The return of ϵ-greedy: sublinear regret for model-free linear quadratic control

Yasin Abbasi-Yadkori
Adobe Research

Nevena Lazić
Google Brain

Csaba Szepesvári
Deepmind

July 13, 2018

Abstract

Model-free approaches for reinforcement learning (RL) and continuous control find policies based only on past states and rewards, without fitting a model of the system dynamics. They are appealing as they are general purpose and easy to implement; however, they also come with fewer theoretical guarantees than model-based RL. In this work, we present a new model-free algorithm for controlling linear quadratic (LQ) systems, and show that its regret scales as $O(T^{3/4})$. The algorithm is based on a reduction of control of Markov decision processes to an expert prediction problem. In practice, it corresponds to a variant of policy iteration with ϵ-greedy exploration, where the policy is greedy with respect to the average of all previous value functions. This is the first model-free algorithm for adaptive control of LQ systems that provably achieves sublinear regret and has a polynomial computation cost. Empirically, our algorithm dramatically outperforms standard policy iteration, but performs worse than a model-based approach.

1 Introduction

Reinforcement learning (RL) algorithms have recently shown impressive performance in many challenging decision making problems, including game playing and various robotic tasks. *Model-based* RL approaches estimate a model of the transition dynamics and rely on the model to plan future actions using approximate dynamic programming. *Model-free* approaches aim to find an optimal policy without explicitly modeling the system transitions; they either estimate state-action value functions or directly optimize a parameterized policy based only on interactions with the environment. Model-free RL is appealing for a number of reasons: 1) it is an "end-to-end" approach, directly optimizing the cost function of interest, 2) it avoids the difficulty of modeling and robust planning, and 3) it is easy to implement. However, model-free algorithms also come with fewer theoretical guarantees than their model-based counterparts, which presents a considerable obstacle in deploying them in real-world physical systems with safety concerns and the potential for expensive failures.

In this work, we propose a model-free algorithm for controlling linear quadratic (LQ) systems with theoretical guarantees. LQ control is one of the most studied problems in control theory (Bertsekas, 1995), and it is also widely used in practice. Its simple formulation and tractability given known dynamics make it an appealing benchmark for studying RL algorithms with continuous states and actions. A common way to analyze the performance of sequential decision making algorithms is to use the notion of regret - the difference between the total cost incurred and the cost of the best policy in hindsight (Cesa-Bianchi and Lugosi, 2006, Hazan, 2016, Shalev-Shwartz, 2012). We show that our model-free LQ control algorithm enjoys a $O(T^{3/4})$ regret bound. Note that existing regret bounds for LQ systems are only available for model-based approaches.

Our algorithm is a modified version of policy iteration with ϵ-greedy exploration. Standard policy iteration estimates the value of the current policy in each round, and sets the next policy to be greedy with respect to the most recent value function. By contrast, we use a policy that is greedy with respect to the *average of all past value functions* in each round. The form of this update is a direct consequence of a reduction of the control of Markov decision processes (MDPs) to expert prediction problems (Even-Dar et al., 2009). In this reduction, each prediction loss corresponds to the value function of the most recent policy, and the next policy is the output of the expert algorithm. The structure of the LQ

control problem allows for an easy implementation of this idea: since the LQ value function is quadratic, the average of all previous value functions is also quadratic.

One major challenge in this approach is reliable estimation of average-cost LQ value functions from data. Most of the existing finite-sample estimation approaches either consider bounded functions or discounted problems, and are not applicable in our setting. Interestingly, as we will show, a carefully tuned ϵ-greedy exploration strategy (Sutton and Barto, 1998) suffices in the LQ case. Another major challenge is showing boundedness of the value functions in our iterative scheme, especially considering that the state and action spaces are unbounded. We are able to do so using the particular problem structure and the boundedness of the value estimation error.

Our main contribution is a model-free algorithm for adaptive control of linear quadratic systems with strong theoretical guarantees. This is the first such algorithm that provably achieves sublinear regret and has a polynomial computation cost. The only other computationally efficient algorithm with sublinear regret is the model-based approach of Dean et al. (2018) (which appeared in parallel to this work). Previous works have either been restricted to one-dimensional LQ problems (Abeille and Lazaric, 2017), or considered the problem in a Bayesian setting (Ouyang et al., 2017). In addition to theoretical guarantees, we demonstrate empirically that our algorithm leads to significantly more stable policies than standard policy iteration.

1.1 Related work

Model-based adaptive control of linear quadratic systems has been studied extensively in control literature. *Open-loop* strategies identify the system in a dedicated exploration phase. Classical asymptotic results in linear system identification are covered in (Ljung and Söderström, 1983); an overview of frequency-domain system identification methods is available in (Chen and Gu, 2000), while identification of auto-regressive time series models is covered in (Box et al., 2015). Non-asymptotic results are limited, and existing studies often require additional stability assumptions on the system (Helmicki et al., 1991, Hardt et al., 2016, Tu et al., 2017). Dean et al. (2017) relate the finite-sample identification error to the smallest eigenvalue of the controllability Gramian.

Closed-loop model-based strategies update the model online while trying to control the system, and are more akin to standard RL. Fiechter (1997) and Szita (2007) study model-based algorithms with PAC-bound guarantees for discounted LQ problems. Asymptotically efficient algorithms are shown in (Lai and Wei, 1982, 1987, Chen and Guo, 1987, Campi and Kumar, 1998, Bittanti and Campi, 2006). Multiple approaches (Campi and Kumar, 1998, Bittanti and Campi, 2006, Abbasi-Yadkori and Szepesvári, 2011, Ibrahimi et al., 2012) have relied on the *optimism in the face of uncertainty* principle. Abbasi-Yadkori and Szepesvári (2011) show an $O(\sqrt{T})$ finite-time regret bound for an optimistic algorithm that selects the dynamics with the lowest attainable cost from a confidence set; however this strategy is somewhat impractical as finding lowest-cost dynamics is computationally intractable. Abbasi-Yadkori and Szepesvári (2015), Abeille and Lazaric (2017), Ouyang et al. (2017) demonstrate similar regret bounds in the Bayesian and one-dimensional settings using Thompson sampling. Dean et al. (2018) show an $O(T^{2/3})$ regret bound using robust control synthesis.

Fewer theoretical results exist for model-free LQ control. The LQ value function can be expressed as a linear function of known features, and is hence amenable to least squares estimation methods. Least squares temporal difference (LSTD) learning has been extensively studied in reinforcement learning, with asymptotic convergence shown by Tsitsiklis and Van Roy (1997), Tsitsiklis and Roy (1999), Yu and Bertsekas (2009), and finite-sample analyses given in Antos et al. (2008), Farahmand et al. (2016), Lazaric et al. (2012), Liu et al. (2015, 2012). Most of these methods assume bounded features and rewards, and hence do not apply to the LQ setting. For LQ control, Bradtke et al. (1994) show asymptotic convergence of Q-learning to optimum under persistently exciting inputs, and Tu and Recht (2017) analyze the finite sample complexity of LSTD for discounted LQ problems. Here we adapt the work of Pires and Szepesvári (2012) and Tu and Recht (2017) to analyze the finite sample estimation error in the average-cost setting. Among other model-free LQ methods, Fazel et al. (2018) analyze policy gradient for deterministic dynamics, and Arora et al. (2018) formulate optimal control as a convex program by relying on a spectral filtering technique for representing linear dynamical systems in a linear basis.

Relevant model-free methods for finite state-action MDPs include the Delayed Q-learning algorithm of Strehl et al. (2006), which is based on the optimism principle and has a PAC bound in the discounted setting. Osband et al. (2017) propose exploration by randomizing value function parameters, an algorithm that is applicable to large state problems.

However the performance guarantees are only shown for finite-state problems.

Our approach is based on a reduction of the MDP control to an expert prediction problem. The reduction was first proposed by Even-Dar et al. (2009) for the online control of finite-state MDPs with changing cost functions. This approach has since been extended to finite MDPs with known dynamics and bandit feedback (Neu et al., 2014), LQ tracking with known dynamics (Abbasi-Yadkori et al., 2014), and linearly solvable MDPs (Neu and Gómez, 2017).

2 Preliminaries

We model the interaction between the agent (i.e. the learning algorithm) and the environment as a Markov decision process (MDP). An MDP is a tuple $\langle \mathcal{X}, \mathcal{A}, c, P \rangle$, where $\mathcal{X} \subset \mathbb{R}^n$ is the state space, $\mathcal{A} \subset \mathbb{R}^d$ is the action space, $c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is a cost function, and $P : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$ is the transition probability distribution that maps each state-action pair to a distribution over states $\Delta_{\mathcal{X}}$. At each discrete time step $t \in \mathbb{N}$, the agent receives the state of the environment $x_t \in \mathcal{X}$, chooses an action $a_t \in \mathcal{A}$ based on x_t and past observations, and suffers a cost $c(x_t, a_t)$. The environment then transitions to the next state according to $x_{t+1} \sim P(x_t, a_t)$. We assume that the agent does not know P, but does know c. A policy is a mapping $\pi : \mathcal{X} \to \mathcal{A}$ from the current state to an action, or a distribution over actions. Following a policy means that in any round upon receiving state x, the action a is chosen according to $\pi(x)$. Let $\mu_\pi(x)$ be the stationary state distribution under policy π, and let $\lambda_\pi = \mathbf{E}_\mu(c(x, \pi(x))$ be the average cost of policy π. Let x_t^π be the state at time step t when policy π is followed. The objective of the agent is to have small regret, defined as

$$\text{Regret}_T = \sum_{t=1}^{T} c(x_t, a_t) - \min_\pi \sum_{t=1}^{T} c(x_t^\pi, \pi(x_t^\pi)) \; .$$

2.1 Linear quadratic control

In a linear quadratic control problem, the state transition dynamics and the cost function are given by

$$x_{t+1} = Ax_t + Ba_t + w_{t+1} \, , \quad c(x_t, a_t) = x_t^\top M x_t + a_t^\top N a_t \; .$$

The state space is $\mathcal{X} = \mathbb{R}^n$ and the action space is $\mathcal{A} = \mathbb{R}^d$. We assume the initial state is zero, $x_1 = 0$. A and B are unknown dynamics matrices of appropriate dimensions, assumed to be controllable[1]. M and N are known positive definite cost matrices. Vectors w_{t+1} correspond to system noise; similarly to previous work, we assume that w_t are drawn i.i.d. from a known Gaussian distribution $\mathcal{N}(0, W)$.

In the infinite horizon setting, it is well-known that the optimal policy $\pi_*(x)$ corresponding to the lowest average cost λ_π is given by constant linear state feedback, $\pi_*(x) = -K_* x$. When following any linear feedback policy $\pi(x) = -Kx$, the system states evolve as $x_{t+1} = (A - BK)x_t + w_{t+1}$. A linear policy is called *stable* if $\rho(A - BK) < 1$, where $\rho(\cdot)$ denotes the spectral radius of a matrix. It is well-known that the value function V_π and state-action value function Q_π of any stable linear policy $\pi(x) = -Kx$ are quadratic functions (see Appendix A for a proof):

$$Q_\pi(x, a) = \begin{pmatrix} x^\top & a^\top \end{pmatrix} G_\pi \begin{pmatrix} x \\ a \end{pmatrix} \, , \quad V_\pi(x) = x^\top \begin{pmatrix} I & -K^\top \end{pmatrix} G_\pi \begin{pmatrix} I \\ -K \end{pmatrix} x = x^\top H_\pi x \, .$$

We call $H_\pi \succ 0$ the value matrix of policy π. Matrix $G_\pi \succ 0$ is the unique solution of the equation

$$G_\pi = \begin{pmatrix} A^\top \\ B^\top \end{pmatrix} \begin{pmatrix} I & -K^\top \end{pmatrix} G_\pi \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} + \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} \, .$$

The greedy policy with respect to Q_π is given by $\operatorname{argmin}_a Q_\pi(x, a) = -G_{\pi,22}^{-1} G_{\pi,21} x = -K_\pi x$. The average expected cost of following a linear policy is $\lambda_\pi = \operatorname{tr}(H_\pi W)$. The stationary state distribution of a stable linear policy is $\mu_\pi(x) = \mathcal{N}(x|0, \Sigma)$, where Σ is the unique solution of the Lyapunov equation $\Sigma = (A - BK)\Sigma(A - BK)^\top + W$.

[1] The linear system is controllable if the matrix $(B \; AB \; \cdots \; A^{n-1}B)$ has full column rank.

Input: Stable policy K_1, number of phases S, value estimation length τ, number of random actions τ', sampling period τ'', initial state x_0

$\pi_1(x) = -K_1 x$
for $i := 1, 2, \ldots, I$ **do**
  $\widehat{Q}_i,\ x_i = \textsc{EstValue}(\pi_i, \tau, \tau', \tau'', x_{i-1})$
  $\pi_{i+1}(x) = \operatorname{argmin}_a \sum_{j=1}^{i} \widehat{Q}_j(x, a) = -K_{i+1} x$
end for

Figure 1: MFLQ: a model-free algorithm for linear quadratic control

3 Model-free control of linear quadratic systems

3.1 MFLQ algorithm

Our model-free linear quadratic control algorithm (MFLQ) is shown in Figure 1. At a high level, the algorithm is a variant of policy iteration with a forced-exploration strategy. We assume that an initial stable suboptimal policy $\pi_1(x) = -K_1 x$ is given. During phase i, we first execute policy π_i for a sufficient number of rounds to estimate V_i. We then execute the policy and take random actions at appropriate intervals to estimate Q_i. The random actions are necessary as otherwise actions may lie in a subspace of $\mathcal{A}$, as a consequence of the linear policy. Estimation details are given Figure 2 and Section 3.2. Having estimated the value function Q_i, we produce the next policy π_{i+1} by computing the greedy policy with respect to the average of all previous value estimates. This step is different than standard policy iteration, which only considers the most recent value estimate. The difference is a result of our reliance on the FOLLOW-THE-LEADER expert algorithm to produce the sequence of policies. MFLQ runs for $S = O(T^{1/4})$ phases of length $\tau = O(T^{3/4})$, and we take $\tau'' = O(T^{1/4})$ random actions; these numbers were chosen to minimize the regret.

The main result of this section is the following theorem.

Theorem 3.1. *For an appropriate constant C, the regret of the MFLQ algorithm is bounded as*

$$\textit{Regret}_T \le C T^{3/4} \log T \ .$$

In the remainder of this section, we prove the above theorem. First, we show a regret decomposition. The regret of an algorithm with respect to a fixed policy π can be written as

$$\text{Regret}_T = \sum_{t=1}^{T} c(x_t, a_t) - \sum_{t=1}^{T} c(x_t^\pi, \pi(x_t^\pi)) = \alpha_T + \beta_T + \gamma_T \ ,$$

$$\text{where } \ \alpha_T = \sum_{t=1}^{T} (c(x_t, a_t) - \lambda_{\pi_t}) \ , \ \ \beta_T = \sum_{t=1}^{T} (\lambda_{\pi_t} - \lambda_\pi) \ , \ \ \gamma_T = \sum_{t=1}^{T} (\lambda_\pi - c(x_t^\pi, \pi(x_t^\pi))) \ .$$

The terms α_T and γ_T represent the difference between instantaneous and average cost of a policy, and can be bounded using mixing properties of policies and MDPs. To bound β_T, first note that using Bellman equations we can show that (see, e.g. Even-Dar et al. (2009))

$$\lambda_{\pi_t} - \lambda_\pi = \mathbf{E}_{x \sim \mu_\pi} (Q_{\pi_t}(x, \pi_t(x)) - Q_{\pi_t}(x, \pi(x))) \ .$$

Let $\widehat{Q}_i$ be an estimate of Q_i, computed from data at the end of phase i. We can write

$$\begin{aligned} Q_i(x, \pi_t(x)) - Q_i(x, \pi(x)) = \widehat{Q}_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi(x)) \\ + Q_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi_i(x)) \end{aligned}$$

Input: Policy π, estimation lengths τ, τ', τ'', initial state x_0

Execute policy π for τ rounds and compute $\widehat{V}_\pi$ using Equation (1)
$\mathcal{Z} = \{\}$
for $j := 1, 2, \ldots, \tau'$ **do**
  Execute policy π for τ'' rounds and let x be the final state
  Sample $a \sim \mathcal{N}(0, I_d)$, observe the next state x', add (x, a, x') to $\mathcal{Z}$
end for
Compute $\widehat{Q}_\pi$ as in Equation (5) using data $\mathcal{Z}$ and $\widehat{V}_\pi$
Return $\widehat{Q}_\pi$

Figure 2: ESTVALUE: estimation of the state-action value function of a linear policy

$$+ \widehat{Q}_i(x, \pi(x)) - Q_i(x, \pi(x)) \; .$$

Since we feed the expert in state x with $\widehat{Q}_i(x, .)$ at the end of each phase, the first term on the RHS can be bounded by the regret bound of the expert algorithm. The remaining terms correspond to the estimation errors. We will show that in the case of linear quadratic control, the value function parameters can be estimated with small error of order $O(1/\sqrt{\tau})$. Given sufficiently small estimation errors, we show that all policies remain stable, and hence states, actions, and value functions remain bounded. Given the boundedness and the quadratic form of the value functions, we use existing regret bounds for the FTL strategy to finish the proof.

3.2 Finite-time analysis of LSTD

In this section, we study least squares temporal difference (LSTD) estimates of the value matrix H_π. We will assume the policies generated by the algorithm have small mixing coefficient (see Appendix B.1 for definition). This condition is shown to hold later in Lemma 3.4 in Section 3.4. Further, we assume that states and actions remain bounded. This condition is also shown to hold in Lemma 3.5 in Section 3.4. The boundedness of states and actions and fast mixing of policies are needed to bound certain tail functions and also the α_T term in regret decomposition.

In order to simplify notation, we will drop π subscripts in this section. In steady-state, we have that

$$V(x_t) = c(x_t, a_t) + \mathbf{E}(V(x_{t+1})|x_t, \pi(x_t)) - \lambda$$
$$x_t^\top H x_t = c(x_t, \pi(x_t)) + \mathbf{E}(x_{t+1}^\top H x_{t+1}|x_t, \pi(x_t)) - \lambda$$

where the expectation is with respect to the noise w_t. Let $\text{VEC}(A)$ denote the vectorized version of a symmetric matrix A, such that $\text{VEC}(A_1)^\top \text{VEC}(A_2) = \text{tr}(A_1 A_2)$, and let $\phi(x) = \text{VEC}(xx^\top)$. We will use the shorthand notation $\phi_t = \phi(x_t)$ and $c_t = c(x_t, a_t)$. We have that

$$\mathbf{E}(\phi_t - \phi_{t+1}|x_t, \pi)^\top \text{VEC}(H) = c_t - \lambda$$

By multiplying both sides with ϕ_t and taking expectations with respect to the steady state distribution,

$$\mathbf{E}(\phi_t(\phi_t - \phi_{t+1})^\top)\text{VEC}(H) = \mathbf{E}(\phi_t(c_t - \lambda)) \; .$$

The average-cost LSTD estimator of H is given by (see e.g. Tsitsiklis and Roy (1999), Yu and Bertsekas (2009)):

$$\text{VEC}(\widehat{H}_\tau) = \left(\sum_{t=1}^{\tau} \phi_t(\phi_t - \phi_{t+1})^\top \right)^\dagger \left(\sum_{t=1}^{\tau} (c_t - \widehat{\lambda}_t)\phi_t \right) \; , \tag{1}$$

where $(.)^\dagger$ denotes the pseudo-inverse and $\widehat{\lambda}_t$ is an estimate of the average cost at step t. We set all $\widehat{\lambda}_t$ to the average of the costs suffered during the episode. Given that the true value matrix is positive definite with $H \succ M$, we additionally project this estimate onto the constraint $H \succeq M$.

We use results of Pires and Szepesvári (2012) to bound the estimation error. Consider the equation $F\theta = b$ with matrix F and vectors θ and b of appropriate dimensions, and assume $F + F^\top \succ 2cI$ for some $c > 0$. Let $(\widehat{F}, \widehat{b})$ be noisy estimates of (F, b). Let $z_{F,\delta} > 0$ and $z_{b,\delta} > 0$ be tail functions such that for any $\delta \in (0,1)$, with probability at least $1-\delta$, $\left\|F - \widehat{F}\right\|_2 \leq z_{F,\delta}$ and $\left\|b - \widehat{b}\right\|_2 \leq z_{b,\delta}$. Let $\textsc{mat}(\theta)$ denote the symmetric matrix version of a vector θ. Let

$$\widehat{\theta}_\gamma \in \operatorname*{argmin}_{\theta:\textsc{mat}(\theta)\succeq M} \left\{ \left\| \widehat{F}\theta - \widehat{b} \right\|_2 + \gamma \left\|\theta\right\| \right\} . \tag{2}$$

Theorem 3.4 of Pires and Szepesvári (2012) shows that for any $\delta \in (0,1)$ and with $\gamma = z_{F,\delta}$, with probability at least $1-\delta$,

$$\left\| F(\widehat{\theta}_\gamma - \theta) \right\| \leq 2 z_{F,\delta} \left\|\theta\right\| + 2 z_{b,\delta} . \tag{3}$$

From $F + F^\top \succ 2cI$, we have that for any vector v, $v^\top F v = v^\top F^\top v > c \left\|v\right\|^2$. Hence

$$\forall v , \qquad \left\|Fv\right\| > c \left\|v\right\| , \qquad \left\|F^\top v\right\| > c \left\|v\right\| . \tag{4}$$

Let λ be an eigenvalue of $F^\top F$ with corresponding eigenvector u. Using (4), we have that $\left\|F^\top F u\right\| > c \left\|Fu\right\| > c^2 \left\|u\right\|$, and thus $\lambda_{min}(F^\top F) > c^2$. By (3),

$$\left\| \widehat{\theta}_\gamma - \theta \right\| \leq \frac{\left\|F\right\|}{c^2} \left(2 z_{F,\delta} \left\|\theta\right\| + 2 z_{b,\delta}\right) .$$

Consider $F = \mathbf{E}(\phi_t(\phi_t - \phi_{t+1})^\top)$ and $b = \mathbf{E}(\phi_t(c_t - \lambda))$. Fast mixing of stable linear systems (see Tu and Recht (2017) and Lemma B.1) implies that tail functions $z_{F,\delta}$ and $z_{b,\delta}$ scale as $O(1/\sqrt{\tau})$ (Yu, 1994), resulting in an error bound of the form

$$\left\| \widehat{H}_\tau - H \right\|_2^2 = O\left(\frac{1}{\tau}\right) .$$

In order to use the above argument, we need to lower-bound the smallest eigenvalue of $F + F^\top$:

$$\mathbf{E}(\phi_t(\phi_t - \phi_{t+1})^\top) + \mathbf{E}(\phi_t(\phi_t - \phi_{t+1})^\top)^\top = \mathbf{E}\left((\phi(x_{t+1}) - \phi(x_t))(\phi(x_{t+1}) - \phi(x_t))^\top\right) .$$

Together with fast mixing of stable linear dynamical systems, this lower bound gives the desired bound on the estimation error.

Lemma 3.2. *Let $(x_k)_{k=1}^\tau$ be a sequence of states generated under policy π. We have that*

$$\lambda_{min}\left(\sum_{k=1}^{\tau} (\phi(x_{k+1}) - \phi(x_k))(\phi(x_{k+1}) - \phi(x_k))^\top\right) \geq \Omega(\tau) .$$

The proof adapts the results of Tu and Recht (2017) and is given in Appendix B.

3.3 Finite-time analysis of LSTD-Q

One simple way to estimate the state-action value matrix G is to use an LSTD-Q algorithm that parallels the LSTD algorithm of the previous section. However analyzing this estimate turns out to be challenging due to the structure of the covariance matrix of state-action pairs under a linear policy. We instead use the procedure described in Figure 2 which relies on $\widehat{H}$ and randomly sampled actions.

Let $z_t^\top = (x_t^\top, a_t^\top)$ and $\psi_t = \textsc{vec}(z_t z_t^\top)$. Suppose that x is sampled from a distribution sufficiently close to μ, and that a is sampled from $\mathcal{N}(0, I_d)$. Then

$$\begin{aligned} Q(z_t) = \operatorname{tr}(z_t z_t^\top G) &= c_t - \lambda + \mathbf{E}(V(x_{t+1})|z_t) \\ \psi_t^\top \textsc{vec}(G) &= c_t - \lambda + \mathbf{E}(V(x_{t+1}|z_t)) - V(x_{t+1}) + V(x_{t+1}) . \end{aligned}$$

Note that $\mathbf{E}\left[\mathbf{E}(V(x_{t+1})|z_t) - V(x_{t+1})|z_t\right] = 0$, so we can use standard arguments to analyze the least-squares estimate of G. To gather appropriate data, we iteratively (1) execute the policy π for τ'' iterations in order to have states distributed sufficiently close to μ, and then (2) sample an action according to $\mathcal{N}(0, I_d)$. We estimate G from τ' such tuples $(x_k, a_k, x_{k+1})_{k=1}^{\tau'}$ and the estimate $\widehat{V}(x)$:

$$\text{VEC}(\widehat{G}_{\tau'}) = \left(\sum_{k=1}^{\tau'} \psi_k \psi_k^\top\right)^{-1} \left(\sum_{k=1}^{\tau'} (c(x_k, a_k) - \widehat{\lambda}_k + \widehat{V}(x_{k+1}))\psi_k\right) . \tag{5}$$

Additionally, we project the above estimate onto the known constraint $\widehat{G}_{\tau'} \succeq \left(\begin{smallmatrix} M & 0 \\ 0 & N \end{smallmatrix}\right)$. If $x_k \sim \mu$ and $\mathbf{E}(\psi\psi^\top) \succ \sigma' I$, then we can use the same techniques as before to show a high probability error bound of the form $\|\widehat{G}_{\tau'} - G\|_2^2 = O\left(\frac{1}{\sigma'^2 \tau'}\right)$. Write $x = \Sigma^{1/2} g$ and $a = g'$, where $g \sim \mathcal{N}(0, I)$ and $g' \sim \mathcal{N}(0, I)$ are independent. Thus, $z = Fh$, where

$$F = \begin{pmatrix} \Sigma^{1/2} & 0 \\ 0 & I \end{pmatrix} , \qquad h = \begin{pmatrix} g \\ g' \end{pmatrix} .$$

Notice that h is a multivariate standard normal random vector. By following the same steps as Tu and Recht (2017) and Appendix B, we can show that $\mathbf{E}(\psi\psi^\top) \succ \sigma' I$ for a positive σ'. The result is summarized next.

Lemma 3.3. *Let $\pi(x) = -Kx$ be a stable linear policy. Let $(x_k, a_k, x_{k+1})_{k=1}^{\tau'}$ be τ' tuples such that each is generated by running policy π for $\tau'' = O(T^{1/4})$ rounds and then taking action $a_k \sim \mathcal{N}(0, I_d)$. Assume the ℓ^2-norm of the value matrix of policy π, and states and actions generated during the phase are bounded by C_H, C_X and C_A, respectively. Let $\delta_1 \in (0, 1)$. There is a least-squares estimate $\widehat{G}_\pi$ based on this data such that with probability at least $1 - \delta_1$, for some appropriate constant C_G,*

$$\left\|G_\pi - \widehat{G}_\pi\right\|_\infty \le C_H C_G \sqrt{\log(C_X C_A/\delta_1)/\tau'} .$$

By a union bound, with probability at least $1 - \delta_1$, for all i,

$$\left\|G_i - \widehat{G}_i\right\|_\infty \le \varepsilon_1 \stackrel{\text{def}}{=} C_H C_G \sqrt{\log(T C_X C_A/\delta_1)/\tau'} . \tag{6}$$

3.4 Analysis of the MFLQ algorithm

In this section, we first show that given sufficiently small estimation error, all policies produced by the MFLQ algorithm remain stable. Consequently the value matrices, states, and actions remain bounded. We then bound the terms α_T, β_T, and γ_T to show the main result. For simplicity, we will assume that $M \succ I$ and $N \succ I$ for the rest of this section; we can always rescale M and N so that this holds true without loss of generality.

By assumption, K_1 is bounded and stable. By the arguments in Section 3.2 and 3.3, the estimation error in the first phase can be made small for sufficiently long phases. In particular, we assume that estimation error in the first phase is bounded by ε_1 as in Lemma 3.3 and that ε_1 satisfies

$$\varepsilon_1 < \left(12 C_1 (\sqrt{n} + C_K \sqrt{d})^2 T^{1/4}\right)^{-1} . \tag{7}$$

Here $C_1 > 1$ is an upper bound on $\|H_1\|$, and $C_K = 2(3C_1 \|B\| \|A\| + 1)$. Note that the estimation error in the first phase is $O(1/\sqrt{\tau}) = O(1/T^{3/8})$, so the error factor $T^{-1/4}$ is valid. We also assume that T is large enough so that the following holds:

$$T^{3/4} \ge \ln(3C_1)/\ln(1 + (6C_1)^{-1}) . \tag{8}$$

We prove the following two lemmas in Appendix C.1 and C.2.

Lemma 3.4. *Let $\{K_i\}_{i=2}^S$ be the sequence of policies produced by the MFLQ algorithm. For all $i \in [S]$, $\|H_i\| \le C_i < C_H := 3C_1$, K_i is stable, $\|K_i\| \le C_K$, and for all $k \in \mathbb{N}$,*

$$\left\|(A - BK_i)^k\right\| \le \sqrt{C_{i-1}}(1 - (6C_1)^{-1})^{k/2} \le \sqrt{C_H}(1 - (2C_H)^{-1})^{k/2} .$$

Given the above lemma, Gelfand's formula, and Lemma B.1, we conclude that all policies have small mixing coefficients.

Lemma 3.5. *Let $\delta_2 \in (0,1)$. For any $t = 1, 2, \ldots, T$ with probability at least $1 - \delta_2$,*

$$\|x_t\| \le C_X := \frac{\sqrt{2n \log(Tn/\delta_2)}}{1 - \sqrt{1 - (2C_H)^{-2}}}, \qquad \|a_t\| \le C_A := \sqrt{C_H} C_X \ .$$

By Lemma 3.3, Inequality (6), and Lemma 3.5, if we choose $\delta_1 = \delta_2 = \delta/2$ we get that with probability at least $1 - \delta$, $\|x_t\| \le C_X$, $\|a_t\| \le C_A$, and $\|G_i - \widehat{G}_i\|_\infty \le \varepsilon_1$. We use $\mathcal{E}$ to denote the event that the above inequalities hold. Finally we bound the terms α_T, β_T, and γ_T.

Lemma 3.6. *Under event $\mathcal{E}$, for appropriate constants C', D', and D'', $\beta_T \le C' T^{3/4} \log(T/\delta)$, $\alpha_T \le D' T^{3/4}$ and $\gamma_T \le D''$.*

The proof is given in Appendix C.3. For β_T, the proof relies on the FTL regret bound of (Cesa-Bianchi and Lugosi, 2006), along with the fact that value functions are quadratic and bounded. The bound on γ_T is a consequence of the fact that states and actions remain bounded, and the state distribution under policy π_* converges to its stationary distribution exponentially fast. The bound on α_T is due to the fact that we have $S = O(T^{1/4})$ policy switches and in each policy execution, we have $\tau' = O(T^{1/2})$ random actions. The main result is a consequence of Lemma 3.6.

4 Experiments

We evaluate our algorithm on two LQ problem instances: the system studied in Dean et al. (2017) and Tu and Recht (2017), and the power system studied in Lewis et al. (2012), Example 11.5-1, with noise $W = 0.1I$. We start at an all-zero initial state x_0 and use an initial stable policy K_1 obtained by solving the control problem with a modified cost $M' = 200M$. We compare MFLQ to the following:

- Least squares policy iteration (LSPI) where the policy π_i in phase i is greedy with respect to the most recent value function estimate $\widehat{Q}_{i-1}$.
- A version RLSVI Osband et al. (2017) where we randomize the value function parameters rather than taking random actions. In particular, we update the mean and covariance of a TD estimate of G after each step, and switch to a policy greedy w.r.t. a parameter sample $\widehat{G}$ every $T^{1/2}$ steps. We project the sample onto the positive semidefinite cone.
- A model-based approach which estimates the dynamics parameters $(\widehat{A}, \widehat{B})$ using ordinary least squares. The policy at the end of each phase is produced by treating the estimate as the true parameters. We execute the policy in each phase as in the model-free case, and update parameters using only the data corresponding to random actions.[2]

We also evaluate modified versions of MFLQ, LSPI, and model-based algorithms, where we estimate unknown quantities using all data rather than only the random actions, and denote these versions by an asterisk (e.g. MFLQ*).

To evaluate stability, we run each algorithm 100 times for 10 phases and compute the fraction of times it produces stable policies throughout. Figure 3 (left) shows the results as a function of phase length. MFLQ is the most stable algorithm, and the model-based approach is similar when using all data.

We evaluate solution cost by running each algorithm until we obtain 100 stable trajectories (if possible), where each trajectory consists of 16 phases of length 4096. We compute both the average cost incurred during each phase i, and true expected cost of each policy π_i (for RLSVI we use policy at the end of each multiple of 4096 steps). The cost median and 25th and 75th percentiles at the end of each phase are shown in Figure 3 (center and right). MFLQ considerably outperforms the other model-free methods; however, the model-based approach achieves the lowest steady-state cost. These results are consistent with the empirical findings of Tu and Recht (2017), where model-based approaches outperform discounted LSTDQ.

[2]Since the system of Dean et al. (2017) is unstable without a controller, running a stable policy between random actions ensures numerical stability during system identification from a single trajectory; Dean et al. (2017) force-reset the state to zero every 6 time steps instead.

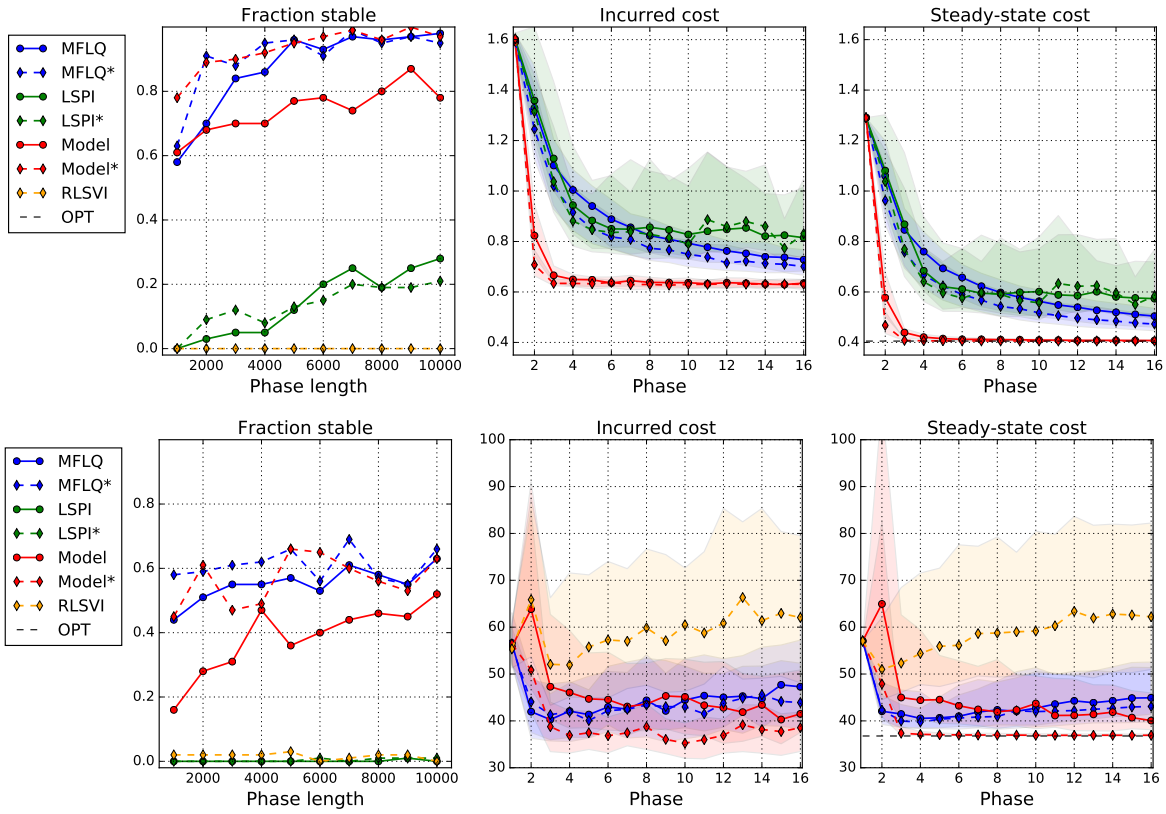

Figure 3: Top row: experimental evaluation on the dynamics of Dean et al. (2017). Bottom row: experimental evaluation on Lewis et al. (2012), Example 11.5.1.

5 Discussion

The simple formulation and wide practical applicability of LQ control make it an idealized benchmark for studying RL algorithms for continuous-valued states and actions. In this work, we have presented MFLQ, an algorithm for model-free control of LQ systems with an $O(T^{3/4})$ regret bound. Empirically, MFLQ considerably improves the performance of standard policy iteration in terms of both solution stability and cost, although it is still not cost-competitive with model-based methods.

Our algorithm is based on a reduction of control of MDPs to an expert prediction problem. In the case of LQ control, the problem structure allows for an efficient implementation and strong theoretical guarantees for a policy iteration algorithm with ϵ-greedy exploration. While ϵ-greedy is known to be suboptimal in unstructured multi-armed bandit problems (Langford and Zhang, 2007), it has been shown to achieve near optimal performance in problems with special structure (Abbasi-Yadkori, 2009, Rusmevichientong and Tsitsiklis, 2010, Bastani and Bayati, 2015), and it is worth considering whether it applies to other structured control problems. However, the same approach might not generalize to other domains. For example, Boltzmann exploration may be more appropriate for MDPs with finite states and actions. We leave this issue, as well as the application of ϵ-greedy exploration to other structured control problems, to future work.

References

Y. Abbasi-Yadkori, P. Bartlett, and V. Kanade. Tracking adversarial targets. In *International Conference on Machine Learning (ICML)*, 2014.

Yasin Abbasi-Yadkori. Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, University of Alberta, 2009.

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, 2011.

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Uncertainty in Artificial Intelligence (UAI)*, pages 1–11, 2015.

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In *AISTATS*, 2017.

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

Sanjeev Arora, Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Towards provable control for unknown linear dynamical systems. *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BygpQlbA-`. rejected: invited to workshop track.

Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates. *SSRN*, 2015.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

S. Bittanti and M.C. Campi. Adaptive control of linear time invariant systems: The bet on the best principle. *Communications in Information and Systems*, 6(4):299–320, 2006.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Steven J Bradtke, B Erik Ydstie, and Andrew G Barto. Adaptive linear quadratic control using policy iteration. In *American Control Conference, 1994*, volume 3, pages 3475–3479. IEEE, 1994.

M. C. Campi and P. R. Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

Nicoló Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

H. Chen and L. Guo. Optimal adaptive control and consistent parameter estimates for ARMAX model with quadratic cost. *SIAM Journal on Control and Optimization*, 25(4):845–867, 1987.

Jie Chen and Guoxiang Gu. *Control-oriented system identification: an H_∞ approach*, volume 19. Wiley-Interscience, 2000.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34 (3):726–736, 2009.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for linearized control problems. *arXiv preprint arXiv:1801.05039*, 2018.

C. Fiechter. PAC adaptive control of linear systems. In *COLT*, 1997.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Arthur J Helmicki, Clas A Jacobson, and Carl N Nett. Control oriented system identification: a worst-case/deterministic approach in H_∞. *IEEE Transactions on Automatic control*, 36(10):1163–1176, 1991.

Morteza Ibrahimi, Adel Javanmard, and Benjamin V. Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems 25*, pages 2636–2644. Curran Associates, Inc., 2012.

V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *arXiv preprint arXiv:1312.3580*, 2013.

T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

T. L. Lai and C. Z. Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25:466–481, 1987.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(Oct):3041–3074, 2012.

Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. *Optimal control*. John Wiley & Sons, 2012.

Bo Liu, Sridhar Mahadevan, and Ji Liu. Regularized off-policy td-learning. In *Advances in Neural Information Processing Systems*, pages 836–844, 2012.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, pages 504–513. Citeseer, 2015.

Lennart Ljung and Torsten Söderström. *Theory and practice of recursive identification*, volume 5. JSTOR, 1983.

G. Neu and V. Gómez. Fast rates for online learning in linearly solvable markov decision processes. In *30th Annual Conference on Learning Theory (COLT)*, 2017.

G. Neu, A. György, Cs. Szepesvári, and A. Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2014.

Ian Osband, Daniel J. Russo, Zheng Wen, and Benjamin Van Roy. Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*, 2017.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with Thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

Bernardo Ávila Pires and Csaba Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2012.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *MATHEMATICS OF OPERATIONS RESEARCH*, 35(2):395–411, 2010.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2): 107–194, 2012.

A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In *ICML*, 2006.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Istvan Szita. *Rewarding Excursions: Extending Reinforcement Learning to Complex Domains*. PhD thesis, Eötvös Loránd University, 2007.

John N. Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35:1799–1808, 1999.

John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081, 1997.

S. Tu, R. Boczar, A. Packard, and B. Recht. Non-Asymptotic Analysis of Robust Control from Coarse-Grained Identification. *arXiv preprint*, 2017.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*, 2017.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

Huizhen Yu and Dimitri P Bertsekas. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

A Q function for linear quadratic systems

Lemma A.1. *Let $\pi(x) = -Kx$ be a linear policy such that $\|A - BK\| < 1$. The state-action value function of policy π has the quadratic form*

$$Q_\pi(x, a) = \begin{pmatrix} x^\top & a^\top \end{pmatrix} G_\pi \begin{pmatrix} x \\ a \end{pmatrix} ,$$

where G_π is the unique symmetric solution of the equation

$$G_\pi = S^\top G_\pi S + \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} , \quad S = \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix} .$$

Proof. We prove the lemma by showing that the given quadratic form is the unique solution of the Bellman equation. Let

$$z = \begin{pmatrix} x \\ a \end{pmatrix} , \qquad z' = \begin{pmatrix} x' \\ a' \end{pmatrix} = \begin{pmatrix} Ax + Ba + w \\ -K(Ax + Ba + w) \end{pmatrix} ,$$

be the current state-action and a random next state-action under policy π. We guess a quadratic form $z^\top G_\pi z + L^\top z$ for the value function and we write

$$\lambda_\pi + \begin{pmatrix} x^\top & a^\top \end{pmatrix} G_\pi \begin{pmatrix} x \\ a \end{pmatrix} + L^\top \begin{pmatrix} x \\ a \end{pmatrix} = \begin{pmatrix} x^\top & a^\top \end{pmatrix} \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} x \\ a \end{pmatrix} + \mathbf{E} \left\{ \begin{pmatrix} x'^\top & a'^\top \end{pmatrix} G_\pi \begin{pmatrix} x' \\ a' \end{pmatrix} + L^\top \begin{pmatrix} x' \\ a' \end{pmatrix} \right\} .$$

By matching terms, we find that the above equation has a solution iff

$$\lambda_\pi = \mathrm{trace}\left(\begin{pmatrix} I \\ -K \end{pmatrix}^\top G_\pi \begin{pmatrix} I \\ -K \end{pmatrix} W \right) ,$$

$$G_\pi = S^\top G_\pi S + \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} , \tag{9}$$

$$L = LS . \tag{10}$$

We have that

$$\left\| \begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} I \\ -K \end{pmatrix} \right\| = \|A - BK\| < 1 .$$

By the associativity of matrix products, this implies that if we form iterations from equations (9) and (10) where the right-hand side is reassigned to the left-hand side, then this iteration converges. By continuity, it follows that the limit will satisfy the respective equations. It also follows that these equations have unique solutions, and in particular $L = 0$. Thus, the quadratic form as stated is the solution of the Bellman equation. □

B Estimating H

B.1 Bounding the smallest eigenvalue of $\mathbf{E}[\phi_t(\phi_t - \phi_{t+1}^\top) + (\phi_t - \phi_{t+1})\phi_t^\top]$

To bound the smallest eigenvalue of $\mathbf{E}[\phi_t(\phi_t - \phi_{t+1}^\top) + (\phi_t - \phi_{t+1})\phi_t^\top]$, we adapt the results of Tu and Recht (2017) who analyze $\mathbf{E}(\phi\phi^\top)$ in the context of LSTD for discounted LQ control. We first state a number of useful results from Tu and Recht (2017). Let $(x_k)_{k=1}^\infty$ be a discrete-time vector-valued process adapted to a filtration $(\mathcal{F}_t)_t$. For an integer k, define the β-mixing coefficient $\beta(k)$ with respect to the steady-state distribution μ as

$$\beta(k) := \sup_{t \geq 1} \mathbf{E}[\left\| \mathbf{P}_{x_{t+k}}(\cdot|\mathcal{F}_t) - \mu \right\|_{\mathrm{tv}}] ,$$

where $\|.\|_{\mathrm{tv}}$ denotes the total-variation norm, and $\mathbf{P}_{x_{t+k}}(\cdot|\mathcal{F}_t)$ is the conditional distribution of x_{t+k} given $\mathcal{F}_t$. We next state mixing time results for a stable linear dynamical system.

Lemma B.1 (Tu and Recht (2017))**.** *Consider a linear dynamical system $x_{t+1} = \Gamma x_t + w_t$, $w_t \sim \mathcal{N}(0, I_n)$, such that $\rho(\Gamma) < 1$, where $\rho(\cdot)$ is the spectral radius of a matrix. Fix any $\alpha \in (\rho(\Gamma), 1)$. For any $k \geq 1$ we have*

$$\beta(k) \leq \frac{\|R_{\alpha^{-1}\Gamma}\|_{\mathcal{H}_\infty}}{2} \sqrt{\mathrm{tr}(\Sigma) + \frac{n}{1 - \alpha^2}} \alpha^k ,$$

where $R_A(a) = (aI - A)^{-1}$, and $\Sigma = \sum_{s=1}^\infty \Gamma^s \Gamma^{s\top}$ is the steady-state covariance of x.

For any positive ω define the small-ball probability $P(\omega)$ as

$$P(\omega) := \inf_{\|v\|=1} \mathbf{P}_\mu(|\langle v, x\rangle| \geq \omega) \ .$$

We use the following theorem of Tu and Recht (2017), which is a generalization of the result of Koltchinskii and Mendelson (2013).

Theorem B.2. *Fix $\delta \in (0,1)$. Suppose that $(x_k)_{k=1}^\infty$ is a discrete-time stochastic process with stationary distribution μ that satisfies $\beta(k) \leq D\alpha^k$ for some $D > 0$ and $\alpha \in (0,1)$. Suppose that there exists a ω satisfying $P(\omega) > 0$. Furthermore, suppose that τ satisfies*

$$\tau \geq \frac{\log(2D\tau/\delta)}{1-\alpha}\left(\max\left\{\frac{1024\mathbf{E}_\mu\|x\|^2}{\omega^2 P^2(\omega)}, \frac{32}{P^2(\omega)}\log\left(\frac{4}{\delta(1-\alpha)}\log\left(\frac{2D\tau}{\delta}\right)\right)\right\}+1\right) \ .$$

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

$$\lambda_{min}\left(\frac{1}{\tau}\sum_{k=1}^{\tau} x_k x_k^\top\right) \geq \frac{\omega^2 P(\omega)}{8} \ .$$

We are now ready to state and prove Lemma 3.2.

Proof. We apply Theorem B.2 to the process $\phi(x_t) - \phi(x_{t+1})$. In Section B.2, we show that the small-ball probability of this process is bounded as

$$\mathbf{P}\left(|\langle v, \phi(x_{t+1}) - \phi(x_t)\rangle| \geq 1\right) \geq 1/324 \ , \tag{11}$$

where $\|v\| = 1$. Next, we upper-bound the second moment,

$$\begin{aligned}
\mathbf{E}\left[\|\phi(x_{t+1}) - \phi(x_t)\|^2\right] &\leq 4\mathbf{E}\left[\|\phi(x_t)\|^2\right] \\
&= 4\mathbf{E}\left[\|x_t\|^4\right] \\
&= 8\left\|\Sigma^{1/2}\right\|_F^2 + 4\,\mathrm{tr}(\Sigma^{1/2})^2 \\
&\leq 12\,\mathrm{tr}(\Sigma^{1/2})^2 \ .
\end{aligned} \tag{12}$$

Given (11) and (12), we get the desired lower bound from Theorem B.2. □

B.2 Bounding the small-ball probability

In this section we bound the small-ball probability $P(\omega) = \inf_{\|v\|=1} \mathbf{P}_\mu(|\langle v, \phi' - \phi\rangle| \geq \omega)$ for $\omega = 1$. We assume identity noise covariance $W = I$ for simplicity; with known $W \succ 0$, we can always change coordinates to whiten the noise. We first state a useful result for fourth moment of multivariate Gaussians. A proof can be found in (Tu and Recht, 2017).

Proposition B.3. *Let $g \sim \mathcal{N}(0, I)$, and let F, F' be two fixed symmetric matrices. We have that*

$$\mathbf{E}\left[g^\top F g g^\top F' v\right] = 2\langle F, F'\rangle + \mathrm{tr}(F)\,\mathrm{tr}(F') \ .$$

Write $x = \Sigma^{1/2} g$ and $x' = \Gamma\Sigma^{1/2} g + g'$, where $g \sim \mathcal{N}(0, I)$ and $g' \sim \mathcal{N}(0, I)$ are independent. For any unit norm vector v of appropriate dimension, let $V = \textsc{mat}(v)$, where MAT is the inverse of the VEC operation. Define

$$\begin{aligned}
f_v(g, g') &= \langle v, \phi' - \phi\rangle = \langle v, \phi(\Gamma\Sigma^{1/2} g + g') - \phi(\Sigma^{1/2} g)\rangle \\
&= (g'^\top + g^\top \Sigma^{1/2}\Gamma^\top) V (\Gamma\Sigma^{1/2} g + g') - g^\top \Sigma^{1/2} V \Sigma^{1/2} g \ .
\end{aligned}$$

Function f_v is a degree two polynomial in g and g'. Next we show the following lower bound for the second moment of f_v,

$$\mathbf{E}(f_v(g, g')^2) \geq 2 \ .$$

We decompose the expectation as $\mathbf{E}(f_v(g, g')^2) = T_1 + T_2 + T_3$, where

$$\begin{aligned}
T_1 &= \mathbf{E}\left[((g'^\top + g^\top \Sigma^{1/2}\Gamma^\top) V (\Gamma\Sigma^{1/2} g + g'))^2\right] \\
T_2 &= \mathbf{E}\left[(g^\top \Sigma^{1/2} V \Sigma^{1/2} g)^2\right] \\
T_3 &= -2\mathbf{E}\left[(g^\top \Sigma^{1/2} V \Sigma^{1/2} g)((g'^\top + g^\top \Sigma^{1/2}\Gamma^\top) V (\Gamma\Sigma^{1/2} g + g'))\right] \ .
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
T_1 &= \mathbf{E}\left[(g'^{\top} V g' + g^{\top}\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} g + 2 g'^{\top} V \Gamma \Sigma^{1/2} g)^2\right] \\
&= \mathbf{E}\left[(g'^{\top} V g')^2\right] + \mathbf{E}\left[(g^{\top}\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} g)^2\right] + 4\mathbf{E}\left[(g'^{\top} V \Gamma \Sigma^{1/2} g)^2\right] \\
&\qquad + 2\mathbf{E}\left[(g'^{\top} V g')(g^{\top}\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} g)\right] + 4\mathbf{E}\left[(g'^{\top} V g')(g'^{\top} V \Gamma \Sigma^{1/2} g)\right] \\
&\qquad + 4\mathbf{E}\left[(g^{\top}\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} g)(g'^{\top} V \Gamma \Sigma^{1/2} g)\right] \\
&= 2\left\|V\right\|_F^2 + \operatorname{tr}(V)^2 + 2\left\|\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2}\right\|_F^2 + \operatorname{tr}(\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2})^2 && \text{Proposition (B.3) and } g \perp g' \\
&\qquad + 4\mathbf{E}\left[(g'^{\top} V \Gamma \Sigma^{1/2} g)^2\right] + 2\mathbf{E}\left[g'^{\top} V g'\right]\mathbf{E}\left[g^{\top}\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} g\right] \\
&= 2\left\|V\right\|_F^2 + \operatorname{tr}(V)^2 + 2\left\|\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2}\right\|_F^2 + \operatorname{tr}(\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2})^2 \qquad (13) \\
&\qquad + 4\left\|\Sigma^{1/2}\Gamma^{\top} V\right\|_F^2 + 2\operatorname{tr}(V)\operatorname{tr}(\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2}) ,
\end{aligned}$$

where in the last step we used

$$\mathbf{E}\left[(g'^{\top} V \Gamma \Sigma^{1/2} g)^2\right] = \mathbf{E}\left[g'^{\top} V \Gamma \Sigma^{1/2} g g^{\top}\Sigma^{1/2}\Gamma^{\top} V g'\right] = \operatorname{tr}(\Sigma^{1/2}\Gamma^{\top} V^2 \Gamma \Sigma^{1/2}) = \left\|\Sigma^{1/2}\Gamma^{\top} V\right\|_F^2 .$$

For the second term, we have

$$T_2 = 2\left\|\Sigma^{1/2} V \Sigma^{1/2}\right\|_F^2 + \operatorname{tr}(\Sigma^{1/2} V \Sigma^{1/2})^2 . \qquad (14)$$

For the third term, we have

$$\begin{aligned}
T_3 &= -2\mathbf{E}\left[(g^{\top}\Sigma^{1/2} V \Sigma^{1/2} g)(g'^{\top} V g')\right] - 2\mathbf{E}\left[(g^{\top}\Sigma^{1/2} V \Sigma^{1/2} g)(g^{\top}\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} g)\right] \\
&\qquad - 4\mathbf{E}\left[(g^{\top}\Sigma^{1/2} V \Sigma^{1/2} g)(g'^{\top} V \Gamma \Sigma^{1/2} g)\right] \\
&= -2\operatorname{tr}(\Sigma^{1/2} V \Sigma^{1/2})\operatorname{tr}(V) - 4\langle \Sigma^{1/2} V \Sigma^{1/2}, \Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2}\rangle - 2\operatorname{tr}(\Sigma^{1/2} V \Sigma^{1/2})\operatorname{tr}(\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2}) . \qquad (15)
\end{aligned}$$

By (13), (14), and (15), we get that

$$\begin{aligned}
\mathbf{E}(f_v(g, g')^2) &= (\operatorname{tr}(\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2}) - \operatorname{tr}(\Sigma^{1/2} V \Sigma^{1/2}) + \operatorname{tr}(V))^2 + 2\left\|\Sigma^{1/2}\Gamma^{\top} V \Gamma \Sigma^{1/2} - \Sigma^{1/2} V \Sigma^{1/2}\right\|_F^2 \\
&\qquad + 2\left\|V\right\|_F^2 + 4\left\|\Sigma^{1/2}\Gamma^{\top} V\right\|_F^2 \\
&\geq 2\left\|V\right\|_F^2 = 2 .
\end{aligned}$$

Using the above lower bound and Lemma 4.6 of Tu and Recht (2017),

$$\mathbf{P}\left(|\langle v, \phi(x_{t+1}) - \phi(x_t)\rangle| \geq 1\right) \geq 1/324 . \qquad (16)$$

C Analysis of the MFLQ algorithm

C.1 Proof of Lemma 3.4

Proof. Let $G^j = \frac{1}{j}\sum_{i=1}^{j} G_i$ and $\widehat{G}^j = \frac{1}{j}\sum_{i=1}^{j} \widehat{G}_i$ be the averages of true and estimated state-action value matrices of policies $K_1, \ldots, K_j$, respectively. Let H^j and $\widehat{H}^j$ be the corresponding value matrices. The greedy policy with respect to $\widehat{G}^j$ is given by:

$$K_{j+1} = \arg\min_K \operatorname{tr}\left(x^{\top}\begin{bmatrix} I & -K^{\top}\end{bmatrix}\widehat{G}^j\begin{bmatrix} I \\ -K\end{bmatrix} x\right) = \arg\min_K \operatorname{tr}\left(\widehat{G}^j X_K\right) , \qquad (17)$$

$$\text{where } X_K = \begin{bmatrix} I \\ -K\end{bmatrix} x x^{\top}\begin{bmatrix} I & -K^{\top}\end{bmatrix} . \qquad (18)$$

Let $|X_K|$ be the matrix obtained from X_K by taking the absolute value of each entry. We have the following:

$$\begin{aligned}
\operatorname{tr}(G_j X_{K_{j+1}}) &\le \operatorname{tr}(\widehat{G}_j X_{K_{j+1}}) + \varepsilon_1 \operatorname{tr}(\mathbf{1}\mathbf{1}^\top |X_{K_{j+1}}|) && (19)\\
&\le \operatorname{tr}(\widehat{G}_j X_{K_j}) + \varepsilon_1 \operatorname{tr}(\mathbf{1}\mathbf{1}^\top |X_{K_{j+1}}|) && (20)\\
&\le \operatorname{tr}(G_j X_{K_j}) + \varepsilon_1 \operatorname{tr}(\mathbf{1}\mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|)) && (21)\\
&= x^\top H_j x + \varepsilon_1 \operatorname{tr}(\mathbf{1}\mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|)) && (22)
\end{aligned}$$

Here, (19) and (21) follow from the error bound, and (22) follows from $\operatorname{tr}(G_j X_{K_j}) = x^\top H_j x$. To see (20), note that K_{i+1} is optimal for $\widehat{G}^i$ and we have:

$$\begin{aligned}
\operatorname{tr}(\widehat{G}^j X_{K_{j+1}}) &= \frac{j-1}{j} \operatorname{tr}(\widehat{G}^{j-1} X_{K_{j+1}}) + \frac{1}{j} \operatorname{tr}(\widehat{G}_j X_{K_{j+1}}) \\
&\le \frac{j-1}{j} \operatorname{tr}(\widehat{G}^{j-1} X_{K_j}) + \frac{1}{j} \operatorname{tr}(\widehat{G}_j X_{K_j}) = \operatorname{tr}(\widehat{G}^j X_{K_j}) \,.
\end{aligned}$$

Since $\operatorname{tr}(\widehat{G}^{j-1} X_{K_j}) \le \operatorname{tr}(\widehat{G}^{j-1} X_{K_{j+1}})$ it follows that $\operatorname{tr}(\widehat{G}_j X_{K_{j+1}}) \le \operatorname{tr}(\widehat{G}_j X_{K_j})$.

Now note that we can rewrite $\operatorname{tr}(G_j X_{K_{j+1}})$ as a function of H_j as follows:

$$\begin{aligned}
\operatorname{tr}(G_j X_{K_{j+1}}) &= x^\top \begin{bmatrix} I & -K_{j+1}^\top \end{bmatrix} G_j \begin{bmatrix} I \\ -K_{j+1} \end{bmatrix} x \\
&= x^\top \Bigg(\begin{bmatrix} I & -K_{j+1}^\top \end{bmatrix} \Bigg(\begin{bmatrix} A^\top \\ B^\top \end{bmatrix} H_j \begin{bmatrix} A & B \end{bmatrix} + \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} \Bigg) \begin{bmatrix} I \\ -K_{j+1} \end{bmatrix} \Bigg) x \\
&= x^\top \Big((A - BK_{j+1})^\top H_j (A - BK_{j+1}) \Big) x + \operatorname{tr} \Bigg(\begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} X_{K_{j+1}} \Bigg)
\end{aligned}$$

Thus we have that

$$x^\top \Big((A - BK_{j+1})^\top H_j (A - BK_{j+1}) \Big) x + \varepsilon_2 \le x^\top H_j x \tag{23}$$

$$\text{where } \ \varepsilon_2 = x^\top (M + K_{j+1}^\top N K_{j+1}) x - \varepsilon_1 \mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|) \mathbf{1}.$$

If the estimation error ε_1 is small enough so that $\varepsilon_2 > 0$ for any unit-norm x and all policies, then $H_j \succ (A - BK_{j+1})^\top H_j (A - BK_{j+1})$ and K_{j+1} is stable by a Lyapunov theorem. Since K_1 is stable and H_1 bounded, all policies remain stable.

In order to have $\varepsilon_2 > 0$, it suffices to have

$$\varepsilon_1 < \big((\sqrt{n} + \|K_j\| \sqrt{d})^2 + (\sqrt{n} + \|K_{j+1}\| \sqrt{d})^2 \big)^{-1}.$$

This follows since $M \succ I$, and since for any unit norm vector $x \in \mathbb{R}^n$, $\mathbf{1}^\top x x^\top \mathbf{1} \le n$, with equality achieved by $x = \frac{1}{\sqrt{n}} \mathbf{1}$. Similarly, $\mathbf{1}^\top K x x^\top K^\top \mathbf{1} \le \|K\|^2 d$, and $\mathbf{1}^\top (|X_{K_j}|) \mathbf{1} \le (\sqrt{n} + \|K_j\| \sqrt{d})^2$.

As we will see, we need a smaller estimation error in phase j:

$$\varepsilon_1 < \frac{1}{6 C_1 S} \big((\sqrt{n} + \|K_j\| \sqrt{d})^2 + (\sqrt{n} + \|K_{j+1}\| \sqrt{d})^2 \big)^{-1}. \tag{24}$$

Here, C_1 is an upper bound on $\|H_1\|$; note that $H_1 \succ M \succ I$, so $C_1 > 1$. The above condition guarantees that

$$\varepsilon_1 \mathbf{1}^\top (|X_{K_j}| + |X_{K_{j+1}}|) \mathbf{1} \le \frac{1}{6 C_1 S} \,.$$

We have that $G_{1,22} \succ N \succ I$ and $G_{1,21} = B^\top H_1 A$. Given that the estimation error (7) is small, we have $\|K_2\| \le 2(\big\|B^\top H_1 A\big\| + 1) \le C_K$. Then (7) implies (24) for $j = 1$, and the above argument shows that K_2 is stable.

Next, we show a bound on $\big\|(A - BK_i)^k\big\|$. Let $\Gamma_i = A - BK_i$ and $L_{i+1} = H_i^{1/2} \Gamma_{i+1} H_i^{-1/2}$. By (23), $M \succ I$, and the error bound,

$$\begin{aligned}
H_1 &\succ \Gamma_2^\top H_1 \Gamma_2 + (M + K_2^\top N K_2) - (6 C_1 S)^{-1} I \\
I &\succ L_2^\top L_2 + H_1^{-1/2} (M + K_2^\top N K_2) H_1^{-1/2} - (6 C_1 S)^{-1} H_1^{-1}
\end{aligned}$$

$$\succ L_2^\top L_2 + H_1^{-1} - (6C_1)^{-1} I$$
$$\succ L_2^\top L_2 + (3C_1)^{-1} I - (6C_1)^{-1} I \; .$$

Thus, $\|L_2\| \leq \sqrt{1-(6C_1)^{-1}}$ and we have that

$$\left\|(A-BK_2)^k\right\| = \left\|(H_1^{-1/2} L_2 H_1^{1/2})^k\right\| \leq \sqrt{C_1}(1-(6C_1)^{-1})^{k/2}.$$

Finally, we show explicit bounds on the value functions. We have that

$$H_2 = \Gamma_2^\top H_2 \Gamma_2 + M + K_2^\top N K_2 \; .$$

Thus

$$H_2 - H_1 \prec \Gamma_2^\top (H_2 - H_1)\Gamma_2 + (6C_1 S)^{-1} I$$

Thus for some $C' \succ 0$,

$$H_2 - H_1 = \Gamma_2^\top (H_2 - H_1)\Gamma_2 + (6C_1 S)^{-1} I - C'$$

From stability of Γ_2, we have

$$H_2 - H_1 = \sum_{k=0}^{\infty} (\Gamma_2^\top)^k \big((6C_1 S)^{-1} I - C'\big) \Gamma_2^k \prec (6C_1 S)^{-1} \sum_{k=0}^{\infty} (\Gamma_2^\top)^k \Gamma_2^k$$
$$\|H_2\| \leq \|H_1\| + \frac{C_1}{6C_1 S(1-\|L_2\|^2)} \leq (1+S^{-1})C_1$$

Thus $C_2 \leq (1+S^{-1})C_1$, and by repeating the same argument,

$$C_i \leq (1+S^{-1})^i C_1 \leq 3C_1 \; . \tag{25}$$

□

C.2 Proof of Lemma 3.5

Proof. First, we show that states remain bounded. Let $\Gamma_i = A - BK_i$, and let $\Gamma_{(t)}$ be the closed-loop matrix at time step t. We have that

$$\begin{aligned} x_{t+1} &= \Gamma_{(t)} x_t + w_t \\ &= \Gamma_{(t)} \Gamma_{(t-1)} x_{t-1} + \Gamma_{(t)} w_{t-1} + w_t \\ &= \dots \\ &= \left(\prod_{s=0}^{t} \Gamma_{(t-s)} \right) x_0 + \sum_{\tau=0}^{t} \left(\prod_{s=0}^{t-\tau-1} \Gamma_{(t-s)} \right) w_\tau \end{aligned}$$

For each policy K_i, we have that $\left\|\Gamma_i^k\right\| \leq \sqrt{C_H}\,(1-(2C_H)^{-1})^{k/2}$. Furthermore, since we run $T^{1/4}$ policies for $T^{3/4}$ time steps each, after t time steps there have been at most $\lfloor t/T^{3/4} \rfloor$ policies. Hence

$$\left\| \prod_{s=1}^{t} \Gamma_{(t+1-s)} \right\| \leq \left({C_H}^{T^{-3/4}} (1-(2C_H)^{-1}) \right)^{t/2} \tag{26}$$

By assumption (8), $T^{3/4} \geq \frac{\ln(C_H)}{\ln(1+(2C_H)^{-1})}$. Hence ${C_H}^{T^{-3/4}} \leq 1 + (2C_H)^{-1}$ and

$$\left\| \prod_{s=1}^{t} \Gamma_{(t+1-s)} \right\| \leq \left(1-(2C_H)^{-2}\right)^{t/2} .$$

Given that all noise terms are smaller than $\sqrt{2n\log(Tn/\delta_2)}$ with probability at least $1-\delta_2$, we get for all $s = 1, \dots, \tau+1$

$$\|x_s\| \leq \frac{\sqrt{2n\log(Tn/\delta_2)}}{1-\sqrt{1-(2C_H)^{-2}}} \; .$$

Next, we show that all actions remain bounded. We have that

$$K_i^\top K_i \prec (A - BK_i)^\top H_i (A - BK_i) + M + K_i^\top N K_i = H_i \ .$$

Thus,

$$\|a_s\|^2 \le C_H \|x_s\|^2 \le \frac{2nC_H \log(Tn/\delta_2)}{(1 - \sqrt{1 - (2C_H)^{-2}})^2} \ .$$

□

C.3 Regret bound

Because we use FTL as our expert algorithm and value functions are quadratic, we can use the following regret bound for the FTL algorithm (Theorem 3.1 in (Cesa-Bianchi and Lugosi, 2006)).

Theorem C.1 (FTL Regret Bound)**.** *Assume that the loss function $f_t(\cdot)$ is convex, is Lipschitz with constant F_1, and is twice differentiable everywhere with Hessian $H \succ F_2 I$. Then the regret of the Follow The Leader algorithm is bounded by*

$$B_T \le \frac{F_1^2}{2F_2}(1 + \log T) \ .$$

We prove Lemma 3.6 next.

Proof. Because we execute $S = O(T^{1/4})$ policies, each for $\tau + \tau'\tau'' = O(T^{3/4})$ rounds,

$$\begin{aligned}
\beta_T &= \sum_{i=1}^{S} (\tau + \tau'\tau'') \mathbf{E}_{x \sim \mu_\pi} (Q_i(x, \pi_i(x)) - Q_i(x, \pi(x))) \\
&= (\tau + \tau'\tau'') \sum_{i=1}^{S} \mathbf{E}_{x \sim \mu_\pi} (\widehat{Q}_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi(x))) \\
&\quad + (\tau + \tau'\tau'') \sum_{i=1}^{S} \mathbf{E}_{x \sim \mu_\pi} (Q_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi_i(x))) \\
&\quad + (\tau + \tau'\tau'') \sum_{i=1}^{S} \mathbf{E}_{x \sim \mu_\pi} (\widehat{Q}_i(x, \pi(x)) - Q_i(x, \pi(x))) \\
&\le C'(\tau + \tau'\tau'') \log T + (\tau + \tau'\tau'') \sum_{i=1}^{S} \mathbf{E}_{x \sim \mu_\pi} (\widehat{Q}_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi(x))) \,,
\end{aligned}$$

where the last inequality holds by Lemma 3.3. Consider the remaining term:

$$E_T = (\tau + \tau'\tau'') \sum_{i=1}^{S} \mathbf{E}_{x \sim \mu_\pi} (\widehat{Q}_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi(x))) \ .$$

We bound this term using the FTL regret bound. We show that the conditions of Theorem C.1 hold for the loss function $f_i(K) = \mathbf{E}_{x \sim \mu_\pi} (\widehat{Q}_i(x, Kx))$. Let Σ_π be the covariance matrix of the steady-state distribution $\mu_\pi(x)$. We have that

$$\begin{aligned}
f_i(K) &= \operatorname{tr}\left(\Sigma_\pi \big(\widehat{G}_{i,11} - K^\top \widehat{G}_{i,21} - \widehat{G}_{i,12} K + K^\top \widehat{G}_{i,22} K \big) \right) \\
\nabla_K f_i(K) &= 2\Sigma_\pi \big(K^\top \widehat{G}_{i,22} - \widehat{G}_{i,12} \big) \\
&= 2\textsc{mat}\big((\widehat{G}_{i,22} \otimes \Sigma_\pi) \textsc{vec}(K) \big) - 2\Sigma_\pi \widehat{G}_{i,12} \\
\nabla^2_{\textsc{vec}(K)} f_i(K) &= 2\widehat{G}_{i,22} \otimes \Sigma_\pi
\end{aligned}$$

Boundedness and Lipschitzness of the loss function $f_i(K_i)$ follow from the boundedness of policies K_i and value matrix estimates $\widehat{G}_i$. By Lemma 3.4, we have that $\|K_i\| \le C_K$. To bound $\|\widehat{G}_i\|$, note that

$$G_i = \begin{pmatrix} A^\top \\ B^\top \end{pmatrix} H_i \begin{pmatrix} A & B \end{pmatrix} + \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix}$$

$$\|G_i\| \le C_H(\|A\| + \|B\|)^2 + \|M\| + \|N\| \qquad \text{(Lemma 3.4)}$$
$$\left\|\widehat{G}_i\right\| \le \|G_i\| + \varepsilon_1\sqrt{n+d} \qquad \text{(Lemma 3.3).}$$

The Hessian lower bound is $\nabla^2_{\mathrm{vec}(K)} f_i(K) \succ F_2 I$, where F_2 is given by two times the product of the minimum eigenvalues of Σ_π and $\widehat{G}_{i,22}$. For any stable policy $\pi(x) = Kx$, the covariance matrix of the stationary distribution satisfies $\Sigma_\pi \succ W$, and we project the estimates $\widehat{G}_i$ onto the constraint $\widehat{G}_i \succeq \left(\begin{smallmatrix} M & 0 \\ 0 & N \end{smallmatrix}\right)$. Therefore the Hessian of the loss is lower-bounded by $2\lambda_{\min}(W)I$.

Thus by Theorem C.1, $E_T \le (\tau + \tau'\tau'')\log S = C''T^{3/4}\log T$ for an appropriate constant C''.

The bound on $\gamma_T = \sum_{t=1}^T(\lambda_\pi - c(x_t^\pi, \pi(x_t^\pi)))$ is a consequence of the fact that states and actions remain bounded, and the state distribution under policy π converges to its stationary distribution exponentially fast. The bound on $\alpha_T = \sum_{t=1}^T(c(x_t, a_t) - \lambda_{\pi_t})$ is due to the fact that we have $S = O(T^{1/4})$ policy switches and in each policy execution, we have $\tau' = O(T^{1/2})$ random actions. □