Regret Bounds for Stochastic Shortest Path Problems with Linear Function Approximation

Daniel Vial [1 2] **Advait Parulekar** [1] **Sanjay Shakkottai** [1] **R. Srikant** [2]

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 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 (MDP) 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:

- *Statistically efficient:* regret independent of S and A, sublinear (ideally, square root) in T or K, and polynomial in d and any other parameters.
- *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 stationary policy settings, by which we mean those settings where stationary policies are guaranteed to be optimal (i.e., stochastic shortest path (SSP) and average/discounted cost problems). To our knowledge, this problem is essentially open: state-of-the-art algorithms are either computationally inefficient, or they sidestep the issue by using non-stationary policies that often require access to unknown MDP parameters. See Section 1.1 for details.

In addition to this theoretical point of interest, there is practical motivation for understanding stationary policies. First, they 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 as SSP problems, where the agent tries to minimize cost before reaching a goal state.

Contributions: Motivated by these theoretical and practical concerns, we provide an algorithm for episodic SSP with LFA that is statistically and computationally efficient while using stationary policies. Beyond SSP, it is the first LFA algorithm with these three desirable properties 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

[1]Department of Electrical and Computer Engineering, University of Texas, Austin, TX, USA [2]Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL, USA. Correspondence to: Daniel Vial <dvial@utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

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 the minimal assumptions. Again, this is a first in the LFA literature, to the best of our knowledge.

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 & Wang, 2020; Zanette et al., 2020a), which (like us) assume linear costs and transitions: $c(s,a) = \phi(s,a)^\mathsf{T}\theta$ and $P(s'|s,a) = \phi(s,a)^\mathsf{T}\mu(s')$ 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 Berstein-style confidence sets for the related linear mixture model (LMM, see, e.g., (Ayoub et al., 2020; Jia et al., 2020)), where $P(s'|s,a) = \varphi(s'|s,a)^\mathsf{T}\vartheta$ for known $\varphi(s'|s,a)$.

Infinite horizon LFA: Comparatively little is known for infinite horizons. (Wei et al., 2021; Wu et al., 2022) 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 approximates the infinite horizon problem with a finite horizon one, so it uses non-stationary policies and requires knowledge of the span of the optimal value function in order to tune the finite horizon approximation. The third relies on even stronger assumptions. (Wu et al., 2022) proved $\sqrt{T}$ regret for the LMM, but the algorithm is inefficient due to computation[1] of $\sum_{s' \in \mathcal{S}} h(s')\varphi(s'|s,a)$ for certain $h \in \mathbb{R}^{\mathcal{S}}$. Analogous algorithms are proposed in (Zhou et al., 2020; 2021b) for discounted costs, which 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. For SSP in the LMM, the concurrent work (Min et al., 2022) establishes $\tilde{O}(\sqrt{B_\star^3 d^2 K/c_{min}})$ regret, where $B_\star$ is the maximal cost-to-go of the optimal policy (see Section 2). However, similar to the above LMM papers, their algorithm is inefficient as it involves summing over $\mathcal{S}$; it also requires access to (an estimate of) the unknown parameter $B_\star$, which we do not. Finally, the concurrent work (Chen et al., 2022) proves $\tilde{O}(\sqrt{B_\star^2 T_\star d^3 K})$ regret for SSP under our linearity assumption, where $T_\star \leq B_\star/c_{min}$ is the expected hitting time for the goal state under the optimal policy. However, their algorithm relies on a finite horizon approximation, so it uses non-stationary policies and requires access to the unknown parameters $B_\star$ and $T_\star$ (again, we do not). They also provide an efficient parameter-free algorithm with worse regret $\tilde{O}(\sqrt{B_\star^3 d^3 K/c_{min}})$, and an inefficient one with $\tilde{O}(\sqrt{B_\star^2 d^7 K})$ regret. These latter two algorithms also use finite horizon approximations and thus non-stationary policies.

Tabular SSP, c_{min}^{-1} dependent: (Tarbouriech et al., 2020) proved $\tilde{O}(D^{3/2}S\sqrt{AK/c_{min}})$ regret, where $D \geq B_\star$ is a measure of the SSP diameter (see their Assumption 2). (Rosenberg et al., 2020) improved this to $\tilde{O}(B_\star^{3/2}S\sqrt{AK/c_{min}})$. Both algorithms use Hoeffding-style confidence sets and can be generalized to the case $c_{min} = 0$, though regret becomes $K^{2/3}$ (see Section 6).

Tabular SSP, c_{min}^{-1} independent: (Rosenberg et al., 2020) also proved $\tilde{O}(B_\star S\sqrt{AK})$ regret when $c_{min} = 0$, and the lower bound $\tilde{\Omega}(B_\star\sqrt{SAK})$. Removing the c_{min}^{-1} dependence required Berstein-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 $\tilde{O}(B_\star S\sqrt{AK})$ 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 $[m] = \{1, \dots, m\}$. We write $\mathbb{1}(\cdot)$ for the indicator function. We let e_i be the vector with j-th element $e_i(j) = \mathbb{1}(i = j)$. For $x \in \mathbb{R}^d$ and positive definite $Y \in \mathbb{R}^{d\times d}$, we set $\|x\|_Y = \sqrt{x^\mathsf{T} Y x}$.

SSP: An SSP instance is defined by $(\mathcal{S}, \mathcal{A}, P, c, s_{goal})$, where $\mathcal{S}$ is a set of $S = |\mathcal{S}| < \infty$ states, $\mathcal{A}$ is a set of

[1] Appendix B of (Zhou et al., 2021b) 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.

$A = |\mathcal{A}| < \infty$ actions, P is the transition kernel, c is the cost vector, and $s_{goal} \in \mathcal{S}$ is an absorbing zero-cost state, i.e., $P(s_{goal}|s_{goal}, a) = 1$ and $c(s_{goal}, a) = 0$ for any $a \in \mathcal{A}$. A stationary and deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ induces a trajectory $\{s_t^\pi\}_{t=1}^\infty$, where s_1^π is an initial state and $s_{t+1}^\pi \sim P(\cdot|s_t^\pi, \pi(s_t^\pi))$ for $t \in \mathbb{N}$. We call π proper if s_{goal} is reached with probability 1 from any $s_1^\pi \in \mathcal{S}$; otherwise, we call it improper. We make the following assumption, 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 $(s, a) \in (\mathcal{S} \setminus \{s_{goal}\}) \times \mathcal{A}$, $c(s, a) \in [c_{min}, 1]$.*

For any $\pi : \mathcal{S} \to \mathcal{A}$, we define the (possibly infinite) cost-to-go function $J^\pi : \mathcal{S} \to \mathbb{R}$ by

$$J^\pi(s) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} c(s_t^\pi, \pi(s_t^\pi)) \middle| s_1^\pi = s\right].$$

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

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

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

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

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

Remark 1 (Positive costs)**.** *We require $c(s, a) \geq c_{min}$ 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 $\{\phi(s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}} \subset \mathbb{R}^d$, unknown $\theta \in \mathbb{R}^d$, and unknown $\{\mu(s')\}_{s'\in\mathcal{S}} \subset \mathbb{R}^d$, such that, for any $(s, a, s') \in (\mathcal{S} \setminus \{s_{goal}\}) \times \mathcal{A} \times \mathcal{S}$,*

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

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

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

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 $\phi(s_{goal}, a) = 0 \; \forall\, a \in \mathcal{A}$.

Remark 2 (Tabular case)**.** *Any SSP with $c(s, a) \in [0, 1]$ satisfies Assumption 2 with $d = SA$, $\phi(s, a) = e_{(s,a)}$, $\theta = c$, and $\mu(s') = \{P(s'|s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.*

Remark 3 (Realizability)**.** *As shown in Appendix E.2, Assumption 2 implies $Q^\star$ is linear in ϕ. Ideally, we would only assume this, but recent work for finite horizons (a special case of SSP) has shown this problem is much 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 $s_{h+1}^k = s_{goal}$, the episode terminates (without taking action a_{h+1}^k). We assume $s_1^k \neq s_{goal}$ without loss of generality but make no further assumptions on the sequence of initial states $\{s_1^k\}_{k=1}^K$. 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 $s_t' = s_{goal}$ 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

$$R(K) = \sum_{t=1}^{T} c(s_t, a_t) - \sum_{k=1}^{K} J^\star(s_1^k),$$

which is the difference between the total cost of the agent and the expected total cost of a "genie" who knows the optimal policy *a priori* and runs it for K episodes.

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.*

3. 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.

Proposition 1 (Feature space fixed point)**.** *Let Assumptions 1 and 2 hold. Define $G : \mathbb{R}^d \to \mathbb{R}^d$ by*

$$Gw = \theta + \sum_{s\in\mathcal{S}} \min_{a\in\mathcal{A}} \phi(s, a)^\mathsf{T} w\mu(s) \; \forall\, w \in \mathbb{R}^d.$$

[2] We reiterate T is random for SSP, unlike the fixed T used in the infinite horizon discussion of Section 1.

Then $w^\star = \theta + \sum_{s\in\mathcal{S}} J^\star(s)\mu(s)$ is a fixed point of G (i.e., $Gw^\star = w^\star$), $J^\star(s) = \min_{a\in\mathcal{A}} \phi(s,a)^\mathsf{T} w^\star$, and

$$\pi^\star(s) \in \arg\min_{a\in\mathcal{A}} \phi(s,a)^\mathsf{T} w^\star \;\forall\, s \in \mathcal{S}. \tag{6}$$

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 $\Lambda_t = I + \sum_{\tau=1}^t \phi(s_\tau,a_\tau)\phi(s_\tau,a_\tau)^\mathsf{T}$. Then the regularized least-squares estimate of Gw is

$$\tilde{G}w = \Lambda_t^{-1} \sum_{\tau=1}^t \phi(s_\tau,a_\tau)\Big(c(s_\tau,a_\tau) + \min_{a\in\mathcal{A}} \phi(s'_\tau,a)^\mathsf{T} w\Big).$$

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 $\tilde{w}^\star$ of $\tilde{G}$ and define policies like (6), with $w^\star$ replaced by $\tilde{w}^\star$. 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 optimistic estimates

$$f_t(s,w) = \min_{a\in\mathcal{A}} \left(\phi(s,a)^\mathsf{T} w - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}}\right), \tag{7}$$

where $\alpha_t > 0$ is the usual 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

$$g_t(s,w) = \min\{\max\{f_t(s,w), 0\}, B_t\}. \tag{8}$$

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 that do not arise in finite horizon LFA, or in finite horizon approximations of infinite horizon problems like the algorithms in (Wei et al., 2021; Chen et al., 2022) (see Section 1.1).

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 α_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 12). 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 essense, 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*

$$f_t(s,w) \leq J^\star(s) \;\forall\, s \in \mathcal{S}, \quad \|\hat{G}_t w - w\|_{\Lambda_t} \leq \alpha_t. \tag{9}$$

Note that by Cauchy-Schwarz, the latter bound implies

$$|\phi(s,a)^\mathsf{T}(\hat{G}_t w - w)| \leq \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}}. \tag{10}$$

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 $U_t w = \theta + \sum_{s\in\mathcal{S}} g_t(s,w)\mu(s)$. More specifically, we show that with high probability, for any bounded w,

$$|\phi(s,a)^\mathsf{T}(\hat{G}_t w - U_t w)| = O(\sqrt{\log t})\|\phi(s,a)\|_{\Lambda_t^{-1}}. \tag{11}$$

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 α_t.

For later use, we also note that by Assumption 2,

$$\phi(s,a)^\mathsf{T} U_t w = c(s,a) + \mathbb{E}_{s'} g_t(s',w), \tag{12}$$

where $\mathbb{E}_{s'}$ is expectation with respect to $s' \sim P(\cdot|s,a)$, i.e.,

$$\mathbb{E}_{s'} g_t(s',w) = \sum_{s'\in\mathcal{S}} g_t(s',w) P(s'|s,a).$$

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 $\{s_\tau, a_\tau, s'_\tau\}_{\tau=1}^t$, returns an OAFP w_t per Definition 1.

Inputs: The inputs are a failure probability δ and a sequence $\{\kappa_t\}_{t=1}^\infty$ that will be used to define the exploration parameter α_t in (7) (which we cannot do *a priori* due to Remark 6).

Remark 8 (Choice of κ_t). *In more detail, the forthcoming Line 22 of Algorithm 1 shows that $\alpha_t = \tilde{O}(\kappa_t)$ (where here the $\tilde{O}$ notation only shows dependence on κ_t), i.e., the algorithm is more explorative for larger κ_t. For now, we keep this parameter general. In the forthcoming theoretical results, we will specify an appropriate κ_t (i.e., an appropriate degree of exploration) under various assumptions.*

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 2-5 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 $w_0 = \alpha_0 = M_0 = 0$ to ensure the forthcoming notation is well-defined.

Episodic protocol: Lines 6-10, 15-16, and 27 implement the protocol from Section 2. Additionally, Line 9 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 16 updates Λ_t. When (or if) the last episode ends, Lines 11-14 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 $f_{M_{l-1}}(s'_t, w_{M_{l-1}}) \le B_\star$ by Definition 1, we know B_{t-1} was *not* an upper bound for $B_\star$, so we double it (Lines 17-18). Otherwise, we let $B_t = B_{t-1}$ (Lines 19-20). Having defined B_t, we use it and the input κ_t to define α_t (Line 22), as alluded to above.

Policy update conditions: Line 23 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 Λ_t doubles. The idea is that, before this doubling occurs,

$$\|\phi(s,a)\|_{\Lambda_t^{-1}} \le \sqrt{2}\|\phi(s,a)\|_{\Lambda_{M_{l-1}}^{-1}} \quad \forall\,(s,a) \in \mathcal{S}\times\mathcal{A},$$

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

Policy update: If any of the conditions are met, Lines 24-25 call the oracle for an OAFP w_{M_l} and end the current interval. Note that in the next interval, the policy in Line 9 will use this OAFP.

We can now present the main result of this section, Theorem 1. It assumes that the input κ_t to Algorithm 1 is $O(t^\lambda)$ for some $\lambda \in [0, \frac{1}{2})$, which implies the exploration parameter α_t is $\tilde{O}(t^\lambda)$ (see Remark 8). Provided this holds, the theorem shows that Algorithm 1 obtains $K^{\frac{1}{2}+\lambda}$ regret, i.e., smaller α_t yields lower regret. The tradeoff is that smaller α_t requires the OAFP to have lower Bellman error and yield sharper cost-to-go estimates (see Definition 1). In other words, larger λ implies larger α_t and thus (possibly unnecessary) exploration (see Remark 8), but it demands less of the oracle (see Definition 1), which will allow for guarantees under weaker assumptions.

Put differently, we would ideally choose $\lambda = 0$ to obtain $\sqrt{K}$ regret, but this choice means the oracle must compute an OAFP with $\tilde{O}(1)$ Bellman error. Computing such OAFPs will present a challenge when we discuss oracle implementation in the next section. However, if we disregard computation and assume access to an oracle that computes such OAFPs (e.g., by exhaustive search over an ϵ-net), choosing $\lambda = 0$ presents no issue (because the theorem also guarantees that such OAFPs exist).

Algorithm 1 Regret minimization with oracle

1: **Input:** $\delta \in (0,1)$, $\{\kappa_t\}_{t=1}^\infty \subset [1,\infty)$
2: $\Lambda_0 = I_d$ (regularizer), $B_0 = c_{min}$ ($B_\star$ bound)
3: $w_0 = 0_d$ (OAFP), $\alpha_0 = 0$ (explore parameter)
4: $M_0 = 0$ (time 0-th interval ended)
5: $t = 1$ (current time), $l = 1$ (current interval)
6: **for** episode $k = 1, \dots, K$ **do**
7: $h = 1$ (current step), observe $s_h^k \in \mathcal{S} \setminus \{s_{goal}\}$
8: **while** $s_h^k \neq s_{goal}$ **do**
9: Choose $a_h^k \in \mathcal{A}$ to minimize
$$\phi(s_h^k, a_h^k)^\mathsf{T} w_{M_{l-1}} - \alpha_{M_{l-1}} \|\phi(s_h^k, a_h^k)\|_{\Lambda_{M_{l-1}}^{-1}}$$
10: Observe $c(s_h^k, a_h^k)$ and $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$
11: **if** $k = K$ and $s_{h+1}^k = s_{goal}$ **then**
12: $L = l$ (total number intervals), $M_L = t$
13: $T = t$ (total time elapsed)
14: **else**
15: $(s_t, a_t, s'_t) = (s_h^k, a_h^k, s_{h+1}^k)$
16: $\Lambda_t = \Lambda_{t-1} + \phi(s_t, a_t)\phi(s_t, a_t)^\mathsf{T}$
17: **if** $f_{M_{l-1}}(s'_t, w_{M_{l-1}}) > B_{t-1}$ **then**
18: $B_t = 2B_{t-1}$
19: **else**
20: $B_t = B_{t-1}$
21: **end if**
22: $\alpha_t = (B_t+1)\kappa_t \sqrt{\log(t(B_t+1)\kappa_t/\delta)}$
23: **if** $t = 1$ or $s'_t = s_{goal}$ or $B_t \neq B_{t-1}$ or $det(\Lambda_t) \ge 2\det(\Lambda_{M_{l-1}})$ **then**
24: Call oracle for OAFP w_t (Def. 1)
25: $M_l = t$, $l \leftarrow l+1$
26: **end if**
27: $t \leftarrow t+1$, $h \leftarrow h+1$
28: **end if**
29: **end while**
30: **end for**

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 $1-\delta$, there exists an OAFP for all $t \in [T]$ and

$$R(K) = \tilde{O}\Big((B_\star^{\frac{3}{2}+\lambda} + B_\star^{\frac{1}{2}+\lambda}) d^{\frac{1}{2}} \Psi (K/c_{min})^{\frac{1}{2}+\lambda} + (B_\star+1)^{\frac{2}{1-2\lambda}} d^{\frac{1}{1-2\lambda}} \Psi^{\frac{2}{1-2\lambda}} c_{min}^{-\frac{1+2\lambda}{1-2\lambda}}\Big).$$

In summary, 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, $B_\star \geq 1$, and $\kappa_t = 9d$. With probability at least $1-\delta$, there exists an OAFP for all $t \in [T]$ and*

$$R(K) = \tilde{O}\Big(\sqrt{B_\star^3 d^3 K / c_{min}} + B_\star^2 d^3 / c_{min}\Big).$$

Note Corollary 1 also assumes $B_\star \geq 1$, which is natural (otherwise, $J^\star$ can arbitrarily smaller than the cost upper bound $1 \geq c(s,a)$). Of course, the $B_\star < 1$ case can be recovered from Theorem 1. Forthcoming results also assume $B_\star \geq 1$, but the appendix contains bounds for the case $B_\star < 1$ as well. Note the regret bound in Corollary 1 matches the only efficient parameter-free bound from the very recent paper (Chen et al., 2022), which does not use stationary policies (see Section 1.1).

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. Finally, we write $\approx$ to denote equalities that hold up to noise and bonus terms that are bounded by standard linear bandit techniques.

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

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

as the per-interval regret, and

$$\tilde{R}_2(\tilde{T}) = \sum_{l=0}^{\tilde{L}-1} J^\star(s_{1+M_l}) - \sum_{k=1}^{\tilde{K}} J^\star(s_1^k)$$

as the "excess regret" from intra-episode updates.

Cost-to-go bound: To bound both terms, we require a bound on B_t. Since $f_{M_{l-1}}(s'_t, w_{M_{l-1}}) \leq B_\star$ by Definition 1, as soon as B_{t-1} exceeds $B_\star$, the condition Line 17 will stop occurring. This implies $B_t \leq 2B_\star$.

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

$$\tilde{R}_1(\tilde{T}) \leq \sum_{l=0}^{\tilde{L}-1} \sum_{t=1+M_l}^{M_{l+1}} c(s_t, a_t) - f_{M_l}(s_{1+M_l}, w_{M_l}). \quad (13)$$

Now fix l and t as the double summation. Then by the chosen policy (Line 9 of Algorithm 1), we know

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

where $\approx$ hides the bonus term $\alpha_{M_l} \|\phi(s_t, a_t)\|_{\Lambda_{M_l}^{-1}}$. Again up to the bonus, (10) and (11) imply

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

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

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

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.[3] Both are $\tilde{O}(\sqrt{T \wedge \tilde{T}})$, because $\alpha_t = \tilde{O}(1)$ by $\lambda = 0$ and the above discussion that $B_t \leq 2B_\star = O(1)$.

Excess regret: By definition, $\tilde{R}_2(\tilde{T}) \leq B_\star(\tilde{L} - \tilde{K})$, 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 $\tilde{T}$).

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

$$(T \wedge \tilde{T}) c_{min} \leq \sum_{t=1}^{T \wedge \tilde{T}} c(s_t, a_t) = \tilde{R}(\tilde{T}) + \sum_{k=1}^{\tilde{K}} J^\star(s_1^k).$$

Combining, we obtain $T \wedge \tilde{T} = O(\sqrt{T \wedge \tilde{T}} + K)$, 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. □

[3] Again, we emphasize that $\approx$ suppresses terms that are handled using standard techniques; the ≈ 0 notation is not intended to suggest that $\tilde{R}_1(\tilde{T})$ is enirely negligible.

Algorithm 2 Computing OAFPs

Set $n = 1$, compute $\hat{G}_t^n 0 = \hat{G}_t 0$ and $\hat{G}_t^{n-1} 0 = 0$
while $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}\|_{\Lambda_t} > \alpha_t$ **do**
  Set $n \leftarrow n + 1$, compute $\hat{G}_t^n 0 = \hat{G}_t(\hat{G}_t^{n-1} 0)$
end while
Return $w_t = \hat{G}_t^{n-1} 0$

Remark 9 (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 10 (Complexity). *Algorithm 1's runtime is dominated by computation of $\{\pi(s_t)\}_{t=1}^T$, which is $O(Ad^2T)$ when Λ_t^{-1} and $det(\Lambda_t)$ are iteratively updated. In the proof, we show T is polynomial in all parameters (see Remark 14 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 $\alpha_t = \Omega(\sqrt{\log t})$,*

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

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

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

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

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

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

$$f_t(s, \hat{G}_t^n 0) \le \min_{a \in \mathcal{A}} Q^\star(s,a) = J^\star(s). \qquad \square$$

We thus propose Algorithm 2 for OAFP computation, which iterates $\hat{G}_t$ until the second inequality in (9) holds. The first inequality in (9) holds by Lemma 1, so the algorithm returns OAFPs *if* it terminates.[4] Our next result shows that, for appropriately chosen κ_t (see Remark 8), it indeed terminates in polynomially many iterations. Combined with Remark 10, 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, $B_\star \ge 1$, $\kappa_t = 54dt^{1/3}$, 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*

$$R(K) = \tilde{O}\Big(B_\star^{\frac{11}{6}} d^{\frac{3}{2}} (K/c_{min})^{\frac{5}{6}} + B_\star^6 d^9 c_{min}^{-5}\Big).$$

Proof sketch. The proof (and those of the forthcoming 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., $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \le \alpha_t$ for some $n = O(dt^{1/6})$. Equivalently, if we ignore the regularizer, then by definition of $\|\cdot\|_{\Lambda_t}$, we aim to show

$$\sum_{\tau=1}^{t} (\phi(s_\tau, a_\tau)^\mathsf{T} (\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0))^2 \le \alpha_t^2. \tag{14}$$

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

U_t converges: Lemma 4.3 of (Bonet, 2007) implies the standard Bellman iterates converge at rate $\frac{SA}{n}$. By (12), $\phi(s,a)^\mathsf{T} U_t^n 0$ 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

$$\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\mathsf{T} (U_t^n 0 - U_t^{n-1} 0)| = O\Big(\frac{d^2}{n}\Big). \tag{15}$$

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

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

On the other hand, (12) implies

$$|\phi(s,a)^\mathsf{T} (U_t x_n - y_{n+1})| \le \mathbb{E}_{s'} |g_t(s', x_n) - g_t(s', y_n)|.$$

Combining and using the triangle inequality, we obtain

$$\begin{aligned} |\phi(s,a)^\mathsf{T} (x_{n+1} - y_{n+1})| &= \tilde{O}(1) |\phi(s,a)\|_{\Lambda_t^{-1}} \\ &\quad + \mathbb{E}_{s'} |g_t(s', x_n) - g_t(s', y_n)|. \end{aligned} \tag{16}$$

Finally, a straightforward calculation yields

$$|g_t(s', x_n) - g_t(s', y_n)| \le \max_{a'} |\phi(s', a')^\mathsf{T} (x_n - y_n)|.$$

This suggests bounding the average in (16) by the max (over $s' \in \mathcal{S}$) and iterating. However, such a bound involves

[4] Experiments suggest that this termination occurs more generally than the setting of Theorem 2 – in particular, even when $\lambda = 0$, under which Corollary 1 promises a $\sqrt{K}$ regret scaling (unlike the forthcoming Theorem 2). Proving this, however, remained surprisingly elusive, and we leave it as an open problem.

$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|_{\Lambda_t^{-1}}$, which is too large for our purposes. The crucial idea of the proof is to take max only over the *explored* states $\mathcal{S}_t$, namely, those $s' \in \mathcal{S}$ for which $\max_{a'} \|\phi(s',a')\|_{\Lambda_t^{-1}} \ll \alpha_t^{-1}$. The key implication is that if s' is *un*explored, then $\alpha_t\|\phi(s',a')\|_{\Lambda_t^{-1}} \gg 1$ for some a', so $f_t(s',x_n) \le 0$ 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 $(s,a) \in \mathcal{S}_t \times \mathcal{A}$, to obtain

$$\max_{(s,a)\in\mathcal{S}_t\times\mathcal{A}} |\phi(s,a)^\mathsf{T}(x_n - y_n)| = \tilde{O}(n/\alpha_t).$$

Plugging into (16) and recalling $x_n = \hat{G}_t^n 0$ and $y_n = U_t^n 0$, this extends to *all* $(s,a) \in \mathcal{S}\times\mathcal{A}$ as follows:

$$|\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - U_t^n 0)| = \tilde{O}\Big(\|\phi(s,a)\|_{\Lambda_t^{-1}} + \frac{n}{\alpha_t}\Big). \quad (17)$$

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

Remark 11 (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 shows we* exploit *it on the high probability event.*

Remark 12 (Convergence)**.** *The proof sketch shows $\|x_t\|_{\Lambda_t} = O(t^{1/3})$, where $x_t = \hat{G}_t^{n_t} 0 - \hat{G}_t^{n_t-1} 0$ is the fixed point error after $n_t = O(t^{1/6})$ iterations. Note the norm equivalence $\|x_t\|_{\Lambda_t} = O(\sqrt{t})\|x_t\|_2$ always holds, so if it is reasonably tight (e.g., 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 technically stronger, it seems perfectly reasonable for, e.g., games that eventually end. The benefit is that the Bellman operator $\mathcal{T} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ given by

$$(\mathcal{T}Q)(s,a) = c(s,a) + \mathbb{E}_{s'} \min_{a'\in\mathcal{A}} Q(s',a') \quad (18)$$

is contractive. More precisely, for some $\rho \in (0,1)$ and $\omega(s) > 0$, if $\|x\| = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega(s)|x(s,a)|$, then

$$\|\mathcal{T}Q_1 - \mathcal{T}Q_2\| \le \rho\|Q_1 - Q_2\|. \quad (19)$$

Define $\chi = \max_{s\in\mathcal{S}} \omega(s) / \min_{s\in\mathcal{S}} \omega(s)$. Assuming nontrivial upper bounds $\bar{\rho} \in [\rho, 1)$ and $\bar{\chi} \in [\chi, \infty)$ are known, our next result establishes $K^{3/4}$ regret. (We soon show how this last assumption can be avoided.)

Theorem 3 (All proper)**.** *Suppose Assumptions 1 and 2 hold, $B_\star \ge 1$, all stationary policies are proper, $\kappa_t = 54dt^{1/4}\sqrt{N_t}$ with $N_t = \log(3t\bar{\chi})/(1-\bar{\rho})$, 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*

$$R(K) = \tilde{O}\Big(B_\star^{\frac{7}{4}} d^{\frac{3}{2}} (K/c_{min})^{\frac{3}{4}} N_t^{1/2} + B_\star^4 d^6 N_t^2 c_{min}^{-3}\Big).$$

Proof sketch. Recall in the Theorem 2 proof sketch, we showed $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}\|_{\Lambda_t} = O(\sqrt{t}(\frac{1}{n} + \frac{n}{\alpha_t}))$, 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 ρ^n. Hence, after N_t iterations, we have $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}\|_{\Lambda_t} = \tilde{O}(\sqrt{t}/\alpha_t) = \tilde{O}(\alpha_t)$ by the choice $\kappa_t = \tilde{O}(t^{1/4})$. □

Finally, we demonstrate a case where Algorithm 2 returns OAFPs for the best case κ_t from Corollary 1. This case generalizes the tabular one – where the features are the elementary basis vectors in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ – to allow for any $d \in \mathbb{N}$ and any orthogonal features. While arguably stylized, the main purpose is to exhibit a computationally efficient algorithm that achieves $\sqrt{K}$ regret beyond the tabular case, and to demonstrate a different proof technique. Moreover, this result will be generalized in the next section.

Theorem 4 (Orthogonal features)**.** *Suppose Assumptions 1 and 2 hold, $B_\star \ge 1$, $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}} \subset \{q_i\}_{i=1}^d$ 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 Λ_t^{-1}, which allows us to show $\hat{G}_t$ 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 $\bar{\chi}$ and $\bar{\rho}$ are unavailable, we can instead set $N_t = t^{2\gamma}$ for some constant $\gamma \in (0, \frac{1}{4})$ and modify Algorithm 2 to terminate after N_t iterations (if it has not already). This approach is efficient by design, returns OAFPs for $t \ge \Gamma = \tilde{O}((\frac{\log\chi}{1-\rho})^{\frac{1}{2\gamma}})$, and (combined with Algorithm 1) achieves the Theorem 1 regret bound with $\lambda = \frac{1}{4} + \gamma$ and an additive Γ 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 $c_{min} = 0$, which is the minimal assumption in tabular SSP (the upper bound $c(s,a) \leq 1$ can be easily generalized to bounded costs). In this setting, we define regret with respect to the optimal *proper* policy $\pi^\star_{\text{prop}}$. We use the same algorithms but replace $c(s,a)$ with $c(s,a) + \eta$ for a small perturbation $\eta > 0$ in the definition of $\hat{G}_t$, 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 $\pi^\star_{\text{prop}}$. With κ_t scaling as t^λ for some $\lambda \in [0, \frac{1}{2})$ (as in Theorem 1) and η as $K^{(2\lambda-1)/(2\lambda+3)}$, this yields $K^{(4\lambda+2)/(2\lambda+3)}$ regret.[5] 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 assumptions*. Note this also works if Assumption 1 holds but c_{min} is small. For example, Corollary 1 only promises linear regret when $c_{min} = K^{-1}$, but choosing $\eta = K^{-1/3}$ ensures $K^{2/3}$ regret.

Remark 13 (c_{min}^{-1} dependence)**.** *As seen above, the c_{min}^{-1} dependence of the leading term in our regret bound inflates the scaling in K when dealing with small costs. This issue also arises for the algorithm in (Min et al., 2022), the only efficient parameter-free algorithm in (Chen et al., 2022), and the earlier tabular algorithms, none of which use Bernstein-style confidence bounds (see Section 1.1). For LFA, these bounds have only been studied recently and only for simple finite horizon problems (see Section 1.1). Thus, given the unique LFA challenges that arise for SSP (see Remarks 4, 5, 6, and 7), we leave such bounds, and the improvement regarding c_{min}^{-1} dependence, for future work.*

7. Conclusion

In this paper, we presented 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.

Acknowledgements

This work was partially supported by ONR Grant N00014-19-1-2566, ARO Grant ARO W911NF-19-1-0379, and NSF Grants 1910112, 2019844, 2112471, 2106801, 1704970, and 1934986.

[5] If K is unknown, we can use a standard doubling trick.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pp. 2312–2320, 2011.

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

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

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

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021.

Chen, L., Jain, R., and Luo, H. Improved no-regret algorithms for stochastic shortest path with linear mdp. *International Conference on Machine Learning*, 2022.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Jafarnia-Jahromi, M., Chen, L., Jain, R., and Luo, H. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

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

Min, Y., He, J., Wang, T., and Gu, Q. Learning stochastic shortest path with linear function approximation. *International Conference on Machine Learning*, 2022.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Tarbouriech, J., Garcelon, E., Valko, M., Pirotta, M., and Lazaric, A. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34, 2021.

Wang, Y., Wang, R., and Kakade, S. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34, 2021.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

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

Wu, Y., Zhou, D., and Gu, Q. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3883–3913. PMLR, 2022.

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

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

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

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021.

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

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 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 $\gamma \in (0, \frac{1}{4})$ and set $\kappa_t = 54dt^{\frac{1}{4}}\sqrt{N_t}$ as in Theorem 3. We show in Appendix D (see Remark 16) that with probability at least $1 - \delta/2$, for any $t \geq (\log(3t\chi)/(1-\rho))^{\frac{1}{2\gamma}}$ 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 $\Gamma = \tilde{O}((\frac{\log \chi}{(1-\rho)})^{\frac{1}{2\gamma}})$. Then for the first Γ time steps, Algorithm 3 need not return an OAFP (though it will terminate, so everything is well-defined) but does thereafter. Using Assumption 1, we bound regret by Γ for the first Γ time steps, and by modifying the proof of Theorem 1, we can bound regret by $K^{\frac{3}{4}+\gamma}$ thereafter (in terms of K). Thus, regret will scale as $K^{\frac{3}{4}+\gamma}$ for this algorithm, with a second-order term Γ in addition to the one from Theorem 1.

Algorithm 3 Computing OAFPs with iteration limit

Set $n = 1$, compute $\hat{G}_t^n 0 = \hat{G}_t 0$ and $\hat{G}_t^{n-1} 0 = 0$
while $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}\|_{\Lambda_t} > \alpha_t$ and $n \leq N_t$ **do**
  Set $n \leftarrow n + 1$, compute $\hat{G}_t^n 0 = \hat{G}_t(\hat{G}_t^{n-1} 0)$
end while
Return $w_t = \hat{G}_t^{n-1} 0$

A.2. Generalizing Theorem 4

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

$$\Phi = \begin{bmatrix}\varphi_1 & \cdots & \varphi_{d'}\end{bmatrix} \in \mathbb{R}^{d\times d'}, \quad \Xi = \begin{bmatrix}\Phi & 0_{d\times(d''-d')}\end{bmatrix} \in \mathbb{R}^{d\times d''}.$$

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

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

so $\tilde{\phi}(s,a)^\mathsf{T}R^\mathsf{T}\theta = \phi(s,a)^\mathsf{T}\theta = c(s,a)$ and $\tilde{\phi}(s,a)^\mathsf{T}R^\mathsf{T}\mu(s') = \phi(s,a)^\mathsf{T}\mu(s') = P(s'|s,a)$, as desired. The first inequality in (4) holds by construction. For the second inequality in (4), note $\varphi_i^\mathsf{T}\theta$ is the cost of some state-action pair and thus lies in $[0,1]$ by Assumption 1. Combined with the fact that $\tilde{\Phi}$ is orthogonal,

$$\|R^\mathsf{T}\theta\|_2^2 = \|\tilde{\Phi}^\mathsf{T}R^\mathsf{T}\theta\|_2^2 = \|\Xi^\mathsf{T}\theta\|_2^2 = \sum_{i=1}^{d'}(\varphi_i^\mathsf{T}\theta)^2 \leq d' \leq d''.$$

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

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

Algorithmically, this means that if Φ is known *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)**.** *There exists a proper policy and* $c(s,a) \in [0,1]\ \forall\ (s,a) \in \mathcal{S} \times \mathcal{A}$.

Now suppose the SSP instance $(\mathcal{S}, \mathcal{A}, P, c, s_{goal})$ 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 $(\mathcal{S}, \mathcal{A}, P, c_\eta, s_{goal})$ 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 $\theta_\eta = \theta + \eta \sum_{s\in\mathcal{S}} \mu(s)$. Then since Assumption 2 holds for the original instance, we have

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

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

$$\sum_{t=1}^{T} c_\eta(s_t,a_t) - \sum_{k=1}^{K} J^\star_\eta(s_1^k) = \tilde{O}\left(\left(B_\star^{\frac{3}{2}+\lambda} + B_\star^{\frac{1}{2}+\lambda}\right) d^{\frac{1}{2}} \Psi (K/\eta)^{\frac{1}{2}+\lambda} + (B_\star+1)^{\frac{2}{1-2\lambda}} d^{\frac{1}{1-2\lambda}} \Psi^{\frac{2}{1-2\lambda}} \eta^{-\frac{1+2\lambda}{1-2\lambda}}\right).$$

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

$$J^\star_\eta(s) - J^\star(s) \le J^{\pi^\star_{\text{prop}}}_\eta(s) - J^{\pi^\star_{\text{prop}}}(s) \le \eta T_\star\ \forall\ s \in \mathcal{S},$$

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

$$\begin{aligned} R(K) &\le \sum_{t=1}^{T} c_\eta(s_t,a_t) - \sum_{k=1}^{K} J^{\pi^\star_{\text{prop}}}_\eta(s_1^k) + \sum_{k=1}^{K}\left(J^{\pi^\star_{\text{prop}}}_\eta(s_1^k) - J^{\pi^\star_{\text{prop}}}(s_1^k)\right) \\ &= \tilde{O}\left(\left(B_\star^{\frac{3}{2}+\lambda} + B_\star^{\frac{1}{2}+\lambda}\right) d^{\frac{1}{2}} \Psi (K/\eta)^{\frac{1}{2}+\lambda} + \eta T_\star K + (B_\star+1)^{\frac{2}{1-2\lambda}} d^{\frac{1}{1-2\lambda}} \Psi^{\frac{2}{1-2\lambda}} \eta^{-\frac{1+2\lambda}{1-2\lambda}}\right). \end{aligned}$$

Choosing η 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 $\frac{2\lambda+1}{2\lambda+3} \le \frac{4\lambda+2}{2\lambda+3} < 1$ 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 $\mathbb{E}_t$ for expectation conditioned on the first $t-1$ state-action-state triples and the t-th state action pair, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\{(s_\tau, a_\tau, s'_\tau\}_{\tau=1}^{t-1} \cup \{s_t, a_t\}]$. We let $\mathbb{E}_{s'_t}$ denote expectation with respect to s'_t. Hence, for $h : \mathcal{S} \to \mathbb{R}$,

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

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 $\mathbb{E}_\tau[g_t(s'_\tau, w)] \ne \sum_{s\in\mathcal{S}} g_t(s,w) P(s|s_\tau,a_\tau)$ for some $w \in \mathbb{R}^d$. On the other hand, $\mathbb{E}_{s'_\tau}[g_t(s'_\tau, w)] = \sum_{s\in\mathcal{S}} g_t(s,w) P(s|s_\tau,a_\tau)$ 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

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

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, $U_t w$ 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:

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

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

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

$$\Pi_{[0,b]}(x) = \min\{\max\{x, 0\}, b\} = \max\{\min\{x, b\}, 0\}. \tag{20}$$

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 Λ_t lie in $[1, t+1]$, and*

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

Proof. Let $\{\lambda_i\}_{i=1}^d$ and $\{q_i\}_{i=1}^d$ be the eigenvalues and (unit) eigenvectors of Λ_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 $0 \le (\phi(s_\tau,a_\tau)^\mathsf{T} q_i)^2 \le \|\phi(s_\tau,a_\tau)\|_2\|q_i\|_2 \le 1$ by Cauchy-Schwarz. For the norm equivalences, we first use the eigenvalue bounds to write

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

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. □

Claim 2 ($\Pi_{[0,b]}$ properties)**.** *For any $b > 0$ and $x, y \in \mathbb{R}$, $\Pi_{[0,b]}(x) \le \max\{x, 0\}$ and $|\Pi_{[0,b]}(x) - \Pi_{[0,b]}(y)| \le |x - y|$.*

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

Claim 3 (g_t bounds)**.** *If Assumption 2 holds, then for any $t \in [T]$, $s \in \mathcal{S}$, and $w_1, w_2 \in \mathbb{R}^d$,*

$$|g_t(s,w_1) - g_t(s,w_2)| \le |f_t(s,w_1) - f_t(s,w_2)| \le \max_{a \in \mathcal{A}} |\phi(s,a)^\mathsf{T}(w_1 - w_2)| \le \|w_1 - w_2\|_2 \le \sqrt{d}\|w_1 - w_2\|_\infty.$$

Proof. The first bound follows from Claim 2. For the second, let $\bar{a} \in \mathcal{A}$ be any action attaining the minimum in the definition of $f_t(s,w_1)$, i.e., $\bar{a} \in \arg\min_{a \in \mathcal{A}}(\phi(s,a)^\mathsf{T} w_1 - \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}})$. Then

$$\begin{aligned} f_t(s,w_1) - f_t(s,w_2) &\ge f_t(s,w_1) - \left(\phi(s,\bar{a})^\mathsf{T} w_2 - \alpha_t\|\phi(s,\bar{a})\|_{\Lambda_t^{-1}}\right) \\ &= \phi(s,\bar{a})^\mathsf{T}(w_1 - w_2) \ge -\max_{a \in \mathcal{A}} |\phi(s,a)^\mathsf{T}(w_1 - w_2)|. \end{aligned}$$

By symmetry, we also have $f_t(s,w_1) - f_t(s,w_2) \le \max_{a \in \mathcal{A}} |\phi(s,a)^\mathsf{T}(w - w')|$; 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 $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|\hat{G}_t w\|_\infty \le \sqrt{td}\left(1 + \max_{s \in \mathcal{S}} g_t(s,w)\right), \quad \|\hat{G}_t 0\|_{\Lambda_t} \le \sqrt{t+1}\|\hat{G}_t 0\|_2 \le 2\sqrt{(t+1)d},$$
$$\|U_t w\|_2 \le \sqrt{d}\left(1 + \max_{s \in \mathcal{S}} g_t(s,w)\right), \quad \phi(s,a)^\mathsf{T} U_t w \in [0, B_t + 1].$$

Proof. First observe that by a standard norm equivalence and Claim 1, for any $w \in \mathbb{R}^d$, we have

$$\|\Lambda_t^{-1} w\|_\infty \le \|\Lambda_t^{-1} w\|_2 = \|\Lambda_t^{-1/2} w\|_{\Lambda_t^{-1}} \le \|\Lambda_t^{-1/2} w\|_2 = \|w\|_{\Lambda_t^{-1}} \le \|w\|_2. \tag{21}$$

Combined with Cauchy-Schwarz and Lemma D.1 of (Jin et al., 2020), we obtain

$$\sum_{\tau=1}^t \|\Lambda_t^{-1}\phi(s_\tau, a_\tau)\|_\infty \le \sum_{\tau=1}^t \|\phi(s_\tau, a_\tau)\|_{\Lambda_t^{-1}} \le \sqrt{t \sum_{\tau=1}^t \|\phi(s_\tau, a_\tau)\|^2_{\Lambda_t^{-1}}} \le \sqrt{td},$$

so the first $\hat{G}_t$ bound follows from the triangle inequality. Next, because $g_t(s,0) = 0 \; \forall \, s \in \mathcal{S}$, we have

$$\hat{G}_t 0 = \Lambda_t^{-1} \sum_{\tau=1}^t \phi(s_\tau, a_\tau) c(s_\tau, a_\tau) = \Lambda_t^{-1} \sum_{\tau=1}^t \phi(s_\tau, a_\tau)\phi(s_\tau, a_\tau)^\mathsf{T}\theta = \Lambda_t^{-1}(\Lambda_t - I)\theta = (I - \Lambda_t^{-1})\theta.$$

Therefore, by Claim 1 and (21), we obtain

$$\|\hat{G}_t 0\|_{\Lambda_t} \le \sqrt{t+1}\|\hat{G}_t 0\|_2 \le \sqrt{t+1}(\|\theta\|_2 + \|\Lambda_t^{-1}\theta\|_2) \le 2\sqrt{t+1}\|\theta\|_2 \le 2\sqrt{(t+1)d}.$$

Finally, the U_t bounds hold by assumption. □

Claim 5 ($B_\star$ estimate)**.** *If Assumption 1 holds, then* $\sup_{t \ge 0} B_t \le 2B_\star$.

Proof. Since $B_0 = c_{min} \le B_\star$ by assumption, it suffices to show $B_\tau \le 2B_\star \; \forall \, \tau \in \mathbb{N}$. Suppose instead that $B_\tau > 2B_\star$ for some such τ. Let $t = \min\{\tau \in \mathbb{N} : B_\tau > 2B_\star\}$ be the first time it occurs. Then by Algorithm 1, we have $B_\star < B_t/2 = B_{t-1} < f_{M_{l-1}}(s'_t, w_{M_{l-1}})$ 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_l} is an OAFP (see Definition 1). □

B.3. Operator concentration

Define the random variables W_t and ε_t, and the event $\mathcal{E}$, by

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

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* $\min_{t \in \mathbb{N}} \kappa_t \ge 9d$*, then* $\mathbb{P}(\mathcal{E}) \ge 1 - \delta/2$.

The error bound ε_t in the lemma is related to the exploration parameter α_t in the following manner.

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

Proof. For the first bound, since $\log x \le x \; \forall \, x \in \mathbb{R}$, we have

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

Combined with $5\sqrt{3/2} \le 9$ and the assumption $\kappa_t \ge 9md$, we obtain

$$\begin{aligned} m\varepsilon_t &= 5md\sqrt{\log(t\alpha_t/\delta)}(B_t + 1) \le 5\sqrt{3/2}md\sqrt{\log(t(B_t + 1)\kappa_t/\delta)}(B_t + 1) \\ &\le 9md\sqrt{\log(t(B_t + 1)\kappa_t/\delta)}(B_t + 1) \le \kappa_t\sqrt{\log(t(B_t + 1)\kappa_t/\delta)}(B_t + 1) = \alpha_t. \end{aligned}$$

For the second bound, simply note $\log(t(B_t + 1)\kappa_t/\delta) \ge \log(9md) \ge 1$ and use the definition of α_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 $\mathcal{E}$,*

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

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,

$$|\phi(s,a)^\mathsf{T}(\hat{G}_t w_t - U_t w_t)| = |\phi(s,a)^\mathsf{T} E_t w_t| \leq \|\phi(s,a)\|_{\Lambda_t^{-1}} \|E_t\|_{\Lambda_t} \leq \|\phi(s,a)\|_{\Lambda_t^{-1}} \alpha_t. \qquad \square$$

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 $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$|\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - U_t(\hat{G}_t^{n-1} 0))| \leq \|\hat{G}_t^n 0 - U_t(\hat{G}_t^{n-1} 0)\|_{\Lambda_t} \|\phi(s,a)\|_{\Lambda_t^{-1}} \leq \varepsilon_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \leq \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} / m.$$

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

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 $t \in [T]$,*

$$0 \leq \cdots \leq \phi(s,a)^\mathsf{T} U_t^{n-1} 0 \leq \phi(s,a)^\mathsf{T} U_t^n 0 \leq \cdots \leq B_t + 1 \; \forall \, (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{22}$$

$$\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(U_t^{n+1} 0 - U_t^n 0)| \leq \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0)| \; \forall \, n \in \mathbb{N}. \tag{23}$$

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

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

Now assuming $\min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0) \geq 0$, we have $\min_{s' \in \mathcal{S}}(g_t(s', U_t^n 0) - g_t(s', U_t^{n-1} 0)) \geq 0$, so

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

Finally, (23) follows from Claim 3:

$$\begin{aligned} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(U_t^{n+1} 0 - U_t^n 0)| &\leq \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s' \in \mathcal{S}} |g_t(s', U_t^n 0) - g_t(s', U_t^{n-1} 0)| P(s'|s,a) \\ &\leq \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0)|. \end{aligned} \qquad \square$$

The proof of Claim 7 shows that U_t is nonexpansive in the induced ℓ_∞ norm $\|\cdot\| = \|\Phi^\mathsf{T} \cdot\|_\infty$, where Φ 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 $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$, $\Upsilon \in \mathbb{R}^{d \times \mathcal{Z}}$, and $r = rank(\Upsilon)$. For any $\mathcal{Z}' \subset \mathcal{Z}$, denote by $\Upsilon(\mathcal{Z}')$ the submatrix of Υ with columns $\mathcal{Z}'$, with $\Upsilon(z) = \Upsilon(\{z\})$ for any $z \in \mathcal{Z}$ for simplicity. Then there exists $\mathcal{Z}' \subset \mathcal{Z}$ such that $|\mathcal{Z}'| = r$ and $\|\Upsilon^\mathsf{T} x\|_\infty \leq r \|\Upsilon(\mathcal{Z}')^\mathsf{T} x\|_\infty$ for any $x \in \mathbb{R}^d$.*

Proof. We first assume $r = d$. Then we can find $\mathcal{Z}'' \subset \mathcal{Z}$ such that $|\mathcal{Z}''| = d$ and $det(\Upsilon(\mathcal{Z}'')) \neq 0$. Let $\mathcal{Z}'$ be whichever such $\mathcal{Z}''$ maximizes $|det(\Upsilon(\mathcal{Z}''))|$. Set $H = \Upsilon^\mathsf{T}(\Upsilon(\mathcal{Z}')^\mathsf{T})^{-1}$, which is well-defined by choice of $\mathcal{Z}'$. 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

$$\|\Upsilon^\mathsf{T} x\|_\infty = \|H\Upsilon(\mathcal{Z}')^\mathsf{T} x\|_\infty \leq \|H\|_\infty \|\Upsilon(\mathcal{Z}')^\mathsf{T} x\|_\infty \leq d \max_{z \in \mathcal{Z}, z' \in \mathcal{Z}'} |H(z, z')| \|\Upsilon(\mathcal{Z}')^\mathsf{T} x\|_\infty.$$

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

$$\Upsilon(\mathcal{Z}' \cup \{z\} \setminus \{z'\}) = \Upsilon(\mathcal{Z}', \Upsilon(z)) = \Upsilon\left(\mathcal{Z}', \sum_{z'' \in \mathcal{Z}'} H(z, z'')\Upsilon(z'')\right).$$

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

$$det\left(\Upsilon\left(\mathcal{Z}', \sum_{z'' \in \mathcal{Z}'} H(z, z'')\Upsilon(z'')\right)\right) = \sum_{z'' \in \mathcal{Z}'} H(z, z'') det(\Upsilon(\mathcal{Z}', \Upsilon(z''))) = H(z, z') det(\Upsilon(\mathcal{Z}')).$$

Combining the previous two identities with the definition of $\mathcal{Z}'$ yields the desired bound:

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

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

$$\Upsilon^\mathsf{T} x = \Upsilon^\mathsf{T} U U^\mathsf{T} x = \begin{bmatrix}\tilde{\Upsilon}^\mathsf{T} & 0\end{bmatrix} \begin{bmatrix}\tilde{x} \\ U_2^\mathsf{T} x\end{bmatrix} = \tilde{\Upsilon}^\mathsf{T}\tilde{x}.$$

Therefore, by the choice of $\mathcal{Z}'$, we have

$$\|\Upsilon^\mathsf{T} x\|_\infty = \|\tilde{\Upsilon}^\mathsf{T}\tilde{x}\|_\infty \leq r\|\tilde{\Upsilon}(\mathcal{Z}')^\mathsf{T}\tilde{x}\|_\infty = r\|\Upsilon(\mathcal{Z}')^\mathsf{T} x\|_\infty. \qquad \square$$

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 $n \in \mathbb{N}$,*

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

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

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

Then combining Claims 7 and 8, we can find $\mathcal{Z}' \subset \mathcal{S} \times \mathcal{A}$ such that, for any $m \in [n]$,

$$\frac{(B_t + 1)|\mathcal{Z}'|}{n} \leq \frac{(B_t + 1)d^2}{nd} < \frac{1}{d} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)^\mathsf{T}(U_t^m 0 - U_t^{m-1} 0)| \leq \max_{(s,a) \in \mathcal{Z}'} |\phi(s, a)^\mathsf{T}(U_t^m 0 - U_t^{m-1} 0)|.$$

Thus, for each $m \in [n]$, we can find $z_m \in \mathcal{Z}'$ with $|\phi(z_m)^\mathsf{T}(U_t^m 0 - U_t^{m-1} 0)| > (B_t + 1)|\mathcal{Z}'|/n$. But by Claim 7,

$$n = \sum_{z \in \mathcal{Z}'} \sum_{m=1}^{n} \mathbb{1}(z_m = z) < \frac{n}{(B_t + 1)|\mathcal{Z}'|} \sum_{z \in \mathcal{Z}'} \sum_{m=1}^{n} \phi(z)^\mathsf{T}(U_t^m 0 - U_t^{m-1} 0) = \frac{n}{(B_t + 1)|\mathcal{Z}'|} \sum_{z \in \mathcal{Z}'} \phi(z)^\mathsf{T} U_t^n 0 \leq n,$$

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 $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\phi(s,a)^\mathsf{T} U_t^{n-1} 0 \leq Q^\star(s,a).$$

Proof. We fix t and use induction on n. For $n=1$, we simply have $\phi(s,a)^\mathsf{T} U_t^{n-1} 0 = 0 \leq Q^\star(s,a)$. Assuming true for $n \in \mathbb{N}$, the Bellman optimality equation (1) implies

$$f_t(s, U_t^{n-1}0) = \min_{a\in\mathcal{A}} \left(\phi(s,a)^\mathsf{T} U_t^{n-1} 0 - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \right) \leq \min_{a\in\mathcal{A}} Q^\star(s,a) = J^\star(s) \;\forall\, s \in \mathcal{S}.$$

Hence, by Claim 2, $g_t(s, U_t^{n-1}0) \leq \max\{J^\star(s), 0\} = J^\star(s)$. Again using Bellman optimality, we thus obtain

$$\phi(s,a)^\mathsf{T} U_t^n 0 = c(s,a) + \sum_{s'\in\mathcal{S}} g_t(s', U_t^{n-1}0) P(s'|s,a) \leq c(s,a) + \sum_{s'\in\mathcal{S}} J^\star(s') P(s'|s,a) = Q^\star(s,a). \qquad \square$$

We now establish existence of OAFPs. First note that by Claim 3 and Lemma 3, for any norm $\|\cdot\|$, we have

$$\|U_t^{n+1}0 - U_t^n 0\| \leq \sqrt{d} \max_{s\in\mathcal{S}} |g_t(s, U_t^n 0) - g_t(s, U_t^{n-1}0)| \sqrt{d} \leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1}0)| \xrightarrow[n\to\infty]{} 0.$$

Hence, $w_t^\star = \lim_{n\to\infty} U_t^n 0$ exists. By continuity, it is a fixed point:

$$w_t^\star = \lim_{n\to\infty} U_t^{n+1} 0 = \lim_{n\to\infty} U_t(U_t^n 0) = U_t\left(\lim_{n\to\infty} U_t^n 0\right) = U_t(w_t^\star).$$

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

$$\|w_t^\star\|_\infty \leq \|w_t^\star\|_2 = \|U_t w_t^\star\|_2 \leq \sqrt{d}\left(1 + \max_{s\in\mathcal{S}} g_t(s, w_t^\star)\right) \leq \sqrt{d}(1+B_t) \leq \alpha_t + \sqrt{td}(B_t+1) = W_t.$$

By Claim 6, on the event $\mathcal{E}$, this implies

$$\|\hat{G}_t w_t^\star - w_t^\star\|_{\Lambda_t} = \|\hat{G}_t w_t^\star - U_t w_t^\star\|_{\Lambda_t} = \|E_t w_t^\star\|_{\Lambda_t} \leq \varepsilon_t \leq \alpha_t.$$

Finally, for any $s \in \mathcal{S}$, using continuity, Lemma 4, and Bellman optimality, we obtain

$$f_t(s, w_t^\star) \leq \min_{a\in\mathcal{A}} \phi(s,a)^\mathsf{T} w_t^\star = \min_{a\in\mathcal{A}} \phi(s,a)^\mathsf{T} \lim_{n\to\infty} U_t^n 0 = \min_{a\in\mathcal{A}} \lim_{n\to\infty} \phi(s,a)^\mathsf{T} U_t^n 0 \leq \min_{a\in\mathcal{A}} Q^\star(s,a) = J^\star(s).$$

Hence, on the event $\mathcal{E}$, for any $t \in [T]$, $w_t^\star$ 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 $\tilde{L} = \min\{l \in [L] : M_l \geq T \wedge \tilde{T}\}$ denote the least number of intervals that encompass the times $1, \ldots, T \wedge \tilde{T}$. Also let $\tilde{K} = |\{t \in [T \wedge \tilde{T}] : s_t' = s_{goal}\}|$ 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

$$\tilde{R}(\tilde{T}) = \sum_{t=1}^{T\wedge\tilde{T}} c(s_t, a_t) - \sum_{k=1}^{\tilde{K}} J^\star(s_1^k) < \infty.$$

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

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

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

$$\sum_{t=1}^{T\wedge\tilde{T}} c(s_t, a_t) \leq 1 + \sum_{t=2}^{T\wedge\tilde{T}} c(s_t, a_t) = 1 + \sum_{l=1}^{\tilde{L}-1} \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} c(s_t, a_t), \tag{24}$$

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

$$\sum_{k=1}^{\tilde{K}} J^\star(s_1^k) = \sum_{l=1}^{\tilde{L}} J^\star(s_{1+M_l}) + \left(\sum_{k=1}^{\tilde{K}} J^\star(s_1^k) - \sum_{l=1}^{\tilde{L}} J^\star(s_{1+M_l}) \right). \tag{25}$$

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 $k \in [\tilde{K}]$, we can find $l \in [\tilde{L}]$ such that $s_1^k = s_{1+M_l}$. Hence, all summands cancel, except those corresponding to intervals $\mathcal{L} = \{l \in [\tilde{L}] : M_l = 1 \text{ or } B_{M_l} = 2B_{M_l-1} \text{ or } det(\Lambda_{M_l}) \geq 2det(\Lambda_{M_{l-1}})\}$, which may not have reached the goal state. Taken together, and since $J^\star(s_{1+M_l}) \leq B_\star$, we obtain

$$\sum_{k=1}^{\tilde{K}} J^\star(s_1^k) - \sum_{l=1}^{\tilde{L}} J^\star(s_{1+M_l}) \geq -\sum_{l\in\mathcal{L}} J^\star(s_{1+M_l}) \geq -B_\star|\mathcal{L}|. \tag{26}$$

It remains to bound $|\mathcal{L}|$. Clearly, $|\mathcal{L}| \leq 1 + \sum_{i=1}^2 |\mathcal{L}_i|$, where $\mathcal{L}_1 = \{l \in [\tilde{L}] : B_{M_l} = 2B_{M_l-1}\}$ and $\mathcal{L}_2 = \{l \in [\tilde{L}] : det(\Lambda_{M_l}) \geq 2det(\Lambda_{M_{l-1}})\}$. 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,

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

For $\mathcal{L}_2$, we have $|\mathcal{L}_2| \leq |\mathcal{L}_2'| + 1$, where $\mathcal{L}_2' = \{l \in [\tilde{L} - 1] : det(\Lambda_{M_l}) \geq 2det(\Lambda_{M_{l-1}})\}$ excludes $\tilde{L} - 1$ if it belongs to $\mathcal{L}_2$. By definition, $M_{\tilde{L}-1} < T \wedge \tilde{T}$, which by Claim 1 implies $det(\Lambda_{M_{\tilde{L}-1}}) \leq (1 + (T \wedge \tilde{T}))^d \leq (2(T \wedge \tilde{T}))^d$. Hence, because $det(\Lambda_{M_{\tilde{L}-1}}) \geq 2^{|\mathcal{L}_2'|} det(\Lambda_0) = 2^{|\mathcal{L}_2'|}$ by definition of $\mathcal{L}_2'$, we obtain

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

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

$$\begin{aligned}\tilde{R}(\tilde{T}) - \sum_{l=1}^{\tilde{L}-1} \left(\sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} c(s_t, a_t) - J^\star(s_{1+M_l}) \right) &\leq 1 + B_\star(\log_2(2B_\star/c_{min}) + d\log_2(2(T \wedge \tilde{T})) + 2) \\ &\leq 2(B_\star + 1) d \log_2(4B_\star(T \wedge \tilde{T})/c_{min}),\end{aligned}$$

where the last inequality uses $B_\star \geq c_{min}$ and $d \geq 2$. □

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 $l \in [\tilde{L} - 1]$,*

$$\sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} c(s_t, a_t) - J^\star(s_{1+M_l}) \leq \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} (g_{M_l}(s_t', w_{M_l}) - \mathbb{E}_t[g_{M_l}(s_t', w_{M_l})]) + 3\alpha_{M_l} \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} \|\phi(s_t, a_t)\|_{\Lambda_{M_l}^{-1}}.$$

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

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

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

$$f_{M_l}(s_{1+M_l}, w_{M_l}) \le J^\star(s_{1+M_l}), \quad \|\hat{G}_{M_l} w_{M_l} - w_{M_l}\|_{\Lambda_{M_l}} \le \alpha_{M_l}. \tag{30}$$

Using the first inequality in (30), we obtain

$$\sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} c(s_t,a_t) - J^\star(s_{1+M_l}) \le \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} c(s_t,a_t) - f_{M_l}(s_{1+M_l}, w_{M_l}) = \gamma_{1+M_l},$$

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

$$\begin{aligned} f_{M_l}(s_\tau, w_{M_l}) &= \phi(s_\tau,a_\tau)^\mathsf{T}\hat{G}_{M_l} w_{M_l} + \phi(s_\tau,a_\tau)^\mathsf{T}(w_{M_l} - \hat{G}_{M_l} w_{M_l}) - \alpha_{M_l}\|\phi(s_\tau,a_\tau)\|_{\Lambda^{-1}_{M_l}} \\ &\ge \phi(s_\tau,a_\tau)^\mathsf{T}\hat{G}_{M_l} w_{M_l} - 2\alpha_{M_l}\|\phi(s_\tau,a_\tau)\|_{\Lambda^{-1}_{M_l}} \ge \phi(s_\tau,a_\tau)^\mathsf{T} U_{M_l} w_{M_l} - 3\alpha_{M_l}\|\phi(s_\tau,a_\tau)\|_{\Lambda^{-1}_{M_l}}, \end{aligned}$$

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

$$\begin{aligned} \phi(s_\tau,a_\tau)^\mathsf{T} U_{M_l} w_{M_l} &= c(s_\tau,a_\tau) + \sum_{s\in\mathcal{S}} g_{M_l}(s, w_{M_l}) P(s|s_\tau,a_\tau) \\ &= c(s_\tau,a_\tau) + g_{M_l}(s'_\tau, w_{M_l}) + \mathbb{E}_\tau[g_{M_l}(s'_\tau, w_{M_l})] - g_{M_l}(s'_\tau, w_{M_l}). \end{aligned}$$

Combining the previous two inequalities and rearranging, we obtain

$$c(s_\tau,a_\tau) - f_{M_l}(s_\tau, w_{M_l}) \le -g_{M_l}(s'_\tau, w_{M_l}) + g_{M_l}(s'_\tau, w_{M_l}) - \mathbb{E}_\tau[g_{M_l}(s'_\tau, w_{M_l})] + 3\alpha_{M_l}\|\phi(s_\tau,a_\tau)\|_{\Lambda^{-1}_{M_l}}. \tag{31}$$

We complete the proof separately in each of two cases.

- If $\tau = M_{l+1}\wedge\tilde{T}$, the left side of (31) is γ_τ and $-g_{M_l}(s'_\tau, w_{M_l}) \le 0 = \gamma_{\tau+1}$, so (31) implies (29).
- Otherwise, an interval did *not* end between times $1+M_l$ and τ (inclusive). This implies (A) $s'_\tau \ne s_{goal}$, so $s'_\tau = s_{\tau+1}$, (B) $B_\tau = B_{M_l}$, and (C) $f_{M_l}(s'_\tau, w_{M_l}) \le B_\tau$. Taken together, (B) and (C) give $f_{M_l}(s'_\tau, w_{M_l}) \le B_{M_l}$, so (D) $-g_{M_l}(s'_\tau, w_{M_l}) \le -f_{M_l}(s'_\tau, w_{M_l})$ by definition. Combining (A) and (D) with (31), we obtain

  $$c(s_\tau,a_\tau) - f_{M_l}(s_\tau, w_{M_l}) \le -f_{M_l}(s_{\tau+1}, w_{M_l}) + g_{M_l}(s'_\tau, w_{M_l}) - \mathbb{E}_\tau[g_{M_l}(s'_\tau, w_{M_l})] + 3\alpha_{M_l}\|\phi(s,a)\|_{\Lambda^{-1}_{M_l}}.$$

  Hence, recalling $\tau < M_{l+1}\wedge\tilde{T}$, we can use the definitions of γ_τ and $\gamma_{\tau+1}$ to obtain

  $$\begin{aligned} \gamma_\tau &= \sum_{t=\tau+1}^{M_{l+1}} c(s_t,a_t) + c(s_\tau,a_\tau) - f_{M_l}(s_\tau, w_{M_l}) \\ &\le \sum_{t=\tau+1}^{M_{l+1}} c(s_t,a_t) - f_{M_l}(s_{\tau+1}, w_{M_l}) + g_{M_l}(s'_\tau, w_{M_l}) - \mathbb{E}_\tau[g_{M_l}(s'_\tau, w_{M_l})] + 3\alpha_{M_l}\|\phi(s,a)\|_{\Lambda^{-1}_{M_l}} \\ &= \gamma_{\tau+1} + g_{M_l}(s'_\tau, w_{M_l}) - \mathbb{E}_\tau[g_{M_l}(s'_\tau, w_{M_l})] + 3\alpha_{M_l}\|\phi(s,a)\|_{\Lambda^{-1}_{M_l}}. \end{aligned}$$

  □

We next bound the martingale difference sequence from Lemma 6.

Lemma 7 (Martingale difference sequence)**.** *Under the assumptions of Theorem 1, for any $\delta > 0$, if we define*

$$\mathcal{F} = \left\{ \sum_{l=1}^{\tilde{L}-1} \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} (g_{M_l}(s'_t, w_{M_l}) - \mathbb{E}_t[g_{M_l}(s'_t, w_{M_l})]) \le 2B_\star \sqrt{(T\wedge\tilde{T}) \log\frac{8(T\wedge\tilde{T})}{\delta}} \right\},$$

then $\mathbb{P}(\mathcal{F}) \ge 1 - \delta/2$.

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

$$\sum_{l=1}^{\tilde{L}-1} 3\alpha_{M_l} \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} \|\phi(s_t,a_t)\|_{\Lambda_{M_l}^{-1}} \le 6\sqrt{(T\wedge\tilde{T})d\log(2(T\wedge\tilde{T}))} \max_{t\in[T\wedge\tilde{T}]} \alpha_t.$$

Proof. For each $l \in [\tilde{L}-1]$ and $t \in \{2+M_l, \dots, M_{l+1}\wedge\tilde{T}\}$, the $(l+1)$-th interval did *not* end at time $t-1$, which implies $det(\Lambda_{t-1}) \le 2det(\Lambda_{M_l})$. By Lemma 12 of (Abbasi-Yadkori et al., 2011) this implies $\Lambda_{M_l} - \Lambda_{t-1}/2$ is positive semidefinite, so $\Lambda_{M_l}^{-1} - 2\Lambda_{t-1}^{-1}$ is negative semidefinite, so

$$\|\phi(s_t,a_t)\|_{\Lambda_{M_l}^{-1}} = \sqrt{\phi(s_t,a_t)^\intercal \Lambda_{M_l}^{-1}\phi(s_t,a_t)} \le \sqrt{2\phi(s_t,a_t)^\intercal \Lambda_{t-1}^{-1}\phi(s_t,a_t)} = \sqrt{2}\|\phi(s_t,a_t)\|_{\Lambda_{t-1}^{-1}}.$$

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 $M_{\tilde{L}} \wedge \tilde{T} \le M_L \wedge \tilde{T} = T \wedge \tilde{T}$ by definition, we thus obtain

$$\begin{aligned}\sum_{l=1}^{\tilde{L}-1} 3\alpha_{M_l} \sum_{t=1+M_l}^{M_{l+1}\wedge\tilde{T}} \|\phi(s_t,a_t)\|_{\Lambda_{M_l}^{-1}} &\le 3\sqrt{2} \max_{t\in[T\wedge\tilde{T}]} \alpha_t \sum_{t=1}^{T\wedge\tilde{T}} \|\phi(s_t,a_t)\|_{\Lambda_{t-1}^{-1}} \\ &\le 3\sqrt{2} \max_{t\in[T\wedge\tilde{T}]} \alpha_t \sqrt{(T\wedge\tilde{T}) \sum_{t=1}^{T\wedge\tilde{T}} \|\phi(s_t,a_t)\|^2_{\Lambda_{t-1}^{-1}}}.\end{aligned}$$

Finally, by Lemma 11 of (Abbasi-Yadkori et al., 2011) and Claim 1, we have

$$\sum_{t=1}^{T\wedge\tilde{T}} \|\phi(s_t,a_t)\|^2_{\Lambda_{t-1}^{-1}} \le 2\log\frac{det(\Lambda_{T\wedge\tilde{T}})}{det(\Lambda_0)} \le 2d\log((T\wedge\tilde{T})+1) \le 2d\log(2(T\wedge\tilde{T})).$$ □

We can now prove the regret bound on $\mathcal{E}\cap\mathcal{F}$. By Lemmas 5, 6, 7, and 8, we know

$$\begin{aligned}\tilde{R}(\tilde{T}) \le 6 \max_{t\in T\wedge\tilde{T}} \alpha_t \sqrt{(T\wedge\tilde{T})d\log(2(T\wedge\tilde{T}))} + 2B_\star\sqrt{(T\wedge\tilde{T})\log(8(T\wedge\tilde{T})/\delta)} \\ + 2(B_\star+1)d\log_2(4B_\star(T\wedge\tilde{T})/c_{min})\end{aligned} \tag{32}$$

Next, recall $B_t + 1 \le 2B_\star + 1 \le 2(B_\star+1)$ by Claim 5. Combined with the assumption that $\kappa_t \le \Psi t^\lambda \log(t+1)$ with $\Psi \ge 9d$ and $\lambda \in [0, \frac{1}{2})$ for any $t \in [T\wedge\tilde{T}]$, we have

$$\alpha_t = (B_t+1)\kappa_t\sqrt{\log(t(B_t+1)\kappa_t/\delta)} \le 2(B_\star+1)\Psi t^\lambda\log(t+1)\sqrt{\log(2(B_\star+1)\Psi t^{1+\lambda}\log(t+1)/\delta)}.$$

Since $(B_\star+1)\Psi t^\gamma \log(t+1)/\delta \ge 9\log 2 \ge 1$, $\log(t+1) \le t$, and $\gamma \le 1/2$, we also have

$$t+1 \le 2t \le 2(B_\star+1)\Psi t^{1+\lambda}\log(t+1)/\delta \le 2(B_\star+1)\Psi t^{5/2}/\delta \le (2(B_\star+1)\Psi t/\delta)^{5/2}.$$

Hence, combining the previous two inequalities, we obtain

$$\begin{aligned}6\alpha_t &\le 6\cdot 2(B_\star+1)\Psi t^\gamma(5/2)\log(2(B_\star+1)\Psi t/\delta)\sqrt{(5/2)\log(2(B_\star+1)\Psi t/\delta)} \\ &= (30\sqrt{5/2})(B_\star+1)\Psi t^\gamma \log^{3/2}(2(B_\star+1)\Psi t/\delta) < 48(B_\star+1)\Psi t^\gamma\log^{3/2}(2(B_\star+1)\Psi t/\delta)\end{aligned}$$

Since the right side is increasing in t, the first summand in (32) can thus be upper bounded by $48\bar{R}(\tilde{T})$, where

$$\bar{R}(\tilde{T}) = (B_\star + 1)\sqrt{d}\Psi(T \wedge \tilde{T})^{\frac{1}{2}+\lambda} \log^2(2(B_\star + 1)\Psi(T \wedge \tilde{T})/(c_{min}\delta)). \tag{33}$$

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

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

$$(T \wedge \tilde{T})c_{min} \leq \sum_{t=1}^{T\wedge\tilde{T}} c(s_t, a_t) = \tilde{R}(\tilde{T}) + \sum_{k=1}^{\tilde{K}} J^\star(s_1^k) \leq 50\bar{R}(\tilde{T}) + KB_\star. \tag{34}$$

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

$$T \wedge \tilde{T} \leq \left(100(B_\star + 1)\sqrt{d}\Psi/c_{min}\right)(T \wedge \tilde{T})^{\frac{1}{2}+\lambda} \log^2\left((2(B_\star + 1)\Psi/(c_{min}\delta))\,(T \wedge \tilde{T})\right).$$

By Claim 9 below, this implies that for some $\iota_1, \iota_2 > 0$ depending only on λ (which, by assumption on λ, means that ι_1, ι_2 are absolute constants), we have

$$T \wedge \tilde{T} \leq T_1 \triangleq \left(\frac{\iota_1(B_\star + 1)\sqrt{d}\Psi}{c_{min}} \log^2\left(\frac{\iota_2(B_\star + 1)d\Psi}{c_{min}\delta}\right)\right)^{\frac{2}{1-2\lambda}}.$$

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

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

$$R(K) = \tilde{O}\left(\left(B_\star^{\frac{3}{2}+\lambda} + B_\star^{\frac{1}{2}+\lambda}\right) d^{\frac{1}{2}}\Psi(K/c_{min})^{\frac{1}{2}+\lambda}\right). \tag{35}$$

If instead only $T \wedge \tilde{T} \leq T_1$ holds, then (33) implies

$$R(K) = \tilde{O}\left((B_\star + 1)\sqrt{d}\Psi T_1^{\frac{1}{2}+\lambda}\right) = \tilde{O}\left((B_\star + 1)^{\frac{2}{1-2\lambda}} d^{\frac{1}{1-2\lambda}} \Psi^{\frac{2}{1-2\lambda}} c_{min}^{-\frac{1+2\lambda}{1-2\lambda}}\right).$$

Finally, bounding $R(K)$ by the max of the cases, then the max by the sum, yields the desired bound.

Remark 14 (Bound for T)**.** *As shown above, for $\tilde{T} \geq T_2$, we have the bound $T \leq T_2$, or (by definition)*

$$T = O\left(\max\left\{\frac{2KB_\star}{c_{min}}, \left(\frac{(B_\star + 1)\sqrt{d}\Psi}{c_{min}} \log^2\left(\frac{(B_\star + 1)d\Psi}{c_{min}\delta}\right)\right)^{\frac{2}{1-2\lambda}}\right\}\right).$$

Remark 15 (Sharpening the tabular case)**.** *When K is large, the bound (35) holds and has $\sqrt{d}\Psi$ dependence on d. We required Ψ (the constant component of κ_t) to be linear in d in order to match the scaling of the error bound ε_t from Appendix B.3. In the tabular case, we can reduce ε_t's dependence to $\sqrt{d}$, (see Remark 17 in Appendix E.1), so we can choose Ψ 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

$$x \leq \frac{16ax_1^c}{(1-c_1)^2}\left(\frac{1-c_1}{4}\log(bx)\right)^2 = \frac{16ax^{c_1}}{(1-c_1)^2}\left(\log\left((bx)^{\frac{1-c_1}{4}}\right)\right)^2 \leq \frac{16ax^{c_1}}{(1-c_1)^2}(bx)^{\frac{1-c_1}{2}} = \frac{16ab^{\frac{1-c_1}{2}}}{(1-c_1)^2}x^{\frac{1+c_1}{2}}.$$

Solving for x, we obtain $x \le (16a/(1-c_1)^2)^{2/(1-c_1)}b$. Plugging back into the log term of the assumed inequality, and since $2 \le 2/(1-c_1)$, we obtain

$$x \le ax^{c_1}\left(\log\left(\left(\frac{16a}{(1-c_1)^2}\right)^{\frac{2}{1-c_1}}b^2\right)\right)^2 \le ax^{c_1}\left(\frac{2}{1-c_1}\log\left(\frac{16ab}{(1-c_1)^2}\right)\right)^2 = c_2 a\log^2(c_3 ab)x^{c_1},$$

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 $\kappa_t \ge 9d$, then on the event $\mathcal{E}$,*

$$\phi(s,a)^\mathsf{T}\hat{G}_t^{n-1}0 - Q^\star(s,a) \le \varepsilon_t\|\phi(s,a)\|_{\Lambda_t^{-1}} \le \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}} \;\forall\,(s,a)\in\mathcal{S}\times\mathcal{A}, n\in\mathbb{N}, t\in[T].$$

Proof. The proof is similar to that of Lemma 4. For $n=1$, $\phi(s,a)^\mathsf{T}\hat{G}_t^{n-1}0 = 0 \le Q^\star(s,a)$, so the result is immediate. Now assume the bound holds for $n\in\mathbb{N}$. Then by the Bellman optimality equation (1), we obtain

$$f_t(s,\hat{G}_t^{n-1}0) = \min_{a\in\mathcal{A}}\left(\phi(s,a)^\mathsf{T}\hat{G}_t^{n-1}0 - \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}}\right) \le \min_{a\in\mathcal{A}} Q^\star(s,a) = J^\star(s). \tag{36}$$

Hence, by Claim 3, $g_t(s,\hat{G}_t^{n-1}0) \le J^\star(s)$ as well. Again by Bellman optimality, this implies

$$\phi(s,a)^\mathsf{T}U_t(\hat{G}_t^{n-1}0) = c(s,a) + \sum_{s'\in\mathcal{S}} g_t(s,\hat{G}_t^{n-1}0)P(s'|s,a) \le c(s,a) + \sum_{s'\in\mathcal{S}} J^\star(s')P(s'|s,a) = Q^\star(s,a).$$

On the other hand, by Corollary 3 and the assumption $\kappa_t \ge 9d$, we have

$$\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - U_t(\hat{G}_t^{n-1}0)) \le \varepsilon_t\|\phi(s,a)\|_{\Lambda_t^{-1}} \le \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}}.$$

Combining the last two inequalities completes the inductive step. □

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 $\kappa_t \ge 9d$, 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 $f_t(s,\hat{G}_t^{n-1}0) \le J^\star(s)$.*

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

The preceding corollary implies the optimism inequality in (9). For the fixed point inequality $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le \alpha_t$, 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 $\kappa_t \ge 18d$, then on the event $\mathcal{E}$,*

$$|\phi(s,a)^\mathsf{T}(\hat{G}_t^{n-1}0 - U_t^{n-1}0)| \le \varepsilon_t\|\phi(s,a)\|_{\Lambda_t^{-1}} + \frac{2(B_t+1)(n-1)\varepsilon_t}{\alpha_t} \;\forall\,(s,a)\in\mathcal{S}\times\mathcal{A}, n\in\mathbb{N}, t\in[T]$$

Proof. First, for any $n\in\mathbb{N}$, we use Corollary 3 and $\kappa_t \ge 18d$ to write

$$\phi(s,a)^\mathsf{T}\hat{G}_t^n 0 \le \phi(s,a)^\mathsf{T}U_t(\hat{G}_t^{n-1}0) + \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}}/2.$$

By Claim 4, we also know $\phi(s,a)^\mathsf{T}U_t(\hat{G}_t^{n-1}0) \le B_t + 1$; combined with the previous inequality, we obtain

$$\phi(s,a)^\mathsf{T}\hat{G}_t^n 0 - \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}} \le B_t + 1 - \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}}/2. \tag{37}$$

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 *un*explored state $s \in \mathcal{S} \setminus \mathcal{S}_t$, (37) implies that for some $a \in \mathcal{A}$,

$$\phi(s,a)^\mathsf{T} \hat{G}_t^n 0 - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \leq B_t + 1 - \frac{\alpha_t}{2} \times \frac{2(B_t+1)}{\alpha_t} = 0.$$

Taking minimum over $a \in \mathcal{A}$ on both sides gives $f_t(s, \hat{G}_t^n 0) \leq 0$, which implies $g_t(s, \hat{G}_t^n 0) = 0$. Again using Claim 4, we similarly obtain that for any $s \in \mathcal{S} \setminus \mathcal{S}_t$ and some $a \in \mathcal{A}$,

$$\phi(s,a)^\mathsf{T} U_t^n 0 - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \leq B_t + 1 - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}} \leq -(B_t+1) < 0,$$

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

$$g_t(s, \hat{G}_t^{n-1} 0) = g_t(s, U_t^{n-1}, 0) = 0 \; \forall \; s \in \mathcal{S} \setminus \mathcal{S}_t, n \in \mathbb{N}.$$

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

$$\begin{aligned}
|\phi(s,a)^\mathsf{T}(U_t(\hat{G}_t^{n-1}0) - U_t(U_t^{n-1}0))| &\leq \sum_{s' \in \mathcal{S}} |g_t(s', \hat{G}_t^{n-1}0) - g_t(s', U_t^{n-1}0)| P(s'|s,a) \\
&\leq \max_{s' \in \mathcal{S}_t} |g_t(s', \hat{G}_t^{n-1}0) - g_t(s', U_t^{n-1}0)| \\
&\leq \max_{(s',a') \in \mathcal{S}_t \times \mathcal{A}} |\phi(s',a')^\mathsf{T}(\hat{G}_t^{n-1}0 - U_t^{n-1}0)|.
\end{aligned}$$

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

$$\begin{aligned}
|\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - U_t^n 0)| &\leq |\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - U_t(\hat{G}_t^{n-1}0))| + |\phi(s,a)^\mathsf{T}(U_t(\hat{G}_t^{n-1}0) - U_t(U_t^{n-1}0))| \\
&\leq \varepsilon_t \|\phi(s,a)\|_{\Lambda_t^{-1}} + \max_{(s',a') \in \mathcal{S}_t \times \mathcal{A}} |\phi(s',a')^\mathsf{T}(\hat{G}_t^{n-1}0 - U_t^{n-1}0)|.
\end{aligned} \tag{38}$$

Thus, taking the maximum over $(s,a) \in \mathcal{S}_t \times \mathcal{A}$ on both sides, by definition of $\mathcal{S}_t$, we have shown

$$\max_{(s,a) \in \mathcal{S}_t \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - U_t^n 0)| \leq \frac{2(B_t+1)\varepsilon_t}{\alpha_t} + \max_{(s,a) \in \mathcal{S}_t \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(\hat{G}_t^{n-1}0 - U_t^{n-1}0)| \; \forall \; n \in \mathbb{N}.$$

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

$$\max_{(s,a) \in \mathcal{S}_t \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(\hat{G}_t^{n-1}0 - U_t^{n-1}0)| \leq \frac{2(B_t+1)\varepsilon_t(n-1)}{\alpha_t} \; \forall \; n \in \mathbb{N}.$$

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

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 $\mathcal{E}$,*

$$\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \leq 5\sqrt{d}\varepsilon_t + \frac{10\sqrt{t}(B_t+1)n\varepsilon_t}{\alpha_t} + 3\sqrt{t} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0)| \; \forall \; n \in \mathbb{N}, t \in [T].$$

Proof. We consider three cases (the last two are corner cases). For the first and most natural case, we assume that $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_2 \leq \|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} / \sqrt{2}$. Then by definition of the induced norm,

$$\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t}^2 \leq 2 \sum_{\tau=1}^t (\phi(s_\tau, a_\tau)^\mathsf{T}(\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0))^2.$$

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

$$(\phi(s,a)^\mathsf{T}(\hat{G}_t^m 0 - U_t^m 0))^2 \le \left(\varepsilon_t \|\phi(s,a)\|_{\Lambda_t^{-1}} + \frac{2(B_t+1)n\varepsilon_t}{\alpha_t}\right)^2 \le 2\varepsilon_t^2 \|\phi(s,a)\|_{\Lambda_t^{-1}}^2 + \frac{8(B_t+1)^2 n^2 \varepsilon_t^2}{\alpha_t^2},$$

which, after another application of Cauchy-Schwarz, gives

$$\begin{aligned}(\phi(s,a)^\mathsf{T}(\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0))^2 &\le 3\sum_{m=n-1}^{n} (\phi(s,a)^\mathsf{T}(\hat{G}_t^m 0 - U_t^m 0))^2 + 3(\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0))^2 \\ &\le 12\varepsilon_t^2 \|\phi(s,a)\|_{\Lambda_t^{-1}}^2 + \frac{48(B_t+1)^2 n^2 \varepsilon_t^2}{\alpha_t^2} + 3 \max_{(s',a') \in \mathcal{S}\times\mathcal{A}} (\phi(s',a')^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0))^2.\end{aligned}$$

By Lemma D.1 of (Jin et al., 2020), we know $\sum_{\tau=1}^t \|\phi(s_\tau, a_\tau)\|_{\Lambda_t^{-1}}^2 \le d$. Combined with previous three inequalities,

$$\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t}^2 \le 24 d\varepsilon_t^2 + \frac{96t(B_t+1)^2 n^2 \varepsilon_t^2}{\alpha_t^2} + 6t \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} (\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0))^2.$$

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

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

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

where we used Claim 1, Corollary 3, and Claim 4. By assumption $\kappa_t \ge 18d$, we also know

$$\varepsilon_t/5 = (B_t+1)d\sqrt{\log(t(B_t+1)\kappa_t \sqrt{\log(t(B_t+1)\kappa_t/\delta)}/\delta)} \ge (B_t+1)d \ge 2. \tag{39}$$

Hence, combining the previous two bounds, we obtain $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \le 24\varepsilon_t/5 \le 5\sqrt{d}\varepsilon_t$.

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

$$\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} < \sqrt{2}\|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_2 = \sqrt{2}\|\hat{G}_t 0\|_2 \le \sqrt{8d} \le 5\sqrt{d}\varepsilon_t. \qquad \square$$

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 $t \in [T]$,*

$$\min_{n \in [\lceil 2dt^{1/6} \rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \le \alpha_t.$$

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

$$\alpha_t \ge \max\{6t^{1/3}\varepsilon_t, (B_t+1)\kappa_t\}. \tag{40}$$

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

$$\frac{\alpha_t}{9\varepsilon_t} = \frac{\kappa_t\sqrt{\log(t(B_t+1)\kappa_t/\delta)}}{45d\sqrt{\log(t\alpha_t/\delta)}} \le \frac{\kappa_t}{40d} = \frac{54t^{1/3}}{45} < 4t^{1/3}.$$

Thus, if we define $N_t = d\sqrt{\alpha_t/\varepsilon_t}/3$, we are guaranteed that $N_t \leq 2dt^{1/6}$, so $\lceil N_t \rceil \in [\lceil 2dt^{1/6} \rceil]$. Combining this result with Lemmas 3 and 10 (we can invoke the latter since $\kappa_t \geq 18d$), we obtain

$$\min_{n \in [\lceil 2dt^{1/6} \rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \leq \|\hat{G}_t^{\lceil N_t \rceil} 0 - \hat{G}_t^{\lceil N_t \rceil - 1} 0\|_{\Lambda_t} \leq 5\sqrt{d}\varepsilon_t + \frac{10\sqrt{t}(B_t+1)\lceil N_t \rceil \varepsilon_t}{\alpha_t} + \frac{3\sqrt{t}(B_t+1)d^2}{\lceil N_t \rceil}. \quad (41)$$

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

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

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

$$\min_{n \in [\lceil 2dt^{1/6} \rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1} 0\|_{\Lambda_t} \leq 5\sqrt{d} \cdot \varepsilon_t + \frac{\alpha_t}{6} \leq 5t^{1/3} \cdot \frac{\alpha_t}{6t^{1/3}} + \frac{\alpha_t}{6} = \alpha_t.$$

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

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

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

$$R(K) = \tilde{O}\left(\left(B_\star^{\frac{11}{6}} + B_\star^{\frac{5}{6}}\right) d^{\frac{3}{2}} (K/c_{min})^{\frac{5}{6}} + (B_\star + 1)^6 d^9 c_{min}^{-5}\right).$$

Finally, Corollaries 4 and 5 imply that on $\mathcal{E}$, 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 $\mathbb{P}(\mathcal{E} \cap \mathcal{F}) \geq 1 - \delta$, 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,*

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

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

$$|\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1} 0)| = |\phi(s,a)^\mathsf{T}\theta| = |c(s,a)| \leq 1 \leq \chi = \chi\rho^{n-1}. \quad (42)$$

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

$$g_t(s, U_t^n 0) = \min_{a \in \mathcal{A}} \Pi_{[0,B_t]} \left(\phi(s,a)^\mathsf{T} U_t^n 0 - \alpha \|\phi(s,a)\|_{\Lambda_t^{-1}}\right).$$

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

$$Q_n(s,a) = \Pi_{[0,B_t]} \left(\phi(s,a)^\mathsf{T} U_t^n 0 - \alpha_t \|\phi(s,a)\|_{\Lambda_t^{-1}}\right),$$

we have $g_t(s, U_t^n 0) = \min_{a\in\mathcal{A}} Q_n(s,a)$. Thus, by definition of U_t, we obtain

$$\phi(s,a)^\mathsf{T}(U_t^{n+1}0 - U_t^n 0) = \sum_{s'\in\mathcal{S}} \left(\min_{a'\in\mathcal{A}} Q_n(s',a') - \min_{a'\in\mathcal{A}} Q_{n-1}(s',a')\right) P(s'|s,a) = (\mathcal{T}Q_n - \mathcal{T}Q_{n-1})(s,a),$$

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

$$\begin{aligned}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega(s)|\phi(s,a)^\mathsf{T}(U_t^{n+1}0 - U_t^n 0)| &\le \rho \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega(s)|(Q_n - Q_{n-1})(s,a)| \\ &\le \rho \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega(s)|\phi(s,a)^\mathsf{T}(U_t^n 0 - U_t^{n-1}0)|,\end{aligned}$$

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

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \omega(s)|\phi(s,a)^\mathsf{T}(U_t^{n+1}0 - U_t^n 0)| \le \rho^n \max_{s\in\mathcal{S}} \omega(s).$$

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

$$|\phi(s,a)^\mathsf{T}(U_t^{n+1}0 - U_t^n 0)| \le \frac{\omega(s)|\phi(s,a)^\mathsf{T}(U_t^{n+1}0 - U_t^n 0)|}{\min_{s'\in\mathcal{S}}\omega(s')} \le \frac{\max_{s'\in\mathcal{S}}\omega(s')\rho^n}{\min_{s'\in\mathcal{S}}\omega(s')} = \chi\rho^n. \qquad \square$$

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 $N_t' \ge \log(3t\chi)/(1-\rho)$, then on the event $\mathcal{E}$, for any $t \in [T]$,*

$$\min_{n\in[\lceil N_t'\rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le \alpha_t.$$

Proof. Because $\chi \ge 1$ and $\rho \in (0,1)$ by definition, we know $N_t' \ge \log 3 \ge 1$, so $1 \le \lceil N_t'\rceil \le 2N_t'$. Combined with Lemmas 10 and 11, we obtain

$$\min_{n\in[\lceil N_t'\rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le \|\hat{G}_t^{\lceil N_t'\rceil}0 - \hat{G}_t^{\lceil N_t'\rceil-1}0\|_{\Lambda_t} \le 5\sqrt{d}\varepsilon_t + \frac{24\sqrt{t}(B_t+1)N_t'\varepsilon_t}{\alpha_t} + 3\sqrt{t}\chi\rho^{N_t'}. \tag{43}$$

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

$$\alpha_t \ge \max\{6t^{1/4}\sqrt{N_t'}\varepsilon_t, (B_t+1)\kappa_t\}. \tag{44}$$

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

$$3\sqrt{t}\chi\rho^{N_t'} \le 3\sqrt{t}\chi e^{-(1-\rho)N_t'} \le 1/\sqrt{t} \le 1 \le \kappa_t(B_t+1)/54 \le \alpha_t/54.$$

For the second term in (43), we again use (44) to obtain

$$\frac{24\sqrt{t}(B_t+1)N_t'\varepsilon_t}{\alpha_t} \le \frac{24\sqrt{t}(B_t+1)N_t'}{6t^{1/4}\sqrt{N_t'}} = 4t^{1/4}\sqrt{N_t'}(B_t+1) \le \frac{4\kappa_t(B_t+1)}{54} \le \frac{4\alpha_t}{54}.$$

Plugging the previous two inequalities into (43), and assuming $\sqrt{d} \le t^{1/4}$, we can use (44) and $N_t' \ge 1$ to obtain

$$\min_{n\in[\lceil N_t\rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le 5\sqrt{d}\cdot\varepsilon_t + \frac{5\alpha_t}{54} \le 5t^{1/4}\cdot\frac{\alpha_t}{6t^{1/4}} + \frac{5\alpha_t}{54} = \frac{50\alpha_t}{54} < \alpha_t.$$

If instead $\sqrt{d} > t^{1/4}$, we simply use Claim 4 and $N_t' \ge 1$ to obtain

$$\min_{n\in[\lceil N_t\rceil]} \|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le \|\hat{G}_t 0\|_{\Lambda_t} \le \sqrt{8td} = \sqrt{8}t^{1/4}t^{1/4}\sqrt{d} \le \sqrt{8}t^{1/4}d \le \kappa_t \le \alpha_t. \qquad \square$$

We now prove Theorem 3. As for Theorem 2, it suffices to prove the guarantees on $\mathcal{E} \cap \mathcal{F}$. Corollaries 4 and 6 establish the OAFP guarantee on $\mathcal{E} \cap \mathcal{F}$ (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

$$\kappa_t = 54dt^{1/4}\sqrt{\log(3t\bar{\chi})/(1-\bar{\rho})} = \Psi t^{\lambda}\sqrt{\log(3t\bar{\chi})/(3\log(3\bar{\chi}))} \le \Psi t^{\gamma}\log(t+1),$$

where the inequality holds because by $\bar{\chi} \ge 1$, we have

$$\frac{\log(3t\bar{\chi})}{3\log(3\bar{\chi})} = \frac{1}{3} + \frac{\log(t)}{3\log(3\bar{\chi})} \le \frac{1+\log(t)}{3} \le \frac{\frac{\log(t+1)}{\log 2} + \log(t)}{3} \le \left(\frac{\frac{1}{\log 2}+1}{3}\right)\log(t+1) \le \log(t+1).$$

Hence, on $\mathcal{E} \cap \mathcal{F}$, we can use the Theorem 1 regret bound with this Ψ and λ to obtain

$$R(K) = \tilde{O}\left(\left(B_\star^{\frac{7}{4}} + B_\star^{\frac{3}{4}}\right) d^{\frac{3}{2}}(K/c_{min})^{\frac{3}{4}} N_t^{1/2} + (B_\star+1)^4 d^6 N_t^2 c_{min}^{-3}\right).$$

Remark 16 (Unknown $\bar{\chi}$ and $\bar{\rho}$)**.** *Suppose $\kappa_t = 54dt^{\frac{1}{4}}\sqrt{N_t}$ as in Theorem 3 but $N_t = t^{2\gamma}$ as in Appendix A.1. Then $\kappa_t \ge 9d$ for any $t \in \mathbb{N}$ and $N_t \ge \log(3t\chi)/(1-\rho)$ as soon as $t \ge (\log(3t\chi)/(1-\rho))^{\frac{1}{2\gamma}}$, so we can use Corollaries 4 and 6 and Lemma 2 to obtain the following: with probability at least $1-\delta/2$, for any $t \ge (\log(3t\chi)/(1-\rho))^{\frac{1}{2\gamma}}$ 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 $\|\cdot\| = \|Q^\mathsf{T}\cdot\|_\infty$, 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*

$$\|Q^\mathsf{T}(\hat{G}_t w_1 - \hat{G}_t w_2)\|_\infty \le e^{-t/(t+1)}\|Q(w_1 - w_2)\|_\infty.$$

Proof. For $(s,a) \in \mathcal{S}\times\mathcal{A}$, 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

$$\Lambda_t = I + \sum_{\tau=1}^t \phi(s_\tau,a_\tau)\phi(s_\tau,a_\tau)^\mathsf{T} = \sum_{i=1}^d q_i q_i^\mathsf{T} + \sum_{i=1}^d d_i q_i q_i^\mathsf{T} = \sum_{i=1}^d (1+d_i) q_i q_i^\mathsf{T} = QDQ^\mathsf{T}.$$

This implies $\Lambda_t^{-1} = QD^{-1}Q^\mathsf{T}$, so for any $i \in [d]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$e_i^\mathsf{T} Q^\mathsf{T}\Lambda_t^{-1}\phi(s,a) = e_i^\mathsf{T} D^{-1}Q^\mathsf{T} q_{i(s,a)} = \frac{e_i^\mathsf{T} Q^\mathsf{T} q_{i(s,a)}}{d_i+1} = \frac{q_i^\mathsf{T} q_{i(s,a)}}{d_i+1} = \frac{\mathbb{1}(i(s,a)=i)}{d_i+1}.$$

Using this identity, we obtain

$$\begin{aligned} e_i^\mathsf{T} Q^\mathsf{T}(\hat{G}_t w_1 - \hat{G}_t w_2) &= \sum_{\tau=1}^t e_i^\mathsf{T} Q^\mathsf{T}\Lambda_t^{-1}\phi(s_\tau,a_\tau)(g_t(s'_\tau,w_1) - g_t(s'_\tau,w_2)) \\ &= \frac{\sum_{\tau\in[t]:i(s_\tau,a_\tau)=i}(g_t(s'_\tau,w_1) - g_t(s'_\tau,w_2))}{d_i+1} \end{aligned}$$

On the other hand, for any $s \in \mathcal{S}$, we know

$$|g_t(s,w_1) - g_t(s,w_2)| \le \max_{a\in\mathcal{A}} |\phi(s,a)^\mathsf{T}(w_1-w_2)| = \max_{a\in\mathcal{A}} |e_{i(s,a)}^\mathsf{T} Q^\mathsf{T}(w_1-w_2)| \le \|Q^\mathsf{T}(w_1-w_2)\|_\infty,$$

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

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

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

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 $t \in [T]$,*

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

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

$$\|Q^\mathsf{T}(\hat{G}_t^n 0 - \hat{G}_t^{n-1}0)\|_\infty \le e^{-\frac{n-1}{t+1}}\|Q^\mathsf{T}\hat{G}_t 0\|_\infty.$$

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

$$\|Q^\mathsf{T}\hat{G}_t 0\|_\infty \le \|Q^\mathsf{T}\hat{G}_t 0\|_2 = \|\hat{G}_t 0\|_2 \le 2\sqrt{d}.$$

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

$$\|w\|_{\Lambda_t} \le \sqrt{(t+1)d}\|w\|_2 = \sqrt{(t+1)d}\|Q^\mathsf{T}w\|_2 \le \sqrt{(t+1)d^2}\|Q^\mathsf{T}w\|_\infty \;\forall\, w \in \mathbb{R}^d.$$

Hence, combining the previous three inequalities, we obtain

$$\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le \sqrt{(t+1)d^2}\|Q^\mathsf{T}(\hat{G}_t^n 0 - \hat{G}_t^{n-1}0)\|_\infty \le \sqrt{(t+1)d^2}e^{-\frac{n-1}{t+1}}\|Q^\mathsf{T}\hat{G}_t 0\|_\infty \le 2\sqrt{(t+1)d^3}e^{-\frac{n-1}{t+1}}.$$

Therefore, if $n \ge 1 + (t+1)\log((t+1)d)/2$, then the previous bound, the assumed choice $\kappa_t = 9d$ in Theorem 4, and Claim 6 imply $\|\hat{G}_t^n 0 - \hat{G}_t^{n-1}0\|_{\Lambda_t} \le 2d \le \kappa_t \le \alpha_t$. □

Similar to the above, on $\mathcal{E}\cap\mathcal{F}$, 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 $\mathbb{P}(\mathcal{B}_{t,b}) \le \delta/(2t(t+1)^2)$.*

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

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

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

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

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 $Y \in \mathbb{R}^{d\times d}_{\succ 0}$ (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

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

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

$$B_t = b \quad \Rightarrow \quad g_t(s,w) = g_{w,\alpha_t^2\Lambda_t^{-1}}(s) \;\forall\, s \in \mathcal{S}, w \in \mathbb{R}^d. \tag{45}$$

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

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

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 $w \in \mathbb{R}^d$,*

$$\|E_t w\|_{\Lambda_t} \le \left\| \sum_{\tau=1}^t \phi_\tau (g_{w,\alpha_t^2\Lambda_t^{-1}}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{w,\alpha_t^2\Lambda_t^{-1}}(s'_\tau)) \right\|_{\Lambda_t^{-1}} + \sqrt{d}(b+1).$$

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

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

The first term can be rewritten as

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

Combining the previous two identities, we obtain

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

Thus, by the triangle inequality, we have

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

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

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 α_t, W_t, and ε_t when $B_t = b$. Thus, $\alpha_t = \alpha_t|B_t$ (and similar for W_t and ε_t). Next, let $\mathcal{X}$ be a $1/(\sqrt{dt})$-net of $[-W_t|b, +W_t|b]^d$ in the ℓ_∞ norm; explicitly, we define

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

Finally, let $\mathcal{Y}$ be a $1/(dt^2)$-net of $\{Y \in \mathbb{R}^{d\times d}_{\succ 0} : |Y(i,j)| \le (\alpha_t|b)^2 \ \forall\, i,j\}$, where we view the matrices as vectors:

$$\mathcal{Y} = \left\{ \left[i_{j_1,j_2}/(dt^2) \right]_{j_1,j_2=1}^d : i_{j_1,j_2} \in \left\{ -\lceil (\alpha_t|b)^2 dt^2 \rceil, \ldots, \lceil (\alpha_t|b)^2 dt^2 \rceil \right\} \ \forall\, j_1, j_2 \right\} \cap \mathbb{R}^{d\times d}_{\succ 0}.$$

Moving forward, we discard the cumbersome $\cdot|b$ notation. However, we emphasize that $\mathcal{X}$ and $\mathcal{Y}$ 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 $\{g_{x,Y} : \mathcal{X} \times \mathcal{Y}\}$.

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

$$\left\|\sum_{\tau=1}^{t} \phi_\tau(g_{w,\alpha_t^2\Lambda_t^{-1}}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{w,\alpha_t^2\Lambda_t^{-1}}(s'_\tau))\right\|_{\Lambda_t^{-1}} \le \left\|\sum_{\tau=1}^{t} \phi_\tau(g_{x,Y}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{x,Y}(s'_\tau))\right\|_{\Lambda_t^{-1}} + 2.$$

Proof. By Claim 1 and a standard spectral norm inequality, we have $|\alpha_t^2\Lambda_t^{-1}(i,j)| \le \alpha_t^2\|\Lambda_t^{-1}\|_2 \le \alpha_t^2 \ \forall\, i,j$. Hence, we can find $Y \in \mathcal{Y}$ such that $\max_{i,j} |Y(i,j) - \alpha_t^2\Lambda_t^{-1}(i,j)| \le 1/(dt^2)$. For such Y and any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we then obtain

$$\begin{aligned}|\phi(s,a)^\intercal(Y - \alpha_t^2\Lambda_t^{-1})\phi(s,a))| &\le \sum_{i,j\in[d]} |\phi_i(s,a)||\phi_j(s,a)||Y(i,j) - \alpha_t^2\Lambda_t^{-1}(i,j)| \\ &\le \frac{\|\phi(s,a)\|_1^2}{dt^2} \le \frac{\|\phi(s,a)\|_2^2}{t^2} \le \frac{1}{t^2},\end{aligned}$$

which implies that

$$\|\phi(s,a)\|_Y \le \sqrt{\alpha_t^2\phi(s,a)^\intercal\Lambda_t^{-1}\phi(s,a) + |\phi(s,a)^\intercal(Y - \alpha_t^2\Lambda_t^{-1})\phi(s,a)|} \le \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}} + 1/t.$$

Hence, by symmetry, we conclude that

$$\left|\|\phi(s,a)\|_Y - \alpha_t\|\phi(s,a)\|_{\Lambda_t^{-1}}\right| \le 1/t. \tag{46}$$

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

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

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

$$\begin{aligned}&\left\|\sum_{\tau=1}^{t} \phi_\tau(g_{w,\alpha_t^2\Lambda_t^{-1}}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{w,\alpha_t^2\Lambda_t^{-1}}(s'_\tau))\right\|_{\Lambda_t^{-1}} \\ &\quad\le \left\|\sum_{\tau=1}^{t} \phi_\tau(g_{x,Y}(s'_\tau) - \mathbb{E}_{s'_\tau} g_{x,Y}(s'_\tau))\right\|_{\Lambda_t^{-1}} + \left\|\sum_{\tau=1}^{t} \phi_\tau(\Delta(s'_\tau) - \mathbb{E}_{s'_\tau}\Delta(s'_\tau))\right\|_{\Lambda_t^{-1}}.\end{aligned}$$

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

$$\left\|\sum_{\tau=1}^{t} \phi_\tau(\Delta(s'_\tau) - \mathbb{E}_{s'_\tau}\Delta(s'_\tau))\right\|_{\Lambda_t^{-1}} \le \sum_{\tau=1}^{t} \|\phi_\tau\|_2|\Delta(s'_\tau) - \mathbb{E}_{s'_\tau}\Delta(s'_\tau)| \le 2. \qquad \square$$

Our final ingredient for proving Claim 11 is the following bound on ε_t.

Claim 15. *Under the assumptions of Lemma 2, if $B_t = b$, then*

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

Proof. We first observe that by assumption $\kappa_t \ge 9d$ and $d \ge 2$, we have

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

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

$$W_t\sqrt{d}t = \alpha_t\sqrt{d}t + (b+1)dt^{3/2} \le \alpha_t\sqrt{\alpha_t/9}t + (\alpha_t/9)t^{3/2} \le 4(\alpha_t t)^{3/2}/9.$$

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

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

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

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

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

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

Combining the previous three inequalities, and since $\delta \ge \delta^{4d^2}$, we obtain

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

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

$$\begin{aligned}
&\sqrt{2b^2 \log\left(\sqrt{\frac{det(\Lambda_t)}{det(\Lambda_0)}}\frac{2t(t+1)^2|\mathcal{X}||\mathcal{Y}|}{\delta}\right)} + \sqrt{d}(b+1) + 2 \\
&\le \sqrt{8b^2d^2\log(\alpha_t t/\delta)} + 2(b+1)d \le (\sqrt{8}+2)(b+1)d\sqrt{\log(\alpha_t t/\delta)} \le 5(b+1)d\sqrt{\log(\alpha_t t/\delta)} = \varepsilon_t. \qquad \square
\end{aligned}$$

Remark 17 (Sharpening the tabular case)**.** *In the tabular case, Λ_t^{-1} is diagonal, so we can replace $\mathcal{Y}$ with $\mathcal{Y}' = \{Y \in \mathcal{Y} : Y \text{ is diagonal}\}$. Since $|\mathcal{Y}'|$ is exponential in d (instead of d^2), we can define ε_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

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

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 Theorem 1 of (Abbasi-Yadkori et al., 2011), we have $\mathbb{P}(\mathcal{C}_{x,Y,b}) \le \delta/(2t(t+1)^2|\mathcal{X}||\mathcal{Y}|)$. Combined with Claims 13, 14, and 15 and the union bound,

$$\mathbb{P}(\mathcal{B}_{t,b}) \le \mathbb{P}(\cup_{(x,Y)\in\mathcal{X}\times\mathcal{Y}}\mathcal{C}_{x,Y} \cap \{B_t = b\}) \le \mathbb{P}(\cup_{(x,Y)\in\mathcal{X}\times\mathcal{Y}}\mathcal{C}_{x,Y}) \le \sum_{(x,Y)\in\mathcal{X}\times\mathcal{Y}} \mathbb{P}(\mathcal{C}_{x,Y}) \le \frac{\delta}{2t(t+1)^2}. \qquad \square$$

E.2. Proof of Proposition 1

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

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

Hence, by the Bellman optimality equations (1),

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

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

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