Regret Bounds for Stochastic Shortest Path Problems with Linear Function Approximation

Daniel Vial\textsuperscript{1}, Advait Parulekar\textsuperscript{2}, Sanjay Shakkottai\textsuperscript{2}, and R. Srikant\textsuperscript{1}

\textsuperscript{1}University of Illinois at Urbana-Champaign
\textsuperscript{2}University of Texas at Austin

October 20, 2021

Abstract

We propose an algorithm that uses linear function approximation (LFA) for stochastic shortest path (SSP). Under minimal assumptions, it obtains sublinear regret, is computationally efficient, and uses stationary policies. To our knowledge, this is the first such algorithm in the LFA literature (for SSP or other formulations). Our algorithm is a special case of a more general one, which achieves regret square root in the number of episodes given access to a certain computation oracle.

1 Introduction

To cope with the massive state spaces of modern reinforcement learning (RL) applications, a plethora of recent papers have studied function approximation. A particularly tractable case is linear function approximation (LFA). Here one assumes the transition kernel and cost vector are linear in known \(d\)-dimensional feature vectors, where typically \(d \ll S\) and \(A\) (the number of states and actions). In the online setting, an agent interacts with the Markov decision process over \(T\) time steps (for infinite horizon average and discounted cost problems) or \(K\) episodes (for finite horizon and stochastic shortest path problems). At a high level, one seeks algorithms with two properties:

- \textit{Statistically efficient}: regret independent of \(S\) and \(A\), sublinear (ideally, square root) in \(T\) or \(K\), and polynomial in \(d\) and any other parameters.
- \textit{Computationally efficient}: time and space complexity independent of \(S\) and polynomial in \(d\), \(A\), \(T\) or \(K\), and any other parameters.

For the finite horizon problem, several algorithms have been shown to achieve both properties. A key question we address is whether such algorithms exist in settings where stationary policies are optimal, i.e., for stochastic shortest path (SSP) and average/discounted cost problems. To our knowledge, this problem is (essentially) open: state-of-the-art algorithms are computationally inefficient for average/discounted costs, and none have been proposed for SSP. (There is one exception for average cost but it uses non-stationary policies.) See Section 1.1 for related work.

\*Corresponding author. Email: vial2@illinois.edu.
In addition to this theoretical point of interest, there is practical motivation for understanding stationary policy settings. First, stationary policies are simpler to deploy and (compared to large, but finite, horizons) less costly to store. Second, they do not require a notion of “time zero,” which may be ill-defined in practice. Third, RL applications like games with a random number of moves are best modeled in the stationary policy setting, in particular SSP, where the agent tries to minimize its cost before reaching a goal state.

**Contributions:** Motivated by these theoretical and practical concerns, we provide the first algorithm for episodic SSP with LFA. More generally, this is the first statistically and computationally efficient LFA algorithm that uses stationary policies (in any setting). In more detail, our contributions are as follows:

- **Optimistic approximate fixed points (OAFPs):** In Section 3, we show that under the LFA assumption, the optimal policy in an SSP can be computed from the fixed point of a $d$-dimensional Bellman operator, denoted by $G$ (see Proposition 1). This is a simple observation, but it leads to an important definition of OAFPs (see Definition 1). Roughly, these are $d$-dimensional vectors that have small Bellman error with respect to a data-driven operator $\hat{G}_t$ that we interpret as an optimistic approximation of $G$.

- **Regret bound with oracle:** In Section 4, we assume access to an oracle that computes OAFPs from trajectories and propose Algorithm 1, which uses the oracle to update its policy. When the LFA assumption holds, the minimal cost for non-goal states $c_{min}$ is positive, and a proper policy exists (see Assumptions 1-2), Theorem 1 shows Algorithm 1 achieves sublinear regret, with the exponent determined by the oracle’s quality ($\sqrt{K}$ in the best case – see Corollary 1). This reduces the problem of regret minimization to that of finding OAFPs (which exist with high probability, by the same theorem).

- **Oracle implementations:** In Section 5, we show how to compute OAFPs. Combined with the results of Section 4, this yields an efficient end-to-end algorithm with the following regret scaling in $K$:
  - $K^{5/6}$ if Assumptions 1-2 hold (Theorem 2).
  - $K^{3/4}$ if Assumptions 1-2 hold and all stationary policies are proper (Theorem 3).
  - $\sqrt{K}$ if Assumptions 1-2 hold and the features are orthogonal in a certain sense (Theorem 4).

- **Extensions:** In Section 6, we provide generalizations of Theorems 3 and 4 and remove the $c_{min} > 0$ assumption. The latter point shows we can obtain sublinear regret and computational efficiency with stationary policies under minimal SSP assumptions.

### 1.1 Related work

**Finite horizon LFA:** Several efficient algorithms have been proposed (of course, the policies are not stationary). To our knowledge, the earliest are Jin et al. (2020); Yang and Wang (2020); Zanette et al. (2020a), which (like us) assume linear costs and transitions: $c(s, a) = \phi(s, a)^T \theta$ and $P(s'|s, a) = \phi(s, a)^T \mu(s')$ for known $\phi(s, a) \in \mathbb{R}^d$. The most relevant is Jin et al. (2020), which proposed an optimistic, least squares version of backward induction; our algorithm is the value iteration analogue. Subsequent work is too vast to survey here, but for later discussion, we note Zhang et al. (2021); Zhou et al. (2021a) proposed Bernstein-style confidence sets for the related linear mixture model (see, e.g., Ayoub et al. (2020); Jia et al. (2020)), where $P(s'|s, a) = \varphi(s'|s, a)^T \theta$ for known $\varphi(s'|s, a)$.
**Infinite horizon LFA:** Comparatively little is known for infinite horizons. Wei et al. (2021); Wu et al. (2021) studied average costs under the minimal assumption that the optimal policy’s long-term average reward is independent of the initial state (see references therein for work with stronger assumptions). The first algorithm in Wei et al. (2021) has $\sqrt{T}$ regret assuming access to a certain fixed point oracle (analogous to our Algorithm 1) but no efficient oracle is provided, the second is computationally efficient with $T^{3/4}$ regret but uses non-stationary policies and requires knowledge of $T$ (unlike ours), and the third requires stronger assumptions. Wu et al. (2021) proved $\sqrt{T}$ regret for the linear mixture model, but the algorithm is inefficient due to computation\(^1\) of $\sum_{s' \in S} h(s')\varphi(s'|s, a)$ for certain $h \in \mathbb{R}^S$. Analogous algorithms are proposed in Zhou et al. (2020, 2021b) for discounted costs, which have $\sqrt{T}$ regret but are inefficient for the same reason. Also, the discounted cost regret formulation is a bit unsatisfying, as it compares to the optimal policy along the algorithm’s trajectory. Thus, one that stays in a bad set of states and only learns on this set can still have low regret. Finally, as mentioned above, we are not aware of any LFA papers that consider SSP.

**Tabular SSP, $c_{\min}^{-1}$ dependent:** Tarbouriech et al. (2020) proved $\hat{O}(D^{3/2}S\sqrt{AK}/c_{\min})$ regret, where $D$ is the SSP diameter (see their Assumption 2). Rosenberg et al. (2020) improved this to $\hat{O}(B^{3/2}S\sqrt{AK}/c_{\min})$, where $B_s \leq D$ is the maximal cost-to-go of the optimal policy. Both algorithms use Hoeffding-style confidence sets and can be generalized to the case $c_{\min} = 0$, though regret increases to $K^{2/3}$ (see Section 6).

**Tabular SSP, $c_{\min}^{-1}$ independent:** Rosenberg et al. (2020) also proved $\hat{O}(B_sS\sqrt{AK})$ regret when $c_{\min} = 0$, and the lower bound $\Omega(B_s\sqrt{SAK})$. Removing the $c_{\min}^{-1}$ dependence required Bernstein-style confidence sets, which Chen et al. (2021); Cohen et al. (2021); Tarbouriech et al. (2021); Jafarnia-Jahromi et al. (2021) also employed. The former three showed UCB-based algorithms achieve the lower bound; the latter showed posterior sampling obtains $\hat{O}(B_sS\sqrt{AK})$ regret. See references therein for prior work on SSP variants (e.g., adversarially changing costs).

## 2 Preliminaries

**Notation:** For $m \in \mathbb{N}$, we let $[m] = \{1, \ldots, m\}$. We write $\mathbbm{1}(\cdot)$ for the indicator function. We let $e_i$ be the vector with $j$-th element $e_i(j) = \mathbbm{1}(i = j)$. For $x \in \mathbb{R}^d$ and positive definite $Y \in \mathbb{R}^{d \times d}$, $\|x\|_Y = \sqrt{x^T Y x}$.

**SSP:** An SSP instance is defined by $(S, A, P, c, s_{goal})$, where $S$ is a set of $S = |S| < \infty$ states, $A$ is a set of $A = |A| < \infty$ actions, $P$ is the transition kernel, $c$ is the cost vector, and $s_{goal} \in S$ is an absorbing zero-cost state, i.e., $P(s_{goal}|s_{goal}, a) = 1$ and $c(s_{goal}, a) = 0$ for any $a \in A$. Any stationary and deterministic policy $\pi: S \rightarrow A$ induces a trajectory $\{s_{t+1}^\pi, \pi(s_{t+1}^\pi)\}_{t=1}^\infty$, where $s_{t+1}^\pi$ is some initial state and $s_{t+1}^\pi \sim P(\cdot|s_t^\pi, \pi(s_t^\pi))$ for $t \in \mathbb{N}$. We call $\pi$ proper if $s_{goal}$ is reached with probability 1 from any $s_t^\pi \in S$; otherwise, we call it improper. We make the following assumption regarding proper policies, which we discuss in Remark 1 below.

**Assumption 1** (Basic properties). There exists at least one proper policy, and for some $c_{\min} > 0$ and any $(s, a) \in (S \setminus \{s_{goal}\}) \times A$, $c(s, a) \in [c_{\min}, 1]$.

\(^1\)(Zhou et al., 2021b, Appendix B) provides a scheme to estimate the sums, but only in some special cases, and the estimation error is not accounted for in the regret analysis.
For any $\pi : S \rightarrow A$, we define the (possibly infinite) cost-to-go function $J^\pi : S \rightarrow \mathbb{R}$ by

$$J^\pi(s) = \lim_{T \to \infty} \mathbb{E} \left[ \sum_{t=1}^{T} c(s_t^\pi, \pi(s_t^\pi)) \right] ,$$

Given Assumption 1, the optimal policy $\pi^*$, i.e., the $\pi$ that minimizes $J^\pi(s)$ over all $s$, is stationary, deterministic, and proper (Bertsekas and Tsitsiklis, 1991). It also satisfies the Bellman optimality equations

$$J^\pi(s) = \min_{a \in A} Q^\pi(s, a), \quad \pi^*(s) \in \arg\min_{a \in A} Q^\pi(s, a), \quad (1)$$

where $J^* = J^{\pi^*}$ and the optimal state-action cost-to-go function $Q^* : S \times A \rightarrow \mathbb{R}$ is given by

$$Q^*(s, a) = c(s, a) + \sum_{s' \in S} J^*(s') P(s' | s, a). \quad (2)$$

Finally, we define $B_* = \max_{s \in S} J^*(s)$.

**Remark 1** (Positive costs). We require $c(s, a) \geq c_{\text{min}}$ to show that episodes incurring finite total cost must terminate in finite time. In Section 6, we remove this assumption while still achieving sublinear regret and computational efficiency with stationary policies.

**Linearity**: As discussed in the introduction, we make the following assumption to enable LFA.

**Assumption 2** (Linearity). For some $d \geq 2$, there exists known $\{\phi(s, a)\}_{(s, a) \in S \times A} \subseteq \mathbb{R}^d$, unknown $\theta \in \mathbb{R}^d$, and unknown $\{\mu(s')\}_{s' \in S} \subseteq \mathbb{R}^d$, such that, for any $(s, a, s') \in (S \setminus \{s_{\text{goal}}\}) \times A \times S$,

$$c(s, a) = \phi(s, a)^T \theta, \quad P(s' | s, a) = \phi(s, a)^T \mu(s'), \quad (3)$$

$$\|\phi(s, a)\|_2 \leq 1, \quad \|\theta\|_2 \leq \sqrt{d}, \quad (4)$$

$$\left\| \sum_{s' \in S} h(s') \mu(s') \right\|_2 \leq \sqrt{d} \|h\|_{\infty} \quad \forall \; h \in \mathbb{R}^S. \quad (5)$$

This assumption naturally generalizes that of Jin et al. (2020) to SSP. We also assume $d \geq 2$, which, given (3), only eliminates a trivial case where $\phi(s, a)$ is independent of $(s, a)$. Finally, we assume without further loss of generality that $\phi(s_{\text{goal}}, a) = 0 \quad \forall \; a \in A$.

**Remark 2** (Tabular case). Any SSP with $c(s, a) \in [0, 1]$ satisfies Assumption 2 with $d = SA$, $\phi(s, a) = c(s, a)\theta$, $\pi^*(s) = \text{arg\,min}_{a \in A} c(s, a)\mu(s)$.

**Remark 3** (Realizability). As shown in Appendix E.2, Assumption 2 implies $Q^*$ is linear in the features. Ideally, we would only assume this, but recent work for finite horizons (a special case of SSP) has shown this problem is fundamentally harder (Du et al., 2020; Wang et al., 2021; Weisz et al., 2021).

**Regret**: We consider a protocol with $K$ episodes. For each $k \in [K]$, the agent begins at step $h = 1$ at initial state $s_h^k$. At step $h$, the agent takes action $a_h^k$, incurs cost $c(s_h^k, a_h^k)$, and transitions to $s_{h+1}^k \sim P(\cdot | s_h^k, a_h^k)$. If $s_{h+1}^k = s_{\text{goal}}$, the episode terminates (without taking action $a_{h+1}^k$). We assume $s_h^k \neq s_{\text{goal}}$ without loss of generality but make no further assumptions on the sequence of initial states $\{s_h^k\}_{k=1}^K$. We let $(s_t, a_t, s'_t)$ denote the $t$-th state-action-state triple observed across all episodes. Hence, for each $t$, $s'_t \sim P(\cdot | s_t, a_t)$, and $s'_t = s_{t+1}$ unless an episode ends at time $t$ (in which case $s'_t = s_{\text{goal}}$ and $s_{t+1} = s_1^{k+1}$, where $k$ is the episode that ended at $t$). We also let $T$
denote the random total number of steps across all \( K \) episodes.\(^2\) As in the tabular SSP literature, we define the regret
\[
R(K) = \sum_{t=1}^{T} c(s_t, a_t) - \sum_{k=1}^{K} J^*(s^k_t),
\]
which is the difference between the total cost of the agent and the expected total cost of a “genie” who knows the optimal policy \emph{a priori} and runs it for \( K \) episodes with the same initial states.

\textbf{Remark 4} (Challenge 1). Unlike finite horizon LFA, no episode is guaranteed to end, since the agent may use improper policies. In this case, \( T = \infty \) and we suffer infinite regret. Thus, we will need to detect improper policies and fix them within episodes, a challenge that does not arise for finite horizon LFA.

\section{Optimistic approximate fixed point}

To motivate the definition of OAFPs, we begin with the simple observation that for a linear SSP, the optimal policy can be computed from a feature space version of the Bellman operator (when the model is known). The proof is elementary; see Appendix E.2.

\textbf{Proposition 1} (Feature space fixed point). Let Assumptions 1 and 2 hold. Define \( G : \mathbb{R}^d \to \mathbb{R}^d \) by
\[
Gw = \theta + \sum_{s \in S} \min_{a \in A} \phi(s, a)^T w \mu(s) \quad \forall \ w \in \mathbb{R}^d.
\]
Then \( w^* = \theta + \sum_{s \in S} J^*(s) \mu(s) \) is a fixed point of \( G \) (i.e., \( Gw^* = w^* \)), \( J^*(s) = \min_{a \in A} \phi(s, a)^T w^* \), and
\[
\pi^*(s) \in \arg \min_{a \in A} \phi(s, a)^T w^* \quad \forall \ s \in S.
\]

When the model is unknown, we instead must estimate \( Gw \) from data. Formally, let \( \{ s_\tau, a_\tau, s'_\tau \}_{\tau=1}^t \) denote the first \( t \) state-action-state triples as in Section 2, and define \( \Lambda_t = I + \sum_{\tau=1}^{t} \phi(s_\tau, a_\tau) \phi(s_\tau, a_\tau)^T \).
Then the regularized least-squares estimate of \( Gw \) is
\[
\tilde{G}w = \Lambda_t^{-1} \sum_{\tau=1}^{t} \phi(s_\tau, a_\tau) \left( c(s_\tau, a_\tau) + \min_{a \in A} \phi(s'_\tau, a)^T w \right).
\]
Due to Assumption 2 and concentration, we should expect \( \tilde{G}w \approx Gw \) for any (bounded) \( w \). Thus, it seems reasonable to find a fixed point \( \tilde{w}^* \) of \( \tilde{G} \) and define policies like (6), with \( w^* \) replaced by \( \tilde{w}^* \).

This is roughly our approach, though we will modify \( \tilde{G} \) in two ways. First, as is common for LFA, we subtract linear bandit-style bonuses (Abbasi-Yadkori et al., 2011) to encourage exploration. Namely, we consider the following optimistic estimate of \( \min_a \phi(s, a)^T w \):
\[
f_t(s, w) = \min_{a \in A} \left( \phi(s, a)^T w - \alpha_t \| \phi(s, a) \| \Lambda_t^{-1} \right),
\]
where \( \alpha_t > 0 \) is an exploration parameter. Second, and again common for LFA, we “clip” this estimate between 0 and some \( B_t > 0 \) (see Remark 5) to ensure bounded random variables, i.e., we define
\[
g_t(s, w) = \min \{ \max \{ f_t(s, w), 0 \}, B_t \}.
\]

\(^2\)We reiterate \( T \) is random for SSP, unlike the fixed \( T \) used in the infinite horizon discussion of Section 1.
This yields the operator \( \hat{G}_t : \mathbb{R}^d \to \mathbb{R}^d \) given by

\[
\hat{G}_t w = \Lambda_t^{-1} \sum_{\tau=1}^{t} \phi(s_{\tau}, a_{\tau}) \left( c(s_{\tau}, a_{\tau}) + g_t(s'_{\tau}, w) \right).
\]

Thus far, everything has naturally generalized finite horizon LFA. However, in the SSP setting, we will encounter several additional challenges.

**Remark 5** (Challenge 2). In finite horizon LFA, one sets \( B_t = H \) (the known horizon). Since the optimal value is \([0, H]\)-valued, clipping as in (8) only improves the optimal value estimate. In contrast, the analogous quantity in SSP is \( B_\star \), which is unknown. Hence, we will need to learn an upper bound \( B_t \geq B_\star \) to ensure the clipping does not distort our \( J_\star \) estimate.

**Remark 6** (Challenge 3). In light of Remark 5, \( B_t \), and thus \( \alpha_t \) (which needs to scale with \( B_t \) to ensure optimism), become trajectory-dependent random variables. This stands in contrast to other LFA settings, where the exploration parameter is deterministic.

**Remark 7** (Challenge 4). In SSP, we need to find fixed points, which we will do by showing the iterates of \( \hat{G}_t \) converge (see Remark 11). In contrast, finite horizon LFA uses a simple backward induction procedure, which basically iterates the operator \( H \) times and does not require any sort of convergence.

To overcome these issues, we break the problem into two parts, which treat Challenges 1-3 and 4, respectively. First, in Section 4, we assume an oracle provides OAFPs, which we use to solve the regret minimization problem. Second, in Section 5, we show how to compute OAFPs.

We define OAFPs as follows. In essence, we require the estimate (7) to be optimistic with respect to \( J_\star \) (when \( w \) in (7) is the OAFP), and the vector to be a fixed point of \( \hat{G}_t \) up to some tolerance.

**Definition 1** (OAFP). We say that \( w \in \mathbb{R}^d \) is an optimistic approximate fixed point (OAFP) if

\[
f_t(s, w) \leq J_\star(s) \quad \forall \ s \in S, \quad \| \hat{G}_t w - w \|_{\Lambda_t} \leq \alpha_t.
\]

Note that by Cauchy-Schwarz, the latter bound implies

\[
| \phi(s, a)^T (\hat{G}_t w - w) | \leq \alpha_t \| \phi(s, a) \|_{\Lambda_t^{-1}}.
\]

Finally, we note that due to the bonuses and clipping, \( \hat{G}_t \) need not concentrate near \( G \). Instead, Lemma 2 in Appendix B shows it concentrates near \( U_t \), where \( U_t w = \theta + \sum_{s \in S} g_t(s, w) \mu(s) \). More specifically, we show that with high probability, for any bounded \( w \),

\[
| \phi(s, a)^T (\hat{G}_t w - U_t w) | = O(\sqrt{\log t} \| \phi(s, a) \|_{\Lambda_t^{-1}}).
\]

To prove (11), we use covering arguments to take union bounds over \( w \), and the random functions \( g_t \). This is similar to Jin et al. (2020), though we have the added complication of random (and dependent) \( B_t \) and \( \alpha_t \). For later use, we also note that by Assumption 2,

\[
\phi(s, a)^T U_t w = c(s, a) + \mathbb{E}_{s'} g_t(s', w),
\]

where \( \mathbb{E}_{s'} \) is expectation with respect to \( s' \sim P(\cdot|s, a) \).
4 Regret minimization with oracle

We can now describe Algorithm 1, which assumes access to an OAFP oracle – i.e., a black box that, given \( \{s_t, a_t, s'_t\}_{t=1}^K \), returns an OAFP \( w_t \) per Definition 1.

**Inputs:** The inputs are a failure probability \( \delta \) and a sequence \( \{\kappa_t\}_{t=1}^\infty \) that will be used to shape \( \alpha_t \) in (7) (we cannot define \( \alpha_t \) a priori due to Remark 6).

**Intervals:** As in tabular SSP, we split time into intervals indexed by \( l \). The \( l \)-th interval will end at time \( M_l \), which will either correspond to the end of an episode or an intra-episode policy update (see Remark 4). At each such \( M_l \), we will call the oracle for an OAFP \( w_{M_l} \), which will define the policy executed in interval \( l+1 \).

**Initialization:** Lines 1-4 initialize the regularizer \( \Lambda_0 = I \), a (candidate) \( B_\star \) upper bound \( B_0 \) (see Remark 5), and the time and interval indices \( t \) and \( l \). We also set \( w_0 = a_0 = M_0 = 0 \) to ensure the forthcoming notation is well-defined.

**Episodic protocol:** Lines 5-9, 14-15, and 24 implement the protocol from Section 2. Additionally, Line 8 chooses the action to minimize the optimistic cost-to-go estimate (7) with respect to the most recent OAFP \( w_{M_{l-1}} \), and Line 15 updates \( \Lambda_t \). When the last episode ends (if it ever does), Lines 10-13 record the total number of intervals \( L \) and the total time \( T \).

**Cost-to-go bound:** If the cost-to-go estimate exceeds \( B_{t-1} \), then since \( f_{M_{l-1}}(s_t, w_{M_{l-1}}) \leq B_\star \) by Definition 1, we know \( B_{t-1} \) was not an upper bound for \( B_\star \), so we double it (Lines 16-17). Otherwise, we let \( B_t = B_{t-1} \) (Lines 18-19). Having defined \( B_t \), we use it and the input \( \kappa_t \) to define \( \alpha_t \) (Line 20).

**Policy update conditions:** Line 21 checks four conditions that require policy updates. The first three cause updates after the first observation, an episode ends, or \( B_{t-1} \) doubles. The fourth, taken from Abbasi-Yadkori et al. (2011), is that the determinant of \( \Lambda_t \) doubles. The idea is that, before this doubling occurs,

\[
\|\phi(s, a)\|_{\Lambda_t^{-1}} \leq \sqrt{2}\|\phi(s, a)\|_{\Lambda_{M_{l-1}}^{-1}} \forall (s, a) \in S \times A,
\]

which is analogous to tabular RL algorithms that wait to update until the number of visits to some \( (s, a) \) double (e.g., Jakobs et al. (2010)).

**Policy update:** If any of the conditions are met, Lines 22-23 call the oracle for an OAFP \( w_{M_0} \) and end the current interval. Note that in the next interval, the policy in Line 8 will use this OAFP.

We now present the main result of this section (Theorem 1), which assumes the input \( \kappa_t \) to Algorithm 1 scales as \( t^\lambda \) for some \( \lambda \in [0, \frac{1}{2}) \). Provided this holds, the theorem shows that Algorithm 1 obtains \( K^{\frac{1}{2}+\lambda} \) regret, i.e., smaller \( \kappa_t \) yields lower regret. The tradeoff is that smaller \( \kappa_t \) means smaller \( \alpha_t \) (see Line 20), so lower regret requires the OAFP to be a tighter fixed point and yield sharper cost-to-go estimates (see Definition 1). Note this is only a computational issue (not a statistical one), because Theorem 1 shows that OAFPs exist even when \( \lambda = 0 \) (where we obtain the optimal \( \sqrt{K} \) rate). Hence, the tradeoff is not too relevant in this section, though it will be in Section 5.

**Theorem 1** (General result). *Suppose Assumptions 1 and 2 hold and \( \kappa_t \in [9d, \Psi t^\lambda \log(t+1)] \) for some \( \Psi > 0 \) independent of \( t \) and some absolute constant \( \lambda \in [0, \frac{1}{2}) \). With probability at least \( 1 - \delta \),*
Algorithm 1: Regret minimization with oracle

**Input:** $\delta \in (0, 1), \{\kappa_t\}_{t=1}^{\infty} \subset [1, \infty)$

1. $A_0 = I_d$ (regularizer), $B_0 = c_{\min}$ ($B_\ast$ bound)
2. $w_0 = 0_d$ (OAFP), $\alpha_0 = 0$ (explore parameter)
3. $M_0 = 0$ (time 0-th interval ended)
4. $t = 1$ (current time), $l = 1$ (current interval)

for episode $k = 1, \ldots, K$ do
   5. $h = 1$ (current step), observe $s_h^k \in S \setminus \{s_{\text{goal}}\}$
   6. while $s_h^k \neq s_{\text{goal}}$ do
      7. Choose $a_h^k \in A$ to minimize
         \[ \phi(s_h^k, a_h^k)^T w_{M_{l-1}} - \alpha_{M_{l-1}} \| \phi(s_h^k, a_h^k) \|_{\Lambda_{M_{l-1}}} \]
      8. Observe $c(s_h^k, a_h^k)$ and $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$
      9. if $k = K$ and $s_{h+1}^k = s_{\text{goal}}$ then
         10. $L = l$ (total number intervals), $M_L = t$
         11. $T = t$ (total time elapsed)
      else
         12. $(s_t, a_t, s'_t) = (s_h^k, a_h^k, s_{h+1}^k)$
         13. $\Lambda_t = \Lambda_{t-1} + \phi(s_t, a_t) \phi(s_t, a_t)^T$
         14. if $f_{M_{l-1}}(s'_t, w_{M_{l-1}}) > B_{t-1}$ then
            15. $B_t = 2B_{t-1}$
         else
            16. $B_t = B_{t-1}$
            17. $\alpha_t = (B_t + 1)\kappa_t \sqrt{\log(t(B_t + 1)\kappa_t/\delta)}$
      18. if $t = 1$ or $s'_t = s_{\text{goal}}$ or $B_t \neq B_{t-1}$ or $\det(\Lambda_t) \geq 2 \det(\Lambda_{M_{l-1}})$ then
         19. Call oracle for OAFP $w_t$ (Def. 1)
         20. $M_l = t, l \leftarrow l + 1$
      21. $t \leftarrow t + 1, h \leftarrow h + 1$
   end while

there exists an OAFP for all $t \in [T]$ and
\[
R(K) = \hat{O}\left( (B_\ast^{2+\lambda} + B_\ast^{2+\lambda})d^2 \Psi(K/c_{\min})^{\frac{1}{2}+\lambda} \\
+ (B_\ast + 1)^{\frac{2}{1-2\lambda}} d^{\frac{2}{1-2\lambda}} \Psi^{\frac{2}{1-2\lambda}} c_{\min}^{\frac{1+2\lambda}{1-2\lambda}} \right).
\]

Thus, Algorithm 1 ensures $\sqrt{K}$ regret when given an oracle that returns OAFPs for $\lambda = 0$. This is analogous to Wei et al. (2021); Zanette et al. (2020b), which provide $\sqrt{K}$ regret for average cost and finite horizon problems when given certain optimization oracles. More specifically, in the best case $\kappa_t = 9d$ permitted by Theorem 1, we have the following corollary.

**Corollary 1** (Best case). Suppose Assumptions 1 and 2 hold, $B_\ast \geq 1$, and $\kappa_t = 9d$. With probability at least $1 - \delta$, there exists an OAFP for all $t \in [T]$ and
\[
R(K) = \hat{O}\left( \sqrt{B_\ast^2 d^3 K/c_{\min}} + B_\ast^2 d^3 / c_{\min} \right).
\]
Note Corollary 1 also assumes $B_* \geq 1$, which is natural (otherwise, $J^*$ can arbitrarily smaller than the cost upper bound $1 \geq c(s, a)$). Of course, the $B_* < 1$ case can be recovered from Theorem 1. Forthcoming results also assume $B_* \geq 1$, but we report bounds for the general case in the appendix.

There are no existing LFA bounds for SSP to compare with, so we consider the tabular case. Here we obtain $R(K) = O(\sqrt{B^2S^3A^3K/c_{\text{min}}})$ for large $K$, which matches the best Hoeffding algorithms in terms of $B_*$, $K$, and $c_{\text{min}}$ (see Section 1.1). Our scaling $(SA)^{3/2}$ is worse but can be improved to within a $\sqrt{A}$ factor, i.e., to $SA$ (see Remark 14 in Appendix C). We note a similar gap arises when specializing Jin et al. (2020)'s bound to the tabular finite horizon setting.

Theorem 1 proof sketch. The proof is in Appendix C but we discuss the key ideas for the regret bound here. For simplicity, we set $\lambda = 0$ and show $R(K) = O(\sqrt{K})$ while hiding terms independent of $K$. Again for simplicity, we use $f_t$ and its clipping $g_t$ interchangeably.

Regret decomposition: Fix $\hat{T} \in \mathbb{N}$ and let $\hat{K}$ and $\hat{L}$ denote the number of episodes and intervals completed by time $T \land \hat{T}$. In light of Remark 4, we will bound regret by time $T \land \hat{T}$, show it is finite, and let $\hat{T} \to \infty$. More specifically, let $\hat{R}(\hat{T}) = \hat{R}_1(\hat{T}) + \hat{R}_2(\hat{T})$, where we define the per-interval regret

$$\hat{R}_1(\hat{T}) = \sum_{l=0}^{\hat{L}-1} \sum_{t=1+M_l}^{M_{l+1}} c(s_t, a_t) - J^*(s_{1+M_l}),$$

and the “excess regret” from intra-episode updates

$$\hat{R}_2(\hat{T}) = \sum_{l=0}^{\hat{L}-1} J^*(s_{1+M_l}) - \sum_{k=1}^{\hat{K}} J^*(s_k^T).$$

Cost-to-go bound: To bound both terms, we require a bound on $B_t$. Since $f_{M_{t-1}}(s_t', w_{M_{t-1}}) \leq B_*$ by Definition 1, as soon as $B_{t-1}$ exceeds $B_*$, the condition Line 16 will stop occurring. This implies $B_t \leq 2B_*$. 

Per-interval regret: First note that by (9),

$$\hat{R}_1(\hat{T}) \leq \sum_{l=0}^{\hat{L}-1} \sum_{t=1+M_l}^{M_{l+1}} c(s_t, a_t) - f_{M_l}(s_{1+M_l}, w_{M_l}).$$

Now fix $l$ and $t$ as the double summation. Then by the chosen policy (Line 8 of Algorithm 1), we know

$$f_{M_l}(s_t, w_{M_l}) \approx \phi(s_t, a_t)^T w_{M_l},$$

where $\approx$ hides the bonus term $\alpha_{M_l}\|\phi(s_t, a_t)\|_{A_{M_l}}$. Again up to the bonus, (10) and (11) imply

$$\phi(s_t, a_t)^T w_{M_l} \approx \phi(s_t, a_t)^T U_t w_{M_l}.$$

Finally, by (12), up to a conditionally zero-mean term,

$$\phi(s_t, a_t)^T U_t w_{M_l} \approx c(s_t, a_t) + f_{M_l}(s_{t+1}, w_{M_l}).$$

Combining the last three inequalities, we obtain

$$c(s_t, a_t) - f_{M_l}(s_t, w_{M_l}) \approx -f_{M_l}(s_{t+1}, w_{M_l}).$$
Iterating in (13), this implies $\tilde{R}_1(\tilde{T}) \approx 0$, where $\approx$ hides a sum of $T \wedge \tilde{T}$ zero-mean terms and bonuses. Both are $O(\sqrt{T \wedge T})$, because $\alpha_t = \tilde{O}(1)$ by $\lambda = 0$ and the above proof that $B_t \leq 2B_\ast = O(1)$.

**Excess regret:** By definition, $\tilde{R}_2(\tilde{T}) \leq B_\ast(\tilde{L} - \tilde{K})$, where $\tilde{L} - \tilde{K}$ is the number of intra-episode episodes, i.e., the number of times $B_t$ or $\det(\Lambda_t)$ double. The former occurs $O(1)$ times since $B_t = O(1)$ and the latter $\tilde{O}(1)$ times since $\det(\Lambda_t) = O(t)$. Thus, $\tilde{R}_2(\tilde{T})$ is dominated by $\tilde{R}_1(\tilde{T})$ (in terms of $T$).

**Completing the proof:** So far, we have argued $\tilde{R}(\tilde{T}) = O(\sqrt{T \wedge T})$. By definition, we also know

$$(T \wedge \tilde{T})c_{\text{min}} \leq \sum_{t=1}^{T \wedge \tilde{T}} c(s_t, a_t) = \tilde{R}(\tilde{T}) + \tilde{O}(\tilde{T}).$$

Combining, we obtain $T \wedge \tilde{T} = O(\sqrt{T \wedge T} + K)$, which implies $T \wedge \tilde{T} = O(K)$. Thus, choosing $\tilde{T} \gg K$, we conclude $T = T \wedge \tilde{T} = O(K)$, so $R(K) = \tilde{R}(\tilde{T})$ and $\sqrt{T \wedge \tilde{T}} = O(\sqrt{K})$. Plugging into the bound $\tilde{R}(\tilde{T}) = O(\sqrt{T \wedge \tilde{T}})$ completes the proof. □

**Remark 8** (Finite $T$). It is tempting to choose $\tilde{T} = \infty$ at the start of the proof, show $T = O(\sqrt{T} + K)$ as above, and conclude $T = O(K) < \infty$. However, such logic is circular: it assumes $T$ is finite (e.g., to justify adding/subtracting $T$ terms) in order to prove it is finite. We point out this mistake (which some tabular SSP papers have made) so future work can avoid it.

**Remark 9** (Complexity). Algorithm 1’s runtime is dominated by computation of $\{\pi(s_t)\}_{t=1}^T$, which is $O(Ad^2T)$ when $\Lambda_t^{-1}$ and $\det(\Lambda_t)$ are iteratively updated. In the proof, we show $T$ is polynomial in all parameters (see Remark 13 in Appendix C), so given an efficient oracle, Algorithm 1 is itself efficient.

## 5 Oracle implementation

We next discuss how to compute OAFPs. The obvious approach is to iterate $\hat{G}_t$. This indeed yields optimistic estimates, i.e., the first inequality in (9) will hold.

**Lemma 1** (Informal version of Corollary 4 from Appendix D). With high probability, if $\alpha_t = \Omega(\sqrt{\log t})$,

$$f_t(s, \hat{G}_t^{n-1}0) \leq J^*(s) \forall s \in \mathcal{S}, n \in \mathbb{N}, t \in [T].$$

**Proof sketch.** When $n = 0$, the bound is immediate, since $f_t(s, \hat{G}_t^00) = f_t(s, 0) \leq 0$. If true for $n$, then

$$g_t(s', \hat{G}_t^{n-1}0) = \max\{f_t(s', \hat{G}_t^{n-1}0), 0\} \leq J^*(s').$$

Thus, by (12) and Bellman optimality (1),

$$\phi(s, a)^T U_t(\hat{G}_t^{n-1}0) \leq c(s, a) + \mathbb{E}_{s'} J^*(s') = Q^*(s, a),$$

so by (11), $\alpha_t = \Omega(\sqrt{\log t})$, and (1),

$$f_t(s, \hat{G}_t^n0) \leq \min_{a \in \mathcal{A}} Q^*(s, a) = J^*(s).$$

□
We thus propose Algorithm 2 for OAFP computation, which iterates \( \hat{G}_t \) until the second inequality in (9) holds (the first holds by Lemma 1). Our next theorem shows that, for appropriate \( \kappa_t \), it terminates in polynomial iterations. Combined with Remark 9, this shows Algorithms 1 and 2 provide an end-to-end statistically/computationally efficient scheme that uses stationary policies – a first in the LFA literature.

**Theorem 2** (End-to-end algorithm). Suppose Assumptions 1 and 2 hold, \( B_* \geq 1, \kappa_t = 54dt^{1/3} \), and Algorithm 2 is the oracle. With probability at least \( 1 - \delta \), Algorithm 2 returns an OAFP within \( O(dt^{1/6}) \) iterations for each \( t \in [T] \) it is called, and

\[
R(K) = \tilde{O}\left(B_*^{11/6}d^3(K/c_{\text{min}})\frac{\kappa}{n} + B_6^6d^0c_{\text{min}}^{-5}\right).
\]

**Proof sketch.** The proof (and those of Theorems 3 and 4) can be found in Appendix D. Given Theorem 1 and Lemma 1, the remaining challenge is to show Algorithm 2 terminates, i.e., \( \| \hat{G}_t^n 0 - \hat{G}_t^{n-1} 0 \|_{\Lambda_t} \leq \alpha_t \) for some \( n = O(dt^{1/6}) \). Equivalently, if we ignore the regularizer, then by definition of \( \| \cdot \|_{\Lambda_t} \), we aim to show

\[
\sum_{\tau=1}^{t} (\phi(s_{\tau}, a_{\tau})^T(\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0))^2 \leq \alpha_t^2.
\]  

To bound the \( \tau \)-th summand, we show \( U_t \) converges, \( \hat{G}_t \) tracks \( U_t \), and use the triangle inequality.

**U_t converges:** (Bonet, 2007, Lemma 4.3) implies the standard Bellman iterates converge at rate \( \frac{SA}{n} \). By (12), \( \phi(s, a)^T U_t^n 0 \) are basically the same iterates (up to bonuses and clipping), which means they converge at rate \( \frac{SA}{n} \) as well. The constant \( SA \) is infeasible, but with a more careful analysis, we can exploit the low rank structure to show (in terms of \( d \) and \( n \))

\[
\max_{(s,a) \in S \times A} |\phi(s, a)^T(U_t^n 0 - U_t^{n-1} 0)| = O\left(\frac{d^2}{n}\right).
\]  

**\( \hat{G}_t \) tracks \( U_t \):** Let \( x_n = \hat{G}_t^n 0 \) and \( y_n = U_t^n 0 \). By (11),

\[
|\phi(s, a)^T(x_{n+1} - U_t x_n)| = \tilde{O}(1) |\phi(s, a)|_{\Lambda_t^{-1}}.
\]

On the other hand, (12) implies

\[
|\phi(s, a)^T(U_t x_n - y_{n+1})| \leq \mathbb{E}_{\nu'} |g_t(s', x_n) - g_t(s', y_n)|.
\]

Combining and using the triangle inequality, we obtain

\[
|\phi(s, a)^T(x_{n+1} - y_{n+1})| = \tilde{O}(1) |\phi(s, a)|_{\Lambda_t^{-1}}
\]

\[
+ \mathbb{E}_{\nu'} |g_t(s', x_n) - g_t(s', y_n)|.
\]  

| Algorithm 2: Computing OAFPs |
|---|
| 1. Set \( n = 1 \), compute \( \hat{G}_1^n 0 = \hat{G}_1 0 \) and \( \hat{G}_1^{n-1} 0 = 0 \) |
| 2. While \( \| \hat{G}_t^n 0 - \hat{G}_t^{n-1} 0 \|_{\Lambda_t} > \alpha_t \) do |
| 3. Set \( n \leftarrow n + 1 \), compute \( \hat{G}_t^n 0 = \hat{G}_t(\hat{G}_t^{n-1} 0) \) |
| 4. Return \( w_t = \hat{G}_t^{n-1} 0 \) |
Finally, a straightforward calculation yields
\[ |g_t(s', x_n) - g_t(s', y_n)| \leq \max_{a'} |\phi(s', a')^T (x_n - y_n)|. \]

This suggests bounding the average in (16) by the max (over \( s' \in S \)) and iterating. However, such a bound involves \( \max_{(s, a) \in S \times A} \|\phi(s, a)\|_{\Lambda_i^{-1}} \), which is too large. The crucial idea is to take max only over explored states, namely, \( s' \) such that \( \max_{a'} \|\phi(s', a')\|_{\Lambda_i^{-1}} \ll \alpha_i^{-1} \). The key implication is that if \( s' \) is unexplored, then \( \alpha_i \|\phi(s', a')\|_{\Lambda_i^{-1}} \gg 1 \) for some \( a' \), so \( f_t(s', x_n) \leq 0 \) by definition and \( g_t(s', x_n) = 0 \) by clipping (and similar for \( y_n \)). This insight allows us to iterate the above, but only over \( (s, a) \in S_t \times A_t \), to obtain
\[ \max_{(s, a) \in S_t \times A_t} \|\phi(s, a)\|^T (x_n - y_n) = \tilde{O}(n/\alpha_t). \]

Plugging into (16) and recalling \( x_n = \hat{G}_t^0 \) and \( y_n = U_t^0 \), this extends to all \( (s, a) \in S \times A \) as follows:
\[ |\phi(s, a)^T (\hat{G}_t^n - U_t^n)| = \tilde{O}(\|\phi(s, a)\|_{\Lambda_i^{-1}} + n/\alpha_t). \quad (17) \]

**Completing the proof:** By (15) and (17), the \( \tau \)-th summand in (14) is \( \tilde{O}(\|\phi(s, a)\|_{\Lambda_i^{-1}} + (n/\alpha_t)^2) \) (in terms of \( n \) and \( t \)). This yields a sum of squared bonuses, which is independent of \( t \), plus \( O(t(\frac{1}{\alpha_t} + \frac{n}{\alpha_t})^2) \). Finally, since \( \alpha_t = O(t^{1/3}) \) by choice of \( \kappa_t \), after \( n = O(t^{1/6}) \) iterations, \( t(\frac{1}{\alpha_t} + \frac{n}{\alpha_t})^2 = O(t^{2/3}) = O(\alpha_t^2) \).

**Remark 10 (Clipping).** Most LFA papers use clipping to show an event like (11) occurs with high probability, then bound regret on this event, after which clipping becomes somewhat of a nuisance. In contrast, the proof sketch exploits it on the high probability event.

**Remark 11 (Convergence).** The proof sketch shows \( \|x_t\|_{\Lambda_t} = O(t^{1/3}) \), where \( x_t = \hat{G}_t^n - \hat{G}_t^{n-1} \) is the fixed point error after \( n_t = O(t^{1/6}) \) iterations. Note the norm equivalence \( \|x_t\|_{\Lambda_t} = O(\sqrt{\tau})\|x_t\|_2 \) always holds, so if it is reasonably tight (i.e., if \( \|x_t\|_2 = o(t^{-1/3})\|x_t\|_{\Lambda_t} \)), then \( x_t \to 0 \) as \( t \to \infty \) (i.e., Algorithm 2 yields a fixed point asymptotically in \( t \)).

If we strengthen Assumption 1 to mandate that all stationary policies are proper, we can improve Theorem 2’s regret bound. While this assumption is arguably strong, it seems perfectly reasonable for, e.g., games that eventually end. The benefit is that the Bellman operator \( T : S \times A \to \mathbb{R} \) given by
\[ (TQ)(s, a) = c(s, a) + \mathbb{E}_{s', a'} \min_{a' \in A} Q(s', a') \quad (18) \]
is contractive. More precisely, for some \( \rho \in (0, 1) \) and \( \omega(s) > 0 \), if \( \|x\| = \max_{(s, a) \in S \times A} \omega(s)|x(s, a)| \), then
\[ \|TQ_1 - TQ_2\| \leq \rho \|Q_1 - Q_2\|. \quad (19) \]

Define \( \chi = \max_{s \in S} \omega(s)/\min_{s \in S} \omega(s) \). Assuming nontrivial upper bounds \( \hat{\rho} \in [\rho, 1) \) and \( \bar{\chi} \in (1, \infty) \) are known, our next result establishes \( K^{3/4} \) regret.

**Theorem 3 (All proper).** Suppose Assumptions 1 and 2 hold, \( B_* \geq 1 \), all stationary policies are proper, \( \kappa_t = 54d t^{1/4} / \sqrt{N_t} \) with \( N_t = \log(3t \bar{\chi})/(1 - \hat{\rho}) \), and Algorithm 2 is the oracle. With probability at least \( 1 - \delta \), Algorithm 2 returns an OAFP within \( N_t \) iterations for each \( t \in [T] \) it is called, and
\[ R(K) = \tilde{O}\left(B_*^7d^3(K/c_{min})^{3/2}N_t^{1/2} + B_*^4d^6N_t^2c_{min}^{-3}\right). \]
Proof sketch. Recall in the Theorem 2 proof sketch, we showed \( \| \tilde{G}_t^0 - \tilde{G}_t^{n-1} \|_{\Lambda_t} = O(\sqrt{t} (\frac{1}{n} + \frac{n}{\alpha t})) \), where \( \frac{1}{n} \) was the \( U_t \) convergence rate. Under the stronger assumption of Theorem 3, \( U_t \) inherits a contraction property from (19), which improves the rate to \( \rho^n \). Hence, after \( N_t \) iterations, we have \( \| \tilde{G}_t^0 - \tilde{G}_t^{n-1} \|_{\Lambda_t} = \tilde{O}(\sqrt{t}/\alpha_t) = \tilde{O}(\alpha_t) \) by the choice \( \kappa_t = \tilde{O}(t^{1/4}) \).

Finally, we demonstrate a nontrivial case where Algorithm 2 returns OAFPs for the best case \( \kappa_t \).

Theorem 4 (Orthogonal features). Suppose Assumptions 1 and 2 hold, \( B_* \geq 1 \), \( \{\phi(s,a)\}_{(s,a) \in S \times A} \subset \{q_i\}_{i=1}^d \) for some orthonormal set \( \{q_i\}_{i=1}^d \), \( \kappa_t = 9d \), and Algorithm 2 is the oracle. With probability at least \( 1 - \delta \), Algorithm 2 returns an OAFP within \( \tilde{O}(t) \) iterations for each \( t \in [T] \) it is called, and regret is bounded as in Corollary 1.

Proof sketch. The additional assumption yields an explicit expression for \( \Lambda_t^{-1} \), which allows us to show \( G_t \) itself is contractive. This enables a direct convergence proof, i.e., without comparing to the iterates of \( U_t \).

6 Extensions

Before closing, we mention some extensions of our results. We defer the details to Appendix A.

Generalizing Theorem 3: When the upper bounds \( \hat{\chi} \) and \( \hat{\rho} \) are unavailable, we can instead set \( N_t = t^{2/3} \) for some absolute constant \( \gamma \in (0, \frac{1}{2}) \) and modify Algorithm 2 to terminate after \( N_t \) iterations (if it has not already). This approach is efficient by design, returns OAFPs for \( t \geq \Gamma = \tilde{O}( (\log n)^{\frac{1}{3\gamma}} ) \), and (combined with Algorithm 1) achieves the Theorem 1 regret bound with \( \lambda = \frac{1}{2} + \gamma \) and an additive \( \Gamma \) term.

Generalizing Theorem 4: When \( \{\phi(s,a)\} \) is not orthogonal but there at most \( d' \) unique features, they can be orthogonalized to recover the \( \sqrt{K} \) regret bound from Theorem 4, with \( d' \) replaced by \( d \). This is efficient if \( d' \ll SA \), which is reminiscent of state aggregation.

Zero/vanishing costs: Suppose we modify Assumption 1 to allow for \( c_{\text{min}} = 0 \), which is the minimal assumption in tabular SSP (the upper bound \( c(s,a) \leq 1 \) can be easily generalized). In this setting, as in the tabular case, we define regret with respect to the optimal proper policy \( \pi_{\text{prop}} \).

We use the same algorithms but replace \( c(s,a) \) with \( c(s,a) + \eta \) for some small perturbation \( \eta > 0 \) in the definition of \( G_t \), invoke Theorem 1 to bound the regret of this algorithm with respect to the optimal policy in the perturbed SSP (which remains linear), and compare the cost-to-go of the latter with that of \( \pi_{\text{prop}}^* \). With \( \kappa_t \) scaling as \( t^\lambda \) for some \( \lambda \in [0, \frac{1}{2}) \) (as in Theorem 1) and \( \eta \) as \( K^{(2\lambda-1)/(2\lambda+3)} \), this yields \( K^{(4\lambda+3)/(2\lambda+3)} \) regret.\(^3\) Since \( \frac{4\lambda+2}{2\lambda+3} < 1 \) for any \( \lambda \in [0, \frac{1}{2}) \), Algorithms 1-2 with the Theorem 2 parameters obtain statistical/computational efficiency with stationary policies under minimal SSP assumptions. Note this approach also works if Assumption 1 holds but \( c_{\text{min}} \) vanishes in \( K \). For example, Corollary 1 only promises linear regret when \( c_{\text{min}} = K^{-1} \), but choosing \( \eta = K^{-1/3} \) ensures \( K^{2/3} \) regret.

Remark 12 (\( c_{\text{min}}^{-1} \) dependence). As seen above, the \( c_{\text{min}}^{-1} \) dependence of the leading term in our regret bound inflates the scaling in \( K \) when dealing with small costs. Consequently, it would be ideal if this term was independent of \( c_{\text{min}}^{-1} \). In the special case of tabular SSP, avoiding this issue seems

\(^3\)If \( K \) is unknown, we can use a standard doubling trick.
to require Bernstein-style confidence sets, which for LFA have only been studied recently and only for simple finite horizon problems (see Section 1.1). Given the unique LFA challenges that arise for SSP (see Remarks 4, 5, 6, and 7), we leave such bounds for future work.

7 Conclusion

In this paper, we presented the first algorithms and regret bounds for SSP with LFA, and more generally, the first efficient LFA algorithm that uses stationary policies. Addressing the remaining statistical/computational gap (i.e., proving \( \sqrt{K} \) regret in general) is an important open problem. Given the modular nature of the paper, one solution approach would be to combine our results with an improved oracle.

Broader societal impact: This paper is theoretical and has no immediate societal impact. Nevertheless, RL focuses on automated decision making, and training data can inject bias into these decisions. Care should be taken to minimize this bias when using our (or any other) algorithms in practice.

Acknowledgements

This work was partially supported by ONR Grant N00014-19-1-2566, ARO Grant ARO W911NF-19-1-0379, NSF/USDA Grant AG 2018-67007-28379, and NSF Grants 1910112, 2019844, 1704970, and 1934986.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In NIPS, volume 11, pages 2312–2320, 2011.

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

Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. Mathematics of Operations Research, 16(3):580–595, 1991.

Blai Bonet. On the speed of convergence of value iteration on stochastic shortest-path problems. Mathematics of Operations Research, 32(2):365–373, 2007.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. arXiv preprint arXiv:2106.08377, 2021.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. arXiv preprint arXiv:2103.13056, 2021.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In International Conference on Learning Representations, 2020.

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. arXiv preprint arXiv:2106.05335, 2021.
Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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

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

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pages 9428–9437. PMLR, 2020.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *arXiv preprint arXiv:2104.11186*, 2021.

Yuanhao Wang, Ruosong Wang, and Sham M Kakade. An exponential lower bound for linearly-realizable mdps with constant suboptimality gap. *arXiv preprint arXiv:2103.12690*, 2021.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

Gellert Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. *arXiv preprint arXiv:2102.07301*, 2021.

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

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

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

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021.
Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020.

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

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.
A Section 6 details

A.1 Generalizing Theorem 3

As discussed in Section 6, we can use the following OAFP oracle, which modifies Algorithm 2 by returning the $N_t$-th iterate if it reaches the $N_t$-th iteration. Let $N_t = t^{2\gamma}$ for some absolute constant $\gamma \in (0, \frac{1}{2})$ and set $\kappa_t = 54dt^5N_t$ as in Theorem 3. We show in Appendix D (see Remark 15) that with probability at least $1 - \delta/2$, for any $t \geq (\log(3t\gamma)/(1 - \rho))^{\frac{1}{2\gamma}}$ that Algorithm 3 is called, it returns an OAFP within $t^{2\gamma}$ iterations.

Now suppose we run Algorithm 1 with Algorithm 3 as the oracle. Let $\Gamma = N_t - N_{t-1}$ hold and suppose with Algorithm 1, we bound regret by $\Gamma$ for the first $\Gamma$ time steps, and by modifying the proof of Theorem 1, we can bound regret by $K^2\gamma$ thereafter (in terms of $K$). Thus, regret will scale as $K^{2\gamma}$ for this algorithm, with a second-order term $\Gamma$ in addition to the one from Theorem 1.

**Algorithm 3:** Computing OAFPs with iteration limit

1. Set $n = 1$, compute $G_{n-1} = G_0$ and $G_{n-1}^{-1} = 0$
2. while $\|\hat{G}_n - G_{n-1}\|_\Lambda > \alpha_t$ and $n \leq N_t$
   3. Set $n \leftarrow n + 1$, compute $G_n = \hat{G}_n(\hat{G}_n^{-1}0)$
4. return $w_t = G_{n-1}^{-1}$

A.2 Generalizing Theorem 4

Let Assumption 1 hold and suppose $\phi(s, a), \theta,$ and $\mu(s')$ satisfy Assumption 2. Denote by $\{\phi_i\}_{i=1}^{d''}$ the unique elements of $\{\phi(s, a)\}_{(s, a) \in (S \setminus \{s_{goal}\}) \times A}$. For any $d' \in \{d', d' + 1, \ldots\}$, define

$\Phi = [\phi_1 \cdots \phi_d'] \in \mathbb{R}^{d \times d'}$, $\Xi = [\Phi \ 0_{d \times (d'' - d')}] \in \mathbb{R}^{d \times d''}$.

Let $\Xi = R\hat{\Phi}$ be an RQ decomposition, i.e., $R \in \mathbb{R}^{d \times d'}$ is upper triangular $\hat{\Phi} \in \mathbb{R}^{d' \times d''}$ is orthogonal. For $(s, a) \in (S \setminus \{s_{goal}\}) \times A$, let $\hat{\phi}(s, a)$ be the $i$-th column of $\hat{\Phi}$, where $i \in [d']$ is such that $\phi(s, a) = \phi_i(s, a)$, and set $\hat{\phi}(s_{goal}, a) = 0 \forall a \in A$. We claim that $\hat{\phi}(s, a), R^T\theta,$ and $R^T\mu(s')$ satisfy Assumption 2. To prove (3), we first observe that for any $(s, a, s') \in (S \setminus \{s_{goal}\}) \times A$,

$$\phi(s, a)^T R^T = e_{i(s,a)}^T \hat{\Phi}^T R^T = e_{i(s,a)}^T \Xi^T = e_{i(s,a)}^T \Phi^T = \phi(s, a)^T,$$

so $\phi(s, a)^T R^T \theta = \phi(s, a)^T \theta = c(s, a)$ and $\phi(s, a)^T R^T \mu(s') = \phi(s, a)^T \mu(s') = P(s'|s, a)$, as desired.

The first inequality in (4) holds by construction. For the second inequality in (4), note $\varphi_i T \theta$ is the cost of some state-action pair and thus lies in $[0, 1]$ by Assumption 1. Combined with the fact that $\hat{\Phi}$ is orthogonal,

$$\|R^T \theta\|^2 = \|\hat{\Phi}^T R^T \theta\|^2 = \|\Xi T \theta\|^2 = \sum_{i=1}^{d'} (\varphi_i T \theta)^2 \leq d' \leq d''.$$ 

Similarly, for (5), since $\varphi_i \mu(\cdot)$ is a probability distribution over $S$, for any $h \in \mathbb{R}^S$, we have

$$\left\| \sum_{s' \in S} h(s') R^T \mu(s') \right\|^2_2 = \sum_{i=1}^{d'} \left( \sum_{s' \in S} h(s') \varphi_i^T \mu(s') \right)^2 \leq d'\|h\|_\gamma^2 \leq d''\|h\|_\gamma^2.$$

17
Algorithmically, this means that if $\Phi$ is known \textit{a priori}, we can set $d'' = d'$, compute $\tilde{\Phi}$, and use features $\tilde{\phi}(s, a) \in \mathbb{R}^{d'}$ instead of $\phi(s, a)$. Alternatively, if a nontrivial bound $d'' = O(d')$ is known, we can iteratively compute $\tilde{\Phi}$ via Gram–Schmidt (computing the $i$-th column when we observe unique features for the $i$-th time), increasing the dimension to $d''$. In the respective cases, our results follow with $d$ replaced by $d'$ and $d''$, respectively.

A.3 Zero/vanishing costs

Finally, we extend our results to the case where only Assumption 2 and the following hold.

**Assumption 3** (Weaker than Assumption 1). \textit{There exists a proper policy and $c(s, a) \in [0, 1] \forall (s, a) \in S \times A$.}

Now suppose the SSP instance $(S, A, P, c, s_{goal})$ only satisfies Assumptions 2 and 3. Let $c_\eta(s, a) = c(s, a) + \eta$ be the perturbed cost discussed in Section 6. Then the instance $(S, A, P, c_\eta, s_{goal})$ satisfies Assumption 1, with $c_{\min} = \eta$ (up to a small constant, since $c_\eta(s, a)$ may be as large as $1 + \eta$). Also define $\theta_\eta = \theta + \eta \sum_{s \in S} \mu(s)$. Then since Assumption 2 holds for the original instance, we have

$$c_\eta(s, a) = c(s, a) + \eta = c(s, a) + \eta \sum_{s' \in S} P(s'|s, a) = \phi(s, a)^T \theta + \eta \sum_{s' \in S} \phi(s, a)^T \mu(s') = \phi(s, a)^T \theta_\eta,$$

so it also holds for the perturbed instance (again, up to a constant, since we can only assert $\|\theta_\eta\|_2 \leq \sqrt{d}(1 + \eta)$). Thus, if we run Algorithm 1 on the original instance but replace $c(s, a)$ with $c_\eta(s, a)$ in the definition of $G_t$, and if $J^*_\eta$ is the optimal cost-to-go function on the perturbed instance, Theorem 1 ensures that

$$\sum_{t=1}^{T} c_\eta(s_t, a_t) - \sum_{k=1}^{K} J^*_\eta(s^k_1) = \tilde{O}\left( \left( B^2 + \frac{1+\lambda}{2} \right) d^2 \Psi(K/\eta)^{\frac{1}{2} + \lambda} + (B_* + 1) \frac{1}{2} \frac{1}{\lambda - 2 \alpha} \frac{1}{\lambda - 2 \alpha} \Psi \frac{1}{\alpha - 2 \lambda} \eta^\frac{1}{2} \frac{1}{\alpha - 2 \lambda} \right).$$

Also, since $J^*_\eta$ is optimal on the perturbed instance and both instances have the same transition kernel, we have

$$J^*_\eta(s) - J^*(s) \leq J^*_{\eta_{prop}}(s) - J^*_{prop}(s) \leq \eta T_* \forall s \in S,$$

where (we recall from Section 6) $\pi^*_{prop}$ is the optimal proper policy and $T_*$ is the maximum expected time it takes $\pi^*_{prop}$ to reach the goal state from any starting state (since this policy is proper, $T_* < \infty$). Therefore, since $c(s, a) \leq c_\eta(s, a)$ by definition, we can bound regret (defined with respect to $\pi^*_{prop}$, as in Section 6) by

$$R(K) \leq \sum_{t=1}^{T} c_\eta(s_t, a_t) - \sum_{k=1}^{K} J^*_{\eta_{prop}}(s^k_1) + \sum_{k=1}^{K} \left( J^*_{\eta_{prop}}(s^k_1) - J^*_{prop}(s^k_1) \right)
= \tilde{O}\left( \left( B^2 + \frac{1+\lambda}{2} \right) d^2 \Psi(K/\eta)^{\frac{1}{2} + \lambda} + \eta T_* K + (B_* + 1) \frac{1}{2} \frac{1}{\lambda - 2 \alpha} \frac{1}{\lambda - 2 \alpha} \Psi \frac{1}{\alpha - 2 \lambda} \eta^\frac{1}{2} \frac{1}{\alpha - 2 \lambda} \right).$$

Choosing $\eta$ to decay as $K^{(2\lambda-1)/(2\lambda+3)}$ ensures the first two terms scale as $K^{(4\lambda+2)/(2\lambda+3)}$ and the third as $K^{(2\lambda+1)/(2\lambda+3)}$. Since $\frac{2\lambda+1}{2\lambda+3} \leq \frac{4\lambda+2}{2\lambda+3} < 1$ for any $\lambda \in [0, \frac{1}{2})$, we thus have sublinear regret.

B Proof preliminaries

In this appendix, we collect some notation and results used in the proofs of multiple theorems.
B.1 Additional notation

We write $E_t$ for expectation conditioned on the first $t - 1$ state-action-state pairs and the $t$-th state action pair, $E_t[\cdot] = \mathbb{E}[\cdot|\{ (s_\tau, a_\tau, s'_\tau) \}_{\tau = 1}^{t-1} \cup \{ s_t, a_t \}]$. We let $E_{s_t}'$ denote expectation with respect to $s'_t$. Hence, for $h: \mathcal{S} \to \mathbb{R}$,

$$
E_t[h(s'_t)] = E_{s'_t}'[h(s'_t)] = \sum_{s \in \mathcal{S}} h(s)P(s|s_t, a_t).
$$

However, we emphasize that since $g_t$ in Definition 1 is a random function of the first $t$ state-action pairs, if $\tau < t$, we may have $E_t[|g_t(s'_t, w)|] \neq \sum_{s \in \mathcal{S}} g_t(s, w)P(s|s_t, a_t)$ for some $w \in \mathbb{R}^d$. On the other hand, $E_{s'_t}'[g_t(s'_t, w)] = \sum_{s \in \mathcal{S}} g_t(s, w)P(s|s_t, a_t)$ does hold for any $w \in \mathbb{R}^d$.

As discussed in Section 3, we also define the (random) operators $U_t, E_t : \mathbb{R}^d \to \mathbb{R}^d$ by

$$
U_tw = \theta + \sum_{s \in \mathcal{S}} g_t(s, w)\mu(s) \forall w \in \mathbb{R}^d, \quad E_t = \hat{G}_t - U_t.
$$

Here $U_t$ can be roughly viewed as the expected value of $\hat{G}_t$, so $E_t$ is the error between $\hat{G}_t$ and its mean. Note, however, that since $g_t$ is a random function, $U_tw$ is a random vector (even for fixed $w \in \mathbb{R}^d$). We also note the following identity, which is an immediate consequence of Assumption 2 and is frequently used:

$$
\phi(s, a)\mathsf{T}U_tw = \phi(s, a)\mathsf{T}\theta + \sum_{s' \in \mathcal{S}} g_t(s', w)\phi(s, a)\mathsf{T}\mu(s') = c(s, a) + \sum_{s' \in \mathcal{S}} g_t(s', w)P(s'|s, a).
$$

As in Section 5, we iterate these operators in the usual way, e.g., $\hat{G}_t^n0 = \hat{G}_t(\hat{G}_t^{n-1}0)$ for $n \in \mathbb{N}$ with $\hat{G}_t^00 = 0$.

Finally, for any $b > 0$, we define the clipping function $\Pi_{[0,b]} : \mathbb{R} \to [0,b]$ by

$$
\Pi_{[0,b]}(x) = \min\{\max\{x, 0\}, b\} = \max\{\min\{x, b\}, 0\}.
$$

(20)

Note that with this notation, we can more compactly write $g_t(\cdot, \cdot) = \Pi_{[0,B_t]}(f_t(\cdot, \cdot))$ in Definition 1.

B.2 Simple results

Claim 1 (Eigenvalues and norms). If Assumption 2 holds, then the eigenvalues of $\Lambda_t$ lie in $[1, t+1]$, and

$$
\|w\|_{\Lambda_t^{-1}} \leq \|w\|_2 \leq \|w\|_{\Lambda_t} \leq \sqrt{t+1}\|w\|_2 \leq \sqrt{(t+1)d}\|w\|_\infty \forall w \in \mathbb{R}^d.
$$

Proof. Let $\{\lambda_i\}_{i=1}^d$ and $\{q_i\}_{i=1}^d$ be the eigenvalues and (unit) eigenvectors of $\Lambda_t$. Then

$$
\lambda_i = \lambda_i q_i^\mathsf{T} q_i = q_i^\mathsf{T} \Lambda_t q_i = q_i^\mathsf{T} q_i + \sum_{\tau = 1}^t (\phi(s_\tau, a_\tau)\mathsf{T} q_i)^2 = 1 + \sum_{\tau = 1}^t (\phi(s_\tau, a_\tau)\mathsf{T} q_i)^2.
$$

The eigenvalue bounds follow, since $0 \leq (\phi(s_\tau, a_\tau)\mathsf{T} q_i)^2 \leq \|\phi(s_\tau, a_\tau)\|_2\|q_i\|_2 \leq 1$ by Cauchy-Schwarz. For the norm equivalences, we first use the eigenvalue bounds to write

$$
\|w\|^2_{\Lambda_t^{-1}} = \sum_{i=1}^d \frac{(q_i^\mathsf{T}w)^2}{\lambda_i} \leq \sum_{i=1}^d (q_i^\mathsf{T}w)^2 \leq \|w\|^2_{\Lambda_t} = \sum_{i=1}^d \lambda_i (q_i^\mathsf{T}w)^2 \leq (t+1) \sum_{i=1}^d (q_i^\mathsf{T}w)^2.
$$

Since $\sum_{i=1}^d (q_i^\mathsf{T}w)^2 = \|w\|^2_2$ by orthogonality, this proves the first three norm bounds. The fourth is standard. $\square$
Claim 2 (\( \Pi_{[0,b]} \) properties). For any \( b > 0 \) and \( x, y \in \mathbb{R} \), \( \Pi_{[0,b]}(x) \leq \max\{x, 0\} \) and \( |\Pi_{[0,b]}(x) - \Pi_{[0,b]}(y)| \leq |x - y| \).

Proof. The first bound holds by (20). For the second, assume without loss of generality that \( x \geq y \). By monotonicity, it suffices to show \( \Pi_{[0,b]}(x) - \Pi_{[0,b]}(y) \leq x - y \). If \( x < 0 \) or \( y > b \), then \( \Pi_{[0,b]}(x) = \Pi_{[0,b]}(y) \), so this is immediate. Otherwise, (20) implies \( \Pi_{[0,b]}(x) - \Pi_{[0,b]}(y) \leq \max\{x, 0\} - \min\{y, b\} = x - y. \)

Claim 3 (\( g_t \) bounds). If Assumption 2 holds, then for any \( t \in [T], s, a \in S \), and \( w_1, w_2 \in \mathbb{R}^d \),
\[
|g_t(s, w_1) - g_t(s, w_2)| \leq |f_t(s, w_1) - f_t(s, w_2)| \leq \max_{a \in A} |\phi(s,a)^T(w_1 - w_2)| \leq \|w_1 - w_2\|_2 \leq \sqrt{d}\|w_1 - w_2\|_\infty.
\]

Proof. The first bound follows from Claim 2. For the second, let \( \hat{a} \in A \) be any action attaining the minimum in the definition of \( f_t(s, w_1) \), i.e., \( \hat{a} \in \arg\min_{a \in A} (\phi(s,a)^T w_1 - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}}) \). Then
\[
f_t(s, w_1) - f_t(s, w_2) \geq f_t(s, w_1) - \left( \phi(s,\hat{a})^T w_2 - \alpha_t \|\phi(s,\hat{a})\|_{\Lambda_t^{-1}} \right) = \phi(s,\hat{a})^T (w_1 - w_2) \geq -\max_{a \in A} |\phi(s,a)^T(w_1 - w_2)|.
\]
By symmetry, we also have \( f_t(s, w_1) - f_t(s, w_2) \leq \max_{a \in A} |\phi(s,a)^T(w - w')| \); the second bound follows. The third follows from Cauchy-Schwarz and the fourth from a standard norm equivalence.

Claim 4 (Operator bounds). If Assumptions 1 and 2 hold, then for any \( t \in [T], w \in \mathbb{R}^d \), and \((s,a) \in S \times A\),
\[
\|\hat{G}_t w\|_\infty \leq \sqrt{td} \left( 1 + \max_{s \in S} g_t(s,w) \right), \quad \|\hat{G}_t 0\|_{\Lambda_t} \leq \sqrt{t+1}\|\hat{G}_t 0\|_2 \leq 2\sqrt{(t+1)d},
\]
\[
\|U_t w\|_2 \leq \sqrt{d} \left( 1 + \max_{s \in S} g_t(s,w) \right), \quad \phi(s,a)^T U_t w \in [0, B_t + 1].
\]

Proof. First observe that by a standard norm equivalence and Claim 1, for any \( w \in \mathbb{R}^d \), we have
\[
\|\Lambda_t^{-1} w\|_\infty \leq \|\Lambda_t^{-1} w\|_2 = \|\Lambda_t^{-1/2} w\|_{\Lambda_t^{-1}} \leq \|\Lambda_t^{-1/2} w\|_2 = \|w\|_{\Lambda_t^{-1}} \leq \|w\|_2.
\]
(21)
Combined with Cauchy-Schwarz and (Jin et al., 2020, Lemma D.1), we obtain
\[
\sum_{\tau=1}^{t} \|\Lambda_t^{-1} \phi(s_{\tau},a_{\tau})\|_\infty \leq \sum_{\tau=1}^{t} \|\phi(s_{\tau},a_{\tau})\|_{\Lambda_t^{-1}} \leq \sqrt{t} \sum_{\tau=1}^{t} \|\phi(s_{\tau},a_{\tau})\|_{\Lambda_t^{-1}}^2 \leq \sqrt{td},
\]
so the first \( \hat{G}_t \) bound follows from the triangle inequality. Next, because \( g_t(s,0) = 0 \ \forall \ s \in S \), we have
\[
\hat{G}_t 0 = \Lambda_t^{-1} \sum_{\tau=1}^{t} \phi(s_{\tau},a_{\tau}) c(s_{\tau},a_{\tau}) = \Lambda_t^{-1} \sum_{\tau=1}^{t} \phi(s_{\tau},a_{\tau}) \phi(s_{\tau},a_{\tau})^T \theta = \Lambda_t^{-1} (\Lambda_t - I) \theta = (I - \Lambda_t^{-1}) \theta.
\]
Therefore, by Claim 1 and (21), we obtain
\[
\|\hat{G}_t 0\|_{\Lambda_t} \leq \sqrt{t+1}\|\hat{G}_t 0\|_2 \leq \sqrt{t+1}(\|\theta\|_2 + \|\Lambda_t^{-1} \theta\|_2) \leq 2\sqrt{t+1}\|\theta\|_2 \leq 2\sqrt{(t+1)d}.
\]
Finally, the \( U_t \) bounds hold by assumption.
Claim 5 (\(B_*\) estimate). If Assumption 1 holds, then \(\sup_{t \geq 0} B_t \leq 2B_*\).

Proof. Since \(B_0 = c_{\text{min}} \leq B_*\) by assumption, it suffices to show \(B_\tau \leq 2B_*\ \forall \ \tau \in \mathbb{N}\). Suppose instead that \(B_\tau > 2B_*\) for some such \(\tau\). Let \(t = \min\{\tau \in \mathbb{N} : B_\tau > 2B_*\}\) be the first time it occurs. Then by Algorithm 1, we have \(B_* < B_t/2 = B_{t-1} < f_{M_{t-1}}(\hat{s}_t', w_{M_{t-1}})\) for some \(l \in \mathbb{N}\). If \(l = 1\), this contradicts the fact that \(w_0 = 0\); otherwise, it contradicts the fact that \(w_{M_t}\) is an OAFP (see Definition 1).

\[\] B.3 Operator concentration

Define the random variables \(W_t\) and \(\varepsilon_t\), and the event \(\mathcal{E}\), by

\[W_t = \alpha_t + \sqrt{td(B_t+1)}, \quad \varepsilon_t = 5(B_t+1)d\sqrt{\log(t\alpha_t/\delta)}, \quad \mathcal{E} = \left\{\sup_{w \in [-W_t+W_t]} \|E_t w\|_{\Lambda_t} \leq \varepsilon_t \ \forall \ t \in [T]\right\}.

The following is our main concentration result; the proof is lengthy so is deferred to Appendix E.1.

Lemma 2 (Error operator tail bound). If Assumptions 1 and 2 hold and \(\min_{t \in \mathbb{N}} \kappa_t \geq 9d\), then\(\mathbb{P}(\mathcal{E}) \geq 1 - \delta/2\).

The error bound \(\varepsilon_t\) in the lemma is related to the exploration parameter \(\alpha_t\) in the following manner.

Claim 6 (Lower bound on \(\alpha_t\)). For any \(t \in \mathbb{N}\), if \(\kappa_t \geq 9md\) for some \(m \geq 1\), then \(\alpha_t \geq \max\{m\varepsilon_t, (B_t + 1)\kappa_t\}\).

Proof. For the first bound, since \(\log x \leq x \ \forall \ x \in \mathbb{R}\), we have

\[\alpha_t = (B_t+1)\kappa_t \sqrt{\log(t(B_t+1)\kappa_t/\delta)} \leq (B_t+1)^{3/2}\kappa_t^{3/2}t^{1/2}/\delta^{1/2}.

Combined with \(5\sqrt{3/2} \leq 9\) and the assumption \(\kappa_t \geq 9md\), we obtain

\[m\varepsilon_t = 5md\sqrt{\log(t\alpha_t/\delta)}(B_t+1) \leq 5\sqrt{3/2md\log(t(B_t+1)\kappa_t/\delta)}(B_t+1) \leq 9md\sqrt{\log(t(B_t+1)\kappa_t/\delta)}(B_t+1) = \alpha_t.

For the second bound, simply note \(\log(t(B_t+1)\kappa_t/\delta) \geq \log(9md) \geq 1\) and use the definition of \(\alpha_t\).

As corollaries, we have the following special cases of operator concentration.

Corollary 2 (Error at OAFP). For any \(t \in \mathbb{N}\), if \(\kappa_t \geq 9d\) and \(w_t\) is an OAFP, then on the event \(\mathcal{E}\),

\[|\phi(s,a)^T(\hat{G}_t w_t - U_t w_t)| \leq \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \ \forall \ (s,a) \in \mathcal{S} \times \mathcal{A}.
\]

Proof. By Claims 1 and 4 and Definition 1, we have

\[\|w_t\|_{\infty} \leq \|w_t - \hat{G}_t w_t\|_2 + \|\hat{G}_t w_t\|_{\infty} \leq \|w_t - \hat{G}_t w_t\|_{\Lambda_t} + \sqrt{td}(B_t + 1) \leq \alpha_t + \sqrt{td}(B_t + 1) = W_t.
\]

Hence, using \(\kappa_t \geq 9d\) and Claim 6, we conclude \(\|E_t w_t\|_{\Lambda_t} \leq \alpha_t\) on \(\mathcal{E}\). Combined with Cauchy-Schwarz,

\[|\phi(s,a)^T(\hat{G}_t w_t - U_t w_t)| = |\phi(s,a)^T E_t w_t| \leq \|\phi(s,a)\|_{\Lambda_t^{-1}} \|E_t\|_{\Lambda_t} \leq \|\phi(s,a)\|_{\Lambda_t^{-1}} \alpha_t.
\]
Corollary 3 (Error at $\hat{G}_t$ iterates). For any $t \in \mathbb{N}$, if $\kappa_t \geq 9md$ for some $m \geq 1$, then on the event $\mathcal{E}$, for any $n \in \mathbb{N}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$|\phi(s, a)^T(\hat{G}_t^n - U_t(\hat{G}_t^{n-1}))| \leq \|\hat{G}_t^n - U_t(\hat{G}_t^{n-1})\|_{\Lambda_t} \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq \varepsilon_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}}/m.$$  

Proof. First note $\hat{G}_t^n - U_t(\hat{G}_t^{n-1}) = E_t(\hat{G}_t^{n-1})$. For $n \geq 2$, $\|\hat{G}_t^{n-1}\|_{\infty} = \|\hat{G}_t(\hat{G}_t^{n-2})\|_{\infty} \leq \sqrt[d]{d}(B_t + 1) \leq W_t$ by Claim 4, and for $n = 1$, $\|\hat{G}_t^{n-1}\|_{\infty} = 0$. Hence, for any $n \in \mathbb{N}$, we have $\|E_t(\hat{G}_t^{n-1})\|_{\Lambda_t} \leq \varepsilon_t$ on $\mathcal{E}$ by definition. The desired bounds follow from Cauchy-Schwarz and Claim 6 similar to Corollary 2. 

B.4 Operator convergence

We next show the operator $U_t$ converges in a certain sense. We begin by proving some basic properties.

Claim 7 ($U_t$ properties). If Assumptions 1 and 2 hold, then for any $t \in [T]$,

$$0 \leq \cdots \leq \phi(s, a)^TU_t^n \leq \phi(s, a)^TU_t^0 \leq \cdots \leq B_t + 1 \forall (s, a) \in \mathcal{S} \times \mathcal{A},$$ (22)

$$\max_{(s, a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)^T(U_t^{n+1} - U_t^n)| \leq \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)^T(U_t^n - U_t^{n-1})| \forall n \in \mathbb{N}. \quad (23)$$

Proof. By Claim 4, we already know $\phi(s, a)^TU_t^0 \in [0, B_t + 1]$. To complete the proof of (22), we show by induction on $n$ that $\min_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a)^T(U_t^n - U_t^{n-1}) \geq 0$. For $n = 1$, we have

$$\min_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a)^T(U_t^1 - U_t^0) = \min_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a)^TU_t0 = \min_{(s, a) \in \mathcal{S} \times \mathcal{A}} c(s, a) \geq 0.$$

Now assuming $\min_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a)^T(U_t^n - U_t^{n-1}) \geq 0$, we have

$$\min_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a)^T(U_t^{n+1} - U_t^{n}) = \min_{(s, a) \in \mathcal{S} \times \mathcal{A}} \sum_{s' \in \mathcal{S}} (g_t(s', U_t^n) - g_t(s', U_t^{n-1}))P(s'|s, a) \geq 0.$$ 

Finally, (23) follows from Claim 3:

$$\max_{(s, a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)^T(U_t^{n+1} - U_t^n)| \leq \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \sum_{s' \in \mathcal{S}} |g_t(s', U_t^n) - g_t(s', U_t^{n-1})|P(s'|s, a) \leq \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)^T(U_t^n - U_t^{n-1})|. \quad \square$$

The proof of Claim 7 shows that $U_t$ is nonexpansive in the induced $\ell_\infty$ norm $\|\cdot\| = \|\Phi^T\cdot\|_\infty$, where $\Phi$ is the matrix with columns $\{\phi(s, a)\}_{(s, a) \in \mathcal{S} \times \mathcal{A}}$. Combined with the claim’s monotonicity result, this is enough to show that $U_t$ converges at rate $1/n$. However, because the induced norm lifts to $|\mathcal{S} \times \mathcal{A}|$-dimensional space, a naive convergence proof yields a constant that scales with $|\mathcal{S} \times \mathcal{A}|$. The next claim will allow us to avoid this.

Claim 8 (A linear algebra result). Let $Z = \mathcal{S} \times \mathcal{A}$, $\Upsilon \in \mathbb{R}^{d \times Z}$, and $r = \text{rank}(\Upsilon)$. For any $Z' \subset Z$, denote by $\Upsilon(Z')$ the submatrix of $\Upsilon$ with columns $Z'$, with $T(z) = \Upsilon(\{z\})$ for any $z \in Z$ for simplicity. Then there exists $Z' \subset Z$ such that $|Z'| = r$ and $\|\Upsilon^T x\|_{\infty} \leq r \|\Upsilon(Z')^T x\|_{\infty}$ for any $x \in \mathbb{R}^d$. 

22
Proof. We first assume \( r = d \). Then we can find \( Z'' \subset \mathcal{Z} \) such that \( |Z''| = d \) and \( \det(\mathcal{Y}(Z'')) \neq 0 \). Let \( Z' \) be whichever such \( Z'' \) maximizes \( |\det(\mathcal{Y}(Z''))| \). Set \( H = \mathcal{Y}^T(\mathcal{Y}(Z')^{-1} \right), which is well-defined by choice of \( Z' \). Then letting \( \|H\|_\infty = \max_{z \in \mathcal{Z}} \sum_{z' \in \mathcal{Z}'} |H(z, z')| \) denote the operator norm, for any \( x \in \mathbb{R}^d \), we obtain

\[
\|Y^T x\|_\infty = \|H Y(Z')^T x\|_\infty \leq \|H\|_\infty \|Y(Z')^T x\|_\infty \leq d \max_{z \in \mathcal{Z}, z' \in \mathcal{Z}'} |H(z, z')| \|Y(Z')^T x\|_\infty.
\]

Thus, it suffices to show \( |H(z, z')| \leq 1 \). Toward this end, for any \( y \in \mathbb{R}^d \), let \( \mathcal{Y}(Z', y) \) be the matrix that results from replacing the \( z' \)-th column of \( \mathcal{Y}(Z') \) with \( y \). Then since \( \mathcal{Y}(z) = \sum_{z'' \in \mathcal{Z}'} H(z, z'') \mathcal{Y}(z'') \), we have

\[
\mathcal{Y}(Z' \cup \{z\} \setminus \{z'\}) = \mathcal{Y}(Z', \mathcal{Y}(z)) = \mathcal{Y}\left(Z', \sum_{z'' \in \mathcal{Z}'} H(z, z'') \mathcal{Y}(z'')\right).
\]

Next, observe that \( \mathcal{Y}(Z', \mathcal{Y}(z'')) \) is rank deficient when \( z'' \neq z' \); otherwise, when \( z'' = z' \), we have \( \mathcal{Y}(Z', \mathcal{Y}(z'')) = \mathcal{Y}(Z') \). Hence, by multilinearity of the determinant, we obtain

\[
\det\left(\mathcal{Y}\left(Z', \sum_{z'' \in \mathcal{Z}'} H(z, z'') \mathcal{Y}(z'')\right)\right) = \sum_{z'' \in \mathcal{Z}'} H(z, z'') \det(\mathcal{Y}(Z', \mathcal{Y}(z''))) = H(z, z') \det(\mathcal{Y}(Z')).
\]

Combining the previous two identities with the definition of \( Z' \) yields the desired bound:

\[
|H(z, z')| = |\det(\mathcal{Y}(Z' \cup \{z\} \setminus \{z'\})|/|\det(\mathcal{Y}(Z'))| \leq 1.
\]

If instead \( r < d \), let \( \mathcal{Y} = U \Sigma V^T \) be the SVD. Then \( \mathcal{Y}^T U = V \Sigma^T = [\hat{Y}^T \ 0] \right), where \( \hat{Y} \in \mathbb{R}^{r \times \mathcal{Z}} \) has full rank. Hence, by the previous case, we can find \( Z' \subset \mathcal{Z} \) such that \( |Z'| = r \) and \( \|\hat{Y}^T \tilde{x}\|_\infty \leq r \|\hat{Y}(Z')^T \tilde{x}\|_\infty \) for any \( \tilde{x} \in \mathbb{R}^r \). Let \( U = [U_1 \ U_2] \) with \( U_1 \in \mathbb{R}^{d \times r} \). Then for any \( x \in \mathbb{R}^d \), if we let \( \tilde{x} = U_1^T x \in \mathbb{R}^r \), we obtain

\[
\mathcal{Y}^T x = \mathcal{Y}^T U U^T x = [\hat{Y}^T \ 0] \begin{bmatrix} \tilde{x} \\ U_2^T x \end{bmatrix} = \hat{Y}^T \tilde{x}.
\]

Therefore, by the choice of \( Z' \), we have

\[
\|\mathcal{Y}^T x\|_\infty = \|\hat{Y}^T \tilde{x}\|_\infty \leq r \|\hat{Y}(Z')^T \tilde{x}\|_\infty = r \|\mathcal{Y}(Z')^T x\|_\infty.
\]

We can now show that \( U_t \) converges at rate \( 1/n \) with a constant depending only \( B_t + 1 \) and \( d \).

**Lemma 3 (U_t convergence).** If Assumptions 1 and 2 hold, then for any \( t \in [T] \) and \( n \in \mathbb{N} \),

\[
\max_{(s,a) \in S \times A} |\phi(s,a)^T (U_t^n 0 - U_t^{n-1} 0)| \leq \frac{(B_t + 1)d^2}{n}.
\]

**Proof.** Suppose instead that for some \( n \in \mathbb{N} \), we have

\[
\max_{(s,a) \in S \times A} |\phi(s,a)^T (U_t^n 0 - U_t^{n-1} 0)| > \frac{(B_t + 1)d^2}{n}.
\]
Then combining Claims 7 and 8, we can find \( Z' \subset S \times A \) such that, for any \( m \in [n] \),
\[
\frac{(B_t + 1)|Z'|}{n} \leq \frac{(B_t + 1)d^2}{nd} < \frac{1}{d} \max_{(s,a) \in S \times A} |\phi(s,a)^T(U_t^m 0 - U_t^{m-1})| \leq \max_{(s,a) \in Z'} |\phi(s,a)^T(U_t^m 0 - U_t^{m-1})|.
\]

Thus, for each \( m \in [n] \), we can find \( z_m \in Z' \) with \(|\phi(z_m)^T(U_t^m 0 - U_t^{m-1})| > (B_t + 1)|Z'|/n \). But by Claim 7,
\[
n = \sum_{z \in Z'} \sum_{m=1}^n 1(z_m = z) < \frac{n}{(B_t + 1)|Z'|} \sum_{z \in Z'} \sum_{m=1}^n \phi(z)^T(U_t^m 0 - U_t^{m-1}) = \frac{n}{(B_t + 1)|Z'|} \sum_{z \in Z'} \phi(z)^TU_t^0 \leq n,
\]
which is a contradiction. \( \square \)

C Proof of Theorem 1

In this appendix, we prove Theorem 1 in two steps. First, in Appendix C.1, we show that OAFPs exist on the event \( \mathcal{E} \). Second, in Appendix C.2, we prove the regret bound on the intersection of \( \mathcal{E} \) and an event \( \mathcal{F} \) defined in Lemma 7 that occurs with probability at least \( 1 - \delta/2 \). Thus, by the union bound and Lemma 2, OAFPs exist and the regret bound holds with probability at least \( 1 - \delta \), which establishes the theorem.

C.1 Existence of OAFPs

We begin with optimism lemma for the operator \( U_t \).

**Lemma 4** (\( U_t \) optimism). Under the assumptions of Theorem 1, for any \( t \in [T] \), \( n \in \mathbb{N} \), and \((s,a) \in S \times A\),
\[
\phi(s,a)^TU_t^{n-1} \leq Q^*(s,a).
\]

**Proof.** We fix \( t \) and use induction on \( n \). For \( n = 1 \), we simply have \( \phi(s,a)^TU_t^{n-1} = 0 \leq Q^*(s,a) \).
Assuming true for \( n \in \mathbb{N} \), the Bellman optimality equation (1) implies
\[
f_t(s,U_t^{n-1}) = \min_{a \in A} \left( \phi(s,a)^TU_t^{n-1} - \alpha_t||\phi(s,a)||_{\Lambda_t} \right) \leq \min_{a \in A} Q^*(s,a) = J^*(s) \forall s \in S.
\]
Hence, by Claim 2, \( g_t(s,U_t^{n-1}) \leq \max\{J^*(s),0\} = J^*(s) \). Again using Bellman optimality, we thus obtain
\[
\phi(s,a)^TU_t^n = c(s,a) + \sum_{s' \in S} g_t(s',U_t^{n-1})P(s'|s,a) \leq c(s,a) + \sum_{s' \in S} J^*(s')P(s'|s,a) = Q^*(s,a). \square
\]

We now establish existence of OAFPs. First note that by Claim 3 and Lemma 3, for any norm \( \| \cdot \| \), we have
\[
\|U_t^{n+1} 0 - U_t^n 0\| \leq \sqrt{d} \max_{s \in S} |g_t(s,U_t^n 0) - g_t(s,U_t^{n-1})| \sqrt{d} \leq \max_{(s,a) \in S \times A} |\phi(s,a)^T(U_t^n 0 - U_t^{n-1})| \frac{\sqrt{d} \delta}{n} \to 0.
\]

24
Hence, \( w^*_t = \lim_{n \to \infty} U^n_t 0 \) exists. By continuity, it is a fixed point:

\[
w^*_t = \lim_{n \to \infty} U^{n+1}_t 0 = \lim_{n \to \infty} U_t(U^n_t 0) = U_t \left( \lim_{n \to \infty} U^n_t 0 \right) = U_t(w^*_t).
\]

Thus, by a standard norm equivalence and Claim 4, we have

\[
\|w^*_t\|_2 \leq \|U_t w^*_t\|_2 \leq \sqrt{d} \left( 1 + \max_{s \in S} g(s, w^*_t) \right) \leq \sqrt{d}(1 + B_t) \leq \alpha_t + \sqrt{td}(B_t + 1) = W_t.
\]

By Claim 6, on the event \( E \), this implies

\[
\|\hat{G}_t w^*_t - w^*_t\|_{\Lambda_t} = \|\hat{G}_t w^*_t - U_t w^*_t\|_{\Lambda_t} = \|E_t w^*_t\|_{\Lambda_t} \leq \varepsilon_t \leq \alpha_t.
\]

Finally, for any \( s \in S \), using continuity, Lemma 4, and Bellman optimality, we obtain

\[
f_t(s, w^*_t) \leq \min_{a \in A} \phi(s, a)^T w^*_t = \min_{a \in A} \phi(s, a)^T U^n_t 0 = \min_{a \in A} \phi(s, a)^T U^n_t 0 \leq \min_{a \in A} Q^*(s, a) = J^*(s).
\]

Hence, on the event \( E \), for any \( t \in [T] \), \( w^*_t \) is an OAFP by the previous two inequalities and Definition 1.

### C.2 Regret bound

Recall from Algorithm 1 that \( M_l \) is the time the \( l \)-th interval ended and \( L \) is the total number of intervals completed. Fix \( \tilde{T} \in \mathbb{N} \) and let \( \bar{L} = \min\{l \in [L] : M_l \geq T \wedge \tilde{T} \} \) denote the least number of intervals that encompass the times \( 1, \ldots, T \wedge \tilde{T} \). Also let \( \bar{K} = |\{t \in [T \wedge \tilde{T}] : s'_t = s_{goal}\}| \) denote the number of episodes completed by time \( T \wedge \tilde{T} \). Finally, define the regret incurred up to time \( T \wedge \tilde{T} \) by

\[
\tilde{R}(\tilde{T}) = \sum_{t=1}^{T \wedge \tilde{T}} c(s_t, a_t) - \sum_{k=1}^{\bar{K}} J^*(s^k_t) < \infty.
\]

**Lemma 5** (Regret decomposition). Under the assumptions of Theorem 1,

\[
\tilde{R}(\tilde{T}) \leq \sum_{l=1}^{\bar{L} - 1} \left( \sum_{t=1+M_l}^{M_{l+1}} c(s_t, a_t) - J^*(s_{1+M_l}) \right) + 2(B_* + 1)d \log_2(4B_*(T \wedge \tilde{T})/c_{min}).
\]

**Proof.** Since \( M_1 = 1 \) in Algorithm 1 and \( c(s_1, a_1) \leq 1 \), we can bound the total cost incurred by

\[
\sum_{t=1}^{T \wedge \tilde{T}} c(s_t, a_t) \leq 1 + \sum_{t=2}^{T \wedge \tilde{T}} c(s_t, a_t) = 1 + \sum_{l=1}^{\bar{L} - 1} \sum_{t=1+M_l}^{M_{l+1}} c(s_t, a_t),
\]

where the equality holds because \( M_{l+1} \wedge \tilde{T} = M_{l+1} \) for \( l < \bar{L} - 1 \) and \( M_{\bar{L}} \wedge \tilde{T} = T \wedge \tilde{T} \) by definition. On the other hand, the expected cost for the optimal policy can be written as

\[
\sum_{k=1}^{\bar{K}} J^*(s^k_t) = \sum_{l=1}^{\bar{L}} J^*(s_{1+M_l}) + \left( \sum_{k=1}^{\bar{K}} J^*(s^k_t) - \sum_{l=1}^{\bar{L}} J^*(s_{1+M_l}) \right).
\]

25
Thus, we seek a lower bound for the term in parentheses. First note that since Algorithm 1 ends an interval each time an episode ends, for each \(k \in [K]\), we can find \(l \in \mathbb{L}\) such that \(s_l^k = s_{1+M_t}\). Hence, all summands cancel, except those corresponding to intervals \(\mathcal{L} = \{l \in \mathbb{L} : M_l = 1\) or \(B_{M_l} = 2B_{M_l-1}\) or \(det(\Lambda_{M_l}) \geq 2det(\Lambda_{M_l-1})\)\), which may not have reached the goal state. Taken together, and since \(J^*(s_{1+M_t}) \leq B_\ast\), we obtain

\[
\sum_{k=1}^{K} J^*(s_l^k) - \sum_{l=1}^{\hat{\mathbb{L}}} J^*(s_{1+M_t}) \geq -\sum_{l \in \mathcal{L}} J^*(s_{1+M_t}) \geq -B_\ast |\mathcal{L}|. \tag{26}
\]

It remains to bound \(|\mathcal{L}|\). Clearly, \(|\mathcal{L}| \leq 1 + \sum_{i=1}^{2} |\mathcal{L}_i|\), where \(\mathcal{L}_1 = \{l \in \mathbb{L} : B_{M_l} = 2B_{M_l-1}\}\) and \(\mathcal{L}_2 = \{l \in \mathbb{L} : det(\Lambda_{M_l}) \geq 2det(\Lambda_{M_l-1})\}\). For \(\mathcal{L}_1\), note \(\sup_{t \geq 0} B_t \geq 2^{\mathcal{L}_1} c_{min}\) in Algorithm 1, so by Claim 5,

\[
|\mathcal{L}_1| = \log_2 \left(2^{\mathcal{L}_1} c_{min}/c_{min}\right) \leq \log_2 \left(\sup_{t \geq 0} B_t/c_{min}\right) \leq \log_2 (2B_\ast/c_{min}). \tag{27}
\]

For \(\mathcal{L}_2\), we have \(|\mathcal{L}_2| \leq |\mathcal{L}_2'| + 1\), where \(\mathcal{L}_2' = \{l \in \mathbb{L} - 1 : det(\Lambda_{M_l}) \geq 2det(\Lambda_{M_l-1})\}\) excludes \(\hat{\mathbb{L}} - 1\) if it belongs to \(\mathcal{L}_2\). By definition, \(M_{l-1} < T \wedge \hat{T}\), which by Claim 1 implies \(det(\Lambda_{M_{l-1}}) \leq (1 + (T \wedge \hat{T})d \leq (2(T \wedge \hat{T})d). Hence, because \(det(\Lambda_{M_{l-1}}) \geq 2^{\mathcal{L}_2'} det(\Lambda_{0}) = 2^{\mathcal{L}_2'}\) by definition of \(\mathcal{L}_2'\), we obtain

\[
|\mathcal{L}_2| \leq \log_2 (2^{\mathcal{L}_2'}) + 1 \leq \log_2 (det(\Lambda_{M_{l-1}})) + 1 \leq d \log_2 (2(T \wedge \hat{T})) + 1. \tag{28}
\]

Recalling \(|\mathcal{L}| \leq 1 + \sum_{i=1}^{2} |\mathcal{L}_i|\) and combining (24), (25), (26), (27), and (28), we obtain

\[
\tilde{R}(\hat{T}) - \sum_{l=1}^{\hat{\mathbb{L}}} \sum_{t=1+M_l}^{M_l \wedge \hat{T}} c(s_t, a_t) - J^*(s_{1+M_l}) \leq 1 + B_\ast(\log_2 (2B_\ast/c_{min}) + d \log_2 (2(T \wedge \hat{T})) + 2) \leq 2(B_\ast + 1)d \log_2 (4B_\ast(T \wedge \hat{T})/c_{min}),
\]

where the last inequality uses \(B_\ast \geq c_{min}\) and \(d \geq 2\). \(\square\)

We next bound the summand in Lemma 5 by a martingale difference sequences and sum of bonuses.

**Lemma 6** (Per-interval regret). Under the assumptions of Theorem 1 and on the event \(\mathcal{E}\), for any \(l \in [\hat{\mathbb{L}} - 1]\),

\[
\sum_{t=1+M_l}^{M_l \wedge \hat{T}} c(s_t, a_t) - J^*(s_{1+M_l}) \leq \sum_{t=1+M_l}^{M_l \wedge \hat{T}} \left( g_{M_l}(s'_t, w_{M_l}) - \mathbb{E}_t[g_{M_l}(s'_t, w_{M_l})] \right) + 3\alpha_{M_l} \sum_{t=1+M_l}^{M_l \wedge \hat{T}} \|\phi(s_t, a_t)\|_{\Lambda_{M_l}}^{-1}.
\]

**Proof.** Define \(\gamma_\tau = \sum_{t=\tau}^{M_l \wedge \hat{T}} c(s_t, a_t) - f_{M_l}(s_{\tau}, w_{M_l})\) for each \(\tau \in \{1 + M_l, \ldots, M_l \wedge \hat{T}\}\) and \(\gamma_{1+M_l \wedge \hat{T}} = 0\). We claim, and will return to prove, that for any \(\tau \in \{1 + M_l, \ldots, M_l \wedge \hat{T}\},\)

\[
\gamma_\tau \leq \gamma_{\tau+1} + g_{M_l}(s'_\tau, w_{M_l}) - \mathbb{E}_\tau[g_{M_l}(s'_\tau, w_{M_l})] + 3\alpha_{M_l} \|\phi(s_{\tau}, a_{\tau})\|_{\Lambda_{M_l}}^{-1}. \tag{29}
\]

Assuming (29) holds, we prove the lemma. First, since \(w_{M_l}\) is an OAFP, Definition 1 implies

\[
f_{M_l}(s_{1+M_l}, w_{M_l}) \leq J^*(s_{1+M_l}), \quad \|\tilde{G}_{M_l} w_{M_l} - w_{M_l}\|_{\Lambda_{M_l}} \leq \alpha_{M_l}. \tag{30}
\]
Using the first inequality in (30), we obtain

\[ \sum_{t=1+M_i}^{M_{i+1} \wedge \hat{T}} c(s_t, a_t) - J^*(s_{1+M_i}) \leq \sum_{t=1+M_i}^{M_{i+1} \wedge \hat{T}} c(s_t, a_t) - f_{M_i}(s_{1+M_i}, w_{M_i}) = \gamma_{1+M_i}, \]

so the lemma follows from recursively applying (29). Hence, it only remains to prove (29). First observe

\[
\begin{align*}
  f_{M_i}(s_t, w_{M_i}) &= \phi(s_t, a_t)^T \hat{G}_{M_i} w_{M_i} + \phi(s_t, a_t)^T (w_{M_i} - \hat{G}_{M_i} w_{M_i}) - \alpha_{M_i} \|\phi(s_t, a_t)\|_{\Lambda_{M_i}^{-1}} \\
  &\geq \phi(s_t, a_t)^T \hat{G}_{M_i} w_{M_i} - 2\alpha_{M_i} \|\phi(s_t, a_t)\|_{\Lambda_{M_i}^{-1}} \geq \phi(s_t, a_t)^T U_{M_i} w_{M_i} - 3\alpha_{M_i} \|\phi(s_t, a_t)\|_{\Lambda_{M_i}^{-1}},
\end{align*}
\]

where the equality holds by the policy update in Algorithm 1 and the inequalities use Cauchy-Schwarz, the second bound in (30), and Corollary 2. Now by definition, we have

\[
\phi(s_t, a_t)^T U_{M_i} w_{M_i} = c(s_t, a_t) + \sum_{s \in S} g_{M_i}(s, w_{M_i}) P(s|s_t, a_t)
\]

\[
= c(s_t, a_t) + \sum_{s \in S} g_{M_i}(s', w_{M_i}) + \mathbb{E}_r[g_{M_i}(s'_t, w_{M_i})] - g_{M_i}(s'_t, w_{M_i}).
\]

Combining the previous two inequalities and rearranging, we obtain

\[
c(s_t, a_t) - f_{M_i}(s_t, w_{M_i}) \leq -g_{M_i}(s'_t, w_{M_i}) + g_{M_i}(s'_t, w_{M_i}) - \mathbb{E}_r[g_{M_i}(s'_t, w_{M_i})] + 3\alpha_{M_i} \|\phi(s_t, a_t)\|_{\Lambda_{M_i}^{-1}}.
\]

We complete the proof separately in each of two cases.

- If \( \tau = M_{i+1} \wedge \hat{T} \), the left side of (31) is \( \gamma_\tau \) and \(-g_{M_i}(s'_t, w_{M_i}) \leq 0 = \gamma_{\tau+1} \), so (31) implies (29).

- Otherwise, an interval did not end between times \( 1 + M_i \) and \( \tau \) (inclusive). This implies (A) \( s'_t \neq s_{goal} \), so \( s'_t = s_{\tau+1} \), (B) \( B_\tau = B_{M_i} \), and (C) \( f_{M_i}(s'_t, w_{M_i}) \leq B_\tau \). Taken together, (B) and (C) give \( f_{M_i}(s'_t, w_{M_i}) \leq B_{M_i} \), so (D) \(-g_{M_i}(s'_t, w_{M_i}) \leq -f_{M_i}(s'_t, w_{M_i}) \) by definition. Combining (A) and (D) with (31), we obtain

\[
c(s_t, a_t) - f_{M_i}(s_t, w_{M_i}) \leq -f_{M_i}(s_{\tau+1}, w_{M_i}) + g_{M_i}(s'_t, w_{M_i}) - \mathbb{E}_r[g_{M_i}(s'_t, w_{M_i})] + 3\alpha_{M_i} \|\phi(s, a)\|_{\Lambda_{M_i}^{-1}}.
\]

Hence, recalling \( \tau < M_{i+1} \wedge \hat{T} \), we can use the definitions of \( \gamma_\tau \) and \( \gamma_{\tau+1} \) to obtain

\[
\begin{align*}
  \gamma_\tau &= \sum_{t=\tau+1}^{M_{i+1}} c(s_t, a_t) + c(s_t, a_t) - f_{M_i}(s_t, w_{M_i}) \\
  &\leq \sum_{t=\tau+1}^{M_{i+1}} c(s_t, a_t) - f_{M_i}(s_{\tau+1}, w_{M_i}) + g_{M_i}(s'_t, w_{M_i}) - \mathbb{E}_r[g_{M_i}(s'_t, w_{M_i})] + 3\alpha_{M_i} \|\phi(s, a)\|_{\Lambda_{M_i}^{-1}} \\
  &= \gamma_{\tau+1} + g_{M_i}(s'_t, w_{M_i}) - \mathbb{E}_r[g_{M_i}(s'_t, w_{M_i})] + 3\alpha_{M_i} \|\phi(s, a)\|_{\Lambda_{M_i}^{-1}}.
\end{align*}
\]

We next bound the martingale difference sequence from Lemma 6.
Lemma 7 (Martingale difference sequence). Under the assumptions of Theorem 1, for any δ > 0, if we define
\[ \mathcal{F} = \left\{ \sum_{t=1}^{M_{i+1} \wedge T} \sum_{t=1+M_i} \left( g_{M_i}(s'_t, w_{M_i}) - \mathbb{E}_t[g_{M_i}(s'_t, w_{M_i})] \right) \leq 2B_* \sqrt{(T \wedge \tilde{T}) \log \frac{8(T \wedge \tilde{T})}{\delta}} \right\}, \]
then \( \mathbb{P}(\mathcal{F}) \geq 1 - \delta/2. \)

Proof. The left side of the inequality is a martingale difference sequence. By definition and Claim 5, each term satisfies \( g_{M_i}(s'_t, w_{M_i}) - \mathbb{E}_t[g_{M_i}(s'_t, w_{M_i})] \leq B_t \leq 2B_* \). The number of terms is \( M_\lambda \wedge \tilde{T} \leq M_L \wedge \tilde{T} = T \wedge \tilde{T} \). The lemma follows from (Rosenberg et al., 2020, Theorem D.1) (an anytime version of Azuma’s inequality).

Finally, we bound the sum of bonuses from Lemma 6.

Lemma 8 (Sum of bonuses). Under the assumptions of Theorem 1,

\[ \sum_{t=1}^{M_{i+1} \wedge T} \sum_{t=1+M_i} \|\phi(s_t, a_t)\|_{\Lambda_{M_i}^{-1}} \leq 6\sqrt{(T \wedge \tilde{T})d \log(2(T \wedge \tilde{T}))} \max_{t \in [T \wedge \tilde{T}]} \alpha_t. \]

Proof. For each \( l \in [\tilde{L} - 1] \) and \( t \in \{2 + M_l, \ldots, M_{l+1} \wedge \tilde{T}\} \), the \((l+1)\)-th interval did not end at time \( t - 1 \), which implies \( \text{det}(\Lambda_{l-1}) \leq 2\text{det}(\Lambda_{M_l}) \). By (Abbasi-Yadkori et al., 2011, Lemma 12) this implies \( \Lambda_{M_l} - \Lambda_{l-1}/2 \) is positive semidefinite, so \( \Lambda_{M_l}^{-1} - 2\Lambda_{l-1}^{-1} \) is negative semidefinite, so
\[ \|\phi(s_t, a_t)\|_{\Lambda_{M_l}^{-1}} = \sqrt{\phi(s_t, a_t)^T \Lambda_{M_l}^{-1} \phi(s_t, a_t)} \leq \sqrt{2\phi(s_t, a_t)^T \Lambda_{l-1}^{-1} \phi(s_t, a_t)} = \sqrt{2}\|\phi(s_t, a_t)\|_{\Lambda_{l-1}^{-1}}. \]

For any \( l \in [\tilde{L} - 1] \), the inequality clearly holds at time \( t = 1 + M_l \) as well. Combined with and Cauchy-Schwarz and the fact that \( M_{\lambda} \wedge \tilde{T} \leq M_L \wedge \tilde{T} = T \wedge \tilde{T} \) by definition, we thus obtain
\[ \sum_{l=1}^{\tilde{L} - 1} \sum_{t=1}^{M_{l+1} \wedge T} \|\phi(s_t, a_t)\|_{\Lambda_{M_l}^{-1}} \leq 3\sqrt{2} \max_{t \in [T \wedge \tilde{T}]} \alpha_t \sum_{l=1}^{T \wedge \tilde{T}} \|\phi(s_t, a_t)\|_{\Lambda_{l-1}^{-1}} \leq 3\sqrt{2} \max_{t \in [T \wedge \tilde{T}]} \alpha_t \left( T \wedge \tilde{T} \sum_{t=1}^{T \wedge \tilde{T}} \|\phi(s_t, a_t)\|_{\Lambda_{l-1}^{-1}}^2 \right). \]

Finally, by (Abbasi-Yadkori et al., 2011, Lemma 11) and Claim 1, we have
\[ \sum_{t=1}^{T \wedge \tilde{T}} \|\phi(s_t, a_t)\|_{\Lambda_{l-1}^{-1}}^2 \leq 2\log \frac{\text{det}(\Lambda_{T \wedge \tilde{T}})}{\text{det}(\Lambda_0)} \leq 2\log((T \wedge \tilde{T}) + 1) \leq 2d \log(2(T \wedge \tilde{T})). \]

We can now prove the regret bound on \( \mathcal{E} \cap \mathcal{F} \). By Lemmas 5, 6, 7, and 8, we know
\[ \hat{R}(\tilde{T}) \leq 6 \max_{t \in T \wedge \tilde{T}} \alpha_t \sqrt{(T \wedge \tilde{T})d \log(2(T \wedge \tilde{T}))} + 2B_* \sqrt{(T \wedge \tilde{T}) \log(8(T \wedge \tilde{T})/\delta)} + 2(B_* + 1)d \log_2(4B_*(T \wedge \tilde{T})/c_{\text{min}}) \] (32)
Next, recall \( B_t + 1 \leq 2B_\ast + 1 \leq 2(B_\ast + 1) \) by Claim 5. Combined with the assumption that \( \kappa_t \leq \Psi t^\lambda \log(t + 1) \) with \( \Psi \geq 9d \) and \( \lambda \in [0, \frac{1}{2}) \) for any \( t \in [T \land \bar{T}] \), we have

\[
\alpha_t = (B_t + 1)\kappa_t \sqrt{\log(t(B_t + 1)\kappa_t/\delta)} \leq 2(B_\ast + 1)\Psi t^\lambda \log(t + 1) \sqrt{\log(2(B_\ast + 1)\Psi t^{1+\lambda} \log(t + 1)/\delta)}.
\]

Since \( (B_\ast + 1)\Psi t^\gamma \log(t + 1)/\delta \geq 9 \log 2 \geq 1 \), \( \log(t + 1) \leq t \), and \( \gamma \leq 1/2 \), we also have

\[
t + 1 \leq 2t \leq 2(B_\ast + 1)\Psi t^{1+\lambda} \log(t + 1)/\delta \leq 2(B_\ast + 1)\Psi t^{5/2}/\delta \leq (2(B_\ast + 1)\Psi t/\delta)^{5/2}.
\]

Hence, combining the previous two inequalities, we obtain

\[
6\alpha_t \leq 6 \cdot 2(B_\ast + 1)\Psi t^\gamma (5/2) \log(2(B_\ast + 1)\Psi t/\delta) \sqrt{(5/2) \log(2(B_\ast + 1)\Psi t/\delta)}
\]

\[
= (30 \sqrt{5/2})(B_\ast + 1)\Psi t^\gamma \log^{3/2}(2(B_\ast + 1)\Psi t/\delta) < 48(B_\ast + 1)\Psi t^\gamma \log^{3/2}(2(B_\ast + 1)\Psi t/\delta)
\]

Since the right side is increasing in \( t \), the first summand in (32) can thus be upper bounded by

\[
50 \bar{R}(\bar{T}) \leq 50 \bar{R}(\bar{T}).
\]

Finally, \( \Psi \geq 9d \) implies the other summands in (32) are bounded by \( 50 \bar{R}(\bar{T}) \), so \( \bar{R}(\bar{T}) \leq 50 \bar{R}(\bar{T}) \).

Next, we show \( R(K) = \bar{R}(\bar{T}) \) when \( \bar{T} \) is large enough. Toward this end, first note that by Assumption 1 and definition of \( \bar{R}(\bar{T}) \) and \( \bar{K} \), the bound \( \bar{R}(\bar{T}) \leq 50 \bar{R}(\bar{T}) \) from the previous paragraph implies

\[
(T \land \bar{T})c_{\min} \leq \sum_{t=1}^{T \land \bar{T}} c(s_t, a_t) = \bar{R}(\bar{T}) + \sum_{k=1}^{\bar{K}} J^*(s_t^k) \leq 50 \bar{R}(\bar{T}) + KB_\ast. \tag{34}
\]

Now consider two cases. First, if \( T \land \bar{T} \geq 100 \bar{R}(\bar{T})/c_{\min} \), then \( 50 \bar{R}(\bar{T}) \leq (T \land \bar{T})c_{\min}/2 \), so \( T \land \bar{T} \leq 2KB_\ast/c_{\min} \) by (34). Otherwise, \( T \land \bar{T} \leq 100 \bar{R}(\bar{T})/c_{\min} \), which by definition (33) implies

\[
T \land \bar{T} \leq \left(100(B_\ast + 1)\sqrt{d\Psi/c_{\min}} \right) (T \land \bar{T})^{\frac{1}{2}+\lambda} \log^2 \left((2(B_\ast + 1)\Psi/(c_{\min}) |T \land \bar{T}\right).
\]

By Claim 9 below, this implies that for some \( t_1, t_2 > 0 \) depending only on \( \lambda \) (which, by assumption on \( \lambda \), means that \( t_1, t_2 \) are absolute constants), we have

\[
T \land \bar{T} \leq T_1 \triangleq \left(\frac{t_1(B_\ast + 1)\sqrt{d\Psi}}{c_{\min}} \log^2 \left(\frac{t_2(B_\ast + 1)d\Psi}{c_{\min}}\right)\right)^{\frac{1}{1-2\lambda}}.
\]

Combining the cases, we conclude \( T \land \bar{T} \leq T_2 \triangleq \max\{2KB_\ast/c_{\min}, T_1\} \leq 50 \bar{R}(\bar{T}) \). Hence, choosing \( \bar{T} \geq T_2 \), we obtain \( T \land \bar{T} = T \) and \( \bar{K} = K \), which together imply \( R(K) = \bar{R}(\bar{T}) < \infty \).

Finally, we establish the bounds of the theorem. Recall we have shown \( R(K) = \bar{R}(\bar{T}) \leq 100 \bar{R}(\bar{T}) \) for large \( \bar{T} \) and \( T \land \bar{T} \leq \max\{2KB_\ast/c_{\min}, T_1\} \). We again consider two cases. First, if the bound \( T \land \bar{T} \leq 2KB_\ast/c_{\min} \) holds, then by \( R(K) = O(\bar{R}(\bar{T})) \) and the definition (33),

\[
R(K) = \hat{O} \left(\left(\frac{B_\ast^{\frac{3}{2}+\lambda} \ast + B_\ast^{\frac{1}{2}+\lambda}}{d^{\frac{1}{2}}\Psi(K/c_{\min})^{\frac{1}{2}+\lambda}}\right)\right). \tag{35}
\]
If instead only $T \wedge \hat{T} \leq T_1$ holds, then (33) implies
\[
R(K) = \tilde{O} \left( (B_\star + 1)\sqrt{d\Psi T_1^{1+1}} \right) = \tilde{O} \left( (B_\star + 1) \frac{2}{1-2\kappa} d^{1/2} \Psi^{1/2} \kappa^{-1/2} \right).
\]
Finally, bounding $R(K)$ by the max of the cases, then the max by the sum, yields the desired bound.

**Remark 13** (Bound for $\hat{T}$). As shown above, for $\hat{T} \geq T_2$, we have the bound $T \leq T_2$, or (by definition)
\[
T = O \left( \max \left\{ \frac{2KB_\star}{c_{\min}}, \left( \frac{(B_\star + 1)\sqrt{d\Psi}}{c_{\min}} \log^2 \left( \frac{(B_\star + 1)d\Psi}{c_{\min}^2} \right) \right) \right\} \right).
\]

**Remark 14** (Sharpening the tabular case). When $K$ is large, the bound (35) holds and has $\sqrt{d}\Psi$ dependence on $d$. We required $\Psi$ (the constant component of $\kappa_t$) to be linear in $d$ in order to match the scaling of the error bound $\varepsilon_t$ from Appendix B.3. In the tabular case, we can reduce $\varepsilon_t$’s dependence to $\sqrt{d}$, (see Remark 16 in Appendix E.1), so we can choose $\Psi$ to scale as $\sqrt{d}$, after which the regret bound’s dependence becomes linear in $d$.

**Claim 9.** Suppose $x \leq ax^{c_1} \log^2(bx)$ for some $c_1 \in (0,1)$ and $a,b,x \geq 1$. Then $x \leq (c_2 a \log^2(c_3 ab))^{1/(1-c_1)}$ for some constants $c_2,c_3 > 0$ that depend only on $c_1$.

**Proof.** By the assumed inequality and log $y \leq y \forall y \geq 0$, we have
\[
x \leq \frac{16ax^{c_1}}{(1-c_1)^2} \left( \frac{1-c_1}{4} \log(bx) \right)^2 = \frac{16ax^{c_1}}{(1-c_1)^2} \left( \log \left( \frac{bx^{1-c_1}}{1+c_1} \right) \right)^2 \leq \frac{16ax^{c_1}}{(1-c_1)^2} \left( \frac{bx^{1-c_1}}{2} \right)^2 = \frac{16a}{(1-c_1)^2} x^{1+c_1}.
\]
Solving for $x$, we obtain $x \leq (16a/(1-c_1)^2)^{2/(1-c_1)} b$. Plugging back into the log term of the assumed inequality, and since $2 \leq 2/(1-c_1)$, we obtain
\[
x \leq ax^{c_1} \left( \log \left( \left( \frac{16a}{(1-c_1)^2} \right)^{2/(1-c_1)} b^2 \right) \right)^2 \leq ax^{c_1} \left( \frac{2}{1-c_1} \log \left( \frac{16ab}{(1-c_1)^2} \right) \right)^2 = c_2 a \log^2(c_3 ab)x^{c_1},
\]
where we define $c_2 = 4/(1-c_1)^2$ and $c_3 = 16/(1-c_1)^2$. Solving for $x$ gives the desired bound. □

### D Proofs of Theorems 2-4

We begin with an optimism lemma used for all three proofs.

**Lemma 9** ($\hat{G}_t$ optimism). If Assumptions 1 and 2 hold and $\kappa_t \geq 9d$, then on the event $\mathcal{E}$,
\[
\phi(s,a)^T \hat{G}_t^{n-1} - Q^*(s,a) \leq \varepsilon_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \leq \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \forall (s,a) \in \mathcal{S} \times \mathcal{A}, n \in \mathbb{N}, t \in [T].
\]

**Proof.** The proof is similar to that of Lemma 4. For $n = 1$, $\phi(s,a)^T \hat{G}_t^{1-1} = 0 \leq Q^*(s,a)$, so the result is immediate. Now assume the bound holds for $n \in \mathbb{N}$. Then by the Bellman optimality equation (1), we obtain
\[
f_t(s,\hat{G}_t^{n-1}) = \min_{a \in \mathcal{A}} \left( \phi(s,a)^T \hat{G}_t^{n-1} - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \right) \leq \min_{a \in \mathcal{A}} Q^*(s,a) = J^*(s).
\]
Hence, by Claim 3, \( g_t(s, \hat{G}^{n-1}_t) \leq J^*(s) \) as well. Again by Bellman optimality, this implies
\[
\phi(s, a)^T U_t(\hat{G}^{n-1}_t) = c(s, a) + \sum_{s' \in S} g_t(s, \hat{G}^{n-1}_t) P(s'|s, a) \leq c(s, a) + \sum_{s' \in S} J^*(s') P(s'|s, a) = Q^*(s, a).
\]
On the other hand, by Corollary 3 and the assumption \( \kappa_t \geq 9d \), we have
\[
\phi(s, a)^T (\hat{G}^{n-1}_t - U_t(\hat{G}^{n-1}_t)) \leq \varepsilon_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}}.
\]
Combining the last two inequalities completes the inductive step. \( \square \)

As a simple corollary, we have the following formal version of Lemma 1 from the main text.

**Corollary 4** (\( \hat{G}_t \) optimism). If Assumptions 1 and 2 hold and \( \kappa_t \geq 9d \), then on the event \( \mathcal{E} \), for any \( t \in [T] \), \( n \in \mathbb{N} \), and \( (s, a) \in \mathcal{S} \times \mathcal{A} \), we have \( f_t(s, \hat{G}^{n-1}_t) \leq J^*(s) \).

**Proof.** Rearrange the Lemma 9 bound and take minimum over \( a \in \mathcal{A} \) as in (36). \( \square \)

The preceding corollary implies the optimism inequality in (9). For the fixed point inequality \( \|\hat{G}^0_t - \hat{G}^{n-1}_t\|_{\Lambda_t} \leq \alpha_t \), we use a similar approach for Theorems 2 and 3, so we will provide a general result (Lemma 12 below) for use in both theorems. Toward this end, we begin with an intermediate claim. Note that, while the bound grows with \( n \) for fixed \( t \), we will later choose \( n \) in terms of \( t \) so that the bound vanishes as \( t \to \infty \).

**Claim 10** (\( \hat{G}_t \) tracks \( U_t \)). If Assumptions 1 and 2 hold and \( \kappa_t \geq 18d \), then on the event \( \mathcal{E} \),
\[
|\phi(s, a)^T (\hat{G}^{n-1}_t - U_t^{n-1})| \leq \varepsilon_t \|\phi(s, a)\|_{\Lambda_t^{-1}} + \frac{2(B_t + 1)(n - 1)\varepsilon_t}{\alpha_t} \forall (s, a) \in \mathcal{S} \times \mathcal{A}, n \in \mathbb{N}, t \in [T]
\]

**Proof.** First, for any \( n \in \mathbb{N} \), we use Corollary 3 and \( \kappa_t \geq 18d \) to write
\[
\phi(s, a)^T \hat{G}^n_t = \phi(s, a)^T U_t(\hat{G}^{n-1}_t) + \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}}/2.
\]
By Claim 4, we also know \( \phi(s, a)^T U_t(\hat{G}^{n-1}_t) \leq B_t + 1 \); combined with the previous inequality, we obtain
\[
\phi(s, a)^T \hat{G}^n_t - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq B_t + 1 - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}}/2.
\]
(37)

Now define the “explored” states at time \( t \) (i.e., those with small bonuses across actions) by
\[
\mathcal{S}_t = \left\{ s \in \mathcal{S} : \max_{a \in \mathcal{A}} \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq \frac{2(B_t + 1)}{\alpha_t} \right\}.
\]

Then for any unexplored state \( s \in \mathcal{S} \setminus \mathcal{S}_t \), (37) implies that for some \( a \in \mathcal{A} \),
\[
\phi(s, a)^T \hat{G}^n_t - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq B_t + 1 - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq -(B_t + 1) < 0.
\]
Taking minimum over \( a \in \mathcal{A} \) on both sides gives \( f_t(s, \hat{G}^n_t) \leq 0 \), which implies \( g_t(s, \hat{G}^n_t) = 0 \). Again using Claim 4, we similarly obtain that for any \( s \in \mathcal{S} \setminus \mathcal{S}_t \) and some \( a \in \mathcal{A} \),
\[
\phi(s, a)^T U_t^n \leq B_t + 1 - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \leq -(B_t + 1) < 0,
\]

31
so \( g_t(s, U_t^n) = 0 \) as well. Since \( n \in \mathbb{N} \) was arbitrary and \( g_t(s, \hat{G}_t^n) = g_t(s, U_t^n) = g(s, 0) = 0 \), we conclude

\[
g_t(s, \hat{G}_t^{n-1}) = g_t(s, U_t^{n-1}) = 0 \quad \forall \, s \in S \setminus S_t, \, n \in \mathbb{N}.
\]

Combined with Claim 3, this implies that for any \( n \in \mathbb{N} \) and \( (s, a) \in S \times A \),

\[
|\phi(s, a)^T(U_t(\hat{G}_t^{n-1}) - U_t(U_t^{n-1}))| \leq \sum_{s' \in S} |g_t(s', \hat{G}_t^{n-1}) - g_t(s', U_t^{n-1})| P(s'|s, a) \\
\leq \max_{s' \in S_t} |g_t(s', \hat{G}_t^{n-1}) - g_t(s', U_t^{n-1})| \\
\leq \max_{(s', a') \in S_t \times A} |\phi(s', a')^T(\hat{G}_t^{n-1} - U_t^{n-1})|.
\]

Hence, again using Corollary 3, for any \( n \in \mathbb{N} \) and \( (s, a) \in S \times A \), we obtain

\[
|\phi(s, a)^T(\hat{G}_t^n - U_t^n)| \leq |\phi(s, a)^T(\hat{G}_t^n - U_t(\hat{G}_t^{n-1}))| + |\phi(s, a)^T(U_t(\hat{G}_t^{n-1}) - U_t(U_t^{n-1}))| \\
\leq \varepsilon_t \|\phi(s, a)\|_{\Lambda_t^{-1}} + \max_{(s', a') \in S_t \times A} |\phi(s', a')^T(\hat{G}_t^{n-1} - U_t^{n-1})|.
\]

Thus, taking the maximum over \( (s, a) \in S_t \times A \) on both sides, by definition of \( S_t \), we have shown

\[
\max_{(s, a) \in S_t \times A} |\phi(s, a)^T(\hat{G}_t^n - U_t^n)| \leq \frac{2(B_t + 1)\varepsilon_t}{\alpha_t} + \max_{(s, a) \in S_t \times A} |\phi(s, a)^T(\hat{G}_t^{n-1} - U_t^{n-1})| \quad \forall \, n \in \mathbb{N}.
\]

Iterating this inequality, and since \( \hat{G}_t^0 = U_t^0 = 0 \), we conclude

\[
\max_{(s, a) \in S_t \times A} |\phi(s, a)^T(\hat{G}_t^{n-1} - U_t^{n-1})| \leq \frac{2(B_t + 1)\varepsilon_t(n - 1)}{\alpha_t} \quad \forall \, n \in \mathbb{N}.
\]

Substituting back into (38) and again using \( \hat{G}_t^0 = U_t^0 = 0 \), we obtain the desired result. \( \square \)

We can now state the aforementioned Lemma 10. Note that while we have already established a polynomial rate of convergence for \( U_t \) in Lemma 3, we keep the rate general here, since Theorem 3 will use an improved rate.

**Lemma 10 (\( \hat{G}_t \) convergence).** If Assumptions 1 and 2 hold and \( \kappa_t \geq 18d \), then on the event \( \mathcal{E} \),

\[
\|\hat{G}_t^n - \hat{G}_t^{n-1}\|_{\Lambda_t} \leq 5\sqrt{d\varepsilon_t} + \frac{10\sqrt{t}(B_t + 1)n\varepsilon_t}{\alpha_t} + 3\sqrt{t} \max_{(s, a) \in S \times A} |\phi(s, a)^T(U_t^n - U_t^{n-1})| \quad \forall \, n \in \mathbb{N}, \, t \in [T].
\]

**Proof.** We consider three cases (the last two are corner cases). For the first and most natural case, we assume that \( \|\hat{G}_t^n - \hat{G}_t^{n-1}\|^2_2 \leq \|\hat{G}_t^n - \hat{G}_t^{n-1}\|_{\Lambda_t}/\sqrt{\mathcal{E}} \). Then by definition of the induced norm,

\[
\|\hat{G}_t^n - \hat{G}_t^{n-1}\|^2_{\Lambda_t} \leq \sum_{\tau=1}^t |\phi(s_\tau, a_\tau)^T(\hat{G}_t^n - \hat{G}_t^{n-1})|^2.
\]

Next, for any \( (s, a) \in S \times A \) and \( m \in \{n-1, n\} \), Claim 10 and Cauchy-Schwarz imply

\[
|\phi(s, a)^T(\hat{G}_t^m - U_t^m)| \leq \left( \varepsilon_t \|\phi(s, a)\|_{\Lambda_t^{-1}} + \frac{2(B_t + 1)n\varepsilon_t}{\alpha_t} \right)^2 \leq 2\varepsilon_t^2 \|\phi(s, a)\|^2_{\Lambda_t^{-1}} + \frac{8(B_t + 1)^2n^2\varepsilon_t^2}{\alpha_t^2},
\]

and
which, after another application of Cauchy-Schwarz, gives

\[
(\phi(s, a)^T (G_t^{m0} - \hat{G}_t^{m-1}0))^2 \leq 3 \sum_{m=n-1}^{n} (\phi(s, a)^T (\hat{G}_t^{m0} - U_t^{m0}))^2 + 3(\phi(s, a)^T (U_t^{m0} - U_t^{m-1}0))^2
\]

\[
\leq 12\varepsilon_t^2\|\phi(s, a)\|_{\Lambda_t}^2 + \frac{48(B_t + 1)^2n^2\varepsilon_t^2}{\alpha_t^2} + 3 \max_{(s', a')\in S\times A} (\phi(s', a')^T (U_t^{m0} - U_t^{m-1}0))^2.
\]

By (Jin et al., 2020, Lemma D.1), we know \(\sum_{t=1}^t \|\phi(s_t, a_t)\|_{\Lambda_t}^2 \leq d\). Combined with previous three inequalities,

\[
\|G_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t}^2 \leq 24d\varepsilon_t^2 + \frac{96t(B_t + 1)^2n^2\varepsilon_t^2}{\alpha_t^2} + 6t \max_{(s, a)\in S\times A} (\phi(s, a)^T (U_t^{m0} - U_t^{m-1}0))^2.
\]

Taking square roots on both sides, bounding the square root of sum by the sum of square roots, and using \(\sqrt{24} \leq 5\), \(\sqrt{96} \leq 100\), and \(\sqrt{6} \leq 3\) yields the desired bound.

For the second case, suppose \(\|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_2 > \|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t}/\sqrt{2}\) and \(n \geq 2\). Then

\[
\|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t} < 2 \sum_{m=n-1}^{n} \left( \|\hat{G}_t^{m0} - U_t^{m0}\|_2 + \|U_t^{m0}\|_2 \right) \leq 4(\varepsilon_t + d(B_t + 1)),
\]

where we used Claim 1, Corollary 3, and Claim 4. By assumption \(\kappa_t \geq 18d\), we also know

\[
\varepsilon_t/5 = (B_t + 1)\sqrt{\log(t(B_t + 1)\kappa_t\sqrt{\log(t(B_t + 1)\kappa_t/\delta)}))/\delta} \geq (B_t + 1)\delta \geq 2.
\]

Hence, combining the previous two bounds, we obtain \(\|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t} \leq 24\varepsilon_t/5 \leq 5\sqrt{d}\varepsilon_t\).

Finally, suppose \(\|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_2 > \|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t}/\sqrt{2}\) and \(n = 1\). Then by Claim 4 and (39),

\[
\|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t} < \sqrt{2}\|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_2 = \sqrt{2}\|\hat{G}_t^{m0}\|_2 \leq \sqrt{2d}\varepsilon_t.
\]

We now proceed to the proofs of the theorems.

**D.1 Proof of Theorem 2**

We begin with a corollary of Lemmas 3 and 10 in the setting of Theorem 2. The proof is mostly algebra.

**Corollary 5** (\(\hat{G}_t\) convergence). Under the Assumptions of Theorem 2 and on the event \(\mathcal{E}\), for any \(t \in [T]\),

\[
\min_{n \in [2dt^{1/6}]} \|\hat{G}_t^{m0} - \hat{G}_t^{m-1}0\|_{\Lambda_t} \leq \alpha_t.
\]

**Proof.** Since \(\kappa_t = 54dt^{1/3} = 9(6t^{1/3})d\) by assumption in Theorem 2, Claim 6 implies

\[
\alpha_t \geq \max\{6t^{1/3}\varepsilon_t, (B_t + 1)\kappa_t\}.
\]

Hence, using the bound \(\alpha_t \geq (B_t + 1)\kappa_t\), we have

\[
\frac{\alpha_t}{9\varepsilon_t} = \frac{\kappa_t\sqrt{\log(t(B_t + 1)\kappa_t/\delta)}}{45d\sqrt{\log(t\alpha_t/\delta)}} \leq \frac{\kappa_t}{40d} = \frac{54t^{1/3}}{45} < 4t^{1/3}.
\]
Thus, if we define $N_t = d \sqrt{\alpha_t / \varepsilon_t}/3$, we are guaranteed that $N_t \leq 2t^{1/6}$, so $\lceil N_t \rceil \in [\lceil 2t^{1/6} \rceil]$. Combining this result with Lemmas 3 and 10 (we can invoke the latter since $\kappa_t \geq 18d$), we obtain

$$
\min_{n \in [\lceil 2t^{1/6} \rceil]} \| G_t^n - \hat{G}_t^n \|_{\Lambda_t} \leq \| G_t^{[N_t]} - \hat{G}_t^{[N_t]} \|_{\Lambda_t} \leq 5\sqrt{d} \varepsilon_t + \frac{10\sqrt{t} (B_t + 1) \lceil N_t \rceil \varepsilon_t}{\alpha_t} + \frac{3\sqrt{t} (B_t + 1) d^2}{\lceil N_t \rceil}.
$$

(41)

For the third term in (41), since $\lceil N_t \rceil \geq N_t$, we have

$$
3\sqrt{t} (B_t + 1) d^2 / \lceil N_t \rceil \leq 3\sqrt{t} (B_t + 1) d^2 / N_t = 9\sqrt{t} (B_t + 1) d / \sqrt{\alpha_t \varepsilon_t}.
$$

For the second term, since $N_t \geq 2\sqrt{6} / 3 \geq 1$ by (40), we have $\lceil N_t \rceil \leq 2N_t$, so

$$
10\sqrt{t} (B_t + 1) \lceil N_t \rceil \varepsilon_t / \alpha_t \leq 21\sqrt{t} (B_t + 1) N_t \varepsilon_t / \alpha_t = 7\sqrt{t} (B_t + 1) d / \sqrt{\alpha_t \varepsilon_t}.
$$

Hence, because $\sqrt{\alpha_t \varepsilon_t} \geq 2t^{1/6}$ and $\alpha_t \geq \kappa_t (B_t + 1)$ by (40), the last two terms in (41) can be bounded by

$$
\frac{10\sqrt{t} (B_t + 1) \lceil N_t \rceil \varepsilon_t}{\alpha_t} + \frac{3\sqrt{t} (B_t + 1) d^2}{\lceil N_t \rceil} \leq \frac{16\sqrt{t} d (B_t + 1)}{\sqrt{\alpha_t \varepsilon_t}} \leq \frac{8t^{1/3} d (B_t + 1)}{54} \leq \frac{8\kappa_t (B_t + 1)}{54} \leq \frac{8\alpha_t}{54} \leq \frac{\alpha_t}{6}.
$$

Substituting into (41), and assuming for the moment that $\sqrt{d} \leq t^{1/3}$, we can use (40) to obtain

$$
\min_{n \in [\lceil 2t^{1/6} \rceil]} \| G_t^n - \hat{G}_t^n \|_{\Lambda_t} \leq 5\sqrt{d} \cdot \varepsilon_t + \frac{\alpha_t}{6} \leq 5t^{1/3} \cdot \frac{\alpha_t}{6t^{1/3}} + \frac{\alpha_t}{6} = \alpha_t.
$$

If instead $\sqrt{d} > t^{1/3}$, then $t^{1/6} < \sqrt{d}$ as well, so we can instead use Claim 4 and (40) to obtain

$$
\min_{n \in [\lceil 2t^{1/6} \rceil]} \| G_t^n - \hat{G}_t^n \|_{\Lambda_t} \leq \| G_t 0 \|_{\Lambda_t} \leq \sqrt{8t d} = \sqrt{8t^{1/3} t^{1/6} d^{1/2}} \leq \sqrt{8t^{1/3} d} \leq \kappa_t \leq \alpha_t. \quad \square
$$

We can now prove Theorem 2. Recall from Appendix C that the regret bound in Theorem 1 holds on $E \cap F$. Hence, on this event, and since $\kappa_t = 60d t^{1/3}$ in Theorem 2, we can set $\Psi = 60d$ and $\lambda = 1/3$ to obtain

$$
R(K) = O \left( (B_{sf}^{1/2} + B_s^{5/2}) d^{3/2} (K/c_{\min})^{5/2} + (B_s + 1)^6 d^6 c_{\min}^{-5} \right).
$$

Finally, Corollaries 4 and 5 imply that on $E$, for any $t \in [T]$ Algorithm 2 is called, it returns an OAFP within $O(dt^{1/6})$ iterations. Together with Lemmas 2 and 7, which ensure $P(E \cap F) \geq 1 - \delta$, this completes the proof.

### D.2 Proof of Theorem 3

As discussed above, we first establish geometric convergence using the contraction property.

**Lemma 11** (Geometric $U_t$ convergence). If Assumptions 1 and 2 hold and all stationary policies are proper,

$$
|\phi(s, a)^T (U_t^n 0 - U_t^{n-1} 0)| \leq \chi \rho^{n-1} \forall (s, a) \in S \times A, n \in \mathbb{N}, t \in [T].
$$

**Proof.** Fix $t$. When $n = 1$, since $\chi \geq 1$ by definition, we simply have

$$
|\phi(s, a)^T (U_t^n 0 - U_t^{n-1} 0)| = |\phi(s, a)^T \theta| = |c(s, a)| \leq 1 \leq \chi = \chi \rho^{n-1}. \quad (42)
$$
It remains to show that \(|\phi(s,a)^T(U_t^{n+1} - U_t^n)| \leq \chi \rho^n\) for all \(n \in \mathbb{N}\). Fix such an \(n\). By monotonicity of \(\Pi_{[0,B_t]}\),

\[
g_t(s, U_t^n) = \min_{a \in A} \Pi_{[0,B_t]} \left( \phi(s,a)^T U_t^n - \alpha\|\phi(s,a)\|_{A_t} \right).
\]

Hence, if we define \(Q_n \in \mathbb{R}^{S \times A}\) to be the matrix with \((s,a)\)-th element

\[
Q_n(s,a) = \Pi_{[0,B_t]} \left( \phi(s,a)^T U_t^n - \alpha\|\phi(s,a)\|_{A_t} \right),
\]

we have \(g_t(s, U_t^n) = \min_{a \in A} Q_n(s,a)\). Thus, by definition of \(U_t\), we obtain

\[
\phi(s,a)^T(U_t^{n+1} - U_t^n) = \sum_{s',a'} \left( \min_{\alpha \in A} Q_n(s',a') - \min_{\alpha' \in A} Q_{n-1}(s',a') \right) P(s'|s,a) = (TQ_n - TQ_{n-1})(s,a),
\]

where \(T\) is the state-action operator defined in (18). By (19), this implies

\[
\max_{(s,a) \in S \times A} \omega(s)|\phi(s,a)^T(U_t^{n+1} - U_t^n)| \leq \rho \max_{(s,a) \in S \times A} \omega(s)|(Q_n - Q_{n-1})(s,a)|
\]

\[
\leq \rho \max_{(s,a) \in S \times A} \omega(s)|\phi(s,a)^T(U_t^n - U_t^{n-1})|,
\]

where the last inequality holds by Claim 2. Iterating the previous inequality and using the bound \(|\phi(s,a)^T(U_t^n - U_t^0)| \leq 1 \text{ from (42)}\), we obtain that for any \(n \in \mathbb{N}\),

\[
\max_{(s,a) \in S \times A} \omega(s)|\phi(s,a)^T(U_t^{n+1} - U_t^n)| \leq \rho^n \max_{s \in S} \omega(s).
\]

Hence, for any \((s,a) \in S \times A\) and \(n \in \mathbb{N}\), we obtain

\[
|\phi(s,a)^T(U_t^{n+1} - U_t^n)| \leq \frac{\omega(s)|\phi(s,a)^T(U_t^{n+1} - U_t^n)|}{\min_{s' \in S} \omega(s')} \leq \frac{\max_{s' \in S} \omega(s') \rho^n}{\min_{s' \in S} \omega(s')} = \chi \rho^n. \quad \Box
\]

Next, we have an analogue of Corollary 5, whose proof is also similar.

**Corollary 6 (\(\hat{G}_t\) convergence).** If Assumptions 1 and 2 hold, all stationary policies are proper, and \(\kappa_t = 54dt^{1/4}\sqrt{N_t'}\) for some \(N_t' \geq \log(3t\chi)/(1-\rho)\), then on the event \(\mathcal{E}\), for any \(t \in [T]\),

\[
\min_{n \in [N_t']} \|\hat{G}_t^n - \hat{G}_t^{n-1}\|_{A_t} \leq \alpha_t.
\]

**Proof.** Because \(\chi \geq 1\) and \(\rho \in (0,1)\) by definition, we know \(N_t' \geq \log 3 \geq 1\), so \(1 \leq [N_t'] \leq 2N_t'\). Combined with Lemmas 10 and 11, we obtain

\[
\min_{n \in [N_t']} \|\hat{G}_t^n - \hat{G}_t^{n-1}\|_{A_t} \leq \|\hat{G}_t^{[N_t']} - \hat{G}_t^{[N_t']-1}\|_{A_t} \leq 5\sqrt{d_{\varepsilon_t} + \frac{24\sqrt{T}(B_t + 1)N_t'}{\alpha_t}} + 3\sqrt{t\chi \rho^{N_t'}}. \quad (43)
\]

On the other hand, since \(6t^{1/4}\sqrt{N_t'} \geq 1\) (recall \(N_t' \geq 1\)), Claim 6 implies

\[
\alpha_t \geq \max\{6t^{1/4}\sqrt{N_t'\varepsilon_t}, (B_t + 1)\kappa_t\}. \quad (44)
\]

Thus, using the assumption \(N_t' \geq \log(3t\chi)/(1-\rho)\), we can bound the third term in (43) by

\[
3\sqrt{t\chi \rho^{N_t'}} \leq 3\sqrt{t\chi e^{-(1-\rho)N_t'}} \leq 1/\sqrt{t} \leq 1 \leq \kappa_t(B_t + 1)/54 \leq \alpha_t/54.
\]
For the second term in (43), we again use (44) to obtain
\[
\frac{24\sqrt{t}N_t'}{\alpha_t} \leq \frac{24\sqrt{t}N_t'}{6t^{1/4}/\sqrt{N_t'}} = 4t^{1/4}\sqrt{N_t'}(B_t + 1) \leq \frac{4\kappa_t(B_t + 1)}{54} \leq \frac{4\alpha_t}{54}.
\]
Plugging the previous two inequalities into (43), and assuming \( \sqrt{d} \leq t^{1/4} \), we can use (44) and \( N_t' \geq 1 \) to obtain
\[
\min_{n \in [N_t]} \| \hat{\mu}^n_{t_0} - \hat{\mu}^{n-10}_{t_0} \|_{\Lambda_t} \leq 5\sqrt{d}:\varepsilon_t + \frac{5\alpha_t}{54} \leq 5t^{1/4} \cdot \frac{\alpha_t}{6t^{1/4}} + \frac{5\alpha_t}{54} = \frac{50\alpha_t}{54} < \alpha_t.
\]
If instead \( \sqrt{d} > t^{1/4} \), we simply use Claim 4 and \( N_t' \geq 1 \) to obtain
\[
\min_{n \in [N_t]} \| \hat{\mu}^n_{t_0} - \hat{\mu}^{n-10}_{t_0} \|_{\Lambda_t} \leq \sqrt{8t^{1/4}d} \leq \sqrt{8t^{1/4}}d \leq \kappa_t \leq \alpha_t.
\]

We now prove Theorem 3. As for Theorem 2, it suffices to prove the guarantees on \( E \cap F \). Corollaries 4 and 6 establish the OAFP guarantee on \( E \cap F \) (we choose \( N_t' = N_t = \log(3t\bar{\chi})/(1 - \bar{\rho}) \) in the latter). Next, setting \( \Psi = 54d\sqrt{3\log(3\bar{\chi})/(1 - \bar{\rho})} \) and \( \lambda = 1/4 \), we have
\[
\kappa_t = 54d t^{1/4}\sqrt{\log(3t\bar{\chi})/(1 - \bar{\rho})} = \Psi t^{\lambda}\sqrt{\log(3t\bar{\chi})/(3\log(3\bar{\chi}))} \leq \Psi t^{\gamma}\log(t + 1),
\]
where the inequality holds because by \( \bar{\chi} \geq 1 \), we have
\[
\frac{\log(3t\bar{\chi})}{3\log(3\bar{\chi})} = \frac{1}{3} + \frac{\log(t)}{3\log(3\bar{\chi})} \leq \frac{1 + \log(t)}{3} \leq \frac{\log(t+1)}{3} \leq \left( \frac{1}{3} \right) \log(t+1) \leq \log(t+1).
\]
Hence, on \( E \cap F \), we can use the Theorem 1 regret bound with this \( \Psi \) and \( \lambda \) to obtain
\[
R(K) = O\left( \left( B^*_{t_0} + B^*_{s} \right)^{3/2} (K/c_{\min})^{3/2} N_t^{1/2} + (B^* + 1)^4 \right).\]

Remark 15 (Unknown \( \bar{\chi} \) and \( \bar{\rho} \)). Suppose \( \kappa_t = 54d t^{1/4} \sqrt{N_t} \) as in Theorem 3 but \( N_t = t^{2\gamma} \) as in Appendix A.1. Then \( \kappa_t \geq 9d \) for any \( t \in \mathbb{N} \) and \( N_t \geq \log(3t\bar{\chi})/(1 - \rho) \) as soon as \( t \geq (\log(3t\bar{\chi})/(1 - \rho))^{1/\gamma} \), so we can use Corollaries 4 and 6 and Lemma 2 to obtain the following: with probability at least \( 1 - \delta/2 \), for any \( t \geq (\log(3t\bar{\chi})/(1 - \rho))^{1/\gamma} \), that Algorithm 3 is called, it returns an OAFP in \( t^{2\gamma} \) iterations.

D.3 Proof of Theorem 4

We begin by showing \( \hat{G}_t \) is a contraction with respect to \( \| \cdot \| = \| Q^T \cdot \|_{\infty} \), where \( Q \) is the orthogonal matrix with columns \( \{ q_i \}_{i=1}^d \). Note \( \| \cdot \| \) is a norm by orthogonality of \( Q \).

Lemma 12 (\( \hat{G}_t \) contraction). Under the assumptions of Theorem 4, for any \( t \in [T] \) and \( w_1, w_2 \in \mathbb{R}^d \), we have
\[
\| Q^T (\hat{G}_t w_1 - \hat{G}_t w_2) \|_{\infty} \leq e^{-t/(t+1)} \| Q(w_1 - w_2) \|_{\infty}.
\]

Proof. For \( (s, a) \in S \times A \), let \( i(s, a) \in [d] \) be such that \( \phi(s, a) = q_{i(s, a)} \) (which exists by assumption). For \( i \in [d] \), define \( d_i = |\{ \tau \in [t] : i(s_\tau, a_\tau) = i \}| \). Let \( D \) be the diagonal matrix with diagonal elements \( \{ d_i + 1 \}_{i=1}^d \). Then
\[
\Lambda_t = I + \sum_{\tau=1}^{t} \phi(s_\tau, a_\tau) \phi(s_\tau, a_\tau)^T = \sum_{i=1}^d q_i q_i^T + \sum_{i=1}^d d_i q_i q_i^T = \sum_{i=1}^d (1 + d_i) q_i q_i^T = Q D Q^T.
\]
This implies \( \Lambda_t^{-1} = QD^{-1}Q^T \), so for any \( i \in [d] \) and \( (s,a) \in S \times A \), we have

\[
e_i^T Q_i^T \Lambda_i^{-1} \phi(s,a) = e_i^T D_i^{-1} Q_i^T q_i(s,a) = \frac{e_i^T Q_i^T q_i(s,a)}{d_i + 1} = \frac{q_i^T q_i(s,a)}{d_i + 1} = \frac{1(i(s,a) = i)}{d_i + 1}.
\]

Using this identity, we obtain

\[
e_i^T Q_i^T (\hat{G}_t w_1 - \hat{G}_t w_2) = \sum_{\tau = 1}^t e_i^T Q_i^T \Lambda_i^{-1} \phi(s_\tau, a_\tau) (g_\tau(s_\tau, w_1) - g_\tau(s_\tau', w_2)) = \sum_{\tau \in [t] : i(s_\tau, a_\tau) = i} (g_\tau(s_\tau', w_1) - g_\tau(s_\tau', w_2))
\]

On the other hand, for any \( s \in S \), we know

\[
|g_\tau(s, w_1) - g_\tau(s, w_2)| \leq \max_{a \in A} |\phi(s, a)^T (w_1 - w_2)| = \max_{a \in A} |e_i^T Q_i^T (w_1 - w_2)| \leq \|Q_i^T (w_1 - w_2)\|_\infty,
\]

where we used Claim 3 for the first inequality. Combining the last two expressions, we obtain

\[
\|Q_i^T (\hat{G}_t w_1 - \hat{G}_t w_2)\|_\infty = \max_{i \in [d]} \left| \sum_{\tau \in [t] : i(s_\tau, a_\tau) = i} (g_\tau(s_\tau', w_1) - g_\tau(s_\tau', w_2)) \right| \leq \max_{i \in [d]} \frac{d_i}{d_i + 1} \|Q_i^T (w_1 - w_2)\|_\infty.
\]

This completes the proof, since \( d_i/(d_i + 1) \leq t/(t + 1) \leq e^{-t/(t+1)}. \)

Using Lemma 12, we can show Algorithm 2 terminates within \( O(t \log(td)) \) iterations.

**Corollary 7.** Under the assumptions of Theorem 4, for any \( t \in [T] \),

\[
\min_{n \in \{1 + (t+1) \log((t+1)d)/2\}} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \leq \alpha_t.
\]

**Proof.** For any \( n \in \mathbb{N} \), we can iterate the bound from Lemma 12 to obtain

\[
\|Q_i^T (\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0)\|_\infty \leq e^{-\frac{n}{t+1}} \|Q_i^T \hat{G}_t 0\|_\infty.
\]

By a standard norm equivalence, orthogonality, and Claim 4, we also have

\[
\|Q_i^T \hat{G}_t 0\|_\infty \leq \|Q_i^T \hat{G}_t 0\|_2 = \|\hat{G}_t 0\|_2 \leq 2\sqrt{d}.
\]

By Claim 1, orthogonality, and a standard equivalence, we also know

\[
\|w\|_{\Lambda_t} \leq \sqrt{\frac{1}{t+1} d} \|w\|_2 = \sqrt{\frac{1}{t+1} d} \|Q_i^T w\|_2 \leq \sqrt{\frac{1}{t+1} d^2} \|Q_i^T w\|_\infty \forall \ w \in \mathbb{R}^d.
\]

Hence, combining the previous three inequalities, we obtain

\[
\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \leq \sqrt{\frac{1}{t+1} d^2} \|Q_i^T (\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0)\|_\infty \leq \sqrt{(t+1)d^2} e^{-\frac{n}{t+1}} \|Q_i^T \hat{G}_t 0\|_\infty \leq 2\sqrt{(t+1)d^2} e^{-\frac{n}{t+1}}.
\]

Therefore, if \( n \geq 1 + (t+1) \log((t+1)d)/2 \), then the previous bound, the assumed choice \( \kappa_t = 9d \) in Theorem 4, and Claim 6 imply

\[
\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \leq 2d \leq \kappa_t \leq \alpha_t.
\]

Similar to the above, on \( E \cap F \), Corollaries 4 and 7 show Algorithm 2 returns OAFPs in \( O(t \log(td)) \) iterations, and since \( \kappa_t = 9d \) in Theorem 4, we obtain the regret bound from Corollary 1.
E Other proofs

E.1 Proof of Lemma 2

For any $t \in \mathbb{N}$ and $b > 0$, define the following bad event:

$$\mathcal{B}_{t,b} = \left\{ \sup_{w \in [-W_t + W_t]^d} \|E_t w\|_{\Lambda_t} > \varepsilon_t \right\} \cap \{ B_t = b \}.$$  

Our main goal is to prove the following claim.

**Claim 11.** Under the assumptions of Lemma 2, for any $t \in \mathbb{N}$ and $b > 0$, we have $\mathbb{P}(\mathcal{B}_{t,b}) \leq \delta/(2(t+1)^2)$.

Before proving the claim, we show it implies the lemma. First note $B_t$ is $\{2^{i-1}c_{\min}\}_{i=1}^t$-valued, so

$$\mathcal{E}^C = \bigcup_{t \in \mathbb{N}} \left\{ \sup_{w \in [-W_t + W_t]^d} \|E_t w\|_{\Lambda_t} > \varepsilon_t \right\} = \bigcup_{t \in \mathbb{N}} \bigcup_{b \in \{2^{i-1}c_{\min}\}_{i=1}^t} \mathcal{B}_{t,b}.$$  

Hence, taking union bounds over $t$ and $b$ and invoking Claim 11, we obtain

$$\mathbb{P}(\mathcal{E}^C) \leq \sum_{t=1}^{\infty} \sum_{b \in \{2^{i-1}c_{\min}\}_{i=1}^t} \mathbb{P}(\mathcal{B}_{t,b}) \leq \frac{\delta}{2} \sum_{t=1}^{\infty} \frac{1}{(t+1)^2} \leq \frac{\delta}{2} \int_{t=1}^{\infty} \frac{dt}{t^2} = \frac{\delta}{2}.$$  

Thus, it only remains to prove Claim 11. We fix $t$ and $b$ for the remainder of this appendix. For $x \in \mathbb{R}^d$ and $Y \in \mathbb{R}_{>0}^{d \times d}$ (the set $d \times d$ positive definite matrices), we define $f_{x,Y} : \mathcal{S} \to \mathbb{R}$ and $g_{x,Y} : \mathcal{S} \to \mathbb{R}$ by

$$f_{x,Y}(s) = \min_{a \in \mathcal{A}} \left( \phi(s,a)^T x - \|\phi(s,a)\|_Y \right), \quad g_{x,Y}(s) = \Pi_{[0,b]}(f_{x,Y}(s)).$$  

Here $\Pi_{[0,b]}(\cdot)$ clips between 0 and $b$ as in (20). Hence, we have the following implication:

$$B_t = b \implies g_t(s,w) = g_{w,\alpha_t^2\Lambda_t^{-1}}(s) \forall s \in \mathcal{S}, w \in \mathbb{R}^d.$$  

**(45)**

**Claim 12.** Under the assumptions of Lemma 2, for any $x_1, x_2 \in \mathbb{R}^d$, $Y_1, Y_2 \in \mathbb{R}_{>0}^{d \times d}$, and $s \in \mathcal{S}$,

$$|g_{x_1,Y_1}(s) - g_{x_2,Y_2}(s)| \leq \sqrt{d} \|w_1 - w_2\|_{\infty} + \max_{a \in \mathcal{A}} \|\phi(s,a)\|_{Y_1} - \|\phi(s,a)\|_{Y_2}.$$  

**Proof.** The proof is almost identical to Claim 3, except the bonus terms $\|\phi(s,a)\|_{Y_i}$ do not cancel.

We now derive a bound on the error operator that removes the bias introduced by the regularizer. Here and moving forward, for any $\tau \in [t]$, we use the shorthand $\phi_{\tau} = \phi(s_{\tau}, a_{\tau}).$

**Claim 13.** Under the assumptions of Lemma 2, if $B_t = b$, then for any $w \in \mathbb{R}^d$,

$$\|E_t w\|_{\Lambda_t} \leq \left\| \sum_{\tau=1}^{t} \phi_{\tau} (g_{w,\alpha_t^2\Lambda_t^{-1}}(s_{\tau}) - E_{s_{\tau}} g_{w,\alpha_t^2\Lambda_t^{-1}}(s_{\tau})) \right\|_{\Lambda_t^{-1}} + \sqrt{d}(b + 1).$$

38
Proof. Fix $w \in \mathbb{R}^d$. If $B_t = b$, then by (45),

$$
\hat{G}_t w = \Lambda_t^{-1} \sum_{\tau=1}^{t} \phi_{\tau}(c(s_{\tau}, a_{\tau}) + \mathbb{E}_{s_t'} g_t(s_t', w)) + \Lambda_t^{-1} \sum_{\tau=1}^{t} \phi_{\tau}(g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}'), - \mathbb{E}_{s_t'} g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}')).
$$

The first term can be rewritten as

$$
\Lambda_t^{-1} \sum_{\tau=1}^{t} \phi_{\tau}^T \left( \theta + \sum_{s \in S} \mu(s) g_t(s, w) \right) = (I - \Lambda_t^{-1}) U_t w.
$$

Combining the previous two identities, we obtain

$$
E_t w = \hat{G}_t w - U_t w = \Lambda_t^{-1} \left( \sum_{\tau=1}^{t} \phi_{\tau}(g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}') - \mathbb{E}_{s_t'} g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}')) - U_t w \right).
$$

Thus, by the triangle inequality, we have

$$
\|E_t w\|_{\Lambda_t} \leq \left\| \sum_{\tau=1}^{t} \phi_{\tau}(g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}') - \mathbb{E}_{s_t'} g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}')) \right\|_{\Lambda_t^{-1}} + \|U_t w\|_{\Lambda_t^{-1}}.
$$

This completes the proof, because when $B_t = b$, $\|U_t w\|_{\Lambda_t^{-1}} \leq \|U_t w\|_2 \leq \sqrt{d}(b + 1)$ by Claims 1 and 4. \qed

Since $g_{w, \alpha_t^2 \Lambda_t^{-1}}$ is a random function that depends on the random state-action pairs before time $t$, we take a union bound over it using a covering argument. Toward this end, let

$$
\alpha_t |b| = (b+1)\kappa_t \sqrt{\log(t(b+1)\kappa_t/\delta)}, \quad W_t |b| = (\alpha_t |b| + \sqrt{td(b+1)}), \quad \varepsilon_t = 5(b+1)d \sqrt{\log(t(\alpha_t |b|)/\delta)},
$$

denote the values of the random variables $\alpha_t, W_t$, and $\varepsilon_t$ when $B_t = b$. Thus, $\alpha_t = \alpha_t |B_t$ (and similar for $W_t$ and $\varepsilon_t$). Next, let $X$ be a $1/((\sqrt{dt})$-net of $[-W_t |b|, +W_t |b|]^d$ in the $\ell_\infty$ norm; explicitly, we define

$$
X = \left\{ \left[ i_j / (\sqrt{dt}) \right]_{j=1}^d : i_j \in \left\{ -\left[ (W_t |b|) \sqrt{dt} \right] , \ldots , \left[ (W_t |b|) \sqrt{dt} \right] \right\} \forall j \right\}.
$$

Finally, let $Y$ be a $1/(dt^2)$-net of $\{Y \in \mathbb{R}^{d \times d} : |Y(i, j)| \leq (\alpha_t |b|)^2 \forall i, j \}$, where we view the matrices as vectors:

$$
Y = \left\{ \left[ i_{j_1,j_2} / (dt^2) \right]_{j_1,j_2=1}^d : i_{j_1,j_2} \in \left\{ -\left[ (\alpha_t |b|)^2 dt^2 \right] , \ldots , \left[ (\alpha_t |b|)^2 dt^2 \right] \right\} \forall j_1, j_2 \right\} \cap \mathbb{R}^{d \times d}.
$$

Moving forward, we discard the cumbersome $\cdot |b$ notation. However, we emphasize that $X$ and $Y$ are deterministic sets, irrespective of the value taken by $B_t$.

We next show that when $B_t = b$, $g_{w, \alpha_t^2 \Lambda_t^{-1}}$ is close to some element of the function class $\{g_{x, Y} : X \times Y\}$.

Claim 14. Under the assumptions of Lemma 2, if $B_t = b$, then for any $w \in [-W_t, +W_t]^d$, there exists $x \in X$ and $Y \in Y$ such that

$$
\left\| \sum_{\tau=1}^{t} \phi_{\tau}(g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}') - \mathbb{E}_{s_t'} g_{w, \alpha_t^2 \Lambda_t^{-1}}(s_{\tau}')) \right\|_{\Lambda_t^{-1}} \leq \left\| \sum_{\tau=1}^{t} \phi_{\tau}(g_{x, Y}(s_{\tau}') - \mathbb{E}_{s_t'} g_{x, Y}(s_{\tau}')) \right\|_{\Lambda_t^{-1}} + 2.
$$
Proof. By Claim 1 and a standard spectral norm inequality, we have \(|\alpha_t^2 \Lambda_t^{-1}(i,j)| \leq \alpha_t^2 \|\Lambda_t^{-1}\|_2 \leq \alpha_t^2 \forall i,j\). Hence, we can find \(Y \in \mathcal{Y}\) such that \(\max_{i,j} |Y(i,j) - \alpha_t^2 \Lambda_t^{-1}(i,j)| \leq 1/(dt^2)\). For such \(Y\) and any \((s, a) \in \mathcal{S} \times \mathcal{A}\), we then obtain

\[
|\phi(s, a)^T (Y - \alpha_t^2 \Lambda_t^{-1}) \phi(s, a)| \leq \sum_{i,j \in [d]} |\phi_i(s, a)||\phi_j(s, a)||Y(i,j) - \alpha_t^2 \Lambda_t^{-1}(i,j)| \\
\leq \frac{\|\phi(s, a)\|_1^2}{dt^2} \leq \frac{\|\phi(s, a)\|_2^2}{t^2} \leq \frac{1}{t^2},
\]

which implies that

\[
\|\phi(s, a)\|_Y \leq \sqrt{\alpha_t^2 \|\phi(s, a)^T \Lambda_t^{-1} \phi(s, a)\| + |\phi(s, a)^T (Y - \alpha_t^2 \Lambda_t^{-1}) \phi(s, a)|} \leq \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}} + 1/t.
\]

Hence, by symmetry, we conclude that

\[
\left\| \phi(s, a) \right\|_{Y - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}}} \leq 1/t. \tag{46}
\]

Also, we can clearly find \(x \in \mathcal{X}\) such that \(\|w - x\|_\infty \leq 1/(\sqrt{d}t)\). Hence, for any \(s \in \mathcal{S}\), we obtain

\[
|g_{x,Y}(s) - g_{w,\alpha_t^2 \Lambda_t^{-1}}(s)| \leq \sqrt{d} \|w - x\|_\infty + \max_{a \in \mathcal{A}} \left| \|\phi(s, a)\|_Y - \alpha_t \|\phi(s, a)\|_{\Lambda_t^{-1}} \right| \leq 2/t, \tag{47}
\]

where we used Claim 12, (46) and the choice of \(x\). Also, defining \(\Delta(s) = g_{w,\alpha_t^2 \Lambda_t^{-1}}(s) - g_{x,Y}(s) \forall s \in \mathcal{S}\), we have

\[
\left\| \sum_{t=1}^T \phi_T(g_{w,\alpha_t^2 \Lambda_t^{-1}}(s'_t) - \mathbb{E}_{\varepsilon'_t} g_{w,\alpha_t^2 \Lambda_t^{-1}}(s'_t)) \right\|_{\Lambda_t^{-1}} \\
\leq \left\| \sum_{t=1}^T \phi_T(g_{x,Y}(s'_t) - \mathbb{E}_{\varepsilon'_t} g_{x,Y}(s'_t)) \right\|_{\Lambda_t^{-1}} + \left\| \sum_{t=1}^T \phi_T(\Delta(s'_t) - \mathbb{E}_{\varepsilon'_t} \Delta(s'_t)) \right\|_{\Lambda_t^{-1}}.
\]

By the triangle inequality, Claim 1, and (47), the second term satisfies

\[
\left\| \sum_{t=1}^T \phi_T(\Delta(s'_t) - \mathbb{E}_{\varepsilon'_t} \Delta(s'_t)) \right\|_{\Lambda_t^{-1}} \leq \sum_{t=1}^T \|\phi_T\|_2 |\Delta(s'_t) - \mathbb{E}_{\varepsilon'_t} \Delta(s'_t)| \leq 2. \tag*{\square}
\]

Our final ingredient for proving Claim 11 is the following bound on \(\varepsilon_t\).

Claim 15. Under the assumptions of Lemma 2, if \(B_t = b\), then

\[
\varepsilon_t \geq \sqrt{2b^2 \log \left( \frac{\det(\Lambda_t)}{\det(\Lambda_0)} \right) + \sqrt{\frac{2t(t+1)^2|\mathcal{X}||\mathcal{Y}|}{\delta}}} + \sqrt{d(b+1) + 2}.
\]

Proof. We first observe that by assumption \(\kappa_t \geq 9d\) and \(d \geq 2\), we have

\[
\alpha_t = \kappa_t(b+1) \sqrt{\log(t(b+1)\kappa_t/\delta)} \geq 9d(b+1) \geq 9d \geq 18. \tag{48}
\]

Using this bound, we (coarsely) bound the sizes of the nets. For \(\mathcal{X}\), we first recall that \(W_t = \alpha_t + \sqrt{d(b+1)}\), so by (48) and \((1/\sqrt{9}) + (1/9) = (1/3) + (1/9) = 4/9\), we have

\[
W_t \sqrt{dt} = \alpha_t \sqrt{dt} + (b+1)dt^{3/2} \leq \alpha_t \sqrt{\alpha_t/9t} + (\alpha_t/9) t^{3/2} \leq 4(\alpha_t t)^{3/2}/9.
\]

40
Again using (48), we have $3 \leq 3(\alpha_t t^{3/2})^2 / 18^{3/2} \leq (\alpha_t t^{3/2})^2 / 9$. Thus, because $d \geq 2$, we obtain

$$|\mathcal{X}| \leq (1 + 2 |W_t \sqrt{dt}|)^d \leq (3 + 2W_t \sqrt{dt})^d \leq (\alpha_t t^{3/2})^d \leq (\alpha_t t)^d.$$ 

For $\mathcal{Y}$, we can use (48) to obtain $3 \leq 3(\alpha_t t^2) / 18^3 \leq (\alpha_t t^2 / 4)$ and $2d \leq \alpha_t / 4$, so

$$|\mathcal{Y}| \leq (3 + 2d^2 \alpha_t^2)^{(2d)} \leq (\alpha_t t^2 / 2)^d = (2d^2 \alpha_t^2)^{(2d^2)}.$$ 

Next, observe $\det(\Lambda_t) / \det(\Lambda_0) \leq (t + 1)^d$ by Claim 1, so again using $d \geq 2$, we have

$$2t(t + 1)^2 \sqrt{\det(\Lambda_t) / \det(\Lambda_0)} \leq 2t(t + 1)^{2+d/2} \leq (2t)^{3+d/2} \leq (2t)^{(3d^2 / 4) + (d^2 / 4)} = (2t)^d.$$ 

Combining the previous three inequalities, and since $\delta \geq \delta d^d$, we obtain

$$2t(t + 1)^2 \sqrt{\det(\Lambda_t) / \det(\Lambda_0)} |\mathcal{X}| |\mathcal{Y}| / \delta \leq (\alpha_t t)^{4d^2}.$$ 

Since $d \geq 2$, we also have $\sqrt{d} (b + 1) + 2 \leq (b + 1) d$. Combined with the previous inequality,

$$\sqrt{2b^2 \log \left( \frac{\sqrt{\det(\Lambda_t)}}{\det(\Lambda_0)} \frac{2t(t + 1)^2 |\mathcal{X}| |\mathcal{Y}|}{\delta} \right)} + \sqrt{d} (b + 1) + 2 \leq \sqrt{8b^2 d^2 \log(\alpha t / \delta)} + 2(b + 1) d \leq (\sqrt{8} + 2) (b + 1) d \sqrt{\log(\alpha t / \delta)} \leq 5(b + 1) d \sqrt{\log(\alpha t / \delta) = \varepsilon_t}.$$

**Remark 16** (Sharpening the tabular case). In the tabular case, $\Lambda_t^{-1}$ is diagonal, so we can replace $\mathcal{Y}$ with $\mathcal{Y}' = \{Y \in \mathcal{Y} : Y$ is diagonal}. Since $|\mathcal{Y}'|$ is exponential in $d$ (instead of $d^2$), we can define $\varepsilon_t$ to have square root (instead of linear) dependence on $d$.

**Proof of Claim 11.** For each $(x, Y) \in \mathcal{X} \times \mathcal{Y}$, define the event

$$C_{x, Y} = \left\{ \left| \sum_{\tau=1}^{t} \phi_\tau (g_{x,Y}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{x,Y}(s'_\tau)) \right|_{\Lambda_t^{-1}} > \sqrt{2b^2 \log \left( \frac{\sqrt{\det(\Lambda_t)}}{\det(\Lambda_0)} \frac{2t(t + 1)^2 |\mathcal{X}| |\mathcal{Y}|}{\delta} \right)} \right\} .$$

Then since $g_{x,Y}$ is a deterministic $[0, b]$-valued function, $g_{x,Y}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{x,Y}(s'_\tau)$ are conditionally zero-mean $[-b, b]$-valued random variables, so are $b$-subgaussian. Hence, by (Abbasi-Yadkori et al., 2011, Theorem 1), we have $\mathbb{P}(C_{x, Y, b}) \leq \delta / (2t(t + 1)^2 |\mathcal{X}| |\mathcal{Y}|)$). Combined with Claims 13, 14, and 15 and the union bound,

$$\mathbb{P}(B_{t, b}) \leq \mathbb{P}(\bigcup_{(x, Y) \in \mathcal{X} \times \mathcal{Y}} C_{x, Y} \cap \{B_t = b\}) \leq \mathbb{P}(\bigcup_{(x, Y) \in \mathcal{X} \times \mathcal{Y}} C_{x, Y}) \leq \sum_{(x, Y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{P}(C_{x, Y}) \leq \frac{\delta}{2t(t + 1)^2}. \square$$

**E.2 Proof of Proposition 1**

By Assumption 2 and the definition of $Q^*$ (2), for any $s \in S$, we have

$$\phi(s, a)^T w^* = \phi(s, a)^T \left( \theta + \sum_{s' \in S} J^*(s') \mu(s') \right) = c(s, a) + \sum_{s' \in S} J^*(s') P(s'|s, a) = Q^*(s, a).$$

Hence, by the Bellman optimality equations (1),

$$J^*(s) = \min_{a \in \mathcal{A}} Q^*(s, a) = \min_{a \in \mathcal{A}} \phi(s, a)^T w^*, \quad \pi^*(s) \in \arg \min_{a \in \mathcal{A}} Q^*(s, a) = \arg \min_{a \in \mathcal{A}} \phi(s, a)^T w^*.$$ 

The first equality also implies that $w^*$ is a fixed point of $G$:

$$Gw^* = \theta + \sum_{s \in S} \mu(s) \min_{a \in \mathcal{A}} \phi(s, a)^T w^* = \theta + \sum_{s \in S} \mu(s) J^*(s) = w^*.$$ 

41