JMLR: Workshop and Conference Proceedings 19 (2011) 1–26 24th Annual Conference on Learning Theory

# Regret Bounds for the Adaptive Control of Linear Quadratic Systems

**Yasin Abbasi-Yadkori** ABBASIYA@CS.UALBERTA.CA
**Csaba Szepesvári** SZEPESVA@CS.UALBERTA.CA
*Department of Computing Science, University of Alberta*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

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

## 1. Introduction

We study the average cost LQ control problem with unknown model parameters, also known as the adaptive control problem in the control community. The problem is to minimize the average cost of a controller that operates in an environment whose dynamics is linear, while the cost is a quadratic function of the state and the control. The optimal solution is a linear feedback controller which can be computed in a closed form from the matrices underlying the dynamics and the cost. In the learning problem, the topic of this paper, the dynamics of the environment is unknown. This problem is challenging since the control actions influence both the cost and the rate at which the dynamics is learned, a topic of adaptive control. The objective in this case is to minimize the *regret* of the controller, i.e. to minimize the difference between the average cost incurred by the learning controller and that of the optimal controller. In this paper, for the first time, we show an adaptive controller and prove that, under some assumptions, its expected regret is bounded by $\tilde{O}(\sqrt{T})$. We build on recent works in online linear estimation and adaptive control design, 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. 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* (Simon, 1956). The idea was to estimate the unknown parameters from observations and then use the estimated parameters as if they were the true parameters to design a controller. 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 (Becker et al., 1985). This in turn might lead to suboptimal performance.

To avoid non-identification problem, methods that actively explore the environment to gather information are developed (Lai and Wei, 1982a, 1987; Chen and Guo, 1987; Chen and Zhang, 1990; Fiechter, 1997; Lai and Ying, 2006; Campi and Kumar, 1998; Bittanti and Campi, 2006). However, only asymptotic results are proven for these methods. One exception is the work of Fiechter (1997) who proposes an algorithm for the "discounted" LQ problem and analyzes its performance in a PAC framework.

Most of the aforementioned methods use forced-exploration schemes to provide the sufficient exploratory information. The idea is to take exploratory actions according to a fixed and appropriately designed schedule. However, the forced-exploration schemes lack strong worst-case regret bounds, even in the simplest problems (see e.g. Dani and Hayes (2006), section 6). Unlike the preceding methods, Campi and Kumar (1998) 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. They call this the Bet On the Best (BOB) principle. The idea 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 whose associated average cost is the smallest. However, Campi and Kumar (1998) only show asymptotic optimality, i.e. the average cost of their algorithm converges to that of the optimal policy in the limit. In this paper, we modify the algorithm and the proof technique of Campi and Kumar (1998) and extend their work to derive a finite time regret bound. Our work also builds upon on the works of Lai and Wei (1982b); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) in analyzing the linear estimation with dependent covariates, although we use a more recent, improved confidence bound (see Theorem 1).

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 (e.g., Bartlett and Tewari 2009; Auer et al. 2010) and in a PAC-learning setting (e.g., Kearns and Singh 1998; Brafman and Tennenholtz 2002; Kakade 2003; Strehl et al. 2006; Szita and Szepesvári 2010). 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, none of these works 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 (2005); 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).

## 2. Notation and conventions

We use $\|\cdot\|$ and $\|.\|_F$ to denote the 2-norm and the Frobenius norm, respectively. For a positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$, the weighted 2-norm $\|.\|_A$ is defined by $\|x\|_A^2 = x^\top A x$, where $x \in \mathbb{R}^d$. The inner product is denoted by $\langle \cdot, \cdot \rangle$. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of the positive semidefinite matrix $A$, respectively. 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. We use $\mathbb{I}_{\{A\}}$ to denote the indicator function of event $A$.

## 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 $t = 0, 1, \dots$, $u_t \in \mathbb{R}^d$ is the control at time $t$, $x_t \in \mathbb{R}^n$ is the state at time $t$, $c_t \in \mathbb{R}$ is the cost at time $t$, $w_{t+1}$ is the "noise", $A_* \in \mathbb{R}^{n\times n}$ and $B_* \in \mathbb{R}^{n\times d}$ are unknown matrices while $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$ are known (positive definite) matrices. At time zero, for simplicity, $x_0 = 0$. The problem is to design a controller based on past observations to minimize the average expected cost

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

Let $J_*$ be the optimal (lowest) average cost. The *regret* up to time $T$ of a controller which incurs a cost of $c_t$ at time $t$ is defined by

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

which is the difference between the performance of the controller and the performance of the optimal controller that has full information about the system dynamics. Thus the regret can be interpreted as a measure of the cost of not knowing the system dynamics.

### 3.1. Assumptions

In this section, we state our assumptions on the noise and the system dynamics. In particular, we assume that the noise is sub-Gaussian and the system is controllable and observable[1]. Define

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

Thus, the state transition can be written as

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

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

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

(ii) 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 ($\mathbb{E}\left[w_{t+1} | \mathcal{F}_t\right] = 0$, $t = 0, 1, \ldots$);

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

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

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

The assumption $\mathbb{E}\left[w_{t+1} w_{t+1}^\top \,|\, \mathcal{F}_t\right] = I_n$ makes the analysis more readable. However, we shall show it in Section 4 that it is in fact not necessary. Our next assumption on the system uncertainty states that the unknown parameter is a member of a bounded set and is such that the system is controllable and observable. This assumption will let us derive a closed form expression for the optimal control law.

**Assumption A2** The unknown parameter $\Theta_*$ is a member of set $\mathcal{S}$ such that

$$\mathcal{S} \subseteq \mathcal{S}_0 \cap \left\{ \Theta \in \mathbb{R}^{n \times (n+d)} \mid \operatorname{trace}(\Theta^\top \Theta) \leq S^2 \right\} ,$$

where

$$\mathcal{S}_0 = \Big\{ \Theta = (A, B) \in \mathbb{R}^{n \times (n+d)} \mid (A, B) \text{ is controllable,} \\ (A, M) \text{ is observable, where } Q = M^\top M \Big\}.$$

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

1. Controllability and observability are defined in Appendix B

### 3.2. Parameter estimation

In order to implement the OFU principle, we need high-probability confidence sets for the unknown parameter matrix. The derivation of the confidence set is based on results from Abbasi-Yadkori et al. (2011) that use techniques from self-normalized processes to estimate the least squares estimation error. 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 $\ell^2$-regularized least-squares estimate of $\Theta_*$ with regularization parameter $\lambda > 0$:

$$\hat{\Theta}_t = \operatorname*{argmin}_{\Theta} \; e(\Theta) = (Z^\top Z + \lambda I)^{-1} Z^\top X, \tag{2}$$

where $Z$ and $X$ are the matrices whose rows are $z_0^\top, \dots, z_{t-1}^\top$ and $x_1^\top, \dots, x_t^\top$, respectively.

**Theorem 1** *Let $(z_0, x_1), \dots, (z_t, x_{t+1})$, $z_i \in \mathbb{R}^{n+d}$, $x_i \in \mathbb{R}^n$ satisfy the linear model Assumption A1 with some $L > 0$, $\Theta_* \in \mathbb{R}^{(n+d)\times n}$, $\operatorname{trace}(\Theta_*^\top \Theta_*) \le S^2$ and let $\mathcal{F} = (\mathcal{F}_t)$ be the associated filtration. Consider the $\ell^2$-regularized least-squares parameter estimate $\hat{\Theta}_t$ with regularization coefficient $\lambda > 0$ (cf. (2)). Let*

$$V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^\top$$

*be the regularized design matrix underlying the covariates. Define*

$$\beta_t(\delta) = \left( nL \sqrt{2 \log\left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right)^2. \tag{3}$$

*Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\operatorname{trace}((\hat{\Theta}_t - \Theta_*)^\top V_t (\hat{\Theta}_t - \Theta_*)) \le \beta_t(\delta).$$

*In particular, $\mathbb{P}\left(\Theta_* \in \mathcal{C}_t(\delta),\ t = 1, 2, \dots\right) \ge 1 - \delta$, where*

$$\mathcal{C}_t(\delta) = \left\{ \Theta \in \mathbb{R}^{n\times(n+d)} : \operatorname{trace}\left\{ (\Theta - \hat{\Theta}_t)^\top V_t (\Theta - \hat{\Theta}_t) \right\} \le \beta_t(\delta) \right\}.$$

### 3.3. The design of the controller

Let $(A, B) = \Theta \in \mathcal{S}_0$, where $\mathcal{S}_0$ is defined in Assumption A2. Then there is a unique solution $P(\Theta)$ in the class of positive semidefinite symmetric matrices to the *Riccati equation*

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

Under the same assumptions, the matrix $A + BK(\Theta)$ is stable, i.e. its norm-2 is less than one, where

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

is the *gain matrix* (Bertsekas, 2001). Further, by the boundedness of $\mathcal{S}$, we also obtain the boundedness of $P(\Theta)$ (Anderson and Moore, 1971). The corresponding constant will be denoted by $D$:

$$D = \sup \{ \|P(\Theta)\| \mid \Theta \in \mathcal{S} \}. \tag{4}$$

The *optimal control law* for a LQ system with parameters $\Theta$ is

$$u_t = K(\Theta) x_t, \tag{5}$$

i.e., this controller achieves the optimal average cost which satisfies $J(\Theta) = \mathrm{trace}(P(\Theta))$ (Bertsekas, 2001). In particular, the average cost of control law (5) with $\Theta = \Theta_*$ is the optimal average cost $J_* = J(\Theta_*) = \mathrm{trace}(P(\Theta_*))$.

We assume that the bound on the norm of the unknown parameter, $S$, and the sub-Gaussianity constant, $L$, are known:

**Assumption A3** Constants $L$ and $S$ in Assumptions A1 and A2 are known.

The algorithm that we propose implements the OFU principle as follows: At time $t$, the algorithm chooses a parameter $\tilde{\Theta}_t$ from $\mathcal{C}_t(\delta) \cap \mathcal{S}$ such that

$$J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}$$

and then uses the optimal feedback controller (5) underlying the chosen parameter. In order to prevent too frequent changes to the controller (which might harm performance), the algorithm changes controllers only after the current parameter estimates are significantly refined. The details of the algorithm are given in Algorithm 1.

## 4. Analysis

In this section we give our main result together with its proof. Before stating the main theorem, we make one more assumption in addition to the assumptions we made before.

**Assumption A4** The set $\mathcal{S}$ is such that $\rho := \sup_{(A,B) \in \mathcal{S}} \|A + BK(A,B)\| < 1$. Further, there exists a positive number $C$ such that $C = \sup_{\Theta \in \mathcal{S}} \|K(\Theta)\| < \infty$.

Our main result is the following theorem:

**Inputs**: $T, S > 0, \delta > 0, Q, L, \lambda > 0$.
Set $V_0 = \lambda I$ and $\hat{\Theta}_0 = 0$.
$(\tilde{A}_0, \tilde{B}_0) = \tilde{\Theta}_0 = \operatorname{argmin}_{\Theta \in \mathcal{C}_0(\delta) \cap S} J(\Theta)$.
**for** $t := 0, 1, 2, \ldots$ **do**
  **if** $\det(V_t) > 2 \det(V_0)$ **then**
    Calculate $\hat{\Theta}_t$ by (2).
    Find $\tilde{\Theta}_t$ such that $J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap S} J(\Theta) + \frac{1}{\sqrt{t}}$.
    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 LQ problem

**Theorem 2** *For any $0 < \delta < 1$, for any time $T$, 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.*[2]

**Remark 3** *The assumption $\mathbb{E}\left[w_{t+1} w_{t+1}^\top | \mathcal{F}_t\right] = I_n$ makes the analysis more readable. 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 $J(\Theta_*, G_*) = \operatorname{trace}(P(\Theta_*) G_*)$. 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 $\Theta_*$ and $G_*$. The rest of the analysis remains identical, provided that an appropriate confidence set is constructed.*

The least squares estimation error from Theorem 1 scales with the size of the state and action vectors. Thus, in order to prove Theorem 2, we first prove a high-probability bound on the norm of the state vector. Given the boundedness of the state, we decompose the regret and analyze each term using appropriate concentration inequalities.

2. Here, $\tilde{O}$ hides logarithmic factors.

### 4.1. Bounding $\|x_t\|$

We choose an error probability, $\delta > 0$. Given this, we define two "good events" in the probability space $\Omega$. In particular, we define the event that the confidence sets hold for $s = 0, \dots, t$,

$$E_t = \{\omega \in \Omega : \forall s \le t, \quad \Theta_* \in \mathcal{C}_s(\delta/4)\},$$

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

$$F_t = \{\omega \in \Omega : \forall s \le t, \quad \|x_s\| \le \alpha_t\}$$

where

$$\begin{aligned}
\alpha_t &= \frac{1}{1-\rho}\left(\frac{\eta}{\rho}\right)^{n+d}\left[GZ_T^{\frac{n+d}{n+d+1}}\beta_t(\delta/4)^{\frac{1}{2(n+d+1)}} + 2L\sqrt{n\log\frac{4nt(t+1)}{\delta}}\right], \\
\eta &= 1 \vee \sup_{\Theta\in\mathcal{S}} \|A_* + B_* K(\Theta)\|, \\
Z_T &= \max_{0\le t\le T} \|z_t\|, \\
G &= 2\left(\frac{2S(n+d)^{n+d+1/2}}{U^{1/2}}\right)^{1/(n+d+1)}, \\
U &= \frac{U_0}{H}, \\
U_0 &= \frac{1}{16^{n+d-2}(1 \vee S^{2(n+d-2)})},
\end{aligned}$$

and $H$ is any number satisfying[3]

$$H > \left(16 \vee \frac{4S^2M^2}{(n+d)U_0}\right),$$

where

$$M = \sup_{Y\ge 0} \frac{\left(nL\sqrt{(n+d)\log\left(\frac{1+TY/\lambda}{\delta}\right)} + \lambda^{1/2}S\right)}{Y}.$$

In what follows, we let $E = E_T$ and $F = F_T$. First, we show that $E \cap F$ holds with high probability and on $E \cap F$, the state vector does not explode.

**Lemma 4** *$\mathbb{P}(E \cap F) \ge 1 - \delta/2$.*

The proof is in Appendix D. It first shows that $\left\|(\Theta_* - \tilde{\Theta}_t)^\top z_t\right\|$ is well controlled except for a small number of occasions. Given this and proper decomposition of the state update equation, we can prove that the state vector $x_t$ stays smaller than $\alpha_t$. Notice that $\alpha_t$ itself depends $\beta_t$ and $Z_T$, which in turn depend on $x_t$. Thus, we need one more step to have a bound on $\|x_t\|$.

3. We use $\wedge$ and $\vee$ to denote the minimum and the maximum, respectively.

**Lemma 5** *For appropriate problem dependent constants $C_1 > 0, C_2 > 0$ (which are independent of $t, \delta, T$), for any $t \geq 0$, it holds that $\mathbb{I}_{\{F_t\}} \max_{1 \leq s \leq t} \|x_s\| \leq X_t$, where*

$$X_t = Y_t^{n+d+1}$$

*and*

$Y_t \stackrel{\text{def}}{=} (e \vee \lambda(n+d)(e-1) \vee 4(C_1 \log(1/\delta) + C_2 \log(t/\delta)) \log^2(4(C_1 \log(1/\delta) + C_2 \log(t/\delta)))$.

**Proof** Fix $t$. On $F_t$, $\hat{X}_t := \max_{0 \leq s \leq t} \|x_s\| \leq \alpha_t$. With appropriate constants, this implies that

$$x \leq D_1 \sqrt{\beta_t(\delta)} \log(t) x^{\frac{n+d}{n+d+1}} + D_2 \sqrt{\log \frac{t}{\delta}},$$

or

$$x \leq \left( D_1 \sqrt{\beta_t(\delta)} \log(t) + D_2 \sqrt{\log \frac{t}{\delta}} \right)^{n+d+1}, \tag{6}$$

holds for $x = \hat{X}_t$. Let $X_t$ be the largest value of $x \geq 0$ that satisfies (6). Thus,

$$X_t \leq \left( D_1 \sqrt{\beta_t(\delta)} \log(t) + D_2 \sqrt{\log \frac{t}{\delta}} \right)^{n+d+1}, \tag{7}$$

Clearly, $\hat{X}_t \leq X_t$. Because $\beta_t(\delta)$ is a function of $\log\det(V_t)$, (7) has the form of

$$X_t \leq f(\log(X_t))^{n+d+1}. \tag{8}$$

Let $a_t = X_t^{1/(n+d+1)}$. Then, (8) is equivalent to

$$a_t \leq f(\log a_t^{n+d+1}) = f((n+d+1)\log a_t).$$

Let $c = \max(1, \max_{1 \leq s \leq t} \|a_s\|)$. Assume that $t \geq \lambda(n+d)$. By the construction of $F_t$, Lemma 10, tedious, but elementary calculations, it can then be shown that

$$c \leq A \log^2(c) + B_t, \tag{9}$$

where $A = G_1 \log(1/\delta)$ and $B_t = G_2 \log(t/\delta)$. From this, further elementary calculations show that the maximum value that $c$ can take on subject to the constraint (9) is bounded from above by $Y_t$. ■

### 4.2. Regret Decomposition

From the Bellman optimality equations for the LQ problem, we get that (Bertsekas, 1987)[V. 2, p. 228–229]

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

where $\tilde{x}^u_{t+1} = \tilde{A}_t x_t + \tilde{B}_t u + w_{t+1}$ and $(\tilde{A}_t, \tilde{B}_t) = \tilde{\Theta}_t$. Hence,

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

where in the one before last equality we have used $x_{t+1} = A_* x_t + B_* u_t + w_{t+1}$ and the martingale property of the noise. 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

$$R_1 = \sum_{t=0}^{T} \left\{ x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E}\left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \middle| \mathcal{F}_t \right] \right\}, \tag{10}$$

$$R_2 = \sum_{t=0}^{T} \mathbb{E}\left[ x_{t+1}^\top (P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})) x_{t+1} \middle| \mathcal{F}_t \right], \tag{11}$$

and

$$R_3 = \sum_{t=0}^{T} \left( (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) - (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t) \right). \tag{12}$$

Thus, on $E \cap F$,

$$\begin{aligned}
\sum_{t=0}^{T} (x_t^\top Q x_t + u_t^\top R u_t) &= \sum_{t=0}^{T} J(\tilde{\Theta}_t) + R_1 - R_2 - R_3 \\
&\le T J(\Theta_*) + R_1 - R_2 - R_3 + 2\sqrt{T},
\end{aligned}$$

where the last inequality follows from the choice of $\tilde{\Theta}_t$ and the fact that on $E$, $\Theta_* \in C_t(\delta)$). Thus, on $E \cap F$,

$$R(T) \le R_1 - R_2 - R_3 + 2\sqrt{T}. \tag{13}$$

In the following subsections, we bound $R_1, R_2$, and $R_3$.

### 4.3. Bounding $\mathbb{I}_{\{E\cap F\}}R_1$

We start by showing that with high probability all noise terms are small.

**Lemma 6** *With probability* $1-\delta/8$*, for any* $k \le T$*,* $\|w_k\| \le Ln\sqrt{2n\log(8nT/\delta)}$*.*

**Proof** From sub-Gaussianity Assumption A1, we have that for any index $1 \le i \le n$ and any time $k$,

$$|w_{k,i}| \le L\sqrt{2\log(8/\delta)}\,.$$

Thus, with probability $1-\delta/8$, for any $k \le T$, $\|w_k\| \le Ln\sqrt{2n\log(8nT/\delta)}$. ■

**Lemma 7** *Let* $R_1$ *be as defined by* (10)*. With probability at least* $1-\delta/2$*,*

$$\mathbb{I}_{\{E\cap F\}}R_1 \le 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B'_\delta},$$

*where* $W = Ln\sqrt{2n\log(8nT/\delta)}$ *and*

$$B'_\delta = \left(v + TD^2S^2X^2(1+C^2)\right)\log\left(\frac{4nv^{-1/2}}{\delta}\left(v + TD^2S^2X^2(1+C^2)\right)^{1/2}\right).$$

**Proof** Let $f_{t-1} = A_*x_{t-1} + B_*u_{t-1}$ and $P_t = P(\tilde{\Theta}_t)$. Write

$$\begin{aligned} R_1 &= x_0^\top P(\tilde{\Theta}_0)x_0 - x_{T+1}^\top P(\tilde{\Theta}_{T+1})x_{T+1} \\ &\quad + \sum_{t=1}^{T}\left(x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E}\left[x_t^\top P(\tilde{\Theta}_t)x_t|\mathcal{F}_{t-1}\right]\right). \end{aligned}$$

Because $P$ is positive semi-definite and $x_0 = 0$, the first term is bounded by zero. The second term can be decomposed as follows

$$\begin{aligned} \sum_{t=1}^{T}\left(x_t^\top P_t x_t - \mathbb{E}\left[x_t^\top P_t x_t|\mathcal{F}_{t-1}\right]\right) &= \sum_{t=1}^{T} f_{t-1}^\top P_t w_t \\ &\quad + \sum_{t=1}^{T}\left(w_t^\top P_t w_t - \mathbb{E}\left[w_t^\top P_t w_t|\mathcal{F}_{t-1}\right]\right). \end{aligned}$$

We bound each term separately. Let $v_t^\top = f_{t-1}^\top P_t$ and

$$G_1 = \mathbb{I}_{\{E\cap F\}}\sum_{t=1}^{T} v_t^\top w_t = \mathbb{I}_{\{E\cap F\}}\sum_{t=1}^{T}\sum_{i=1}^{n} v_{k,i}w_{k,i} = \sum_{i=1}^{n}\mathbb{I}_{\{E\cap F\}}\sum_{t=1}^{T} v_{k,i}w_{k,i}.$$

Let $M_{T,i} = \sum_{t=1}^{T} v_{k,i} w_{k,i}$. By Theorem 16, on some event $G_{\delta,i}$ that holds with probability at least $1 - \delta/(4n)$, for any $T \geq 0$,

$$M_{T,i}^2 \leq 2R^2 \left( v + \sum_{t=1}^{T} v_{t,i}^2 \right) \log \left( \frac{4nv^{-1/2}}{\delta} \left( v + \sum_{t=1}^{T} v_{t,i}^2 \right)^{1/2} \right) = B_{\delta,i}.$$

On $E \cap F$, $\|v_t\| \leq DSX\sqrt{1+C^2}$ and thus, $v_{t,i} \leq DSX\sqrt{1+C^2}$. Thus, on $G_{\delta,i}$, $\mathbb{I}_{\{E\cap F\}} M_{t,i}^2 \leq B'_\delta$. Thus, we have $G_1 \leq \sum_{i=1}^{n} \sqrt{B'_{\delta,i}}$ on $\cap_{i=1}^{n} G_{\delta,i}$, that holds w.p. $1 - \delta/4$.

Define $X_t = w_t^\top P_t w_t - \mathbb{E}\left[ w_t^\top P_t w_t | \mathcal{F}_{t-1} \right]$ and its truncated version $\tilde{X}_t = X_t \mathbb{I}_{\{X_t \leq 2DW^2\}}$. Define $G_2 = \sum_{t=1}^{T} X_t$ and $\tilde{G}_2 = \sum_{t=1}^{T} \tilde{X}_t$. By Lemma 14,

$$\mathbb{P}\left( G_2 > 2DW^2 \sqrt{2T \log \frac{8}{\delta}} \right) \leq \mathbb{P}\left( \max_{1\leq t\leq T} X_t \geq 2DW^2 \right) + \mathbb{P}\left( \tilde{G}_2 > 2DW^2 \sqrt{2T \log \frac{8}{\delta}} \right).$$

By Lemma 6 and Azuma's inequality, each term on the right hand side is bounded by $\delta/8$. Thus, w.p. $1 - \delta/4$,

$$G_2 \leq 2DW^2 \sqrt{2T \log \frac{8}{\delta}}$$

Summing up the bounds on $G_1$ and $G_2$ gives the result that holds w.p. at least $1 - \delta/2$,

$$\mathbb{I}_{\{E\cap F\}} R_1 \leq 2DW^2 \sqrt{2T \log \frac{8}{\delta}} + n\sqrt{B'_\delta}.$$

■

### 4.4. Bounding $\mathbb{I}_{\{E\cap F\}} |R_2|$

We can bound $\mathbb{I}_{\{E\cap F\}} |R_2|$ by simply showing that Algorithm 1 rarely changes the policy, and hence most terms in (11) are zero.

**Lemma 8** *On the event $E \cap F$, Algorithm 1 changes the policy at most*

$$(n+d) \log_2 \left( 1 + TX_T^2 (1+C^2)/\lambda \right)$$

*times up to time $T$.*

**Proof** If we have changed the policy $K$ times up to time $T$, then we should have that $\det(V_T) \geq \lambda^{n+d} 2^K$. On the other hand, we have

$$\lambda_{\max}(V_T) \leq \lambda + \sum_{t=0}^{T-1} \|z_t\|^2 \leq \lambda + TX_T^2 (1+C^2),$$

where $C$ is the bound on the norm of $K(.)$ as defined in Assumption A4. Thus, it holds that

$$\lambda^{n+d}2^K \leq (\lambda + TX_T^2(1+C^2))^{n+d}.$$

Solving for $K$, we get

$$K \leq (n+d)\log_2\left(1 + \frac{TX_T^2(1+C^2)}{\lambda}\right).$$

■

**Lemma 9** *Let $R_2$ be as defined by Equation (11). Then we have*

$$\mathbb{I}_{\{E\cap F\}}|R_2| \leq 2DX_T^2(n+d)\log_2\left(1 + TX_T^2(1+C^2)/\lambda\right).$$

**Proof** On event $E \cap F$, we have at most $K = (n+d)\log_2\left(1 + TX_T^2(1+C^2)/\lambda\right)$ policy changes up to time $T$. So at most $K$ terms in the summation (11) are non-zero. Each term in the summation is bounded by $2DX_T^2$. Thus,

$$\mathbb{I}_{\{E\cap F\}}|R_2| \leq 2DX_T^2(n+d)\log_2\left(1 + TX_T^2(1+C^2)/\lambda\right).$$

■

### 4.5. Bounding $\mathbb{I}_{\{E\cap F\}}|R_3|$

The summation $\sum_{t=0}^{T}\left\|(\Theta_* - \tilde{\Theta}_t)^\top z_t\right\|^2$ will appear in the analysis while bounding $|R_3|$. So we first bound this summation, whose analysis requires the following two results.

**Lemma 10** *The following holds for any $t \geq 1$:*

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

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

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

The proof of the lemma can be found in Abbasi-Yadkori et al. (2011).

**Lemma 11** *Let $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times m}$ be positive semi-definite matrices such that $A \succeq B$. Then, we have*

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

The proof of this lemma is in Appendix C.

**Lemma 12** *On $E \cap F$, it holds that*

$$\sum_{t=0}^{T} \left\| (\Theta_* - \tilde{\Theta}_t)^\top z_t \right\|^2 \le \frac{16}{\lambda}(1+C^2)X_T^2 \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)}.$$

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

$$\|s_t\| \le \left\| (\Theta_* - \hat{\Theta}_\tau)^\top z_t \right\| + \left\| (\hat{\Theta}_\tau - \tilde{\Theta}_\tau)^\top z_t \right\|. \tag{14}$$

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

$$\begin{aligned}
\left\| (\Theta - \hat{\Theta}_\tau)^\top z_t \right\| &\le \left\| V_t^{1/2}(\Theta - \hat{\Theta}_\tau) \right\| \|z_t\|_{V_t^{-1}} && \text{(Cauchy-Schwartz inequality)} \\
&\le \left\| V_\tau^{1/2}(\Theta - \hat{\Theta}_\tau) \right\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} && \text{(Lemma 11)} \\
&\le \sqrt{2} \left\| V_\tau^{1/2}(\Theta - \hat{\Theta}_\tau) \right\| \|z_t\|_{V_t^{-1}} && \text{(Choice of } \tau) \\
&\le \sqrt{2\beta_\tau(\delta/4)} \|z_t\|_{V_t^{-1}}, && (\lambda_{\max}(M) \le \mathrm{trace}(M) \text{ for } M \succeq 0)
\end{aligned}$$

Applying the last inequality to $\Theta_*$ and $\tilde{\Theta}_\tau$, together with (14) gives

$$\|s_t\|^2 \le 8\beta_\tau(\delta/4) \|z_t\|^2_{V_t^{-1}}.$$

Now, by Assumption A4 and the fact that $\tilde{\Theta}_t \in \mathcal{S}$ we have that

$$\|z_t\|^2_{V_t^{-1}} \le \frac{\|z_t\|^2}{\lambda} \le \frac{(1+C^2)X_T^2}{\lambda}.$$

It follows then that

$$\begin{aligned}
\sum_{t=0}^{T} \|s_t\|^2 &\le \frac{8}{\lambda}(1+C^2)X_T^2 \beta_T(\delta/4) \sum_{t=0}^{T} (\|z_t\|^2_{V_t^{-1}} \wedge 1) \\
&\le \frac{16}{\lambda}(1+C^2)X_T^2 \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)}. && \text{(Lemma 10).}
\end{aligned}$$

■

Now, we are ready to bound $R_3$.

**Lemma 13** *Let $R_3$ be as defined by Equation* (12)*. Then we have*

$$\mathbb{I}_{\{E\cap F\}} |R_3| \le \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}.$$

**Proof** We have that

$$\begin{aligned}
\mathbb{I}_{\{E\cap F\}}\,|R_3| &\le \mathbb{I}_{\{E\cap F\}} \sum_{t=0}^{T} \left| \left\| P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{1/2}\Theta_*^\top z_t \right\|^2 \right| && \text{(Tri. ineq.)} \\
&\le \mathbb{I}_{\{E\cap F\}} \left( \sum_{t=0}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{1/2}\Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} && \text{(C.-S. ineq.)} \\
&\quad \times \left( \sum_{t=0}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2}\Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\le \mathbb{I}_{\{E\cap F\}} \left( \sum_{t=0}^{T} \left\| P(\tilde{\Theta}_t)^{1/2}(\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{1/2} && \text{(Tri. ineq.)} \\
&\quad \times \left( \sum_{t=0}^{T} \left( \left\| P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2}\Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
&\le \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}. && \text{((4), L. 12)}
\end{aligned}$$

■

Now we are ready to prove Theorem 2.

### 4.6. Putting Everything Together

**Proof** [Proof of Theorem 2] By (13) and Lemmas 7, 9, 13 we have that with probability at least $1-\delta/2$,

$$\begin{aligned}
\mathbb{I}_{\{E\cap F\}}(R_1 - R_2 - R_3) \le\ & 2DX_T^2(n+d)\log_2\left(1 + TX_T^2(1+C^2)/\lambda\right) \\
& + 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B'_\delta} \\
& + \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}.
\end{aligned}$$

Thus, on $E\cap F$, with probability at least $1-\delta/2$,

$$\begin{aligned}
R(T) \le\ & 2DX_T^2(n+d)\log_2\left(1 + TX_T^2(1+C^2)/\lambda\right) \\
& + 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B'_\delta} \\
& + \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}.
\end{aligned}$$

Further, on $E\cap F$, by Lemmas 5 and 10,

$$\log\det V_T \le (n+d)\log\left(\frac{\lambda(n+d) + T(1+C^2)X_T^2}{\lambda(n+d)}\right) + \log\det\lambda I.$$

Plugging in this gives the final bound, which, by Lemma 4, holds with probability $1-\delta$. ■

## References


Y. Abbasi-Yadkori, D. Pal, and Cs. Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. Arxiv preprint http://arxiv.org/abs/1102.2670, 2011.

B. D. O. Anderson and J. B. Moore. *Linear Optimal Control*. Prentice-Hall, 1971.

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

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.

P. L. Bartlett and A. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009.

A. Becker, P. R. Kumar, and C. Z. Wei. Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Trans. on Automatic Control*, 30(4):330–338, 1985.

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

D. 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.

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.

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.

H. Chen and J. 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 and T. P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 937–943, 2006.

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

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

S. M. 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.

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. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 681–690, 2008.

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. 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, 1982a.

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, 1982b.

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

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

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

H. A. Simon. dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24(1):74–81, 1956.

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

A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In *ICML*, pages 881–888, 2006.

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

## Appendix A. Tools from Probability Theorem

**Lemma 14** *Let $X_1, \ldots, X_t$ be random variables. Let $a \in \mathbb{R}$. Let $S_t = \sum_{s=1}^t X_s$ and $\tilde{S}_t = \sum_{s=1}^t X_s \mathbb{I}_{\{X_s \le a\}}$. Then it holds that*

$$\mathbb{P}\left(S_t > x\right) \le \mathbb{P}\left(\max_{1\le s\le t} X_s \ge a\right) + \mathbb{P}\left(\tilde{S}_t > x\right).$$

**Proof**

$$\begin{aligned}\mathbb{P}\left(S_t \ge x\right) &\le \mathbb{P}\left(\max_{1\le s\le t} X_s \ge a\right) + \mathbb{P}\left(S_t \ge x,\ \max_{1\le s\le t} X_s \le a\right)\\ &\le \mathbb{P}\left(\max_{1\le s\le t} X_s \ge a\right) + \mathbb{P}\left(\tilde{S}_t \ge x\right).\end{aligned}$$

■

**Theorem 15 (Azuma's inequality)** *Assume that $(X_s; s \ge 0)$ is a supermartingale and $|X_s - X_{s-1}| \le c_s$ almost surely. Then for all $t > 0$ and all $\epsilon > 0$,*

$$P(|X_t - X_0| \ge \epsilon) \le 2\exp\left(\frac{-\epsilon^2}{2\sum_{s=1}^t c_s^2}\right).$$

**Theorem 16 (Self-normalized bound for vector-valued martingales)** *Let $(\mathcal{F}_k; k \ge 0)$ be a filtration, $(m_k; k \ge 0)$ be an $\mathbb{R}^d$-valued stochastic process adapted to $(\mathcal{F}_k)$, $(\eta_k; k \ge 1)$ be a real-valued martingale difference process adapted to $(\mathcal{F}_k)$. Assume that $\eta_k$ is conditionally sub-Gaussian with constant $R$. Consider the martingale*

$$S_t = \sum_{k=1}^t \eta_k m_{k-1}$$

*and the matrix-valued processes*

$$V_t = \sum_{k=1}^{t} m_{k-1} m_{k-1}^\top, \qquad \overline{V}_t = V + V_t, \quad t \geq 0,$$

*Then for any* $0 < \delta < 1$*, with probability* $1 - \delta$*,*

$$\forall t \geq 0, \qquad \|S_t\|^2_{\overline{V}_t^{-1}} \leq 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

## Appendix B. Controllability and Observability

**Definition 1 (Bertsekas (2001))** *A pair* $(A, B)$*, where* $A$ *is an* $n \times n$ *matrix and* $B$ *is an* $n \times d$ *matrix, is said to be controllable if the* $n \times nd$ *matrix*

$$[B \; AB \; \ldots \; A^{n-1}B]$$

*has full rank. A pair* $(A, C)$*, where* $A$ *is an* $n \times n$ *matrix and* $C$ *is an* $d \times n$ *matrix, is said to be observable if the pair* $(A^\top, C^\top)$ *is controllable.*

## Appendix C. Proof of Lemma 11

**Proof** [Proof of Lemma 11]

We consider first a simple case. Let $A = B + mm^\top$, $B$ positive definite. Let $X \neq 0$ be an arbitrary matrix. Using the Cauchy-Schwartz inequality and the fact that for any matrix $M$, $\left\|M^\top M\right\| = \|M\|^2$, we get

$$\left\|X^\top m m^\top X\right\| = \left\|m^\top X\right\|^2 = \left\|m^\top B^{-1/2} B^{1/2} X\right\|^2 \leq \left\|m^\top B^{-1/2}\right\|^2 \left\|B^{1/2} X\right\|^2.$$

Thus,

$$\begin{aligned}\left\|X^\top (B + mm^\top) X\right\| &\leq \left\|X^\top B X\right\| + \left\|m^\top B^{-1/2}\right\|^2 \left\|B^{1/2} X\right\|^2 \\ &= \left(1 + \left\|m^\top B^{-1/2}\right\|^2\right) \left\|B^{1/2} X\right\|^2\end{aligned}$$

and so

$$\frac{\left\|X^\top A X\right\|}{\|X^\top B X\|} \leq 1 + \left\|m^\top B^{-1/2}\right\|^2.$$

We also have that

$$\det(A) = \det(B + mm^\top) = \det(B) \det(I + B^{-1/2} m (B^{-1/2} m)^\top) = \det(B)(1 + \|m\|^2_{B^{-1}}),$$

thus finishing the proof of this case.

More generally, if $A = B + m_1 m_1^\top + \cdots + m_{t-1} m_{t-1}^\top$ then define $V_s = B + m_1 m_1^\top + \cdots + m_{s-1} m_{s-1}^\top$ and use

$$\frac{\left\|X^\top A X\right\|}{\|X^\top B X\|} = \frac{\left\|X^\top V_t X\right\|}{\|X^\top V_{t-1} X\|} \frac{\left\|X^\top V_{t-1} X\right\|}{\|X^\top V_{t-2} X\|} \cdots \frac{\left\|X^\top V_2 X\right\|}{\|X^\top B X\|}.$$

By the above argument, since all the terms are positive, we get

$$\frac{\left\|X^\top A X\right\|}{\|X^\top B X\|} \le \frac{\det(V_t)}{\det(V_{t-1})} \frac{\det(V_{t-1})}{\det(V_{t-2})} \cdots \frac{\det(V_2)}{\det(B)} = \frac{\det(V_t)}{\det(B)} = \frac{\det(A)}{\det(B)},$$

the desired inequality.

Finally, by SVD, if $C \succ 0$, $C$ can be written as the sum of at most $m$ rank-one matrices, finishing the proof for the general case. ∎

## Appendix D. Bounding $\|x_t\|$

We show that $E \cap F$ holds with high probability.

**Proof** [Proof of Lemma 4] Let $M_t = \Theta_* - \tilde{\Theta}_t$. On event $E$, for any $t \le T$ we have that

$$\operatorname{trace}\left( M_t^\top \left( \sum_{s=0}^{t-1} \lambda I + z_s z_s^\top \right) M_t \right) \le \beta_t(\delta/4).$$

Since $\lambda > 0$ we get that,

$$\operatorname{trace}\left( \sum_{s=0}^{t-1} M_t^\top z_s z_s^\top M_t \right) \le \beta_t(\delta/4).$$

Thus,

$$\sum_{s=0}^{t-1} \operatorname{trace}(M_t^\top z_s z_s^\top M_t) \le \beta_t(\delta/4).$$

Since $\lambda_{\max}(M) \le \operatorname{trace}(M)$ for $M \succeq 0$, we get that

$$\sum_{s=0}^{t-1} \left\| M_t^\top z_s \right\|^2 \le \beta_t(\delta/4).$$

Thus, for all $t \ge 1$,

$$\max_{0 \le s \le t-1} \left\| M_t^\top z_s \right\| \le \beta_t(\delta/4)^{1/2} \le \beta_T(\delta/4)^{1/2}. \tag{15}$$

Choose

$$H > \left( 16 \vee \frac{4 S^2 M^2}{(n+d) U_0} \right),$$

where

$$U_0 = \frac{1}{16^{n+d-2}(1 \vee S^{2(n+d-2)})},$$

and

$$M = \sup_{Y \geq 0} \frac{\left( nL\sqrt{(n+d)\log\left(\frac{1+TY/\lambda}{\delta}\right)} + \lambda^{1/2} S \right)}{Y}.$$

Fix a real number $0 \leq \epsilon \leq 1$, and consider the time horizon $T$. Let $\pi(v, \mathcal{B})$ and $\pi(M, \mathcal{B})$ be the projections of vector $v$ and matrix $M$ onto subspace $\mathcal{B} \subset \mathbb{R}^{(n+d)}$, where the projection of matrix $M$ is done column-wise. Let $\mathcal{B} \oplus \{v\}$ be the span of $\mathcal{B}$ and $v$. Let $\mathcal{B}^\perp$ be the subspace orthogonal to $\mathcal{B}$ such that $\mathbb{R}^{(n+d)} = \mathcal{B} \oplus \mathcal{B}^\perp$.

Define a sequence of subspaces $\mathcal{B}_t$ as follows: Set $\mathcal{B}_{T+1} = \emptyset$. For $t = T, \ldots, 1$, initialize $\mathcal{B}_t = \mathcal{B}_{t+1}$. Then while $\left\|\pi(M_t, \mathcal{B}_t^\perp)\right\|_F > (n+d)\epsilon$, choose a column of $M_t$, $v$, such that $\left\|\pi(v, \mathcal{B}_t^\perp)\right\|_F > \epsilon$ and update $\mathcal{B}_t = \mathcal{B}_t \oplus \{v\}$. After finishing with timestep $t$, we will have

$$\left\|\pi(M_t, \mathcal{B}_t^\perp)\right\| \leq \left\|\pi(M_t, \mathcal{B}_t^\perp)\right\|_F \leq (n+d)\epsilon. \tag{16}$$

Let $\mathcal{T}_T$ be the set of timesteps at which subspace $\mathcal{B}_t$ expands. The cardinality of this set, $m$, is at most $n+d$. Denote these timesteps by $t_1 > t_2 > \cdots > t_m$. Let $i(t) = \max\{1 \leq i \leq m : t_i \geq t\}$.

**Lemma 17** *For any vector $x \in \mathbb{R}^{n+d}$*

$$U\epsilon^{2(n+d)} \left\|\pi(x, \mathcal{B}_t)\right\|^2 \leq \sum_{i=1}^{i(t)} \left\|M_{t_i}^\top x\right\|^2,$$

*where $U = U_0/H$.*

**Proof** Let $N = \{v_1, \ldots, v_m\}$ be the set of vectors that are added to $\mathcal{B}_t$ during the expansion timesteps. By construction, $N$ is a subset of the set of all columns of $M_{t_1}, M_{t_2}, \ldots, M_{t_{i(t)}}$. Thus, we have that

$$\sum_{i=1}^{i(t)} \left\|M_{t_i}^\top x\right\|^2 \geq x^\top (v_1 v_1^\top + \cdots + v_m v_m^\top) x.$$

Thus, in order to prove the statement of the lemma, it is enough to show that

$$\forall x,\, \forall j \in \{1, \ldots, m\},\, \sum_{k=1}^{j} \langle v_k, x\rangle^2 \geq \frac{\epsilon^4}{H} \frac{\epsilon^{2(j-2)}}{16^{j-2}(1 \vee S^{2(j-2)})} \left\|\pi(x, B_j)\right\|^2, \tag{17}$$

where $B_j = \operatorname{span}(v_1, \ldots, v_j)$ for any $1 \leq j \leq m$. We have $B_m = \mathcal{B}_t$. We can write $v_k = w_k + u_k$, where $w_k \in B_{k-1}$, $u_k \perp B_{k-1}$, $\|u_k\| \geq \epsilon$, and $\|v_k\| \leq 2S$.

The inequality (17) is proven by induction. First, we prove the induction base for $j = 1$. Without loss of generality, assume that $x = Cv_1$. From condition $H > 16$, we get that

$$16^{-1} H(1 \vee S^{-1}) \geq 1.$$

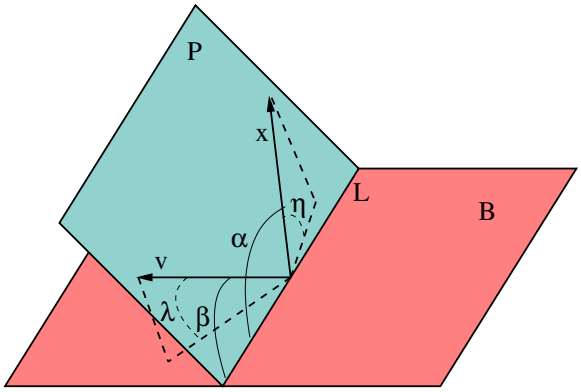


Figure 1: The geometry used in the inductive step. $v = v_{l+1}$ and $B = B_l$.

Thus,

$$\epsilon^2 \geq \frac{\epsilon^2}{16^{-1} H (1 \vee S^{-1})}.$$

Thus,

$$C^2 \|v_1\|^4 \geq \frac{\epsilon^2 C^2 \|v_1\|^2}{16^{-1} H (1 \vee S^{-1})},$$

where we have used the fact that $\|v_1\| \geq \epsilon$. Thus,

$$\langle v_1, x \rangle^2 \geq \frac{\epsilon^4}{H} \frac{\epsilon^{-2}}{16^{-1}(1 \vee S^{-2})} \|\pi(x, B_1)\|^2 ,$$

which establishes the base of induction.

Next, we prove that if the inequality (17) holds for $j = l$, then it also holds for $j = l+1$. Figure 1 contains all relevant quantities that are used in the following argument.

Assume that the inequality (17) holds for $j = l$. Without loss of generality, assume that $x$ is in $B_{l+1}$, and thus $\|x\| = \|\pi(x, B_{l+1})\|$. Let $P \subset B_{l+1}$ be the 2-dimensional subspace that passes through $x$ and $v_{l+1}$. The 2-dimensional subspace $P$ and the $l$-dimensional subspace $B_l$ can, respectively, be identified by $l-1$ and one equations in $B_{l+1}$. Because $P$ is not a subset of $B_l$, the intersection of $P$ and $B_l$ is a line in $B_{l+1}$. Let's call this line $L$. The line $L$ creates two half-planes on $P$. Without loss of generality, assume that $x$ and $v_{l+1}$ are on the same half-plane (notice that we can always replace $x$ by $-x$ in (17)).

Let $0 \leq \beta \leq \pi/2$ be the angle between $v_{l+1}$ and $L$. Let $0 < \lambda < \pi/2$ be the orthogonal angle between $v_{l+1}$ and $B_l$. We know that $\beta > \lambda$, $u_{l+1}$ is orthogonal to $B_l$, and $\|u_{l+1}\| \geq \epsilon$. Thus, $\beta \geq \arcsin(\epsilon / \|v_{l+1}\|)$. Let $0 \leq \alpha \leq \pi$ be the angle between $x$ and $L$ ($\alpha < \pi$, because $x$ and $v_{l+1}$ are on the same half-plane). The direction of $\alpha$ is chosen so that it is consistent with the direction of $\beta$. Finally, let $0 \leq \eta \leq \pi/2$ be the orthogonal angle between $x$ and $B_l$.

By the induction assumption

$$\sum_{k=1}^{l+1}\langle v_k, x\rangle^2 = \langle v_{l+1}, x\rangle^2 + \sum_{k=1}^{l}\langle v_k, x\rangle^2 \geq \langle v_{l+1}, x\rangle^2 + \frac{\epsilon^4}{H}\frac{\epsilon^{2(l-2)}}{16^{l-2}(1\vee S^{2(l-2)})}\left\|\pi(x, B_l)\right\|^2.$$

If $\alpha < \pi/2 + \beta/2$ or $\alpha > \pi/2 + 3\beta/2$, then

$$|\langle v_{l+1}, x\rangle| = |\|v_{l+1}\|\,\|x\|\cos\angle(v_{l+1}, x)| \geq \left|\|v_{l+1}\|\,\|x\|\sin\left(\frac{\beta}{2}\right)\right| \geq \frac{\epsilon\,\|x\|}{4}.$$

Thus,

$$\langle v_{l+1}, x\rangle^2 \geq \frac{\epsilon^2\,\|x\|^2}{16} \geq \frac{\epsilon^4}{H}\frac{\epsilon^{2(l-1)}}{16^{l-1}(1\vee S^{2(l-1)})}\left\|\pi(x, B_{l+1})\right\|^2,$$

where we use $0 \leq \epsilon \leq 1$ and $x \in B_{l+1}$ in the last inequality.

If $\pi/2 + \beta/2 < \alpha < \pi/2 + 3\beta/2$, then $\eta < \pi/2 - \beta/2$. Thus,

$$\|\pi(x, B_l)\| = \|x\|\,|\cos(\eta)| \geq \|x\|\left|\sin\left(\frac{\beta}{2}\right)\right| \geq \frac{\epsilon\,\|x\|}{4S}.$$

Thus,

$$\|\pi(x, B_l)\|^2 \geq \frac{\epsilon^2\,\|x\|^2}{16S^2},$$

and

$$\frac{\epsilon^4}{H}\frac{\epsilon^{2(l-2)}}{16^{l-2}(1\vee S^{2(l-2)})}\left\|\pi(x, B_l)\right\|^2 \geq \frac{\epsilon^4}{H}\frac{\epsilon^{2(l-1)}}{16^{l-1}(1\vee S^{2(l-1)})}\left\|x\right\|^2,$$

which finishes the proof. ■

Next we show that $\left\|M_t^\top z_t\right\|$ is well controlled except when $t \in \mathcal{T}_T$.

**Lemma 18** *We have that for any* $0 \leq t \leq T$,

$$\max_{s\leq t, s\notin\mathcal{T}_t}\left\|M_s^\top z_s\right\| \leq G Z_t^{\frac{n+d}{n+d+1}}\beta_t(\delta/4)^{\frac{1}{2(n+d+1)}},$$

*where*

$$G = 2\left(\frac{2S(n+d)^{n+d+1/2}}{U^{1/2}}\right)^{1/(n+d+1)},$$

*and*

$$Z_t = \max_{s\leq t}\|z_s\|.$$

**Proof** From Lemma 17 we get that

$$\sqrt{U}\epsilon^{n+d}\left\|\pi(z_s,\mathcal{B}_s)\right\| \le \sqrt{i(s)}\max_{1\le i\le i(s)}\left\|M_{t_i}^\top z_s\right\|,$$

which implies that

$$\left\|\pi(z_s,\mathcal{B}_s)\right\| \le \sqrt{\frac{n+d}{U}}\frac{1}{\epsilon^{n+d}}\max_{1\le i\le i(s)}\left\|M_{t_i}^\top z_s\right\|. \tag{18}$$

Now we can write

$$\begin{aligned}
\left\|M_s^\top z_s\right\| &= \left\|(\pi(M_s,\mathcal{B}_s^\perp)+\pi(M_s,\mathcal{B}_s))^\top(\pi(z_s,\mathcal{B}_s^\perp)+\pi(z_s,\mathcal{B}_s))\right\| \\
&= \left\|\pi(M_s,\mathcal{B}_s^\perp)^\top\pi(z_s,\mathcal{B}_s^\perp)+\pi(M_s,\mathcal{B}_s)^\top\pi(z_s,\mathcal{B}_s)\right\| \\
&\le \left\|\pi(M_s,\mathcal{B}_s^\perp)^\top\pi(z_s,\mathcal{B}_s^\perp)\right\| + \left\|\pi(M_s,\mathcal{B}_s)^\top\pi(z_s,\mathcal{B}_s)\right\| \\
&\le (n+d)\epsilon\left\|z_s\right\| + 2S\sqrt{\frac{n+d}{U}}\frac{1}{\epsilon^{n+d}}\max_{1\le i\le i(s)}\left\|M_{t_i}^\top z_s\right\|. \qquad \text{by ((18) and (16))} \quad (19)
\end{aligned}$$

Thus,

$$\max_{s\le t,s\notin\mathcal{T}_t}\left\|M_s^\top z_s\right\| \le (n+d)\epsilon Z_t + 2S\sqrt{\frac{n+d}{U}}\frac{1}{\epsilon^{n+d}}\max_{s\notin\mathcal{T}_t,s\le t}\max_{1\le i\le i(s)}\left\|M_{t_i}^\top z_s\right\|.$$

From $1\le i\le i(s), s\notin\mathcal{T}_t$, we conclude that $s<t_i$. Thus,

$$\max_{s\le t,s\notin\mathcal{T}_t}\left\|M_s^\top z_s\right\| \le (n+d)\epsilon Z_t + 2S\sqrt{\frac{n+d}{U}}\frac{1}{\epsilon^{n+d}}\max_{0\le s<t}\left\|M_t^\top z_s\right\|.$$

By (15) we get that

$$\max_{s\le t,s\notin\mathcal{T}_t}\left\|M_s^\top z_s\right\| \le (n+d)\epsilon Z_t + 2S\sqrt{\frac{n+d}{U}}\frac{1}{\epsilon^{n+d}}\beta_t(\delta/4)^{1/2}.$$

Now if we choose

$$\epsilon = \left(\frac{2S\beta_t(\delta/4)^{1/2}}{Z_t(n+d)^{1/2}U^{1/2}H}\right)^{1/(n+d+1)}$$

we get that

$$\begin{aligned}
\max_{s\le t,s\notin\mathcal{T}_t}\left\|M_s^\top z_s\right\| &\le 2\left(\frac{2S\beta_t(\delta/4)^{1/2}Z_t^{n+d}(n+d)^{n+d+1/2}}{U^{1/2}}\right)^{1/(n+d+1)} \\
&= GZ_t^{\frac{n+d}{n+d+1}}\beta_t(\delta/4)^{\frac{1}{2(n+d+1)}}.
\end{aligned}$$

Finally, we show that this choice of $\epsilon$ satisfies $\epsilon<1$. From the chose of $H$, we have that

$$H > \frac{4S^2M^2}{(n+d)U_0}.$$

Thus,

$$\left(\frac{4S^2M^2}{(n+d)U_0H}\right)^{\frac{1}{2(n+d+1)}} < 1.$$

Thus,

$$\epsilon = \left(\frac{2S\beta_t(\delta/4)}{Z_t(n+d)^{1/2}U_0^{1/2}H^{1/2}}\right)^{\frac{1}{n+d+1}} < 1.$$

■

We can write the state update as

$$x_{t+1} = \Gamma_t x_t + r_{t+1},$$

where

$$\Gamma_{t+1} = \begin{cases} \tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t) & t \notin \mathcal{T}_T \\ A_* + B_* K(\tilde{\Theta}_t) & t \in \mathcal{T}_T \end{cases}$$

and

$$r_{t+1} = \begin{cases} M_t^\top z_t + w_{t+1} & t \notin \mathcal{T}_T \\ w_{t+1} & t \in \mathcal{T}_T \end{cases}$$

Hence we can write

$$\begin{aligned} x_t &= \Gamma_{t-1}x_{t-1} + r_t = \Gamma_{t-1}(\Gamma_{t-2}x_{t-2} + r_{t-1}) + r_t = \Gamma_{t-1}\Gamma_{t-2}x_{t-2} + r_t + \Gamma_{t-1}r_{t-1} \\ &= \Gamma_{t-1}\Gamma_{t-2}\Gamma_{t-3}x_{t-3} + r_t + \Gamma_{t-1}r_{t-1} + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} = \cdots = \Gamma_{t-1}\dots\Gamma_{t-t}x_{t-t} \\ &+ r_t + \Gamma_{t-1}r_{t-1} + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} + \cdots + \Gamma_{t-1}\Gamma_{t-2}\dots\Gamma_{t-(t-1)}r_{t-(t-1)} \\ &= \sum_{k=1}^{t}\left(\prod_{s=k}^{t-1}\Gamma_s\right)r_k. \end{aligned}$$

From Section 4, we have that

$$\eta \geq \max_{t\leq T}\left\|A_* + B_*K(\tilde{\Theta}_t)\right\|, \ \rho \geq \max_{t\leq T}\left\|\tilde{A}_t + \tilde{B}_tK(\tilde{\Theta}_t)\right\|.$$

So we have that

$$\prod_{s=k}^{t-1}\|\Gamma_s\| \leq \eta^{n+d}\rho^{t-k-(n+d)+1}.$$

Hence, we have that

$$\begin{aligned} \|x_t\| &\leq \left(\frac{\eta}{\rho}\right)^{n+d}\sum_{k=1}^{t}\rho^{t-k+1}\|r_{k+1}\| \\ &\leq \frac{1}{1-\rho}\left(\frac{\eta}{\rho}\right)^{n+d}\max_{0\leq k\leq t-1}\|r_{k+1}\|. \end{aligned}$$

Now, $\|r_{k+1}\| \leq \left\|M_k^\top z_k\right\| + \|w_{k+1}\|$ when $k \notin \mathcal{T}_T$, and $\|r_{k+1}\| = \|w_{k+1}\|$, otherwise. Hence,

$$\max_{k<t}\|r_{k+1}\| \leq \max_{k<t, k\notin\mathcal{T}_t}\left\|M_k^\top z_k\right\| + \max_{k<t}\|w_{k+1}\|.$$

The first term can be bounded by Lemma 18. The second term can be bounded as follows: notice that from the sub-Gaussianity Assumption A1, we have that for any index $1 \le i \le n$ and any time $k \le t$, with probability $1 - \delta/(t(t+1))$

$$|w_{k,i}| \le L\sqrt{2\log\frac{t(t+1)}{\delta}}\,.$$

As a result, with a union bound argument, on some event $H$ with $\mathbb{P}(H) \ge 1 - \delta/4$, $\|w_t\| \le 2L\sqrt{n\log\frac{4nt(t+1)}{\delta}}$. Thus, on $H \cap E$,

$$\|x_t\| \le \frac{1}{1-\rho}\left(\frac{\eta}{\rho}\right)^{n+d}\left[GZ_T^{\frac{n+d}{n+d+1}}\beta_t(\delta/4)^{\frac{1}{2(n+d+1)}} + 2L\sqrt{n\log\frac{4nt(t+1)}{\delta}}\right] = \alpha_t.$$

By the definition of $F$, $H \cap E \subset F \cap E$. Since, by the union bound, $\mathbb{P}(H \cap E) \ge 1 - \delta/2$, $\mathbb{P}(E \cap F) \ge 1 - \delta/2$ also holds, finishing the proof. ■