# Leveraging Offline Data in Online Reinforcement Learning

**Andrew Wagenmaker** [1] **Aldo Pacchiano** [2]

## Abstract

Two central paradigms have emerged in the reinforcement learning (RL) community: online RL and offline RL. In the online RL setting, the agent has no prior knowledge of the environment, and must interact with it in order to find an $\epsilon$-optimal policy. In the offline RL setting, the learner instead has access to a fixed dataset to learn from, but is unable to otherwise interact with the environment, and must obtain the best policy it can from this offline data. Practical scenarios often motivate an intermediate setting: if we have some set of offline data and may also interact with the environment, how can we best use the offline data to minimize the number of online interactions necessary to learn an $\epsilon$-optimal policy? In this work, we consider this setting, which we call the FineTuneRL setting, for MDPs with linear structure. We characterize the necessary number of online samples needed in this setting given access to some offline dataset, and develop an algorithm, FTPEDEL, which is provably optimal, up to $H$ factors. We show through an explicit example that combining offline data with online interactions can lead to a provable improvement over either purely offline or purely online RL. Finally, our results illustrate the distinction between *verifiable* learning, the typical setting considered in online RL, and *unverifiable* learning, the setting often considered in offline RL, and show that there is a formal separation between these regimes.

## 1. Introduction

Many important learning problems in adaptive environments can be mapped into the reinforcement learning (RL) paradigm. Recent years have seen an impressive set of results deploying RL algorithms in a variety of domains such as healthcare (Yu et al., 2021), robotics (Kober et al., 2013), and games (Silver et al., 2016). Typically, in such RL settings, the goal of the learner is to find a *policy* that maximizes the expected reward that can be obtained in a given environment. Motivated by the practical successes of RL, developing efficient approaches to *policy optimization* has been a question of much interest in the machine learning community in recent years. Broadly speaking, policy optimization algorithms can be divided into two camps: online RL, where the learner has no prior knowledge of the environment and must simply interact with it to learn a good policy, and offline RL, where the agent has access to a set of offline data from interactions with the environment, but is otherwise unable to interact with it.

Online methods (Brafman & Tennenholtz, 2002; Azar et al., 2017; Auer et al., 2008) rely on the continuous deployment of policies to collect data. These policies are computed at every step by the algorithm and make use of the information that has been collected so far. Unfortunately, online methods can be very data inefficient. In the absence of a sufficiently exploratory baseline policy, they may require an extremely large number of samples to gather sufficient data to learn a near-optimal policy. Offline RL methods (Levine et al., 2020) mitigate some of these shortcomings. For example, if the offline data was originally collected by running a hand-crafted expert policy in the environment, or by running a known safe exploration strategy in a production system, it could be sufficient to learn a near-optimal policy without requiring any additional interactions with the environment. Unfortunately, offline methods are very sensitive to the coverage of the logged data. Namely, the quality of the candidate policy generated by an offline algorithm will strongly depend on how well the available data covers the true optimal policy's support (Zhan et al., 2022). Moreover, since no further interactions with the environment are allowed, offline RL algorithms may not have any way of knowing whether or not their candidate policy is near-optimal—they cannot *verify* the optimality of the policy.

In this work, we aim to bridge the gap between online and offline reinforcement learning and consider an intermediate setting, which we call FineTuneRL, where the algorithm has access to an offline dataset but can also augment this for

[1]University of Washington, Seattle. [2]Work done while at Microsoft Research, New York. Current Affiliation: Broad Institute of MIT and Harvard and Boston University. Correspondence to: Andrew Wagenmaker <ajwagen@cs.washington.edu>, Aldo Pacchiano <aldopacchiano@aldopacchiano.ai>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

further fine-tuning via online interactions with the environment. Here, the goal is to minimize the number of *online* interactions—we assume the offline data is available "for free". We believe that this is often a more realistic form of training than the rigid set of assumptions imposed by pure online or offline scenarios, and has been the subject of much recent attention in the applied RL literature (Ball et al., 2023; Nakamoto et al., 2023; Zheng et al., 2023). This reflects the fact that in practical problems, we may indeed have access to large amounts of cheaply available offline data, which we wish to use to minimize the number of—much more difficult to acquire—online interactions. In FineTuneRL, then, the offline data can be used to bootstrap the online exploration procedure, reducing the online complexity, and the online interaction rounds can be used to optimally refine the candidate policy that would have resulted from only using the offline data.

The FineTuneRL paradigm setting takes some inspiration from the emerging need to devise optimal ways to fine-tune large models. Just as in the case in large language models (Brown et al., 2020) or in image generation tasks (Ramesh et al., 2021), accessibility to large amounts of offline data may allow for the creation of pre-trained models that must be adapted online to solve specific tasks. We hope that by introducing this RL paradigm and by laying the groundwork for analyzing the complexity of FineTuneRL, more work can be dedicated to this important yet relatively unexplored feedback model in RL.

### 1.1. Summary of Contributions

In addition to introducing the FineTuneRL setting, we make the following contributions:

1. We introduce a new notion of concentrability coefficient in the setting of linear MDPs, which we call the *Offline-to-Online Concentrability Coefficient* (Definition 4.1), defined as, for each step $h$:

$$C^h_{\mathsf{o2o}}(\mathfrak{D}_{\mathrm{off}}, \epsilon, T) := \inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi_{\mathsf{lsm}}} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\boldsymbol{\Lambda}^h_{\pi_{\mathrm{exp}}} + \boldsymbol{\Lambda}^h_{\mathrm{off}})^{-1}}}{(V^\star_0 - V^\pi_0)^2 \vee \epsilon^2}.$$

   Here $\boldsymbol{\phi}_{\pi,h}$ denotes the "average feature vector" of policy $\pi$ at step $h$, $\boldsymbol{\Lambda}^h_{\mathrm{off}}$ are the offline covariates at step $h$ for offline dataset $\mathfrak{D}_{\mathrm{off}}$, $\boldsymbol{\Lambda}^h_{\pi_{\mathrm{exp}}}$ denotes the expected covariates induced by policy $\pi_{\mathrm{exp}}$ at step $h$, and $T$ denotes the number of episodes of online exploration. This quantifies the total coverage of our combined offline and online data, if we run for $T$ online episodes, playing the exploration policy $\pi_{\mathrm{exp}}$ that optimally explores the regions of feature-space left unexplored by the offline data.

2. We show there exists an algorithm, FTPEDEL, which, up to lower-order terms, only collects, for each step $h$,

$$\min_{T_{\mathrm{on}}} T_{\mathrm{on}} \quad \text{s.t.} \quad C^h_{\mathsf{o2o}}(\mathfrak{D}_{\mathrm{off}}, \epsilon, T_{\mathrm{on}}) \le \frac{1}{\beta}$$

   online episodes—the minimal number of online episodes which ensures the offline-to-online concentrability coefficient is sufficiently small—and returns a policy that is $\epsilon$-optimal. Furthermore, we show that this complexity is necessary—no algorithm can collect fewer online samples and return a policy guaranteed to be $\epsilon$-optimal.

3. Finally, we study the question of *verifiability* in RL. We note that many existing approaches in offline RL, especially those relying on pessimism, give guarantees that are *unverifiable*—an algorithm may return a near-optimal policy but it has no way of verifying it is near-optimal. We show that coverage conditions necessary for unverifiable RL are insufficient for verifiable RL, and propose stronger coverage conditions to ensure verifiability.

While our work focuses on understanding the statistical complexity of online RL with access to offline data, it motivates a simple, intuitive, and broadly applicable algorithmic principle: direct online exploration to cover (relevant) regions of the feature space not covered by the offline data. Our algorithm, FTPEDEL, instantiates this principle in the setting of linear MDPs, and we hope inspires further work in leveraging offline data in online RL in more general settings.

## 2. Related Work

Three lines of existing work are particularly relevant to our work: Online RL, Offline RL, and works that lie at the intersection of these regimes.

**Online RL.** Much work has been dedicated to designing sample efficient algorithms for online RL. A significant portion of this work has focused on designing algorithms for tabular MDPs with finitely many states and actions (Brafman & Tennenholtz, 2002; Azar et al., 2017; Auer et al., 2008; Jin et al., 2018; Dann et al., 2017; Kearns & Singh, 2002; Agrawal & Jia, 2017; Simchowitz & Jamieson, 2019; Pacchiano et al., 2021; Wagenmaker et al., 2022c). Moving beyond the tabular setting (Yang & Wang, 2020; Jin et al., 2020) propose sample efficient no-regret algorithms for MDPs with linear features, work that has been subsequently built on by a large number of additional works on RL with function approximation (Zanette et al., 2020a;b; Ayoub et al., 2020; Weisz et al., 2021; Zhou et al., 2020; 2021; Du et al., 2021; Jin et al., 2021a; Foster et al., 2021; Wagenmaker et al., 2022b).

Not all MDP instances are equally difficult. The majority of existing work in RL has focused on obtaining algorithms that are *worst-case* optimal, scaling with the complexity of the hardest instance in a given problem class. Such guarantees, however, fail to take into account that some instances may be significantly "easier" than others. While several classical works consider *instance-dependent* bounds (Auer et al., 2008; Tewari & Bartlett, 2007)—bounds scaling with the difficulty of learning on a given problem instance—the

last several years have witnessed significant progress in obtaining such guarantees, both in the tabular setting (Ok et al., 2018; Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019; Yang et al., 2021; Dann et al., 2021; Xu et al., 2021; Wagenmaker et al., 2022c) as well as the function approximation setting (He et al., 2020; Wagenmaker et al., 2022a; Wagenmaker & Jamieson, 2022; Wagenmaker & Foster, 2023). Our work builds on this line of work, in particular (Wagenmaker & Jamieson, 2022), and we aim to obtain an instance-dependent guarantee in the FineTuneRL setting.

**Offline RL.** Early theoretical works in offline RL focus on the setting where the offline data is assumed to have global coverage. This is the case for algorithms such as FQI (Munos & Szepesvári, 2008; Chen & Jiang, 2019) or DAgger for Agnostic MBRL (Ross & Bagnell, 2012). While these approaches are shown to find near-optimal policies, with the aid of either a least squares or a model-fitting oracle, they require that the logged data covers all states and actions.

Towards relaxing such strong coverage conditions, more recent works have developed algorithms for offline RL where the offline data has only partial coverage. This is addressed by either imposing constraints at the policy level, preventing the policy from visit states and actions where the offline data coverage is poor (Fujimoto et al., 2019; Liu et al., 2020; Kumar et al., 2019; Wu et al., 2019), or by relying on the principle of "pessimism" and acting conservatively when learning from offline data (Kumar et al., 2020; Yu et al., 2020; Kidambi et al., 2020; Jin et al., 2021b; Yin et al., 2021; Rashidinejad et al., 2021). In algorithms relying on pessimism, the error of the resulting candidate policy scales with a quantity characterizing the coverage of the offline dataset under the visitation distribution of the optimal policy. Recent works have also extended the study of offline RL to more general function approximation settings (Jiang & Huang, 2020; Uehara & Sun, 2021; Xie et al., 2021a; Chen & Jiang, 2022; Zhan et al., 2022; Yin et al., 2022), as well as in the presence of corrupted data (Zhang et al., 2022b), or offline imitation learning (Chang et al., 2021).

**Bridging Online and Offline RL.** While there exist empirical works considering the setting of online RL where the learner also has some form of access to logged data (Rajeswaran et al., 2017; Nair et al., 2018; Hester et al., 2018; Ball et al., 2023; Nakamoto et al., 2023; Zheng et al., 2023), to our knowledge, only several works offer formal guarantees in this setting (Ross & Bagnell, 2012; Xie et al., 2021b; Song et al., 2022; Tennenholtz et al., 2021). (Tennenholtz et al., 2021) consider a linear bandit setting where they have access to offline data, but where the offline data features are only partially observed, a different setting than what we consider. Both (Ross & Bagnell, 2012) and (Xie et al., 2021b) consider the setting where the learner has access to a logging policy $\mu$ rather than a fixed set of logged data, and at the start of every episode can choose whether to play $\mu$, or to play any other online policy. In many respects, this setting is much more akin to online RL than offline RL. All data available to the learner is collected in an online fashion, either by rolling out $\mu$ or another policy, and the sample complexity bounds are then obtained in terms of the total number of rollouts—both of $\mu$ or alternate online policies played—and are shown to scale with the coverage of $\mu$. (Xie et al., 2021b) prove that in a minimax sense, in this setting there does not exist an approach which can have a strictly better sample complexity than either using purely online RL algorithms (ignoring $\mu$), or collecting data only by playing $\mu$. In contrast to this setting, in this work we assume the learner is simply given access to some offline dataset which could be generated arbitrarily, rather than being given access to a logging policy, and is then able to interact with the environment in an online fashion by playing any policy it desires, ultimately using the combination of the offline data and online interactions to learn a near-optimal policy. Our goal is to minimize the number of online interactions—the offline data is "free", and does not count towards the total number of samples collected. In contrast to (Xie et al., 2021b), we show that there is a provable gain in combining offline data with online interactions in this setting, over either purely offline or purely online RL (Proposition 1).

Concurrent to this work, (Song et al., 2022) propose a setting similar to ours, which they call "Hybrid RL". They propose the Hybrid Q-learning algorithm (Hy-Q), a simple adaptation of fitted Q-iteration for low bilinear rank MDPs (Du et al., 2021). Letting $d_{\mathrm{bi}}$ denote the MDP's bilinear rank, up to logarithmic factors, Hy-Q can be used to find an $\epsilon$-optimal policy in a total number of samples (online + offline) of order $\mathcal{O}\left(\max\{(C^\star)^2, 1\} \cdot \mathrm{poly}(H, d_{\mathrm{bi}})/\epsilon^2\right)$, where $C^\star$ serves as a measure of how well the offline data covers the optimal policy. In the setting where $C^\star$ is large—the offline data does not cover a near-optimal policy—the guarantee obtained by Hy-Q could be much worse than that obtained in the purely online setting (Du et al., 2021). One might hope to obtain a guarantee never worse than the purely online guarantee and, even in the setting of poor offline data coverage, that some useful information may still be extracted from the offline data—precisely the guarantee we obtain.

Finally, we mention the recent work (Xie et al., 2022). While this work is purely online and does not assume access to offline data, it shows that online guarantees can be obtained in terms of the concentrability coefficient parameter introduced in the offline RL literature, providing a bridge between the analysis techniques of offline and online RL.

## 3. Preliminaries

**Notation.** We let $\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}}^2 = \boldsymbol{v}^\top \boldsymbol{\Lambda} \boldsymbol{v}$. $a \vee b$ denotes $\max\{a, b\}$. $\mathcal{S}^{d-1}$ denotes the unit sphere in $d$ dimensions. $\triangle$ denotes the simplex. $\mathsf{logs}(\cdot)$ denotes some function

which depends at most logarithmically on its arguments: $\mathsf{logs}(x_1, \ldots, x_n) = \sum_{i=1}^n c_i \log(e + x_i)$ for $x_i \geq 0$ and absolute constants $c_i \geq 0$. We let $\mathbb{P}_{\mathcal{M}}[\cdot]$ and $\mathbb{E}_{\mathcal{M}}[\cdot]$ denote the measure and expectation induced by MDP $\mathcal{M}$, and $\mathbb{P}_\pi[\cdot]$ and $\mathbb{E}_\pi[\cdot]$ the measure and expectation induced playing policy $\pi$ on our MDP. $C$ and $c$ denote universal constants.

**Markov Decision Processes.** In this work we study episodic Markov Decision Processes (MDPs). In the episodic setting, an MDP is denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{\nu_h\}_{h=1}^H)$ for $\mathcal{S}$ the set of states, $\mathcal{A}$ the set of actions, $H$ the horizon, $\{P_h\}_{h=1}^H$ the probability transition kernels, and $\{\nu_h\}_{h=1}^H$ the reward distributions, which we assume are supported on $[0, 1]$. Each episode begins at some fixed state $s_1$. The agent then takes some action $a_1 \in \mathcal{A}$, transitions to $s_2 \sim P_1(\cdot|s_1, a_1)$, and receives reward $r_1(s_1, a_1) \sim \nu_1(s_1, a_1)$. This repeats for $H$ steps at which point the episode terminates and the process restarts. We assume $\{P_h\}_{h=1}^H$ and $\{\nu_h\}_{h=1}^H$ are initially unknown.

We denote a policy by $\pi : \mathcal{S} \times [H] \to \triangle_{\mathcal{A}}$, and the $Q$-value function for policy $\pi$ as

$$Q_h^\pi(s, a) = \mathbb{E}_\pi\big[\textstyle\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\big].$$

We define the value function as $V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_h^\pi(s, a)]$. We denote the *value of a policy* by $V_0^\pi := V_1^\pi(s_1)$, the total expected reward policy $\pi$ will acquire, and $V_0^\star := \sup_\pi V_0^\pi$. We let $\pi^\star$ denote an optimal policy—any policy with $V_0^\pi = V_0^\star$.

### 3.1. Linear MDPs

In order to allow for efficient learning in MDPs with large state spaces—where $|\mathcal{S}|$ is extremely large or even infinite—further assumptions must be made on the structure of the MDP. One such formulation is the *linear MDP* setting, which we consider in this work.

**Definition 3.1** (Linear MDPs (Jin et al., 2020))**.** We say that an MDP is a *d-dimensional linear MDP*, if there exists some (known) feature map $\boldsymbol{\phi}(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $H$ (unknown) signed vector-valued measures $\boldsymbol{\mu}_h \in \mathbb{R}^d$ over $\mathcal{S}$, and $H$ (unknown) reward vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that:

$$P_h(\cdot|s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\mu}_h(\cdot)\rangle, \quad \mathbb{E}[\nu_h(s, a)] = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h\rangle.$$

We will assume $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$ for all $s, a$; and for all $h$, $\||\boldsymbol{\mu}_h|(\mathcal{S})\|_2 = \|\int_{s\in\mathcal{S}} |\mathrm{d}\boldsymbol{\mu}_h(s)|\|_2 \leq \sqrt{d}$ and $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$.

Note that our definition of linear MDPs allows the reward to be random. While linear MDPs encompass settings such as tabular MDPs—where $\boldsymbol{\phi}(s, a)$ are then simply taken to be the standard basis vectors—they also encompass more complex settings where generalization across states is possible. Indeed, several recent works have demonstrated that complex real-world environments can be modeled as linear MDPs to allow for sample efficient learning (Ren et al., 2022; Zhang et al., 2022a).

We introduce several additional pieces of a notation in the linear MDP setting. For policy $\pi$, let $\boldsymbol{\phi}_{\pi,h} := \mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)]$ denote the expected feature vector at step $h$, which generalizes the state-action visitation frequencies often found in the tabular RL literature. Denote the expected covariates at step $h$ generated by playing policy $\pi$ as $\boldsymbol{\Lambda}_{\pi,h} := \mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top]$. Let $\lambda_{\min}^\star = \min_h \sup_\pi \lambda_{\min}(\boldsymbol{\Lambda}_{\pi,h})$, the largest achievable minimum eigenvalue. We assume the following.

**Assumption 1** (Full Rank Covariates)**.** $\lambda_{\min}^\star > 0$.

Note that Assumption 1 is similar to other explorability assumptions found in the RL literature (Zanette et al., 2020c; Hao et al., 2021; Agarwal et al., 2021; Wagenmaker & Jamieson, 2022; Yin et al., 2022). For the remainder of this work, we assume Assumption 1 holds for the MDP under consideration.

We will be interested in optimizing over covariance matrices in this work, and to this end, define

$$\boldsymbol{\Omega}_h := \{\mathbb{E}_{\pi\sim\omega}[\boldsymbol{\Lambda}_{\pi,h}] \ : \ \omega \in \boldsymbol{\Omega}_\pi\} \tag{3.1}$$

for $\boldsymbol{\Omega}_\pi$ the set of all valid distributions over Markovian policies (both deterministic and stochastic). $\boldsymbol{\Omega}_h$ then denotes the set of all covariance matrices realizable by distributions over policies at step $h$.

**Policy Cover.** The learning approach we propose is *policy-based*, in that it learns over a set of policies $\Pi$, with the goal of finding the best policy in the class. In particular, we consider the class of *linear softmax policies*.

**Definition 3.2** (Linear Softmax Policy)**.** A policy $\pi$ is a *linear softmax policy* with parameters $\eta$ and $\{\boldsymbol{w}_h\}_{h=1}^H$ if

$$\pi_h(a|s) = \frac{e^{\eta\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_h\rangle}}{\sum_{a'\in\mathcal{A}} e^{\eta\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_h\rangle}}, \quad \forall s, a, h.$$

It can be shown that there exists some choice of $\eta$ and set of parameter vectors $\mathcal{W}$ such that the set of linear softmax policies corresponding to $\eta$ and $\mathcal{W}$ contains an $\epsilon$-optimal policy on any linear MDP. Henceforth we refer to this set of policies as $\Pi_{\mathsf{lsm}}$.

### 3.2. Offline Reinforcement Learning and FineTuneRL

In this work we are interested in the setting where we have access to some set of offline data. Let $z^\tau := (s_{h(\tau)}^\tau, a_{h(\tau)}^\tau, r_{h(\tau)}^\tau, s_{h(\tau)+1}^\tau)$ and denote such a dataset by $\mathfrak{D}_{\text{off}} = \{z^\tau\}_{\tau=1}^{T_{\text{off}}}$, where here $h(\tau)$ denotes the step of the $\tau$th sample. We make the following assumption on this data.

**Assumption 2** (Offline Data)**.** *Let $\mathfrak{D}_{\text{off}}$ be an offline dataset and $\mathcal{M}$ our underlying MDP. Then for each $t \leq T_{\text{off}}$:*

$$\begin{aligned}&\mathbb{P}_{\mathfrak{D}_{\text{off}}}[(r_{h(t)}^t, s_{h(t)+1}^t) \in A \times B \mid \{z^\tau\}_{\tau=1}^{t-1}, s_{h(t)}^t, a_{h(t)}^t]\\ &\quad = \mathbb{P}_{\mathcal{M}}[(r_{h(t)}(s_h, a_h), s_{h(t)+1}) \in A \times B \mid s_{h(t)}^t, a_{h(t)}^t]\end{aligned}$$

*for all $A \subseteq [0, 1]$ and $B \subseteq \mathcal{S}$.*

Assumption 2 is similar to existing assumptions on offline data found in the offline RL literature, for instance the *compliance* condition of (Jin et al., 2021b). Assumption 2 implies that the distribution of the reward and next state in $\mathfrak{D}_{\text{off}}$ matches the distribution induces by our MDP $\mathcal{M}$. However, it allows for correlations between steps $\tau$ (e.g. the data could be collected by an adaptive policy) and, additionally, does not even require that the dataset contain full trajectories. For some dataset $\mathfrak{D}_{\text{off}}$, we define $\mathfrak{D}^h_{\text{off}} := \cup_{\tau=1,h(\tau)=h}^{T_{\text{off}}}\{s^\tau_{h(\tau)}, a^\tau_{h(\tau)}, r^\tau_{h(\tau)}, s^\tau_{h(\tau)+1})\}$ and, for $\boldsymbol{\phi}_\tau := \boldsymbol{\phi}(s^\tau_{h(\tau)}, a^\tau_{h(\tau)})$,

$$\boldsymbol{\Lambda}^h_{\text{off}} = \textstyle\sum_{\tau=1}^{T_{\text{off}}} \mathbb{I}\{h(\tau) = h\} \cdot \boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top$$

the covariates collected at step $h$. Finally, we recall the definition of the concentrability coefficient, $C^\pi$, from the offline RL literature (Zanette et al., 2021):

$$C^\pi(\mathfrak{D}_{\text{off}}) := \textstyle\sum_{h=1}^H \|\boldsymbol{\phi}_{\pi,h}\|_{(\boldsymbol{\Lambda}^h_{\text{off}})^{-1}}. \tag{3.2}$$

We let $C^\star(\mathfrak{D}_{\text{off}}) := C^{\pi^\star}(\mathfrak{D}_{\text{off}})$. Existing work has shown that having $C^\star(\mathfrak{D}_{\text{off}})$ bounded is a necessary and sufficient condition to obtain a near-optimal using offline data (Zanette et al., 2021; Jin et al., 2021b).

**Bridging Offline and Online RL.** Given the previous definitions, we are now ready to formally define our learning setting, FineTuneRL.

**Problem Definition (FineTuneRL).** For any linear MDP $\mathcal{M}$ satisfying Assumption 1, given access to some dataset $\mathfrak{D}_{\text{off}}$ which satisfies Assumption 2 on MDP $\mathcal{M}$ as well as the ability to interact online with $\mathcal{M}$, return some policy $\widehat{\pi}$ such that $\mathbb{P}_{\mathcal{M}}[V_0^{\widehat{\pi}} \geq V_0^\star - \epsilon] \geq 1 - \delta$, using as few online interactions as possible.

## 4. Main Results

In the setting of FineTuneRL, we are interested in generalizing the notion of concentrability to account not just for the offline data available, but how this data can be augmented by online exploration to improve coverage. To this end, we introduce the following notion of *offline-to-online concentrability*:

**Definition 4.1** (Offline-to-Online Concentrability Coefficient)**.** Given step $h$, offline dataset $\mathfrak{D}_{\text{off}}$, desired tolerance of learning $\epsilon > 0$, and number of online samples $T$, we define the *offline-to-online concentrability coefficient* as:

$$C^h_{\mathsf{o2o}}(\mathfrak{D}_{\text{off}}, \epsilon, T) := \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h} \max_{\pi \in \Pi_{\mathsf{lsm}}} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^h_{\text{off}})^{-1}}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2}.$$

Intuitively, we can think of $C^h_{\mathsf{o2o}}(\mathfrak{D}_{\text{off}}, \epsilon, T)$ as generalizing offline concentrability to the setting where we can augment the offline data by collecting $T$ samples of online data as well, thereby improving the coverage of the data. In particular, we note that the coverage term, $\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^h_{\text{off}})^{-1}}$, bears a direct resemblance to the offline concentrability coefficient, (3.2), but instead of scaling only with the offline data $\boldsymbol{\Lambda}^h_{\text{off}}$, it also scales with $T$ samples of online data—denoted by $T\boldsymbol{\Lambda}$, the covariates we can collect in $T$ online interactions with the environment. Note that the online-to-offline concentrability scales with the *best-case* online covariates realizable on our MDP—the best possible online data we could collect to cover our policy space given our offline data and $T$ episodes of online exploration.

We weight this coverage term by the optimality of the policy under consideration, scaling it by the minimum of inverse squared gap for policy $\pi$, $V_0^\star - V_0^\pi$, and $\epsilon^{-2}$. This quantifies the fact that for very suboptimal policies, we should not need to collect a significant amount of data, as they can easily be shown to be suboptimal. Note that $C^h_{\mathsf{o2o}}$ corresponds to a somewhat stronger notion of coverage than what has recently been considered in the offline RL literature—rather than simply covering the optimal policy, the offline-to-online concentrability scales with the coverage of *every* policy, weighted by each policy's optimality. As we discuss in more detail in Section 5, this stronger notion of coverage is necessary if we care about *verifiable* learning.

Existing work in the offline RL literature shows that efficient learning is possible if the concentrability coefficient is bounded. We will take a similar approach in this work, and aim to collect enough online data so that the offline-to-online concentrability coefficient is sufficiently small. To this end, we introduce the following notion of minimal online exploration for coverage, quantifying the minimal number of online samples, $T$, that must be collected in order to ensure the offline-to-online concentrability is less than some desired threshold.

**Definition 4.2** (Minimal Online Samples for Coverage)**.** For some desired tolerance $\beta$, we define the *minimal online samples for coverage* as:

$$T^h_{\mathsf{o2o}}(\mathfrak{D}_{\text{off}}, \epsilon; \beta) := \min_T T \quad \text{s.t.} \quad C^h_{\mathsf{o2o}}(\mathfrak{D}_{\text{off}}, \epsilon, T) \leq \frac{1}{\beta}.$$

Note that if our goal is to bound the offline-to-online concentrability coefficient at step $h$, $T^h_{\mathsf{o2o}}(\mathfrak{D}_{\text{off}}, \epsilon; \beta)$ is essentially the minimum number of online interactions that would be required to do so. Intuitively, this corresponds to the minimum number of online interactions needed to cover relevant regions of the feature space not sufficiently covered by the offline data. The following lower bound shows that this quantity is fundamental.

**Theorem 1.** *Fix $T_{\text{off}} \geq 0$ and $\epsilon > 0$. Then there exists some class of MDPs $\mathfrak{M}$ and some offline data $\mathfrak{D}_{\text{off}}$ with $|\mathfrak{D}_{\text{off}}| = T_{\text{off}}$, $\lambda_{\min}(\boldsymbol{\Lambda}^h_{\text{off}}) \geq \Omega(T_{\text{off}}/d), \forall h \in [H]$, such that any algorithm must collect at least*

$$\textstyle\sum_{h=1}^H T^h_{\mathsf{o2o}}(\mathfrak{D}_{\text{off}}, \epsilon; c \cdot dH)$$

*online episodes on some instance $\mathcal{M} \in \mathfrak{M}$ in order to identify an $\epsilon$-optimal policy with constant probability on $\mathcal{M}$, for universal constant $c$.*

Theorem 1 illustrates that there exists a setting where collecting at least $T^h_{\mathsf{o2o}}$ online samples is necessary if our goal is to learn an $\epsilon$-optimal policy—given tolerance $\epsilon > 0$ and offline dataset size $T_{\mathrm{off}} \geq 0$, we can construct a class of instances and offline dataset of size $T_{\mathrm{off}}$ such that any algorithm must collect at least $\sum_{h=1}^{H} T^h_{\mathsf{o2o}}(\mathfrak{D}_{\mathrm{off}}, \epsilon; c \cdot dH)$ online episodes on some instance in the class in order to learn an $\epsilon$-optimal policy. The construction used and proof of Theorem 1 are given in Appendix F.2.

### 4.1. Efficient Learning in FineTuneRL

The following result shows that $T^h_{\mathsf{o2o}}$ is also a sufficient measure of the number of samples needed to ensure learning.

**Theorem 2.** *Fix $\epsilon > 0$, and assume we have access to some offline dataset $\mathfrak{D}_{\mathrm{off}}$ satisfying Assumption 2. Then there exists an algorithm,* FTPEDEL*, which with probability at least $1 - \delta$ returns an $\epsilon$-optimal policy and collects at most*

$$\iota_0 \cdot \sum_{h=1}^{H} T^h_{\mathsf{o2o}}(\mathfrak{D}_{\mathrm{off}}, \epsilon; \beta) + \frac{C_{\mathsf{lot}}}{\epsilon^{8/5}}$$

*online episodes, for lower-order constant $C_{\mathsf{lot}} := \mathrm{poly}\big(d, H, \log\frac{1}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log\frac{1}{\epsilon}, \log T_{\mathrm{off}}\big)$, $\iota_0 = \mathcal{O}(\log\frac{1}{\epsilon})$, and*

$$\beta := dH^6 \cdot \mathsf{logs}\left(d, H, T_{\mathrm{off}}, \frac{1}{\lambda^\star_{\min}}, \frac{1}{\epsilon}, \log\frac{1}{\delta}\right) + cH^4 \cdot \log\frac{1}{\delta}.$$

Theorem 2 shows that, up to $H$ factors and lower-order terms, $T^h_{\mathsf{o2o}}$ is a sufficient measure for the number of online samples that must be collected, and that our algorithm, FTPEDEL, achieves this complexity. Furthermore, Theorem 1 shows that the leading-order term in Theorem 2 is unimprovable in a minimax sense (up to $H$ factors). In addition, in Appendix F.1 we present an *instance-dependent* lower bound for the $\epsilon = 0$ case—rather than scaling with the worst-case complexity over a class of instances, it scales with the complexity necessary on a particular instance—which holds for *any* offline dataset and shows that the $\log 1/\delta$ dependence of Theorem 2 is also necessary. We emphasize that, while the sample complexity of FTPEDEL corresponds to realizing a stronger coverage condition than simply ensuring $C^\star(\mathfrak{D}_{\mathrm{off}})$ is bounded, our lower bounds show that this stronger condition is necessary if our goal is verifiable learning. The proof of Theorem 2 is given in Appendix B, and a description of our algorithm FTPEDEL is given in Section 4.3.

We next provide the following guarantee which shows that, even in the case when $\mathfrak{D}_{\mathrm{off}} = \emptyset$ or when $\mathfrak{D}_{\mathrm{off}}$ has poor coverage, FTPEDEL does essentially no worse than the PEDEL algorithm of (Wagenmaker & Jamieson, 2022) which, up to $H$ factors and lower-order terms, is the tightest known complexity bound for online PAC RL in linear MDPs.

**Corollary 1.** *Regardless of $\mathfrak{D}_{\mathrm{off}}$, with probability at least $1 - \delta$,* FTPEDEL *collects at most*

$$\iota_0\beta \cdot \sum_{h=1}^{H} \inf_{\mathbf{\Lambda} \in \mathbf{\Omega}_h} \max_{\pi \in \Pi_{\mathsf{lsm}}} \frac{\|\phi_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{(V^\star_0 - V^\pi_0)^2 \vee \epsilon^2} + \frac{C_{\mathsf{lot}}}{\epsilon^{8/5}} \tag{4.1}$$

*online episodes. Furthermore, (4.1) is always bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{dH^5(dH^2 + \log 1/\delta)}{\epsilon^2} + \frac{C_{\mathsf{lot}}}{\epsilon^{8/5}}\right). \tag{4.2}$$

The main complexity measure given in Corollary 1 matches almost exactly the complexity measure of PEDEL given in (Wagenmaker & Jamieson, 2022), up to log factors and lower-order terms, implying that FTPEDEL loses virtually nothing from incorporating offline data, as compared to a purely online approach. Furthermore, (4.2) shows that FTPEDEL hits the worst-case optimal online rate, up to $H$ factors and lower-order terms (Wagenmaker et al., 2022b).

**Remark 4.1** (Comparison to (Song et al., 2022))**.** Instantiating the bound given in (Song et al., 2022) in our setting, we see that, in order to find an $\epsilon$-optimal policy, they require collecting at least $\widetilde{\mathcal{O}}(\frac{\max\{(C^\star)^2, 1\} \cdot d^3 H^5 \log 1/\delta}{\epsilon^2})$ online episodes. In comparison, our worst-case bound, Corollary 1, improves on this complexity by a factor of $d$ (though is a factor of $H^2$ worse) implying that, *even when we have access to no offline data*, FTPEDEL obtains a better online sample complexity than the algorithm of (Song et al., 2022), up to a factor of $H^2$, even if the algorithm of (Song et al., 2022) has access to an arbitrarily large amount of offline data. We remark as well that the dependence on $C^\star$ can only hurt the sample complexity given in (Song et al., 2022)—since their complexity scales as $\max\{(C^\star)^2, 1\}$, their approach is unable to benefit from small $C^\star$, while if $C^\star$ is large (the offline data coverage is poor), their complexity could be significantly worse than, for example, our worst-case complexity of Corollary 1. In contrast, our complexity will only improve as the coverage of the offline data improves.

**Remark 4.2** (Scaling of $T^h_{\mathsf{o2o}}$ and $\epsilon$ Dependence)**.** Note that $T^h_{\mathsf{o2o}}(\mathfrak{D}_{\mathrm{off}}, \epsilon; \beta)$ will typically scale linearly in $\beta$ and, except in cases when the offline data coverage is extremely rich, as $\mathcal{O}(\frac{1}{\epsilon^2})$. In general, then, the $\frac{C_{\mathsf{lot}}}{\epsilon^{8/5}}$ term will be lower-order, scaling with a smaller power of $\epsilon$. Intuitively, this term corresponds to the cost of *learning to explore*—learning the set of actions that must be taken to obtain the optimal online covariates which reduce uncertainty. We leave further reducing the $\epsilon$ dependence in this term for future work.

### 4.2. Leveraging Offline Data Yields a Provable Improvement

We next show that there exist settings where complementing the offline data with online exploration yields a provable

improvement over either (a) relying purely on the offline data without online exploration or (b) ignoring the offline data and using only data collected online.

**Proposition 1.** *Fix $\epsilon \leq 1/20$. Then there exist two MDPs $\mathcal{M}^1$ and $\mathcal{M}^2$, and some dataset $\mathfrak{D}_{\text{off}}$ that satisfies Assumption 2 on both $\mathcal{M}^1$ and $\mathcal{M}^2$, such that:*

- *Any algorithm which returns some policy $\widehat{\pi}$ without further online exploration must have:*

$$\max_{i \in \{1,2\}} \mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^i}[V_0^\star(\mathcal{M}^i) - V_0^{\widehat{\pi}}(\mathcal{M}^i)] \geq \Omega(\sqrt{\epsilon}).$$

- *To identify an $\epsilon$-optimal policy on either $\mathcal{M}^1$ or $\mathcal{M}^2$ with constant probability, any algorithm which does not use $\mathfrak{D}_{\text{off}}$ must collect at least $\Omega(\frac{1}{\epsilon^2})$ online samples.*

- FTPEDEL *will return an $\epsilon$-optimal policy with constant probability after collecting at most $\mathcal{O}(\frac{1}{\epsilon^{8/5}})$ online episodes.*

As Proposition 1 shows, we can construct a dataset which does not contain enough information itself to allow us to identify an $\epsilon$-optimal policy, but coupled with a small amount of online exploration, reduces the cost of pure online RL needed to identify an $\epsilon$-optimal policy by a factor of $1/\epsilon^{2/5}$. This illustrates that, for example, using offline data we can beat the standard $\Omega(1/\epsilon^2)$ online lower bounds found throughout the RL literature. Furthermore, it shows that FTPEDEL is able to properly leverage this offline data to reduce the number of online samples it must collect.

*Proof Sketch of Proposition 1.* We briefly sketch the proof of Proposition 1—see Appendix D for a full proof. Proposition 1 is proved by constructing a family of MDPs with three states and three actions: $s_0$ is a fixed starting state, from which the learner transitions to either $s_1$ or $s_2$, and the episode terminates. To identify the optimal action in $s_1$, at least $\Omega(1/\epsilon^2)$ samples are needed from each action; however, the offline dataset contains enough information from $s_1$ that the optimal action in this state can be identified from only the offline data. In contrast, in state $s_2$, the optimal action can be identified by playing each action $\Omega(1)$ times, yet the offline data only contains observations from one of the three actions in $s_2$. Thus, using only the offline data, the learner is unable to find the optimal action in $s_2$, and will therefore be unable to find a policy that $\epsilon$-optimal. However, by using the offline data, it does not need to collect any additional samples from $s_1$ (avoiding the $\Omega(1/\epsilon^2)$ samples it would otherwise need to collect from $s_1$), and must only collect a constant number of samples from $s_2$ to identify an $\epsilon$-optimal policy, reducing the sample complexity of Theorem 2 to only the cost of learning-to-explore. □

### 4.3. Algorithm Description

Our algorithm, FTPEDEL, is based on the PEDEL algorithm of (Wagenmaker & Jamieson, 2022). We provide a brief description of FTPEDEL in what follows and a full definition in Appendix B. We refer the reader to (Wagenmaker & Jamieson, 2022) for an in-depth discussion of PEDEL.

**Algorithm 1** **F**ine-**T**uning **P**olicy Learning via **E**xperiment **De**sign in **L**inear MDPs (FTPEDEL, informal)

**input:** tolerance $\epsilon$, confidence $\delta$, offline data $\mathfrak{D}_{\text{off}}$
$\Pi_0 \leftarrow \Pi_{\mathsf{lsm}}$, $\epsilon_\ell \leftarrow 2^{-\ell}$
**for** $\ell = 1, 2, \ldots, \mathcal{O}(\log \frac{1}{\epsilon})$ **do**
  **for** $h = 1, 2, \ldots, H$ **do**
    Solve online experiment design (Algorithm 4) to collect online covariates $\mathbf{\Lambda}_{h,\ell}$ satisfying (4.3)
    **for** $\pi \in \Pi_\ell$ **do**
      $\widehat{\boldsymbol{\phi}}^\ell_{\pi,h+1} \leftarrow$ estimate of feature visitation for $\pi$
    $\widehat{\boldsymbol{\theta}}^\ell_h \leftarrow$ estimate of reward vector
    $\widehat{V}_0^\pi \leftarrow \sum_{h=1}^H \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi,h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle$
  $\Pi_{\ell+1} \leftarrow \Pi_\ell \backslash \{\pi \in \Pi_\ell \,:\, \widehat{V}_0^\pi < \sup_{\pi' \in \Pi_\ell} \widehat{V}_0^{\pi'} - 2\epsilon_\ell\}$
  **if** $|\Pi_{\ell+1}| = 1$ **then return** $\pi \in \Pi_{\ell+1}$
**return** any $\pi \in \Pi_{\ell+1}$

FTPEDEL is a *policy elimination*-based algorithm. It begins with some initial set of policies $\Pi_0$, and then gradually refines this set—at epoch $\ell$, maintaining a set of policies $\Pi_\ell$ which are guaranteed to be at most $\mathcal{O}(\epsilon_\ell)$-suboptimal. The key property of FTPEDEL is its ability to carefully explore, directing online exploration only to regions that are both relevant to learn about the set of active policies, and that have not yet been sufficiently covered by the offline data. In particular, at step $h$ of epoch $\ell$, it aims to collect online covariates $\mathbf{\Lambda}_{h,\ell}$ that satisfy

$$\max_{\pi \in \Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{(\mathbf{\Lambda}_{h,\ell} + \mathbf{\Lambda}^h_{\text{off}})^{-1}} \leq H^2 \epsilon_\ell^2 / \beta. \tag{4.3}$$

We show that if our collected covariates satisfy (4.3), our estimate of the value of each $\pi \in \Pi_\ell$, $\widehat{V}_0^\pi$, will be within a factor of $\mathcal{O}(\epsilon_\ell)$ of its true value, allowing us to safely eliminate policies more than $\mathcal{O}(\epsilon_\ell)$-suboptimal. Note that to satisfy (4.3) we only need to collect data in directions for which the offline data is not sufficiently rich, and that are relevant to the active policies at epoch $\ell$, $\Pi_\ell$. To efficiently achieve this we rely on an online *experiment design* procedure (Algorithm 4) originally developed in (Wagenmaker & Jamieson, 2022), which is able to collect covariates satisfying (4.3) at a near-optimal rate.

We remark that FTPEDEL is in many respects similar to the PEDEL algorithm of (Wagenmaker & Jamieson, 2022)—both rely on policy elimination strategies and on the same experiment design routine to collect data. The key difference is that FTPEDEL initializes its data buffer with the available

offline data, which allows it to then focus exploration on regions not covered by the offline data. We emphasize the simplicity of this modification—efficiently incorporating offline data does not require entirely new algorithmic approaches; offline data can be naturally used to warm-start online RL algorithms and speed up learning.

## 5. The Cost of Verifiability

As noted, the coverage condition implied by $C^h_{\mathsf{o2o}}$ is stronger than that required by many recent offline RL results. This stronger coverage arises due to the difference between *verifiable* learning and *unverifiable* learning. Informally, an algorithm is *verifiable* if, upon termination, it can *guarantee* that with probability at least $1-\delta$ the returned policy is $\epsilon$-optimal. We contrast this with algorithms that are *unverifiable*—though they may return a near-optimal policy, they cannot guarantee it is near optimal.

In online RL, existing work typically considers the verifiable setting. In contrast, though it is rarely explicitly stated, the offline RL literature has more recently focused on unverifiable complexity results—an algorithm may output an $\epsilon$-optimal policy, but often cannot verify this is the case, unless the data coverage is sufficiently rich. Towards making this distinction formal, we introduce the following definitions of verifiable and unverifiable reinforcement learning, inspired by work in verifiable vs unverifiable learning in the bandit setting (Katz-Samuels & Jamieson, 2020).

**Definition 5.1** (Verifiable RL)**.** We say an algorithm is $(\epsilon,\delta)$-PAC verifiable over some class of MDPs $\mathfrak{M}$ if, for any MDP $\mathcal{M}\in\mathfrak{M}$, it terminates after $\tau_{\mathsf{ver}}$ episodes and returns some policy $\widehat{\pi}$ such that $\mathbb{P}_{\mathcal{M}}[V_0^{\widehat{\pi}}\geq V_0^\star-\epsilon]\geq 1-\delta$. We let $\mathbb{E}_{\mathcal{M}}[\tau_{\mathsf{ver}}]$ denote the expected $(\epsilon,\delta)$-verifiable sample complexity on $\mathcal{M}$.

Note that the above definition of verifiable learning coincides with the standard definition of the PAC RL problem given in online RL. We contrast this with the following definition of unverifiable learning.

**Definition 5.2** (Unverifiable RL)**.** Consider an algorithm which at each step $k$ outputs some $\widehat{\pi}_k$. Then we say that this algorithm has expected unverifiable sample complexity $\mathbb{E}_{\mathcal{M}}[\tau_{\mathsf{uver}}]$ on instance $\mathcal{M}$ if $\tau_{\mathsf{uver}}$ is the minimum stopping time such that $\mathbb{P}_{\mathcal{M}}[\forall k\geq\tau_{\mathsf{uver}}\ :\ V_0^{\widehat{\pi}_k}\geq V_0^\star-\epsilon]\geq 1-\delta$.

Note that implicit in both definitions is that the algorithm *learns*. It is possible that, according to these definitions, an algorithm is neither verifiable or unverifiable with finite sample complexity if, for example, it always outputs policies that are more that $\epsilon$-suboptimal no matter how many samples it collects. As a simple example illustrating verifiability, we consider a multi-armed bandit instance.

**Example 5.1** (Data Coverage in Multi-Armed Bandits)**.** Fix $\epsilon>0$ and consider an $A$-armed multi-armed bandit where arm 1 is optimal, and has a mean of $\mu_1=1$, and every other arm is suboptimal and has a mean $\mu_2=1-3\epsilon$. Let $\mathfrak{D}_{\mathrm{off}}$ be a dataset containing $\mathrm{poly}(1/\epsilon,\log 1/\delta)$ samples from arm 1, but 0 samples from arms 2 to $A$.

In this example we have $C^\star(\mathfrak{D}_{\mathrm{off}})\leq\frac{1}{\mathrm{poly}(\frac{1}{\epsilon},\log\frac{1}{\delta})}$, so by standard offline RL guarantees, $\mathfrak{D}_{\mathrm{off}}$ is sufficient to obtain a policy $\widehat{\pi}$ that is $\epsilon$-optimal. However, as the following result shows, while $\mathfrak{D}_{\mathrm{off}}$ is sufficient in an *unverifiable* fashion, it is not able to identify the best arm in a *verifiable* fashion:

**Proposition 2.** *In the instance of Example 5.1, any $(\epsilon,\delta)$-PAC verifiable algorithm must collect at least $\Omega(\frac{1}{\epsilon^2}\cdot\log\frac{1}{\delta})$ additional samples from every arm 2 to $A$.*

Proposition 2 shows that in order to provide a *guarantee* that $\widehat{\pi}$ is $\epsilon$-optimal, we need to collect a potentially large number of additional samples. This is very intuitive—if we have no samples from arms 2 to $A$ in $\mathfrak{D}_{\mathrm{off}}$, we have no way of knowing whether or not they are better or worse than arm 1. To guarantee that they are in fact worse than arm 1, we need to sample them sufficiently many times to show that they are suboptimal. Note then that in this example, the number of samples needed for verifiable learning is at least a factor of $A$ larger than that needed for unverifiable learning.

Example 5.1 illustrates that simply covering the optimal policy—the condition typically given for pessimistic learning—is not sufficient for verifiable learning, and motivates the coverage condition of $C^h_{\mathsf{o2o}}$. Next, we provide a *sufficient* condition for offline data coverage to guarantee that verifiable learning is possible.

**Theorem 3.** *Assume that our offline data $\mathfrak{D}_{\mathrm{off}}$ satisfies, for each $h\in[H]$:*

$$\max_{\pi\in\Pi_{\mathsf{lsm}}}\frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\boldsymbol{\Lambda}^h_{\mathrm{off}})^{-1}}}{(V_0^\star-V_0^\pi)^2\vee\epsilon^2}\leq\frac{1}{\beta},\quad\forall h\in[H]\tag{5.1}$$

*and $\min_{h\in[H]}\lambda_{\min}(\boldsymbol{\Lambda}^h_{\mathrm{off}})\geq\frac{d^2}{H^2}\cdot\beta$, for*

$$\beta:=dH^6\cdot\mathsf{logs}\left(d,H,T_{\mathrm{off}},\tfrac{1}{\epsilon},\log\tfrac{1}{\delta}\right)+c\log\tfrac{1}{\delta}.$$

*Then there exists an $(\epsilon,\delta)$-PAC verifiable algorithm that returns an $\epsilon$-optimal policy with probability at least $1-\delta$ using only $\mathfrak{D}_{\mathrm{off}}$.*

We prove Theorem 3 and state the $(\epsilon,\delta)$-PAC verifiable algorithm that realizes it in Appendix E. Theorem 3 applies in the purely offline setting, giving a condition when verifiable learning is possible. The condition given in Theorem 3, (5.1), essentially requires that every policy is covered by the offline data. However, it does not require that every policy is covered uniformly well—policies that are very suboptimal need significantly less coverage than near-optimal policies. While this is stronger than only requiring $C^\star(\mathfrak{D}_{\mathrm{off}})$ is bounded, it is a much weaker condition than the uniform concentrability condition—that $\max_\pi C^\pi(\mathfrak{D}_{\mathrm{off}})$ is bounded—required by many works (Antos et al., 2008; Chen & Jiang,

2019; Xie & Jiang, 2021). Our lower bound, Theorem 1, shows that, in a minimax sense, (5.1), is not in general improvable for verifiable learning from purely offline data.

We emphasize here that, in situations where the conditions of Theorem 3 are not met, unverifiable learning may be more appropriate. Theorem 3 simply gives a sufficient coverage condition in settings where the goal is verifiable learning.

### 5.1. Verifying the Performance of a Policy

We turn now to a related question on the cost of verification—given some policy $\widehat{\pi}$, and some data $\mathfrak{D}_{\mathrm{off}}$ which may or may not have been used to obtain $\widehat{\pi}$, what is the cost of verifying whether or not $\widehat{\pi}$ is $\epsilon$-optimal? This setting can model, for example, scenarios in which some prior information about the system may be available, allowing us to obtain a guess at a near-optimal policy, but where we wish to verify that this is indeed the case before deploying the policy in the wild. To facilitate learning in this setting, we make the following assumption on the policy we wish to verify, $\widehat{\pi}$.

**Assumption 3.** *Assume that $\widehat{\pi} \in \Pi$ for some $\Pi$ which can be efficiently covered in the sense that, for any $\gamma > 0$, there exists some set $\Pi^{\gamma}_{\mathsf{cov}} \subseteq \Pi$ with cardinality bounded by $\mathsf{N}_{\mathsf{cov}}(\Pi, \gamma)$ such that for any $\pi \in \Pi$, there exists $\widetilde{\pi} \in \Pi_{\mathsf{cov}}$ satisfying: $\|\boldsymbol{\phi}_{\pi,h}(s) - \boldsymbol{\phi}_{\widetilde{\pi},h}(s)\|_2 \leq \gamma$, for all $s, h$.*

As we show in Appendix E.2, this assumption is met by standard policy classes, for example the class of linear-softmax policies. We obtain the following result.

**Corollary 2.** *Fix $\epsilon > 0$, and assume we have access to some offline dataset $\mathfrak{D}_{\mathrm{off}}$ satisfying Assumption 2, and some policy $\widehat{\pi}$ satisfying Assumption 3. Then there exists some algorithm which with probability at least $1 - \delta$ will verify whether or not $\widehat{\pi}$ is $\epsilon$-optimal and collect at most*

$$\iota_0 \cdot \textstyle\sum_{h=1}^{H} T^h_{\mathsf{ver}}(\mathfrak{D}_{\mathrm{off}}, \epsilon_{\mathsf{ver}}; \beta, \widehat{\pi}) + \frac{C_{\mathsf{lot}}}{\epsilon_{\mathsf{ver}}^{8/5}}$$

*online episodes, for $\epsilon_{\mathsf{ver}} = \epsilon \vee (V_0^{\star} - V_0^{\widehat{\pi}})$, $T^h_{\mathsf{ver}}(\mathfrak{D}_{\mathrm{off}}, \epsilon_{\mathsf{ver}}; \beta, \widehat{\pi})$ the value of the optimization:*

$$\min_T T \quad s.t. \quad \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h} \max_{\pi \in \Pi_{\mathsf{lsm}} \cup \{\widehat{\pi}\}} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^h_{\mathrm{off}})^{-1}}}{(V_0^{\star} - V_0^{\pi})^2 \vee \epsilon^2_{\mathsf{ver}}} \leq \frac{1}{\beta},$$

$$\beta := dH^6 \mathsf{logs}(d, H, T_{\mathrm{off}}, \tfrac{1}{\lambda^{\star}_{\min}}, \tfrac{1}{\epsilon}, \log \tfrac{1}{\delta}) + cH^4 \log \tfrac{\mathsf{N}_{\mathsf{cov}}(\Pi, \gamma)}{\delta},$$

*and $\gamma = \mathrm{poly}(d, H, T_{\mathrm{off}}, \frac{1}{\lambda^{\star}_{\min}}, \frac{1}{\epsilon}, \log \frac{1}{\delta})$.*

Corollary 2 shows that it is possible to verify the quality of a policy with complexity similar to that given in Theorem 2, yet scaling with $\epsilon_{\mathsf{ver}}$ rather than $\epsilon$. This difference in $\epsilon$ dependence arises because, if $\widehat{\pi}$ is very suboptimal, we only need to learn the performance of $\widehat{\pi}$ up to a tolerance that is in the order of its policy gap. Note that $\epsilon_{\mathsf{ver}}$ does not need to be known by the algorithm in advance—our procedure is able to adapt to the value of $\epsilon_{\mathsf{ver}}$.

## 6. Conclusion

This work takes a first step towards understanding the statistical complexity of online RL when the learner is given access to a set of logged data. We introduce the FineTuneRL setting and develop matching upper and lower bounds on the number of online episodes needed to obtain an $\epsilon$-optimal policy given access to an offline dataset, and an algorithm, FTPEDEL, able to achieve it. We believe our work opens several interesting directions for future work.

**Improving FTPEDEL.** FTPEDEL inherits several weakness from the original PEDEL algorithm on which it is based: it requires enumerating a policy class, which in general will be exponentially large in $d$ and $H$, rendering it computationally infeasible for all but the smallest problems, and requires that our MDP has full rank covariates, Assumption 1. Developing a computationally efficient algorithm that achieves the same complexity as given in Theorem 2 but free of $\lambda^{\star}_{\min}$ is an exciting future direction. In addition, Theorem 2 exhibits a lower-order $\mathcal{O}(\frac{1}{\epsilon^{8/5}})$ term, arising from the complexity required to learn to explore. While a more careful analysis can refine this term somewhat, in general it is not clear how to remove this term completely. We believe removing this completely may require new algorithmic techniques, which we leave for future work.

**Verifiability in RL.** Our results point to an important distinction between verifiable and unverifiable RL, and show that verifiable RL can be significantly more difficult than unverifiable RL. We are not the first to observe this separation: (Katz-Samuels & Jamieson, 2020) show that in the multi-armed bandit setting the cost of unverifiable learning is smaller than verifiable learning, and very recently (Tirinzoni et al., 2022) show a similar result in online RL. We believe a more thorough exploration of verifiability in RL is an interesting direction for future work with many open questions.

**Towards a General Theory of Offline-to-Online RL.** While our work answers the question of how to leverage offline data in online RL, it applies only in PAC setting and requires the underlying MDP structure to be linear. Recently, both the offline and online RL literature has devoted significant attention to RL with more general function approximation. Developing algorithms that address the offline-to-online setting with general function approximation while preserving our optimality guarantees is of much interest. In addition, understanding how to best leverage offline data in different RL settings—for example, regret minimization—is an exciting direction for future work.

### Acknowledgements

AW is supported in part by NSF TRIPODS II-DMS 2023166. AP would like to thank the support of the Broad Institute of MIT and Harvard and Boston University.

## References


Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Agarwal, N., Chaudhuri, S., Jain, P., Nagaraj, D., and Netrapalli, P. Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv preprint arXiv:2110.08440*, 2021.

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Ball, P. J., Smith, L., Kostrikov, I., and Levine, S. Efficient online reinforcement learning with offline data. *arXiv preprint arXiv:2302.02948*, 2023.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chang, J., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34:965–979, 2021.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chen, J. and Jiang, N. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. *arXiv preprint arXiv:2203.13935*, 2022.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 1–12, 2021.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Hao, B., Lattimore, T., Szepesvári, C., and Wang, M. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 316–324. PMLR, 2021.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2011.11566*, 2020.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Katz-Samuels, J. and Jamieson, K. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1781–1791. PMLR, 2020.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6292–6299. IEEE, 2018.

Nakamoto, M., Zhai, Y., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A., and Levine, S. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *arXiv preprint arXiv:2303.05479*, 2023.

Ok, J., Proutiere, A., and Tranos, D. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K., and Roberts, S. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 1413–1423. PMLR, 2021.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Ren, T., Zhang, T., Szepesvári, C., and Dai, B. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, pp. 1686–1696. PMLR, 2022.

Ross, S. and Bagnell, J. A. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pp. 3–24. PMLR, 2013.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Tennenholtz, G., Shalit, U., Mannor, S., and Efroni, Y. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pp. 430–439. PMLR, 2021.

Tewari, A. and Bartlett, P. Optimistic linear programming gives logarithmic regret for irreducible mdps. *Advances in Neural Information Processing Systems*, 20, 2007.

Tirinzoni, A., Al-Marjani, A., and Kaufmann, E. Optimistic pac reinforcement learning: the instance-dependent view. *arXiv preprint arXiv:2207.05852*, 2022.

Tsybakov, A. B. Introduction to nonparametric estimation., 2009.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Wagenmaker, A. and Foster, D. J. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. *arXiv preprint arXiv:2304.12466*, 2023.

Wagenmaker, A. and Jamieson, K. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *arXiv preprint arXiv:2207.02575*, 2022.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022a.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pp. 22430–22456. PMLR, 2022b.

Wagenmaker, A. J., Simchowitz, M., and Jamieson, K. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pp. 358–418. PMLR, 2022c.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Xu, H., Ma, T., and Du, S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34: 7677–7688, 2021.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33:11756–11766, 2020c.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022a.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022b.

Zheng, H., Luo, X., Wei, P., Song, X., Li, D., and Jiang, J. Adaptive policy learning for offline-to-online reinforcement learning. *arXiv preprint arXiv:2303.07693*, 2023.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021.

## A. Technical Results

**Proposition 3** (Theorem 1 of (Abbasi-Yadkori et al., 2011)). *Let $(\mathcal{F}_t)_{t=0}^{\infty}$ be a filtration and $(\eta_t)_{t=1}^{\infty}$ be a real-valued stochastic process such that $\eta_t$ is $\mathcal{F}_t$ measurable and $\eta_t|\mathcal{F}_{t-1}$ is mean 0 and $\sigma^2$-subgaussian. Let $\{\boldsymbol{\phi}_t\}_{t=1}^{\infty}$ be an $\mathcal{R}^d$-valued stochastic process such that $\boldsymbol{\phi}_t$ is $\mathcal{F}_{t-1}$-measurable. Let $\boldsymbol{\Lambda}_0 \succ 0$ and define*

$$\boldsymbol{\Lambda}_t = \boldsymbol{\Lambda}_0 + \sum_{s=1}^{t} \boldsymbol{\phi}_s \boldsymbol{\phi}_s^\top.$$

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

$$\left\| \sum_{s=1}^{t} \boldsymbol{\phi}_s \eta_s \right\|_{\boldsymbol{\Lambda}_t^{-1}} \leq \sigma \sqrt{\log \frac{\det \boldsymbol{\Lambda}_t}{\det \boldsymbol{\Lambda}_0} + 2 \log \frac{1}{\delta}}.$$

**Lemma A.1.** *Assume that $B, D \geq 1$. Then if*

$$x \geq 3C \log(3B \max\{C, D\}) + D,$$

*we have that $x \geq C \log(Bx) + D$.*

*Proof.* Take $x = 3C \log(3B \max\{C, D\}) + D$. Then,

$$\begin{aligned}
C \log(Bx) + D &= C \log(3BC \log(3B \max\{C, D\}) + BD) + D \\
&\leq C \log(9B^2 C \max\{C, D\} + BD) + D \\
&\leq C \log(18B \max\{BC^2, BCD, D\}) + D \\
&\leq C \log(18B^2 \max\{C^2, D^2\}) + D \\
&\overset{(a)}{\leq} C \log(3^3 B^3 \max\{C^3, D^3\}) + D \\
&\leq 3C \log(3B \max\{C, D\}) + D \\
&= x.
\end{aligned}$$

Note that $(a)$ holds even if $C < 1$, since we have assumed $D \geq 1$, so in this case $\max\{C, D\} = D$. □

## B. Proof of Upper Bound

**Algorithm 2** **F**ine-**T**uning **P**olicy Learning via **E**xperiment **De**sign in **L**inear MDPs (FTPEDEL)

1: **input:** tolerance $\epsilon$, confidence $\delta$, policy set $\Pi$ (default: $\Pi_{\mathsf{lsm}}$)
2: **for** $i = 1, 2, 3, \ldots$ **do**
3: $\quad \pi \leftarrow$ FTPEDEL-SE $(\epsilon, \delta/2i^2, \Pi, 2^i)$
4: $\quad$ **if** $\pi \neq \emptyset$ **then**
5: $\quad\quad$ **return** $\pi$
6:
7: **function:** FTPEDEL-SE
8: **input:** tolerance $\epsilon$, confidence $\delta$, policy set $\Pi$, upper bound on number of online samples $\bar{T}_{\text{on}}$
9: $\Pi_1 \leftarrow \Pi$, $\widehat{\boldsymbol{\phi}}^1_{\pi,1} \leftarrow \mathbb{E}_{a\sim\pi_1(\cdot|s_1)}[\boldsymbol{\phi}(s_1, a)], \forall \pi \in \Pi$, $T_{\text{on}} \leftarrow 0$, $\lambda \leftarrow 1/d$
10: **for** $\ell = 1, 2, \ldots, \lceil \log \frac{4}{\epsilon} \rceil$ **do**
11: $\quad \epsilon_\ell \leftarrow 2^{-\ell}$, $\beta_\ell \leftarrow H^4 \left( 2\sqrt{d \log \frac{\lambda + (T_{\text{off}} + \bar{T}_{\text{on}})/d}{\lambda}} + 2 \log \frac{2H^2|\Pi|\ell^2}{\delta} + \sqrt{d\lambda} \right)^2$
12: $\quad$ **for** $h = 1, 2, \ldots, H$ **do**
13: $\quad\quad$ Run procedure outlined in Theorem 4 with parameters

$$\epsilon_{\text{exp}} \leftarrow \frac{\epsilon_\ell^2}{\beta_\ell}, \quad \delta \leftarrow \frac{\delta}{2H\ell^2}, \quad \underline{\lambda} \leftarrow \log \frac{4H^2|\Pi|\ell^2}{\delta}, \quad \Phi \leftarrow \Phi_{\ell,h} := \{\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} \ : \ \pi \in \Pi_\ell\},$$

$\quad\quad$ to collect data $\mathfrak{D}_{h,\ell} \leftarrow \{(s_{h,\tau}, a_{h,\tau}, r_{h,\tau}, s_{h+1,\tau})\}_{\tau=1}^{K_{h,\ell}}$ such that:

$$\max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{(\boldsymbol{\Lambda}_{h,\ell}+\boldsymbol{\Lambda}^h_{\text{off}})^{-1}} \leq \epsilon_\ell^2/\beta_\ell \quad \text{and} \quad \lambda_{\min}(\boldsymbol{\Lambda}_{h,\ell} + \boldsymbol{\Lambda}^h_{\text{off}}) \geq \log \frac{4H^2|\Pi|\ell^2}{\delta}$$
$$\text{for} \quad \boldsymbol{\Lambda}_{h,\ell} \leftarrow \textstyle\sum_{\tau=1}^{K_{h,\ell}} \boldsymbol{\phi}_{h,\tau}\boldsymbol{\phi}_{h,\tau}^\top + 1/d \cdot I \quad \text{and} \quad \boldsymbol{\phi}_{h,\tau} = \boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau})$$

$\quad\quad$ terminate procedure early if $T_{\text{on}}+$ total number online episodes collected $> \bar{T}_{\text{on}}$
14: $\quad\quad T_{\text{on}} \leftarrow T_{\text{on}}+$ total number online episodes collected on Line 13
15: $\quad\quad$ **if** $T_{\text{on}} \geq \bar{T}_{\text{on}}$ **then**
16: $\quad\quad\quad$ **return** $\emptyset$
17: $\quad\quad$ **for** $\pi \in \Pi_\ell$ **do** `// Estimate feature-visitations for active policies`
18: $\quad\quad\quad \widehat{\boldsymbol{\phi}}^\ell_{\pi,h+1} \leftarrow \left( \sum_{(s_h,a_h,r_h,s_{h+1})\in\mathfrak{D}^h_{\text{off}}\cup\mathfrak{D}_{h,\ell}} \boldsymbol{\phi}_{\pi,h+1}(s_{h+1})\boldsymbol{\phi}(s_h, a_h)^\top (\boldsymbol{\Lambda}_{h,\ell} + \boldsymbol{\Lambda}^h_{\text{off}})^{-1} \right) \widehat{\boldsymbol{\phi}}^\ell_{\pi,h}$
19: $\quad\quad \widehat{\boldsymbol{\theta}}^\ell_h \leftarrow (\boldsymbol{\Lambda}_{h,\ell} + \boldsymbol{\Lambda}^h_{\text{off}})^{-1} \sum_{(s_h,a_h,r_h,s_{h+1})\in\mathfrak{D}^h_{\text{off}}\cup\mathfrak{D}_{h,\ell}} \boldsymbol{\phi}(s_h, a_h) r_h$
`// Remove provably suboptimal policies from active policy set`
20: $\quad$ Update $\Pi_\ell$:

$$\Pi_{\ell+1} \leftarrow \Pi_\ell \backslash \left\{ \pi \in \Pi_\ell \ : \ \widehat{V}_0^\pi < \sup_{\pi'\in\Pi_\ell} \widehat{V}_0^{\pi'} - 2\epsilon_\ell \right\} \quad \text{for} \quad \widehat{V}_0^\pi := \textstyle\sum_{h=1}^H \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi,h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle$$

21: $\quad$ **if** $|\Pi_{\ell+1}| = 1$ **then return** $\pi \in \Pi_{\ell+1}$
22: **return** any $\pi \in \Pi_{\ell+1}$

The proof of Theorem 2 follows closely the proof of Theorem 1 in (Wagenmaker & Jamieson, 2022). As such, we omit details where appropriate and refer the interested reader to the cited results in (Wagenmaker & Jamieson, 2022).

### B.1. Estimation of Feature Visitations

**Lemma B.1.** *Assume that we have some dataset $\mathfrak{D} = \{(s_{h-1,\tau}, a_{h-1,\tau}, r_{h-1,\tau}, s_{h,\tau})\}_{\tau=1}^K$ satisfying Assumption 2. Denote $\boldsymbol{\phi}_{h-1,\tau} = \boldsymbol{\phi}(s_{h-1,\tau}, a_{h-1,\tau})$ and $\boldsymbol{\Lambda}_{h-1} = \sum_{\tau=1}^K \boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top + \lambda I$. Fix $\pi$ and let*

$$\widehat{\mathcal{T}}_{\pi,h} = \left( \sum_{\tau=1}^K \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\boldsymbol{\phi}_{h-1,\tau}^\top \right) \boldsymbol{\Lambda}_{h-1}^{-1}.$$

*Fix $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{v} \in \mathbb{R}^d$ satisfying $|\boldsymbol{v}^\top \boldsymbol{\phi}_{\pi,h}(s)| \le 1$ for all $s$. Then with probability at least $1-\delta$, we can bound*

$$|\boldsymbol{v}^\top(\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\boldsymbol{u}| \le \left(2\sqrt{\log \frac{\det \boldsymbol{\Lambda}_{h-1}}{\lambda^d} + 2\log\frac{1}{\delta}} + \sqrt{\lambda}\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}\|_2\right) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}.$$

*Proof.* Define the $\sigma$-algebra

$$\mathcal{F}_{h,\tau} = \sigma(\{(s_{h,j}, a_{h,j})\}_{j=1}^{\tau+1} \cup \{(r_{h,j}, s_{h+1,j})\}_{j=1}^{\tau}).$$

Then $(s_{h,\tau}, a_{h,\tau})$ is $\mathcal{F}_{h,\tau-1}$-measurable, and $(r_{h,\tau}, a_{h,\tau})$ is $\mathcal{F}_{h,\tau}$-measurable. Since $\mathfrak{D}$ satisfies Assumption 2, we have that

$$\begin{aligned}\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_h) \mid s_{h-1} = s_{h-1,\tau}, a_{h-1} = a_{h-1,\tau}] &= \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_h) \mid \{s_{h-1,j}, a_{h-1,j}\}_{j=1}^{\tau} \cup \{r_{h-1,j}, s_{h,j}\}_{j=1}^{\tau-1}] \\ &= \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_h) \mid \mathcal{F}_{h-1,\tau-1}]\end{aligned}$$

It follows that, by Definition 3.1,

$$\begin{aligned}\mathcal{T}_{\pi,h} &= \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \\ &= \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \left(\sum_{\tau=1}^K \boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top\right)\boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1} \\ &= \sum_{\tau=1}^K \left(\int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\phi}_{h-1,\tau}\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1} \\ &= \sum_{\tau=1}^K \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|s_{h-1} = s_{h-1,\tau}, a_{h-1} = a_{h-1,\tau}]\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1} \\ &= \sum_{\tau=1}^K \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}]\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1} \\ &= \sum_{\tau=1}^K \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}]\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \mathcal{T}_{\pi,h}\boldsymbol{\Lambda}_{h-1}^{-1}\end{aligned}$$

so

$$|\boldsymbol{v}^\top(\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\boldsymbol{u}| \le \underbrace{\left|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}\right|}_{(a)} + \underbrace{\left|\lambda \boldsymbol{v}^\top \mathcal{T}_{\pi,h}\boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}\right|}_{(b)}.$$

We can bound

$$(a) \le \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} \left\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\right\|_2.$$

By assumption $|\boldsymbol{v}^\top \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})| \le 1$, so $\boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)$ is, conditioned on $\mathcal{F}_{h-1,\tau-1}$, mean 0 and 2-subgaussian. Proposition 3 then gives that, with probability at least $1-\delta$,

$$\left\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\right\|_2 \le 2\sqrt{\log\frac{\det \boldsymbol{\Lambda}_{h-1}}{\lambda^d} + 2\log\frac{1}{\delta}}.$$

We can also bound

$$(b) \le \sqrt{\lambda}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}\|_2.$$

Combining these gives the result. □

**Lemma B.2.** *Assume that we have some dataset $\mathfrak{D} = \{(s_{h,\tau}, a_{h,\tau}, r_{h,\tau}, s_{h+1,\tau})\}_{\tau=1}^K$ satisfying Assumption 2. Denote $\boldsymbol{\phi}_{h,\tau} = \boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau})$ and $\boldsymbol{\Lambda}_h = \sum_{\tau=1}^K \boldsymbol{\phi}_{h,\tau}\boldsymbol{\phi}_{h,\tau}^\top + \lambda I$. Let*

$$\widehat{\boldsymbol{\theta}}_h = \arg\min_{\boldsymbol{\theta}} \sum_{\tau=1}^K (r_{h,\tau} - \langle \boldsymbol{\phi}_{h,\tau}, \boldsymbol{\theta}\rangle)^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

*and fix $\boldsymbol{u} \in \mathbb{R}^d$. Then with probability at least $1-\delta$:*

$$|\langle \boldsymbol{u}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle| \le \left(2\sqrt{\log\frac{\det \boldsymbol{\Lambda}_h}{\lambda^d} + 2\log\frac{1}{\delta}} + \sqrt{d\lambda}\right) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

*Proof.* We construct a filtration as in Lemma B.1. By construction we have

$$\widehat{\boldsymbol{\theta}}_h = \boldsymbol{\Lambda}_h^{-1} \sum_{\tau=1}^K \boldsymbol{\phi}_{h,\tau} r_{h,\tau}.$$

Furthermore, by Definition 3.1:

$$\boldsymbol{\theta}_h = \boldsymbol{\Lambda}_h^{-1}\boldsymbol{\Lambda}_h\boldsymbol{\theta}_h = \boldsymbol{\Lambda}_h^{-1}\sum_{\tau=1}^K \boldsymbol{\phi}_{h,\tau}\mathbb{E}[r_{h,\tau}|\mathcal{F}_{h,\tau-1}] + \lambda\boldsymbol{\Lambda}_h^{-1}\boldsymbol{\theta}_h.$$

Thus,

$$|\langle \boldsymbol{u}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle| \le \underbrace{\left|\sum_{\tau=1}^K \boldsymbol{u}^\top\boldsymbol{\Lambda}_h^{-1}\boldsymbol{\phi}_{h,\tau}(r_{h,\tau} - \mathbb{E}[r_{h,\tau}|\mathcal{F}_{h,\tau-1}])\right|}_{(a)} + \underbrace{\left|\lambda\boldsymbol{u}^\top\boldsymbol{\Lambda}_h^{-1}\boldsymbol{\theta}_h\right|}_{(b)}.$$

We can bound

$$(a) \le \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}} \left\|\sum_{\tau=1}^K \boldsymbol{\Lambda}_h^{-1/2}\boldsymbol{\phi}_{h,\tau}(r_{h,\tau} - \mathbb{E}[r_{h,\tau}|\mathcal{F}_{h,\tau-1}])\right\|_2.$$

Since rewards are bounded in [0,1] almost surely, we have that $r_{h,\tau} - \mathbb{E}[r_{h,\tau}|\mathcal{F}_{h,\tau-1}]$ is conditionally mean 0 and 2-subgaussian. Therefore, applying Theorem 1 of (Abbasi-Yadkori et al., 2011), with probability at least $1-\delta$,

$$\left\|\sum_{\tau=1}^K \boldsymbol{\Lambda}_h^{-1/2}\boldsymbol{\phi}_{h,\tau}(r_{h,\tau} - \mathbb{E}[r_{h,\tau}|\mathcal{F}_{h-1,\tau}])\right\|_2 \le 2\sqrt{\log\frac{\det\boldsymbol{\Lambda}_h}{\lambda^d} + 2\log\frac{1}{\delta}}$$

By Definition 3.1, we can also bound

$$(b) \le \sqrt{\lambda}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}\|\boldsymbol{\theta}_h\|_2 \le \sqrt{d\lambda}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

Combining these proves the result. □

**Lemma B.3.** *Let $\mathcal{E}_{\text{est}}^{\ell,h}$ denote the event on which, for all $\pi \in \Pi_\ell$:*

$$|\langle \boldsymbol{\theta}_{h+1}, \widehat{\boldsymbol{\phi}}_{\pi,h+1}^\ell - \boldsymbol{\phi}_{\pi,h+1}\rangle| \le \sum_{i=1}^h \left(2\sqrt{\log\frac{\det\boldsymbol{\Lambda}_{i,\ell}}{\lambda^d} + 2\log\frac{2H^2|\Pi|\ell^2}{\delta}} + \sqrt{d\lambda}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,i}^\ell\|_{\boldsymbol{\Lambda}_{i,\ell}^{-1}},$$

$$\|\widehat{\boldsymbol{\phi}}_{\pi,h+1}^\ell - \boldsymbol{\phi}_{\pi,h+1}\|_2 \le d\sum_{i=1}^h \left(2\sqrt{\log\frac{\det\boldsymbol{\Lambda}_{i,\ell}}{\lambda^d} + 2\log\frac{2H^2 d|\Pi|\ell^2}{\delta}} + \sqrt{d\lambda}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,i}^\ell\|_{\boldsymbol{\Lambda}_{i,\ell}^{-1}},$$

$$|\langle \widehat{\boldsymbol{\phi}}_{\pi,h}^\ell, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle| \le \left(2\sqrt{\log\frac{\det\boldsymbol{\Lambda}_{h,\ell}}{\lambda^d} + 2\log\frac{2H^2|\Pi|\ell^2}{\delta}} + \sqrt{d\lambda}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,h}^\ell\|_{\boldsymbol{\Lambda}_{h,\ell}^{-1}}.$$

*Then $\mathbb{P}[(\mathcal{E}_{\text{est}}^{\ell,h})^c] \le \frac{\delta}{2H\ell^2}$.*

*Proof.* This follows analogously to Lemma B.5 of (Wagenmaker & Jamieson, 2022) and using Lemma B.1 and Lemma B.2 in place of Lemma B.1 and B.4 of (Wagenmaker & Jamieson, 2022).

We apply Lemma B.1 and Lemma B.2 at step $h$ and round $\ell$ with the dataset

$$\mathfrak{D} \leftarrow \mathfrak{D}_{\text{off}}^{h-1} \cup \mathfrak{D}_{h-1,\ell}.$$

By assumption, the offline data in $\mathfrak{D}$ satisfies Assumption 2. In addition, it is easy to see that given our collection procedure, the online data also satisfies Assumption 2. Thus, $\mathfrak{D}$ satisfies Assumption 2. □

### B.2. Policy Elimination via FTPEDEL-SE

**Lemma B.4.** *Let $\mathcal{E}_{\text{exp}}^{\ell,h}$ denote the event on which:*

- *The exploration procedure on Appendix B terminates after running for at most*

$$\max\left\{\min_{T} C\cdot T \quad s.t. \quad \inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\boldsymbol{\phi}\in\Phi_{\ell,h}} \boldsymbol{\phi}^\top(T(\boldsymbol{\Lambda}+\bar{\lambda}I)+\boldsymbol{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi} \le \frac{\epsilon_\ell^2}{6\beta_\ell}, \quad \frac{\text{poly}(d,H,\frac{1}{\lambda_{\min}^\star},\log 1/\delta,\log|\Pi|)}{\epsilon_\ell^{8/5}}\right\}$$

  *episodes*

- *The covariates returned by Appendix B for any $(h,\ell)$, $\boldsymbol{\Lambda}_{h,\ell}$, satisfy*

$$\max_{\boldsymbol{\phi}\in\Phi_{\ell,h}} \|\boldsymbol{\phi}\|^2_{(\boldsymbol{\Lambda}_{h,\ell}+\boldsymbol{\Lambda}_{\text{off}})^{-1}} \le \frac{\epsilon_\ell^2}{\beta_\ell}, \qquad \lambda_{\min}(\boldsymbol{\Lambda}_{h,\ell}) \ge \log\frac{4H^2|\Pi_\ell|\ell^2}{\delta}.$$

*Then* $\mathbb{P}[(\mathcal{E}_{\text{exp}}^{\ell,h})^c \cap \mathcal{E}_{\text{est}}^{\ell,h-1} \cap (\cap_{i=1}^{h-1}\mathcal{E}_{\text{exp}}^{\ell,i})] \le \frac{\delta}{2H\ell^2}$.

*Proof.* This follows from Theorem 4 and a union bound. □

**Lemma B.5.** *Define $\mathcal{E}_{\text{exp}} = \cap_\ell \cap_h \mathcal{E}_{\text{exp}}^{\ell,h}$ and $\mathcal{E}_{\text{est}} = \cap_\ell \cap_h \mathcal{E}_{\text{est}}^{\ell,h}$. Then $\mathbb{P}[\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}] \ge 1-2\delta$ and on $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, for all $h,\ell$, and $\pi \in \Pi_\ell$,*

$$\begin{aligned}
&|\langle \boldsymbol{\theta}_{h+1}, \widehat{\boldsymbol{\phi}}_{\pi,h+1}^\ell - \boldsymbol{\phi}_{\pi,h+1}\rangle| \le \epsilon_\ell/2H,\\
&\|\widehat{\boldsymbol{\phi}}_{\pi,h+1}^\ell - \boldsymbol{\phi}_{\pi,h+1}\|_2 \le d\epsilon_\ell/2H,\\
&|\langle \widehat{\boldsymbol{\phi}}_{\pi,h}^\ell, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle| \le \epsilon_\ell/2H.
\end{aligned}$$

*Proof.* This follows analogously to the proof of Lemma B.8 of (Wagenmaker & Jamieson, 2022). □

**Lemma B.6.** *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, for all $\ell > \ell_0$, every policy $\pi \in \Pi_\ell$ satisfies $V_0^\star(\Pi) - V_0^\pi \le 4\epsilon_\ell$ and $\widetilde{\pi}^\star \in \Pi_\ell$, for $\widetilde{\pi}^\star = \arg\max_{\pi\in\Pi} V_0^\pi$.*

*Proof.* This follows analogously to the proof of Lemma B.9 of (Wagenmaker & Jamieson, 2022). □

**Lemma B.7.** *With probability at least $1-2\delta$,* FTPEDEL-SE *run with parameters $(\epsilon,\delta,\bar{T}_{\text{on}})$ will terminate after collecting at most*

$$\min\left\{C\cdot\sum_{h=1}^{H}\sum_{\ell=1}^{\iota_0} T_{h,\ell}(\bar{T}_{\text{on}}) + \text{poly}\left(d,H,\log\frac{1}{\delta},\frac{1}{\lambda_{\min}^\star},\log|\Pi|,\log\frac{1}{\epsilon}\right)\cdot\frac{\beta_{\iota_0}(\bar{T}_{\text{on}})}{\epsilon^{8/5}\vee\Delta_{\min}(\Pi)^{8/5}}, \bar{T}_{\text{on}}\right\}$$

*episodes for $\iota_0 := \min\{\lceil\log\frac{4}{\epsilon}\rceil, \log\frac{4}{\Delta_{\min}(\Pi)}\}$ and*

$$T_{h,\ell}(\bar{T}_{\text{on}}) = \min_N N \quad s.t. \quad \inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\pi\in\Pi(4\epsilon_\ell)} \|\boldsymbol{\phi}_{\pi,h}\|^2_{(N\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{\text{off}}^h)^{-1}} \le \frac{\epsilon_\ell^2}{48\beta_\ell(\bar{T}_{\text{on}})}$$

*where here $\Pi(4\epsilon_\ell) = \{\pi\in\Pi \,:\, V_0^\pi \ge \max_{\pi\in\Pi} V_0^\pi - 4\epsilon_\ell\}$.*

*Proof.* Note that FTPEDEL-SE will terminate and output $\emptyset$ if the total number of online episodes it has collected reaches $\bar{T}_{\text{on}}$, so it follows that $\bar{T}_{\text{on}}$ is always an upper bound on the number of online episodes that will be collected. Henceforth we assume that we are in a situation where $\bar{T}_{\text{on}}$ is greater than the number of online episodes collected.

By Lemma B.5 the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$ occurs with probability at least $1 - 2\delta$. Henceforth we assume we are on this event.

By Lemma B.4, the complexity of exploration at round $\ell$, step $h$, is bounded by

$$\max\left\{\min_T C \cdot T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\boldsymbol{\phi}\in\Phi_{\ell,h}} \boldsymbol{\phi}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi} \le \frac{\epsilon_\ell^2}{6\beta_\ell(\bar{T}_{\text{on}})}, \right.$$
$$\left. \frac{\text{poly}(d, H, \frac{1}{\lambda^\star_{\min}}, \log 1/\delta, \log|\Pi|)\cdot \beta_\ell(\bar{T}_{\text{on}})^{3/4}}{\epsilon_\ell^{8/5}}\right\}.$$

On $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, by Lemma B.5, for each $\pi \in \Pi_\ell$, we have $\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2 \le d\epsilon_\ell/2H$. As $\Phi_{\ell,h} = \{\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} : \pi \in \Pi_\ell\}$, it follows that we can upper bound

$$\begin{aligned}
&\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\boldsymbol{\phi}\in\Phi_{\ell,h}} \boldsymbol{\phi}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi} \\
&\le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} 2\boldsymbol{\phi}_{\pi,h}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h} + 2(\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h})^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}(\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}) \\
&\le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} 2\boldsymbol{\phi}_{\pi,h}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h} + \frac{d^2\epsilon_\ell^2}{2H^2\lambda_{\min}(\mathbf{\Lambda})} \\
&\le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} 4\boldsymbol{\phi}_{\pi,h}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h} + \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \frac{d^2\epsilon_\ell^2}{TH^2\lambda_{\min}(\mathbf{\Lambda})} \\
&\le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} 4\boldsymbol{\phi}_{\pi,h}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h} + \frac{d^2\epsilon_\ell^2}{TH^2\lambda^\star_{\min}} \\
&\le \max\left\{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} 8\boldsymbol{\phi}_{\pi,h}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h}, \frac{2d^2\epsilon_\ell^2}{TH^2\lambda^\star_{\min}}\right\}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\min_T C\cdot T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\boldsymbol{\phi}\in\Phi_{\ell,h}} \boldsymbol{\phi}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi} \le \frac{\epsilon_\ell^2}{6\beta_\ell(\bar{T}_{\text{on}})} \\
&\le \min_T C\cdot T \quad \text{s.t.} \quad \max\left\{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} 8\boldsymbol{\phi}_{\pi,h}^\top (T(\mathbf{\Lambda}+\bar{\lambda}I)+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h}, \frac{2d^2\epsilon_\ell^2}{TH^2\lambda^\star_{\min}}\right\} \le \frac{\epsilon_\ell^2}{6\beta_\ell(\bar{T}_{\text{on}})} \\
&\le \left[\min_T C\cdot T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi_\ell} \boldsymbol{\phi}_{\pi,h}^\top (T\mathbf{\Lambda}+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h} \le \frac{\epsilon_\ell^2}{48\beta_\ell(\bar{T}_{\text{on}})}\right] + \frac{12\beta_\ell(\bar{T}_{\text{on}})d^2}{H^2\lambda^\star_{\min}} \\
&\le \underbrace{\left[\min_T C\cdot T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi(4\epsilon_\ell)} \boldsymbol{\phi}_{\pi,h}^\top (N\mathbf{\Lambda}+\mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}_{\pi,h} \le \frac{\epsilon_\ell^2}{48\beta_\ell(\bar{T}_{\text{on}})}\right]}_{=:T_{h,\ell}(\bar{T}_{\text{on}})} + \frac{12\beta_\ell(\bar{T}_{\text{on}})d^2}{H^2\lambda^\star_{\min}}
\end{aligned}$$

where the last inequality follows since, by Lemma B.6, for $\ell > \ell_0$, every policy $\pi \in \Pi_\ell$ will be $4\epsilon_\ell$ optimal, so we therefore have

$$\Pi_\ell \subseteq \{\pi \in \Pi \,:\, V_0^\pi \ge V_0^{\widetilde{\pi}^\star} - 4\epsilon_\ell\} =: \Pi(4\epsilon_\ell).$$

By Lemma B.6, if $4\epsilon_\ell < \Delta_{\min}(\Pi)$, we must have that $\Pi_\ell = \{\widetilde{\pi}^\star\}$, and will therefore terminate on Appendix B since $|\Pi_\ell| = 1$. Thus, we can bound the number of number of epochs by

$$\iota_0 := \min\{\lceil \log \frac{4}{\epsilon}\rceil, \log \frac{4}{\Delta_{\min}(\Pi)}\}.$$

It follows that the total complexity is bounded as

$$T_{\text{se}}(\bar{T}_{\text{on}}) := \sum_{h=1}^{H}\sum_{\ell=1}^{\iota_0} T_{h,\ell}(\bar{T}_{\text{on}}) + \text{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \log\frac{1}{\epsilon}\right)\cdot \frac{\beta_{\iota_0}(\bar{T}_{\text{on}})}{\epsilon^{8/5} \vee \Delta_{\min}(\Pi)^{8/5}}. \tag{B.1}$$

□

### B.3. Completing the Proof: FTPEDEL

**Lemma B.8.** FTPEDEL *will terminate after running for at most*

$$C \cdot \sum_{h=1}^{H} \sum_{\ell=1}^{\iota_0} T_{h,\ell}(\widetilde{T}_{\mathrm{on}}) + \mathrm{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \log\frac{1}{\epsilon}, \log T_{\mathrm{off}}\right) \cdot \frac{1}{\epsilon^{8/5} \vee \Delta_{\min}(\Pi)^{8/5}}$$

*episodes, for*

$$\widetilde{T}_{\mathrm{on}} := \mathrm{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \frac{1}{\epsilon}, \log(T_{\mathrm{off}})\right).$$

*Furthermore, it will output a policy $\widehat{\pi}$ such that*

$$V_0^{\widehat{\pi}} \geq \max_{\pi\in\Pi} V_0^{\pi} - \epsilon.$$

*Proof.* By Lemma B.10 of (Wagenmaker & Jamieson, 2022), we can bound

$$T_{h,\ell}(\bar{T}_{\mathrm{on}}) \leq \frac{48Cd\beta_\ell(\bar{T}_{\mathrm{on}})}{\epsilon_\ell^2}.$$

From Lemma B.7, it follows that the total complexity of running FTPEDEL-SE with parameters $(\epsilon, \delta/2i^2, \bar{T}^i_{\mathrm{on}})$, for $\bar{T}^i_{\mathrm{on}} = 2^i$, is bounded as

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{\ell=1}^{\iota_0}\left(\frac{48Cd\beta_\ell(\bar{T}^i_{\mathrm{on}})}{\epsilon_\ell^2} + \mathrm{poly}\left(d, H, \log\frac{i}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \log\frac{1}{\epsilon}\right) \cdot \frac{\beta_{\iota_0}(\bar{T}^i_{\mathrm{on}})}{\epsilon^{8/5} \vee \Delta_{\min}(\Pi)^{8/5}}\right) \\
&\leq \mathrm{poly}\left(d, H, \log\frac{i}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \frac{1}{\epsilon}\right) \cdot \sqrt{\log(T_{\mathrm{off}} + \bar{T}^i_{\mathrm{on}})} \\
&\leq \mathrm{poly}\left(d, H, \log\frac{\bar{T}^i_{\mathrm{on}}}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \frac{1}{\epsilon}\right) \cdot \left(\log(T_{\mathrm{off}}) + \log(\bar{T}^i_{\mathrm{on}})\right).
\end{aligned}$$

To ensure that $\bar{T}^i_{\mathrm{on}}$ is sufficiently large, we then need only that

$$\mathrm{poly}\left(d, H, \log\frac{\bar{T}^i_{\mathrm{on}}}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \frac{1}{\epsilon}\right) \cdot \left(\log(T_{\mathrm{off}}) + \log(\bar{T}^i_{\mathrm{on}})\right) \leq \bar{T}^i_{\mathrm{on}}.$$

To achieve this it suffices that

$$\bar{T}^i_{\mathrm{on}} \geq \mathrm{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \frac{1}{\epsilon}, \log(T_{\mathrm{off}})\right) =: \widetilde{T}_{\mathrm{on}}.$$

Note that FTPEDEL-SE will terminate and output $\emptyset$ if the total number of online episodes it has collected reaches $\bar{T}_{\mathrm{on}}$, so it follows that $\bar{T}_{\mathrm{on}}$ is an upper bound on the number of online episodes that will be collected. Thus, the total number of episodes is bounded as

$$\min_i \sum_{j=1}^{i} 2^j \quad \text{s.t.} \quad T_{\mathrm{se}}(2^i) \leq 2^i$$

for $T_{\mathrm{se}}$ as defined in (B.1). Note that a feasible solution to this is $i = \log_2(\widetilde{T}_{\mathrm{on}})$, so

$$\begin{aligned}
\min_i \sum_{j=1}^{i} 2^j \quad \text{s.t.} \quad T_{\mathrm{se}}(2^i) \leq 2^i &= \min_{i \leq \log_2(\widetilde{T}_{\mathrm{on}})} \sum_{j=1}^{i} 2^j \quad \text{s.t.} \quad T_{\mathrm{se}}(2^i) \leq 2^i \\
&\leq \min_{i \leq \log_2(\widetilde{T}_{\mathrm{on}})} \sum_{j=1}^{i} 2^j \quad \text{s.t.} \quad T_{\mathrm{se}}(\widetilde{T}_{\mathrm{on}}) \leq 2^i
\end{aligned}$$

$$\leq 2T_{\mathrm{se}}(\widetilde{T}_{\mathrm{on}})$$

where the first inequality uses that $T_{\mathrm{se}}(2^i)$ is increasing in $i$. The result follows.

Correctness follows by Lemma B.6, since upon termination, $\Pi_\ell$ will only contain policies $\pi$ satisfying $V_0^\pi \geq \max_{\pi\in\Pi} V_0^\pi - \epsilon$ (and will contain at least 1 policy since $\widetilde{\pi}^\star \in \Pi_\ell$ for all $\ell$). Furthermore, by Lemma B.6, if $4\epsilon_\ell < \Delta_{\min}(\Pi)$, we must have that $\Pi_\ell = \{\widetilde{\pi}^\star\}$, and will therefore terminate on Appendix B since $|\Pi_\ell| = 1$. □

*Proof of Theorem 2.* By the definition of $\Pi(4\epsilon_\ell)$, for each $\pi \in \Pi(4\epsilon_\ell)$ we have

$$\epsilon_\ell^2 = \frac{1}{16}\left((V_0^\star(\Pi) - V_0^\pi)^2 \vee (4\epsilon_\ell)^2\right).$$

We can therefore bound

$$\begin{aligned} T_{h,\ell}(\widetilde{T}_{\mathrm{on}}) &\leq \min_T T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi(4\epsilon_\ell)} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\mathbf{\Lambda}+\mathbf{\Lambda}^h_{\mathrm{off}})^{-1}}}{(V_0^\star(\Pi) - V_0^\pi)^2 \vee \epsilon_\ell^2} \leq \frac{c}{\beta_\ell(\widetilde{T}_{\mathrm{on}})} \\ &\leq \min_T T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\mathbf{\Lambda}+\mathbf{\Lambda}^h_{\mathrm{off}})^{-1}}}{(V_0^\star(\Pi) - V_0^\pi)^2 \vee \epsilon^2 \vee \Delta_{\min}(\Pi)^2} \leq \frac{c}{\beta_\ell(\widetilde{T}_{\mathrm{on}})}. \end{aligned} \tag{B.2}$$

We next set $\Pi$ to be the set of linear softmax policies defined in Lemma A.14 of (Wagenmaker & Jamieson, 2022), and note that this set is guaranteed to contain a policy which is $\epsilon$-optimal as compared to the best possible policy. Furthermore, we can bound $\log|\Pi| \leq \mathcal{O}(dH^2 \cdot \log\frac{1}{\epsilon})$. Using this and that $\ell \leq \iota_0$, we can then bound

$$(\text{B.2}) \leq \min_T T \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\pi\in\Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(T\mathbf{\Lambda}+\mathbf{\Lambda}^h_{\mathrm{off}})^{-1}}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \leq \frac{c}{\beta}$$

for

$$\beta := dH^6 \cdot \mathsf{logs}(d, H, T_{\mathrm{off}}, \frac{1}{\lambda^\star_{\min}}, \frac{1}{\epsilon}, \log\frac{1}{\delta}) + H^4 \cdot \log\frac{1}{\delta}$$

We can therefore bound the complexity given in Lemma B.8 as

$$C\iota_0 \cdot \sum_{h=1}^H T^h_{\mathsf{o2o}}(\mathfrak{D}_{\mathrm{off}}, \epsilon; \beta) + \mathrm{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\lambda^\star_{\min}}, \log|\Pi|, \log\frac{1}{\epsilon}, \log T_{\mathrm{off}}\right) \cdot \frac{1}{\epsilon^{8/5} \vee \Delta_{\min}(\Pi)^{8/5}}.$$

□

## C. Online Experiment Design

The results in this section largely follow those presented in Appendices C and D of (Wagenmaker & Jamieson, 2022), with several minor modifications. As such, we omit calculations that would be identical to those in (Wagenmaker & Jamieson, 2022) and refer the reader to (Wagenmaker & Jamieson, 2022) for more in-depth proofs.

**Algorithm 3** Online Frank-Wolfe via Regret Minimization (FWREGRET)

1: **input**: function to optimize $f$, number of iterates $T$, episodes per iterate $K$
2: Play any policy for $K$ episodes, denote collected covariates as $\mathbf{\Gamma}_0$, collected data as $\mathfrak{D}_0$
3: $\mathbf{\Lambda}_1 \leftarrow K^{-1}\mathbf{\Gamma}_0$
4: **for** $t = 1, 2, \ldots, T$ **do**
5: Set $\gamma_t \leftarrow \frac{1}{t+1}$
6: Run FORCE (Wagenmaker et al., 2022a) on reward $r_h^t(s,a) = \mathrm{tr}(\Xi_{\mathbf{\Lambda}_t} \cdot \boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top)/M$ for $K$ episodes, denote collected covariates as $\mathbf{\Gamma}_t$, collected data as $\mathfrak{D}_t$
7: $\mathbf{\Lambda}_{t+1} \leftarrow (1-\gamma_t)\mathbf{\Lambda}_t + \gamma_t K^{-1}\mathbf{\Gamma}_t$
8: **return** $\mathbf{\Lambda}_{T+1}, \cup_{t=0}^T \mathfrak{D}_t$

**Lemma C.1.** *Consider running Algorithm 3 with a function $f$ satisfying Definition 5.1 of (Wagenmaker & Jamieson, 2022). Then, we have that, with probability at least $1-\delta$,*

$$f(\mathbf{\Lambda}_{T+1}) - \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} f(\mathbf{\Lambda}) \le \frac{\beta R^2(\log T+3)}{2(T+1)} + \sqrt{\frac{8M^2\log(4T/\delta)}{K}} + \sqrt{\frac{c_1 M^2 d^4 H^4 \log^3(2HKT/\delta)}{K}} + \frac{c_2 M d^4 H^3 \log^{7/2}(2HKT/\delta)}{K}$$

*for $R = \sup_{\pi,\pi'} \|\mathbf{\Lambda}_\pi - \mathbf{\Lambda}_{\pi'}\|$.*

*Proof.* This follows analogously to the proof of Lemma C.3 of (Wagenmaker & Jamieson, 2022), but without requiring that $K$ be large enough to upper bound the $\mathcal{O}(1/\sqrt{K})$ and $\mathcal{O}(1/K)$ terms by $\frac{\beta R^2(\log T+3)}{2(T+1)}$, and instantiating REGMIN with the FORCE algorithm of (Wagenmaker et al., 2022a). □

**Algorithm 4** Collect Optimal Covariates (OPTCOV)

1: **input**: functions to optimize $(f_i)_i$, constraint tolerance $\epsilon$, confidence $\delta$
2: **for** $i = 1, 2, 3, \ldots$ **do**
3: $T_i \leftarrow 2^i$, $K_i \leftarrow 2^i$
4: $\widehat{\mathbf{\Lambda}}_i, \mathfrak{D}_i \leftarrow$ FWREGRET$(f_i, T_i - 1, K_i)$
5: **if** $f_i(\widehat{\mathbf{\Lambda}}_i) \le K_i T_i \epsilon$ **then**
6: **return** $\widehat{\mathbf{\Lambda}}$, $K_i T_i$, $\mathfrak{D}_i$

**Lemma C.2.** *Let*

$$f_i(\mathbf{\Lambda}) = \frac{1}{\eta_i}\log\left(\sum_{\phi\in\Phi} e^{\eta_i \|\phi\|^2_{\mathbf{A}_i(\mathbf{\Lambda})^{-1}}}\right), \quad \mathbf{A}_i(\mathbf{\Lambda}) = \mathbf{\Lambda} + \frac{1}{T_i K_i}\mathbf{\Lambda}_{0,i} + \frac{1}{T_i K_i}\mathbf{\Lambda}_{\text{off}}$$

*for some $\mathbf{\Lambda}_{0,i}$ satisfying $\mathbf{\Lambda}_{0,i} \succeq \mathbf{\Lambda}_0$ for all $i$, and $\eta_i = 2^{2i/5}$. Let $(\beta_i, M_i)$ denote the smoothness and magnitude constants for $f_i$. Let $(\beta, M)$ be some values such that $\beta_i \le \eta_i\beta$, $M_i \le M$ for all $i$.*
*Then, if we run Algorithm 4 on $(f_i)_i$ with constraint tolerance $\epsilon$ and confidence $\delta$, we have that with probability at least $1-\delta$, it will run for at most*

$$\max\left\{\min_N 16N \quad s.t. \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\phi\in\Phi} \phi^\top (N\mathbf{\Lambda} + \mathbf{\Lambda}_0 + \mathbf{\Lambda}_{\text{off}})^{-1}\phi \le \frac{\epsilon}{6}, \quad \frac{\text{poly}(\beta, R, d, H, M, \log 1/\delta, \log|\Phi|)}{\epsilon^{4/5}}\right\}.$$

*episodes, and will return data $\{\phi_\tau\}_{\tau=1}^N$ with covariance $\widehat{\mathbf{\Sigma}}_N = \sum_{\tau=1}^N \phi_\tau\phi_\tau^\top$ such that*

$$f_{\widehat{i}}(N^{-1}\widehat{\mathbf{\Sigma}}_N) \le N\epsilon,$$

*where $\widehat{i}$ is the iteration on which* OPTCOV *terminates.*

*Proof.* Our goal is to simply find a setting of $i$ that is sufficiently large to guarantee the condition $f_i(\widehat{\mathbf{\Lambda}}_i) \le K_i T_i \epsilon$ is met. By Lemma C.1, we have with probability at least $1 - \delta/(2i^2)$:

$$f_i(\widehat{\mathbf{\Lambda}}_i) \le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) + \frac{\beta_i R^2(\log T_i + 3)}{2T_i} + \sqrt{\frac{4M^2\log(8i^2T_i/\delta)}{K_i}} + \sqrt{\frac{c_1 M^2 d^4 H^4 \log^3(8i^2 HK_iT_i/\delta)}{K_i}} + \frac{c_2 M d^4 H^3 \log^{7/2}(4i^2 HK_iT_i/\delta)}{K_i}$$

$$\leq 3\max\left\{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}} f_i(\boldsymbol{\Lambda}), \frac{\beta_i R^2(\log T_i+3)}{2T_i},\right.$$
$$\left.\sqrt{\frac{4M^2\log(8i^2T_i/\delta)}{K_i}}+\sqrt{\frac{c_1M^2d^4H^4\log^3(8i^2HK_iT_i/\delta)}{K_i}}+\frac{c_2Md^4H^3\log^{7/2}(4i^2HK_iT_i/\delta)}{K_i}\right\}.$$

So a sufficient condition for $f_i(\widehat{\boldsymbol{\Lambda}}_i)\leq K_iT_i\epsilon$ is that

$$K_iT_i \geq \frac{3}{\epsilon}\max\left\{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}} f_i(\boldsymbol{\Lambda}), \frac{\beta_i R^2(\log T_i+3)}{2T_i},\right.$$
$$\left.\sqrt{\frac{4M^2\log(8i^2T_i/\delta)}{K_i}}+\sqrt{\frac{c_1M^2d^4H^4\log^3(8i^2HK_iT_i/\delta)}{K_i}}+\frac{c_2Md^4H^3\log^{7/2}(4i^2HK_iT_i/\delta)}{K_i}\right\}. \tag{C.1}$$

Recall that

$$f_i(\boldsymbol{\Lambda})=\frac{1}{\eta_i}\log\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta_i\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\boldsymbol{\Lambda})^{-1}}}\right),\quad \mathbf{A}_i(\boldsymbol{\Lambda})=\boldsymbol{\Lambda}+\frac{1}{T_iK_i}\boldsymbol{\Lambda}_{0,i}+\frac{1}{T_iK_i}\boldsymbol{\Lambda}_{\text{off}}.$$

By Lemma D.1 of (Wagenmaker & Jamieson, 2022), we can bound

$$\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\boldsymbol{\Lambda})^{-1}}\leq f_i(\boldsymbol{\Lambda})\leq\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\boldsymbol{\Lambda})^{-1}}+\frac{\log|\Phi|}{\eta_i}.$$

Thus,

$$\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}} f_i(\boldsymbol{\Lambda})\leq\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\boldsymbol{\Lambda})^{-1}}+\frac{\log|\Phi|}{\eta_i}$$
$$=\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}T_iK_i\boldsymbol{\phi}^\top(T_iK_i\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{0,i}+\boldsymbol{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}+\frac{\log|\Phi|}{\eta_i}$$

By our choice of $\eta_i=2^{2i/5}$, and $K_i=2^i$, $T_i=2^i$, we can ensure that

$$K_iT_i\geq\frac{6}{\epsilon}\frac{\log|\Phi|}{\eta_i}$$

as long as $i\geq\frac{2}{5}\log_2[\frac{6\log|\Phi|}{\epsilon}]$. To ensure that

$$T_iK_i\geq\frac{6}{\epsilon}\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}T_iK_i\boldsymbol{\phi}^\top(T_iK_i\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{0,i}+\boldsymbol{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}$$

it suffices to take

$$i\geq\arg\min_i i\quad\text{s.t.}\quad\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\boldsymbol{\phi}^\top(2^{3i}\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{0,i}+\boldsymbol{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}\leq\frac{\epsilon}{6}.$$

Since we assume that we can lower bound $\boldsymbol{\Lambda}_{0,i}\succeq\boldsymbol{\Lambda}_0$ for each $i$, so this can be further simplified to

$$i\geq\arg\min_i i\quad\text{s.t.}\quad\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\boldsymbol{\phi}^\top(2^{3i}\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_0+\boldsymbol{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi}\leq\frac{\epsilon}{6}. \tag{C.2}$$

We next want to show that

$$T_iK_i\geq\frac{3}{\epsilon}\cdot\frac{\beta_iR^2(\log T_i+3)}{2T_i}.$$

Bounding $\beta_i \le \eta_i \beta$, a sufficient condition for this is that

$$i \ge \frac{2}{5}\left(\log_2(12\beta R^2 i) + \log_2 \frac{1}{\epsilon}\right).$$

By Lemma A.1, it suffices to take

$$i \ge \frac{6}{5}\log_2(9\beta R^2 \log_2 \frac{1}{\epsilon}) + \frac{2}{5}\log_2 \frac{1}{\epsilon} \tag{C.3}$$

to meet this condition (this assumes that $12\beta R^2 \ge 1$ and $\frac{1}{4}\log_2 \frac{1}{\epsilon} \ge 1$—if either of these is not the case we can just replace them with 1 without changing the validity of the final result).

Finally, we want to ensure that

$$T_i K_i \ge \frac{3}{\epsilon}\left(\sqrt{\frac{4M^2 \log(8i^2 T_i/\delta)}{K_i}} + \sqrt{\frac{c_1 M^2 d^4 H^4 \log^3(8i^2 H K_i T_i/\delta)}{K_i}} + \frac{c_2 M d^4 H^3 \log^{7/2}(4i^2 H K_i T_i/\delta)}{K_i}\right).$$

To guarantee this, it suffices that

$$2^{5i/2} \ge \frac{c}{\epsilon}\sqrt{M^2 d^4 H^4 i^3 \log^3(iH/\delta)}, \quad 2^{3i} \ge \frac{c}{\epsilon} \cdot M d^4 H^3 i^{7/2} \log^{7/2}(iH/\delta)$$

or, equivalently,

$$i \ge \frac{4}{5}\log_2(cMdHi\log(H/\delta)) + \frac{2}{5}\log_2 \frac{1}{\epsilon}, \quad i \ge \frac{4}{3}\log_2(cMdH\log(H/\delta)) + \frac{1}{3}\log_2 \frac{1}{\epsilon}.$$

By Lemma A.1, it then suffices to take

$$\begin{aligned} i &\ge \frac{12}{5}\log(cMdH\log(H/\delta)\log_2 1/\epsilon) + \frac{2}{5}\log_2 \frac{1}{\epsilon}, \\ i &\ge 4\log_2(cMdH\log(H/\delta)\log_2 1/\epsilon) + \frac{1}{3}\log_2 \frac{1}{\epsilon} \end{aligned} \tag{C.4}$$

Thus, a sufficient condition to guarantee (C.1) is that $i$ is large enough to satisfy (C.2), (C.3), and (C.4) and $i \ge \frac{2}{5}\log_2[\frac{6\log|\Phi|}{\epsilon}]$.

If $\widehat{i}$ is the final round, the total complexity scales as

$$\sum_{i=1}^{\widehat{i}} T_i K_i = \sum_{i=1}^{\widehat{i}} 2^{3i} \le 2 \cdot 2^{3\widehat{i}}.$$

Using the sufficient condition on $i$ given above, we can bound the total complexity as

$$\max\left\{\min_N 16N \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\boldsymbol{\phi}\in\Phi} \boldsymbol{\phi}^\top (N\mathbf{\Lambda} + \mathbf{\Lambda}_0 + \mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi} \le \frac{\epsilon}{6}, \right.$$
$$\left. \frac{\text{poly}(\beta, R, d, H, M, \log 1/\delta, \log|\Phi|)}{\epsilon^{4/5}}\right\}.$$

□

**Theorem 4.** *Consider running* OptCov *with some $\epsilon_{\text{exp}} > 0$ and functions $f_i$ as defined in Lemma C.2, for $\mathbf{\Lambda}_{i,0}$ the matrix returned by running* ConditionedCov *(Wagenmaker & Jamieson, 2022) with $N \leftarrow T_i K_i$, $\delta \leftarrow \delta/(2i^2)$, and some $\underline{\lambda} \ge 0$. Then with probability $1 - 2\delta$, this procedure will collect at most*

$$\max\left\{\min_N C \cdot N \quad \textit{s.t.} \quad \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} \max_{\boldsymbol{\phi}\in\Phi} \boldsymbol{\phi}^\top (N(\mathbf{\Lambda} + \bar{\lambda} I) + \mathbf{\Lambda}_{\text{off}})^{-1}\boldsymbol{\phi} \le \frac{\epsilon_{\text{exp}}}{6}, \right.$$

$$\frac{\mathrm{poly}(d, H, \frac{1}{\lambda^\star_{\min}}, \log 1/\delta, \underline{\lambda}, \log|\Phi|)}{\epsilon_{\exp}^{4/5}}\Bigg\}$$

*episodes, where*

$$\bar{\lambda} = \min\left\{\frac{(\lambda^\star_{\min})^2}{d}, \frac{\lambda^\star_{\min}}{d^3 H^3 \log^{7/2} 1/\delta}\right\} \cdot \mathrm{poly}\log\left(\frac{1}{\lambda^\star_{\min}}, d, H, \underline{\lambda}, \log\frac{1}{\delta}\right)^{-1},$$

*and will produce covariates* $\widehat{\mathbf{\Sigma}}$ *such that*

$$\max_{\phi\in\Phi} \|\phi\|^2_{(\widehat{\mathbf{\Sigma}}+\mathbf{\Lambda}_{i,0}+\mathbf{\Lambda}_{\mathrm{off}})^{-1}} \le \epsilon_{\exp}$$

*and*

$$\lambda_{\min}(\widehat{\mathbf{\Sigma}} + \mathbf{\Lambda}_{i,0} + \mathbf{\Lambda}_{\mathrm{off}}) \ge \max\{d\log 1/\delta, \underline{\lambda}\}.$$

*Proof.* This proof follows closely the proof of Theorem 9 of (Wagenmaker & Jamieson, 2022).
Note that the total failure probability of our calls to CONDITIONEDCOV is at most

$$\sum_{i=1}^{\infty} \frac{\delta}{2i^2} = \frac{\pi^2}{12}\delta \le \delta.$$

For the remainder of the proof, we will then assume that we are on the success event of CONDITIONEDCOV, as defined in Lemma D.8 of (Wagenmaker & Jamieson, 2022).
By Lemma D.5 of (Wagenmaker & Jamieson, 2022), it suffices to bound the smoothness constants of $f_i(\mathbf{\Lambda})$ by

$$L_i = \|\mathbf{\Lambda}_{i,0}^{-1}\|^2_{\mathrm{op}}, \quad \beta_i = 2\|\mathbf{\Lambda}_{i,0}^{-1}\|^3_{\mathrm{op}}(1+\eta_i\|\mathbf{\Lambda}_{i,0}^{-1}\|_{\mathrm{op}}), \quad M_i = \|\mathbf{\Lambda}_{i,0}^{-1}\|^2_{\mathrm{op}}.$$

By Lemma D.8 of (Wagenmaker & Jamieson, 2022), on the success event of CONDITIONEDCOV we have that

$$\lambda_{\min}(\mathbf{\Lambda}_{i,0}) \ge \min\left\{\frac{(\lambda^\star_{\min})^2}{d}, \frac{\lambda^\star_{\min}}{d^3 H^3 \log^{7/2} 1/\delta}\right\} \cdot \mathrm{poly}\log\left(\frac{1}{\lambda^\star_{\min}}, d, H, \underline{\lambda}, i, \log\frac{1}{\delta}\right)^{-1} =: \bar{\lambda}.$$

Thus, we can bound, for all $i$:

$$L_i = M_i \le \max\left\{\frac{d^2}{(\lambda^\star_{\min})^4}, \frac{d^6 H^6 \log^7 1/\delta}{(\lambda^\star_{\min})^2}\right\} \cdot \mathrm{poly}\log\left(\frac{1}{\lambda^\star_{\min}}, d, H, \underline{\lambda}, i, \log\frac{1}{\delta}\right),$$
$$\beta_i \le \eta_i \cdot \mathrm{poly}\left(d, H, \log 1/\delta, \frac{1}{\lambda^\star_{\min}}, \underline{\lambda}, i, \log|\Phi|\right).$$

Assume that the termination condition of OPTCOV is met for $\widehat{i}$ satisfying

$$\widehat{i} \le \log\left(\mathrm{poly}\left(\frac{1}{\epsilon_{\exp}}, d, H, \log 1/\delta, \frac{1}{\lambda^\star_{\min}}, \underline{\lambda}, \log|\Phi|\right)\right). \tag{C.5}$$

We assume this holds and justify it at the conclusion of the proof. For notational convenience, define

$$\iota := \mathrm{poly}\left(\log\frac{1}{\epsilon_{\exp}}, d, H, \log 1/\delta, \frac{1}{\lambda^\star_{\min}}, \underline{\lambda}, \log|\Phi|\right).$$

Given this upper bound on $\widehat{i}$, set

$$L = M := \max\left\{\frac{d^2}{(\lambda^\star_{\min})^4}, \frac{d^6 H^6 \log^7 1/\delta}{(\lambda^\star_{\min})^2}\right\} \cdot \mathrm{poly}\log\iota, \qquad \beta := \iota.$$

With this choice of $L, M, \beta$, we have $L_i \le L, M_i \le M, \beta_i \le \eta_i\beta$ for all $i \le \widehat{i}$.

Since on the success event of CONDITIONEDCOV we have have $\mathbf{\Lambda}_{i,0} \succeq \bar{\lambda} \cdot I$, we can apply Lemma C.2 with $\mathbf{\Lambda}_0 = \bar{\lambda} \cdot I$ and get that, with probability at least $1 - \delta$, OPTCOV terminates after at most

$$\max\left\{ \min_N 16N \quad \text{s.t.} \quad \inf_{\mathbf{\Lambda} \in \mathbf{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \boldsymbol{\phi}^\top (N\mathbf{\Lambda} + \bar{\lambda} \cdot I + \mathbf{\Lambda}_{\text{off}})^{-1} \boldsymbol{\phi} \leq \frac{\epsilon_{\text{exp}}}{6}, \right.$$
$$\left. \frac{\text{poly}(d, H, \underline{\lambda}, 1/\lambda^\star_{\min}, \log 1/\epsilon_{\text{exp}}, \log 1/\delta, \log |\Phi|)}{\epsilon_{\text{exp}}^{4/5}} \right\}.$$

episodes, and returns data $\{\boldsymbol{\phi}_\tau\}_{\tau=1}^N$ with covariance $\widehat{\mathbf{\Sigma}} = \sum_{\tau=1}^N \boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top$ such that

$$f_{\widehat{i}}(N^{-1}\widehat{\mathbf{\Sigma}}) \leq N\epsilon_{\text{exp}},$$

where $\widehat{i}$ is the iteration on which OPTCOV terminates, and choosing $\mathbf{\Lambda}_0 = \bar{\lambda} \cdot I$

By Lemma D.1 of (Wagenmaker & Jamieson, 2022) we have

$$N \cdot \max_{\boldsymbol{\phi} \in \Phi} \|\boldsymbol{\phi}\|^2_{(\widehat{\mathbf{\Sigma}} + \mathbf{\Lambda}_{\widehat{i},0} + \mathbf{\Lambda}_{\text{off}})^{-1}} \leq f_{\widehat{i}}(N^{-1}\widehat{\mathbf{\Sigma}}).$$

The final upper bound on the number of episodes collected and the lower bound on the minimum eigenvalue of the covariates follows from Lemma D.8 of (Wagenmaker & Jamieson, 2022).

It remains to justify our bound on $\widehat{i}$, (C.5). Note that by definition of OPTCOV, if we run for a total of $\bar{N}$ episodes, we can bound $\widehat{i} \leq \frac{1}{4}\log_2(\bar{N})$. However, we see that the bound on $\widehat{i}$ given in (C.5) upper bounds $\frac{1}{4}\log_2(\bar{N})$ for $\bar{N}$ the upper bound on the number of samples collected by OPTCOV stated above. Thus, our bound on $\widehat{i}$ is valid. □

## D. Leveraging Offline Data Yields a Provable Improvement

*Proof of Proposition 1.* Let $\mathcal{M}^i$, $i \in \{1, 2\}$, denote the MDP with three states—an initial state $s_0$ and two other states $s_1, s_2$, and three actions—and:

$$\begin{aligned}
&r_0(s_0, a_1) = 1, \quad r_0(s_0, a_2) = r_0(s_0, a_3) = 0 \\
&P_0(s_1|s_0, a_1) = 1 - p, P_0(s_2|s_0, a_1) = p, \\
&P_0(s_1|s_0, a_2) = 1, P_0(s_2|s_0, a_2) = 0 \\
&P_0(s_1|s_0, a_3) = 0, P_0(s_2|s_0, a_3) = 1
\end{aligned}$$

and

$$\begin{aligned}
&r_1(s_1, a_1) = 1/2 + \Delta, \quad r_1(s_1, a_2) = r_1(s_1, a_3) = 1/2 \\
&r_1(s_2, a_i) = 1/2, \quad r_1(s_2, a_j) = 0, j \neq i.
\end{aligned}$$

After taking an action at $h = 1$, the MDP terminates. We set $p = \sqrt{\epsilon}$ and $\Delta = 6\epsilon$. As this is a tabular MDP, we can simply take the feature vectors to be the standard basis vectors and will have $d = 9$. In addition, as every state and action are easily reachable, we have $\lambda^\star_{\min} = \Omega(1)$.

Let $\mathfrak{D}_{\text{off}}$ be a dataset formed by playing the logging policy $\pi^{\log}$ for $\text{poly}(1/\epsilon)$ episodes, for $\pi^{\log}$ specified as:

$$\begin{aligned}
&\pi_0^{\log}(a_1|s_0) = \pi_0^{\log}(a_2|s_0) = \pi_0^{\log}(a_3|s_0) = 1/3, \\
&\pi_1^{\log}(a_1|s_1) = \pi_1^{\log}(a_2|s_1) = \pi_1^{\log}(a_3|s_1) = 1/3, \\
&\pi_1^{\log}(a_1|s_2) = \pi_1^{\log}(a_2|s_2) = 0, \pi_1^{\log}(a_3|s_2) = 1.
\end{aligned}$$

Note that with this logging policy $\mathfrak{D}_{\text{off}}$ will contain $\text{poly}(1/\epsilon)$ samples from every state-action pair except for $(s_2, a_1)$ and $(s_2, a_2)$, for which it contains 0 samples. Note also that $\mathbb{P}_{\mathcal{M}^1, \pi^{\log}}[\mathcal{E}] = \mathbb{P}_{\mathcal{M}^2, \pi^{\log}}[\mathcal{E}]$ for any event $\mathcal{E}$—the measures induced by playing $\pi^{\log}$ on $\mathcal{M}^1$ and $\mathcal{M}^2$ are identical.

The first two conclusions are then an immediate consequence of this construction, Lemma D.1, and Lemma D.2.

Note that on both $\mathcal{M}^1$ and $\mathcal{M}^2$, the suboptimality of any policy $\pi$ is at least $1 - \pi_0(a_1|s_0)$. Furthermore, we can bound

$$[\boldsymbol{\phi}_{\pi,1}]_{(s_2,a)} \leq p\pi_0(a_1|s_0) + 1 - \pi_0(a_1|s_0) \leq \sqrt{\epsilon} + 1 - \pi_0(a_1|s_0) \leq 2\max\{\sqrt{\epsilon}, 1 - \pi_0(a_1|s_0)\}.$$

Thus, using our setting of $\mathfrak{D}_{\text{off}}$, we have, for any $\pi$,

$$\frac{\|\phi_{\pi,1}\|^2_{(N\mathbf{\Lambda}+\mathbf{\Lambda}^1_{\text{off}})^{-1}}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \leq \frac{4\max\{\epsilon, (1-\pi_0(a_1|s_0))^2\}}{\min_i N_{s_2,a_i}} \cdot \frac{1}{(1-\pi_0(a_1|s_0))^2 \vee \epsilon^2} + \mathcal{O}(1)$$
$$\leq \frac{4}{\min_i N_{s_2,a_i}} \cdot \max\{1, \frac{1}{\epsilon}\} + \mathcal{O}(1).$$

To ensure that this is $\mathcal{O}(1)$, we therefore only need to make $N_{s_2,a_i} = \Omega(1/\epsilon)$ for each $i$. Therefore, the additional online samples required to find an $\epsilon$-optimal policy is $\mathcal{O}(1/\epsilon)$. The final conclusion then follows from Theorem 2. □

**Lemma D.1.** *Fix $\epsilon \leq 1/4$ and assume that $\Delta > \frac{16\epsilon}{3}$. For $\mathcal{M}^1$ and $\mathcal{M}^2$ constructed as in the proof of Proposition 1, any purely online algorithm must take at least $\Omega(\frac{1}{\Delta^2} \cdot \log \frac{1}{2.4\delta})$ episodes to identify an $\epsilon$-optimal policy with probability at least $1-\delta$.*

*Proof.* Assume our algorithm has returned some policy $\widehat{\pi}$. Let $\widehat{i} = \arg\max_a \widehat{\pi}_1(s_1, a)$. We show that if $\widehat{\pi}$ is $\epsilon$-optimal, then it must be the case that $\widehat{i} = a_1$.

Note that for $\epsilon \leq 1/4$, for $\widehat{\pi}$ to be $\epsilon$-optimal we must have that $\widehat{\pi}_0(a_1|s_0) \geq 3/4$ since, even if $\widehat{\pi}$ plays optimally in $s_1$ and $s_2$, if $\widehat{\pi}_0(a_1|s_0) < 3/4$ it will be at least $1/4$-suboptimal.

Now assume that $\widehat{\pi}$ is suboptimal in state $s_1$ by $\xi$. Then it follows that the total contribution to the suboptimality from state $s_1$ is at least $\frac{3}{4}(1-p)\xi$. Thus, if we assume the policy is $\epsilon$-optimal, it follows that $\xi \leq \frac{4\epsilon}{3(1-p)} \leq \frac{8\epsilon}{3}$ where the last inequality holds assuming that $p \leq 1/2$. Note that

$$\xi = \Delta(1 - \widehat{\pi}_1(a_1|s_1))$$

so it follows that

$$\Delta(1 - \widehat{\pi}_1(a_1|s_1)) \leq \frac{8\epsilon}{3} \iff \widehat{\pi}_1(a_1|s_1) \geq 1 - \frac{8\epsilon}{3\Delta}.$$

It follows that for $\Delta > \frac{16\epsilon}{3}$, we have $\widehat{\pi}_1(a_1|s_1) > 1/2$, which implies that $\widehat{i} = a_1$. In other words, if $\widehat{\pi}$ is $\epsilon$-optimal with probability at least $1-\delta$, $\widehat{i} = a_1$ with probability at least $1-\delta$.

By standard lower bounds on bandits (see e.g. (Kaufmann et al., 2016)), the complexity of identifying the best action in state $s_1$ with probability at least $1-\delta$ scales as $\Omega(\frac{1}{\Delta^2} \cdot \log \frac{1}{2.4\delta})$. As the above procedure is able to identify the best arm in $s_1$ with probability at least $1-\delta$, it follows that we must have collected at least $\Omega(\frac{1}{\Delta^2} \cdot \log \frac{1}{2.4\delta})$ samples from $s_1$, which serves as a lower on the total complexity. □

**Lemma D.2.** *For $\mathcal{M}^1$, $\mathcal{M}^2$, and $\mathfrak{D}_{\text{off}}$ constructed as in the proof of Proposition 1, any algorithm which returns some policy $\widehat{\pi}$ without further exploration must have:*

$$\max_{i\in\{1,2\}} \mathbb{E}_{\mathfrak{D}_{\text{off}}\sim\mathcal{M}^i}[V_0^\star(\mathcal{M}^i) - V_0^{\widehat{\pi}}(\mathcal{M}^i)] \geq \frac{3\sqrt{\epsilon}}{16}.$$

*Proof.* Note that for $\epsilon \leq 1/4$, for $\widehat{\pi}$ to be $\epsilon$-optimal we must have that $\widehat{\pi}_0(a_1|s_0) \geq 3/4$ since, even if $\widehat{\pi}$ plays optimally in $s_1$ and $s_2$, if $\widehat{\pi}_0(a_1|s_0) < 3/4$ it will be at least $1/4$-suboptimal. With $p = \sqrt{\epsilon}$, this implies that we will transition to $s_2$ with probability at least $\frac{3}{4}p = \frac{3}{4}\sqrt{\epsilon}$. Note that on $\mathcal{M}^i$ the suboptimality of policy $\widehat{\pi}$ in state $s_2$ is given by

$$\frac{1}{2}(1 - \widehat{\pi}_1(a_i|s_2))$$

so the total suboptimality that $s_2$ contributes is at least

$$\frac{3}{8}\sqrt{\epsilon}(1 - \widehat{\pi}_1(a_i|s_2)).$$

It follows that we can lower bound

$$\max_{i\in\{1,2\}} \mathbb{E}_{\mathfrak{D}_{\text{off}}\sim\mathcal{M}^i}[V_0^\star(\mathcal{M}^i) - V_0^{\widehat{\pi}}(\mathcal{M}^i)] \geq \frac{3\sqrt{\epsilon}}{16}\Big(\mathbb{E}_{\mathfrak{D}_{\text{off}}\sim\mathcal{M}^1}[1 - \widehat{\pi}_1(a_1|s_2)] + \mathbb{E}_{\mathfrak{D}_{\text{off}}\sim\mathcal{M}^2}[1 - \widehat{\pi}_1(a_2|s_2)]\Big).$$

Define the function $\psi(\mathfrak{D}_{\text{off}}) = \mathbb{I}\{a \neq a_1\}$ for $a \sim \widehat{\pi}_1(\cdot \mid s_2)$. Then

$$\mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^1}[1 - \widehat{\pi}_1(a_1|s_2)] = \mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^1}[\psi(\mathfrak{D}_{\text{off}}) = 1]$$

and

$$\mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^2}[1 - \widehat{\pi}_1(a_2|s_2)] \geq \mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^2}[\psi(\mathfrak{D}_{\text{off}}) = 0],$$

so the total suboptimality is lower bounded by

$$\frac{3\sqrt{\epsilon}}{16}\Big(\mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^1}[\psi(\mathfrak{D}_{\text{off}}) = 1] + \mathbb{E}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^2}[\psi(\mathfrak{D}_{\text{off}}) = 0]\Big) \geq \frac{3\sqrt{\epsilon}}{16}\Big(1 - \mathrm{TV}(\mathbb{P}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^1}, \mathbb{P}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^2})\Big).$$

However, by the construction of $\mathfrak{D}_{\text{off}}$, the distribution of $\mathfrak{D}_{\text{off}}$ is identical under $\mathcal{M}^1$ and $\mathcal{M}^2$, so $\mathrm{TV}(\mathbb{P}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^1}, \mathbb{P}_{\mathfrak{D}_{\text{off}} \sim \mathcal{M}^2}) = 0$, which proves the result. □

## E. The Cost of Verifiability

*Proof of Proposition 2.* Assume that some procedure returns a possibly random policy $\widehat{\pi}$. Let $\widehat{i} = \arg\max_i \widehat{\pi}(i)$. Then if $\widehat{\pi}$ is $\epsilon$-optimal, it follows that $\widehat{i} = 1$, since the suboptimality of a policy is given by $(1 - \widehat{\pi}(1)) \cdot 3\epsilon$, so if $\widehat{\pi}$ is $\epsilon$-optimal, it follows that $\widehat{\pi}(1) \geq 2\epsilon/3$.

We can then apply standard lower bounds for multi-armed bandits (see e.g. (Kaufmann et al., 2016)) to get that any $(\epsilon, \delta)$-PAC algorithm must collect at least $\Omega(\frac{1}{9\epsilon^2} \log \frac{1}{2.4\delta})$ samples from each arm $2 - A$. □

### E.1. Proof of Theorem 3

**Algorithm 5** Verifiable Offline RL

**input:** tolerance $\epsilon$, confidence $\delta$, policy set $\Pi$
$\Pi_1 \leftarrow \Pi$, $\widehat{\boldsymbol{\phi}}^1_{\pi,1} \leftarrow \mathbb{E}_{a \sim \pi_1(\cdot|s_1)}[\boldsymbol{\phi}(s_1, a)], \forall \pi \in \Pi$, $\lambda \leftarrow 1/d$
**for** $h = 1, 2, \ldots, H$ **do**
  **for** $\pi \in \Pi_1$ **do**
    $\widehat{\boldsymbol{\phi}}_{\pi,h+1} \leftarrow \Big(\sum_{(s_h,a_h,r_h,s_{h+1}) \in \mathfrak{D}^h_{\text{off}}} \boldsymbol{\phi}_{\pi,h+1}(s_{h+1})\boldsymbol{\phi}(s_h, a_h)^\top (\boldsymbol{\Lambda}^h_{\text{off}})^{-1}\Big)\widehat{\boldsymbol{\phi}}_{\pi,h}$
  $\widehat{\boldsymbol{\theta}}_h \leftarrow (\boldsymbol{\Lambda}^h_{\text{off}})^{-1} \sum_{(s_h,a_h,r_h,s_{h+1}) \in \mathfrak{D}^h_{\text{off}}} \boldsymbol{\phi}(s_h, a_h) r_h$
**for** $\ell = 1, 2, \ldots, \lceil \log \frac{4}{\epsilon} \rceil$ **do**
  $\epsilon_\ell \leftarrow 2^{-\ell}$, $\beta_\ell \leftarrow H^4 \left(2\sqrt{d \log \frac{\lambda + T_{\text{off}}/d}{\lambda}} + 2 \log \frac{2H^2|\Pi|\ell^2}{\delta} + \sqrt{d\lambda}\right)^2$
  **for** $h = 1, 2, \ldots, H$ **do**
    **if** $\max_{\pi \in \Pi_\ell} \|\widehat{\boldsymbol{\phi}}_{\pi,h}\|^2_{(\boldsymbol{\Lambda}^h_{\text{off}})^{-1}} > \epsilon_\ell^2/\beta_\ell$ or $\lambda_{\min}(\boldsymbol{\Lambda}^h_{\text{off}}) < \log \frac{4H^2|\Pi|\ell^2}{\delta}$ **then**
      **return** $\emptyset$
  Set

$$\Pi_{\ell+1} \leftarrow \Pi_\ell \backslash \Big\{\pi \in \Pi_\ell \ : \ \widehat{V}^\pi_0 < \sup_{\pi' \in \Pi_\ell} \widehat{V}^{\pi'}_0 - 2\epsilon_\ell\Big\} \quad \text{for} \quad \widehat{V}^\pi_0 := \textstyle\sum_{h=1}^H \langle \widehat{\boldsymbol{\phi}}_{\pi,h}, \widehat{\boldsymbol{\theta}}_h \rangle$$

  **if** $|\Pi_{\ell+1}| = 1$ **then return** $\pi \in \Pi_{\ell+1}$
**return** any $\pi \in \Pi_{\ell+1}$

*Proof of Theorem 3.* We prove Theorem 3 for the algorithm outlined in Algorithm 5.

Note that the condition

$$\max_{\pi \in \Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{(\boldsymbol{\Lambda}^h_{\text{off}})^{-1}} \leq \epsilon_\ell^2/\beta_\ell \quad \text{and} \quad \lambda_{\min}(\boldsymbol{\Lambda}^h_{\text{off}}) \geq \log \frac{4H^2|\Pi|\ell^2}{\delta}$$

is precisely the condition required on Appendix B of epoch $\ell$ and step $h$ of FTPEDEL. Thus, if Algorithm 5 returns some policy $\widehat{\pi} \neq \emptyset$, it is $\epsilon$-optimal with probability at least $1 - \delta$ by an argument identical to the proof of Theorem 2. We omit

details for the sake of brevity. As we do not need to run any additional exploration, the $\mathcal{O}(1/\epsilon^{3/2})$ term present in Theorem 2 is no longer incurred here.

It remains to show that the coverage condition given in Theorem 3, (5.1), suffices to ensure that the if statement on Appendix E.1 is never true. That $\lambda_{\min}(\mathbf{\Lambda}^h_{\text{off}}) \geq \log \frac{4H^2|\Pi|\ell^2}{\delta}$ for all $h$ is immediate by assumption. We next prove that $\max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} \leq \epsilon_\ell^2/\beta_\ell$ inductively. The base case is immediate by assumption. Now assume that, for all $h' < h$,

$$\max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h'}\|^2_{(\mathbf{\Lambda}^{h'}_{\text{off}})^{-1}} \leq \epsilon_\ell^2/\beta_\ell.$$

Then by Lemma B.3, we have that on the event $\mathcal{E}^{\ell,h}_{\text{est}}$,

$$\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2 \leq d\sum_{i=1}^h \left(2\sqrt{\log\frac{\det \mathbf{\Lambda}_{i,\ell}}{\lambda^d} + 2\log\frac{2H^2 d|\Pi|\ell^2}{\delta}} + \sqrt{d\lambda}\right)\cdot \frac{\epsilon_\ell^2}{\beta_\ell} \leq \frac{d\epsilon_\ell}{2H}$$

where the last inequality follows by our choice of $\beta_\ell$.

Now note that

$$\begin{aligned}
\max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} &\leq \max_{\pi\in\Pi_\ell} 2\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} + 2\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} \\
&\leq \max_{\pi\in\Pi_\ell} 2\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} + \frac{2\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2^2}{\lambda_{\min}(\mathbf{\Lambda}^h_{\text{off}})} \\
&\leq \max_{\pi\in\Pi_\ell} 2\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} + \frac{d^2\epsilon_\ell^2}{2H^2\lambda_{\min}(\mathbf{\Lambda}^h_{\text{off}})} \\
&\leq \max_{\pi\in\Pi(4\epsilon_\ell)} 2\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} + \frac{d^2\epsilon_\ell^2}{2H^2\lambda_{\min}(\mathbf{\Lambda}^h_{\text{off}})}
\end{aligned}$$

where the last inequality follows by Lemma B.6, which gives that $\Pi_\ell \subseteq \Pi(4\epsilon_\ell)$. It follows that as long as

$$\max_{\pi\in\Pi(4\epsilon_\ell)} 2\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} \leq \frac{\epsilon_\ell^2}{2\beta_\ell} \quad \text{and} \quad \frac{d^2\epsilon_\ell^2}{2H^2\lambda_{\min}(\mathbf{\Lambda}^h_{\text{off}})} \leq \frac{\epsilon_\ell^2}{2\beta_\ell}$$

the inductive hypothesis holds. The latter condition follows by assumption. The former also holds since

$$\begin{aligned}
\max_{\pi\in\Pi(4\epsilon_\ell)} 2\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} \leq \frac{\epsilon_\ell^2}{2\beta_\ell} &\iff \max_{\pi\in\Pi(4\epsilon_\ell)} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}}}{\epsilon_\ell^2} \leq \frac{1}{4\beta_\ell} \\
&\iff \max_{\pi\in\Pi(4\epsilon_\ell)} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}}}{\epsilon_\ell^2 \vee (V_0^\star - V_0^\pi)^2/16} \leq \frac{1}{4\beta_\ell}
\end{aligned}$$

but this is implied by our coverage assumption, (5.1). By a union bound, it follows that (5.1) is sufficient to guarantee that for each $h$ and $\ell$, with high probability,

$$\max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{(\mathbf{\Lambda}^h_{\text{off}})^{-1}} \leq \epsilon_\ell^2/\beta_\ell \quad \text{and} \quad \lambda_{\min}(\mathbf{\Lambda}^h_{\text{off}}) \geq \log\frac{4H^2|\Pi|\ell^2}{\delta}.$$

□

### E.2. Policy Verification

**Lemma E.1.** *Assume that we have some dataset $\mathfrak{D} = \{(s_{h-1,\tau}, a_{h-1,\tau}, r_{h-1,\tau}, s_{h,\tau})\}_{\tau=1}^K$ satisfying Assumption 2. Denote $\boldsymbol{\phi}_{h-1,\tau} = \boldsymbol{\phi}(s_{h-1,\tau}, a_{h-1,\tau})$ and $\mathbf{\Lambda}_{h-1} = \sum_{\tau=1}^K \boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top + \lambda I$.*

*Assume that $\Pi$ satisfies Assumption 3, and let $\Pi^\gamma_{\mathsf{cov}}$ denote the corresponding cover of $\Pi$. For each $\pi$, let*

$$\widehat{\mathcal{T}}_{\pi,h} = \left(\sum_{\tau=1}^K \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\boldsymbol{\phi}_{h-1,\tau}^\top\right)\mathbf{\Lambda}_{h-1}^{-1}.$$

*Consider some $\widehat{\pi} \in \Pi$, where $\widehat{\pi}$ might be correlated with $\mathfrak{D}$. Fix $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{v} \in \mathbb{R}^d$ satisfying $|\boldsymbol{v}^\top \boldsymbol{\phi}_{\pi,h}(s)| \le 1$ for all $s$. Then with probability at least $1-\delta$, we can bound, for all $\pi \in \Pi^\gamma_{\mathsf{cov}} \cup \{\widehat{\pi}\}$:*

$$|\boldsymbol{v}^\top(\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\boldsymbol{u}| \le \left(2\sqrt{\log\frac{\det \boldsymbol{\Lambda}_{h-1}}{\lambda^d} + 2\log\frac{\mathsf{N}_{\mathsf{cov}}(\Pi,\gamma)}{\delta}} + \sqrt{\lambda}\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}\|_2\right)\cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} + \frac{2K\cdot\gamma}{\sqrt{\lambda}}.$$

*Proof.* Construct a filtration as in Lemma B.1. Let $\mathcal{E}$ denote the event that, for each $\pi \in \Pi^\gamma_{\mathsf{cov}}$,

$$\left\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\right\|_2 \le 2\sqrt{\log\frac{\det \boldsymbol{\Lambda}_{h-1}}{\lambda^d} + 2\log\frac{\mathsf{N}_{\mathsf{cov}}(\Pi,\gamma)}{\delta}}.$$

Then using the same argument as in the proof of Lemma B.1, we have $\mathbb{P}[\mathcal{E}] \ge 1-\delta$, and on $\mathcal{E}$, we can immediately bound

$$|\boldsymbol{v}^\top(\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\boldsymbol{u}| \le \left(2\sqrt{\log\frac{\det \boldsymbol{\Lambda}_{h-1}}{\lambda^d} + 2\log\frac{\mathsf{N}_{\mathsf{cov}}(\Pi,\gamma)}{\delta}} + \sqrt{\lambda}\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}\|_2\right)\cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}, \quad \forall \pi \in \Pi^\gamma_{\mathsf{cov}}.$$

It remains to bound the error in the estimate $\widehat{\mathcal{T}}_{\widehat{\pi},h}$. Following the proof of Lemma B.1, we have that

$$|\boldsymbol{v}^\top(\mathcal{T}_{\widehat{\pi},h} - \widehat{\mathcal{T}}_{\widehat{\pi},h})\boldsymbol{u}| \le \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} \cdot \underbrace{\left\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\widehat{\pi},h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\widehat{\pi},h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\right\|_2}_{(a)} + \left|\lambda \boldsymbol{v}^\top \mathcal{T}_{\widehat{\pi},h}\boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}\right|.$$

Let $\widetilde{\pi} \in \Pi^\gamma_{\mathsf{cov}}$ denote the policy satisfying

$$\|\boldsymbol{\phi}_{\widehat{\pi},h}(s) - \boldsymbol{\phi}_{\widetilde{\pi},h}(s)\|_2 \le \gamma, \quad \forall s, h.$$

Note that such a $\widetilde{\pi}$ is guaranteed to exist under Assumption 3. We can bound

$$\begin{aligned}
(a) &= \Bigg\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2} \\
&\quad + \sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\widehat{\pi},h}(s_{h,\tau}) - \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\widehat{\pi},h}(s_{h,\tau}) + \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\Bigg\|_2 \\
&\le \Bigg\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\Bigg\|_2 \\
&\quad + \|\boldsymbol{\Lambda}_{h-1}^{-1/2}\|_{\mathrm{op}} \sum_{\tau=1}^K \left(\mathbb{E}[\|\boldsymbol{\phi}_{\widehat{\pi},h}(s_{h,\tau}) - \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})\|_2|\mathcal{F}_{h-1,\tau-1}] + \|\boldsymbol{\phi}_{\widehat{\pi},h}(s_{h,\tau}) - \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})\|_2\right) \\
&\le \Bigg\|\sum_{\tau=1}^K \boldsymbol{v}^\top\left(\mathbb{E}[\boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau-1}] - \boldsymbol{\phi}_{\widetilde{\pi},h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1/2}\Bigg\|_2 + \frac{2\gamma K}{\sqrt{\lambda}}.
\end{aligned}$$

The result then follows since, on $\mathcal{E}$, this is bounded, since $\widetilde{\pi} \in \Pi^\gamma_{\mathsf{cov}}$. □

*Proof of Corollary 2.* Corollary 2 can be proved almost identically to Theorem 2, but using Lemma E.1 to bound the error in the estimate of $\widehat{\boldsymbol{\phi}}_{\widehat{\pi},h}$ in place of Lemma B.1. In addition, for the $i$th call to FTPEDEL-SE, we set $\gamma = \frac{\sqrt{\lambda}}{2(T_{\mathrm{off}}+\bar{T}^i_{\mathrm{on}})}\cdot\epsilon$, as $T_{\mathrm{off}} + \bar{T}^i_{\mathrm{on}}$ will upper bound the total number of samples in the $i$th call, resulting in the approximation error in Lemma E.1 to be $\epsilon$.

As the goal is simply to determine whether or not $\widehat{\pi}$ is $\epsilon$-optimal, we terminate early if it is determined to be suboptimal, which yields the modified scaling in the policy gap.

#### E.2.1. Common Policy Classes Satisfy Assumption 3

We remark briefly on Assumption 3 and which policy classes satisfy it. A common policy class we might consider is one parameterized by some vectors $\boldsymbol{w} := (\boldsymbol{w}_h)_{h=1}^H$, and given by

$$\pi_h^{\boldsymbol{w}}(s) = \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h \rangle.$$

Note that the optimal policy takes this form (Jin et al., 2020). Policies of this form are in fact non-smooth, and therefore do not obviously satisfy Assumption 3. However, as shown in Section A.3 of (Wagenmaker & Jamieson, 2022), they can be approximated arbitrarily well with linear softmax policies. As linear softmax policies are smooth, it is straightforward to show that they satisfy Assumption 3 with $\mathsf{N}_{\mathsf{cov}}(\Pi, \gamma) = \widetilde{\mathcal{O}}(dH^2 \cdot \log \frac{1}{\gamma})$.

Another common policy class found in the literature (see e.g. (Jin et al., 2020) or (Jin et al., 2021b)) takes the form

$$\pi_h^{\boldsymbol{w},\boldsymbol{\Lambda}}(s) = \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h \rangle + \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}},$$

for some $\boldsymbol{\Lambda}_h \succeq 0$. This policy is again non-smooth, but can similarly be approximated arbitrarily well using a softmax policy class (where now the softmax is taken over $\{\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h \rangle + \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}\}_{a \in \mathcal{A}}$). As this policy class has $\mathcal{O}(d^2)$ parameters, we now have that $\mathsf{N}_{\mathsf{cov}}(\Pi, \gamma)$ scales with $d^2$.

More generally, any policy class that is "smooth" in $\boldsymbol{\phi}(s,a)$, or can be approximated by a policy class that is smooth in $\boldsymbol{\phi}(s,a)$, and that has a finite number of parameters will satisfy Assumption 3.

□

## F. Lower Bounds

### F.1. Instance-Dependent Lower Bound

**Theorem 5.** *Let $\mathfrak{M}$ denote the set of linear MDPs with reward vectors in $\mathbb{R}^d$. Fix some $h \in [H]$ and consider some algorithm that is $(0,\delta)$-PAC on $\mathfrak{M}$ and has knowledge of the dynamics of each MDP, and let $\tau$ denote the number of online samples collected by this algorithm. Then there exists some MDP $\mathcal{M} \in \mathfrak{M}$ with $\mathcal{O}(1)$ states and actions such that, for any set of offline data $\mathfrak{D}_{\mathrm{off}}$ satisfying Assumption 2 on $\mathcal{M}$, we must have $\mathbb{E}^{\mathcal{M}}[\tau] \geq \Omega(N_h^{\mathrm{det}}(\mathfrak{D}_{\mathrm{off}}, 0; \log \frac{1}{\delta}))$, for*

$$N_h^{\mathrm{det}}(\mathfrak{D}_{\mathrm{off}}) := \min_N N \quad \textit{s.t.} \quad \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h} \max_{\pi \in \Pi_{\mathsf{det}}, \pi \neq \pi^\star} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{(N\boldsymbol{\Lambda} + \boldsymbol{\Lambda}_{\mathrm{off}}^h)^{-1}}}{(V_0^\star - V_0^\pi)^2} \leq \frac{1}{\log 1/\delta},$$

*where $\Pi_{\mathsf{det}}$ is the set of all deterministic policies.*

**Remark F.1** ($\Pi_{\mathsf{det}}$ vs $\Pi_{\mathsf{lsm}}$ Dependence)**.** Note that Theorem 5 depends on $\Pi_{\mathsf{det}}$ and not $\Pi_{\mathsf{lsm}}$, the policy class our upper bound Theorem 2 scales with. Our upper bounds were in fact proved for a generic policy class $\Pi$, and provide a guarantee on finding a policy $\widehat{\pi}$ that is $\epsilon$-optimal with respect to the best policy in $\Pi$. In the case of linear MDPs, to construct a policy class that is guaranteed to have a policy $\epsilon$-optimal with respect to the best possible policy on the MDP, we rely on the linear softmax policy construction. However, it is well known that in tabular MDPs, there always exists a deterministic optimal policy. Thus, in tabular settings, it suffices to choose $\Pi \leftarrow \Pi_{\mathsf{det}}$ to guarantee we have a globally $\epsilon$-optimal policy in our class. Given this, in the setting of Theorem 5 with a finite number of states and actions, we can simply run FTPEDEL with $\Pi \leftarrow \Pi_{\mathsf{det}}$ and will obtain the same guarantee as in Theorem 2 but with $\Pi_{\mathsf{lsm}}$ replaced by $\Pi_{\mathsf{det}}$, matching the guarantee of Theorem 5.

#### F.1.1. Linear Bandits with Random Arms

Towards proving a lower bound for linear MDPs, we first consider the setting of linear bandits with random arms. In this setting, assume we have some set of arms $\mathcal{A}$. When arm $a \in \mathcal{A}$ is played, a vector $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$ is sampled from some distribution $\xi_a$, which we assume is known to the learner. The learner then observes $\boldsymbol{z}$ as well as $y = \boldsymbol{z}^\top \boldsymbol{\theta}^\star + \eta$ for some noise $\eta$, and unknown $\boldsymbol{\theta}^\star \in \mathbb{R}^d$.

Note that this can simply be thought of as a multi-armed bandit where arm $a$ has expected reward $\mathbb{E}_\eta \mathbb{E}_{\boldsymbol{z} \sim \xi_a}[\boldsymbol{z}^\top \boldsymbol{\theta}^\star + \eta]$. Define $\boldsymbol{\Lambda}_a := \mathbb{E}_{\boldsymbol{z} \sim \xi_a}[\boldsymbol{z}\boldsymbol{z}^\top]$. For some set $\mathcal{X} \subseteq \mathbb{R}^d$, our goal is to identify $\boldsymbol{x}^\star(\boldsymbol{\theta}^\star) \in \mathcal{X}$ such that

$$\boldsymbol{x}^\star(\boldsymbol{\theta}^\star)^\top \boldsymbol{\theta}^\star \geq \max_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{x}^\top \boldsymbol{\theta}^\star.$$

In the offline-to-online setting, we assume the learner has access to some set of data $\{(a_s, \boldsymbol{z}_s, y_s)\}_{s=1}^{T_{\mathrm{off}}}$.

Finally, we say that an algorithm is $\delta$-PAC if it returns $\boldsymbol{x}^\star(\boldsymbol{\theta}^\star)$ with probability at least $1 - \delta$.

**Proposition 4.** *Assume we are in the linear bandit with random arms settings defined above, for noise distribution $\eta \sim \mathcal{N}(0,1)$. Then any $\delta$-PAC strategy with online stopping time $\tau$ must have*

$$\mathbb{E}_{\boldsymbol{\theta}^\star}[\tau] \geq \min_t \sum_{a\in\mathcal{A}} t_a \quad s.t. \quad \max_{\boldsymbol{x}\neq \boldsymbol{x}^\star(\boldsymbol{\theta}^\star)} \frac{\|\boldsymbol{x}-\boldsymbol{x}^\star(\boldsymbol{\theta}^\star)\|^2_{\mathbf{A}(t)^{-1}}}{((\boldsymbol{x}^\star(\boldsymbol{\theta}^\star)-\boldsymbol{x})^\top\boldsymbol{\theta}^\star)^2} \leq \frac{1}{\log\frac{1}{2.4\delta}}$$

*for* $\mathbf{A}(t) = \sum_{a\in\mathcal{A}} t_a \boldsymbol{\Lambda}_a + \sum_{s=1}^{T_{\text{off}}} \boldsymbol{z}_s\boldsymbol{z}_s^\top$.

*Proof.* We follow the proof of Theorem 1 of (Fiez et al., 2019), with some small modifications. Let $\Theta^\star_{\text{alt}} = \{\boldsymbol{\theta}\in\mathbb{R}^d : \boldsymbol{x}^\star(\boldsymbol{\theta}) \neq \boldsymbol{x}^\star(\boldsymbol{\theta}^\star)\}$ denote the set of alternate instances. Let $\nu_{\boldsymbol{\theta},\boldsymbol{z}} = \mathcal{N}(\boldsymbol{\theta}^\top\boldsymbol{z},1)$, $\nu_{\boldsymbol{\theta},a}$ the reward distribution of action $a$, and let $t_a^{\text{on}}$ the total number of (online) pulls of arm $a$.

Given some alternate $\boldsymbol{\theta}$, the KL divergence between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$ is given by

$$\sum_{s=1}^{T_{\text{off}}} \text{KL}(\nu_{\boldsymbol{\theta}^\star,\boldsymbol{z}_s}, \nu_{\boldsymbol{\theta},\boldsymbol{z}_s}) + \sum_{a\in\mathcal{A}} \mathbb{E}_{\boldsymbol{\theta}^\star}[t_a^{\text{on}}]\text{KL}(\nu_{\boldsymbol{\theta}^\star,a}, \nu_{\boldsymbol{\theta},a}).$$

Then Lemma 1 of (Kaufmann et al., 2016) and the fact that the algorithm is $\delta$-PAC give that, for any $\boldsymbol{\theta}\in\Theta^\star_{\text{alt}}$,

$$\sum_{s=1}^{T_{\text{off}}} \text{KL}(\nu_{\boldsymbol{\theta}^\star,\boldsymbol{z}_s}, \nu_{\boldsymbol{\theta},\boldsymbol{z}_s}) + \sum_{a\in\mathcal{A}} \mathbb{E}_{\boldsymbol{\theta}^\star}[t_a^{\text{on}}]\text{KL}(\nu_{\boldsymbol{\theta}^\star,a}, \nu_{\boldsymbol{\theta},a}) \geq \log\frac{1}{2.4\delta}.$$

Note that, by the convexity of the KL divergence, we have

$$\begin{aligned}
\text{KL}(\nu_{\boldsymbol{\theta}^\star,a}, \nu_{\boldsymbol{\theta},a}) &\leq \sum_{\boldsymbol{z}\in\mathcal{Z}} \xi_a(\boldsymbol{z})\text{KL}(\nu_{\boldsymbol{\theta}^\star,\boldsymbol{z}}, \nu_{\boldsymbol{\theta},\boldsymbol{z}}) \\
&= \sum_{\boldsymbol{z}\in\mathcal{Z}} \xi_a(\boldsymbol{z})(\boldsymbol{z}^\top(\boldsymbol{\theta}^\star-\boldsymbol{\theta}))^2 \\
&= (\boldsymbol{\theta}^\star-\boldsymbol{\theta})^\top\left(\sum_{\boldsymbol{z}\in\mathcal{Z}} \xi_a(\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^\top\right)(\boldsymbol{\theta}^\star-\boldsymbol{\theta}) \\
&= (\boldsymbol{\theta}^\star-\boldsymbol{\theta})^\top\boldsymbol{\Lambda}_a(\boldsymbol{\theta}^\star-\boldsymbol{\theta}).
\end{aligned}$$

Since $\tau = \sum_{a\mathcal{A}} t_a^{\text{on}}$, we have $\mathbb{E}_{\boldsymbol{\theta}^\star}[\tau] \geq \sum_{a\in\mathcal{A}} t_a$ for $t$ the solution to

$$\min_t \sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \quad \text{s.t.} \quad \min_{\boldsymbol{\theta}\in\Theta^\star_{\text{alt}}} \sum_{s=1}^{T_{\text{off}}} \text{KL}(\nu_{\boldsymbol{\theta}^\star,\boldsymbol{z}_s}, \nu_{\boldsymbol{\theta},\boldsymbol{z}_s}) + \sum_{a\in\mathcal{A}} t_a(\boldsymbol{\theta}^\star-\boldsymbol{\theta})^\top\boldsymbol{\Lambda}_a(\boldsymbol{\theta}^\star-\boldsymbol{\theta}) \geq \log\frac{1}{2.4\delta}. \tag{F.1}$$

Fix $\epsilon > 0$ and $t$ and, for $\boldsymbol{x}\in\mathcal{X}$, $\boldsymbol{x}\neq\boldsymbol{x}^\star(\boldsymbol{\theta}^\star)$, let

$$\boldsymbol{\theta}_{\boldsymbol{x}}(\epsilon,t) = \boldsymbol{\theta}^\star - \frac{(\boldsymbol{u}_{\boldsymbol{x}}^\top\boldsymbol{\theta}^\star+\epsilon)\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}}}{\boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}}}$$

for $\boldsymbol{u}_{\boldsymbol{x}} = \boldsymbol{x}^\star(\boldsymbol{\theta}^\star)-\boldsymbol{x}$. Note that $\boldsymbol{u}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_{\boldsymbol{x}}(\epsilon,t) = -\epsilon < 0$ so $\boldsymbol{\theta}_{\boldsymbol{x}}(\epsilon,t)\in\Theta^\star_{\text{alt}}$. Furthermore, we have

$$\text{KL}(\nu_{\boldsymbol{\theta}^\star,\boldsymbol{z}}, \nu_{\boldsymbol{\theta}_{\boldsymbol{x}}(\epsilon,t),\boldsymbol{z}}) = \boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\frac{(\boldsymbol{u}_{\boldsymbol{x}}^\top\boldsymbol{\theta}^\star+\epsilon)^2\boldsymbol{z}\boldsymbol{z}^\top}{(\boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}})^2}\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}}$$

and

$$(\boldsymbol{\theta}^\star-\boldsymbol{\theta})^\top\boldsymbol{\Lambda}_a(\boldsymbol{\theta}^\star-\boldsymbol{\theta}) = \boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\frac{(\boldsymbol{u}_{\boldsymbol{x}}^\top\boldsymbol{\theta}^\star+\epsilon)^2\boldsymbol{\Lambda}_a}{(\boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}})^2}\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}}.$$

It follows that

$$\begin{aligned}
\text{(F.1)} &\geq \min_t \sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \quad \text{s.t.} \quad \min_{\boldsymbol{\theta}\in\Theta^\star_{\text{alt}}} \frac{(\boldsymbol{u}_{\boldsymbol{x}}^\top\boldsymbol{\theta}^\star+\epsilon)^2}{(\boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}})^2}\cdot\boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\Big(\sum_{s=1}^{T_{\text{off}}}\boldsymbol{z}_s\boldsymbol{z}_s^\top + \sum_{a\in\mathcal{A}} t_a\boldsymbol{\Lambda}_a\Big)\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}} \geq \log\frac{1}{2.4\delta} \\
&= \min_t \sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \quad \text{s.t.} \quad \min_{\boldsymbol{\theta}\in\Theta^\star_{\text{alt}}} \frac{(\boldsymbol{u}_{\boldsymbol{x}}^\top\boldsymbol{\theta}^\star+\epsilon)^2}{\boldsymbol{u}_{\boldsymbol{x}}^\top\mathbf{A}(t)^{-1}\boldsymbol{u}_{\boldsymbol{x}}} \geq \log\frac{1}{2.4\delta}.
\end{aligned}$$

Taking $\epsilon\to 0$ and rearranging proves the result. □

### F.1.2. LINEAR MDPS

To obtain a lower bound for linear MDPs, we show a reduction from linear MDPs to linear bandits with random arms. In the following we will denote $\Pi_{\mathrm{exp}}$ the set of all possible exploration policies (this could include any possible policy, though in practice it suffices to just take a policy cover). We will also let $\Pi_{\mathsf{det}}$ denote the set of all deterministic policies.

Consider the following linear bandit with random arms construction:

- Choose some set of transition kernels $\{P_h\}_{h=1}^H$ on an MDP with $|\mathcal{S}| =: S < \infty$ states and $|\mathcal{A}| =: A < \infty$ actions.
- Choose some set of feature vectors $\boldsymbol{\phi}(s, a)$ defined on $\mathcal{S} \times \mathcal{A}$.
- Let the set of actions in the linear bandit be $\Pi_{\mathrm{exp}}$, the set of all possible policies on the MDP constructed above.
- For some $h$, let $\xi_\pi$ denote the distribution of $\boldsymbol{\phi}(s_h, a_h)$ under policy $\pi \in \Pi_{\mathrm{exp}}$.
- Set $\boldsymbol{\theta}^\star$ as desired, and $\mathcal{X} = \{\boldsymbol{\phi}_{\pi,h} \ : \ \pi \in \Pi_{\mathsf{det}}\}$.

We furthermore assume that there is a unique optimal policy in $\Pi_{\mathsf{det}}$. This is a well-specified linear bandit with random arms, so the lower bound from Proposition 4 applies to finding the best decision on $\mathcal{X}$. To make this reduction precise, we need the following result.

**Lemma F.1.** *For any (possibly random) policy $\widetilde{\pi}$ and fixed $h$, there exists some $p \in \triangle_{\Pi_{\mathsf{det}}}$ such that $\sum_{\pi \in \Pi_{\mathsf{det}}} p_\pi \boldsymbol{\phi}_{\pi,h} = \boldsymbol{\phi}_{\widetilde{\pi},h}$.*

*Proof.* First, note that for any policy $\pi$, we can write

$$w_{h+1}^\pi = \pi_{h+1} P_h w_h^\pi$$

for $w_h^\pi \in \mathbb{R}^{SA}$ the visitation probabilities, $P_h \in \mathbb{R}^{S \times SA}$ the transition matrix at step $h$, and $\pi_{h+1} \in \mathbb{R}^{SA \times S}$ the policy's probabilities, where $[\pi_{h+1}]_{(sa),s} = \pi_{h+1}(a|s)$ and 0 otherwise.

Fix some policy $\pi$. Let us assume that $w_h^\pi = \sum_{\widetilde{\pi} \in \Pi_{\mathsf{det}}} p_{\widetilde{\pi}}^h w_h^{\widetilde{\pi}}$ for some $p^h \in \triangle_{\Pi_{\mathsf{det}}}$. Note that there exists some $\widetilde{p} \in \triangle_{\Pi_{\mathsf{det}}}$ such that $\pi_{h+1} = \sum_{\pi' \in \Pi_{\mathsf{det}}} \widetilde{p}_{\pi'} \pi'_{h+1}$. Then we have

$$w_{h+1}^\pi = \pi_{h+1} P_h w_h = \sum_{\pi' \in \Pi_{\mathsf{det}}} \widetilde{p}_{\pi'} \pi'_h \cdot P_h \sum_{\widetilde{\pi} \in \Pi_{\mathsf{det}}} p_{\widetilde{\pi}}^h w_h^{\widetilde{\pi}} = \sum_{\pi', \widetilde{\pi} \in \Pi_{\mathsf{det}}} \widetilde{p}_{\pi'} p_{\widetilde{\pi}}^h \pi'_h P_h w_h^{\widetilde{\pi}}.$$

Note that $\pi'_{h+1} P_h w_h^{\widetilde{\pi}} = w_{h+1}^{\pi''}$ for some $\pi'' \in \Pi_{\mathsf{det}}$ (namely the concatenation of $\widetilde{\pi}$ up to step $h$ with $\pi'$ at step $h-1$). It follows that the above can be written as

$$\sum_{\pi', \widetilde{\pi} \in \Pi_{\mathsf{det}}} \widetilde{p}_{\pi'} p_{\widetilde{\pi}}^h w_{h+1}^{\pi''(\widetilde{\pi}, \pi')}$$

where $\pi''(\widetilde{\pi}, \pi')$ denotes the concatenation of $\widetilde{\pi}$ and $\pi'$ given above. However, this itself can be written in terms of some $p'$ as $\sum_{\pi'' \in \Pi_{\mathsf{det}}} p'_{\pi''} w_{h+1}^{\pi''}$, which proves the inductive hypothesis.

The result then follows since $\boldsymbol{\phi}_{\pi,h}$ is completely specified by the visitation probabilities of policy $\pi$. □

Assume we have defined a linear bandit with random actions as above and that we have access to some linear MDP algorithm that will return an $\epsilon$-optimal policy with probability at least $1 - \delta$. (possibly one that uses knowledge of the dynamics). Consider running our linear bandit on the linear MDP with transitions $\{P_h\}_{h=1}^H$ and rewards $\boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_{h-1} = \boldsymbol{\theta}_{h+1} = \ldots = \boldsymbol{\theta}_H = \mathbf{0}$ and $\boldsymbol{\theta}_h = \boldsymbol{\theta}^\star$, for reward distribution $r_h(s, a) \sim \mathcal{N}(\boldsymbol{\phi}(s, a)^\top \boldsymbol{\theta}^\star, 1)$[1]. Note that this linear MDP can be completely simulated by running our linear bandit with random actions. Assume that our linear MDP algorithm returns some policy $\widehat{\pi}$ that is $\epsilon$-optimal. Consider the following procedure:

1. Find weights $p \in \Pi_{\mathsf{det}}$ such that $\boldsymbol{\phi}_{\widehat{\pi},h} = \sum_{\pi \in \Pi_{\mathsf{det}}} p_\pi \boldsymbol{\phi}_{\pi,h}$ (note that this is possible by Lemma F.1 and since we assume the dynamics are known).
2. Set $\widetilde{\pi} = \arg\max_{\pi \in \Pi_{\mathsf{det}}} p_\pi$.

[1] Note that allowing rewards to be normally distributed violates the linear MDP definition as rewards could now fall outside $[0, 1]$. This is done only for simplicity—at the expense of a more complicated calculation, all results in this section can be shown to hold for slightly different constants if the noise is instead Bernoulli (see e.g. Lemma E.1 of (Wagenmaker & Jamieson, 2022) for an example).

We have the following result.

**Lemma F.2.** *Assume that $\epsilon < \Delta_{\min}/2$ for $\Delta_{\min} = \boldsymbol{\phi}_{\pi^\star,h}^\top \boldsymbol{\theta}^\star - \max_{\pi \in \Pi_{\mathsf{det}}, \pi \neq \pi^\star} \boldsymbol{\phi}_{\pi,h}^\top \boldsymbol{\theta}^\star$ and $\pi^\star = \arg\max_{\pi \in \Pi} \boldsymbol{\phi}_{\pi,h}^\top \boldsymbol{\theta}^\star$. Then $\widetilde{\pi} = \pi^\star$ as long as $\widehat{\pi}$ is $\epsilon$-optimal.*

*Proof.* Let $V^\star = \boldsymbol{\phi}_{\pi^\star,h}^\top \boldsymbol{\theta}^\star$ denote the value of the optimal policy. Note that

$$\sum_{\pi \in \Pi_{\mathsf{det}}} p_\pi \boldsymbol{\phi}_{\pi,h}^\top \boldsymbol{\theta}^\star = \boldsymbol{\phi}_{\widehat{\pi},h}^\top \boldsymbol{\theta}^\star \geq V^\star - \epsilon.$$

However, we also have that for any $\pi \neq \pi^\star$, $\boldsymbol{\phi}_{\pi,h}^\top \boldsymbol{\theta}^\star \leq V^\star - \Delta_{\min}$, so it follows that

$$\sum_{\pi \in \Pi_{\mathsf{det}}} p_\pi \boldsymbol{\phi}_{\pi,h}^\top \boldsymbol{\theta}^\star \leq p_{\pi^\star} V^\star + (1 - p_{\pi^\star})(V^\star - \Delta_{\min}) = V^\star - (1 - p_{\pi^\star})\Delta_{\min}.$$

Putting these together we have

$$V^\star - (1 - p_{\pi^\star})\Delta_{\min} \geq V^\star - \epsilon \implies p_{\pi^\star} \geq 1 - \frac{\epsilon}{\Delta_{\min}}.$$

Thus, it follows that if $\epsilon < \Delta_{\min}/2$, then $p_{\pi^\star} > 1/2$, which implies that $\widetilde{\pi} = \pi^\star$. □

**Lemma F.3.** *Consider the linear MDP setting outlined above, and assume there is a unique optimal policy in $\Pi_{\mathsf{det}}$. For $\epsilon < \Delta_{\min}/2$, any $(\epsilon, \delta)$-PAC algorithm given access to some dataset $\mathfrak{D}_{\mathrm{off}}$ must have*

$$\mathbb{E}[\tau] \geq \min_t \sum_{\pi \in \Pi_{\exp}} t_\pi \quad \textit{s.t.} \quad \max_{\pi \neq \pi^\star, \pi \in \Pi_{\mathsf{det}}} \frac{\|\boldsymbol{\phi}_{\pi,h} - \boldsymbol{\phi}_{\pi^\star,h}\|_{\mathbf{A}(t)^{-1}}^2}{((\boldsymbol{\phi}_{\pi^\star,h} - \boldsymbol{\phi}_{\pi,h})^\top \boldsymbol{\theta}_h)^2} \leq \frac{1}{\log 1/2.4\delta}$$

*where $\pi^\star = \arg\max_{\pi \in \Pi} V_0^\pi$ and $\mathbf{A}(t) = \sum_{\pi \in \Pi_{\exp}} t_\pi \boldsymbol{\Lambda}_{\pi,h} + \sum_{n=1}^{T_{\mathrm{off}}} \boldsymbol{\phi}(s_h^n, a_h^n)\boldsymbol{\phi}(s_h^n, a_h^n)^\top$.*

*Proof.* By Lemma F.2, we can simulate a linear MDP with a linear bandit with random arms, and use any $(\epsilon, \delta)$-PAC algorithm to identify the optimal arm in the linear bandit with probability at least $1 - \delta$, as long as $\epsilon < \Delta_{\min}/2$. It follows that the lower bound from Proposition 4 applies with action set $\mathcal{A} = \Pi_{\exp}$, and $\mathcal{X}$ the set $\{\boldsymbol{\phi}_{\pi,h}\}_{\pi \in \Pi_{\mathsf{det}}}$. □

#### F.1.3. Removing Differences

The upper bound we obtain scales as $\|\boldsymbol{\phi}_{\pi,h}\|_{\mathbf{A}(t)^{-1}}^2$ instead of $\|\boldsymbol{\phi}_{\pi,h} - \boldsymbol{\phi}_{\pi^\star,h}\|_{\mathbf{A}(t)^{-1}}^2$. The following result shows that we can construct linear MDPs where this difference is not too significant.

**Lemma F.4.** *Let*

$$t^\star = \arg\min_t \sum_{\pi \in \Pi_{\exp}} t_\pi \quad \textit{s.t.} \quad \max_{\pi \neq \pi^\star, \pi \in \Pi_{\mathsf{det}}} \frac{\|\boldsymbol{\phi}_{\pi,h} - \boldsymbol{\phi}_{\pi^\star,h}\|_{\mathbf{A}(t)^{-1}}^2}{((\boldsymbol{\phi}_{\pi^\star,h} - \boldsymbol{\phi}_{\pi,h})^\top \boldsymbol{\theta}_h)^2} \leq \frac{1}{\log 1/2.4\delta}.$$

*Assume that every state in level $h$ is reached with probability at least $p$ by every policy. Then there exists settings of the feature vectors such that*

$$(1 + \frac{1}{p^2})\|t^\star\|_1 \geq \min_t \sum_{\pi \in \Pi_{\exp}} t_\pi \quad \textit{s.t.} \quad \max_{\pi \neq \pi^\star, \pi \in \Pi_{\mathsf{det}}} \frac{\|\boldsymbol{\phi}_{\pi,h}\|_{\mathbf{A}(t)^{-1}}^2}{((\boldsymbol{\phi}_{\pi^\star,h} - \boldsymbol{\phi}_{\pi,h})^\top \boldsymbol{\theta}_h)^2} \leq \frac{1}{\log 1/2.4\delta}.$$

*Proof.* Consider, for example, the case where for each $s$, $\boldsymbol{\phi}(s, \pi_h^\star(s)) = \boldsymbol{e}_1$, and for every $a \neq \pi_h^\star(s)$, $\boldsymbol{\phi}(s,a)^\top \boldsymbol{e}_1 = 0$. In this case, $\mathbf{A}(t)$ will have the form $a\boldsymbol{e}_1\boldsymbol{e}_1^\top + \mathbf{B}(t)$, for some $a$ and $\mathbf{B}(t)$ whose first column and row are entirely 0. It follows that $\boldsymbol{\phi}_{\pi^\star,h} = \boldsymbol{e}_1$. Note that $|[\boldsymbol{\phi}_{\pi,h}]_1| \leq 1$ for any other $\pi$ as well.

Assume that every reachable state at level $h$ is reached with probability at least $p$ for every roll-in policy. Then we have that $|[\boldsymbol{\phi}_{\pi^\star,h}]_1 - [\boldsymbol{\phi}_{\pi,h}]_1| \geq p$ for every $\pi \in \Pi_{\mathsf{det}}$. We have

$$\|\boldsymbol{\phi}_{\pi,h} - \boldsymbol{\phi}_{\pi^\star,h}\|_{\mathbf{A}(t^\star)^{-1}}^2 = \frac{([\boldsymbol{\phi}_{\pi,h}]_1 - [\boldsymbol{\phi}_{\pi^\star,h}])^2}{a} + \|\widetilde{\boldsymbol{\phi}}_{\pi,h}\|_{\mathbf{B}(t^\star)^{-1}}^2, \quad \|\boldsymbol{\phi}_{\pi,h}\|_{\mathbf{A}(t^\star)^{-1}}^2 = \frac{([\boldsymbol{\phi}_{\pi,h}]_1)^2}{a} + \|\widetilde{\boldsymbol{\phi}}_{\pi,h}\|_{\mathbf{B}(t^\star)^{-1}}^2.$$

Consider playing $t^\star + t$ for some $t$ such that every sample goes to $\boldsymbol{e}_1$. In other words, $\mathbf{A}(t^\star + t) = \mathbf{A}(t^\star) + t\boldsymbol{e}_1\boldsymbol{e}_1^\top$. Then if we take $t \geq \frac{a}{p^2} - a$, we have

$$\begin{aligned}
\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{A}(t^\star+t)^{-1}} &= \frac{([\boldsymbol{\phi}_{\pi,h}]_1)^2}{a+t} + \|\widetilde{\boldsymbol{\phi}}_{\pi,h}\|^2_{\mathbf{B}(t^\star)^{-1}} \\
&\leq \frac{1}{a+t} + \|\widetilde{\boldsymbol{\phi}}_{\pi,h}\|^2_{\mathbf{B}(t^\star)^{-1}} \\
&\leq \frac{p}{a} + \|\widetilde{\boldsymbol{\phi}}_{\pi,h}\|^2_{\mathbf{B}(t^\star)^{-1}} \\
&\leq \frac{([\boldsymbol{\phi}_{\pi,h}]_1 - [\boldsymbol{\phi}_{\pi^\star,h}])^2}{a} + \|\widetilde{\boldsymbol{\phi}}_{\pi,h}\|^2_{\mathbf{B}(t^\star)^{-1}}.
\end{aligned}$$

It follows that $t^\star + t$ is a feasible solution. Since $\|t^\star\|_1 \geq a$, the result follows.

□

*Proof of Theorem 5.* This follows directly from Lemma F.3 and Lemma F.4. To make the construction explicit, we take $\boldsymbol{\theta}^\star = \boldsymbol{e}_1$, and construct the feature vectors as in Lemma F.4. The transition kernel can then be any transition kernel such that, regardless of the policy we play, we end up in each state $s \in \mathcal{S}$ at step $h$ with probability $\Omega(1)$. □

### F.2. Minimax Lower Bound (Theorem 1)

Consider the following *multi-dimensional linear bandit* setting. Let $\boldsymbol{\theta}_h \in \Theta := \{-\mu, \mu\}^d$ for some $\mu$ to be chosen, and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H]$. At step $t$, we can query points $\boldsymbol{\phi}_{ht} \in \mathcal{S}^{d-1}$ for each $h \in [H]$, and observe

$$y_{ht} \sim \text{Bernoulli}(\langle \boldsymbol{\phi}_{ht}, \boldsymbol{\theta}_h \rangle + 1/2).$$

Our goal will then be to find some $(\widehat{\boldsymbol{\phi}}_h)_{h=1}^H$, $\widehat{\boldsymbol{\phi}}_h \in \mathcal{S}^{d-1}$, which is $\epsilon$-optimal in the sense that

$$\sum_{h=1}^H \langle \widehat{\boldsymbol{\phi}}_h, \boldsymbol{\theta}_h \rangle \geq \sup_{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_H \in \mathcal{S}^{d-1}} \sum_{h=1}^H \langle \boldsymbol{\phi}_h, \boldsymbol{\theta}_h \rangle - \epsilon = \sum_{h=1}^H \|\boldsymbol{\theta}_h\|_2 - \epsilon.$$

We start with the follow regression lower bound.

**Lemma F.5.** *Assume that $\mu \in (0, \frac{1}{20\sqrt{d}}]$ and consider the multi-dimensional linear bandit setting outlined above. Let $\widehat{\boldsymbol{\theta}}$ be some estimator of $\boldsymbol{\theta}$, and $\pi$ some query strategy. Then, if we query for $T$ steps, we have*

$$\inf_{\widehat{\boldsymbol{\theta}},\pi} \sup_{\boldsymbol{\theta} \in \Theta^H} \mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2] \geq \frac{dH\mu^2}{2}\left(1 - \sqrt{\frac{20T\mu^2}{d}}\right).$$

*Proof.* The proof of this result follows closely the proof of Theorem 5 of (Wagenmaker et al., 2022b), which is itself based on the proof of Theorem 3 of (Shamir, 2013).

We have

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in \Theta^H} \mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2] &= \sup_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_H \in \Theta} \sum_{h=1}^H \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^d (\boldsymbol{\theta}_{hi} - \widehat{\boldsymbol{\theta}}_{hi})^2\right] \\
&\geq \mathbb{E}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_H \sim \text{unif}(\Theta)} \sum_{h=1}^H \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^d (\boldsymbol{\theta}_{hi} - \widehat{\boldsymbol{\theta}}_{hi})^2\right] \\
&\geq \sum_{h=1}^H \mathbb{E}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_H \sim \text{unif}(\Theta)} \mathbb{E}_{\boldsymbol{\theta}}\left[\mu^2 \sum_{i=1}^d \mathbb{I}\{\boldsymbol{\theta}_{hi}\widehat{\boldsymbol{\theta}}_{hi} < 0\}\right].
\end{aligned}$$

**Lemma F.1** (Lemma 4 of (Shamir, 2013))**.** *Let $\boldsymbol{\theta}_h$ be a random vector, none of whose coordinates is supported on 0, and let $(r_1^t, \ldots, r_H^t)_{t=1}^T$ be a sequence of observations obtained by a query strategy where $\boldsymbol{\phi}_h^t$ is a deterministic function*

*of* $\mathcal{F}_{h,t} := \{(r_1^s, \ldots, r_H^s)_{s=1}^{t-1}, (\phi_1^s, \ldots, \phi_H^s)_{s=1}^{t-1}, (r_1^t, \ldots, r_{h-1}^t), (\phi_1^t, \ldots, \phi_{h-1}^t)\}$. *Let* $\widehat{\boldsymbol{\theta}}_h$ *be some estimator that is a deterministic function of* $(r_1^t, \ldots, r_H^t)_{t=1}^T$ *and* $(\phi_1^t, \ldots, \phi_H^t)_{t=1}^T$. *Then we have*

$$\mathbb{E}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_H \sim \mathrm{unif}(\Theta)} \mathbb{E}_{\boldsymbol{\theta}} \left[ \sum_{i=1}^d \mathbb{I}\{\boldsymbol{\theta}_{hi} \widehat{\boldsymbol{\theta}}_{hi} < 0\} \right] \geq \frac{d}{2} \left( 1 - \sqrt{\frac{1}{d} \sum_{i=1}^d \sum_{t=1}^T U_{t,i}^h} \right)$$

*where*

$$U_{t,i}^h = \sup_{\boldsymbol{\theta}_{hj}, j \neq i; \boldsymbol{\theta}_{h'}, h' \neq h} \mathrm{KL}\Big( \mathbb{P}(r_h^t | \boldsymbol{\theta}_i > 0, \{\boldsymbol{\theta}_{hj}\}_{j \neq i}, \{\boldsymbol{\theta}_{h'}\}_{h' \neq h}, \mathcal{F}_{h,t}) || \mathbb{P}(r_h^t | \boldsymbol{\theta}_i < 0, \{\boldsymbol{\theta}_{hj}\}_{j \neq i}, \{\boldsymbol{\theta}_{h'}\}_{h' \neq h}, \mathcal{F}_{h,t}) \Big).$$

Since our rewards are Bernoulli, we have

$$U_{t,i}^h = \sup_{\boldsymbol{\theta}_{hj}, j \neq i} \mathrm{KL}\Big( \mathrm{Bernoulli}(1/2 + \sum_{j \neq i} \boldsymbol{\theta}_{hj} \phi_{hj}^t + \mu \phi_{hi}^t) || \mathrm{Bernoulli}(1/2 + \sum_{j \neq i} \boldsymbol{\theta}_{hj} \phi_{hj}^t - \mu \phi_{hi}^t) \Big).$$

Note that:

**Lemma F.2** (Lemma 2.7 of (Tsybakov, 2009))**.**

$$\mathrm{KL}(\mathrm{Bernoulli}(p) || \mathrm{Bernoulli}(q)) \leq \frac{(p-q)^2}{q(1-q)}.$$

Applying this, we have

$$U_{t,i}^h \leq \frac{(2\mu \phi_{hi}^t)^2}{(1/2 + \sum_{j \neq i} \boldsymbol{\theta}_{hj} \phi_{hj}^t - \mu \phi_{hi}^t)(1 - 1/2 - \sum_{j \neq i} \boldsymbol{\theta}_{hj} \phi_{hj}^t + \mu \phi_{hi}^t)} \leq 20 \mu^2 (\phi_{hi}^t)^2$$

where we make use of the fact that $\mu \leq \frac{1}{20\sqrt{d}}$ and $\phi_h^t \in \mathcal{S}^{d-1}$, which implies that $1/2 + \sum_{j \neq i} \boldsymbol{\theta}_{hj} \phi_{hj}^t - \mu \phi_{hi}^t \geq 9/20$ and $1 - 1/2 - \sum_{j \neq i} \boldsymbol{\theta}_{hj} \phi_{hj}^t + \mu \phi_{hi}^t \geq 9/20$.
We can then lower bound

$$\begin{aligned}
\sum_{h=1}^H \mathbb{E}_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_H \sim \mathrm{unif}(\Theta)} \mathbb{E}_{\boldsymbol{\theta}} \left[ \mu^2 \sum_{i=1}^d \mathbb{I}\{\boldsymbol{\theta}_{hi} \widehat{\boldsymbol{\theta}}_{hi} < 0\} \right] &\geq \sum_{h=1}^H \frac{d\mu^2}{2} \left( 1 - \sqrt{\frac{1}{d} \sum_{i=1}^d \sum_{t=1}^T U_{t,i}^h} \right) \\
&\geq \sum_{h=1}^H \frac{d\mu^2}{2} \left( 1 - \sqrt{\frac{1}{d} \sum_{i=1}^d \sum_{t=1}^T 20 \mu^2 (\phi_{hi}^t)^2} \right) \\
&= \frac{dH\mu^2}{2} \left( 1 - \sqrt{\frac{20 T \mu^2}{d}} \right)
\end{aligned}$$

where we have used that $\phi_{ht} \in \mathcal{S}^{d-1}$. This proves the result. □

**Lemma F.6.** *Assume that* $(\widehat{\phi}_h)_{h=1}^H$ *is* $\epsilon$*-optimal. Then,*

$$\sum_{h=1}^H \| \sqrt{d\mu^2} \cdot \widehat{\phi}_h - \boldsymbol{\theta}_h \|_2^2 \leq 2\sqrt{d\mu^2} \cdot \epsilon$$

*Proof.* Denote $\widehat{\phi} = [\widehat{\phi}_1, \ldots, \widehat{\phi}_H]$, so that $\sum_{h=1}^H \langle \widehat{\phi}_h, \boldsymbol{\theta}_h \rangle = \langle \widehat{\phi}, \boldsymbol{\theta} \rangle$. If $\widehat{\phi}$ is $\epsilon$-optimal, then this implies that

$$\langle \widehat{\phi}, \boldsymbol{\theta} \rangle \geq \sum_{h=1}^H \| \boldsymbol{\theta}_h \|_2 - \epsilon = \sqrt{H} \| \boldsymbol{\theta} \|_2 - \epsilon \tag{F.2}$$

where the last inequality follows since, by construction of $\boldsymbol{\theta}_h$, we have $\|\boldsymbol{\theta}_h\|_2 = \sqrt{d\mu^2}$ and $\|\boldsymbol{\theta}\|_2 = \sqrt{dH\mu^2}$. Now note that

$$\begin{aligned}
\|\frac{1}{\sqrt{H}}\|\boldsymbol{\theta}\|_2\widehat{\boldsymbol{\phi}} - \boldsymbol{\theta}\|_2^2 &= \frac{1}{H}\|\boldsymbol{\theta}\|_2^2\|\widehat{\boldsymbol{\phi}}\|_2^2 + \|\boldsymbol{\theta}\|_2^2 - \frac{2}{\sqrt{H}}\|\boldsymbol{\theta}\|_2 \cdot \langle\widehat{\boldsymbol{\phi}}, \boldsymbol{\theta}\rangle \\
&\overset{(a)}{=} 2\|\boldsymbol{\theta}\|_2^2 - \frac{2}{\sqrt{H}}\|\boldsymbol{\theta}\|_2 \cdot \langle\widehat{\boldsymbol{\phi}}, \boldsymbol{\theta}\rangle \\
&\overset{(b)}{\leq} \frac{2}{\sqrt{H}}\|\boldsymbol{\theta}\|_2 \cdot \epsilon \\
&\overset{(c)}{=} 2\sqrt{d\mu^2} \cdot \epsilon
\end{aligned}$$

where $(a)$ uses that $\|\widehat{\boldsymbol{\phi}}\|_2^2 = \sum_{h=1}^H \|\widehat{\boldsymbol{\phi}}_h\|_2^2 = H$, $(b)$ follows from (F.2), and $(c)$ uses $\|\boldsymbol{\theta}\|_2 = \sqrt{dH\mu^2}$. □

**Lemma F.7.** *Fix $\epsilon > 0$, $d > 1$, and $T \geq d^2$. Consider running some (possibly adaptive) algorithm which stops at some (possibly random) stopping time $\tau$ and outputs $(\widehat{\boldsymbol{\phi}}_h)_{h=1}^H$. Let $\mathcal{E}$ denote the event*

$$\mathcal{E} := \{\tau \leq T \text{ and } (\widehat{\boldsymbol{\phi}}_h)_{h=1}^H \text{ is } \epsilon\text{-optimal}\}.$$

*Then unless $T \geq c \cdot \frac{d^2H^2}{\epsilon^2}$, there exists some instance $\boldsymbol{\theta} \in \Theta^H$ such that $\mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}] \geq 1/10$.*

*Proof.* Set $\mu = \sqrt{d/700T}$ and let $\widehat{\boldsymbol{\theta}}_h = \sqrt{d\mu^2} \cdot \widehat{\boldsymbol{\phi}}_h$. Note that $\|\widehat{\boldsymbol{\theta}}\|_2 \leq d\sqrt{H/700T}$. Then

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2] &= \mathbb{E}_{\boldsymbol{\theta}}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \cdot \mathbb{I}\{\mathcal{E}\} + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \cdot \mathbb{I}\{\mathcal{E}^c\}] \\
&\leq 2\sqrt{\frac{d^2}{700T}} \cdot \epsilon + \frac{4d^2H}{700T} \cdot \mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c]
\end{aligned}$$

where the inequality follows from Lemma F.6 since on $\mathcal{E}$, $\widehat{\boldsymbol{\phi}}$ is $\epsilon$-optimal. However, by Lemma F.5, we know that there exists some $\boldsymbol{\theta}$ such that

$$\mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2] \geq \frac{dH\mu^2}{2}\left(1 - \sqrt{\frac{20T\mu^2}{d}}\right) \geq 0.00059 \cdot \frac{d^2H}{T}.$$

However, this is a contradiction unless

$$2\sqrt{\frac{d^2}{700T}} \cdot \epsilon + \frac{4d^2H}{700T} \cdot \mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \geq 0.00059 \cdot \frac{d^2H}{T} \iff \mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \geq 0.10325 - \frac{\sqrt{700T}}{2dH} \cdot \epsilon.$$

It follows that it

$$0.10325 - \frac{\sqrt{700T}}{2dH} \cdot \epsilon \geq 0.1 \iff (\frac{2 \cdot 0.00325}{\sqrt{700}})^2 \cdot \frac{d^2H^2}{\epsilon^2} \geq T$$

we have that $\mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \geq 0.1$. □

### F.2.1. Mapping to Linear MDPs

Consider a single-state, $H$-step, $d+1$-dimensional linear MDP specified by:

$$\begin{aligned}
&\boldsymbol{\phi}(s, \boldsymbol{a}) = [\boldsymbol{a}/2, 1/2], \quad \boldsymbol{a} \in \mathcal{S}^{d-1} \cup \{\mathbf{0}\} \\
&\widetilde{\boldsymbol{\theta}}_h = [\boldsymbol{\theta}_h, 1], \quad \boldsymbol{\theta}_h \in \Theta := \{-\mu, \mu\}^d \\
&\boldsymbol{\mu}_h(s) = [\mathbf{0}, 2].
\end{aligned}$$

We take the reward distribution to be $r_h(s, a) \sim \text{Bernoulli}(\langle\boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h\rangle)$. It is straightforward to see that this satisfies the definition of a linear MDP, as given in Definition 3.1.

Assume we have access to the regression setting defined above. Then we can simulate this linear MDP with the regression setting by simply, at episode $k$ step $h$, choosing some action $\boldsymbol{a}$, and using the regression setting to observe a reward $r \sim \text{Bernoulli}(\langle\boldsymbol{a}, \boldsymbol{\theta}_h\rangle + 1/2)$, and then transitioning to state $s$ and step $h+1$. The following lemma shows that a policy which is near-optimal on the linear MDP induces a near-optimal $\widehat{\boldsymbol{\phi}}$ in the regression setting.

**Lemma F.8.** *Assume that $\widehat{\pi}$ is $\epsilon$-optimal in the linear MDP defined above, with rewards $(\widetilde{\boldsymbol{\theta}}_h)_{h=1}^H$. Then $(\widehat{\boldsymbol{\phi}}_{\widehat{\pi},h})_{h=1}^H$ is $\epsilon$-optimal in the multi-dimensional linear bandit setting with rewards $(\boldsymbol{\theta}_h)_{h=1}^H$.*

*Proof.* Note that the value of any policy $\pi$ is given by

$$\sum_{h=1}^H \langle \boldsymbol{\phi}_{\pi,h}, \widetilde{\boldsymbol{\theta}}_h \rangle = H/2 + \frac{1}{2}\sum_{h=1}^H \langle \boldsymbol{a}_{\pi,h}, \boldsymbol{\theta}_h \rangle$$

where $\boldsymbol{a}_{\pi,h}$ denotes the first $d$ coordinates of $\boldsymbol{\phi}_{\pi,h}$. It follows that it $\widehat{\pi}$ is $\epsilon$-optimal, it must be the case that

$$\frac{1}{2}\sum_{h=1}^H \langle \boldsymbol{a}_{\widehat{\pi},h}, \boldsymbol{\theta}_h \rangle \geq \sup_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_H \in \mathcal{S}^{d-1}} \frac{1}{2}\sum_{h=1}^H \langle \boldsymbol{a}_h, \boldsymbol{\theta}_h \rangle - \epsilon.$$

However, this is precisely the definition of a $2\epsilon$-optimal policy in the multi-dimensional linear bandit setting with rewards $(\boldsymbol{\theta}_h)_{h=1}^H$. □

**Lemma F.9.** *Fix $\epsilon > 0$, $d > 1$, and $K \geq d^2$. Consider running some (possibly adaptive) algorithm which stops at some (possibly random) stopping time $\tau$ in a $(d+1)$-dimensional linear MDP, and outputs some policy $\widehat{\pi}$. Let $\mathcal{E}$ denote the event*

$$\mathcal{E} := \{\tau \leq K \text{ and } \widehat{\pi} \text{ is } \epsilon\text{-optimal}\}.$$

*Then unless $K \geq c \cdot \frac{d^2 H^2}{\epsilon^2}$, there exists some MDP $\mathcal{M}$ such that $\mathbb{P}_{\mathcal{M}}[\mathcal{E}] \geq 1/10$.*

*Proof.* As noted, we can simulate the linear MDP defined above using only access to a multi-dimensional linear bandit setting. Since Lemma F.8 implies that finding an $\epsilon$-optimal policy in our linear MDP is equivalent to find a $2\epsilon$-optimal set of vectors in our multi-dimensional linear bandit, the lower bound for multi-dimensional linear bandits, Lemma F.7, must hold here, which immediately gives the result. □

*Proof of Theorem 1.* Let $\mathfrak{D}_{\text{off}}$ be the dataset obtained from running $T_{\text{off}}$ times on the linear MDP defined above, and at each step $h$ playing each $[\boldsymbol{e}_i/2, 1/2]$ $T_{\text{off}}/2d$ times, and $\boldsymbol{e}_{d+1}/2$ $T_{\text{off}}/2d$ times. Note that in this case we have

$$\boldsymbol{\Lambda}_{\text{off}}^h = \frac{T_{\text{off}}}{8d}\begin{bmatrix} I & \mathbf{1} \\ \mathbf{1}^\top & 2d \end{bmatrix}.$$

Note that

$$\begin{aligned} N_h(\mathfrak{D}_{\text{off}}) &\leq \min_{N\geq 0} N \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h} \sup_{\pi} \|\boldsymbol{\phi}_{\pi,h}\|^2_{(N\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{\text{off}}^h)^{-1}} \leq \frac{c'\cdot\epsilon^2}{dH} \\ &\leq \min_{N\geq 0} N \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h} \sup_{\pi} \|\boldsymbol{\phi}_{\pi,h}\|_2^2 \|(N\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{\text{off}}^h)^{-1}\|_{\text{op}} \leq \frac{c'\cdot\epsilon^2}{dH} \\ &\leq \min_{N\geq 0} N \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h} \|(N\boldsymbol{\Lambda}+\boldsymbol{\Lambda}_{\text{off}}^h)^{-1}\|_{\text{op}} \leq \frac{c'\cdot\epsilon^2}{dH}. \end{aligned} \tag{F.3}$$

Now take

$$\boldsymbol{\Lambda} = \frac{1}{8d}\begin{bmatrix} I & \mathbf{1} \\ \mathbf{1}^\top & 2d \end{bmatrix},$$

(note that this is a valid setting of $\boldsymbol{\Lambda}$, and can be constructed analogously to $\mathfrak{D}_{\text{off}}$), and with this choice we can bound

$$\text{(F.3)} \leq \min_{N\geq 0} N \quad \text{s.t.} \quad \frac{8d}{N+T_{\text{off}}}\left\|\begin{bmatrix} I & \mathbf{1} \\ \mathbf{1}^\top & 2d \end{bmatrix}^{-1}\right\|_{\text{op}} \leq \frac{c'\cdot\epsilon^2}{dH}. \tag{F.4}$$

Using the formula for the inverse of a block matrix, we have

$$\begin{bmatrix} I & \mathbf{1} \\ \mathbf{1}^\top & 2d \end{bmatrix}^{-1} = \begin{bmatrix} I + \frac{1}{d}\mathbf{1}\mathbf{1}^\top & -\frac{1}{d}\mathbf{1} \\ -\frac{1}{d}\mathbf{1}^\top & \frac{1}{d} \end{bmatrix}.$$

Take $\boldsymbol{v} \in \mathcal{S}^{d-1}$, and write $\boldsymbol{v} = [\widetilde{\boldsymbol{v}}, b]$. Then,

$$\boldsymbol{v}^\top \begin{bmatrix} I + \frac{1}{d}\mathbf{1}\mathbf{1}^\top & -\frac{1}{d}\mathbf{1} \\ -\frac{1}{d}\mathbf{1}^\top & \frac{1}{d} \end{bmatrix} \boldsymbol{v} = \widetilde{\boldsymbol{v}}^\top \widetilde{\boldsymbol{v}} + \frac{1}{d}(\mathbf{1}^\top \widetilde{\boldsymbol{v}})^2 - \frac{2b}{d}\mathbf{1}^\top \widetilde{\boldsymbol{v}} + \frac{b^2}{d} \leq 5.$$

Using this, we can bound

$$(\text{F.4}) \leq \min_{N \geq 0} N \quad \text{s.t.} \quad \frac{40d}{N + T_{\text{off}}} \leq \frac{c' \cdot \epsilon^2}{dH} = \max\{\frac{40d^2 H}{c'\epsilon^2} - T_{\text{off}}, 0\}.$$

It follows that

$$\sum_{h=1}^{H} N_h(\mathfrak{D}_{\text{off}}) \leq H \max\{\frac{40d^2 H}{c'\epsilon^2} - T_{\text{off}}, 0\} = \max\{\frac{40d^2 H^2}{c'\epsilon^2} - HT_{\text{off}}, 0\} \leq \max\{\frac{40d^2 H^2}{c'\epsilon^2} - T_{\text{off}}, 0\}.$$

By Lemma F.9, we know that we need to collect at least $\frac{cd^2H^2}{\epsilon^2}$ episodes on some MDP in our class or we will fail to return an $\epsilon$-optimal policy with constant probability. Note that this lower bound is agnostic to how these episodes were collected—they could be either offline (which we can think of as equivalent to just running a non-adaptive query policy) or online. Thus, if we have already collected $T_{\text{off}}$ offline episodes, we must collect at least $\max\{\frac{cd^2H^2}{\epsilon^2} - T_{\text{off}}, 0\}$ online episodes. From what we have just shown, though,

$$\sum_{h=1}^{H} N_h(\mathfrak{D}_{\text{off}}) \leq \max\{\frac{40d^2 H^2}{c'\epsilon^2} - T_{\text{off}}, 0\},$$

so the result follows by Lemma F.9 and proper setting of $c'$. □