$

which is (479). □

Lemma D.10 (Expected reciprocal count)**.** *For every state-action pair,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{\mu^{\pi_t}(s,a)}{N_t^+(s,a)}\right] \le 2+\log(eT)\,. \tag{484}$$

Proof. Fix (s,a) and define the episode-t visit indicator $I_t \coloneqq \mathbb{1}\{(s_{t,h(s)},a_{t,h(s)})=(s,a)\}$. The denominator $N_t^+(s,a)$ is $\mathcal{F}_{t-1}$-measurable, and the conditional visit probability is $\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] = \mu^{\pi_t}(s,a)$. Therefore the tower property gives, for each t,

$$\mathbb{E}\left[\frac{\mu^{\pi_t}(s,a)}{N_t^+(s,a)}\right] = \mathbb{E}\left[\frac{\mathbb{E}[I_t \mid \mathcal{F}_{t-1}]}{N_t^+(s,a)}\right] = \mathbb{E}\left[\frac{I_t}{N_t^+(s,a)}\right]\,. \tag{485}$$

Let $M \coloneqq \sum_{t=1}^{T} I_t$ be the total number of visits. On every sample path, the value of $N_t(s,a)$ immediately before the j-th visit is $j-1$. Hence the successive denominators $N_t^+(s,a)$ at visits are $1,1,2,\dots,M-1$, and

$$\begin{aligned}\sum_{t=1}^{T}\frac{I_t}{N_t^+(s,a)} &= \begin{cases} 0, & M=0\,,\\ 1, & M=1\,,\\ 2+\displaystyle\sum_{j=2}^{M-1}\frac{1}{j}, & M\ge 2\,,\end{cases}\\ &\le 2+\sum_{j=2}^{\max\{2,M-1\}}\frac{1}{j} \le 2+\log(eT)\,.\end{aligned}$$

where the last inequality uses $M \le T$ and the harmonic-sum bound. Taking expectations and using (485) proves (484). □

For kernels P, P' and a policy π, let $\mu^{P',\pi}(s \mid w)$ be the probability of reaching s from w when the future kernel is P' and the future policy is π. More precisely, when $w \in \mathcal{S}_j$ and $s \in \mathcal{S}_k$ with $j \le k$, $\mu^{P',\pi}(s \mid w)$ is the conditional probability of the event $s_k = s$ given $s_j = w$; it equals $\mathbb{1}\{w=s\}$ when $j=k$ and is defined as zero when $j>k$. Consequently, every next-state index w associated with a predecessor u in the occupancy-difference identity below ranges over

$$w \in \mathcal{S}_{h(u)+1}\,. \tag{486}$$

Lemma D.11 (Occupancy-difference identity)**.** *For every state s,*

$$\mu^{P',\pi}(s)-\mu^{P,\pi}(s) = \sum_{\substack{u:\,h(u)<h(s)\\ a,w}} \mu^{P,\pi}(u,a)\big(P'(w\mid u,a)-P(w\mid u,a)\big)\mu^{P',\pi}(s\mid w)\,. \tag{487}$$

Proof. Let $k \coloneqq h(s)$. For $j \in \{0,\dots,k\}$, define a nonstationary layered kernel $P^{(j)}$ that uses P for transitions out of layers $0,\dots,j-1$ and uses P' for transitions out of layers $j,\dots,k-1$. Thus, for the purpose of reaching the target state s, $\mu^{P^{(0)},\pi}(s) = \mu^{P',\pi}(s)$, $\mu^{P^{(k)},\pi}(s) = \mu^{P,\pi}(s)$. The telescoping decomposition is therefore

$$\mu^{P',\pi}(s)-\mu^{P,\pi}(s) = \sum_{j=0}^{k-1}\Big(\mu^{P^{(j)},\pi}(s)-\mu^{P^{(j+1)},\pi}(s)\Big)\,. \tag{488}$$

The kernels $P^{(j)}$ and $P^{(j+1)}$ agree at every transition except the transition out of layer j. Their state-action distribution at layer j is generated entirely by the common prefix kernel P, so it equals $\mu^{P,\pi}(u,a)$ for $u \in \mathcal{S}_j$. After the next state w is selected, both hybrids use P' on all remaining transitions. Conditioning on (u,a,w) therefore gives

$$\mu^{P^{(j)},\pi}(s)-\mu^{P^{(j+1)},\pi}(s) = \sum_{u\in\mathcal{S}_j}\sum_{a\in\mathcal{A}}\sum_{w\in\mathcal{S}_{j+1}} \mu^{P,\pi}(u,a)\big(P'(w\mid u,a)-P(w\mid u,a)\big)\mu^{P',\pi}(s\mid w)\,. \tag{489}$$

Substituting (489) into (488) and replacing the layer index j by the equivalent condition $h(u)<h(s)$ yields (487). □

Lemma D.12 (Pointwise occupancy confidence radius)**.** *On $\mathcal{E}$, for every policy π, every state s, every t, and every $\widetilde{P} \in \mathcal{P}_t$,*

$$|\mu^{\widetilde{P},\pi}(s) - \mu^{P,\pi}(s)| \le c_0 \sum_{\substack{u:\, h(u)<h(s) \\ a,\ w\in\mathcal{S}_{h(u)+1}}} \mu^{P,\pi}(u,a)\sqrt{\frac{P(w \mid u,a)\iota}{N_t^+(u,a)}}\mu^{\widetilde{P},\pi}(s \mid w) + c_0 S \sum_{\substack{u:\, h(u)<h(s) \\ a}} \mu^{P,\pi}(u,a)\frac{\iota}{N_t^+(u,a)}\,, \tag{490}$$

where c_0 is universal.

Proof. Fix t, π, s, and $\widetilde{P} \in \mathcal{P}_t$, and work on $\mathcal{E}$. Equation (487), with $P' = \widetilde{P}$, and the range convention (486) give

$$|\mu^{\widetilde{P},\pi}(s) - \mu^{P,\pi}(s)| \le \sum_{\substack{u:\, h(u)<h(s) \\ a,\ w\in\mathcal{S}_{h(u)+1}}} \mu^{P,\pi}(u,a)|\widetilde{P}(w \mid u,a) - P(w \mid u,a)|\mu^{\widetilde{P},\pi}(s \mid w)\,. \tag{491}$$

On $\mathcal{E}$, the true-radius inequality (479) reads

$$|\widetilde{P}(w \mid u,a) - P(w \mid u,a)| \le 8\sqrt{\frac{P(w \mid u,a)\iota}{N_t^+(u,a)}} + \frac{56\iota}{N_t^+(u,a)}\,. \tag{492}$$

Substituting the two terms in (492) directly into (491) gives

$$|\mu^{\widetilde{P},\pi}(s) - \mu^{P,\pi}(s)| \le 8 \sum_{\substack{u:\, h(u)<h(s) \\ a,\ w\in\mathcal{S}_{h(u)+1}}} \mu^{P,\pi}(u,a)\sqrt{\frac{P(w \mid u,a)\iota}{N_t^+(u,a)}}\mu^{\widetilde{P},\pi}(s \mid w) \tag{493}$$

$$+ 56 \sum_{\substack{u:\, h(u)<h(s) \\ a,\ w\in\mathcal{S}_{h(u)+1}}} \mu^{P,\pi}(u,a)\frac{\iota}{N_t^+(u,a)}\mu^{\widetilde{P},\pi}(s \mid w)\,. \tag{494}$$

For every fixed predecessor pair (u,a), the index w in (494) belongs only to the single next layer $\mathcal{S}_{h(u)+1}$, and $0 \le \mu^{\widetilde{P},\pi}(s \mid w) \le 1$. Therefore

$$\begin{aligned} &56 \sum_{\substack{u:\, h(u)<h(s) \\ a,\ w\in\mathcal{S}_{h(u)+1}}} \mu^{P,\pi}(u,a)\frac{\iota}{N_t^+(u,a)}\mu^{\widetilde{P},\pi}(s \mid w) \\ &\le 56 \sum_{\substack{u:\, h(u)<h(s) \\ a}} \mu^{P,\pi}(u,a)\frac{\iota}{N_t^+(u,a)}|\mathcal{S}_{h(u)+1}| \\ &\le 56S \sum_{\substack{u:\, h(u)<h(s) \\ a}} \mu^{P,\pi}(u,a)\frac{\iota}{N_t^+(u,a)}\,. \end{aligned}$$

The factor S comes from the single next-state sum. Combining this bound with (493)–(494) and choosing $c_0 \ge 56$ proves (490). □

Lemma D.13 (Weighted occupancy mismatch)**.** *Let $g_t(s) \in [0, g_{\max}]$ be $\mathcal{F}_{t-1}$-measurable. For every t, s, let $\widetilde{p}_{t,s} \in \mathcal{P}_t$ be an arbitrary measurable selector, possibly depending on s. Define*

$$X_g := \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s)g_t(s)^2\right]\,. \tag{495}$$

Then

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_{s} g_t(s)\left|\mu^{\widetilde{p}_{t,s},\pi_t}(s) - \mu^{\pi_t}(s)\right|\right] \le \sqrt{\mathsf{X}_T X_g} + g_{\max}\mathsf{Y}_T\,. \tag{496}$$

The same conclusion holds when $\widetilde{p}_{t,s}$ is a maximizing or a minimizing occupancy selector.

Proof. If $g_{\max} = 0$, then $g_t(s) = 0$ for every t, s, and (496) is immediate. Assume $g_{\max} > 0$ and work on $\mathcal{E}$.

Step 1: A True-suffix Pointwise Refinement The direct estimate in Lemma D.12 retains the selector-dependent suffix occupancy $\mu^{\widetilde{p}_{t,s},\pi_t}(s \mid w)$. The sharp weighted estimate is obtained by replacing that suffix occupancy with the true suffix occupancy and charging the replacement as a second-order transition term. Fix t and s, and define $\kappa_t(u,a) := \frac{\iota}{N_t^+(u,a)}$, $r_t(u,a,w) := \sqrt{P(w \mid u,a)\kappa_t(u,a)}$. On $\mathcal{E}$, (479) and the elementary bound between two probability coordinates imply

$$|\widetilde{p}_{t,s}(w \mid u,a) - P(w \mid u,a)| \le c_\rho\big(r_t(u,a,w) + \kappa_t(u,a)\big)\,, \qquad |\widetilde{p}_{t,s}(w \mid u,a) - P(w \mid u,a)| \le 1\,, \tag{497}$$

where c_ρ is universal. Add and subtract the true suffix occupancy inside (491). With all sums below restricted to $h(u) < h(s)$ and $w \in \mathcal{S}_{h(u)+1}$, this gives

$$\begin{aligned}
&\left|\mu^{\widetilde{p}_{t,s},\pi_t}(s) - \mu^{\pi_t}(s)\right| \\
&\le \underbrace{\sum_{u,a,w} \mu^{\pi_t}(u,a)|\widetilde{p}_{t,s}(w \mid u,a) - P(w \mid u,a)|\mu^{\pi_t}(s \mid w)}_{\mathcal{T}_{t,s}^{(1)}} \\
&\quad + \underbrace{\sum_{u,a,w} \mu^{\pi_t}(u,a)|\widetilde{p}_{t,s}(w \mid u,a) - P(w \mid u,a)|\left|\mu^{\widetilde{p}_{t,s},\pi_t}(s \mid w) - \mu^{\pi_t}(s \mid w)\right|}_{\mathcal{T}_{t,s}^{(2)}}\,.
\end{aligned} \tag{498}$$

Direct substitution of the first inequality in (497) into $\mathcal{T}_{t,s}^{(1)}$ yields

$$\mathcal{T}_{t,s}^{(1)} \le c_\rho \sum_{u,a,w} \mu^{\pi_t}(u,a) r_t(u,a,w)\mu^{\pi_t}(s \mid w) + c_\rho S \sum_{\substack{u:\, h(u)<h(s) \\ a}} \mu^{\pi_t}(u,a)\kappa_t(u,a)\,. \tag{499}$$

where the factor S again bounds the next-state sum.

For every w with $h(w) < h(s)$, apply Lemma D.11 to the conditional process starting from w. If (x,b,z) ranges over $h(w) \le h(x) < h(s)$ and $z \in \mathcal{S}_{h(x)+1}$, then

$$\begin{aligned}
&\left|\mu^{\widetilde{p}_{t,s},\pi_t}(s \mid w) - \mu^{\pi_t}(s \mid w)\right| \\
&\le \sum_{x,b,z} \mu^{\pi_t}(x,b \mid w)|\widetilde{p}_{t,s}(z \mid x,b) - P(z \mid x,b)|\mu^{\widetilde{p}_{t,s},\pi_t}(s \mid z)\,.
\end{aligned} \tag{500}$$

Define

$$\mathcal{M}_{t,s} := \sum_{\substack{u:\, h(u)<h(s) \\ a}} \mu^{\pi_t}(u,a)\kappa_t(u,a)\,. \tag{501}$$

We now bound the second-order quantity $\mathcal{T}_{t,s}^{(2)}$ in (498). Split the first transition error according to the two summands in (497). For the square-root part, substitute (500); for the linear part, use the probability bound $|\mu^{\widetilde{p}_{t,s},\pi_t}(s \mid w) - \mu^{\pi_t}(s \mid w)| \le 1$. This gives

$$\mathcal{T}_{t,s}^{(2)} \le c_\rho \mathcal{T}_{rr} + c_\rho \mathcal{T}_{r\ell} + c_\rho S \mathcal{M}_{t,s}\,, \tag{502}$$

where

$$\mathcal{T}_{rr} := \sum_{u,a,w,x,b,z} \mu^{\pi_t}(u,a) r_t(u,a,w)\mu^{\pi_t}(x,b \mid w)\min\{c_\rho r_t(x,b,z), 1\}\,, \tag{503}$$

$$\mathcal{T}_{r\ell} := \sum_{u,a,w,x,b,z} \mu^{\pi_t}(u,a) r_t(u,a,w)\mu^{\pi_t}(x,b \mid w)\min\{c_\rho \kappa_t(x,b), 1\}\,. \tag{504}$$

The terminal suffix factor $\mu^{\widetilde{p}_{t,s},\pi_t}(s \mid z)$ has been bounded by one in (503)–(504). Constants c_ρ inside the minima can be absorbed into a universal multiplicative constant below.

For $\mathcal{T}_{rr}$, remove the cap $\min\{c_\rho r_t(x,b,z), 1\} \le c_\rho r_t(x,b,z)$ and factor each summand as

$$\mu^{\pi_t}(u,a)\mu^{\pi_t}(x,b \mid w)\sqrt{P(w \mid u,a)\kappa_t(u,a)}\sqrt{P(z \mid x,b)\kappa_t(x,b)}$$

$$= \sqrt{\mu^{\pi_t}(u,a)P(z \mid x,b)\mu^{\pi_t}(x,b \mid w)\kappa_t(u,a)} \times \sqrt{\mu^{\pi_t}(u,a)P(w \mid u,a)\mu^{\pi_t}(x,b \mid w)\kappa_t(x,b)}\,. \tag{505}$$

Cauchy–Schwarz applied to (505) yields

$$\begin{aligned}\mathcal{T}_{rr} &\le c\left(\sum_{u,a,w,x,b,z}\mu^{\pi_t}(u,a)P(z \mid x,b)\mu^{\pi_t}(x,b \mid w)\kappa_t(u,a)\right)^{\frac{1}{2}} \\ &\quad\times\left(\sum_{u,a,w,x,b,z}\mu^{\pi_t}(u,a)P(w \mid u,a)\mu^{\pi_t}(x,b \mid w)\kappa_t(x,b)\right)^{\frac{1}{2}} \\ &\le c\sqrt{HS\mathcal{M}_{t,s}}\sqrt{HS\mathcal{M}_{t,s}} = cHS\mathcal{M}_{t,s}\,.\end{aligned} \tag{506}$$

where the first squared factor is bounded by summing $P(z \mid x,b)$ to one, then summing the conditional occupancy over at most H future layers and at most S starting states w. For the second factor, occupancy flow through (u,a,w) gives the true occupancy of (x,b), with at most H predecessor layers and S choices of z.

For $\mathcal{T}_{r\ell}$, use $\sqrt{pq} \le p+q$ for $p,q \ge 0$ and $\min\{c_\rho\kappa,1\} \le c\min\{\kappa,1\}$. Summing first over z contributes at most S, and therefore

$$\begin{aligned}\mathcal{T}_{r\ell} &\le cS\sum_{u,a,w,x,b}\mu^{\pi_t}(u,a)\big(P(w \mid u,a)+\kappa_t(u,a)\big)\mu^{\pi_t}(x,b \mid w)\min\{\kappa_t(x,b),1\} \\ &\le cHS\sum_{x,b}\mu^{\pi_t}(x,b)\kappa_t(x,b) + cHS^2\sum_{u,a}\mu^{\pi_t}(u,a)\kappa_t(u,a) \le cHS^2\mathcal{M}_{t,s}\,.\end{aligned} \tag{507}$$

where the second inequality uses occupancy flow over at most H predecessor layers for the $P(w \mid u,a)$ term. For the $\kappa_t(u,a)$ term, the sum over w contributes at most S, and the capped conditional occupancy sums to at most H. The last inequality uses (501).

Substituting (506) and (507) into (502) gives

$$\mathcal{T}^{(2)}_{t,s} \le cHS^2\mathcal{M}_{t,s}\,. \tag{508}$$

Combining (499) and (508) in the complete decomposition (498) proves

$$\begin{aligned}\left|\mu^{\widetilde{p}_{t,s},\pi_t}(s)-\mu^{\pi_t}(s)\right| &\le c_1\sum_{\substack{u:\,h(u)<h(s)\\ a,\ w\in\mathcal{S}_{h(u)+1}}}\mu^{\pi_t}(u,a)\sqrt{P(w \mid u,a)\kappa_t(u,a)}\mu^{\pi_t}(s \mid w) \\ &\quad + c_1HS^2\sum_{\substack{u:\,h(u)<h(s)\\ a}}\mu^{\pi_t}(u,a)\kappa_t(u,a)\,,\end{aligned} \tag{509}$$

where c_1 is universal. The HS^2 remainder pays for replacing the selector-dependent suffix by the true suffix; the direct linear radius in (490) contributes only S.

Step 2: Weighted Square-root Term Multiply (509) by $g_t(s)$, sum over t and s, and separate predecessor layers. This yields

$$\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\sum_{s}g_t(s)\left|\mu^{\widetilde{p}_{t,s},\pi_t}(s)-\mu^{\pi_t}(s)\right| \le c_1\mathbb{1}\{\mathcal{E}\}\sum_{h=0}^{H-1}\mathsf{R}_h + c_1\mathbb{1}\{\mathcal{E}\}\mathsf{L}\,, \tag{510}$$

where

$$\mathsf{R}_h := \sum_{t=1}^{T}\sum_{\substack{u\in\mathcal{S}_h,\ a\in\mathcal{A},\ w\in\mathcal{S}_{h+1}\\ s\in\mathcal{S}}}\mu^{\pi_t}(u,a)\sqrt{P(w \mid u,a)\kappa_t(u,a)}\mu^{\pi_t}(s \mid w)g_t(s)\,, \tag{511}$$

$$\mathsf{L} := HS^2\sum_{t=1}^{T}\sum_{s}g_t(s)\sum_{\substack{u:\,h(u)<h(s)\\ a}}\mu^{\pi_t}(u,a)\kappa_t(u,a)\,.$$

For a fixed h and $\lambda > 0$, factor each summand of (511) as the product of

$$X_{t,u,a,w,s} := \sqrt{\mu^{\pi_t}(u,a)P(w \mid u,a)\mu^{\pi_t}(s \mid w)}\, g_t(s)\,,$$
$$Y_{t,u,a,w,s} := \sqrt{\mu^{\pi_t}(u,a)\kappa_t(u,a)\mu^{\pi_t}(s \mid w)}\,.$$

Applying $xy \le \lambda x^2/2 + y^2/(2\lambda)$ to all tuples gives

$$\mathsf{R}_h \le \frac{\lambda}{2}\mathsf{A}_h + \frac{1}{2\lambda}\mathsf{B}_h\,, \tag{512}$$
$$\mathsf{A}_h := \sum_{t,u,a,w,s} \mu^{\pi_t}(u,a)P(w \mid u,a)\mu^{\pi_t}(s \mid w)g_t(s)^2\,,$$
$$\mathsf{B}_h := \sum_{t,u,a,w,s} \mu^{\pi_t}(u,a)\kappa_t(u,a)\mu^{\pi_t}(s \mid w)\,. \tag{513}$$

All sums in (513) use the index ranges in (511). True occupancy flow gives the layer-restricted identity

$$\mathsf{A}_h = \sum_{t=1}^{T}\sum_{s:\,h(s)>h} \mu^{\pi_t}(s)g_t(s)^2 \le \sum_{t=1}^{T}\sum_{s} \mu^{\pi_t}(s)g_t(s)^2\,,$$
$$\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{A}_h] \le X_g\,.$$

For $w \in \mathcal{S}_{h+1}$, the sum of $\mu^{\pi_t}(s \mid w)$ over all target states s is at most H, because there is total conditional occupancy one at each later layer. Hence

$$\mathsf{B}_h \le H|\mathcal{S}_{h+1}|\iota \sum_{t=1}^{T}\sum_{u\in\mathcal{S}_h,a} \frac{\mu^{\pi_t}(u,a)}{N_t^+(u,a)}\,,$$
$$\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{B}_h] \le H|\mathcal{S}_h||\mathcal{S}_{h+1}|A\mathcal{L}_T\,,$$

where the second line applies Lemma D.10 separately to the $|\mathcal{S}_h|A$ predecessor pairs. Taking expectations in (512) and substituting the two bounds above gives

$$\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{R}_h] \le \frac{\lambda}{2}X_g + \frac{H|\mathcal{S}_h||\mathcal{S}_{h+1}|A\mathcal{L}_T}{2\lambda}\,.$$

Optimizing in λ yields

$$\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{R}_h] \le \sqrt{H|\mathcal{S}_h||\mathcal{S}_{h+1}|A\mathcal{L}_T\, X_g}\,.$$

Summing over predecessor layers and using $\sqrt{xy} \le (x+y)/2$ gives

$$\sum_{h=0}^{H-1}\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{R}_h] \le \sqrt{HA\mathcal{L}_T X_g}\sum_{h=0}^{H-1}\sqrt{|\mathcal{S}_h||\mathcal{S}_{h+1}|}$$
$$\le (S+1)\sqrt{HA\mathcal{L}_T X_g} \le 2\sqrt{HS^2A\mathcal{L}_T X_g}\,. \tag{514}$$

Choosing $c_{\mathrm{occ}} \ge 4c_1^2$ in (197) and multiplying (514) by c_1 give

$$c_1\sum_{h=0}^{H-1}\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{R}_h] \le \sqrt{\mathsf{X}_T X_g}\,. \tag{515}$$

Step 3: Weighted Linear Term Using $0 \le g_t(s) \le g_{\max}$, dropping the restriction $h(u) < h(s)$, and then applying Lemma D.10 to all SA state-action pairs gives

$$\mathsf{L} \le HS^2\sum_{t=1}^{T}\left(\sum_s g_t(s)\right)\sum_{u,a}\mu^{\pi_t}(u,a)\kappa_t(u,a) \le HS^3 g_{\max}\iota\sum_{t=1}^{T}\sum_{u,a}\frac{\mu^{\pi_t}(u,a)}{N_t^+(u,a)}\,,$$
$$\mathbb{E}[\mathbb{1}\{\mathcal{E}\}\mathsf{L}] \le HS^3 g_{\max}\iota\, SA(2+\log(eT)) = HS^4A\mathcal{L}_T g_{\max}\,. \tag{516}$$

Choosing c_{occ} also large enough that $c_1 H S^4 A \mathcal{L}_T \le \mathsf{Y}_T$ and substituting (515) and (516) into (510) yields (496).

The use of the true occupancy in (495) and the scale $\mathsf{X}_T = \mathcal{O}(HS^2A\mathcal{L}_T)$ are the same structural choices as in Lemma C.4 (Lemma 16 in the arXiv numbering) of Dann, Wei, and Zimmert (Dann et al., 2023b). A proof that stops at the direct selector-suffix bound (490) must instead control a hybrid occupancy by $\overline{\mu}_t$ and counts target states once more, producing the looser $S^{3/2}$ dependence. The true-suffix expansion (498) moves that selector dependence into the second-order linear remainder; consequently the leading square-root term uses μ^{π_t} and has only linear dependence on S after taking the square root. □

Proof of Theorem C.2. Fix a state-action pair (s,a). For each episode t, let $P^+_{t,s}$ be the fixed measurable selector satisfying $\mu^{P^+_{t,s},\pi_t}(s) = \overline{\mu}_t(s)$. On $\mathcal{E}$, the true kernel belongs to $\mathcal{P}_t$, and therefore

$$\overline{\mu}_t(s) - \mu^{\pi_t}(s) = \left|\mu^{P^+_{t,s},\pi_t}(s) - \mu^{\pi_t}(s)\right| .$$

Substituting $\widetilde{\mu}_t(s) = \overline{\mu}_t(s) + \gamma_t$ into the definition of $\widetilde{n}_T(s,a)$ gives

$$\begin{aligned}\widetilde{n}_T(s,a) &= \mathbb{E}\left[\sum_{t=1}^{T}\big(\overline{\mu}_t(s)+\gamma_t\big)p_{t,s,a}\right] \\ &= n_T(s,a) + \mathbb{E}\left[\sum_{t=1}^{T} p_{t,s,a}\big(\overline{\mu}_t(s) - \mu^{\pi_t}(s)\big)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t p_{t,s,a}\right] . \end{aligned} \tag{517}$$

Splitting the occupancy difference in (517) according to $\mathcal{E}$ and $\mathcal{E}^c$, using $0 \le p_{t,s,a}[\overline{\mu}_t(s) - \mu^{\pi_t}(s)]_+ \le 1$, $\mathbb{P}(\mathcal{E}^c) \le \delta$, and $\sum_t \gamma_t \le \mathsf{G}_T$, yields

$$\begin{aligned}\widetilde{n}_T(s,a) &\le n_T(s,a) + \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T} p_{t,s,a}\big(\overline{\mu}_t(s) - \mu^{\pi_t}(s)\big)\right] + T\delta + \mathsf{G}_T \\ &\le n_T(s,a) + \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T} p_{t,s,a}\big(\overline{\mu}_t(s) - \mu^{\pi_t}(s)\big)\right] + \mathsf{G}_T + 1 . \end{aligned} \tag{518}$$

The last line uses $T\delta \le 1$.

Apply Lemma D.13 to $P^+_{t,s}$ with the predictable weight $g_t(x) := \mathbb{1}\{x = s\}p_{t,s,a}$. For the true-occupancy definition (495), the quadratic quantity is bounded directly by the actual count:

$$X_g = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T}\mu^{\pi_t}(s){p_{t,s,a}}^2\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)p_{t,s,a}\right] = n_T(s,a) .$$

where the inequality uses $0 \le p_{t,s,a} \le 1$. Thus Lemma D.13 gives

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{t=1}^{T} p_{t,s,a}\big(\overline{\mu}_t(s) - \mu^{\pi_t}(s)\big)\right] \le \sqrt{\mathsf{X}_T n_T(s,a)} + \mathsf{Y}_T . \tag{519}$$

Substituting (519) into (518) gives

$$\begin{aligned}\widetilde{n}_T(s,a) &\le n_T(s,a) + \sqrt{\mathsf{X}_T n_T(s,a)} + \mathsf{Y}_T + \mathsf{G}_T + 1 \\ &\le n_T(s,a) + \sqrt{\mathsf{X}_T n_T(s,a)} + \mathsf{Z}_T , \end{aligned} \tag{520}$$

where the last inequality uses $\mathsf{Z}_T = 2(\mathsf{Y}_T + \mathsf{G}_T + 1)$, proving (220).

Finally, $0 < \alpha < 1$, so $x \mapsto x^\alpha$ is subadditive on $\mathbb{R}_+$. Applying subadditivity to (520) gives

$$\begin{aligned}\widetilde{n}_T(s,a)^\alpha &\le n_T(s,a)^\alpha + \left(\sqrt{\mathsf{X}_T n_T(s,a)}\right)^\alpha + \mathsf{Z}_T^\alpha \\ &= n_T(s,a)^\alpha + \mathsf{X}_T^{\frac{\alpha}{2}} n_T(s,a)^{\frac{\alpha}{2}} + \mathsf{Z}_T^\alpha , \end{aligned}$$

which is (221). □