Primal-Dual Representation Learning for Low-Rank Constrained MDPs

Chenhao Zhou **Chao Zhang*** **Wensong Bai** **Hanbin Zhao** **Hui Qian**
College of Computer Science and Technology, Zhejiang University
zhouchenhao@zju.edu.cn zczju@zju.edu.cn

Abstract

Representation learning has substantially improved the efficiency of unconstrained continuous-state Markov Decision Processes (MDPs), but its extension to Constrained MDPs (CMDPs) is difficult because the transition representation must be learned while constraint violation is controlled. In this paper, we study representation learning for low-rank CMDPs with continuous state spaces, and propose REP-PD-TEG for the soft cumulative-violation setting. In the feature learning period, the algorithm interacts with the environment using a decaying-ϵ-greedy mixture execution policy, and uses the collected samples to learn transition features by maximum likelihood. Based on the learned features, it constructs uncertainty bonuses and plans with an optimistic Lagrangian objective to obtain the return policy. We prove regret and soft-violation guarantees for the return policies, and quantify the influence of the decaying-ϵ-greedy exploration. In particular, REP-PD-TEG achieves the optimal $\widetilde{O}(\sqrt{K})$ return policy bound, and gives a $\widetilde{O}(K^{3/4})$ overall sample bound when exploratory actions are charged. For the stricter hard cumulative violation, where feasibility must be verified episode by episode, we propose REP-PD-H-TEG. It plans under a bonus-corrected lower-confidence utility certificate with the learned representation, and performs a stabilized dual search to obtain the return policy, yielding regret and hard-violation bounds. Together, these results give the first theoretically guaranteed representation-learning method for low-rank continuous-state CMDPs with both soft cumulative guarantees and certified hard-violation control.

1 Introduction

Constrained Markov Decision Processes (CMDPs) provide a principled framework for constrained reinforcement learning with both reward objectives and safety requirements [4, 22]. This framework is especially relevant in safety-critical domains, and a growing line of provably constrained RL algorithms has been developed within this framework [8, 9, 16, 24, 27, 39]. These works primarily focus on the tabular setting, where guarantees scale with the cardinality of the state space and become unsuitable for continuous or high-dimensional environments. Linear CMDP methods [5, 13, 17, 18, 26, 42] partially address this issue by assuming that transitions are linearly represented by known low-dimensional features, yielding state-dimension-independent bounds. However, existing tabular/linear CMDP methods all rely on a priori knowledge of the latent feature representations. In the absence of such prior feature knowledge, their guarantees no longer apply, leaving CMDPs with unknown features insufficiently addressed.

Representation learning has been effective for discovering low-dimensional transition features in unconstrained MDPs [2, 29, 31, 36, 37, 41, 44]. Extending this idea to CMDPs is more challenging because representation error now affects both reward optimization and constraint evaluation. The

*Corresponding author.

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

learner must identify latent dynamics from online data, use the learned representation to guide exploration, and still control cumulative constraint violation [2, 16]. This creates a coupling between representation coverage and constraint-aware policy optimization that is absent from unconstrained low-rank RL and is not handled by existing CMDP results with known features.

In this paper, we develop primal-dual representation-learning algorithms for low-rank continuous-state CMDPs with unknown transition features. During feature learning, the algorithms interact with the environment via an ϵ-greedy mixture *execution policy* and use the collected samples to learn transition features by maximum-likelihood estimation (MLE). For soft violation criterion, where only the total constraint shortfall across episodes is charged, our algorithm, REP-PD-TEG, uses the learned features to construct uncertainty bonuses, and plan with the bonus-augmented Lagrangian objective that couples reward maximization, utility optimization, and exploration. The algorithm outputs the *return policies* with a primal-dual update, and admits both return-policy and execution-policy guarantees. For the stricter per-episode hard-violation metric, where each episode's shortfall is charged separately, we develop REP-PD-H-TEG, which keeps the representation-learning procedure but plans under a tight lower-confidence utility certificate and conducts a stabilized dual search to obtain the return policy. Specifically, we make the following contributions:

- We propose the first theoretically guaranteed representation-learning method for low-rank continuous-state CMDPs with unknown features. Starting from the occupancy of the current return policy, it uses ϵ-greedy auxiliary probes for action coverage, fits transition features by MLE, and constructs a merged-covariance uncertainty bonus. We quantify the effect of decaying-ϵ exploration for both return and execution policies.
- For the aggregate soft-violation criterion, we propose REP-PD-TEG, which updates the return policy through an optimistic Lagrangian whose reward and dual-weighted utility terms are both bonus-augmented based on the learned features. We show that REP-PD-TEG achieves return policy regret and soft violation of order $\tilde{O}(K^{1/2+p})$ sample complexity under the schedule $\epsilon_k = k^{-p}$, where ϵ_k is the exploration level in episode k. In particular, this recovers the optimal $\tilde{O}(\sqrt{K})$ return policy rate when $p = 0$, which corresponds to uniform auxiliary probing over actions during the whole procedure. We also prove an execution policy overall bound: executing the ϵ_k-greedy exploration adds $O(K^{1-p})$ complexity, and the balanced choice $p = 1/4$ yields a $\tilde{O}(K^{3/4})$ execution policy regret and soft violation up to problem-dependent factors.
- For the stricter per-episode hard-violation criterion, we propose REP-PD-H-TEG, which uses the learned features to conduct a stabilized dual search and certifies feasibility through a bonus-corrected lower-confidence utility certificate. For return policies, we obtain regret of order $\tilde{O}(K^{3/4+p})$ and hard violation of order $\tilde{O}(K^{1/2+p})$; taking $p = 0$ again gives uniform probing and the clean $\tilde{O}(\sqrt{K})$ hard-violation dependence. For execution policies, we show that the cost of ϵ_k-greedy exploration is also $O(K^{1-p})$ and choosing $p = 1/8$ gives a $\tilde{O}(K^{7/8})$ regret and hard violation up to problem-dependent factors.

Appendix H provides the empirical study of REP-PD-TEG on a simulated job-scheduling task. Besides, we also implement an approximate neural network version of REP-PD-TEG on the Safety Gym task to validate the effectiveness of our representation learning procedure for a complex state-action space.

Remark on the probing schedule. In tabular and linear CMDPs, the state or feature representation is known, so the analyzed policy is the execution policy. In low-rank CMDPs, the transition features must be learned online, and auxiliary ϵ_k-greedy probing is needed to obtain coverage over the state-action space for representation learning. This creates a distinction between the return-policy sequence and the execution policy. The latter is especially relevant in constrained RL because exploratory behavior can incur real training-time costs—for example, robot collisions or excessive contact forces—even when the final return policy is safe. The two REP-PD-TEG rows in Table 1 report, respectively, the sample complexity of the best return policy specialization ($p = 0$) and the best execution policy specialization ($p = 1/4$), where p denotes the probing strength with $\epsilon_k = k^{-p}$. Similarly, the two REP-PD-H-TEG rows in Table 2 show the best return policy complexity with $p = 0$ and the balanced execution policy complexity under the hard specialization $p = 1/8$, respectively.

Table 1: Comparison of soft-violation bounds. The two REP-PD-TEG rows report the complexity of the best return policy specialization and the best execution policy specialization, respectively.

MDP	Method	Known Feature	Regret	Violation
Tabular	Efroni et al. [16]	Yes	$\tilde{O}\left(\sqrt{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert H^4K}\right)$	$\tilde{O}\left(\sqrt{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert H^4K}\right)$
	Liu et al. [27]	Yes	$\tilde{O}\left(\sqrt{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert H^6K}\right)$	0
Linear	Ding et al. [13]	Yes	$\tilde{O}\left(\sqrt{d^2H^6K}\right)$	$\tilde{O}\left(\sqrt{d^2H^6K}\right)$
	Ghosh et al. [17]	Yes	$\tilde{O}\left(\sqrt{d^3H^3K}\right)$	$\tilde{O}\left(\sqrt{d^3H^3K}\right)$
	Ghosh et al. [17]	Yes	$\tilde{O}\left(\sqrt{d^3H^5K}\right)$	0
Low-rank	**REP-PD-TEG**	**No**	$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2\sqrt{K}(1-\gamma)^{-2}\right)$ [return policy]	$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2\sqrt{K}(1-\gamma)^{-2}\right)$ [return policy]
			$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2K^{3/4}(1-\gamma)^{-2}\right)$ [execution policy]	$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2K^{3/4}(1-\gamma)^{-2}\right)$ [execution policy]

Table 2: Comparison of hard-violation bounds. The two REP-PD-H-TEG rows report the complexity of the best return policy specialization and the balanced execution policy specialization, respectively.

MDP	Method	Known Feature	Known Safe Action Set	Regret	Violation
Tabular	Efroni et al. [16]	Yes	Yes	$\tilde{O}\left(\sqrt{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert H^4K}\right)$	$\tilde{O}\left(\sqrt{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert H^4K}\right)$
	Bura et al. [9]	Yes	Yes	$\tilde{O}\left(\sqrt{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert H^6K}\right)$	0
Linear	Amani et al. [5]	Yes	Yes	$\tilde{O}\left(\sqrt{d^3H^4K}\right)$	0
	Kitamura et al. [26]	Yes	Yes	$\tilde{O}\left(\sqrt{d^5H^8K}\right)$	0
	Wei et al. [42]	Yes	No	$\tilde{O}\left(\sqrt{d^3H^4K}\right)$	$\tilde{O}\left(\sqrt{dH^2K}\right)$
	Ghosh et al. [18]	Yes	No	$\tilde{O}\left(\sqrt{d^3H^4K}\right)$	$\tilde{O}\left(\sqrt{d^3H^4K}\right)$
Low-rank	**REP-PD-H-TEG**	**No**	**No**	$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2K^{3/4}(1-\gamma)^{-2}\right)$ [return policy]	$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2\sqrt{K}(1-\gamma)^{-2}\right)$ [return policy]
				$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2K^{7/8}(1-\gamma)^{-2}\right)$ [execution policy]	$\tilde{O}\left(\lvert\mathcal{A}\rvert d^2K^{7/8}(1-\gamma)^{-2}\right)$ [execution policy]

1.1 Related Work

Provably Constrained RL CMDPs provide a rigorous framework for safe RL, balancing reward maximization with cumulative cost constraints [4, 19]. Several practical constrained RL methods have been developed on this foundation, including policy-gradient, Lagrangian, shielding, and model-based approaches [1, 6, 7, 11, 21, 35]. On the theoretical side, there has been a growing body of work on provably safe reinforcement learning. Many works [8, 9, 14, 16, 24, 25, 27, 39] consider the tabular setting. Among the best known regret and constraint-violation guarantees are of order $\tilde{O}(\sqrt{K})$ [16, 27]. However, their polynomial dependence on the cardinality of state space $|\mathcal{S}|$ makes them unsuitable for large or continuous state spaces.

Recent advances in linear CMDPs [5, 13, 17, 18, 26, 30, 42, 43] achieve dimension-independent sample complexity by assuming known low-dimensional features $\phi^\star(s,a)$. They enforce safety through confidence intervals, optimistic bonuses, primal-dual updates, pessimistic constraints, or safe baselines. However, their reliance on predefined latent features, assumed to be known a priori or handcrafted, limits applicability when transition structures are unknown and must be learned from data. Our setting instead has unknown latent features that must be learned online. To our knowledge, the only prior work on low-rank CMDPs is [20], which studies the sample complexity of identifying a near-optimal feasible policy after training. In contrast, our work analyzes the entire learning trajectory and provides regret and constraint-violation guarantees.

Representation Learning for Unconstrained MDPs A growing body of research has focused on theoretically grounded representation learning in RL, aiming to derive low-dimensional features that efficiently capture environment dynamics. For example, works such as [15, 28, 46] investigated representation learning in block MDPs, leveraging hierarchical or modular state decompositions to improve sample efficiency. [38] studied representations under a Gaussian noise model, establishing guarantees for feature learning in noisy environments. Low-Rank MDPs have emerged as a pivotal framework for representation learning in structured environments. Within this regime, model-based approaches [2, 36, 41, 44, 45] learn latent transition features by leveraging model classes of transition probabilities, achieving sample-efficient exploration under realizability assumptions. In contrast, model-free methods like [29] eschew explicit transition models and show efficiency under minimal reachability conditions. The framework has recently expanded into reward-free exploration [10, 20] and multi-agent Markov games [31, 34, 47, 48]. However, these works do not address online safety constraints in CMDPs. Consequently, this work gives the first representation-learning framework for low-rank CMDPs that bridges feature discovery and provable online safety.

2 Background

We consider an episodic constrained MDP, denoted by $(\mathcal{S}, \mathcal{A}, P^*, r, g, \gamma, d_1)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P^* is the transition function, r is the reward function, g is the utility function, γ is the discounted factor, and d_1 is the initial distribution. Following prior work [23, 40, 41], we assume trajectory reward is normalized, i.e., for any trajectory $\{s_h, a_h\}_{h=1}^{\infty}$, we have $\sum_{h=1}^{\infty} \gamma^h r(s_h, a_h) \in [0, 1]$ and $\sum_{h=1}^{\infty} \gamma^h g(s_h, a_h) \in [0, 1]$.

In an infinite-horizon discounted CMDP, each episode starts with state $s_1 \sim d_1$. Then at each step h, the agent observes state $s_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, and receives a reward $r(s_h, a_h)$ and a utility $g(s_h, a_h)$. The MDP then transitions to s_{h+1} which is drawn from $P^*(\cdot|s_h, a_h)$. In this paper, we assume access to reward and utility function feedback, but does not require access to the transition model. For any $s \in \mathcal{S}$, $\pi(a|s)$ denotes the probability of selecting action $a \in \mathcal{A}$ when the state is s under policy π. Let $d^{\pi}_{P^*}(s)$ be the state occupancy measure

$$d^{\pi}_{P^*}(s) = (1-\gamma) \sum_{h=1}^{\infty} \gamma^{h-1} P^{\star}_{\pi}(s_h = s | s_1 \sim d_1),$$

where $P^*_{\pi}(s_h = s|s_1 \sim d_1)$ is the probability that $s_h = s$, after starting at state $s_1 \sim d_1$ and following π thereafter. Also, let $d^{\pi}_{P^*}(s, a) = \pi(a|s) d^{\pi}_{P^*}(s)$ be the state-action occupation measure. We define the expected value function for the discounted reward under policy π starting from state s as:

$$V^{\pi}_{P^*, r}(s) = \mathbb{E}_{P^*, \pi} \left[\sum_{h=1}^{\infty} \gamma^h r(s_h, a_h) | s_1 = s \right], \tag{1}$$

where $\mathbb{E}$ is taken with respect to the policy π and the transition function P^*. Similarly, we define the action-value function for the reward as:

$$Q^{\pi}_{P^*, r}(s, a) = \mathbb{E}_{P^*, \pi} \left[\sum_{h=1}^{\infty} \gamma^h r(s_h, a_h) | s_1 = s, a_1 = a \right]. \tag{2}$$

For simplicity, when the context is clear, we sometimes abbreviate $V^{\pi}_{P^\star, r}$ and $Q^{\pi}_{P^\star, r}$ as V^{π}_r and Q^{π}_r, respectively. We also define the value function for the utility $V^{\pi}_g(s)$ and the action-value function for the utility $Q^{\pi}_g(s, a)$ analogously by replacing r with g in (1) and (2), respectively. We denote $V^{\pi}_j(s)$, and $Q^{\pi}_j(s, a)$ for $j = r, g$.

Definition 2.1. For brevity, we denote $P^* V^{\pi}_j(s, a) = \mathbb{E}_{s' \sim P^*(\cdot|s,a)} V^{\pi}_j(s')$ for $j = r, g$.

Using this notation, the Bellman's equation associated with the policy π becomes

$$Q^{\pi}_j(s, a) = (j + \gamma P^* V^{\pi}_j)(s, a), \qquad j \in \{r, g\}.$$

Note that $V^{\pi}_j(s) = \langle \pi(\cdot|s), Q^{\pi}_j(s, \cdot) \rangle_{\mathcal{A}}$, where $\langle \pi(\cdot|s), Q^{\pi}_j(s, \cdot) \rangle_{\mathcal{A}} = \sum_{a \in \mathcal{A}} \pi_h(a|s) Q^{\pi}_j(s, a)$.

The objective of the learning agent is to find an optimal policy that maximizes the expected reward while satisfying a safety constraint on the expected utility:

$$\max_{\pi} V^{\pi}_r(s_1), \qquad \text{subject to } V^{\pi}_g(s_1) \geq b. \tag{3}$$

where $b \in (0, 1]$ is a given threshold. While we focus on a single constraint for simplicity, our method naturally extends to handle multiple constraints. We denote the optimal policy as π^* which solves the above optimization problem.

In the absence of prior knowledge about the constraint, the agent must balance exploration and exploitation while ensuring that the policy satisfies the constraint. To this end, we allow the policy to occasionally violate the constraint and aim to minimize the regret and the total constraint violations over K episodes. This approach is consistent with existing literature on constrained RL [12, 13, 16].

To formally measure the performance of the policy sequence $\{\pi_k\}_1^K$, we define the regret as

$$\text{Regret}(K) = \sum_{k=1}^{K} V_r^{\pi^*}(s_1) - V_r^{\pi_k}(s_1),$$

and the soft/hard constraint violations as

$$\text{Violation}_{\text{soft}}(K) = \left[\sum_{k=1}^{K}(b - V_g^{\pi_k}(s_1))\right]_+, \text{Violation}_{\text{hard}}(K) \quad = \sum_{k=1}^{K}\left[b - V_g^{\pi_k}(s_1)\right]_+,$$

where $[z]_+ = \max\{z, 0\}$. The regret measures the gap between the total reward value by following the optimal policy π^*, and the total reward value obtained by following agent's policy π_k at episode k over K episodes. The soft constraint violation quantifies the cumulative difference between the desired constraint threshold and the achieved utility, allowing for both positive and negative deviations over time. In contrast, the hard constraint violation is a stricter metric that measures only the cumulative positive deviations. As shown in [18], even zero soft violation may lead to hard constraint violation that grows linearly with the number of episodes, while soft constraint violation can only grow sub-linearly with the number of episodes if the hard constraint violation grows sub-linearly.

These metrics capture cumulative notions of safety rather than strict zero-violation safe exploration. In particular, soft violation is the standard online-CMDP notion for budget-style or long-run average constraints, while hard violation is a stricter cumulative metric that counts only positive deviations. Throughout the paper, our guarantees concern cumulative violation control; they do not imply per-episode safety during learning.

Dual problem and Slater's Condition. We then introduce some additional notation that will be used throughout this paper.

Definition 2.2. For any policy π and Lagrangian multiplier Y, we define the composite value function and Q-function as $V^{\pi,Y}(s) = V_r^\pi(s) + YV_g^\pi(s)$ and $Q^{\pi,Y}(s,a) = Q_r^\pi(s,a) + YQ_g^\pi(s,a)$.

We can cast problem (3) as a saddle point problem $\max_\pi \min_Y \mathcal{L}(\pi, Y)$, with Lagrangian $\mathcal{L}(\pi, Y) = V_r^\pi(s_1) + Y(V_g^\pi(s_1) - b) = V^{\pi,Y} - Yb$. Although the Lagrangian is non-concave in π [3], the strong duality holds [32]. Hence, there exists an optimal Lagrangian multiplier Y^* such that $\max_\pi \mathcal{L}(\pi, Y^*)$ recovers the optimal reward value function.

We assume the following Slater's condition in this paper.

Assumption 2.3 (Slater's Condition)**.** There exists $\theta > 0$, and a policy $\bar{\pi}$, such that $V_g^{\bar{\pi}}(s_1) \geq b + \theta$.

The Slater's condition is mild in practice and widely used in prior work [13, 16, 33]; it also implies that the optimal dual variable is bounded.

Lemma 2.4 (Boundedness of Y^*)**.** *The optimal dual-variable* $Y^* \leq \dfrac{V_r^{\pi^*}(s_1) - V_r^{\bar{\pi}}(s_1)}{\theta} \leq \dfrac{1}{\theta}$.

Low-rank CMDP. In this paper, we study low-rank CMDPs, which are characterized by a low-rank decomposition of the transition function. Specifically, we assume the following:

Assumption 2.5 (low-rank CMDP)**.** The transition function of the CMDP $P^* : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ admits a low rank decomposition with rank $d \in \mathbb{N}$ if there exists two embedding functions $\phi^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d, \mu^* : \mathcal{S} \to \mathbb{R}^d$ such that

$$\forall s, s' \in \mathcal{S}, a \in \mathcal{A} : P^*(s'|s,a) = \langle \phi^*(s,a), \mu^*(s') \rangle.$$

For normalization, we assume $\|\phi^*(s,a)\| \leq 1$ for all (s,a) and for any functions $f : \mathcal{S} \to [0,1]$, $\|\int \mu^*(s) f(s)\, ds\|_2 \leq \sqrt{d}$.

Remark 2.6. The low-rank CMDP assumption, characterized by the existence of low-dimensional embedding functions $\phi^*(s,a)$ and $\mu^*(s')$, is widely adopted structural assumption in RL [2, 37, 41]. This assumption encompasses several common settings as special cases, including tabular CMDPs, linear CMDPs, and block CMDPs.

Stochastic-oracle feedback model. Following prior low-rank representation-learning work, we study an occupancy-based stochastic-oracle setting. The learner may query states distributed as $d^{\pi}_{P^\star}$ together with auxiliary transitions that are used only for representation learning and bonus construction. Practically, a discounted-occupancy state can be sampled by resetting to the initial distribution and rolling in under π for a geometrically distributed number of steps. Here K counts outer oracle rounds, not raw environment transitions; implementing K queries this way uses $O(K/(1-\gamma)+3K)$ transitions in expectation. Our regret and violation bounds do not charge the complete roll-in trajectories. Likewise, the execution-policy corollaries transfer values to the stationary ϵ-mixture policy rather than giving an end-to-end guarantee for the entire three-transition probing trajectory.

Function approximation setup and MLE oracle. Since the transition function P^* is unknown, we employ a function class $\mathcal{M} = \{(\mu,\phi) : \mu \in \Theta, \phi \in \Gamma\}$ to characterize $\mu^\star$ and $\phi^\star$, and by extension, to approximate P^*. Our formal guarantees use a finite realizable class, $|\mathcal{M}| < \infty$ and $P^\star \in \mathcal{M}$; this restricts the candidate representations rather than the cardinality of the state space. Finite realizability and logarithmic dependence on $|\mathcal{M}|$ are standard in low-rank representation-learning analyses [2, 41]. To select the appropriate transition $\hat{P}$ from the function class $\mathcal{M}$, we assume access to a supervised learning-style maximum-likelihood estimation (MLE) oracle.

Definition 2.7 (MLE Oracle)**.** Given a model class $\mathcal{M}$ and a dataset $\mathcal{D}$ in the form of (s,a,s'), the MLE oracle returns the maximum likelihood estimator $\hat{P} = (\hat{\mu},\hat{\phi}) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}} \ln(\mu(s')^\top \phi(s,a))$.

3 Soft Violation via Cumulative Primal-Dual Control

We first study low-rank CMDPs under the soft cumulative-violation metric, where constraint deficits and surpluses are aggregated before taking the positive part. The learner must identify an unknown low-rank transition representation while updating policies through a primal-dual procedure: return-policy data may lack action coverage, whereas always-uniform actions are less policy-adaptive. REP-PD-TEG (**Rep**resentation-learning **P**rimal-**D**ual with **T**ertiary decaying-$\boldsymbol{\epsilon}$- **G**reedy probing) resolves this tension by keeping the probing kernel close to the previous return policy while retaining a uniform action floor; the tertiary transition supplies the coverage needed for optimism. Unless stated otherwise, regret and soft violation are evaluated on the return policies $\{\pi_k\}_{k=1}^K$; auxiliary tuples are used only for model estimation and bonus construction. We later quantify the cost of charging the execution policy.

3.1 REP-PD-TEG Algorithm

Algorithm 1 gives REP-PD-TEG. At episode k, the algorithm interacts with the environment using the execution policy $q_k(\cdot \mid s) = q'_k(\cdot \mid s) = q''_k(\cdot \mid s) = (1-\epsilon_k)\pi_{k-1}(\cdot \mid s) + \epsilon_k U(\mathcal{A})$, where $U(\mathcal{A})$ is the uniform distribution over the action space and $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_K > 0$. The oracle returns one tertiary tuple $(s,a,s',a',s'',a'',\bar{s})$: it first samples $s \sim d^{\pi_{k-1}}_{P^\star}$, then alternates actions from q_k, q'_k, q''_k with transitions from $P^\star$. The resulting three transitions are stored in separate replay buffers denoted as D_k, D'_k and D''_k. After updating the replay buffers, REP-PD-TEG fits $\widehat{P}_k = (\widehat{\mu}_k, \widehat{\phi}_k)$ by MLE on $D_k + D'_k + D''_k$. The third stream is required by the present *merged-covariance* design, not by MLE alone. Since the covariance contains both D_k and D'_k, the one-step-back proof produces propagated quadratic terms under both primary and secondary designs: D'_k controls the former, while the one-step pushforward D''_k supplies an MLE-controlled denominator for the latter. We therefore do not place D''_k in the covariance, which would create another propagated level. The learned features are then utilized to calculate the empirical covariance matrix $\widehat{\Sigma}^{\mathrm{mix}}_k$, which quantifies the uncertainty associated with these features. Based on the empirical covariance matrix, an exploration bonus, $\widehat{b}_k(x,u)$, is constructed as a scaled Mahalanobis norm of the learned feature $\widehat{\phi}_k$ with respect to $\widehat{\Sigma}^{\mathrm{mix}}_k$. This bonus induces uncertainty-aware planning under the learned model by augmenting both the reward and utility value estimates, after which the return policy and dual variable are updated according to the final two steps of Algorithm 1.

Algorithm 1 REP-PD-TEG

Input: Parameters $\{\lambda_k\}_{k=1}^K$, $\{\alpha_k\}_{k=1}^K$, η, ξ, $\{\epsilon_k\}_{k=1}^K$, model class $\mathcal{M} = \{(\mu, \phi) : \mu \in \Theta, \phi \in \Gamma\}$, number of episodes K.
Initialize return policy $\pi_0(\cdot \mid s) = U(\mathcal{A})$, $Y_1 = 0$, and $D_0 = D_0' = D_0'' = \emptyset$.
for $k = 1, \ldots, K$ **do**
  Set execution policy $q_k = q_k' = q_k'' = (1 - \epsilon_k)\pi_{k-1} + \epsilon_k U(\mathcal{A})$.
  Query one auxiliary tertiary tuple $(s, a, s', a', s'', a'', \bar{s})$, where $s \sim d_{P^\star}^{\pi_{k-1}}$, $a \sim q_k(\cdot \mid s)$, $s' \sim P^\star(\cdot \mid s, a)$, $a' \sim q_k'(\cdot \mid s')$, $s'' \sim P^\star(\cdot \mid s', a')$, $a'' \sim q_k''(\cdot \mid s'')$, $\bar{s} \sim P^\star(\cdot \mid s'', a'')$.
  Update $D_k = D_{k-1} \cup \{(s, a, s')\}$, $D_k' = D_{k-1}' \cup \{(s', a', s'')\}$, $D_k'' = D_{k-1}'' \cup \{(s'', a'', \bar{s})\}$.
  Fit the low-rank transition model by MLE:

$$\widehat{P}_k = (\widehat{\mu}_k, \widehat{\phi}_k) \in \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{D_k + D_k' + D_k''} \big[\log \mu(x')^\top \phi(x, u) \big].$$

  Form the merged empirical covariance

$$\widehat{\Sigma}_k^{\mathrm{mix}} = \sum_{(x,u,x')\in D_k} \widehat{\phi}_k(x, u)\widehat{\phi}_k(x, u)^\top + \sum_{(x,u,x')\in D_k'} \widehat{\phi}_k(x, u)\widehat{\phi}_k(x, u)^\top + \lambda_k I.$$

  Set the covariance mixture bonus

$$\widehat{b}_k(x, u) = \min\left\{ \alpha_k \sqrt{\widehat{\phi}_k(x, u)^\top (\widehat{\Sigma}_k^{\mathrm{mix}})^{-1} \widehat{\phi}_k(x, u)}, \bar{b} \right\}.$$

  Update the return policy $\pi_k \in \arg\max_\pi \left(Q^\pi_{\widehat{P}_k, r+\widehat{b}_k} + Y_k Q^\pi_{\widehat{P}_k, g+\widehat{b}_k} \right)$.
  Update the dual variable $Y_{k+1} = \Pi_{[0,\xi]} \left(Y_k + \eta\big(b - V^{\pi_k}_{\widehat{P}_k, g+\widehat{b}_k}(s_1)\big) \right)$.
end for
Return $\pi_1, \ldots, \pi_K$.

The dual variable is projected into $[0, \xi]$ and we set $\xi = 2/\theta$ in the analysis. Reward optimism promotes exploration, while utility optimism is valid for soft violation because episode-level deficits and surpluses are aggregated. Thus one episode consists of auxiliary tertiary sampling, representation learning, bonus construction, optimistic primal update, and projected dual update.

Algorithm 1 uses exact MLE and exact optimistic-Lagrangian planning to isolate the statistical challenge of learning unknown representations. The proof exposes a conditional interface for other learners: a representation method may replace MLE if it establishes the same prediction event under all three probing designs, with any excess prediction error added to the representation-error term. Likewise, full-planning suboptimality enters through the cumulative planning term. A raw MLE objective gap need not imply the required prediction event, so the present theorems do not claim guarantees for arbitrary approximate oracles, misspecified models, or infinite classes. Spectral representation learning [37] is an example of a non-MLE alternative, but its three-design event would have to be verified separately here.

3.2 Polynomial-Schedule Guarantees

We state the guarantee for polynomial schedules $\epsilon_k = k^{-p}$, which make the exploration–deployment tradeoff explicit. The return-policy bound captures the statistical cost of learning the representation from auxiliary probing data, while the execution-policy statement adds the standard performance-difference cost of actually deploying the mixed probing policy.

Theorem 3.1 (Return policy and execution policy guarantees for polynomial probing)**.** *Suppose Assumptions 2.3 and 2.5 hold and the finite model class is realizable, $|\mathcal{M}| < \infty$ and $P^\star \in \mathcal{M}$. Let $\epsilon_k = k^{-p}$ for some $p \in [0, 1/2)$. Run Algorithm 1 with $\alpha_k = \widetilde{O}(\sqrt{(|\mathcal{A}|k^p + d^2)\gamma \log(|\mathcal{M}|k/\delta)})$, $\lambda_k = \widetilde{\Theta}(((|\mathcal{A}|k^p + d^2)/d) \log(|\mathcal{M}|k/\delta))$, $\bar{b} \geq \max_{k\leq K} \alpha_k/\sqrt{\lambda_k}$, $\eta = 2(1-\gamma)/(\theta\bar{b}\sqrt{K})$, $\xi = 2/\theta$. Then, with probability at least $1 - \delta$,*

$$\mathrm{Regret}^{\mathrm{return}}(K) \leq \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2 K^{1/2+p}}{\theta(1-\gamma)^2} \right), \qquad \mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) \leq \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2 K^{1/2+p}}{(1-\gamma)^2} \right).$$

Moreover, for the execution policy $\widetilde{\pi}_k = (1-\epsilon_k)\pi_k + \epsilon_k U(\mathcal{A})$,

$$\mathrm{Regret}^{\mathrm{exec}}(K) \le \mathrm{Regret}^{\mathrm{return}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K} k^{-p},$$

$$\mathrm{Violation}^{\mathrm{exec}}_{\mathrm{soft}}(K) \le \mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K} k^{-p}.$$

The proof is given in Appendix D. Generally, our analyses combine a tuple-level MLE event, a pathwise-to-averaged bridge, tertiary-aware optimism for the merged bonus, and a cumulative overestimation bound for the dual update. The schedule parameter p trades coverage for deployment cost: smaller p improves the return-policy rate, while larger p reduces the cost of executing the exploratory mixture. Since $\sum_{k=1}^{K} k^{-p} \le K^{1-p}/(1-p)$, the return-policy term scales as $\widetilde{O}(K^{1/2+p})$, whereas execution adds $O(K^{1-p}/(1-\gamma))$.

Corollary 3.2 (Best return policy and execution policy rates)**.** *Under the same setting as Theorem 3.1, the following two specializations hold.*

- *First, if* $p = 0$*, equivalently* $\epsilon_k \equiv 1$*, then the probing kernel in Algorithm 1 becomes uniform:* $q_k = q'_k = q''_k = U(\mathcal{A})$*. The return policies satisfy*

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2\sqrt{K}}{\theta(1-\gamma)^2}\right), \quad \mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) \le \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2\sqrt{K}}{(1-\gamma)^2}\right).$$

  This is the best return policy specialization, and reduces to uniform tertiary probing.

- *Second, if* $p = 1/4$*, then the return policies satisfy*

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2K^{3/4}}{\theta(1-\gamma)^2}\right), \quad \mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) \le \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2K^{3/4}}{(1-\gamma)^2}\right).$$

  The execution policy gap is also $O(K^{3/4}/(1-\gamma))$*, and the overall sample complexities of the execution policies become*

$$\mathrm{Regret}^{\mathrm{exec}}(K) \le \widetilde{O}(K^{3/4}), \mathrm{Violation}^{\mathrm{exec}}_{\mathrm{soft}}(K) \le \widetilde{O}(K^{3/4})$$

  up to logarithmic and problem-dependent factors.

Corollary 3.2 highlights the schedule tradeoff: $p = 0$ is statistically sharp for return policies, while $p = 1/4$ balances representation learning with the cost of using the execution policy. Thus the return-policy theorem isolates representation-learning cost, and the execution policy corollary quantifies the extra price of deploying the probes as real behavior.

4 Hard Violation via Per-Episode Certification

We now consider the hard cumulative-violation metric. Unlike soft violation, hard violation does not allow constraint surplus in one episode to offset constraint deficit in another. Therefore the optimistic utility estimate used by the soft primal-dual update is insufficient for certifying feasibility at the episode level: an optimistic estimate may exceed the threshold b even when the true utility remains below it. REP-PD-H-TEG (the **H**ard-violation version of REP-PD-TEG) therefore keeps the same representation-learning procedure but changes the policy decision: policy search remains optimistic, while a tight feasibility test on the lower-confidence utility of the policy is explicitly conducted.

4.1 Lower-Certificate Planning in REP-PD-H-TEG

As shown in Algorithm 2 in Appendix F, REP-PD-H-TEG uses the same tertiary data collection, MLE representation update, merged covariance, and bonus construction as REP-PD-TEG. The hard-specific change lies in how the policy is selected after the bonus is built. In REP-PD-TEG, the optimistic utility value can be used in the dual update because violation is aggregated across episodes. In the hard violation setting, however, the policy executed in the current episode must be

certified feasible on its own. Thus REP-PD-H-TEG introduces a lower-confidence utility certificate. Write $\Gamma_k := |\mathcal{A}|/\epsilon_k$ and $\zeta_k := \ln(|\mathcal{M}|k/\delta)/k$, and choose a sufficiently large universal calibration constant C_{cert}. The one-sided statistical correction and the certificate are

$$\ell_k := C_{\mathrm{cert}} \sqrt{\frac{\Gamma_k \zeta_k}{1-\gamma}}, \qquad \underline{\mathcal{U}}_k(\pi) := \left[V^{\pi}_{\widehat{P}_k, g-\widehat{b}_k}(s_1) - \ell_k \right]_+ ,$$

where $[x]_+ = \max\{x, 0\}$. On the representation-learning good event, the one-step-back argument gives, simultaneously for every queried policy,

$$\underline{\mathcal{U}}_k(\pi) \le V^{\pi}_{P^\star, g}(s_1).$$

Hence $\underline{\mathcal{U}}_k(\pi) \ge b$ implies true feasibility. The corresponding two-sided width is needed only in the analysis of the selected policy and is

$$w_k(\pi) := 2\ell_k + 2V^{\pi}_{\widehat{P}_k, \widehat{b}_k}(s_1), \qquad V^{\pi}_{P^\star, g}(s_1) \le \underline{\mathcal{U}}_k(\pi) + w_k(\pi).$$

Separating ℓ_k from $w_k(\pi)$ is essential: the former makes the stopping rule conservative, while the latter accounts for the pessimistic bonus and is summed only along the returned policy sequence.

Instead of choosing a single policy at the current dual variable, REP-PD-H-TEG performs a capped dual search within the episode in the planning step. For each queried dual level Y, it solves the optimistic Lagrangian planning problem

$$\widehat{\pi}_{k,Y} \in \arg\max_{\pi} \left(Q^{\pi}_{\widehat{P}_k, r+\widehat{b}_k} + Y Q^{\pi}_{\widehat{P}_k, g+\widehat{b}_k} \right),$$

and then forms a statewise softmax candidate from the corresponding optimistic reward-plus-utility scores. The bonus makes planning optimistic and uncertainty-aware, exactly as in REP-PD-TEG; the stopping rule is different. REP-PD-H-TEG deploys the first candidate whose lower-confidence certificate satisfies $\underline{\mathcal{U}}_k(\pi_{k,Y}) \ge b$. If no candidate is certified before the cap $Y = K^{1/4}$, the algorithm deploys the final capped-dual candidate. Thus optimistic bonus-augmented values are used to search for good policies, while the pessimistic bonus-based certificate decides whether a candidate is safe enough to execute.

Capped certificate search in one outer round.

1: For each grid level $Y \le Y_{\max}$, compute the optimistic candidate $\pi_{k,Y}$.
2: Evaluate $\underline{\mathcal{U}}_k(\pi_{k,Y})$; if it is at least b, deploy $\pi_{k,Y}$ and stop.
3: Otherwise increase Y and repeat the optimistic-plan/certificate test.
4: If no earlier candidate passes, deploy the candidate at $Y_{\max}$.

4.2 Hard-Violation Guarantees

We state the sample complexity result of REP-PD-H-TEG in this section. The proof decomposes into a front-end regret bound, a lower-certificate reduction, and a Slater-induced search-success argument; the full proof is given in Appendices F and G.

Assumption/interface	Role in the proof	Results
Low rank + finite realizability	MLE prediction event and one-step-back control	Soft + hard
Three-stream coverage	Controls both propagated terms from the merged covariance	Soft + hard
Slater margin $\theta > 0$	Bounds the dual variable and cap-induced deficit	Soft + hard
Positive score gap	Controls the hard-search softmax approximation	Hard only

The positive score gap is not used to learn the representation or to establish the lower-certificate inequality; it is used by the current hard proof to control the softmax approximation along the capped dual search.

Theorem 4.1 (Return policy and execution policy hard guarantees for REP-PD-H-TEG)**.** *Suppose Assumptions 2.3, 2.5, and C.1 hold, the finite model class is realizable* ($|\mathcal{M}| < \infty$, $P^\star \in \mathcal{M}$)*, and fix* $\delta \in (0,1)$*. Run Algorithm 2 with* $\epsilon_k = k^{-p}$ *for some* $p \in [0, 1/2)$*, and set* $\alpha_k = \widetilde{\Theta}(\sqrt{(|\mathcal{A}|k^p + d^2)\gamma \log(|\mathcal{M}|k/\delta)})$ *with a sufficiently large leading constant,* $\lambda_k = \widetilde{\Theta}(((|\mathcal{A}|k^p +$

$d^2)/d)\log(|\mathcal{M}|k/\delta))$, $\bar{b} \geq \max_{k\leq K} \alpha_k/\sqrt{\lambda_k}$, $\ell_k = C_{\mathrm{cert}}\sqrt{\Gamma_k\zeta_k/(1-\gamma)}$ *for a sufficiently large universal* C_{cert}, $Y_{\max} = K^{1/4}$, $\tau_h = \Theta(1/((1-\gamma)\sqrt{K}\log|\mathcal{A}|))$, $\eta_h = \Theta(\tau_h(1-\gamma)/K^{3/4})$. *Then, with probability at least* $1-\delta$, *the return policies satisfy*

$$\mathrm{Regret}^{\mathrm{return}}(K) \leq \widetilde{O}\left(\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{3/4+p}\right),$$

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{hard}}(K) \leq \widetilde{O}\left(\frac{1}{\theta}\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{1/2+p}\right) + O(\theta^{-4}).$$

The hard-violation bound is sublinear for $p < 1/2$, *while the regret bound is sublinear only when* $p < 1/4$. *Moreover, for the execution policy* $\widetilde{\pi}_k = (1-\epsilon_k)\pi_k + \epsilon_k U(\mathcal{A})$, *we have* $\mathrm{Regret}^{\mathrm{exec}}(K) \leq \mathrm{Regret}^{\mathrm{return}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K} k^{-p}$ *and* $\mathrm{Violation}^{\mathrm{exec}}_{\mathrm{hard}}(K) \leq \mathrm{Violation}^{\mathrm{return}}_{\mathrm{hard}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K} k^{-p}$.

The $K^{1/2+p}$ hard-violation term is the representation-learning price, and the extra $K^{1/4}$ in regret comes from capped in-episode dual search; the execution policy bounds follow similar performance-difference argument from Section 3.

Corollary 4.2 (Best return policy and execution policy hard rates)**.** *Under the same setting as Theorem 4.1, the following two specializations hold.*

- *First, if* $p = 0$, *equivalently* $\epsilon_k \equiv 1$, *then REP-PD-H-TEG uses* $q_k = q'_k = q''_k = U(\mathcal{A})$ *and the return policies satisfy*

$$\mathrm{Regret}^{\mathrm{return}}(K) \leq \widetilde{O}\left(\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{3/4}\right),$$

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{hard}}(K) \leq \widetilde{O}\left(\frac{1}{\theta}\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] \sqrt{K}\right) + O(\theta^{-4}).$$

- *Second, if* $p = 1/8$, *then the return policies satisfy*

$$\mathrm{Regret}^{\mathrm{return}}(K) \leq \widetilde{O}\left(\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{7/8}\right),$$

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{hard}}(K) \leq \widetilde{O}\left(\frac{1}{\theta}\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{5/8}\right) + O(\theta^{-4}).$$

  For $p = 1/8$, *the execution-policy gap is* $O(K^{7/8}/(1-\gamma))$*; hence both execution-policy regret and hard violation are* $\widetilde{O}(K^{7/8})$ *up to problem-dependent factors, with exploratory actions determining the balanced rate.*

Thus $p = 0$ is sharp for return policies, whereas $p = 1/8$ balances the hard-regret front end $K^{3/4+p}$ with the execution gap K^{1-p}.

5 Conclusion

We develop representation-learning algorithms for low-rank continuous-state CMDPs with unknown transition features. REP-PD-TEG combines auxiliary probing, covariance bonuses, and optimistic primal-dual planning for soft-violation control; REP-PD-H-TEG adds a lower-confidence utility certificate and stabilized dual search for hard-violation control. Both provide return- and execution-policy guarantees.

Limitations. The theorems assume finite realizability and exact MLE/planning; alternative learners require the same three-design prediction guarantee. The hard result additionally uses a positive optimistic-score gap. The occupancy oracle is implementable with resets and geometric roll-ins, but our bounds do not charge full roll-in trajectories. Finally, the experiments illustrate the mechanism rather than exhaustively evaluating hard certification.

Acknowledgments and Disclosure of Funding

Funding. There was no funding in direct support of this work.

Competing interests. The authors declare no competing interests.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[2] 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.

[3] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[4] Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

[5] 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.

[6] Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models. *arXiv preprint arXiv:2201.09802*, 2022.

[7] Arko Banerjee, Kia Rahmani, Joydeep Biswas, and Isil Dillig. Dynamic model predictive shielding for provably safe reinforcement learning. *Advances in Neural Information Processing Systems*, 37:100131–100159, 2024.

[8] Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

[9] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in neural information processing systems*, 35:1047–1059, 2022.

[10] Yuan Cheng, Ruiquan Huang, Jing Yang, and Yingbin Liang. Improved sample complexity for reward-free reinforcement learning under low-rank mdps. *arXiv preprint arXiv:2303.10859*, 2023.

[11] Juntao Dai, Jiaming Ji, Long Yang, Qian Zheng, and Gang Pan. Augmented proximal policy optimization for safe reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7288–7295, 2023.

[12] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *NeurIPS*, 2020.

[13] 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.

[14] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36:66138–66200, 2023.

[15] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

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

[17] 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.

[18] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2024.

[19] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[20] Ruiquan Huang, Jing Yang, and Yingbin Liang. Safe exploration incurs nearly no additional sample complexity for reward-free rl. *arXiv preprint arXiv:2206.14057*, 2022.

[21] Weidong Huang, Jiaming Ji, Chunhe Xia, Borong Zhang, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. *arXiv preprint arXiv:2307.07176*, 2023.

[22] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[23] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[24] Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8030–8037, 2021.

[25] Toshinori Kitamura, Tadashi Kozuno, Masahiro Kato, Yuki Ichihara, Soichiro Nishimori, Akiyoshi Sannai, Sho Sonoda, Wataru Kumagai, and Yutaka Matsuo. A policy gradient primal-dual algorithm for constrained mdps with uniform pac guarantees. *arXiv preprint arXiv:2401.17780*, 2024.

[26] Toshinori Kitamura, Arnob Ghosh, Tadashi Kozuno, Wataru Kumagai, Kazumi Kasaura, Kenta Hoshino, Yohei Hosoe, and Yutaka Matsuo. Provably efficient rl under episode-wise safety in linear cmdps. *arXiv preprint arXiv:2502.10138*, 2025.

[27] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

[28] 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.

[29] 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.

[30] Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

[31] Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

[32] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 68(3):1321–1336, 2022.

[33] Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual optimization: Stochastically constrained markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.

[34] Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pages 18168–18210. PMLR, 2022.

[35] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

[36] Tongzheng Ren, Chenjun Xiao, Tianjun Zhang, Na Li, Zhaoran Wang, Sujay Sanghavi, Dale Schuurmans, and Bo Dai. Latent variable representation for reinforcement learning. *arXiv preprint arXiv:2212.08765*, 2022.

[37] Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022.

[38] 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.

[39] Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[40] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

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

[42] Honghao Wei, Xin Liu, and Lei Ying. Safe reinforcement learning with instantaneous constraints: the role of aggressive exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21708–21716, 2024.

[43] Yunchang Yang, Tianhao Wu, Han Zhong, Evrard Garcelon, Matteo Pirotta, Alessandro Lazaric, Liwei Wang, and Simon S Du. A reduction-based framework for conservative bandits and reinforcement learning. *arXiv preprint arXiv:2106.11692*, 2021.

[44] Hongming Zhang, Tongzheng Ren, Chenjun Xiao, Dale Schuurmans, and Bo Dai. Provable representation with efficient planning for partially observable reinforcement learning. *arXiv preprint arXiv:2311.12244*, 2023.

[45] Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

[46] 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.

[47] Chenhao Zhou and Jiyu Wei. Toward online robust zero-sum markov games with function approximation. *arXiv preprint arXiv:2610.26286*, 2026. Accepted at NeurIPS 2026.

[48] Chenhao Zhou, Zebang Shen, Hanbin Zhao, Hui Qian, et al. Solving zero-sum markov games with continuous state via spectral dynamic embedding. *Advances in Neural Information Processing Systems*, 37:68622–68657, 2024.

Appendix Roadmap

The appendix is organized as follows. Appendix A fixes notation and records the basic covariance concentration tool. Appendix B collects the MLE, elliptical-potential, simulation, and primal-dual lemmas used throughout the proofs. Appendix C states the technical ingredients specific to the hard-violation analysis. Appendix D proves the tertiary-aware decaying-ϵ representation-learning result for REP-PD-TEG, while Appendix E gives the $p = 0$ uniform specialization. Appendix F derives the one-sided hard certificate and proves the decaying-schedule result, while Appendix G records its $p = 0$ specialization. Appendix H contains the experimental details.

A Notation and Auxiliary Concentration

This appendix fixes the occupancy notation used throughout the proof and recalls the empirical-to-population covariance comparison that underlies the uncertainty bonus. Later appendices replace the uniform probing distribution below by the tertiary decaying-ϵ design, but the same comparison principle is repeatedly used.

We use $d^{\pi}_{P^\star}$ for the normalized discounted state occupancy and $d^{\pi}_{P^\star}(s,a) = d^{\pi}_{P^\star}(s)\pi(a \mid s)$ for the state-action occupancy. For the uniform probing baseline, let $\rho_k(s) = k^{-1}\sum_{i=0}^{k-1} d^{\pi_i}_{P^\star}(s)$, and let ρ'_k be the marginal distribution of s' generated by first drawing $s \sim \rho_k$, then $a \sim U(\mathcal{A})$, and finally $s' \sim P^\star(\cdot \mid s,a)$. For a feature map ϕ, define

$$\Sigma_{\rho_k \times U(\mathcal{A}),\phi} := k\mathbb{E}_{s\sim\rho_k, a\sim U(\mathcal{A})}[\phi(s,a)\phi(s,a)^\top] + \lambda_k I.$$

Lemma A.1 (Concentration of the bonus term)**.** *Let*

$$\hat{\Sigma}_{k,\phi} := \sum_{i=0}^{k-1} \phi(s^{(i)},a^{(i)})\phi(s^{(i)},a^{(i)})^\top + \lambda_k I, \qquad \lambda_k = \Theta(d\log(k|\Phi|/\delta)).$$

Then, with probability at least $1-\delta$, simultaneously for all $k \in \mathbb{K}^+$, all $\phi \in \Phi$, and all (s,a),

$$c_1\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_k\times U(\mathcal{A}),\phi}} \le \|\phi(s,a)\|_{\hat{\Sigma}^{-1}_{k,\phi}} \le c_2\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_k\times U(\mathcal{A}),\phi}},$$

where $c_1, c_2 > 0$ are universal constants.

Proof. This is the standard empirical-to-population covariance comparison for the uniform probing design; see Lemma 11 of Uehara et al. [41]. In the repaired analysis we apply the same concentration argument to block increments generated by the predictable auxiliary streams. □

B Auxiliary Lemmas

This appendix collects standard auxiliary tools used by both the soft and hard analyses. We first recall the MLE guarantee for low-rank transition models, then state the elliptical-potential and simulation lemmas used to aggregate model error and convert value differences into regret and violation bounds. The MLE statement is a direct consequence of Theorem 21 of Agarwal et al. [2], specialized to the notation of our algorithms.

Lemma B.1 (MLE guarantee)**.** *For a fixed episode k, with probability $1-\delta$,*

$$\mathbb{E}_{s\sim\{0.5\rho_k+0.5\rho'_k\},a\sim U(\mathcal{A})}[\|\hat{P}_k(\cdot\,|\,s,a) - P^\star(\cdot\,|\,s,a)\|_1^2] \lesssim \zeta,$$

where $\zeta := \frac{\ln(|\mathcal{M}|/\delta)}{k}$. As a straightforward corollary, with probability $1-\delta$, $\forall k \in \mathbb{K}^+$,

$$\mathbb{E}_{s\sim\{0.5\rho_k+0.5\rho'_k\},a\sim U(\mathcal{A})}[\|\hat{P}_k(\cdot\,|\,s,a) - P^\star(\cdot\,|\,s,a)\|_1^2] \lesssim 0.5\zeta_k,$$

where $\zeta_k := \frac{\ln(|\mathcal{M}|k/\delta)}{k}$.

The next two lemmas are the standard determinant-trace inequalities used in the elliptical-potential argument; see Lemma G.2 of Agarwal et al. [2].

Lemma B.2. *Consider the following process. For $k = 1, \cdots, K$, $M_k = M_{k-1} + G_k$ with $M_0 = \lambda_0 I$ and G_k being a positive semidefinite matrix with eigenvalues upper-bounded by 1. We have that:*

$$2\ln\det(M_K) - 2\ln\det(\lambda_0 I) \geq \sum_{k=1}^{K} \mathit{Tr}(G_k M_{k-1}^{-1}).$$

Lemma B.3 (Potential function lemma)**.** *Suppose $\mathit{Tr}(G_k) \leq B^2$.*

$$2\ln\det(M_K) - 2\ln\det(\lambda_0 I) \leq d\ln\left(1 + \frac{KB^2}{d\lambda_0}\right).$$

Lemma B.4 (Simulation lemma)**.** *For any $c \in \{r, g\}$, two CMDPs $(P', c+b)$ and (P, c), and any policy π, we have*

$$V^{\pi}_{P',c+b} - V^{\pi}_{P,c} = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi}_{P'}}\left[b(s,a) + \gamma\mathbb{E}_{P'}Q^{\pi}_{P,c}(s',\pi) - \gamma\mathbb{E}_{P}Q^{\pi}_{P,c}(s',\pi)\right],$$
$$V^{\pi}_{P',c+b} - V^{\pi}_{P,c} = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi}_{P}}\left[b(s,a) + \gamma\mathbb{E}_{P'}Q^{\pi}_{P,c+b}(s',\pi) - \gamma\mathbb{E}_{P}Q^{\pi}_{P',c+b}(s',\pi)\right].$$

Finally, we use the following primal-dual reduction, which is Lemma 9 of Ding et al. [13] in our notation.

Lemma B.5. *Let Y^* be the optimal dual variable, and $C \geq 2Y^*$, then, if*

$$V_r^{\pi^*}(s_1) - V_r^{\tilde{\pi}}(s_1) + C[b - V_g^{\tilde{\pi}}(s_1)]_+ \leq \delta$$

then

$$[b - V_g^{\tilde{\pi}}(s_1)]_+ \leq \frac{2\delta}{C}.$$

C Hard-Constraint Technical Lemmas

This appendix collects the hard-constraint technical ingredients that are reused by the lower-certificate analyses in Appendices G and F. The hard-violation guarantees in this draft are stated through lower-confidence certificates and Slater-induced control of cap episodes; no optimistic feasibility check is used as a hard certificate.

Assumption C.1 (Positive action gap)**.** For every episode k, queried dual level Y, and state s, the maximizer of the optimistic Lagrangian action score used by Algorithm 2 is separated from every suboptimal action by at least $\delta_{\min} > 0$. The constant is uniform over k, Y, s.

Lemma C.2 (Softmax sensitivity)**.** *Let π_k^Y and $\pi_k^{Y+\eta}$ be the softmax policies produced by two neighboring dual levels Y and $Y + \eta$ with temperature τ. Then*

$$\left|V_g^{\pi_k^Y}(s_1) - V_g^{\pi_k^{Y+\eta}}(s_1)\right| \leq O\left(\frac{\eta}{\tau(1-\gamma)}\right).$$

Proof. Let $\Delta = \sup_s |V_g^{\pi_k^Y}(s) - V_g^{\pi_k^{Y+\eta}}(s)|$. Comparing the two Bellman equations gives

$$\Delta \leq \gamma\Delta + \sup_s \|\pi_k^Y(\cdot \mid s) - \pi_k^{Y+\eta}(\cdot \mid s)\|_1.$$

The softmax map is $O(1/\tau)$-Lipschitz in its input. Increasing the dual level by η changes the softmax score by $O(\eta)$ under the normalized utility scale, and therefore

$$\sup_s \|\pi_k^Y(\cdot \mid s) - \pi_k^{Y+\eta}(\cdot \mid s)\|_1 \leq O(\eta/\tau).$$

Rearranging the Bellman inequality proves the claim. □

D Decaying ϵ-Greedy Probing: Tertiary-Aware Analysis

This appendix proves the decaying-ϵ tertiary-aware result used in Section 3. The proof is organized around three steps:

1. the naive two-level decaying-ϵ probing rule is not enough to close the optimism step;
2. a minimal tertiary-aware repair restores the missing denominator;
3. for this repaired algorithm, one can prove a regret guarantee for the *return policy sequence* $\{\pi_k\}_{k=1}^K$.

The guarantees in this appendix are for the return policies of Algorithm 1. The execution-policy transfer in Section 3 is obtained afterward through a performance-difference argument. The same tertiary-aware front end also supports soft-violation control once the cumulative optimistic overestimation of the utility is bounded explicitly. The hard-violation analysis in later appendices uses the same representation-learning front end, but replaces optimistic feasibility checks with a lower-confidence certificate.

D.1 Why the Naive Two-Level Variant Is Not Enough

The natural first attempt is the two-level decaying-ϵ rule

$$q_i(\cdot \mid s) = q_i'(\cdot \mid s) = (1-\epsilon_i)\pi_{i-1}(\cdot \mid s) + \epsilon_i U(\mathcal{A}).$$

This already improves the sampling rule relative to pure uniform probing, but it does not fully solve the coverage problem needed by the optimism proof.

Proposition D.1 (Two-level obstruction)**.** *There is no universal constant $C < \infty$ such that the pointwise comparison*

$$\|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{pri}})^{-1}} \le C\,\|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}}, \qquad \forall (s,a),$$

holds for every episode k under the naive two-level decaying-ϵ probing rule.

Proof. Consider a deterministic two-state MDP with states s_0, s_1 and two actions $a_{\mathrm{stay}}, a_{\mathrm{jump}}$. Let the deployed policy always choose a_{stay} at s_0, let s_1 be absorbing, and let a_{jump} be the only action that can move the process from s_0 to s_1. Choose a bounded feature map such that

$$\hat{\phi}_k(s_0,a) = e_1, \qquad \hat{\phi}_k(s_1,a) = e_2, \qquad \forall a \in \mathcal{A}.$$

Then all primary samples stay at s_0, so

$$\Sigma_k^{\mathrm{pri}} = k e_1 e_1^\top + \lambda_k I.$$

However, the secondary stream reaches s_1 with probability of order $\epsilon_k/|\mathcal{A}|$, hence

$$\Sigma_k^{\mathrm{mix}} \succeq k e_1 e_1^\top + \frac{k\epsilon_k}{|\mathcal{A}|} e_2 e_2^\top + \lambda_k I.$$

Evaluating the two norms at any point with feature vector e_2 gives

$$\|e_2\|_{(\Sigma_k^{\mathrm{pri}})^{-1}} = \frac{1}{\sqrt{\lambda_k}}, \qquad \|e_2\|_{(\Sigma_k^{\mathrm{mix}})^{-1}} \le \frac{1}{\sqrt{\lambda_k + k\epsilon_k/|\mathcal{A}|}}.$$

Therefore

$$\frac{\|e_2\|_{(\Sigma_k^{\mathrm{pri}})^{-1}}}{\|e_2\|_{(\Sigma_k^{\mathrm{mix}})^{-1}}} \ge \sqrt{1 + \frac{k\epsilon_k}{|\mathcal{A}|\lambda_k}},$$

which rules out a universal constant bridge. □

Proposition D.1 explains why the clean theorem is not for the naive two-level rule. The minimal repair for the present merged-covariance design is to add one more probe level. This obstruction does not rule out a primary-only alternative that fits on $D + D'$ but constructs its covariance from D alone, as in the architecture of Uehara et al. [41]. Such a design removes the secondary covariance term but requires a different potential argument; the current regret aggregation cannot be transferred to it directly.

D.2 The Tertiary-Aware Repaired Algorithm

For each round $i \in [K]$, define the primary, secondary, and tertiary state marginals by

$$\rho_i(s) := d_{P^\star}^{\pi_{i-1}}(s),$$

$$\rho_i'(s') := \mathbb{E}_{(s,a)\sim\rho_i\times q_i}\big[P^\star(s' \mid s,a)\big],$$

$$\rho_i''(s'') := \mathbb{E}_{(s,a)\sim\rho_i'\times q_i'}\big[P^\star(s'' \mid s,a)\big].$$

The corresponding probing distributions are

$$\nu_i(s,a) := \rho_i(s)q_i(a \mid s), \quad \nu_i'(s,a) := \rho_i'(s)q_i'(a \mid s), \quad \nu_i''(s,a) := \rho_i''(s)q_i''(a \mid s).$$

We write

$$\bar\nu_k^{\mathrm{pri}} := \frac{1}{k}\sum_{i=1}^k \nu_i, \qquad \bar\nu_k^{\mathrm{sec}} := \frac{1}{k}\sum_{i=1}^k \nu_i', \qquad \bar\nu_k^{\mathrm{ter}} := \frac{1}{k}\sum_{i=1}^k \nu_i'',$$

and

$$\bar\nu_k^{\mathrm{all}} := \frac{1}{3}\big(\bar\nu_k^{\mathrm{pri}} + \bar\nu_k^{\mathrm{sec}} + \bar\nu_k^{\mathrm{ter}}\big).$$

The repaired probing kernels are

$$q_k(\cdot \mid s) = q_k'(\cdot \mid s) = q_k''(\cdot \mid s) = (1-\epsilon_k)\pi_{k-1}(\cdot \mid s) + \epsilon_k U(\mathcal{A}),$$

where

$$\epsilon_1 \ge \epsilon_2 \ge \cdots \ge \epsilon_K > 0.$$

We also write

$$\Gamma_k := \frac{|\mathcal{A}|}{\epsilon_k}, \qquad \Lambda_k := \frac{1}{1-\epsilon_k+\epsilon_k/|\mathcal{A}|}.$$

The coverage coefficients used below are domination constants between the target distributions appearing in the proof and the three empirical probing streams. For two state-action distributions η, ν, define

$$\mathsf{D}(\eta\|\nu) := \inf\left\{c \ge 0 : \mathbb{E}_\eta[g^2] \le c\,\mathbb{E}_\nu[g^2] \text{ for all } g\right\}.$$

When densities exist, this is the essential supremum of $d\eta/d\nu$. For a state-action distribution ν, let $\mathsf{T}_\pi\nu$ denote the one-step pushforward obtained by drawing $(\tilde s,\tilde a) \sim \nu$, $s \sim P^\star(\cdot \mid \tilde s,\tilde a)$, and $a \sim \pi(\cdot \mid s)$. With this notation,

$$\begin{aligned}
\mathfrak{C}_k^{\mathrm{init}}(\pi) &:= \mathsf{D}\Big((1-\gamma)d_1 \otimes \pi \;\Big\|\; \bar\nu_k^{\mathrm{pri}}\Big),\\
\mathfrak{C}_k^{\mathrm{next}}(\pi) &:= \mathsf{D}\Big(\mathsf{T}_\pi\bar\nu_k^{\mathrm{pri}} \;\Big\|\; \bar\nu_k^{\mathrm{sec}}\Big),\\
\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi) &:= \mathfrak{C}_k^{\mathrm{next}}(\pi),\\
\mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi) &:= \mathsf{D}\big(\mathsf{T}_\pi\bar\nu_k^{\mathrm{sec}} \;\big\|\; \bar\nu_k^{\mathrm{ter}}\big),\\
\mathfrak{C}_k^{\mathrm{tri,next}}(\pi) &:= \max\Big\{\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi), \mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi)\Big\}.
\end{aligned}$$

Thus $\mathfrak{C}_k^{\mathrm{init}}$ compares the initial discounted mass to the primary stream, $\mathfrak{C}_k^{\mathrm{next}}$ compares the primary-to-secondary one-step design, and $\mathfrak{C}_k^{\mathrm{tri,next}}$ is the maximum of the two one-step comparisons needed for the tertiary one-step-back argument.

The algorithm analyzed in this appendix is the main-text Algorithm 1. The repair relative to the earlier two-stream protocol is entirely in the auxiliary sampling design: it adds one more probe-action / transition pair, while keeping the bonus covariance merged over the primary and secondary tuples.

D.3 Supporting Lemmas

We now record the small collection of lemmas needed for the final theorem.

Lemma D.2 (Three-stream MLE event at tuple level)**.** *Under realizability $P^\star \in \mathcal{M}$, with probability at least $1-\delta$, simultaneously for all $k \in [K]$,*

$$\mathbb{E}_{(s,a)\sim\widetilde{\nu}_k^{\mathrm{all}}}\left[\|\hat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2\right] \lesssim \frac{\ln(|\mathcal{M}|k/\delta)}{k} =: \zeta_k.$$

Here

$$\widetilde{\nu}_k^{\mathrm{sec}} := \frac{1}{k}\sum_{i=1}^{k} \delta_{s_i'} q_i'(\cdot \mid s_i'), \qquad \widetilde{\nu}_k^{\mathrm{ter}} := \frac{1}{k}\sum_{i=1}^{k} \delta_{s_i''} q_i''(\cdot \mid s_i''),$$

and

$$\widetilde{\nu}_k^{\mathrm{all}} := \frac{1}{3}\big(\bar{\nu}_k^{\mathrm{pri}} + \widetilde{\nu}_k^{\mathrm{sec}} + \widetilde{\nu}_k^{\mathrm{ter}}\big).$$

Proof. For episode i, write the realized tertiary tuple as $(s_i, a_i, s_i', a_i', s_i'', a_i'', \bar{s}_i)$, where $\bar{s}_i \sim P^\star(\cdot \mid s_i'', a_i'')$ is the next state of the third probe transition. Define

$$x_{3i-2} := (s_i, a_i), \qquad y_{3i-2} := s_i',$$
$$x_{3i-1} := (s_i', a_i'), \qquad y_{3i-1} := s_i'',$$

and

$$x_{3i} := (s_i'', a_i''), \qquad y_{3i} := \bar{s}_i.$$

Let $\mathcal{F}_{3i-3}$ be the sigma-field generated by all data collected before episode i starts. Recursively define

$$\mathcal{F}_{3i-2} := \sigma(\mathcal{F}_{3i-3}, x_{3i-2}, y_{3i-2}), \quad \mathcal{F}_{3i-1} := \sigma(\mathcal{F}_{3i-2}, x_{3i-1}, y_{3i-1}),$$

and

$$\mathcal{F}_{3i} := \sigma(\mathcal{F}_{3i-1}, x_{3i}, y_{3i}).$$

Then the predictable covariate designs are

$$D_{3i-2} := \mathrm{Law}(x_{3i-2} \mid \mathcal{F}_{3i-3}) = \nu_i,$$
$$D_{3i-1} := \mathrm{Law}(x_{3i-1} \mid \mathcal{F}_{3i-2}) = \delta_{s_i'} q_i'(\cdot \mid s_i'),$$
$$D_{3i} := \mathrm{Law}(x_{3i} \mid \mathcal{F}_{3i-1}) = \delta_{s_i''} q_i''(\cdot \mid s_i'').$$

Hence the full covariate sequence

$$x_1, x_2, \ldots, x_{3k}$$

is a predictable design sequence with respect to the tuple-level filtration $\{\mathcal{F}_t\}_{t=0}^{3k}$ in the sense of [2, Theorem 21]. Moreover, the responses satisfy

$$y_t \sim P^\star(\cdot \mid x_t) \qquad \text{for every } t \le 3k,$$

and realizability gives $P^\star \in \mathcal{M}$.

Apply Theorem 21 of [2] to the first $3k$ samples, with confidence level δ/k^2. This yields, for every fixed k,

$$\frac{1}{3k}\sum_{t=1}^{3k} \mathbb{E}_{x\sim D_t}\left[\|\hat{P}_k(\cdot \mid x) - P^\star(\cdot \mid x)\|_1^2\right] \lesssim \frac{\ln(|\mathcal{M}|k/\delta)}{k},$$

where D_t denotes the predictable design of the t-th tuple. By the definitions above,

$$\frac{1}{3k}\sum_{t=1}^{3k} D_t = \frac{1}{3}\big(\bar{\nu}_k^{\mathrm{pri}} + \widetilde{\nu}_k^{\mathrm{sec}} + \widetilde{\nu}_k^{\mathrm{ter}}\big) = \widetilde{\nu}_k^{\mathrm{all}},$$

as a random probability measure. Therefore the left-hand side is precisely

$$\mathbb{E}_{(s,a)\sim\widetilde{\nu}_k^{\mathrm{all}}}\left[\|\hat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2\right].$$

Finally, a union bound over all $k \in [K]$ turns the fixed-k statement into the simultaneous event claimed in the lemma, with the extra $\ln k$ absorbed into ζ_k. □

Remark D.3 (Route choice: pathwise-to-averaged bridge)**.** Lemma D.2 is the formally correct tuple-level statement: the secondary and tertiary designs are the pathwise predictable kernels $\delta_{s'_i} q'_i(\cdot \mid s'_i)$ and $\delta_{s''_i} q''_i(\cdot \mid s''_i)$, not the averaged covariates ν'_i, ν''_i directly. For the remainder of this appendix, we keep the coverage coefficients and theorem statements in terms of the deterministic averaged designs $\bar{\nu}_k^{\mathrm{pri}}$, $\bar{\nu}_k^{\mathrm{sec}}$, and $\bar{\nu}_k^{\mathrm{ter}}$, rather than rewriting every later lemma in pathwise notation. We therefore choose the *pathwise-to-averaged bridge* route: whenever the tuple-level MLE event is used under $\bar{\nu}_k^{\mathrm{sec}}$ or $\bar{\nu}_k^{\mathrm{ter}}$, we apply the bridge lemma below to the relevant squared mismatch class. This ensures simultaneous control under the random pathwise designs $\widetilde{\nu}_k^{\mathrm{sec}}, \widetilde{\nu}_k^{\mathrm{ter}}$ and their conditional means $\bar{\nu}_k^{\mathrm{sec}}, \bar{\nu}_k^{\mathrm{ter}}$. The covariance comparison lemma below already avoids this issue by working at the episode-block level. The only bridge actually needed later in this appendix is for the squared MLE-mismatch class, which we state next.

Lemma D.4 (Pathwise-to-averaged bridge for the MLE mismatch class)**.** *For each candidate model $P \in \mathcal{M}$, define*

$$h_P(s,a) := \|P(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2.$$

Then there is a universal constant $C < \infty$ such that, with probability at least $1-\delta$, simultaneously for all $k \in [K]$ and all $P \in \mathcal{M}$,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h_P] \le 2\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{sec}}}[h_P] + C\frac{\ln(|\mathcal{M}|k/\delta)}{k},$$

and

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{ter}}}[h_P] \le 2\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{ter}}}[h_P] + C\frac{\ln(|\mathcal{M}|k/\delta)}{k}.$$

Consequently, on the intersection of this event with Lemma D.2, we have simultaneously for all $k \in [K]$,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[f_k^2] \lesssim \zeta_k, \qquad \mathbb{E}_{\bar{\nu}_k^{\mathrm{ter}}}[f_k^2] \lesssim \zeta_k.$$

In particular, this lemma is used later only to transfer the internal MLE-mismatch terms from the pathwise designs to the averaged designs. The generic test function g in the one-step-back and optimism lemmas is handled directly under the deterministic averaged measures $\bar{\nu}_k^{\mathrm{sec}}, \bar{\nu}_k^{\mathrm{ter}}$ through the coverage coefficients, so no stronger uniform bridge over an adaptive class is invoked.

Proof. Fix a candidate model $P \in \mathcal{M}$ and write $h := h_P$. Since $\|P(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1 \le 2$, we have

$$0 \le h(s,a) \le 4 \qquad \forall (s,a).$$

We first treat the secondary stream. For each episode i, define

$$X_i^{\mathrm{sec}}(h) := h(s'_i, a'_i) - \mathbb{E}_{\bar{\nu}_i^{\mathrm{sec}}}[h] = h(s'_i, a'_i) - \mathbb{E}_{\nu'_i}[h].$$

Let $\mathcal{G}_i$ denote the sigma-field generated by all tuples up to the end of episode i. Because (s'_i, a'_i) has conditional mean design ν'_i given $\mathcal{G}_{i-1}$, we have

$$\mathbb{E}[X_i^{\mathrm{sec}}(h) \mid \mathcal{G}_{i-1}] = 0.$$

Moreover,

$$|X_i^{\mathrm{sec}}(h)| \le 4,$$

and, using $h^2 \le 4h$ on $[0,4]$,

$$\mathbb{E}\Big[\big(X_i^{\mathrm{sec}}(h)\big)^2 \mid \mathcal{G}_{i-1}\Big] \le \mathbb{E}\big[h(s'_i,a'_i)^2 \mid \mathcal{G}_{i-1}\big] \le 4\mathbb{E}[h(s'_i,a'_i) \mid \mathcal{G}_{i-1}] = 4\mathbb{E}_{\nu'_i}[h].$$

Hence the martingale $S_k^{\mathrm{sec}}(h) := \sum_{i=1}^k X_i^{\mathrm{sec}}(h)$ has predictable quadratic variation bounded by

$$V_k^{\mathrm{sec}}(h) \le 4\sum_{i=1}^k \mathbb{E}_{\nu'_i}[h] = 4k\,\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h].$$

Apply Freedman's inequality with confidence level $\delta_{k,P}^{\mathrm{sec}} := \delta/(4|\mathcal{M}|k^2)$, and write $\ell_k^{\mathrm{sec}} := \ln(1/\delta_{k,P}^{\mathrm{sec}})$. Then, with probability at least $1-\delta_{k,P}^{\mathrm{sec}}$,

$$-S_k^{\mathrm{sec}}(h) \le \sqrt{2V_k^{\mathrm{sec}}(h)\ell_k^{\mathrm{sec}}} + \frac{4}{3}\ell_k^{\mathrm{sec}}.$$

Dividing by k and recalling that

$$\frac{-S_k^{\mathrm{sec}}(h)}{k} = \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h] - \mathbb{E}_{\widetilde{\nu}_k^{\mathrm{sec}}}[h],$$

we obtain

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h] - \mathbb{E}_{\widetilde{\nu}_k^{\mathrm{sec}}}[h] \le \sqrt{\frac{8\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h]\ell_k^{\mathrm{sec}}}{k}} + \frac{4}{3}\frac{\ell_k^{\mathrm{sec}}}{k}.$$

Using $2\sqrt{ab} \le \frac{a}{2} + 2b$ with $a = \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h]$ and $b = 8\ell_k^{\mathrm{sec}}/k$ gives

$$\sqrt{\frac{8\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h]\ell_k^{\mathrm{sec}}}{k}} \le \frac{1}{2}\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h] + 16\frac{\ell_k^{\mathrm{sec}}}{k}.$$

Therefore

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[h] \le 2\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{sec}}}[h] + \frac{104}{3}\frac{\ell_k^{\mathrm{sec}}}{k}.$$

Since $\ell_k^{\mathrm{sec}} = \ln(4|\mathcal{M}|k^2/\delta) \lesssim \ln(|\mathcal{M}|k/\delta)$, this yields the desired secondary bound after adjusting the universal constant C.

The tertiary bound is identical. Indeed, defining

$$X_i^{\mathrm{ter}}(h) := h(s_i'', a_i'') - \mathbb{E}_{\nu_i''}[h],$$

we again have

$$\mathbb{E}[X_i^{\mathrm{ter}}(h) \mid \mathcal{G}_{i-1}] = 0, \qquad |X_i^{\mathrm{ter}}(h)| \le 4,$$

and

$$\mathbb{E}\Big[\big(X_i^{\mathrm{ter}}(h)\big)^2 \mid \mathcal{G}_{i-1}\Big] \le 4\mathbb{E}_{\nu_i''}[h].$$

Running the same Freedman argument and taking a union bound over the secondary and tertiary streams, over all $P \in \mathcal{M}$, and over all $k \in [K]$, gives the first two displays in the lemma.

Finally, set $P = \hat{P}_k$. Since $\hat{P}_k \in \mathcal{M}$, we may plug $h_{\hat{P}_k} = f_k^2$ into the simultaneous event. By Lemma D.2,

$$\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{all}}}[f_k^2] \lesssim \zeta_k.$$

Because

$$\widetilde{\nu}_k^{\mathrm{all}} = \frac{1}{3}\big(\bar{\nu}_k^{\mathrm{pri}} + \widetilde{\nu}_k^{\mathrm{sec}} + \widetilde{\nu}_k^{\mathrm{ter}}\big),$$

and all three component expectations are nonnegative,

$$\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{sec}}}[f_k^2] \le 3\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{all}}}[f_k^2] \lesssim \zeta_k,$$

and likewise

$$\mathbb{E}_{\widetilde{\nu}_k^{\mathrm{ter}}}[f_k^2] \lesssim \zeta_k.$$

Substituting these into the secondary and tertiary bridge inequalities and absorbing the extra $\ln(|\mathcal{M}|k/\delta)/k$ term into ζ_k proves the final display. □

Lemma D.5 (Primary-covariance deployed-to-probe transfer)**.** *Fix an episode k and define*

$$\Sigma_{k,\star}^{\mathrm{pri}} := k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[\phi^\star(s,a)\phi^\star(s,a)^\top] + \lambda_k I.$$

Then for every policy π and every measurable $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $\|g\|_\infty \le B$,

$$\begin{aligned}\Big|\mathbb{E}_{d_{P^\star}^\pi}[g]\Big| \lesssim& \sqrt{\gamma}\,\mathbb{E}_{d_{P^\star}^\pi}\|\phi^\star(\tilde{s},\tilde{a})\|_{(\Sigma_{k,\star}^{\mathrm{pri}})^{-1}}\sqrt{k\mathfrak{C}_k^{\mathrm{next}}(\pi)\,\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[g^2] + \lambda_k dB^2} \\ &+ \sqrt{(1-\gamma)\mathfrak{C}_k^{\mathrm{init}}(\pi)\,\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[g^2]}.\end{aligned}$$

Proof. Define

$$m_g(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[g(s,a)], \quad T_\star^\pi g(\tilde{s},\tilde{a}) := \mathbb{E}_{s\sim P^\star(\cdot|\tilde{s},\tilde{a})}[m_g(s)].$$

Then $|m_g(s)| \le B$, and the discounted-occupancy recursion gives

$$\mathbb{E}_{d^\pi_{P^\star}}[g] = \gamma \mathbb{E}_{d^\pi_{P^\star}}[T_\star^\pi g] + (1-\gamma)\mathbb{E}_{s\sim d_1}[m_g(s)]. \tag{4}$$

The initialization term is bounded by Cauchy-Schwarz and the definition of $\mathfrak{C}_k^{\mathrm{init}}(\pi)$:

$$(1-\gamma)\,|\mathbb{E}_{s\sim d_1}[m_g(s)]| \le \sqrt{(1-\gamma)\mathfrak{C}_k^{\mathrm{init}}(\pi)\,\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[g^2]}.$$

Next write $v_g^\star := \int \mu^\star(s) m_g(s)\,ds$. Since $T_\star^\pi g(\tilde{s},\tilde{a}) = \phi^\star(\tilde{s},\tilde{a})^\top v_g^\star$, Cauchy-Schwarz gives

$$\mathbb{E}_{d^\pi_{P^\star}}[T_\star^\pi g] \le \mathbb{E}_{d^\pi_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}} \cdot \|v_g^\star\|_{\Sigma^{\mathrm{pri}}_{k,\star}}.$$

Moreover, $\|v_g^\star\|_2 \le \sqrt{d}\,B$, and

$$\|v_g^\star\|^2_{\Sigma^{\mathrm{pri}}_{k,\star}} = k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\left[(T_\star^\pi g)^2\right] + \lambda_k \|v_g^\star\|_2^2.$$

By Jensen and the definition of $\mathfrak{C}_k^{\mathrm{next}}(\pi)$,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\left[(T_\star^\pi g)^2\right] \le \mathfrak{C}_k^{\mathrm{next}}(\pi)\,\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[g^2].$$

Substituting these bounds into (4) gives the claim. □

Lemma D.6 (Merged covariance comparison)**.** *For each episode k, define*

$$\Sigma_k^{\mathrm{mix}} := k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[\hat{\phi}_k\hat{\phi}_k^\top] + k\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[\hat{\phi}_k\hat{\phi}_k^\top] + \lambda_k I.$$

There exist universal constants $0 < c_1 \le c_2 < \infty$ such that, with probability at least $1-\delta$, simultaneously for all $k \in [K]$ and all (s,a),

$$c_1\|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}} \le \|\hat{\phi}_k(s,a)\|_{(\hat{\Sigma}_k^{\mathrm{mix}})^{-1}} \le c_2\|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}}.$$

Proof. Fix an episode k and first work with an arbitrary candidate feature map $\phi \in \Gamma$. Define the episode-block matrix

$$H_i(\phi) := \phi(s_i,a_i)\phi(s_i,a_i)^\top + \phi(s_i',a_i')\phi(s_i',a_i')^\top, \qquad i \in [k].$$

Let $\mathcal{G}_i$ be the sigma-field generated by all tuples collected up to the end of episode i, with $\mathcal{G}_0$ the history before episode 1. Define the predictable episode-level conditional mean $\bar{H}_i(\phi) := \mathbb{E}[H_i(\phi) \mid \mathcal{G}_{i-1}]$. By construction, conditioned on $\mathcal{G}_{i-1}$ the primary tuple is sampled from ν_i, while the secondary tuple has conditional mean design ν_i' after marginalizing over the primary transition and the secondary probe action. Therefore

$$\bar{H}_i(\phi) = \mathbb{E}_{\nu_i}[\phi\phi^\top] + \mathbb{E}_{\nu_i'}[\phi\phi^\top], \qquad \sum_{i=1}^k \bar{H}_i(\phi) = k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[\phi\phi^\top] + k\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[\phi\phi^\top].$$

Now define

$$\Sigma_k^{\mathrm{mix}}(\phi) := \sum_{i=1}^k \bar{H}_i(\phi) + \lambda_k I, \qquad \hat{\Sigma}_k^{\mathrm{mix}}(\phi) := \sum_{i=1}^k H_i(\phi) + \lambda_k I.$$

Under the bounded-feature assumption used throughout the paper, $H_i(\phi)$ is positive semidefinite with operator norm at most 2. Hence the matrix martingale

$$M_t(\phi) := \sum_{u=1}^t \big(H_u(\phi) - \bar{H}_u(\phi)\big)$$

satisfies the standard bounded-increment conditions for episode-block matrix concentration with respect to $\{\mathcal{G}_i\}_{i=0}^k$. In particular, the same self-normalized matrix martingale argument used in the

proof of Lemma A.1 applies to these block increments. After choosing $\lambda_k \gtrsim d \ln(|\mathcal{M}|k/\delta)$, there is an event of probability at least $1-\delta$ on which

$$\left\| \left(\Sigma_k^{\mathrm{mix}}(\phi)\right)^{-1/2} M_k(\phi) \left(\Sigma_k^{\mathrm{mix}}(\phi)\right)^{-1/2} \right\|_{\mathrm{op}} \le \frac{1}{2},$$

simultaneously for all $k \in [K]$ and all $\phi \in \Gamma$. Equivalently,

$$-\frac{1}{2}\Sigma_k^{\mathrm{mix}}(\phi) \preceq M_k(\phi) \preceq \frac{1}{2}\Sigma_k^{\mathrm{mix}}(\phi).$$

Since $\hat{\Sigma}_k^{\mathrm{mix}}(\phi) = \Sigma_k^{\mathrm{mix}}(\phi) + M_k(\phi)$, this immediately yields

$$\frac{1}{2}\Sigma_k^{\mathrm{mix}}(\phi) \preceq \hat{\Sigma}_k^{\mathrm{mix}}(\phi) \preceq \frac{3}{2}\Sigma_k^{\mathrm{mix}}(\phi).$$

Specializing to $\phi = \hat{\phi}_k$ and inverting the PSD inequalities gives

$$\frac{1}{2}\Sigma_k^{\mathrm{mix}} \preceq \hat{\Sigma}_k^{\mathrm{mix}} \preceq \frac{3}{2}\Sigma_k^{\mathrm{mix}},$$
$$\frac{2}{3}(\Sigma_k^{\mathrm{mix}})^{-1} \preceq (\hat{\Sigma}_k^{\mathrm{mix}})^{-1} \preceq 2(\Sigma_k^{\mathrm{mix}})^{-1}.$$

Applying the resulting quadratic-form comparison to $\hat{\phi}_k(s,a)$ proves

$$\sqrt{\frac{2}{3}} \|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}} \le \|\hat{\phi}_k(s,a)\|_{(\hat{\Sigma}_k^{\mathrm{mix}})^{-1}} \le \sqrt{2}\, \|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}}.$$

Thus the claim holds with any universal constants $c_1 \le \sqrt{2/3}$ and $c_2 \ge \sqrt{2}$. □

Proposition D.7 (Concrete tertiary coverage bound)**.** *Under Algorithm 1, for every episode k and every policy π,*

$$\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi) \le \Gamma_k, \qquad \mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi) \le \Gamma_k, \qquad \mathfrak{C}_{k}^{\mathrm{tri,next}}(\pi) \le \Gamma_k.$$

Proof. Because the schedule is nonincreasing,

$$\epsilon_i \ge \epsilon_k \qquad \forall i \le k.$$

Hence every probing kernel among the first k episodes satisfies

$$q_i(a \mid s) \ge \frac{\epsilon_k}{|\mathcal{A}|}, \qquad q_i'(a \mid s) \ge \frac{\epsilon_k}{|\mathcal{A}|}, \qquad q_i''(a \mid s) \ge \frac{\epsilon_k}{|\mathcal{A}|}.$$

For the first tertiary coefficient, let

$$\mu_{k,1}^{\pi}(s,a) := \mathbb{E}_{\substack{(\tilde{s},\tilde{a}) \sim \bar{\nu}_k^{\mathrm{pri}} \\ s \sim P^\star(\cdot \mid \tilde{s},\tilde{a})}} \left[\pi(a \mid s) \mathbf{1}\{s \in ds\}\right]$$

denote the pushforward state-action measure appearing in the definition of $\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi)$. Since

$$\bar{\nu}_k^{\mathrm{sec}}(s,a) = \frac{1}{k}\sum_{i=1}^{k} \rho_i'(s) q_i'(a \mid s) \ge \frac{\epsilon_k}{|\mathcal{A}|} \cdot \frac{1}{k}\sum_{i=1}^{k} \rho_i'(s),$$

we have, pointwise,

$$\mu_{k,1}^{\pi}(s,a) \le \frac{|\mathcal{A}|}{\epsilon_k} \bar{\nu}_k^{\mathrm{sec}}(s,a) = \Gamma_k \bar{\nu}_k^{\mathrm{sec}}(s,a).$$

Therefore for every nonnegative measurable g^2,

$$\mathbb{E}_{\mu_{k,1}^{\pi}}[g^2] \le \Gamma_k \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[g^2],$$

which proves

$$\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi) \le \Gamma_k.$$

The proof for $\mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi)$ is identical. If

$$\mu_{k,2}^{\pi}(s,a) := \mathbb{E}_{\substack{(\tilde{s},\tilde{a})\sim\bar{\nu}_k^{\mathrm{sec}} \\ s\sim P^\star(\cdot|\tilde{s},\tilde{a})}}\big[\pi(a \mid s)\mathbf{1}\{s \in ds\}\big],$$

then

$$\bar{\nu}_k^{\mathrm{ter}}(s,a) \geq \frac{\epsilon_k}{|\mathcal{A}|} \cdot \frac{1}{k}\sum_{i=1}^{k} \rho_i''(s),$$

while by construction

$$\mu_{k,2}^{\pi}(s,a) = \frac{1}{k}\sum_{i=1}^{k} \rho_i''(s)\pi(a \mid s).$$

Using $\pi(a \mid s) \leq 1$ and the lower bound on $\bar{\nu}_k^{\mathrm{ter}}$, we obtain the pointwise domination

$$\mu_{k,2}^{\pi}(s,a) \leq \Gamma_k \bar{\nu}_k^{\mathrm{ter}}(s,a).$$

Therefore, for every nonnegative measurable g^2,

$$\mathbb{E}_{\mu_{k,2}^{\pi}}[g^2] \leq \Gamma_k \mathbb{E}_{\bar{\nu}_k^{\mathrm{ter}}}[g^2],$$

which proves

$$\mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi) \leq \Gamma_k.$$

Taking the maximum of the two coefficients completes the proof. □

Lemma D.8 (Tertiary-aware mixed one-step-back inequality)**.** *Fix an episode k and define $f_k(s,a) := \|\hat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$ and*

$$\Sigma_k^{\mathrm{mix}} := k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[\hat{\phi}_k\hat{\phi}_k^\top] + k\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[\hat{\phi}_k\hat{\phi}_k^\top] + \lambda_k I.$$

Also define

$$M_k^{\mathrm{mix}}(\pi) := \mathbb{E}_{d_{\hat{P}_k}^{\pi}}\Big[\|\hat{\phi}_k(\tilde{s},\tilde{a})\|_{(\Sigma_k^{\mathrm{mix}})^{-1}}\Big], \qquad G_{k,23}(g) := \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[g^2] + \mathbb{E}_{\bar{\nu}_k^{\mathrm{ter}}}[g^2].$$

Then for every policy π and every measurable $g : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ with $\|g\|_\infty \leq B$,

$$\begin{aligned}\left|\mathbb{E}_{d_{\hat{P}_k}^{\pi}}[g]\right| \lesssim & \sqrt{\gamma}\, M_k^{\mathrm{mix}}(\pi)\sqrt{k\mathfrak{C}_k^{\mathrm{tri,next}}(\pi)\, G_{k,23}(g) + \lambda_k dB^2 + kB^2\zeta_k} \\ & + \sqrt{(1-\gamma)\mathfrak{C}_k^{\mathrm{init}}(\pi)\, \mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[g^2]},\end{aligned}$$

on the events of Lemmas D.2 and D.4.

Proof. Define

$$\begin{aligned} m_g(s) &:= \mathbb{E}_{a\sim\pi(\cdot|s)}[g(s,a)], \\ T_{\hat{P}_k}^{\pi} g(\tilde{s},\tilde{a}) &:= \mathbb{E}_{s\sim\hat{P}_k(\cdot|\tilde{s},\tilde{a})}[m_g(s)], \\ T_{\star}^{\pi} g(\tilde{s},\tilde{a}) &:= \mathbb{E}_{s\sim P^\star(\cdot|\tilde{s},\tilde{a})}[m_g(s)]. \end{aligned}$$

Since $\|g\|_\infty \leq B$, we also have $\|m_g\|_\infty \leq B$. The discounted occupancy recursion under $\hat{P}_k$ gives

$$\mathbb{E}_{d_{\hat{P}_k}^{\pi}}[g] = \gamma\mathbb{E}_{d_{\hat{P}_k}^{\pi}}[T_{\hat{P}_k}^{\pi} g] + (1-\gamma)\mathbb{E}_{s\sim d_1}[m_g(s)].$$

We treat the two terms on the right separately.

For the initialization term, write

$$\mathbb{E}_{s\sim d_1}[m_g(s)] = \sum_{s,a} d_1(s)\pi(a \mid s)g(s,a).$$

By the definition of $\mathfrak{C}_k^{\mathrm{init}}(\pi)$,

$$(1-\gamma)d_1(s)\pi(a \mid s) \leq \mathfrak{C}_k^{\mathrm{init}}(\pi)\, \bar{\nu}_k^{\mathrm{pri}}(s,a) \qquad \forall (s,a)$$

on the support of $\bar{\nu}_k^{\mathrm{pri}}$. Therefore Cauchy-Schwarz gives

$$(1-\gamma)\,|\mathbb{E}_{s\sim d_1}[m_g(s)]| \le \sqrt{(1-\gamma)\mathfrak{C}_k^{\mathrm{init}}(\pi)\,\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[g^2]}.$$

For the propagated term, define $v_g^{\hat{P}} := \int \hat{\mu}_k(s) m_g(s)\,ds$. Because $T^\pi_{\hat{P}_k} g(\tilde{s},\tilde{a}) = \hat{\phi}_k(\tilde{s},\tilde{a})^\top v_g^{\hat{P}}$, Cauchy-Schwarz under $(\Sigma_k^{\mathrm{mix}})^{-1}$ yields

$$\mathbb{E}_{d^\pi_{\hat{P}_k}}[T^\pi_{\hat{P}_k} g] \le M_k^{\mathrm{mix}}(\pi)\,\|v_g^{\hat{P}}\|_{\Sigma_k^{\mathrm{mix}}}.$$

Moreover, $\|v_g^{\hat{P}}\|_2 \le \sqrt{d}\,B$, so

$$\|v_g^{\hat{P}}\|^2_{\Sigma_k^{\mathrm{mix}}} = k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\big[(T^\pi_{\hat{P}_k} g)^2\big] + k\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}\big[(T^\pi_{\hat{P}_k} g)^2\big] + \lambda_k \|v_g^{\hat{P}}\|_2^2.$$

We now control the two quadratic terms. By definition of f_k,

$$\big|T^\pi_{\hat{P}_k} g(\tilde{s},\tilde{a}) - T^\pi_\star g(\tilde{s},\tilde{a})\big| \le B f_k(\tilde{s},\tilde{a}).$$

Hence, using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\big[(T^\pi_{\hat{P}_k} g)^2\big] \lesssim \mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\big[(T^\pi_\star g)^2\big] + B^2\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[f_k^2],$$
$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}\big[(T^\pi_{\hat{P}_k} g)^2\big] \lesssim \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}\big[(T^\pi_\star g)^2\big] + B^2\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[f_k^2].$$

On the events of Lemmas D.2 and D.4,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[f_k^2] \lesssim \zeta_k, \qquad \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[f_k^2] \lesssim \zeta_k,$$

because the primary term is one component of the pathwise average $\widetilde{\nu}_k^{\mathrm{all}}$, while the secondary term is provided directly by Lemma D.4. This is the only place where a pathwise-to-averaged transfer is used: the generic g-dependence remains entirely inside $G_{k,23}(g) = \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[g^2] + \mathbb{E}_{\bar{\nu}_k^{\mathrm{ter}}}[g^2]$ and is handled through the deterministic coverage coefficients.

Next, by Jensen's inequality,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\big[(T^\pi_\star g)^2\big] \le \mathbb{E}_{\mu^\pi_{k,1}}[g^2] \le \mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi)\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[g^2],$$
$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}\big[(T^\pi_\star g)^2\big] \le \mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi)\mathbb{E}_{\bar{\nu}_k^{\mathrm{ter}}}[g^2].$$

Combining the last four displays and using

$$\mathfrak{C}_k^{\mathrm{tri,next}}(\pi) = \max\Big\{\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi), \mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi)\Big\},$$

we obtain

$$\|v_g^{\hat{P}}\|^2_{\Sigma_k^{\mathrm{mix}}} \lesssim k\mathfrak{C}_k^{\mathrm{tri,next}}(\pi)\,G_{k,23}(g) + \lambda_k d B^2 + kB^2\zeta_k.$$

Substituting this and the initialization bound into the discounted-occupancy recursion proves the claimed inequality. □

Lemma D.9 (Direct optimism for the repaired merged bonus)**.** *Fix an episode k and a policy π. Define*

$$b_{k,\mathrm{mix,pop}}(s,a) := \alpha_k^{\mathrm{mix}}\|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}},$$
$$b_{k,\mathrm{mix,emp}}(s,a) := \alpha_k^{\mathrm{emp}}\|\hat{\phi}_k(s,a)\|_{(\hat{\Sigma}_k^{\mathrm{mix}})^{-1}}.$$

Assume the events of Lemmas D.2, D.4, and the covariance comparison event of Lemma D.6 all hold. If

$$\alpha_k^{\mathrm{mix}} \gtrsim \sqrt{\gamma\Big(k\mathfrak{C}_k^{\mathrm{tri,next}}(\pi)\zeta_k + \lambda_k d + k\zeta_k\Big)},$$

and

$$\alpha_k^{\mathrm{emp}} \ge \alpha_k^{\mathrm{mix}}/c_1,$$

then

$$V^\pi_{\hat{P}_k, r+b_{k,\mathrm{mix,emp}}} - V^\pi_{P^\star, r} \gtrsim -\sqrt{\frac{\mathfrak{C}_k^{\mathrm{init}}(\pi)\zeta_k}{1-\gamma}}.$$

Under the nonbinding-cap condition $\hat{b}_k = b_{k,\mathrm{mix,emp}}$, the same bound holds with $b_{k,\mathrm{mix,emp}}$ replaced by the actual algorithmic bonus $\hat{b}_k$.

Proof. Apply the first form of the simulation lemma (Lemma B.4) to the pair $(\hat{P}_k, r + b_{k,\mathrm{mix,pop}})$ and $(P^\star, r)$. This gives

$$(1-\gamma)\big(V^\pi_{\hat{P}_k, r+b_{k,\mathrm{mix,pop}}} - V^\pi_{P^\star,r}\big) \quad = \mathbb{E}_{d^\pi_{\hat{P}_k}}\Big[b_{k,\mathrm{mix,pop}}(s,a) + \gamma \mathbb{E}_{\hat{P}_k}[V^\pi_{P^\star,r}(s')] - \gamma \mathbb{E}_{P^\star}[V^\pi_{P^\star,r}(s')]\Big].$$

Because the total discounted reward is normalized to lie in $[0,1]$, we have

$$\|V^\pi_{P^\star,r}\|_\infty \le 1, \qquad \Big|\mathbb{E}_{\hat{P}_k}[V^\pi_{P^\star,r}(s')] - \mathbb{E}_{P^\star}[V^\pi_{P^\star,r}(s')]\Big| \le f_k(s,a).$$

Therefore

$$(1-\gamma)\big(V^\pi_{\hat{P}_k, r+b_{k,\mathrm{mix,pop}}} - V^\pi_{P^\star,r}\big) \quad \ge \mathbb{E}_{d^\pi_{\hat{P}_k}}\big[b_{k,\mathrm{mix,pop}}(s,a)\big] - \gamma \mathbb{E}_{d^\pi_{\hat{P}_k}}[f_k].$$

For brevity, define

$$\Omega_k(\pi) := k\mathfrak{C}_k^{\mathrm{tri,next}}(\pi)\zeta_k + \lambda_k d + k\zeta_k.$$

Now apply Lemma D.8 with $g = f_k$ and $B = 2$. On the events assumed in the lemma,

$$\mathbb{E}_{d^\pi_{\hat{P}_k}}[f_k] \lesssim \sqrt{\gamma}\, M_k^{\mathrm{mix}}(\pi)\sqrt{\Omega_k(\pi)} + \sqrt{(1-\gamma)\mathfrak{C}_k^{\mathrm{init}}(\pi)\zeta_k}.$$

By the choice of α_k^{mix}, $\alpha_k^{\mathrm{mix}} \gtrsim \sqrt{\gamma\Omega_k(\pi)}$, so the first term on the right is absorbed by $\mathbb{E}_{d^\pi_{\hat{P}_k}}[b_{k,\mathrm{mix,pop}}] = \alpha_k^{\mathrm{mix}} M_k^{\mathrm{mix}}(\pi)$. Consequently,

$$V^\pi_{\hat{P}_k, r+b_{k,\mathrm{mix,pop}}} - V^\pi_{P^\star,r} \gtrsim -\sqrt{\frac{\mathfrak{C}_k^{\mathrm{init}}(\pi)\zeta_k}{1-\gamma}}.$$

Next, Lemma D.6 gives

$$\|\hat{\phi}_k(s,a)\|_{(\hat{\Sigma}_k^{\mathrm{mix}})^{-1}} \ge c_1 \|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}},$$

and therefore $\alpha_k^{\mathrm{emp}} \ge \alpha_k^{\mathrm{mix}}/c_1$ implies the pointwise domination $b_{k,\mathrm{mix,emp}}(s,a) \ge b_{k,\mathrm{mix,pop}}(s,a)$. Monotonicity of the simulation lemma in the bonus term then yields

$$V^\pi_{\hat{P}_k, r+b_{k,\mathrm{mix,emp}}} - V^\pi_{P^\star,r} \gtrsim -\sqrt{\frac{\mathfrak{C}_k^{\mathrm{init}}(\pi)\zeta_k}{1-\gamma}}.$$

Finally, under the nonbinding-cap condition, $\hat{b}_k = b_{k,\mathrm{mix,emp}}$, so the same bound holds for the actual algorithmic bonus. □

D.4 Final Aggregation

We now combine the optimism step with the deployed-to-probe transfer and the elliptical-potential argument.

Lemma D.10 (Tertiary-aware primal-dual aggregation)**.** *Fix an episode k and define*

$$\begin{aligned}
\Delta_k &:= \big(V^{\pi^\star}_{P^\star,r}(s_1) - V^{\pi_k}_{P^\star,r}(s_1)\big) + Y_k\big(V^{\pi^\star}_{P^\star,g}(s_1) - V^{\pi_k}_{P^\star,g}(s_1)\big),\\
A_k &:= \mathbb{E}_{d^{\pi_k}_{P^\star}}\Big[\|\phi^\star(\tilde{s},\tilde{a})\|_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big],\\
f_k(s,a) &:= \|\hat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1.
\end{aligned}$$

Assume, for every episode k,

1. *the cap is nonbinding:*
$$\bar{b} \ge \max_{1\le j\le K} \frac{\alpha_j^{\mathrm{emp}}}{\sqrt{\lambda_j}};$$
2. *the events of Lemmas D.2 and D.4 hold;*

3. Lemma D.9 holds with $\pi = \pi^\star$ for both the reward objective r and the utility objective g.

Then, under Algorithm 1,

$$\begin{aligned}\Delta_k \lesssim & \frac{1+\xi}{1-\gamma} A_k \sqrt{\gamma d \Gamma_k (\alpha_k^{\mathrm{emp}})^2 + \gamma \lambda_k d \bar{b}^2} + \frac{(1+\xi)\bar{b}}{(1-\gamma)^2} A_k \sqrt{\gamma\big(k\Gamma_k\zeta_k + \lambda_k d\big)} \\ & + \frac{1+\xi}{1-\gamma} \alpha_k^{\mathrm{emp}} \sqrt{\frac{\Gamma_k d(1-\gamma)}{k}} + \frac{(1+\xi)\bar{b}}{(1-\gamma)^2} \sqrt{\Gamma_k \zeta_k (1-\gamma)} + (1+\xi)\sqrt{\frac{\Gamma_k \zeta_k}{1-\gamma}}.\end{aligned}$$

Consequently,

$$\begin{aligned}\sum_{k=1}^K \Delta_k \lesssim & \frac{1+\xi}{1-\gamma} \sum_{k=1}^K A_k \sqrt{\gamma d \Gamma_k (\alpha_k^{\mathrm{emp}})^2 + \gamma \lambda_k d \bar{b}^2} + \frac{(1+\xi)\bar{b}}{(1-\gamma)^2} \sum_{k=1}^K A_k \sqrt{\gamma\big(k\Gamma_k\zeta_k + \lambda_k d\big)} \\ & + \frac{1+\xi}{1-\gamma} \sum_{k=1}^K \alpha_k^{\mathrm{emp}} \sqrt{\frac{\Gamma_k d(1-\gamma)}{k}} \\ & + \frac{(1+\xi)\bar{b}}{(1-\gamma)^2} \sum_{k=1}^K \sqrt{\Gamma_k \zeta_k (1-\gamma)} + (1+\xi) \sum_{k=1}^K \sqrt{\frac{\Gamma_k \zeta_k}{1-\gamma}}.\end{aligned}$$

Proof. Fix an episode k. By the definition of Δ_k,

$$\Delta_k = \big(V_{P^\star,r}^{\pi^\star} - V_{P^\star,r}^{\pi_k}\big) + Y_k\big(V_{P^\star,g}^{\pi^\star} - V_{P^\star,g}^{\pi_k}\big).$$

Because $0 \le Y_k \le \xi$ and Lemma D.9 holds at $\pi^\star$ for both the reward and utility objectives,

$$V_{P^\star,r}^{\pi^\star} \le V_{\hat{P}_k, r+\hat{b}_k}^{\pi^\star} + \sqrt{\frac{\mathfrak{C}_k^{\mathrm{init}}(\pi^\star)\zeta_k}{1-\gamma}}, \qquad V_{P^\star,g}^{\pi^\star} \le V_{\hat{P}_k, g+\hat{b}_k}^{\pi^\star} + \sqrt{\frac{\mathfrak{C}_k^{\mathrm{init}}(\pi^\star)\zeta_k}{1-\gamma}}.$$

Since every primary probing distribution has the uniform floor $\epsilon_k/|\mathcal{A}|$ and $\rho_i(s) = d_{P^\star}^{\pi_{i-1}}(s) \ge (1-\gamma)d_1(s)$, we have $\bar{\nu}_k^{\mathrm{pri}}(s,a) \ge (1-\gamma)d_1(s)\epsilon_k/|\mathcal{A}|$, hence $\mathfrak{C}_k^{\mathrm{init}}(\pi^\star) \le \Gamma_k$. Therefore

$$\Delta_k \le \widetilde{\Delta}_k + (1+\xi)\sqrt{\frac{\Gamma_k\zeta_k}{1-\gamma}},$$

where

$$\widetilde{\Delta}_k := \big(V_{\hat{P}_k, r+\hat{b}_k}^{\pi^\star} - V_{P^\star,r}^{\pi_k}\big) + Y_k\big(V_{\hat{P}_k, g+\hat{b}_k}^{\pi^\star} - V_{P^\star,g}^{\pi_k}\big).$$

By the definition of π_k as the maximizer of the optimistic empirical Lagrangian,

$$V_{\hat{P}_k, r+\hat{b}_k}^{\pi^\star} + Y_k V_{\hat{P}_k, g+\hat{b}_k}^{\pi^\star} \le V_{\hat{P}_k, r+\hat{b}_k}^{\pi_k} + Y_k V_{\hat{P}_k, g+\hat{b}_k}^{\pi_k}.$$

Hence

$$\widetilde{\Delta}_k \le \big(V_{\hat{P}_k, r+\hat{b}_k}^{\pi_k} - V_{P^\star,r}^{\pi_k}\big) + Y_k\big(V_{\hat{P}_k, g+\hat{b}_k}^{\pi_k} - V_{P^\star,g}^{\pi_k}\big).$$

Define

$$W_{k,r} := V_{\hat{P}_k, r+\hat{b}_k}^{\pi_k}, \qquad W_{k,g} := V_{\hat{P}_k, g+\hat{b}_k}^{\pi_k},$$
$$\Delta_{\hat{P}_k, P^\star} W(s,a) := \mathbb{E}_{\hat{P}_k}[W(s')] - \mathbb{E}_{P^\star}[W(s')].$$

Apply the second form of the simulation lemma (Lemma B.4) to each bracket. For the reward term,

$$V_{\hat{P}_k, r+\hat{b}_k}^{\pi_k} - V_{P^\star,r}^{\pi_k} = \frac{1}{1-\gamma} \mathbb{E}_{d_{P^\star}^{\pi_k}}\Big[\hat{b}_k(s,a) + \gamma\, \Delta_{\hat{P}_k, P^\star} W_{k,r}(s,a)\Big].$$

The analogous identity holds for the utility term. Since $\|\hat{b}_k\|_\infty \le \bar{b}$, the normalized reward and utility assumptions imply, up to universal constants,

$$\|V_{\hat{P}_k, r+\hat{b}_k}^{\pi_k}\|_\infty \lesssim \frac{\bar{b}}{1-\gamma}, \qquad \|V_{\hat{P}_k, g+\hat{b}_k}^{\pi_k}\|_\infty \lesssim \frac{\bar{b}}{1-\gamma}.$$

Thus

$$\widetilde{\Delta}_k \lesssim \frac{1+\xi}{1-\gamma}\mathbb{E}_{d_{P^\star}^{\pi_k}}[\hat{b}_k] + \frac{(1+\xi)\bar{b}}{(1-\gamma)^2}\mathbb{E}_{d_{P^\star}^{\pi_k}}[f_k].$$

Combining this with the previous reduction gives

$$\Delta_k \lesssim \frac{1+\xi}{1-\gamma}\mathbb{E}_{d_{P^\star}^{\pi_k}}[\hat{b}_k] + \frac{(1+\xi)\bar{b}}{(1-\gamma)^2}\mathbb{E}_{d_{P^\star}^{\pi_k}}[f_k] + (1+\xi)\sqrt{\frac{\Gamma_k\zeta_k}{1-\gamma}}.$$

We now bound the two expectations separately.

For the bonus term, Lemma D.6 implies

$$\hat{b}_k(s,a) \lesssim \alpha_k^{\mathrm{emp}}\|\hat{\phi}_k(s,a)\|_{(\Sigma_k^{\mathrm{mix}})^{-1}}.$$

Moreover,

$$k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}\left[\|\hat{\phi}_k\|^2_{(\Sigma_k^{\mathrm{mix}})^{-1}}\right] \le d,$$

and the same bound holds with $\bar{\nu}_k^{\mathrm{pri}}$ replaced by $\bar{\nu}_k^{\mathrm{sec}}$, because each term is upper-bounded by

$$\mathrm{Tr}\Big(\big(k\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[\hat{\phi}_k\hat{\phi}_k^\top] + k\mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[\hat{\phi}_k\hat{\phi}_k^\top]\big)(\Sigma_k^{\mathrm{mix}})^{-1}\Big) \le d.$$

Applying Lemma D.5 with $g = \hat{b}_k$ and $B = \bar{b}$, and using $\mathfrak{C}_k^{\mathrm{next}}(\pi_k) \le \Gamma_k$ and $\mathfrak{C}_k^{\mathrm{init}}(\pi_k) \le \Gamma_k$, gives

$$\mathbb{E}_{d_{P^\star}^{\pi_k}}[\hat{b}_k] \lesssim A_k\sqrt{\gamma d\Gamma_k(\alpha_k^{\mathrm{emp}})^2 + \gamma\lambda_k d\bar{b}^2} + \alpha_k^{\mathrm{emp}}\sqrt{\frac{\Gamma_k d(1-\gamma)}{k}}.$$

For the model-error term, apply Lemma D.5 with $g = f_k$ and $B = 2$. By Lemmas D.2 and D.4,

$$\mathbb{E}_{\bar{\nu}_k^{\mathrm{pri}}}[f_k^2] \lesssim \zeta_k, \qquad \mathbb{E}_{\bar{\nu}_k^{\mathrm{sec}}}[f_k^2] \lesssim \zeta_k,$$

and again $\mathfrak{C}_k^{\mathrm{next}}(\pi_k) \le \Gamma_k$ and $\mathfrak{C}_k^{\mathrm{init}}(\pi_k) \le \Gamma_k$. Hence

$$\mathbb{E}_{d_{P^\star}^{\pi_k}}[f_k] \lesssim A_k\sqrt{\gamma\big(k\Gamma_k\zeta_k + \lambda_k d\big)} + \sqrt{\Gamma_k\zeta_k(1-\gamma)}.$$

Substituting the last two displays into the bound on Δ_k yields the episode-wise inequality claimed in the lemma. Summing over $k = 1, \dots, K$ proves the cumulative bound. □

Lemma D.11 (Tertiary-aware deployed-to-probe potential reduction)**.** *Under Algorithm 1,*

$$\Psi_K := \sum_{k=1}^K A_k \lesssim \frac{1}{\sqrt{\lambda_1}} + \sqrt{d\ln\left(1 + \frac{K}{d\lambda_1}\right)\sum_{k=2}^K \Lambda_k}.$$

In particular, since $\Lambda_k \le \Gamma_k$,

$$\Psi_K \lesssim \frac{1}{\sqrt{\lambda_1}} + \sqrt{d\ln\left(1 + \frac{K}{d\lambda_1}\right)\sum_{k=2}^K \Gamma_k}.$$

Proof. The first episode is trivial:

$$A_1 \le \frac{1}{\sqrt{\lambda_1}},$$

because $(\Sigma_{1,\star}^{\mathrm{pri}})^{-1} \preceq \lambda_1^{-1}I$ and $\|\phi^\star(s,a)\|_2 \le 1$.

Now fix $k \le K-1$. The relation

$$q_{k+1}(a \mid s) = (1-\epsilon_{k+1})\pi_k(a \mid s) + \epsilon_{k+1}U(a)$$

implies the pointwise ratio bound

$$\frac{\pi_k(a \mid s)}{q_{k+1}(a \mid s)} \le \Lambda_{k+1}.$$

Therefore

$$A_k \le \sqrt{\Lambda_{k+1}}\sqrt{\mathbb{E}_{(s,a)\sim\nu_{k+1}}\Big[\|\phi^\star(s,a)\|^2_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big]}.$$

Indeed,

$$\begin{aligned}
A_k &= \mathbb{E}_{(s,a)\sim d^{\pi_k}_{P^\star}}\Big[\|\phi^\star(s,a)\|_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big] \\
&\le \sqrt{\mathbb{E}_{d^{\pi_k}_{P^\star}}\Big[\frac{d^{\pi_k}_{P^\star}(s,a)}{\nu_{k+1}(s,a)}\Big]}\sqrt{\mathbb{E}_{(s,a)\sim\nu_{k+1}}\Big[\|\phi^\star(s,a)\|^2_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big]} \\
&\le \sqrt{\Lambda_{k+1}}\sqrt{\mathbb{E}_{(s,a)\sim\nu_{k+1}}\Big[\|\phi^\star(s,a)\|^2_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big]}.
\end{aligned}$$

For each $k \le K-1$, define

$$G_{k+1} := \mathbb{E}_{(s,a)\sim\nu_{k+1}}\big[\phi^\star(s,a)\phi^\star(s,a)^\top\big].$$

Then

$$\Sigma^{\mathrm{pri}}_{k,\star} = \lambda_k I + \sum_{i=1}^{k} G_i.$$

Since $\lambda_{k+1} \ge \lambda_k$, we may apply Lemma B.2 to the PSD increments $G_2, \dots, G_K$ and obtain

$$\sum_{k=1}^{K-1} \mathrm{Tr}\big(G_{k+1}(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}\big) \lesssim \ln\det(\Sigma^{\mathrm{pri}}_{K,\star}) - \ln\det(\lambda_1 I).$$

Because

$$\mathbb{E}_{(s,a)\sim\nu_{k+1}}\Big[\|\phi^\star(s,a)\|^2_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big] = \mathrm{Tr}\big(G_{k+1}(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}\big),$$

we deduce

$$\sum_{k=1}^{K-1} \mathbb{E}_{(s,a)\sim\nu_{k+1}}\Big[\|\phi^\star(s,a)\|^2_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big] \lesssim \ln\det(\Sigma^{\mathrm{pri}}_{K,\star}) - \ln\det(\lambda_1 I).$$

Lemma B.3 and the bound $\|\phi^\star(s,a)\|_2 \le 1$ imply

$$\ln\det(\Sigma^{\mathrm{pri}}_{K,\star}) - \ln\det(\lambda_1 I) \lesssim d\ln\left(1 + \frac{K}{d\lambda_1}\right).$$

Finally, Cauchy-Schwarz gives

$$\sum_{k=1}^{K-1} A_k \le \sqrt{\sum_{k=2}^{K}\Lambda_k}\sqrt{\sum_{k=1}^{K-1}\mathbb{E}_{(s,a)\sim\nu_{k+1}}\Big[\|\phi^\star(s,a)\|^2_{(\Sigma^{\mathrm{pri}}_{k,\star})^{-1}}\Big]} \lesssim \sqrt{d\ln\left(1+\frac{K}{d\lambda_1}\right)\sum_{k=2}^{K}\Lambda_k}.$$

Adding the bound on A_1 proves the result. □

Lemma D.12 (Summable optimistic overestimation of the utility)**.** *Fix an episode k and define*

$$\hat{U}_k := V^{\pi_k}_{\hat{P}_k, g+\hat{b}_k}(s_1), \qquad U_k := V^{\pi_k}_{P^\star, g}(s_1).$$

Assume Algorithm 1, the nonbinding-cap condition, and the events of Lemmas D.2, D.4, and D.6. Then, episode-wise,

$$\begin{aligned}
\big[\hat{U}_k - U_k\big]_+ \lesssim\ & \frac{1}{1-\gamma}\,A_k\alpha^{\mathrm{emp}}_k\sqrt{\gamma d\Gamma_k} + \frac{\bar{b}}{1-\gamma}\,A_k\sqrt{\gamma\lambda_k d} \\
&+ \frac{\bar{b}}{(1-\gamma)^2}\,A_k\sqrt{\gamma\big(k\Gamma_k\zeta_k + \lambda_k d\big)} \\
&+ \frac{1}{1-\gamma}\alpha^{\mathrm{emp}}_k\sqrt{\frac{\Gamma_k d(1-\gamma)}{k}} + \frac{\bar{b}}{(1-\gamma)^2}\sqrt{\Gamma_k\zeta_k(1-\gamma)}.
\end{aligned}$$

Consequently, under the tuning of Proposition D.13,

$$\sum_{k=1}^{K}\big[\hat{U}_k - U_k\big]_+ \le \tilde{\mathcal{O}}\Big(\frac{\bar{b}\,d^2}{(1-\gamma)^2}\Big)\left[\sqrt{\Gamma_K\sum_{k=2}^{K}\Lambda_k} + \sum_{k=1}^{K}\frac{\Gamma_k}{\sqrt{k}}\right].$$

Proof. Apply the second form of the simulation lemma (Lemma B.4) to the utility objective:

$$\hat{U}_k - U_k = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_k}_{P^\star}}\left[\hat{b}_k(s,a) + \gamma\Delta_{\hat{P}_k,P^\star}W_{k,g}(s,a)\right],$$

where

$$W_{k,g} := V^{\pi_k}_{\hat{P}_k,g+\hat{b}_k}, \quad \Delta_{\hat{P}_k,P^\star}W(s,a) := \mathbb{E}_{\hat{P}_k}[W(s')] - \mathbb{E}_{P^\star}[W(s')].$$

Since $\|\hat{b}_k\|_\infty \le \bar{b}$ and the utility is normalized to lie in $[0,1]$, we have $\|W_{k,g}\|_\infty \lesssim \bar{b}/(1-\gamma)$. Therefore

$$[\hat{U}_k - U_k]_+ \le \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_k}_{P^\star}}[\hat{b}_k] + \frac{\bar{b}}{(1-\gamma)^2}\mathbb{E}_{d^{\pi_k}_{P^\star}}[f_k].$$

The bonus term is controlled exactly as in the proof of Lemma D.10:

$$\mathbb{E}_{d^{\pi_k}_{P^\star}}[\hat{b}_k] \lesssim A_k\sqrt{\gamma d\Gamma_k(\alpha^{\mathrm{emp}}_k)^2 + \gamma\lambda_k d\bar{b}^2} + \alpha^{\mathrm{emp}}_k\sqrt{\frac{\Gamma_k d(1-\gamma)}{k}},$$

and the model-mismatch term obeys

$$\mathbb{E}_{d^{\pi_k}_{P^\star}}[f_k] \lesssim A_k\sqrt{\gamma\big(k\Gamma_k\zeta_k + \lambda_k d\big)} + \sqrt{\Gamma_k\zeta_k(1-\gamma)}.$$

Substituting the last two displays proves the episode-wise bound.

For the cumulative statement, one substitutes the same choices of $\alpha^{\mathrm{emp}}_k, \lambda_k, \zeta_k$ as in Proposition D.13. The two A_k-weighted sums are then of order

$$\tilde{\mathcal{O}}\left(\frac{\bar{b}\,d^2}{(1-\gamma)^2}\sqrt{\Gamma_K}\,\Psi_K\right), \qquad \Psi_K := \sum_{k=1}^K A_k,$$

while the two remaining sums are of order

$$\tilde{\mathcal{O}}\left(\frac{\bar{b}\,d^2}{(1-\gamma)^2}\sum_{k=1}^K\frac{\Gamma_k}{\sqrt{k}}\right).$$

Finally Lemma D.11 gives

$$\sqrt{\Gamma_K}\,\Psi_K \le \tilde{\mathcal{O}}\left(\sqrt{\Gamma_K\sum_{k=2}^K\Lambda_k}\right),$$

which proves the cumulative bound. □

Proposition D.13 (Theorem-level regret and soft-violation dependence)**.** *Assume Algorithm 1 and suppose the high-probability events required by Lemmas D.9, D.10, and D.11 hold simultaneously. Choose*

$$\alpha^{\mathrm{mix}}_k = \alpha^{\mathrm{emp}}_k = \tilde{\mathcal{O}}\Big(\sqrt{(\Gamma_k + d^2)\gamma\ln(|\mathcal{M}|k/\delta)}\Big),$$

$$\lambda_k = \tilde{\Theta}\left(\frac{\Gamma_k + d^2}{d}\,\ln(|\mathcal{M}|k/\delta)\right),$$

and choose

$$\eta = \frac{2(1-\gamma)}{\theta\bar{b}\sqrt{K}}, \qquad \xi = \tilde{\Theta}(1/\theta),$$

and assume the cap is nonbinding:

$$\bar{b} \ge \max_{1\le j\le K}\frac{\alpha^{\mathrm{emp}}_j}{\sqrt{\lambda_j}}.$$

For a nonnegative sequence $z_2,\dots,z_K$, define

$$\Xi_K(z) := \sqrt{\Gamma_K\sum_{k=2}^K z_k} + \sum_{k=1}^K\frac{\Gamma_k}{\sqrt{k}}.$$

Then

$$\sum_{k=1}^{K} \Delta_k \le \tilde{\mathcal{O}}\Big(\frac{(1+\xi)\bar{b}\, d^2}{(1-\gamma)^2} \Xi_K(\Lambda) \Big).$$

Consequently, taking the comparator dual level to be $Y = 0$ yields

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \tilde{\mathcal{O}}\Big(\frac{\bar{b}\, d^2}{\theta(1-\gamma)^2} \Xi_K(\Lambda) \Big).$$

Moreover,

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) \le \tilde{\mathcal{O}}\Big(\frac{\bar{b}\, d^2}{(1-\gamma)^2} \Xi_K(\Lambda) \Big).$$

In particular, since $\Lambda_k \le \Gamma_k = |\mathcal{A}|/\epsilon_k$,

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \tilde{\mathcal{O}}\Big(\frac{\bar{b}\, d^2}{\theta(1-\gamma)^2} \Xi_K(\Gamma) \Big).$$

Likewise,

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) \le \tilde{\mathcal{O}}\Big(\frac{\bar{b}\, d^2}{(1-\gamma)^2} \Xi_K(\Gamma) \Big).$$

For $\epsilon_k = k^{-p}$ with $p \in (0, 1/2)$, this becomes

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \tilde{\mathcal{O}}\Big(K^{1/2+p} \Big), \quad \mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) \le \tilde{\mathcal{O}}\Big(K^{1/2+p} \Big).$$

up to problem-dependent prefactors, with the regret bound carrying the extra θ^{-1} dependence. Moreover, under the displayed tuning of λ_k,

$$\frac{\alpha_k^{\mathrm{emp}}}{\sqrt{\lambda_k}} = \tilde{\mathcal{O}}(\sqrt{d}),$$

so the nonbinding-cap condition is compatible with any $\bar{b} = \tilde{\Omega}(\sqrt{d})$ and does not introduce any additional K-dependence.

Proof. Start from Lemma D.10 and substitute the tunings

$$\alpha_k^{\mathrm{emp}} = \tilde{\mathcal{O}}\Big(\sqrt{(\Gamma_k + d^2)\gamma \ln(|\mathcal{M}|k/\delta)} \Big), \quad \lambda_k = \tilde{\Theta}\Big(\frac{\Gamma_k + d^2}{d} \ln(|\mathcal{M}|k/\delta) \Big),$$

together with

$$\zeta_k = \tilde{\mathcal{O}}(1/k).$$

Under these choices, $\lambda_k d = \tilde{\Theta}((\Gamma_k + d^2)\ln(|\mathcal{M}|k/\delta))$, so every place in Lemma D.10 where $\lambda_k d$ appears is of the same order as the dominant $(\Gamma_k + d^2)$ term already present in $\alpha_k^{\mathrm{emp}\,2}$. Therefore enlarging λ_k to make the cap nonbinding does not change the final K-dependence.

We also check the cap condition explicitly:

$$\frac{\alpha_k^{\mathrm{emp}}}{\sqrt{\lambda_k}} = \tilde{\mathcal{O}}\left(\sqrt{\frac{(\Gamma_k + d^2)\ln(|\mathcal{M}|k/\delta)}{((\Gamma_k + d^2)/d)\ln(|\mathcal{M}|k/\delta)}} \right) = \tilde{\mathcal{O}}(\sqrt{d}).$$

Hence any choice $\bar{b} = \tilde{\Omega}(\sqrt{d})$ is sufficient to ensure the nonbinding-cap condition uniformly over $k \le K$.

Because Γ_k is nondecreasing, the dominant dependence in every episode-wise term may be upper bounded by Γ_K inside the coefficients that multiply A_k. Consequently the two A_k-weighted sums in Lemma D.10 are both of order

$$\tilde{\mathcal{O}}\Big(\frac{(1+\xi)\bar{b}\, d^2}{(1-\gamma)^2} \sqrt{\Gamma_K}\, \Psi_K \Big), \qquad \Psi_K := \sum_{k=1}^{K} A_k,$$

The three remaining sums in Lemma D.10 contain no factor A_k; after substituting the same choices of $\alpha_k^{\mathrm{emp}}, \lambda_k, \zeta_k$, each of them is bounded by

$$\tilde{\mathcal{O}}\left(\frac{(1+\xi)\bar{b}\,d^2}{(1-\gamma)^2}\sum_{k=1}^{K}\frac{\Gamma_k}{\sqrt{k}}\right).$$

Now apply Lemma D.11, which gives

$$\Psi_K \lesssim \frac{1}{\sqrt{\lambda_1}} + \sqrt{d\ln\left(1+\frac{K}{d\lambda_1}\right)\sum_{k=2}^{K}\Lambda_k}.$$

After absorbing the harmless initial term $1/\sqrt{\lambda_1}$ into the polylogarithmic factor, we obtain

$$\sqrt{\Gamma_K}\,\Psi_K \le \tilde{\mathcal{O}}\left(\sqrt{\Gamma_K\sum_{k=2}^{K}\Lambda_k}\right).$$

This is exactly the quantity $\Xi_K(\Lambda)$ defined in the proposition, so

$$\sum_{k=1}^{K}\Delta_k \le \tilde{\mathcal{O}}\left(\frac{(1+\xi)\bar{b}\,d^2}{(1-\gamma)^2}\Xi_K(\Lambda)\right).$$

The projection step in Algorithm 1 is identical to that of the main algorithm, so the proof of Lemma 2.4 applies verbatim and gives, for every $Y \in [0,\xi]$,

$$\sum_{k=1}^{K}(Y-Y_k)\big(b - V^{\pi_k}_{\hat{P}_k,g+\hat{b}_k}(s_1)\big) \le \frac{Y^2}{2\eta} + \frac{\eta\bar{b}^2K}{2(1-\gamma)^2}.$$

Setting $Y=0$ and combining this with Lemma D.10 yields

$$\sum_{k=1}^{K}\big(V^{\pi^\star}_{P^\star,r}(s_1) - V^{\pi_k}_{P^\star,r}(s_1)\big) \lesssim \sum_{k=1}^{K}\Delta_k + \frac{\eta\bar{b}^2K}{2(1-\gamma)^2}.$$

Substituting the bound on $\sum_k \Delta_k$ gives a right-hand side

$$\tilde{\mathcal{O}}\left(\frac{\bar{b}\,d^2}{\theta(1-\gamma)^2}\Xi_K(\Lambda)\right) + \mathcal{O}\left(\frac{\bar{b}\sqrt{K}}{\theta(1-\gamma)}\right),$$

$\eta = \frac{2(1-\gamma)}{\theta\bar{b}\sqrt{K}}$. The second term is lower order than the first whenever $\epsilon_k = k^{-p}$ with $p>0$, since then $\Xi_K(\Lambda) = \tilde{\Omega}(K^{1/2+p})$. This proves the displayed regret bound.

We now turn to the soft-violation term. Define

$$\hat{U}_k := V^{\pi_k}_{\hat{P}_k,g+\hat{b}_k}(s_1), \qquad U_k := V^{\pi_k}_{P^\star,g}(s_1), \qquad S_K := \sum_{k=1}^{K}(b-U_k).$$

Because $V^{\pi^\star}_{P^\star,g}(s_1) \ge b$, we have

$$V^{\pi^\star}_{P^\star,r}(s_1) - V^{\pi_k}_{P^\star,r}(s_1) + Y_k(b-U_k) \le \Delta_k \qquad \forall k\in[K].$$

Hence, for every $Y\in[0,\xi]$,

$$\mathrm{Regret}^{\mathrm{return}}(K) + YS_K \le \sum_{k=1}^{K}\Delta_k + \sum_{k=1}^{K}(Y-Y_k)(b-\hat{U}_k) + \sum_{k=1}^{K}(Y-Y_k)(\hat{U}_k-U_k).$$

Choose $Y=\xi$ when $S_K>0$ and $Y=0$ otherwise. Then

$$\mathrm{Regret}^{\mathrm{return}}(K) + \xi[S_K]_+ \le \sum_{k=1}^{K}\Delta_k + \frac{\xi^2}{2\eta} + \frac{\eta\bar{b}^2K}{2(1-\gamma)^2} + \xi\sum_{k=1}^{K}[\hat{U}_k-U_k]_+.$$

Now use Lemma D.12 together with the bound on $\sum_k \Delta_k$. This gives

$$\mathrm{Regret}^{\mathrm{return}}(K) + \xi[S_K]_+ \le \tilde{\mathcal{O}}\left(\frac{\xi \bar{b}\, d^2}{(1-\gamma)^2}\Xi_K(\Lambda)\right) + \mathcal{O}\left(\frac{\xi \bar{b}\sqrt{K}}{1-\gamma}\right).$$

By linearity of discounted occupancy measures, there exists a policy π' whose reward and utility values equal the episode averages:

$$V^{\pi'}_{P^\star,r}(s_1) = \frac{1}{K}\sum_{k=1}^{K} V^{\pi_k}_{P^\star,r}(s_1), \qquad V^{\pi'}_{P^\star,g}(s_1) = \frac{1}{K}\sum_{k=1}^{K} V^{\pi_k}_{P^\star,g}(s_1).$$

Therefore

$$V^{\pi^\star}_{P^\star,r}(s_1) - V^{\pi'}_{P^\star,r}(s_1) + \xi\big[b - V^{\pi'}_{P^\star,g}(s_1)\big]_+ \le \frac{1}{K}\Big(\mathrm{Regret}^{\mathrm{return}}(K) + \xi[S_K]_+\Big).$$

By Lemma 2.4, the optimal dual variable satisfies $Y^\star \le 1/\theta$, so the constant hidden in $\xi = \tilde{\Theta}(1/\theta)$ may be chosen so that $\xi \ge 2Y^\star$. Applying Lemma B.5 yields

$$[S_K]_+ \le \frac{2}{\xi}\Big(\mathrm{Regret}^{\mathrm{return}}(K) + \xi[S_K]_+\Big).$$

Since $\xi^{-1} = \tilde{\mathcal{O}}(\theta)$ and $\Xi_K(\Lambda) = \tilde{\Omega}(\sqrt{K})$, the lower-order $\mathcal{O}(\xi\bar{b}\sqrt{K}/(1-\gamma))$ term is absorbed, and we obtain

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) = [S_K]_+ \le \tilde{\mathcal{O}}\left(\frac{\bar{b}\, d^2}{(1-\gamma)^2}\Xi_K(\Lambda)\right).$$

This proves the displayed soft-violation bound. □

Proposition D.14 (Execution policy corollary for a stationary-mixture variant)**.** *The guarantees in Proposition D.13 are stated for the return policy sequence $\{\pi_k\}_{k=1}^K$ under the discounted-occupancy oracle protocol. The present proposition records a* separate *fully online interpretation in which episode k executes the stationary mixed policy*

$$\tilde{\pi}_k(\cdot \mid s) := (1-\epsilon_k)\pi_k(\cdot \mid s) + \epsilon_k U(\mathcal{A}).$$

This online mixture is not the probing protocol analyzed above; it is included only to translate return policy guarantees into execution policy guarantees.

Define

$$\mathrm{Regret}^{\mathrm{exec}}(K) := \sum_{k=1}^{K}\big(V_r^{\pi^\star}(s_1) - V_r^{\tilde{\pi}_k}(s_1)\big),$$

$$\mathrm{Violation}^{\mathrm{exec}}_{\mathrm{soft}}(K) := \left[\sum_{k=1}^{K}\big(b - V_g^{\tilde{\pi}_k}(s_1)\big)\right]_+.$$

Then

$$\mathrm{Regret}^{\mathrm{exec}}(K) \le \mathrm{Regret}^{\mathrm{return}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K}\epsilon_k,$$

$$\mathrm{Violation}^{\mathrm{exec}}_{\mathrm{soft}}(K) \le \mathrm{Violation}^{\mathrm{return}}_{\mathrm{soft}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K}\epsilon_k.$$

In particular, if $\epsilon_k = k^{-p}$ with $p \in [1/4, 1/2)$, then Proposition D.13 implies

$$\mathrm{Regret}^{\mathrm{exec}}(K),\ \mathrm{Violation}^{\mathrm{exec}}_{\mathrm{soft}}(K) \le \tilde{\mathcal{O}}\Big(K^{1/2+p}\Big),$$

which matches the asymptotic orders of the return policy bounds.

Proof. Fix an episode k. For $j \in \{r, g\}$, define $A_j^{\pi_k}(s,a) := Q_j^{\pi_k}(s,a) - V_j^{\pi_k}(s)$. By the trajectory-level normalization introduced in the background section, $|A_j^{\pi_k}(s,a)| \le 1$ for both $j = r$ and $j = g$. The performance-difference lemma with normalized discounted occupancy measure gives

$$(1-\gamma)\big(V_j^{\tilde{\pi}_k}(s_1) - V_j^{\pi_k}(s_1)\big) = \mathbb{E}_{s\sim d_{P^\star}^{\tilde{\pi}_k}}\Big[\sum_{a\in\mathcal{A}}\big(\tilde{\pi}_k(a \mid s) - \pi_k(a \mid s)\big)A_j^{\pi_k}(s,a)\Big].$$

Hence

$$D_{k,j}(s) := \sum_{a\in\mathcal{A}} \big|\tilde{\pi}_k(a \mid s) - \pi_k(a \mid s)\big| \cdot \big|A_j^{\pi_k}(s,a)\big|.$$

Therefore

$$\big|V_j^{\tilde{\pi}_k}(s_1) - V_j^{\pi_k}(s_1)\big| \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{P^\star}^{\tilde{\pi}_k}}\big[D_{k,j}(s)\big] \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{P^\star}^{\tilde{\pi}_k}}\Big[\|\tilde{\pi}_k(\cdot \mid s) - \pi_k(\cdot \mid s)\|_1\Big].$$

Now

$$\tilde{\pi}_k(\cdot \mid s) - \pi_k(\cdot \mid s) = \epsilon_k\big(U(\mathcal{A}) - \pi_k(\cdot \mid s)\big), \quad \|\tilde{\pi}_k(\cdot \mid s) - \pi_k(\cdot \mid s)\|_1 \le \epsilon_k \|U(\mathcal{A}) - \pi_k(\cdot \mid s)\|_1 \le 2\epsilon_k.$$

Substituting yields, for both $j = r$ and $j = g$,

$$\big|V_j^{\tilde{\pi}_k}(s_1) - V_j^{\pi_k}(s_1)\big| \le \frac{2\epsilon_k}{1-\gamma}.$$

Applying this bound with $j = r$, we obtain

$$\begin{aligned}\mathrm{Regret}^{\mathrm{exec}}(K) =& \sum_{k=1}^{K}\big(V_r^{\pi^\star}(s_1) - V_r^{\pi_k}(s_1)\big) + \sum_{k=1}^{K}\big(V_r^{\pi_k}(s_1) - V_r^{\tilde{\pi}_k}(s_1)\big)\\ \le& \mathrm{Regret}^{\mathrm{return}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K}\epsilon_k.\end{aligned}$$

Applying the same bound with $j = g$, together with $[x+y]_+ \le [x]_+ + y$ for $y \ge 0$, yields

$$\begin{aligned}\mathrm{Violation}_{\mathrm{soft}}^{\mathrm{exec}}(K) =& \left[\sum_{k=1}^{K}\big(b - V_g^{\tilde{\pi}_k}(s_1)\big)\right]_+\\ \le& \left[\sum_{k=1}^{K}\big(b - V_g^{\pi_k}(s_1)\big)\right]_+ + \sum_{k=1}^{K}\big|V_g^{\tilde{\pi}_k}(s_1) - V_g^{\pi_k}(s_1)\big|\\ \le& \mathrm{Violation}_{\mathrm{soft}}^{\mathrm{return}}(K) + \frac{2}{1-\gamma}\sum_{k=1}^{K}\epsilon_k.\end{aligned}$$

Finally, if $\epsilon_k = k^{-p}$ with $p \in [1/4, 1/2)$, then Proposition D.13 gives

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \tilde{\mathcal{O}}(K^{1/2+p}), \qquad \sum_{k=1}^{K}\epsilon_k = \sum_{k=1}^{K}k^{-p} = \mathcal{O}(K^{1-p}).$$

Since $p \ge 1/4$ is equivalent to $1 - p \le 1/2 + p$, the additional execution term is of no larger order than the returned-policy terms. Hence $\mathrm{Regret}^{\mathrm{exec}}(K) \le \tilde{\mathcal{O}}(K^{1/2+p})$. The same comparison applies to $\mathrm{Violation}_{\mathrm{soft}}^{\mathrm{exec}}(K)$, because the returned-policy soft-violation term is already of order $\tilde{\mathcal{O}}(K^{1/2+p})$. Therefore $\mathrm{Violation}_{\mathrm{soft}}^{\mathrm{exec}}(K) \le \tilde{\mathcal{O}}(K^{1/2+p})$. □

E Uniform Exploration as a Tertiary-Aware Repaired Variant

This appendix specializes the decaying-ϵ analysis in Appendix D to the constant schedule $\epsilon_k \equiv 1$. This case is useful as a clean benchmark: all probe actions are drawn uniformly, while the tertiary probe level and merged bonus covariance are retained so that the repaired optimism proof still applies.

Throughout this appendix, we set

$$q_k(\cdot \mid s) = q_k'(\cdot \mid s) = q_k''(\cdot \mid s) = U(\mathcal{A}), \qquad \forall k \in [K].$$

Equivalently, this is Algorithm 1 with $\epsilon_k \equiv 1$. Therefore

$$\Gamma_k = \Lambda_k = |\mathcal{A}|, \qquad \forall k \in [K].$$

E.1 Uniform Specialization

For each round $i \in [K]$, define the primary, secondary, and tertiary state marginals by

$$\rho_i(s) := d_{P^\star}^{\pi_{i-1}}(s), \quad \rho_i'(s') := \mathbb{E}_{\rho_i \times U(\mathcal{A})} P^\star(s' \mid s, a), \quad \rho_i''(s'') := \mathbb{E}_{\rho_i' \times U(\mathcal{A})} P^\star(s'' \mid s, a).$$

The corresponding probing distributions are

$$\nu_i(s,a) := \rho_i(s)U(a), \quad \nu_i'(s,a) := \rho_i'(s)U(a), \quad \nu_i''(s,a) := \rho_i''(s)U(a).$$

We write

$$\bar{\nu}_k^{\mathrm{pri}} := \frac{1}{k}\sum_{i=1}^k \nu_i, \qquad \bar{\nu}_k^{\mathrm{sec}} := \frac{1}{k}\sum_{i=1}^k \nu_i', \qquad \bar{\nu}_k^{\mathrm{ter}} := \frac{1}{k}\sum_{i=1}^k \nu_i''.$$

No separate pseudocode is needed: this is exactly Algorithm 1 with the constant schedule $\epsilon_k \equiv 1$. Relative to the general algorithm, the only change is that all probing kernels are fixed to $U(\mathcal{A})$; the tertiary tuple, merged covariance, model fitting step, and primal-dual update remain the same.

E.2 Supporting Specializations

The tuple-level MLE event, the pathwise-to-averaged bridge for the squared MLE-mismatch class, and the merged covariance comparison from Appendix D apply verbatim to Algorithm 1 with $\epsilon_k \equiv 1$. The only new point is that all coverage coefficients collapse to explicit constants.

Proposition E.1 (Uniform coverage specialization)**.** *Under Algorithm 1 with $\epsilon_k \equiv 1$, for every episode k and every policy π,*

$$\mathfrak{C}_k^{\mathrm{init}}(\pi) \le |\mathcal{A}|, \qquad \mathfrak{C}_k^{\mathrm{next}}(\pi) \le |\mathcal{A}|,$$

and

$$\mathfrak{C}_{k,1}^{\mathrm{tri,next}}(\pi) \le |\mathcal{A}|, \qquad \mathfrak{C}_{k,2}^{\mathrm{tri,next}}(\pi) \le |\mathcal{A}|, \qquad \mathfrak{C}_k^{\mathrm{tri,next}}(\pi) \le |\mathcal{A}|.$$

Proof. Because every probe kernel is uniform,

$$\bar{\nu}_k^{\mathrm{pri}}(s,a) = \frac{1}{|\mathcal{A}|} \cdot \frac{1}{k}\sum_{i=1}^k \rho_i(s),$$

and every discounted occupancy satisfies

$$\rho_i(s) = d_{P^\star}^{\pi_{i-1}}(s) \ge (1-\gamma) d_1(s).$$

Hence

$$\bar{\nu}_k^{\mathrm{pri}}(s,a) \ge \frac{(1-\gamma)d_1(s)}{|\mathcal{A}|},$$

which implies

$$\mathfrak{C}_k^{\mathrm{init}}(\pi) \le |\mathcal{A}|$$

for every policy π.

The bound on $\mathfrak{C}_k^{\mathrm{next}}(\pi)$ follows exactly as in the uniform version of the original proof: the next-step pushforward state-action measure has density at most $|\mathcal{A}|$ times $\bar{\nu}_k^{\mathrm{sec}}$, because the action component of $\bar{\nu}_k^{\mathrm{sec}}$ is uniform.

The tertiary coefficients are the specialization of Proposition D.7 to the constant schedule $\epsilon_k \equiv 1$, for which $\Gamma_k = |\mathcal{A}|$. □

Proposition E.2 (Uniform potential reduction)**.** *Under Algorithm 1 with $\epsilon_k \equiv 1$,*

$$\Psi_K := \sum_{k=1}^K A_k \lesssim \frac{1}{\sqrt{\lambda_1}} + \sqrt{|\mathcal{A}|\, dK \ln\left(1 + \frac{K}{d\lambda_1}\right)}.$$

Proof. This is the specialization of Lemma D.11 to $\Lambda_k = |\mathcal{A}|$ for all k:

$$\Psi_K \lesssim \frac{1}{\sqrt{\lambda_1}} + \sqrt{d \ln\left(1 + \frac{K}{d\lambda_1}\right) \sum_{k=2}^K \Lambda_k}.$$

Since $\sum_{k=2}^K \Lambda_k = (K-1)|\mathcal{A}|$, the claim follows. □

E.3 Theorem-Level Derivation

Theorem E.3 (Return policy regret and soft violation for $p = 0$)**.** *Assume Algorithm 1 is run with $\epsilon_k \equiv 1$, and suppose the high-probability events required by Lemmas D.9, D.10, and D.11 hold simultaneously. Choose*

$$\alpha_k^{\mathrm{mix}} = \alpha_k^{\mathrm{emp}} = \tilde{\mathcal{O}}\Big(\sqrt{(|\mathcal{A}| + d^2)\gamma \ln(|\mathcal{M}|k/\delta)}\Big),$$

$$\lambda_k = \tilde{\Theta}\left(\frac{|\mathcal{A}| + d^2}{d} \ln(|\mathcal{M}|k/\delta)\right),$$

and

$$\eta = \frac{2(1-\gamma)}{\theta \bar{b}\sqrt{K}}, \qquad \xi = \tilde{\Theta}(1/\theta),$$

and assume the cap is nonbinding:

$$\bar{b} \geq \max_{1 \leq j \leq K} \frac{\alpha_j^{\mathrm{emp}}}{\sqrt{\lambda_j}}.$$

Then, with probability at least $1 - \delta$,

$$\mathrm{Regret}^{\mathrm{return}}(K) \leq \tilde{\mathcal{O}}\left(\frac{|\mathcal{A}|\bar{b}\, d^2\sqrt{K}}{\theta(1-\gamma)^2}\right), \quad \mathrm{Violation}_{\mathrm{soft}}^{\mathrm{return}}(K) \leq \tilde{\mathcal{O}}\left(\frac{|\mathcal{A}|\bar{b}\, d^2\sqrt{K}}{(1-\gamma)^2}\right).$$

Proof. This is a direct specialization of Proposition D.13 to the constant schedule $\epsilon_k \equiv 1$. Under this choice,

$$\Gamma_k = \Lambda_k = |\mathcal{A}|, \qquad \forall k \in [K].$$

Therefore the quantity

$$\Xi_K(\Lambda) := \sqrt{\Gamma_K \sum_{k=2}^{K} \Lambda_k} + \sum_{k=1}^{K} \frac{\Gamma_k}{\sqrt{k}}$$

satisfies

$$\Xi_K(\Lambda) = \sqrt{|\mathcal{A}| \cdot (K-1)|\mathcal{A}|} + |\mathcal{A}| \sum_{k=1}^{K} \frac{1}{\sqrt{k}} \lesssim |\mathcal{A}|\sqrt{K}.$$

Substituting this into Proposition D.13 yields

$$\mathrm{Regret}^{\mathrm{return}}(K) \leq \tilde{\mathcal{O}}\left(\frac{\bar{b}\, d^2}{\theta(1-\gamma)^2}\Xi_K(\Lambda)\right) \leq \tilde{\mathcal{O}}\left(\frac{|\mathcal{A}|\bar{b}\, d^2\sqrt{K}}{\theta(1-\gamma)^2}\right),$$

and

$$\mathrm{Violation}_{\mathrm{soft}}^{\mathrm{return}}(K) \leq \tilde{\mathcal{O}}\left(\frac{\bar{b}\, d^2}{(1-\gamma)^2}\Xi_K(\Lambda)\right) \leq \tilde{\mathcal{O}}\left(\frac{|\mathcal{A}|\bar{b}\, d^2\sqrt{K}}{(1-\gamma)^2}\right).$$

It remains only to note that the displayed tuning of λ_k makes the cap condition compatible with a K-independent choice of $\bar{b}$. Indeed,

$$\frac{\alpha_k^{\mathrm{emp}}}{\sqrt{\lambda_k}} = \tilde{\mathcal{O}}\left(\sqrt{\frac{(|\mathcal{A}| + d^2)\ln(|\mathcal{M}|k/\delta)}{((|\mathcal{A}| + d^2)/d)\ln(|\mathcal{M}|k/\delta)}}\right) = \tilde{\mathcal{O}}(\sqrt{d}),$$

so any $\bar{b} = \tilde{\Omega}(\sqrt{d})$ is sufficient, exactly as in Appendix D. □

Constraint discussion. The regret proof uses the same representation-learning and optimism ingredients as Appendix D. For soft violation, one additional ingredient is needed: the cumulative overestimation lemma controls how much the optimistic utility used by the dual update can exceed the true utility. This term is summable under the repaired tertiary design, so the standard projected dual argument yields the stated soft-violation bound.

Remark E.4 (Relation to the $p = 0$ specialization)**.** The uniform protocol is exactly the $\epsilon_k \equiv 1$ specialization of Algorithm 1 in Section 3. Consequently, this appendix gives the fully explicit return-policy theorem for the $p = 0$ schedule, with the familiar $\tilde{\mathcal{O}}(|\mathcal{A}|d^2\sqrt{K})$ dependence up to the same problem-dependent prefactors and logarithmic terms.

F Hard Constraints Under Decaying Tertiary Probing

This appendix gives the repaired hard-violation analysis. The key distinction is between the one-sided correction used by the algorithm and the two-sided width used only in the proof. In particular, we do not assume an external certificate event and we do not require a summable radius uniformly over all policies queried by the inner loop.

F.1 Algorithm and Certificate

The representation-learning front end is the same as in Algorithm 1. At episode k, let

$$q_k = q'_k = q''_k := (1-\epsilon_k)\pi_{k-1} + \epsilon_k U(\mathcal{A}), \qquad \Gamma_k := \frac{|\mathcal{A}|}{\epsilon_k},$$

and let

$$\zeta_k := \frac{\ln(|\mathcal{M}|k/\delta)}{k}.$$

After fitting $\widehat{P}_k = (\widehat{\mu}_k, \widehat{\phi}_k)$ on the three streams and forming $\widehat{\Sigma}_k^{\mathrm{mix}}$, define

$$\widehat{b}_k(s,a) := \min\Big\{\alpha_k \|\widehat{\phi}_k(s,a)\|_{(\widehat{\Sigma}_k^{\mathrm{mix}})^{-1}}, \bar{b}\Big\}.$$

For a sufficiently large universal calibration constant C_{cert}, set

$$\ell_k := C_{\mathrm{cert}}\sqrt{\frac{\Gamma_k \zeta_k}{1-\gamma}}, \qquad \underline{\mathcal{U}}_k(\pi) := \Big[V^{\pi}_{\widehat{P}_k, g-\widehat{b}_k}(s_1) - \ell_k\Big]_+. \tag{G.1}$$

The quantity ℓ_k is the one-sided statistical correction used by the algorithm. For analysis, define the policy-dependent selected width

$$w_k(\pi) := 2\ell_k + 2V^{\pi}_{\widehat{P}_k, \widehat{b}_k}(s_1). \tag{G.2}$$

Unlike the correction ℓ_k, the width $w_k(\pi)$ is not required as an algorithmic input.

Let $Y_{\max} := K^{1/4}$, and use the capped grid

$$\mathcal{Y}_K := \{0, \eta_h, 2\eta_h, \ldots\} \cap [0, Y_{\max}] \ \cup \ \{Y_{\max}\}.$$

Thus the cap itself is always queried even when $Y_{\max}/\eta_h$ is not an integer.

F.2 Derived Certificate Event

Let $\mathcal{E}_{\mathrm{rep}}$ be the intersection of the tuple-level MLE, pathwise bridge, and covariance-comparison events in Lemmas D.2, D.4, and D.6. Replacing δ by a constant fraction in the individual events makes $\mathbb{P}(\mathcal{E}_{\mathrm{rep}}) \geq 1-\delta$.

Lemma F.1 (One-sided certificate and selected-policy interval)**.** *Assume $\mathcal{E}_{\mathrm{rep}}$, the nonbinding-cap condition*

$$\bar{b} \geq \max_{j \leq K} \frac{\alpha_j}{\sqrt{\lambda_j}},$$

and the parameter lower bound on α_k in Lemma D.9 with $\mathfrak{C}_k^{\mathrm{tri,next}}(\pi) \leq \Gamma_k$. For a sufficiently large C_{cert}, simultaneously for every episode k and every policy π,

$$\underline{\mathcal{U}}_k(\pi) \leq V^{\pi}_{P^\star, g}(s_1) \leq \underline{\mathcal{U}}_k(\pi) + w_k(\pi). \tag{G.3}$$

Proof. We first verify that the residual in the one-step-back argument is uniform in π. Since every primary occupancy contains the initial component $(1-\gamma)d_1$, and every probing kernel up to episode k has action floor $\epsilon_k/|\mathcal{A}|$,

$$\bar{\nu}_k^{\mathrm{pri}}(s,a) \geq (1-\gamma)d_1(s)\frac{\epsilon_k}{|\mathcal{A}|}.$$

Consequently

$$\mathfrak{C}_k^{\mathrm{init}}(\pi) = \mathsf{D}\Big((1-\gamma)d_1 \otimes \pi \ \Big\|\ \bar{\nu}_k^{\mathrm{pri}}\Big) \leq \Gamma_k$$

Algorithm 2 REP-PD-H-TEG with an explicit one-sided certificate

Input: Parameters $\{\lambda_k, \alpha_k, \epsilon_k\}_{k=1}^K$, confidence δ, calibration C_{cert}, temperature τ_h, grid step η_h, model class $\mathcal{M}$, and episodes K.
Initialize $\pi_0(\cdot \mid s) = U(\mathcal{A})$ and $D_0 = D_0' = D_0'' = \emptyset$.
for $k = 1, \dots, K$ **do**
  Set $q_k = q_k' = q_k'' = (1-\epsilon_k)\pi_{k-1} + \epsilon_k U(\mathcal{A})$.
  Query one tertiary tuple and update D_k, D_k', D_k'' exactly as in Algorithm 1.
  Fit

$$\widehat{P}_k = (\widehat{\mu}_k, \widehat{\phi}_k) \in \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{D_k + D_k' + D_k''}[\log \mu(x')^\top \phi(x,u)].$$

  Form $\widehat{\Sigma}_k^{\mathrm{mix}}$ and $\widehat{b}_k$ as above, and set

$$\ell_k = C_{\mathrm{cert}} \sqrt{\frac{|\mathcal{A}| \ln(|\mathcal{M}|k/\delta)}{\epsilon_k k(1-\gamma)}}.$$

  Set found $\leftarrow$ false.
  for Y in the increasing grid $\mathcal{Y}_K$ **do**
    Compute

$$\widehat{\pi}_{k,Y} \in \arg\max_{\pi} \left(Q^{\pi}_{\widehat{P}_k, r+\widehat{b}_k} + Y Q^{\pi}_{\widehat{P}_k, g+\widehat{b}_k} \right).$$

    Form the statewise softmax candidate

$$\pi_{k,Y}(\cdot \mid x) = \mathrm{softmax}_{\tau_h}\left(Q^{\widehat{\pi}_{k,Y}}_{\widehat{P}_k, r+\widehat{b}_k}(x,\cdot) + Y Q^{\widehat{\pi}_{k,Y}}_{\widehat{P}_k, g+\widehat{b}_k}(x,\cdot) \right).$$

    Compute

$$\underline{\mathcal{U}}_k(\pi_{k,Y}) = \left[V^{\pi_{k,Y}}_{\widehat{P}_k, g-\widehat{b}_k}(s_1) - \ell_k \right]_+.$$

    if $\underline{\mathcal{U}}_k(\pi_{k,Y}) \geq b$ **then**
      Set $(\pi_k, Y_k) \leftarrow (\pi_{k,Y}, Y)$ and found $\leftarrow$ true.
      break
    end if
  end for
  if found $=$ false **then**
    Set $(\pi_k, Y_k) \leftarrow (\pi_{k,Y_{\max}}, Y_{\max})$.
  end if
end for
Return $\pi_1, \dots, \pi_K$.

for every policy π. Proposition D.7 likewise gives $\mathfrak{C}_k^{\mathrm{tri,next}}(\pi) \leq \Gamma_k$ uniformly in π.

Write $f_k(s,a) = \|\widehat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$. The first form of the simulation lemma gives

$$\begin{aligned}(1-\gamma)\left(V^{\pi}_{\widehat{P}_k, g-\widehat{b}_k}(s_1) - V^{\pi}_{P^\star, g}(s_1) \right) \\ = \mathbb{E}_{d^{\pi}_{\widehat{P}_k}}\left[-\widehat{b}_k(s,a) + \gamma \left(\mathbb{E}_{\widehat{P}_k} - \mathbb{E}_{P^\star} \right) V^{\pi}_{P^\star, g}(s') \right].\end{aligned} \tag{G.4}$$

The normalized utility assumption implies $\|V^{\pi}_{P^\star,g}\|_\infty \leq 1$. Applying Lemma D.8 with $g = f_k$, using $\mathfrak{C}_k^{\mathrm{init}}(\pi), \mathfrak{C}_k^{\mathrm{tri,next}}(\pi) \leq \Gamma_k$, and choosing α_k to absorb the covariance-dependent term yields

$$V^{\pi}_{\widehat{P}_k, g-\widehat{b}_k}(s_1) - V^{\pi}_{P^\star, g}(s_1) \leq C_{\mathrm{cert}} \sqrt{\frac{\Gamma_k \zeta_k}{1-\gamma}} = \ell_k. \tag{G.5}$$

The proof is deterministic on $\mathcal{E}_{\mathrm{rep}}$, so no additional union bound over the adaptively queried policies is required. Since the true utility is nonnegative, (G.5) proves the first inequality in (G.3) after positive-part clipping.

Repeating the direct-optimism argument of Lemma D.9 for the normalized utility signal g gives

$$V^{\pi}_{P^\star, g}(s_1) \leq V^{\pi}_{\widehat{P}_k, g+\widehat{b}_k}(s_1) + \ell_k. \tag{G.6}$$

For the same model and policy, discounted value is linear in the signal:

$$V^{\pi}_{\widehat{P}_k,g+\widehat{b}_k}(s_1) - V^{\pi}_{\widehat{P}_k,g-\widehat{b}_k}(s_1) = 2V^{\pi}_{\widehat{P}_k,\widehat{b}_k}(s_1). \tag{G.7}$$

Combining (G.6), (G.7), and the raw lower endpoint $V^{\pi}_{\widehat{P}_k,g-\widehat{b}_k} - \ell_k$ proves the upper inequality with $w_k(\pi) = 2\ell_k + 2V^{\pi}_{\widehat{P}_k,\widehat{b}_k}$. Replacing a negative raw endpoint by zero only tightens this upper bound because $V^{\pi}_{P^\star,g} \geq 0$. □

For the selected policy sequence, define

$$\mathcal{B}_K := \widetilde{O}\left(\frac{\bar{b}d^2}{(1-\gamma)^2}\left[\sqrt{\Gamma_K \sum_{k=2}^{K}\Lambda_k} + \sum_{k=1}^{K}\frac{\Gamma_k}{\sqrt{k}}\right]\right), \tag{G.8}$$

where

$$\Lambda_k := \frac{1}{1-\epsilon_k+\epsilon_k/|\mathcal{A}|}.$$

Lemma F.2 (Selected-policy width summability)**.** *On $\mathcal{E}_{\mathrm{rep}}$, the policy sequence returned by Algorithm 2 satisfies*

$$\sum_{k=1}^{K} w_k(\pi_k) \leq \mathcal{B}_K + 2\sum_{k=1}^{K}\ell_k. \tag{G.9}$$

Proof. Condition on the history before episode k. Apply the second form of the simulation lemma to the signal $\widehat{b}_k$, with no additional bonus:

$$V^{\pi_k}_{\widehat{P}_k,\widehat{b}_k}(s_1) - V^{\pi_k}_{P^\star,\widehat{b}_k}(s_1) = \frac{\gamma}{1-\gamma}\mathbb{E}_{d^{\pi_k}_{P^\star}}\left[\left(\mathbb{E}_{\widehat{P}_k} - \mathbb{E}_{P^\star}\right)V^{\pi_k}_{\widehat{P}_k,\widehat{b}_k}(s')\right].$$

Because $0 \leq \widehat{b}_k \leq \bar{b}$, $\|V^{\pi_k}_{\widehat{P}_k,\widehat{b}_k}\|_\infty \leq \bar{b}/(1-\gamma)$, and hence

$$V^{\pi_k}_{\widehat{P}_k,\widehat{b}_k}(s_1) \leq \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_k}_{P^\star}}[\widehat{b}_k] + \frac{\gamma\bar{b}}{(1-\gamma)^2}\mathbb{E}_{d^{\pi_k}_{P^\star}}[f_k]. \tag{G.10}$$

The two expectations in (G.10) are exactly the bonus and model-error terms bounded in the proof of Lemma D.10; their cumulative bound follows from Lemma D.11 and is $\mathcal{B}_K$ up to universal constants. Summing (G.2) and using (G.10) proves (G.9). Importantly, this step is applied only to the returned sequence $\{\pi_k\}$, for which the deployed-to-probe coupling used by Lemma D.11 is valid. □

For a polynomial schedule $\epsilon_k = k^{-p}$,

$$\sum_{k=1}^{K}\ell_k \leq \widetilde{O}\left(\sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\,K^{(1+p)/2}\right), \tag{G.11}$$

while

$$\mathcal{B}_K \leq \widetilde{O}\left(\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2}K^{1/2+p}\right). \tag{G.12}$$

Thus the correction term is no larger in K than the selected-width term.

F.3 Hard Front End

For the selected dual level Y_k, define the Lagrangian front end

$$\mathfrak{F}_K := \sum_{k=1}^{K}\left[V^{\pi^\star}_{P^\star,r}(s_1) - V^{\pi_k}_{P^\star,r}(s_1) + Y_k\left(V^{\pi^\star}_{P^\star,g}(s_1) - V^{\pi_k}_{P^\star,g}(s_1)\right)\right]. \tag{G.13}$$

Let σ^{sm}_k be the value suboptimality introduced by replacing the exact optimistic planner at the selected dual level with its statewise softmax candidate. Under Assumption C.1,

$$\sum_{k=1}^{K}\sigma^{\mathrm{sm}}_k \leq \widetilde{O}\left(\frac{K|\mathcal{A}|}{(1-\gamma)^2}\exp(-\delta_{\min}/\tau_h)\right). \tag{G.14}$$

Theorem F.3 (Decaying hard front end)**.** *Assume $\mathcal{E}_{\mathrm{rep}}$, the nonbinding bonus cap, and Assumption C.1. With the same α_k, λ_k as in the soft analysis and $0 \le Y_k \le Y_{\max}$,*

$$\mathfrak{F}_K \le \widetilde{O}\left(\frac{\bar{b}d^2 Y_{\max}}{(1-\gamma)^2}\left[\sqrt{\Gamma_K \sum_{k=2}^{K} \Lambda_k} + \sum_{k=1}^{K} \frac{\Gamma_k}{\sqrt{k}}\right]\right) + \sum_{k=1}^{K} \sigma_k^{\mathrm{sm}}. \tag{G.15}$$

Proof. At each selected dual level, the optimistic Lagrangian comparison is the same as in Lemma D.10, with the utility-side statistical terms multiplied by $Y_k \le Y_{\max}$. The reward-side terms are unchanged. The returned policy is a softmax approximation of the optimistic planner, so its value loss contributes σ_k^{sm}. Applying Lemma D.11 to the returned sequence and summing gives (G.15). No certificate width is used in this front-end argument. □

F.4 Selected-Certificate Reduction

Let $\mathcal{K}_{\mathrm{br}}$ contain episodes that stop at a certified grid point, and let $\mathcal{K}_{\mathrm{cap}}$ contain episodes that reach $Y_{\max}$ without certification. Define the certificate grid sensitivity

$$s_{\mathrm{grid}} := C_{\mathrm{grid}} \frac{\eta_h}{\tau_h(1-\gamma)}, \tag{G.16}$$

where C_{grid} is the universal constant from the softmax-sensitivity Lemma C.2. The positive-part clipping is 1-Lipschitz, so neighboring certificates differ by at most s_{grid}.

Proposition F.4 (Selected-certificate regret reduction)**.** *On $\mathcal{E}_{\mathrm{rep}}$,*

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \mathfrak{F}_K + Y_{\max}\left(\sum_{k=1}^{K} w_k(\pi_k) + K s_{\mathrm{grid}}\right), \tag{G.17}$$

and

$$\mathrm{Violation}_{\mathrm{hard}}^{\mathrm{return}}(K) \le \sum_{k \in \mathcal{K}_{\mathrm{cap}}} \left[b - V_{P^\star,g}^{\pi_k}(s_1)\right]_+. \tag{G.18}$$

Proof. If $k \in \mathcal{K}_{\mathrm{br}}$, then $\underline{\mathcal{U}}_k(\pi_k) \ge b$, so Lemma F.1 gives $V_{P^\star,g}^{\pi_k}(s_1) \ge b$. Break episodes therefore contribute zero hard violation, proving (G.18).

To convert the Lagrangian front end into reward regret, use $V_{P^\star,g}^{\pi^\star}(s_1) \ge b$:

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \mathfrak{F}_K + \sum_{k=1}^{K} Y_k \left[V_{P^\star,g}^{\pi_k}(s_1) - b\right]_+. \tag{G.19}$$

If a break occurs at $Y_k > 0$, the preceding grid point failed the certificate. Grid sensitivity and Lemma F.1 therefore imply

$$\left[V_{P^\star,g}^{\pi_k}(s_1) - b\right]_+ \le s_{\mathrm{grid}} + w_k(\pi_k).$$

When $Y_k = 0$, the corresponding term in (G.19) is zero. On a cap episode, the certificate at the explicitly queried cap failed, so

$$\left[V_{P^\star,g}^{\pi_k}(s_1) - b\right]_+ \le w_k(\pi_k).$$

Using $Y_k \le Y_{\max}$ proves (G.17). □

F.5 Hard Violation at the Executed Cap Policy

The remaining argument is performed at the cap policy that is actually returned and subsequently enters the probing kernel; it does not use an unexecuted fixed-Y_θ candidate.

For $c \in \{r, g\}$, define the selected-policy optimistic overestimation

$$o_{k,c} := \left[V_{\widehat{P}_k, c+\widehat{b}_k}^{\pi_k}(s_1) - V_{P^\star,c}^{\pi_k}(s_1)\right]_+. \tag{G.20}$$

The proof of Lemma D.12 applies to the returned sequence of the hard algorithm and to either normalized signal r or g; hence

$$\sum_{k=1}^{K} o_{k,r} \le \mathcal{B}_K, \qquad \sum_{k=1}^{K} o_{k,g} \le \mathcal{B}_K. \tag{G.21}$$

Lemma F.5 (Cap-level Slater pressure)**.** *Let $\bar{\pi}$ satisfy the Slater condition $V^{\bar{\pi}}_{P^\star,g}(s_1) \ge b + \theta$. For any $k \in \mathcal{K}_{\mathrm{cap}}$, if $Y_{\max}\theta \ge 1$, then*

$$\Big[b - V^{\pi_k}_{P^\star,g}(s_1)\Big]_+ \le \frac{o_{k,r}}{Y_{\max}} + o_{k,g} + \left(1 + \frac{1}{Y_{\max}}\right)\ell_k + \frac{\sigma_k^{\mathrm{sm}}}{Y_{\max}}. \tag{G.22}$$

Proof. At the cap, softmax planning gives

$$\begin{aligned} &V^{\pi_k}_{\widehat{P}_k, r+\widehat{b}_k}(s_1) + Y_{\max} V^{\pi_k}_{\widehat{P}_k, g+\widehat{b}_k}(s_1) \\ &\quad \ge V^{\bar{\pi}}_{\widehat{P}_k, r+\widehat{b}_k}(s_1) + Y_{\max} V^{\bar{\pi}}_{\widehat{P}_k, g+\widehat{b}_k}(s_1) - \sigma_k^{\mathrm{sm}}. \end{aligned} \tag{G.23}$$

Direct optimism, with $\mathfrak{C}_k^{\mathrm{init}}(\bar{\pi}) \le \Gamma_k$, yields

$$V^{\bar{\pi}}_{\widehat{P}_k, c+\widehat{b}_k}(s_1) \ge V^{\bar{\pi}}_{P^\star,c}(s_1) - \ell_k \qquad (c \in \{r, g\}). \tag{G.24}$$

The definition of $o_{k,c}$ upper-bounds the left-hand optimistic values in (G.23) by the corresponding true values plus $o_{k,c}$. Consequently,

$$\begin{aligned} &V^{\pi_k}_{P^\star,r}(s_1) + Y_{\max} V^{\pi_k}_{P^\star,g}(s_1) + o_{k,r} + Y_{\max} o_{k,g} \\ &\quad \ge V^{\bar{\pi}}_{P^\star,r}(s_1) + Y_{\max} V^{\bar{\pi}}_{P^\star,g}(s_1) - (1 + Y_{\max})\ell_k - \sigma_k^{\mathrm{sm}}. \end{aligned} \tag{G.25}$$

If $v_k = [b - V^{\pi_k}_{P^\star,g}(s_1)]_+ > 0$, the Slater condition and the normalization $|V^{\pi_k}_{P^\star,r} - V^{\bar{\pi}}_{P^\star,r}| \le 1$ imply

$$Y_{\max}(\theta + v_k) \le 1 + o_{k,r} + Y_{\max} o_{k,g} + (1 + Y_{\max})\ell_k + \sigma_k^{\mathrm{sm}}.$$

Since $Y_{\max}\theta \ge 1$, subtracting this term and dividing by $Y_{\max}$ proves (G.22). The claim is trivial when $v_k = 0$. □

Proposition F.6 (Decaying Slater-induced hard violation)**.** *Assume the hypotheses of Proposition F.4. If $Y_{\max}\theta \ge 1$, then*

$$\mathrm{Violation}^{\mathrm{return}}_{\mathrm{hard}}(K) \le \left(1 + \frac{1}{Y_{\max}}\right)\left(2\mathcal{B}_K + \sum_{k=1}^{K}\ell_k + \sum_{k=1}^{K}\sigma_k^{\mathrm{sm}}\right) \tag{G.26}$$

up to universal constants. If $Y_{\max}\theta < 1$, the trivial bound $\mathrm{Violation}^{\mathrm{return}}_{\mathrm{hard}}(K) \le K \le \theta^{-4}$ applies.

Proof. Combine (G.18) with Lemma F.5, sum over cap episodes, and use (G.21). Extending each nonnegative sum from $\mathcal{K}_{\mathrm{cap}}$ to all episodes gives (G.26). The second statement follows from $Y_{\max} = K^{1/4}$. □

F.6 Polynomial Schedules

Corollary F.7 (Polynomial hard rates)**.** *Let $\epsilon_k = k^{-p}$ with $p \in [0, 1/2)$, and choose*

$$\alpha_k = \widetilde{\Theta}\Big(\sqrt{(\Gamma_k + d^2)\gamma \log(|\mathcal{M}|k/\delta)}\Big) \quad \textit{with a sufficiently large leading constant},$$

$$\lambda_k = \widetilde{\Theta}\left(\frac{\Gamma_k + d^2}{d}\log(|\mathcal{M}|k/\delta)\right).$$

$$\tau_h = \Theta\left(\frac{1}{(1-\gamma)\sqrt{K}\log|\mathcal{A}|}\right), \qquad \eta_h = \Theta\left(\frac{\tau_h(1-\gamma)}{K^{3/4}}\right).$$

Under the assumptions of Theorem F.3 and Proposition F.6, with probability at least $1 - \delta$,

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \widetilde{O}\left(\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right]K^{3/4+p}\right), \tag{G.27}$$

and

$$\mathrm{Violation}_{\mathrm{hard}}^{\mathrm{return}}(K) \le \widetilde{O}\left(\frac{1}{\theta}\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{1/2+p}\right) + O(\theta^{-4}). \tag{G.28}$$

Proof. For polynomial schedules,

$$\sqrt{\Gamma_K \sum_{k=2}^{K} \Lambda_k} + \sum_{k=1}^{K} \frac{\Gamma_k}{\sqrt{k}} \le \widetilde{O}(|\mathcal{A}|K^{1/2+p}).$$

Equations (G.11)–(G.12), Theorem F.3, Proposition F.4, and $Y_{\max} = K^{1/4}$ give (G.27); the grid contribution is

$$Y_{\max} K s_{\mathrm{grid}} = \widetilde{O}(K^{1/2}).$$

Proposition F.6 gives (G.28). We display the harmless factor θ^{-1} to match the theorem's problem-dependent Slater convention; for normalized utility $\theta \le 1$. The exponential softmax residual under the stated temperature is lower order. □

G The $p = 0$ Hard Specialization

This appendix specializes the derived-certificate proof in Appendix F to uniform tertiary probing. It is included separately to make the clean return-policy rate transparent.

When $p = 0$, $\epsilon_k \equiv 1$, and hence

$$q_k = q'_k = q''_k = U(\mathcal{A}), \qquad \Gamma_k = |\mathcal{A}|, \qquad \Lambda_k = |\mathcal{A}|.$$

The algorithmic lower correction becomes

$$\ell_k = C_{\mathrm{cert}} \sqrt{\frac{|\mathcal{A}| \ln(|\mathcal{M}|k/\delta)}{k(1-\gamma)}}, \tag{H.1}$$

and Lemma F.1 gives

$$\underline{\mathcal{U}}_k(\pi) \le V_{P^\star,g}^{\pi}(s_1) \le \underline{\mathcal{U}}_k(\pi) + w_k(\pi)$$

for every queried policy. The algorithm uses only the first inequality; the second is summed only for the returned sequence.

Lemma G.1 ($p = 0$ selected correction and width)**.** *On the representation-learning good event,*

$$\sum_{k=1}^{K} \ell_k \le \widetilde{O}\left(\sqrt{\frac{|\mathcal{A}|K}{1-\gamma}}\right),$$

and

$$\sum_{k=1}^{K} w_k(\pi_k) \le \widetilde{O}\left(\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right]\sqrt{K}\right). \tag{H.2}$$

Proof. The first display follows by summing $\ell_k = \widetilde{O}(k^{-1/2})$. The second follows from Lemma F.2 with $\Gamma_k = \Lambda_k = |\mathcal{A}|$. □

Proposition G.2 ($p = 0$ certificate reduction)**.** *Under the hypotheses of Proposition F.4,*

$$\mathrm{Regret}^{\mathrm{return}}(K) \le \widetilde{O}\left(\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right] K^{3/4}\right), \tag{H.3}$$

and hard violation is supported only on cap episodes.

Proof. For $p = 0$, the hard front end in Theorem F.3 is $\widetilde{O}(|\mathcal{A}|\bar{b}d^2K^{3/4}/(1-\gamma)^2)$. Apply Proposition F.4 and Lemma G.1. With $\eta_h = \Theta(\tau_h(1-\gamma)/K^{3/4})$, the grid term is

$$Y_{\max} K s_{\mathrm{grid}} = \widetilde{O}(K^{1/2}),$$

and is lower order. □

Theorem G.3 (Slater-induced hard violation for $p = 0$)**.** *Suppose Assumptions 2.3, 2.5, and C.1 hold, $|\mathcal{M}| < \infty$, $P^\star \in \mathcal{M}$, and the representation-learning good event holds. Then Algorithm 2 with $\epsilon_k \equiv 1$ satisfies*

$$\mathrm{Violation}_{\mathrm{hard}}^{\mathrm{return}}(K) \le \widetilde{O}\left(\frac{1}{\theta}\left[\frac{|\mathcal{A}|\bar{b}d^2}{(1-\gamma)^2} + \sqrt{\frac{|\mathcal{A}|}{1-\gamma}}\right]\sqrt{K}\right) + O(\theta^{-4}). \tag{H.4}$$

Proof. Break episodes are feasible by the one-sided inequality in Lemma F.1. At a cap episode, the returned policy is the actually queried $Y_{\max}$ policy, so Lemma F.5 applies and its overestimation terms are summable along the returned sequence. Proposition F.6 and Lemma G.1 give (H.4). If $K^{1/4}\theta < 1$, the trivial $K \le \theta^{-4}$ bound accounts for the last term. □

H Experiments

H.1 Job scheduling.

To validate the convergence of REP-PD-TEG in CMDPs, we first evaluate it in a simulated job scheduling environment. This toy domain lets us visualize how the learned representation, optimistic bonus, and primal-dual update interact under unknown dynamics.

Following [17, 18], we design a constrained MDP with state space $\mathcal{S} = \{0, 1, \ldots, 9\}$, where s represents the number of pending jobs. The agent chooses either to withhold jobs ($a = 0$) or submit them ($a = 1$) over a horizon $H = 10$. The transition dynamics follow:

$$s_{h+1} = \begin{cases} \max\{s_h - 2a_h, 0\} & \text{with probability } 0.8 \\ \max\{s_h - a_h, 0\} & \text{with probability } 0.1 \\ s_h & \text{otherwise.} \end{cases}$$

We assume that at time steps from 3 to 6, the reward is $1-0.9a$, while at all other time steps the reward is $1 - 0.2a$. This mimics a system in which job processing becomes temporarily expensive. The utility function $g(s, a, s') = \frac{s-s'}{2}$ quantifies job-processing efficiency, with per-episode constraint $\sum_{h=0}^{H-1} g(s_h, a_h, s_{h+1}) \ge b = 2$.

We do not hand-craft transition features in this domain. Instead, the appendix implementation learns latent transition factors $\hat{\phi}(s, a)$ and $\hat{\mu}(s')$ directly from the collected transitions using one-hot encodings of the discrete states and actions. After each episode, the newly collected transitions are appended to the dataset, the representation learner is updated on the accumulated data for multiple gradient steps, and the primal-dual policy update is then executed with the induced exploration bonus.

To demonstrate the effectiveness of our constraint-aware design, we compare against REP-UCB [41], an unconstrained low-rank MDP approach using similar representation learning but without dual-variable updates, and a linear method [17] where the transition feature is known. As shown in Figure 1, REP-PD-TEG has higher regret than REP-UCB on this toy instance but substantially lower cumulative constraint violation, illustrating the tradeoff induced by constraint-aware learning. Despite not knowing the transition feature, REP-PD-TEG is comparable in scale to the known-feature linear method of Ghosh et al. [17] on both metrics. Because this illustrative plot reports a single run without uncertainty bands, we do not infer statistical significance or an empirical asymptotic rate from it.

H.2 Practical Neural Implementation

Neural Safety Gym implementation. To probe scalability beyond the toy domain, we instantiate a practical neural variant, denoted **REP-PD-NN** for neural REP-PD-TEG. The theoretical algorithm uses a per-round full-batch MLE representation-learning step, which is difficult to scale directly to continuous-control domains. REP-PD-NN therefore replaces this step with a mini-batch SPEDER-style stochastic representation objective [37], while preserving the primal-dual safe RL structure. The learned state-action representation is used as the input to neural reward and cost critics, and the policy is trained with a Lagrangian-style constrained policy update. We evaluate this variant on Safety Gym tasks [35].

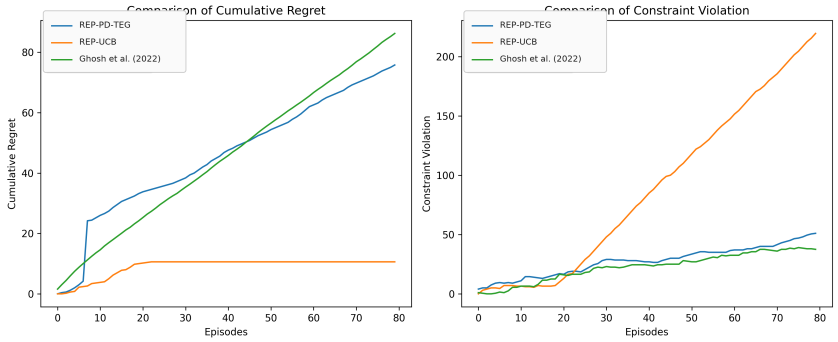

Figure 1: Illustrative single-run comparison of REP-PD-TEG, REP-UCB, and the known-feature linear method of Ghosh et al. [17] on the job-scheduling CMDP.

Algorithm 3 REP-PD-NN

Input: CMDP environment $\mathcal{E}$, cost budget d_{cost}, replay buffer $\mathcal{D}$, warm-up steps W, gradient steps per update G, dual stepsize ρ.
Parameters: representation η, actor θ, reward critics $\{\psi_i\}_{i=1}^2$, cost critics $\{\omega_i\}_{i=1}^2$, multiplier λ.
Initialize target networks $\bar{\eta} \leftarrow \eta$, $\bar{\psi}_i \leftarrow \psi_i$, and $\bar{\omega}_i \leftarrow \omega_i$.
Collect W transitions with a random or stochastic policy and store them in $\mathcal{D}$.
for $t = 1, \ldots, T$ **do**
  Sample $a_t \sim \pi_\theta(\cdot|s_t)$, execute a_t in $\mathcal{E}$, and store $(s_t, a_t, r_t, c_t, s_{t+1}, m_t)$ in $\mathcal{D}$.
  if an episode terminates with cumulative cost $\widehat{J}_c$ **then**
    Update $\lambda \leftarrow [\lambda + \rho(\widehat{J}_c - d_{\text{cost}})]_+$.
  end if
  for $g = 1, \ldots, G$ **do**
    Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$.
    Update the representation network by minimizing $\mathcal{J}_{\text{rep}}(\eta)$ on $\mathcal{B}$.
    Encode current pairs with $\phi_\eta(s, a)$ and target pairs with $\phi_{\bar{\eta}}(s', a')$.
    Update reward critics by minimizing the SAC reward Bellman loss.
    Update cost critics by minimizing the cost Bellman loss.
    Update the actor by minimizing $\mathcal{J}_\pi(\theta)$ with the current multiplier λ.
    Optionally update the entropy temperature α.
    Soft-update target networks $\bar{\eta}$, $\{\bar{\psi}_i\}_{i=1}^2$, and $\{\bar{\omega}_i\}_{i=1}^2$.
  end for
end for
Return policy π_θ and representation ϕ_η.

Let $\phi_\eta(s, a)$ denote the learned state-action representation, and $J_{\text{rep}}(\eta)$ denote the objective of representation learning. The actor is trained with the Lagrangian objective

$$\mathcal{J}_\pi(\theta) = \mathbb{E}\Big[\alpha \log \pi_\theta(a|s) - \min_i Q^r_{\psi_i}(\phi_\eta(s, a)) + \lambda \min_i Q^c_{\omega_i}(\phi_\eta(s, a))\Big],$$

where λ is the Lagrange multiplier for the cost constraint.

Raw-state baseline. For the raw-state baseline, we use the same Lagrangian training skeleton but remove the representation update. The critics and actor receive raw state-action pairs (s, a) instead of $\phi_\eta(s, a)$. Thus, any performance difference is attributable to the learned representation rather than to a different safe RL backbone.

Table 3 reports the completed positive results after 1M interaction steps and 3 random seeds, using the final deterministic evaluation checkpoint (`eval_return` / `eval_cost`). Across the three reported tasks, **REP-PD-NN** substantially improves final evaluation return and also reduces final evaluation cost relative to the raw baseline. These results indicate that a learned representation can improve both task performance and safety on a meaningful subset of continuous-control safe RL tasks.

Table 3: Safety Gym results for the practical neural implementation **REP-PD-NN**. Metrics are the mean and standard deviation over 3 seeds at the final deterministic evaluation checkpoint after 1M environment steps.

Environment	Raw Ret.	Raw Cost	REP-PD-NN Ret.	REP-PD-NN Cost
PointGoal1	-1.38 ± 0.58	27.85 ± 18.99	$\mathbf{-0.14 \pm 0.03}$	$\mathbf{16.45 \pm 12.69}$
PointPush1	-1.40 ± 0.60	32.17 ± 14.66	$\mathbf{-0.17 \pm 0.12}$	$\mathbf{12.13 \pm 6.95}$
PointPush2	-1.40 ± 0.71	13.00 ± 8.17	$\mathbf{-0.03 \pm 0.01}$	$\mathbf{10.90 \pm 3.85}$

Table 4: Approximate compute required to reproduce the reported experiments. Runtime denotes wall-clock time for one run. Total compute is over the reported methods and seeds.

Experiment	Worker	Steps	Runs	Time/run	Total compute
Job scheduling	CPU ($<$ 1GB RAM)	–	1	$<$ 1 min	$<$ 1 CPU-hour
Safety Gym PointGoal1	1 RTX 4090 (24GB)	10^6	6	8–20 h	48–120 GPU-hours
Safety Gym PointPush1	1 RTX 4090 (24GB)	10^6	6	8–20 h	48–120 GPU-hours
Safety Gym PointPush2	1 RTX 4090 (24GB)	10^6	6	8–20 h	48–120 GPU-hours

H.3 Computational Resources

All experiments were run on a local institutional server equipped with eight NVIDIA RTX 4090 GPUs, each with 24GB of memory, and standard x86_64 CPUs. The tabular job-scheduling experiment is lightweight and can be reproduced on a CPU-only machine. Each Safety Gym training run uses one GPU; the eight-GPU server was used only to run independent seeds and methods in parallel, and is not required for reproducing any single run. The implementation uses Python with PyTorch, CUDA-enabled GPU execution, MuJoCo, and Safety Gym.

Table 4 summarizes the approximate compute required to reproduce the reported experiments. For Safety Gym, each environment is evaluated with two methods, `raw` and `REP-PD-NN`, and three random seeds, resulting in six runs per environment and eighteen final training runs in total.

Thus, the reported Safety Gym results require approximately 144–360 GPU-hours in total. When all eight GPUs are used in parallel, the final runs can be completed in roughly 18–45 wall-clock hours. During method development, we also ran preliminary hyperparameter sweeps and diagnostic experiments that are not required to reproduce the reported tables. These preliminary runs used additional GPU time of the same order as the final reported experiment set; failed or preliminary runs were excluded from the reported results.

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: See Introduction.

   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: See Conclusion.

   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: See Section 3 and Section 4.

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: See Appendix H.

   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: [Yes]

Justification: We release code of our experiment and it can reproduce experimental results.

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: See Appendix H.

   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: See Appendix H.

   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 H.3 reports the compute worker type, memory, per-run runtime, total compute for the reported runs, and additional preliminary compute used during method development.

   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 is theoretical and simulation-based, does not involve human subjects, and does not use private or sensitive personal data. The submission is anonymized in accordance with conference policy.

   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: There is no societal impact of the work performed. The contribution is foundational and focused on theory and controlled simulations.

    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 poses no such risks.

    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: [N/A]

    Justification: The paper does not use existing assets.

    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.

    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 nor research with human subjects.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

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.