Offline–Online Reinforcement Learning for Linear Mixture MDPs

Zhongjun Zhang
Northwestern University
zhongjun@u.northwestern.edu

Sean R. Sinclair
Northwestern University
sean.sinclair@northwestern.edu

Abstract

We study offline–online reinforcement learning in linear mixture Markov decision processes (MDPs) under environment shift. In the offline phase, data are collected by an unknown behavior policy and may come from a mismatched environment, while in the online phase the learner interacts with the target environment. We propose an algorithm that adaptively leverages offline data. When the offline data are informative, either due to sufficient coverage or small environment shift, the algorithm provably improves over purely online learning. When the offline data are uninformative, it safely ignores them and matches the online-only performance. We establish regret upper bounds that explicitly characterize when offline data are informative, together with nearly matching lower bounds. Numerical experiments further corroborate our theoretical findings.

1 Introduction

Recent progress in reinforcement learning (RL) has demonstrated strong performance in complex domains such as large-scale video games [46, 37, 52]. However, these successes rely on massive amounts of online interaction, often enabled by high-fidelity simulators. In many real-world applications, such interaction is limited or simulators are unavailable. For example, in inventory planning for new products, decisions must be made under uncertain demand without sufficient historical data to simulate future outcomes, making it difficult to optimize policies through online experimentation alone. A natural approach is to leverage pre-collected data (i.e., offline data) from related environments. In retail, for instance, historical data from existing products can inform decisions for new products with limited data [5]. Despite its promise, leveraging offline data is challenging due to two fundamental issues. First, offline data may be generated under a similar but not identical environment, leading to bias under environment shift [36]. Second, the offline data is collected under a behavior policy, which shapes the observed offline data that may not align with the online target objective [34].

While both purely online and offline RL are well understood for many canonical models, such as bandits and tabular MDPs, leveraging offline data in online settings (i.e., offline–online learning) remains less explored. Existing work on offline–online learning often relies on simplifying assumptions on either environment shift or the offline behavior policy. Some works assume no environment shift [28, 53], while others impose restrictive assumptions on the offline behavior policy, such as fixed sampling in bandits [21] or uniform visitation in tabular MDPs [20]. These assumptions obscure the joint role of environment shift and the behavior policy in the problem, and do not extend naturally to large or continuous state–action spaces. Linear mixture MDPs provide a natural framework to study this gap, capturing rich function approximation while retaining tractable structure [19, 67]. Notably, the inventory example can be modeled as a special case of a linear mixture MDP [54].

Addressing these challenges is nontrivial. Naively combining offline and online data without accounting for environment shift can introduce bias and lead to linear regret [48]. This motivates the need for algorithms that *safely* incorporate offline data, achieving improved regret when the offline data is

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

informative, while matching purely online guarantees otherwise. At the same time, the presence of environment shift complicates the coverage conditions and pessimism-based estimation techniques commonly used in offline RL [34, 61, 42], suggesting that these principles must be reconsidered in the offline–online setting. Motivated by these challenges, we investigate the following:

How can offline data be safely leveraged in online learning under environment shift? How do environment shift and the offline behavior policy jointly govern the informativeness of offline data?

1.1 Our contributions

Offline–online learning algorithm to safely leverage offline data. Our first contribution is an algorithm, O-O UCRL-VTR (Algorithm 1), for offline–online learning in linear mixture MDPs under environment shift, built upon the backbone of the UCRL-VTR algorithm in Jia et al. [29]. The key difference is that in the online phase we maintain two estimators of the true parameter θ^* that governs the online transition dynamics: an *online-only* estimator based solely on online interactions, and an *offline–online* estimator that combines both offline data and online interactions. Our algorithm adaptively relies on the more reliable estimator, leveraging offline data only when it is informative. We establish a regret upper bound, which depends on a quantity τ that measures the coverage of the offline behavior policy. τ is defined by the minimum eigenvalue of the offline design matrix, and is closely related to uniform coverage conditions in the offline RL literature [61]. The regret bound in Theorem 4.2 exhibits a "min-of-two" structure: it is always safeguarded by the regret of the purely online algorithm [29], and can be strictly smaller when the offline data is informative. We note that similar min-of-two structures have been established in the bandit literature [21, 64]. However, in linear mixture MDPs one needs to determine how the offline data can be leveraged to obtain estimators of the underlying unknown parameter, which is absent from standard bandits. The information extracted from each offline data is thus an important design choice (see Section 4.3), allowing the learner to adjust how individual offline observations contribute to the coverage of the feature space, and is closely related to estimator design in offline RL [17].

Characterization of offline data informativeness and coverage τ. In Section 4.2 we provide (sufficient) conditions under which offline data is informative, i.e., yields $o(\sqrt{K})$ regret in the online learning process. We first consider the "sufficient coverage" case $\tau = \Theta(1)$. In this regime, offline data is informative if: (i) the offline sample size M^{off} is sufficiently large relative to the online interaction budget K; and (ii) the environment shift Δ is below a threshold determined by K. We further show in Section 4.3 that $\tau = \Theta(1)$ holds under the conventional uniform coverage condition in the offline RL literature [42] under appropriate design of offline value function estimates. Notably, the offline value function estimates we propose (see Proposition 4.10 and Proposition 4.12) differ from conventional pessimism-based principles in the offline RL literature, as the optimal policies in the offline and online environments can differ substantially due to environment shift. We next consider the complementary case $\tau = o(1)$. In this case, offline data is informative if: (i) τ exceeds a threshold determined by K and M^{off}; and (ii) the ratio Δ/τ is below a threshold determined by K. Condition (ii) implies a *linear* relationship between Δ and τ, so that larger τ permits proportionally larger environment shift. This regime captures a less explored setting in the offline RL literature, where the offline behavior policy is *adaptive* [42]. We provide a bandit example in Section C.6 showing that offline data can remain informative when the behavior policy follows UCB1 [6].

Nearly matching regret lower bound. We provide a nearly matching lower bound in Section 4.4. The lower bound considers the case when $\tau = \Theta(1)$ and exhibits a similar "min-of-two" structure, and nearly matches the upper bound in Theorem 4.2 in terms of d, M^{off}, K and Δ, in both regimes where offline data is informative and where it is not. A discussion of the regret lower bound for the complementary case $\tau = o(1)$ is provided in Section C.7.

1.2 Related literature

Online RL. The classical model for online RL is the tabular setting. For tabular settings, two main algorithmic paradigms dominate: model-based methods [15, 10, 7, 23, 47] and Q-learning-based approaches [32]. Alternative perspectives have also been explored, including posterior sampling [2] and policy-level optimization under structural assumptions such as convexity or counterfactual estimability [3, 65]. Beyond the tabular regime, increasing attention has been devoted to function approximation in continuous state and action spaces [24, 26, 68]. Under linear realizability assumptions,

much of the recent progress concerns linear MDPs [60, 33, 62], linear mixture MDPs [38, 29, 67, 9], and analyses based on controlling Bellman errors [31, 63].

This work builds on online RL for linear mixture MDPs with an explicit incorporation of an offline data phase. In the absence of offline data, our regret bound reduces to $\tilde{O}(dH^2\sqrt{K})$, matching the guarantees in Jia et al. [29] and Ayoub et al. [9]. A tighter bound of $\tilde{O}(dH^{3/2}\sqrt{K})$ is obtained by Zhou et al. [67] via a Bernstein-style concentration argument, at the cost of a substantially more involved analysis. We leave such refinements in our model to future work.

Offline RL. In offline RL the learner cannot interact with the environment and must rely entirely on a fixed, pre-collected dataset. As a result, coverage plays a central role in offline RL. Early work focused on *uniform* coverage, requiring the logged data to adequately explore the entire state–action space [39, 18]. More recent studies shift to *partial* coverage, examining whether the behavior policy sufficiently covers the optimal policy's state–action distribution [34, 57, 42, 30]. In structured models such as linear MDPs and linear mixture MDPs, coverage is also characterized in the feature space via eigenvalue conditions on the design matrix, as in Yin et al. [61] and Uehara and Sun [50].

Among the existing literature, the works most closely related to our setting include Yin et al. [61], Uehara and Sun [50], Nguyen-Tang et al. [42]. Yin et al. [61] study linear MDPs under a uniform coverage assumption in the feature space, formalized via a uniformly positive minimum eigenvalue of the design matrix. Their coverage condition is closely related to ours as in Definition 4.1. Uehara and Sun [50] and Nguyen-Tang et al. [42] consider offline RL for linear mixture MDPs under partial coverage, with the former focusing on data-space coverage and the latter on feature-space coverage. A detailed comparison between our coverage condition and those in the literature is given in Section B.3. All of these literature achieve a suboptimality gap of order $\tilde{O}(1/\sqrt{M^{\text{off}}})$, where M^{off} denotes the offline sample size. This rate is recovered as a special case of our results when $K = 1$ and the environment shift is negligible.

Offline–online RL. Building on the extensive literature on online and offline RL, offline–online RL has recently emerged as an active research direction. The central question is whether and how logged offline data can be leveraged to improve online learning. Beyond offline policy coverage, environment shift also becomes a key factor. When there is no environment shift and the offline and online environments coincide, it is well understood in both RL and bandits that offline data can be leveraged to reduce both the bias and the uncertainty in value and model estimation [28, 53, 45, 12, 4, 44, 48].

In the presence of environment shift, the offline and online environments, along with their corresponding optimal policies, may differ. Cheung and Lyu [21], Yang et al. [59] analyze offline–online bandit model under environment shift, but assume fixed numbers of offline pulls for each arm, so the offline behavior policy and coverage do not play an explicit role. In tabular RL, Chen et al. [20] study the offline–online setting, but again neither the offline behavior policy nor coverage is explicitly characterized. In contrast, our work extends to linear mixture MDPs, which subsume both bandits and tabular MDPs (see Section B), and proposes an algorithm whose performance is always safeguarded by a purely online baseline. Moreover, we provide a detailed characterization of how environment shift and coverage interact to determine when offline data is informative.

Related lines of work also include settings where the offline behavior policy can be queried during the online phase [43, 58], as well as non-stationary RL and change-point detection [56, 8]. There is also a growing body of empirical studies on offline–online RL [11, 40, 66]. Specifically, Zheng et al. [66] study an offline–online RL setting and propose an algorithm that combines pessimism-based estimation in the offline phase with optimistic exploration in the online phase. In contrast, our analysis indicates that our proposed value function estimators for the offline phase (see Proposition 4.10, Proposition 4.12, and Figure 2b) can outperform those based on pessimism in our setting.

2 Preliminary

Offline–online MDP. We study an RL problem consisting of two phases: an offline phase and an online phase. Each phase is modeled as an episodic MDP. Specifically, the offline MDP is given by the tuple $(\mathcal{S}, \mathcal{A}, P^{\text{off}}, r^{\text{off}}, H, s_1^{\text{off}})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P^{off} is the transition kernel, $r^{\text{off}} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic known reward function, H is the episode length, and s_1^{off} is the initial state. The online MDP is defined analogously by $(\mathcal{S}, \mathcal{A}, P, r, H, s_1)$. We assume that the state space $\mathcal{S}$, action space $\mathcal{A}$, and horizon H are identical in the offline and online

MDPs. We refer to the pair $\{(\mathcal{S}, \mathcal{A}, P^{\text{off}}, r^{\text{off}}, H, s_1^{\text{off}}), (\mathcal{S}, \mathcal{A}, P, r, H, s_1)\}$ as the corresponding *offline–online MDP*. We restrict attention to homogeneous linear mixture MDPs [29, 67, 42]:

Definition 2.1. There exist unknown vectors $\theta^{\text{off},*}, \theta^* \in \mathbb{R}^d$ such that $\forall s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$P^{\text{off}}(s' \mid s, a) = \langle \theta^{\text{off},*}, \phi(s' \mid s, a) \rangle, \qquad P(s' \mid s, a) = \langle \theta^*, \phi(s' \mid s, a) \rangle, \tag{1}$$

where $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is a known feature mapping.[1]

We assume that the feature mapping satisfies $\sum_{s'} |\phi_j(s' \mid s, a)| \le 1$, for all (s, a) and $j \in [d]$, where $\phi(s' \mid s, a) = [\phi_1(s' \mid s, a), \ldots, \phi_d(s' \mid s, a)]^\top$. Moreover, we assume that $\max\{\|\theta^{\text{off},*}\|_2, \|\theta^*\|_2\} \le B$, for a known constant $B > 0$ [29]. We define $\phi(\cdot \mid s, a) = [\phi(s' \mid s, a)]_{s' \in \mathcal{S}} \in \mathbb{R}^{d \times |\mathcal{S}|}$ and, for each $j \in [d]$, $\phi_j(\cdot \mid s, a) = [\phi_j(s' \mid s, a)]_{s' \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$.

Remark 2.2. Any tabular MDP can be expressed as a linear mixture MDP with dimension $d = |\mathcal{S}|^2|\mathcal{A}|$. To see this, we let $\phi(s' \mid s, a) = e_{(s',s,a)}$ be a one-hot encoding of the tuple (s', s, a). Then if $\theta_{(s',s,a)} = P(s' \mid s, a)$ we trivially have that $P(s' \mid s, a) = \theta^\top \phi(s' \mid s, a)$.

Finally, we assume the discrepancy between θ^* and $\theta^{\text{off},*}$ is *known* and bounded, as is standard in offline–online learning settings [22, 20, 14, 21]. See Section B (comparisons with environment shift assumptions in the literature) and Section C.2 (the case where Δ is unknown or misspecified) for more details.

Assumption 2.3. $\|\theta^* - \theta^{\text{off},*}\|_2 \le \Delta$, where $\Delta \ge 0$ is a known constant.

Offline data and behavior policy. Learning begins with an offline phase. In this phase, a total number of M^{off} trajectories are collected from the offline environment. The trajectories are generated sequentially under a sequence of behavior policies $\{\pi_m^{\text{off}}\}_{m=1}^{M^{\text{off}}}$. For each $m \in [M^{\text{off}}]$, the behavior policy $\pi_m^{\text{off}} = \{\pi_{m,h}^{\text{off}}\}_{h=1}^H$ consists of stage-wise mappings $\pi_{m,h}^{\text{off}} : \mathcal{S} \to \mathcal{A}$. Under the behavior policy π_m^{off}, the environment generates a length-H trajectory $\mathcal{H}_m^{\text{off}}$ according to the offline transition dynamics P^{off}, starting from the initial state s_1^{off}:

$$\mathcal{H}_m^{\text{off}} = \big\{(s_{m,1}^{\text{off}}, a_{m,1}^{\text{off}}, r^{\text{off}}(s_{m,1}^{\text{off}}, a_{m,1}^{\text{off}})), \ldots, (s_{m,H}^{\text{off}}, a_{m,H}^{\text{off}}, r^{\text{off}}(s_{m,H}^{\text{off}}, a_{m,H}^{\text{off}}))\big\},$$

where at stage h, upon observing state $s_{m,h}^{\text{off}}$, the action is $a_{m,h}^{\text{off}} = \pi_{m,h}^{\text{off}}(s_{m,h}^{\text{off}})$, the received reward is $r^{\text{off}}(s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}})$, and the next state satisfies $s_{m,h+1}^{\text{off}} \sim P^{\text{off}}(\cdot \mid s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}})$. The resulting offline dataset is $\mathcal{D}_{\text{off}} = \{\mathcal{H}_m^{\text{off}}\}_{m=1}^{M^{\text{off}}}$. The index m in $\{\pi_m^{\text{off}}\}_{m=1}^{M^{\text{off}}}$ indicates that the behavior policies may be adaptive during the offline phase [42]; that is, the policy π_m^{off} may vary across m.

Online interaction and regret. After observing the offline dataset $\mathcal{D}_{\text{off}}$, the learner interacts with the online MDP for K episodes, each of horizon H, starting from the designated initial state s_1. At the beginning of episode k, the learner selects a deterministic policy π_k based on $\mathcal{D}_{\text{off}}$ and all past online observations, and executes π_k throughout the episode. Similar to the offline behavior policies, the policy $\pi_k = \{\pi_{k,h}\}_{h=1}^H$ consists of stage-wise mappings $\pi_{k,h} : \mathcal{S} \to \mathcal{A}$. At stage h, upon observing state $s_{k,h}^{\text{on}}$, the learner takes action $a_{k,h}^{\text{on}} = \pi_{k,h}(s_{k,h}^{\text{on}})$, receives reward $r(s_{k,h}^{\text{on}}, a_{k,h}^{\text{on}})$, and transitions to $s_{k,h+1}^{\text{on}} \sim P(\cdot \mid s_{k,h}^{\text{on}}, a_{k,h}^{\text{on}})$. For any policy $\pi = \{\pi_h\}_{h=1}^H$, the Q functions and value functions in the online phase are defined as

$$Q_h^\pi(s, a) = \mathbb{E}_\pi\Big[\sum_{i=h}^H r(s_i, a_i) \;\Big|\; s_h = s, a_h = a\Big], \qquad V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)), \tag{2}$$

with the convention $V_{H+1}^\pi(s) = 0$ for all $s \in \mathcal{S}$. Let π^* denote an optimal policy for the online MDP. The performance metric is the (expected) regret over K episodes, defined as

$$\mathsf{Regret}(K) = \sum_{k=1}^K \mathbb{E}\Big[V_1^{\pi^*}(s_1) - V_1^{\pi_k}(s_1)\Big], \tag{3}$$

where the expectation is taken over both the offline data generation and the online interaction.

[1] We assume a shared feature mapping ϕ across offline and online MDPs. Our theoretical results extend naturally to settings with different offline and online feature mappings; the main results in Section 4, including the nearly matching regret bounds in Theorems 4.2 and 4.13, remain unchanged.

3 Algorithm design

This section presents the details of our algorithm (Algorithm 1). The algorithm consists of two components: estimating the online transition parameter θ^*, and using this estimate to compute an optimistic approximation of the optimal value function $V_1^{\pi^*}$ and the corresponding policy π^*. A central feature of Algorithm 1 is that θ^* can be estimated in two distinct ways. The first is an *online-only* estimator that uses data collected from online interaction alone, while the second is an *offline–online* estimator that aggregates both offline and online data. By appropriately leveraging these two estimators, the algorithm effectively uses the more reliable estimate at each stage.

Estimation of θ^*. We estimate θ^* via *value-targeted regression* [29]. We begin with the *online-only* estimator (line 11 of Algorithm 1). At episode k and stage h, define

$$x_{k,h}^{\mathrm{on}} = \phi(\cdot \mid s_{k,h}^{\mathrm{on}}, a_{k,h}^{\mathrm{on}})\widehat{V}_{k,h+1}, \qquad y_{k,h}^{\mathrm{on}} = \widehat{V}_{k,h+1}(s_{k,h+1}^{\mathrm{on}}), \tag{4}$$

where $\widehat{V}_{k,h+1} \in \mathbb{R}^{|\mathcal{S}|}$ is an estimate of the optimal value function $V_{h+1}^{\pi^*}$ (specified later in Eq. (10)). Under the linear mixture model, we have

$$\theta^{*\top} x_{k,h}^{\mathrm{on}} = \mathbb{E}\big[\widehat{V}_{k,h+1}(s_{k,h+1}^{\mathrm{on}}) \mid s_{k,h}^{\mathrm{on}}, a_{k,h}^{\mathrm{on}}\big], \tag{5}$$

and $y_{k,h}^{\mathrm{on}}$ is a corresponding realization. This yields a standard linear regression problem, allowing θ^* to be estimated via ridge regression:

$$\widehat{\theta}_k^{\mathrm{on}} = (M_k^{\mathrm{on}})^{-1} w_k^{\mathrm{on}}, \tag{6}$$

where $M_k^{\mathrm{on}} = \lambda I + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} x_{k',h}^{\mathrm{on}} \big(x_{k',h}^{\mathrm{on}}\big)^\top$, $w_k^{\mathrm{on}} = \sum_{k'=1}^{k-1}\sum_{h=1}^{H} x_{k',h}^{\mathrm{on}} y_{k',h}^{\mathrm{on}}$. Here $\lambda > 0$ is a ridge regularization parameter to be specified later. The regression in Eq. (5) holds for any non-anticipatory $\widehat{V}_{k,h+1}$, but the regret guarantee requires a specific data-adaptive optimistic choice (defined in Eq. (10)) to control error propagation from parameter estimation to value functions. In addition to the online-only estimator, we also construct an *offline–online* estimator that additionally incorporates offline data. This is enabled by Assumption 2.3, which bounds the discrepancy between θ^* and $\theta^{\mathrm{off},*}$. Specifically, for offline samples we define

$$x_{m,h}^{\mathrm{off}} = \phi(\cdot \mid s_{m,h}^{\mathrm{off}}, a_{m,h}^{\mathrm{off}})\widetilde{V}_{m,h+1}, \qquad y_{m,h}^{\mathrm{off}} = \widetilde{V}_{m,h+1}(s_{m,h+1}^{\mathrm{off}}), \tag{7}$$

where $\{\widetilde{V}_{m,h} \in \mathbb{R}^{|\mathcal{S}|}\}$ are auxiliary value functions. The offline–online estimator is then

$$\widehat{\theta}_k^{\mathrm{all}} = (M_k^{\mathrm{all}})^{-1} w_k^{\mathrm{all}}, \tag{8}$$

where $M_k^{\mathrm{all}} = M_k^{\mathrm{on}} + \sum_{m=1}^{M^{\mathrm{off}}}\sum_{h=1}^{H} x_{m,h}^{\mathrm{off}} \big(x_{m,h}^{\mathrm{off}}\big)^\top$, $w_k^{\mathrm{all}} = w_k^{\mathrm{on}} + \sum_{m=1}^{M^{\mathrm{off}}}\sum_{h=1}^{H} x_{m,h}^{\mathrm{off}} y_{m,h}^{\mathrm{off}}$. Unlike the online phase, the offline phase does not require $\{\widetilde{V}_{m,h}\}$ to be chosen for regret minimization. This provides flexibility in designing $\{\widetilde{V}_{m,h}\}$ to optimize the convergence rate of $\|\widehat{\theta}_k^{\mathrm{all}} - \theta^*\|$. A detailed discussion of the principles behind the design of $\{\widetilde{V}_{m,h}\}$ is deferred to Section 4.3.

Optimistic planning. In each episode k, the algorithm follows a standard optimistic planning approach based on confidence set elimination. We construct two confidence sets for θ^*:

$$\mathcal{C}_k = \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_k^{\mathrm{on}}\|_{M_k^{\mathrm{on}}} \le \beta_k\Big\}, \qquad \mathcal{B}_k = \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_k^{\mathrm{all}}\|_{M_k^{\mathrm{all}}} \le \gamma_k\Big\}, \tag{9}$$

where β_k and γ_k are appropriately chosen confidence radii. The value function estimates $\widehat{V}_{k,h}$ are then computed via optimistic planning over

$$\mathcal{D}_k = \begin{cases} \mathcal{B}_k \cap \mathcal{C}_k, & \mathcal{B}_k \cap \mathcal{C}_k \neq \varnothing, \\ \mathcal{B}_k, & \mathcal{B}_k \cap \mathcal{C}_k = \varnothing. \end{cases}$$

For all $h \le H$, we define

$$\widehat{Q}_{k,h}(s,a) = r(s,a) + \max_{\theta \in \mathcal{D}_k} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}, \qquad \widehat{V}_{k,h}(s) = \max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s,a), \tag{10}$$

with terminal condition $\widehat{Q}_{k,H+1} \equiv 0$. [2] Optimizing over $\mathcal{C}_k \cap \mathcal{B}_k$ (with high probability, $\mathcal{C}_k \cap \mathcal{B}_k$ is

nonempty and hence $\mathcal{D}_k = \mathcal{C}_k \cap \mathcal{B}_k$; see Lemma D.1) is central to the algorithm. When offline data are informative, $\mathcal{B}_k$ tightens the confidence region and improves efficiency; when they are biased or uninformative, $\mathcal{C}_k$ safeguards performance, ensuring regret no worse than ignoring offline data. The implemented policy π_k is then obtained via a greedy search with respect to the optimistic value function estimates $\{\widehat{V}_{k,h}\}_{h=1}^H$ (line 17 of Algorithm 1).

4 Regret bounds and characterization of offline data informativeness

4.1 Regret upper bound

We first introduce a parameter τ that quantifies the coverage of the behavior policy. Define the offline design matrix $G_{\text{off}} = \sum_{m=1}^{M^{\text{off}}} \sum_{h=1}^{H} x_{m,h}^{\text{off}} \big(x_{m,h}^{\text{off}}\big)^\top$. Intuitively, $\lambda_{\min}(\lambda I + G_{\text{off}})$ characterizes the offline data informativeness in the least informative direction of the feature space $\mathbb{R}^d$.

Definition 4.1. For a fixed $\lambda > 0$, we define the δ**-effective coverage** as:

$$\tau(\delta) := \sup\left\{ z \geq 0 \;\middle|\; \mathbb{P}\Big(\frac{1}{M^{\text{off}}} \lambda_{\min}(\lambda I + G_{\text{off}}) \geq z\Big) \geq 1 - \delta \right\}, \qquad \delta \in (0,1). \tag{11}$$

For any $\delta \in (0,1)$, $\tau(\delta)$ depends on the choice of $\{\widetilde{V}_{m,h}\}$ and the offline behavior policy; we elaborate on this dependence in Section 4.3. Throughout the paper, we write $\tau = \tau\big(\frac{1}{2KM^{\text{off}}}\big)$ (so $\delta = 1/(2KM^{\text{off}})$) for notational convenience. Regardless of the offline data, one always has $\lambda_{\min}(\lambda I + G_{\text{off}}) \geq \lambda$, and hence $\tau \geq \lambda / M^{\text{off}}$ holds trivially. Since the offline regression vectors $x_{m,h}^{\text{off}}$ are uniformly upper bounded, τ is also upper bounded. Definition 4.1 is closely related to uniform coverage assumptions in offline RL [42, 25, 61] (see Section C.3). In Section C.3.2 we also show that conventional partial coverage conditions are insufficient in the offline–online setting. We next state our main regret upper bound; the proof is provided in Section D.1.

Theorem 4.2. *Under Assumptions 2.3 and Definition 4.1, let* $\lambda = H^2 d$ *and define* $\gamma_k = H\sqrt{2\log\Big(\frac{2\det(M_k^{\text{all}})^{1/2} K}{\det(\lambda I)^{1/2}}\Big)} + \Delta \frac{\lambda_{\max}(G_{\text{off}})}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} + \sqrt{\lambda} B$, $\beta_k = H\sqrt{2\log\Big(\frac{2\det(M_k^{\text{on}})^{1/2} K}{\det(\lambda I)^{1/2}}\Big)} + \sqrt{\lambda} B$. *Then, when running Algorithm 1, the expected regret* $\mathsf{Regret}(K)$ *is upper bounded by*

$$\tilde{O}\left(\min\left\{ \left[BH^2 d\sqrt{K} + H^4 d^{\frac{3}{2}} \Delta \sqrt{K} \sqrt{\frac{M^{\text{off}}}{\tau}} \right] \cdot \sqrt{\log\Big(1 + \frac{H^3 K}{\tau M^{\text{off}}}\Big)}, BH^2 d\sqrt{K} \right\} \right), \tag{12}$$

where $\tilde{O}(\cdot)$ *hides logarithmic factors in* K *and* M^{off}.

The regret bound is with a "min-of-two" form. The first term corresponds to the offline–online estimator (term (A)), while the second term (term (B)) corresponds to the online-only estimator.

4.2 When is offline data informative

The regret bound in Theorem 4.2 is upper bounded by term (B), $\tilde{O}(BH^2 d\sqrt{K})$, which matches the online-only regret guarantee [29]. This highlights the robustness of our algorithm. When the offline data are uninformative (due to either poor coverage τ or a large environment shift Δ), the algorithm provably performs no worse than ignoring offline data altogether. We next characterize regimes in which offline data is informative and yields a strict improvement over the online-only baseline.

Definition 4.3. Given $\epsilon \in (0, \frac{1}{2}]$, we say that the offline data is **informative** with respect to ϵ if the regret satisfies $\mathsf{Regret}(K) = \tilde{O}(BH^2 d K^{\frac{1}{2}-\epsilon})$.

Any such improvement must be realized through term (A). The contribution of offline data improves monotonically as Δ decreases and τ increases. If no nontrivial lower bound on τ can be established, namely when $\tau \leq \lambda / M^{\text{off}}$, then offline data becomes uninformative once K is sufficiently large. The following corollary gives a sufficient condition under which offline data is informative.

[2] We assume access to an optimization oracle that solves the maximization over the action space $\mathcal{A}$ in Eq. (10).

Corollary 4.4. *If there exists* $\epsilon \in (0, \frac{1}{2}]$ *such that* $\tau = \Omega\Big(H^3 \frac{K^{1+2\epsilon}}{M^{\text{off}}}\Big)$ *and* $\frac{\Delta}{\tau} = O\Big(\frac{B}{H^{7/2} d^{1/2} K^{\frac{1}{2}+\epsilon}}\Big)$, *then the offline data is informative with respect to* ϵ.

The proof is deferred to Section D.2. We next discuss how the coverage parameter τ and the environment shift Δ affect the regret bound, treating K and M^{off} as fixed.

Case 1: $\tau = \Theta(1)$. Here the behavior policy exhibits uniform coverage over the entire state–action space. Correspondingly, Corollary 4.4 simplifies to the conditions $M^{\text{off}} \gtrsim K^{1+2\epsilon}$ and $\Delta \lesssim K^{-1/2-\epsilon}$. That is, when M^{off} is suitably large, and Δ is suitably small, the offline data become informative. A noteworthy implication is that offline data need not be excessively large to yield substantial gains. For any $\epsilon \in (0, \frac{1}{2}]$, when $M^{\text{off}} = \Theta(K^2)$, whether offline data is informative depends solely on Δ. In particular, under a sufficiently small shift, e.g., $\Delta \lesssim K^{-1}$, the regret reduces to $\tilde{O}(\log(K))$.

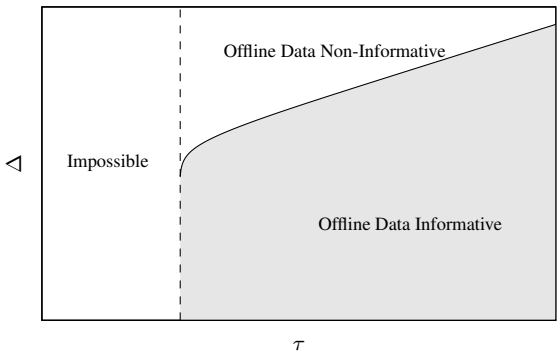


Figure 1: Graphical illustration of sufficient conditions for offline data informativeness from Eq. (12). For small τ, no Δ yields informative offline data. Beyond a threshold, larger τ admits larger Δ, with an approximately linear dependence; deviations near the threshold arise from logarithmic terms.

Case 2: $\tau = o(1)$. We now consider regimes where τ may decrease to 0 as M^{off} increases. Such behavior can arise when the behavior policy is adaptive [42, 33]. As M^{off} grows, the policy may concentrate more heavily on certain state–action pairs and reduce visits to others, thereby decreasing τ. The first condition in Corollary 4.4, $\tau = \Omega(H^3 K^{1+2\epsilon}/M^{\text{off}})$ imposes a minimum coverage requirement, i.e., τ must exceed a threshold. The second condition, $\Delta/\tau = O(B\big(H^{7/2} d^{1/2} K^{1/2+\epsilon}\big)^{-1})$, shows that the admissible magnitude of Δ scales *linearly* with τ (see Figure 1). Together, these conditions indicate that stronger effective coverage (i.e., a slower decay of τ with respect to M^{off}) linearly enlarges the range of environment shifts under which offline data is informative. A similar argument suggests an upper bound on the offline data size M^{off} that is sufficient to attain a $\tilde{O}(\log(K))$ regret, analogous to Case 1.

4.3 Dependence of τ on underlying factors

Other than Δ, Theorem 4.2 critically depends on τ. We now examine how τ depends on the feature mapping ϕ, the behavior policy, and $\{\widetilde{V}_{m,h}\}$. We first study a necessary condition on ϕ. We then consider the uniform data coverage on the behavior policy, under which we discuss the designs of $\{\widetilde{V}_{m,h}\}$ and derive lower bounds on τ. We defer the discussion of adaptive behavior policies to Section C.6. We assume throughout this subsection that state and action spaces are finite.

The feature mapping ϕ. We begin by isolating a structural property of the feature mapping ϕ that is necessary for τ to admit a nontrivial lower bound.

Definition 4.5. The feature mapping ϕ of the offline MDP is **uniformly learnable** if there exist L probe vectors $\{V_1, \ldots, V_L\}$ with $V_l \in [-1, 1]^{|\mathcal{S}|}$ for all $l \in [L]$, and state–action pairs $\{(s_i, a_i)\}_{i=1}^I$, such that $\lambda_{\min}\Big(\sum_{i=1}^I \sum_{l=1}^L \phi(\cdot \mid s_i, a_i) V_l (\phi(\cdot \mid s_i, a_i) V_l)^\top\Big) \geq \kappa$ for some constant $\kappa > 0$, where I and L are fixed positive integers.

Notably, the probe vectors $\{V_l\}_{l=1}^L$ are constructed from the known feature mapping ϕ. Thus, they are known to the learner. Also, uniform learnability does not depend on the behavior policy or $\{\widetilde{V}_{m,h}\}$. The following proposition shows that it is necessary for the offline data to be informative. The proof is deferred to Section D.3.

Proposition 4.6. *If* ϕ *is not uniformly learnable, then for any behavior policy and any design of* $\{\widetilde{V}_{m,h}\}$, *we have* $\tau(\delta) = \frac{\lambda}{M^{\text{off}}}$ *for any* $\delta \in (0, 1)$. *In this case, the offline data is uninformative.*

Remark 4.7. The one-hot feature mapping for any offline–online tabular MDP (see Remark 2.2) satisfies uniform learnability with $I = |\mathcal{S}||\mathcal{A}|$, $L = |\mathcal{S}|$, $V_{s'} = e_{s'}$ for all $s' \in \mathcal{S}$, and $\kappa = 1$, where $e_{s'} \in \mathbb{R}^{|\mathcal{S}|}$ denotes the canonical basis vector corresponding to state s' in the one-hot representation. In this case, for any (s, a) we have $\phi(\cdot \mid s, a) V_{s'} V_{s'}^\top \phi(\cdot \mid s, a)^\top = e_{s'} e_{s'}^\top$.

Offline behavior policy. Even if uniform learnability holds, the behavior policy must visit the specific (s, a) pairs in Definition 4.5 to obtain a nontrivial lower bound on τ implied by Definition 4.5. Inspired by the offline RL literature [42, 34, 61, 50], we focus on the canonical uniform data coverage assumption [41, 42]. As briefly mentioned in Section 4.1, conventional partial coverage conditions are not suitable in our setting (see Section C.3). Formally, for any behavior policy π^{off}, define the marginal state–action probability measure at stage h in the offline MDP as $d_h^{\pi^{\text{off}}}(s, a) := \mathbb{P}(s_h = s, a_h = a)$, where the probability is taken with respect to the trajectory distribution induced by policy π^{off} and the offline environment dynamics. The uniform data coverage assumption is stated as follows.

Assumption 4.8. The offline behavior policy is non-adaptive; that is, there exists a policy π^{off} such that $\pi_m^{\text{off}} = \pi^{\text{off}}$ for all $m \in [M^{\text{off}}]$. Moreover, the probability of visiting any (h, s, a) triple is lower bounded by $p_0 > 0$, i.e., $\inf_{h,s,a} d_h^{\pi^{\text{off}}}(s, a) \geq p_0$.

Design of $\{\widetilde{V}_{m,h}\}$. Under uniform learnability and uniform data coverage, however, a well-designed $\{\widetilde{V}_{m,h}\}$ is *necessary* to achieve a nontrivial value of τ. We illustrate this via a counterexample. Consider an offline–online tabular MDP. Then uniform learnability is satisfied, and one can obtain $\tau = \Theta(1)$, as in Remark 4.7. However, if one sets $\widetilde{V}_{m,h} = \mathbf{1}_{|\mathcal{S}|}$ for all m, h, the resulting offline design matrix G_{off} is not full rank, yielding $\tau = \frac{\lambda}{M^{\text{off}}}$ regardless of the behavior policy.

To maximize τ, a natural principle is to select $\{\widetilde{V}_{m,h}\}$ so as to increase $\lambda_{\min}(G_{\text{off}})$. Formally, one may attempt to choose $\widetilde{V}_{m,h+1}$ sequentially and greedily according to

$$\arg\max_{V \in [-1,1]^{|\mathcal{S}|}} \lambda_{\min}\Big(\sum_{(m',h')<(m,h)} x_{m',h'}^{\text{off}}(\widetilde{V}_{m',h'+1}) \big(x_{m',h'}^{\text{off}}(\widetilde{V}_{m',h'+1})\big)^\top + x_{m,h}^{\text{off}}(V) \big(x_{m,h}^{\text{off}}(V)\big)^\top \Big) \quad (13)$$

for each realized sample path, where $x_{m,h}^{\text{off}}(V) = \phi(\cdot \mid s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}}) V$ and $\widetilde{V}_{m,H+1} = 0$ for all m. However, Eq. (13) is neither convex nor concave in V (see Section C.4). We propose two tractable alternatives, a stochastic design and a deterministic design of $\{\widetilde{V}_{m,h}\}$. In principle, the stochastic design aims to maximize $\lambda_{\min}(G_{\text{off}})$ by exploring the feature space in an "even" manner.

Definition 4.9. For each offline index (m, h), let $\xi_{m,h} = (\xi_{m,h,1}, \ldots, \xi_{m,h,L})$ be a vector of i.i.d. Rademacher random variables, i.e., $\mathbb{P}(\xi_{m,h,l} = 1) = \mathbb{P}(\xi_{m,h,l} = -1) = \frac{1}{2}$ for all $l \in [L]$, independent of the offline data. We define the stochastic design via $\widetilde{V}_{m,h} = \frac{1}{L} \sum_{l=1}^{L} \xi_{m,h,l} V_l$. By construction, $\widetilde{V}_{m,h} \in [-1, 1]^{|\mathcal{S}|}$ for all (m, h).

Under Definition 4.5, the L probe vectors span the entire feature space. Thus, for each visited (s, a) pair in the offline data, we construct $\widetilde{V}_{m,h}$ by randomly combining these probe vectors. We have:

Proposition 4.10. *Under uniform learnability (Definition 4.5) and uniform data coverage (Assumption 4.8), and under the stochastic design of $\{\widetilde{V}_{m,h}\}$ (Definition 4.9), for any $\delta > 0$, the effective coverage $\tau(\delta)$ satisfies $\tau(\delta) \geq \frac{p_0 \kappa}{L^2} - \sqrt{\frac{2d p_0 \kappa}{L^2 M^{\text{off}}} \log\big(\frac{d}{\delta}\big)}$. In particular, when $M^{\text{off}} \geq \frac{8dL^2}{p_0 \kappa} \log(2dKM^{\text{off}})$, we have $\tau = \tau\big(\frac{1}{2KM^{\text{off}}}\big) \geq \frac{p_0 \kappa}{2L^2}$.*

Next up we consider the tabular setting under the natural embedding in Remark 2.2. In this case, we can obtain a simpler approximation of Eq. (13) and provide a deterministic design of $\{\widetilde{V}_{m,h}\}$.

Definition 4.11. In the offline–online tabular MDP setting, the deterministic design of $\{\widetilde{V}_{m,h}\}$ is given by $\widetilde{V}_{m,h} = e_{s'}$, where $s' \in \arg\min_{x \in \mathcal{S}} N_{m,h}(x, s, a)$. Here, $N_{m,h}(x, s, a)$ denotes the number of times the triple (x, s, a) has been selected in the offline data up to (m, h).

In tabular setting, Definition 4.11 coincides with Eq. (13), except that the optimization is restricted to $V \in \{e_1, \ldots, e_{|\mathcal{S}|}\}$ rather than $V \in [-1, 1]^{|\mathcal{S}|}$. Thus the value of τ obtained via Definition 4.11 provides a lower bound. Similar to Proposition 4.10, the deterministic design implies that $\tau = \Theta(1)$.

Proposition 4.12. *In offline–online tabular MDP setting, under deterministic design of* $\{\widetilde{V}_{m,h}\}$ *(Definition 4.11) and Assumption 4.8, for any* $\delta > 0$*, we have* $\tau(\delta) \geq \frac{p_0}{|\mathcal{S}|} - \frac{1}{|\mathcal{S}|}\big(\frac{1}{2M^{\text{off}}} \log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})\big)^{1/2}$*. When* $M^{\text{off}} \geq \frac{2}{p_0^2} \log(2|\mathcal{S}||\mathcal{A}|KM^{\text{off}})$*, we have that* $\tau = \tau\big(\frac{1}{2KM^{\text{off}}}\big) \geq \frac{p_0}{2|\mathcal{S}|}$.

The proofs of Proposition 4.10 and Proposition 4.12 are in Section D.3. Both Proposition 4.10 and Proposition 4.12 imply that $\tau = \Theta(1)$. Thus Case 1 in Section 4.2 applies and under appropriate values of M^{off}, K, and Δ, the offline data could be informative. Our results recover several existing results in the literature (see Section B), including offline–online tabular and linear bandits [21, 64], offline–online tabular MDPs [20], and offline linear mixture MDPs [42].

We conclude this section by noting that our design of $\{\widetilde{V}_{m,h}\}$ and the resulting coverage τ need not be optimal; incorporating Δ into the estimator design, for example through sample weighting as in Chen et al. [20], may improve finite-sample performance. The intuition behind their weighting scheme is that a larger known environment shift Δ calls for placing less weight on offline observations relative to online ones. More broadly, optimal estimator design remains largely open in the offline–online setting, given that using all offline samples before online learning need not be strictly optimal even in bandits without environment shift [12].

4.4 Regret lower bound

We complement Theorem 4.2 with a nearly matching lower bound when $\tau = \Theta(1)$. The regime $\tau = o(1)$ yields a different lower bound, and we defer the discussion to Section C.7.

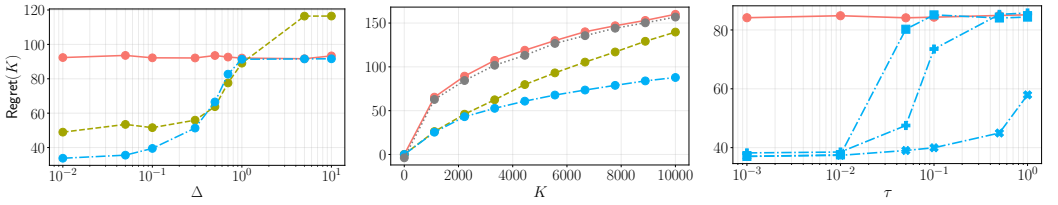
Theorem 4.13. *Under environment shift (Assumption 2.3) and coverage condition (Definition 4.1), for any* $\Delta, H \geq 2$*, and* $d \geq 2$*, for any sufficiently large* K *and* M^{off}*, and any offline–online algorithm, there exists an offline–online linear mixture MDP instance such that, whenever the offline data induces* $\tau = \Theta(1)$*, any algorithm has regret at least*

$$\Omega\left(\min\left\{dH\sqrt{K}, dH\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \frac{\Delta}{\sqrt{d}}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)\right\}\right). \quad (14)$$

The proof of Theorem 4.13 is in Section D.4. The lower bound in Eq. (14) exhibits a "min-of-two" structure, mirroring the upper bound. It matches the upper bound in Eq. (12) in the dependence on d and K when the online-only term $dH\sqrt{K}$ dominates, i.e., when the offline data is not informative. The gap in the dependence on H is consistent with literature [67, 29]. When the offline–online term dominates, i.e., when the offline data is informative, the condition $\tau = \Theta(1)$ ensures that $\lambda_{\min}(\mathbb{E}[G_{\text{off}}]) = \Theta(M^{\text{off}})$ (see Eq. (56) in Appendix), under which the lower bound matches Eq. (12) in the dependence on K, M^{off}, and Δ. The dependence on d is also nearly matched, up to a gap between $d^{\frac{3}{2}}$ and $\sqrt{d}$ in the Δ-dependent term. This gap, together with the gap in the dependence on H, stems from self-normalized concentration (Lemma E.1) in the upper bound and from measuring offline data through the worst-direction eigenvalue in Definition 4.1, which is consistent with the offline RL literature (see, e.g., Table 1 of Nguyen-Tang et al. [42]). It's worth mentioning that concurrent work on offline–online tabular MDPs achieves an exact minimax-optimal online safeguard under a similar min-of-two framework [55], suggesting that exact minimax regret guarantees may also be attainable in our model. However, under function approximation, even in the linear mixture models studied here, the design of offline estimators complicates the characterization of the value of offline data. Existing bandit and tabular MDP analyses therefore do not directly yield minimax guarantees for our model. We leave such refinements for future work.

5 Numerical simulations

We further support Section 4 with numerical experiments on a synthetic tabular MDP. We implement two versions of O-O UCRL-VTR: (i) O-O UCRL-VTR (Optimal), in which the design of $\{\widetilde{V}_{m,h}\}$ follows Definition 4.11; (ii) O-O UCRL-VTR (Pessimistic), in which a pessimism-based version of $\{\widetilde{V}_{m,h}\}$ is adopted following Algorithm 2 of Nguyen-Tang et al. [42]. We compare them with two baseline algorithms: UCRL [29] and COMPLETE (Algorithm 2), which closely follows Chen et al. [20], assuming $\Delta = 0$ and merges offline data into online learning. See the code base for



(a) Regret(K) versus Δ. (b) Cumulative regret over K. (c) Regret(K) versus τ for different $\Delta(\tau)$.

Figure 2: Comparison of UCRL, COMPLETE, and O-O UCRL-VTR under different environment shifts Δ, coverage τ, K, and M^{off}. The three O-O UCRL-VTR (Optimal) curves in Figure 2c correspond to different $\Delta(\tau)$ scaling regimes: the cross markers indicate linear scaling of Δ with τ, while the plus and square markers represent quadratic and cubic scaling, respectively. (subset of simulation results; full results and detailed hyperparameters are in Section F, see Figure 3)

implementation details. Additional comparisons with more baseline algorithms, along with further simulation results and details on the behavior policy, baseline implementations, and experimental settings, are provided in Section F. We also include an additional simulation of a linear mixture MDP with continuous state and action spaces in Section F.4.

Fig. 2a plots the cumulative regret as Δ increases. UCRL remains insensitive to Δ since it does not use offline data. When Δ is small, O-O UCRL-VTR outperforms UCRL, indicating that offline data is effectively leveraged. As Δ grows, the regret of O-O UCRL-VTR increases and approaches that of UCRL, demonstrating the safeguarding property in Theorem 4.2. Fig. 2b compares two constructions of $\{\widetilde{V}_{m,h}\}$. Both O-O UCRL-VTR (Optimal) and O-O UCRL-VTR (Pessimistic) outperform the purely online baseline UCRL. However, O-O UCRL-VTR (Optimal) consistently attains substantially lower regret than O-O UCRL-VTR (Pessimistic). Figure 2c compares O-O UCRL-VTR (Optimal) under three Δ–τ scaling regimes. When $\Delta(\tau) \propto \tau$, offline data remain informative even as Δ increases. When $\Delta(\tau) \propto \tau^2$ or τ^3, the benefit of offline data quickly vanishes once τ exceeds 0.1. These trends align with the dependence predicted by Corollary 4.4. In practical settings where larger environment shift is expected, one can proactively increase coverage (e.g., by broadening the behavior policy) so that τ scales accordingly. When coverage grows commensurately with the shift (τ need not be optimized to the maximum possible level, i.e., $\tau = \Theta(1)$), regret guarantees continue to improve and offline data remain informative.

6 Conclusion

In this work, we study offline–online learning for linear mixture MDPs under environment shift and an uncontrollable offline behavior policy. We develop an algorithm O-O UCRL-VTR for this setting, together with a regret upper bound and a nearly matching lower bound, and characterize when offline data is informative. Our results show that offline data is informative when the environment shift Δ is small and the coverage parameter τ is large, and that these two factors can compensate for each other. In particular, offline data can remain informative even under non-negligible environment shift, as long as Δ scales linearly with τ. We further disentangle the factors that determine τ, including the feature mapping of the linear mixture MDP, the offline behavior policy, and the design of offline value function estimates $\{\widetilde{V}_{m,h}\}$.

Several directions for future work arise from our results. One natural question is whether the interplay between environment shift and offline behavior policy extends to broader function approximation settings in RL. Another direction is to investigate empirical methods for estimating Δ when it is unknown, such as the cross-validation procedure in Chen et al. [20]. It would also be valuable to go beyond the scalar quantity τ and study the full spectrum of the offline design matrix, which may yield a more refined characterization of offline data informativeness and potentially unify the lower bounds across different coverage regimes.

Acknowledgments and Disclosure of Funding

The authors received no external funding for this work and declare no competing interests.

References

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

[2] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in neural information processing systems*, 30, 2017.

[3] Shipra Agrawal and Randy Jia. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Operations Research*, 70(3):1646–1664, 2022.

[4] Shubhada Agrawal, Sandeep Juneja, Karthikeyan Shanmugam, and Arun Sai Suggala. Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048*, 2023.

[5] Amazon.com. Solving some of the largest, most complex operations problems. https://www.amazon.science/latest-news/solving-some-of-the-largest-most-complex-operations-problems, 2022. accessed March 21, 2026.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[7] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[8] Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on learning theory*, pages 138–158. PMLR, 2019.

[9] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[10] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[11] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

[12] Siddhartha Banerjee, Sean R Sinclair, Milind Tambe, Lily Xu, and Christina Lee Yu. Artificial replay: a meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

[13] Xiaowei Bao, Nikolaos V Sahinidis, and Mohit Tawarmalani. Semidefinite relaxations for quadratically constrained quadratic programming: A review and comparisons. *Mathematical programming*, 129(1):129–157, 2011.

[14] Omar Besbes, Will Ma, and Omar Mouchtaki. Beyond iid: data-driven decision-making in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35: 23979–23991, 2022.

[15] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[16] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

[17] Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.

[18] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR, 2019.

[19] Xiaoyu Chen, Jiachen Hu, Lin F. Yang, and Liwei Wang. Near-optimal reward-free exploration for linear mixture MDPs with plug-in solver. In *International Conference on Learning Representations*, 2022.

[20] Xinyun Chen, Pengyi Shi, and Shanwen Pu. Data-pooling reinforcement learning for preventative healthcare intervention. *Management Science*, 72(9):7870–7888, 2025. doi: 10.1287/mnsc.2023.03880.

[21] Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 8286–8309. PMLR, 2024.

[22] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of machine learning research*, 9(8), 2008.

[23] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[24] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[25] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[26] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[27] Qiyang Han, Koulik Khamaru, and Cun-Hui Zhang. Ucb algorithms for multi-armed bandits: precise regret and adaptive inference. *arXiv preprint arXiv:2412.06126*, 2024.

[28] Botao Hao, Rahul Jain, Tor Lattimore, Benjamin Van Roy, and Zheng Wen. Leveraging demonstrations to improve online learning: Quality matters. In *International Conference on Machine Learning*, pages 12527–12545. PMLR, 2023.

[29] Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

[30] Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 40(4):570–596, 2025.

[31] 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.

[32] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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

[34] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pages 5084–5096. PMLR, 2021.

[35] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[36] Russell Mendonca, Xinyang Geng, Chelsea Finn, and Sergey Levine. Meta-reinforcement learning robust to distributional shift via model identification and experience relabeling. *arXiv preprint arXiv:2006.07178*, 2020.

[37] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[38] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[39] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[40] Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *Advances in Neural Information Processing Systems*, 36:62244–62269, 2023.

[41] Thanh Nguyen-Tang, Sunil Gupta, Hung Tran-The, and Svetha Venkatesh. On sample complexity of offline reinforcement learning with deep reLU networks in besov spaces. *Transactions on Machine Learning Research*, 2022.

[42] Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, and Raman Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9310–9318, 2023.

[43] Stephane Ross and J. Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1905–1912, 2012.

[44] Flore Sentenac, Ilbin Lee, and Csaba Szepesvari. Balancing optimism and pessimism in offline-to-online learning. *arXiv preprint arXiv:2502.08259*, 2025.

[45] Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial intelligence and statistics*, pages 1046–1054. PMLR, 2012.

[46] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[47] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

[48] Yuda Song, Yifei Zhou, Ayush Sekhari, Drew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

[49] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[50] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[51] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[52] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[53] Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, pages 35300–35338. PMLR, 2023.

[54] Jia Wan, Sean R Sinclair, Devavrat Shah, and Martin J Wainwright. Exploiting exogenous structure for sample-efficient reinforcement learning. *arXiv preprint arXiv:2409.14557*, 2024.

[55] Zheshun Wu, Renjie Zheng, Jinhang Zuo, Zenglin Xu, and Fang Kong. A unified algorithmic framework for hybrid reinforcement learning in tabular mdps with shifted transition dynamics. *arXiv preprint arXiv:2607.25207*, 2026.

[56] Jize Xie, Yi Chen, and Rachel Q Zhang. Online stochastic optimization with offline data. *Available at SSRN 5573580*, 2025.

[57] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[58] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[59] Le Yang, Vincent YF Tan, and Wang Chi Cheung. Best arm identification with possibly biased offline data. In *Conference on Uncertainty in Artificial Intelligence*, pages 4715–4730. PMLR, 2025.

[60] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pages 6995–7004. PMLR, 2019.

[61] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[62] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[63] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[64] Yixuan Zhang, Ruihao Zhu, and Qiaomin Xie. Contextual online pricing with (biased) offline data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[65] Zhongjun Zhang, Shipra Agrawal, Ilan Lobel, Sean R Sinclair, and Christina Lee Yu. Reinforcement learning in mdps with information-ordered policies. *arXiv preprint arXiv:2508.03904*, 2025.

[66] Han Zheng, Xufang Luo, Pengfei Wei, Xuan Song, Dongsheng Li, and Jing Jiang. Adaptive policy learning for offline-to-online reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11372–11380, 2023.

[67] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[68] Hanlin Zhu, Ruosong Wang, and Jason Lee. Provably efficient reinforcement learning via surprise bound. In *International Conference on Artificial Intelligence and Statistics*, pages 4006–4032. PMLR, 2023.

Appendix contents

A Pseudocode for O-O UCRL-VTR **18**

B Comparisons to settings in the literature **18**

B.1 Offline–online tabular and linear bandits [21, 64] 19

B.1.1 Comparison of environment shift assumptions 19

B.1.2 Comparison of coverage assumptions and the induced τ 20

B.1.3 Comparison of regret bounds 20

B.2 Offline–online tabular MDPs [20] 21

B.2.1 Comparison of environment shift assumptions 21

B.2.2 Comparison of coverage assumptions and the induced τ 22

B.2.3 Comparison of regret bounds 22

B.2.4 Regret bound for offline–online tabular MDPs in Chen et al. [20] 22

B.3 Offline linear mixture MDPs [42] 23

B.3.1 Comparison of coverage assumptions and the induced τ 23

B.3.2 Comparison of suboptimality gap 23

C Omitted discussion from Section 2 to Section 4 **23**

C.1 Environment shift in terms of value functions 24

C.2 Impossibility of unknown Δ and misspecification of Δ 25

C.3 Comparison of Definition 4.1 and Assumption 4.8 with other coverage assumptions 26

C.3.1 Comparison with feature coverage assumptions 26

C.3.2 Comparison with partial data coverage assumptions 26

C.4 Discussion on the convexity of Eq. (13) 27

C.5 Specializing Theorem 4.2 to the purely offline regime 28

C.6 Adaptive offline behavior policy and decaying coverage τ 28

C.7 Regret lower bound when $\tau = o(1)$ 29

D Proofs **29**

D.1 Proofs in Section 4.1 29

D.1.1 Proof of Theorem C.4 38

D.2 Proof in Section 4.2 41

D.3 Proof in Section 4.3 41

D.4 Proofs in Section 4.4 45

D.4.1 Proof of Lemma D.8 47

E Useful facts **53**

F Omitted details in Section 5 **56**

F.1 MDP setting 56

F.2 Complete simulation results 57

F.3 Behavior policy, implementation of O-O UCRL-VTR, and baseline algorithms . . 57
F.4 Additional experiments with continuous state and action spaces 60

Table 1: Common notation.

Symbol	Definition
	Problem setting specifications
$\mathcal{S}, \mathcal{A}$	State space and action space
H, K, d	Episode horizon, number of online episodes, and feature dimension
$\phi(s' \mid s, a)$	Feature map $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$
$\phi_j(s' \mid s, a)$	The jth coordinate of the feature vector $\phi(s' \mid s, a)$
P^{off}, P	Offline and online transition probability
$\theta^{\mathrm{off},*}, \theta^*$	Offline and online transition parameters in the linear mixture model
r^{off}, r	Offline and online reward functions
s_1^{off}, s_1	Initial states of the offline and online MDPs
B	Known bound on $\|\theta^*\|_2$ and $\|\theta^{\mathrm{off},*}\|_2$
Δ	Upper bound over the environment shift, $\|\theta^* - \theta^{\mathrm{off},*}\|_2 \le \Delta$
	Offline data
π_m^{off}	Offline behavior policy for the m-th trajectory
$\mathcal{D}_{\mathrm{off}}$	Offline dataset $\{\mathcal{H}_m^{\mathrm{off}}\}_{m=1}^{M^{\mathrm{off}}}$
M^{off}	Number of offline trajectories
$s_{m,h}^{\mathrm{off}}, a_{m,h}^{\mathrm{off}}$	State and action at stage h of offline trajectory m
$\widetilde{V}_{m,h}$	Auxiliary value function used in offline regression
$x_{m,h}^{\mathrm{off}}$	Offline feature $\phi(\cdot \mid s_{m,h}^{\mathrm{off}}, a_{m,h}^{\mathrm{off}})\widetilde{V}_{m,h+1}$ in the regression
$y_{m,h}^{\mathrm{off}}$	Offline target $\widetilde{V}_{m,h+1}(s_{m,h+1}^{\mathrm{off}})$ in the regression
G_{off}	Offline design matrix $\sum_{m,h} x_{m,h}^{\mathrm{off}}(x_{m,h}^{\mathrm{off}})^\top$
τ	Offline data coverage parameter
	Online interaction
π^*	An optimal policy
π_k	Policy executed in episode k
$s_{k,h}^{\mathrm{on}}, a_{k,h}^{\mathrm{on}}$	State and action at stage h of episode k
V_h^π, Q_h^π	Value and Q functions at stage h under policy π
$x_{k,h}^{\mathrm{on}}$	Online feature $\phi(\cdot \mid s_{k,h}, a_{k,h})\widehat{V}_{k,h+1}$ in the regression
$y_{k,h}^{\mathrm{on}}$	Online target $\widehat{V}_{k,h+1}(s_{k,h+1})$ in the regression
	Algorithm design
λ	Ridge regularization parameter
$M_k^{\mathrm{all}}, M_k^{\mathrm{on}}$	Offline–online design matrix and online design matrix up to episode k
$w_k^{\mathrm{all}}, w_k^{\mathrm{on}}$	Response vectors in the offline–online and online-only ridge regressions up to episode k
$\widehat{\theta}_k^{\mathrm{all}}, \widehat{\theta}_k^{\mathrm{on}}$	Offline–online and online-only ridge estimator
γ_k, β_k	Confidence radius for $\widehat{\theta}_k^{\mathrm{all}}$ and $\widehat{\theta}_k^{\mathrm{on}}$
$\mathcal{B}_k, \mathcal{C}_k$	Offline–online confidence set and online-only confidence set in episode k
$\mathcal{D}_k$	Planning set $\mathcal{B}_k \cap \mathcal{C}_k$ if nonempty, and $\mathcal{B}_k$ otherwise
$\widehat{V}_{k,h}, \widehat{Q}_{k,h}$	Optimistic estimation of value and Q-functions in episode k
	Regret and analysis
$\mathsf{Regret}(K)$	Expected cumulative regret over K episodes
$\lambda_{\min}(\cdot), \lambda_{\max}(\cdot)$	Minimum and maximum eigenvalues
$\|\cdot\|_A$	Weighted norm $\|x\|_A = \sqrt{x^\top A x}$ (A is positive definite)
$\tilde{O}(\cdot)$	Big-O notation hiding logarithmic factors

A Pseudocode for O-O UCRL-VTR

This section presents the detailed pseudocode for O-O UCRL-VTR (Algorithm 1).

Algorithm 1 O-O UCRL-VTR

Require: Online MDP, dimension d, number of online episodes K, offline dataset $\mathcal{D}_{\text{off}}$, auxiliary value functions $\{\widetilde{V}_{m,h}\}$, confidence level δ, radii $\{\gamma_k\}_{k\le K}$ and $\{\beta_k\}_{k\le K}$, ridge regression coefficient λ.

Initialize online-only design matrix and offline–online design matrix: $M_1^{\text{on}} \leftarrow \lambda I, w_1^{\text{on}} \leftarrow 0 \in \mathbb{R}^d, M_1^{\text{all}} \leftarrow \lambda I, w_1^{\text{all}} \leftarrow 0 \in \mathbb{R}^d$.
 # Offline phase
for $m = 1 : M_{\text{off}}$ **do**
  for $h = 1 : H$ **do**
    Form the regression pair

$$x_{m,h}^{\text{off}} \leftarrow \phi(\cdot \mid s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}})\widetilde{V}_{m,h+1} \in \mathbb{R}^d, y_{m,h}^{\text{off}} \leftarrow \widetilde{V}_{m,h+1}\big(s_{m,h+1}^{\text{off}}\big).$$

    Update statistics $M_1^{\text{all}} \leftarrow M_1^{\text{all}} + x_{m,h}^{\text{off}}(x_{m,h}^{\text{off}})^\top, w_1^{\text{all}} \leftarrow w_1^{\text{all}} + x_{m,h}^{\text{off}} y_{m,h}^{\text{off}}$.
  end for
end for
 # Online phase
for $k = 1 : K$ **do**
  Compute estimators $\widehat{\theta}_k^{\text{all}} \leftarrow (M_k^{\text{all}})^{-1} w_k^{\text{all}}, \widehat{\theta}_k^{\text{on}} \leftarrow (M_k^{\text{on}})^{-1} w_k^{\text{on}}$.
  Define confidence sets[3]

$$\mathcal{B}_k \leftarrow \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_k^{\text{all}}\|_{M_k^{\text{all}}} \le \gamma_k\Big\}, \mathcal{C}_k \leftarrow \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_k^{\text{on}}\|_{M_k^{\text{on}}} \le \beta_k\Big\}.$$

  Also define

$$\mathcal{D}_k = \begin{cases} \mathcal{B}_k \cap \mathcal{C}_k, & \mathcal{B}_k \cap \mathcal{C}_k \neq \varnothing, \\ \mathcal{B}_k, & \mathcal{B}_k \cap \mathcal{C}_k = \varnothing. \end{cases}$$

  Set $\widehat{Q}_{k,H+1}(\cdot,\cdot) \leftarrow 0$, and for $h = H, \dots, 1$ define
  // Compute optimistic estimates $\widehat{Q}$ and $\widehat{V}$

$$\widehat{V}_{k,h}(s) \leftarrow \max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s,a), \widehat{Q}_{k,h}(s,a) \leftarrow r(s,a) + \max_{\theta\in\mathcal{D}_k} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}.$$

  Set $s_{k,1}^{\text{on}} = s_1$.
  for $h = 1 : H$ **do**
    Choose $a_{k,h}^{\text{on}} \in \arg\max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s_{k,h}^{\text{on}}, a)$ and observe $s_{k,h+1}^{\text{on}} \sim P(\cdot \mid s_{k,h}^{\text{on}}, a_{k,h}^{\text{on}})$.
    Form the online regression pair $x_{k,h}^{\text{on}} \leftarrow \phi(\cdot \mid s_{k,h}^{\text{on}}, a_{k,h}^{\text{on}})\widehat{V}_{k,h+1}, y_{k,h}^{\text{on}} \leftarrow \widehat{V}_{k,h+1}\big(s_{k,h+1}^{\text{on}}\big)$.
    Update online-only statistics $M_k^{\text{on}} \leftarrow M_k^{\text{on}} + x_{k,h}^{\text{on}}(x_{k,h}^{\text{on}})^\top, w_k^{\text{on}} \leftarrow w_k^{\text{on}} + x_{k,h}^{\text{on}} y_{k,h}^{\text{on}}$.
    Update offline–online statistics $M_k^{\text{all}} \leftarrow M_k^{\text{all}} + x_{k,h}^{\text{on}}(x_{k,h}^{\text{on}})^\top, w_k^{\text{all}} \leftarrow w_k^{\text{all}} + x_{k,h}^{\text{on}} y_{k,h}^{\text{on}}$.
  end for
  $M_{k+1}^{\text{on}} \leftarrow M_k^{\text{on}}, M_{k+1}^{\text{all}} \leftarrow M_k^{\text{all}}, w_{k+1}^{\text{on}} \leftarrow w_k^{\text{on}}, w_{k+1}^{\text{all}} \leftarrow w_k^{\text{all}}$.
end for

B Comparisons to settings in the literature

The offline–online linear mixture MDP setting studied in this paper subsumes several canonical models of interest, including offline–online tabular and linear bandits, offline–online tabular MDPs, and offline linear mixture MDPs. In this section, we show how these three settings can be embedded into our offline–online linear mixture MDP framework, and we compare our main result in Theorem 4.2 with the corresponding guarantees established for each setting in the literature.

[3] Both $\mathcal{B}_k$ and $\mathcal{C}_k$ are ellipsoids for each k. Since the objective is linear in θ, the resulting optimization is a quadratically constrained quadratic program (QCQP) [13], which is tractable.

B.1 Offline–online tabular and linear bandits [21, 64]

We start by examining the connection between Theorem 4.2 and prior work on offline–online tabular bandits [21] and offline–online linear bandits [64]. Since tabular bandits can be expressed as linear bandits through one-hot mapping, similar to Remark 2.2, both models can be viewed as special cases of the offline–online linear mixture MDP framework. We show that (i) the environment shift assumption in Assumption 2.3 is no stronger than those adopted in [21, 64]; (ii) how the coverage assumption in Definition 4.1 connects with the coverage assumptions in these literature; and (iii) in benign regimes, the regret bound in Theorem 4.2 matches the order of the regret guarantee in Cheung and Lyu [21]. Note that Zhang et al. [64] consider a contextual model, and therefore the regret guarantees are not directly comparable.

B.1.1 Comparison of environment shift assumptions

In stochastic tabular bandits with A arms [21], the environment shift is characterized by an arm-wise reward discrepancy: for any arm a,

$$|\mu(a) - \mu^{\text{off}}(a)| \le \Delta_1 \tag{15}$$

where $\mu^{\text{off}}(a)$ and $\mu(a)$ denote the expected rewards of arm a in the offline and online environments, respectively. Similarly, in stochastic linear bandits [64], the expected reward under action $a \in \mathbb{R}^d$ is modeled as $\mu(a) = \langle \theta_B^*, a\rangle$ for the online environment, and $\mu^{\text{off}}(a) = \langle \theta_B^{\text{off},*}, a\rangle$ for the offline environment. The environment shift is then quantified by:

$$\|\theta_B^{\text{off},*} - \theta_B^*\|_2 \le \Delta_2. \tag{16}$$

Similar to MDPs, a tabular bandit can be represented as a linear bandit via a one-hot feature mapping where $\theta = (\mu(a_1), \ldots, \mu(a_A)) \in \mathbb{R}^A$. Then the induced linear bandit environment shift,

$$\Delta_2(\Delta_1) := \|\theta_B^{\text{off},*} - \theta_B^*\|_2 \le \sqrt{A}\Delta_1 \tag{17}$$

via the corresponding feature representation, where A denotes the number of arms. We therefore focus on embedding linear bandits satisfying Eq. (16) into offline–online linear mixture MDPs, which directly induces the relationship between Eq. (15) in a tabular bandit model and the environment shift Δ in Assumption 2.3. We now describe an embedding of linear bandits into linear mixture MDPs. For simplicity, we consider a stochastic Bernoulli offline–online linear bandit. Given an instance with parameters $\theta_B^{\text{off},*}, \theta_B^* \in \mathbb{R}^{d-1}$, we construct an offline–online linear mixture MDP with the same action space, state space $\mathcal{S} = \{s_1, s_2, s_3\}$, and horizon $H = 2$. At stage $h = 1$, the initial state is always s_1 and the reward is zero. Upon taking action a, the system transitions to s_2 with probability $\langle \theta_B^{\text{off},*}, a\rangle$ in the offline environment and $\langle \theta_B^*, a\rangle$ in the online environment. With the remaining probability, it transitions to s_3. At stage $h = 2$, the reward at s_2 is 1, while the reward at s_3 is 0. To express these transitions in linear mixture form, define $\theta^{\text{off},*} = (\theta_B^{\text{off},*}, 1)$ and $\theta^* = (\theta_B^*, 1)$, with the known feature mapping $\phi(s_2 \mid s_1, a) = (a, 0)$ and $\phi(s_3 \mid s_1, a) = (-a, 1)$.

It is easy to see that this offline–online linear mixture MDP satisfies all assumptions in the main text. In particular, the reward functions r^{off} and r are known, and the induced stochastic process is equivalent to that of the original linear bandit problem. As a result, we obtain the following relationship between the environment shifts.

Proposition B.1. *Consider a stochastic offline–online linear bandit problem for which Eq. (16) holds. Let (P^{off}, r) and (P, r) be the equivalent offline and online linear mixture MDPs constructed above, with parameter vectors $\theta^{\text{off},*}$ and θ^*. We define $\Delta(\Delta_2)$ as the induced environment shift bound under the constructed offline–online linear mixture MDP. Then Assumption 2.3 holds with $\Delta(\Delta_2) \le \Delta_2$.*

Similarly, for any offline–online tabular bandit problem satisfying Eq. (15), there exists an equivalent offline–online linear mixture MDP representation. Let $\Delta(\Delta_1)$ denote the induced environment shift bound under the resulting offline–online linear mixture MDP. We then directly obtain that $\Delta(\Delta_1) \le \Delta_2(\Delta_1) \le \sqrt{A}\,\Delta_1$, where the second inequality is from Eq. (17). Consequently, for any offline–online tabular bandit problem under Eq. (15), or offline–online linear bandit problem under Eq. (16), Theorem 4.2 applies.

B.1.2 Comparison of coverage assumptions and the induced τ

In this subsection, we compare Definition 4.1 with the coverage assumptions adopted in Cheung and Lyu [21], Zhang et al. [64].

Coverage assumption in Cheung and Lyu [21]. In Cheung and Lyu [21], the authors assume that the number of offline pulls for each arm is known. We now explain how this assumption specifies the corresponding value of τ in Definition 4.1 in our framework.

Consider a bandit problem with A arms indexed by $\{a_1, \dots, a_A\}$. Any finite-armed bandit can be embedded into a linear mixture MDP with three states, as described in Section B.1.1. Since each tabular bandit is a special case of a tabular MDP, the deterministic construction of $\{\widetilde{V}_{m,h}\}$ in Section 4.3 applies directly. Here we use the tabular embedding in Remark 2.2; thus, for each $i \in [A]$, $\phi(\cdot \mid s_1, a_i)$ has exactly one nonzero entry in each column. Thus under the deterministic design of $\{\widetilde{V}_{m,h}\}$, namely by selecting the auxiliary value functions $\widetilde{V}_m(s_1, a_i)$ cyclically from the standard basis $\{(1,0,0),(0,1,0),(0,0,1)\}$ for each arm a_i, the resulting offline design matrix has the form

$$G_{\text{off}} = \operatorname{diag}\Big(\frac{N(a_1)}{3}, \frac{N(a_1)}{3}, \frac{N(a_1)}{3}, \frac{N(a_2)}{3}, \dots, \frac{N(a_A)}{3}\Big), \tag{18}$$

where $N(a_i)$ denotes the number of times arm a_i is pulled in the offline data. For simplicity, we assume that $N(a_i)$ is divisible by 3 for all $i \in [A]$. As a result, we have that $\tau \geq \min_{i\in[A]} \frac{N(a_i)+\lambda}{3M^{\text{off}}}$. Therefore, the value of τ depends on the allocation of offline pulls. When the pulls are distributed uniformly across arms, we obtain $\tau \geq \frac{1}{3A}$. In contrast, under highly uneven allocations, one can only guarantee the trivial lower bound $\tau \geq \frac{1+\lambda}{3M^{\text{off}}}$ in the worst case.

Coverage assumption in Zhang et al. [64]. Although Zhang et al. [64] study a contextual model that differs from our linear mixture MDP framework, and therefore a direct relation between the two settings is not available as in the previous case, their coverage assumption is closely related at a high level to Definition 4.1. In the offline phase, they assume a dataset of size M^{off}, where each sample indexed by $m \in [M^{\text{off}}]$ is associated with a contextual feature vector x_m (we adopt our notation for consistency). Their coverage assumption is

$$\lambda_{\min}\left(\sum_{m=1}^{M^{\text{off}}} x_m x_m^\top\right) \geq M^{\text{off}}\tau \qquad \text{a.s.}, \tag{19}$$

which implies that $\tau = \Theta(1)$ under our definition of τ. This assumption is structurally similar to Definition 4.1, and is in fact stronger in two respects. First, it requires an almost sure guarantee, whereas Definition 4.1 is stated in a high-probability sense. Second, it captures only the ideal regime in which $\tau = \Theta(1)$ (Case 1 in Section 4.2), while in our framework under Definition 4.1, we also allow τ to decay as M^{off} increases (Case 2 in Section 4.2). Although the underlying models and feature spaces differ, both Eq. (19) and Definition 4.1 impose a form of uniform coverage over the feature space during the offline phase.

B.1.3 Comparison of regret bounds

We finally compare Theorem 4.2 with the offline–online tabular bandit setting studied in Cheung and Lyu [21]. The regret bound in Zhang et al. [64] is not directly comparable, as that work considers a contextual model.

By Section B.1.1, any offline–online tabular bandit problem with environment shift Δ_1 (see Eq. (15)), when embedded into the offline–online linear mixture MDP framework, satisfies $\Delta(\Delta_1) \leq \Delta_2(\Delta_1) \leq \sqrt{A}\,\Delta_1$, where the first inequality follows from Proposition B.1, and the second follows from the relation between $\Delta_2(\Delta_1)$ and Δ_1 under the one-hot embedding. Furthermore, due to the discussion in Section B.1.2, we have that $\tau \geq \min_{i\in[A]} \frac{N(a_i)+\lambda}{3M^{\text{off}}}$, where $N(a_i)$ denotes the number of times arm a_i is pulled in the offline data. Moreover, under this embedding, only transitions from the initial state s_1 are nontrivial; the transitions at s_2 and s_3 are deterministic and can therefore be ignored. Thus the dimension $d = 3A$ is sufficient. Applying Theorem 4.2 with $d = 3A$ then gives

$$\mathsf{Regret}(K) = \tilde{O}\left(\min\left\{\frac{AK}{\sqrt{\tau M^{\text{off}}}} + A^2\frac{\Delta_1 K}{\tau}, A\sqrt{K}\right\}\right).$$

When the offline data are evenly distributed across arms, we have that $\tau = \frac{1}{3A} + \frac{\lambda}{3M^{\text{off}}}$, which yields

$$\textsf{Regret}(K) = \tilde{O}\left(\min\left\{\frac{A^{3/2}K}{\sqrt{M^{\text{off}}}} + A^3\Delta_1 K, A\sqrt{K}\right\}\right).$$

This matches the dependence on K, M^{off} and Δ_1 established in Cheung and Lyu [21] (see Theorem 4.9 of Cheung and Lyu [21]). However, if the offline samples are highly imbalanced across arms, e.g., some arm receives only $O(1)$ pulls as $M^{\text{off}} \to \infty$, then $\tau = O(1/M^{\text{off}})$ and the bound in Theorem 4.2 reduces to the purely online bound $\tilde{O}(A\sqrt{K})$. This reflects that our guarantee in Theorem 4.2 is governed by the worst-covered direction through τ. In contrast, the guarantees in Cheung and Lyu [21] are expressed in terms of per-arm counts, and may still yield improvements under imbalanced offline data. These differences arise because our analysis is designed to cover general linear mixture models beyond the tabular setting.

B.2 Offline–online tabular MDPs [20]

By Remark 2.2, any offline–online tabular MDP can be embedded into the offline–online linear mixture MDP framework. Analogous to Section B.1, we examine the relation between Theorem 4.2 and prior work on offline–online tabular MDPs [20], focusing on the respective environment shift assumptions, coverage assumptions, and regret guarantees.

B.2.1 Comparison of environment shift assumptions

In this subsection, we relate Assumption 2.3 to the environment shift assumption in tabular MDPs proposed by Chen et al. [20], which assumes that for all (s, a),

$$\left\|P(\cdot \mid s, a) - P^{\text{off}}(\cdot \mid s, a)\right\|_1 \le \Delta_3. \tag{20}$$

Using the standard equivalence that represents any tabular MDP as a linear mixture MDP (see Remark 2.2), we show that Eq. (20) implies Assumption 2.3 with an explicit conversion between the two environment shifts.

Proposition B.2. *Consider a tabular offline–online MDP for which Eq. (20) holds. Let (P^{off}, r) and (P, r) be rewritten as the equivalent offline and online linear mixture MDPs following Remark 2.2, with parameter vectors θ^* and $\theta^{\text{off},*}$ of dimension d. We define $\Delta(\Delta_3)$ as the induced environment shift bound under the corresponding offline–online linear mixture MDP. Then Assumption 2.3 holds with[4] $\Delta(\Delta_3) \le \sqrt{SA}\Delta_3 \le \sqrt{d}\Delta_3$, where $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, and $\Delta(\Delta_3)$ is defined analogously to $\Delta_2(\Delta_1)$.*

Proof. Under the equivalence that represents a tabular MDP as a linear mixture MDP (see Remark 2.2), each coordinate of the parameter vector θ corresponds to a transition probability $P(s' \mid s, a)$. Consequently, for each (s, a), the vector of $\theta^* - \theta^{\text{off},*}$ associated with (s, a) is exactly $P(\cdot \mid s, a) - P^{\text{off}}(\cdot \mid s, a)$. Therefore,

$$\|\theta^* - \theta^{\text{off},*}\|_2^2 = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left\|P(\cdot \mid s, a) - P^{\text{off}}(\cdot \mid s, a)\right\|_2^2.$$

Using the inequality $\|\cdot\|_2 \le \|\cdot\|_1$ and the tabular environment shift assumption (20), we obtain $\|\theta^* - \theta^{\text{off},*}\|_2^2 \le \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Delta_3^2 = SA\Delta_3^2$. Finally, since $d = S^2A$ under the tabular equivalence, we have $\sqrt{SA} \le \sqrt{d}$, which completes the proof. □

Remark B.3. Note that Eq. (20) captures only part of the environment shift assumption in Chen et al. [20] (their Assumption 1). In particular, Chen et al. [20] also allow discrepancies between r and r^{off}, treating rewards as stochastic and unknown, and consider time-inhomogeneous transition models. Here we restate their assumption in the form of Eq. (20) to align with the setting in our main text. Extensions of Proposition B.2 to unknown reward functions and time-inhomogeneous transitions are conceptually straightforward and therefore omitted.

[4] Proposition B.2 shows that $\Delta(\Delta_3) \le O(\Delta_3)$. To obtain the reverse direction and conclude that $\Delta(\Delta_3) = \Theta(\Delta_3)$, one can instead define $\Delta(\Delta_3)$ and Δ_3 to measure the actual discrepancy between the offline and online environments, rather than an upper bound on that discrepancy.

B.2.2 Comparison of coverage assumptions and the induced τ

In Chen et al. [20], their coverage assumption requires that, for every (h, s, a), the number of offline visits equals $M^{\text{off}}/(|\mathcal{S}||\mathcal{A}|)$ (see Theorem 2 of Chen et al. [20]). Under this uniform allocation, the deterministic construction of $\{\widetilde{V}_{m,h}\}$ in Section 4.3 applies directly. The resulting offline design matrix satisfies

$$G_{\text{off}} = \text{diag}\Big(\frac{HM^{\text{off}}}{|\mathcal{S}|^2|\mathcal{A}|}, \ldots, \frac{HM^{\text{off}}}{|\mathcal{S}|^2|\mathcal{A}|}\Big) \qquad \text{a.s.}, \tag{21}$$

so that $\lambda_{\min}(G_{\text{off}}) = \frac{HM^{\text{off}}}{|\mathcal{S}|^2|\mathcal{A}|}$ almost surely. Consequently, $\tau \geq \frac{H}{|\mathcal{S}|^2|\mathcal{A}|}$. In other words, the coverage assumption in Chen et al. [20] guarantees a nontrivial lower bound on τ in our framework, namely $\tau = \Theta(1)$. Furthermore, similar to the comparison between Definition 4.1 and the coverage assumption in Zhang et al. [64] in the previous section, our framework under Definition 4.1 also accommodates the regime in which τ decays as M^{off} increases.

B.2.3 Comparison of regret bounds

Combining the discussions in Section B.2.1 and Section B.2.2, we obtain the following corollary.

Corollary B.4. *Under the assumptions of Chen et al. [20] where the environment shift is bounded by Δ_3 (see Eq. (20)), we have $\tau \geq H/(|\mathcal{S}|^2|\mathcal{A}|)$, and the regret bound in Theorem 4.2 becomes*

$$\textit{Regret}(K) = \tilde{O}\left(\min\left\{|\mathcal{S}|^3|\mathcal{A}|^{3/2}K\sqrt{\frac{1}{M^{\text{off}}}} + |\mathcal{S}|^6|\mathcal{A}|^3\Delta_3 K, |\mathcal{S}|^2|\mathcal{A}|\sqrt{K}\right\}\right),$$

where the dependence on H and B is omitted.

Proof of Corollary B.4. Note that $\tau \geq \frac{H}{|\mathcal{S}|^2|\mathcal{A}|}$ due to Section B.2.2. Substituting $d = |\mathcal{S}|^2|\mathcal{A}|$ (see Remark 2.2) and $\Delta(\Delta_3) \leq \sqrt{d}\Delta_3$ into the regret bound in Theorem 4.2 directly yields

$$\textsf{Regret}(K) = \tilde{O}\left(\min\left\{|\mathcal{S}|^3|\mathcal{A}|^{3/2}K\sqrt{\frac{1}{M^{\text{off}}}} + |\mathcal{S}|^6|\mathcal{A}|^3\Delta_3 K, |\mathcal{S}|^2|\mathcal{A}|\sqrt{K}\right\}\right).$$

□

In comparison, a tight upper bound of the regret bound of Chen et al. [20] scales as $|\mathcal{S}||\mathcal{A}|^{1/2}\frac{K}{\sqrt{M^{\text{off}}}} + \Delta_3 K$, again omitting the dependence on H (within the regime of interest; see Section B.2.4). When M^{off}/K is sufficiently large, the two bounds agree in their dependence on K, Δ_3, and M^{off}.

B.2.4 Regret bound for offline–online tabular MDPs in Chen et al. [20]

We briefly discuss how the regret bound of Chen et al. [20] leads to the scaling $|\mathcal{S}||\mathcal{A}|^{1/2}\frac{K}{\sqrt{M^{\text{off}}}} + \Delta_3 K$, which we use for comparison in Section B.2.3.

Although Chen et al. [20] do not explicitly state a closed-form regret bound, their Theorem 2 provides a bound expressed in terms of confidence radii that depend on the environment shift. In their notation, three confidence radii are defined: $\epsilon_R^{\text{DP}}(n, N), \epsilon_P^{\text{DP}}(n, N), \epsilon_V^{\text{DP}}(n, N)$, where $n \in \{0, \ldots, K/(|\mathcal{S}||\mathcal{A}|)\}$ and $N = M^{\text{off}}/(|\mathcal{S}||\mathcal{A}|)$. These radii are given by

$$\epsilon_R^{\text{DP}}(n, N) = \sqrt{\frac{\mu^2}{2n} + \frac{(1-\mu)^2}{2N}} + \Delta_3(1-\mu),$$

$$\epsilon_P^{\text{DP}}(n, N) = \sqrt{2|\mathcal{S}|\left(\frac{\mu^2}{2n} + \frac{(1-\mu)^2}{2N}\right)} + \Delta_3(1-\mu), \epsilon_V^{\text{DP}}(n, N) = H\epsilon_R^{\text{DP}}(n, N),$$

where $\mu \in [0, 1]$ is a weighting parameter (see Section 4.2 of Chen et al. [20] for details).

As noted in Chen et al. [20], the function $\sqrt{\frac{\mu^2}{2n} + \frac{(1-\mu)^2}{2N}} + \Delta_3(1-\mu)$ is convex in μ. Therefore, by setting $\mu = 0$ and applying Equation (EC.11) of Chen et al. [20], the regret satisfies

$$\textsf{Regret}(K) \leq |\mathcal{S}||\mathcal{A}| \sum_{n=0}^{K/(|\mathcal{S}||\mathcal{A}|)} \Big(\epsilon_R^{\text{DP}}(n, N) + \epsilon_P^{\text{DP}}(n, N) + \epsilon_V^{\text{DP}}(n, N)\Big).$$

Substituting the above expressions and summing over n yields

$$\mathsf{Regret}(K) = \tilde{O}\left(|\mathcal{S}||\mathcal{A}|^{1/2}\sqrt{K^2/M^{\mathrm{off}}} + \Delta_3 K\right),$$

where we omit logarithmic factors and constants, as well as the dependence on H. We emphasize that setting $\mu = 0$ and exploiting convexity does not materially loosen the bound derived in Chen et al. [20]. As long as the optimal choice of μ satisfies $\mu < 1$ for all n (which holds in most regimes of interest), the gap between the tightest bound obtainable from Chen et al. [20] and the simplified bound used in our discussion is at most a constant factor. Note that the algorithm in Chen et al. [20] also has an online-only safeguard. Its choice of μ depends on the problem parameters, including Δ, and equals 1 when Δ is large. Substituting $\mu = 1$ into the equations above directly yields the online-only regret bound.

B.3 Offline linear mixture MDPs [42]

We finally compare Theorem 4.2 with the suboptimality guarantees for offline linear mixture MDPs studied in Nguyen-Tang et al. [42]. When $K = 1$ and $\Delta = 0$, our regret bound reduces to a purely offline suboptimality bound; see Section C.5 for details. Since this regime involves no environment shift, we focus on comparing the respective coverage assumptions and suboptimality gap guarantees in Theorem 4.2 and Nguyen-Tang et al. [42]. As in Section 4.3, for clarity of exposition, we assume in this section that the state and action spaces are finite, so all probability measures are treated as probability mass functions.

B.3.1 Comparison of coverage assumptions and the induced τ

Although the primary contribution of Nguyen-Tang et al. [42] is to establish suboptimality guarantees under a partial coverage assumption (Assumptions 4.1 and 4.2 in Nguyen-Tang et al. [42]), the uniform data coverage assumption, Assumption 4.8, as discussed in Nguyen-Tang et al. [42], is also sufficient to obtain their suboptimality results for linear mixture MDPs (Theorem 7 and Theorem 9 in Nguyen-Tang et al. [42]). Under the stochastic construction of $\{\widetilde{V}_{m,h}\}$ in Definition 4.9 and the uniform learnability assumption (Definition 4.5), the uniform data coverage condition in Assumption 4.8 implies that $\tau = \Theta(1)$. This argument is formalized in Proposition 4.10 in Section 4.3. We defer the discussion of the partial coverage assumptions to Section C.3.2 and show that these assumptions are not applicable for the offline–online learning setting considered here.

B.3.2 Comparison of suboptimality gap

With discussions in Section B.3.1, we directly obtain the following corollary.

Corollary B.5. *Under Definition 4.9, Definition 4.5 and Assumption 4.8, when $K = 1$ and $\Delta = 0$, we have $\tau = \Theta(1)$, and the regret bound in Theorem 4.2 reduces to an offline suboptimality gap of order $\tilde{O}(\frac{1}{\sqrt{M^{\mathrm{off}}}})$.*

Corollary B.5 follows directly from combining Proposition 4.10 with Theorem 4.2. The $\tilde{O}(\frac{1}{\sqrt{M^{\mathrm{off}}}})$ rate matches the rate obtained in Theorems 7 and Theorem 9 of Nguyen-Tang et al. [42], up to logarithmic factors, and is statistically optimal. This shows that, in the purely offline regime, our bound is sharp.

C Omitted discussion from Section 2 to Section 4

In this section, we supplement Section B with additional discussions beyond comparisons to existing literature. Specifically, we show (i) the environment shift Δ and the resulting discrepancy between offline and online value functions are nearly of the same order (Section C.1); (ii) the necessity of the assumption that Δ is known in Assumption 2.3 (Section C.2); (iii) a comparison of Definition 4.1 and other coverage assumptions in the literature (Section C.3); (iv) Eq. (13) is neither convex nor concave in general (Section C.4); (v) the technical details omitted in translating Theorem 4.2 into a suboptimality gap bound for comparison with Nguyen-Tang et al. [42] in Section B.3 (Section C.5); (vi) the discussion about adaptive offline behavior policies (Section C.6); and (vii) the regret lower bound when $\tau = o(1)$ (Section C.7).

C.1 Environment shift in terms of value functions

We start off by studying Assumption 2.3 from a value function perspective. Although one could alternatively define the environment shift Δ directly in terms of value function discrepancies between the offline and online MDPs, we show that Assumption 2.3 already implies an explicit bound on the difference between value functions induced by the same policy in the offline and online environments. Moreover, there exist instances in which Δ exactly matches this value function gap. We therefore retain the parameter-space definition of Δ in Assumption 2.3, which more naturally reflects the model geometry and supports the regret analysis.

Denote $V_h^{\mathrm{off},\pi}$ as the value function at stage h under policy π in the offline MDP, defined analogously to V_h^π in the online MDP in Eq. (2). We start by showing that when the reward functions r and r^{off} are identical, the value difference at the initial state is upper bounded at the same order of $\|\theta^* - \theta^{\mathrm{off},*}\|_2$, up to a polynomial dependence on the horizon H and the feature dimension d.

Proposition C.1. *Assume that the reward functions are identical, i.e., $r = r^{\mathrm{off}}$. Under Definition 2.1 and Assumption 2.3, for any fixed policy π, we have $|V_1^\pi(s_1) - V_1^{\mathrm{off},\pi}(s_1)| \le \frac{H(H-1)}{2}\Delta\sqrt{d}$.*

Proof. Fix an arbitrary policy π. For any stage h, since the reward functions are identical, the value functions satisfy the Bellman recursions

$$V_h^\pi(s) = r(s,\pi_h(s)) + \sum_{s'} P(s' \mid s,\pi_h(s)) V_{h+1}^\pi(s'),$$
$$V_h^{\mathrm{off},\pi}(s) = r(s,\pi_h(s)) + \sum_{s'} P^{\mathrm{off}}(s' \mid s,\pi_h(s)) V_{h+1}^{\mathrm{off},\pi}(s'),$$

with terminal condition $V_{H+1}^\pi \equiv V_{H+1}^{\mathrm{off},\pi} \equiv 0$. Subtracting the two equations and adding and subtracting $\sum_{s'} P^{\mathrm{off}}(s' \mid s,\pi_h(s)) V_{h+1}^\pi(s')$, we obtain

$$\begin{aligned} V_h^\pi(s) - V_h^{\mathrm{off},\pi}(s) = &\sum_{s'} \big(P - P^{\mathrm{off}}\big)\big(s' \mid s,\pi_h(s)\big) V_{h+1}^\pi(s') \\ &+ \sum_{s'} P^{\mathrm{off}}\big(s' \mid s,\pi_h(s)\big)\big(V_{h+1}^\pi(s') - V_{h+1}^{\mathrm{off},\pi}(s')\big). \end{aligned}$$

We bound the two terms separately. For the first term, by the linear mixture structure,

$$\sum_{s'} \big(P - P^{\mathrm{off}}\big)\big(s' \mid s,a\big) V_{h+1}^\pi(s') = \sum_{j=1}^d \big(\theta_j^* - \theta_j^{\mathrm{off},*}\big) \sum_{s'} \phi_j(s' \mid s,a) V_{h+1}^\pi(s').$$

Noticing that (i) $0 \le V_{h+1}^\pi(s') \le H - h$ for all s', and that (ii) $\sum_{s'} |\phi_j(s' \mid s,a)| \le 1$, we have $\big|\sum_{s'} \phi_j(s' \mid s,a) V_{h+1}^\pi(s')\big| \le H - h$. Applying Cauchy–Schwarz yields

$$\left|\sum_{s'} \big(P - P^{\mathrm{off}}\big)\big(s' \mid s,a\big) V_{h+1}^\pi(s')\right| \le \|\theta^* - \theta^{\mathrm{off},*}\|_2 \cdot \sqrt{d}(H-h) \le \Delta\sqrt{d}(H-h).$$

For the second term, since $P^{\mathrm{off}}(\cdot \mid s,a)$ is a probability distribution,

$$\left|\sum_{s'} P^{\mathrm{off}}\big(s' \mid s,a\big)\big(V_{h+1}^\pi(s') - V_{h+1}^{\mathrm{off},\pi}(s')\big)\right| \le \sup_{s'} \big|V_{h+1}^\pi(s') - V_{h+1}^{\mathrm{off},\pi}(s')\big|.$$

Define $D_h := \sup_s \big|V_h^\pi(s) - V_h^{\mathrm{off},\pi}(s)\big|$. Combining the above bounds gives the recursion

$$D_h \le \Delta\sqrt{d}(H-h) + D_{h+1}, \qquad D_{H+1} = 0.$$

Unrolling the recursion yields $D_1 \le \Delta\sqrt{d}\sum_{h=1}^H (H-h) = \Delta\sqrt{d} \cdot \frac{H(H-1)}{2}$. Since $\big|V_1^\pi(s_1) - V_1^{\mathrm{off},\pi}(s_1)\big| \le D_1$, the proof is complete. □

We now show that the above relationship is nearly tight in a worst-case sense. Specifically, there exist instances for which the discrepancy between the offline and online value functions scales linearly with the parameter shift $\|\theta^* - \theta^{\mathrm{off},*}\|_2$.

Proposition C.2. *For any $\Delta > 0$, there exist parameters θ^* and $\theta^{\mathrm{off},*}$ with $\|\theta^* - \theta^{\mathrm{off},*}\|_2 = \Delta$ such that $\sup_\pi \left|V_1^\pi(s_1) - V_1^{\mathrm{off},\pi}(s_1)\right| = \Omega(\Delta)$.*

Proof. Consider a linear bandit instance, which can be equivalently represented as a linear mixture MDP as described in Section B.1.1. Under this construction, the parameter vectors of the induced linear mixture MDP coincide exactly with the bandit parameters (ignoring the additional constant coordinate), so $\Delta = \|\theta^* - \theta^{\mathrm{off},*}\|_2$. Moreover, for any policy $\pi \in \mathbb{R}^d$ (which is valid in the linear bandit setting), the corresponding value functions satisfy $V_1^\pi(s_1) - V_1^{\mathrm{off},\pi}(s_1) = \langle \theta^* - \theta^{\mathrm{off},*}, \pi \rangle$. Taking $\pi = \frac{\theta^* - \theta^{\mathrm{off},*}}{\|\theta^* - \theta^{\mathrm{off},*}\|_2}$ yields $\sup_\pi \left|V_1^\pi(s_1) - V_1^{\mathrm{off},\pi}(s_1)\right| = \|\theta^* - \theta^{\mathrm{off},*}\|_2 = \Delta$, which completes the proof. □

C.2 Impossibility of unknown Δ and misspecification of Δ

We briefly discuss the offline–online linear mixture MDP setting in which no explicit upper bound on $\|\theta^* - \theta^{\mathrm{off},*}\|_2$ (or any equivalent measure of environment shift) is given, i.e., Δ is unknown. The main takeaway is that, without knowing Δ, any algorithm that achieves a regret rate strictly better than $\sqrt{K}$ in the no-shift case $\theta^* = \theta^{\mathrm{off},*}$ must necessarily incur regret strictly worse than $\sqrt{K}$-scale on some shifted instance. Thus, knowledge of an upper bound on Δ is essential for obtaining a worst-case guarantee that uniformly improves upon the online-only rate while remaining strictly better in the no-shift regime. Thus, seemingly appealing approaches, such as estimating Δ in the online phase as more online data are collected, cannot improve the regret guarantees from the minimax perspective.

Proposition C.3. *Consider the class of offline–online linear mixture MDP instances for which all assumptions in the main text hold except Assumption 2.3, i.e., the environment shift is not assumed to be bounded by a known constant. Suppose an algorithm satisfies $\mathsf{Regret}(K) = o(\sqrt{K})$ over the instance class when $\theta^* = \theta^{\mathrm{off},*}$, then there exists another instance in the same class with $\theta^* \neq \theta^{\mathrm{off},*}$ for which the algorithm incurs regret $\mathsf{Regret}(K) > \Omega(\sqrt{K})$.*

We omit the proof of Proposition C.3, as Cheung and Lyu [21] provides an explicit hard instance construction in the tabular bandit setting (see Proposition 3.1 in Cheung and Lyu [21]). By the equivalence discussed in Section B.1.1, such instances can be embedded as a special case of offline–online linear mixture MDPs.

We next show that when Δ is misspecified, i.e., the value $\tilde{\Delta}$ used by the algorithm is smaller than the true environment shift Δ, Algorithm 1 remains implementable, and the offline–online regret bound in Theorem 4.2 (term (A)) continues to hold under the true Δ, up to lower-order terms in the regime of interest. However, the *min-of-two safeguard is no longer guaranteed.* Concretely, we have the following statement.

Theorem C.4. *Suppose that all assumptions in Theorem 4.2 hold except Assumption 2.3. Let $\Delta = \|\theta^{\mathrm{off},*} - \theta^*\|_2$ denote the true environment shift. Run Algorithm 1 with $\lambda = H^2 d$ and the confidence radii specified in Theorem 4.2, replacing Δ by a misspecified value $\tilde{\Delta}$ satisfying $0 \le \tilde{\Delta} < \Delta$. Then the algorithm remains implementable without knowledge of Δ, and we have*

$$\mathsf{Regret}(K) \le \tilde{O}\left(\left[BH^2 d\sqrt{K} + H^4 d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\mathrm{off}}}{\tau}} \right] \sqrt{\log\left(1 + \frac{H^3 K}{\tau M^{\mathrm{off}}}\right)} + \frac{BH^3 dK}{\sqrt{\tau M^{\mathrm{off}}}} + \frac{\Delta H^5 d^{3/2} K}{\tau} \right). \tag{22}$$

We defer the proof of Theorem C.4 to Section D.1.1, as it largely follows the proof of Theorem 4.2 in Section D.1. The bound in Eq. (22) consists of three terms, the first of which is the offline–online term in Theorem 4.2. Note that, the last term is proportional to ΔK, consistent with the standard linear regret under model misspecification. Furthermore, when $H^3 K/(\tau M^{\mathrm{off}}) = O(1)$, we have $\log\left(1 + \frac{H^3 K}{\tau M^{\mathrm{off}}}\right) = \Theta\left(\frac{H^3 K}{\tau M^{\mathrm{off}}}\right)$. The two additional terms can then be absorbed into the first, so the regret bound above under Δ misspecification exactly recovers the offline–online term in Theorem 4.2.

C.3 Comparison of Definition 4.1 and Assumption 4.8 with other coverage assumptions

C.3.1 Comparison with feature coverage assumptions

Beyond the uniform data coverage assumption (Assumption 4.8) discussed in Section 4.3, feature coverage assumptions, often formulated in a form similar to Definition 4.1, are also widely adopted in offline RL models with linear structure. In this subsection, we compare Definition 4.1 with other feature coverage assumptions used in related models in the literature.

Comparison with Yin et al. [61]. Although Yin et al. [61] study linear MDPs, whose feature space differs from that of linear mixture MDPs, the underlying coverage assumption is essentially analogous to Definition 4.1. Specifically, they assume that, with high probability, $\lambda_{\min}\left(\sum_{m=1}^{M^{\text{off}}} \phi_m \phi_m^\top\right) \geq \tau M^{\text{off}}$, where ϕ_m denotes the feature vector associated with the m-th offline sample (see Assumption 2.2 and Lemma H.5 in Yin et al. [61]). This condition ensures that the empirical design matrix grows linearly in the number of offline samples in every direction in the feature space.

In our linear mixture MDP setting, the role of ϕ_m is played by the induced feature $\phi(\cdot \mid s,a)\widetilde{V}$. While linear MDPs do not require the design of $\{\widetilde{V}_{m,h}\}$ due to their structural properties, both Yin et al. [61] and Definition 4.1 formulate coverage in terms of a lower bound on the minimum eigenvalue of an empirical design matrix. However, the scope of Definition 4.1 is strictly broader. The assumption in Yin et al. [61] enforces a non-trivial bound $\tau = \Theta(1)$, corresponding to the ideal case discussed in Section 4.2. In contrast, Definition 4.1 accommodates more general regimes in which τ may decay as M^{off} increases (Case 2 in Section 4.2).

C.3.2 Comparison with partial data coverage assumptions

Unlike in offline RL, canonical partial data coverage assumptions may not be directly applicable in the offline–online setting due to the presence of environment shift. In this subsection we illustrate this issue using the partial data coverage assumption in Nguyen-Tang et al. [41] (Assumption 4.1 therein).

Concretely, the partial data coverage assumption in Nguyen-Tang et al. [41] states that for any (h,s,a) tuple, $d_h^{\pi^{\text{off},*}}(s,a) > 0 \Rightarrow d_h^{\pi^{\text{off}}}(s,a) > 0$, where $\pi^{\text{off},*}$ denotes an optimal policy of the offline MDP, and $d_h^{\pi^{\text{off},*}}$ denotes the marginal state–action probability measure under $\pi^{\text{off},*}$, defined analogously to $d_h^{\pi^{\text{off}}}$ in Section 4.3.

We now present a simple counterexample showing that, for any M^{off}, K, and $\Delta > 0$, this assumption is not sufficient to guarantee that the offline data is informative according to Definition 4.3. Consider an offline–online tabular bandit with three arms. Suppose the mean rewards of the arms satisfy

$$\mu(1) = 0.5,\ \mu^{\text{off}}(1) = 0.5,\ \mu(2) = 0.5 - \frac{1}{\sqrt{K}},$$
$$\mu^{\text{off}}(2) = 0.5 - \frac{1}{\sqrt{K}}, \mu(3) = 0.5 - \frac{1}{4}\Delta,\ \mu^{\text{off}}(3) = 0.5 + \frac{1}{4}\Delta,$$

where μ and μ^{off} follow the definitions in Section B.1. As discussed in Proposition B.1, this instance satisfies Assumption 2.3.

In the offline environment, arm 3 is optimal. Suppose the offline behavior policy always pulls arm 3. Then the partial coverage assumption above is satisfied, since the optimal offline policy also selects arm 3. However, in the online environment arm 1 is optimal. Since the offline data contains no information about arm 1 and arm 2, the learner must explore arm 2 during the online phase. Consequently, the regret remains of order $\sqrt{K}$ regardless of the value of M^{off}, and the offline data is therefore not informative. This completes the argument, since under the uniform data coverage assumption (Assumption 4.8), the offline data can be informative for appropriate values of M^{off}, K, and $\Delta > 0$, as discussed in Section B.1.

If the partial coverage assumption were instead defined with respect to the optimal policy of the online environment, one might obtain positive guarantees when the offline and online MDPs share the same state and action spaces. However, in general offline–online settings the two environments may differ substantially, and such a definition may not even be well-defined. We leave the investigation of such coverage conditions for future work.

C.4 Discussion on the convexity of Eq. (13)

In this section we provide a counterexample showing that, in general, Eq. (13) is neither concave nor convex with respect to V. Consequently, one cannot expect a closed-form solution or to leverage convex optimization algorithms to solve it in general.

For the counterexample, we consider a bandit embedded linear mixture MDP with $|\mathcal{A}| = 1$. Following the discussion in Section B.1, the corresponding offline–online linear mixture MDP has $|\mathcal{A}| = 1$, $|\mathcal{S}| = 2$, and $H = 2$. Choosing $d = 2$ is sufficient as discussed in Section B.1. The feature mapping satisfies $\phi(\cdot \mid s, a) = I \in \mathbb{R}^{2\times 2}$ for any (s, a).

We show that for the adaptively obtained real symmetric PSD matrix $A_{m,h}$, which corresponds to

$$\sum_{(m',h')<(m,h)} x^{\text{off}}_{m',h'}(\widetilde{V}_{m',h'+1})\big(x^{\text{off}}_{m',h'}(\widetilde{V}_{m',h'+1})\big)^\top$$

in Eq. (13), the problem

$$\min_{V\in[-1,1]^{|\mathcal{S}|}} \lambda_{\min}(A_{m,h}+\phi(\cdot|s_{m,h},a_{m,h})V(\phi(\cdot|s_{m,h},a_{m,h})V)^\top) = \min_{V\in[-1,1]^{|\mathcal{S}|}} \lambda_{\min}(A_{m,h}+VV^\top)$$

is neither concave nor convex with respect to V.

We first consider the step $(m, h) = (1, 1)$. In this case the matrix in Eq. (13) satisfies $A_{1,1} = H^2 dI_2$ following Theorem 4.2. For any $V \in \mathbb{R}^2$, the matrix $VV^\top$ has rank at most one. Therefore no choice of V can increase the minimum eigenvalue of $A_{1,1} + VV^\top$. Thus it's valid to choose $V = (1, 0)^\top$.

After this step, the matrix becomes

$$A_{2,1} = A_{1,1} + VV^\top = \begin{pmatrix} 9 & 0 \\ 0 & 8 \end{pmatrix}.$$

Since we consider a bandit problem, the horizon index can be omitted in the following discussion.

Define $f(V) = \lambda_{\min}\big(A_{2,1} + VV^\top\big)$ and $V \in [-1, 1]^2$. We first show that $f(V)$ is not convex. Let

$$V_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad V_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \bar{V} = \frac{1}{2}(V_1 + V_2) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Then

$$A_{2,1} + V_1V_1^\top = \begin{pmatrix} 10 & 1 \\ 1 & 9 \end{pmatrix}, \qquad A_{2,1} + V_2V_2^\top = \begin{pmatrix} 10 & -1 \\ -1 & 9 \end{pmatrix}.$$

Both matrices have smallest eigenvalue $f(V_1) = f(V_2) = \frac{19-\sqrt{5}}{2}$. On the other hand,

$$A_{2,1} + \bar{V}\bar{V}^\top = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}, \qquad f(\bar{V}) = 9.$$

Hence $f(\bar{V}) > \frac{f(V_1)+f(V_2)}{2}$, which violates convexity.

Next we show that $f(V)$ is not concave. Let

$$W_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad W_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \bar{W} = \frac{1}{2}(W_1 + W_2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then

$$A_{2,1} + W_1W_1^\top = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}, \qquad A_{2,1} + W_2W_2^\top = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}.$$

Hence $f(W_1) = f(W_2) = 9$. On the other hand,

$$A_{2,1} + \bar{W}\bar{W}^\top = A_{2,1} = \begin{pmatrix} 9 & 0 \\ 0 & 8 \end{pmatrix}, \qquad f(\bar{W}) = 8.$$

Therefore $f(\bar{W}) < \frac{f(W_1)+f(W_2)}{2}$, which shows that $f(V)$ is not concave.

C.5 Specializing Theorem 4.2 to the purely offline regime

It is not fully precise to directly conclude from the statement of Theorem 4.2 that the suboptimality gap is of order $1/\sqrt{M^{\mathrm{off}}}$ by simply setting $K = 1$ and $\Delta = 0$. Indeed, the proof of Theorem 4.2 implicitly operates in a regime where K is sufficiently large and several constant terms are absorbed into the $\tilde{O}(\cdot)$ notation (see Eq. (26)).

To make the comparison rigorous, we return to the explicit regret upper bound for any K and M^{off} in Eq. (26):

$$\begin{aligned}\mathsf{Regret}(K) \le \min\Bigg\{ & \mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg[H\sqrt{2\left(\log\left(\frac{1}{\delta}\right)+\frac{1}{2}d\log\left(1+\frac{\lambda_{\max}(G_{\mathrm{off}})}{H^2d}+\frac{HK}{d}\right)\right)} \\ & + H\sqrt{d}B + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\tau(\delta')M^{\mathrm{off}}}}\Bigg] \times \sqrt{d\log\left(1+\frac{H^3K}{\tau(\delta')M^{\mathrm{off}}}\right)} \;\Big|\; \mathcal{E}_2\Bigg], \\ & \mathbb{E}\Bigg[2\sqrt{2H^2K}\left(H\sqrt{2\left(\log\left(\frac{1}{\delta}\right)+\frac{d}{2}\log\left(1+\frac{HK}{d}\right)\right)}+H\sqrt{d}B\right) \\ & \times\sqrt{d\log\left(1+\frac{HK}{d}\right)}\;\Big|\;\mathcal{E}_2\Bigg]\Bigg\} + H^2K(2\delta+\delta') + HK(2\delta+\delta').\end{aligned}$$

Setting $K = 1$ and $\Delta = 0$, and choosing $\delta = \delta' = 1/M^{\mathrm{off}}$, due to $\lambda_{\max}(G_{\mathrm{off}}) \le H^3 dM^{\mathrm{off}}$, we obtain

$$\mathsf{Regret}(K) = \tilde{O}\left(BH^2d\sqrt{\log\left(1+\frac{H^3}{\tau(\delta')M^{\mathrm{off}}}\right)}\right).$$

In particular, under the uniform coverage condition of Proposition 4.10, where $\tau(\delta') = \Theta(1)$, this bound simplifies to $\tilde{O}\left(\frac{1}{\sqrt{M^{\mathrm{off}}}}\right)$, which is directly comparable to the suboptimality gap established in Theorems 7 and 9 of Nguyen-Tang et al. [42].

C.6 Adaptive offline behavior policy and decaying coverage τ

Although the analysis in Section 4.3 mainly focuses on the regime $\tau = \Theta(1)$ under the uniform data coverage assumption (Assumption 4.8), this assumption can be overly idealized, as it essentially requires the offline behavior policies $\{\pi_m^{\mathrm{off}}\}_{m=1}^{M^{\mathrm{off}}}$ to be designed for exploration. In practice, the offline data may instead be generated by learning algorithms that aim to minimize regret in the offline environment. In this case the behavior policy becomes adaptive, and the visitation distribution gradually concentrates around the optimal policy of the offline MDP rather than exploring the state–action space uniformly.

We show that even in this setting the offline data can remain informative under our algorithm O-O UCRL-VTR. This corresponds to Case 2 of Section 4.2, where the coverage parameter τ may decay with M^{off}. To illustrate this phenomenon, we consider an offline–online tabular bandit instance with A arms, where the offline behavior policies follow the UCB1 algorithm [6]. In this instance, let $\mu^{\mathrm{off}}(i)$ and $\mu(i)$ denote the expected rewards of arm $i \in [A]$ in the offline and online environments, respectively. In the offline environment, the reward obtained at $m \in [M^{\mathrm{off}}]$ by pulling arm $i \in [A]$ follows $\mu^{\mathrm{off}}(i) + \xi_m$, where $\{\xi_m\}_{m\in[M^{\mathrm{off}}]}$ are i.i.d. $N(0,1)$ noises. The reward model in the online environment is defined analogously. Let $\mu^{\mathrm{off},*}$ and μ^* denote the expected rewards of the optimal arms in the offline and online environments, respectively; the optimal arms in the two environments may differ. Define

$$\Delta_{\mathrm{gap}}^{\mathrm{off}} = \sup_{i\in[A]} \mu^{\mathrm{off},*} - \mu^{\mathrm{off}}(i)$$

as the maximum suboptimality gap in the offline environment. Such an instance admits an equivalent offline–online linear mixture MDP reformulation (see Remark 2.2). For this reformulation, we assume that Assumption 2.3 holds, namely that the environment shift is upper bounded by Δ and is known. The following corollary characterizes the scaling of τ and shows that the offline data can

remain informative under appropriate values of M^{off}, K, and Δ, whose admissible values depend on $\Delta_{\text{gap}}^{\text{off}}$.

Proposition C.5. *Consider the aforementioned offline–online tabular bandit setting, where the offline behavior policies follow the* UCB1 *algorithm with exploration rate* $\gamma_{M^{\text{off}}} = \sqrt{2\log(M^{\text{off}})}$ *(See Algorithm 1 of Han et al. [27]). If* $K = o((\log(M^{\text{off}}))^{49})$,[5] *and there exists* $p \in (0, \frac{1}{2}]$ *such that* $\Delta_{gap}^{\text{off}} = O((M^{\text{off}})^{-p})$, *then under the deterministic design of* $\{\widetilde{V}_{m,h}\}$ *(Definition 4.11), running* O-O UCRL-VTR *yields that* $\tau = \Omega((M^{\text{off}})^{2p-1})$. *Furthermore, if there exists* $\epsilon \in (0, \frac{1}{2}]$ *such that*

$$(M^{\text{off}})^{2p} = \Omega(K^{1+2\epsilon}), \qquad \Delta = O\left(\frac{(M^{\text{off}})^{2p-1}}{K^{\frac{1}{2}+\epsilon}}\right), \tag{23}$$

then the offline data is informative.

The proof of Proposition C.5 is provided in Section D.3. When $\Delta_{\text{gap}}^{\text{off}}$ decreases, meaning that the expected rewards of different arms in the offline bandit problem become closer, UCB1 explores suboptimal arms more frequently, which leads to a larger effective coverage parameter τ. As a result, a larger value of Δ can be tolerated while the offline data remains informative. This illustrates the relation between τ and Δ in Figure 1 when the regime $\tau = \Theta(1)$ no longer holds. Note that $\Delta_{\text{gap}}^{\text{off}}$ affects the online phase in two ways. A smaller $\Delta_{\text{gap}}^{\text{off}}$ (together with a small Δ) makes the online bandit problem harder due to smaller suboptimality gaps, while it also induces more exploration in the offline phase under UCB1 and leads to a larger τ, making the offline data more informative. Intuitively, this reveals a somewhat counterintuitive phenomenon: harder online instances may allow the offline data to provide greater benefit.

C.7 Regret lower bound when $\tau = o(1)$

We discuss the lower bound in the regime $\tau = o(1)$ and show that this case leads to a different regret lower bound, as the dependence on d in Theorem 4.13 no longer applies. To illustrate, consider an offline–online bandit problem with A arms (see Section B.1) with no environment shift (so $d = \Theta(A)$). Suppose that all but one arm are well-covered in the offline data, while the remaining arm is never explored in the offline phase. In this case, we have $\tau = O\big(\frac{1}{M^{\text{off}}}\big)$ and $\lambda_{\min}(\mathbb{E}[G_{\text{off}}]) = O(1)$, as discussed in Definition 4.1. Substituting these quantities and $\Delta = 0$ into Eq. (14) would yield a lower bound of order $dH\sqrt{K}$ from Theorem 4.13, which is incorrect; In this instance, the problem reduces to identifying a single poorly explored arm in the online phase, and the regret need not scale with d. This example highlights that, when $\tau = o(1)$, the lower bound is governed more by the geometry of the offline data, as captured by $\lambda_{\min}(\mathbb{E}[G_{\text{off}}])$, rather than by the dimension d. However, the relation $\lambda_{\min}(\mathbb{E}[G_{\text{off}}]) = \Theta(M^{\text{off}})$ no longer holds in this regime. In principle, one can derive lower bounds for $\tau = o(1)$ following the same framework as in the proof of Theorem 4.13, under additional conditions on the offline behavior policy that characterize $\lambda_{\min}(\mathbb{E}[G_{\text{off}}])$. We do not pursue this direction here, as it requires a detailed, instance-dependent characterization of the behavior policy, which is beyond the scope of this paper.

D Proofs

D.1 Proofs in Section 4.1

This section provides a complete proof of Theorem 4.2.

Theorem 4.2. *Under Assumptions 2.3 and Definition 4.1, let* $\lambda = H^2 d$ *and define* $\gamma_k = H\sqrt{2\log\Big(\frac{2\det(M_k^{\text{all}})^{1/2}K}{\det(\lambda I)^{1/2}}\Big)} + \Delta\frac{\lambda_{\max}(G_{\text{off}})}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} + \sqrt{\lambda}B$, $\beta_k = H\sqrt{2\log\Big(\frac{2\det(M_k^{\text{on}})^{1/2}K}{\det(\lambda I)^{1/2}}\Big)} + \sqrt{\lambda}B$. *Then, when running Algorithm 1, the expected regret* Regret(K) *is upper bounded by*

$$\tilde{O}\left(\min\left\{\left[BH^2d\sqrt{K} + H^4d^{\frac{3}{2}}\Delta\sqrt{K}\sqrt{\frac{M^{\text{off}}}{\tau}}\right]\cdot\sqrt{\log\left(1+\frac{H^3K}{\tau M^{\text{off}}}\right)}, BH^2d\sqrt{K}\right\}\right), \tag{12}$$

[5] This condition arises from the high-probability bound on the number of pulls in a bandit model provided by Proposition 3.7 of Han et al. [27]. The exponent 49 is not intrinsic to our framework and results from matching the failure probability with the confidence level in the definition of τ. A sharper analysis could potentially relax this requirement.

where $\tilde{O}(\cdot)$ hides logarithmic factors in K and M^{off}.

Before presenting the proof of Theorem 4.2, we introduce a filtration that will play a central role in the analysis. The filtration jointly accounts for the offline and online phases. To this end, we introduce a unified indexing scheme that orders all samples from both phases. Specifically, consider the collection of offline indices $\mathcal{I}_{\mathrm{off}} := \{(m,h) : m = 1,\dots,M^{\mathrm{off}}; h = 1,\dots,H\}$, and the collection of online indices $\mathcal{I}_{\mathrm{on}} := \{(k,h) : k = 1,\dots,K; h = 1,\dots,H\}$. We consider the unique one-one mapping $t \longleftrightarrow (i_t, h_t) \in \mathcal{I}_{\mathrm{off}} \cup \mathcal{I}_{\mathrm{on}}$ that orders all samples, such that all offline indices in $\mathcal{I}_{\mathrm{off}}$ precede those in $\mathcal{I}_{\mathrm{on}}$. Note that the mapping is unique as (i) within each trajectory, samples are indexed by $h = 1,\dots,H$; (ii) online trajectories are indexed by $k = 1,\dots,K$; and (iii) offline trajectories are indexed by $m = 1,\dots,M^{\mathrm{off}}$, which induces a unique order due to the adaptive behavior policy.

For each time index t, we define the filtration $\mathcal{F}_t$ as follows.

- If t corresponds to an offline index $(m,h) \in \mathcal{I}_{\mathrm{off}}$, then $\mathcal{F}_t$ is the σ-algebra generated by the offline observations up to and including (m,h), that is,
$$\mathcal{F}_t = \sigma\big(\{(S^{\mathrm{off}}_{m',h'}, A^{\mathrm{off}}_{m',h'}, r^{\mathrm{off}}(S^{\mathrm{off}}_{m',h'}, A^{\mathrm{off}}_{m',h'})) : m' < m \text{ or } (m' = m, h' \le h)\}\big).$$
- If t corresponds to an online index $(k,h) \in \mathcal{I}_{\mathrm{on}}$, then $\mathcal{F}_t$ is the σ-algebra generated by the entire offline dataset together with the online observations up to and including stage h of episode k, namely
$$\mathcal{F}_t = \sigma\big(\{\mathcal{H}^{\mathrm{off}}_m\}_{m=1}^{M^{\mathrm{off}}}, \{(s^{k'}_{h'}, a^{k'}_{h'}, r(s^{k'}_{h'}, a^{k'}_{h'})) : k' < k \text{ or } (k' = k, h' \le h)\}\big).$$

We use $\mathcal{F}_t$ and $\mathcal{F}_{m,h}$ (or $\mathcal{F}_{k,h}$) interchangeably when there is no ambiguity, since the indices are in one-to-one correspondence.

To prove Theorem 4.2, we first establish a key concentration result for the value-targeted regression introduced in Section 3. Specifically, we rely on a self-normalized concentration inequality adapted from Abbasi-Yadkori et al. [1].

Lemma D.1. *For any $\delta \in (0,1)$, with probability at least $1-2\delta$, for all $k \ge 1$,*
$$\big\|\widehat{\theta}^{\mathrm{all}}_k - \theta^*\big\|_{M^{\mathrm{all}}_k} \le \gamma_k, \qquad \big\|\widehat{\theta}^{\mathrm{on}}_k - \theta^*\big\|_{M^{\mathrm{on}}_k} \le \beta_k,$$
where
$$\gamma_k = H\sqrt{2\log\Big(\frac{\det(M^{\mathrm{all}}_k)^{1/2}}{\det(\lambda I)^{1/2}\delta}\Big)} + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(\lambda I + G_{\mathrm{off}})}} + \sqrt{\lambda}B,$$
$$\beta_k = H\sqrt{2\log\Big(\frac{\det(M^{\mathrm{on}}_k)^{1/2}}{\det(\lambda I)^{1/2}\delta}\Big)} + \sqrt{\lambda}B.$$
Equivalently, define the good event
$$\mathcal{E}_0 = \big\{\theta^* \in \mathcal{B}_k \cap \mathcal{C}_k, \text{ for all } k \ge 1\big\},$$
where $\mathcal{B}_k = \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}^{\mathrm{all}}_k\|_{M^{\mathrm{all}}_k} \le \gamma_k\Big\}, \mathcal{C}_k = \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}^{\mathrm{on}}_k\|_{M^{\mathrm{on}}_k} \le \beta_k\Big\}$. Then the event $\mathcal{E}_0$ holds with probability at least $1-2\delta$. Also, on $\mathcal{E}_0$, we have $\mathcal{D}_k = \mathcal{B}_k \cap \mathcal{C}_k$ for all $k \ge 1$.

Proof. We discuss the offline–online estimator first. Note that M^{all}_k is symmetric and satisfies $M^{\mathrm{all}}_k \succeq \lambda I$, hence it is positive definite and invertible. Therefore, using the fact that $M^{\mathrm{all}}_k = M^{\mathrm{on}}_k + G_{\mathrm{off}}$,
$$\begin{aligned}
M^{\mathrm{all}}_k\big(\widehat{\theta}^{\mathrm{all}}_k - \theta^*\big) &= w^{\mathrm{all}}_k - M^{\mathrm{all}}_k\theta^* = \big(w^{\mathrm{all}}_k - M^{\mathrm{on}}_k\theta^*\big) - G_{\mathrm{off}}\theta^* \\
&= \big(w^{\mathrm{all}}_k - M^{\mathrm{on}}_k\theta^*\big) - G_{\mathrm{off}}\theta^{\mathrm{off},*} + (G_{\mathrm{off}}\theta^{\mathrm{off},*} - G_{\mathrm{off}}\theta^*) \\
&= \sum_{k'=1}^{k-1}\sum_{h=1}^{H} x^{\mathrm{on}}_{k',h}\Big(y^{\mathrm{on}}_{k',h} - \langle x^{\mathrm{on}}_{k',h}, \theta^*\rangle\Big) - \lambda\theta^* + \sum_{m=1}^{M^{\mathrm{off}}}\sum_{h=1}^{H} x^{\mathrm{off}}_{m,h}\Big(y^{\mathrm{off}}_{m,h} - \langle x^{\mathrm{off}}_{m,h}, \theta^{\mathrm{off},*}\rangle\Big) \\
&\quad + \sum_{m=1}^{M^{\mathrm{off}}}\sum_{h=1}^{H} x^{\mathrm{off}}_{m,h}\big(x^{\mathrm{off}}_{m,h}\big)^\top(\theta^{\mathrm{off},*} - \theta^*).
\end{aligned}$$

Define

$$\eta^{\mathrm{on}}_{k',h} = y^{\mathrm{on}}_{k',h} - \langle x^{\mathrm{on}}_{k',h}, \theta^* \rangle, \eta^{\mathrm{off}}_{m,h} = y^{\mathrm{off}}_{m,h} - \langle x^{\mathrm{off}}_{m,h}, \theta^{\mathrm{off},*} \rangle, b^{\mathrm{off}} = \sum_{m=1}^{M^{\mathrm{off}}} \sum_{h=1}^{H} x^{\mathrm{off}}_{m,h} \big(x^{\mathrm{off}}_{m,h}\big)^\top (\theta^{\mathrm{off},*} - \theta^*),$$

and let

$$S_k = \sum_{m=1}^{M^{\mathrm{off}}} \sum_{h=1}^{H} x^{\mathrm{off}}_{m,h} \eta^{\mathrm{off}}_{m,h} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} x^{\mathrm{on}}_{k',h} \eta^{\mathrm{on}}_{k',h}.$$

Here both $\{\eta^{\mathrm{on}}_{k',h}\}$ and $\{\eta^{\mathrm{off}}_{m,h}\}$ form martingale difference sequences with respect to the filtration $\{\mathcal{F}_t\}$, and S_k collects the corresponding self-normalized noise terms that can be controlled via a self-normalized concentration inequality. Then the decomposition becomes $M^{\mathrm{all}}_k \big(\widehat{\theta}^{\mathrm{all}}_k - \theta^*\big) = S_k + b^{\mathrm{off}} - \lambda\theta^*$. For the online-only estimator, following the same approach, we have that

$$M^{\mathrm{on}}_k \big(\widehat{\theta}^{\mathrm{on}}_k - \theta^*\big) = \sum_{k'=1}^{k-1} \sum_{h=1}^{H} x^{\mathrm{on}}_{k',h} \eta^{\mathrm{on}}_{k',h} - \lambda\theta^*.$$

Since $0 \le \widehat{V}_{k',h}, \widetilde{V}_{m,h} \le H$, for any h, k', m, both $\eta^{\mathrm{on}}_{k',h}$ and $\eta^{\mathrm{off}}_{m,h}$ are uniformly bounded almost surely:

$$|\eta^{\mathrm{on}}_{k',h}| \le H, \qquad |\eta^{\mathrm{off}}_{m,h}| \le H \quad \text{a.s..}$$

Recall the definition of the filtration $\mathcal{F}_{k,h}$ and $\mathcal{F}_{m,h}$, we correspondingly have that $\mathbb{E}\Big[\eta^{\mathrm{on}}_{k',h} \mid \mathcal{F}_{k',h}\Big] = 0, \mathbb{E}\Big[\eta^{\mathrm{off}}_{m,h} \mid \mathcal{F}_{m,h}\Big] = 0$, for all k', h, m. Then all requirements of Lemma E.1 (see Section E) are satisfied. By applying Lemma E.1 to the offline–online estimator, we have that with probability at least $1-\delta$, for all $k \ge 1$,

$$\|S_k\|_{(M^{\mathrm{all}}_k)^{-1}} \le H \sqrt{2 \log\Big(\frac{\det(M^{\mathrm{all}}_k)^{1/2}}{\det(\lambda I)^{1/2}\delta}\Big)}.$$

Similarly, by applying Lemma E.1 to the online-only estimator, we have that with probability at least $1-\delta$,

$$\Big\| \sum_{k'=1}^{k-1} \sum_{h=1}^{H} x^{\mathrm{on}}_{k',h} \eta^{\mathrm{on}}_{k',h} \Big\|_{(M^{\mathrm{on}}_k)^{-1}} \le H \sqrt{2 \log\Big(\frac{\det(M^{\mathrm{on}}_k)^{1/2}}{\det(\lambda I)^{1/2}\delta}\Big)}.$$

Since M^{all}_k and M^{on}_k are positive definite and symmetric, $\|\cdot\|_{M^{\mathrm{all}}_k}$ and $\|\cdot\|_{M^{\mathrm{on}}_k}$ are well-defined norms. Thus by the triangle inequality,

$$\|\widehat{\theta}^{\mathrm{all}}_k - \theta^*\|_{M^{\mathrm{all}}_k} = \|M^{\mathrm{all}}_k(\widehat{\theta}^{\mathrm{all}}_k - \theta^*)\|_{(M^{\mathrm{all}}_k)^{-1}} \le \|S_k\|_{(M^{\mathrm{all}}_k)^{-1}} + \|b^{\mathrm{off}}\|_{(M^{\mathrm{all}}_k)^{-1}} + \|\lambda\theta^*\|_{(M^{\mathrm{all}}_k)^{-1}},$$

also,

$$\|\widehat{\theta}^{\mathrm{on}}_k - \theta^*\|_{M^{\mathrm{on}}_k} = \|M^{\mathrm{on}}_k(\widehat{\theta}^{\mathrm{on}}_k - \theta^*)\|_{(M^{\mathrm{on}}_k)^{-1}} \le \Big\| \sum_{k'=1}^{k-1} \sum_{h=1}^{H} x^{\mathrm{on}}_{k',h} \eta^{\mathrm{on}}_{k',h} \Big\|_{(M^{\mathrm{on}}_k)^{-1}} + \|\lambda\theta^*\|_{(M^{\mathrm{on}}_k)^{-1}}.$$

Furthermore, by definition,

$$\|b^{\mathrm{off}}\|_{(M^{\mathrm{all}}_k)^{-1}} \le \|b^{\mathrm{off}}\|_{(M^{\mathrm{all}}_1)^{-1}} \le \sqrt{\lambda_{\max}\big((M^{\mathrm{all}}_1)^{-1}\big)}\|b^{\mathrm{off}}\|_2 = \frac{1}{\sqrt{\lambda_{\min}(M^{\mathrm{all}}_1)}}\|b^{\mathrm{off}}\|_2,$$

where the first inequality follows from $M^{\mathrm{all}}_k \succeq M^{\mathrm{all}}_1$ and Lemmas E.3 and E.4 (see Section E), which together imply $(M^{\mathrm{all}}_k)^{-1} \preceq (M^{\mathrm{all}}_1)^{-1}$, the second inequality is from Lemma E.5 (Section E), and the final equality follows from the spectral identity $\lambda_{\max}(A^{-1}) = 1/\lambda_{\min}(A)$ for any symmetric positive definite matrix. Thus

$$\begin{aligned}
\|b^{\mathrm{off}}\|_{(M^{\mathrm{all}}_k)^{-1}} &\le \frac{1}{\sqrt{\lambda_{\min}(M^{\mathrm{all}}_1)}}\|b^{\mathrm{off}}\|_2 \le \frac{1}{\sqrt{\lambda_{\min}(M^{\mathrm{all}}_1)}}\lambda_{\max}(G_{\mathrm{off}})\|\theta^{\mathrm{off},*} - \theta^*\|_2 \\
&\le \Delta \frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(\lambda I + G_{\mathrm{off}})}},
\end{aligned}$$

where the first inequality again follows from Lemma E.5, noting that the bound holds for any symmetric positive semi-definite matrix, even though the induced quantity $\|\cdot\|_A$ is only a seminorm when A is not strictly positive definite. The last inequality follows directly from Assumption 2.3, which ensures $\|\theta^{\mathrm{off},*} - \theta^*\|_2 \le \Delta$, together with the definition of M_1^{all}.

Finally, since $M_k^{\mathrm{all}} \succeq \lambda I$ and $M_k^{\mathrm{on}} \succeq \lambda I$ for all k,

$$\|\lambda\theta^*\|_{(M_k^{\mathrm{all}})^{-1}} \le \sqrt{\lambda}\|\theta^*\|_2 \le \sqrt{\lambda}B, \|\lambda\theta^*\|_{(M_k^{\mathrm{on}})^{-1}} \le \sqrt{\lambda}\|\theta^*\|_2 \le \sqrt{\lambda}B,$$

where we use the bound $\max\{\|\theta^{\mathrm{off},*}\|_2, \|\theta^*\|_2\} \le B$ from Definition 2.1. The proof is complete with a union bound. □

With the self-normalized concentration bounds Lemma D.1 in place, we next establish the key optimism property of the value estimates produced by Algorithm 1, which is crucial for the subsequent regret analysis.

Lemma D.2. *Conditional on $\mathcal{E}_0$, for any episode k, the online optimistic Q-function and value function satisfy*

$$Q_h^{\pi^*}(s,a) \le \widehat{Q}_{k,h}(s,a), \qquad V_h^{\pi^*}(s) \le \widehat{V}_{k,h}(s),$$

for all states s, actions a, and stages $h = 1, \dots, H+1$.

Proof. We prove the result by induction. For the base case $h = H+1$, by definition,

$$Q_{H+1}^{\pi^*}(s,a) = 0, \qquad \widehat{Q}_{k,H+1}(s,a) = 0,$$

for all s, a. Hence the claim holds trivially. Assume that for some $h+1 \le H$,

$$Q_{h+1}^{\pi^*}(s,a) \le \widehat{Q}_{k,h+1}(s,a) \quad \text{for all } s,a.$$

Then, by the definition of the value function,

$$V_{h+1}^{\pi^*}(s) = \max_{a\in\mathcal{A}} Q_{h+1}^{\pi^*}(s,a) \le \max_{a\in\mathcal{A}} \widehat{Q}_{k,h+1}(s,a) = \widehat{V}_{k,h+1}(s),$$

for all states s. Furthermore, for any state–action pair (s,a), we have

$$Q_h^{\pi^*}(s,a) = r(s,a) + P(\cdot \mid s,a)^\top V_{h+1}^{\pi^*}.$$

By the linear mixture model,

$$P(\cdot \mid s,a) = \langle \theta^*, \phi(\cdot \mid s,a)\rangle,$$

hence

$$P(\cdot \mid s,a)^\top V_{h+1}^{\pi^*} = \sum_{j=1}^d \theta_j^* \phi_j(\cdot \mid s,a)^\top V_{h+1}^{\pi^*} \le \sum_{j=1}^d \theta_j^* \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}.$$

Since $\theta^* \in \mathcal{B}_k \cap \mathcal{C}_k$, we further have via the induction hypothesis that

$$Q_h^{\pi^*}(s,a) = r(s,a) + P(\cdot \mid s,a)^\top V_{h+1}^{\pi^*} \le r(s,a) + \sum_{j=1}^d \theta_j^* \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}$$

$$\le r(s,a) + \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k} \sum_{j=1}^d \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1} = \widehat{Q}_{k,h}(s,a).$$

Taking the maximum over a on both sides gives $V_h^{\pi^*}(s) \le \widehat{V}_{k,h}(s)$, which completes the proof. □

We next bound the gap between the optimistic value estimates produced by the algorithm and the true value functions induced by the executed policy π_k. The analysis follows a standard telescoping argument along the trajectory within an episode: the gap at stage h is related to the gap at stage $h+1$, plus a martingale difference term accounting for the randomness of the next state. The remaining term is controlled by the uncertainty in the transition parameter, which is quantified through the confidence radii associated with the sets $\mathcal{C}_k$ and $\mathcal{B}_k$.

Lemma D.3. *Let π_k be the implemented policy at episode k. Then conditional on $\mathcal{E}_0$ and $\mathcal{F}_{k,h+1}$,*

$$\widehat{V}_{k,h}\big(s^{\mathrm{on}}_{k,h}\big) - V^{\pi_k}_h\big(s^{\mathrm{on}}_{k,h}\big) \le \widehat{V}_{k,h+1}\big(s^{\mathrm{on}}_{k,h+1}\big) - V^{\pi_k}_{h+1}\big(s^{\mathrm{on}}_{k,h+1}\big) + \xi^{\mathrm{on}}_{k,h+1} + 2\min\left\{\gamma_k\left\|\phi(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})\widehat{V}_{k,h+1}\right\|_{(M^{\mathrm{all}}_k)^{-1}}, \beta_k\left\|\phi(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})\widehat{V}_{k,h+1}\right\|_{(M^{\mathrm{on}}_k)^{-1}}\right\},$$

where

$$\xi^{\mathrm{on}}_{k,h+1} = P(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})^\top\big(\widehat{V}_{k,h+1} - V^{\pi_k}_{h+1}\big) - \Big(\widehat{V}_{k,h+1}(s^{\mathrm{on}}_{k,h+1}) - V^{\pi_k}_{h+1}(s^{\mathrm{on}}_{k,h+1})\Big).$$

Proof. Recall that $a^{\mathrm{on}}_{k,h} = \pi_{k,h}(s^{\mathrm{on}}_{k,h})$ and $V^{\pi_k}_h(s) = Q^{\pi_k}_h(s, \pi_{k,h}(s))$. Hence

$$\widehat{V}_{k,h}\big(s^{\mathrm{on}}_{k,h}\big) - V^{\pi_k}_h\big(s^{\mathrm{on}}_{k,h}\big) = \widehat{Q}_{k,h}\big(s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h}\big) - Q^{\pi_k}_h\big(s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h}\big).$$

Recall also that

$$x^{\mathrm{on}}_{k,h} = \phi(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})\widehat{V}_{k,h+1}.$$

By the definition of $\widehat{Q}_{k,h}$ and the linear mixture model,

$$\begin{aligned}
&\widehat{Q}_{k,h}\big(s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h}\big) - Q^{\pi_k}_h\big(s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h}\big) \\
&\quad= r\big(s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h}\big) + \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}\theta^\top x^{\mathrm{on}}_{k,h} - \Big(r\big(s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h}\big) + (\theta^*)^\top\phi(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})V^{\pi_k}_{h+1}\Big) \\
&\quad= (\theta^*)^\top x^{\mathrm{on}}_{k,h} - (\theta^*)^\top\phi(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})V^{\pi_k}_{h+1} + \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}(\theta-\theta^*)^\top x^{\mathrm{on}}_{k,h} \\
&\quad= P(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})^\top\big(\widehat{V}_{k,h+1} - V^{\pi_k}_{h+1}\big) + \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}(\theta-\theta^*)^\top x^{\mathrm{on}}_{k,h},
\end{aligned}$$

where the first equality is from the definition of $\widehat{Q}_{k,h}$. Since $\theta^* \in \mathcal{B}_k\cap\mathcal{C}_k$, $\mathcal{B}_k = \{\theta : \|\theta-\widehat{\theta}^{\mathrm{all}}_k\|_{M^{\mathrm{all}}_k} \le \gamma_k\}$, and $\mathcal{C}_k = \big\{\theta\in\mathbb{R}^d : \|\theta - \widehat{\theta}^{\mathrm{on}}_k\|_{M^{\mathrm{on}}_k} \le \beta_k\big\}$,

$$\begin{aligned}
\max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}(\theta-\theta^*)^\top x^{\mathrm{on}}_{k,h} &\le \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}\big|(\theta-\widehat{\theta}^{\mathrm{all}}_k)^\top x^{\mathrm{on}}_{k,h}\big| + \big|(\theta^*-\widehat{\theta}^{\mathrm{all}}_k)^\top x^{\mathrm{on}}_{k,h}\big| \\
&\le \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}\|(M^{\mathrm{all}}_k)^{\frac12}(\theta-\widehat{\theta}^{\mathrm{all}}_k)\|_2\|(M^{\mathrm{all}}_k)^{-\frac12}x^{\mathrm{on}}_{k,h}\|_2 \\
&\quad+ \|(M^{\mathrm{all}}_k)^{\frac12}(\theta^*-\widehat{\theta}^{\mathrm{all}}_k)\|_2\|(M^{\mathrm{all}}_k)^{-\frac12}x^{\mathrm{on}}_{k,h}\|_2 \\
&\le \gamma_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}} + \gamma_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}} = 2\gamma_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}}.
\end{aligned}$$

where the second inequality is from Cauchy-Schwarz inequality. Similarly, we have that

$$\begin{aligned}
\max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}(\theta-\theta^*)^\top x^{\mathrm{on}}_{k,h} &\le \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}\big|(\theta-\widehat{\theta}^{\mathrm{on}}_k)^\top x^{\mathrm{on}}_{k,h}\big| + \big|(\theta^*-\widehat{\theta}^{\mathrm{on}}_k)^\top x^{\mathrm{on}}_{k,h}\big| \\
&\le \max_{\theta\in\mathcal{B}_k\cap\mathcal{C}_k}\|(M^{\mathrm{on}}_k)^{\frac12}(\theta-\widehat{\theta}^{\mathrm{on}}_k)\|_2\|(M^{\mathrm{on}}_k)^{-\frac12}x^{\mathrm{on}}_{k,h}\|_2 \\
&\quad+ \|(M^{\mathrm{on}}_k)^{\frac12}(\theta^*-\widehat{\theta}^{\mathrm{on}}_k)\|_2\|(M^{\mathrm{on}}_k)^{-\frac12}x^{\mathrm{on}}_{k,h}\|_2 \\
&\le \beta_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{on}}_k)^{-1}} + \beta_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{on}}_k)^{-1}} = 2\beta_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{on}}_k)^{-1}}.
\end{aligned}$$

The proof is complete with the definition of $\xi^{\mathrm{on}}_{k,h+1}$. □

$\sum_{k=1}^K\sum_{h=1}^H\gamma_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}}$ and $\sum_{k=1}^K\sum_{h=1}^H\beta_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{on}}_k)^{-1}}$ can be bounded with the next lemma to further simplify the results in Lemma D.3.

Lemma D.4. *For all $k \geq 1$, with $\lambda = H^2 d$, we have that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\gamma_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}} \leq \sqrt{2H^2K}\Bigg(H\sqrt{2\log\Big(\frac{\det(M^{\mathrm{all}}_K)^{1/2}}{\det(H^2dI)^{1/2}\delta}\Big)} + \sqrt{H^2d}B + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(H^2dI + G_{\mathrm{off}})}}\Bigg) \times \sqrt{\Big(\log(\det(M^{\mathrm{all}}_{K+1})) - \log(\det(M^{\mathrm{all}}_1))\Big)},$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{on}}_k)^{-1}} \leq \sqrt{2H^2K}\Bigg(H\sqrt{2\log\Big(\frac{\det(M^{\mathrm{on}}_K)^{1/2}}{\det(H^2dI)^{1/2}\delta}\Big)} + \sqrt{H^2d}B\Bigg) \times \sqrt{\Big(\log(\det(M^{\mathrm{on}}_{K+1})) - \log(\det(H^2dI))\Big)}.$$

Proof. Note that due to

$$M^{\mathrm{all}}_k = \lambda I + \sum_{m=1}^{M^{\mathrm{off}}}\sum_{h=1}^{H} x^{\mathrm{off}}_{m,h}\big(x^{\mathrm{off}}_{m,h}\big)^\top + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} x^{\mathrm{on}}_{k',h}\big(x^{\mathrm{on}}_{k',h}\big)^\top,\ M^{\mathrm{on}}_k = \lambda I + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} x^{\mathrm{on}}_{k',h}\big(x^{\mathrm{on}}_{k',h}\big)^\top,$$

we have that

$$\begin{aligned}
\|x^{\mathrm{on}}_{k,h}\|^2_{(M^{\mathrm{all}}_k)^{-1}} &= (x^{\mathrm{on}}_{k,h})^\top (M^{\mathrm{all}}_k)^{-1} x^{\mathrm{on}}_{k,h} \leq (x^{\mathrm{on}}_{k,h})^\top (\lambda I)^{-1} x^{\mathrm{on}}_{k,h} = \frac{1}{\lambda}\|x^{\mathrm{on}}_{k,h}\|_2^2 \\
&= \frac{1}{\lambda}\sum_{j=1}^{d}\big(\phi_j(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})^\top \widehat{V}_{k,h+1}\big)^2 \\
&\leq \frac{1}{\lambda}\sum_{j=1}^{d}\|\phi_j(\cdot \mid s^{\mathrm{on}}_{k,h}, a^{\mathrm{on}}_{k,h})\|_1^2 \|\widehat{V}_{k,h+1}\|_\infty^2 \leq \frac{H^2 d}{\lambda} = 1, \qquad (24)
\end{aligned}$$

where the first inequality is from Lemma E.3, and the last inequality is from the assumption that $\|\phi_j(\cdot \mid s, a)\|_1 \leq 1$ for all (s, a, j). Also,

$$\|x^{\mathrm{on}}_{k,h}\|^2_{(M^{\mathrm{on}}_k)^{-1}} = (x^{\mathrm{on}}_{k,h})^\top (M^{\mathrm{on}}_k)^{-1} x^{\mathrm{on}}_{k,h} \leq (x^{\mathrm{on}}_{k,h})^\top (\lambda I)^{-1} x^{\mathrm{on}}_{k,h} = \frac{1}{\lambda}\|x^{\mathrm{on}}_{k,h}\|_2^2 \leq \frac{H^2 d}{\lambda} = 1.$$

Thus by denoting

$$\gamma_{k,1} = H\sqrt{2\log\Big(\frac{\det(M^{\mathrm{all}}_k)^{1/2}}{\det(\lambda I)^{1/2}\delta}\Big)} + \sqrt{\lambda}B, \qquad \gamma_\Delta = \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(\lambda I + G_{\mathrm{off}})}},$$

it's easy to see that $\gamma_k = \gamma_{k,1} + \gamma_\Delta$, and (1) $\gamma_{k,1}$ is non-decreasing w.r.t. k; (2) γ_Δ is a constant after observing all offline data. Thus,

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\gamma_k\|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}} &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\gamma_k \min\{1, \|x^{\mathrm{on}}_{k,h}\|_{(M^{\mathrm{all}}_k)^{-1}}\} \\
&\leq (\gamma_{K,1} + \gamma_\Delta)\sqrt{HK}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \|x^{\mathrm{on}}_{k,h}\|^2_{(M^{\mathrm{all}}_k)^{-1}}\}} \\
&\leq (\gamma_{K,1} + \gamma_\Delta)\sqrt{2HK}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\log(1 + \|x^{\mathrm{on}}_{k,h}\|^2_{(M^{\mathrm{all}}_k)^{-1}})} \\
&\leq (\gamma_{K,1} + \gamma_\Delta)\sqrt{2HK}\sqrt{\sum_{k=1}^{K} H\log(1 + \sum_{h=1}^{H}\|x^{\mathrm{on}}_{k,h}\|^2_{(M^{\mathrm{all}}_k)^{-1}})},
\end{aligned}$$

where the first inequality is from Eq. (24), the second inequality is from Cauchy-Schwarz inequality and two arguments over γ_k above, the third inequality is from that $\min\{1, x\} \leq 2\log(1+x)$ for all $x \geq 0$, and the last inequality is from that for any $\{x_h\}_{h=1}^H$ with $x_h \geq 0$ for all $h \in [H]$,

$$1 + \sum_{h=1}^{H} x_h = \frac{\sum_{h=1}^{H}(1+Hx_h)}{H} \geq \Pi_{h=1}^{H}(1+Hx_h)^{1/H} \geq \Pi_{h=1}^{H}(1+x_h)^{1/H},$$

thus we have that $\sum_{h=1}^{H}\log(1+x_h) \leq H\log(1+\sum_{h=1}^{H}x_h)$. The same argument applies to the term $\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_k\|x_{k,h}^{\mathrm{on}}\|_{(M_k^{\mathrm{on}})^{-1}}$, and we have that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_k\|x_{k,h}^{\mathrm{on}}\|_{(M_k^{\mathrm{on}})^{-1}} \leq \beta_K\sqrt{2HK}\sqrt{\sum_{k=1}^{K}H\log(1+\sum_{h=1}^{H}\|x_{k,h}^{\mathrm{on}}\|^2_{(M_k^{\mathrm{on}})^{-1}})}.$$

Furthermore, note that

$$\begin{aligned}1 + \sum_{h=1}^{H}\|x_{k,h}^{\mathrm{on}}\|^2_{(M_k^{\mathrm{all}})^{-1}} &= 1 + \mathrm{tr}(\sum_{h=1}^{H}(M_k^{\mathrm{all}})^{-\frac{1}{2}}x_{k,h}^{\mathrm{on}}(x_{k,h}^{\mathrm{on}})^\top(M_k^{\mathrm{all}})^{-\frac{1}{2}}) \\ &\leq \det(I + \sum_{h=1}^{H}(M_k^{\mathrm{all}})^{-\frac{1}{2}}x_{k,h}^{\mathrm{on}}(x_{k,h}^{\mathrm{on}})^\top(M_k^{\mathrm{all}})^{-\frac{1}{2}}) = \frac{\det(M_{k+1}^{\mathrm{all}})}{\det(M_k^{\mathrm{all}})},\end{aligned}$$

where the inequality is due to the fact that $\Pi_i(1+x_i) \geq 1 + \sum_i x_i$ for all $x \geq 0$, and the last equality is due to the definition of M_k^{all}. Similarly, for the online-only estimator, we have

$$1 + \sum_{h=1}^{H}\|x_{k,h}^{\mathrm{on}}\|^2_{(M_k^{\mathrm{on}})^{-1}} \leq \det(I + \sum_{h=1}^{H}(M_k^{\mathrm{on}})^{-\frac{1}{2}}x_{k,h}^{\mathrm{on}}(x_{k,h}^{\mathrm{on}})^\top(M_k^{\mathrm{on}})^{-\frac{1}{2}}) = \frac{\det(M_{k+1}^{\mathrm{on}})}{\det(M_k^{\mathrm{on}})}.$$

Correspondingly, we have that

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\gamma_k\|x_{k,h}^{\mathrm{on}}\|_{(M_k^{\mathrm{all}})^{-1}} \leq (\gamma_{K,1}+\gamma_\Delta)\sqrt{2HK}\sqrt{H\log(\det(M_{K+1}^{\mathrm{all}})) - H\log(\det(M_1^{\mathrm{all}}))} \\ &= \sqrt{2H^2K}\left(H\sqrt{2\log\Big(\frac{\det(M_K^{\mathrm{all}})^{1/2}}{\det(H^2dI)^{1/2}\delta}\Big)} + \sqrt{\lambda}B + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(H^2dI + G_{\mathrm{off}})}}\right) \\ &\quad\times\sqrt{\log(\det(M_{K+1}^{\mathrm{all}})) - \log(\det(M_1^{\mathrm{all}}))}.\end{aligned}$$

The same argument applies to the term $\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_k\|x_{k,h}^{\mathrm{on}}\|_{(M_k^{\mathrm{on}})^{-1}}$, which completes the proof. □

With the previous lemmas in hand we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. Recall that the good event $\mathcal{E}_0$ occurs with probability at least $1-2\delta$ due to Lemma D.1, where $\mathcal{E}_0 = \{\theta^* \in \mathcal{B}_k \cap \mathcal{C}_k, \text{ for all } k \geq 1\}$. Furthermore, for any fixed $\delta' \in (0,1)$, we denote $\mathcal{E}_1 = \{\frac{1}{M^{\mathrm{off}}}\lambda_{\min}(\lambda I + G_{\mathrm{off}}) \geq \tau(\delta')\}$. From Definition 4.1, we have that $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta'$.

Now we define $\mathcal{E}_2 = \mathcal{E}_0 \cap \mathcal{E}_1$. Then $\mathbb{P}(\mathcal{E}_2) \geq 1 - 2\delta - \delta'$ through a union bound. Thus we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^{\pi^*}(s_1) - V_1^{\pi_k}(s_1)\right) \,\middle|\, \mathcal{E}_2\right] \leq \mathbb{E}\left[\sum_{k=1}^{K}\left(\widehat{V}_{k,1}(s_1) - V_1^{\pi_k}(s_1)\right) \,\middle|\, \mathcal{E}_2\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\widehat{V}_{k,h}(s_{k,h}^{\mathrm{on}}) - V_h^{\pi_k}(s_{k,h}^{\mathrm{on}}) - \widehat{V}_{k,h+1}(s_{k,h+1}^{\mathrm{on}}) + V_{h+1}^{\pi_k}(s_{k,h+1}^{\mathrm{on}})\right) \,\middle|\, \mathcal{E}_2\right] \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}[\mathbb{E}[\widehat{V}_{k,h}(s_{k,h}^{\mathrm{on}}) - V_h^{\pi_k}(s_{k,h}^{\mathrm{on}}) - \widehat{V}_{k,h+1}(s_{k,h+1}^{\mathrm{on}}) + V_{h+1}^{\pi_k}(s_{k,h+1}^{\mathrm{on}}) \mid \{\mathcal{E}_2, \mathcal{F}_{k,h+1}\}] \mid \mathcal{E}_2] \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}[\mathbb{E}[\xi_{k,h+1}^{\mathrm{on}} + 2\min\{\gamma_k\|\phi(\cdot \mid s_{k,h}^{\mathrm{on}}, a_{k,h}^{\mathrm{on}})\widehat{V}_{k,h+1}\|_{(M_k^{\mathrm{all}})^{-1}}, \\
&\quad \beta_k\|\phi(\cdot \mid s_{k,h}^{\mathrm{on}}, a_{k,h}^{\mathrm{on}})\widehat{V}_{k,h+1}\|_{(M_k^{\mathrm{on}})^{-1}}\} \mid \{\mathcal{E}_2, \mathcal{F}_{k,h+1}\}] \mid \mathcal{E}_2],
\end{aligned}$$

where the first inequality is directly from Lemma D.2 and that $\mathcal{E}_2 \subset \mathcal{E}_0$, and the second inequality is from Lemma D.3 and that $\mathcal{E}_2 \subset \mathcal{E}_0$. Thus we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^{\pi^*}(s_1) - V_1^{\pi_k}(s_1)\right) \,\middle|\, \mathcal{E}_2\right] \leq \min\Bigg\{\mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg(H\sqrt{2\log\left(\frac{\det(M_K^{\mathrm{all}})^{1/2}}{\det(H^2dI)^{1/2}\delta}\right)} \\
&\quad + \sqrt{H^2d}B + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(H^2dI + G_{\mathrm{off}})}}\Bigg) \times \sqrt{\log(\det(M_{K+1}^{\mathrm{all}})) - \log(\det(M_1^{\mathrm{all}}))} \,\Bigg|\, \mathcal{E}_2\Bigg], \\
&\mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg(H\sqrt{2\log\left(\frac{\det(M_K^{\mathrm{on}})^{1/2}}{\det(H^2dI)^{1/2}\delta}\right)} + \sqrt{H^2d}B\Bigg) \\
&\quad \times \sqrt{\log(\det(M_{K+1}^{\mathrm{on}})) - \log(\det(H^2dI))} \,\Bigg|\, \mathcal{E}_2\Bigg]\Bigg\} + \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_{k,h+1}^{\mathrm{on}} \,\middle|\, \mathcal{E}_2\right],
\end{aligned}$$

where the second inequality is from Lemma D.4. From $\mathbb{E}[\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_{k,h+1}^{\mathrm{on}}] = 0$ and $|\xi_{k,h+1}^{\mathrm{on}}| \leq H$ a.s. for all k, h, $H^2K(2\delta + \delta')$ is a trivial upper bound over $\mathbb{E}[\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_{k,h+1}^{\mathrm{on}}|\mathcal{E}_2]\mathbb{P}(\mathcal{E}_2)$. Thus

$$\begin{aligned}
\mathsf{Regret}(K) \leq \min\Bigg\{&\mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg(H\sqrt{2\log\left(\frac{\det(M_K^{\mathrm{all}})^{1/2}}{\det(H^2dI)^{1/2}\delta}\right)} + \sqrt{H^2d}B \\
&+ \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(H^2dI + G_{\mathrm{off}})}}\Bigg) \times \sqrt{\log(\det(M_{K+1}^{\mathrm{all}})) - \log(\det(M_1^{\mathrm{all}}))} \,\Bigg|\, \mathcal{E}_2\Bigg], \\
&\mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg(H\sqrt{2\log\left(\frac{\det(M_K^{\mathrm{on}})^{1/2}}{\det(H^2dI)^{1/2}\delta}\right)} + \sqrt{H^2d}B\Bigg) \\
&\times \sqrt{\log(\det(M_{K+1}^{\mathrm{on}})) - \log(\det(H^2dI))} \,\Bigg|\, \mathcal{E}_2\Bigg]\Bigg\} + H^2K(2\delta + \delta') \\
&+ HK(2\delta + \delta'),
\end{aligned}$$

where the term $HK(2\delta + \delta')$ accounts for the contribution of the failure event $\mathcal{E}_2^c$ to the expected regret, using the facts that $\mathsf{Regret}(K) \leq HK$ almost surely and $\mathbb{P}(\mathcal{E}_2^c) \leq 2\delta + \delta'$.

Recall that $M_1^{\mathrm{all}} = H^2dI + G_{\mathrm{off}}$. Similar to G^{off}, we define $G_k^{\mathrm{on}} = \sum_{k'=1}^{k}\sum_{h=1}^{H}x_{k',h}^{\mathrm{on}}(x_{k',h}^{\mathrm{on}})^\top$. Thus $M_{K+1}^{\mathrm{all}} = M_1^{\mathrm{all}} + G_K^{\mathrm{on}}$. Applying Lemma E.6 (Section E) with $A = M_1^{\mathrm{all}}$ and $B = G_K^{\mathrm{on}}$, we obtain

$$\log(\det(M_{K+1}^{\mathrm{all}})) - \log(\det(M_1^{\mathrm{all}})) \leq d\log\left(1 + \frac{1}{d}\operatorname{tr}\left((M_1^{\mathrm{all}})^{-1}G_K^{\mathrm{on}}\right)\right).$$

Since $M_1^{\mathrm{all}} \succ 0$ and $G_K^{\mathrm{on}} \succeq 0$, the trace term can be bounded as

$$\mathrm{tr}\big((M_1^{\mathrm{all}})^{-1} G_K^{\mathrm{on}}\big) \le \lambda_{\max}\big((M_1^{\mathrm{all}})^{-1}\big)\,\mathrm{tr}(G_K^{\mathrm{on}}) = \lambda_{\min}(M_1^{\mathrm{all}})^{-1}\,\mathrm{tr}(G_K^{\mathrm{on}}),$$

where the first inequality is from Lemma E.7 (Section E). Moreover, by definition of G_K^{on},

$$\mathrm{tr}(G_K^{\mathrm{on}}) = \sum_{k=1}^{K}\sum_{h=1}^{H} \|x_{k,h}^{\mathrm{on}}\|_2^2 \le H^3 dK, \tag{25}$$

where the inequality follows from the boundedness $\|x_{k,h}^{\mathrm{on}}\|_2^2 \le H^2 d$ and Eq. (24). Combining the above bounds yields

$$\log(\det(M_{K+1}^{\mathrm{all}})) - \log(\det(M_1^{\mathrm{all}})) \le d\log\big(1 + H^3 K \lambda_{\min}(H^2 dI + G_{\mathrm{off}})^{-1}\big).$$

Furthermore, using Lemma E.6 (Section E) again with $A = H^2 dI$ and $B = G_{\mathrm{off}} + G_K^{\mathrm{on}}$, and noting that $\mathrm{tr}(G_{\mathrm{off}}) \le d\lambda_{\max}(G_{\mathrm{off}})$, we obtain that

$$\begin{aligned}
\log\left(\frac{\det\big(M_K^{\mathrm{all}}\big)^{1/2}}{\det(H^2 dI)^{1/2}\delta}\right) &\le \log\left(\frac{1}{\delta}\right) + \frac{1}{2} d\log\left(1 + \frac{1}{d}\,\mathrm{tr}\left(\frac{G_{\mathrm{off}} + G_K^{\mathrm{on}}}{H^2 d}\right)\right) \\
&= \log\left(\frac{1}{\delta}\right) + \frac{1}{2} d\log\left(1 + \frac{1}{H^2 d^2}\,\mathrm{tr}(G_{\mathrm{off}} + G_K^{\mathrm{on}})\right) \\
&\le \log\left(\frac{1}{\delta}\right) + \frac{1}{2} d\log\left(1 + (H^2 d)^{-1}\lambda_{\max}(G_{\mathrm{off}}) + \frac{HK}{d}\right),
\end{aligned}$$

where the last inequality follows from the additivity of the trace, Eq. (25), and the fact that $\mathrm{tr}(G_{\mathrm{off}}) \le d\lambda_{\max}(G_{\mathrm{off}})$. Similarly, we have that

$$\log\left(\frac{\det\big(M_{K+1}^{\mathrm{on}}\big)}{\det(H^2 dI)}\right) \le d\log\left(1 + \frac{HK}{d}\right).$$

Thus we obtain that

$$\begin{aligned}
\mathsf{Regret}(K) \le \min\Bigg\{ &\mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg[H\sqrt{2\left(\log\left(\frac{1}{\delta}\right) + \frac{1}{2}d\log\left(1 + \frac{\lambda_{\max}(G_{\mathrm{off}})}{H^2 d} + \frac{HK}{d}\right)\right)} \\
&\quad + H\sqrt{d}B + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(H^2 dI + G_{\mathrm{off}})}}\Bigg] \\
&\quad \times \sqrt{d\log\left(1 + \frac{H^3 K}{\lambda_{\min}(H^2 dI + G_{\mathrm{off}})}\right)} \;\Big|\; \mathcal{E}_2\Bigg], \\
&\mathbb{E}\left[2\sqrt{2H^2K}\left(H\sqrt{2\left(\log\left(\frac{1}{\delta}\right) + \frac{d}{2}\log\left(1 + \frac{HK}{d}\right)\right)} + H\sqrt{d}B\right)\right. \\
&\quad \left.\times \sqrt{d\log\left(1 + \frac{HK}{d}\right)} \;\Big|\; \mathcal{E}_2\right]\Bigg\} + H^2 K(2\delta + \delta') + HK(2\delta + \delta').
\end{aligned}$$

Since $\mathcal{E}_2 \subset \mathcal{E}_1$, by the definition of $\mathcal{E}_1$, we obtain

$$\begin{aligned}
\mathsf{Regret}(K) \le \min\Bigg\{ &\mathbb{E}\Bigg[2\sqrt{2H^2K}\Bigg[H\sqrt{2\left(\log\left(\frac{1}{\delta}\right) + \frac{1}{2}d\log\left(1 + \frac{\lambda_{\max}(G_{\mathrm{off}})}{H^2 d} + \frac{HK}{d}\right)\right)} \\
&\quad + H\sqrt{d}B + \Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\tau(\delta')M^{\mathrm{off}}}}\Bigg] \times \sqrt{d\log\left(1 + \frac{H^3 K}{\tau(\delta')M^{\mathrm{off}}}\right)} \;\Big|\; \mathcal{E}_2\Bigg], \\
&\mathbb{E}\left[2\sqrt{2H^2K}\left(H\sqrt{2\left(\log\left(\frac{1}{\delta}\right) + \frac{d}{2}\log\left(1 + \frac{HK}{d}\right)\right)} + H\sqrt{d}B\right)\right. \\
&\quad \left.\times \sqrt{d\log\left(1 + \frac{HK}{d}\right)} \;\Big|\; \mathcal{E}_2\right]\Bigg\} + H^2 K(2\delta + \delta') + HK(2\delta + \delta').
\end{aligned} \tag{26}$$

Finally, we note that $\lambda_{\max}(G_{\text{off}}) \le \operatorname{tr}(G_{\text{off}}) \le H^3 d M^{\text{off}}$, where the second inequality follows from Equation (24). We take $\delta = \frac{1}{2K}$ and $\delta' = \frac{1}{2KM^{\text{off}}}$. As $\tau = \tau(\frac{1}{2KM^{\text{off}}})$, substituting these choices into the expression above yields that

$$\begin{aligned}\textsf{Regret}(K) \le \min\Bigg\{ \tilde{O}\Bigg(& 2\sqrt{2H^2K}\Bigg[H\sqrt{2\Big(\log(2K) + \frac{1}{2}d\log\Big(1 + \frac{H^3 d M^{\text{off}}}{H^2 d} + \frac{HK}{d}\Big)\Big)} \\ & + H\sqrt{d}B + \Delta H^3 d\sqrt{\frac{M^{\text{off}}}{\tau}}\Bigg] \times \sqrt{d\log\Big(1 + \frac{H^3 K}{\tau M^{\text{off}}}\Big)}\Bigg), \\ & 2\sqrt{2H^2K}\Bigg(H\sqrt{2\Big(\log(2K) + \frac{d}{2}\log\Big(1 + \frac{HK}{d}\Big)\Big)} + H\sqrt{d}B\Bigg) \\ & \times \sqrt{d\log\Big(1 + \frac{HK}{d}\Big)}\Bigg\} + H^2(1 + \frac{1}{2M^{\text{off}}}) + H(1 + \frac{1}{2M^{\text{off}}}). \end{aligned} \tag{27}$$

Equivalently, up to logarithmic factors, we have

$$\begin{aligned}&\textsf{Regret}(K) \\ &= \tilde{O}\Bigg(\min\Bigg\{\Bigg[BH^2 d\sqrt{K} + H^4 d\Delta\sqrt{dK}\sqrt{\frac{M^{\text{off}}}{\tau}}\Bigg] \times \sqrt{\log\Big(1 + \frac{H^3 K}{\tau M^{\text{off}}}\Big)}, BH^2 d\sqrt{K}\Bigg\}\Bigg),\end{aligned} \tag{28}$$

where $\tilde{O}(\cdot)$ hides logarithmic factors in K and M^{off}. □

D.1.1 Proof of Theorem C.4

Theorem C.4. *Suppose that all assumptions in Theorem 4.2 hold except Assumption 2.3. Let $\Delta = \|\theta^{\text{off},*} - \theta^*\|_2$ denote the true environment shift. Run Algorithm 1 with $\lambda = H^2 d$ and the confidence radii specified in Theorem 4.2, replacing Δ by a misspecified value $\tilde{\Delta}$ satisfying $0 \le \tilde{\Delta} < \Delta$. Then the algorithm remains implementable without knowledge of Δ, and we have*

$$\begin{aligned}\textit{Regret}(K) \le \tilde{O}\Bigg(& \Bigg[BH^2 d\sqrt{K} + H^4 d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\text{off}}}{\tau}}\Bigg]\sqrt{\log\Big(1 + \frac{H^3 K}{\tau M^{\text{off}}}\Big)} \\ & + \frac{BH^3 dK}{\sqrt{\tau M^{\text{off}}}} + \frac{\Delta H^5 d^{3/2} K}{\tau}\Bigg).\end{aligned} \tag{22}$$

Proof. Under the implemented value $\tilde{\Delta}$, for all $k \ge 1$, note that the planning set is defined as

$$\tilde{\mathcal{D}}_k = \begin{cases} \tilde{\mathcal{B}}_k \cap \mathcal{C}_k, & \tilde{\mathcal{B}}_k \cap \mathcal{C}_k \neq \varnothing, \\ \tilde{\mathcal{B}}_k, & \tilde{\mathcal{B}}_k \cap \mathcal{C}_k = \varnothing, \end{cases}$$

where $\tilde{\mathcal{B}}_k = \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_k^{\text{all}}\|_{M_k^{\text{all}}} \le \tilde{\gamma}_k\}$ and $\tilde{\gamma}_k$ is obtained by replacing Δ by $\tilde{\Delta}$ in γ_k. For clarity, we denote this implemented set by $\tilde{\mathcal{D}}_k$ throughout the proof. The proof follows the same structure as that of Theorem 4.2. We first present the counterparts of Lemmas D.1 to D.4, explaining how each lemma changes under Δ misspecification.

We first note that Lemma D.1 remains valid, except that $\tilde{\mathcal{D}}_k = \tilde{\mathcal{B}}_k \cap \mathcal{C}_k$ need not hold for all $k \ge 1$ on $\mathcal{E}_0$; recall that $\mathcal{E}_0 = \{\theta^* \in \mathcal{B}_k \cap \mathcal{C}_k, \text{ for all } k \ge 1\}$. Indeed, the self-normalized concentration in Lemma E.1 still applies, and we still have $\|\theta^{\text{off},*} - \theta^*\|_2 \le \Delta$. However, $\theta^* \in \mathcal{B}_k \cap \mathcal{C}_k$ on $\mathcal{E}_0$ (as in Lemma D.1) no longer implies $\theta^* \in \tilde{\mathcal{B}}_k \cap \mathcal{C}_k$, since $\tilde{\Delta} < \Delta$ yields $\tilde{\gamma}_k \le \gamma_k$ for all $k \ge 1$. Thus, $\tilde{\mathcal{B}}_k \cap \mathcal{C}_k$ may be empty.

As a counterpart of Lemma D.2, we have the following.

Lemma D.5. *Conditional on $\mathcal{E}_0$, for any episode k, we have*

$$Q_h^{\pi^*}(s,a) \le \widehat{Q}_{k,h}(s,a) + (H-h+1)2\gamma_k \frac{H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}},$$
$$V_h^{\pi^*}(s) \le \widehat{V}_{k,h}(s) + (H-h+1)2\gamma_k \frac{H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}},$$

for all states s, actions a, and stages $h = 1, \dots, H+1$.

Proof. As in Lemma D.2, we prove the result by backward induction. For the case $h = H+1$, we have $Q_{H+1}^{\pi^*}(s,a) = \widehat{Q}_{k,H+1}(s,a) = 0$ for all s, a, so the claim holds trivially. Assume that for some $h \in [H]$, $Q_{h+1}^{\pi^*}(s,a) \le \widehat{Q}_{k,h+1}(s,a) + (H-(h+1)+1)2\gamma_k \frac{H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I+G_{\text{off}})}}$ for all s, a. Then,

$$\begin{aligned} V_{h+1}^{\pi^*}(s) &= \max_{a\in\mathcal{A}} Q_{h+1}^{\pi^*}(s,a) \le \max_{a\in\mathcal{A}} \widehat{Q}_{k,h+1}(s,a) + (H-h)2\gamma_k \frac{H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \\ &= \widehat{V}_{k,h+1}(s) + (H-h)2\gamma_k \frac{H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \end{aligned} \tag{29}$$

for all states s.

For any state–action pair (s,a), we have

$$\begin{aligned} Q_h^{\pi^*}(s,a) &= r(s,a) + P(\cdot \mid s,a)^\top V_{h+1}^{\pi^*} \le r(s,a) + \sum_{j=1}^{d} \theta_j^* \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1} \\ &\quad + \frac{2(H-h)\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \\ &\le r(s,a) + \max_{\theta\in\tilde{\mathcal{D}}_k} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1} \\ &\quad + \max_{\theta\in\tilde{\mathcal{D}}_k} \left| \sum_{j=1}^{d} (\theta_j^* - \theta_j)\phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1} \right| + \frac{2(H-h)\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \\ &\le \widehat{Q}_{k,h}(s,a) + \max_{\theta\in\tilde{\mathcal{D}}_k} \left(\|\theta^* - \widehat{\theta}_k^{\text{all}}\|_{M_k^{\text{all}}} + \|\theta - \widehat{\theta}_k^{\text{all}}\|_{M_k^{\text{all}}} \right) \left\| \phi(\cdot \mid s,a)\widehat{V}_{k,h+1} \right\|_{(M_k^{\text{all}})^{-1}} \\ &\quad + \frac{2(H-h)\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \\ &\le \widehat{Q}_{k,h}(s,a) + (\gamma_k + \tilde{\gamma}_k) \left\| \phi(\cdot \mid s,a)\widehat{V}_{k,h+1} \right\|_{(M_k^{\text{all}})^{-1}} + \frac{2(H-h)\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \\ &\le \widehat{Q}_{k,h}(s,a) + \frac{2\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} + \frac{2(H-h)\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}} \\ &= \widehat{Q}_{k,h}(s,a) + \frac{2(H-h+1)\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}}. \end{aligned}$$

The first inequality follows from Eq. (29) and the fact that $\sum_{s'} P(s' \mid s,a) = 1$. The third follows from Cauchy–Schwarz and the triangle inequality, see the proof of Lemma D.3. The fourth follows from the definition of $\mathcal{E}_0$ and that $\tilde{\mathcal{D}}_k \subset \tilde{\mathcal{B}}_k$. The fifth uses $\tilde{\gamma}_k \le \gamma_k$, $M_k^{\text{all}} \succeq \lambda I + G_{\text{off}}$, and $\|\phi(\cdot \mid s,a)\widehat{V}_{k,h+1}\|_2 \le H\sqrt{d}$; see Eq. (24). Taking the maximum over a yields $V_h^{\pi^*}(s) \le \widehat{V}_{k,h}(s) + 2(H-h+1)\gamma_k H\sqrt{d}/\sqrt{\lambda_{\min}(\lambda I + G_{\text{off}})}$, which completes the proof. □

Note that the proof of Lemma D.5 establishes

$$\max_{\theta\in\tilde{\mathcal{D}}_k}\left|\sum_{j=1}^{d}(\theta_j^*-\theta_j)\phi_j(\cdot\mid s,a)^\top\widehat{V}_{k,h+1}\right|\le 2\gamma_k\left\|\phi(\cdot\mid s,a)\widehat{V}_{k,h+1}\right\|_{(M_k^{\mathrm{all}})^{-1}}\le\frac{2\gamma_k H\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I+G_{\mathrm{off}})}}. \tag{30}$$

This is precisely the error term controlled by Lemma D.3 to bound the regret when there is no misspecification. Thus, the same argument yields the following lemma, whose proof is omitted.

Lemma D.6. *Let π_k be the implemented policy at episode k. Conditional on $\mathcal{E}_0$ and $\mathcal{F}_{k,h+1}$, we have*

$$\begin{aligned}\widehat{V}_{k,h}\big(s_{k,h}^{\mathrm{on}}\big)-V_h^{\pi_k}\big(s_{k,h}^{\mathrm{on}}\big)\le\widehat{V}_{k,h+1}\big(s_{k,h+1}^{\mathrm{on}}\big)-V_{h+1}^{\pi_k}\big(s_{k,h+1}^{\mathrm{on}}\big)+\xi_{k,h+1}^{\mathrm{on}}\\+2\gamma_k\left\|\phi(\cdot\mid s_{k,h}^{\mathrm{on}},a_{k,h}^{\mathrm{on}})\widehat{V}_{k,h+1}\right\|_{(M_k^{\mathrm{all}})^{-1}},\end{aligned}$$

where

$$\xi_{k,h+1}^{\mathrm{on}}=P(\cdot\mid s_{k,h}^{\mathrm{on}},a_{k,h}^{\mathrm{on}})^\top\big(\widehat{V}_{k,h+1}-V_{h+1}^{\pi_k}\big)-\big(\widehat{V}_{k,h+1}(s_{k,h+1}^{\mathrm{on}})-V_{h+1}^{\pi_k}(s_{k,h+1}^{\mathrm{on}})\big),$$

which is the same as in Lemma D.3.

We note that the difference between Lemma D.6 and Lemma D.3 identifies one point at which the online-only safeguard is lost (the other is in Lemma D.5, which is straightforward); conditional on $\mathcal{E}_0$, $\tilde{\mathcal{D}}_k\subset\mathcal{C}_k$ is no longer guaranteed, so the bound in Lemma D.6 no longer has the min-of-two structure that incorporates the online-only estimator.

Finally, Lemma D.4 remains unchanged, since its proof is only about sample-path-wise algebra. The proof of Theorem C.4 therefore follows the same steps as that of Theorem 4.2, with the additional term

$$\sum_{k=1}^{K}\frac{2\gamma_k H^2\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I+G_{\mathrm{off}})}}, \tag{31}$$

arising from Lemma D.5 and Lemma D.6 at $h=1$, see Eq. (30). Substituting the definition of γ_k and following the proof of Theorem 4.2 (see Eq. (28)), we obtain

$$\begin{aligned}\mathsf{Regret}(K)&\le\tilde{O}\left(\left[BH^2d\sqrt{K}+H^4d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\mathrm{off}}}{\tau}}\right]\sqrt{\log\left(1+\frac{H^3K}{\tau M^{\mathrm{off}}}\right)}\right)\\&\quad+\mathbb{E}\left[\sum_{k=1}^{K}\frac{2\gamma_k H^2\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I+G_{\mathrm{off}})}}\;\Big|\;\mathcal{E}_2\right]\\&=\tilde{O}\left(\left[BH^2d\sqrt{K}+H^4d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\mathrm{off}}}{\tau}}\right]\sqrt{\log\left(1+\frac{H^3K}{\tau M^{\mathrm{off}}}\right)}\right)\\&\quad+\mathbb{E}\left[\sum_{k=1}^{K}\frac{2H^2\sqrt{d}}{\sqrt{\lambda_{\min}(\lambda I+G_{\mathrm{off}})}}\left(H\sqrt{2\log\left(\frac{2\det(M_k^{\mathrm{all}})^{1/2}K}{\det(\lambda I)^{1/2}}\right)}\right.\right.\\&\quad\left.\left.+\Delta\frac{\lambda_{\max}(G_{\mathrm{off}})}{\sqrt{\lambda_{\min}(\lambda I+G_{\mathrm{off}})}}+\sqrt{\lambda}B\right)\;\Big|\;\mathcal{E}_2\right]\\&\le\tilde{O}\left(\left[BH^2d\sqrt{K}+H^4d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\mathrm{off}}}{\tau}}\right]\sqrt{\log\left(1+\frac{H^3K}{\tau M^{\mathrm{off}}}\right)}\right.\\&\quad\left.+\frac{BH^3dK}{\sqrt{\tau M^{\mathrm{off}}}}+\frac{\Delta H^5d^{3/2}K}{\tau}\right).\end{aligned}$$

The first and second terms on the right-hand side of the first line follow from Eqs. (28) and (31), respectively. The second line substitutes the definition of γ_k. The third line uses $\lambda=H^2d$, as

specified in the statement of Theorem 4.2; the bound $\lambda_{\min}(\lambda I + G_{\text{off}}) \geq \tau M^{\text{off}}$ on $\mathcal{E}_1$; and the bounds $\lambda_{\max}(G_{\text{off}}) \leq H^3 d M^{\text{off}}$ and

$$\log\left(\frac{2\det(M_k^{\text{all}})^{1/2}K}{\det(\lambda I)^{1/2}}\right) \leq \log(2K) + \frac{d}{2}\log\left(1 + HM^{\text{off}} + \frac{HK}{d}\right)$$

for all $k \geq 1$. All of these were established in the proof of Theorem 4.2. Recall that for any fixed $\delta' \in (0,1)$, we denote $\mathcal{E}_1 = \left\{\frac{1}{M^{\text{off}}}\lambda_{\min}(\lambda I + G_{\text{off}}) \geq \tau(\delta')\right\}$, as in the proof of Theorem 4.2. We again let $\mathcal{E}_2 = \mathcal{E}_0 \cap \mathcal{E}_1$ and pick the same choices of δ and δ' as in the proof of Theorem 4.2; the failure-event contributions are handled in the same way. □

D.2 Proof in Section 4.2

Corollary 4.4. *If there exists $\epsilon \in (0, \frac{1}{2}]$ such that $\tau = \Omega\left(H^3 \frac{K^{1+2\epsilon}}{M^{\text{off}}}\right)$ and $\frac{\Delta}{\tau} = O\left(\frac{B}{H^{7/2}d^{1/2}K^{\frac{1}{2}+\epsilon}}\right)$, then the offline data is informative with respect to ϵ.*

Proof. Recall from Theorem 4.2 that

$$\text{term } (A) = \left[BH^2 d\sqrt{K} + H^4 d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\text{off}}}{\tau}}\right]\sqrt{\log\left(1 + \frac{H^3 K}{\tau M^{\text{off}}}\right)}, \text{term } (B) = BH^2 d\sqrt{K}.$$

By definition, the offline data is informative with respect to ϵ if term $(A) \leq \tilde{O}(\text{term } (B) \cdot K^{-\epsilon})$. Thus from the definition of term (A), we directly obtain that

$$\begin{aligned}
\text{term } (A) &\leq \left[BH^2 d\sqrt{K} + H^4 d^{3/2}\Delta\sqrt{K}\sqrt{\frac{M^{\text{off}}}{\tau}}\right]\sqrt{\frac{H^3 K}{\tau M^{\text{off}}}} = \frac{BH^{\frac{7}{2}}dK}{\sqrt{\tau M^{\text{off}}}} + \frac{H^{\frac{11}{2}}d^{\frac{3}{2}}\Delta K}{\tau} \\
&\leq \frac{BH^{\frac{7}{2}}dK}{\sqrt{H^3 K^{1+2\epsilon}}} + H^{\frac{11}{2}}d^{\frac{3}{2}}K\frac{B}{H^{7/2}d^{1/2}K^{\frac{1}{2}+\epsilon}} = BH^2 dK^{\frac{1}{2}-\epsilon} + BH^2 dK^{\frac{1}{2}-\epsilon} \\
&= \tilde{O}(\text{term } (B) \cdot K^{-\epsilon}),
\end{aligned}$$

where the first inequality follows from the bound $\log(1+x) \leq x$ for all $x > 0$, and the second line follows from the conditions in Corollary 4.4, with absolute constants omitted. □

D.3 Proof in Section 4.3

Proposition 4.6. *If ϕ is not uniformly learnable, then for any behavior policy and any design of $\{\widetilde{V}_{m,h}\}$, we have $\tau(\delta) = \frac{\lambda}{M^{\text{off}}}$ for any $\delta \in (0,1)$. In this case, the offline data is uninformative.*

Proof. We prove the contrapositive of the claim. Suppose there exist a behavior policy and a design of $\{\widetilde{V}_{m,h}\}$ such that for some $\delta \in (0,1)$, we have $\tau(\delta) > \frac{\lambda}{M^{\text{off}}}$. By the definition of $\tau(\delta)$ in Definition 4.1, this implies that, with probability at least $1-\delta$,

$$\lambda_{\min}(G_{\text{off}}) = \lambda_{\min}\left(\sum_{m=1}^{M^{\text{off}}}\sum_{h=1}^{H}\phi(\cdot \mid s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}})\widetilde{V}_{m,h+1}\left(\phi(\cdot \mid s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}})\widetilde{V}_{m,h+1}\right)^\top\right) > 0. \quad (32)$$

We now show that this implies the feature mapping ϕ of the offline MDP is uniformly learnable. If Eq. (32) holds with probability at least $1-\delta$, then there exists a realization of the offline data, denoted as $\left\{\left(s_{m,h}^{\text{off,sample}}, a_{m,h}^{\text{off,sample}}\right)\right\}_{(m,h)}$, together with the corresponding realization of the designed vectors $\left\{\widetilde{V}_{m,h}^{\text{sample}}\right\}_{(m,h)}$, such that

$$\lambda_{\min}\left(\sum_{m=1}^{M^{\text{off}}}\sum_{h=1}^{H}\phi\left(\cdot \mid s_{m,h}^{\text{off,sample}}, a_{m,h}^{\text{off,sample}}\right)\widetilde{V}_{m,h+1}^{\text{sample}}\left(\phi\left(\cdot \mid s_{m,h}^{\text{off,sample}}, a_{m,h}^{\text{off,sample}}\right)\widetilde{V}_{m,h+1}^{\text{sample}}\right)^\top\right) > 0.$$

Define $\{(s_i, a_i)\}_{i=1}^{I}$ as the collection of $\left\{\left(s_{m,h}^{\text{off,sample}}, a_{m,h}^{\text{off,sample}}\right)\right\}_{(m,h)}$, and let $\{V_1, \ldots, V_L\}$ be the collection of $\left\{\widetilde{V}_{m,h+1}^{\text{sample}}\right\}_{(m,h)}$ as in Definition 4.5. Then

$$\begin{aligned}
&\lambda_{\min}\left(\sum_{i=1}^{I}\sum_{l=1}^{L}\phi(\cdot \mid s_i, a_i)V_l\big(\phi(\cdot \mid s_i, a_i)V_l\big)^\top\right) \\
&\quad \geq \lambda_{\min}\left(\sum_{m=1}^{M^{\text{off}}}\sum_{h=1}^{H}\phi\Big(\cdot \mid s_{m,h}^{\text{off,sample}}, a_{m,h}^{\text{off,sample}}\Big)\widetilde{V}_{m,h+1}^{\text{sample}}\Big(\phi\Big(\cdot \mid s_{m,h}^{\text{off,sample}}, a_{m,h}^{\text{off,sample}}\Big)\widetilde{V}_{m,h+1}^{\text{sample}}\Big)^\top\right) \\
&\quad > 0,
\end{aligned}$$

where the first inequality holds because the omitted terms are rank-one symmetric positive semidefinite matrices. Hence, the feature mapping ϕ of the offline MDP is uniformly learnable. □

Proposition 4.10. *Under uniform learnability (Definition 4.5) and uniform data coverage (Assumption 4.8), and under the stochastic design of $\{\widetilde{V}_{m,h}\}$ (Definition 4.9), for any $\delta > 0$, the effective coverage $\tau(\delta)$ satisfies $\tau(\delta) \geq \frac{p_0\kappa}{L^2} - \sqrt{\frac{2dp_0\kappa}{L^2 M^{\text{off}}}\log\big(\frac{d}{\delta}\big)}$. In particular, when $M^{\text{off}} \geq \frac{8dL^2}{p_0\kappa}\log(2dKM^{\text{off}})$, we have $\tau = \tau\big(\frac{1}{2KM^{\text{off}}}\big) \geq \frac{p_0\kappa}{2L^2}$.*

Proof. Fix any $h' \in [H]$. For each offline trajectory $m \in [M^{\text{off}}]$, define the random PSD matrix

$$X_m = \phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})\widetilde{V}_{m,h'+1}\big(\phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})\widetilde{V}_{m,h'+1}\big)^\top.$$

Note that $\{X_m\}_{m=1}^{M^{\text{off}}}$ are independent. Let $V_s = [V_1, \ldots, V_L] \in \mathbb{R}^{|\mathcal{S}|\times L}$, where $\{V_1, \ldots, V_L\}$ denote the probe vectors in Definition 4.5. Using the definition of $\widetilde{V}_{m,h'}$ and conditioning on $(s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})$, we obtain

$$\begin{aligned}
\mathbb{E}[X_m] &= \mathbb{E}\Big[\mathbb{E}\Big[\phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})\widetilde{V}_{m,h'+1}\big(\phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})\widetilde{V}_{m,h'+1}\big)^\top \,\Big|\, (s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})\Big]\Big] \\
&= \frac{1}{L^2}\mathbb{E}\left[\sum_{l=1}^{L}\phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})V_l\big(\phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})V_l\big)^\top\right],
\end{aligned}$$

where the last equality follows from taking the expectation over $\widetilde{V}_{m,h'}$ conditional on the visitation of $(s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})$.

Since the feature mapping of the offline MDP is uniformly learnable, Definition 4.5 implies that there exists a set of state–action pairs $\{(s_i, a_i)\}_{i=1}^{I}$ for which Definition 4.5 holds. By Assumption 4.8, for each state–action pair (s_i, a_i), $i \in [I]$, $d_{h'}^{\pi^{\text{off}}}(s_i, a_i) \geq p_0$. Restricting the expectation to the set $\{(s_i, a_i)\}_{i=1}^{I}$ yields

$$\mathbb{E}[X_m] \succeq \frac{p_0}{L^2}\sum_{i=1}^{I}\sum_{l=1}^{L}\phi(\cdot \mid s_i, a_i)V_l\big(\phi(\cdot \mid s_i, a_i)V_l\big)^\top \succeq \frac{p_0\kappa}{L^2}I,$$

where the last inequality is from Definition 4.5. Consequently, $\mu_{\min} := \lambda_{\min}\Big(\sum_{m=1}^{M^{\text{off}}}\mathbb{E}[X_m]\Big) \geq M^{\text{off}}\frac{p_0\kappa}{L^2}$.

Since $\widetilde{V}_{m,h'} \in [-1, 1]^{|\mathcal{S}|}$ for all m and $\sum_{s'}|\phi_j(s' \mid s, a)| \leq 1$ for all j and all (s, a), for all m,

$$\lambda_{\max}(X_m) = \|\phi(\cdot \mid s_{m,h'}^{\text{off}}, a_{m,h'}^{\text{off}})\widetilde{V}_{m,h'+1}\|_2^2 \leq d \quad \text{almost surely.}$$

Applying the matrix Chernoff inequality (Lemma E.2) with $\lambda_{\max}(X_m) \leq d$ a.s., for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{m=1}^{M^{\text{off}}}X_m\right) \leq (1-\delta)\mu_{\min}\right) \leq d\exp\left(-\frac{\delta^2}{2d}\mu_{\min}\right).$$

By rearranging terms, we obtain that with probability at least $1-\delta$,

$$\frac{1}{M^{\mathrm{off}}}\lambda_{\min}\left(\sum_{m=1}^{M^{\mathrm{off}}} X_m\right) \geq \frac{p_0\kappa}{L^2} - \sqrt{\frac{2dp_0\kappa}{L^2 M^{\mathrm{off}}}\log\left(\frac{d}{\delta}\right)}.$$

Finally, since

$$G^{\mathrm{off}} = \sum_{h=1}^{H}\sum_{m=1}^{M^{\mathrm{off}}} \phi(\cdot \mid s_{m,h}^{\mathrm{off}}, a_{m,h}^{\mathrm{off}})\widetilde{V}_{m,h+1}\big(\phi(\cdot \mid s_{m,h}^{\mathrm{off}}, a_{m,h}^{\mathrm{off}})\widetilde{V}_{m,h+1}\big)^\top \succeq \sum_{m=1}^{M^{\mathrm{off}}} X_m,$$

the same lower bound applies to $\lambda_{\min}(G^{\mathrm{off}})/M^{\mathrm{off}}$. The proof of Proposition 4.10 is complete by letting $\delta = \frac{1}{2KM^{\mathrm{off}}}$ and requiring that $\sqrt{\frac{2dp_0\kappa}{L^2M^{\mathrm{off}}}\log\big(\frac{d}{\delta}\big)} \leq \frac{1}{2}\frac{p_0\kappa}{L^2}$. □

Proposition 4.12. *In offline–online tabular MDP setting, under deterministic design of $\{\widetilde{V}_{m,h}\}$ (Definition 4.11) and Assumption 4.8, for any $\delta > 0$, we have $\tau(\delta) \geq \frac{p_0}{|\mathcal{S}|} - \frac{1}{|\mathcal{S}|}\big(\frac{1}{2M^{\mathrm{off}}}\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})\big)^{1/2}$. When $M^{\mathrm{off}} \geq \frac{2}{p_0^2}\log(2|\mathcal{S}||\mathcal{A}|KM^{\mathrm{off}})$, we have that $\tau = \tau\big(\frac{1}{2KM^{\mathrm{off}}}\big) \geq \frac{p_0}{2|\mathcal{S}|}$.*

Proof. Note that under the deterministic design of $\{\widetilde{V}_{m,h}\}$ in Proposition 4.12, the corresponding block of G_{off} associated with any fixed (s,a) takes the form

$$\mathrm{diag}\left(\frac{N(s,a)}{|\mathcal{S}|}, \ldots, \frac{N(s,a)}{|\mathcal{S}|}\right), \tag{33}$$

where $N(s,a)$ denotes the number of visits to the state–action pair (s,a) in the offline phase. For simplicity, we assume that $|\mathcal{S}|$ divides $N(s,a)$ for every (s,a). Fix any $h' \in [H]$. For any fixed (s,a) pair, for any $m \in [M^{\mathrm{off}}]$, let

$$X_{m,h'}^{(s,a)} = \mathbf{1}\{(s_{m,h'}^{\mathrm{off}}, a_{m,h'}^{\mathrm{off}}) = (s,a)\},$$

where $\mathbf{1}$ denotes the indicator function. By assumption, $\mathbb{P}(X_{m,h'}^{(s,a)} = 1) \geq p_0$, and the random variables $\{X_{m,h'}^{(s,a)}\}_{m=1}^{M^{\mathrm{off}}}$ are independent across trajectories. Hence, $N_{h'}(s,a) := \sum_{m=1}^{M^{\mathrm{off}}} X_{m,h'}^{(s,a)}$ is a sum of independent Bernoulli random variables with mean at least $M^{\mathrm{off}}p_0$.

By Hoeffding's inequality (Chap 2 of Vershynin [51]), for any $\epsilon > 0$,

$$\mathbb{P}\big(N_{h'}(s,a) \leq M^{\mathrm{off}}p_0 - \epsilon\big) \leq \exp\left(-\frac{2\epsilon^2}{M^{\mathrm{off}}}\right).$$

Setting $\epsilon = \sqrt{\frac{M^{\mathrm{off}}}{2}\log\big(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\big)}$ for any fixed $\delta \in (0,1)$, we obtain

$$\mathbb{P}\left(N_{h'}(s,a) \leq M^{\mathrm{off}}p_0 - \sqrt{\frac{M^{\mathrm{off}}}{2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)}\right) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|}.$$

Applying a union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we conclude that with probability at least $1-\delta$,

$$N_{h'}(s,a) \geq M^{\mathrm{off}}p_0 - \sqrt{\frac{M^{\mathrm{off}}}{2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)} \tag{34}$$

holds simultaneously for all (s,a). Note that $N(s,a) \geq N_{h'}(s,a)$ almost surely for all (s,a) pairs. The proof of Proposition 4.12 follows by combining Eqs. (33) and (34), setting $\delta = \frac{1}{2KM^{\mathrm{off}}}$ in, and requiring that $\frac{1}{|\mathcal{S}|}\sqrt{\frac{1}{2M^{\mathrm{off}}}\log\big(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\big)} \leq \frac{1}{2}\frac{p_0}{|\mathcal{S}|}$. □

Proposition C.5. *Consider the aforementioned offline–online tabular bandit setting, where the offline behavior policies follow the UCB1 algorithm with exploration rate $\gamma_{M^{\mathrm{off}}} = \sqrt{2\log(M^{\mathrm{off}})}$ (See*

Algorithm 1 of Han et al. [27]). If $K = o((\log(M^{\text{off}}))^{49})$,[6] *and there exists* $p \in (0, \frac{1}{2}]$ *such that* $\Delta^{\text{off}}_{gap} = O((M^{\text{off}})^{-p})$, *then under the deterministic design of* $\{\widetilde{V}_{m,h}\}$ *(Definition 4.11), running* **O-O UCRL-VTR** *yields that* $\tau = \Omega((M^{\text{off}})^{2p-1})$. *Furthermore, if there exists* $\epsilon \in (0, \frac{1}{2}]$ *such that*

$$(M^{\text{off}})^{2p} = \Omega(K^{1+2\epsilon}), \qquad \Delta = O\left(\frac{(M^{\text{off}})^{2p-1}}{K^{\frac{1}{2}+\epsilon}}\right), \tag{23}$$

then the offline data is informative.

The proof of Proposition C.5 follows Proposition 3.7 of Han et al. [27], which provides a high probability bound on the number of pulls of each arm under UCB1. We restate their result in Lemma D.7, replacing T by M^{off} to reflect that UCB1 is applied in the offline phase. In particular, the number of pulls of arm $i \in [A]$ is denoted by $n_{i;M^{\text{off}}}$, which is the random quantity to be controlled.

The bound in Lemma D.7 depends on several absolute constants, including $n_*(\Theta)$, $\{n^*_{i;M^{\text{off}}}\}_{i\in[A]}$, D_*, and $\text{err}(\Theta)$ [27]. We briefly explain their roles. Let $\Delta^{\text{off}}_i = \mu^{\text{off},*} - \mu^{\text{off}}(i)$ and $\Delta^{\text{off}} = [\Delta^{\text{off}}_i]_{i\in[A]}$. The quantity $n_*(\Theta) \in [0, M^{\text{off}}]$ serves as a reference scale for the number of pulls, determined by the equation

$$\sum_{i\in[A]} n_*(\Theta)\left(1 + n_*^{1/2}(\Theta)\frac{\Delta^{\text{off}}_i}{\gamma_{M^{\text{off}}}}\right)^{-2} = M^{\text{off}}, \tag{35}$$

where $\gamma_{M^{\text{off}}} = \sqrt{2\log(M^{\text{off}})}$ corresponds to the exploration rate in UCB1, matching the $\sqrt{2\log T}$ factor in Auer et al. [6]. Intuitively, $n_*(\Theta)$ captures the overall exploration budget allocated across arms. Based on $n_*(\Theta)$, the constants

$$n^*_{i;M^{\text{off}}} = n_*(\Theta)\left(1 + n_*^{1/2}(\Theta)\frac{\Delta^{\text{off}}_i}{\gamma_{M^{\text{off}}}}\right)^{-2}, \qquad i \in [A]. \tag{36}$$

represent the target number of pulls for each arm. Arms with smaller gaps Δ^{off}_i receive larger $n^*_{i;M^{\text{off}}}$, reflecting increased exploration. Lemma D.7 shows that the random variables $n_{i;M^{\text{off}}}$ concentrates around $n^*_{i;M^{\text{off}}}$ with high probability. The quantity

$$D_* := \frac{1}{M^{\text{off}}}\sum_{i\in[A]}\left(\frac{n^*_{i;M^{\text{off}}}}{n_*(\Theta)}\right)^{1/2} n^*_{i;M^{\text{off}}}$$

is a normalization factor. It is bounded within $[A^{-1/3}, 1]$ (Lemma 6.1 of Han et al. [27]), ensuring it does not degenerate. Finally, $\text{err}(\Theta)$ quantifies the estimation error. It is given by

$$\text{err}(\Theta) = \frac{\sqrt{\log(\gamma_{M^{\text{off}}})} + \sqrt{\log(\log(M^{\text{off}}))}}{\gamma_{M^{\text{off}}}} + \frac{A}{M^{\text{off}}} + \frac{\|\Delta^{\text{off}}\|^2_\infty}{\gamma^2_{M^{\text{off}}}}. \tag{37}$$

We now state Lemma D.7 formally.

Lemma D.7 (Proposition 3.7 of Han et al. [27])**.** *For the offline–online bandit model in Section 4.3, fixing* $\gamma \geq 1$, *there exists a universal constant* $C > 0$ *such that if*

$$\text{err}(\Theta) \vee \left(\frac{\gamma}{\gamma_{M^{\text{off}}}}\right) \leq \frac{1}{C}, \tag{38}$$

then

$$\max_{i\in[A]} \mathbb{P}\left(\left|\frac{n_{i;M^{\text{off}}}}{n^*_{i;M^{\text{off}}}} - 1\right| \geq C\left(D_*^{-1}\text{err}(\Theta) + \frac{\gamma}{\gamma_{M^{\text{off}}}}\right)\right) \leq C\left(\gamma^{-100}_{M^{\text{off}}} + \log(M^{\text{off}})\cdot\gamma e^{-\gamma^2/2}\right). \tag{39}$$

[6] This condition arises from the high-probability bound on the number of pulls in a bandit model provided by Proposition 3.7 of Han et al. [27]. The exponent 49 is not intrinsic to our framework and results from matching the failure probability with the confidence level in the definition of τ. A sharper analysis could potentially relax this requirement.

Based on Lemma D.7, we now establish Proposition C.5.

Proof of Proposition C.5. We begin by showing that in our model, we have the constant $n_*(\Theta) \geq \frac{M^{\text{off}}}{A}$. By Eq. (35), for any $i \in [A]$, $\left(1 + n_*^{1/2}(\Theta)\frac{\Delta_i^{\text{off}}}{\gamma_{M^{\text{off}}}}\right)^{-2} \leq 1$. Thus

$$M^{\text{off}} = \sum_{i\in[A]} n_*(\Theta)\left(1 + n_*^{1/2}(\Theta)\frac{\Delta_i^{\text{off}}}{\gamma_{M^{\text{off}}}}\right)^{-2} \leq An_*(\Theta),$$

which leads to $n_*(\Theta) \geq \frac{M^{\text{off}}}{A}$. Under the assumption that there exists $p \in (0, \frac{1}{2}]$ such that $\Delta_{\text{gap}}^{\text{off}} = O((M^{\text{off}})^{-p})$, and noting that the right-hand side of Eq. (36) is non-increasing in Δ_i^{off}, we obtain that for each arm $i \in [A]$, $n^*_{i;M^{\text{off}}} = \Omega((M^{\text{off}})^{2p})$, up to logarithmic factors arising from $\gamma_{M^{\text{off}}} = \sqrt{2\log(M^{\text{off}})}$.

For the fixed constant C in Eq. (38), the condition Eq. (38) is satisfied by choosing $\gamma = \gamma_{M^{\text{off}}}/4C$, since all three terms in Eq. (37) converge to 0 as $M^{\text{off}} \to \infty$. Thus by combining Eq. (38) and Eq. (39), we obtain that when K and M^{off} are sufficiently large,

$$\begin{aligned}\max_{i\in[A]}\mathbb{P}\left(n_{i;M^{\text{off}}} \leq \frac{1}{2}n^*_{i;M^{\text{off}}}\right) &\leq C\left(\gamma_{M^{\text{off}}}^{-100} + \log(M^{\text{off}})\cdot\gamma e^{-\gamma^2/2}\right)\\ &= C((2\log(M^{\text{off}}))^{-50} + \log(M^{\text{off}})\cdot\frac{\sqrt{2\log(M^{\text{off}})}}{4C}\cdot(M^{\text{off}})^{-\frac{1}{16C^2}}) \leq \frac{1}{K\log(M^{\text{off}})}, \quad (40)\end{aligned}$$

where the second line follows by substituting $\gamma = \gamma_{M^{\text{off}}}/4C$, and the third line uses the assumption $K = o((\log(M^{\text{off}}))^{49})$.

Define the event $\mathcal{E}_B = \{\forall i \in [A], n_{i;M^{\text{off}}} \geq \frac{1}{2}n^*_{i;M^{\text{off}}}\}$. Then by Eq. (40), $\mathbb{P}(\mathcal{E}_B) \geq 1 - \frac{A}{K\log(M^{\text{off}})}$. Conditional on $\mathcal{E}_B$, using the deterministic construction of $\{\widetilde{V}_{m,h}\}$ in Section B.1.2, the offline design matrix satisfies

$$G_{\text{off}} = \text{diag}\left(\frac{n_{1;M^{\text{off}}}}{3}, \frac{n_{1;M^{\text{off}}}}{3}, \frac{n_{1;M^{\text{off}}}}{3}, \frac{n_{2;M^{\text{off}}}}{3}, \ldots, \frac{n_{A;M^{\text{off}}}}{3}\right).$$

By the definition of $\tau(\delta)$ in Definition 4.1, we directly obtain that $\tau\left(\frac{A}{K\log(M^{\text{off}})}\right) \geq \Omega\left(\frac{(M^{\text{off}})^{2p}}{M^{\text{off}}}\right) = \Omega((M^{\text{off}})^{2p-1})$.

Note that $\frac{A}{K\log(M^{\text{off}})}$ differs from the conventional choice of δ used throughout this paper, namely $\frac{1}{2KM^{\text{off}}}$. Nevertheless, returning to the explicit regret upper bound for arbitrary K and M^{off} in Eq. (26) (see the proof of Theorem 4.2), and choosing $\delta = \frac{1}{2KM^{\text{off}}}$ and $\delta' = \frac{A}{K\log(M^{\text{off}})}$ yields the same guarantee as in Theorem 4.2. The bound in Equation (23) then follows by the same argument as in Corollary 4.4, completing the proof. □

D.4 Proofs in Section 4.4

In this section we provide a detailed proof of Theorem 4.13.

Theorem 4.13. *Under environment shift (Assumption 2.3) and coverage condition (Definition 4.1), for any $\Delta, H \geq 2$, and $d \geq 2$, for any sufficiently large K and M^{off}, and any offline–online algorithm, there exists an offline–online linear mixture MDP instance such that, whenever the offline data induces $\tau = \Theta(1)$, any algorithm has regret at least*

$$\Omega\left(\min\left\{dH\sqrt{K}, dH\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}(1 + \frac{\Delta}{\sqrt{d}}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])})\right)\right\}\right). \quad (14)$$

Proof. We organize the proof in two parts. First, we establish the regret lower bound for an offline–online linear bandit instance (Step 1). Second, we embed this offline–online linear bandit instance into an offline–online linear mixture MDP and complete the proof (Step 2).

Step 1: regret lower bound for offline–online linear bandit instance. We begin with the offline–online linear bandit instance. Inspired by Lattimore and Szepesvári [35], for any given dimension d and $\tilde{\Delta}$, and for any fixed K and M^{off}, we consider a class of offline–online linear bandit instances in which the environment shift is defined under the $\|\cdot\|_\infty$ norm and bounded by $\tilde{\Delta}$. Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d \times \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d$, where $c_{\tilde{\Delta}}$ depends on $\tilde{\Delta}$, K, and M^{off}. The value of $c_{\tilde{\Delta}}$ will be specified later. Concretely, for any instance $\theta = (\theta^{\text{off},*}, \theta^*) \in \Theta$, assume that for all $i \in [d]$,

$$|\theta_i^{\text{off},*} - \theta_i^*| \le \tilde{\Delta}. \tag{41}$$

Note that this defines $\tilde{\Delta}$ in terms of the ℓ_∞ norm, but this relates to the ℓ_2 definition in Assumption 2.3 up to a $\sqrt{d}$ factor. Furthermore, for any action $a \in \mathcal{A}$ and instance $\theta \in \Theta$, the reward of the offline bandit is a Bernoulli random variable with mean $\frac{1}{2} + \langle a, \theta^{\text{off},*}\rangle$, while the reward of the online bandit is a Bernoulli random variable with mean $\frac{1}{2} + \langle a, \theta^*\rangle$.

We next establish the lower bound for this instance class Θ:

Lemma D.8. *Under any fixed offline behavior policy and for any offline–online algorithm, for any sufficiently large K and M^{off}, there exists an offline–online linear bandit instance $\theta = (\theta^{\text{off},*}, \theta^*) \in \Theta$ such that, as long as the induced coverage satisfies $\tau = \Theta(1)$, the expected regret is at least*

$$\mathbb{E}[\textit{Regret}(K)] = \Omega\left(\min\left\{d\sqrt{K}, d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])})\right)\right\}\right). \tag{42}$$

The proof of Lemma D.8 largely follows the lower bound construction for linear bandits in Lattimore and Szepesvári [35]. A key difference in our setting is the need to capture the "min-of-two" structure in Eq. (42), which requires constructing different hard instances under different regimes where distinct terms dominate the bound. A detailed discussion of each regime, along with the full proof of Lemma D.8, is provided in Section D.4.1.

Step 2: additional H factor by embedding in a linear mixture MDP. We next prove Eq. (14) by embedding the offline–online linear bandit instances constructed via Lemma D.8 into an offline–online linear mixture MDP. Concretely, for any H, Δ, and $d \ge 2$, consider the following offline–online linear mixture MDP instance class. The state space is $\mathcal{S} = \{s_1^*, s_2^*, s_3^*\}$ and the action space is $\mathcal{A} = [-1, 1]^{d-1}$. The initial state of every episode is $s_1 = s_1^*$. For any action $a \in \mathcal{A}$ and any parameter $\theta = (\theta^{\text{off},*}, \theta^*) \in \Theta$, where Θ is the $(d-1)$-dimensional offline–online linear bandit instance class constructed above, we define the transition probabilities as

$$P^{\text{off}}(s_2^* \mid s_1^*, a) = \frac{1}{2} + \langle \theta^{\text{off},*}, a\rangle, \qquad P^{\text{off}}(s_3^* \mid s_1^*, a) = \frac{1}{2} - \langle \theta^{\text{off},*}, a\rangle, \tag{43}$$

$$P(s_2^* \mid s_1^*, a) = \frac{1}{2} + \langle \theta^*, a\rangle, \qquad P(s_3^* \mid s_1^*, a) = \frac{1}{2} - \langle \theta^*, a\rangle. \tag{44}$$

Thus at step $h = 2$ in both offline environment and online environment, the system transitions to either s_2^* or s_3^*. We further set

$$P^{\text{off}}(s_2^* \mid s_2^*, a) = 1, \quad P^{\text{off}}(s_3^* \mid s_3^*, a) = 1, \quad P(s_2^* \mid s_2^*, a) = 1, \quad P(s_3^* \mid s_3^*, a) = 1, \quad \forall a \in \mathcal{A}.$$

The reward is 1 at state s_2^* and 0 at state s_3^*.

By construction, the transition from s_1^* determines whether the process enters s_2^* or s_3^*, and the obtained reward is then repeated for the remaining $H - 1$ steps. Therefore this MDP interaction is equivalent to an offline–online linear bandit interaction whose reward is scaled by a factor H. Let $\mathscr{A}_{\text{MDP}}$ denote the class of offline–online algorithms for the linear mixture MDP problem, and $\mathscr{A}_{\text{B}}$ denote the class of algorithms for the offline–online linear bandit problem. Also let Θ denote the $(d-1)$-dimensional offline–online linear bandit instance class constructed in the first part of the proof. Based on Θ, consider any offline–online linear mixture MDP instance class Θ_0 that contains the subset of instances constructed above following Eq. (43) and Eq. (44). Any algorithm $\in \mathscr{A}_{\text{MDP}}$

induces an algorithm in $\mathscr{A}_{\mathrm{B}}$ when restricted to this embedded bandit interaction. Consequently,

$$\begin{aligned}
\inf_{\in \mathscr{A}_{\mathrm{MDP}}} \sup_{\theta \in \Theta_0} \mathbb{E}[\mathsf{Regret}(K)] &\geq H \cdot \inf_{\in \mathscr{A}_{\mathrm{B}}} \sup_{\theta \in \Theta} \mathbb{E}[\mathsf{Regret}(K)] \\
&= \Omega\Bigg(\min\Bigg\{ (d-1)H\sqrt{K}, (d-1)H\sqrt{K}\Bigg(\sqrt{\log\Bigg(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\mathrm{off}}])}\Bigg)} \\
&\quad (1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\mathrm{off}}])})\Bigg)\Bigg\}\Bigg) \\
&= \Omega\Bigg(\min\Bigg\{ dH\sqrt{K}, dH\sqrt{K}\Bigg(\sqrt{\log\Bigg(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\mathrm{off}}])}\Bigg)} \\
&\quad (1 + \frac{\Delta}{\sqrt{d}}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\mathrm{off}}])})\Bigg)\Bigg\}\Bigg),
\end{aligned}$$

where the second line follows from Eq. (42) applied to the $(d-1)$-dimensional linear bandit instance class, and the last line uses the relation between the ℓ_∞ and ℓ_2 environment shift parameters. For any Δ defined under the $\|\cdot\|_2$ norm, there exists an instance with $\tilde{\Delta} = \Delta/\sqrt{d-1}$ under the $\|\cdot\|_\infty$ norm that satisfies the same shift constraint. □

D.4.1 Proof of Lemma D.8

Proof. Note that for any given θ^*, the optimal action is $a^* = [\mathrm{sign}(\theta_i^*)]_{i\in[d]}$, since $\mathcal{A} = [-1,1]^d$. Now, fix $i \in [d]$ and $\theta = (\theta^{\mathrm{off},*}, \theta^*) \in \Theta$, and define

$$p_{\theta i} = \mathbb{P}_\theta\left(\sum_{k=1}^K \mathbf{1}\{\mathrm{sign}(a_{ki}) \neq \mathrm{sign}(\theta_i^*)\} \geq \frac{K}{2} \right), \tag{45}$$

where a_{ki} denotes the i-th coordinate of the action a_k. Thus, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}[\mathsf{Regret}(K)] &= \mathbb{E}_\theta\left[\sum_{k=1}^K \sum_{i=1}^d (\mathrm{sign}(\theta_i^*) - a_{ki})\theta_i^* \right] \\
&\geq c_{\tilde{\Delta}} \sum_{i=1}^d \mathbb{E}_\theta\left[\sum_{k=1}^K \mathbf{1}\{\mathrm{sign}(a_{ki}) \neq \mathrm{sign}(\theta_i^*)\} \right] \\
&\geq \frac{K c_{\tilde{\Delta}}}{2} \sum_{i=1}^d \mathbb{P}_\theta\left(\sum_{k=1}^K \mathbf{1}\{\mathrm{sign}(a_{ki}) \neq \mathrm{sign}(\theta_i^*)\} \geq \frac{K}{2} \right) \\
&= \frac{K c_{\tilde{\Delta}}}{2} \sum_{i=1}^d p_{\theta i},
\end{aligned} \tag{46}$$

where the first inequality is from the definition of Θ and $\mathcal{A}$, the third line follows from Markov's inequality [35], and the fourth line follows from substituting the definition of $p_{\theta i}$.

Next, we lower bound $\sum_{i=1}^d p_{\theta i}$ via a pairing argument. For each $i \in [d]$, define θ' by flipping the i-th coordinate of θ^* while keeping all other coordinates fixed, and choose $\theta'^{\mathrm{off},*} \in \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d$ so that $\theta' \in \Theta$. (We suppress the dependence of θ' on i in the notation.) Restricting attention to a subset of Θ that is closed under this mapping, we obtain a one-to-one pairing between θ and θ'. For each such pair, we bound $p_{\theta i} + p_{\theta' i}$, and summing over $i \in [d]$ yields a lower bound on $f(\theta) + f(\theta')$, where $f(\theta) := \sum_{i=1}^d p_{\theta i}$. Averaging over all θ in this subset and using the one-to-one pairing, we conclude that there exists $\theta \in \Theta$ such that $\sum_{i=1}^d p_{\theta i}$ satisfies the desired lower bound.

To bound $p_{\theta i} + p_{\theta' i}$, let P_θ and $P_{\theta'}$ denote the probability measures over outcomes under θ and θ', respectively. Concretely, under any fixed algorithm and environment $\theta = (\theta^{\mathrm{off},*}, \theta^*)$, let P_θ denote the joint measure over

$$\Big(\underbrace{(\ldots, a_{M^{\mathrm{off}}}, r_{M^{\mathrm{off}}})}_{\text{governed by } \theta^{\mathrm{off},*}}, \underbrace{(\ldots, a_K, r_K)}_{\text{governed by } \theta^*} \Big),$$

where we slightly abuse notations and use a_m and a_k to denote the actions taken at episode m in the offline phase and episode k in the online phase, respectively. $P_{\theta'}$ is defined analogously. Let $D_{KL}(P\|Q)$ denote the Kullback–Leibler divergence between two probability measures P and Q. Suppose that for all $m \in [M^{\text{off}}]$ and $k \in [K]$,

$$\frac{1}{4} < \frac{1}{2} - |\langle a_m, \theta^{\text{off},*}\rangle| \le \frac{1}{2} + |\langle a_m, \theta^{\text{off},*}\rangle| < \frac{3}{4}, \frac{1}{4} < \frac{1}{2} - |\langle a_k, \theta^*\rangle| \le \frac{1}{2} + |\langle a_k, \theta^*\rangle| < \frac{3}{4}. \tag{47}$$

Then

$$\begin{aligned}
D_{KL}(P_\theta \| P_{\theta'}) &= \mathbb{E}_\theta \Bigg[\sum_{m=1}^{M^{\text{off}}} D_{KL}(B(\frac{1}{2} + \langle a_m, \theta^{\text{off},*}\rangle) \,\|\, B(\frac{1}{2} + \langle a_m, \theta'^{\text{off},*}\rangle)) \\
&\quad + \sum_{k=1}^{K} D_{KL}(B(\frac{1}{2} + \langle a_k, \theta^*\rangle) \,\|\, B(\frac{1}{2} + \langle a_k, \theta'^*\rangle)) \Bigg] \\
&= \mathbb{E}_\theta \Bigg[\sum_{m=1}^{M^{\text{off}}} \left(\frac{1}{2} + \langle a_m, \theta^{\text{off},*}\rangle\right) \log\left(\frac{\frac{1}{2} + \langle a_m, \theta^{\text{off},*}\rangle}{\frac{1}{2} + \langle a_m, \theta'^{\text{off},*}\rangle}\right) \\
&\quad + \left(\frac{1}{2} - \langle a_m, \theta^{\text{off},*}\rangle\right) \log\left(\frac{\frac{1}{2} - \langle a_m, \theta^{\text{off},*}\rangle}{\frac{1}{2} - \langle a_m, \theta'^{\text{off},*}\rangle}\right) \\
&\quad + \sum_{k=1}^{K} \left(\frac{1}{2} + \langle a_k, \theta^*\rangle\right) \log\left(\frac{\frac{1}{2} + \langle a_k, \theta^*\rangle}{\frac{1}{2} + \langle a_k, \theta'^*\rangle}\right) + \left(\frac{1}{2} - \langle a_k, \theta^*\rangle\right) \log\left(\frac{\frac{1}{2} - \langle a_k, \theta^*\rangle}{\frac{1}{2} - \langle a_k, \theta'^*\rangle}\right) \Bigg] \qquad (48) \\
&\le \frac{8}{3} \mathbb{E}_\theta \left[\sum_{m=1}^{M^{\text{off}}} \langle a_m, (\theta^{\text{off},*} - \theta'^{\text{off},*})\rangle^2 + \sum_{k=1}^{K} \langle a_k, (\theta^* - \theta'^*)\rangle^2 \right], \qquad (49)
\end{aligned}$$

where the first equality follows from the standard decomposition of Kullback–Leibler divergence (see Section 4.7 of Lattimore and Szepesvári [35]), and the second equality follows from the definition of the Kullback–Leibler divergence. The inequality is from the following bound on the Kullback–Leibler divergence between Bernoulli distributions: for two Bernoulli random variables with means μ_1 and μ_2, if $\frac{1}{4} \le \mu_1 \le \frac{3}{4}$ and $\frac{1}{4} \le \mu_2 \le \frac{3}{4}$, then $D_{KL}(B(\mu_1)\|B(\mu_2)) \le \frac{8}{3}(\mu_1 - \mu_2)^2$. We omit the detailed proof.

Recall that $\theta'^*_j = \theta^*_j$ for $j \ne i$, $\theta'^*_i = -\theta^*_i$, and $\theta'^{\text{off},*} \in \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d$. By the Bretagnolle–Huber inequality (Theorem 14.2 of Lattimore and Szepesvári [35]) and Eq. (49),

$$p_{\theta i} + p_{\theta' i} \ge \frac{1}{2} \exp\left(-\frac{8}{3} \mathbb{E}_\theta \left[\sum_{m=1}^{M^{\text{off}}} \langle a_m, (\theta^{\text{off},*} - \theta'^{\text{off},*})\rangle^2 + \sum_{k=1}^{K} \langle a_k, (\theta^* - \theta'^*)\rangle^2 \right] \right). \tag{50}$$

To complete the proof of Lemma D.8, we consider the following four cases.

- **Case 1**: $d\sqrt{K} \ge d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, as well as $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} < 1$.
- **Case 2**: $d\sqrt{K} \ge d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, as well as $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} \ge 1$.
- **Case 3**: $d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, as well as $\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} \ge \frac{1}{2}$.

- **Case 4**: $d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, as well as $\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} < \frac{1}{2}$.

Cases 1 and 2 correspond to regimes where the offline–online term dominates in Eq. (42). Cases 3 and 4 correspond to regimes where the online-only term dominates. For each case, we verify that (i) the condition in Eq. (47) holds, and (ii) a lower bound on $\sum_{i=1}^{d} p_{\theta i}$ can be established via Eq. (50). The desired lower bound on the expected regret then follows from Eq. (46).

Case 1: $d\sqrt{K} \geq d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, together with that $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} < 1$. In this case, we set $c_{\tilde{\Delta}} = \frac{1}{\sqrt{K+M^{\text{off}}}}$. We focus on a subset of Θ where the offline environment is identical to the online environment. Denote

$$\Theta_1 = \{\theta \in \Theta \mid \theta^{\text{off},*} = \theta^*\}. \tag{51}$$

As a consequence, the environment shift condition Eq. (41) is automatically satisfied.

We first verify the condition in Eq. (47). In our setting, for any actions a_m and a_k and any $\theta \in \Theta_1$,

$$|\langle a_m, \theta^{\text{off},*}\rangle| \leq \frac{d}{\sqrt{K+M^{\text{off}}}}, \qquad |\langle a_k, \theta^*\rangle| \leq \frac{d}{\sqrt{K+M^{\text{off}}}}.$$

Therefore, for sufficiently large M^{off} and K satisfying $K + M^{\text{off}} \geq 16d^2$,

$$\frac{1}{4} < \frac{1}{2} - |\langle a_m, \theta^{\text{off},*}\rangle| \leq \frac{1}{2} + |\langle a_m, \theta^{\text{off},*}\rangle| < \frac{3}{4}, \frac{1}{4} < \frac{1}{2} - |\langle a_k, \theta^*\rangle| \leq \frac{1}{2} + |\langle a_k, \theta^*\rangle| < \frac{3}{4}. \tag{52}$$

Thus Eq. (47) holds.

For any $i \in [d]$, by Eq. (50),

$$p_{\theta i} + p_{\theta' i} \geq \frac{1}{2}\exp\left(-\frac{8}{3}\mathbb{E}_\theta\left[\sum_{m=1}^{M^{\text{off}}}\langle a_m, (\theta^* - \theta'^*)\rangle^2 + \sum_{k=1}^{K}\langle a_k, (\theta^* - \theta'^*)\rangle^2\right]\right) \geq \frac{1}{2}\exp\left(-\frac{32}{3}\right), \tag{53}$$

where the second inequality follows from $c_{\tilde{\Delta}} = \frac{1}{\sqrt{K+M^{\text{off}}}}$ and $\mathcal{A} = [-1,1]^d$. Summing over $\theta \in \Theta_1$ and $i \in [d]$, we obtain

$$\sum_{\theta\in\Theta_1}\frac{1}{|\Theta_1|}\sum_{i=1}^{d} p_{\theta i} = \frac{1}{|\Theta_1|}\sum_{i=1}^{d}\sum_{\theta\in\Theta_1} p_{\theta i} \geq \frac{d}{4}\exp\left(-\frac{32}{3}\right). \tag{54}$$

Therefore, there exists $\theta \in \Theta_1$ such that $\sum_{i=1}^{d} p_{\theta i} \geq \frac{d}{4}\exp(-\frac{32}{3})$. For this specific choice of θ, following Eq. (46),

$$\mathbb{E}[\mathsf{Regret}(K)] \geq \frac{Kc_{\tilde{\Delta}}}{2}\sum_{i=1}^{d} p_{\theta i} = \frac{K}{2\sqrt{K+M^{\text{off}}}}\sum_{i=1}^{d} p_{\theta i} \geq \frac{d}{8}\exp\left(-\frac{32}{3}\right)\frac{K}{\sqrt{K+M^{\text{off}}}}, \tag{55}$$

where the equality follows from substituting the definition of $c_{\tilde{\Delta}}$.

Under the condition $\tau = \Theta(1)$, for any constant H, there exists a constant $c > 0$ such that for sufficiently large M^{off},

$$\mathbb{P}(\lambda_{\min}(G_{\text{off}}) \geq cM^{\text{off}} - H^2 d) \geq 1 - \frac{1}{2KM^{\text{off}}} \Rightarrow \mathbb{P}(\lambda_{\min}(G_{\text{off}}) \geq \frac{c}{2}M^{\text{off}}) \geq 1 - \frac{1}{2KM^{\text{off}}}.$$

Note that G_{off} is positive semidefinite, and therefore $\lambda_{\min}(G_{\text{off}}) \geq 0$ a.s.. Furthermore,

$$\lambda_{\min}(\mathbb{E}[G_{\text{off}}]) \geq \mathbb{E}[\lambda_{\min}(G_{\text{off}})] \geq \frac{c}{2}M^{\text{off}}\left(1 - \frac{1}{2KM^{\text{off}}}\right), \tag{56}$$

where the first inequality follows from Lemma E.8, and the second inequality follows from the above high-probability bound together with $\lambda_{\min}(G_{\text{off}}) \geq 0$ a.s..

Note that the assumption in Case 1,

$$d\sqrt{K} \geq d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$$

and $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} < 1$, imply that

$$\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} \leq 1 \Rightarrow K \leq (e-1)\lambda_{\min}(\mathbb{E}[G_{\text{off}}]), \tag{57}$$

since $1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} \geq 1$ holds trivially. Combining Eq. (55) and Eq. (56),

$$\begin{aligned}
\mathbb{E}[\mathsf{Regret}(K)] &\geq \frac{d}{8}\exp\left(-\frac{32}{3}\right)\frac{K}{\sqrt{K+\frac{2\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}{c\left(1-\frac{1}{2KM^{\text{off}}}\right)}}} = \Omega\left(d\frac{K}{\sqrt{K+\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}}\right) \\
&= \Omega\left(d\sqrt{K}\sqrt{\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}}\right) = \Omega\left(d\sqrt{K}\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\right) \\
&= \Omega\left(\min\left\{d\sqrt{K}, d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)\right\}\right),
\end{aligned} \tag{58}$$

where the second line uses Eq. (57), which holds by the condition defining Case 1. The third equality uses the inequality $x \geq \log(1+x)$ for all $x > -1$, and the last line follows from the assumption $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} < 1$.

Case 2: $d\sqrt{K} \geq d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, together with that $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} \geq 1$. Here the instance class setting is similar to that in Case 1, with the only difference being the choice of $c_{\tilde{\Delta}}$ and the environment shift. In this case, we set $c_{\tilde{\Delta}} = \frac{\tilde{\Delta}}{2}$. Thus for any $\theta = (\theta^{\text{off},*}, \theta^*) \in \Theta = \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d \times \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d$, $\|\theta^{\text{off},*} - \theta^*\|_\infty \leq \tilde{\Delta}$.

Denote

$$\Theta_2 = \{\theta = (\theta^{\text{off},*}, \theta^*) \in \Theta \mid \theta_i^{\text{off},*} = c_{\tilde{\Delta}}, \forall i \in [d]\}, \tag{59}$$

that is, the offline environment is identical for all $\theta \in \Theta_2$.

The condition defining Case 2,

$$d\sqrt{K} \geq d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right),$$

implies

$$\begin{aligned}
&\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} \leq 1 \\
&\Rightarrow \tilde{\Delta} \leq \frac{1}{\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}}.
\end{aligned} \tag{60}$$

In turn, Eq. (60) ensures that the condition in Eq. (47) is satisfied when $K \geq \frac{32}{\log(3)}d^2$, as

$$\begin{aligned}
\max_{m,k}\{|\langle a_m, \theta^{\text{off},*}\rangle|, |\langle a_k, \theta^*\rangle|\} \leq c_{\tilde{\Delta}}d = \frac{\tilde{\Delta}}{2}d &\leq \frac{d}{2}\frac{1}{\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}} \\
&\leq \frac{d}{\sqrt{\frac{\log(3)}{2}K}} \leq \frac{1}{4},
\end{aligned} \tag{61}$$

where the first inequality is from the definition of Θ and $\mathcal{A}$, the equality is from $c_{\tilde{\Delta}} = \frac{\tilde{\Delta}}{2}$, the second inequality is from Eq. (60), and the third inequality is from the bound $x/\log(1+x) \leq 2/\log(3)$ for all $x \leq 2$, together with that $\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} \leq 1$ (similar to Eq. (57)), which implies that $\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} \leq e - 1 \leq 2$.

For any $i \in [d]$, by Eq. (50),

$$\begin{aligned} p_{\theta i} + p_{\theta' i} &\geq \frac{1}{2}\exp\left(-\frac{8}{3}\mathbb{E}_\theta\left[\sum_{k=1}^{K}\langle a_k, (\theta^* - \theta'^*)\rangle^2\right]\right) \geq \frac{1}{2}\exp\left(-\frac{8}{3}\tilde{\Delta}^2 K\right) \\ &\geq \frac{1}{2}\exp\left(-\frac{8}{3}\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\right) \\ &\geq \frac{1}{2}\exp\left(-\frac{8}{3}\frac{2}{\log(3)}\right) = \frac{1}{2}\exp\left(-\frac{16}{3\log(3)}\right), \end{aligned}$$

where the first line follows from the definition of Θ_2 together with $\mathcal{A} = [-1, 1]^d$. The second line then follows from Eq. (60). The third line again follows from

$$\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} \leq 1,$$

as in Eq. (61).

Thus, similar to Eq. (54), there exists a specific $\theta \in \Theta_2$ such that $\sum_{i=1}^{d} p_{\theta i} \geq \frac{d}{4}\exp\left(-\frac{16}{3\log(3)}\right)$. For this choice of θ, following Eq. (46), we obtain

$$\begin{aligned} \mathbb{E}[\mathsf{Regret}(K)] &\geq \frac{K c_{\tilde{\Delta}}}{2}\sum_{i=1}^{d} p_{\theta i} = \frac{K\tilde{\Delta}}{4}\sum_{i=1}^{d} p_{\theta i} \geq \exp\left(-\frac{16}{3\log(3)}\right)\cdot\frac{K\tilde{\Delta}d}{16} \\ &= \Omega\left(d\sqrt{K}\sqrt{\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}}(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])})\right) \\ &= \Omega\left(\min\left\{d\sqrt{K}, d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)\right\}\right), \end{aligned} \tag{62}$$

where the second line directly follows from the assumption in Case 2 that $\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} \geq 1$, and the third line also follows from the assumption in Case 2 that

$$d\sqrt{K} \geq d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right),$$

together with the inequality that $x \geq \log(1+x)$ for all $x > 0$.

Case 3: $d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1 + \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, together with that $\sqrt{\log\left(1 + \frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} \geq \frac{1}{2}$. We consider the same setting as in Case 1. We set $c_{\tilde{\Delta}} = \frac{1}{\sqrt{K + M^{\text{off}}}}$ and focus on the instance class Θ_1 defined in Eq. (51). Following the argument in Case 1, Eq. (47) holds when $K + M^{\text{off}} \geq 16d^2$, and there exists $\theta \in \Theta_1$ such that

$$\mathbb{E}[\mathsf{Regret}(K)] = \Omega\left(d\frac{K}{\sqrt{K + \lambda_{\min}(\mathbb{E}[G_{\text{off}}])}}\right),$$

which follows from Eq. (58).

Under the condition defining Case 3, $\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} \geq \frac{1}{2}$, which implies that $K \geq (e^{\frac{1}{4}}-1)\lambda_{\min}(\mathbb{E}[G_{\text{off}}])$. Consequently,

$$\begin{aligned}\mathbb{E}[\mathsf{Regret}(K)] &= \Omega\left(d\frac{K}{\sqrt{K+\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}}\right) = \Omega\left(d\sqrt{K}\right) \\ &= \Omega\left(\min\left\{d\sqrt{K}, d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)\right\}\right),\end{aligned}$$

where the second line follows from the condition defining Case 3 that

$$d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right).$$

Case 4: $d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$, together with that $\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} < \frac{1}{2}$. We consider the same setting as in Case 2 with a different choice of $c_{\tilde{\Delta}}$, since the condition defining Case 4 implicitly restricts the choice of $\tilde{\Delta}$ in a different way. Note that the condition defining Case 4,

$$d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$$

implies that

$$\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} + \sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} > 1.$$

Since $\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} < \frac{1}{2}$ (also see the condition defining Case 4), we obtain

$$\tilde{\Delta}\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])} > \frac{1}{2}.$$

Thus

$$\begin{aligned}\tilde{\Delta}\sqrt{K} &= \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\sqrt{\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}} \\ &\geq \tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)} > \frac{1}{2},\end{aligned}$$

where the first inequality follows from $x \geq \log(1+x)$ for all $x > 0$. Therefore, by choosing

$$c_{\tilde{\Delta}} = \frac{1}{4\sqrt{K}}, \tag{63}$$

for any $\theta = (\theta^{\text{off},*}, \theta^*) \in \Theta = \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d \times \{-c_{\tilde{\Delta}}, c_{\tilde{\Delta}}\}^d$, $\|\theta^{\text{off},*} - \theta^*\|_\infty \leq \tilde{\Delta}$. This allows us to consider Θ_2 as defined in Eq. (59).

For any $i \in [d]$, following Eq. (50), and noting that Eq. (63) guarantees that the condition Eq. (47) holds when $K \geq d^2$, we obtain

$$p_{\theta i} + p_{\theta' i} \geq \frac{1}{2}\exp\left(-\frac{8}{3}\mathbb{E}_\theta\left[\sum_{k=1}^K \langle a_k, (\theta^* - \theta'^*)\rangle^2\right]\right) \geq \frac{1}{2}\exp\left(-\frac{2}{3}\right).$$

Similar to Eq. (54), there exists $\theta \in \Theta_2$ such that $\sum_{i=1}^{d} p_{\theta i} \geq \frac{d}{4}\exp\left(-\frac{2}{3}\right)$. For this choice of θ, following Eq. (46),

$$\mathbb{E}[\mathsf{Regret}(K)] \geq \frac{K c_{\tilde{\Delta}}}{2} \sum_{i=1}^{d} p_{\theta i} = \Omega(d\sqrt{K})$$

$$= \Omega\left(\min\left\{d\sqrt{K}, d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)\right\}\right),$$

where the second line follows from Eq. (63), and the last line follows from the condition defining Case 4 that $d\sqrt{K} < d\sqrt{K}\left(\sqrt{\log\left(1+\frac{K}{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)}\left(1+\tilde{\Delta}\sqrt{\lambda_{\min}(\mathbb{E}[G_{\text{off}}])}\right)\right)$. □

E Useful facts

In this section we collect several auxiliary results that are used throughout the proofs in Section D. These include self-normalized concentration inequalities adopted from the literature, a matrix Chernoff bound, as well as basic algebraic inequalities that are invoked repeatedly. We begin with the self-normalized concentration inequality of Abbasi-Yadkori et al. [1].

Lemma E.1 (Theorem 1 of Abbasi-Yadkori et al. [1])**.** *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that x_t is F_{t-1}-measurable. Assume that W is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$\bar{W}_t = W + \sum_{s=1}^{t} x_s x_s^{\top}, \qquad S_t = \sum_{s=1}^{t} \eta_s x_s.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

$$\|S_t\|^2_{\bar{W}_t^{-1}} \leq 2R^2 \log\left(\frac{\det(\bar{W}_t)^{1/2}\det(W)^{-1/2}}{\delta}\right).$$

We refer the reader to Abbasi-Yadkori et al. [1] for a detailed proof of Lemma E.1. Our analysis also relies on a matrix concentration inequality, for which we invoke standard results from Tropp [49].

Lemma E.2 (Corollay 5.2 of Tropp [49])**.** *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices that satisfy $X_k \succeq 0$ and $\lambda_{\max}(X_k) \leq R$ a.s.. Compute the minimum eigenvalue of the sum of expectations $\mu_{\min} = \lambda_{\min}(\sum_k \mathbb{E}X_k)$. Then for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_k X_k\right) \leq (1-\delta)\mu_{\min}\right) \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R}. \tag{64}$$

where $X_k \in \mathbb{R}^{d\times d}$ for any k. Furthermore, the bound above admits the following simplified form:

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_k X_k\right) \leq (1-\delta)\mu_{\min}\right) \leq d e^{-\frac{\delta^2}{2}\frac{\mu_{\min}}{R}}. \tag{65}$$

Proof. We refer the reader to Tropp [49] for the proof of the first part of Lemma E.2, namely Eq. (64). Building on this result, we now prove Eq. (65). Note that from Eq. (64), it suffices to show that for every $\delta \in (0,1)$, $\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \leq \exp\left(-\frac{\delta^2}{2}\right)$. Taking logarithms, this is equivalent to $-\delta-(1-\delta)\log(1-\delta) \leq -\frac{\delta^2}{2}$, or, equivalently, $g(\delta) := -\delta-(1-\delta)\log(1-\delta)+\frac{\delta^2}{2} \leq 0$. We verify that $g(\delta) \leq 0$ on $(0,1)$ using concavity. A direct calculation gives that for any $\delta \in (0,1)$, $g'(\delta) = \log(1-\delta)+\delta$ and $g''(\delta) = -\frac{\delta}{1-\delta} \leq 0$. Hence g is concave on $(0,1)$, so its derivative g' is nonincreasing. Moreover, since $g'(0) = 0$, we have that for every $\delta \in (0,1)$, $g'(\delta) \leq g'(0) = 0$, which implies that g is nonincreasing on $(0,1)$. Since $g(0) = 0$, we conclude that $g(\delta) \leq g(0) = 0$ for all $\delta \in (0,1)$. So far the proof is complete. □

Next, we present several elementary lemmas concerning algebraic properties of the weighted norm $\|\cdot\|_A$, where A is a symmetric and positive definite matrix.

Lemma E.3. *Let $A \in \mathbb{R}^{d\times d}$ be symmetric and positive definite, and let $B = A + \sum_i x_i x_i^\top$ for some vectors $\{x_i\} \subset \mathbb{R}^d$. Then for any $y \in \mathbb{R}^d$, $\|y\|_A \le \|y\|_B$.*

Proof. Note that $y^\top B y = y^\top A y + \sum_i (x_i^\top y)^2 \ge y^\top A y$. Taking square roots on both sides yields $\|y\|_B \ge \|y\|_A$, as claimed. □

Lemma E.4. *Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric and positive definite matrices. If $A \succeq B \succ 0$, that is, $A - B$ is positive semidefinite, then $B^{-1} \succeq A^{-1}$.*

Proof. Since $A \succeq B \succ 0$, we can write $A = B^{1/2}(I + C)B^{1/2}$, where $C \succeq 0$. Taking inverses gives $A^{-1} = B^{-1/2}\big(I + C\big)^{-1} B^{-1/2}$. Because $C \succeq 0$, all eigenvalues of $I + C$ are at least 1, hence $(I + C)^{-1} \preceq I$. Therefore, $A^{-1} \preceq B^{-1/2} I B^{-1/2} = B^{-1}$, which proves the claim. □

Lemma E.5. *Let $A \in \mathbb{R}^{d\times d}$ be symmetric and positive definite. Then for any $x \in \mathbb{R}^d$, $\|x\|_A \le \sqrt{\lambda_{\max}(A)}\|x\|_2$.*

Proof. Since A is symmetric positive definite, it admits an eigendecomposition $A = U\Lambda U^\top$, where U is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ with $0 < \lambda_i \le \lambda_{\max}(A)$ for all i. Let $z = U^\top x$. Then $\|z\|_2 = \|x\|_2$ and

$$\begin{aligned}\|x\|_A^2 = x^\top A x = x^\top U\Lambda U^\top x = z^\top \Lambda z = \sum_{i=1}^d \lambda_i z_i^2 \le \lambda_{\max}(A)\sum_{i=1}^d z_i^2 \\ = \lambda_{\max}(A)\|z\|_2^2 = \lambda_{\max}(A)\|x\|_2^2.\end{aligned}$$

Taking square roots on both sides yields the claim. □

Lemma E.6. *Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite and $B \in \mathbb{R}^{d\times d}$ be symmetric positive definite. Then*

$$\log(\det(A + B)) - \log(\det(A)) \le d \log\left(1 + \frac{1}{d}\operatorname{tr}(A^{-1}B)\right).$$

Proof. We write

$$\log(\det(A + B)) - \log(\det(A)) = \log\det\left(I + A^{-1/2}BA^{-1/2}\right).$$

Let $\widetilde{B} := A^{-1/2}BA^{-1/2}$. Since that $B \succeq 0$ and $A^{-1/2}$ is symmetric, $\widetilde{B}$ is symmetric PSD. Let $\nu_1, \dots, \nu_d \ge 0$ be the eigenvalues of $\widetilde{B}$. Then $\log\det(I + \widetilde{B}) = \sum_{i=1}^d \log(1 + \nu_i)$. Thus,

$$\prod_{i=1}^d (1 + \nu_i) \le \left(\frac{1}{d}\sum_{i=1}^d (1 + \nu_i)\right)^d,$$

hence

$$\log(\det(I + \widetilde{B})) \le d\log\left(1 + \frac{1}{d}\sum_{i=1}^d \nu_i\right) = d\log\left(1 + \frac{1}{d}\operatorname{tr}(\widetilde{B})\right).$$

Finally, the proof is complete with $\operatorname{tr}(\widetilde{B}) = \operatorname{tr}(A^{-1/2}BA^{-1/2}) = \operatorname{tr}(A^{-1}B)$. □

Lemma E.7. *Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite and let $B \in \mathbb{R}^{d\times d}$ be symmetric positive semi-definite. Then $\operatorname{tr}(AB) \le \lambda_{\max}(A)\operatorname{tr}(B)$.*

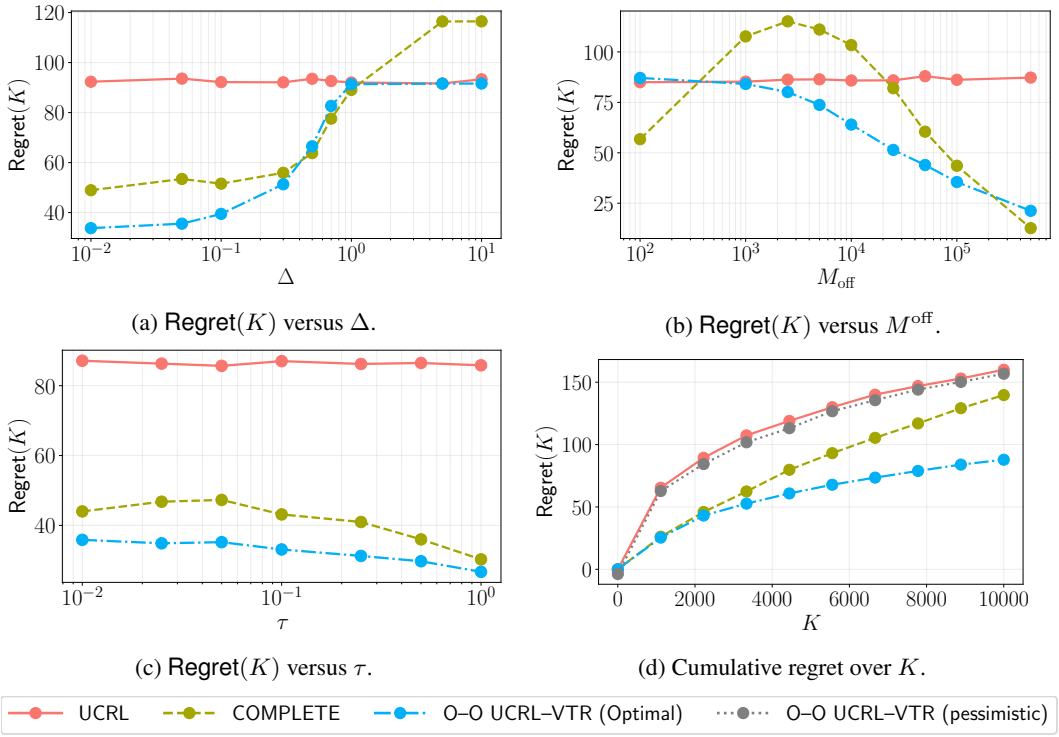
Proof. Since $A \succ 0$ is symmetric, it admits an eigendecomposition $A = U\Lambda U^\top$, where U is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $0 < \lambda_i \leq \lambda_{\max}(A)$ for all $i \in [d]$. By cyclicity of the trace, $\mathrm{tr}(AB) = \mathrm{tr}(U\Lambda U^\top B) = \mathrm{tr}(\Lambda U^\top B U)$. Let $\widetilde{B} \coloneqq U^\top B U$. Since $B \succeq 0$ and U is orthogonal, we have $\widetilde{B} \succeq 0$, which implies $\widetilde{B}_{ii} \geq 0$ for all i. Therefore,

$$\mathrm{tr}(AB) = \sum_{i=1}^{d} \lambda_i \widetilde{B}_{ii} \leq \lambda_{\max}(A) \sum_{i=1}^{d} \widetilde{B}_{ii} = \lambda_{\max}(A)\,\mathrm{tr}(\widetilde{B}) = \lambda_{\max}(A)\,\mathrm{tr}(B),$$

where the last equality again follows from cyclicity of the trace. □

Lemma E.8. *Let $A \in \mathbb{R}^{d\times d}$ be a random positive semidefinite matrix. Then $\mathbb{E}[\lambda_{\min}(A)] \leq \lambda_{\min}(\mathbb{E}[A])$.*

Proof. Recall that for any symmetric matrix A, $\lambda_{\min}(A) = \min_{\|x\|_2=1} x^\top A x$. Hence, for any deterministic unit vector $x \in \mathbb{R}^d$, $\lambda_{\min}(A) \leq x^\top A x$. Taking expectation on both sides yields $\mathbb{E}[\lambda_{\min}(A)] \leq x^\top \mathbb{E}[A] x$. Since the above inequality holds for every unit vector x, taking the minimum over all $\|x\|_2 = 1$ gives $\mathbb{E}[\lambda_{\min}(A)] \leq \min_{\|x\|_2=1} x^\top \mathbb{E}[A] x = \lambda_{\min}(\mathbb{E}[A])$. □



(a) Regret(K) versus Δ.

(b) Regret(K) versus M^{off}.

(c) Regret(K) versus τ.

(d) Cumulative regret over K.

Figure 3: Comparison of UCRL, COMPLETE, O-O UCRL-VTR (Optimal), and O-O UCRL-VTR (Pessimistic) under different environment shifts Δ, coverage τ, K, and M^{off}. This figure provides the complete version of Figure 2.

F Omitted details in Section 5

F.1 MDP setting

We consider a synthetic episodic MDP with $|\mathcal{S}| = 5$, $|\mathcal{A}| = 10$, $H = 3$, and $B = 2$. We follow the setting $\lambda = H^2 d$ as in the main text. Under the standard embedding of tabular MDPs into linear mixture MDPs (Remark 2.2), the feature dimension is $d = |\mathcal{S}|^2|\mathcal{A}| = 250$. For each (s, a), the reward $r(s, a)$ is sampled independently from the uniform distribution $\text{Unif}(0, 1)$. For simplicity, we redefine Δ through the tabular transition discrepancy $\|P(\cdot \mid s, a) - P^{\text{off}}(\cdot \mid s, a)\|_1 \le \Delta$. As discussed in Section B.2.1, this definition is consistent with Assumption 2.3 up to norm equivalence. Each experiment is repeated over 50 independent runs.

Online and offline transition kernel. In the online transition kernel P, each $P(\cdot \mid s, a)$ is sampled from a Dirichlet distribution with parameter $(1, 1, \ldots, 1)$ over the probability simplex, yielding uniform density over transitions. The offline transition kernel P^{off} is constructed to maximize the discrepancy between the offline and online environments subject to the budget Δ. Given a fixed online transition kernel P and reward function r, we first compute, for each state $s \in \mathcal{S}$, the value $V(s)$ induced by the worst policy under the online transition kernel P. For each state–action pair (s, a), we then solve the following constrained optimization problem:

$$P^{\text{off}}(\cdot \mid s, a) = \arg \min_{P^{\text{off}}(\cdot|s,a)} \mathbb{E}_{s' \sim P^{\text{off}}(\cdot|s,a)}[V(s')] \quad \text{s.t.} \quad \|P^{\text{off}}(s, a) - P(s, a)\|_1 \le \Delta. \tag{66}$$

Intuitively, states with smaller values of $V(s)$ are regarded as worse under the online kernel P. Under constraint Δ, the construction Eq. (66) shifts probability mass of $P^{\text{off}}(\cdot \mid s, a)$ toward such states for every (s, a). Note that, although this procedure yields a substantially shifted offline environment, it does not necessarily correspond to the most adversarial design of P^{off}.

F.2 Complete simulation results

Robustness to environment shift Δ. We first examine how performance varies with the environment shift Δ. Fig. 3a plots the cumulative regret as Δ increases, with $\tau = \Theta(1)$, $K = 4 \times 10^3$, and $M^{\text{off}} = 10^5$ fixed. As expected, UCRL remains insensitive to Δ since it does not use any historical data. When Δ is small, we see that O-O UCRL-VTR significantly outperforms UCRL, demonstrating how our algorithm is able to effectively leverage offline data. As Δ grows, the regret of O-O UCRL-VTR increases smoothly and approaches that of UCRL, demonstrating the safeguarding property in Theorem 4.2. When offline data is uninformative, the algorithm gracefully falls back to purely online learning. In contrast, methods that directly merge offline and online data without explicit robustness mechanisms, such as COMPLETE, exhibit substantial performance degradation when Δ becomes large. This phenomenon underscores the robustness of O-O UCRL-VTR.

Impact of offline sample size M^{off}. We next vary the offline sample size M^{off} while fixing $\Delta = 0.05$, $\tau = \Theta(1)$, and $K = 4 \times 10^3$. Fig. 3b shows that once $M^{\text{off}} \gtrsim 10^3$, O-O UCRL-VTR begins to outperform UCRL, indicating that the offline data have become informative relative to the online horizon K. As M^{off} increases further, both offline–online approaches improve. This behavior aligns with Corollary 4.4, which predicts that the offline data become informative once their effective information content is comparable to that of online samples.

An interesting observation is that for small M^{off}, the regret of COMPLETE first increases and then decreases as M^{off} grows, unlike O-O UCRL-VTR. Since COMPLETE ignores the environment shift, when M^{off} is small, increasing M^{off} makes the estimator rely more heavily on the offline phase and become increasingly confident in a potentially misspecified model, thereby amplifying bias and increasing regret at K. As M^{off} becomes large, COMPLETE effectively commits to the offline model and performs little to no exploration in the online phase. Given that the environment shift $\Delta = 0.05$ is small, this lack of exploration does not incur substantial loss for $K = 4 \times 10^3$, leading to a decrease in regret.

Role of coverage τ. We next investigate the effect of the coverage parameter τ, fixing $\Delta = 0.05$, $K = 4 \times 10^3$, and $M^{\text{off}} = 2 \times 10^5$. As shown in Fig. 3c, the advantage of O-O UCRL-VTR over UCRL increases as τ grows. This observation is consistent with the analysis in Section 4. When τ is small, O-O UCRL-VTR does not coincide with UCRL. This is also consistent with the theory in Section 4, since the regret bound depends on the worst-direction coverage, whereas in finite samples, certain well-covered directions may still provide empirical gains.

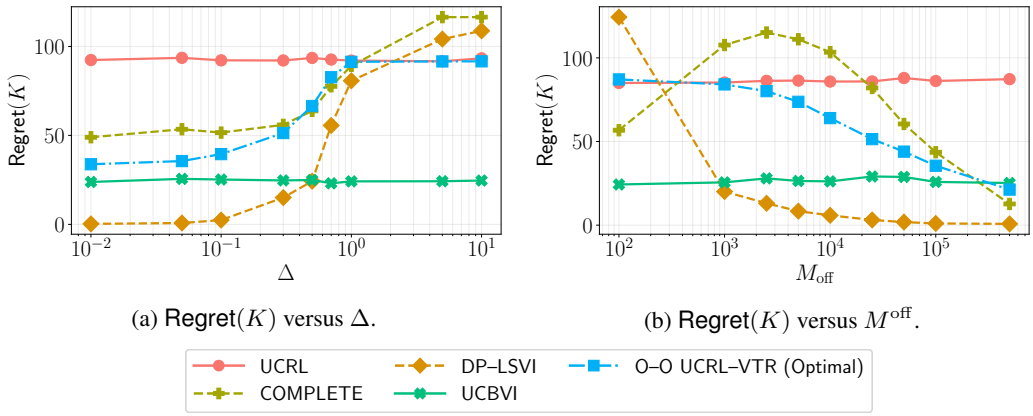
Design of the offline value function estimates $\{\widetilde{V}_{m,h}\}$. Fig. 3d compares two constructions of $\{\widetilde{V}_{m,h}\}$ under $\Delta = 0.05$, $\tau = \Theta(1)$, and $M^{\text{off}} = 10^5$. Both O-O UCRL-VTR (Optimal) and O-O UCRL-VTR (Pessimistic) outperform the purely online baseline UCRL. However, O-O UCRL-VTR (Optimal) consistently attains substantially lower regret than O-O UCRL-VTR (Pessimistic). This suggests that directly improving the statistical geometry of the offline Gram matrix G_{off} can be more effective than adopting a pessimism-driven construction in this setting. The result empirically validates the design principle proposed in Section 4.3.

Interaction between environment shift and coverage. The interaction between environment shift and coverage is illustrated in Figure 2c; we refer to Section 5 for analysis and provide only the experimental details here. Specifically, we set $K = 4 \times 10^3$ and $M^{\text{off}} = 10^5$. The linear scaling (cross marker) in Figure 2c corresponds to $\Delta(\tau) = 0.5\tau$, the quadratic scaling (plus marker) to $\Delta(\tau) = \frac{1}{200}(100\tau)^2$, and the cubic scaling (square marker) to $\Delta(\tau) = \frac{1}{200}(100\tau)^3$. All curves are calibrated to match the same value of Δ at $\tau = 0.01$.

F.3 Behavior policy, implementation of O-O UCRL-VTR, and baseline algorithms

Computing infrastructure. The experiments were conducted on a server equipped with an Intel i7-14700K 20-core processor and 32GB of RAM. No GPUs were required. Each simulation, which evaluates all algorithms presented in the figures, took approximately 150 hours to complete.

Behavior policy and control of τ. In all experiments, the offline behavior policy is the random policy. In Fig. 2a, 2b, 3a, 3b, and 3d, where $\tau = \Theta(1)$, we adopt the designs of $\{\widetilde{V}_{m,h}\}$ described in Section 5 for O-O UCRL-VTR (Optimal) and O-O UCRL-VTR (Pessimistic). In Fig. 2c and Fig. 3c, where we vary τ, the behavior policy is fixed and τ is controlled through the design



(a) Regret(K) versus Δ.

(b) Regret(K) versus M^{off}.

Figure 4: Supplementary results to Figure 2, including DP-LSVI and UCBVI. The setting in Figure 4a matches that of Figure 3a, and the setting in Figure 4b matches that of Figure 3b.

of $\{\widetilde{V}_{m,h}\}$. Specifically, let $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5\}$. For any target value of τ, whenever the state at (m, h) is s_1 or s_2, we set $\mathbb{P}(\widetilde{V}_{m,h} = 0) = 1 - \tau$, and with probability τ, we apply the designs of $\{\widetilde{V}_{m,h}\}$ described in Section 5.

Computational simplification of O-O UCRL-VTR. In line 15 of Algorithm 1, the algorithm requires solving

$$\widehat{Q}_{k,h}(s,a) = r(s,a) + \max_{\theta \in \mathcal{D}_k} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}.$$

Recall that $\mathcal{B}_k$ and $\mathcal{C}_k$ (which determine $\mathcal{D}_k$) are ellipsoids. Since the objective is linear in θ, this problem is a quadratically constrained quadratic program (QCQP) [13] and is tractable.

For computational efficiency in simulations, when $\mathcal{B}_k \cap \mathcal{C}_k$ is nonempty, we adopt the following relaxation:

$$\widehat{Q}_{k,h}(s,a) = r(s,a) + \min\Big\{\max_{\theta \in \mathcal{B}_k} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}, \max_{\theta \in \mathcal{C}_k} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}\Big\}.$$

Following Jia et al. [29], this admits the closed-form expression

$$\hat{Q}_{k,h}(s,a) = r(s,a) + \min\Big\{(x^{\text{on}}_{k,h})^\top \hat{\theta}^{\text{on}}_k + \beta_k \|x^{\text{on}}_{k,h}\|_{(M^{\text{on}}_k)^{-1}}, (x^{\text{on}}_{k,h})^\top \hat{\theta}^{\text{all}}_k + \gamma_k \|x^{\text{on}}_{k,h}\|_{(M^{\text{all}}_k)^{-1}}\Big\}. \tag{67}$$

In our implementation of O-O UCRL-VTR, we use Eq. (67) for computational simplicity. The case where $\mathcal{B}_k \cap \mathcal{C}_k$ is empty is straightforward and therefore omitted.

Baseline algorithms. In addition to the baselines introduced in the main text, we also include comparisons between O-O UCRL-VTR and two tabular MDP algorithms, DP-LSVI and UCBVI.

- DP-LSVI (Algorithm 2 in Chen et al. [20]): The offline–online tabular MDP algorithm proposed by Chen et al. [20].
- UCBVI [10]: A standard purely online algorithm for tabular MDPs.

The comparison between O-O UCRL-VTR and the two tabular MDP algorithms is presented in Figure 4.

The implementations of UCRL, DP-LSVI, UCBVI, O-O UCRL-VTR (Optimal), and O-O UCRL-VTR (Pessimistic) are straightforward. For implementation details of UCRL, DP-LSVI, UCBVI, and the offline phase of O-O UCRL-VTR (Pessimistic), we refer the reader to Jia et al. [29], Chen et al. [20], Azar et al. [10], and Nguyen-Tang et al. [42], respectively. Following the spirit of the COMPLETE algorithm proposed in Chen et al. [20], we present the full description of our COMPLETE baseline in Algorithm 2. Algorithm 2 closely mirrors Algorithm 1, with only two

Algorithm 2 COMPLETE

Require: Online MDP, dimension d, number of online episodes K, offline dataset $\mathcal{D}_{\text{off}}$, auxiliary value functions $\{\widetilde{V}_{m,h}\}$, confidence level δ, radius $\{\beta_k\}_{k\le K}$, ridge regression coefficient λ.

Initialize online-only design matrix and offline–online design matrix: $M_1^{\text{on}} \leftarrow \lambda I, w_1^{\text{on}} \leftarrow 0 \in \mathbb{R}^d, M_1^{\text{all}} \leftarrow \lambda I, w_1^{\text{all}} \leftarrow 0 \in \mathbb{R}^d$.
for $m = 1 : M_{\text{off}}$ **do**
  for $h = 1 : H$ **do**
    Form the regression pair

$$x_{m,h}^{\text{off}} \leftarrow \phi(\cdot \mid s_{m,h}^{\text{off}}, a_{m,h}^{\text{off}})\widetilde{V}_{m,h+1} \in \mathbb{R}^d, y_{m,h}^{\text{off}} \leftarrow \widetilde{V}_{m,h+1}\big(s_{m,h+1}^{\text{off}}\big).$$

    Update statistics $M_1^{\text{all}} \leftarrow M_1^{\text{all}} + x_{m,h}^{\text{off}}(x_{m,h}^{\text{off}})^\top, w_1^{\text{all}} \leftarrow w_1^{\text{all}} + x_{m,h}^{\text{off}} y_{m,h}^{\text{off}}$.
  end for
end for
for $k = 1 : K$ **do**
  Compute estimators $\widehat{\theta}_k^{\text{all}} \leftarrow (M_k^{\text{all}})^{-1} w_k^{\text{all}}$.
  Define confidence set $\mathcal{B}_k' \leftarrow \Big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_k^{\text{all}}\|_{M_k^{\text{all}}} \le \beta_k\Big\}$.
  Set $\widehat{Q}_{k,H+1}(\cdot,\cdot) \leftarrow 0$, and for $h = H, \dots, 1$ define

$$\widehat{V}_{k,h}(s) \leftarrow \max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s,a), \widehat{Q}_{k,h}(s,a) \leftarrow r(s,a) + \max_{\theta\in\mathcal{B}_k'} \sum_{j=1}^{d} \theta_j \phi_j(\cdot \mid s,a)^\top \widehat{V}_{k,h+1}.$$

  Set $s_{k,1}^{\text{on}} = s_1$.
  for $h = 1 : H$ **do**
    Choose $a_{k,h}^{\text{on}} \in \arg\max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s_{k,h}^{\text{on}}, a)$ and observe $s_{k,h+1}^{\text{on}} \sim P(\cdot \mid s_{k,h}^{\text{on}}, a_{k,h}^{\text{on}})$.
    Form the online regression pair $x_{k,h}^{\text{on}} \leftarrow \phi(\cdot \mid s_{k,h}^{\text{on}}, a_{k,h}^{\text{on}})\widehat{V}_{k,h+1}, y_{k,h}^{\text{on}} \leftarrow \widehat{V}_{k,h+1}\big(s_{k,h+1}^{\text{on}}\big)$.
    Update offline–online statistics $M_k^{\text{all}} \leftarrow M_k^{\text{all}} + x_{k,h}^{\text{on}}(x_{k,h}^{\text{on}})^\top, w_k^{\text{all}} \leftarrow w_k^{\text{all}} + x_{k,h}^{\text{on}} y_{k,h}^{\text{on}}$.
  end for
  $M_{k+1}^{\text{all}} \leftarrow M_k^{\text{all}}, w_{k+1}^{\text{all}} \leftarrow w_k^{\text{all}}$.
end for

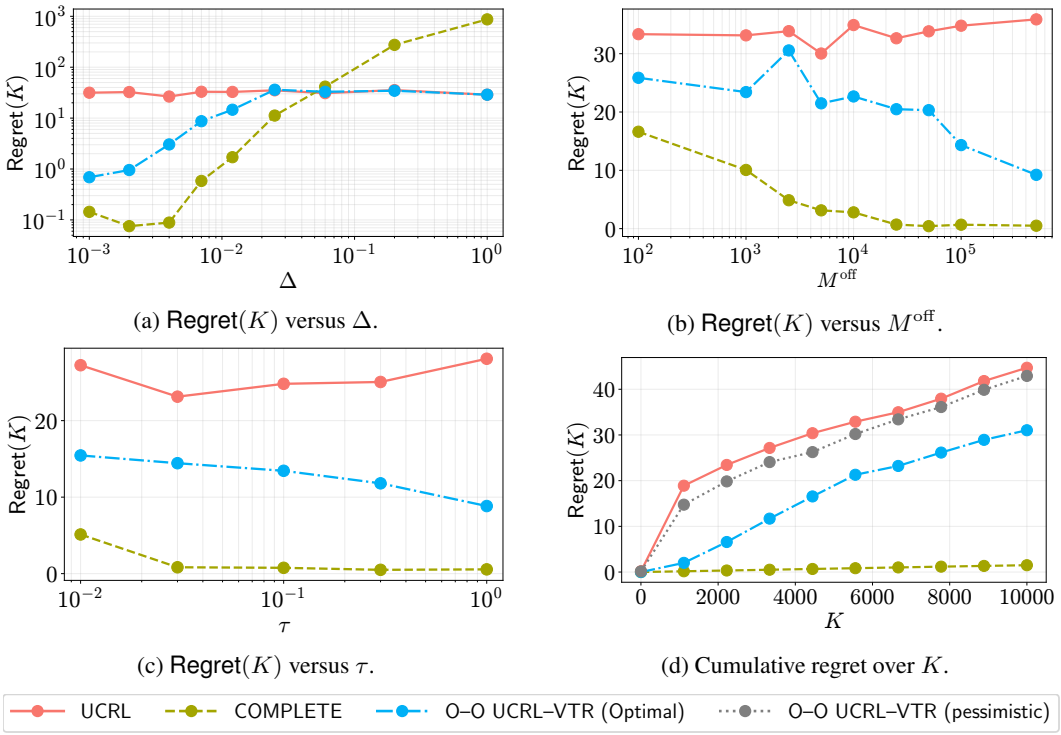
differences: (i) it does not employ the "min-of-two" safeguard (since assuming no environment shift), and (ii) the confidence set is defined as $\mathcal{B}_k' = \Big\{\theta \in \mathbb{R}^d : \|\theta - \theta_k^{\text{all}}\|_{M_k^{\text{all}}} \le \beta_k\Big\}$, also due to the fact that we assume no environment shift. Recall that the confidence radius β_k is defined in Theorem 4.2 as $\beta_k = H\sqrt{2\log\Big(\frac{2\det(M_k^{\text{on}})^{1/2}K}{\det(\lambda I)^{1/2}}\Big)} + \sqrt{\lambda}B$. Here we adopt the same deterministic design of $\{\widetilde{V}_{m,h}\}$ as in Section 4.3, and choose $\lambda = H^2 d$. Conceptually, COMPLETE leverages offline data under the assumption of no environment shift, in the same spirit as classical offline–online approaches for bandits (e.g., Shivaswamy and Joachims [45]).

Hyperparameter tuning. For UCRL, O-O UCRL-VTR (Optimal), and the online phase of O-O UCRL-VTR (Pessimistic), we set λ, β_k, and γ_k according to Theorem 4.2. The offline phase of O-O UCRL-VTR (Pessimistic) involves an additional uncertainty parameter, which is chosen following Lemma D.1 of Nguyen-Tang et al. [42]. All hyperparameters of DP-LSVI follow their original specification in Chen et al. [20]; in particular, we set $\xi_t = 1$ (as noted to be valid in Chen et al. [20]). The bonus term of UCBVI follows Algorithm 3 of Azar et al. [10].

Since the convergence behaviors of different algorithms vary, we introduce an additional bonus scaling parameter b_s to ensure a fair comparison [16]. For example, in O-O UCRL-VTR, the estimate in Eq. (67) becomes

$$\widehat{Q}_{k,h}(s,a) = r(s,a) + \min\Big\{(x_{k,h}^{\text{on}})^\top \widehat{\theta}_k^{\text{on}} + b_s\beta_k\|x_{k,h}^{\text{on}}\|_{(M_k^{\text{on}})^{-1}}, (x_{k,h}^{\text{on}})^\top \widehat{\theta}_k^{\text{all}} + b_s\gamma_k\|x_{k,h}^{\text{on}}\|_{(M_k^{\text{all}})^{-1}}\Big\}.$$

The bonus scales for UCRL, O-O UCRL-VTR (Optimal), O-O UCRL-VTR (Pessimistic), DP-LSVI, UCBVI, and COMPLETE are selected via grid search over $[0, 1]$. We use the same b_s for UCRL, O-O UCRL-VTR (Optimal), and O-O UCRL-VTR (Pessimistic).



(a) Regret(K) versus Δ.

(b) Regret(K) versus M^{off}.

(c) Regret(K) versus τ.

(d) Cumulative regret over K.

Figure 5: Simulation results for the linear mixture MDP with continuous state and action spaces described in Section F.4.

Additional simulation results including DP-LSVI and UCBVI. We conclude this section by presenting additional simulation results that include comparisons with DP-LSVI and UCBVI.

As shown in Fig. 4, UCRL and UCBVI remain insensitive to Δ since they do not use any offline data. DP-LSVI and UCBVI outperform UCRL, O-O UCRL-VTR, and COMPLETE across many regimes. This is expected, as DP-LSVI and UCBVI are designed for tabular settings, whereas O-O UCRL-VTR is developed for the more general linear mixture MDP framework.

However, in certain regimes of M^{off} and Δ (e.g., large Δ in Figure 4a and large M^{off} in Figure 4b), O-O UCRL-VTR can outperform UCBVI and DP-LSVI. This suggests that a well-designed integration of offline data can compensate for, and in some regimes even overcome, the dimensional disadvantage inherent in the linear mixture formulation, thereby validating the effectiveness of O-O UCRL-VTR.

F.4 Additional experiments with continuous state and action spaces

We further evaluate our method in a linear mixture MDP with continuous state and action spaces and feature dimension $d = 10$. Specifically, we consider $\mathcal{S} = \mathcal{A} = [0, 1)$. For each $j \in [10]$, we also define

$$\phi_j(s' \mid s, a) \propto \exp\left[-\frac{1}{2}\left(\frac{f\left(s', \left[0.55s + 0.35(a - 0.5) + \frac{j-1}{10}\right] \bmod 1\right)}{0.0275}\right)^2\right],$$

where $f(x, y) = \min\{|x - y|, 1 - |x - y|\}$. Note that each kernel $\phi_j(\cdot \mid s, a)$ is continuous. The reward function is $r(s, a) = \exp[-\frac{1}{2}(g(s, 0.75)/0.18)^2]$, where $g(x, c) = (x - c + 0.5) \bmod 1 - 0.5$. The constants are chosen so that the kernels ϕ_j remain sufficiently separated. As in Section F.1, the online transition parameter θ^* is sampled from $\text{Dirichlet}(1, \ldots, 1)$. The offline transition kernel is constructed using the continuous-space analogue of the adversarial design in Eq. (66). The offline behavior policy remains random, with actions sampled uniformly from $[0, 1)$. Unless otherwise specified, all remaining implementation details and hyperparameters follow the preceding experimental settings.

For this continuous setting, we repeat the experiments corresponding to all four panels of Figure 3, with results shown in Figure 5. Whenever K or M^{off} is held fixed in Figure 3, we use the same fixed value here. We set $\Delta = 0.01$ except in Figure 5a. The conclusions remain unchanged, so we omit further discussion.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and the contributions in Section 1 summarize all the claims made in the paper, including the algorithmic contributions, regret bounds, characterization of offline data informativeness, and the simulation results.

   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: The limitations are explicitly discussed in their corresponding sections. Following Assumption 2.3, the assumption of knowing Δ is introduced, with further details provided in Section C.2. Additionally, the gap between the upper and lower bounds is explicitly addressed in Section 4.4.

   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: All proofs are provided in the supplementary materials.

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: The code used to conduct the numerical experiments is included in the supplementary materials. Details for replication are provided in Section F. All experiments were run with a fixed random seed to ensure reproducibility.

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: The code is included in the supplementary materials, along with the exact commands and conda environments required to reproduce the 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: Simulation details, along with a description of the hyperparameter optimization procedures, are provided in Section F.

   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: [No]

   Justification: Although each experiment is repeated 50 times and the average is reported, error bars are not included. Reducing variance between experiments would require larger values of M^{off}, K, and related parameters, which is computationally expensive. Therefore, consistent with prior offline–online literature (e.g., Cheung and Lyu [21]), we omit them. Moreover, this works mainly focuses on the theory.

   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: Computing infrastructure details are provided in Section F.

   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: Research complies with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

Justification: This work is theoretical and foundational, and is not tied to any specific application. Applying the proposed algorithm in a particular context would require separate consideration of its potential impacts.

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: [N/A]

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The code package requirements are specified in the supplementary materials and referenced in Section F.

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: [Yes]

    Justification: The code is properly documented and included in the supplementary materials. Experimental details are provided in Section F.

    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: [N/A]

    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: [N/A]

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used solely for formatting purposes and therefore do not require reporting.

    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.