CS 542: Statistical Reinforcement Learning
Project Report

Alisina Bayati

1 Introduction

This report reproduces and analyzes the key results from the paper "Regret Bounds for the Adaptive Control of Linear Quadratic Systems" by Abbasi-Yadkóri and Szepesvári [2]. It focuses on presenting the main theorems, lemmas, and the algorithm designed to solve Linear Quadratic (LQ) control problems with unknown model parameters—commonly referred to as adaptive control—and aims to minimize regret. The paper introduces a high-probability confidence set-based method for estimating the unknown parameters and proposes an algorithm that achieves a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$.

2 Problem Setup

Before formally defining the problem setup, let us briefly describe the notations and conventions used in the paper.

2.1 Notations and Conventions

Notation	**Definition**
$\|\cdot\|$	2-norm
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_A$	Weighted 2-norm, defined by $\|x\|_A^2 = x^\top Ax$, where $x \in \mathbb{R}^d$
$\langle\cdot,\cdot\rangle$	Inner product
$\lambda_{\min}(A)$	Minimum eigenvalue of A
$\lambda_{\max}(A)$	Maximum eigenvalue of A
$A \succ 0$	A is positive definite
$A \succeq 0$	A is positive semidefinite
$\mathbb{I}_{\{A\}}$	Indicator function of event A

Table 1: Notations and Conventions

2.2 Mathematical Formulation

Consider the discrete-time, infinite-horizon Linear Quadratic (LQ) control problem defined by the system dynamics and cost function:

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

for each time step $t \in \{0, 1, 2, \dots\}$. Here, $u_t \in \mathbb{R}^d$ is the control input, $x_t \in \mathbb{R}^n$ is the system state, and $c_t \in \mathbb{R}$ is the cost incurred at time t. The term w_{t+1} represents noise. The matrices

$A_* \in \mathbb{R}^{n\times n}$ and $B_* \in \mathbb{R}^{n\times d}$ are unknown, while $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$ are known and positive definite. We assume the initial state $x_0 = 0$ for simplicity.

The objective is to design a controller that uses past observations to minimize the long-term average expected cost:

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

Let J_* denote the optimal average cost achievable with full knowledge of the system parameters. The regret $R(T)$ up to time T for a controller is defined as:

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

where c_t denotes the incurred costs. Regret measures the total difference between the controller's performance and that of the optimal controller with complete system dynamics information.

3 Summary of Main Results

The paper presents two major contributions. First, it adresses the challenge of estimating the system dynamics parameters A_* and B_* using observations collected up to the current time. Under specific assumptions, it constructs high-probability confidence sets for these parameters. Second, the paper introduces an algorithm for controller design that achieves a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$. These contributions are formalized through two theorems and are detailed in Sections 3.2 and 3.3, respectively.

The proof of Theorem 1, which establishes a confidence set for the system parameters, is **not** explicitly provided in the paper. Instead, the authors reference their prior work [1] for a justification. In this report, I provide an independent sketch of the proof for this theorem.

Theorem 2 is proved within the paper using several supporting lemmas. For the sake of brevity, I state some of these lemmas without including their detailed proofs.

Before presenting these results, let us outline the key assumptions that form the basis of the analysis.

3.1 Assumptions

1. The noise w_t is component-wise sub-Gaussian with known constant L.

2. The system is controllable. In other words, the pair (A_*, B_*) is controllable.

3. The unknown parameter $\Theta^* = (A^*, B^*)$ lies within a bounded set, i.e.

$$\mathcal{S} \subseteq \{\Theta \in \mathbb{R}^{n\times(n+d)} : \text{trace}(\Theta^\top \Theta) \le S^2\},$$

where S is known and

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

(Refer to Appendix 5.1 for detailed definitions of controllability and observability.)

3.2 Parameter Estimation

Define

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

Therefore,

$$x_{t+1} = \begin{pmatrix} A_* & B_* \end{pmatrix} \begin{pmatrix} x_t \\ u_t \end{pmatrix} + w_{t+1} = \Theta_*^\top z_t + w_{t+1}.$$

The paper aims to create high-probability confidence sets for Θ_* using an ℓ_2-regularized least-squares approach, similar to the ridge regression lectures in class. The loss function is defined as:

$$\begin{aligned} e(\Theta) &= \lambda \operatorname{trace}(\Theta^\top \Theta) + \sum_{t=0}^{t-1} \|x_{t+1} - \Theta^\top z_t\|_2^2 \\ &= \lambda \operatorname{trace}\left(\Theta^\top \Theta\right) + \sum_{t=0}^{t-1} \operatorname{trace}\left(\left(x_{t+1} - \Theta^\top z_t\right)\left(x_{t+1} - \Theta^\top z_t\right)^\top\right). \end{aligned}$$

The estimator $\hat{\Theta}_t$ is obtained by minimizing the loss function:

$$\hat{\Theta}_t = \arg\min_{\Theta} e(\Theta) = \left(Z^\top Z + \lambda I\right)^{-1} Z^\top X, \tag{2}$$

where:

- Z is the matrix with rows $z_0, z_1, \ldots, z_{t-1}$.
- X is the matrix with rows $x_1, x_2, \ldots, x_T$.

Now, let us state the following theorem.

Theorem 1

Let $(z_0, x_1), \ldots, (z_t, x_{t+1})$, $z_i \in \mathbb{R}^{n+d}$, $x_i \in \mathbb{R}^n$ be generated by the linear model described earlier and the assumptions in 3.1 hold for some $L > 0$, $\Theta_* \in \mathbb{R}^{(n+d)\times n}$, $\operatorname{trace}(\Theta_*^\top \Theta_*) \leq S^2$. Consider the ℓ_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 $1 - \delta$,

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

In particular,

$$\mathbb{P}(\Theta_* \in \mathcal{C}_t(\delta), t = 1, 2, \ldots) \geq 1 - \delta,$$

where

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

Proof of Theorem 1: First, let us show that the estimator $\hat{\Theta}_t$ given in equation (2) minimizes $e(\Theta)$. We set the derivative of the loss function with respect to Θ to zero:

$$\frac{\partial e(\Theta)}{\partial \Theta} = 2\lambda\Theta - 2\sum_{i=0}^{t-1} z_i(x_{i+1} - \Theta^\top z_i)^\top = 0.$$

Simplifying the expression:

$$2\lambda\Theta - 2\left(\sum_{i=0}^{t-1} z_i x_{i+1}^\top - \sum_{i=0}^{t-1} z_i z_i^\top \Theta\right) = 0.$$

Rewriting:

$$\left(\lambda I + \sum_{i=0}^{t-1} z_i z_i^\top\right)\Theta = \sum_{i=0}^{t-1} z_i x_{i+1}^\top.$$

Defining $Z^\top = [z_0\ z_1\ \cdots\ z_{t-1}]$ and $X^\top = [x_1\ x_2\ \cdots\ x_t]$, we simplify to:

$$\left(\lambda I + Z^\top Z\right)\Theta = Z^\top X \implies \boxed{\hat{\Theta}_t = \left(\lambda I + Z^\top Z\right)^{-1} Z^\top X}$$

Thus, the estimator $\hat{\Theta}_t$ is given by equation (2).

Next, we present a sketch of the proof demonstrating that, for any $0 < \delta < 1$, with probability at least $1 - \delta$:

$$\text{trace}\left((\hat{\Theta}_t - \Theta_*)^\top V_t(\hat{\Theta}_t - \Theta_*)\right) \leq \beta_t(\delta),$$

where

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

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

Claim 1: The expression we aim to bound,

$$\text{trace}\left((\hat{\Theta}_t - \Theta_*)^\top V_t(\hat{\Theta}_t - \Theta_*)\right),$$

can be decomposed into three distinct terms, denoted as A, B, and C, defined below:

$$\text{Term A} := \lambda^2\,\text{trace}\left(\Theta_*^\top V_t^{-1}\Theta_*\right),$$

$$\text{Term B} := -2\lambda\,\text{trace}\left(\Theta_*^\top V_t^{-1}\sum_{j=0}^{t-1} z_j w_{j+1}^\top\right),$$

$$\text{Term C} := \text{trace}\left(\left(\sum_{i=0}^{t-1} w_{i+1} z_i^\top\right) V_t^{-1} \left(\sum_{j=0}^{t-1} z_j w_{j+1}^\top\right)\right).$$

Proof of Claim 1: Please refer to 5.2 in Appendix.
Now we bound each term individually.

- **Term A:** Note that term A is deterministic. Since $V_t \succeq \lambda I$, it follows that $V_t^{-1} \preceq \lambda^{-1} I$. Therefore:

$$\text{Term A} \leq \lambda^2\,\text{trace}\left(\Theta_*^\top(\lambda^{-1}I)\Theta_*\right) = \lambda\,\text{trace}\left(\Theta_*^\top\Theta_*\right) \leq \lambda S^2.$$

- **Term B:** To bound Term B, we first present the following claim.

  Claim 2: Term B can be expressed as

  $$\text{Term B} = -2\lambda \sum_{j=0}^{t-1} w_{j+1}^\top \Theta_*^\top V_t^{-1} z_j,$$

  which forms a martingale difference sequence. Consequently, the Azuma-Hoeffding inequality can be applied to establish a high-probability bound for this term.

  Proof of Claim 2: Please refer to 5.3 in Appendix.

- **Term C:** We now present the following claim.

  Claim 3: With probability at least $1 - \delta$,

  $$\text{Term C} \leq 2nL^2 \log\left(\frac{n \ \det(V_t)^{1/2} \ \det(\lambda I)^{-1/2}}{\delta}\right).$$

  Proof of Claim 3: Please refer to 5.4 in Appendix.

Finally, by putting together all the bounds derived for each individual term, it can be shown that with probability at least $1 - \delta$,

$$\text{Term A} + \text{Term B} + \text{Term C} \leq \left(nL\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}S\right)^2. \qquad \blacksquare$$

3.3 Controller Design

Consider the system parameters $(A, B) = \Theta \in \mathcal{S}_0$ as defined in 3.1. For each Θ, there exists a unique positive semidefinite solution $P(\Theta)$ to the discrete-time algebraic Riccati equation:

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

Then, the optimal control law for the linear-quadratic system [3] is given by

$$u_t = K(\Theta)x_t, \tag{4}$$

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

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

The matrix $P(\Theta)$ is uniformly bounded on $\mathcal{S}$, i.e. there exists $D > 0$ such that $\|P(\Theta)\| \leq D$ for all $\Theta \in \mathcal{S}$. Moreover, under the above control, the closed-loop matrix $A + BK(\Theta)$ is stable (i.e., $\|A + BK(\Theta)\|_2 < 1$) and the average cost of control law (4) with $\Theta = \Theta_*$ is the optimal average cost $J_* = J(\Theta_*) = \text{trace}(P(\Theta_*))$.

The paper's main contribution lies in proposing the following algorithm for control design within this problem setting, along with its regret analysis, as summarized in Theorem 2.

Algorithm 1 The proposed adaptive algorithm for the LQ problem

```
Inputs: $T, S > 0, \delta > 0, Q, L, \lambda > 0$.
Set $V_0 = \lambda I$ and $\hat{\Theta}_0 = 0$.
$(\hat{A}_0, \hat{B}_0) = \hat{\Theta}_0 = \arg\min_{\Theta \in \mathcal{C}_0(\delta) \cap \mathcal{S}} J(\Theta)$.
for $t = 0, 1, 2, \dots$ do
    if $\det(V_t) > 2\det(V_0)$ then
        Calculate $\hat{\Theta}_t$ by Equation (2).
        Find $\hat{\Theta}_t$ such that $J(\hat{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}$.
        Let $V_0 = V_t$.
    else
        $\hat{\Theta}_t = \hat{\Theta}_{t-1}$.
    end if
    Calculate $u_t$ based on the current parameters, $u_t = K(\hat{\Theta}_t) x_t$.
    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
```

Theorem 2

Under the following assumptions:

1. $\rho := \sup_{(A,B) \in \mathcal{S}} \|A + BK(A,B)\| < 1$,
2. There exists a constant $C > 0$ such that $C = \sup_{\Theta \in \mathcal{S}} \|K(\Theta)\| < \infty$,

for any $0 < \delta < 1$ and any time horizon T, the regret of Algorithm 1 is bounded with probability at least $1 - \delta$ by:

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

where the constant in the bound depends on the problem, and $\tilde{O}$ hides logarithmic factors.

Proof of Theorem 2: From [3], assuming for simplicity that the covariance matrix of the process noise, given by $\mathbb{E}[w_{t+1} w_{t+1}^\top \mid x_t, u_t]$, is the identity matrix I_n, the Bellman optimality equation for the LQ problem is as follows:

$$\operatorname{trace}\left(P(\Theta)\right) + x_t^\top P(\Theta) x_t = \min_u \left\{ x_t^\top Q x_t + u^\top R u + \mathbb{E}\left[x_{t+1}^{u\ \top} P(\Theta_t) x_{t+1}^u \mid x_t, u \right] \right\}.$$

Thus, given the system dynamics parameters at time t, denoted by $\tilde{\Theta}_t^\top = (\tilde{A}, \tilde{B})$, the right-hand side of the Bellman optimality equation is expressed as:

$$\begin{aligned}
&\min_u \left\{ x_t^\top Q x_t + u^\top R u + \mathbb{E}\left[\tilde{x}_{t+1}^{u\ \top} P(\tilde{\Theta}_t) \tilde{x}_{t+1}^u \mid x_t, u \right] \right\} \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[\tilde{x}_{t+1}^{u\ \top} P(\tilde{\Theta}_t) \tilde{x}_{t+1}^u \mid x_t, u_t \right], \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[(\tilde{A}_t x_t + \tilde{B}_t u_t + w_{t+1})^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t u_t + w_{t+1}) \mid x_t, u_t \right], \\
&= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E}\left[(\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t u_t) \mid x_t, u_t \right] + \mathbb{E}\left[w_{t+1}^\top P(\tilde{\Theta}_t) w_{t+1} \mid x_t, u_t \right], \\
&= x_t^\top Q x_t + u_t^\top R u_t + (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t) (\tilde{A}_t x_t + \tilde{B}_t u_t) + \mathbb{E}\left[w_{t+1}^\top P(\tilde{\Theta}_t) w_{t+1} \mid x_t, u_t \right],
\end{aligned}$$

in the above equation, replacing the noise term w_{t+1} with

$$w_{t+1} = x_{t+1} - A_* x_t - B_* u_t$$

simplifies $\mathbb{E}\left[w_{t+1}^\top P(\tilde{\Theta}_t) w_{t+1} \mid x_t, u_t\right]$ to

$$\begin{aligned}
&\mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mid x_t, u_t\right] - \mathbb{E}\left[(A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t) \mid x_t, u_t\right], \\
&= \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mid x_t, u_t\right] - (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t)
\end{aligned}$$

Therefore, the RHS is expressed as:

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

Moreover, the LHS of the Bellman optimality equation is written as:

$$\operatorname{trace}\left(P(\tilde{\Theta}_t)\right) + x_t^\top P(\tilde{\Theta}_t) x_t = J(\tilde{\Theta}_t) + x_t^\top P(\tilde{\Theta}_t) x_t.$$

Therefore, equating both sides gives:

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

Hence,

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

where we added and subtracted $\mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mid x_t, u_t\right]$ in the last line.

Thus, the total cost incurred by Algorithm 1 up to time T is computed as:

$$\sum_{t=1}^{T} (x_t^\top Q x_t + u_t^\top R u_t) = \sum_{t=1}^{T} J(\tilde{\Theta}_t) + R_1 - R_2 - R_3, \tag{5}$$

where

$$\begin{aligned}
R_1 &= \sum_{t=0}^{T} \left(x_t^\top P(\tilde{\Theta}_t) x_t - \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mid x_t, u_t\right]\right), \\
R_2 &= \sum_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\top \left(P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})\right) x_{t+1} \mid x_t, u_t\right],
\end{aligned}$$

and

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

Now, let us choose an error probability $\delta > 0$ and define the following "good events":

$$E_t = \left\{\omega \in \Omega : \forall s \le t, \Theta_* \in \mathcal{C}_s\left(\frac{\delta}{4}\right)\right\}, \quad \text{and} \quad E = E_T,$$

$$F_t = \{\omega \in \Omega : \forall s \leq t, \|x_s\| \leq \alpha_t\}, \quad \text{and} \quad F = F_T.$$

where

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

$$\eta = 1 \vee \sup_{\Theta \in \mathcal{S}} \|A_* + B_* K(\Theta)\|,$$

$$Z_T = \max_{0 \leq t \leq T} \|z_t\|,$$

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

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

and H is any number satisfying[1]

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

where

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

In the above definitions, E_t ensures that the true parameters Θ_* remain within the confidence sets $\mathcal{C}_s\left(\frac{\delta}{4}\right)$ for all $s \leq t$, while F_t ensures that the state vectors x_s remain bounded by α_t over the same time horizon.

In Algorithm 1, at each time step t, $\tilde{\Theta}_t$ is selected to satisfy (line 7):

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

On $E \cap F$, since $\Theta_* \in \mathcal{C}_t(\delta)$, it follows that:

$$\inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) \leq J(\Theta_*).$$

Summing over all time steps $t = 1$ to T, we obtain:

$$\sum_{t=1}^{T} J(\tilde{\Theta}_t) \leq TJ(\Theta_*) + \sum_{t=1}^{T} \frac{1}{\sqrt{t}}.$$

Claim 4:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}.$$

Proof of Claim 4: See 5.5 in the Appendix for details.

Therefore, combining these results:

$$\sum_{t=1}^{T} J(\tilde{\Theta}_t) \leq TJ(\Theta_*) + 2\sqrt{T}.$$

[1] We use $\wedge$ and $\vee$ to denote the minimum and the maximum, respectively.

Substituting this into (5) yields:

$$\boxed{R(T) := \sum_{t=1}^{T}(x_t^\top Q x_t + u_t^\top R u_t) - TJ(\Theta_*) \leq 2\sqrt{T} + R_1 - R_2 - R_3,} \tag{6}$$

where $R(T)$ denotes the regret.

In the remainder of the paper, the authors focus on deriving high-probability bounds for R_1, R_2 and R_3. For the sake of brevity, I present these bounds as lemmas without providing detailed proofs and then add them together to demonstrate that the regret of the proposed algorithm is $\tilde{\mathcal{O}}(\sqrt{T})$.

- **Bound R_1 on $E \cap F$:**

Lemma 1 (stated without proof)

Lemma 7 of the paper: With probability at least $1 - \delta/2$,

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

where $W = Ln\sqrt{2n\log\frac{8nT}{\delta}}$ and

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

- **Bound $|R_2|$ on $E \cap F$:**

Lemma 2

Lemma 9 of the paper:

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

where X_T is defined in the following lemma.

Lemma 3 (stated without proof)

Lemma 5 of the paper: For appropriate problem-dependent constants $C_1 > 0$, $C_2 > 0$ (which are independent of t, δ, 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

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

Proof of Lemma 2: Recall the definition:

$$R_2 = \sum_{t=0}^{T}\mathbb{E}\left[x_{t+1}^\top\left(P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})\right)x_{t+1} \mid x_t, u_t\right].$$

The summation is non-zero only at time steps where $\tilde{\Theta}_t$ (and consequently the policy) changes. Let K denote the number of policy changes up to time T. Each non-zero term is bounded by:

$$\begin{aligned}\left|x_{t+1}^\top \left(P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})\right) x_{t+1}\right| &\le \|x_{t+1}\|^2 \cdot \left\|P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})\right\| \\ &\le X_T^2 \left(\|P(\tilde{\Theta}_t)\| + \|P(\tilde{\Theta}_{t+1})\|\right) \le 2DX_T^2.\end{aligned}$$

Thus, on $E \cap F$:

$$|R_2| \le K \cdot 2DX_T^2.$$

To bound K, note that the algorithm updates the policy when $V_T = \lambda I + \sum_{t=0}^{T-1} z_t z_t^\top$ grows significantly. More precisely, as denoted in line 5 of Algorithm 1, the policy is updated when $\det(V_t) > 2\det(V_0)$. Thus, if K policy changes have occurred, then:

$$\det(V_T) \ge \det(\lambda I) \cdot 2^K = \lambda^{n+d} 2^K.$$

On the other hand, on $E \cap F$,

$$\|z_t\|^2 = \|x_t\|^2 + \|u_t\|^2 = \|x_t\|^2 + \|K(\tilde{\Theta}_t)x_t\|^2 \le \|x_t\|^2(1+C^2) \le X_T^2(1+C^2)$$

Hence:

$$\lambda_{\max}(V_T) \le \lambda + TX_T^2(1+C^2).$$

Thus:

$$\det(V_T) = \prod_{i=1}^{n+d} \lambda_i(V_T) \le (\lambda_{\max}(V_t))^{n+d} \le (\lambda + TX_T^2(1+C^2))^{n+d}.$$

Combining these inequalities:

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

Taking the $(n+d)$-th root:

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

Rearranging:

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

Taking $\log_2$:

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

Finally:

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

Substituting K into the earlier inequality:

$$|R_2| \le K \cdot 2DX_T^2 \le 2DX_T^2(n+d)\log_2\left(1 + \frac{TX_T^2(1+C^2)}{\lambda}\right). \qquad \blacksquare$$

which concludes the proof of Lemma 2.

- **Bound $|R_3|$ on $E \cap F$:**

Lemma 4 (stated without proof)

Lemma 13 of the paper: Let R_3 be as defined by Equation (12). Then we have

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

Thus, from (6) and Lemmas 1, 2, and 4, with probability at least $1 - \delta/2$, on the event $E \cap F$,

$$R(T) \leq 2DW^2 \sqrt{2T \log\left(\frac{8}{\delta}\right)} + n\sqrt{B'_\delta} + 2DX_T^2(n+d)\log_2\left(1 + \frac{TX_T^2(1+C^2)}{\lambda}\right)$$

$$+ \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD\left(\beta_T\left(\frac{\delta}{4}\right)\log\frac{\det(V_T)}{\det(\lambda I)}\right)^{1/2}\sqrt{T}.$$

To establish that the regret $R(T)$ satisfies $R(T) = \tilde{\mathcal{O}}(\sqrt{T})$, we analyze each term in the final bound, focusing on their scaling to T while ignoring logarithmic factors.

1. **First Term:**

$$2DX_T^2(n+d)\log_2\left(1 + \frac{TX_T^2(1+C^2)}{\lambda}\right) = \tilde{\mathcal{O}}(1)$$

Reasoning: Given $X_T = Y_T^{n+d+1}$ and $Y_T = \mathcal{O}(\log^k T)$ for some constant k, it follows that:

$$X_T^2 = \mathcal{O}\left(\log^{2k(n+d+1)} T\right)$$

The logarithmic term inside the log scales as:

$$\log_2\left(1 + \frac{TX_T^2(1+C^2)}{\lambda}\right) = \mathcal{O}(\log T)$$

Therefore, the first term is poly-logarithmic in T, which is absorbed into $\tilde{\mathcal{O}}(1)$.

2. **Second Term:**

$$2DW^2\sqrt{2T\log\left(\frac{8}{\delta}\right)} = \tilde{\mathcal{O}}(\sqrt{T})$$

Reasoning: This term directly scales with $\sqrt{T}$, contributing linearly to the $\sqrt{T}$ component of the regret bound.

3. **Third Term:**

$$n\sqrt{B'_\delta} = \tilde{\mathcal{O}}(\sqrt{T})$$

Reasoning: Since $X_T = \mathcal{O}(\log^{k(n+d+1)} T)$, we have:

$$B'_\delta = \mathcal{O}\left(T\log^{2k(n+d+1)+1} T\right)$$

Taking the square root:

$$\sqrt{B'_\delta} = \mathcal{O}\left(\sqrt{T}\log^{k(n+d+1)+0.5} T\right)$$

Therefore, the third term scales as:

$$\tilde{\mathcal{O}}(\sqrt{T})$$

4. **Fourth Term:**

$$\frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD\left(\beta_T\left(\frac{\delta}{4}\right)\log\frac{\det(V_T)}{\det(\lambda I)}\right)^{1/2}\sqrt{T} = \tilde{\mathcal{O}}(\sqrt{T})$$

Reasoning: Given $\beta_T(\delta) = \mathcal{O}(\log T)$ and $\log\det(V_T) = \mathcal{O}(\log T)$, we have:

$$\left(\beta_T\left(\frac{\delta}{4}\right)\log\frac{\det(V_T)}{\det(\lambda I)}\right)^{1/2} = \mathcal{O}(\log T)$$

Therefore, the fourth term scales as:

$$\tilde{\mathcal{O}}(\sqrt{T})$$

Combining all contributions, the dominant scaling is $\tilde{O}(\sqrt{T})$. Therefore, the cumulative regret $R(T)$ on the event $E \cap F$ satisfies:

$$R(T) = \tilde{O}(\sqrt{T}).$$

Thus, for any $0 < \delta < 1$ and any time horizon T, on the event $E \cap F$, we have shown that with probability at least $1 - \frac{\delta}{2}$, the regret of Algorithm 1 is bounded by $\tilde{O}(\sqrt{T})$.

Now, consider the following lemma:

Lemma 5 (stated without proof)

Lemma 4 of the paper: It holds that $P(E \cap F) \geq 1 - \frac{\delta}{2}$.

Since the probability that both events E and F occur simultaneously is at least $1 - \frac{\delta}{2}$, and conditioned on $E \cap F$ the regret is at most $\tilde{O}(\sqrt{T})$ with probability at least $1 - \frac{\delta}{2}$, we now combine these statements as below

$$P\left(\text{Regret} \leq \tilde{O}(\sqrt{T})\right) \geq P\left(\text{Regret} \leq \tilde{O}(\sqrt{T}) \middle| E \cap F\right) \cdot P(E \cap F) \geq (1 - \delta/2)^2 \geq 1 - \delta$$

Thus, with probability at least $1 - \delta$, the regret of Algorithm 1 is at most $\tilde{O}(\sqrt{T})$. This completes the proof of the main theorem presented in the paper.

4 Conclusion

In this report, I reproduced and analyzed the main results from Abbasi-Yadkóri and Szepesvári's study on regret bounds for adaptive control in Linear Quadratic (LQ) systems.

Firstly, they established a high-probability confidence bound for the system parameters, ensuring that the true parameters lie within the constructed confidence sets with probability at least $1 - \delta$.

Secondly, they proposed an algorithm that achieves a regret bound of $\tilde{O}(\sqrt{T})$, demonstrating its effectiveness in minimizing cumulative regret over time.

5 Appendix

5.1 Definition of Controllability and Observability

Controllability: A system is controllable if the state x_t can be driven from any initial state to any final state within a finite time using an appropriate control input u_t. This is determined by the controllability matrix:

$$\mathcal{C} = \begin{bmatrix} B_* & A_* B_* & A_*^2 B_* & \cdots & A_*^{n-1} B_* \end{bmatrix}.$$

The system is controllable if $\mathcal{C}$ has full rank n, where n is the dimension of x_t.

Observability: A system is observable if the initial state x_0 can be uniquely determined from output measurements over a finite time. For a pair (A, M), the observability matrix is:

$$\mathcal{O} = \begin{bmatrix} M \\ MA \\ MA^2 \\ \vdots \\ MA^{n-1} \end{bmatrix}.$$

The system is observable if $\mathcal{O}$ has full column rank n.

5.2 proof of claim 1

From the derivation of the ℓ_2-regularized least-squares estimator, we have:

$$\hat{\Theta}_t = V_t^{-1} \left(\sum_{i=0}^{t-1} z_i x_{i+1}^\top \right).$$

Since $x_{i+1} = \Theta_*^\top z_i + w_{i+1}$, we have:

$$\begin{aligned}
\hat{\Theta}_t &= V_t^{-1} \left(\sum_{i=0}^{t-1} z_i (\Theta_*^\top z_i + w_{i+1})^\top \right) \\
&= V_t^{-1} \left(\sum_{i=0}^{t-1} z_i z_i^\top \Theta_* + \sum_{i=0}^{t-1} z_i w_{i+1}^\top \right) \\
&= V_t^{-1} \left((V_t - \lambda I)\Theta_* + \sum_{i=0}^{t-1} z_i w_{i+1}^\top \right) \\
&= \Theta_* - V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{i=0}^{t-1} z_i w_{i+1}^\top.
\end{aligned}$$

Therefore, the estimation error is:

$$\hat{\Theta}_t - \Theta_* = -V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{i=0}^{t-1} z_i w_{i+1}^\top.$$

We are interested in bounding:

$$\text{trace}\left((\hat{\Theta}_t - \Theta_*)^\top V_t (\hat{\Theta}_t - \Theta_*) \right).$$

Substituting the expression for $\hat{\Theta}_t - \Theta_*$, we have:

$$\begin{aligned}
&\text{trace}\left((\hat{\Theta}_t - \Theta_*)^\top V_t (\hat{\Theta}_t - \Theta_*) \right) \\
&= \text{trace}\left(\left(-V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{i=0}^{t-1} z_i w_{i+1}^\top \right)^\top V_t \left(-V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top \right) \right) \\
&= \text{trace}\left(\left(-\lambda \Theta_*^\top V_t^{-1} + \sum_{i=0}^{t-1} w_{i+1} z_i^\top V_t^{-1} \right) V_t \left(-V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top \right) \right) \\
&= \text{trace}\left(\left(-\lambda \Theta_*^\top V_t^{-1} V_t + \sum_{i=0}^{t-1} w_{i+1} z_i^\top V_t^{-1} V_t \right) \left(-V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top \right) \right) \\
&= \text{trace}\left(\left(-\lambda \Theta_*^\top + \sum_{i=0}^{t-1} w_{i+1} z_i^\top \right) \left(-V_t^{-1} \lambda \Theta_* + V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top \right) \right).
\end{aligned}$$

We now expand the product inside the trace:

$$\begin{aligned}
&\text{trace}\left(\lambda \Theta_*^\top V_t^{-1} \lambda \Theta_* - \lambda \Theta_*^\top V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top - \sum_{i=0}^{t-1} w_{i+1} z_i^\top V_t^{-1} \lambda \Theta_* + \sum_{i=0}^{t-1} w_{i+1} z_i^\top V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top \right) \\
&= \underbrace{\lambda^2 \text{trace}\left(\Theta_*^\top V_t^{-1} \Theta_* \right)}_{A} \underbrace{-2\lambda\, \text{trace}\left(\Theta_*^\top V_t^{-1} \sum_{j=0}^{t-1} z_j w_{j+1}^\top \right)}_{B} + \underbrace{\text{trace}\left(\left(\sum_{i=0}^{t-1} w_{i+1} z_i^\top \right) V_t^{-1} \left(\sum_{j=0}^{t-1} z_j w_{j+1}^\top \right) \right)}_{C},
\end{aligned}$$

completing the proof.

5.3 Proof of Claim 2

$$\begin{aligned}\text{Term B} &= -2\lambda \operatorname{trace}\left(\underbrace{\Theta_*^\top V_t^{-1}}_{K} \sum_{j=0}^{t-1} z_j w_{j+1}^\top\right) = -2\lambda \sum_{j=0}^{t-1} w_{j+1}^\top K z_j \\ &= -2\lambda \sum_{j=0}^{t-1} w_{j+1}^\top \Theta_*^\top V_t^{-1} z_j := \sum_{j=0}^{t-1} a_j,\end{aligned}$$

where we define the sequence $\{a_j\}_{j=0}^{t-1}$ as:

$$a_j := -2\lambda\, w_{j+1}^\top \Theta_*^\top V_t^{-1} z_j.$$

Each a_j satisfies

$$\mathbb{E}[a_j \mid z_1, z_2, \ldots, z_j] = 0.$$

Therefore, $\{a_j\}_{j=0}^{t-1}$ forms a martingale difference sequence, allowing us to apply the Azuma-Hoeffding inequality to bound Term B with high probability.

5.4 Proof of Claim 3

To bound Term C, we rely on Corollary 1 from Section 2 of [1], presented below without its proof for brevity.

Lemma – Corollary 1 of [1] (Modified)

Let $S_t = \sum_{k=1}^t \eta_k m_{k-1}$, where:

- η_k is a martingale difference sequence, i.e., $\mathbb{E}[\eta_k \mid \mathcal{F}_{k-1}] = 0$, where $\mathcal{F}_{k-1}$ is the filtration representing the information available up to time $k-1$.
- $m_k \in \mathbb{R}^d$ is a vector-valued process adapted to the filtration $\mathcal{F}_k$.

Let $V_t = V + \sum_{k=1}^t m_{k-1} m_{k-1}^\top$, where:

- V is a fixed positive definite matrix in $\mathbb{R}^{d \times d}$.

Assume:

- η_k is sub-Gaussian and $L^2 > 0$, such that $\forall \gamma \in \mathbb{R}$,

$$\mathbb{E}[e^{\gamma \eta_k} \mid \mathcal{F}_{k-1}] \leq e^{\frac{\gamma^2 L^2}{2}}.$$

Then, for any $t \geq 0$, with probability $1 - \delta$,

$$\|S_t\|_{V_t^{-1}}^2 \leq 2L^2 \log\left(\frac{\det(V_t)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

Recall that:

$$\text{Term C} := \operatorname{trace}\left(\left(\sum_{i=0}^{t-1} w_{i+1} z_i^\top\right) V_t^{-1} \left(\sum_{j=0}^{t-1} z_j w_{j+1}^\top\right)\right).$$

Define, for each $j = 1, \ldots, n$:

$$\mathbf{m}_j = \sum_{i=0}^{t-1} w_{i+1,j} z_i,$$

where $w_{i+1,j}$ denotes the j-th component of the noise vector w_{i+1}. Thus, Term C can be expressed as:

$$\text{Term C} = \sum_{j=1}^{n} \mathbf{m}_j^\top V_t^{-1} \mathbf{m}_j = \sum_{j=1}^{n} \|\mathbf{m}_j\|_{V_t^{-1}}^2.$$

We apply the provided lemma to each $\mathbf{m}_j$. For each j, define:

- $\eta_k^{(j)} = w_{k,j}$, the j-th component of the noise vector w_k.
- $m_{k-1} = z_{k-1}$.
- $V = \lambda I$.

With these definitions, for each j, $\mathbf{m}_j = \sum_{i=0}^{t-1} w_{i+1,j} z_i$ corresponds to $S_t = \sum_{k=1}^{t} \eta_k^{(j)} m_{k-1}$ in the lemma. Therefore, by the lemma, with probability at least $1 - \frac{\delta}{n}$, we have:

$$\|\mathbf{m}_j\|_{V_t^{-1}}^2 \leq 2L^2 \log \left(\frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\frac{\delta}{n}} \right).$$

To ensure that this bound holds for all $j = 1, \ldots, n$ simultaneously, we apply the union bound. Therefore, with probability at least $1 - \delta$, the above inequality holds for all j, and thus:

$$\text{Term C} = \sum_{j=1}^{n} \|\mathbf{m}_j\|_{V_t^{-1}}^2 \leq 2nL^2 \log \left(\frac{n \det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right),$$

which completes the proof.

5.5 Proof of Claim 4

We prove by induction that:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}.$$

Base Case: For $T = 1$,

$$\sum_{t=1}^{1} \frac{1}{\sqrt{t}} = \frac{1}{\sqrt{1}} = 1, \quad \text{and} \quad 2\sqrt{1} = 2.$$

Inductive Step: Assume for some $k \geq 1$,

$$\sum_{t=1}^{k} \frac{1}{\sqrt{t}} \leq 2\sqrt{k}.$$

We need to prove:

$$\sum_{t=1}^{k+1} \frac{1}{\sqrt{t}} \leq 2\sqrt{k+1}.$$

Expanding the sum:

$$\sum_{t=1}^{k+1} \frac{1}{\sqrt{t}} = \sum_{t=1}^{k} \frac{1}{\sqrt{t}} + \frac{1}{\sqrt{k+1}}.$$

By the inductive hypothesis:

$$\sum_{t=1}^{k} \frac{1}{\sqrt{t}} \leq 2\sqrt{k}.$$

Thus:

$$\sum_{t=1}^{k+1} \frac{1}{\sqrt{t}} \leq 2\sqrt{k} + \frac{1}{\sqrt{k+1}}.$$

It suffices to show:

$$2\sqrt{k} + \frac{1}{\sqrt{k+1}} \leq 2\sqrt{k+1}.$$

Using $\sqrt{k+1} - \sqrt{k} = \frac{1}{\sqrt{k+1}+\sqrt{k}}$, we get:

$$\frac{1}{\sqrt{k+1}} \leq \frac{2}{\sqrt{k+1}+\sqrt{k}} \leq 2(\sqrt{k+1} - \sqrt{k}),$$

which is true for all $k \geq 1$. Thus, By induction, the inequality holds for all $T \geq 1$:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}. \qquad \blacksquare$$

References

[1] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online least squares estimation with self-normalized processes: An application to bandit problems, 2011.

[2] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

[3] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.