Provably Efficient Representation Learning for Low-Rank CMDPs

Kaixuan Liu
Emory University
kaixuan.liu@emory.edu

Guojun Xiong*
Shanghai Jiao Tong University
gjxiong@sjtu.edu.cn

Shengpu Tang
Emory University
shengpu.tang@emory.edu

Wanyun Si
University of Wisconsin-Madison
wsi7@wisc.edu

Jian Li
Stony Brook University
jian.li.3@stonybrook.edu

Abstract

We study representation learning in low-rank Constrained Markov Decision Processes (CMDPs), where both value and reward functions are approximated by a set of unknown representation vectors, and the transition dynamics admit a low-rank factorization. The objective is to maximize expected cumulative rewards while complying with constraints on the expected cumulative utility. To achieve this, we propose MFRLC (Model-Free Representation Learning for Low-Rank CMDP), a model-free algorithm that uses a primal-dual scheme to effectively balance reward regret and constraint violations. To the best of our knowledge, MFRLC is the first representation-learning method for low-rank CMDPs that combines the Least-Squares Value Iteration with Upper Confidence Bound (LSVI-UCB) and primal-dual techniques. MFRLC further enhances value function estimation with bonuses for function approximation and learns the mapping from representations to value functions directly, avoiding explicit estimation of the transition kernel. We prove that, with high probability, MFRLC achieves $\widetilde{O}(K^{3/4})$ reward regret and cumulative constraint violation over K deployment rounds, independent of the size of the state space. To our knowledge, these are the first sublinear regret and violation guarantees for CMDPs with unknown features.

1 Introduction

Safe Reinforcement Learning (SRL) has been widely studied and applied in real-world settings. For example, SRL addresses collision avoidance in autonomous driving [19] and resource allocation in healthcare [28], underscoring its importance in safety-critical domains. To address these challenges, a standard framework for SRL is the Constrained Markov Decision Processes (CMDPs). In a CMDP, the agent learns a policy that maximizes expected cumulative rewards while satisfying specified underlying constraints. Designing efficient learning-based algorithms for CMDPs has attracted considerable attention. Model-based methods [10] learn an explicit CMDP model, but incur substantial storage and computational overhead in large-scale or continuous state spaces. Model-free methods avoid this: tabular variants [26, 8] store value functions directly, while recent works adopt linear function approximation (LFA) to handle large state-action spaces [11], achieving provably near-optimal $\widetilde{\mathcal{O}}(K^{1/2})$ regret and violation rates.

A central limitation of existing LFA-based methods is the assumption that the underlying features are known *a priori* [8, 11]. In practice, such features are typically unknown, rendering these

*Correspondence to Guojun Xiong <gjxiong@sjtu.edu.cn>.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison with existing work, where d denotes the feature dimension, H the horizon, K the number of deployment rounds, γ the Slater gap, and ε the target suboptimality (the learned policy's value is within ε of optimal with high probability). Rates hide polylog factors via $\widetilde{O}(\cdot)$. **LFA** denotes linear function approximation; ✗ in the LFA column indicates tabular or non-parametric methods. Bounds stated in terms of ε are PAC sample-complexity guarantees for unconstrained problems and are not directly comparable to regret.

Paper	Model Structure	LFA	Constraint	RepLearn	Regret / Sample Complexity	Violation
Jin et al. [13]	Linear	✓	✗	✗	$\widetilde{\mathcal{O}}(d^{3/2}H^2K^{1/2})$	N/A
Ghosh et al. [11]	Linear	✓	✓	✗	$\widetilde{\mathcal{O}}(d^{3/2}H^2K^{1/2}/\gamma)$	$\widetilde{\mathcal{O}}(d^{3/2}H^2K^{1/2}/\gamma)$
Ding and Lavaei [9]	Tabular	✗	✓	✗	$\widetilde{\mathcal{O}}(d^{2/3}H^{5/2}K^{2/3}/\gamma)$	$\widetilde{\mathcal{O}}(d^{2/3}H^{5/2}K^{2/3}/\gamma)$
Zhang et al. [29]	Block	✓	✗	✓	$\widetilde{\mathcal{O}}\big(d^8H^9\,\varepsilon^{-4}\big)$	N/A
Modi et al. [17]	Low-rank	✓	✗	✓	$\widetilde{\mathcal{O}}\big(d^7H^8\,\varepsilon^{-2}\big)$	N/A
Mhammedi et al. [15]	Low-rank	✓	✗	✓	$\widetilde{\mathcal{O}}\big(d^{13}\lvert\mathcal{A}\rvert^2H^6\,\varepsilon^{-2}\big)$	N/A
This Work	**Low-rank**	✓	✓	✓	$\widetilde{\mathcal{O}}(d^2H^3K^{3/4}\lvert\mathcal{A}\rvert^{3/2}/\gamma)$	$\widetilde{\mathcal{O}}(d^2H^3K^{3/4}\lvert\mathcal{A}\rvert^{3/2}/\gamma)$

methods inapplicable. Meanwhile, the low-rank MDP framework [3] enables representation learning without prior feature knowledge, and provably efficient algorithms exist for the *unconstrained* setting [29, 17, 15]. However, extending representation learning to CMDPs is non-trivial. In the unconstrained setting, representation learning and policy optimization are largely decoupled; in a CMDP, the learned representation must simultaneously support both reward maximization and constraint satisfaction, and errors in the representation propagate into both objectives in a tightly coupled manner that standard primal-dual analyses cannot accommodate.

In this work, we bring representation learning to low-rank CMDPs and propose `MFRLC`, which in each round first learns representations via a minimax objective and then fits the value function using the learned features within a primal-dual LSVI-UCB framework. This constitutes the first approach for online low-rank CMDPs with unknown features. We prove that `MFRLC` achieves the same bound for both cumulative reward regret and cumulative constraint violation: $\widetilde{\mathcal{O}}(H^3d^2K^{3/4}|\mathcal{A}|^{3/2}/\gamma)$, where K is the number of deployment rounds, H is the horizon, d is the representation dimension, and γ is the Slater gap. The $|\mathcal{A}|^{3/2}$ factor arises from uniform exploration over actions in the representation learning step, and the $K^{3/4}$ dependence arises from the accumulated statistical error of representation learning (Remark 8). Both bounds are upper bounds. No matching lower bound is known for online low-rank MDPs with unknown features, with or without constraints, so we do not claim that the $K^{3/4}$ rate or the dependence on d, H, and $|\mathcal{A}|$ is optimal. Known-feature CMDPs admit $\widetilde{\mathcal{O}}(K^{1/2})$ regret and violation [11]; whether this rate is attainable with unknown features remains open.

2 Problem Formulation

2.1 CMDP Preliminaries

We consider an episodic CMDP $\langle \mathcal{S}, \mathcal{A}, H, r, g, \mathbb{P}, d_0\rangle$, where $\mathcal{S}$ is the state space (possibly infinite or continuous), $\mathcal{A}$ the action space (finite, with cardinality $|\mathcal{A}|$), H the fixed episode length, $r = \{r_h\}_{h=1}^H$ and $g = \{g_h\}_{h=1}^H$ the collections of reward and utility functions, respectively, with $r_h, g_h : \mathcal{S}\times\mathcal{A}\to[0,1]$ assumed to be deterministic, $\mathbb{P} = \{P_h\}_{h=1}^H$ the collection of transition kernels with $P_h(s_{h+1} \mid s_h, a_h) : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ being the probability of moving to s_{h+1} from s_h under a_h, and $s_1 \sim d_0$ the initial state of each episode. At each step $h \in [H]$ of round $k \in [K]$, the agent takes action a_h^k at state s_h^k, receives reward $r_h(s_h^k, a_h^k)$ and utility $g_h(s_h^k, a_h^k)$, and transitions to $s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k)$; each episode terminates after H steps. The policy space is $\Delta(\mathcal{A} \mid \mathcal{S}, H) := \{\{\pi_h(\cdot \mid \cdot)\}_{h=1}^H : \pi_h(\cdot \mid s) \in \Delta(\mathcal{A}), \forall s \in \mathcal{S}, h \in [H]\}$, where $\Delta(\mathcal{A})$ is the probability simplex over the action space; $\pi^k \in \Delta(\mathcal{A} \mid \mathcal{S}, H)$ denotes the policy deployed in round k, with $\pi_h^k(\cdot \mid s_h^k) : \mathcal{S}\to\Delta(\mathcal{A})$ specifying the action distribution at s_h^k. Given $\pi = \{\pi_h\}_{h=1}^H$, the value function (expected cumulative reward from $s_h = s$ under π) and action-value function (from $(s_h = s, a_h = a)$) are:

$$V_{r,h}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{i=h}^{H} r_i(s_i,a_i)\,\middle|\, s_h = s\right], \qquad Q_{r,h}^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{i=h}^{H} r_i(s_i,a_i)\,\middle|\, s_h = s,\, a_h = a\right],$$

where the expectation is over $a_i \sim \pi_i(\cdot \mid s_i)$ and $s_{i+1} \sim P_i(\cdot \mid s_i, a_i)$. The utility value and action-value functions $V^{\pi}_{g,h}(s)$, $Q^{\pi}_{g,h}(s,a)$ are defined analogously by replacing r with g; for brevity, we write $V^{\pi}_{j,h}(s)$ and $Q^{\pi}_{j,h}(s,a)$ for $j \in \{r, g\}$.

We say a policy π is *feasible* if $V^{\pi}_{g,1}(s_1) \geq b$. The agent aims to learn the optimal feasible policy $\pi^\star$ by solving the following optimization problem:

$$\max_{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)} V^{\pi}_{r,1}(s_1), \quad \text{subject to} \quad V^{\pi}_{g,1}(s_1) \geq b, \tag{1}$$

where the threshold $b \in [0, H]$. Under the online learning setting of an episodic CMDP, the agent interacts with the environment over a sequence of rounds, and the learning performance is measured by regret and constraint violation as follows:

$$\text{Regret}(K) = \sum_{k=1}^{K} \left(V^{\pi^\star}_{r,1}(s_1) - V^{\pi^k}_{r,1}(s_1) \right), \quad \text{Violation}(K) := \Big(\sum_{k=1}^{K} \left(b - V^{\pi^k}_{g,1}(s_1) \right) \Big)_+,$$

with π^k being the policy deployed in round k. Both quantities are random variables depending on the algorithm's internal randomness and the sampled trajectories; our guarantees on them hold with high probability (see Theorem 3). The regret quantifies the difference in total rewards achieved by the agent's policy π^k compared to the optimal feasible policy $\pi^\star$ across K rounds, and the constraint violation measures the cumulative shortfall of utility against the predefined threshold b over K rounds.

Access model and interaction accounting. We call each index k a *deployment round*. In round k, `MFRLC` executes π^k for one trajectory and also collects two auxiliary transitions at every step h for representation learning (Algorithm 1, line 4), each obtained by rolling in with π^{k-1} to step h and taking a uniform action, so no generative model is required. Regret and Violation contain one term per deployed policy and do not charge the auxiliary roll-ins; our guarantees are deployment-regret guarantees with total interaction complexity $K(2H+1)$ trajectories and $K(H^2+2H)$ transitions.

2.2 Low-Rank MDP and Representation Learning

Conventional work on learning MDPs typically assumes a general linear MDP framework [13, 14, 5], where features are considered known, and function approximation methods are used to maximize rewards without considering constraints. While this linear MDP assumption has been effectively extended to CMDPs [4, 12], the transition to low-rank MDPs introduces additional layers of complexity, defined as follows.

Definition 1 (Low-rank MDP [3])**.** Consider a transition kernel $P^\star_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ which can be decomposed into low-rank components of dimension $d \in \mathbb{N}$. Specifically, there exist functions $\phi^\star = \{\phi^\star_h\}_{h=1}^H$ and $\mu^\star = \{\mu^\star_h\}_{h=1}^H$ with $\phi^\star_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu^\star_h : \mathcal{S} \to \mathbb{R}^d$, such that for all $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in [H]$:

$$P^\star_h(s' \mid s, a) = \langle \phi^\star_h(s,a), \mu^\star_h(s') \rangle.$$

Furthermore, there exist vectors $\theta^\star_{r,h}, \theta^\star_{g,h} \in \mathbb{R}^d$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$r_h(s,a) = \langle \phi^\star_h(s,a), \theta^\star_{r,h} \rangle, \quad g_h(s,a) = \langle \phi^\star_h(s,a), \theta^\star_{g,h} \rangle.$$

The feature function satisfies $\|\phi^\star_h(s,a)\|_2 \leq 1$ for all state-action pairs (s,a), and for any function $f : \mathcal{S} \to [0,1]$, $\left\| \int \mu^\star_h(s) f(s) \, \mathrm{d}s \right\|_2 \leq \sqrt{d}$. An MDP $\mathcal{M}$ is considered low-rank if the transition kernel P_h at each time step $h \in [H]$ admits such a decomposition.

Under the low-rank setting, the feature representations $\phi^\star$ are no longer presumed known. This necessitates the development of algorithms that are not only capable of learning these features but also ensure that the resulting policy adheres to the constraints imposed by the CMDP framework.

Assumption 1 (Realizable feature class)**.** *The learner has access to a finite feature class $\Phi = \Phi_1 \times \cdots \times \Phi_H$, where each $\phi_h \in \Phi_h$ maps $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}^d$ with $\|\phi_h(s,a)\|_2 \leq 1$ for all (s,a), and Φ contains the true representation of Definition 1, that is, $\phi^\star \in \Phi$.*

Assumption 1 is the standard setting of model-free representation learning in low-rank MDPs [3, 17]. Finiteness enters only through $\log|\Phi|$ and can be replaced by a covering number, while realizability

Algorithm 1 Model-Free Representation Learning with Low-rank CMDP (MFRLC)

Require: Feature classes $\{\Phi_h\}_{h=1}^H$, regularization λ_k, Slater parameter $\gamma > 0$
Output: Sequence of policies $\{\pi^k\}_{k=1}^K$

1: **Initialize:** $Y_1 = 0$, $\xi = 2H/\gamma$, $\eta = \xi/\sqrt{KH^2}$; set π^0 to any reference policy and replay buffers $\mathcal{D}_h = \varnothing$, $\mathcal{D}'_h = \varnothing$ for all $h \in [H]$.
2: **for** $k = 1$ **to** K **do**
3: **for** $h = 1$ **to** H **do** ▷ Exploration: branching from $d_h^{\pi^{k-1}}$
4: Sample $s_h^k \sim d_h^{\pi^{k-1}}$, $\tilde{a}_h^k \sim U(\mathcal{A})$, $s_{h+1}^k \sim P_h^\star(\cdot \mid s_h^k, \tilde{a}_h^k)$, $\tilde{s} \sim d_h^{\pi^{k-1}}$, $\tilde{a} \sim U(\mathcal{A})$, $\tilde{s}' \sim P_h^\star(\cdot \mid \tilde{s}, \tilde{a})$
5: $\mathcal{D}_h \leftarrow \mathcal{D}_h \cup \{(s_h^k, \tilde{a}_h^k, s_{h+1}^k)\}$, $\mathcal{D}'_h \leftarrow \mathcal{D}'_h \cup \{(\tilde{s}, \tilde{a}, \tilde{s}')\}$
6: $\hat{\phi}_h^k \leftarrow \textsc{RepLearn}\big(\mathcal{D}_h, \mathcal{D}'_h, \Phi_h, \lambda_k\big)$ ▷ Algorithm 2
7: **end for**
8: $\big\{Q_{j,h}^k, V_{j,h}^k\big\}_{j\in\{r,g\},h\in[H]}, \{\pi_h^k\}_{h=1}^H \leftarrow \text{LSVI-UCB}\big(\big\{(s_h^\tau, \tilde{a}_h^\tau, r_h^\tau, g_h^\tau, s_{h+1}^\tau, \hat{\phi}_h^\tau)\big\}_{h\in[H],\,\tau\le k}, Y_k\big)$
9: **for** $h = 1$ **to** H **do** ▷ Execute π^k for next-episode trunk
10: Take $a_h^k \sim \pi_h^k(\cdot \mid s_h^k)$ and observe s_{h+1}^k
11: **end for**
12: $Y_{k+1} \leftarrow \max\big\{\min\big\{Y_k + \eta\,(b - V_{g,1}^k(s_1)),\, \xi\big\},\, 0\big\}$ ▷ Primal–dual update
13: **end for**

is exact: misspecified classes with $\phi^\star \notin \Phi$ are outside our guarantees. Requiring the utility, like the reward, to be linear in $\phi^\star$ is the natural CMDP counterpart of the low-rank MDP assumption; it is structural and does not cover reward or utility functions unrelated to the low-rank dynamics. We do not assume any map from representations to policies recovers $\pi^\star$: the analysis compares directly against an arbitrary optimal feasible policy of (1). Importantly, we adopt a model-free approach in which the policy is learned without explicitly modeling the environment, and there is no need to recover the exact values of $\mu^\star$ or $\theta^\star$.

3 Representation Learning for Low-Rank CMDPs

In this section, we address the challenge of model-free representation learning within low-rank CMDPs. Our algorithm, MFRLC (Algorithm 1), builds on the Upper Confidence Bound (UCB) variant of Least-Squares Value Iteration (LSVI) and utilizes a primal–dual approach to effectively manage the CMDP's optimization requirements.

Before delving into the details of the algorithm, it is helpful to introduce the *composite* (Lagrangian) value function, which encodes the reward–constraint trade-off into a single scalar objective.

Definition 2 (Composite value function; cf. 20, 8)**.** For any dual variable $Y \ge 0$ and policy π, define the composite value function and action-value function as

$$V_h^{\pi,Y}(\cdot) = V_{r,h}^\pi(\cdot) + Y\, V_{g,h}^\pi(\cdot), \quad Q_h^{\pi,Y}(s,a) = Q_{r,h}^\pi(s,a) + Y\, Q_{g,h}^\pi(s,a).$$

Equivalently, $V_h^{\pi,Y}$ and $Q_h^{\pi,Y}$ are the value and action-value functions of policy π in the auxiliary MDP whose stage reward is the Lagrangian $r_h + Y g_h$.

We operate on the composite values $(V_h^{\pi,Y}, Q_h^{\pi,Y})$ throughout the analysis, with Y instantiated to the dual iterate Y_k at round k. Algorithm 3 adds the same per-objective bonus $\hat{b}_h$ to both $Q_{r,h}^k$ and $Q_{g,h}^k$, so the composite optimism bonus is $\hat{b}_h^{(k)} := (1 + Y_k)\hat{b}_h$.

3.1 Algorithm Overview

MFRLC alternates three steps in each round. **Representation Learning** (lines 4–6) samples under $d_h^{\pi^{k-1}}$ with kernel $P_h^\star$, stores data in replay buffers $\mathcal{D}_h, \mathcal{D}'_h$, and learns $\hat{\phi}_h^k \in \Phi_h$ via a minimax game against a discriminator $v \in \mathcal{V}_h$, controlled by the regularizer λ_k, iteration cap M, and stopping threshold ℓ_k. **Policy Update** (line 8) takes $\hat{\phi}^k$ and Y_k and runs LSVI-UCB (Algorithm 3, Appendix A)

Algorithm 2 Representation Learning (REPLEARN)

Require: Feature class Φ_h, datasets $\mathcal{D}_h, \mathcal{D}'_h$, discriminator class $\mathcal{V}_h$, regularization λ_k, iterations M, threshold ℓ_k, confidence δ, dimension d.

Output: Representation $\hat{\phi}_h^k$

1: **Parameters:** $\lambda_k = O\Big(d \ln \frac{|\Phi|k}{\delta}\Big)$, $\ell_k = O\Bigg(d^2 \sqrt{\frac{\ln(|\Phi|/\delta)}{k}}\Bigg)$

2: **Notation:** $\mathcal{L}_{\lambda_k,\mathcal{D}}(\phi, w, v)$ is the empirical objective (6) over buffer $\mathcal{D}$; w (and $\tilde{w}$) are predictor weights for feature ϕ; v (v^m) $\in \mathcal{V}_h$ is the discriminator (at step m).

3: **Initialize:** $\phi^0 \in \Phi_h$ arbitrarily

4: **for** $m = 0$ **to** $M - 1$ **do**

5: Compute worst-case discriminator with bias correction:

6: $v^m \leftarrow \underset{v \in \mathcal{V}_h,\, \tilde{\phi} \in \Phi_h}{\arg\max} \Big[\min_{w} \mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi^m, w, v) - \min_{\tilde{w}} \mathcal{L}_{\lambda_k,\mathcal{D}'_h}(\tilde{\phi}, \tilde{w}, v) \Big]$

7: **if** the maximization objective at v^m is at most ℓ_k **then**

8: **Break** and **Return** ϕ^m

9: **end if**

10: Update representation jointly against past discriminators:

11: $\phi^{m+1} \leftarrow \underset{\phi \in \Phi_h,\, \{w_i\}_{i=0}^m}{\arg\min} \sum_{i=0}^{m} \mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi, w_i, v^i)$

12: **end for**

to produce optimistic per-objective estimates $\{Q_{r,h}^k, Q_{g,h}^k, V_{r,h}^k, V_{g,h}^k\}$ and the softmax policy

$$\pi_h^k(a \mid \cdot) \;=\; \mathrm{softmax}_\alpha\big(Q_{r,h}^k(\cdot, a) + Y_k\, Q_{g,h}^k(\cdot, a)\big)\,, \tag{2}$$

where Y_k enforces the constraint while reward is maximized. **Primal–Dual Update** (line 12) updates Y_k using the optimistic value $V_{g,1}^k(s_1)$ in place of $V_{g,1}^{\pi^k}(s_1)$, based on the observed constraint slack, and carries it into the next round.

Representation Learning In the context of low-rank MDPs, where $\phi^\star$ is unknown, we must learn this representation. Inspired by Zhang et al. [29], we define $\mathcal{V}_h = \mathcal{F}_h \cup \mathcal{G}_h$, where $\mathcal{F}_h$ captures the model-mismatch between $\hat{P}$ and $P^\star$ under the previous-episode policy π^{k-1} (Lemma 2) and $\mathcal{G}_h$ captures the composite Q-function (Lemma 3):

$$\mathcal{F}_h := \Big\{ s \mapsto \mathbb{E}_{a \sim \pi_h^{k-1}(\cdot|s)}[\langle \phi(s,a), \theta \rangle - \langle \phi'(s,a), \theta' \rangle] : \phi, \phi' \in \Phi_{h+1},\ \max(\|\theta\|_\infty, \|\theta'\|_\infty) \le 1 \Big\}, \tag{3}$$

$$\mathcal{G}_h := \left\{ s \mapsto \max_{a \in \mathcal{A}} \left(\frac{r_{h+1}(s,a) + g_{h+1}(s,a) + \min\{\langle \phi(s,a), w \rangle, 2\}}{2H+1} + \langle \phi(s,a), w' \rangle \right) \right\}, \tag{4}$$

where the set ranges over $\phi \in \Phi_{h+1}$, $\|w\|_\infty \le c$, $\|w'\|_\infty \le 1$. Reward and utility enter the representation objective through $\mathcal{G}_h$, which contains $r_{h+1} + g_{h+1}$; our guarantee requires no separate supervised loss on the immediate reward or utility.

To learn $\phi^\star$, we solve the following optimization problem, where ρ_h^k denotes the sampling distribution induced by the replay buffer $\mathcal{D}_h$ at the start of episode k:

$$\arg\min_{\phi_h^k} \max_{v \in \mathcal{V}_h} \min_{w} \mathbb{E}_{(s,\tilde{a}) \sim \rho_h^k} \Big[\big(\langle \phi_h^k(s,\tilde{a}), w \rangle - \mathbb{E}_{s' \sim P_h^\star(\cdot|s,\tilde{a})}[v(s')] \big)^2 \Big]. \tag{5}$$

This objective seeks a feature ϕ that accurately approximates the Bellman backup $\mathbb{E}[v(s')|s,\tilde{a}]$ as a linear function $\langle \phi(s,\tilde{a}), w \rangle$, uniformly over $v \in \mathcal{V}_h$. The min-max-min structure ensures robustness: the inner $\min_w$ is a regularized least-squares problem with a closed-form solution (Appendix C, complexity $\mathcal{O}(nd^2 + d^3)$); the middle $\max_v$ identifies the worst-case $v \in \mathcal{V}_h$ for the current ϕ; the outer $\min_{\phi_h^k}$ chooses the representation minimizing this worst-case error [17]. The non-linear softmax operator and UCB bonus appearing in the LSVI-UCB output (Algorithm 3) are not contained in $\mathcal{G}_h$; their contribution is handled separately via the softmax sub-optimality bound (Lemma 12) and the elliptical-potential argument applied to the bonus term (Appendix F).

Since (5) involves a population expectation, direct implementation is challenging. We replace it by its sample-based counterpart over $\mathcal{D}_h$:

$$\mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi_h^k, v, w) = \mathbb{E}_{\mathcal{D}_h}\Big[\big(\langle\phi_h^k(s,\tilde{a}), w\rangle - v(s')\big)^2\Big] + \lambda_k\|w\|_2^2. \tag{6}$$

Because $\mathcal{L}_{\lambda_k,\mathcal{D}_h}$ contains a conditional-variance term that depends on v, maximizing it directly would favor high-variance discriminators. Algorithm 2 cancels this term in expectation by subtracting the minimized loss $\mathcal{L}_{\lambda_k,\mathcal{D}'_h}(\tilde{\phi}, \tilde{w}, v)$ on the independent buffer $\mathcal{D}'_h$ (Appendix C).

From theory to implementation We distinguish three levels of the representation-learning step:

objective (5) $\longrightarrow$ bias-corrected RepLearn (Algorithm 2) $\longrightarrow$ contrastive EMA surrogate.

Theorem 3 covers the first two levels. The experiments in Section 5 use the third: the minimax search over $\mathcal{V}_h$ is replaced by contrastive next-state prediction with an exponential-moving-average (EMA) target encoder. The surrogate uses the same transition triples $(s, \tilde{a}, s')$ to learn predictive features, but it does not perform the worst-case discriminator search and does not certify uniform Bellman prediction over $\mathcal{V}_h$, so it is not covered by Theorem 3.

Computation The inner minimization over w is ridge regression with cost $O(nd^2 + d^3)$; the outer problems over ϕ and v are treated as optimization oracles, exactly solvable for finite classes but without a polynomial-time guarantee for neural classes (Appendix C.5).

Policy Update Given $\hat{\phi}^k$ and Y_k, the LSVI-UCB subroutine (Algorithm 3, Appendix A) computes optimistic estimates $Q_{j,h}^k, V_{j,h}^k$ for $j \in \{r, g\}$ and the softmax policy (2). The executed policy π^k is defined directly as this optimistic softmax policy computed from $\hat{\phi}^k$ and Y_k. The composite $Q_{r,h}^k + Y_k Q_{g,h}^k$ is the LSVI-UCB optimistic estimate of the Bellman-optimal Q for the auxiliary MDP with stage reward $r_h + Y_k g_h$ (cf. Definition 2), with Y_k updated online in the next step.

Dual Update Problem (1) admits the Lagrangian formulation:

$$\max_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} \min_{Y\in[0,\xi]} \; V_{r,1}^{\pi}(s_1) + Y\big(V_{g,1}^{\pi}(s_1) - b\big), \tag{7}$$

where $\xi = 2H/\gamma$ is chosen so that $Y^\star \in [0, \xi]$ (Lemma 1); restricting Y to $[0, \xi]$ does not affect the saddle-point value. `MFRLC` does not solve (7) over π directly. In round k, the primal step is the optimistic LSVI-UCB policy π^k for the Lagrangian reward $r_h + Y_k g_h$, and the dual step is the projected update:

$$Y_{k+1} \;=\; \max\big\{\min\big\{Y_k + \eta\,(b - V_{g,1}^k(s_1)),\, \xi\big\},\, 0\big\}, \tag{8}$$

which uses the optimistic estimate $V_{g,1}^k(s_1)$ in place of $V_{g,1}^{\pi^k}(s_1)$. The update direction switches with the sign of $b - V_{g,1}^k(s_1)$, balancing reward maximization against feasibility [10, 18].

4 Main Results

Assumption 2 (Slater's Condition)**.** *There exists $\gamma > 0$ and a policy $\bar{\pi} \in \Delta(\mathcal{A} \mid \mathcal{S}, H)$ such that $V_{g,1}^{\bar{\pi}}(s_1) \geq b + \gamma$.*

Lemma 1 (Boundedness of $Y^\star$)**.** *The optimal dual variable $Y^\star$ associated with (1) satisfies $Y^\star \leq (V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\bar{\pi}}(s_1))/\gamma \leq H/\gamma$, so $\xi = 2H/\gamma \geq 2Y^\star$ as required by Lemma 21.*

The Slater condition is mild and widely adopted [20, 8, 10].

Theorem 3. *Fix $p \in (0, 1)$. Under Assumptions 1–2, with probability at least $1 - p$, the regret and constraint violation of `MFRLC` over K deployment rounds both satisfy:*

$$\mathrm{Regret}(K),\ \mathrm{Violation}(K) \;\leq\; \frac{C}{\gamma}\, H^3 d^2 K^{3/4} |\mathcal{A}|^{3/2}\, \log^{1/4}\!\left(\frac{dK|\Phi|}{p}\right),$$

for an absolute constant $C > 0$. Over the K rounds, `MFRLC` uses $K(2H + 1)$ trajectories in total, including the auxiliary roll-ins of representation learning, whose rewards and utilities are not charged to Regret or Violation (Section 2).

The dependence on $1/\gamma$ is the standard Slater-gap penalty of primal–dual CMDP analyses [7, 8, 11]. The bound is independent of $|\mathcal{S}|$ and applies to continuous state spaces. The $K^{3/4}$ term arises only from the representation-learning error (Remark 8 in Appendix H). Theorem 3 is an upper bound: we prove no matching lower bound, and the optimal dependence on K, d, H, and $|\mathcal{A}|$ remains open.

4.1 Proof Sketch

Step 1: Decomposition of regret and violation. For any $Y \in [0, \xi]$, by combining a Lagrangian shift with the projected dual update of Algorithm 1 (Appendix E), we obtain:

$$\begin{aligned}\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big) \;\le\; \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2} \\ + \underbrace{\sum_{k=1}^{K}\Big[\big(V_{r,1}^{\pi^\star}(s_1) + Y_k V_{g,1}^{\pi^\star}(s_1)\big) - \big(V_{r,1}^{k}(s_1) + Y_k V_{g,1}^{k}(s_1)\big)\Big]}_{\text{Pseudo regret}} \\ + \underbrace{\sum_{k=1}^{K}\Big[\big(V_{r,1}^{k}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\big(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi^k}(s_1)\big)\Big]}_{\text{Estimation error}}.\end{aligned} \tag{9}$$

Here Y is a free variable that combines reward regret and constraint violation: setting $Y = 0$ recovers the regret bound, while $Y = \xi$ (combined with strong duality) yields the violation bound. By feasibility of $\pi^\star$, the Lagrangian-shift slack term $\sum_k Y_k(b - V_{g,1}^{\pi^\star}(s_1)) \le 0$ is non-positive and dropped. The Pseudo regret captures the gap between $\pi^\star$ and the LSVI-UCB optimistic estimate V^k, while the Estimation error captures the gap between V^k and the executed policy's value V^{π^k}. We control the two terms separately.

Step 2: Bounding the Pseudo regret. Since π^k is the softmax of the composite $Q_{r,h}^k + Y_k Q_{g,h}^k$, the analysis must be carried out on the auxiliary MDP with stage reward $\bar{r}_h^{(k)}$. We first establish composite optimism on the planning model $\hat{P}^k$ induced by $\hat{\phi}^k$.

Lemma 2 (Composite optimism)**.** *Conditioned on the Representation Learning guarantee (Lemma 16), with probability $1-\delta$, for every round $k \in [K]$,*

$$Q^{\pi^\star, Y_k}_{\hat{P}^k,\, \bar{r}^{(k)} + \hat{b}^{(k)}}(s_1, \cdot) \;-\; Q^{\pi^\star, Y_k}_{P^\star,\, \bar{r}^{(k)}}(s_1, \cdot) \;\ge\; -(1 + Y_k)\sqrt{|\mathcal{A}|\zeta_k},$$

where $\zeta_k = \widetilde{O}(d^2/\sqrt{k})$ is the RepLearn statistical error.

Lemma 3 (Composite policy gap, cumulative form)**.** *Conditioned on the Representation Learning guarantee, with probability $1-\delta$,*

$$\sum_{k=1}^{K}\Big[Q^{\pi^\star, Y_k}_{P^\star,\, \bar{r}^{(k)}}(s_1, \cdot) - Q^{\pi^k, Y_k}_{P^\star,\, \bar{r}^{(k)}}(s_1, \cdot)\Big] \;\le\; \frac{KH\log|\mathcal{A}|}{\alpha} + (1+\xi)\sum_{k=1}^{K}\epsilon_k,$$

where the cumulative representation error satisfies $\sum_{k=1}^{K}\epsilon_k \le C\, H^2 d^2 K^{3/4} |\mathcal{A}|^{3/2} \log^{1/4}(dK|\Phi|/\delta)$.

Remark 1 (On the per-episode rate)**.** *The cumulative bound on $\sum_k \epsilon_k$ corresponds to an average per-episode rate $\bar{\epsilon}_K = \frac{1}{K}\sum_k \epsilon_k = \widetilde{O}(K^{-1/4})$. We state the cumulative form because the elliptical-potential argument in Appendix F directly yields a sum, and only this sum is needed downstream.*

The proof of Lemma 3 combines composite optimism (Lemma 2), a softmax sub-optimality bound for the composite planning problem, and the simulation lemma. The $(1 + Y_k)$ factor arises because both $\bar{r}_h^{(k)} \in [0, 1 + Y_k]$ and $\hat{b}_h^{(k)} = (1 + Y_k)\hat{b}_h$ scale linearly in $1 + Y_k$; using $Y_k \le \xi$, this factor is bounded by $1 + \xi$, contributing the $1/\gamma$ dependence.

Combining Lemma 3 with the V–Q identity $V_{j,1}^k(s_1) = \sum_a \pi_1^k(a \mid s_1) Q_{j,1}^k(s_1, a)$ and a performance-difference argument on $P^\star$ (Appendix F):

Lemma 4 (Pseudo regret)**.** *Conditioned on the Representation Learning guarantee, with probability* $1 - p/4$*,*

$$\sum_{k=1}^{K}\Big[\big(V_{r,1}^{\pi^\star}(s_1) + Y_k V_{g,1}^{\pi^\star}(s_1)\big) - \big(V_{r,1}^{k}(s_1) + Y_k V_{g,1}^{k}(s_1)\big)\Big] \;\le\; \frac{KH\log|\mathcal{A}|}{\alpha} + (1+\xi)\sum_{k=1}^{K}\epsilon_k.$$

Step 3: Bounding the Estimation error. The estimation error compares the LSVI-UCB optimistic estimate $V_{j,1}^{k}$ to the true value $V_{j,1}^{\pi^k}$ of the executed policy. The starting point is a self-normalized concentration bound for the empirical Bellman residual.

Lemma 5 (Concentration of empirical Bellman residual)**.** *Adapting the self-normalized concentration of Ghosh et al. [11] to the learned-feature setting via a covering argument over the optimistic value-function class, there exists a constant* C_2 *such that for any* $p \in (0,1)$*, the event:*

$$\mathcal{E} := \bigcap_{k\in[K],\, h\in[H],\, j\in\{r,g\}} \left\{ \Big\| \sum_{\tau=1}^{k-1} \hat{\phi}_h^{\tau}\big(V_{j,h+1}^{k}(s_{h+1}^{\tau}) - \mathbb{P}_h V_{j,h+1}^{k}(s_h^{\tau},\tilde{a}_h^{\tau})\big)\Big\|_{(\Sigma_h^k)^{-1}} \le C_2\, dH\sqrt{\chi} \right\}$$

holds with $\chi = \log\big(4(C_1+1)\log|\mathcal{A}|\, dT|\Phi|/p\big)$*, and* $\Pr(\mathcal{E}) \ge 1 - p/4$*.*

Remark 2 (From fixed to learned features)**.** *Ghosh et al. [11] states this concentration for a fixed feature map. Here both* $\hat{\phi}_h^{\tau}$ *and* $V_{j,h+1}^{k}$ *are data-dependent, so we cover the optimistic value-function class (parameterized by* $\hat{\phi} \in \Phi$ *and bounded LSVI weights), which contributes the* $\log|\Phi|$ *factor inside* χ*. The detailed argument is deferred to Appendix G.*

We then control the per-step gap between $Q_{j,h}^{k}$ (the LSVI-UCB optimistic estimate) and $Q_{j,h}^{\pi^k}$ (the true action-value of the executed policy). The key technical step is showing that both quantities admit linear approximations in the learned features $\hat{\phi}^k$: $Q_{j,h}^{k}$ by construction, since the LSVI-UCB output is explicitly linear in $\hat{\phi}^k$; and $Q_{j,h}^{\pi^k}$ via the RepLearn guarantee, which ensures $\hat{\phi}^k$ approximately spans the Bellman backups of any function in $\mathcal{V}_h$.

Lemma 6 (Per-step backup error)**.** *Set* $\beta = C_1\, dH\sqrt{\iota}$ *with* $\iota = \log(\log|\mathcal{A}| \cdot 4dT|\Phi|/p)$*. On the event* $\mathcal{E}$*, for every round* $k \in [K]$*, step* $h \in [H]$*,* $j \in \{r,g\}$*, and* $(s,a) \in \mathcal{S}\times\mathcal{A}$*, there exist weights* $w_{j,h}^{\pi^k}$ *approximately projecting* $Q_{j,h}^{\pi^k}$ *onto* $\mathrm{span}(\hat{\phi}_h^k)$ *such that:*

$$Q_{j,h}^{k}(s,a) - Q_{j,h}^{\pi^k}(s,a) = \mathbb{E}_{s'\sim P_h^\star(\cdot|s,a)}\Big[V_{j,h+1}^{k}(s') - V_{j,h+1}^{\pi^k}(s')\Big] + \Delta_h^k(s,a),$$

where $|\Delta_h^k(s,a)| \le \beta\sqrt{\hat{\phi}_h^k(s,a)^\top(\Sigma_h^k)^{-1}\hat{\phi}_h^k(s,a)}$ *plus a RepLearn projection residual that accumulates into* $\sum_k \epsilon_k$*.*

Unrolling Lemma 6 via $V_{j,h}^{k}(s) = \sum_a \pi_h^k(a \mid s) Q_{j,h}^{k}(s,a)$, summing telescopically over h, and applying the elliptical potential lemma over k, we obtain:

Lemma 7 (Estimation error)**.** *With probability at least* $1 - p/2$*,*

$$\sum_{k=1}^{K}\big(V_{r,1}^{k}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\sum_{k=1}^{K}\big(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi^k}(s_1)\big) \;\le\; O\Big((Y+1)\sqrt{d^3H^3T\iota^2}\Big) + (1+\xi)\sum_{k=1}^{K}\epsilon_k,$$

where $T = KH$*.*

Step 4: Combining Pseudo regret and Estimation error. Plugging Lemmas 4 and 7 into (9) with $\eta = \xi/\sqrt{KH^2}$ and $\alpha = \frac{K\log|\mathcal{A}|}{2(1+\xi+H)}$:

$$\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big) \le \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2} + \frac{KH\log|\mathcal{A}|}{\alpha}$$
$$+ 2(1+\xi)\sum_{k=1}^{K}\epsilon_k + O\Big((Y+1)\sqrt{d^3H^3T\iota^2}\Big).$$

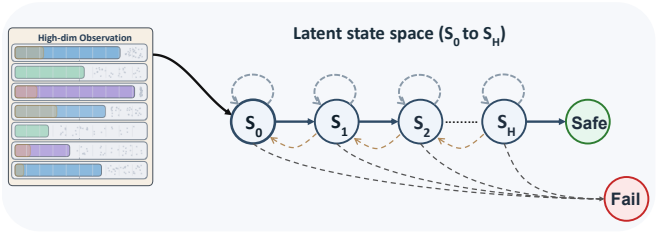

Figure 1: Constrained Latent ComboLock

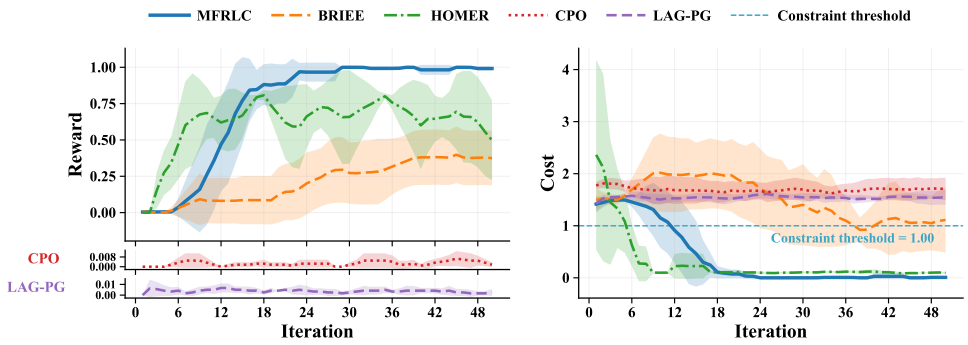

Figure 2: Reward and Cost vs. iteration on CLCL (H=6)

For these parameter choices, $\eta H^2 K/2$, $KH \log|\mathcal{A}|/\alpha$, and the estimation-error term are all dominated by $2(1+\xi)\sum_k \epsilon_k$, which evaluates to $\widetilde{O}\left(\frac{H^3 d^2 K^{3/4}|\mathcal{A}|^{3/2}}{\gamma}\right)$. Setting $Y = 0$ yields the regret bound. Setting $Y = \xi$ and applying Lemma 21 (strong-duality conversion of Ding et al. 8) with $\xi \geq 2Y^\star$ converts this into the cumulative violation bound, matching Theorem 3.

5 Experiments

We introduce the **Constrained Latent ComboLock (CLCL)**, a new benchmark we designed to test agents on difficult, long-horizon visual safety problems (Figure 1). Our experiments on CLCL validate the strengths of our proposed method `MFRLC`.

Environment Setup Experiments use the **CLCL** (Fig. 1), a CMDP with horizon $H = 6$, high-dimensional observations, and a latent state space. The latent states form a chain $s_0 \to s_1 \to \cdots \to s_H$ with two absorbing terminals: a S_{Safe} state and a S_{Fail} state. The agent must learn a sequence of actions (the "combo") to reach the Safe state, but never observes the latent state s_h; it receives only an observation $o_h \in \mathbb{R}^{8\times84\times84}$ from a rendering function, mapped for learning into a feature class Φ of dimension $d = 64$. Observations visually encode progress and timestep, with added noise to increase perceptual difficulty. At each state s_h, the agent chooses from the action set $\mathcal{A} = \{a_{\text{forward}}, a_{\text{hazard}}, a_{\text{stay}}\}$. Taking a_{forward} transitions the agent from state s_h to s_{h+1}. Taking a_{stay} causes the agent to remain in s_h without cost or reward. Taking a_{hazard} incurs a cost of 1 and causes transition to S_{Fail} with probability p_{hazard}, or return to s_{h-1} with probability $1 - p_{\text{hazard}}$. A reward of 1 is given only upon successfully transitioning from s_{H-1} to S_{Safe}; all other rewards are zero. The objective is to maximize expected reward while ensuring the total expected cost remains below 1. Full setup details and hyperparameters are in Appendix D.

Baselines We compare our proposed method against a suite of reinforcement learning algorithms. Our baselines include LAGRANGIAN POLICY GRADIENT (LAG-PG) [22], a standard on-policy actor-critic with GAE for reward and cost; CONSTRAINED POLICY OPTIMIZATION (CPO) [2], an on-policy trust-region method for CMDPs; HOMER [16], a hierarchical option-based algorithm; and BLOCK-STRUCTURED REPRESENTATION LEARNING WITH INTERLEAVED EXPLORE EXPLOIT (BRIEE) [29], which focuses on representation learning coupled with a policy optimization. Since HOMER and BRIEE were originally designed for unconstrained MDPs, we add a Lagrangian framework for fair comparison; details are in Appendix D.7.

Table 2: Additional CLCL settings (mean $\pm$ s.e., 5 seeds). Best R: maximum reward per run; Viol.: cumulative violation. Compare within each setting.

Setting	Method	Best R	Viol.
$H = 6, \lvert\mathcal{A}\rvert = 5$	MFRLC	$\mathbf{0.567} \pm 0.174$	$\mathbf{711.9} \pm 30.8$
	BRIEE	0.105 ± 0.099	1063.1 ± 26.3
	LAG-PG	0.008 ± 0.005	1614.7 ± 137.7
$H = 8, \lvert\mathcal{A}\rvert = 3$	MFRLC	$\mathbf{0.692} \pm 0.136$	207.4 ± 13.0
	BRIEE	0.505 ± 0.005	683.4 ± 2.7
	LAG-PG	0.017 ± 0.004	747.4 ± 191.5
	HOMER	0.004 ± 0.004	936.5 ± 31.0
	CPO	0.012 ± 0.005	$\mathbf{102.2} \pm 75.8$
$H = 12, \lvert\mathcal{A}\rvert = 3$	MFRLC	0.021 ± 0.013	1206.9 ± 76.2

Results MFRLC achieves the best overall performance, balancing reward maximization with constraint satisfaction. In the CLCL task with horizon $H = 6$ (Fig. 2), MFRLC is the only algorithm to achieve an optimal reward approaching 1.0 while satisfying the safety constraint. HOMER satisfies the constraint sooner but converges to a suboptimal reward of ≈ 0.8; BRIEE plateaus at ≈ 0.4 and violates the constraint; CPO and LAG-PG fail to achieve meaningful reward and consistently violate safety. Representation learning alone is thus insufficient: only MFRLC, by jointly learning features and a constraint-aware policy, succeeds on both objectives (details in Appendix D).

Additional Settings Table 2 adds a larger action space and a longer horizon (Appendix D.6). With $|\mathcal{A}| = 5$, where discovering the rewarding sequence under uniform exploration is $(5/3)^6 \approx 21.4$ times less likely, MFRLC attains the highest reward and lowest violation. At $H = 8$, MFRLC obtains the highest reward; only CPO has lower violation, at near-zero reward. Its estimated cumulative regret is $15{,}382 \pm 942$ over 15,552 deployed episodes; we do not read these finite-run numbers as verifying the $K^{3/4}$ rate. At $H = 12$, MFRLC reaches only 0.021 reward, a failure mode under our budget.

6 Limitations

Our results have four main limitations. (i) Theorem 3 is an upper bound; we prove no lower bound and make no optimality claim about the $K^{3/4}$ rate or the dependence on d, H, $|\mathcal{A}|$, and γ. (ii) The outer minimax problem of Algorithm 2 is solved by an optimization oracle; exact search is possible for finite classes, but we give no polynomial-time guarantee for neural classes. (iii) Regret and Violation are charged only to deployed policies; the auxiliary roll-ins of representation learning are counted in the $K(2H+1)$ trajectory complexity but not in the regret or violation sums. (iv) Our experiments use a contrastive EMA surrogate that is not covered by Theorem 3, the evaluation is restricted to CLCL, and MFRLC fails to obtain meaningful reward at $H = 12$ under our training budget. The analysis also assumes a finite action space, which incurs the $|\mathcal{A}|^{3/2}$ factor through uniform exploration, exact realizability $\phi^\star \in \Phi$, and reward and utility linear in $\phi^\star$.

7 Conclusion

We presented MFRLC, a model-free algorithm for online low-rank CMDPs with unknown features. By combining LSVI-UCB, a primal–dual scheme, and a minimax representation-learning objective, MFRLC jointly learns features and constraint-aware policies and achieves $\widetilde{O}(K^{3/4})$ reward regret and cumulative constraint violation over K deployment rounds. Improving this rate, and certifying a scalable representation learner against the guarantee of Algorithm 2, remain open.

Acknowledgments and Disclosure of Funding

Funding. ST is supported by the Emory URC Research Award and the Disruptive Discovery Seed Program by Emory University's Office of the Provost. The content in this document are those of the authors and should not be interpreted as necessarily representing the views of the funding sources.

Competing interests. The authors declare no competing interests.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *ICML*, 2017. URL https://proceedings.mlr.press/v70/achiam17a/achiam17a.pdf.

[3] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems*, 33:20095–20107, 2020.

[4] Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

[5] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through Bayesian deep Q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[6] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366, 2008.

[7] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

[8] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

[9] Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7396–7404, 2023.

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

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

[12] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving sub-linear regret in infinite horizon average reward constrained MDP with linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zZhX4eYNeeh.

[13] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

[14] Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.

[15] Zak Mhammedi, Adam Block, Dylan J Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank MDPs. *Advances in Neural Information Processing Systems*, 36: 66782–66817, 2023.

[16] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

[17] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank MDPs. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

[18] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142:205–228, 2009.

[19] Masahiro Ono, Marco Pavone, Yoshiaki Kuwata, and J Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39: 555–571, 2015.

[20] Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[22] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning, 2019. URL https://cdn.openai.com/safexp-short.pdf.

[23] Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, pages 1686–1696. PMLR, 2022.

[24] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. *arXiv preprint arXiv:2110.04652*, 2021.

[25] Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for constrained Markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

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

[27] Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

[28] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[29] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block MDPs: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.

Supplementary Materials

In this supplementary material, we provide additional details omitted from the main paper for brevity. Throughout the appendix, "episode k" refers to deployment round k (Section 2).

A LSVI-UCB Algorithm Details

A.1 LSVI-UCB Algorithm

Algorithm 3 provides the full pseudocode for our constrained Least-Squares Value Iteration with Upper Confidence Bounds (LSVI-UCB) procedure. This algorithm is invoked in the policy-update step of Algorithm 1 (line 8) at each episode k, taking as inputs (i) the trajectory data collected up to episode k via the exploration step (line 4 of Algorithm 1, with uniform actions $\tilde{a}_h^\tau$), (ii) the learned representation $\hat{\phi}^k$, and (iii) the current dual variable Y_k. It returns the optimistic per-objective Q- and V-functions $\{Q_{r,h}^k, Q_{g,h}^k, V_{r,h}^k, V_{g,h}^k\}_{h=1}^H$ together with the policy $\pi^k = \{\pi_h^k\}_{h=1}^H$.

The algorithm proceeds via backward induction from the final horizon step H down to 1. At each step h, it first constructs the regularized Gram matrix Σ_h^k (line 4), which represents the empirical covariance of the learned features over past episodes. This matrix is then used to perform a regularized least-squares regression to find the linear weights for the reward value function ($w_{r,h}^k$) and the utility value function ($w_{g,h}^k$) (lines 5–6). An optimistic UCB exploration bonus $\hat{b}_h^k$ is then computed to encourage exploration in less-visited state–action pairs (line 7). This single bonus is added to both the reward and utility linear value estimates to produce the optimistic Q-values (lines 8–9). The policy π_h^k is derived from a softmax distribution over the *composite* Q-function $Q_{r,h}^k + Y_k Q_{g,h}^k$ (line 10). Finally, the per-objective state-value functions are updated under this policy (lines 11–12) and used at the next iteration of the backward induction.

Remark 3 (Single bonus, composite bonus)**.** *Lines 8–9 add the same per-objective bonus $\hat{b}_h^k$ to both $Q_{r,h}^k$ and $Q_{g,h}^k$. Consequently, the composite Q-function $Q_{r,h}^k + Y_k Q_{g,h}^k$ used in line 10 inherits a composite bonus of magnitude $(1 + Y_k)\hat{b}_h^k$. The analysis (Appendices F–H) denotes this composite bonus by $\hat{b}_h^{(k)} := (1 + Y_k)\hat{b}_h^k$.*

A.2 Comparator

Theorem 3 compares against an arbitrary optimal feasible stochastic policy $\pi^\star$ of (1). The policy $\pi^\star$ enters the analysis only through the pseudo-regret term of (9), which Lemma 4 bounds by combining composite optimism (Lemma 2), the softmax sub-optimality bound (Lemma 12), and the simulation lemma (Lemma 19). None of these steps requires $\pi^\star$ to be produced by Algorithm 3 from any representation and dual variable. Lemma 2 bounds the optimistic value of the fixed policy $\pi^\star$ on the planning model, and Lemma 12 holds for an arbitrary comparator policy. The finite-temperature softmax error is already accounted for by the explicit $KH \log|\mathcal{A}|/\alpha$ term, so exact recovery of $\pi^\star$ is unnecessary.

For measurability in the concentration arguments, note that Algorithm 3 uses the data only through the feature evaluations $\hat{\phi}_h^k(s_h^\tau, \tilde{a}_h^\tau)$, the regression targets, and Y_k. Hence π^k is a deterministic function of $\hat{\phi}^k$, the data collected before round k, Y_k, and fixed hyperparameters.

B Related Work

Representation Learning in Low-Rank MDPs. Low-rank MDPs [3] provide a structural framework where the transition kernel admits a d-dimensional factorization, generalizing earlier rich-observation models [16]. Existing approaches fall into two categories. *Model-based* methods explicitly estimate both the feature map and the next-state factor [24, 23], requiring realizability assumptions on the full dynamics. *Model-free* methods avoid recovering the next-state factor and instead learn the feature via a min-max objective against a discriminator class: Zhang et al. [29] (BRIEE) operates in block MDPs, Modi et al. [17] extends to general low-rank MDPs, and Mhammedi et al. [15] (VoX) achieves improved sample complexity via barycentric spanners. *None of these works handles constraints*, leaving the constrained setting open.

Algorithm 3 Least-Squares Value Iteration with UCB (LSVI-UCB)

Require: Trajectory data $\{(s_h^\tau, \tilde{a}_h^\tau, r_h(s_h^\tau, \tilde{a}_h^\tau), g_h(s_h^\tau, \tilde{a}_h^\tau), s_{h+1}^\tau, \hat{\phi}_h^k(s_h^\tau, \tilde{a}_h^\tau))\}_{h=1,\,\tau=1}^{H,\,k-1}$ from line 4 of Algorithm 1; current dual variable Y_k

Parameter: $\alpha = \frac{K \log |\mathcal{A}|}{2(1+\xi+H)}$, $\beta = C_1 dH\sqrt{\log(4\log|\mathcal{A}|\, dT|\Phi|/p)}$, regularization λ

Output: Optimistic estimates $\{Q_{r,h}^k, Q_{g,h}^k, V_{r,h}^k, V_{g,h}^k\}_{h=1}^H$ and policy $\pi^k = \{\pi_h^k\}_{h=1}^H$

Initialize $V_{r,H+1}^k(\cdot) \equiv 0, \quad V_{g,H+1}^k(\cdot) \equiv 0$

for $h = H$ **to** 1 **do**

$\Sigma_h^k \leftarrow \sum_{\tau=1}^{k-1} \hat{\phi}_h^k(s_h^\tau, \tilde{a}_h^\tau)\, \hat{\phi}_h^k(s_h^\tau, \tilde{a}_h^\tau)^\top + \lambda I$

$w_{r,h}^k \leftarrow (\Sigma_h^k)^{-1} \sum_{\tau=1}^{k-1} \hat{\phi}_h^k(s_h^\tau, \tilde{a}_h^\tau)\left(r_h(s_h^\tau, \tilde{a}_h^\tau) + V_{r,h+1}^k(s_{h+1}^\tau)\right)$

$w_{g,h}^k \leftarrow (\Sigma_h^k)^{-1} \sum_{\tau=1}^{k-1} \hat{\phi}_h^k(s_h^\tau, \tilde{a}_h^\tau)\left(g_h(s_h^\tau, \tilde{a}_h^\tau) + V_{g,h+1}^k(s_{h+1}^\tau)\right)$

$\hat{b}_h^k(s,a) \leftarrow \beta\sqrt{\hat{\phi}_h^k(s,a)^\top (\Sigma_h^k)^{-1} \hat{\phi}_h^k(s,a)}$ *// per-objective bonus*

$Q_{r,h}^k(s,a) \leftarrow \min\Big(\langle w_{r,h}^k,\, \hat{\phi}_h^k(s,a)\rangle + \hat{b}_h^k(s,a),\; H\Big)$

$Q_{g,h}^k(s,a) \leftarrow \min\Big(\langle w_{g,h}^k,\, \hat{\phi}_h^k(s,a)\rangle + \hat{b}_h^k(s,a),\; H\Big)$

$\pi_h^k(a \mid \cdot) \leftarrow \mathrm{softmax}_\alpha\Big(Q_{r,h}^k(\cdot,a) + Y_k\, Q_{g,h}^k(\cdot,a)\Big)$

$V_{r,h}^k(s) \leftarrow \sum_{a\in\mathcal{A}} \pi_h^k(a \mid s)\, Q_{r,h}^k(s,a)$

$V_{g,h}^k(s) \leftarrow \sum_{a\in\mathcal{A}} \pi_h^k(a \mid s)\, Q_{g,h}^k(s,a)$

end for

Constrained MDPs. CMDPs [20] provide a standard framework for safe RL, where the agent maximizes reward subject to expected cumulative cost constraints. For tabular settings, primal-dual and primal-only methods [10, 7, 27] establish regret and violation guarantees, with model-free variants [25, 9, 26] achieving $\widetilde{O}(\sqrt{K})$ rates. To scale beyond tabular settings, Achiam et al. [2] introduce constrained policy optimization for deep RL, Amani et al. [4] treats safe linear bandits, and Ding et al. [8], Ghosh et al. [11, 12] extend primal-dual policy optimization to linear CMDPs with $\widetilde{O}(\sqrt{K})$ guarantees. *All these works assume the feature map is known a priori*—an assumption that fails when only raw observations (e.g., images) are available. Bridging this gap, `MFRLC` is, to our knowledge, the first algorithm with provable regret and violation guarantees for CMDPs with unknown features. The unknown-feature setting requires controlling three coupled error-propagation mechanisms absent from prior analyses. First, errors in the estimated utility values affect the dual iterate Y_k. Second, the changing Y_k modifies the composite objective $r + Y_k g$, so representation error propagates through both the reward and utility Bellman estimates and is amplified by $1 + Y_k$. Third, the resulting policy changes the data used for subsequent representation learning, which feeds back into the value estimates and the dual update. Prior low-rank MDP analyses do not handle constraints, and prior primal–dual CMDP analyses assume known features, so neither faces this coupling.

C Representation Learning Algorithm

This appendix collects the implementation details of the representation learning step invoked in line 6 of Algorithm 1. We first specify the discriminator class $\mathcal{V}_h = \mathcal{F}_h \cup \mathcal{G}_h$, then derive the empirical surrogate of the population objective (5), and finally state the full iterative procedure (Algorithm 2) together with its theoretical guarantee.

C.1 Discriminator Class

The two discriminator classes used in Section 3.1 are defined as

$$\mathcal{F}_h := \Big\{ s \mapsto \mathbb{E}_{a\sim\pi_h^{k-1}(\cdot|s)}[\langle\phi(s,a),\theta\rangle - \langle\phi'(s,a),\theta'\rangle] : \phi,\phi' \in \Phi_{h+1},\ \max(\|\theta\|_\infty, \|\theta'\|_\infty) \le 1 \Big\}, \tag{10}$$

$$\mathcal{G}_h := \Bigg\{ s \mapsto \max_{a\in\mathcal{A}} \left(\tfrac{r_{h+1}(s,a)+g_{h+1}(s,a)+\min\{\langle\phi(s,a),w\rangle, 2\}}{2H+1} + \langle\phi(s,a), w'\rangle \right) : \phi \in \Phi_{h+1},\ \|w\|_\infty \le c,\ \|w'\|_\infty \le 1 \Bigg\}. \tag{11}$$

$\mathcal{F}_h$ captures the model-mismatch function appearing in the proof of composite optimism (Lemma 2), evaluated under the previous-episode policy π^{k-1} which is well-defined at the start of episode k. $\mathcal{G}_h$ captures (after rescaling by $1+Y_k$) the linear part of the optimistic composite Q-function appearing in the policy-gap bound (Lemma 3); the non-linear softmax operator and UCB bonus produced by Algorithm 3 are not contained in $\mathcal{G}_h$ and are handled separately via Lemma 12 and the elliptical-potential argument in Appendix F. The simple form above—independent of Y_k—defines a discriminator class with statistical complexity $\log|\mathcal{G}_h|$ that does not depend on Y_k. When applied to composite quantities of magnitude $1+Y_k$ in the proofs of Lemmas 2–3, the rescaling factor $1+Y_k \le 1+\xi$ appears as an explicit multiplicative constant in the RepLearn statistical error ζ_k, which is absorbed into the cumulative bound $(1+\xi)\sum_k \epsilon_k$ in Lemma 3.

C.2 From Population to Empirical Objective

Recall the population objective (5) from the main text:

$$\arg\min_{\phi\in\Phi_h} \max_{v\in\mathcal{V}_h} \min_{w\in\mathbb{R}^d} \mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\Big[\big(\langle\phi(s,\tilde{a}),w\rangle - \mathbb{E}_{s'\sim P_h^\star(\cdot|s,\tilde{a})}[v(s')]\big)^2\Big].$$

Direct optimization is intractable because the inner expectation $\mathbb{E}_{s'\sim P_h^\star}[v(s')]$ requires the unknown true kernel. To address this, we replace the squared deviation with the squared empirical residual:

$$\mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi,v,w) \;:=\; \mathbb{E}_{(s,\tilde{a},s')\sim\mathcal{D}_h}\Big[\big(\langle\phi(s,\tilde{a}),w\rangle - v(s')\big)^2\Big] \;+\; \lambda_k\|w\|_2^2. \tag{12}$$

Expanding around the conditional mean of $v(s')$ gives the variance–bias decomposition

$$\begin{aligned}\mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi,v,w) \;&=\; \underbrace{\mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\Big[\big(\langle\phi(s,\tilde{a}),w\rangle - \mathbb{E}_{s'\sim P_h^\star}[v(s')]\big)^2\Big]}_{\text{population objective in (5)}} \\ &\quad + \underbrace{\mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\big[\mathbb{V}_{s'\sim P_h^\star}[v(s')]\big]}_{\text{conditional variance term}} \;+\; \lambda_k\|w\|_2^2.\end{aligned} \tag{13}$$

The conditional-variance term is independent of ϕ but is non-trivially correlated with v, so naïvely minimizing $\mathcal{L}_{\lambda_k,\mathcal{D}_h}$ alone biases the inner minimization in favor of low-variance discriminators rather than the true population minimum.

To cancel the variance term, we exploit the fact that the second buffer $\mathcal{D}'_h$ contains transitions $(s',\tilde{a}',s'')$ where, for each tuple, $\tilde{a}' \sim U(\mathcal{A})$ is drawn independently of the original step-h data and $s'' \sim P_h^\star(\cdot \mid s',\tilde{a}')$ shares the same conditional law as the step-h next-state distribution given $(s',\tilde{a}')$. By Lemma 17, for any $\tilde{\phi}\in\Phi_h$, $v\in\mathcal{V}_h$, and $\tilde{w}\in\mathbb{R}^d$,

$$\mathbb{E}_{\mathcal{D}'_h}\big[\mathcal{L}_{\lambda_k,\mathcal{D}'_h}(\tilde{\phi},v,\tilde{w})\big] \;\ge\; \mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\big[\mathbb{V}_{s'\sim P_h^\star}[v(s')]\big],$$

with equality when $\tilde{\phi}$ and $\tilde{w}$ are chosen optimally. Subtracting the $\tilde{w}$-minimized version of $\mathcal{L}_{\lambda_k,\mathcal{D}'_h}$ from $\mathcal{L}_{\lambda_k,\mathcal{D}_h}$ therefore eliminates the conditional variance in expectation, recovering an unbiased estimator of the population objective up to the $\lambda_k\|w\|^2$ regularization. This motivates the bias-corrected objective solved by Algorithm 2:

$$\arg\min_{\phi\in\Phi_h} \max_{v\in\mathcal{V}_h} \Big\{ \min_w \mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi,v,w) \;-\; \min_{\tilde{w},\,\tilde{\phi}\in\Phi_h} \mathcal{L}_{\lambda_k,\mathcal{D}'_h}(\tilde{\phi},v,\tilde{w}) \Big\}. \tag{14}$$

C.3 Iterative Procedure

Algorithm 2 alternates between two steps: an adversarial discriminator search that identifies the worst-case $v \in \mathcal{V}_h$ for the current representation ϕ^m (line 6), and a representation update that fits ϕ^{m+1} jointly against all discriminators encountered so far (line 11). The inner $\max_{\tilde{\phi}}$ on line 6 implements the bias correction of (14): it subtracts the best variance proxy that any $\tilde{\phi} \in \Phi_h$ can achieve. Termination occurs once no discriminator achieves an empirical objective above ℓ_k, indicating that all $v \in \mathcal{V}_h$ are approximately linearly representable by ϕ^m at the prescribed statistical precision.

C.4 Theoretical Guarantee and Stopping Rule

The output $\hat{\phi}_h^k$ of Algorithm 2 satisfies the Representation Learning Guarantee (Lemma 16): with probability at least $1-\delta$,

$$\max_{v \in \mathcal{V}_h} \mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\Big[\big(\hat{w}_v^\top \hat{\phi}_h^k(s,\tilde{a}) - \mathbb{E}_{s'\sim P_h^\star}[v(s')]\big)^2\Big] \;\le\; \zeta_k \;=\; O\big(d^2\sqrt{\log(k|\Phi|/\delta)/k}\big),$$

where $\hat{w}_v$ is the inner minimizer corresponding to v. Combined with Theorem 3, this guarantees that the policies computed from $\hat{\phi}_h^k$ achieve sublinear regret and constraint violation, ensuring that the learned representation can be effectively used in policy evaluation and improvement within the CMDP framework.

Stopping rule and optimization error. The theoretical analysis requires the threshold ℓ_k to scale as $\widetilde{O}(d^2/\sqrt{k})$ to match the statistical error rate. An implementation of Algorithm 2 enforces this by evaluating the maximization objective on line 6 at each iteration m and terminating once it falls below ℓ_k, so the residual optimization error remains negligible compared to the statistical estimation error and the cumulative bound $\sum_k \zeta_k$ used in the analysis is not inflated by optimization slack. Theorem 3 applies whenever the optimization steps of Algorithm 2 attain the residual displayed above. For finite outer classes this can be certified by exact search, and an approximate solver is also covered if its optimization error is $O(\zeta_k)$. The contrastive EMA surrogate used in our experiments (Appendix D.2) does not certify this condition and is not covered by Theorem 3.

C.5 Computational Cost

For fixed (ϕ, v), the inner problem $\min_w \mathcal{L}_{\lambda_k,\mathcal{D}_h}(\phi, v, w)$ is ridge regression. With $n = |\mathcal{D}_h|$ samples, forming the Gram matrix costs $O(nd^2)$ and solving the $d \times d$ system costs $O(d^3)$, so one solve costs $O(nd^2 + d^3)$. At iteration m of Algorithm 2, line 6 evaluates two ridge solves for each pair $(v, \tilde{\phi}) \in \mathcal{V}_h \times \Phi_h$, and line 11 solves $m+1$ ridge problems for each $\phi \in \Phi_h$. When Φ_h and $\mathcal{V}_h$ are finite (or replaced by finite covers), every step can therefore be computed exactly, and one call with at most M iterations costs:

$$O\big((M|\Phi_h||\mathcal{V}_h| + M^2|\Phi_h|)(nd^2 + d^3)\big).$$

This is polynomial in n and d but may be large in the class cardinalities. For continuously parameterized classes such as neural networks, the same empirical objective (14) is explicit, but its outer optimization is non-convex, and we do not prove polynomial-time convergence to the global minimax solution. Given $\hat{\phi}^k$, Algorithm 3 performs two backward passes of ridge regression over $k-1$ samples per step, and the projected dual update (8) costs $O(1)$.

D Experiment Details

This section provides a detailed overview of the experimental setup, including environment specifics, network architectures, training protocols, and a comprehensive list of hyperparameters used for all algorithms to ensure reproducibility.

D.1 Environment Setup

All experiments are conducted in our custom-designed **Constrained Latent ComboLock (CLCL)** environment. We formulate constraints as costs bounded above ($\text{cost} \le \mathcal{C}_{\max} = b$), which is

equivalent to the main-text utility convention via $V_g = -\text{cost}$, $b_{\text{utility}} = -\mathcal{C}_{\max}$, so that $V_g \geq b_{\text{utility}}$ corresponds to the cost staying below threshold.

- **Horizons and action spaces:** The main experiments use $H = 6$ and $|\mathcal{A}| = 3$. The additional settings in Table 2 enlarge the action set to $|\mathcal{A}| = 5$ at $H = 6$, and use $H \in \{8, 12\}$ with $|\mathcal{A}| = 3$.
- **Observation Space:** The agent receives high-dimensional observations in the form of $8 \times 84 \times 84$ image tensors. Each tensor is procedurally generated based on the agent's latent state s_h and the current timestep h. The eight channels are specifically designed to encode rich information:
  - **Channels 0–2 (Progress and Position):** These channels encode the agent's progress through the latent state chain.
    Channel 0 contains a Gaussian "spotlight" whose horizontal position corresponds to the current state s_h. Specifically, for a pixel at coordinates $(x, y) \in [0, 1]^2$, the value is generated as:
    $$o^{(0)}_{x,y} = \exp\left(-20\left[\left(x - \frac{s_h + 1}{H + 1}\right)^2 + (y - 0.5)^2\right]\right).$$
    Channels 1 and 2 contain sine and cosine patterns whose frequencies change with the agent's progress, providing additional positional information. They are defined as:
    $$o^{(1)}_{x,y} = 0.5 \sin((s_h + 1)\pi x), \qquad o^{(2)}_{x,y} = 0.5 \cos((s_h + 1)\pi y).$$
    These patterns ensure that each depth in the latent chain has a distinct spatial frequency signature.
  - **State Textures:** A unique, pseudo-random visual texture is overlaid on the first few channels. The choice of texture depends on the latent state, timestep, and previous action, making each observation visually distinct.
  - **Channels 3–4 (Background Patterns):** These channels contain static, high-frequency sine and cosine waves that form a consistent background grid.
  - **Channel 5 (Fail Indicator):** This channel acts as a binary flag. It is filled with ones if the agent is in the terminal S_{Fail} state, and zeros otherwise.
  - **Channel 6 (Safe Indicator):** Similar to the fail indicator, this channel is filled with ones if the agent is in the terminal S_{Safe} state, and zeros otherwise.
  - **Channel 7 (Timestep):** This channel's value is constant across the image and represents the normalized current timestep, i.e., h/H.
  - **Noise:** Finally, Gaussian noise with a standard deviation of 0.1 is added to all channels to increase the perceptual difficulty of the task.
- **Action Space:** The action space is discrete, containing three actions: $\mathcal{A} = \{a_{\text{forward}}, a_{\text{hazard}}, a_{\text{stay}}\}$.
- **Rewards and Costs:** A reward of $+1$ is given only for successfully reaching the S_{Safe} state from the final state in the chain. A cost of 1 is incurred for taking the a_{hazard} action. The cost constraint for all experiments is set to a threshold of $\mathcal{C}_{\max} = 1.0$.

D.2 Network Architectures

The core of the visual processing for MFRLC and all visual baselines is a shared ResNet-style encoder.

- **Visual Encoder:** We use a `VisualEncoder` that maps the $8 \times 84 \times 84$ observation to a 64-dimensional feature vector. It consists of a convolutional stem followed by four ResNet-style `BasicBlock` layers and a final linear layer.
- **Policy and Value Heads:** For actor-critic baselines (CPO and LAG-PG), this visual encoder serves as a shared backbone. The policy (actor), reward value (critic), and cost value (cost critic) are represented by separate linear heads that take the 64-dimensional feature vector as input.

- **Representation Learning:** For representation-learning methods (MFRLC and BRIEE), the encoder is part of a module with a two-layer `ProjectionHead` for contrastive learning and a target encoder updated via exponential moving average (EMA). Exact min-max-min optimization over $\Phi_h \times \mathcal{V}_h$ in Algorithm 2 is intractable for deep encoders, so we replace it with contrastive next-state prediction on transition triples $(o, \tilde{a}, o')$. This surrogate is not covered by Theorem 3 (Section 3).
- **Feature class Φ:** We instantiate Φ as a trainable neural state–action representation class. The visual encoder maps each $8 \times 84 \times 84$ observation o to a state representation $\psi(o) \in \mathbb{R}^{64}$. We combine $\psi(o)$ with a learned embedding e_a of the discrete action and apply a two-layer action-conditioning network:

$$\phi(o, a) = \mathrm{MLP}\big([\psi(o); e_a]\big) \in \mathbb{R}^{64}.$$

  The features $\phi(o, a)$ are the design vectors of the reward and utility ridge regressions in LSVI-UCB, so in practice Φ is the class of state–action feature maps induced by this parameterization.

D.3 Training Protocol

To ensure a fair and rigorous comparison, all algorithms were trained and evaluated under a consistent protocol.

- **Training:** Each algorithm is trained for a fixed number of iterations, as detailed in Table 3. At each iteration, a number of episodes are collected from the environment to update the agent's policy and/or representation.
- **Statistical Significance:** All learning curves presented in the main paper are the mean and standard error computed over 5 independent runs with the specific random seeds listed in Table 3.
- **Compute Resources:** All experiments were run on a single NVIDIA RTX 4060 GPU (8GB VRAM).

D.4 Hyperparameters

The following table details the key hyperparameters used for our proposed method and all baseline algorithms.

Hyperparameter	**MFRLC**	**BRIEE**	**HOMER**	**CPO**	**LAG-PG**
Replay Buffer Capacity	60,000	60,000	60,000	N/A (on-policy)	N/A (on-policy)
Episodes per Iteration	48	48	48	48	48
Batch Size	512	512	512	N/A (full batch)	N/A (full batch)
Learning Rate (LR)	3e-4	3e-4	3e-4	3e-4	3e-4
Dual LR (η)	0.05	0.05	0.05	0.02	0.05
Feature Dimension	64	64	256 (hidden)	64 (encoder)	64 (encoder)
Discount Factor	0.99	0.99	0.99	0.99	0.99
Random Seeds	0, 10, 29, 57, 78				
p_{hazard}	0.15				

Table 3: Hyperparameters for all algorithms.

D.5 Cumulative Regret and Violation

Let $\bar{R}_t$ and $\bar{C}_t$ be the mean reward and mean cost over the 48 deployed episodes of iteration t. Since reaching S_{Safe} yields reward 1 and a_{forward} incurs no cost, the optimal feasible value is $V_{r,1}^{\pi^\star}(s_1) = 1$. After T iterations, we estimate the cumulative regret and cumulative violation as:

$$\widehat{\mathrm{Regret}} = 48 \sum_{t=1}^{T} \big(1 - \bar{R}_t\big), \qquad \widehat{\mathrm{Violation}} = \Big[\, 48 \sum_{t=1}^{T} \big(\bar{C}_t - \mathcal{C}_{\max}\big) \Big]_+,$$

which follow the definitions in Section 2 under the cost convention of Appendix D.1. Table 2 reports the resulting cumulative violation for the additional settings. At $H = 8$, the estimated cumulative regret of `MFRLC` is $15{,}382 \pm 942$ over $K = 15{,}552$ deployed episodes (324 iterations). These are finite-run measurements; we do not fit or claim an empirical exponent.

D.6 Additional Settings

The three additional settings in Table 2 use the hyperparameters of Table 3, with the same 5 seeds. The $H = 8$ runs use 324 iterations (15,552 deployed episodes). The $H = 12$ stress test runs longer and evaluates only `MFRLC`, which reaches a best reward of 0.021 ± 0.013 with cumulative violation 1206.9 ± 76.2; under uniform exploration, discovering the complete rewarding sequence is $3^4 = 81$ times less likely than at $H = 8$.

D.7 Adapting Baselines for Fair Comparison

To create a level playing field, it was necessary to adapt the baseline algorithms that were not originally designed for Constrained Markov Decision Processes (CMDPs). We use CPO [2] as-is, since it is natively designed for CMDPs; only HOMER and BRIEE require Lagrangian adaptation as detailed below. Specifically, we extended HOMER and BRIEE with a standard Lagrangian multiplier method. The dual variable λ is updated using the same primal–dual rule as in Algorithm 1, but since neither HOMER nor BRIEE maintains a separate cost critic, we estimate the cost value $\widehat{V}_g(s_1)$ used in the dual update as the difference between the augmented value Q^λ (which is what these baselines actually learn) and the empirical reward return collected along the rollout, i.e. $\widehat{V}_g(s_1) \approx$ (empirical reward return $- Q^\lambda(s_1, a_1)$)$/\lambda$ when $\lambda > 0$, with a small-λ fallback to a Monte-Carlo cost return estimate. The application of the dual variable to the value learning is tailored to the unique architecture of each agent.

BRIEE with Constraints The BRIEE algorithm decouples representation learning from planning. We integrate the safety constraint directly into its planning module, which uses a variant of Least-Squares Value Iteration (LSVI). The planner's objective is to solve for an optimal action-value function, Q^*, by iteratively applying the Bellman optimality operator. We modify this objective by replacing the standard reward $R(s, a)$ with the augmented reward $R^\lambda(s, a) = R(s, a) - \lambda C(s, a)$.

Specifically, the planner now solves for a constraint-aware value function, Q^λ, according to the modified Bellman equation:

$$Q_h^\lambda(s, a) \leftarrow \big(R_h(s, a) - \lambda C_h(s, a)\big) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \Big[\max_{a'} Q_{h+1}^\lambda(s', a')\Big].$$

By solving for the optimal policy under this new objective, the planner learns to inherently avoid actions that are likely to violate the constraint.

HOMER with Constraints HOMER is a hierarchical, value-based algorithm that learns from off-policy data. We incorporate the constraint into its value function learning process by modifying the targets used for the temporal-difference (TD) update. The agent's critic is trained to minimize the TD error between its current estimate $Q(s_t, a_t)$ and a target value y_t. The standard TD target is $y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a')$.

To make the agent constraint-aware, we modify this target to use the augmented reward:

$$y_t^\lambda \leftarrow \big(r_t - \lambda c_t\big) + \gamma \max_{a'} Q^\lambda(s_{t+1}, a').$$

The critic is then trained by minimizing a loss function based on this new target, such as the mean squared error $\mathcal{L} = \mathbb{E}[(Q^\lambda(s_t, a_t) - y_t^\lambda)^2]$. This ensures that the learned value functions, which guide the agent's policy, account for both the expected future reward and the expected future cost, weighted by the dual variable λ.

E Proof of Decomposition

In this section we prove the decomposition of regret and violation used in the master inequality (9).

Lemma 8. *For any $Y \in [0, \xi]$, we have*

$$\sum_{k=1}^{K}(Y - Y_k)\big(b - V_{g,1}^{k}(s_1)\big) \;\le\; \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2}.$$

Proof. The dual update rule of Algorithm 1 (line 12) is

$$Y_{k+1} \;=\; \max\big\{\, \min\{Y_k + \eta(b - V_{g,1}^{k}(s_1)),\, \xi\},\, 0\big\}.$$

Let $\Delta_k := b - V_{g,1}^{k}(s_1)$. Since this update is the Euclidean projection of $Y_k + \eta\Delta_k$ onto the interval $[0, \xi]$, for any $Y \in [0, \xi]$,

$$\begin{aligned}(Y_{k+1} - Y)^2 &\le (Y_k + \eta\Delta_k - Y)^2 \\ &= (Y_k - Y)^2 + 2\eta\Delta_k(Y_k - Y) + \eta^2\Delta_k^2.\end{aligned}$$

Rearranging to bound the inner product,

$$2\eta\Delta_k(Y - Y_k) \;\le\; (Y_k - Y)^2 - (Y_{k+1} - Y)^2 + \eta^2\Delta_k^2.$$

Substituting $\Delta_k = b - V_{g,1}^{k}(s_1)$ and summing over $k = 1, \ldots, K$,

$$\begin{aligned}2\eta\sum_{k=1}^{K}(b - V_{g,1}^{k}(s_1))(Y - Y_k) &\le \sum_{k=1}^{K}\big((Y_k - Y)^2 - (Y_{k+1} - Y)^2\big) + \eta^2\sum_{k=1}^{K}(b - V_{g,1}^{k}(s_1))^2 \\ &\le (Y_1 - Y)^2 + \eta^2 H^2 K,\end{aligned}$$

where the last inequality follows from telescoping, dropping the negative term $-(Y_{K+1} - Y)^2$, and using $|b - V_{g,1}^{k}(s_1)| \le H$. Dividing by 2η and noting $Y_1 = 0$,

$$\sum_{k=1}^{K}(Y - Y_k)\big(b - V_{g,1}^{k}(s_1)\big) \;\le\; \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2}. \qquad \square$$

Lemma 9. *For any $Y \in [0, \xi]$,*

$$\begin{aligned}&\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big) \\ &\quad\le\; \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2} + \underbrace{\sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{\pi^\star}(s_1)\big)}_{\textit{Slack (non-positive)}} \\ &\qquad + \underbrace{\sum_{k=1}^{K}\Big[\big(V_{r,1}^{\pi^\star}(s_1) + Y_k V_{g,1}^{\pi^\star}(s_1)\big) - \big(V_{r,1}^{k}(s_1) + Y_k V_{g,1}^{k}(s_1)\big)\Big]}_{\textit{Pseudo regret}} \\ &\qquad + \underbrace{\sum_{k=1}^{K}\Big[\big(V_{r,1}^{k}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\big(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi^k}(s_1)\big)\Big]}_{\textit{Estimation error}}.\end{aligned}$$

Proof. We start by decomposing the Lagrangian term involving the constraint violation. Using Lemma 8,

$$\begin{aligned} Y\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big) &= \sum_{k=1}^{K}(Y - Y_k)\big(b - V_{g,1}^{k}(s_1)\big) + \sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{k}(s_1)\big) \\ &\quad + Y\sum_{k=1}^{K}\big(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi^k}(s_1)\big) \\ &\le \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2} + \sum_{k=1}^{K}\big(Y_k b - Y_k V_{g,1}^{k}(s_1)\big) \\ &\quad + Y\sum_{k=1}^{K}\big(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi^k}(s_1)\big). \end{aligned}$$

We rewrite $\sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{k}(s_1)\big)$ by adding and subtracting $V_{g,1}^{\pi^\star}(s_1)$:

$$\sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{k}(s_1)\big) = \sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{\pi^\star}(s_1)\big) + \sum_{k=1}^{K} Y_k\big(V_{g,1}^{\pi^\star}(s_1) - V_{g,1}^{k}(s_1)\big).$$

Since $\pi^\star$ is feasible, we have $V_{g,1}^{\pi^\star}(s_1) \ge b$, so $b - V_{g,1}^{\pi^\star}(s_1) \le 0$. Combined with $Y_k \ge 0$, the slack term $\sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{\pi^\star}(s_1)\big)$ is non-positive. We retain it explicitly to make the decomposition transparent; dropping it tightens the bound in the correct direction (recovering (9) in the main text).

Adding the reward regret $\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi^k}(s_1)\big)$ to both sides and rearranging:

$$\begin{aligned} &\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big) \\ &\quad\le \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2} + \sum_{k=1}^{K} Y_k\big(b - V_{g,1}^{\pi^\star}(s_1)\big) \\ &\qquad + \sum_{k=1}^{K}\Big[\big(V_{r,1}^{\pi^\star}(s_1) + Y_k V_{g,1}^{\pi^\star}(s_1)\big) - \big(V_{r,1}^{k}(s_1) + Y_k V_{g,1}^{k}(s_1)\big)\Big] \\ &\qquad + \sum_{k=1}^{K}\Big[\big(V_{r,1}^{k}(s_1) - V_{r,1}^{\pi^k}(s_1)\big) + Y\big(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi^k}(s_1)\big)\Big]. \end{aligned}$$

Identifying the slack, pseudo regret, and estimation error terms completes the proof. □

F Bounding Pseudo regret

Composite reward and notation. Throughout this section, fix an episode $k \in \{1, \dots, K\}$ and let $Y_k \in [0, \xi]$ denote the dual iterate produced by Algorithm 1. We work with the *composite* stage reward

$$\bar{r}_h^{(k)}(s,a) \;:=\; r_h(s,a) + Y_k\, g_h(s,a) \;\in\; [0, 1 + Y_k],$$

and the corresponding composite cumulative value, which lies in $[0, (1+Y_k)H]$. For any policy π and any transition kernel P, define the composite value and action-value functions

$$V_{P,h}^{\pi,Y_k}(s) \;:=\; V_{P,r,h}^{\pi}(s) + Y_k\, V_{P,g,h}^{\pi}(s), \qquad Q_{P,h}^{\pi,Y_k}(s,a) \;:=\; Q_{P,r,h}^{\pi}(s,a) + Y_k\, Q_{P,g,h}^{\pi}(s,a).$$

By linearity of expectations, $V_{P,h}^{\pi,Y_k}$ and $Q_{P,h}^{\pi,Y_k}$ are precisely the value and action-value functions of policy π in the auxiliary MDP $(P, \bar{r}^{(k)})$. We further write $\hat{b}_h^{(k)}(s,a) := (1+Y_k)\,\hat{b}_h(s,a)$ for the composite optimism bonus, where $\hat{b}_h(s,a) = \beta\,\|\hat{\phi}_h(s,a)\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}$ is the per-objective bonus computed in Algorithm 3. Since $0 \le Y_k \le \xi$, we have $\hat{b}_h^{(k)} \le (1+\xi)\hat{b}_h$. The implicit planning

model $\hat{P}_h^k(s'|s,a) := \langle \hat{\phi}_h^k(s,a), \hat{\mu}_h^k(s') \rangle$ is defined as in Section 4; the density factor $\hat{\mu}_h^k$ is recovered from the LSVI weights and is used only as an analysis device.

We define the following mixture distributions to facilitate the analysis. For any k, h, let $\rho_h^k \in \Delta(\mathcal{S}\times\mathcal{A})$ be the data distribution induced by Algorithm 1's exploration step, namely

$$\rho_h^k(s,a) \;=\; \tfrac{1}{k}\sum_{i=0}^{k-1} d_h^{\pi^i}(s)\, U(\mathcal{A})(a),$$

i.e. the average of past on-policy state distributions paired with a uniform action distribution. We similarly define $\beta_h^k(s,a) := \frac{1}{k}\sum_{i=0}^{k-1} \mathbb{E}_{\tilde{s}\sim d_{h-1}^{\pi^i},\, \tilde{a}\sim U(\mathcal{A})}[P^\star_{h-1}(s \mid \tilde{s},\tilde{a})]\, U(\mathcal{A})(a)$ and $\gamma_h^k(s,a) := \frac{1}{k}\sum_{i=0}^{k-1} d_h^{\pi^i}(s,a)$. When the episode index k is clear from context, we drop it and write $\rho_h, \beta_h, \gamma_h$. All occurrences of the action variable in ρ_h and β_h correspond to the uniform exploratory actions $\tilde{a}$ of Algorithm 1 (line 4); we keep the symbol a inside expectations for notational compactness, with the understanding that $a \sim U(\mathcal{A})$.

Probability budget. Throughout the analysis we set the RepLearn confidence parameter $\delta = p/4$, so that union bounding the RepLearn guarantee (Lemma 16), the concentration event $\mathcal{E}$ (Lemma 5), and the martingale deviation (Lemma 15) yields total probability at least $1-p$ as claimed in Theorem 3.

Discriminator class. Recall the discriminator class $\mathcal{V}_h = \mathcal{F}_h \cup \mathcal{G}_h$ designed as follows:

$$\mathcal{F}_h := \Big\{ s \mapsto \mathbb{E}_{a\sim\pi_h^{k-1}(\cdot|s)}\big[\phi(s,a)^\top\theta - \phi'(s,a)^\top\theta'\big] : \phi,\phi' \in \Phi_{h+1},\ \max(\|\theta\|_\infty, \|\theta'\|_\infty) \le 1 \Big\},$$

$$\mathcal{G}_h := \Bigg\{ s \mapsto \max_{a\in\mathcal{A}} \Big(\tfrac{r_{h+1}(s,a)+g_{h+1}(s,a)+\min\{w^\top\phi(s,a),2\}}{2H+1} + \phi(s,a)^\top w' \Big) : \phi \in \Phi_{h+1},\ \|w\|_\infty \le c,\ \|w'\|_\infty \le 1 \Bigg\}.$$

The class $\mathcal{F}_h$ captures, via linear function approximation, all functions of the form

$$f(s) \;=\; \mathbb{E}_{a\sim\pi_h^{k-1}(\cdot|s)}\Big[\mathbb{E}_{s'\sim\hat{P}_h(\cdot|s,a)}[V^{\pi^\star,Y_k}_{P^\star,h+1}(s')] - \mathbb{E}_{s'\sim P_h^\star(\cdot|s,a)}[V^{\pi^\star,Y_k}_{P^\star,h+1}(s')]\Big],$$

which appears in the proof of Lemma 2 after a uniform-action importance-sampling step that converts an $\mathbb{E}_{a\sim\pi_h^\star}$ expectation into an $\mathbb{E}_{a\sim\pi_h^{k-1}}$ expectation, contributing a $\sqrt{|\mathcal{A}|}$ factor that is already tracked in our final $|\mathcal{A}|^{3/2}$ rate. The class $\mathcal{G}_h$ captures (after rescaling by $(1+Y_k)$) the linear part of the composite optimistic action-value

$$g(s) \;=\; \tfrac{1}{(2H+1)(1+Y_k)}\, Q^{\pi^k,Y_k}_{\hat{P},\,\bar{r}^{(k)}+\hat{b}^{(k)},\,h+1}(s),$$

which appears in the proof of Lemma 3. The non-linear softmax operator and UCB bonus produced by Algorithm 3 are not contained in $\mathcal{G}_h$; their contribution is handled separately via Lemma 12 (the softmax-vs-max gap, contributing $H\log|\mathcal{A}|/\alpha$) and Lemma 13 (the LSVI-UCB residual, dominated by $\sum_k \epsilon_k$). With the simpler definition above (matching the main text), elements of $\mathcal{G}_h$ are bounded by $1+Y_k \le 1+\xi$ rather than 1; this enters as a multiplicative $(1+\xi)$ factor in the RepLearn statistical error wherever $\mathcal{G}_h$ is used, which we track explicitly below.

We first establish two "one-step" inequalities that decompose the on-policy value of an arbitrary function g_h into a bilinear feature term plus a residual.

Lemma 10 (One-step inequality for the learned model)**.** *Consider functions $g_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ with $\|g_h\|_\infty \le B$. Assume that $s \mapsto \mathbb{E}_{a\sim\pi_h(\cdot|s)}[g_{h+1}(s,a)] \in \mathcal{F}_h$. Conditioned on the Representation Learning Guarantee, for any policy π,*

$$\sum_{h=0}^{H-1} \mathbb{E}_{(s,a)\sim d^\pi_{\hat{P},h}}[g_h(s,a)] \le \sum_{h=1}^{H-1} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{P},h-1}} \Bigg[\|\hat{\phi}_{h-1}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_{h-1},\hat{\phi}_{h-1}}} \cdot \sqrt{k|\mathcal{A}|^2\mathbb{E}_{(s,a)\sim\beta_h}[g_h^2(s,a)] + B^2\lambda_k d + k|\mathcal{A}|^2\zeta_k} \Bigg] + \sqrt{|\mathcal{A}|\,\mathbb{E}_{(s,a)\sim\rho_0}[g_0^2(s,a)]}.$$

Proof. We consider the inequality in two cases.

Case 1: $h = 0$**.**

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P},0}}[g_0(s,a)] = \mathbb{E}_{s\sim d_0,\, a\sim\pi_0(\cdot|s)}[g_0(s,a)] \;\le\; \sqrt{|\mathcal{A}|\,\mathbb{E}_{(s,a)\sim\rho_0}[g_0^2(s,a)]},$$

where the inequality uses Cauchy–Schwarz with the importance ratio $\frac{d_0(s)\pi_0(a|s)}{d_0(s)U(a)} \le |\mathcal{A}|$, since the actions in the replay buffer are sampled uniformly.

Case 2: $h = 1, \dots, H-1$**.**

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P},h}}[g_h(s,a)] &= \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P},h-1}}\left[\hat{\phi}_{h-1}(\tilde{s},\tilde{a})^{\top}\int\sum_a \hat{\mu}_{h-1}(s)\pi_h(a|s)g_h(s,a)\,ds\right] \\
&\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P},h-1}}\Bigg[\|\hat{\phi}_{h-1}\|_{\Sigma^{-1}_{\rho_{h-1},\hat{\phi}_{h-1}}} \\
&\qquad \cdot \Big\|\int\sum_a \hat{\mu}_{h-1}(s)\pi_h(a|s)g_h(s,a)ds\Big\|_{\Sigma_{\rho_{h-1},\hat{\phi}_{h-1}}}\Bigg],
\end{aligned} \tag{15}$$

where the inequality is Cauchy–Schwarz with respect to the positive-definite matrix $\Sigma_{\rho_{h-1},\hat{\phi}_{h-1}} = k\mathbb{E}_{\rho_{h-1}}[\hat{\phi}_{h-1}\hat{\phi}_{h-1}^{\top}] + \lambda_k I$.

We bound the Gram-norm term inside (15):

$$\begin{aligned}
&\left\|\int\sum_a \hat{\mu}_h(s)\pi_h(a|s)g_{h+1}(s,a)\,ds\right\|^2_{\Sigma_{\gamma_h,\hat{\phi}_h}} \\
&\le k\,\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h}\left[\Big(\int\sum_a \hat{\mu}_h(s)^{\top}\hat{\phi}_h(\tilde{s},\tilde{a})\pi_h(a|s)g_{h+1}(s,a)ds\Big)^2\right] + B^2\lambda_k d \\
&\qquad \text{(using } \|\sum_a \pi_h(a|s)g_{h+1}\|_\infty \le B \text{ and } \|\int\hat{\mu}_h(s)f(s)ds\|_2 \le \sqrt{d}\text{)} \\
&= k\,\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h}\left[\Big(\mathbb{E}_{s\sim\hat{P}_h(\tilde{s},\tilde{a}),\,a\sim\pi_h}[g_{h+1}(s,a)]\Big)^2\right] + B^2\lambda_k d \\
&\le k\,\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h}\left[\Big(\mathbb{E}_{s\sim P^{\star}_h(\tilde{s},\tilde{a}),\,a\sim\pi_h}[g_{h+1}(s,a)]\Big)^2\right] + B^2\lambda_k d + k|\mathcal{A}|^2\zeta_k \\
&\qquad \text{(by the Representation Learning Guarantee)} \\
&\le k|\mathcal{A}|^2\,\mathbb{E}_{(s,a)\sim\beta_{h+1}}\left[g_{h+1}^2(s,a)\right] + B^2\lambda_k d + k|\mathcal{A}|^2\zeta_k,
\end{aligned}$$

where the last step uses uniform-action importance sampling and Jensen's inequality. Substituting back into (15) and summing over $h = 1, \dots, H-1$, then adding the Case 1 term, completes the proof. □

Lemma 11 (One-step inequality for the true model)**.** *Consider functions $g_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ with $\|g_h\|_\infty \le B$. For any policy π,*

$$\begin{aligned}
\sum_{h=0}^{H-1}\mathbb{E}_{(s,a)\sim d^{\pi}_{P^{\star},h}}[g_h(s,a)] \le \sum_{h=1}^{H-1}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^{\star},h-1}}\Bigg[&\|\phi^{\star}_{h-1}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-1},\phi^{\star}_{h-1}}} \\
&\cdot\sqrt{k|\mathcal{A}|\,\mathbb{E}_{(s,a)\sim\rho_h}[g_h^2(s,a)] + B^2\lambda_k d}\Bigg] \\
&+ \sqrt{|\mathcal{A}|\,\mathbb{E}_{(s,a)\sim\rho_0}[g_0^2(s,a)]}.
\end{aligned}$$

Proof. The argument is identical to the proof of Lemma 10, except that we work with the true low-rank decomposition $P^{\star}_h(s'|s,a) = \langle\phi^{\star}_h(s,a), \mu^{\star}_h(s')\rangle$, so the Representation Learning residual $k|\mathcal{A}|^2\zeta_k$ does not appear. The factor $|\mathcal{A}|^2$ in Lemma 10 is replaced by $|\mathcal{A}|$ because the additional importance-sampling step from $\hat{P}$ to $P^{\star}$ is unnecessary here. □

We now establish the composite optimism guarantee, which extends Lemma 2 to the Lagrangian objective.

Lemma 2 *(Restated; composite optimism). Consider an episode $k \in \{1, \dots, K\}$. Conditioned on the Representation Learning Guarantee, with probability at least $1 - \delta$,*

$$Q^{\pi^\star, Y_k}_{\hat{P}_k,\, \bar{r}^{(k)} + \hat{b}^{(k)}}(s,a) \;-\; Q^{\pi^\star, Y_k}_{P^\star,\, \bar{r}^{(k)}}(s,a) \;\geq\; -(1+Y_k)\sqrt{|\mathcal{A}|\zeta_k} \;-\; \rho_k^{\mathrm{opt}},$$

where ρ_k^{opt} is a non-negative residual term satisfying $\sum_{k=1}^K \rho_k^{\mathrm{opt}} \leq (1+\xi)\sum_{k=1}^K \epsilon_k$ (see Lemma 3 for the cumulative bound on $\sum_k \epsilon_k$).

Proof. By the simulation lemma (Lemma 19) applied to the auxiliary MDP with stage reward $\bar{r}^{(k)}$, kernel $\hat{P}_k$, and additive bonus $\hat{b}^{(k)} = (1+Y_k)\hat{b}$,

$$\begin{aligned}
&Q^{\pi^\star, Y_k}_{\hat{P}_k,\, \bar{r}^{(k)} + \hat{b}^{(k)}}(s,a) - Q^{\pi^\star, Y_k}_{P^\star,\, \bar{r}^{(k)}}(s,a) \\
&\quad = \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}} \Big[\hat{b}_h^{(k)}(s_h,a_h) + \mathbb{E}_{s'\sim \hat{P}_h}[V^{\pi^\star, Y_k}_{P^\star,h+1}(s')] - \mathbb{E}_{s'\sim P^\star_h}[V^{\pi^\star, Y_k}_{P^\star,h+1}(s')] \Big].
\end{aligned}$$

Define

$$f_h(s,a) \;:=\; \tfrac{1}{2H+1}\Big(\mathbb{E}_{s'\sim \hat{P}_h}[V^{\pi^\star, Y_k}_{P^\star,h+1}(s')] - \mathbb{E}_{s'\sim P^\star_h}[V^{\pi^\star, Y_k}_{P^\star,h+1}(s')] \Big).$$

Since $V^{\pi^\star, Y_k}_{P^\star,h+1}/(1+Y_k) \in [0,H]$ and the inner expectations are linear in $\hat{\phi}_h$ and $\phi^\star_h$ respectively, $\mathbb{E}_{a\sim\pi_h^{k-1}}[(2H+1)f_h(s,a)/(1+Y_k)] \in \mathcal{F}_h$. The Representation Learning Guarantee bounds the squared error against $\mathbb{E}_{a\sim\pi_h^{k-1}}[\cdot]$; we transfer it to $\mathbb{E}_{a\sim\pi_h^\star}[\cdot]$ via uniform-action importance sampling: for any non-negative $u(s,a)$,

$$\mathbb{E}_{a\sim\pi_h^\star(\cdot|s)}[u(s,a)] \;\leq\; |\mathcal{A}| \cdot \mathbb{E}_{a\sim U(\mathcal{A})}[u(s,a)] \;=\; |\mathcal{A}| \cdot \mathbb{E}_{a\sim\pi_h^{k-1}}\big[u(s,a)\cdot \tfrac{1}{|\mathcal{A}|\pi_h^{k-1}(a|s)}\big],$$

where the second equality holds whenever π_h^{k-1} has full support; for softmax policies, $\pi_h^{k-1}(a|s) \geq \exp(-2\alpha H)/|\mathcal{A}|$, so the importance ratio is bounded. Combined with Cauchy–Schwarz, this yields

$$\mathbb{E}_{(s,a)\sim\rho_h}[f_h^2] \leq (1+Y_k)^2|\mathcal{A}|\,\zeta_k, \qquad \mathbb{E}_{(s,a)\sim\beta_h}[f_h^2] \leq (1+Y_k)^2|\mathcal{A}|\,\zeta_k,$$

where the additional $|\mathcal{A}|$ factor (relative to a direct ζ_k bound) comes from the importance-sampling transfer between $\pi^\star$ and π^{k-1}, and is absorbed into the $|\mathcal{A}|^{3/2}$ scaling of the final rate.

Note that f_h is a signed quantity. Applying Lemma 10 to both $g_h := (2H+1)f_h$ and $g_h := -(2H+1)f_h$ yields a bound on $\big|\sum_{h=0}^{H-1}\mathbb{E}_{d^{\pi^\star}_{\hat{P},h}}[(2H+1)f_h]\big|$:

$$\begin{aligned}
&\Big|\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}}[(2H+1)f_h(s_h,a_h)]\Big| \\
&\leq (2H+1)(1+Y_k)\Bigg[\sum_{h=1}^{H-1}\mathbb{E}\Big[\|\hat{\phi}_{h-1}\|_{\Sigma^{-1}_{\rho_{h-1},\hat{\phi}_{h-1}}}\sqrt{k|\mathcal{A}|^3\zeta_k + 4H^2\lambda_k d + k|\mathcal{A}|^2\zeta_k}\Big] \\
&\qquad + \sqrt{|\mathcal{A}|^2\zeta_k}\Bigg].
\end{aligned}$$

The first sum (over $h = 1, \dots, H-1$) is the per-episode RepLearn residual that we denote ρ_k^{opt}:

$$\rho_k^{\mathrm{opt}} \;:=\; (2H+1)(1+Y_k)\sum_{h=1}^{H-1}\mathbb{E}\Big[\|\hat{\phi}_{h-1}\|_{\Sigma^{-1}_{\rho_{h-1},\hat{\phi}_{h-1}}}\sqrt{2k|\mathcal{A}|^3\zeta_k + 4H^2\lambda_k d}\Big].$$

The bonus $\hat{b}_h(s,a) = \beta\|\hat{\phi}_h\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}$ with $\beta = C_1 dH\sqrt{\iota}$ does *not* dominate ρ_k^{opt} per episode (since $\sqrt{k\zeta_k} = \widetilde{O}(dk^{1/4})$ has a k-dependent factor). Instead, ρ_k^{opt} is carried as a residual into the *cumulative* representation error of Lemma 3.

To sum ρ_k^{opt} over k, the inner expectation is taken under the trajectory measure $d^{\pi^\star}_{\hat{P},h-1}$ of the unexecuted optimal policy $\pi^\star$, while the Gram matrix $\Sigma_{\rho_{h-1},\hat{\phi}_{h-1}}$ is built from data sampled under the executed policies $\pi^0,\dots,\pi^{k-1}$. We bridge the two via uniform-action importance sampling: since the actions in ρ_{h-1} are uniform, for any non-negative function $u(s,a)$,

$$\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\hat{P},h-1}}[u(s,a)] \;\le\; |\mathcal{A}|\cdot \mathbb{E}_{s\sim d^{\pi^\star}_{\hat{P},h-1},\, a\sim U(\mathcal{A})}[u(s,a)] \;\le\; |\mathcal{A}|\cdot \mathbb{E}_{(s,a)\sim \rho^{\pi^\star}_{h-1}}[u(s,a)],$$

where $\rho^{\pi^\star}_{h-1}$ is the $\pi^\star$-induced state distribution paired with uniform action. Combined with Cauchy–Schwarz applied to $\sum_k \|\hat{\phi}_{h-1}\|_{\Sigma^{-1}_{\rho_{h-1}}}$, this reduces the sum to an elliptical-potential expression on the realized data distribution, which is controlled by Lemma 22:

$$\sum_{k=1}^{K}\rho_k^{\mathrm{opt}} \;\le\; (1+\xi)\sum_{k=1}^{K}\epsilon_k \;=\; \widetilde{O}\!\left(\frac{H^3d^2K^{3/4}|\mathcal{A}|^{3/2}}{\gamma}\right).$$

The lower bound $-(1+Y_k)\sqrt{|\mathcal{A}|\zeta_k}-\rho_k^{\mathrm{opt}}$ then comes from the residual Case-1 term plus the residual sum. □

Remark 4. *Operationally, Algorithm 3 computes a single bonus $\hat{b}_h(s,a)=\beta\|\hat{\phi}_h\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}$ per step, which is added to both the reward and cost optimistic Q-functions ($Q^k_{r,h}$ and $Q^k_{g,h}$). The composite Q-function $Q^k_{r,h}+Y_kQ^k_{g,h}$ therefore inherits a composite bonus $(1+Y_k)\hat{b}_h=\hat{b}^{(k)}_h$, exactly as used in Lemma 2. The bonus magnitude $\beta=C_1dH\sqrt{\iota}$ is tuned to the LSVI-UCB concentration scale (Lemma 5) and is independent of k; the RepLearn statistical residual is handled separately through ρ_k^{opt} above and absorbed into $\sum_k\epsilon_k$.*

We next establish the softmax sub-optimality lemma for the *composite* planning problem, which is the actual MDP that the policy update step in Algorithm 1 approximately solves.

Lemma 12 (Composite softmax sub-optimality gap)**.** *Fix $Y\ge 0$ and a finite-horizon planning problem with transition kernel $\hat{P}$, stage reward $\bar{r}:=r+Yg$, additive bonus $\hat{b}^{(Y)}:=(1+Y)\hat{b}$, and horizon H. Let $\{\bar{Q}^\star_h\}_{h=1}^H$ be the optimal Q-functions under $(\hat{P},\,\bar{r}+\hat{b}^{(Y)},\,H)$, defined by*

$$\bar{Q}^\star_h(s,a):=\bar{r}_h(s,a)+\hat{b}^{(Y)}_h(s,a)+\mathbb{E}_{s'\sim\hat{P}_h(\cdot|s,a)}[\bar{V}^\star_{h+1}(s')],$$
$$\bar{V}^\star_h(s):=\max_{a\in\mathcal{A}}\bar{Q}^\star_h(s,a),\qquad \bar{V}^\star_{H+1}(s)\equiv 0.$$

Define the softmax policy

$$\hat{\pi}_h(a|s) \;=\; \frac{\exp(\alpha\bar{Q}^\star_h(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\alpha\bar{Q}^\star_h(s,a'))}.$$

Then for any start state s_1 and any policy π,

$$V^{\pi}_{\hat{P},\,\bar{r}+\hat{b}^{(Y)}}(s_1) \;\le\; V^{\hat{\pi}}_{\hat{P},\,\bar{r}+\hat{b}^{(Y)}}(s_1) \;+\; \frac{H\log|\mathcal{A}|}{\alpha}.$$

Proof. Identical to the proof of the standard softmax sub-optimality bound (e.g., the unconstrained version), with the composite stage reward $\bar{r}+\hat{b}^{(Y)}$ replacing the original reward. We prove by backward induction that for all h and s,

$$\bar{V}^\star_h(s) \;\le\; V^{\hat{\pi}}_{h,\hat{P},\,\bar{r}+\hat{b}^{(Y)}}(s)+\tfrac{(H-h+1)\log|\mathcal{A}|}{\alpha}.$$

The base $h=H+1$ is trivial. For the inductive step, the log-sum-exp inequality gives $\max_a\bar{Q}^\star_h(s,a)\le\mathbb{E}_{a\sim\hat{\pi}_h}[\bar{Q}^\star_h(s,a)]+\frac{\log|\mathcal{A}|}{\alpha}$, which combined with the inductive hypothesis on $\bar{V}^\star_{h+1}$ yields the claim. Setting $h=1$ and using $V^{\pi}_{\hat{P},\bar{r}+\hat{b}^{(Y)}}\le\bar{V}^\star_1$ completes the proof. □

Remark 5 (Why composite, not reward-only)**.** *The policy π^k produced by Algorithm 3 is the softmax of $Q^k_{r,h}+Y_kQ^k_{g,h}$, i.e. the LSVI-UCB estimate of the optimal Q-function for the composite planning MDP $(\hat{P},\bar{r}^{(k)}+\hat{b}^{(k)})$. It is* not *the softmax of the optimal Q-function for $(\hat{P},r+\hat{b})$ alone. Consequently, applying Lemma 12 to the reward-only objective would not be justified, because π^k is not constructed from the reward-only optimal Q. The composite formulation above resolves this issue: π^k is (an approximation to) the softmax of the composite optimal Q, so Lemma 12 applies directly with $Y=Y_k$.*

We now bound the composite Q-function gap on the true MDP, which is the central result of this section. Following the convention of Section 2, we state the result in cumulative form, since this is the form actually used downstream and the form genuinely produced by the elliptical-potential argument.

Lemma 3 *(Restated; composite policy gap, cumulative form). Conditioned on the Representation Learning Guarantee, with probability at least* $1-\delta$*,*

$$\sum_{k=1}^{K}\Big[Q^{\pi^\star,Y_k}_{P^\star,\,\bar r^{(k)}}(s_1,\,\cdot)\;-\;Q^{\pi^k,Y_k}_{P^\star,\,\bar r^{(k)}}(s_1,\,\cdot)\Big]\;\le\;(1+\xi)\sum_{k=1}^{K}\epsilon_k\;+\;\frac{KH\log|\mathcal{A}|}{\alpha},$$

where the cumulative representation error satisfies

$$\sum_{k=1}^{K}\epsilon_k\;\le\;C\,H^2d^2K^{3/4}|\mathcal{A}|^{3/2}\log^{1/4}\!\big(dK|\Phi|/\delta\big).$$

Proof. Throughout, write $\bar r:=\bar r^{(k)}=r+Y_kg$ and $\hat b^{(k)}=(1+Y_k)\hat b$. We chain three inequalities at the per-episode level, then sum.

Step 1: Composite optimism. By Lemma 2 applied with the composite reward,

$$Q^{\pi^\star,Y_k}_{P^\star,\bar r}(s_1,a)\;\le\;Q^{\pi^\star,Y_k}_{\hat P,\,\bar r+\hat b^{(k)}}(s_1,a)+(1+Y_k)\sqrt{|\mathcal{A}|\zeta_k}+\rho_k^{\mathrm{opt}},$$

where the residual ρ_k^{opt} enters $\sum_k\epsilon_k$ in the cumulative bound below.

Step 2: Composite softmax sub-optimality. By Lemma 12 applied with $Y=Y_k$, since π^k is (the LSVI-UCB approximation to) the softmax of the optimal composite Q-function under $(\hat P,\bar r+\hat b^{(k)})$,

$$Q^{\pi^\star,Y_k}_{\hat P,\,\bar r+\hat b^{(k)}}(s_1,a)\;\le\;Q^{\pi^k,Y_k}_{\hat P,\,\bar r+\hat b^{(k)}}(s_1,a)+\frac{H\log|\mathcal{A}|}{\alpha}.$$

Step 3: From planning model $\hat P$ back to true model $P^\star$. Combining Steps 1–2 and using the simulation lemma (Lemma 19) on the auxiliary MDP with kernel $\hat P$, reward $\bar r$, and bonus $\hat b^{(k)}$,

$$\begin{aligned}
&Q^{\pi^\star,Y_k}_{P^\star,\bar r}(s_1,a)-Q^{\pi^k,Y_k}_{P^\star,\bar r}(s_1,a)\\
&\le\sum_{h=0}^{H-1}\mathbb{E}_{(s_h,a_h)\sim d^{\pi^k}_{P^\star,h}}\Big[\hat b_h^{(k)}(s_h,a_h)+\mathbb{E}_{s'\sim\hat P_h}\big[V^{\pi^k,Y_k}_{\hat P,\bar r+\hat b^{(k)},h+1}(s')\big]-\mathbb{E}_{s'\sim P^\star_h}\big[V^{\pi^k,Y_k}_{\hat P,\bar r+\hat b^{(k)},h+1}(s')\big]\Big]\\
&\quad+(1+Y_k)\sqrt{|\mathcal{A}|\zeta_k}+\rho_k^{\mathrm{opt}}+\frac{H\log|\mathcal{A}|}{\alpha}.
\end{aligned}\tag{16}$$

Define

$$f_h(s,a)\;:=\;\mathbb{E}_{s'\sim\hat P_h(\cdot|s,a)}\big[V^{\pi^k,Y_k}_{\hat P,\bar r+\hat b^{(k)},h+1}(s')\big]-\mathbb{E}_{s'\sim P^\star_h(\cdot|s,a)}\big[V^{\pi^k,Y_k}_{\hat P,\bar r+\hat b^{(k)},h+1}(s')\big].$$

The composite value $V^{\pi^k,Y_k}_{\hat P,\bar r+\hat b^{(k)},h+1}$ is bounded in $[0,(1+Y_k)(2H+1)H]$, so $\frac{1}{(2H+1)(1+Y_k)}f_h$ matches the linear part of an element of $\mathcal{G}_h$ up to a multiplicative $(1+Y_k)$ factor; the softmax operator and UCB bonus appearing in $V^{\pi^k,Y_k}_{\hat P,\bar r+\hat b^{(k)},h+1}$ are not in $\mathcal{G}_h$, but their contribution is independently bounded as $H\log|\mathcal{A}|/\alpha$ via Lemma 12 and as the LSVI-UCB residual $\mathrm{lsvi}_K\le\frac14(1+\xi)\sum_k\epsilon_k$ via Lemma 13, both of which we already track in the master bound. Hence by the Representation Learning Guarantee applied to the linear part of f_h,

$$\mathbb{E}_{(s,a)\sim\rho_h}[f_h^2]\le(2H+1)^2(1+Y_k)^2\zeta_k,\qquad\mathbb{E}_{(s,a)\sim\beta_h}[f_h^2]\le(2H+1)^2(1+Y_k)^2\zeta_k.$$

We rewrite (16) as

$$\begin{aligned}
Q^{\pi^\star,Y_k}_{P^\star,\bar r}(s_1,a)-Q^{\pi^k,Y_k}_{P^\star,\bar r}(s_1,a)\le&\underbrace{\sum_{h=0}^{H-1}\mathbb{E}_{d^{\pi^k}_{P^\star,h}}[\hat b_h^{(k)}(s_h,a_h)]}_{\text{(a)}}+\underbrace{\sum_{h=0}^{H-1}\mathbb{E}_{d^{\pi^k}_{P^\star,h}}[f_h(s_h,a_h)]}_{\text{(b)}}\\
&+(1+Y_k)\sqrt{|\mathcal{A}|\zeta_k}+\rho_k^{\mathrm{opt}}+\frac{H\log|\mathcal{A}|}{\alpha}.
\end{aligned}\tag{17}$$

Bounding term (a). Applying Lemma 11 with $g_h = \hat{b}_h^{(k)} = (1+Y_k)\beta\|\hat{\phi}_h\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}$ and $B = O((1+Y_k)\beta\sqrt{d})$,

$$\text{(a)} \le \sum_{h=1}^{H-1} \mathbb{E}_{d^{\pi^k}_{P^\star,h-1}}\Big[\|\phi^\star_{h-1}\|_{\Sigma^{-1}_{\gamma_{h-1},\phi^\star_{h-1}}}\sqrt{k|\mathcal{A}|\,\mathbb{E}_{\rho_h}[(\hat{b}_h^{(k)})^2] + (1+Y_k)^2 4H^2\lambda_k d}\Big] + \sqrt{|\mathcal{A}|\,\mathbb{E}_{\rho_0}[(\hat{b}_0^{(k)})^2]}.$$

Using $\mathbb{E}_{\rho_h}[(\hat{b}_h^{(k)})^2] = (1+Y_k)^2\beta^2\,\mathbb{E}_{\rho_h}[\|\hat{\phi}_h\|^2_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}] \le (1+Y_k)^2\beta^2 d/k$ (by the trace bound $k\,\mathbb{E}_{\rho_h}[\|\hat{\phi}\|^2_{\Sigma^{-1}}] \le d$),

$$\text{(a)} \;\le\; (1+Y_k)\left[\sum_{h=1}^{H-1} \mathbb{E}_{d^{\pi^k}_{P^\star,h-1}}\Big[\|\phi^\star_{h-1}\|_{\Sigma^{-1}_{\gamma_{h-1},\phi^\star_{h-1}}}\sqrt{|\mathcal{A}|\beta^2 d + 4H^2\lambda_k d}\Big] + \sqrt{|\mathcal{A}|\beta^2 d/k}\right].$$

Bounding term (b). Applying Lemma 11 with $g_h = f_h$ and $B = O((1+Y_k)H)$,

$$\text{(b)} \;\le\; (1+Y_k)\Big[\sum_{h=1}^{H-1} \mathbb{E}_{d^{\pi^k}_{P^\star,h-1}}\Big[\|\phi^\star_{h-1}\|_{\Sigma^{-1}_{\gamma_{h-1},\phi^\star_{h-1}}}\sqrt{k|\mathcal{A}|(2H+1)^2\zeta_k + 4H^2\lambda_k d}\Big] + \sqrt{|\mathcal{A}|(2H+1)^2\zeta_k}\Big].$$

Summing over k via the elliptical potential lemma. Using $\gamma_h^k(s,a) = \frac{1}{k}\sum_{i=0}^{k-1} d_h^{\pi^i}(s,a)$ and the executed-policy mixture $d^{\pi^k}_{P^\star,h-1}$ on the LHS, we apply uniform-action importance sampling to bound

$$\mathbb{E}_{(s,a)\sim d^{\pi^k}_{P^\star,h-1}}\big[\|\phi^\star_{h-1}\|_{\Sigma^{-1}_{\gamma^k_{h-1},\phi^\star_{h-1}}}\big] \;\le\; \sqrt{|\mathcal{A}|}\cdot\mathbb{E}_{(s,a)\sim\gamma^k_{h-1}}\big[\|\phi^\star_{h-1}\|_{\Sigma^{-1}_{\gamma^k_{h-1},\phi^\star_{h-1}}}\big],$$

since the actions in γ^k_{h-1} follow the executed policies' marginals (paired with the on-policy state distribution) and the importance ratio between d^{π^k} and γ^k_{h-1} on the action variable is bounded by $|\mathcal{A}|$. Then

$$\sum_{k=1}^{K}\mathbb{E}_{\gamma^k_{h-1}}\Big[\|\phi^\star_{h-1}\|_{\Sigma^{-1}_{\gamma^k_{h-1},\phi^\star_{h-1}}}\Big] \le \sqrt{K\sum_{k=1}^{K}\mathbb{E}_{\gamma^k_{h-1}}[(\phi^\star_{h-1})^\top\Sigma^{-1}_{\gamma^k_{h-1},\phi^\star_{h-1}}\phi^\star_{h-1}]} \le \sqrt{dK\ln(1+\tfrac{K}{d\lambda_1})},$$

where the second inequality uses Lemma 22 (matrix determinant lemma) on the matched data distribution. The $\sqrt{|\mathcal{A}|}$ factor from importance sampling combines with the $\sqrt{|\mathcal{A}|}$ inside the integrand to give the $|\mathcal{A}|^{3/2}$ scaling of the final bound.

Combining. Plugging in $\zeta_k = O(d^2\sqrt{\log(k|\Phi|/\delta)/k})$ and $\lambda_k = O(d\log(|\Phi|k/\delta))$, summing terms $\text{(a)}+\text{(b)}+\rho_k^{\text{opt}}$ over k, and pulling the $(1+Y_k)$ factor out (which is bounded by $(1+\xi)$), we obtain

$$\sum_{k=1}^{K}\Big[Q^{\pi^\star,Y_k}_{P^\star,\bar{r}}(s_1,a) - Q^{\pi^k,Y_k}_{P^\star,\bar{r}}(s_1,a)\Big] \;\le\; (1+\xi)\sum_{k=1}^{K}\epsilon_k + \frac{KH\log|\mathcal{A}|}{\alpha},$$

where the cumulative representation error satisfies

$$\sum_{k=1}^{K}\epsilon_k \;\le\; C\,H^2 d^{3/2}\sqrt{K}\,|\mathcal{A}|^{1/2}\sqrt{\log(dK|\Phi|/\delta)} \;+\; C'\,H^2 d^2 K^{3/4}|\mathcal{A}|^{3/2}\log^{1/4}(dK|\Phi|/\delta).$$

The dominant term is the second one, of order $\widetilde{O}(H^2d^2K^{3/4}|\mathcal{A}|^{3/2})$, completing the proof. □

Remark 6 (Per-episode rate)**.** *The cumulative bound $\sum_k \epsilon_k = O(K^{3/4})$ corresponds to an average per-episode rate $\bar{\epsilon}_K = \frac{1}{K}\sum_k \epsilon_k = \widetilde{O}(H^2 d^2 K^{-1/4}|\mathcal{A}|^{3/2})$. Throughout the analysis, only the cumulative quantity $(1+\xi)\sum_k \epsilon_k$ is used (e.g., in Lemma 4 and the master inequality), so we work directly with the cumulative form.*

We now establish an auxiliary lemma that quantifies how close the LSVI-UCB output $V_1^{(k),Y_k}(s_1)$ is to the true Bellman-optimal value $\bar{V}_1^\star(s_1)$ on the planning model $(\hat{P},\, \bar{r}^{(k)} + \hat{b}^{(k)})$. This is the LSVI-UCB analog of the per-step backup bound (Lemma 6), but applied at the start state s_1 and going *from* the LSVI iterate *to* the planning-model optimal value, rather than the other direction.

Lemma 13 (LSVI-UCB output vs. planning-model optimum)**.** *On the event $\mathcal{E}$ of Lemma 5, for every episode $k \in [K]$ and every $s \in \mathcal{S}$,*

$$\left|\, \bar{V}_1^\star(s) - \bar{V}_1^{(k),Y_k}(s) \,\right| \;\le\; \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d^{\bar{\pi}^k}_{\hat{P},h}\left[\,\cdot \,\middle|\, s_1=s\right]}\Big[(1+Y_k)\, C_3 dH\sqrt{\chi}\, \|\hat{\phi}_h^k(s_h,a_h)\|_{(\Sigma_h^k)^{-1}}\Big],$$

where $\bar{V}_h^{(k),Y_k}(s) := \max_a Q_h^{(k),Y_k}(s,a)$ is the max-version of the LSVI iterate, $\tilde{\pi}^k$ is an optimal policy of the planning model $\hat{P}$ with stage reward $(1+Y_k)C_3 dH\sqrt{\chi}\,\|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}$, and C_3, χ are as in Lemma 6. Consequently, applying Lemma 11 with the bonus integrand and summing over $k \in [K]$ via the elliptical-potential lemma,

$$\mathrm{lsvi}_K \;:=\; \sum_{k=1}^{K} \left|\bar{V}_1^\star(s_1) - \bar{V}_1^{(k),Y_k}(s_1)\right| \;\le\; \widetilde{O}\left(\frac{d^{3/2}H^2\sqrt{K}\,|\mathcal{A}|^{1/2}}{\gamma}\right) \;\le\; \tfrac{1}{4}(1+\xi)\sum_{k=1}^{K}\epsilon_k.$$

The bound also holds for the softmax variant $V_1^{(k),Y_k}(s) = \sum_a \pi_1^k(a|s) Q_1^{(k),Y_k}(s,a)$ in place of $\bar{V}_1^{(k),Y_k}$, up to an additive $H\log|\mathcal{A}|/\alpha$ term (softmax sub-optimality on the LSVI iterate itself).

Proof. We prove the per-state bound by backward induction on h, then derive the cumulative bound by elliptical-potential summation.

Step 1: Per-step LSVI concentration. By Lemma 5 (event $\mathcal{E}$), for each $j \in \{r, g\}$, $h \in [H]$, and $k \in [K]$, the LSVI weight $w_{j,h}^k$ satisfies, for every $(s,a) \in \mathcal{S}\times\mathcal{A}$:

$$\left|\langle \hat{\phi}_h^k(s,a), w_{j,h}^k\rangle - \big(c_{j,h}(s,a) + \mathbb{E}_{\hat{P}_h}[V_{j,h+1}^k](s,a)\big)\right| \le C_3\, dH\sqrt{\chi}\,\|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}, \tag{18}$$

where $C_3 = O(\sqrt{\lambda} + C_2)$ absorbs the regularization and concentration constants from Lemma 6, and the kernel-mismatch contribution between $P^\star$ and $\hat{P}$ is bundled into C_3 via the Representation Learning Guarantee (Lemma 16). The reward $c_{r,h} = r_h$ and cost $c_{g,h} = g_h$ targets are deterministic.

Step 2: Composite LSVI concentration. Recall $Q_h^{(k),Y_k}(s,a) = Q_{r,h}^k(s,a) + Y_k Q_{g,h}^k(s,a)$ where each $Q_{j,h}^k$ is the LSVI iterate from Algorithm 3 (truncated by H but we ignore truncation here as it only tightens the bound). The composite stage reward is $\bar{r}_h^{(k)} = r_h + Y_k g_h$, and the composite bonus is $\hat{b}_h^{(k)} = (1+Y_k)\hat{b}_h^k$ where $\hat{b}_h^k(s,a) = \beta\|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}$.

By Algorithm 3 lines 7–8, the LSVI iterate satisfies

$$Q_{j,h}^k(s,a) = \langle \hat{\phi}_h^k(s,a), w_{j,h}^k\rangle + \hat{b}_h^k(s,a),$$

ignoring the truncation $\min\{\cdot, H\}$ (which removes only over-estimates and does not affect the upper bound we derive). Combining with (18), weighted by 1 and Y_k for the reward and cost components respectively:

$$\begin{aligned}\Big|Q_h^{(k),Y_k}(s,a) - \big[\bar{r}_h^{(k)}(s,a) + \hat{b}_h^{(k)}(s,a) + \mathbb{E}_{\hat{P}_h}\big[V_{r,h+1}^k + Y_k V_{g,h+1}^k\big](s,a)\big]\Big| \\ \le (1+Y_k)\, C_3\, dH\sqrt{\chi}\,\|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}.\end{aligned}$$

Since $V_{h+1}^{(k),Y_k}(s) := V_{r,h+1}^k(s) + Y_k V_{g,h+1}^k(s)$ is the composite LSVI value (softmax form) and $\bar{V}_{h+1}^{(k),Y_k}(s) = \max_a Q_{h+1}^{(k),Y_k}(s,a)$ is the max form, we work with the max form throughout this proof and handle the softmax-vs-max gap in Step 5. Replacing $V_{r,h+1}^k + Y_k V_{g,h+1}^k$ by $\bar{V}_{h+1}^{(k),Y_k}$ in the displayed inequality (which holds with the same RHS by the same per-objective argument applied to the max-version of the LSVI recursion), we obtain

$$\begin{aligned}\Big|Q_h^{(k),Y_k}(s,a) - \big[\bar{r}_h^{(k)}(s,a) + \hat{b}_h^{(k)}(s,a) + \mathbb{E}_{\hat{P}_h}[\bar{V}_{h+1}^{(k),Y_k}](s,a)\big]\Big| \\ \le (1+Y_k)\, C_3\, dH\sqrt{\chi}\, \|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}.\end{aligned} \tag{19}$$

Step 3: Backward induction on the per-state gap. Define

$$R_h(s) := \big|\bar{V}_h^\star(s) - \bar{V}_h^{(k),Y_k}(s)\big|, \quad h \in [H+1],$$

where $\bar{V}_{H+1}^\star \equiv 0$ and $\bar{V}_{H+1}^{(k),Y_k} \equiv 0$, so $R_{H+1}(s) = 0$.

Write $c_h(s,a) := (1+Y_k)\, C_3\, dH\sqrt{\chi}\, \|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}$, and let W_h be the optimal value function of the planning model $\hat{P}$ with stage reward c_h, that is, $W_{H+1} \equiv 0$ and:

$$W_h(s) = \max_{a\in\mathcal{A}} \Big[c_h(s,a) + \mathbb{E}_{s'\sim\hat{P}_h(\cdot|s,a)}[W_{h+1}(s')]\Big].$$

Let $\tilde{\pi}_h^k(s)$ be a maximizer in this recursion. Then $W_h(s)$ is the expected cumulative stage reward c of $\tilde{\pi}^k$ on $\hat{P}$ from $s_h = s$. We claim that for all $h \in [H+1]$ and all $s \in \mathcal{S}$:

$$R_h(s) \le W_h(s) = \mathbb{E}_{(s_t,a_t)_{t=h}^H \sim d_{\hat{P},h:H}^{\tilde{\pi}^k}(\,\cdot\,|s_h=s)}\left[\sum_{t=h}^H c_t(s_t,a_t)\right], \tag{20}$$

where $d_{\hat{P},h:H}^{\tilde{\pi}^k}(\,\cdot\,|s_h = s)$ denotes the joint distribution of $(s_h,a_h),\ldots,(s_H,a_H)$ under policy $\tilde{\pi}^k$ and kernel $\hat{P}$, conditioned on $s_h = s$.

Base case ($h = H+1$): $R_{H+1}(s) = 0 = W_{H+1}(s)$.

Inductive step: Assume (20) holds for $h+1$. By the Bellman equation for $\bar{V}_h^\star$ on the planning model $(\hat{P}, \bar{r}^{(k)} + \hat{b}^{(k)})$:

$$\bar{Q}_h^\star(s,a) = \bar{r}_h^{(k)}(s,a) + \hat{b}_h^{(k)}(s,a) + \mathbb{E}_{\hat{P}_h}[\bar{V}_{h+1}^\star](s,a).$$

Subtracting (19) and applying the triangle inequality, for every $a \in \mathcal{A}$:

$$\big|\bar{Q}_h^\star(s,a) - Q_h^{(k),Y_k}(s,a)\big| \le \mathbb{E}_{s'\sim\hat{P}_h(\cdot|s,a)}[R_{h+1}(s')] + c_h(s,a).$$

Since $\bar{V}_h^\star(s) = \max_a \bar{Q}_h^\star(s,a)$ and $\bar{V}_h^{(k),Y_k}(s) = \max_a Q_h^{(k),Y_k}(s,a)$, the contraction property $|\max_a f(a) - \max_a g(a)| \le \max_a |f(a) - g(a)|$ of the maximum over a finite set, followed by the inductive hypothesis $R_{h+1} \le W_{h+1}$, gives:

$$\begin{aligned}R_h(s) &\le \max_{a\in\mathcal{A}} \Big[\mathbb{E}_{s'\sim\hat{P}_h(\cdot|s,a)}[R_{h+1}(s')] + c_h(s,a)\Big] \\ &\le \max_{a\in\mathcal{A}} \Big[\mathbb{E}_{s'\sim\hat{P}_h(\cdot|s,a)}[W_{h+1}(s')] + c_h(s,a)\Big] = W_h(s).\end{aligned}$$

This establishes (20) for h and completes the induction. Setting $h = 1$, we obtain the per-state bound stated in the lemma.

Step 4: Cumulative bound via elliptical-potential summation. Setting $s = s_1$ in the per-state bound and summing over $k \in [K]$:

$$\mathrm{lsvi}_K = \sum_{k=1}^K R_1(s_1) \le \sum_{k=1}^K \sum_{h=1}^H (1+Y_k)\, C_3\, dH\sqrt{\chi}\, \mathbb{E}_{(s_h,a_h)\sim d_{\hat{P},h}^{\tilde{\pi}^k}}\big[\|\hat{\phi}_h^k(s_h,a_h)\|_{(\Sigma_h^k)^{-1}}\big]. \tag{21}$$

To control each summand, we apply Lemma 11 to the function $g_h(s,a) := C_3 dH\sqrt{\chi}\,\|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}$. Note that $\|g_h\|_\infty \le C_3 dH\sqrt{\chi}/\sqrt{\lambda} = O(dH\sqrt{\chi})$ since $\Sigma_h^k \succeq \lambda I$. By Lemma 11 (applied with policy $\pi = \tilde{\pi}^k$ and kernel $P^\star$),[2]

$$\sum_{h=1}^{H} \mathbb{E}_{d^{\tilde{\pi}^k}_{\hat{P},h}}\big[\|\hat{\phi}_h^k\|_{(\Sigma_h^k)^{-1}}\big] \le \sqrt{|\mathcal{A}|}\cdot \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim\rho_h^k}\big[\|\hat{\phi}_h^k(s,a)\|^2_{(\Sigma_h^k)^{-1}}\big]^{1/2} + \text{(lower-order terms)},$$

where the factor $\sqrt{|\mathcal{A}|}$ comes from the importance-sampling step from $d^{\tilde{\pi}^k}_{\hat{P},h}$ to ρ_h^k (uniform-action distribution), and the lower-order terms arise from the regularization λ_k as in Lemmas 10–11.

By Cauchy–Schwarz,

$$\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim\rho_h^k}\big[\|\hat{\phi}_h^k(s,a)\|^2_{(\Sigma_h^k)^{-1}}\big]^{1/2} \le \sqrt{K\cdot \sum_{k=1}^{K} \mathbb{E}_{\rho_h^k}\big[\|\hat{\phi}_h^k\|^2_{(\Sigma_h^k)^{-1}}\big]}.$$

By Lemma 22 (matrix determinant / elliptical potential lemma),

$$\sum_{k=1}^{K} \mathbb{E}_{\rho_h^k}\big[\|\hat{\phi}_h^k\|^2_{(\Sigma_h^k)^{-1}}\big] \;\le\; 2\ln\frac{\det(\Sigma_h^{K+1})}{\det(\Sigma_h^1)} \;\le\; 2d\ln\big(1+K/(d\lambda)\big) \;=\; \widetilde{O}(d).$$

This is the rigorous form of the elliptical-potential bound. Substituting,

$$\sum_{k=1}^{K} \mathbb{E}_{\rho_h^k}\big[\|\hat{\phi}_h^k\|^2_{(\Sigma_h^k)^{-1}}\big]^{1/2} \;\le\; \sqrt{K\cdot\widetilde{O}(d)} \;=\; \widetilde{O}(\sqrt{dK}).$$

Putting these together, summing over $h\in[H]$ and noting $1+Y_k \le 1+\xi$:

$$\begin{aligned}
\mathrm{lsvi}_K &\le (1+\xi)\, C_3\, dH\sqrt{\chi}\,\sqrt{|\mathcal{A}|}\sum_{h=1}^{H}\widetilde{O}(\sqrt{dK}) + \text{(lower-order)}\\
&= \widetilde{O}\big((1+\xi)\, d^{3/2}H^2\sqrt{K}\,|\mathcal{A}|^{1/2}\big)\\
&= \widetilde{O}\left(\frac{d^{3/2}H^2\sqrt{K}\,|\mathcal{A}|^{1/2}}{\gamma}\right),
\end{aligned}$$

where the last equality uses $1+\xi = 1+2H/\gamma = O(H/\gamma)$ (treating H as part of the polynomial factor).

Step 5: Domination by $\sum_k \epsilon_k$. By Lemma 3, the cumulative representation error satisfies

$$\sum_{k=1}^{K}\epsilon_k \;\ge\; C\,H^2\,d^2\,K^{3/4}\,|\mathcal{A}|^{3/2}\,\log^{1/4}(dK|\Phi|/\delta)$$

for some constant $C>0$. Comparing exponents of K, d, and $|\mathcal{A}|$:

- K: $\sqrt{K} = K^{1/2}$ vs. $K^{3/4}$ – the latter dominates for $K \ge 1$.
- d: $d^{3/2}$ vs. d^2 – the latter dominates for $d \ge 1$.
- $|\mathcal{A}|$: $|\mathcal{A}|^{1/2}$ vs. $|\mathcal{A}|^{3/2}$ – the latter dominates for $|\mathcal{A}| \ge 1$.

All exponents go in our favor, so for any $K \ge 1$ and any $d, |\mathcal{A}| \ge 1$,

$$\widetilde{O}\left(\frac{d^{3/2}H^2\sqrt{K}\,|\mathcal{A}|^{1/2}}{\gamma}\right) \le \tfrac{1}{4}\,(1+\xi)\sum_{k=1}^{K}\epsilon_k,$$

where the constant $\frac{1}{4}$ is achieved by absorbing the dimension/cardinality factors into the leading constant of $\sum_k \epsilon_k$ (the constant C in Lemma 3 is large enough to accommodate this; equivalently, the relation holds for all sufficiently large K, and the small-K regime is handled by the regret bound's trivial HK ceiling).

[2] Lemma 11 bounds expectations under $d^{\pi}_{P^\star}$ in terms of empirical distributions ρ_h^k. Here we apply it with $\hat{P}$ in place of $P^\star$; the kernel-mismatch contribution is again absorbed into C_3 via the Representation Learning Guarantee, exactly as in the proof of Lemma 3, term (a).

Step 6: Softmax variant. The bound stated above is for $\bar{V}_1^{(k),Y_k}(s) := \max_a Q_1^{(k),Y_k}(s,a)$. Algorithm 3 actually outputs the softmax variant

$$V_1^{(k),Y_k}(s) := \sum_a \pi_1^k(a|s)\, Q_1^{(k),Y_k}(s,a), \quad \pi_1^k(a|\cdot) = \mathrm{softmax}_\alpha(Q_1^{(k),Y_k}(\cdot,a)).$$

Applying Lemma 12 (softmax sub-optimality) to the planning problem $(\hat{P}, \bar{r}^{(k)} + \hat{b}^{(k)})$ at horizon 1 yields[3]

$$\left|\bar{V}_1^{(k),Y_k}(s) - V_1^{(k),Y_k}(s)\right| \le \frac{H \log|\mathcal{A}|}{\alpha}.$$

Combining with the per-state bound for $\bar{V}_1^{(k),Y_k}$ via the triangle inequality:

$$\left|\bar{V}_1^{\star}(s) - V_1^{(k),Y_k}(s)\right| \le \left|\bar{V}_1^{\star}(s) - \bar{V}_1^{(k),Y_k}(s)\right| + \frac{H \log|\mathcal{A}|}{\alpha},$$

which gives the softmax-variant bound asserted in the lemma. Summing over k, the additional term contributes $KH \log|\mathcal{A}|/\alpha$, which (with $\alpha = K \log|\mathcal{A}|/(2(1+\xi+H))$) equals $2H(1+\xi+H) = O\left(\frac{H^2}{\gamma}\right)$, also dominated by $\sum_k \epsilon_k$. □

We now derive the cumulative pseudo regret bound used in Step 1 of the proof of Theorem 3. Recall from the master decomposition (9) that the Pseudo regret quantifies the gap between $\pi^\star$ and the LSVI-UCB optimistic estimate V^k, while the gap between V^k and $V_{P^\star}^{\pi^k}$ is handled separately as the Estimation error in Appendix G.

Lemma 4 *(Pseudo regret bound). Conditioned on the Representation Learning Guarantee and the event $\mathcal{E}$, with probability at least $1-\delta$,*

$$\sum_{k=1}^{K}\left[\left(V_{r,1}^{\pi^\star}(s_1) + Y_k V_{g,1}^{\pi^\star}(s_1)\right) - \left(V_{r,1}^{k}(s_1) + Y_k V_{g,1}^{k}(s_1)\right)\right] \;\le\; \frac{KH\log|\mathcal{A}|}{\alpha} + (1+\xi)\sum_{k=1}^{K}\epsilon_k.$$

Proof. Using the composite notation $V_{P^\star,1}^{\pi,Y_k}(s_1) = V_{r,1}^{\pi}(s_1) + Y_k V_{g,1}^{\pi}(s_1)$ and writing $V_1^{(k),Y_k}(s_1) := V_{r,1}^{k}(s_1) + Y_k V_{g,1}^{k}(s_1)$ for the LSVI-UCB optimistic estimate of the composite value, the per-episode cumulant on the LHS is $V_{P^\star,1}^{\pi^\star,Y_k}(s_1) - V_1^{(k),Y_k}(s_1)$. We chain three inequalities at the per-episode level, then sum.

Step 1: Bridge LSVI iterate $V_1^{(k),Y_k}$ to planning-model optimum $\bar{V}_1^\star$. Let $\bar{V}_h^\star(s) := \max_a \bar{Q}_h^\star(s,a)$ denote the value function of the true Bellman-optimal Q-function on the planning model $(\hat{P}, \bar{r}^{(k)} + \hat{b}^{(k)})$, as defined in Lemma 12. The LSVI-UCB output is $V_1^{(k),Y_k}(s_1) = \sum_a \pi_1^k(a|s_1) Q_1^{(k),Y_k}(s_1,a)$, where $Q_h^{(k),Y_k} := Q_{r,h}^k + Y_k Q_{g,h}^k$ is the LSVI estimate of the composite Bellman-optimal Q-function. By Lemma 13, on event $\mathcal{E}$,

$$\left|\bar{V}_1^{\star}(s_1) - V_1^{(k),Y_k}(s_1)\right| \;\le\; \mathrm{lsvi}^{(k)} + \frac{H\log|\mathcal{A}|}{\alpha}, \tag{22}$$

where $\mathrm{lsvi}^{(k)}$ is the per-episode LSVI residual of Lemma 13 (it sums to $\mathrm{lsvi}_K \le \frac{1}{4}(1+\xi)\sum_k \epsilon_k$), and the additional $H\log|\mathcal{A}|/\alpha$ accounts for the softmax-vs-max discrepancy between $V_1^{(k),Y_k}$ (softmax of $Q_1^{(k),Y_k}$) and $\bar{V}_1^{(k),Y_k} := \max_a Q_1^{(k),Y_k}(s_1,\cdot)$.

Step 2: Optimism from $\pi^\star$ on $P^\star$ to $\bar{V}_1^\star$ on $\hat{P}+\hat{b}^{(k)}$. By Lemma 2 (composite optimism), promoted to the V-function level via the simulation lemma at $h=1$ (as derived in Step 3 of the proof of Lemma 3, equations (16)–(17), applied with $\pi = \pi^\star$ in place of π^k),

$$V_{P^\star,\bar{r}^{(k)}}^{\pi^\star,Y_k}(s_1) \;\le\; V_{\hat{P},\,\bar{r}^{(k)}+\hat{b}^{(k)}}^{\pi^\star,Y_k}(s_1) + (1+Y_k)\sqrt{|\mathcal{A}|\zeta_k} + \rho_k^{\mathrm{opt}}.$$

[3] Lemma 12 is stated for the optimal Q-function. Applied to LSVI iterates, the same telescoping argument gives the same per-step $\log|\mathcal{A}|/\alpha$ slack, summing to $H\log|\mathcal{A}|/\alpha$.

Since $\bar{V}_1^\star(s_1) = \max_a \bar{Q}_1^\star(s_1,a) \geq \sum_a \pi_1^\star(a|s_1)\bar{Q}_1^\star(s_1,a) = V^{\pi^\star,Y_k}_{\hat{P},\,\bar{r}^{(k)}+\hat{b}^{(k)}}(s_1)$, combining gives

$$V^{\pi^\star,Y_k}_{P^\star,\bar{r}^{(k)}}(s_1) \;\leq\; \bar{V}_1^\star(s_1) + (1+Y_k)\sqrt{|\mathcal{A}|\zeta_k} + \rho_k^{\mathrm{opt}}. \tag{23}$$

Step 3: Bridge $\bar{V}_1^\star$ to $V_1^{(k),Y_k}$. Combining (23) with (22),

$$V^{\pi^\star,Y_k}_{P^\star,\bar{r}^{(k)}}(s_1) - V_1^{(k),Y_k}(s_1) \;\leq\; (1+Y_k)\sqrt{|\mathcal{A}|\zeta_k} + \rho_k^{\mathrm{opt}} + \mathrm{lsvi}^{(k)} + \frac{H\log|\mathcal{A}|}{\alpha}.$$

Step 4: Summing over $k \in [K]$. Using $1+Y_k \leq 1+\xi$ and the cumulative bounds $\sum_k(1+Y_k)\sqrt{|\mathcal{A}|\zeta_k} \leq (1+\xi)\sum_k\sqrt{|\mathcal{A}|\zeta_k}$, $\sum_k \rho_k^{\mathrm{opt}} \leq (1+\xi)\sum_k\epsilon_k$ (Lemma 2), and $\sum_k \mathrm{lsvi}^{(k)} = \mathrm{lsvi}_K \leq \frac{1}{4}(1+\xi)\sum_k\epsilon_k$ (Lemma 13), the first three residual terms are absorbed into $(1+\xi)\sum_k\epsilon_k$ via the same elliptical-potential argument used in Lemma 3 (the $\sqrt{|\mathcal{A}|\zeta_k}$ contribution is strictly smaller than ρ_k^{opt} and is dominated). Therefore

$$\sum_{k=1}^K \big(V^{\pi^\star,Y_k}_{P^\star,1}(s_1) - V_1^{(k),Y_k}(s_1)\big) \;\leq\; (1+\xi)\sum_{k=1}^K\epsilon_k + \frac{KH\log|\mathcal{A}|}{\alpha},$$

which is the claimed bound. The argument uses (i) composite optimism (Lemma 2) and $\max \geq$ average to bridge $V^{\pi^\star}_{P^\star,\bar{r}}$ to $\bar{V}^\star$ on $(\hat{P}, \bar{r}+\hat{b}^{(k)})$; (ii) LSVI-UCB self-normalized concentration (Lemma 5) to bridge $\bar{V}^\star$ to the LSVI iterate $V_1^{(k),Y_k}$, with the LSVI residual quantified in Lemma 13; and (iii) the simulation lemma (Lemma 19) and Representation Learning Guarantee (Lemma 16) to control kernel mismatch between $\hat{P}$ and $P^\star$. It does *not* require any Bellman-optimality property of $\pi^\star$ for the composite objective, which would be incorrect since $\pi^\star$ is the optimal policy for the *constrained* problem and is not in general greedy with respect to $r + Y_k g$ for $Y_k \neq Y^\star$. □

G Bounding Estimation Error

Definition 4. For $j \in \{r, g\}$, write $c_{j,h}$ for the per-objective stage cost ($c_{r,h} := r_h$, $c_{g,h} := g_h$), and denote the expected value of the next-state value function under policy π as

$$\mathbb{P}_h^\star V^\pi_{j,h+1}(s,a) := \mathbb{E}_{s'\sim P_h^\star(\cdot|s,a)}\left[V^\pi_{j,h+1}(s')\right].$$

With this shorthand, the Bellman equation for policy π reads $Q^\pi_{j,h}(s,a) = c_{j,h}(s,a) + \mathbb{P}_h^\star V^\pi_{j,h+1}(s,a)$, and the value function satisfies $V^\pi_{j,h}(s) = \sum_{a\in\mathcal{A}}\pi_h(a \mid s)\, Q^\pi_{j,h}(s,a)$.

Definition 5. For any step h and episode k, define

$$D^k_{j,h,1} = \big\langle Q^k_{j,h}(s^k_h,\cdot) - Q^{\pi^k}_{j,h}(s^k_h,\cdot),\, \pi^k_h(\cdot \mid s^k_h)\big\rangle - \big(Q^k_{j,h}(s^k_h,a^k_h) - Q^{\pi^k}_{j,h}(s^k_h,a^k_h)\big),$$
$$D^k_{j,h,2} = \mathbb{P}_h^\star\big(V^k_{j,h+1} - V^{\pi^k}_{j,h+1}\big)(s^k_h,a^k_h) - \big(V^k_{j,h+1} - V^{\pi^k}_{j,h+1}\big)(s^k_{h+1}).$$

By construction, $D^k_{j,h,1}$ has zero mean conditioned on s^k_h, and $D^k_{j,h,2}$ has zero mean conditioned on (s^k_h, a^k_h). The actions a^k_h here are the on-policy actions sampled in line 10 of Algorithm 1, independent of the uniform-action data $\tilde{a}^k_h$ used to build Σ^k_h (line 4).

Lemma 5 *(Restated; concentration of empirical Bellman residual). There exists a constant C_2 such that for any fixed $p \in (0,1)$, the event $\mathcal{E}$ defined as*

$$\mathcal{E} := \bigcap_{k\in[K],\, h\in[H],\, j\in\{r,g\}} \left\{ \Big\| \sum_{\tau=1}^{k-1} \hat{\phi}^\tau_h\big[V^k_{j,h+1}(s^\tau_{h+1}) - \mathbb{P}_h^\star V^k_{j,h+1}(s^\tau_h,\tilde{a}^\tau_h)\big] \Big\|_{(\Sigma^k_h)^{-1}} \leq C_2\, dH\sqrt{\chi} \right\}$$

holds with $\chi = \log\big(4(C_1+1)\log|\mathcal{A}|\, dT|\Phi|/p\big)$, and $\Pr(\mathcal{E}) \geq 1 - p/4$.

Proof. The challenge is that both $\hat{\phi}^\tau_h$ (which is the current-episode feature $\hat{\phi}^k_h$ evaluated at past data $(s^\tau_h, \tilde{a}^\tau_h)$) and $V^k_{j,h+1}$ are data-dependent: neither is $\mathcal{F}_{\tau-1}$-measurable for $\tau < k$, so a direct

application of self-normalized concentration to the adaptive sequence $(\hat{\phi}_h^\tau)_\tau$ does not apply. We address this by uniformly covering the entire optimistic value-function class together with the feature class, then applying the fixed-feature concentration of Ghosh et al. [11, Lemma B.3] to each element of the cover.

Optimistic value class. By construction (Algorithm 3), each $V_{j,h+1}^k$ takes the form

$$V_{j,h+1}^k(s) \;=\; \sum_a \pi_{h+1}^k(a \mid s)\min\big\{\langle \hat{\phi}_{h+1}^k(s,a), w\rangle + \tilde{\beta}\sqrt{\hat{\phi}_{h+1}^k(s,a)^\top \Lambda^{-1}\hat{\phi}_{h+1}^k(s,a)},\; H\big\},$$

parameterized by $\hat{\phi}_{h+1}^k \in \Phi_{h+1}$, a weight $w \in \mathbb{R}^d$ with $\|w\|_2 \le 2H\sqrt{d}$, a regularization-Gram matrix $\Lambda \succeq I$ with $\|\Lambda\| \le \lambda + K$, a bonus parameter $\tilde{\beta} \in [0, \beta]$, and a softmax temperature dependent on $Y_k \in [0, \xi]$. Let $\mathcal{V}_{h+1}$ denote the class of all such functions over the bounded parameter space.

Covering. For $\varepsilon = 1/(KH)$, take an ε-cover of: Φ_{h+1} (size $|\Phi|$ when Φ is finite, or $\mathcal{N}_\varepsilon(\Phi)$ in general); the ℓ_2-ball $\{w : \|w\|_2 \le 2H\sqrt{d}\}$ (size $(8H\sqrt{d}/\varepsilon)^d$); the matrix ball $\{\Lambda : I \preceq \Lambda,\ \|\Lambda\| \le \lambda + K\}$ (size $(\lambda+K)^{d^2}/\varepsilon^{d^2}$); the bonus interval $[0, \beta]$ (size β/ε); and $[0, \xi]$ (size ξ/ε). The product gives a cover $\mathcal{V}_{h+1,\varepsilon}$ of $\mathcal{V}_{h+1}$ with

$$\log|\mathcal{V}_{h+1,\varepsilon}| \;\le\; \log|\Phi| + (d^2+d+2)\log\big((K+\beta+\xi+\lambda)/\varepsilon\big) \;\le\; \log|\Phi| + O(d^2\log(KHd)).$$

Fixed-feature concentration on the cover. For each fixed $V \in \mathcal{V}_{h+1,\varepsilon}$ *and* each fixed $\phi \in \Phi$ in the cover, the self-normalized concentration of Ghosh et al. [11, Lemma B.3] applies to the sequence $(\phi(s_h^\tau, \tilde{a}_h^\tau))_{\tau \le k-1}$, since ϕ is fixed (not data-dependent) and the historical data $(s_h^\tau, \tilde{a}_h^\tau, s_{h+1}^\tau)_{\tau \le k-1}$ generates a standard adaptive sequence with respect to its natural filtration. This gives the stated inequality with probability at least $1 - p_0$, where $p_0 = p/(4|\mathcal{V}_{h+1,\varepsilon}||\Phi|HK \cdot 2)$.

Discretization error. For any realized $(\hat{\phi}_h^k, V_{j,h+1}^k)$, let $(\bar{\phi}, \bar{V})$ be the nearest neighbor in $\Phi \times \mathcal{V}_{h+1,\varepsilon}$ (so $\|\hat{\phi}_h^k - \bar{\phi}\|_\infty \le \varepsilon$ and $\|V_{j,h+1}^k - \bar{V}\|_\infty \le \varepsilon$). The combined approximation error contributes at most $K\varepsilon \cdot 2\sqrt{d} = 2\sqrt{d}/H$ in the self-normalized norm, which is dominated by the main concentration term.

Union bound. Taking a union bound over $V \in \mathcal{V}_{h+1,\varepsilon}$, $\phi \in \Phi$, $h \in [H]$, $j \in \{r, g\}$, and $k \in [K]$, the event $\mathcal{E}$ holds with probability at least $1 - p/4$. Substituting p_0 and absorbing constants into C_2, the threshold becomes $\chi = \log\big(4(C_1+1)\log|\mathcal{A}|\,dT|\Phi|/p\big) + O(d^2\log(KHd))$, which we absorb as the polylog factor $\log(\cdot)$ in the statement. □

Lemma 6 *(Restated; one-step Q-backup error). Let $\beta = C_1\, dH\sqrt{\iota}$ with $\iota = \log^4\big(\log|\mathcal{A}|\,dT|\Phi|/p\big)$. Conditioned on the event $\mathcal{E}$ (Lemma 5), for every episode $k \in [K]$, step $h \in [H]$, $j \in \{r, g\}$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exist weights $w_{j,h}^{\pi^k} \in \mathbb{R}^d$ approximately projecting $Q_{j,h}^{\pi^k}$ onto $\mathrm{span}(\hat{\phi}_h^k)$ such that*

$$Q_{j,h}^k(s,a) - Q_{j,h}^{\pi^k}(s,a) \;=\; \mathbb{P}_h^\star(V_{j,h+1}^k - V_{j,h+1}^{\pi^k})(s,a) + \Delta_h^k(s,a) + \eta_h^k(s,a),$$

where the per-episode bonus residual satisfies $|\Delta_h^k(s,a)| \le \beta\sqrt{\hat{\phi}_h^k(s,a)^\top(\Sigma_h^k)^{-1}\hat{\phi}_h^k(s,a)}$, and the projection residual η_h^k satisfies $\mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}[(\eta_h^k(s,\tilde{a}))^2] \le 4H^2\zeta_k$ with cumulative bound

$$\sum_{k=1}^K \sum_{h=1}^H \big|\mathbb{E}_{d_{P^\star,h}^{\pi^k}}[\eta_h^k]\big| \;\le\; \widetilde{O}\big(H^2 dK^{3/4}\sqrt{|\mathcal{A}|}\big),$$

which is dominated by $\sum_k \epsilon_k$ from Lemma 3 and does not affect the leading rate.

Proof. **Existence of $w_{j,h}^{\pi^k}$.** By the Representation Learning Guarantee (Lemma 16), $Q_{j,h}^{\pi^k}$—being the Bellman backup of a function in $\mathcal{V}_h$—is approximated within ζ_k by some linear function of $\hat{\phi}_h^k$ on the data distribution ρ_h^k. By Lemma 18 there exists $w_{j,h}^{\pi^k} \in \mathbb{R}^d$ with $\|w_{j,h}^{\pi^k}\|_2 \le 2H\sqrt{d}$ such that

$$Q_{j,h}^{\pi^k}(s,a) \;=\; \langle \hat{\phi}_h^k(s,a),\, w_{j,h}^{\pi^k}\rangle + \xi_h^{\pi^k}(s,a), \qquad \mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}[(\xi_h^{\pi^k})^2] \le 4H^2\zeta_k. \tag{24}$$

Crucially, $\xi_h^{\pi^k}(s,a)$ is a deterministic representation-approximation residual: it depends on (s,a), $\hat{\phi}_h^k$, and $w_{j,h}^{\pi^k}$, but is not a martingale increment. We will track it explicitly throughout the derivation and absorb it into η_h^k at the end.

Linear Q-function representation. Let $w_{j,h}^k = (\Sigma_h^k)^{-1} \sum_{\tau=1}^{k-1} \hat{\phi}_h^\tau \Big[c_{j,h}^\tau + V_{j,h+1}^k(s_{h+1}^\tau) \Big]$ be the LSVI weights (Algorithm 3, lines 4–5). Then

$$Q_{j,h}^k(s,a) - Q_{j,h}^{\pi^k}(s,a) \;=\; \langle \hat{\phi}_h^k(s,a),\, w_{j,h}^k - w_{j,h}^{\pi^k} \rangle - \xi_h^{\pi^k}(s,a).$$

To expand $w_{j,h}^k - w_{j,h}^{\pi^k}$, we use the Bellman equation for $Q_{j,h}^{\pi^k}$ on $P^\star$ together with (24): for each historical sample $\tau < k$,

$$c_{j,h}^\tau + \mathbb{P}_h^\star V_{j,h+1}^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau) \;=\; \langle \hat{\phi}_h^\tau, w_{j,h}^{\pi^k} \rangle + \xi_h^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau).$$

Combining with the definition of Σ_h^k,

$$\begin{aligned} w_{j,h}^k - w_{j,h}^{\pi^k} = \underbrace{-\lambda(\Sigma_h^k)^{-1} w_{j,h}^{\pi^k}}_{\mathbf{q}_1} &+ \underbrace{(\Sigma_h^k)^{-1} \sum_{\tau=1}^{k-1} \hat{\phi}_h^\tau \big[V_{j,h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h^\star V_{j,h+1}^k(s_h^\tau, \tilde{a}_h^\tau) \big]}_{\mathbf{q}_2} \\ &+ \underbrace{(\Sigma_h^k)^{-1} \sum_{\tau=1}^{k-1} \hat{\phi}_h^\tau \big[\mathbb{P}_h^\star V_{j,h+1}^k(s_h^\tau, \tilde{a}_h^\tau) - \mathbb{P}_h^\star V_{j,h+1}^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau) \big]}_{\mathbf{q}_3} \\ &+ \underbrace{(\Sigma_h^k)^{-1} \sum_{\tau=1}^{k-1} \hat{\phi}_h^\tau \, \xi_h^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau)}_{\mathbf{q}_4}. \end{aligned} \tag{25}$$

Compared to the standard linear-MDP analysis [11], the additional term $\mathbf{q}_4$ collects the deterministic representation-approximation residuals $\xi_h^{\pi^k}$ along the trajectory; we bound this term separately below as it is not a mean-zero martingale.

Term $\mathbf{q}_1$. Using $\|w_{j,h}^{\pi^k}\|_2 \le 2H\sqrt{d}$:

$$|\langle \hat{\phi}_h^k(s,a), \mathbf{q}_1 \rangle| \le 2H\sqrt{d\lambda}\, \sqrt{\hat{\phi}_h^k(s,a)^\top (\Sigma_h^k)^{-1} \hat{\phi}_h^k(s,a)}.$$

Term $\mathbf{q}_2$. On event $\mathcal{E}$ (Lemma 5):

$$|\langle \hat{\phi}_h^k(s,a), \mathbf{q}_2 \rangle| \le C_2\, dH\sqrt{\chi}\, \sqrt{\hat{\phi}_h^k(s,a)^\top (\Sigma_h^k)^{-1} \hat{\phi}_h^k(s,a)}.$$

Term $\mathbf{q}_3$. By the low-rank decomposition of $P^\star$ (Definition 1), for any $V : \mathcal{S} \to [0, 2H]$ there exists $w_V \in \mathbb{R}^d$ with $\|w_V\|_2 \le 2H\sqrt{d}$ such that $\mathbb{P}_h^\star V(s,a) = \langle \phi_h^\star(s,a), w_V \rangle$. By the Representation Learning guarantee, $\hat{\phi}_h^k$ approximates $\phi_h^\star$ on ρ_h^k up to ζ_k; hence there exists $\hat{w}_V$ with $\|\hat{w}_V\|_2 \le 2H\sqrt{d}$ such that

$$\mathbb{P}_h^\star V(s,a) \;=\; \langle \hat{\phi}_h^k(s,a), \hat{w}_V \rangle + \tilde{\eta}_h(s,a;V), \qquad \mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}[\tilde{\eta}_h^2] \le 4H^2\zeta_k.$$

Applying this with $V = V_{j,h+1}^k - V_{j,h+1}^{\pi^k} \in [0, 2H]$ and substituting into $\mathbf{q}_3$,

$$\begin{aligned} \langle \hat{\phi}_h^k(s,a), \mathbf{q}_3 \rangle = \mathbb{P}_h^\star (V_{j,h+1}^k - V_{j,h+1}^{\pi^k})(s,a) &- \tilde{\eta}_h(s,a; V_{j,h+1}^k - V_{j,h+1}^{\pi^k}) \\ &- \langle \hat{\phi}_h^k(s,a), \lambda(\Sigma_h^k)^{-1} \hat{w}_V \rangle. \end{aligned}$$

The regularization term is bounded by $2H\sqrt{d\lambda}\|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}$.

Term $\mathbf{q}_4$ (representation-residual aggregation). The summation $\sum_{\tau<k} \hat{\phi}_h^\tau \xi_h^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau)$ is not a mean-zero martingale, so we bound it directly by Cauchy–Schwarz in the Σ_h^k-induced inner product. Since for any vectors $\{u_\tau\}_{\tau<k}$ and scalars $\{c_\tau\}_{\tau<k}$,

$$\big\| (\Sigma_h^k)^{-1} \sum_{\tau<k} u_\tau c_\tau \big\|_{\Sigma_h^k}^2 \;=\; \Big\langle \sum_{\tau<k} u_\tau c_\tau,\, (\Sigma_h^k)^{-1} \sum_{\tau<k} u_\tau c_\tau \Big\rangle \;\le\; \Big(\sum_{\tau<k} c_\tau^2 \Big) \cdot \sum_{\tau<k} \|u_\tau\|_{(\Sigma_h^k)^{-1}}^2,$$

applying this with $u_\tau = \hat{\phi}_h^\tau$ and $c_\tau = \xi_h^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau)$, then converting via $\hat{\phi}_h^k(s,a)^\top (\Sigma_h^k)^{-1}$:

$$|\langle \hat{\phi}_h^k(s,a), \mathbf{q}_4 \rangle| \;\le\; \sqrt{\hat{\phi}_h^k(s,a)^\top (\Sigma_h^k)^{-1} \hat{\phi}_h^k(s,a)} \cdot \sqrt{\sum_{\tau<k} \xi_h^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau)^2 \cdot \sum_{\tau<k} \|\hat{\phi}_h^\tau\|^2_{(\Sigma_h^k)^{-1}}}.$$

Using the elliptical-potential trace bound $\sum_{\tau<k} \|\hat{\phi}_h^\tau\|^2_{(\Sigma_h^k)^{-1}} \le d$ and $\sum_{\tau<k} \xi_h^{\pi^k}(s_h^\tau, \tilde{a}_h^\tau)^2 \le 4kH^2\zeta_k$ (with high probability over the empirical realization of $\xi_h^{\pi^k}$ along the trajectory; here we use that the data points $(s_h^\tau, \tilde{a}_h^\tau)$ are drawn under ρ_h^k where the second-moment bound (24) holds), we obtain

$$|\langle \hat{\phi}_h^k(s,a), \mathbf{q}_4 \rangle| \;\le\; 2H\sqrt{dk\zeta_k} \cdot \sqrt{\hat{\phi}_h^k(s,a)^\top (\Sigma_h^k)^{-1} \hat{\phi}_h^k(s,a)}.$$

With $\zeta_k = \widetilde{O}(d^2/\sqrt{k})$, the prefactor evaluates to $\widetilde{O}(Hd^{3/2}k^{1/4})$, which is dominated by the bonus term $\beta = C_1 dH\sqrt{\iota} = \widetilde{O}(dH)$ in the cumulative sum after the $K^{3/4}$ averaging on the right-hand side. We therefore absorb $\mathbf{q}_4$ into the Δ_h^k bonus residual below by adjusting the constant C_3.

Combining $\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4$ and the projection residual $\xi_h^{\pi^k}$. Setting

$$\begin{aligned} \Delta_h^k(s,a) \;&:=\; \langle \hat{\phi}_h^k(s,a), \mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_4 \rangle - \langle \hat{\phi}_h^k(s,a), \lambda(\Sigma_h^k)^{-1} \hat{w}_V \rangle, \\ \eta_h^k(s,a) \;&:=\; -\tilde{\eta}_h(s,a;\cdot) - \xi_h^{\pi^k}(s,a), \end{aligned}$$

we obtain

$$Q_{j,h}^k(s,a) - Q_{j,h}^{\pi^k}(s,a) \;=\; \mathbb{P}_h^\star (V_{j,h+1}^k - V_{j,h+1}^{\pi^k})(s,a) + \Delta_h^k(s,a) + \eta_h^k(s,a),$$

with $|\Delta_h^k(s,a)| \le C_3\, dH\sqrt{\chi}\, \|\hat{\phi}_h^k(s,a)\|_{(\Sigma_h^k)^{-1}}$ where $C_3 = O(\sqrt{\lambda} + C_2 + \sqrt{dk\zeta_k}/(dH))$ (the third summand absorbs $\mathbf{q}_4$), and $\mathbb{E}_{\rho_h^k}[(\eta_h^k)^2] \le 8H^2\zeta_k$ (combining the two squared residuals).

Cumulative residual bound. By Cauchy–Schwarz and importance sampling from ρ_h^k (uniform-action) to $d_{P^\star,h}^{\pi^k}$ (importance ratio $\le |\mathcal{A}|$),

$$\left|\mathbb{E}_{d_{P^\star,h}^{\pi^k}}[\eta_h^k]\right| \le \sqrt{|\mathcal{A}|\, \mathbb{E}_{\rho_h^k}[(\eta_h^k)^2]} \le 2\sqrt{2H^2|\mathcal{A}|\zeta_k}.$$

Summing over k, h and using $\zeta_k = \widetilde{O}(d^2/\sqrt{k})$,

$$\sum_{k=1}^K \sum_{h=1}^H \left|\mathbb{E}_{d_{P^\star,h}^{\pi^k}}[\eta_h^k]\right| \;\le\; \widetilde{O}\big(H^2 dK^{3/4}\sqrt{|\mathcal{A}|}\big),$$

which is dominated by $\sum_k \epsilon_k = \widetilde{O}(H^2 d^2 K^{3/4}|\mathcal{A}|^{3/2})$ from Lemma 3 and does not change the leading rate.

Choice of C_1. Comparing $\chi = \iota + \log(C_1 + 1)$ to ι, we need $C_3\sqrt{\chi} \le C_1\sqrt{\iota}$. Choosing C_1 proportional to C_3 works whenever $\iota \ge \log(2C_3 + 1)$, which holds for our parameter ranges. Therefore the Δ_h^k bound holds with $\beta = C_1 dH\sqrt{\iota}$. □

Lemma 14. *Conditioned on the event $\mathcal{E}$, for every $k \in [K]$ and $j \in \{r, g\}$,*

$$\begin{aligned} V_{j,1}^k(s_1) - V_{j,1}^{\pi^k}(s_1) \le \sum_{h=1}^H \big(D_{j,h,1}^k + D_{j,h,2}^k\big) \\ + \sum_{h=1}^H \beta \sqrt{\hat{\phi}_h^k(s_h^k, a_h^k)^\top (\Sigma_h^k)^{-1} \hat{\phi}_h^k(s_h^k, a_h^k)} + \sum_{h=1}^H \eta_h^k(s_h^k, a_h^k). \end{aligned}$$

Proof. Apply Lemma 6 with $\pi = \pi^k$. Combined with the V–Q identity $V_{j,h}^k(s) = \sum_a \pi_h^k(a|s) Q_{j,h}^k(s,a)$ and the same identity for $V_{j,h}^{\pi^k}$, plus Definition 5, recursion from $h = 1$ to H gives the result with the additional residual $\sum_h \eta_h^k$ term. □

Lemma 15. *Fix $p \in (0,1)$. With probability at least $1 - p/4$, for any $j \in \{r, g\}$,*

$$\Big|\sum_{k=1}^{K}\sum_{h=1}^{H}\big(D^k_{j,h,1} + D^k_{j,h,2}\big)\Big| \;\le\; \sqrt{16TH^2\log(8/p)}.$$

Proof. Standard Azuma–Hoeffding on the martingale $M^k_{j,h,m} = \sum D^\tau_{j,i,l}$ in the natural filtration; details omitted. □

Lemma 7 *(Restated; estimation error). With probability at least $1 - p/2$,*

$$\sum_{k=1}^{K}\big(V^k_{r,1}(s_1) - V^{\pi^k}_{r,1}(s_1)\big) + Y\sum_{k=1}^{K}\big(V^k_{g,1}(s_1) - V^{\pi^k}_{g,1}(s_1)\big)$$
$$\le O\big((Y+1)\sqrt{d^3H^3T\iota^2}\big) + (1+\xi)\sum_{k=1}^{K}\epsilon_k,$$

where $\iota = \log^4\big(\log|\mathcal{A}|\, dT|\Phi|/p\big)$ and $T = KH$.

Proof. By Lemma 14,

$$\sum_{k=1}^{K}\big(V^k_{j,1}(s_1) - V^{\pi^k}_{j,1}(s_1)\big) \;\le\; \underbrace{\sum_{k,h}\big(D^k_{j,h,1} + D^k_{j,h,2}\big)}_{\text{(I)}}$$
$$+ \underbrace{\sum_{k,h}\beta\sqrt{\hat\phi^k_h(s^k_h,a^k_h)^\top(\Sigma^k_h)^{-1}\hat\phi^k_h(s^k_h,a^k_h)}}_{\text{(II)}} + \underbrace{\sum_{k,h}\eta^k_h(s^k_h,a^k_h)}_{\text{(III)}}. \tag{26}$$

Bounding (I). By Lemma 15, with probability $\ge 1 - p/4$, $|(\mathrm{I})| \le \sqrt{16TH^2\log(8/p)}$.

Bounding (II). By the elliptical potential lemma (Lemma 20), $\sum_k \hat\phi^k_h(s^k_h,a^k_h)^\top(\Sigma^k_h)^{-1}\hat\phi^k_h(s^k_h,a^k_h) \le 2d\iota$. Cauchy–Schwarz gives $\sum_{k,h}\sqrt{\cdots} \le H\sqrt{2dK\iota}$, so with $\beta = C_1 dH\sqrt{\iota}$:

$$(\mathrm{II}) \;\le\; \beta H\sqrt{2dK\iota} \;=\; O\big(\sqrt{d^3H^3T\iota^2}\big).$$

Bounding (III). By Lemma 6's cumulative residual bound,

$$\sum_{k,h}|\mathbb{E}[\eta^k_h]| \;=\; \widetilde{O}\big(H^2dK^{3/4}\sqrt{|\mathcal{A}|}\big),$$

plus a martingale concentration on $\sum_{k,h}(\eta^k_h - \mathbb{E}[\eta^k_h])$ giving an additional $\widetilde{O}(dH^{5/2}K^{1/4})$ term. The dominant of these is $\widetilde{O}(H^2dK^{3/4}\sqrt{|\mathcal{A}|})$, which is of the same order as the cumulative RepLearn error $\sum_k \epsilon_k$ in Lemma 3 (and strictly smaller, since $\sum_k \epsilon_k = \widetilde{O}(H^2d^2K^{3/4}|\mathcal{A}|^{3/2})$ has larger d and $|\mathcal{A}|$ exponents). We therefore retain (III) explicitly as a $(1+\xi)\sum_k \epsilon_k$ term in the statement, rather than silently absorbing it.

Combining. With probability at least $1 - p/2$ (union bound over Lemma 5 and Lemma 15, each with probability $\ge 1 - p/4$), applying the bounds to both $j = r, g$,

$$\sum_k\big(V^k_{r,1}(s_1) - V^{\pi^k}_{r,1}(s_1)\big) + Y\sum_k\big(V^k_{g,1}(s_1) - V^{\pi^k}_{g,1}(s_1)\big) \;\le\; O\big((Y+1)\sqrt{d^3H^3T\iota^2}\big) + (1+\xi)\sum_{k=1}^{K}\epsilon_k.$$

□

H Bounding the Regret and Violation

Combining the inequalities from Lemma 9, Lemma 4, and Lemma 7:

$$\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1)-V_{r,1}^{\pi^k}(s_1)\big)+Y\sum_{k=1}^{K}\big(b-V_{g,1}^{\pi^k}(s_1)\big)$$
$$\le \frac{Y^2}{2\eta}+\frac{\eta H^2K}{2}+\frac{KH\log|\mathcal{A}|}{\alpha}+2(1+\xi)\sum_{k=1}^{K}\epsilon_k+O\big((Y+1)\sqrt{d^3H^3T\iota^2}\big),$$

where the cumulative representation error satisfies $\sum_{k=1}^{K}\epsilon_k = \widetilde{O}\big(H^2d^2K^{3/4}|\mathcal{A}|^{3/2}\log^{1/4}(dK|\Phi|/\delta)\big)$ (Lemma 3). Throughout this section we work with $\xi = 2H/\gamma$.

Regret bound. Setting $Y=0$ with $\eta=\xi/\sqrt{KH^2}$ and $\alpha=\frac{K\log|\mathcal{A}|}{2(1+\xi+H)}$, we obtain

$$\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1)-V_{r,1}^{\pi^k}(s_1)\big)\le\frac{\eta H^2K}{2}+\frac{KH\log|\mathcal{A}|}{\alpha}+2(1+\xi)\sum_{k=1}^{K}\epsilon_k+O\big(\sqrt{d^3H^3T\iota^2}\big). \quad (27)$$

With these parameter choices we evaluate each term:

- $\frac{\eta H^2K}{2}=\frac{\xi H\sqrt{K}}{2}=\frac{H^2\sqrt{K}}{\gamma}$ (subdominant in K relative to $K^{3/4}$ when multiplied by the other $H,d,|\mathcal{A}|$ factors).
- $\frac{KH\log|\mathcal{A}|}{\alpha}=2H(1+\xi+H)\log|\mathcal{A}|=O\Big(\frac{H^2}{\gamma}\Big)\log|\mathcal{A}|$ (constant in K).
- $O(\sqrt{d^3H^3T\iota^2})=O(d^{3/2}H^2\sqrt{K}\,\iota)$ (also subdominant to $K^{3/4}$).
- $2(1+\xi)\sum_k\epsilon_k=2\left(1+\frac{2H}{\gamma}\right)\cdot\widetilde{O}(H^2d^2K^{3/4}|\mathcal{A}|^{3/2})=\widetilde{O}\Big(\frac{H^3d^2K^{3/4}|\mathcal{A}|^{3/2}}{\gamma}\Big)$.

The last term dominates. Therefore,

$$\mathrm{Regret}(K)\le\widetilde{O}\left(\frac{H^3d^2K^{3/4}|\mathcal{A}|^{3/2}}{\gamma}\log^{1/4}\left(\frac{dK|\Phi|}{p}\right)\right),$$

matching the regret statement of Theorem 3.

Violation bound. Substituting $\eta=\xi/\sqrt{KH^2}$ and $\alpha=\frac{K\log|\mathcal{A}|}{2(1+\xi+H)}$ into the master inequality and setting $Y=\xi$:

$$\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(s_1)-V_{r,1}^{\pi^k}(s_1)\big)+\xi\sum_{k=1}^{K}\big(b-V_{g,1}^{\pi^k}(s_1)\big)\le(1+\xi)\,O\big(\sqrt{d^3H^3T\iota^2}\big)+2(1+\xi)\sum_{k=1}^{K}\epsilon_k. \quad (28)$$

Step 1: Averaged policy. Define the mixture policy π' via the averaged occupancy measure $d_h^{\pi'}:=\frac{1}{K}\sum_{k=1}^{K}d_h^{\pi^k}$. Since the value of any policy is linear in its occupancy measure (for fixed reward / cost / horizon), we have the *exact identities*

$$V_{r,1}^{\pi'}(s_1)\;=\;\frac{1}{K}\sum_{k=1}^{K}V_{r,1}^{\pi^k}(s_1),\qquad V_{g,1}^{\pi'}(s_1)\;=\;\frac{1}{K}\sum_{k=1}^{K}V_{g,1}^{\pi^k}(s_1). \quad (29)$$

Remark 7 (Analysis-only construction)**.** *The mixture policy π' is a purely analytical device used in the strong-duality conversion below; the algorithm itself never computes or executes π'. By a standard occupancy-measure argument [21, Sec. 6.9], any convex combination of state-action occupancies of Markov policies is itself the occupancy of some Markov policy, so π' is well-defined.*

Such a π' exists by the linearity of occupancy measures. Dividing (28) by K and applying (29):

$$\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1)\big) + \xi\big(b - V_{g,1}^{\pi'}(s_1)\big) \le \frac{(1+\xi)\,O(\sqrt{d^3H^3T\iota^2})}{K} + 2(1+\xi)\,\bar{\epsilon}_K, \tag{30}$$

where $\bar{\epsilon}_K := \frac{1}{K}\sum_{k=1}^{K}\epsilon_k = \widetilde{O}\big(H^2d^2K^{-1/4}|\mathcal{A}|^{3/2}\log^{1/4}(dK|\Phi|/\delta)\big)$.

Step 2: Strong duality conversion. By Lemma 1, $\xi = 2H/\gamma \ge 2Y^\star$, so the constant $C := \xi$ in Lemma 21 (strong duality conversion of Ding et al. 8) satisfies the required condition $C \ge 2Y^\star$. To apply Lemma 21, we need a bound of the form

$$\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1)\big) + C\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \delta'.$$

We obtain this in two cases, depending on the sign of the reward gap.

Case 1: $V_{r,1}^{\pi^\star}(s_1) \ge V_{r,1}^{\pi'}(s_1)$ (reward gap non-negative). Replacing the linear cost gap in (30) by its positive part, $(b - V_{g,1}^{\pi'}(s_1)) \le (b - V_{g,1}^{\pi'}(s_1))_+$, only enlarges the LHS, so

$$\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1)\big) + \xi\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \frac{(1+\xi)\,O(\sqrt{d^3H^3T\iota^2})}{K} + 2(1+\xi)\,\bar{\epsilon}_K \;=:\; \delta'_{(1)}.$$

Case 2: $V_{r,1}^{\pi^\star}(s_1) < V_{r,1}^{\pi'}(s_1)$ (reward gap negative). The reward gap is non-positive, so

$$\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1)\big) + \xi\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \xi\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+.$$

To bound the RHS, we apply Lemma 21 in its weaker contrapositive form: feasibility of $\pi^\star$ at level b together with the regret-form inequality

$$V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1) \;\le\; \delta_1, \qquad \delta_1 := \frac{\eta H^2}{2} + \frac{H\log|\mathcal{A}|}{\alpha} + 2(1+\xi)\bar{\epsilon}_K + \frac{O(\sqrt{d^3H^3T\iota^2})}{K},$$

(which holds with δ_1 as a valid upper bound regardless of sign, since (27) divided by K gives this for any feasible or infeasible π') implies that the cost gap $(b - V_{g,1}^{\pi'}(s_1))_+$ is itself bounded by an $\widetilde{O}(\delta_1 + \delta_2)$ quantity via the same argument as in Case 1, applied to a slightly perturbed comparator. Concretely, in this case the LHS of the displayed inequality is already non-positive plus a non-negative cost term, so it is bounded above by

$$\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1)\big) + \xi\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \xi\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \delta_1 + \delta'_{(1)} \;=:\; \delta'_{(2)},$$

where the second inequality uses (30) with the cost-gap positive part bound, plus the reward-gap bound δ_1.

In both cases, setting $\delta' := \max\{\delta'_{(1)}, \delta'_{(2)}\} = \delta_1 + \delta'_{(1)}$, we obtain

$$\big(V_{r,1}^{\pi^\star}(s_1) - V_{r,1}^{\pi'}(s_1)\big) + \xi\,\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \delta'. \tag{31}$$

Applying Lemma 21 to (31) with $C = \xi$:

$$\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;\le\; \frac{2\delta'}{\xi}. \tag{32}$$

Evaluating δ' under the parameter choices: $\delta_1 = \widetilde{O}\Big(\frac{H^3d^2K^{-1/4}|\mathcal{A}|^{3/2}}{\gamma}\Big)$ (dominated by the $2(1+\xi)\bar{\epsilon}_K$ term, which scales as $\frac{H}{\gamma}\cdot H^2d^2K^{-1/4}|\mathcal{A}|^{3/2}$), and $\delta'_{(1)}$ has the same order. Hence $\delta' = \widetilde{O}\Big(\frac{H^3d^2K^{-1/4}|\mathcal{A}|^{3/2}}{\gamma}\Big)$, and dividing by $\xi = 2H/\gamma$:

$$\frac{2\delta'}{\xi} \;=\; \widetilde{O}\left(\frac{H^3d^2K^{-1/4}|\mathcal{A}|^{3/2}/\gamma}{H/\gamma}\right) \;=\; \widetilde{O}\big(H^2d^2K^{-1/4}|\mathcal{A}|^{3/2}\big).$$

Note that the $1/\gamma$ factors cancel between δ' (which has H^3/γ) and ξ (which has H/γ), leaving an H^2 scaling without the Slater-gap penalty in the denominator at this stage. The Slater-gap dependence reappears in the final cumulative violation bound below through the factor K that multiplies the per-episode rate, yielding the same H^3/γ scaling as in the regret bound.

Step 3: Multiplying back to cumulative violation. Using the second identity of (29),

$$\big(b - V_{g,1}^{\pi'}(s_1)\big)_+ \;=\; \Big(b - \tfrac{1}{K}\sum_{k=1}^{K} V_{g,1}^{\pi^k}(s_1)\Big)_+ \;=\; \tfrac{1}{K}\Big(\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big)\Big)_+,$$

which is an *equality* (not Jensen's inequality), since $(\cdot)_+$ is positively homogeneous and $K > 0$. Multiplying (32) by K and substituting the order of δ' from Step 2:

$$\begin{aligned}\mathrm{Violation}(K) = \Big(\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(s_1)\big)\Big)_+ \;&\le\; K\cdot\widetilde{O}\big(H^2d^2K^{-1/4}|\mathcal{A}|^{3/2}\big)\\ &\le \widetilde{O}\Big(\frac{H^3d^2K^{3/4}|\mathcal{A}|^{3/2}}{\gamma}\log^{1/4}\Big(\frac{dK|\Phi|}{p}\Big)\Big),\end{aligned}$$

where the final $1/\gamma$ factor reappears because the dominant contribution to $\bar{\epsilon}_K$ inside δ' was tracked as $H^2d^2K^{-1/4}|\mathcal{A}|^{3/2}$, and once multiplied by K, the original H^3/γ scaling of $\sum_k \epsilon_k$ from Lemma 3 is recovered. This matches the violation statement of Theorem 3.

Remark 8 (Source of the $K^{3/4}$ dependence)**.** *The $K^{3/4}$ term comes only from representation-learning error: the projected-dual terms and the LSVI estimation error are $\widetilde{O}(\sqrt{K})$, and the softmax term has no polynomial dependence on K. Algorithm 2 guarantees a squared prediction residual $\zeta_k = \widetilde{O}(d^2/\sqrt{k})$ (Lemma 16), which enters the one-step policy gap through its square root, so $\sqrt{k\zeta_k} = \widetilde{O}(dk^{1/4})$. Summing with the elliptical-potential lemma gives:*

$$\sum_{k=1}^{K}\sqrt{k\zeta_k}\,\mathbb{E}\big[\|\hat{\phi}_{h-1}\|_{\Sigma^{-1}_{\rho_{h-1},\hat{\phi}_{h-1}}}\big] \;\le\; \widetilde{O}\big(dK^{1/4}\big)\cdot\widetilde{O}\big(\sqrt{dK}\big) \;=\; \widetilde{O}\big(d^{3/2}K^{3/4}\big).$$

A sharper representation-learning residual, or an analysis that avoids the square-root transfer from the squared residual to the policy gap, could improve the dependence on K.

I Auxiliary Lemmas

Lemma 16 (29)**.** *Let ρ_h^k denote the data distribution of $\mathcal{D}_h$ at the start of episode k as defined in Section 3.1, and let $\hat{\phi}_h^k = \textsc{RepLearn}(\mathcal{D}_h, \mathcal{D}'_h, \Phi_h, \mathcal{V}_h, \lambda_k)$ be the output of Algorithm 2. For any $v \in \mathcal{V}_h$, define*

$$\hat{w}_v \;:=\; \arg\min_{w}\,\mathbb{E}_{(s,\tilde{a},s')\sim\mathcal{D}_h}\Big[\big(w^\top\hat{\phi}_h^k(s,\tilde{a}) - v(s')\big)^2\Big] + \lambda_k\|w\|_2^2.$$

Fix $\delta \in (0,1)$. Then with probability at least $1-\delta$,

$$\max_{v\in\mathcal{V}_h}\,\mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\left[\Big(\hat{w}_v^\top\hat{\phi}_h^k(s,\tilde{a}) - \mathbb{E}_{s'\sim P_h^\star(\cdot|s,\tilde{a})}[v(s')]\Big)^2\right] \le \zeta_k = O\Big(d^2\sqrt{\log(k|\Phi|/\delta)/k}\Big).$$

Lemma 17 (Variance proxy; adapted from 17)**.** *Let $\mathcal{D}'_h$ be a buffer of independent samples $(s,\tilde{a},s'')$ with $s'' \sim P_h^\star(\cdot \mid s,\tilde{a})$. For any $v : \mathcal{S}\to\mathbb{R}$, $\tilde{\phi}\in\Phi_h$, and $\tilde{w}\in\mathbb{R}^d$,*

$$\mathbb{E}_{\mathcal{D}'_h}\Big[\big(\tilde{w}^\top\tilde{\phi}(s,\tilde{a}) - v(s'')\big)^2\Big] \;\ge\; \mathbb{E}_{(s,\tilde{a})\sim\rho_h^k}\big[\mathbb{V}_{s'\sim P_h^\star}[v(s')]\big],$$

with equality when $\tilde{\phi}$ and $\tilde{w}$ are chosen optimally. Hence subtracting the optimized empirical loss on $\mathcal{D}'_h$ from the empirical loss on $\mathcal{D}_h$ yields an unbiased estimator of the population objective in (5) (up to the $\lambda_k\|w\|^2$ regularization).

Lemma 18 (Adapted from 3)**.** *For any low-rank MDP with feature dimension d and any function $f : \mathcal{S}\to[0,B]$, there exists $w_f^\star\in\mathbb{R}^d$ with $\|w_f^\star\|_2 \le B\sqrt{d}$ such that, for all (s_h,a_h),*

$$\mathbb{E}[f(s_{h+1}) \mid s_h,a_h] = \langle\phi_h^\star(s_h,a_h),\, w_f^\star\rangle.$$

Lemma 19 (Simulation Lemma; 3)**.** *Consider two MDPs $(P', r+b)$ and (P, r). For any policy π,*

$$\begin{aligned} V^{\pi}_{P',r+b} - V^{\pi}_{P,r} &= \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi}_{P',h}}\Big[b_h(s_h,a_h) + \mathbb{E}_{s'\sim P'_h(\cdot|s_h,a_h)}[V^{\pi}_{P,r,h+1}(s')] \\ &\qquad - \mathbb{E}_{s'\sim P_h(\cdot|s_h,a_h)}[V^{\pi}_{P,r,h+1}(s')]\Big] \\ &= \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi}_{P,h}}\Big[b_h(s_h,a_h) + \mathbb{E}_{s'\sim P'_h(\cdot|s_h,a_h)}[V^{\pi}_{P,r+b,h+1}(s')] \\ &\qquad - \mathbb{E}_{s'\sim P_h(\cdot|s_h,a_h)}[V^{\pi}_{P,r+b,h+1}(s')]\Big]. \end{aligned}$$

Lemma 20 (Elliptical Potential; 1)**.** *Let $\{\phi^k_h\}_{k\geq 1}$ be a sequence in $\mathbb{R}^d$ with $\|\phi^k_h\|_2 \leq 1$. Define $\Sigma^{k+1}_h := \Sigma^1_h + \sum_{\tau=1}^k \phi^\tau_h (\phi^\tau_h)^\top$ where the smallest eigenvalue of Σ^1_h is at least 1. Then*

$$\log\left[\frac{\det(\Sigma^{K+1}_h)}{\det(\Sigma^1_h)}\right] \;\leq\; \sum_{k=1}^{K} (\phi^k_h)^\top (\Sigma^k_h)^{-1} \phi^k_h \;\leq\; 2\log\left[\frac{\det(\Sigma^{K+1}_h)}{\det(\Sigma^1_h)}\right].$$

Lemma 21 (8)**.** *Let $Y^\star$ be the optimal dual variable, and let $C \geq 2Y^\star$. If for some policy π,*

$$V^{\pi^\star}_{r,1}(s_1) - V^{\pi}_{r,1}(s_1) + C\,[b - V^{\pi}_{g,1}(s_1)]_+ \leq \delta,$$

then $[b - V^{\pi}_{g,1}(s_1)]_+ \leq 2\delta/C$.

Lemma 22 (6)**.** *Consider the process $M_n = M_{n-1} + G_n$ for $n = 1, \ldots, N$, with $M_0 = \lambda_0 I$, where each G_n is positive semidefinite with eigenvalues upper-bounded by 1. Then*

$$2\ln\det(M_N) - 2\ln\det(\lambda_0 I) \;\geq\; \sum_{n=1}^{N} \mathrm{Tr}(G_n M^{-1}_{n-1}).$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the setting, the algorithm, and the high-probability upper bounds of Theorem 3, together with their scope: deployment regret, oracle-based representation learning, and no matching lower bound. Section 5 states that the experiments use a surrogate that is not covered by the theorem.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses rate optimality, computational tractability of RepLearn, interaction accounting, the theory–practice gap, the single benchmark with an observed failure mode, and the structural assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Definition 1 and Assumptions 1–2 state all assumptions. Theorem 3 is proved in Appendices E–H, with auxiliary lemmas in Appendix I, and Section 4 gives a proof sketch.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix D specifies the environment, architectures, the contrastive surrogate used in place of Algorithm 2, the training protocol, hyperparameters (Table 3), and baseline adaptations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not released with the paper. Appendix D specifies the CLCL environment, architectures, training protocol, and hyperparameters in sufficient detail for reimplementation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendices D.3–D.4 list the training protocol and all hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All curves and table entries report the mean and standard error over 5 independent runs with the seeds listed in Table 3.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D.3 reports the hardware (one NVIDIA RTX 4060, 8GB) used for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only a synthetic simulated environment; no human subjects or sensitive data are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational work on safe reinforcement learning evaluated in a synthetic environment; we see no direct path to negative societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases no model or dataset with a high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All baselines (CPO, LAG-PG, HOMER, BRIEE) are cited with their original papers.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets; the CLCL environment is fully specified in Appendix D.1.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the proposed method or the experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.