Policy Optimization for Adversarial Linear Mixture MDPs with Unknown Transitions and Bandit Feedback

Yutian Cheng[*1] **Canzhe Zhao**[*1] **Shuai Li**[†1,2]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Shanghai Innovation Institute, Shanghai, China

Abstract

We study episodic reinforcement learning with adversarial losses and bandit feedback in MDPs with unknown transitions under a linear-mixture model. In this setting, prior studies are typically built upon global optimization methods with the notion of occupancy measure, rather than the purely local policy optimization-based method, which avoids solving a constrained convex optimization problem and is thus more preferred in practice. In this work, we develop the first policy optimization-based algorithm that runs online mirror descent locally on each state. The main challenge is to interface local updates with transition learning: the dilated-bonus analysis requires predictable reachability surrogates and robust one-step maximizations over a confidence set, but the statistically natural VLS confidence region is a coupled ellipsoid over a shared parameter and is not (s, a)-rectangular. We address this by rectangularizing the VLS ellipsoid into local per-(s, a) kernel sets, which enables robust dynamic programming to compute occupancy envelopes and to drive the implicit-exploration estimator and dilated bonuses. Based on this technique, we prove a high-probability regret bound $\widetilde{\mathcal{O}}\big(H^{3/2}\sqrt{SAK} + d\,H^{3/2}S^{3/2}\sqrt{K}\big)$, which nearly matches the best-known bound achieved by occupancy-measure-based methods.

1 INTRODUCTION

Reinforcement learning (RL) studies how an agent can learn to make sequential decisions under uncertainty from interaction data, and has become a central tool for modeling and optimizing interactive systems [Sutton and Barto, 2018]. The core difficulty is that the loss and transition dynamics are unknown and depend on the agent's actions, so the agent must trade off gathering informative data for exploration and minimizing cumulative loss for exploitation, and naive myopic learning can be brittle. Markov decision processes (MDPs) provide the standard mathematical framework for reasoning about this sequential structure, separating the unknown environment dynamics from the agent's policy and enabling regret-based performance guarantees [Puterman, 1994].

In the tabular setting, a long line of work develops algorithms with near-optimal regret guarantees by maintaining confidence sets over the transition model and planning optimistically, such as UCRL2 and UCBVI [Jaksch et al., 2010, Azar et al., 2017]. However, when the state space is large or continuous, treating each state-action pair independently becomes statistically infeasible. This motivates RL with *function approximation*, where the learner exploits known structure to generalize across states and actions. Among the most widely studied structured models with linear function approximation are *linear MDPs* [Jin et al., 2020b] and *linear-mixture MDPs* [Ayoub et al., 2020, Zhou et al., 2021]. These are distinct modeling assumptions and are generally treated as incomparable in the literature: without additional restrictions, neither model is a special case of the other [Jin et al., 2020b, Zhou et al., 2021]. In this paper, we focus on the *linear-mixture* model, which posits that each transition kernel lies in the span of a known feature map with an unknown low-dimensional parameter. This model strikes a balance between expressiveness and tractability and has led to near-minimax guarantees in the stochastic setting.

Adversarial RL studies sequential decision making under non-stationary, potentially worst-case losses, and has been studied extensively in episodic MDPs [Even-Dar et al., 2009, Yu et al., 2009, Rosenberg and Mansour, 2019]. More recently, this line of work has been extended to linear-mixture models under stronger feedback assumptions (e.g., full-information access to losses), which enables global

[*] These authors contributed equally.
[†] Correspondence to: Shuai Li <shuaili8@sjtu.edu.cn>

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1249–1270.

planning-style updates over a confidence set [He et al., 2022, Dai et al., 2023]. In this paper, we focus on the more challenging *bandit-feedback* regime with *unknown transitions* under the linear-mixture model: the learner observes only the losses along the realized trajectory in each episode. This setting was recently investigated by Zhao et al. [2023] and Li et al. [2024].

Existing approaches for adversarial linear-mixture MDPs with bandit feedback and unknown transitions predominantly rely on *global* planning and optimization over a transition confidence set, such as optimistic dynamic programming or occupancy-measure-based formulations [Zhao et al., 2023, Li et al., 2024]. While powerful, such global updates typically require solving a planning subproblem or a large-scale optimization problem over the entire state space in every episode, which can be computationally burdensome and less preferred in practice. This motivates the following question: *can we obtain sharp regret guarantees for adversarial linear-mixture MDPs using a purely policy-optimization algorithm with local updates?*

Policy-optimization methods for adversarial linear-mixture MDPs have so far been developed primarily under stronger feedback assumptions, such as full-information access to losses [He et al., 2022, Dai et al., 2023]. In contrast, in the bandit-feedback regime with unknown transitions, existing results are based on occupancy-measure or other global planning formulations over a transition confidence set [Zhao et al., 2023, Li et al., 2024].

We answer this question affirmatively by developing a purely policy-optimization algorithm with local updates. The main technical challenge is to interface per-state OMD with transition learning in a way that supports predictable reachability surrogates and robust planning primitives, despite the coupled and non-rectangular structure of the statistically natural VLS confidence region. Building on the dilated-bonus template of Luo et al. [2021], we resolve this by rectangularizing the VLS confidence set [Li et al., 2024] and using robust dynamic programming to compute occupancy envelopes and to support the dilated-bonus recursion. As a result, we prove a high-probability regret bound $\widetilde{\mathcal{O}}\big(H^{3/2}\sqrt{SAK} + d\,H^{3/2}S^{3/2}\sqrt{K}\big)$, matching the best-known $S^{3/2}$ state dependence in the transition-uncertainty term.

2 RELATED WORK

Adversarial MDPs. Prior work on adversarial MDPs can be organized along two axes: whether the transition dynamics are known or must be learned, and whether the feedback is full-information or bandit. With known transitions, early works obtain regret guarantees by coupling online learning updates with dynamic programming [Even-Dar et al., 2009, Yu et al., 2009, Neu et al., 2010, Zimin and Neu, 2013]. When transitions are unknown, the dominant approach is to maintain confidence sets and perform optimistic global planning, often via occupancy-measure formulations; this applies under full-information feedback [Rosenberg and Mansour, 2019] as well as under bandit feedback [Arora et al., 2012, Jin et al., 2020a, Zhao et al., 2023, Li et al., 2024]. This prevalence of global planning is natural because uncertainty about transitions couples exploration across time, and optimism is often implemented by solving a worst-case or optimistic planning problem over the entire state space. In contrast, policy-optimization methods emphasize *local* updates of per-state action distributions and seek exploration mechanisms that avoid solving a global planning problem in every episode. This line includes O-REPS under full-information feedback [Zimin and Neu, 2013] and the dilated-bonus and FTPL-style local-update frameworks under bandit feedback [Luo et al., 2021, Dai et al., 2022].

Function approximation. Function approximation is essential for scaling RL beyond tabular problems [Sutton and Barto, 2018], and in linear models it is useful to distinguish assumptions on the dynamics. In the stochastic setting, both linear MDPs [Jin et al., 2020b] and linear-mixture MDPs [Ayoub et al., 2020, Zhou et al., 2021] admit near-optimal exploration guarantees, but rely on different structural conditions. In the adversarial setting, the strongest results for linear-mixture dynamics under full-information feedback are obtained by planning over a confidence set [He et al., 2022, Dai et al., 2023]. For the bandit-feedback and unknown-transition regime, existing guarantees for adversarial linear-mixture MDPs similarly proceed through global planning/occupancy-measure optimization [Zhao et al., 2023, Li et al., 2024]. From a methodological perspective, these results highlight a recurring tradeoff: occupancy-measure/global planning makes it easier to integrate transition confidence sets with exploration, while local policy updates require additional machinery to propagate uncertainty and future reachability into statewise updates. Our contribution is to bridge this gap for adversarial linear-mixture MDPs with bandit feedback and unknown transitions by providing an explicit *policy-optimization* interface that couples local OMD updates with transition learning through rectangularized confidence sets.

3 PROBLEM SETTING

Layered episodic MDP. We consider an episodic MDP with horizon H, and adopt the standard layered (loop-free) formulation used in regret analyses of episodic adversarial MDPs. See, e.g., [Zimin and Neu, 2013, Jin et al., 2020a, Zhao et al., 2023, Li et al., 2024]. The state space is partitioned into layers $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$ where transitions only go from $\mathcal{S}_h$ to $\mathcal{S}_{h+1}$. Let $S \triangleq \sum_{h=0}^{H} |\mathcal{S}_h|$ denote the total number of states (including the terminal layer) and let $n := \max_h |\mathcal{S}_h|$ be the maximum number of states per layer.

Table 1: Comparison of representative results that match our setting: adversarial losses with bandit feedback and unknown linear-mixture transitions. The table highlights how different algorithmic interfaces (global optimistic planning versus local policy updates) trade off regret scaling and modularity.

Reference	Regret bound	Optimization style
Lower bound [Zhao et al., 2023]	$\Omega\big(\sqrt{HSAK} + d\,H\sqrt{K}\big)$	information-theoretic
LSUOB-REPS [Zhao et al., 2023]	$\widetilde{\mathcal{O}}\big(\sqrt{HSAK} + d\,S^2\sqrt{K}\big)$	global optimization
VLSUOB-REPS [Li et al., 2024]	$\widetilde{\mathcal{O}}\big(\sqrt{HSAK} + d\sqrt{HS^3K}\big)$	global optimization
Ours	$\widetilde{\mathcal{O}}\big(H^{3/2}\sqrt{SAK} + d\,H^{3/2}S^{3/2}\sqrt{K}\big)$	local optimization

The action space $\mathcal{A}$ is finite with $|\mathcal{A}| = A$. The initial state $s_0 \in \mathcal{S}_0$ is fixed and known.

The layered assumption rules out within-episode cycles and supports backward recursions for value functions and occupancy measures.

Adversarial losses and bandit feedback. For each episode $k = 1, 2, \ldots, K$, an adaptive adversary chooses a loss function $\ell_k : \mathcal{S} \times \mathcal{A} \to [0, 1]$ before the learner acts in episode k. The learner selects a (possibly non-stationary) policy $\pi_k = \{\pi_{k,h}(\cdot \mid s)\}_{h=0}^{H-1}$ and rolls out a trajectory $(s_{k,0}, a_{k,0}, \ldots, s_{k,H})$ under the unknown transition kernel $P^\star$. Bandit feedback means the learner only observes $\ell_k(s_{k,h}, a_{k,h})$ along the trajectory. See, e.g., [Neu et al., 2010, Zimin and Neu, 2013, Jin et al., 2020a, Luo et al., 2021].

Value functions, Q-functions, and occupancies. For any policy π and transition kernel P, define the cost-to-go and action-value functions for episode k by

$$V_{k,h}^{\pi,P}(s) := \mathbb{E}_{\pi,P}\Big[\sum_{t=h}^{H-1} \ell_k(s_t, a_t) \;\Big|\; s_h = s\Big],$$

$$Q_{k,h}^{\pi,P}(s,a) := \mathbb{E}_{\pi,P}\Big[\sum_{t=h}^{H-1} \ell_k(s_t, a_t) \;\Big|\; s_h = s, a_h = a\Big].$$

Let $d_h^{\pi,P}(s)$ be the visitation probability of $s \in \mathcal{S}_h$, and define the state-action occupancy $q_h^{\pi,P}(s,a) \triangleq d_h^{\pi,P}(s)\,\pi_h(a \mid s)$. These quantities are the basic building blocks for both policy updates (they determine the weights in regret decompositions) and bandit estimators.

Linear mixture transitions. For each layer $h \in \{0, \ldots, H-1\}$, we are given a known feature map $\varphi_h(\cdot \mid s, a) : \mathcal{S}_{h+1} \to \mathbb{R}^d$. The unknown transition kernel satisfies the *linear mixture model*: there exists an unknown parameter vector $\theta_h^\star \in \mathbb{R}^d$ such that for all $(s, a, s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$,

$$P_h^\star(s' \mid s, a) = \langle \varphi_h(s' \mid s, a),\, \theta_h^\star \rangle\,. \tag{1}$$

We assume $\|\theta_h^\star\|_2 \le B$ and $\|\varphi_h(s' \mid s, a)\|_2 \le 1$, and that (1) defines a valid probability distribution.

This model is a standard way to formalize low-dimensional structure in transition dynamics and can be viewed as a function-approximation analogue of the tabular model. It reduces to the tabular case by choosing indicator features, and it enables statistically efficient exploration when $d \ll S$ [Ayoub et al., 2020, Zhou et al., 2021].

Value and regret. The episode value is $V_k^{\pi,P}(s_0) = V_{k,0}^{\pi,P}(s_0)$. Regret is measured against the best fixed comparator policy $\pi^\star$:

$$\mathrm{Reg}(K) \triangleq \sum_{k=1}^{K} \big(V_k^{\pi_k,P^\star}(s_0) - V_k^{\pi^\star,P^\star}(s_0)\big).$$

4 ALGORITHM

We present a policy-optimization algorithm that performs online mirror descent (OMD) updates on the per-state action distributions and achieves exploration under bandit feedback by augmenting the loss estimator with a *dilated* bonus [Luo et al., 2021]. To handle *linear-mixture* unknown transitions, the algorithm maintains an (s,a)-rectangular confidence set that supports efficient robust dynamic programming, which yields a modular interface between transition learning and the local policy-update loop.

Algorithm 1 summarizes the full procedure; we introduce its main components below when they are needed.

4.1 RECTANGULAR CONFIDENCE SETS

We now describe our main novelty: transition-learning primitives that are specific to the linear-mixture setting. We first construct a VLS estimator and confidence ellipsoid for the shared linear-mixture parameter (Eq. (3) and Eq. (5)). We next *rectangularize* this coupled uncertainty into an (s,a)-rectangular kernel set $\mathcal{P}_k$ (Eq. (9)), so that robust dynamic programming admits locally-assembled maximizers (Lemma 11). Finally, we compute upper/lower occupancy envelopes $(u_{k,h}, \underline{u}_{k,h})$ (Eq. (10)); controlling their cumulative gap under VLS (Lemma 16) yields the improved $S^{3/2}$ dependence in the transition-uncertainty term.

Algorithm 1 Policy optimization for adversarial linear-mixture MDPs with bandit feedback.

Require: step size $\eta > 0$, regularization $\lambda > 0$, confidence δ

1: Initialize $\pi_{1,h}(\cdot \mid s)$ uniformly over $\mathcal{A}$ for all h, s, and set $\mathcal{P}_0$ to contain all valid kernels.
2: **for** $k = 1, \dots, K$ **do**
3: Construct $\mathcal{P}_{k-1}$ from past data (with $\mathcal{P}_0$ as initialization) using (3)–(9).
4: Compute occupancy envelopes $(u_{k,h}, \underline{u}_{k,h})$ via (10) (robust DP over $\mathcal{P}_{k-1}$).
5: Roll out episode k under π_k, observe trajectory $(s_{k,h}, a_{k,h})_{h=0}^{H-1}$ and losses $\ell_k(s_{k,h}, a_{k,h})$.
6: Form $Q^b_{k,h}$ by (11).
7: Compute local bonuses $b_{k,h}$ and dilated bonuses $B_{k,h}$ using (12) and (13).
8: Update π_{k+1} by (14).
9: **end for**

VLS confidence sets for linear-mixture transitions. Fix a layer $h \in \{0, \dots, H-1\}$. Recall the linear-mixture parameterization $P^\star_h(\cdot \mid s, a) = \Phi_h(s, a)\theta^\star_h$, where $\Phi_h(s, a) \in \mathbb{R}^{|\mathcal{S}_{h+1}| \times d}$ stacks the row features $\varphi_h(s' \mid s, a)^\top$ for $s' \in \mathcal{S}_{h+1}$. Given a transition sample $s_{i,h+1}$, the learner observes the one-hot vector $e_{s_{i,h+1}} \in \mathbb{R}^{|\mathcal{S}_{h+1}|}$. Vector least squares (VLS) fits $\theta^\star_h$ by ridge-regressing these one-hots against $\Phi_h(s_{i,h}, a_{i,h})$ under a squared loss [Li et al., 2024]:

$$\hat{\theta}^{\mathrm{vls}}_{k,h} \in \arg\min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{i=1}^{k} \left\| \Phi_h(s_{i,h}, a_{i,h})\theta - e_{s_{i,h+1}} \right\|_2^2. \tag{2}$$

Writing $x_{i,h}(s') := \varphi_h(s' \mid s_{i,h}, a_{i,h}) \in \mathbb{R}^d$, expanding (2) yields the normal equation $(\lambda I + \sum_{i=1}^{k} \sum_{s'} x_{i,h}(s') x_{i,h}(s')^\top)\hat{\theta}^{\mathrm{vls}}_{k,h} = \sum_{i=1}^{k} x_{i,h}(s_{i,h+1})$, i.e.,

$$\begin{aligned} \Lambda_{k,h} &:= \lambda I + \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{h+1}} x_{i,h}(s') x_{i,h}(s')^\top, \\ \hat{\theta}^{\mathrm{vls}}_{k,h} &:= \Lambda^{-1}_{k,h} \sum_{i=1}^{k} x_{i,h}(s_{i,h+1}). \end{aligned} \tag{3}$$

A key feature of VLS is that *each visit contributes* $\Phi_h(s_{i,h}, a_{i,h})^\top \Phi_h(s_{i,h}, a_{i,h}) = \sum_{s'} x_{i,h}(s') x_{i,h}(s')^\top$ *to the design matrix*, so the estimator leverages the known features of *all* next states, not only the realized $s_{i,h+1}$.

For a confidence level $\delta' \in (0, 1)$, define the radius

$$\beta^{\mathrm{vls}}_{k,h}(\delta') := \sqrt{\lambda}\, B + \sqrt{2 \ln \frac{1}{\delta'} + d \ln\left(4 + \frac{4Sk}{\lambda d}\right)}, \tag{4}$$

and the ellipsoid

$$\Theta^{\mathrm{vls}}_{k,h}(\delta') := \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}^{\mathrm{vls}}_{k,h} \right\|_{\Lambda_{k,h}} \le \beta^{\mathrm{vls}}_{k,h}(\delta') \right\}. \tag{5}$$

The radius in (4) follows from a self-normalized martingale inequality for linear regression [Abbasi-Yadkori et al., 2011]. A high-probability event guaranteeing $\theta^\star_h \in \Theta^{\mathrm{vls}}_{k,h}(\delta/H)$ for all k, h is stated as Lemma 12.

VLS also enables a uniform control of transition deviations for all bounded continuation vectors, which is the key ingredient in our improved bound on the cumulative envelope gap; see Lemma 16.

Rectangularization and robust planning primitives. The statistically natural confidence set for linear-mixture transitions is a coupled ellipsoid over the shared parameter: one maintains an ellipsoid $\Theta^{\mathrm{vls}}_{k,h}(\delta/H)$ and considers all kernels whose transition probabilities are induced by some $\theta \in \Theta^{\mathrm{vls}}_{k,h}(\delta/H)$ through the linear-mixture parametrization. This coupled set is standard in linear mixture MDPs and linear bandits [Abbasi-Yadkori et al., 2011, Ayoub et al., 2020, Zhou et al., 2021]. It is also the starting point of recent adversarial results [Li et al., 2024].

Formally, this corresponds to the ellipsoid-induced (coupled) kernel set

$$\mathcal{P}^{\mathrm{ell}}_{k,h} := \Big\{ P_h : \exists \theta \in \Theta^{\mathrm{vls}}_{k,h}(\delta/H) \text{ s.t. } P_h(\cdot \mid s, a) = \Phi_h(s, a)\theta \ \ \forall (s, a) \Big\}. \tag{6}$$

The key point is that a *single* parameter θ must simultaneously explain *all* (s, a) pairs at layer h, which couples the feasible next-state distributions.

However, this "ellipsoid-induced" kernel set is generally *not* (s, a)-rectangular: the choice of next-state distributions at different (s, a) pairs is coupled through the shared parameter θ. This coupling is incompatible with the dilated-bonus recursion, whose robust backups involve a sequence of per-step maximizations and require a dynamically consistent *joint* maximizer that can be assembled across (s, a) pairs. To restore this property, we introduce a rectangular relaxation that decouples the uncertainty across (s, a) while still being implied by the ellipsoidal parameter uncertainty.

Specifically, the dilated recursion repeatedly uses robust one-step lookaheads of the form

$$\max_{P \in \mathcal{P}_{k-1}} \mathbb{E}_{s' \sim P_h(\cdot \mid s, a)} \left[\bar{B}_{k,h+1}(s') \right], \tag{7}$$

for many (h, s, a) and continuation functions $\bar{B}_{k,h+1}$ induced by the current policy and bonus-to-go. When $\mathcal{P}_{k-1}$ is not (s, a)-rectangular (e.g., (6)), the maximizer in (7) may couple different (s, a) choices. In particular, maximizing (7) separately for different (s, a) pairs can lead to different "best" kernels, so there may be no *single* kernel that simultaneously realizes all these local maximizers. This is exactly the obstruction to a backward recursion: the dilated update requires that the robust lookaheads at all (h, s, a) can be certified by one joint optimistic kernel (formalized in Lemma 11).

Accordingly, for each (s,a) at layer h, we convert the coupled ellipsoid (5) into a local set of next-state distributions. Let $\Phi_h(s,a) \in \mathbb{R}^{|\mathcal{S}_{h+1}|\times d}$ be the feature matrix whose s'-th row equals $\varphi_h(s' \mid s,a)^\top$.

$$\mathcal{C}^{\mathrm{vls}}_{k,h}(s,a) := \Big\{ p \in \Delta(\mathcal{S}_{h+1}) : \exists \theta \in \Theta^{\mathrm{vls}}_{k,h}(\delta/H) \text{ s.t. } p = \Phi_h(s,a)\theta \Big\}. \tag{8}$$

$$\mathcal{P}_k := \{P : \ P_h(\cdot \mid s,a) \in \mathcal{C}^{\mathrm{vls}}_{k,h}(s,a)\ \forall (h,s,a)\}. \tag{9}$$

For brevity, write $\mathcal{P}_k := \mathcal{P}^{\mathrm{vls}}_k(\delta)$. The set $\mathcal{P}^{\mathrm{vls}}_k(\delta)$ is (s,a)-rectangular by construction: nature can choose the conditional distribution at each (h,s,a) independently. This enlargement is the price we pay to make robust dynamic programming and the dilated recursion tractable.

Moreover, the rectangularized set is a relaxation of the coupled ellipsoid-induced set: if $P_h \in \mathcal{P}^{\mathrm{ell}}_{k,h}$, then for each (s,a) we have $P_h(\cdot \mid s,a) = \Phi_h(s,a)\theta \in \mathcal{C}^{\mathrm{vls}}_{k,h}(s,a)$, hence $\mathcal{P}^{\mathrm{ell}}_{k,h} \subseteq \mathcal{P}_k$ at each layer.

Remark 1 (Why the relaxation does not worsen regret rates). Although $\mathcal{P}_k$ is a relaxation of the coupled ellipsoid-induced set, our final regret bound in Theorem 1 attains the same statistical rate as the best-known results in this setting. The reason is that the analysis uses the confidence set only through per-(s,a) deviation control, not through the coupling constraint "the same θ must explain all (s,a)." Concretely, each $P_h(\cdot \mid s,a) \in \mathcal{P}_k$ is realized by some witness parameter $\theta_{s,a} \in \Theta^{\mathrm{vls}}_{k,h}(\delta/H)$, so Lemma 13 yields an ℓ_1 radius $r_{k,h}(s,a)$ for every (h,s,a):

$$\sup_{P \in \mathcal{P}_{k-1}} \big\| P_h(\cdot \mid s,a) - P^\star_h(\cdot \mid s,a) \big\|_1 \le r_{k,h}(s,a).$$

Consequently, for any bounded continuation V with $\|V\|_\infty \le H$, we have for every (s,a)

$$\Big| \max_{P \in \mathcal{P}_{k-1}} \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')] - \mathbb{E}_{s' \sim P^\star(\cdot|s,a)}[V(s')] \Big| \le H\, r_{k,h}(s,a),$$

and similarly for the minimum, so the one-step optimistic/pessimistic width is controlled by the same radius:

$$\max_{P \in \mathcal{P}_{k-1}} \mathbb{E}_P[V] - \min_{P \in \mathcal{P}_{k-1}} \mathbb{E}_P[V] \le 2H\, r_{k,h}(s,a).$$

In our analysis, these one-step width bounds propagate through robust dynamic programming and enter the regret bound through the cumulative occupancy-envelope gap (defined in Section 4.1). Lemma 16 shows that under VLS this cumulative gap remains of the same order as in the coupled model, namely

$$\sum_{k=1}^{K} \sum_{h=0}^{H-1} \sum_{s,a} \big(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a) \big) \le \widetilde{\mathcal{O}}\Big(d\, S^{3/2} \sqrt{HK} \Big).$$

Thus, rectangularization does not deteriorate the statistical rate that drives Theorem 1.

Remark 2 (Why rectangularization is (usually) unnecessary in tabular and linear-MDP settings). In Luo et al. [2021], the dilated-bonus recursion is developed for both tabular MDPs and the (standard) linear-MDP model, and in these settings the transition-confidence sets used for robust backups do not require an additional rectangularization operation. In contrast, in linear-mixture MDPs the statistically natural confidence set is the coupled ellipsoid-induced family (6), where a single shared parameter must explain all (s,a) pairs at a layer; this coupling can break joint maximizers across (s,a) and motivates our rectangular relaxation.

In tabular MDPs, the standard confidence set already controls each conditional distribution separately, e.g.,

$$\mathcal{P}^{\mathrm{tab}}_{k-1}(\delta) := \Big\{ P : \forall (h,s,a), \ \big\| P_h(\cdot \mid s,a) - \hat{P}_{k-1,h}(\cdot \mid s,a) \big\|_1 \le \varepsilon_{k-1,h}(s,a) \Big\},$$

which factorizes over (h,s,a) and is therefore (s,a)-rectangular by construction.

In the (standard) linear-MDP model, the optimistic continuation can be summarized by a single global vector that is independent of (x,a): if $P_h(x' \mid x,a) = \big\langle \phi_h(x,a),\, \nu_h^{x'} \big\rangle$, then for any continuation V,

$$\mathbb{E}_{x' \sim P_h(\cdot|x,a)}[V(x')] = \Big\langle \phi_h(x,a),\, \underbrace{\sum_{x'} \nu_h^{x'} V(x')}_{\Lambda_h(V)} \Big\rangle.$$

The key point is that $\Lambda_h(V)$ does not depend on (x,a), so robust backups do not face the joint-maximizer obstruction. In contrast, for linear-mixture MDPs we only have

$$\mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')] = \big\langle \psi_h^V(s,a),\, \theta_h \big\rangle,$$
$$\psi_h^V(s,a) = \sum_{s' \in \mathcal{S}_{h+1}} \varphi_h(s' \mid s,a)\, V(s'),$$

and the coefficient vector $\psi_h^V(s,a)$ generally varies with (s,a). Thus the "optimistic direction" in parameter space changes across (s,a), which is exactly what can break joint maximizers over the coupled set (6); we therefore use the rectangular relaxation (8)–(9) to recover decomposability.

Robust occupancy envelopes. We now connect the rectangular confidence set $\mathcal{P}_{k-1}$ to the occupancy measures defined in the problem setting. Given the current policy π_k, we upper and lower bound its state-action occupancy $q_h^{\pi_k,P}(s,a)$ over all plausible kernels $P \in \mathcal{P}_{k-1}$. These bounds are computed via robust dynamic programming enabled by (s,a)-rectangularity [Iyengar, 2005, Nilim and El Ghaoui, 2005].

Define the upper and lower occupancy envelopes under the previous confidence set:

$$u_{k,h}(s,a) := \max_{P\in\mathcal{P}_{k-1}} q_h^{\pi_k,P}(s,a),$$
$$\underline{u}_{k,h}(s,a) := \min_{P\in\mathcal{P}_{k-1}} q_h^{\pi_k,P}(s,a). \tag{10}$$

On the realizability event $\{P^\star \in \mathcal{P}_{k-1}\}$, the true occupancy is sandwiched: $\underline{u}_{k,h}(s,a) \le q_h^{\pi_k,P^\star}(s,a) \le u_{k,h}(s,a)$. Moreover, $u_{k,h}$ is predictable (measurable w.r.t. past episodes), since it depends only on π_k and $\mathcal{P}_{k-1}$. This predictability is crucial for bandit concentration when we normalize importance-weighted estimators.

4.2 DILATED BONUSES AND OMD UPDATE

Building on the predictable occupancy envelopes $(u_{k,h}, \underline{u}_{k,h})$ defined in Section 4.1, we now specify the implicit-exploration (IX) estimator, the associated bonuses, and the exponential-weights (OMD) policy update. Let $L_{k,h} := \sum_{t=h}^{H-1} \ell_k(s_{k,t}, a_{k,t})$ be the observed loss-to-go along the trajectory of episode k. For a learning rate $\eta > 0$, set the IX parameter $\gamma := 2\eta H$ [Kocák et al., 2014, Neu, 2015]. Define the bandit return estimator

$$Q^b_{k,h}(s,a) := \frac{L_{k,h}\,\mathbb{1}\{(s_{k,h},a_{k,h}) = (s,a)\}}{u_{k,h}(s,a)+\gamma}. \tag{11}$$

We now describe how we turn the stabilized bandit surrogate Q^b into a full policy update. There are two sources of error that must be explicitly compensated. First, implicit exploration introduces a controlled bias through the normalization $u_{k,h}(s,a)+\gamma$. Second, because transitions are unknown, the occupancy envelope gap $u_{k,h} - \underline{u}_{k,h}$ quantifies how much the state-action visitation probabilities may vary across plausible kernels. Our bonuses are designed so that these two effects enter additively in the regret analysis.

We now define the bonuses and the policy update that together compensate the two errors identified above (IX bias and transition uncertainty). We now define the local bonus. The local bonus provides a one-step compensation that couples the IX bias (via γ) and the transition uncertainty (via the envelope gap) [Kocák et al., 2014, Neu, 2015]:

$$\widetilde{b}_{k,h}(s,a) := \frac{3\gamma H + H\big(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)\big)}{u_{k,h}(s,a)+\gamma},$$
$$b_{k,h}(s) := \sum_{a\in\mathcal{A}} \pi_{k,h}(a\mid s)\,\widetilde{b}_{k,h}(s,a). \tag{12}$$

We next define the dilated bonus. The local bonus only accounts for errors at a fixed (h,s,a). To ensure upstream decisions also pay for downstream uncertainty under the correct reachability weights, we apply the dilated-bonus recursion of Luo et al. [2021]. This performs a robust backward propagation that repeatedly applies the robust lookahead (7) under $\mathcal{P}_{k-1}$. Let $\alpha := 1 + 1/H$ and define $B_{k,H} \equiv 0$ and, for $h = H-1,\dots,0$,

$$B_{k,h}(s,a) := b_{k,h}(s) + \alpha \max_{P\in\mathcal{P}_{k-1}} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\mathbb{E}_{a'\sim\pi_{k,h+1}(\cdot|s')}\big[B_{k,h+1}(s',a')\big]. \tag{13}$$

Finally, we run an OMD update using the stabilized bandit return $Q^b_{k,h}$ together with the dilated bonus-to-go $B_{k,h}$. We use the gradient surrogate $g_{k,h}(s,a) := Q^b_{k,h}(s,a) - B_{k,h}(s,a)$:

$$\pi_{k+1,h}(a\mid s) \propto \pi_{k,h}(a\mid s)\exp\big(-\eta\, g_{k,h}(s,a)\big). \tag{14}$$

This exponential-weights (OMD) update is standard in online learning [Cesa-Bianchi and Lugosi, 2006, Hazan, 2016] and is widely used in online/adversarial MDP algorithms, including O-REPS and recent adversarial linear-mixture MDP methods [Zimin and Neu, 2013, Zhao et al., 2023, Li et al., 2024].

Computational aspects. The main computational distinction is at the **policy-update / planning** layer. In our method, after the statistical estimates and robust surrogates are computed, Eq.(14) gives a closed-form local OMD update, thus the policy update costs $\mathcal{O}(HnA)$ over all state-action pairs and decomposes over (h,s). The robust lookahead terms in Eqs. (10) and (13) are also local: they consist of $\mathcal{O}(HnA)$ independent problems of the form $\max_{p\in\mathcal{C}_{k,h}(s,a)}\langle p, V\rangle$. Under the VLS rectangularization, each local problem has effective dimension about $d+n$. With a generic dense interior-point solver, the dominant robust-backup cost is $\widetilde{\mathcal{O}}\big(HnA(d+n)^{3.5}\big)$, followed by the $\mathcal{O}(HnA)$ closed-form OMD update. Lower-order terms, such as VLS rank-one updates and feature/radius arithmetic, depend on the implementation and do not change the main comparison.

By contrast, global occupancy-measure planning methods obtain the next policy by solving one coupled optimization problem over the whole MDP. A lifted flow formulation uses variables $q_{h,s,a}$ and $y_{h,s,a,s'}$, giving global dimension $D = \mathcal{O}(HnA + Hn^2A) = \mathcal{O}(Hn^2A)$, with Bellman-flow and transition-consistency constraints coupling all layers. A generic dense interior-point implementation has per-update cost $\widetilde{\mathcal{O}}(D^{3.5}) = \widetilde{\mathcal{O}}((Hn^2A)^{3.5}) = \widetilde{\mathcal{O}}(H^{3.5}n^7A^{3.5})$ [Boyd and Vandenberghe, 2004, Nemirovskii and Nesterov, 1994]. This is a solver-model comparison rather than an implementation-independent lower bound: specialized sparse solvers may improve constants or exploit additional structure. The formal distinction is that our method replaces a single coupled global update of dimension $\widetilde{\mathcal{O}}(Hn^2A)$ by $\widetilde{\mathcal{O}}(HnA)$ independent local robust backups of dimension about $d+n$, followed by an $\widetilde{\mathcal{O}}(HnA)$ closed-form policy update.

5 MAIN RESULT

Our main theorem gives a high-probability regret bound for Algorithm 1. The main text highlights the high-level proof roadmap; all technical details and auxiliary lemmas are deferred to the supplementary material.

Theorem 1 (High-probability regret bound)**.** *Set* $\eta := \min\left\{\frac{1}{24H^3}, \sqrt{\frac{\log A}{HSAK}}\right\}$ *and* $\gamma := 2\eta H$*. Under the linear mixture model* (1)*, with probability at least* $1-\delta$*, Algorithm 1 satisfies*

$$\mathrm{Reg}(K) \le \widetilde{\mathcal{O}}\Big(H^{3/2}\sqrt{SAK} \;+\; d\,H^{3/2}S^{3/2}\sqrt{K}\Big).$$

Remark 3 (Comparison to prior work). In adversarial linear-mixture MDPs with bandit feedback and unknown transitions, Zhao et al. [2023] provides the information-theoretic lower bound $\Omega\big(\sqrt{HSAK} + d\,H\sqrt{K}\big)$. For layered MDPs in this setting, representative upper bounds include $\widetilde{\mathcal{O}}(\sqrt{HSAK} + d\,S^2\sqrt{K})$ via optimistic dynamic programming [Zhao et al., 2023] and the sharper transition-uncertainty term $\widetilde{\mathcal{O}}(d\sqrt{HS^3K})$ based on VLS and refined occupancy-difference analysis [Li et al., 2024].

Up to logarithmic factors, Theorem 1 matches the best-known *state* dependence in the transition-uncertainty term, namely $d\,S^{3/2}\sqrt{K}$. On the other hand, our horizon dependence is worse than the sharpest existing bounds by roughly a factor of H. This gap can be traced to the policy-optimization interface under bandit feedback. The IX estimator and the dilated-bonus recursion introduce surrogate losses of magnitude up to order H. Controlling the exponential-weights update therefore requires a conservative step size to maintain uniform boundedness of the updates, which is reflected in the cap $\eta \le 1/(24H^3)$ in Theorem 1. By contrast, global planning and occupancy-measure-based methods can implement optimism and transition learning through value or occupancy objects without the same need to backpropagate bandit-normalization error through local mirror-descent updates, and they can avoid this particular accumulation of horizon factors. Closing this horizon gap, while retaining a purely local policy-optimization loop with dilated bonuses, as opposed to global optimization over value/occupancy objects, remains an interesting open problem.

Regret decomposition (analysis outline). Work on the intersection of the VLS confidence event (Lemma 12) and the IX concentration event (Lemma 2). Let $d_h^\star(s) := d_h^{\pi^\star,P^\star}(s)$. Lemma 1 yields the policy-gradient-style identity

$$\mathrm{Reg}(K) = \sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot\mid s) - \pi_h^\star(\cdot\mid s),\; Q_{k,h}^{\pi_k,P^\star}(s,\cdot)\Big\rangle. \tag{15}$$

Add and subtract the bandit estimator Q^b and the dilated bonus-to-go $B_{k,h}$ to obtain

$$\begin{aligned}
&\mathrm{Reg}(K)\\
&= \underbrace{\sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot\mid s) - \pi_h^\star(\cdot\mid s),\; \big(Q_{k,h}^b - B_{k,h}\big)(s,\cdot)\Big\rangle}_{\text{REG-TERM}}\\
&\quad + \underbrace{\sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot\mid s),\; \big(Q_{k,h}^{\pi_k,P^\star} - Q_{k,h}^b\big)(s,\cdot)\Big\rangle}_{\text{BIAS-1}}\\
&\quad + \underbrace{\sum_{k,h,s} d_h^\star(s)\Big\langle \pi_h^\star(\cdot\mid s),\; \big(Q_{k,h}^b - Q_{k,h}^{\pi_k,P^\star}\big)(s,\cdot)\Big\rangle}_{\text{BIAS-2}}\\
&\quad + \underbrace{\sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot\mid s) - \pi_h^\star(\cdot\mid s),\; B_{k,h}(s,\cdot)\Big\rangle}_{\text{BONUS-COUPLING}}.
\end{aligned} \tag{16}$$

Step 1 (OMD + IX): control REG-TERM **and the two bias terms.** The first term REG-TERM is exactly the OMD regret against the per-round surrogate $a \mapsto Q_{k,h}^b(s,a) - B_{k,h}(s,a)$ under comparator $\pi_h^\star(\cdot\mid s)$, weighted by $d_h^\star(s)$. Online mirror descent yields

$$\begin{aligned}
&\text{REG-TERM}\\
&\le \widetilde{\mathcal{O}}\Big(\frac{H\log A}{\eta}\Big) + \sum_{k,h,s,a} d_h^\star(s)\,\pi_{k,h}(a\mid s)\frac{\gamma H}{u_{k,h}(s,a)+\gamma}\\
&\quad + \frac{1}{H}\sum_{k,h,s,a} d_h^\star(s)\,\pi_{k,h}(a\mid s)\,B_{k,h}(s,a)
\end{aligned} \tag{17}$$

See Lemma 5. The two middle terms in (16) capture the bias introduced by replacing the true return $Q_{k,h}^{\pi_k,P^\star}$ with its stabilized bandit estimate $Q_{k,h}^b$. The IX concentration and predictability of $u_{k,h}$ imply

$$\begin{aligned}
&\text{BIAS-1} + \text{BIAS-2}\\
&\le \widetilde{\mathcal{O}}\Big(\frac{H^2}{\gamma}\Big)\\
&\quad + H\sum_{k,h,s,a} d_h^\star(s)\,\pi_{k,h}(a\mid s)\cdot\frac{\Delta u_{k,h}(s,a)}{u_{k,h}(s,a)+\gamma}.
\end{aligned} \tag{18}$$

Here $\Delta u_{k,h}(s,a) \triangleq u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)$ denotes the occupancy-envelope gap. See Lemmas 6 and 3. Combining (16)–(18) and using the local bonus definition (12) gives

$$\begin{aligned}
&\mathrm{Reg}(K)\\
&\le \widetilde{\mathcal{O}}\Big(\frac{H\log A}{\eta} + \frac{H^2}{\gamma}\Big) + \sum_{k,h,s} d_h^\star(s)\,b_{k,h}(s)\\
&\quad + \frac{1}{H}\sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a\mid s)\,B_{k,h}(s,a)\\
&\quad + \sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot\mid s) - \pi_h^\star(\cdot\mid s),\; B_{k,h}(s,\cdot)\Big\rangle.
\end{aligned}$$

Step 2 (dilated-bonus reduction): The inequality (19) contains a local occupancy bonus term $\sum d_h^\star(s)\, b_{k,h}(s)$ and a dilated coupling term involving $B_{k,h}$. Lemma 8 shows that, by the dilated recursion (13) and (s,a)-rectangularity, these can be upper bounded by a single robust value of the bonus cost under the optimistic kernel P_k^b. Concretely, Lemma 8 upper bounds the two B-dependent terms in (19) by a single optimistic bonus value,

$$\mathrm{Reg}(K) \le \widetilde{\mathcal{O}}\left(\frac{H\log A}{\eta} + \frac{H^2}{\gamma}\right) + 3\sum_{k=1}^{K} V_k^{\pi_k, P_k^b}(s_0; b_k). \tag{19}$$

Step 3 (bonus value bound): It remains to upper bound $\sum_k V_k^{\pi_k,P_k^b}(s_0;b_k)$ in terms of explicit quantities that can be controlled. Recall the local bonus definition (12): $\widetilde{b}_{k,h}(s,a) = \frac{3\gamma H}{u_{k,h}(s,a)+\gamma} + \frac{H\,\Delta u_{k,h}(s,a)}{u_{k,h}(s,a)+\gamma}$, which separates the IX smoothing contribution (the first fraction) from transition uncertainty through the envelope gap (the second fraction). Define $b_{k,h}(s) = \sum_a \pi_{k,h}(a \mid s)\,\widetilde{b}_{k,h}(s,a)$ as in (12). Expanding the definition of the bonus value and using that $q_h^{\pi_k,P_k^b}(s,a) \le u_{k,h}(s,a)$ for $P_k^b \in \mathcal{P}_{k-1}$,

$$\begin{aligned}
&\sum_{k=1}^{K} V_k^{\pi_k,P_k^b}(s_0;b_k) \\
&\le \sum_{k,h,s,a} u_{k,h}(s,a)\left(\frac{3\gamma H}{u_{k,h}(s,a)+\gamma} + \frac{H\,\Delta u_{k,h}(s,a)}{u_{k,h}(s,a)+\gamma}\right) \\
&\le 3\gamma H \sum_{k,h,s,a} 1 + H\sum_{k,h,s,a} \Delta u_{k,h}(s,a) \\
&= 3\gamma HSAK + H\sum_{k,h,s,a} \Delta u_{k,h}(s,a).
\end{aligned} \tag{20}$$

In the penultimate inequality we used $u/(u+\gamma) \le 1$. This derivation is stated and proved formally in Lemma 9. Under VLS, Lemma 16 bounds the cumulative envelope gap by

$$\sum_{k,h,s,a} \Delta u_{k,h}(s,a) \le \widetilde{\mathcal{O}}\left(d\,S^{3/2}\sqrt{HK}\right). \tag{21}$$

This sharper $S^{3/2}$ dependence follows from combining the layered occupancy recursion with a refined control of one-step transition deviations. The envelope gap $\Delta u_{k,h}(s,a)$ is driven by deviations $\|P_h(\cdot \mid s,a) - P_h^\star(\cdot \mid s,a)\|_1$, and these deviations propagate forward through the occupancy recursion so that their cumulative effect is weighted by the reachability under $P^\star$. Under the VLS construction, each visit contributes the full feature covariance $\Phi_h(s,a)^\top \Phi_h(s,a)$ to the design matrix, which enables an elliptical-potential-type summation of the resulting radii over visited (s,a) pairs. This mechanism yields the improved cumulative envelope-gap bound in Lemma 16, building on the refined VLS analysis of Li et al. [2024]. Substituting (20)–(21) into (19) yields

$$\begin{aligned}
&\mathrm{Reg}(K) \\
&\le \widetilde{\mathcal{O}}\left(\frac{H\log A}{\eta} + \frac{H^2}{\gamma}\right) + \widetilde{\mathcal{O}}(\gamma HSAK) \\
&\quad + \widetilde{\mathcal{O}}\left(d\,H\,S^{3/2}\sqrt{HK}\right).
\end{aligned} \tag{22}$$

Substituting $\gamma = 2\eta H$, the leading terms take the form $\widetilde{\mathcal{O}}(H\log A/\eta) + \widetilde{\mathcal{O}}(\eta H^2 SAK)$. Optimizing gives $\eta \asymp \sqrt{\log A/(HSAK)}$ (up to constants and logs), which yields Theorem 1.

6 EXPERIMETS

We conduct one structured experiment to illustrate this computational distinction. It compares our local policy-optimization method with two global occupancy-based REPS baselines, VLSUOB-REPS [Li et al., 2024] and LSUOB-REPS [Zhao et al., 2023], on the same randomly generated adversarial linear-mixture instances. We sweep $H \in \{2,3,4\}$ using shared-prefix instances generated from one $H_{\max} = 4$ instance; each time layer has 4 states and each state has 4 actions, so the total layered state/state-action space scales with H. We use linear-mixture dimension $d = 4, K = 200$ episodes, and 5 random seeds. The experimental results are shown in Figure 1.

The regret curves are comparable in scale, but we do not claim that the local method dominates the baselines in regret. The runtime gap is the main point: across horizons and baselines, our method reduces the mean per-episode total compute by 15.6–30.9x and reduces the update/global-solve component by 77.7–83.7x. The runtime breakdown separates occupancy-envelope/bound construction from the actual local update or global occupancy solve, which directly reflects the algorithmic difference emphasized in the paper.

7 CONCLUSION

We studied policy optimization for adversarial linear-mixture MDPs with bandit feedback and unknown transitions. Our analysis isolates a modular interface—predictable and (s,a)-rectangular confidence sets with robust occupancy envelopes—that allows one to plug structured transition uncertainty into a dilated-bonus OMD policy-optimization loop. Instantiating this interface with VLS confidence sets yields improved control of the cumulative occupancy-envelope gap and leads to the $S^{3/2}$ state dependence in Theorem 1. While our bound matches the best known state dependence up to logarithmic factors, it comes with a worse dependence on the horizon H compared to the sharpest existing rates. An important direction is to tighten the horizon dependence within the policy-optimization framework, and to extend the interface beyond

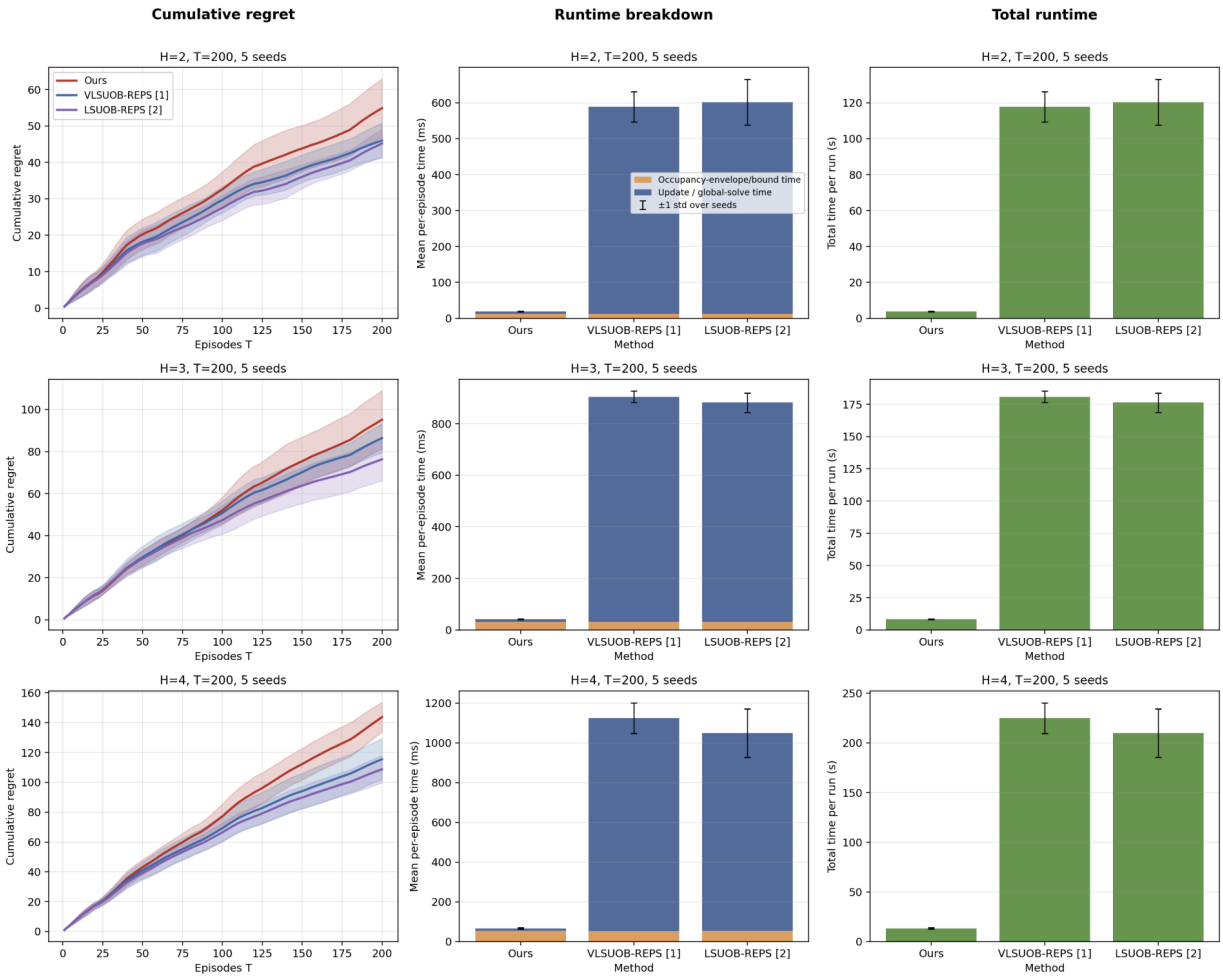

Figure 1: Experimental results

linear-mixture transitions to more general function approximation and uncertainty models.

ACKNOWLEDGMENT

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (62376154).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320, 2011.

Sanjeev Arora, Ofer Dekel, and Ambuj Tewari. Deterministic MDPs with adversarial rewards and bandit feedback. *CoRR*, abs/1210.4843, 2012. doi: 10.48550/arXiv.1210.4843.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR, 13–18 Jul 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Convex Optimization, 2004.

Nicolo Cesa-Bianchi and G'abor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yan Dai, Haipeng Luo, and Lin Chen. Follow-the-perturbed-leader for adversarial markov decision processes with

bandit feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 11437–11449. Curran Associates, Inc., 2022.

Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial MDPs with linear function approximation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6726–6759. PMLR, 23–29 Jul 2023.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

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

Elad Hazan. *Introduction to Online Convex Optimization*. Princeton University Press, 2016.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4259–4280. PMLR, 28–30 Mar 2022.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 13–18 Jul 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020b.

Tomás Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, volume 27, pages 613–621, 2014.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Improved algorithm for adversarial linear mixture MDPs with bandit feedback and unknown transition. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3061–3069. PMLR, 02–04 May 2024.

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

A. S. Nemirovskii and Y. E. Nesterov. Interior-point polynomial algorithms in convex programming. *studies in applied mathematics philadelphia siam*, 1994.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015. doi: 10.48550/arXiv.1506.03271.

Gergely Neu, Andás Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, volume 23, pages 1804–1812. Curran Associates, Inc., 2010.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Avrim Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 09–15 Jun 2019.

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

Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, and Shuai Li. Learning adversarial linear mixture markov decision

processes with bandit feedback and unknown transition. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, 2023.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR, 15–19 Aug 2021.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26, pages 1583–1591, 2013.

Policy Optimization for Adversarial Linear Mixture MDPs (Supplementary Material)

Yutian Cheng[*1] **Canzhe Zhao**[*1] **Shuai Li**[†1,2]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Shanghai Innovation Institute, Shanghai, China

A PROOFS AND TECHNICAL DETAILS

This supplementary material contains all lemma statements and proofs used in the proof of Theorem 1.

A.1 A COMPARATOR-WEIGHTED PERFORMANCE DIFFERENCE LEMMA

Lemma 1 (Comparator-weighted performance difference lemma)**.** *Fix an episode k and a transition kernel P. Let $Q_{k,h}^{\pi,P}(s,a)$ be the Q-function under policy π and transitions P for loss ℓ_k. Then for any two policies π and $\pi^\star$,*

$$V_k^{\pi,P}(s_0) - V_k^{\pi^\star,P}(s_0) = \sum_{h=0}^{H-1} \sum_{s\in\mathcal{S}_h} d_h^{\pi^\star,P}(s)\Big\langle \pi_h(\cdot \mid s) - \pi_h^\star(\cdot \mid s),\ Q_{k,h}^{\pi,P}(s,\cdot)\Big\rangle.$$

Proof. Fix (k,P) and abbreviate $Q_h(\cdot,\cdot) \triangleq Q_{k,h}^{\pi,P}(\cdot,\cdot)$ and $V_h(s) \triangleq \langle \pi_h(\cdot \mid s), Q_h(s,\cdot)\rangle$. By definition, $V_0(s_0) = V_k^{\pi,P}(s_0)$.

For any trajectory (regardless of the generating policy), the Bellman identity holds pointwise:

$$Q_h(s_h,a_h) = \ell_k(s_h,a_h) + \mathbb{E}\big[V_{h+1}(s_{h+1}) \mid s_h,a_h\big].$$

Taking expectation over (s_h,a_h) generated by $(\pi^\star,P)$ yields

$$\mathbb{E}_{\pi^\star,P}[\ell_k(s_h,a_h)] = \mathbb{E}_{\pi^\star,P}[Q_h(s_h,a_h)] - \mathbb{E}_{\pi^\star,P}[V_{h+1}(s_{h+1})].$$

Summing from $h=0$ to $H-1$ telescopes and uses $V_H \equiv 0$:

$$V_k^{\pi^\star,P}(s_0) = \mathbb{E}_{\pi^\star,P}\Big[\sum_{h=0}^{H-1}\ell_k(s_h,a_h)\Big] = V_0(s_0) - \sum_{h=0}^{H-1}\mathbb{E}_{\pi^\star,P}\Big[V_h(s_h) - \mathbb{E}_{a\sim\pi_h^\star(\cdot|s_h)}Q_h(s_h,a)\Big].$$

Rearranging and writing expectations using $d_h^{\pi^\star,P}$ gives

$$V_k^{\pi,P}(s_0) - V_k^{\pi^\star,P}(s_0) = \sum_{h=0}^{H-1}\sum_{s} d_h^{\pi^\star,P}(s)\Big(\langle \pi_h(\cdot \mid s), Q_h(s,\cdot)\rangle - \langle \pi_h^\star(\cdot \mid s), Q_h(s,\cdot)\rangle\Big),$$

which is the claim. □

[*] These authors contributed equally.
[†] Correspondence to: Shuai Li <shuaili8@sjtu.edu.cn>

A.2 AUXILIARY CONCENTRATION: IMPLICIT EXPLORATION

Lemma 2 (Implicit exploration, per-layer form)**.** *Fix a finite set $\mathcal{U}$ and a predictable sequence of distributions $p_t \in \Delta(\mathcal{U})$. At round t, sample $U_t \sim p_t$ and define indicators $I_t(u) = \mathbb{1}\{U_t = u\}$. Let $Z_t(u) \in [0, R]$ be random and $z_t(u) \in [0, R]$ be predictable such that $\mathbb{E}[Z_t(u) \mid \mathcal{F}_{t-1}] = z_t(u)$ for all u. Then for any $\gamma > 0$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \sum_{u \in \mathcal{U}} \left(\frac{I_t(u) Z_t(u)}{p_t(u) + \gamma} - z_t(u) \right) \le \frac{R}{\gamma} \ln \frac{1}{\delta}.$$

Proof. The argument is standard (see, e.g., Luo et al. [2021, Lemma A.2]); we include it for completeness.

Fix $\lambda \triangleq \gamma / R$ and define

$$X_t \triangleq \sum_{u \in \mathcal{U}} \left(\frac{I_t(u) Z_t(u)}{p_t(u) + \gamma} - z_t(u) \right) = \frac{Z_t(U_t)}{p_t(U_t) + \gamma} - \sum_{u \in \mathcal{U}} z_t(u).$$

One can verify that $\mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_{t-1}] \le 1$. Therefore $\exp(\lambda \sum_{t=1}^{T} X_t)$ is a supermartingale, and Markov's inequality gives

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t \ge \frac{1}{\lambda} \ln \frac{1}{\delta} \right) \le \delta.$$

Since $1/\lambda = R/\gamma$, this yields the stated bound. □

A.3 REGRET DECOMPOSITION AND THE THREE MAIN TERMS

Applying Lemma 1 with $P = P^\star$ and $\pi = \pi_k$ and summing over k yields

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} \sum_{h=0}^{H-1} \sum_{s \in \mathcal{S}_h} d_h^\star(s) \Big\langle \pi_{k,h}(\cdot \mid s) - \pi_h^\star(\cdot \mid s),\ Q_{k,h}^{\pi_k, P^\star}(s, \cdot) \Big\rangle,$$

where $d_h^\star(\cdot) = d_h^{\pi^\star, P^\star}(\cdot)$.

Add and subtract the stabilized bandit estimator Q^b and the dilated bonus-to-go $B_{k,h}$ to obtain

$$\begin{aligned}
&\mathrm{Reg}(K) \\
&= \underbrace{\sum_{k,h,s} d_h^\star(s) \Big\langle \pi_{k,h}(\cdot \mid s) - \pi_h^\star(\cdot \mid s),\ \big(Q_{k,h}^b - B_{k,h}\big)(s, \cdot) \Big\rangle}_{\text{REG-TERM}} \\
&\quad + \underbrace{\sum_{k,h,s} d_h^\star(s) \Big\langle \pi_{k,h}(\cdot \mid s),\ \big(Q_{k,h}^{\pi_k, P^\star} - Q_{k,h}^b\big)(s, \cdot) \Big\rangle}_{\text{BIAS-1}} \\
&\quad + \underbrace{\sum_{k,h,s} d_h^\star(s) \Big\langle \pi_h^\star(\cdot \mid s),\ \big(Q_{k,h}^b - Q_{k,h}^{\pi_k, P^\star}\big)(s, \cdot) \Big\rangle}_{\text{BIAS-2}} \\
&\quad + \underbrace{\sum_{k,h,s} d_h^\star(s) \Big\langle \pi_{k,h}(\cdot \mid s) - \pi_h^\star(\cdot \mid s),\ B_{k,h}(s, \cdot) \Big\rangle}_{\text{BONUS-COUPLING}}.
\end{aligned} \tag{23}$$

We control REG-TERM, BIAS-1, and BIAS-2 in Lemmas 5, 6, and 3, respectively. The remaining BONUS-COUPLING term is handled by the dilated-bonus reduction in Lemma 8.

A.4 BOUNDING BIAS-2

Lemma 3 (BIAS-2 bound)**.** *On the event that $P^\star \in \mathcal{P}_{k-1}$ for all k, with probability at least $1-\delta$,*

$$\text{BIAS-2} \le \widetilde{\mathcal{O}}\left(\frac{H^2}{\gamma}\right).$$

Proof. We expand

$$\text{BIAS-2} = \sum_{k,h,s} d_h^\star(s)\Big\langle \pi_h^\star(\cdot \mid s),\ Q_{k,h}^b(s,\cdot) - Q_{k,h}^{\pi_k,P^\star}(s,\cdot)\Big\rangle.$$

Fix a layer h and define $\mathcal{U} = \mathcal{S}_h \times \mathcal{A}$. Let

$$p_k(s,a) \triangleq q_h^{\pi_k,P^\star}(s,a), \qquad I_k(s,a) \triangleq \mathbb{1}\{(s_{k,h},a_{k,h}) = (s,a)\}.$$

Define

$$z_{k,h}(s,a) \triangleq d_h^\star(s)\pi_h^\star(a \mid s)\, Q_{k,h}^{\pi_k,P^\star}(s,a),$$

and the "filled-in" random variable

$$Z_{k,h}(s,a) \triangleq d_h^\star(s)\pi_h^\star(a \mid s)\Big(I_k(s,a)\, L_{k,h} + (1 - I_k(s,a))\, Q_{k,h}^{\pi_k,P^\star}(s,a)\Big).$$

Then $0 \le Z_{k,h}(s,a) \le H$ and $\mathbb{E}[Z_{k,h}(s,a) \mid \mathcal{F}_{k-1}] = z_{k,h}(s,a)$.

Also note

$$Q_{k,h}^b(s,a) = \frac{L_{k,h}\, I_k(s,a)}{u_{k,h}(s,a) + \gamma}.$$

Thus for fixed h,

$$\sum_k \sum_{s,a} d_h^\star(s)\pi_h^\star(a \mid s)\, Q_{k,h}^b(s,a) = \sum_k \sum_{s,a} \frac{I_k(s,a) Z_{k,h}(s,a)}{u_{k,h}(s,a) + \gamma}.$$

On the realizability event, $u_{k,h}(s,a) \ge p_k(s,a)$, hence $(u_{k,h}+\gamma)^{-1} \le (p_k+\gamma)^{-1}$. Apply Lemma 2 with $R = H$ and failure probability δ/H, and union bound over h. This yields, with probability at least $1-\delta$,

$$\sum_k \sum_{s,a} \left(\frac{I_k(s,a) Z_{k,h}(s,a)}{p_k(s,a) + \gamma} - z_{k,h}(s,a)\right) \le \frac{H}{\gamma} \ln \frac{H}{\delta}.$$

Rearranging and summing over h gives BIAS-2 $\le \widetilde{\mathcal{O}}(H^2/\gamma)$. □

A.5 BOUNDING REG-TERM

Lemma 4 (Exponential-weights regret bound, per-state)**.** *Let $\pi_t \in \Delta(\mathcal{A})$ be updated via $\pi_{t+1}(a) \propto \pi_t(a) e^{-\eta \ell_t(a)}$ and assume $|\eta \ell_t(a)| \le 1$. Then for any $\pi^\star \in \Delta(\mathcal{A})$,*

$$\sum_{t=1}^T \sum_{a \in \mathcal{A}} \big(\pi_t(a) - \pi^\star(a)\big)\ell_t(a) \le \frac{\ln A}{\eta} + \eta \sum_{t=1}^T \sum_{a \in \mathcal{A}} \pi_t(a) \ell_t(a)^2.$$

Proof. This is the standard multiplicative-weights analysis (see, e.g., Cesa-Bianchi and Lugosi [2006, Chapter 2]). Let $w_t(a)$ be the unnormalized weights with $\pi_t(a) = w_t(a)/\sum_{a'} w_t(a')$. The update implies $\ln \frac{\sum_a w_{t+1}(a)}{\sum_a w_t(a)} = \ln \mathbb{E}_{a \sim \pi_t}[e^{-\eta \ell_t(a)}]$. Using $e^{-x} \le 1 - x + x^2$ for $|x| \le 1$ and $\ln(1+u) \le u$ yields

$$\ln \frac{\sum_a w_{t+1}(a)}{\sum_a w_t(a)} \le -\eta \langle \pi_t, \ell_t \rangle + \eta^2 \langle \pi_t, \ell_t^2 \rangle.$$

Summing over t and comparing to any fixed action distribution $\pi^\star$ (via a standard relative-entropy argument) gives $\sum_t \langle \pi_t - \pi^\star, \ell_t \rangle \le \frac{\ln A}{\eta} + \eta \sum_t \langle \pi_t, \ell_t^2 \rangle$, completing the proof. □

Lemma 5 (REG-TERM bound)**.** *On the realizability event and with probability at least* $1-\delta$*,*

$$\text{REG-TERM} \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)\left(\frac{\gamma H}{u_{k,h}(s,a)+\gamma} + \frac{1}{H}B_{k,h}(s,a)\right).$$

Proof. Apply Lemma 4 state-wise for each (h,s) to the sequence $\ell_{k,h}(\cdot) = g_{k,h}(s,\cdot) = Q^b_{k,h}(s,\cdot) - B_{k,h}(s,\cdot)$. For each fixed (h,s),

$$\sum_{k=1}^{K}\Big\langle \pi_{k,h}(\cdot \mid s) - \pi_h^\star(\cdot \mid s),\ Q^b_{k,h}(s,\cdot) - B_{k,h}(s,\cdot)\Big\rangle \le \frac{\ln A}{\eta} + \eta\sum_{k,a}\pi_{k,h}(a \mid s)\big(Q^b_{k,h}(s,a) - B_{k,h}(s,a)\big)^2,$$

provided $|\eta(Q^b_{k,h} - B_{k,h})| \le 1$, verified in Lemma 7.

Multiply by $d_h^\star(s)$ and sum over (h,s) (using $\sum_{s\in\mathcal{S}_h} d_h^\star(s) = 1$):

$$\text{REG-TERM} \le \frac{H\ln A}{\eta} + \eta\sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)\big(Q^b_{k,h}(s,a) - B_{k,h}(s,a)\big)^2.$$

Use $(x-y)^2 \le 2x^2 + 2y^2$ to split the square. For the B^2 term, Lemma 7 implies $\eta B_{k,h}(s,a) \le 1/(2H)$, hence $2\eta B_{k,h}(s,a)^2 \le \frac{1}{H}B_{k,h}(s,a)$.

For the $(Q^b)^2$ term, note $(Q^b_{k,h}(s,a))^2 \le H^2\,\mathbb{1}\{(s_{k,h},a_{k,h}) = (s,a)\}/(u_{k,h}(s,a)+\gamma)^2$. Applying Lemma 2 with $R = 1/\gamma$ as in the original analysis of Luo et al. [2021] gives

$$2\eta\sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)(Q^b_{k,h}(s,a))^2 \le \sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)\frac{\gamma H}{u_{k,h}(s,a)+\gamma} + \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right),$$

using $\gamma = 2\eta H$. Collecting terms proves the claim. □

A.6 BOUNDING BIAS-1

Lemma 6 (BIAS-1 bound)**.** *On the realizability event and with probability at least* $1-\delta$*,*

$$\text{BIAS-1} \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)\left(\frac{2\gamma H}{u_{k,h}(s,a)+\gamma} + \frac{H\big(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)\big)}{u_{k,h}(s,a)+\gamma}\right).$$

Proof. We expand BIAS-1 and separate a martingale term and a deterministic bias term.

Write

$$\text{BIAS-1} = \sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot \mid s),\ Q^{\pi_k,P^\star}_{k,h}(s,\cdot)\Big\rangle - \sum_{k,h,s} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot \mid s),\ Q^b_{k,h}(s,\cdot)\Big\rangle.$$

Define

$$Y_{k,h} \triangleq \sum_{s\in\mathcal{S}_h} d_h^\star(s)\Big\langle \pi_{k,h}(\cdot \mid s),\ Q^b_{k,h}(s,\cdot)\Big\rangle = d_h^\star(s_{k,h})\pi_{k,h}(a_{k,h} \mid s_{k,h})\frac{L_{k,h}}{u_{k,h}(s_{k,h},a_{k,h})+\gamma}.$$

Conditioned on $\mathcal{F}_{k-1}$, the trajectory of episode k is generated by $(\pi_k, P^\star)$, so

$$\mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k-1}] = \sum_{s,a} d_h^\star(s)\pi_{k,h}(a \mid s)\,\frac{q_h^{\pi_k,P^\star}(s,a)}{u_{k,h}(s,a)+\gamma}\,Q^{\pi_k,P^\star}_{k,h}(s,a).$$

Hence

$$\text{BIAS-1} = \underbrace{\sum_{k,h}\big(\mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k-1}] - Y_{k,h}\big)}_{\text{(I) martingale term}} + \underbrace{\sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)Q^{\pi_k,P^\star}_{k,h}(s,a)\left(1 - \frac{q_h^{\pi_k,P^\star}(s,a)}{u_{k,h}(s,a)+\gamma}\right)}_{\text{(II) deterministic term}}.$$

Term I: Freedman with self-bounding variance. Define $Y_k \triangleq \sum_h Y_{k,h}$, $\mu_k \triangleq \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}]$, and the martingale difference $X_k \triangleq \mu_k - Y_k$. Since $0 \le L_{k,h} \le H$ and $u_{k,h} + \gamma \ge \gamma$, we have $0 \le Y_{k,h} \le H/\gamma$ and thus $0 \le Y_k \le H^2/\gamma$, implying $|X_k| \le H^2/\gamma$.

Moreover, by $(\sum_h y_h)^2 \le H \sum_h y_h^2$ and using $(d_h^\star \pi_{k,h})^2 \le d_h^\star \pi_{k,h}$,

$$\mathrm{Var}(X_k \mid \mathcal{F}_{k-1}) \le \mathbb{E}[Y_k^2 \mid \mathcal{F}_{k-1}] \le H^3 \sum_{h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s) \frac{1}{u_{k,h}(s,a)+\gamma},$$

where we used $q_h^{\pi_k, P^\star}(s,a) \le u_{k,h}(s,a)$ on the realizability event. Applying Freedman's inequality (Freedman 1975) and linearizing the resulting square root via AM–GM yields, with probability at least $1-\delta$,

$$\sum_{k,h} \big(\mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k-1}] - Y_{k,h}\big) \le \sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s) \frac{\gamma H}{u_{k,h}(s,a)+\gamma} + \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right),$$

using $\gamma = 2\eta H$.

Term II: deterministic bias. Since $Q_{k,h}^{\pi_k, P^\star}(s,a) \le H$,

$$Q_{k,h}^{\pi_k, P^\star}(s,a)\left(1 - \frac{q_h^{\pi_k, P^\star}(s,a)}{u_{k,h}(s,a)+\gamma}\right) \le H \, \frac{u_{k,h}(s,a) + \gamma - q_h^{\pi_k, P^\star}(s,a)}{u_{k,h}(s,a)+\gamma}.$$

On the realizability event, $q_h^{\pi_k, P^\star}(s,a) \ge \underline{u}_{k,h}(s,a)$ and $u_{k,h}(s,a) \ge q_h^{\pi_k, P^\star}(s,a)$, so $u_{k,h} + \gamma - q_h^{\pi_k, P^\star} \le (u_{k,h} - \underline{u}_{k,h}) + \gamma$. Therefore

$$Q_{k,h}^{\pi_k, P^\star}(s,a)\left(1 - \frac{q_h^{\pi_k, P^\star}(s,a)}{u_{k,h}(s,a)+\gamma}\right) \le H \left(\frac{u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)}{u_{k,h}(s,a)+\gamma} + \frac{\gamma}{u_{k,h}(s,a)+\gamma}\right).$$

Combining the γ-fraction here with the martingale term yields the displayed $\frac{2\gamma H}{u+\gamma}$ contribution. Summing over (k,h,s,a) completes the proof. □

A.7 BOUNDEDNESS NEEDED FOR EXPONENTIAL-WEIGHTS UPDATES

Lemma 7 (Boundedness of Q^b, $\widetilde{b}$, b, and B)**.** *Assume $\gamma = 2\eta H$ and $\eta \le 1/(24H^3)$. Then for all (k,h,s,a),*

$$\eta Q_{k,h}^b(s,a) \le \frac{1}{2}, \qquad \eta B_{k,h}(s,a) \le \frac{1}{2H}, \qquad B_{k,h}(s,a) \le 12H^2.$$

Moreover,

$$\widetilde{b}_{k,h}(s,a) \le 4H, \qquad b_{k,h}(s) \le 4H.$$

Proof. We have $Q_{k,h}^b(s,a) \le H/\gamma$, hence $\eta Q_{k,h}^b(s,a) \le \eta H/\gamma = 1/2$.

Using $u_{k,h}(s,a) + \gamma \ge \gamma$ and $0 \le u_{k,h} - \underline{u}_{k,h} \le u_{k,h}$,

$$\widetilde{b}_{k,h}(s,a) = \frac{3\gamma H}{u_{k,h}(s,a)+\gamma} + \frac{H\big(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)\big)}{u_{k,h}(s,a)+\gamma} \le 3H + H\frac{u_{k,h}(s,a)}{u_{k,h}(s,a)+\gamma} \le 4H.$$

Taking expectation over $a \sim \pi_{k,h}(\cdot \mid s)$ yields $b_{k,h}(s) \le 4H$.

Backward induction in the dilated recursion with $\alpha = 1 + 1/H$ yields $B_{k,h}(s,a) \le 12H^2$. Thus $\eta B_{k,h}(s,a) \le 12\eta H^2 \le 1/(2H)$. □

A.8 DILATED-BONUS REDUCTION

Lemma 8 (Dilated reduction (state-only local bonus form))**.** *Suppose that on some event we have*

$$\mathrm{Reg}(K) \le o(K) + \sum_{k,h,s} d_h^\star(s)\, b_{k,h}(s) + \frac{1}{H} \sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a \mid s)\, B_{k,h}(s,a) + \sum_{k,h,s,a} d_h^\star(s)\big(\pi_{k,h}(a \mid s) - \pi_h^\star(a \mid s)\big)\, B_{k,h}(s,a).$$

Assume further that $B_{k,h}$ satisfies the dilated recursion in Algorithm 1 with $B_{k,H} \equiv 0$, and that $\mathcal{P}_{k-1}$ is (s,a)-rectangular so the per-(s,a) maximizers can be assembled into a single joint kernel $P_k^b \in \mathcal{P}_{k-1}$. Then

$$\mathrm{Reg}(K) \le o(K) + 3\sum_{k=1}^{K} V_k^{\pi_k, P_k^b}(s_0; b_k),$$

where $V_k^{\pi_k, P_k^b}(s_0; b_k)$ denotes the value of the state-only bonus cost $b_k = \{b_{k,h}(s)\}$ under policy π_k and kernel P_k^b.

Proof. Since the uncertainty set is (s,a)-rectangular, we can construct a single kernel $P_k^b \in \mathcal{P}_{k-1}$ that simultaneously attains all maxima in the definition of $B_{k,h}(s,a)$ by selecting each conditional distribution independently at its argmax.

Define the π_k-averaged dilated bonus-to-go $\bar{B}_{k,h}(s) \triangleq \mathbb{E}_{a\sim\pi_{k,h}(\cdot|s)}[B_{k,h}(s,a)]$. Taking expectation over $a \sim \pi_{k,h}(\cdot \mid s)$ in the recursion gives

$$\bar{B}_{k,h}(s) = b_{k,h}(s) + \alpha \mathbb{E}_{s'\sim P_{k,h}^{b,\pi_k}(\cdot|s)}[\bar{B}_{k,h+1}(s')],$$

where $P_{k,h}^{b,\pi_k}(\cdot \mid s)$ is the one-step transition induced by first sampling $a \sim \pi_{k,h}(\cdot \mid s)$ and then $s' \sim P_{k,h}^b(\cdot \mid s,a)$. Unrolling yields

$$\bar{B}_{k,0}(s_0) = \mathbb{E}\Big[\sum_{h=0}^{H-1} \alpha^h b_{k,h}(s_{k,h}) \,\Big|\, (\pi_k, P_k^b)\Big] \le \alpha^H V_k^{\pi_k, P_k^b}(s_0; b_k) \le 3\, V_k^{\pi_k, P_k^b}(s_0; b_k),$$

since $\alpha^H \le e < 3$.

Next, because $P^\star \in \mathcal{P}_{k-1}$ on the realizability event and $B_{k,h}$ is defined by a max over $\mathcal{P}_{k-1}$, we have the optimism inequality

$$B_{k,h}(s,a) \ge b_{k,h}(s) + \alpha \mathbb{E}_{s'\sim P_h^\star(\cdot|s,a)} \mathbb{E}_{a'\sim\pi_{k,h+1}(\cdot|s')}[B_{k,h+1}(s',a')].$$

Taking expectation over $a \sim \pi_h^\star(\cdot \mid s)$, multiplying by $d_h^\star(s)$, and summing over s yields a telescoping inequality. Summing over h and rearranging shows that the three B-terms in the premise are bounded by $\alpha \bar{B}_{k,0}(s_0)$. Combining with the bound on $\bar{B}_{k,0}(s_0)$ completes the proof. □

A.9 BOUNDING THE CUMULATIVE BONUS VALUE

Lemma 9 (Cumulative bonus value)**.** *On the realizability event,*

$$\sum_{k=1}^{K} V_k^{\pi_k, P_k^b}(s_0; b_k) \le 3\gamma H\, SA\, K + H \sum_{k,h,s,a} \Big(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)\Big).$$

Proof. Since $b_{k,h}(s) = \sum_a \pi_{k,h}(a \mid s)\widetilde{b}_{k,h}(s,a)$, we can rewrite

$$V_k^{\pi_k, P_k^b}(s_0; b_k) = \sum_{h,s} d_h^{\pi_k, P_k^b}(s) b_{k,h}(s) = \sum_{h,s,a} q_h^{\pi_k, P_k^b}(s,a)\, \widetilde{b}_{k,h}(s,a).$$

Plug in $\widetilde{b}_{k,h}$ and split. For the first part,

$$\sum_{h,s,a} q_h^{\pi_k, P_k^b}(s,a) \frac{3\gamma H}{u_{k,h}(s,a) + \gamma} \le 3\gamma H \sum_{h,s,a} \frac{u_{k,h}(s,a)}{u_{k,h}(s,a) + \gamma} \le 3\gamma H \sum_{h=0}^{H-1} \sum_{s\in\mathcal{S}_h} \sum_{a\in\mathcal{A}} 1 \le 3\gamma H\, SA,$$

using $q^{\pi_k,P_k^b} \le u$ and $u/(u+\gamma) \le 1$. Summing over k gives $3\gamma H\,SA\,K$.

For the second part,

$$\sum_{h,s,a} q_h^{\pi_k,P_k^b}(s,a)\frac{H(u_{k,h}(s,a)-\underline{u}_{k,h}(s,a))}{u_{k,h}(s,a)+\gamma} \le H\sum_{h,s,a}\big(u_{k,h}(s,a)-\underline{u}_{k,h}(s,a)\big),$$

again using $q^{\pi_k,P_k^b} \le u$. Summing over k completes the proof. □

A.10 A PLUG-IN PRINCIPLE FOR CONFIDENCE-SET REPLACEMENT

Lemma 10 (A plug-in principle for confidence-set replacement)**.** *Fix any sequence of transition confidence sets $\{\mathcal{P}_{k-1}\}_{k\ge1}$ and define the UOB/LOB envelopes by*

$$u_{k,h}(s,a) = \max_{P\in\mathcal{P}_{k-1}} q_h^{\pi_k,P}(s,a), \qquad \underline{u}_{k,h}(s,a) = \min_{P\in\mathcal{P}_{k-1}} q_h^{\pi_k,P}(s,a).$$

Assume the following hold on some event $\mathcal{E}$:

(C1) (Predictability) $\mathcal{P}_{k-1}$ is $\mathcal{F}_{k-1}$-measurable.

(C2) (Realizability) $P^\star \in \mathcal{P}_{k-1}$ for all k.

(C3) (Rectangularity) $\mathcal{P}_{k-1}$ is (s,a)-rectangular.

Then, on $\mathcal{E}$, Lemmas 6–9 and Lemma 8 continue to hold verbatim. In particular, the local bonus form and dilated recursion remain valid after replacing the confidence sets.

Proof. The bounds in Lemmas 6, 3, and 5 only require (i) $u_{k,h}(s,a) \ge q_h^{\pi_k,P^\star}(s,a)$ and $\underline{u}_{k,h}(s,a) \le q_h^{\pi_k,P^\star}(s,a)$, and (ii) predictability of $u, \underline{u}$ to apply martingale concentration. Both follow from (C1)–(C2) and the definitions.

Lemma 8 uses (i) realizability to obtain optimism in the recursion and (ii) rectangularity to guarantee the existence of a joint maximizer P_k^b. Lemma 9 additionally uses $q_h^{\pi_k,P_k^b}(s,a) \le u_{k,h}(s,a)$, which holds because $P_k^b \in \mathcal{P}_{k-1}$ by (C3). No step depends on how $\mathcal{P}_{k-1}$ is constructed. □

A.11 VLS TRANSITION ESTIMATION AND CONFIDENCE SETS

We now instantiate the plug-in principle with the VLS estimator of Li et al. [2024].

VLS estimator. For each layer h, define the regularized VLS objective

$$\hat\theta_{k,h}^{\mathrm{vls}} \in \arg\min_{\theta\in\mathbb{R}^d}\left\{\lambda\|\theta\|_2^2 + \sum_{i=1}^{k}\sum_{s'\in\mathcal{S}_{h+1}}\big(\langle\varphi_h(s'\mid s_{i,h},a_{i,h}),\,\theta\rangle - \mathbb{1}\{s_{i,h+1}=s'\}\big)^2\right\}.$$

The closed form is given in Algorithm 1.

Ellipsoids and rectangularization. For a confidence level $\delta'\in(0,1)$, define $\beta_{k,h}^{\mathrm{vls}}(\delta')$ as in Eq. (4) and the ellipsoid $\Theta_{k,h}^{\mathrm{vls}}(\delta') \triangleq \{\theta : \big\|\theta-\hat\theta_{k,h}^{\mathrm{vls}}\big\|_{\Lambda_{k,h}} \le \beta_{k,h}^{\mathrm{vls}}(\delta')\}$. For each (s,a) at layer h, define the local confidence set

$$\mathcal{C}_{k,h}^{\mathrm{vls}}(s,a) \triangleq \big\{p\in\Delta(\mathcal{S}_{h+1}) :\ \exists\theta\in\Theta_{k,h}^{\mathrm{vls}}(\delta') \text{ s.t. } p(s') = \langle\varphi_h(s'\mid s,a),\,\theta\rangle\ \ \forall s'\in\mathcal{S}_{h+1}\big\},$$

and the (s,a)-rectangular set $\mathcal{P}_k \triangleq \{P : P_h(\cdot\mid s,a)\in\mathcal{C}_{k,h}^{\mathrm{vls}}(s,a)\ \forall(h,s,a)\}$.

Lemma 11 (Robust dynamic programming under rectangular sets)**.** *Let $\mathcal{P} = \{P : P_h(\cdot \mid s,a) \in \mathcal{C}_h(s,a)\ \forall (h,s,a)\}$ be (s,a)-rectangular and let π be any fixed policy. For any bounded stage-wise function $g = \{g_h\}_{h=0}^{H-1}$, define*

$$V^{\pi,P}(s;g) := \mathbb{E}^{\pi,P}\left[\sum_{h=0}^{H-1} g_h(s_h,a_h) \,\middle|\, s_0 = s\right].$$

Then the robust value admits a backward recursion: $V_H(s) = 0$ and, for $h = H-1,\dots,0$,

$$V_h(s) = \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\Big(g_h(s,a) + \max_{p\in\mathcal{C}_h(s,a)} \mathbb{E}_{s'\sim p}[V_{h+1}(s')]\Big). \tag{24}$$

Moreover, there exists a kernel $P^b \in \mathcal{P}$ attaining the maximum, i.e., $\max_{P\in\mathcal{P}} V^{\pi,P}(s;g) = V^{\pi,P^b}(s;g)$ for all s. An analogous statement holds for $\min_{P\in\mathcal{P}}$ with $\max$ replaced by $\min$.

Proof. Fix π and g. For $h \in \{0,\dots,H\}$ define the robust-to-go value

$$V_h^\star(s) := \max_{P\in\mathcal{P}} \mathbb{E}^{\pi,P}\left[\sum_{t=h}^{H-1} g_t(s_t,a_t) \,\middle|\, s_h = s\right],$$

with $V_H^\star \equiv 0$. By the tower property, for any $P \in \mathcal{P}$,

$$\mathbb{E}^{\pi,P}\left[\sum_{t=h}^{H-1} g_t(s_t,a_t) \,\middle|\, s_h = s\right] = \sum_a \pi_h(a \mid s)\left(g_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot\mid s,a)}\left[\mathbb{E}^{\pi,P}\left[\sum_{t=h+1}^{H-1} g_t(s_t,a_t) \,\middle|\, s_{h+1} = s'\right]\right]\right).$$

Maximizing the inner conditional expectation over $P \in \mathcal{P}$ gives $\mathbb{E}[\cdot] \le V_{h+1}^\star(s')$, so

$$V_h^\star(s) \le \sum_a \pi_h(a \mid s)\Big(g_h(s,a) + \max_{p\in\mathcal{C}_h(s,a)} \mathbb{E}_{s'\sim p}[V_{h+1}^\star(s')]\Big),$$

where we used rectangularity to decouple the choice of $P_h(\cdot \mid s,a)$ from other state-action pairs and future layers. Conversely, choose for each (s,a) a maximizer $p_{h,s,a} \in \arg\max_{p\in\mathcal{C}_h(s,a)} \mathbb{E}_{s'\sim p}[V_{h+1}^\star(s')]$, and define a kernel P^b by setting $P_h^b(\cdot \mid s,a) = p_{h,s,a}$ for all (h,s,a). Then $P^b \in \mathcal{P}$ and plugging it into the tower expansion yields equality, establishing the recursion (24) and the existence of an attaining kernel. The proof for the minimization case is identical with $\max$ replaced by $\min$. □

Lemma 12 (VLS confidence event)**.** *With probability at least $1-\delta$, simultaneously for all $k \ge 0$ and all layers h,*

$$\theta_h^\star \in \Theta_{k,h}^{\mathrm{vls}}(\delta/H).$$

Proof. This follows from the self-normalized concentration for the VLS estimator. The proof is given (with the same constants) in Li et al. [2024]; we adapt it to our notation. For completeness, we sketch the argument.

Let $\xi_{i,h}(s') \triangleq \mathbb{1}\{s_{i,h+1} = s'\} - P_h^\star(s' \mid s_{i,h}, a_{i,h})$. The noise vector $(\xi_{i,h}(s'))_{s'\in\mathcal{S}_{h+1}}$ is conditionally mean-zero given $\mathcal{F}_{i-1}$ but correlated across s'. Li et al. [2024] prove a vector self-normalized inequality that controls $\left\|\sum_{i\le k} \varphi_h(\cdot \mid s_{i,h},a_{i,h})\xi_{i,h}\right\|_{\Lambda_{k,h}^{-1}}$ uniformly over k. Combining with the ridge term $\sqrt{\lambda}\,\|\theta_h^\star\|_2$ yields (4). □

Lemma 13 (VLS ℓ_1 deviation bound)**.** *On the event of Lemma 12, for any k,h and any $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, every $P \in \mathcal{P}_{k-1}$ satisfies*

$$\|P_h(\cdot \mid s,a) - P_h^\star(\cdot \mid s,a)\|_1 \le 2\beta_{k-1,h}^{\mathrm{vls}}(\delta/H) \sum_{s'\in\mathcal{S}_{h+1}} \|\varphi_h(s' \mid s,a)\|_{\Lambda_{k-1,h}^{-1}}.$$

Proof. Fix k, h, s, a and $P \in \mathcal{P}_{k-1}$. By definition of $\mathcal{P}_{k-1}^{\mathrm{vls}}$, there exists $\theta \in \Theta_{k-1,h}^{\mathrm{vls}}(\delta/H)$ such that $P_h(s' \mid s,a) = \langle \varphi_h(s' \mid s,a),\, \theta \rangle$ for all s'. Thus

$$\begin{aligned}\|P_h(\cdot \mid s,a) - P_h^\star(\cdot \mid s,a)\|_1 &= \sum_{s' \in \mathcal{S}_{h+1}} \left| \langle \varphi_h(s' \mid s,a),\, \theta - \theta_h^\star \rangle \right| \\ &\le \sum_{s' \in \mathcal{S}_{h+1}} \|\varphi_h(s' \mid s,a)\|_{\Lambda_{k-1,h}^{-1}} \, \|\theta - \theta_h^\star\|_{\Lambda_{k-1,h}} \\ &\le \|\theta - \theta_h^\star\|_{\Lambda_{k-1,h}} \sum_{s' \in \mathcal{S}_{h+1}} \|\varphi_h(s' \mid s,a)\|_{\Lambda_{k-1,h}^{-1}} .\end{aligned}$$

The inequality in the second line uses Cauchy–Schwarz in the Mahalanobis norms: for any positive definite matrix M and vectors x, y, $|\langle x,\, y \rangle| \le \|x\|_{M^{-1}} \|y\|_M$.

Since both θ and $\theta_h^\star$ lie in the ellipsoid $\Theta_{k-1,h}^{\mathrm{vls}}(\delta/H)$,

$$\|\theta - \theta_h^\star\|_{\Lambda_{k-1,h}} \le \left\|\theta - \hat{\theta}_{k-1,h}^{\mathrm{vls}}\right\|_{\Lambda_{k-1,h}} + \left\|\theta_h^\star - \hat{\theta}_{k-1,h}^{\mathrm{vls}}\right\|_{\Lambda_{k-1,h}} \le \beta_{k-1,h}^{\mathrm{vls}}(\delta/H) + \beta_{k-1,h}^{\mathrm{vls}}(\delta/H) = 2\beta_{k-1,h}^{\mathrm{vls}}(\delta/H),$$

where the second inequality uses ellipsoid membership for θ and $\theta_h^\star$ (the latter from Lemma 12). Substituting this bound into the previous display yields the claim. □

A.12 UOB–LOB WIDTH BOUND

Define the (clipped) VLS radius

$$r_{k,h}^{\mathrm{vls}}(s,a) \triangleq 2 \;\wedge\; \left(2\beta_{k-1,h}^{\mathrm{vls}}(\delta/H) \sum_{s' \in \mathcal{S}_{h+1}} \|\varphi_h(s' \mid s,a)\|_{\Lambda_{k-1,h}^{-1}} \right).$$

Then Lemma 13 implies $\|P_h(\cdot \mid s,a) - P_h^\star(\cdot \mid s,a)\|_1 \le r_{k,h}^{\mathrm{vls}}(s,a)$.

Lemma 14 (UOB–LOB width recursion with VLS radius)**.** *Fix an episode k and policy π_k. On the event of Lemma 12, for any layer h and any state $s \in \mathcal{S}_h$,*

$$u_{k,h}(s) - \underline{u}_{k,h}(s) \le 2 \sum_{t=0}^{h-1} \sum_{s' \in \mathcal{S}_t} \sum_{a \in \mathcal{A}} q_t^{\pi_k, P^\star}(s',a)\, r_{k,t}^{\mathrm{vls}}(s',a).$$

Consequently,

$$\sum_{h=0}^{H-1} \sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} \big(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)\big) \le 2S \sum_{t=0}^{H-1} \sum_{s' \in \mathcal{S}_t} \sum_{a \in \mathcal{A}} q_t^{\pi_k, P^\star}(s',a)\, r_{k,t}^{\mathrm{vls}}(s',a).$$

Proof. The proof follows the standard occupancy-recursion argument. Fix episode k and layer h. For any kernel $P \in \mathcal{P}_{k-1}$, let $d_{k,t}^{\pi_k,P}$ and $d_{k,t}^{\pi_k,P^\star}$ be the state occupancies induced by (π_k, P) and $(\pi_k, P^\star)$. By the occupancy recursion and triangle inequality,

$$\left\| d_{k,h+1}^{\pi_k,P} - d_{k,h+1}^{\pi_k,P^\star} \right\|_1 \le \left\| d_{k,h}^{\pi_k,P} - d_{k,h}^{\pi_k,P^\star} \right\|_1 + \sum_{s,a} q_h^{\pi_k,P^\star}(s,a)\, \|P_h(\cdot \mid s,a) - P_h^\star(\cdot \mid s,a)\|_1 .$$

Iterating from $t = 0$ to $t = h-1$ yields

$$\left\| d_{k,h}^{\pi_k,P} - d_{k,h}^{\pi_k,P^\star} \right\|_1 \le \sum_{t=0}^{h-1} \sum_{s,a} q_t^{\pi_k,P^\star}(s,a)\, \|P_t(\cdot \mid s,a) - P_t^\star(\cdot \mid s,a)\|_1 .$$

On the event of Lemma 12, Lemma 13 bounds the ℓ_1 deviations by $r_{k,t}^{\mathrm{vls}}(s,a)$. Using the definitions of $u_{k,h}$ and $\underline{u}_{k,h}$ as maxima/minima over $P \in \mathcal{P}_{k-1}$, we can upper bound the envelope gap at layer h by comparing each $q_h^{\pi_k,P}$ to $q_h^{\pi_k,P^\star}$: for any state $s \in \mathcal{S}_h$, $u_{k,h}(s) - \underline{u}_{k,h}(s) \le 2 \max_{P \in \mathcal{P}_{k-1}} \left| d_{k,h}^{\pi_k,P}(s) - d_{k,h}^{\pi_k,P^\star}(s) \right| \le 2 \max_{P \in \mathcal{P}_{k-1}} \left\| d_{k,h}^{\pi_k,P} - d_{k,h}^{\pi_k,P^\star} \right\|_1$, where we used $|x(s)| \le \|x\|_1$. Combining with the previous display yields the first inequality. Summing over $s \in \mathcal{S}_h$ and then over h gives the second. □

Lemma 15 (Generalized elliptical potential)**.** *Fix a layer h and define $A_{k,h} \triangleq \sum_{s' \in \mathcal{S}_{h+1}} \|\varphi_h(s' \mid s_{k,h}, a_{k,h})\|^2_{\Lambda^{-1}_{k-1,h}}$. Then for any $K \geq 1$,*

$$\sum_{k=1}^{K} \left(1 \wedge A_{k,h}\right) \leq 2 \ln \frac{\det(\Lambda_{K,h})}{\det(\lambda I)} \leq 2d \ln \left(1 + \frac{K|\mathcal{S}_{h+1}|}{\lambda d}\right).$$

Proof. Let $X_{k,h} \in \mathbb{R}^{|\mathcal{S}_{h+1}| \times d}$ be the feature matrix at episode k whose rows are $\varphi_h(s' \mid s_{k,h}, a_{k,h})^\top$. Then $\Lambda_{k,h} = \Lambda_{k-1,h} + X_{k,h}^\top X_{k,h}$ and the matrix determinant lemma gives

$$\frac{\det(\Lambda_{k,h})}{\det(\Lambda_{k-1,h})} = \det\left(I + X_{k,h}\Lambda^{-1}_{k-1,h}X_{k,h}^\top\right).$$

Since $M \succeq 0$ implies $\det(I + M) \geq 1 + \operatorname{tr}(M)$,

$$\ln \frac{\det(\Lambda_{k,h})}{\det(\Lambda_{k-1,h})} \geq \ln\left(1 + \operatorname{tr}(X_{k,h}\Lambda^{-1}_{k-1,h}X_{k,h}^\top)\right) = \ln(1 + A_{k,h}).$$

Using $a \wedge 1 \leq 2\ln(1+a)$ for $a \geq 0$ and summing over k yields the first inequality. The second follows from $\operatorname{tr}(\Lambda_{K,h}) \leq \lambda d + K|\mathcal{S}_{h+1}|$ and AM–GM on eigenvalues. □

Lemma 16 (UOB–LOB gap for VLS confidence sets)**.** *On the event of Lemma 12, with probability at least $1-\delta$,*

$$\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}} \left(u_{k,h}(s,a) - \underline{u}_{k,h}(s,a)\right) \leq \widetilde{\mathcal{O}}\Big(d\, S^{3/2}\sqrt{H\,K}\Big).$$

Proof. Combine Lemma 14 with the shorthand

$$\mu_{k,h} \triangleq \sum_{s,a} q_h^{\pi_k, P^\star}(s,a)\, r^{\mathrm{vls}}_{k,h}(s,a) = \mathbb{E}\big[r^{\mathrm{vls}}_{k,h}(s_{k,h}, a_{k,h}) \mid \mathcal{F}_{k-1}\big].$$

Summing Lemma 14 over (k,h,s,a) yields

$$\sum_{k,h,s,a} (u_{k,h} - \underline{u}_{k,h}) \leq 2S \sum_{k=1}^{K}\sum_{h=0}^{H-1} \mu_{k,h}.$$

Step 1: replace $\sum \mu_{k,h}$ by realized radii. Let $R_{k,h} \triangleq r^{\mathrm{vls}}_{k,h}(s_{k,h}, a_{k,h}) \in [0,2]$. Define $R_k \triangleq \sum_h R_{k,h}$ and $\mu_k \triangleq \sum_h \mu_{k,h}$, and the martingale difference $X_k \triangleq \mu_k - R_k$. Then $|X_k| \leq 2H$. By Hoeffding–Azuma (or Freedman [Freedman, 1975]), with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \mu_k \leq \sum_{k=1}^{K} R_k + \widetilde{\mathcal{O}}\left(H\sqrt{K \ln \frac{1}{\delta}}\right).$$

Step 2: bound $\sum_{k,h} R_{k,h}$ via Lemma 15. Monotonicity gives $\beta^{\mathrm{vls}}_{k-1,h} \leq \beta^{\mathrm{vls}}_{K,h} =: \beta^{\mathrm{vls}}_K$. Also, for any $x \geq 0$ and $\beta \geq 1$, $2 \wedge (2\beta x) \leq 4\beta(1 \wedge x)$, hence

$$R_{k,h} \leq 4\beta^{\mathrm{vls}}_K \left(1 \wedge \sum_{s' \in \mathcal{S}_{h+1}} \|\varphi_h(s' \mid s_{k,h}, a_{k,h})\|_{\Lambda^{-1}_{k-1,h}}\right).$$

By Cauchy–Schwarz, $\sum_{s'} \|\varphi_h(s')\|_{\Lambda^{-1}} \leq \sqrt{|\mathcal{S}_{h+1}|}\,\sqrt{A_{k,h}}$, so $1 \wedge \sum_{s'} \|\varphi_h(s')\|_{\Lambda^{-1}} \leq \sqrt{|\mathcal{S}_{h+1}|}\,\sqrt{1 \wedge A_{k,h}}$. Summing over k and applying Cauchy–Schwarz again yields

$$\sum_{k=1}^{K} R_{k,h} \leq 4\beta^{\mathrm{vls}}_K \sqrt{|\mathcal{S}_{h+1}|}\,\sqrt{K \sum_{k=1}^{K} (1 \wedge A_{k,h})}.$$

Lemma 15 bounds $\sum_k (1 \wedge A_{k,h}) \leq \widetilde{\mathcal{O}}(d)$, hence $\sum_k R_{k,h} \leq \widetilde{\mathcal{O}}\big(\beta^{\mathrm{vls}}_K \sqrt{|\mathcal{S}_{h+1}|\, d\, K}\big)$. Summing over h and using $\sum_{h=0}^{H-1} |\mathcal{S}_{h+1}| \leq S$ gives $\sum_{k,h} R_{k,h} \leq \widetilde{\mathcal{O}}\big(\beta^{\mathrm{vls}}_K \sqrt{d\, S\, H\, K}\big)$. Since $\beta^{\mathrm{vls}}_K = \widetilde{\mathcal{O}}(\sqrt{d})$ up to logs, we get $\sum_{k,h} R_{k,h} \leq \widetilde{\mathcal{O}}(d\sqrt{SHK})$. Plugging into Step 1 and then into the first display yields $\sum_{k,h,s,a}(u_{k,h} - \underline{u}_{k,h}) \leq \widetilde{\mathcal{O}}(d\, S^{3/2}\sqrt{HK})$. □

A.13 PROOF OF THEOREM 1

Proof of Theorem 1. We combine the regret decomposition bounds from Lemmas 6, 3, and 5 with the dilated reduction in Lemma 8 and the bonus bound in Lemma 9. By Lemma 10, these lemmas apply to the VLS confidence sets as long as (C1)–(C3) hold. By construction, $\mathcal{P}_{k-1}$ is $\mathcal{F}_{k-1}$-measurable and (s,a)-rectangular, and by Lemma 12 it contains $P^\star$ with probability at least $1-\delta$. We work on this event.

Step 1: from decomposition to bonus values. The decomposition yields, with probability at least $1-\mathcal{O}(\delta)$,

$$\mathrm{Reg}(K) \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{k,h,s,a} d_h^\star(s)\pi_{k,h}(a\mid s)\left(\frac{9\gamma H}{u_{k,h}(s,a)+\gamma} + \frac{3}{H}B_{k,h}(s,a) + \frac{H\big(u_{k,h}(s,a)-\underline{u}_{k,h}(s,a)\big)}{u_{k,h}(s,a)+\gamma}\right) + \sum_{k,h,s,a} d_h^\star(s)\big(\pi_{k,h}-\pi_h^\star\big)B_{k,h}(s,a).$$

Using $u/(u+\gamma)\le 1$ and $\frac{u-\underline{u}}{u+\gamma}\le 1$, the bracketed terms are upper bounded by constants times $b_{k,h}(s)$ and $B_{k,h}(s,a)$, which matches the premise of Lemma 8. Applying Lemma 8 yields

$$\mathrm{Reg}(K) \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + 3\sum_{k=1}^{K} V_k^{\pi_k,P_k^b}(s_0;b_k).$$

Step 2: bound the cumulative bonus. Lemma 9 gives

$$\sum_{k=1}^{K} V_k^{\pi_k,P_k^b}(s_0;b_k) \le 3\gamma HSAK + H\sum_{k,h,s,a}(u_{k,h}-\underline{u}_{k,h}).$$

Finally, Lemma 16 bounds the width term by $\widetilde{\mathcal{O}}(dS^{3/2}\sqrt{HK})$. Putting everything together,

$$\mathrm{Reg}(K) \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \gamma HSAK + H\cdot dS^{3/2}\sqrt{HK}\right).$$

Setting $\gamma = 2\eta H$ and optimizing η yields $\widetilde{\mathcal{O}}\big(H^{3/2}\sqrt{SAK} + d\,H^{3/2}S^{3/2}\sqrt{K}\big)$, proving the theorem. □