Model-Free Linear Quadratic Control via Reduction to Expert Prediction

Yasin Abbasi-Yadkori
Adobe Research

Nevena Lazić
Google Brain

Csaba Szepesvári
Deepmind

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 $O(T^{\xi+2/3})$ for any small $\xi > 0$ if time horizon satisfies $T > C^{1/\xi}$ for a constant C. 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 forced exploration, where the policy in each phase 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 $O(T^{\xi+2/3})$ 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 exploration similar to ϵ-greedy, but performed at a fixed schedule. 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 value function is quadratic, the average of all previous value functions is also quadratic.

One major challenge in this work is the finite-time analy-

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

sis of the value function estimation error. Existing finite-sample results either consider bounded functions or discounted problems, and are not applicable in our setting. Our analysis relies on the contractiveness of stable policies, as well as the fact that our algorithm takes exploratory actions. Another 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 by showing that the policies produced by our algorithm are stable assuming a sufficiently small 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 have 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 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 $c_t = c(x_t, a_t)$. 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:

$$\lambda_\pi := \lim_{T \to +\infty} \mathbf{E}_\pi \left[\frac{1}{T} \sum_{t=1}^{T} c(x_t, a_t) \right] ,$$

which does not depend on the initial state in the problems that we consider in this paper. The corresponding bias function, also called value function in this paper, associated with a stationary policy π is given by:

$$V_\pi(x) := \lim_{T \to +\infty} \mathbf{E}_\pi \left[\sum_{t=1}^{T} (c(x_t, a_t) - \lambda_\pi(x_t)) \right] .$$

The average cost λ_π and value V_π satisfy the following evaluation equation for any state $x \in \mathcal{X}$,

$$V_\pi(x) = c(x, \pi(x)) - \lambda_\pi + \mathbf{E}_{x' \sim P(.|x,\pi(x))}(V_\pi(x')).$$

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

$$x_{t+1} = A x_t + B a_t + w_{t+1} , \quad c_t = x_t^\top M x_t + a_t^\top N a_t .$$

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 $\mathcal{N}(0, W)$.

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

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 e.g. Abbasi-Yadkori et al. (2014)):

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

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

where $H_\pi \succ 0$ and $G_\pi \succ 0$. We call H_π the value matrix of policy π. The matrix G_π is the unique solution of the equation

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

The greedy policy with respect to Q_π is given by

$$\pi'(x) = \operatorname*{argmin}_a Q_\pi(x, a) = -G_{\pi,22}^{-1} G_{\pi,21} x = -Kx .$$

Here, $G_{\pi,ij}$ for $i, j \in \{1, 2\}$ refers to (i, j)'s block of matrix G_π where block structure is based on state and action dimensions. The average expected cost of following a linear policy is $\lambda_\pi = \mathrm{tr}(H_\pi W)$. 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

$$\Sigma = (A - BK)\Sigma(A - BK)^\top + W .$$

3 MODEL-FREE LQ CONTROL

Our model-free linear quadratic control algorithm (MFLQ) is shown in algorithm 1, where V1 and V2 indicate different versions. At a high level, MFLQ is a variant of policy iteration with a deterministic exploration schedule. 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 fixed number of rounds, and compute a value function estimate $\widehat{V}_i$. We then estimate Q_i from $\widehat{V}_i$ and a dataset $\mathcal{Z} = \{(x_t, a_t, x_{t+1})\}$ which includes exploratory actions. We set π_{i+1} to the greedy policy with respect to the average of all previous estimates $\widehat{Q}_1, ..., \widehat{Q}_i$. This step is different than standard policy iteration (which only considers the most recent value estimate), and a consequence of using the FOLLOW-THE-LEADER expert algorithm (Cesa-Bianchi and Lugosi, 2006).

The dataset $\mathcal{Z}$ is generated by executing the policy and taking a random action every T_s steps. In MFLQV1, we generate $\mathcal{Z}$ at the beginning, and reuse it in all phases, while

Algorithm 1 MFLQ

MFLQ (stable policy π_1, trajectory length T, initial state x_0, exploration covariance Σ_a)

v1: $S = T^{1/3-\xi} - 1, T_s = const, T_v = T^{2/3+\xi}$

v1: $\mathcal{Z} = \textsc{CollectData}(\pi_1, T_v, T_s, \Sigma_a)$

v2: $S = T^{1/4}, T_s = T^{1/4-\xi}, T_v = 0.5T^{3/4}$

for $i = 1, 2, \dots, S$ **do**

Execute π_i for T_v rounds and compute $\widehat{V}_i$ using (2)

v2: $\mathcal{Z} = \textsc{CollectData}(\pi_i, T_v, T_s, \Sigma_a)$

Compute $\widehat{Q}_i$ from $\mathcal{Z}$ and $\widehat{V}_i$ using (7)

$\pi_{i+1}(x) = \text{argmin}_a \sum_{j=1}^{i} \widehat{Q}_j(x, a) = -K_{i+1}x$

end for

$\textsc{CollectData}$ (policy π, traj. length τ, exploration period s, cov. Σ_a):

$\mathcal{Z} = \{\}$

for $k = 1, 2, \dots, \lfloor \tau/s \rfloor$ **do**

Execute the policy π for $s - 1$ rounds and let x be the final state

Sample $a \sim \mathcal{N}(0, \Sigma_a)$, observe next state x_+, add (x, a, x_+) to $\mathcal{Z}$

end for

return $\mathcal{Z}$

in v2 we generate a new dataset $\mathcal{Z}$ in each phase following the execution of the policy. While MFLQ as described stores long trajectories in each phase, this requirement can be removed by updating parameters of V_i and Q_i in an online fashion. However, in the case of MFLQv1, we need to store the dataset $\mathcal{Z}$ throughout (since it gets reused), so this variant is more memory demanding.

Assume the initial policy is stable and let C_1 be the norm of its value matrix. Our main result are the following two theorems.

Theorem 3.1. *For any $\delta, \xi > 0$, appropriate constants C and $\overline{C}$, and $T > \overline{C}^{1/\xi}$, the regret of the* MFLQv1 *algorithm is bounded as*

$$\textit{Regret}_T \le C T^{2/3+\xi} \log T \ ,$$

where C and $\overline{C}$ scale as $\text{poly}(n, d, \|H_1\|, \log(1/\delta))$.

Theorem 3.2. *For any $\delta, \xi > 0$, appropriate constants C and $\overline{C}$, and $T > \overline{C}^{1/\xi}$, the regret of the* MFLQv2 *algorithm is bounded as*

$$\textit{Regret}_T \le C T^{3/4+\xi} \log T \ ,$$

where C and $\overline{C}$ scale as $\text{poly}(n, d, \|H_1\|, \log(1/\delta))$.

To prove the above theorems, we rely on the following regret decomposition. The regret of an algorithm with respect to a fixed policy π can be written as

$$\text{Regret}_T = \sum_{t=1}^{T} c_T - \sum_{t=1}^{T} c_t^\pi = \alpha_T + \beta_T + \gamma_T \ ,$$

$$\alpha_T = \sum_{t=1}^{T} c_t - \lambda_{\pi_{(t)}}, \ \beta_T = \sum_{t=1}^{T} \lambda_{\pi_{(t)}} - \lambda_\pi, \ \gamma_T = \sum_{t=1}^{T} \lambda_\pi - c_t^\pi$$

where $\pi_{(t)}$ is the policy played by the algorithm at time t. 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 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

$$\begin{aligned} Q_i(x, \pi_i(x)) - Q_i(x, \pi(x)) &= \widehat{Q}_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi(x)) \\ &+ Q_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi_i(x)) \\ &+ \widehat{Q}_i(x, \pi(x)) - Q_i(x, \pi(x)) \ . \end{aligned} \tag{1}$$

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. Given sufficiently small estimation errors, we show that all policies remain stable, and hence all 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.

4 VALUE FUNCTION ESTIMATION

4.1 State value function

In this section, we study least squares temporal difference (LSTD) estimates of the value matrix H_π. In order to simplify notation, we will drop π subscripts in this section. In steady state, with $a_t = \pi(x_t)$, we have the following:

$$\begin{aligned} V(x_t) &= c(x_t, a_t) + \mathbf{E}(V(x_{t+1})|x_t, a_t) - \lambda \\ x_t^\top H x_t &= c(x_t, a_t) + \mathbf{E}(x_{t+1}^\top H x_{t+1}|x_t, a_t) - \text{tr}(WH) \ . \end{aligned}$$

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 $c_t = c(x_t, a_t)$. The vectorized version of the Bellman equation is

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

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

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

We estimate H from data generated by following the policy for τ rounds. Let Φ be a $\tau \times n^2$ matrix whose rows are vectors $\phi_1, ..., \phi_\tau$, and similarly let Φ_+ be a matrix whose rows are $\phi_2, ..., \phi_{\tau+1}$. Let $\mathbf{W}$ be a $\tau \times n^2$ matrix whose each row is $\text{VEC}(W)$. Let $\mathbf{c} = [c_1, ..., c_\tau]^\top$. The LSTD estimator of H is given by (see e.g. Tsitsiklis and Roy (1999), Yu and Bertsekas (2009)):

$$\text{VEC}(\widehat{H}) = \left(\Phi^\top(\Phi - \Phi_+ + \mathbf{W})\right)^\dagger \Phi^\top \mathbf{c}\,, \qquad (2)$$

where $(.)^\dagger$ denotes the pseudo-inverse. Given that $H \succ M$, we project our estimate onto the constraint $\widehat{H} \succ M$. Note that this step can only decrease the estimation error, since an orthogonal projection onto a closed convex set is contractive.

Remark. Since the average cost $\text{tr}(WH)$ cannot be computed from value function parameters H alone, assuming known noise covariance W seems necessary. However, if W is unknown in practice, we can use the following estimator instead, which relies on the empirical average cost $\overline{c} = \frac{1}{\tau}\sum_{t=1}^{\tau} c_t$:

$$\text{VEC}(\widetilde{H}_\tau) = \left(\Phi^\top(\Phi - \Phi_+)\right)^\dagger \Phi^\top(\mathbf{c} - \overline{c}\mathbf{1}). \qquad (3)$$

Lemma 4.1. *Let Σ_t be the state covariance at time step t, and let Σ_π be the steady-state covariance of the policy. Let $\Sigma_V = \Sigma_\pi + \Sigma_0$ and let $\Gamma = A - BK$. With probability at least $1 - \delta$, we have*

$$\|H - \widehat{H}\|_F = \frac{\|H\|^2}{\lambda_{\min}(M)^2} O\left(\tau^{-1/2}\left\|W\widehat{H}\right\|_F \text{tr}(\Sigma_V)\right.$$
$$\left.\left\|\Gamma\Sigma_V^{1/2}\right\| \text{polylog}(\tau, 1/\delta)\right). \qquad (4)$$

The proof is similar to the analysis of LSTD for the discounted LQ setting by Tu and Recht (2017); however, instead of relying on the contractive property of the discounted Bellman operator, we use the contractiveness of stable policies.

Proof. Let $\overline{\Phi}_+$ be a $\tau \times n^2$ matrix, whose rows correspond to vectors $\mathbf{E}[\phi_2|\phi_1, \pi], \ldots, \mathbf{E}[\phi_{\tau+1}|\phi_\tau, \pi]$. Let $P_\Phi = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$ be the orthogonal projector onto Φ. The value function estimate $\hat{h} = \text{VEC}(\widehat{H})$ and the true value function parameters $h = \text{VEC}(h)$ satisfy the following:

$$\Phi\hat{h} = P_\Phi(\mathbf{c} + (\Phi_+ - \mathbf{W})\hat{h})$$
$$\Phi h = \mathbf{c} + (\overline{\Phi}_+ - \mathbf{W})h\,.$$

Subtracting the two previous equations, and adding and subtracting $\overline{\Phi}_+\hat{h}$, we have:

$$P_\Phi(\Phi - \overline{\Phi}_+ + \mathbf{W})(h - \hat{h}) = P_\Phi(\overline{\Phi}_+ - \Phi_+)\hat{h}\,.$$

In Appendix A.1, we show that the left-hand side can be equivalently be written as

$$P_\Phi(\Phi - \overline{\Phi}_+ + \mathbf{W})(h - \widehat{h}) = \Phi(I - \Gamma\otimes\Gamma)^\top \text{VEC}(h - \hat{h}).$$

Using $\|\Phi v\| \geq \sqrt{\lambda_{\min}(\Phi^\top\Phi)}\,\|v\|$ on the l.h.s., and $P_\Phi v \leq \left\|\Phi^\top v\right\| / \sqrt{\lambda_{\min}(\Phi^\top\Phi)}$ on the r.h.s.,

$$\left\|(I - \Gamma\otimes\Gamma)^\top(h - \hat{h})\right\| \leq \frac{\left\|\Phi^\top(\overline{\Phi}_+ - \Phi_+)\hat{h}\right\|}{\lambda_{\min}(\Phi^\top\Phi)} \qquad (5)$$

Lemma 4.4 of Tu and Recht (2017) shows that for a sufficiently long trajectory, $\lambda_{\min}(\Phi^\top\Phi) = O\big(\tau\lambda^2_{\min}(\Sigma_\pi)\big)$. Lemma 4.8 of Tu and Recht (2017) (adapted to the average-cost setting with noise covariance W) shows that for a sufficiently long trajectory, with probability at least $1 - \delta$,

$$\left\|\Phi^\top(\overline{\Phi}_+ - \Phi_+)\hat{h}\right\| \leq$$
$$O\big(\sqrt{\tau}\,\text{tr}(\Sigma_V)\left\|W\widehat{H}\right\|_F \left\|\Gamma\Sigma_V^{1/2}\right\| \text{polylog}(\tau, 1/\delta)\big). \qquad (6)$$

Here, $\Sigma_V = \Sigma_0 + \Sigma_\pi$ is a simple upper bound on the state covariances Σ_t obtained as follows:

$$\begin{aligned}\Sigma_t &= \Gamma\Sigma_{t-1}\Gamma^\top + W \\ &= \Gamma^t\Sigma_0{\Gamma^t}^\top + \sum_{k=0}^{t-1}\Gamma^k W{\Gamma^k}^\top \\ &\prec \Sigma_0 + \sum_{k=0}^{\infty}\Gamma^k W{\Gamma^k}^\top = \Sigma_0 + \Sigma_\pi.\end{aligned}$$

In Appendix A.1, we show that $\left\|(I - \Gamma\otimes\Gamma)^\top(h - \hat{h})\right\|$ is lower-bounded by $\lambda_{\min}(M)^2\,\|H\|^{-2}\,\|H - \widehat{H}\|_F$. The result follows from applying the bounds to (5) and rearranging terms. □

While the error bound depends on $\widehat{H}$, for large τ we have

$$\|\widehat{H}\|_F - \|H\|_F \leq \|H - \widehat{H}\|_F = c\|\widehat{H}\|_F$$

for some $c < 1$. Therefore $\|\widehat{H}\|_F \leq (1-c)^{-1}\,\|H\|_F$.

4.2 State-action value function

Let $z^\top = (x^\top, a^\top)$, and let $\psi = \text{VEC}(zz^\top)$. The state-action value function corresponds to the cost of deviating from the policy, and satisfies

$$Q(x, a) = z^\top G z = c(x, a) + \mathbf{E}(x_+^\top H x_+|z) - \text{tr}(HW)$$
$$\psi^\top \text{VEC}(G) = c(x, a) + \big(\mathbf{E}(\phi_+|z) - \text{VEC}(W)\big)^\top \text{VEC}(H)\,.$$

We estimate G based on the above equation, using the value function estimate $\widehat{H}$ of the previous section in place of H and randomly sampled actions. Let Ψ be a $\tau \times (n+d)^2$

matrix whose rows are vectors $\psi_1, ..., \psi_\tau$, and let $\mathbf{c} = [c_1, ..., c_\tau]^\top$ be the vector of corresponding costs. Let Φ_+ be the $\tau \times n^2$ matrix containing the next-state features after each random action, and let $\overline{\Phi}_+$ be its expectation. We estimate G as follows:

$$\text{VEC}(\widehat{G}) = (\Psi^\top \Psi)^{-1} \Psi^\top (\mathbf{c} + (\Phi_+ - \mathbf{W})\hat{h}), \quad (7)$$

and additionally project the estimate onto the constraint $\widehat{G} \succeq \left(\begin{smallmatrix} M & 0 \\ 0 & N \end{smallmatrix}\right)$.

To gather appropriate data, we iteratively (1) execute the policy π for T_s iterations in order to get sufficiently close to the steady-state distribution,[2] (2) sample a random action $a \sim \mathcal{N}(0, \Sigma_a)$, (3) observe the cost c and next state x_+, and (4) add the tuple (x, a, x_+) to our dataset $\mathcal{Z}$. We collect $\tau = 0.5T^{1/2+\xi}$ such tuples in each phase of MFLQv2, and $\tau = T^{2/3+\xi}$ such tuples in the first phase of MFLQv1.

Lemma 4.2. *Let $\Sigma_{G,\pi} = A\Sigma_\pi A^\top + B\Sigma_a B^\top$. With probability at least $1 - \delta_1$, we have*

$$\left\| G - \widehat{G} \right\|_F = O\bigg(\text{tr}(\Sigma_{G,\pi}) \left\| H - \widehat{H} \right\| + \tau^{-1/2} \left\| W\widehat{H} \right\|_F \text{tr}(\Sigma_\pi + \Sigma_a) \left\| \Sigma_{G,\pi} \right\|_F \times \text{polylog}(n^2, \frac{1}{\delta_1}, \tau) \bigg). \quad (8)$$

The proof is given in Appendix A.2 and similar to that of Lemma 4.1. One difference is that we now require a lower bound on $\lambda_{\min}(\Psi^\top \Psi)$, where $\Psi^\top \Psi$ is a function of both states and actions. Since actions may lie in a subspace of $\mathcal{A}$ when following a linear policy, we estimate G using only the exploratory dataset $\mathcal{Z}$. Another difference is that we rely on $\widehat{H}$, so the error includes a $\|H - \widehat{H}\|$ term.

Let $\Sigma_{\max}$ be an upper bound on the state covariance matrices in the p.s.d. sense, and let C_H be an upper bound on $\|H_i\|$ (see Appendix C.3.1 and Lemma 5.1 for concrete values for these bounds). By a union bound, with probability at least $1 - S(\delta + \delta_1)$, the estimation error in any phase i of MFLQ with τ_v steps of value estimation and τ_q random actions is bounded as

$$\left\| G_i - \widehat{G}_i \right\|_F \leq \varepsilon_1 := C_h \tau_v^{-1/2} \,\text{tr}(\Sigma_{G,\max}) \frac{C_H^3 \left\| W \right\|_F \text{tr}(\Sigma_{\max})}{\lambda_{\min}(M)} \left\| \Gamma \Sigma_{\max}^{1/2} \right\| + C_g {\tau_q}^{-1/2} C_H \left\| W \right\|_F (\text{tr}(\Sigma_{\max}) + \text{tr}(\Sigma_a)) \left\| \Sigma_{G,\max} \right\|_F \quad (9)$$

for appropriate constants C_h and C_g.

[2]Note that stable linear systems mix exponentially fast; see Tu and Recht (2017) for details.

5 ANALYSIS OF MFLQ

In this section, we first show that given sufficiently small estimation errors, 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. We analyze MFLQv2; the analysis of MFLQv1 is similar and obtained by a different choice of constants.

By assumption, K_1 is bounded and stable. By the arguments in Section 4, the estimation error in the first phase can be made small for sufficiently long phases. In particular, we assume that estimation error in each phase is bounded by ε_1 as in Equation (9) and that ε_1 satisfies

$$\varepsilon_1 < \left(12C_1(\sqrt{n} + C_K\sqrt{d})^2 S \right)^{-1} . \quad (10)$$

Here $C_1 > 1$ is an upper bound on $\|H_1\|$, and $C_K = 2(3C_1 \|B\| \|A\| + 1)$. Since we take $S^2 T^\xi$ random actions in the first phase, the error factor S^{-1} is valid as long as $T > \overline{C}^{1/\xi}$ for a constant $\overline{C}$ that can be derived from (9) and (10). We prove the following lemma in Appendix B.1.

Lemma 5.1. *Let $\{K_i\}_{i=2}^S$ be the sequence of policies produced by the MFLQ algorithm. For all $i \in [S]$, $\|H_i\| \leq C_i < C_H := 3C_1$, K_i is stable, $\|K_i\| \leq C_K$, and for all $k \in \mathbb{N}$,*

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

To prove the lemma, we first show that the value matrices H_j and closed-loop matrices $\Gamma_j = A - BK_j$ satisfy

$$x^\top \left(\Gamma_{j+1}^\top H_j \Gamma_{j+1} \right) x + \varepsilon_2 \leq x^\top H_j x \quad (11)$$

where

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

$$X_K = \begin{bmatrix} I \\ -K \end{bmatrix} x x^\top \begin{bmatrix} I & -K^\top \end{bmatrix} ,$$

and $|X_K|$ is the matrix obtained from X_K by taking the absolute value of each entry. If the estimation error ε_1 is small enough so that $\varepsilon_2 > 0$ for any unit-norm x and all policies, then $H_j \succ \Gamma_{j+1}^\top H_j \Gamma_{j+1}$ and K_{j+1} is stable by a Lyapunov theorem. Since K_1 is stable and H_1 bounded, all policies remain stable. If estimation errors are bounded as in (10), we can show that the policies and value matrices are bounded as in Lemma 5.1.

Let $\mathcal{E}$ denote the event that all errors are bounded as in Equation (10). We bound the terms α_T, β_T, and γ_T for MFLQv2 next.

Lemma 5.2. *Under the event $\mathcal{E}$, for appropriate constants C', D', and D'',*

$$\beta_T \leq C' T^{3/4} \log(T/\delta),\ \alpha_T \leq D' T^{3/4+\xi},\ \gamma_T \leq D'' T^{1/2} .$$

The proof is given in Appendix C. For β_T, we rely on the decomposition shown in Equation (1). Because we execute S policies, each for $\tau = T/S$ rounds (where $S = T^{1/3-\xi}$ and $\tau = T^{1/4}$ for MFLQv1 and MFLQv2, respectively),

$$\begin{aligned}\beta_T &= \sum_{i=1}^{S} \tau \mathbf{E}(Q_i(x, \pi_i(x)) - Q_i(x, \pi(x))) \\ &= \tau \sum_{i=1}^{S} \Big(\mathbf{E}(\widehat{Q}_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi(x))) \\ &\qquad + \mathbf{E}(Q_i(x, \pi_i(x)) - \widehat{Q}_i(x, \pi_i(x))) \\ &\qquad + \mathbf{E}(\widehat{Q}_i(x, \pi(x)) - Q_i(x, \pi(x))) \Big)\end{aligned}$$

where the expectations are w.r.t. $x \sim \mu_\pi$. We bound the first term using the FTL regret bound of (Cesa-Bianchi and Lugosi, 2006) (Theorem 3.1), by showing that the theorem conditions hold for the loss function $f_i(K) = \mathbf{E}_{x\sim\mu_\pi}(\widehat{Q}_i(x, Kx))$. We bound the second and third term (corresponding to estimation errors) using Lemma 4.2. This results in the following bound on β_T for constants C' and C'':

$$\beta_T \leq T/SC'(1 + \log S) + C'\sqrt{ST} \log T .$$

To bound $\gamma_T = \sum_{t=1}^{T} \lambda_\pi - c(x_t, \pi(x_t))$, we first decompose the cost terms as follows. Let Σ_π be the steady-state covariance, and let Σ_t be the covariance of x_t. Let $D_t = \Sigma_t^{1/2}(M + K^\top NK)\Sigma_t^{1/2}$ and $\lambda_t = \mathrm{tr}(D_t)$. We have

$$\begin{aligned}\lambda_\pi - c(x_t, \pi(x_t)) &= \lambda_\pi - \lambda_t + \lambda_t - c(x_t^\pi, \pi(x_t^\pi)) \\ &= \mathrm{tr}((\Sigma_\pi - \Sigma_t)(M + K^\top NK)) \\ &\quad + \mathrm{tr}(D_t) - u_t^\top D_t u_t\end{aligned}$$

where $u_t \sim \mathcal{N}(0, I_n)$. We show that the second term $\sum_{t=1}^{T} \mathrm{tr}(D_t) - u_t^\top D_t u_t$ scales as $\sqrt{T}$ with high probability using the Hanson-Wright inequality. The first term can be bounded by $\mathrm{tr}(H_\pi)\,\mathrm{tr}(\Sigma_\pi)$ as follows. Note that

$$\Sigma_\pi - \Sigma_t = \Gamma(\Sigma_\pi - \Sigma_{t-1})\Gamma^\top = \Gamma^t(\Sigma_\pi - \Sigma_0)(\Gamma^t)^\top .$$

Hence we have

$$\begin{aligned}&\sum_{t=0}^{T} \mathrm{tr}((M + K^\top NK)(\Sigma_\pi - \Sigma_t)) \\ &\quad = \sum_{t=0}^{T} \mathrm{tr}\left((M + K^\top NK)\Gamma^t(\Sigma_\pi - \Sigma_0)(\Gamma^t)^\top\right) \\ &\quad \leq \mathrm{tr}(\Sigma_\pi - \Sigma_0)\,\mathrm{tr}\left(\sum_{t=0}^{\infty}(\Gamma^t)^\top(M + K^\top NK)\Gamma^t\right) \\ &\quad = \mathrm{tr}(\Sigma_\pi - \Sigma_0)\,\mathrm{tr}(H_\pi) .\end{aligned} \tag{12}$$

The bound on $\alpha_T = \sum_{t=1}^{T} c(x_t, a_t) - \lambda_{\pi_t}$ is similar; however, in addition to bounding the cost of following a policy, we need to account for the cost of random actions, and the changes in state covariance due to random actions and policy switches.

Theorem 3.2 is a consequence of Lemma 5.2. The proof of Theorem 3.1 is similar and is obtained by different choice of constants.

6 EXPERIMENTS

We evaluate our algorithm on two LQ problem instances: (1) the system studied in Dean et al. (2017) and Tu and Recht (2017), and (2) the power system studied in Lewis et al. (2012), Example 11.5-1, with noise $W = I$. We start all experiments from an all-zero initial state $x_0 = 0$, and set the initial stable policy K_1 to the optimal controller for a system with a modified cost $M' = 200M$. For simplicity we set $\xi = 0$ and $T_s = 10$ for MFLQv1. We set the exploration covariance to $\Sigma_a = I$ for (1) and $\Sigma_a = 10I$ for (2).

In addition to the described algorithms, we also evaluate MFLQv3, an algorithm identical to MFLQv2 except that the generated datasets $\mathcal{Z}$ include all data, not just random actions. 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 $\widehat{Q}_{i-1}$. We use the same estimation procedure as for MFLQ.
- 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 μ_Q and covariance Σ_Q of a TD estimate of G after each step, and switch to a policy greedy w.r.t. a parameter sample $\widehat{G} \sim (\mu_Q, 0.2\Sigma_Q)$ every $T^{1/2}$ steps. We project the sample onto the constraint $G \succ \left(\begin{smallmatrix} M & 0 \\ 0 & N \end{smallmatrix}\right)$.
- 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 (this approach is called *certainty equivalence* in optimal control). We use the same strategy as in the model-free case, i.e. we execute the policy for some number of iterations, followed by running the policy and taking random actions.

To evaluate stability, we run each algorithm 100 times and compute the fraction of times it produces stable policies in

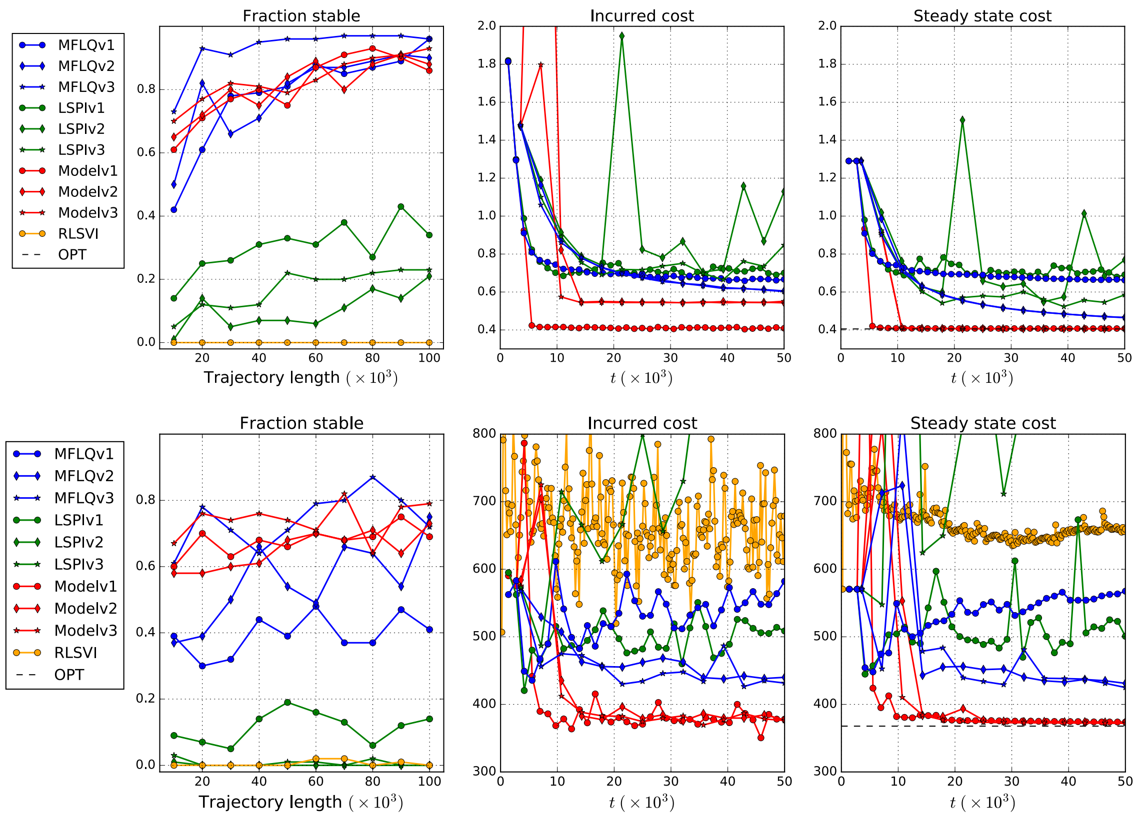

Figure 1: Top row: experimental evaluation on the dynamics of Dean et al. (2017). Bottom row: experimental evaluation on Lewis et al. (2012), Example 11.5.1.

all phases. Figure 1 (left) shows the results as a function of trajectory length. MFLQV3 is the most stable among model-free algorithms, with performance comparable to the model-based approach.

We evaluate solution cost by running each algorithm until we obtain 100 stable trajectories (if possible), where each trajectory is of length 50,000. We compute both the average cost incurred during each phase i, and true expected cost of each policy π_i. The average cost at the end of each phase is shown in Figure 1 (center and right). Overall, MFLQV2 and MFLQV3 achieve lower costs than MFLQV1, and the performance of MFLQV1 and LSPI is comparable. The lowest cost is achieve by the model-based approach. These results are consistent with the empirical findings of Tu and Recht (2017), where model-based approaches outperform discounted LSTDQ.

7 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 $O(T^{2/3+\xi})$ 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 exploration similar to ϵ-greedy (but performed at a fixed schedule). 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.

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.

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.

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.

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.

Fuzhen Zhang and Qingling Zhang. Eigenvalue inequalities for matrix product. *IEEE Transactions on Automatic Control*, 51(9):1506–1509, 2006.