Regret Bounds for the Adaptive Control of Linear Quadratic Systems

Yasin Abbasi-Yadkori
abbasiya@cs.ualberta.ca
Department of Computing Science
University of Alberta

Csaba Szepesvári
szepesva@cs.ualberta.ca
Department of Computing Science
University of Alberta

Abstract

We study the average cost Linear Quadratic (LQ) problem with unknown model parameters, also known as the adaptive control problem in the control community. We design an algorithm and prove that its regret up to time T is $O(\sqrt{T})$ apart from logarithmic factors. Unlike many classical approaches that use a forced-exploration scheme to provide the sufficient exploratory information for parameter estimation, we construct a high-probability confidence set around the model parameters and design an algorithms that plays optimistically with respect to this confidence set. The construction of the confidence set is based on the new results from online least-squares estimation and leads to improved worst-case regret bound for the proposed algorithm. To best of our knowledge this is the the first time that a regret bound is derived for the LQ problem.

1 Introduction

The Linear Quadratic problem is probably the most widely studied problem in the control literature. The problem is to minimize the average cost of a controller that operates in an environment whose dynamics is linear, while the costs are a quadratic function of the state and the control. The optimal solution is a linear feedback controller which can be explicitly computed from the matrices describing the dynamics and the cost. In the learning problem, the topic of this paper, the dynamics of the environment is unknown. The problem then becomes more challenging since the control actions influence both the cost and the rate at which the dynamics is learned, resulting in a so-called learning, or *adaptive* controller. The objective in this case is to minimize the *regret* of the controller, i.e. to minimize the difference between the average cost of the learning controller and that of the optimal controller. In this paper, for the first time, we show an adaptive controller and we prove that the expected regret of this controller is bounded by $O(\sqrt{T})$. We build on recent works in online linear estimation and also in the adaptive controller literature, the latter of which we survey next.

When the model parameters are known and the state is fully observed, one can use the principles of dynamic programming to obtain the optimal controller, which is known as the Linear Quadratic Regulator (LQR). As mentioned before, the version of the problem that deals with the unknown model parameters is called the adaptive control problem. The early attempts to solve this problem relied on the *certainty equivalence principle*. The idea was to estimate the unknown parameters from observations and then use the estimated parameters as if they are the true parameters to design a LQR. It was soon realized that the certainty equivalence principle does not necessarily provide enough information to reliably estimate the parameters and the estimated parameters can converge to incorrect values with positive probability. This in turn might lead to a suboptimal performance.

In order to avoid this identifiably problem, researchers developed methods that actively explore the environment to gather information (Lai and Wei, 1982, 1987, Chen and Guo, 1987, Chen and Zhang, 1990, Fiechter, 1997, Lai and Ying, 2006, Bittanti and Campi, 2006). However, only asymptotic results are proven for these methods. One exception is the work of Fiechter (1997) that proposes an algorithm for the "discounted" LQ problem and analyzes its performance in a PAC framework.

Another problem with most of the aforementioned methods is that they use forced-exploration schemes to provide the sufficient exploratory information. The idea is to take exploratory actions with a fixed and appropriately designed rate. However, the forced-exploration schemes lack strong worst case regret bounds, even in the simplest problems (Langford and Zhang, 2007). Unlike the preceding methods, Bittanti and Campi (2006) proposes an algorithm that uses the Optimism in the Face of Uncertainty (OFU) principle, which goes back to the work of Lai and Robbins (1985), to deal with the exploration/exploitation dilemma. The idea behind the OFU principle, applied to the adaptive control problem, is to construct high-probability confidence sets around the model parameters, find the optimal controller for each member of the confidence set, and finally choose the controller that gives the smallest cost among these controllers. However, Bittanti and Campi (2006) only show that the average cost of the algorithm converges to that of the optimal policy in the limit. We extend their work to derive a finite time regret bound for the LQ problem. Our proof is based on the proof Bittanti and Campi (2006), although with significant differences, due to the difference in the goals of the papers.

Note that the OFU principle has been applied very successfully to a number of challenging learning and control situations. Lai and Robbins (1985), who invented the principle, used it to address learning in bandit problems (i.e., when there is no state) and later this work was picked up and modified by Auer et al. (2002) to make it work in nonparametric bandits. The OFU principle has also been applied to learning in *finite* Markov Decision Processes, both in a regret minimization (see Bartlett and Tewari 2009, Auer et al. 2010 and the references therein) and in a PAC-learning setting (see (Kearns and Singh, 1998, Brafman and Tennenholtz, 2002, Kakade, 2003, Strehl et al., 2006, Szita and Szepesvári, 2010) and the refererences therein). In the PAC-MDP framework there has been some work to extend the OFU principle to infinite Markov Decision Problems under various assumptions. For example, Lipschitz assumptions have been used by Kakade et al. (2003), while Strehl and Littman (2008) explored linear models. However, these works do not consider both continuous state and action spaces. (Continuous action spaces in the context of bandits have been explored in a number of works, such as the works of Kleinberg (2004), Auer et al. (2007), Kleinberg et al. (2008) and in a linear setting by Auer (2003), Dani et al. (2008) and Rusmevichientong and Tsitsiklis (2010).)

As far as we know the regret criterion has not been considered previously in this literature in the continuous state (and control) setting. Although the class of systems considered here in this paper is simple, we think that this class is interesting enough on its own due to its numerous practical applications and, in fact, its simplicity. In fact, we find it remarkable that no regret bounds have been available for this simple setting so far, though we explain this that until the recent works of Dani et al. (2008) and Rusmevichientong and Tsitsiklis (2010), the understanding of linear estimation with dependent covariates was not very well developed. In fact, our work also builds upon on these works, although we use a more recent, improved confidence bound (see Theorem 1).

Further, we think that our work might have implications beyond the problem considered here, as it clearly demonstrates that the OFU principle can be successfully applied even to continuous space and action control problems.

2 Notation and conventions

We use $\|\cdot\|$ to denote the 2-norm. For a positive definite matrix $A \in \mathbb{R}^{d\times d}$, the weighted 2-norm is defined by $\|x\|_A^2 = x^\top Ax$, where $x \in \mathbb{R}^d$. The inner product is denoted by $\langle \cdot, \cdot \rangle$ and the weighted inner-product $x^\top Ay = \langle x, y\rangle_A$. We use $\lambda_{\min}(A)$ to denote the minimum eigenvalue of the positive definite matrix A. We use $A \succ 0$ to denote that A is positive definite, while we use $A \succeq 0$ to denote that it is positive semidefinite. The same notation is used to denote the Loewner partial order of matrices. We shall use $\mathbf{e}_i$ to denote the i^{th} unit vector, i.e., for all $j \neq i$, $\mathbf{e}_{ij} = 0$ and $\mathbf{e}_{ii} = 1$.

3 The Linear Quadratic Problem

We consider the discrete-time infinite-horizon linear quadratic (LQ) control problem:

$$\begin{aligned} x_{t+1} &= A_* x_t + B_* u_t + w_{t+1} \\ c_t &= x_t^\top Q x_t + u_t^\top R u_t, \end{aligned}$$

where $u_t \in \mathbb{R}^d$ is the control at time t, $x_t \in \mathbb{R}^n$ is the state at time t, $A_* \in \mathbb{R}^{n\times n}$ and $B_* \in \mathbb{R}^{n\times d}$ are unknown matrices and $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$ are known (positive definite) matrices. At time zero, for simplicity, $x_0 = 0$. Let

$$\Theta_*^\top = (A_*\,, \quad B_*) \qquad \text{and} \qquad z_t = \begin{pmatrix} x_t \\ u_t \end{pmatrix} .$$

Thus, the state transition can be written as

$$x_{t+1} = \Theta_*^\top z_t + w_{t+1}\,.$$

The objective is to minize the average cost

$$J(u_0, u_1, \dots) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}\left[c_t\right]. \tag{1}$$

Our assumption on the data is as follows:

Assumption A1 Let $(\mathcal{F}_t)$ be a filtration such that for the random variables (z_0, x_1), ..., (z_t, x_{t+1}) the following hold:

(i) $(z_t, x_{t+1}) \in \mathbb{R}^{n+d} \times \mathbb{R}^n$;

(ii) z_t, x_t are $\mathcal{F}_t$-measurable;

(iii) for $\Theta_* \in \mathbb{R}^{(n+d)\times n}$, for any $t \geq 0$,

$$\mathbb{E}\left[x_{t+1}|\mathcal{F}_t\right] = z_t^\top \Theta_*\,,$$

i.e., $w_{t+1} = x_{t+1} - z_t^\top \theta_*$ is a martingale difference sequence ($\mathbb{E}\left[w_{t+1}|\mathcal{F}_t\right] = 0$, $t = 0, 1, \dots$);

(iv) $\mathbb{E}\left[w_{t+1}w_{t+1}^\top \,|\, \mathcal{F}_t\right] = I_n$;

(v) The random variables w_t are component-wise sub-Gaussian in the sense that there exists a known $R > 0$ such that for any $\gamma \in \mathbb{R}$, and index j,

$$\mathbb{E}\left[\exp(\gamma w_{t+1,j})|\mathcal{F}_t\right] \leq \exp(\gamma^2 R^2/2)\,.$$

Our assumptions on the system uncertainty under which we will prove our regret bounds is as follows:

Assumption A2 The unknown parameter is such that Θ_*

$$\operatorname{trace}(\Theta_*^\top \Theta_*)^{1/2} \leq S \tag{2}$$

with $S > 0$ known. Further, for all $\Theta = (A, B) \in \mathcal{S}$, (A, B) is reachable and $(A, Q^{1/2})$ is observable.

We will denote the set of parameters Θ_* such that (2) holds by $\mathcal{S}$:

$$\mathcal{S} = \{\Theta_* \,:\, \operatorname{trace}(\Theta_*^\top \Theta_*)^{1/2} \leq S\,\}\,.$$

In what follows we shall always assume that the above two assumptions are valid.

The assumption $\mathbb{E}\left[w_{t+1}w_{t+1}^\top \,|\, \mathcal{F}_t\right] = I_n$ makes the analysis clean and simple. However, we shall show it later that it is in fact not necessary. If the second assumption can be removed (or relaxed) is left for future work. We think that this might be possible by using a dove-tailing technique.

3.1 Parameter estimation

Define

$$e(\Theta) = \lambda \operatorname{trace}(\Theta^\top \Theta) + \sum_{s=0}^{t-1} \operatorname{trace}((x_{s+1} - \Theta^\top z_s)(x_{s+1} - \Theta^\top z_s)^\top).$$

Let $\hat{\Theta}_t$ be the ℓ^2-regularized least-squares estimate of Θ_* with regularization parameter $\lambda > 0$:

$$\hat{\Theta}_t = \operatorname*{argmin}_{\Theta} \; e(\Theta) = (Z^\top Z + \lambda I)^{-1} Z^\top X, \tag{3}$$

where Z and X are the matrices whose rows are $z_0^\top, \ldots, z_{t-1}^\top$ and $x_1^\top, \ldots, x_t^\top$, respectively. With an argument similar to the one used in Abbasi-Yadkori et al. (2011), we can construct a high-probability confidence set around Θ_*:

Theorem 1. *Let $(z_0, x_1), \ldots, (z_t, x_{t+1})$, $z_i \in \mathbb{R}^{n+d}$, $x_i \in \mathbb{R}^n$ satisfy the linear model Assumption A1 with some $R > 0$, $\Theta_* \in \mathbb{R}^{(n+d)\times n}$ and let $(\mathcal{F}_t)$ be the associated filtration. Consider the ℓ^2-regularized least-squares parameter estimate $\hat{\Theta}_t$ with regularization coefficient $\lambda > 0$ (cf. (3)). Let $V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^\top$ be the regularized design matrix underlying the covariates. Then, for any $0 < \delta < 1$, any stopping time $\tau \geq 1$ w.r.t. $(\mathcal{F}_t)$, with probability at least $1 - \delta$,*

$$\operatorname{trace}((\hat{\Theta}_\tau - \Theta_*)^\top V_\tau (\hat{\Theta}_\tau - \Theta_*)) \leq \left(nR \sqrt{2 \log \left(\frac{\det(V_\tau)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right)^2 . \tag{4}$$

Also, when the covariates satisfy $\|z_t\| \leq c_m$, $t \geq 0$ with some $c_m > 0$ w.p.1 then, with probability at least $1 - \delta$,

$$\operatorname{trace}((\hat{\Theta}_\tau - \Theta_*)^\top V_\tau (\hat{\Theta}_\tau - \Theta_*)) \leq \left(nR \sqrt{(n+d) \log \left(\frac{1 + \frac{\tau c_m}{\lambda}}{\delta} \right)} + \lambda^{1/2} S \right)^2 .$$

In what follows we shall denote the left-hand side of (4) by $\beta_\tau(\delta)$:

$$\beta_\tau(\delta) = \left(nR \sqrt{2 \log \left(\frac{\det(V_\tau)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right)^2 .$$

We see from this theorem that the confidence about Θ_* depends on the behavior of the ratio of two determinants. The following two technical lemmas will be useful in dealing with these terms. The lemmas are taken from Abbasi-Yadkori et al. (2011).

Lemma 2. *The following holds for any $t \geq 1$:*

$$\sum_{k=0}^{t-1} \left(\|z_k\|^2_{V_k^{-1}} \cap 1 \right) \leq 2 \log \frac{\det(V_t)}{\det(\lambda I)} .$$

Further, when the covariates satisfy $\|z_t\| \leq c_m$, $t \geq 0$ with some $c_m > 0$ w.p.1 then

$$2 \log \frac{\det(V_t)}{\det(\lambda I)} \leq 2(n+d) \log \left(\frac{\lambda(n+d) + t c_m^2}{\lambda(n+d)} \right) .$$

Lemma 3. *Let A, B and C be positive semi-definite matrices such that $A = B + C$. Then, we have that*

$$\sup_{X \neq 0} \frac{\|X^\top A X\|}{\|X^\top B X\|} \leq \frac{\det(A)}{\det(B)}.$$

3.2 The controller

The objective (1) for the system with parameters $\Theta = (A, B)$ can also be written as (Chen and Guo, 1987)

$$J(u_0, u_1, \ldots) = \operatorname{trace}(P(\Theta)) + \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} (u_t - K(\Theta) x_t)^\top (R + B^\top P(\Theta) B)(u_t - K(\Theta) x_t), \tag{5}$$

```
**Inputs**: $T, S > 0, \delta > 0, Q, R$.
Set $V_0 = I$ and $\hat{\Theta}_0 = 0$.
$(\tilde{A}_0, \tilde{B}_0) = \tilde{\Theta}_0 = \operatorname{argmin}_{\Theta \in \mathcal{C}_0} J(\Theta)$.
**for** $t := 0, 1, 2, \ldots$ **do**
  **if** $\det(V_t) > 2\det(V_0)$ **then**
    Calculate $\hat{\Theta}_t$ by (3).
    $\tilde{\Theta}_t = \operatorname{argmin}_{\Theta \in \mathcal{C}_t} J(\Theta)$.
    Let $V_0 = V_t$.
  **else**
    $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$.
  **end if**
  Calculate $u_t$ based on the current parameters, $u_t = K(\tilde{\Theta}_t)x_t$.
  Execute control, observe new state $x_{t+1}$.
  Save $(z_t, x_{t+1})$ into the dataset, where $z_t^\top = (x_t^\top, u_t^\top)$.
  $V_{t+1} := V_t + z_t z_t^\top$.
**end for**
```

Table 1: The proposed adaptive algorithm for the LQR problem

where $P(\Theta)$ is the unique solution to the *Ricatti equation*

$$P(\Theta) = Q + A^\top P(\Theta)A - A^\top P(\Theta)B(B^\top P(\Theta)B + R)^{-1}B^\top P(\Theta)A$$

and the *gain matrix* $K(\Theta)$ is defined by

$$K(\Theta) = -(B^\top P(\Theta)B + R)^{-1}B^\top P(\Theta)A.$$

Let $J(\Theta)$ denote the optimal average cost if the model parameters were Θ. From (5), it is clear that $J(\Theta) = \operatorname{trace}(P(\Theta))$. Thus, as it is well known (e.g., Bertsekas, 2001) the *optimal control law* for a system with parameters Θ is

$$u_t = K(\Theta)x_t\,. \tag{6}$$

In particular, the average cost of this control law with $\Theta = \Theta_*$ is the optimal average cost $J_* = J(\Theta_*) = \operatorname{trace}(P(\Theta_*))$. The *regret* up to time T of a controller which suffers a cost of c_t at time t is defined by

$$R(T) = \sum_{t=0}^{T}(c_t - J_*)\,.$$

Minimizing the regret is equivalent to minimizing $J(u_0, u_1, \ldots)$ and the regret measures the cost of not knowing the system dynamics.

From Theorem 1, we can obtain the following $(1-\delta)$-probability confidence set around Θ_*:

$$\mathcal{C}_t(\delta) = \left\{\Theta \in \mathcal{S} : \operatorname{trace}\left\{(\Theta - \hat{\Theta}_t)^\top V_t(\Theta - \hat{\Theta}_t)\right\} \le \beta_t(\delta)\right\}. \tag{7}$$

The algorithm that we propose implements the OFU principle at follows: At time t, the algorithm chooses a parameter Θ from $\mathcal{C}_t(\delta)$ and then uses the optimal feedback controller (6) underlying the chosen parameter. The actual algorithm that we propose differs from this controller in that it changes controllers only after the current parameter estimates are significantly refined. This prevent too frequent changes to the controller (which might harm performance) and it also saves computation. The pseudocode of our proposed algorithm is shown as Algorithm 1.

4 Analysis

In this section we give our main result and the proof of this result, where we bound the regret of Algorithm 1 with high probability.

However, first we need some additional preparation. In particular, in addition to the assumption we made before, we shall also assume that the following assumption holds true in the rest of the article:

Assumption A3 Let $\rho = \sup_{\Theta \in \mathcal{S}} \|A + BK(\Theta)\|$. Then$\rho < 1$. Further, there exists a positive ϵ such that $\lambda_d(R) > \epsilon$, or equivalently, there exist $C > 0$ such that $\sup_{\Theta \in \mathcal{S}} \|K(\Theta)\| < C$.

Note that this assumption was used before by Bittanti and Campi (2006) and other works which considered asymptotic consistency.By the boundedness of $\mathcal{S}$ from this assumption, we also obtain the boundedness of $P(\Theta)$. The corresponding constant will be denoted by D:

$$\Theta \in \mathcal{S} \quad \Longrightarrow \quad \|P(\Theta)\| \leq D\,.$$

We will choose two error probabilities, $\delta_E > 0$ and $\delta_F > 0$. Given these, we define two "good events" in the probability space Ω. In particular, we define the event that the confidence sets hold for $s = 0, \dots, t$,

$$E(t, \delta_E) = \{\omega \in \Omega \,:\, \forall s \leq t, \quad \Theta_* \in \mathcal{C}_s(\delta_E)\,\}\ ,$$

and the event that the state vector stay "small":

$$F(t, \delta_F, \delta_E) = \{\omega \in \Omega \,:\, \forall s \leq t, \quad \|x_s\| \leq \alpha_T(\delta_F, \delta_E)\,\}$$

where

$$\alpha_t(\delta_F, \delta_E) = \frac{1}{\sqrt{1-\rho^2}}\,(8(n+d)\beta_t(\delta_E)\log(\det(V_t)))^{1/2} + \frac{R}{1-\rho}\,\sqrt{2n\log(nt/\delta_F)}\,.$$

In what follows, we let $E = E(T, \delta_E)$ and $F = F(T, \delta_F)$, where $\delta_E > 0$, $\delta_F > 0$ will be chosen later.

Our main result is the following theorem:

Theorem 4. *With probability at least $1 - \delta$, the regret of Algorithm 1 is bounded as follows:*

$$R(T) = \tilde{O}\left(\sqrt{T\log(1/\delta)}\right)\ ,$$

where the constant hidden is a problem dependent constant.[1]

Remark 5. The assumption $\mathbb{E}\left[w_{t+1}w_{t+1}^\top | \mathcal{F}_t\right] = I_n$ makes the analysis clean and simple. Alternatively, we could assume that $\mathbb{E}\left[w_{t+1}w_{t+1}^\top | \mathcal{F}_t\right] = G_*$ and G_* be unknown. Then the optimal average cost becomes $J(\Theta_*, G_*) = \mathrm{trace}(P(\Theta_*)G_*)$. The only change in Algorithm 1 is in the computation of $\tilde{\Theta}_t$, which will have the following form:

$$(\tilde{\Theta}_t, \tilde{G}) = \operatorname*{argmin}_{(\Theta, G) \in \mathcal{C}_t} J(\Theta),$$

where $\mathcal{C}_t$ is now a confidence set over Θ_* and G_*. The rest of the analysis remains identical, provided that an appropriate confidence set is constructed.

4.1 Proof

From average cost dynamic programming (Bertsekas, 1987)[Volume 2, pages 228–229] we have that

$$\begin{aligned}
J(\tilde{\Theta}_t) + x_t^\top P(\tilde{\Theta}_t)x_t &= \min_u\{x_t^\top Q x_t + u^\top R u + \mathbb{E}\left[\tilde{x}_{t+1}^{uT} P(\tilde{\Theta}_t)\tilde{x}_{t+1}^{u} | \mathcal{F}_t\right]\} \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[\tilde{x}_{t+1}^{u_t T} P(\tilde{\Theta}_t)\tilde{x}_{t+1}^{u_t} | \mathcal{F}_t\right],
\end{aligned}$$

where $\tilde{x}_{t+1}^u = \tilde{A}_t x_t + \tilde{B}_t u + w_{t+1}$. Hence,

$$\begin{aligned}
J(\tilde{\Theta}_t) + x_t^\top P(\tilde{\Theta}_t)x_t &= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[(\tilde{A}_t x_t + \tilde{B}_t u_t + w_{t+1})^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t + w_{t+1}) | \mathcal{F}_t\right] \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[(\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) | \mathcal{F}_t\right] \\
&+ \mathbb{E}\left[w_{t+1}^\top P(\tilde{\Theta}_t) w_{t+1} | \mathcal{F}_t\right] \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[(\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) | \mathcal{F}_t\right] \\
&+ \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} | \mathcal{F}_t\right] - \mathbb{E}\left[(A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t) | \mathcal{F}_t\right] \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} | \mathcal{F}_t\right] \\
&+ (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) - (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t),
\end{aligned}$$

[1] Here, $\tilde{O}$ hides logarithmic factors.

where in the equality before the last one we have used $x_{t+1} = A_* x_t + B_* u_t + w_{t+1}$ and the fact that $\mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} | \mathcal{F}_t\right] = \mathbb{E}\left[(A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t) | \mathcal{F}_t\right] + \mathbb{E}\left[w_{t+1}^\top P(\tilde{\Theta}_t) w_{t+1} | \mathcal{F}_t\right]$. Hence,

$$\sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 = \sum_{t=0}^{T} \left(x_t^\top Q x_t + u_t^\top R u_t\right) + R_2 + R_3,$$

where

$$R_1 = \sum_{t=0}^{T} (x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} | \mathcal{F}_t\right]) \tag{8}$$

and

$$R_2 = \sum_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\top (P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})) x_{t+1} | \mathcal{F}_t\right] \tag{9}$$

and

$$R_3 = \sum_{t=0}^{T} \left((\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) - (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t)\right). \tag{10}$$

Thus,

$$\begin{aligned} \sum_{t=0}^{T} (x_t^\top Q x_t + u_t^\top R u_t) &= \sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 - R_2 - R_3 \\ &\le T J(\Theta_*) + R_1 - R_2 - R_3. \end{aligned}$$

Thus,

$$R(T) \le R_1 - R_2 - R_3 \le \mathbb{I}_{\{E \cap F\}}(R_1 - R_2 - R_3) + \mathbb{I}_{\{\overline{E} \cup \overline{F}\}}(R_1 - R_2 - R_3). \tag{11}$$

First we prove the following lemmas that will be used in the proof of Theorem 4.

Lemma 6. *Consider Algorithm 1. Assume that $E(T, \delta_E)$ holds. Then we have*

$$\sum_{t=0}^{T} \left\|(\Theta_* - \tilde{\Theta}_t)^\top z_t\right\|^2 \le 8(n+d)\beta_T(\delta_E) \log(\det(V_T)).$$

Proof. Consider timestep t. Let $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$. Let $\tau \le t$ be the last timestep when we have changed the policy. So $s_t = (\Theta_* - \tilde{\Theta}_\tau)^\top z_t$. We have

$$\|s_t\| \le \left\|(\Theta_* - \hat{\Theta}_\tau)^\top z_t\right\| + \left\|(\hat{\Theta}_\tau - \tilde{\Theta}_\tau)^\top z_t\right\|.$$

For all $\Theta \in \mathcal{C}_\tau$,

$$\begin{aligned} \left\|(\Theta - \hat{\Theta}_\tau)^\top z_t\right\| &\le \left\|V_t^{1/2}(\Theta - \hat{\Theta}_\tau)\right\| \|z_t\|_{V_t^{-1}} \\ &\le \left\|V_\tau^{1/2}(\Theta - \hat{\Theta}_\tau)\right\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} \\ &\le \sqrt{2} \left\|V_\tau^{1/2}(\Theta - \hat{\Theta}_\tau)\right\| \|z_t\|_{V_t^{-1}} \\ &\le \sqrt{2\beta_\tau(\delta_E)} \|z_t\|_{V_t^{-1}}, \end{aligned}$$

where the first step follows from Cauchy-Schwartz inequality, the second step follows from Lemma 3, the third step follows from the fact that at iteration t we have $\det(V_t) < 2\det(V_\tau)$, and the last step follows from the definition of $\beta_\tau(\delta_E)$ and the fact that $\lambda_{\max}(M) \le \mathrm{trace}(M)$ for $M \succeq 0$. Thus,

$$\|s_t\|^2 \le 8\beta_\tau(\delta_E) \|z_t\|_{V_t^{-1}}.$$

Now, we have that

$$\|z_t\|_{V_t^{-1}}^2 = z_t^\top (V_{t-1} + z_t z_t^\top)^{-1} z_t \le z_t^\top (\lambda I + z_t z_t^\top)^{-1} z_t.$$

We also have that

$$\begin{aligned} z_t^\top(\lambda I + z_t z_t^\top)^{-1} z_t &= \operatorname{trace}(z_t^\top(\lambda I + z_t z_t^\top)^{-1} z_t) \\ &= \operatorname{trace}((\lambda I + z_t z_t^\top)^{-1}(z_t z_t^\top + \lambda I - \lambda I)) \\ &= \operatorname{trace}(I - \lambda(\lambda I + z_t z_t^\top)^{-1}) \\ &= n + d - \lambda \operatorname{trace}((\lambda I + z_t z_t^\top)^{-1}) \le n + d. \end{aligned}$$

Thus, $\|z_t\|^2_{V_t^{-1}} \le n + d$ and, also, $\|z_t\|^2_{V_t^{-1}} \le \min\{\|z_t\|^2_{V_t^{-1}}, n + d\}$. It follows then that

$$\|z_t\|^2_{V_t^{-1}} \le (n+d)\min\{\|z_t\|^2_{V_t^{-1}}, 1\}.$$

As a result,

$$\begin{aligned} \sum_{t=0}^{T} \|s_t\|^2 &\le 8(n+d)\beta_T(\delta_E)\sum_{t=0}^{T}\min\{\|z_t\|^2_{V_t^{-1}}, 1\} \\ &\le 8(n+d)\beta_T(\delta_E)\log(\det(V_T)). \end{aligned}$$

□

Lemma 7. $\mathbb{P}(E \cap F) \ge 1 - (\delta_E + \delta_F)$.

Proof. Define

$$\begin{aligned} s_t &= (A_* - \tilde{A}_t)x_t + (B_* - \tilde{B}_t)u_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t, \\ r_t &= s_t + w_{t+1}, \\ \rho_t &= \tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t), \\ c_k &= \prod_{s=0}^{k-1} \rho_{t-s-1}. \end{aligned}$$

Therefore, we can write

$$x_t = \rho_{t-1}x_{t-1} + r_{t-1} = \rho_{t-1}\rho_{t-2}x_{t-2} + r_{t-1} + \rho_{t-1}r_{t-2} = \dots = \sum_{k=0}^{t-1} r_{t-k-1}\prod_{s=0}^{k-1}\rho_{t-s-1} = \sum_{k=0}^{t-1} c_k r_{t-k-1}.$$

Thus,

$$\|x_t\| \le \left\|\sum_{k=0}^{t-1} c_k s_{t-k-1}\right\| + \left\|\sum_{k=0}^{t-1} c_k w_{t-k}\right\|. \tag{12}$$

The first term on the right hand side can be bounded as follows:

$$\begin{aligned} \left\|\sum_{k=0}^{t-1} c_k s_{t-k-1}\right\| &\le \sum_{k=0}^{t-1} \rho^k \|s_{t-k-1}\| \\ &\le \frac{1}{\sqrt{1-\rho^2}}\left(\sum_{k=0}^{t-1}\|s_k\|^2\right)^{1/2} \\ &\le \frac{1}{\sqrt{1-\rho^2}}\left(8(n+d)\beta_T(\delta_E)\log(\det(V_T))\right)^{1/2}, \end{aligned}$$

where the second step follows from the Cauchy-Schwarz Inequality and the third step holds on E and follows from Lemma 6.

In order to bound the second term on the right hand side of (12), notice that from Assumption A1, we have that for any index $1 \le i \le n$ and any time k,

$$|w_{i,k}| \le R\sqrt{2\log(1/\delta)}\,.$$

Thus, $\|c_k w_{k,i}\| \le \rho^k \|w_{k,i}\| \le R\rho^k \sqrt{2n\log(nT/\delta)}$ holds for all i, k, with probability $1-\delta$. As a result, on an event G with $\mathbb{P}(G) \ge 1-\delta_F$,

$$\left\| \sum_{k=0}^{t-1} c_k w_{k,i} \right\| \le \frac{R\sqrt{2n\log(nT/\delta_F)}}{1-\rho} .$$

Now, on $G \cap E$,

$$\|x_t\| \le \frac{1}{\sqrt{1-\rho^2}} \left(8(n+d)\beta_T(\delta_E)\log(\det(V_T))\right)^{1/2} + \frac{R}{1-\rho}\sqrt{2n\log(nT/\delta_F)} .$$

Thus, we have $G \cap E \subset F \cap E$. Since, by the union bound, $\mathbb{P}(G\cap E) \ge 1-\delta_E-\delta_F$, the lemma is proved. □

Lemma 8. *On the event $E \cap F$, Algorithm 1 changes the policy at most $(n+d)\log(T\alpha_T(\delta_F,\delta_E)^2(1+C^2))$ times up to time T.*

Proof. If we have changed the policy K times up to time T, then we should have that $\det(V_T) \ge 2^K$. On the other hand, we have

$$\lambda_1(V_T) \le \sum_{t=0}^{T-1} \|z_t\|^2 \le T\alpha_T(\delta_F,\delta_E)^2(1+C^2).$$

Thus, it holds that

$$2^K \le (T\alpha_T(\delta_F,\delta_E)^2(1+C^2))^{n+d}.$$

As a result, we have

$$K \le (n+d)\log(T\alpha_T(\delta_F,\delta_E)^2(1+C^2)).$$

□

Next, we bound $\mathbb{I}_{\{E\cap F\}}R_1$. However, before doing this we need a bound on $\mathbb{I}_{\{F\}}\max_{1\le t\le T}\|x_t\|$.

Lemma 9. *We have, for appropriate constants $C_1 > 0, C_2 > 0$ which depend on n, d, λ, ρ only, that for any $t \ge 0$, $\mathbb{I}_{\{F\}}\max_{1\le s\le t}\|x_s\| \le X_t$ where*

$$X_t \overset{\text{def}}{=} \max(e, \lambda(n+d)(e-1), 4(C_1\log(1/\delta)+C_2\log(t/\delta))\log^2(4(C_1\log(1/\delta)+C_2\log(t/\delta))) .$$

Proof. Consider events on F. Let $c = \max(1, \max_{1\le s\le t}\|x_s\|)$.[2] Assume that $t \ge \lambda(n+d)$. By the construction of F, Lemma 2, tedious, but elementary calculations, it can then be shown that

$$c \le A\log^2(c) + B_t, \tag{13}$$

where $A = C_1\log(1/\delta)$ and $B_t = C_2\log(t/\delta)$. From this, further elementary calculations show that the maximum value that c can take on subject to the constraint (13) is bounded from above by the statement. □

Now, let us return to bounded $\mathbb{I}_{\{E\cap F\}}R_1$.

Lemma 10. *Let R_1 be as defined by* (8). *With probability at least $1-\delta/2$,*

$$\mathbb{I}_{\{E\cap F\}}R_1 \le D(X_T \vee \alpha_T(\delta_F,\delta_E)^2)(\sqrt{8T\log 2/\delta}+2).$$

Proof. Write

$$\begin{aligned}\mathbb{I}_{\{E\cap F\}}R_1 &= \mathbb{I}_{\{E\cap F\}}(x_0^\top P(\tilde\Theta_0)x_0 - x_{T+1}^\top P(\tilde\Theta_{T+1})x_{T+1}) \\ &\quad + \mathbb{I}_{\{E\cap F\}}\sum_{t=1}^{T}\left(x_t^\top P(\tilde\Theta_t)x_t - \mathbb{E}\left[x_t^\top P(\tilde\Theta_t)x_t|\mathcal{F}_{t-1}\right]\right).\end{aligned}$$

[2] We use both max and $\vee$ to denote the maximum.

By the boundedness of P, the first term is bounded by $2D\alpha_T(\delta_F, \delta_E)^2$. Define $E_t = E(t, \delta_E)$, $F_t = F(t, \delta_F)$. Note that $E_{t+1} \subset E_t$ and $F_{t+1} \subset F_t$, and so $\mathbb{I}_{\{E_{t+1} \cap F_{t+1}\}} \leq \mathbb{I}_{\{E_t \cap F_t\}}$, and in particular, since $E = E_T$, $F = F_T$, $\mathbb{I}_{\{E \cap F\}} \leq \mathbb{I}_{\{E_t \cap F_t\}}$ holds for any $t \leq T$. Now, the second term is bounded as follows:

$$\begin{aligned} &\mathbb{I}_{\{E \cap F\}} \sum_{t=1}^{T} \left(x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E}\left[x_t^\top P(\tilde{\Theta}_t) x_t | \mathcal{F}_{t-1} \right] \right) \\ &\qquad \leq \sum_{t=1}^{T} \mathbb{I}_{\{E_t \cap F_t\}} \left(x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E}\left[x_t^\top P(\tilde{\Theta}_t) x_t | \mathcal{F}_{t-1} \right] \right). \end{aligned}$$

Define the martingale

$$M_\tau = \sum_{t=1}^{\tau} \mathbb{I}_{\{E_t \cap F_t\}} \left(x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E}\left[x_t^\top P(\tilde{\Theta}_t) x_t | \mathcal{F}_{t-1} \right] \right), \, M_0 = 0\,.$$

This is a martingale, since E_t and F_t are $\mathcal{F}_{t-1}$ measurable. We have that

$$\begin{aligned} |M_\tau - M_{\tau-1}| &\leq \mathbb{I}_{\{E_t \cap F_t\}} \left| x_\tau^\top P(\tilde{\Theta}_\tau) x_\tau - \mathbb{E}\left[x_\tau^\top P(\tilde{\Theta}_\tau) x_\tau | \mathcal{F}_{\tau-1} \right] \right| \\ &\leq 2D \mathbb{I}_{\{E_t \cap F_t\}} \left\| x_t \right\|^2 \\ &\leq 2D \mathbb{I}_{\{E_t \cap F_t\}} \, X_t\,. \\ &\leq 2D \, X_T\,. \end{aligned}$$

By Azuma's inequality, $P(M_T - M_0 \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{8TD^2X_T^2})$. Thus, w.p. at least $1 - \delta/2$,

$$\mathbb{I}_{\{E \cap F\}} R_1 \leq D(X_T \vee \alpha_T(\delta_F, \delta_E)^2) \left(\sqrt{8T \log 2/\delta} + 2 \right).$$

□

Next, we bound $\mathbb{I}_{\{E \cap F\}} |R_3|$.

Lemma 11. *Let R_3 be as defined by Equation* (10)*. Then we have*

$$\mathbb{I}_{\{E \cap F\}} |R_3| \leq \left(8(n+d)\beta_T(\delta_E) \log(\det(V_T)) \right)^{1/2} (2\alpha_T(\delta_F, \delta_E)\sqrt{DT}).$$

Proof. We have that

$$\begin{aligned} \mathbb{I}_{\{E \cap F\}} |R_3| &\leq \mathbb{I}_{\{E \cap F\}} \sum_{t=0}^{T} \left| \left\| P(\tilde{\Theta})^{1/2} \tilde{\Theta}^\top z_t \right\|^2 - \left\| P(\tilde{\Theta})^{1/2} \Theta_*^\top z_t \right\|^2 \right| \\ &\leq \mathbb{I}_{\{E \cap F\}} \left(\sum_{t=0}^{T} \left(\left\| P(\tilde{\Theta})^{1/2} \tilde{\Theta}^\top z_t \right\| - \left\| P(\tilde{\Theta})^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\ &\times \left(\sum_{t=0}^{T} \left(\left\| P(\tilde{\Theta})^{1/2} \tilde{\Theta}^\top z_t \right\| + \left\| P(\tilde{\Theta})^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\ &\leq \mathbb{I}_{\{E \cap F\}} \left(\sum_{t=0}^{T} \left\| P(\tilde{\Theta}_t)^{1/2} (\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{1/2} \\ &\times \left(\sum_{t=0}^{T} \left(\left\| P(\tilde{\Theta})^{1/2} \tilde{\Theta}^\top z_t \right\| + \left\| P(\tilde{\Theta})^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\ &\leq \left(8(n+d)\beta_T(\delta_E) \log(\det(V_T)) \right)^{1/2} (2\alpha_T(\delta_F, \delta_E)\sqrt{DT}), \end{aligned}$$

where the first step holds by the Cauchy-Schwarz Inequality, the second step holds by the triangle inequality, and the third step holds by Lemma 6 and boundedness of the matrices. □

Now we are ready to prove Theorem 4.

Proof of Theorem 4. We have at most $(n+d)\log(T\alpha_T(\delta_F,\delta_E)^2(1+C^2))$ policy changes up to time T. So $|R_2| \le (n+d)\log(T\alpha_T(\delta_F,\delta_E)^2(1+C^2))$. By (11) and Lemmas 10 and 11, we have that with probability at least $1-\delta/2$,

$$\mathbb{I}_{\{E\cap F\}}(R_1 - R_2 - R_3) \le (n+d)\log(T\alpha_T(\delta_F,\delta_E)^2(1+C^2)) + D(X_T \vee \alpha_T(\delta_F,\delta_E)^2)\left(\sqrt{8T\log 2/\delta}+2\right) \\ + \left(8(n+d)\beta_T(\delta_E)\log(\det(V_T))\right)^{1/2}\left(2\alpha_T(\delta_F,\delta_E)\sqrt{DT}\right).$$

Thus, on $E \cap F$,

$$R(T) \le (n+d)\log(T\alpha_T(\delta/4,\delta/4)^2(1+C^2)) + D(X_T \vee \alpha_T(\delta/4,\delta/4)^2)\left(\sqrt{8T\log 2/\delta}+2\right) \\ + \left(8(n+d)\beta_T(\delta/4)\log(\det(V_T))\right)^{1/2}\left(2\alpha_T(\delta/4,\delta/4)\sqrt{DT}\right)$$

Further, on $E\cap F$, by Lemma 2 and our earlier bound on $\max_{1\le t\le T}\|x_t\|$, $\log\det V_T \le (n+d)\log\left(\frac{\lambda(n+d)+TX_T^2}{\lambda(n+d)}\right) + \log\det\lambda I$. Plugging in this and the definition of X_T gives the final bound, which, by Lemma 7, holds with probability $1-\delta$, provided that we choose $\delta_E = \delta_F = \delta/4$. □

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online least squares estimation with self-normalized processes: An application to bandit problems. Arxiv preprint http://arxiv.org/submit/0195808/pdf, 2011.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

P. Auer, R. Ortner, and Cs. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT-07)*, pages 454–468, 2007.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563—1600, 2010.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003. ISSN 1533-7928.

P. L. Bartlett and A. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI 2009*, 2009.

D. Bertsekas. *Dynamic Programming.* Prentice-Hall, 1987.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 2nd edition, 2001.

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.

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

Han-Fu Chen and Lei 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.

Han-Fu Chen and Ji-Feng Zhang. Identification and adaptive control for systems with unknown orders, delay, and coefficients. *Automatic Control, IEEE Transactions on*, 35(8):866 –877, August 1990.

V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. *COLT-2008*, pages 355–366, 2008.

Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *in Proceedings of the 10th Annual Conference on Computational Learning Theory, ACM*, pages 72–80. Press, 1997.

S. Kakade, M. J. Kearns, and J. Langford. Exploration in metric state spaces. In T. Fawcett and N. Mishra, editors, *ICML 2003*, pages 306–312. AAAI Press, 2003.

S.M. Kakade. *On the sample complexity of reinforcement learning.* PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

M. Kearns and S. P. Singh. Near-optimal performance for reinforcement learning in polynomial time. In J. W. Shavlik, editor, *ICML 1998*, pages 260–268. Morgan Kauffmann, 1998.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008.

R.D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS-2004*, 2004.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

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

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):pp. 154–166, 1982.

Tze Leung Lai and Zhiliang Ying. Efficient recursive estimation and adaptive control in stochastic regression and armax models. *Statistica Sinica*, 16:741–772, 2006.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2007.

P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

A. L. Strehl and M. L. Littman. Online linear regression and its application to model-based reinforcement learning. In *NIPS-20*, pages 1417–1424. MIT Press, 2008.

A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In W. W. Cohen and A. Moore, editors, *ICML 2006*, pages 881–888. ACM, 2006. doi: http://doi.acm.org/10.1145/1143844.1143955.

I. Szita and Cs. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML 2010*, pages 1031–1038, 2010.