Regret Lower Bounds for Unbiased Adaptive Control of Linear Quadratic Regulators

Ingvar Ziemann, Henrik Sandberg

Regret Lower Bounds for Unbiased Adaptive Control of Linear Quadratic Regulators

Ingvar Ziemann, Henrik Sandberg

Abstract— **We present lower bounds for the regret of adaptive control of the linear quadratic regulator. These are given in terms of problem specific expected regret lower bounds valid for unbiased policies linear in the state. Our approach is based on the insight that the adaptive control problem can, given our assumptions, be reduced to a sequential estimation problem. This enables the use of the Cramér-Rao information inequality which yields a scaling limit lower bound of logarithmic order. The bound features both information-theoretic and control-theoretic quantities. By leveraging existing results, we are able to show that the bound is tight in a special case.**

I. Introduction

In this paper, we study (expected) regret lower bounds for adaptive control of the linear quadratic regulator (LQR) with unknown parametrization. We consider a class of algorithms which are, in particular, linear and unbiased and therefore lend themselves to an estimation-theoretic interpretation. As considered here, the performance of an adaptive policy – depending only on the observations – is measured in terms of its (expected) regret, which compares the difference in cost between a particular adaptive policy sequence and the optimal sequence having knowledge of the system's parameters. It is desirable to provide a tight lower bound on the regret as this informs us what the optimal rate of convergence for an adaptive policy might be. Note that the optimal policy is allowed to depend on the system's parameters whereas the adaptive policy is not. The exact problem formulation is explained in detail in Section II, with a concise statement found as Problem 1. The problem of finding good adaptive controllers for an unknown system has appeared in various incarnations since the 1950s, see for instance [1], [2], [3] or [4] for an overview.

Although considerable attention has been devoted to adaptive control within control engineering, there has been recent revitilization of interest in the topic as an analytically tractable prototype for reinforcement learning in general[1] state and action spaces, [5]. Irrespective of its apparent simplicity, little is known about the instance specific optimal performance (measured in terms of its regret) a policy could or should achieve on this problem. In a simplified setting, with additional assumptions and where the cost only depends on the state, a logarithmic (in time) lower bound was already known and attained asymptotically in the 1980s by Lai in [6]. For SISO systems, Guo establishes upper bounds on the regret of least squares algorithms in [7] of logarithmic order. A non-asymptotic scalar analysis attaining the logarithmic order of magnitude was also executed by Rantzer in [8], giving further evidence that logarithmic rates are attainable. However, to best of our knowledge, neither upper nor lower bounds are known for the expected regret (henceforth just the regret) for adaptive control of the more general linear quadratic regulator considered here. In particular, no lower bounds beyond the case of cheap control and invertible B-matrix as in [6] exist to this date.

In light of this, our contribution is to provide lower bounds for adaptive control of the general case of LQR for linear and unbiased policies. Our lower bounds depend crucially on the parameters of the model and thus lend themselves to control-theoretic interpretation. Notably, our bounds are obtained by regarding the adaptive control regret as *sequential estimation* errors. This both provides a parallel and a contrast to bandit problems, which is perhaps the other most studied reinforcement learning problem, and in which regret is often analyzed as *sequential testing* errors, [9]. These results are presented in Theorem 4.5 which gives an information lower bound for regret, R_T, asymptotically of the form

$$R_T \geq \mathcal{C}(A, B, f) \times \ln T$$

valid for the linear unbiased control laws considered here. Here $\mathcal{C}$ is a constant depending on the system's parameters, A and B, and the distribution of the noise sequence w, sampled indepently and indentically distributed according to a density f. T is the time horizon. In the sequel, we shall see how this constant depends on *both* control-theoretic and information-theoretic quantities. Finally, our lower bound is applied to the special cases of the LQR in [6] and [7]. We are able to establish that our lower bound agrees with upper bounds found in these works, indicating that our analysis is tight in these cases.

A. Related Work

A number of papers, see for instance [10], [11] and [12], have presented policies attaining regret of the order of magnitude[2] $\sqrt{T}$. However, to date, there are no matching lower bounds of order $\sqrt{T}$ and as mentioned, in several special cases ([6], [7], [8]), regret of order $\ln T$ has been achieved.

Further, [13] proves lower bounds for adaptive control of more general Markov decision processes which subsume our

Ingvar Ziemann (ziemann@kth.se) and Henrik Sandberg (hsan@kth.se) are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. This work was supported in part by the Swedish Research Council (grant 2016-00861), and the Swedish Foundation for Strategic Research (Project CLAS). The authors are also indebted to four anonymous referees for their helpful comments.

[1] As opposed to discrete.

[2] Modulo logarithmic factors.

problem. However, the explicit form of their bound requires the solution of an optimization problem which in general does not seem analytically tractable. Their bounds are based on a classical reduction to testing theory argument. By contrast, we appeal to the structure of the linear quadratic problem to give simple bounds based on estimation theory.

Perhaps most similar in structure to our bounds are those available for the multi-armed bandit problem, which is a classical reinforcement learning problem. These lower bounds were first produced in [9] and [14] provides a concise approach. Understanding the fundamental limitiations in terms of regret any policy must face has led to optimal adaptive policies, [15]. It is thus our hope that the present analysis also sheds some light on the fundamental hardness in terms of the structure of the adaptive control problem at hand, as to further our understanding of how to devise optimal policies for this problem. One can for instance imagine that the matrices appearing in our lower bound can inspire the choice of preconditioning for gradient based adaptive controllers, [16].

II. Problem Formulation

We are interested in fundamental limitations of adaptive control performance of systems – linear quadratic regulators – of the following form

$$x_{t+1} = Ax_t + Bu_t + w_{t+1}, \qquad x_0 = w_0 \tag{1}$$

where the state x_t is a vector in $\mathbb{R}^n$ with $A \in \mathbb{R}^{n\times n}$ and the input u_t a vector in $\mathbb{R}^m$ with $B \in \mathbb{R}^{n\times m}$. Here (w_t) is a sequence of independent random variables with covariance matrix $\Sigma \succ 0$. Further, it is assumed that the w_t have a density with finite Fisher information (defined below). The control input, (u_t), is allowed to depend on the current and all past states and control inputs. The pair (A, B) is assumed stabilizable but unknown. We consider the per stage cost

$$\mathbf{E}x_t^\top Q x_t + \mathbf{E}u_t^\top R u_t \tag{2}$$

with $Q \succeq 0, Q \in \mathbb{R}^{n\times n}$ and $R \succeq 0, R \in \mathbb{R}^{m\times m}$, with the inequalities taken in the semi-definite order. Note that the weighting matrices are assumed to be known. We shall further assume that the matrices A, B, Q, R, Σ are chosen such that there exists a unique solution to the Riccati equation associated to (1)–(2) and its steady state limit. In this case the optimal policy is to choose $u_t = K_t x_t$, where K_t is defined implicitly via the Riccati equation. We refer the reader to [17] for conditions sufficient for existence and uniqueness. In what follows we will be more interested in the so-called asymptotically optimal policy given by selecting $u_t = Kx_t$ where K is the limit of K_t.

A. Background and Notation

As mentioned in the introduction, our analysis rests on the notion of regret as a sequential estimation error. Since we are concerned with quadratic performance criteria, it is natural to attempt to find lower bounds based on the Fisher information, [18]. For a family of probability densities $\{f_\theta\}$ parametrized by $\theta \in \Theta$ for some set Θ, this is defined as

$$I_\theta = \int \nabla_\theta \ln f_\theta(x) \left[\nabla_\theta \ln f_\theta(x)\right]^\top f_\theta(x) dx$$

whenever the integral exists. In our case, Θ will be the set of possible A and B matrices for the dynamics (1). We will often use the notation $I_K^t = I_{K(A,B)}^t$, by which we denote the Fisher information (with respect to K) corresponding to the density of the random variable $(x_1, \dots, x_t)$ given by the closed loop dynamics

$$x_{t+1} = (A + BK)x_t + w_{t+1}.$$

This also justifies the notation $K = K(A, B)$ for K the asymptotically optimal linear feedback matrix. Due to the stability of the (asymptotically) optimal closed loop $A + BK$, the underlying Markov chain is mixing and we may apply the Ergodic Theorem (cf. [19] and [20]) to the corresponding sequence of likelihoods to conclude that

$$\bar{I}_K := \lim_{t\to\infty} \frac{1}{t} I_K^t \tag{3}$$

exists. We shall in the sequel see that this quantity is of fundamental importance as it asymptotically measures the information per sample obtained by any "good" policy. Moreover, for any matrix K resulting in stable eigenvalues of $A + BK$, we define the closed loop Gramian

$$\Gamma_K = \sum_{t=0}^{\infty} \left[(A + BK)(A + BK)^\top\right]^t$$

measuring the asymptotic noise-to-cost relationship.

We will require some notions from linear algebra. If A and B are two $n\times n$ matrices, we denote by $A\otimes B$ their Kroenecker product and by $\operatorname{vec} A$ or for short $\vec{A}$ the vectorization of A, which is a vector in $\mathbb{R}^{n^2}$ with the same entries as A. The Moore-Penrose pseudoinverse of A is denoted $A^\dagger$. We also remind the reader of the following useful identity relating vectorization, Kroenecker produts and traces. Namely, for $n \times n$ matrices A, B, C one has that

$$\operatorname{tr}(ABC) = \operatorname{tr}\left[(I \otimes B)\vec{C}\vec{A}^\top\right] \tag{4}$$

Observe that the trace on the left is of an $n\times n$ matrix whereas on the right the trace is that of an $n^2 \times n^2$ matrix.

We will also make frequent use of asymptotic notation. A quantity $f(t)$ is said to be "little-oh" of $g(t)$ if $\limsup |f(t)/g(t)| = 0$ and is in short written as $f = o(g)$. If instead $\limsup |f(t)/g(t)| \le C, C > 0$, we write $f = O(g)$. In general, these limits will be for large times, usually indexed by t or T.

B. Regret and Learning

A natural measure of performance of any adaptive algorithm, is its (expected) regret, defined as

$$R_T = \sum_{t=0}^{T-1} \left(\mathbf{E}x_{t+1}^\top Q x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top Q \tilde{x}_{t+1}\right) + \sum_{t=0}^{T-1} \left(\mathbf{E}u_t^\top R u_t - \mathbf{E}\tilde{u}_t^\top R \tilde{u}_t\right). \tag{5}$$

The variables $\tilde{x}, \tilde{u}$ are the (asymptotically) optimal control and state trajectories given knowledge of A and B. That is, $\tilde{u}_t = K_t \tilde{x}_t$ where K_t solves the appropriate Riccati equation. The regret thus measures the difference between the cumulative cost of a strategy (u_t) and the cumulative optimal cost, given knowledge of A, B. In other words, it is the price of uncertainty about the model's parameters.

Now, if K is the limiting optimal feedback matrix $K_t \to K$ and since the difference in cost between using this and K_t is exponentially small, instead setting

$$\tilde{u}_t = K\tilde{x}_t$$

yields the same asymptotic regret[3] and it thus suffices to compare any policy to $\tilde{u}_t = K\tilde{x}_t$. Due the linear form of the optimal law, we shall restrict ourselves to consider linear controllers, of the form

$$u_t = \hat{K}_t x_t$$

where $\hat{K}_t$ is a random matrix, constituting the decision variable and allowed to depend only on the present and past state observations and inputs. Given this, we shall henceforth refer to $\hat{K}_t$ as the policy itself, where it is tacitly understood that the control law is $u_t = \hat{K}_t x_t$. Note that without further restriction $\hat{K}_t = K$ is an admissible policy which of course leads to zero asymptotic regret. To mitigate this we introduce the following notion.

Definition 2.1: A policy $(\hat{K}_t)$ is (asymptotically) unbiased if $\mathbf{E}\hat{K}_t = K + o((t \ln t)^{-1})$ for all t and any (A, B).

This allows the use of the Cramér-Rao inequality. As in classical statistics from which the notion originates, this is a restriction on the applicability of our bound. It is however clear that some restriction such as unbiasedness needs to be made to enforce adaptivity.

Example 2.2: The maximum likelihood estimator typically[4] has bias of order $1/t$, [21] which is slightly larger than $(t \ln t)^{-1})$. However, this can be reduced to the order t^{-2} via the Jackknife without jeopardizing the rate of convergence, [22].

As we are investigating the rate of convergence of adaptive policies, we shall also need to specify what precisely we mean by convergence.

Definition 2.3: A policy $(\hat{K}_t)$ is said to be consistently convergent if $\hat{K}_t \to K$ in probability and $\mathbf{E}x_t x_t^\top \to \mathbf{E}\tilde{x}_t\tilde{x}_t^\top$.

Next, we need to capture the fact that a good estimator $\hat{K}_t$ for K should use information from all the t samples obtained so far.

Definition 2.4: A policy $(\hat{K}_t)$ is said to be sample stable if

$$\mathbf{E}p(x_k)q(B\hat{K}_t) \quad = \quad \mathbf{E}p(x_k)\mathbf{E}q(B\hat{K}_t) \; \times \; (1 + o(1)) \quad (6)$$

for all polynomials p and q of order 0 to 2 and all $k \le t$.

Note that, assuming that the limit exists, it always holds that $\mathbf{E}pq = \mathbf{E}p\mathbf{E}q + o(1)$ whenever $\hat{K}_t$ converges in probability to a constant. Roughly speaking, the assumption then says that the decay of correlation between policy and sample is proportional to the rate of convergence of the estimator. To see that this definition holds with little loss of generality, consider the scalar case and write

$$\mathbf{E}pq = \mathbf{E}p\mathbf{E}q + \rho_{p,q}\sqrt{\mathbf{V}p\mathbf{V}q} \qquad (7)$$

where $\rho_{p,q}$ is the correlation between p and q. As long as $\rho_{p,q}\sqrt{\mathbf{V}q}$ decays in t as fast as $\mathbf{E}q$, the policy $\hat{K}_t$ will be sample stable. In particular, any consistent and stabilizing policy will have $\rho_{p,q} = o(1)$ and $\mathbf{V}p = O(1)$. Further assuming that $\sqrt{\mathbf{V}q}$ decays at the same rate as $\mathbf{E}q$ the definition applies. Example 2.5, discusses consistent convergence and sample stability in the context of the scalar least squares estimator.

Having stated these definitions, the central problem treated here can now be formulated.

Problem 1: Analyze the asymptotics of regret (5) of linear, asymptotically unbiased, consistently convergent adaptive policies for the linear quadratic regulator (1)–(2).

An example to which this theory applies is presented below.

Example 2.5: Suppose that the dimension of the state space is 1, that $A = a$ is a scalar, that $Q = 1$, $R = 0$, that $B = 1$ is known and that the distribution of w_t is Gaussian with mean zero and variance 1. Since least squares is Gaussian maximum likelihood, the least squares estimator in [8] has bias of order $1/t$, which can be reduced to $O(t^{-2}) = o((t \ln t)^{-1})$ by using the jackknife technique. We now show that the least squares estimator is both consistently convergent and sample stable.

Theorem 4 and Theorem 6 of [8] imply that $\mathbf{V}\hat{a}_t = O(1/t)$ and that $\mathbf{E}x_t^2 = 1 + O(1/t)$ respectively. Since convergence in L^2 implies convergence in probability and since the optimal cost attained by the policy $-a$ is 1 per stage, it follows that $\hat{K}_t = -\hat{a}_t$ is consistently convergent.

To prove that $\hat{K}_t = -\hat{a}_t$ is sample stable, consider (7). Here, we may restrict attention to polynomials $q(z) = |z + a|^{\deg q}, \deg q = 0, 1, 2$ since any other polynomial r, $\deg r \le 2$, is in their span. Clearly $\mathbf{V}p = O(1)$ and by Theorem 4 in [8] $\mathbf{V}q = O(1/t^{\deg q})$. Incidentally, since $\hat{K}_t$ has small bias, this is also the asymptotically optimal order of convergence of $\mathbf{E}q$. Since $\rho_{p,q} = o(1)$ one has that $\rho_{p,q}\sqrt{\mathbf{V}p\mathbf{V}q} = o(\mathbf{E}p\mathbf{E}q) = o(1) \times \mathbf{E}p\mathbf{E}q$ due to the availability of a lower bound of the same order on $\mathbf{E}q$.

Remark 2.6: In general, similar arguments can most likely be applied to policies in n-dimensional state spaces, given supporting results such as those in [8].

III. Regret Analysis by Decoupling and Decomposition

Let us first look at the component of regret corresponding to being at the wrong state, $\mathbf{E}x_t^\top x_t - \mathbf{E}\tilde{x}_t^\top \tilde{x}_t$. Once this is done, we shall see that the analysis of the control cost comes nearly for free.

Now, if we expand both cost sequences in terms of their

[3] Modulo factors $o(\ln T)$ which do not enter our analysis.

[4] This has to be qualified with sufficiently excited inputs.

dynamics, we see that

$$\begin{aligned}
&\mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1}\\
&\quad= \mathbf{E}x_t^\top (A-B\hat{K}_t)^\top (A-B\hat{K}_t)x_t + \mathbf{E}w_{t+1}^\top w_{t+1}\\
&\quad+ \mathbf{E}x_t^\top (A-B\hat{K}_t)^\top w_{t+1} + \mathbf{E}w_{t+1}^\top (A-B\hat{K}_t)x_t\\
&\quad- \mathbf{E}\tilde{x}_t^\top (A-BK)^\top (A-BK)\tilde{x}_t - \mathbf{E}w_{t+1}^\top w_{t+1}\\
&\quad- \mathbf{E}\tilde{x}_t^\top (A-BK)^\top w_{t+1} - \mathbf{E}w_{t+1}^\top (A-BK)\tilde{x}_t\\
&\quad= \mathbf{E}x_t^\top (A-B\hat{K}_t)^\top (A-B\hat{K}_t)x_t\\
&\quad- \mathbf{E}\tilde{x}_t^\top (A-BK)^\top (A-BK)\tilde{x}_t.
\end{aligned}\tag{8}$$

since the mixed terms have expectation zero. Even though this formula is very simple, it is rather revealing. Thinking heuristically about the distribution of x_t and $\tilde{x}_t$ for large t, these clearly have to be very close if we are to have low regret. In particular, both random variables should be stochastically bounded. This is obviously true for $\tilde{x}_t$ since the optimal control law is stabilizing. Similarly, it "ought" to be true for x_t since a "good" adaptive law, should mimic the optimal law. If this is true, this means that the majority of the regret is incurred due to the difference between

$$(A-B\hat{K}_t)^\top (A-B\hat{K}_t) \text{ and } (A-BK)^\top (A-BK).$$

A. Decoupling

Much of the hardness in identification arises since control and estimation are not easily decoupled. To deal with this, we now present a lemma which allows us to approximately decouple past mistakes from present mistakes.

Lemma 3.1 (Regret Decoupling): Assume that the policy $\hat{K}_t$ is asymptotically unbiased and sample stable. Then

$$\begin{aligned}
&\mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1}\\
&\quad= \Big(\operatorname{tr}\Big(\mathbf{E}\left[x_t x_t^\top\right] \mathbf{E}\Big[(BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top\Big]\Big)\\
&\quad+ \operatorname{tr}\mathbf{E}\left[(x_t x_t^\top - \tilde{x}_t\tilde{x}_t^\top)(A-BK)(A-BK)^\top\right] \Big)\\
&\quad\times (1+o(1)) + o((t\ln t)^{-1}).
\end{aligned}$$

Proof: Observe first by expanding the squares that

$$\begin{aligned}
(A-B\hat{K}_t)(A-B\hat{K}_t)^\top &= (A-BK)(A-BK)^\top\\
&+ (BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top\\
&- A(B\hat{K}_t)^\top - B\hat{K}_t A^\top + A(BK)^\top + BKA^\top\\
&+ 2BK(BK)^\top - BK(B\hat{K}_t)^\top - B\hat{K}_t(BK)^\top.
\end{aligned}$$

The last two lines above have expectation $o((t\ln t)^{-1})$. Thus, using first sample stability and then unbiasedness, one has

$$\begin{aligned}
&\mathbf{E}\left[x_t x_t^\top (A-B\hat{K}_t)(A-B\hat{K}_t)^\top\right]\\
&\quad= \mathbf{E}[x_t x_t^\top]\mathbf{E}\left[(A-B\hat{K}_t)(A-B\hat{K}_t)^\top\right](1+o(1))\\
&\quad= \Big(\mathbf{E}[x_t x_t^\top]\mathbf{E}\left[(A-BK)(A-BK)^\top\right]\\
&\quad+ \mathbf{E}[x_t x_t^\top]\left[(BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top\right] + o((t\ln t)^{-1})\Big)\\
&\quad\times (1+o(1)).
\end{aligned}$$

Using the representation devised in (8) together with what we just derived, we find that

$$\begin{aligned}
&\mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1}\\
&\quad= \Bigg(\operatorname{tr}\Big(\mathbf{E}[x_t x_t^\top]\mathbf{E}\Big[(BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top\Big]\Big)\\
&\quad+ \operatorname{tr}\mathbf{E}\left[x_t x_t^\top (A-BK)(A-BK)^\top\right]\\
&\quad- \operatorname{tr}\mathbf{E}\left[\tilde{x}_t\tilde{x}_t^\top (A-BK)(A-BK)^\top\right]\Bigg)\\
&\quad\times (1+o(1)) + o((t\ln t)^{-1})\\
&\quad= \Big(\operatorname{tr}\Big(\mathbf{E}\left[x_t x_t^\top\right] \mathbf{E}\Big[(BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top\Big]\Big)\\
&\quad+ \operatorname{tr}\mathbf{E}\left[(x_t x_t^\top - \tilde{x}_t\tilde{x}_t^\top)(A-BK)(A-BK)^\top\right] \Big)\\
&\quad\times (1+o(1)) + o((t\ln t)^{-1}).
\end{aligned}$$

Above, we also used $o((t\ln t)^{-1}) \times (1+o(1)) = o((t\ln t)^{-1})$. ■

B. Decomposition

The term

$$\operatorname{tr}\Big(\mathbf{E}\left[x_t x_t^\top\right] \mathbf{E}\Big[(BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top\Big]\Big),$$

is a weighted estimation error for K. The second term

$$\operatorname{tr}\mathbf{E}\left[(x_t x_t^\top - \tilde{x}_t\tilde{x}_t^\top)(A-BK)(A-BK)^\top\right]$$

strikingly is linear in $\mathbf{E}x_t x_t^\top - \mathbf{E}\tilde{x}_t\tilde{x}_t^\top$ which corresponds to the regret of the previous stage and thus yields a lower bounding recursion. We now solve this recursion.

Lemma 3.2 (Regret Decomposition): Assume that (A, B) is stabilizable and that the policy $\hat{K}_t$ is asymptotically unbiased, consistently convergent and sample stable. For ρ the spectral radius of $A-BK$ and $\tau = \log\rho^2 \times \log t$, one has

$$\begin{aligned}
&\mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1}\\
&\quad= \Bigg(\sum_{k=t-\tau}^{t} \operatorname{tr}\Bigg\{ \left[(A-BK)(A-BK)^\top\right]^{t-k}\\
&\quad\times \mathbf{E}(BK-B\hat{K}_k)(BK-B\hat{K}_k)^\top \mathbf{E}[x_k x_k^\top]\Bigg\}\Bigg)\\
&\quad\times (1+o(1)) + o(t^{-1}).
\end{aligned}\tag{9}$$

Proof: By iterating Lemma 3.1 one obtains

$$\begin{aligned}
&\mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1} =\\
&\quad\Bigg(\sum_{k=t-\tau}^{t} \operatorname{tr}\Bigg\{ \left[(A-BK)(A-BK)^\top\right]^{t-k} \times\\
&\mathbf{E}(BK-B\hat{K}_t)(BK-B\hat{K}_t)^\top \mathbf{E}[x_k x_k^\top] + o(((t-k)\log k)^{-1})\Bigg\}\\
&\quad+ \operatorname{tr}\Bigg(\left[(A-BK)(A-BK)^\top\right]^{\tau}\\
&\quad\times \mathbf{E}[x_{t-\tau-1}x_{t-\tau-1}^\top - \tilde{x}_{t-\tau-1}\tilde{x}_{t-\tau-1}^\top]\Bigg)\Bigg) \times (1+o(1)).
\end{aligned}$$

Since $\sum_{k=t-\tau}^{t} o(((t-k)\log k)^{-1}) = o(t^{-1})$ by our choice of τ, this is (9) except for the final term. Next, note that

$$\mathbf{E}[x_{t-\tau-1}x_{t-\tau-1}^\top - \tilde{x}_{t-\tau-1}\tilde{x}_{t-\tau-1}^\top] = o(1).$$

Moreover, by stability of the optimal closed loop

$$\left[(A-BK)(A-BK)^\top\right]^\tau = O\left(\rho^{2\tau}\right)$$

for $\rho \in (0,1)$, the spectral radius of the optimal closed loop. Since $\log \rho^2 < 0$, and $\tau = -\log\rho^2 \times \log t$ we have $\rho^\tau = t^{-1}$. Now, our statement is true if $\rho \in (0,1)$ since then

$$o(1) \times O\left(\rho^\tau\right) = o(1)O(t^{-1}) = o(t^{-1}).$$

If $\rho = 0$, then $A - BK = 0$, so the proof is complete. ■

IV. Fundamental Limitations

Now that we have a decomposition for the state component of regret via (9), the stage is almost set for application of the Cramér-Rao inequality.

A. Applying the Information Inequality

If one attemps to compute the Fisher information $I_{\hat{K}_t}^t$ and differentiate with respect to A and B, one soon notices that (as should be) the information is policy dependent[5]. Fortunately, it is not hard to prove that any asymptotically efficient $\hat{K}_t$ yields a trajectory (x_t) of samples with asymptotically similar information to that corresponding to usage of the optimal policy K.

Lemma 4.1 (Information Comparison): Assume (A, B) is stabilizable and that $(\hat{K}_t)$ is consistently convergent. Then

$$I_{\hat{K}_t}^t = I_K^t + o(t) = t\bar{I}_K + o(t).$$

Proof: The first equality follows by continuity (which is a consequence of the differentiability hypothesis on the density), whereas the second follows by ergodicity of the Markov chain associated with the optimal closed loop $A - BK$, cf. (3). ■

We are now ready to state the main Theorem, which allows us to bound the regret pertaining the to the state of the system.

Theorem 4.2: Assume that (A, B) is stabilizable and that the policy $\hat{K}_t$ is linear, asymptotically unbiased, consistently convergent and sample stable. Then

$$\liminf_{T\to\infty} \frac{\sum_{t=0}^{T-1} \mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1}}{\ln T} \geq \mathrm{tr}\left(\left[I \otimes \left(B^\top \sqrt{\Gamma_K}\Sigma\sqrt{\Gamma_K}\Gamma_K B\right)\right](\bar{I}_K)^\dagger\right).$$

Proof: Observe that

$$\sum_{k=\lceil t/2 \rceil}^{t} \left[(A-BK)(A-BK)^\top\right]^{t-k} = \Gamma_K + o(1) \tag{10}$$

and similarly

$$\mathbf{E}x_k x_k^\top = \sqrt{\Gamma_K}\Sigma\sqrt{\Gamma_K} + o(1), \tag{11}$$

valid for $k \geq \lceil t/2 \rceil$. Since by monotonicity of Fisher information $(I_{\hat{K}}^k)^{-1} \succeq (I_{\hat{K}}^t)^{-1}$, applying Cramér-Rao and Lemma 4.1 we have that for any $k \geq \lceil t/2 \rceil$

$$\begin{aligned}\mathbf{E}\left[\mathrm{vec}(K-\hat{K}_k)\,\mathrm{vec}(K-\hat{K}_k)^\top\right] &\succeq (I_{\hat{K}}^k)^\dagger \times (1+o(1)) \\ &\succeq (I_{\hat{K}}^t)^\dagger \times (1+o(1)) = (t\bar{I}_K + o(t))^\dagger \times (1+o(1)).\end{aligned} \tag{12}$$

Thus, we have asymptotics or lower bounds for each quantity given by Lemma 3.2:

$$\begin{aligned}&\underbrace{\sum_{k=t-\tau}^{t} \mathrm{tr}\Bigg\{\left[(A-BK)(A-BK)^\top\right]^{t-k}}_{iii.\ \text{Apply (10).}} \\ &\times B\underbrace{\mathbf{E}(K-\hat{K}_k)(K-\hat{K}_k)^\top}_{ii.\ \text{Apply (12).}} B^\top \underbrace{\mathbf{E}[x_k x_k^\top]}_{i.\ \text{Apply (11).}}\Bigg\}\end{aligned} \tag{13}$$

Using first (4) to vectorize, and then following the above steps in order *i-iii.* together with the trace cyclic property,

$$\begin{aligned}&\mathbf{E}x_{t+1}^\top x_{t+1} - \mathbf{E}\tilde{x}_{t+1}^\top \tilde{x}_{t+1} \geq \\ &\mathrm{tr}\left(\left(I \otimes \left(B^\top \sqrt{\Gamma_K}\Sigma\sqrt{\Gamma_K}\Gamma_K B\right)\right)(t\bar{I}_K + o(t))^{-1})\right) \\ &\qquad\qquad \times (1 + o(1)) + o(t^{-1}).\end{aligned}$$

Summing this from $t = 0$ to $t = T$, dividing by $\ln T$ and then taking limits yields the result. ■

Example 4.3: Theorem 4.2 may be used to compute regret lower bounds for the cheap cost LQR by setting $Q = I$ and $R = 0$. Assume that B is equal to I and that the underlying noise distribution of the increments is Gaussian with covariance Σ. Then $\Gamma_K = I$ and one may compute

$$\bar{I}_K = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{E}x_t x_t^\top \otimes \Sigma^{-1} = \Sigma \otimes \Sigma^{-1}$$

since $\mathbf{E}x_t x_t^\top \to \Sigma$ for the optimal policy when B is invertible. The bound thus reduces to

$$\liminf_{T\to\infty} \frac{R_T}{\ln T} \geq n\,\mathrm{tr}\,\Sigma. \tag{14}$$

Note that this agrees with the bound in Lai's paper [6], where he also constructs a matching upper bound under slightly different assumptions. At least in this particular case, this suggests that our bound is tight.

Specializing further to the scalar system of Example 2.5, the bound reduces to $\liminf_{T\to\infty} R_T/\ln T \geq \sigma^2$ where σ^2 is the variance of the noise. A special case of Theorem 7.2 in [7] shows that the Åström-Wittenmark self-tuning regulator in [2] attains this bound. This is also true more generally for the SISO tracking problem in [7] with lag polynomials of degree p and q in the output and input respectively. The author there proves that under a persistency of excitation condition, $\limsup R_T/\ln T \leq (p+q)\sigma^2$ with σ^2 as above which agrees with our lower bound. This can easily be seen since a SISO system of order (p, q) has a state space representation of dimension $n = p + q$. Comparing with (14) again indicates that the bound is tight.

[5] This is essentially the consequence of the fact that an unknown B requires persistence of excitation to be identified, cf. [23].

Remark 4.4: Note that the LQR is comparable to the tracking problem in [7] since $\mathbf{E}(x_t - r_t)^2 - \mathbf{E}(\tilde{x}_t - r_t)^2 = \mathbf{E}x_t^2 - \mathbf{E}\tilde{x}_t^2$ whenever the reference signal r_t is deterministic. Since this is the key quantity providing our lower bound, the analysis can be carried out mutatis mutandis.

B. The Case with General Cost

By noting that $u_t, \tilde{u}_t$ can be expanded in terms of $x_t, \tilde{x}_t$, we can immediately generalize our lower bound in Theorem 4.2 to regret itself.

Theorem 4.5: Assume that (A, B) is stabilizable and that the policy $\hat{K}_t$ is linear, asymptotically unbiased, consistently convergent and sample stable. Then for any $Q \succ 0, R \succeq 0$,

$$\begin{aligned}&\liminf_{T\to\infty} \frac{R_T}{\ln T} \\ &\quad \geq \operatorname{tr}\Big(\Big[I \otimes \Big((Q + K^\top R K) B^\top \sqrt{\Gamma_K} \Sigma \sqrt{\Gamma_K} \Gamma_K B \Big) \Big] \\ &\qquad\qquad\qquad\qquad \times (\bar{I}_K)^\dagger \Big).\end{aligned}$$

Proof: Observe that

$$\begin{aligned}&x_t^\top Q x_t = \operatorname{tr}[Q x_t x_t^\top] \text{ and that} \\ &u_t^\top R u_t = \operatorname{tr}[\hat{K}_t^\top R \hat{K}_t x_t x_t^\top] = \operatorname{tr}[(K^\top R K + o(1)) x_t x_t^\top]\end{aligned}$$

and argue as in Theorem 4.2 by utilizing (13) and that essentially all terms are linear in the variables $x_t x_t^\top$. ■

Remark 4.6: As the example above shows, the bound can be written explicitly when $B = I$ and under an additional normality assumption on the noise. However, in general, the quantity $\bar{I}_K$ may depend on the distribution of the noise and the parameters of the system.

V. Discussion

As advertised, our bound depends on an information-theoretic quantity, $\bar{I}_K$, the average Fisher information, and a control-theoretic quantity, Γ_K, the Gramian pertaining to the optimal closed loop. Our lower bound thus captures the hardness both in terms of estimation and control. Moreover, by referring to special cases in the literature, we have been able to establish that in each of these cases our bound is tight. In general however, the design of matching upper and lower bounds for adaptive control of LQR remains open. Given Example 4.3, it would be interesting to see if one can find precise conditions on the system (A, B) such that the logarithmic lower bound presented here is attainable.

Notwithstanding, it should be said that our emphasis has not been on presenting the theory in the greatest possible generality but rather to push the viewpoint of regret in large state spaces as a sequence of estimation errors and illustrate how this viewpoint naturally leads to bounds analogous to those in the bandit literature. In particular, one should note that the assumption of the policy $\hat{K}_t$ being asymptotically unbiased is with loss of generality. The typical way to resolve this issue in statistics is to consider local asymptotic minimax bounds (see for instance [24]). Using this local asymptotic theory, it can for instance be shown that the Cramér-Rao inequality still holds in an asymptotic sense for a much more general class of estimators. Using this to weaken or remove our unbiasedness assumption could provide an exciting direction for future work.

References

[1] R. E. Kalman, “Design of self-optimizing control system,” *Trans. ASME*, vol. 80, pp. 468–478, 1958.

[2] K. J. Åström and B. Wittenmark, “On self tuning regulators,” *Automatica*, vol. 9, no. 2, pp. 185–199, 1973.

[3] B. Wittenmark, “Adaptive dual control methods: An overview,” in *Adaptive Systems in Control and Signal Processing 1995*, Elsevier, 1995, pp. 67–72.

[4] N. Matni, A. Proutiere, A. Rantzer, and S. Tu, “From self-tuning regulators to reinforcement learning and back again,” *arXiv preprint arXiv:1906.11392*, 2019.

[5] B. Recht, “A tour of reinforcement learning: The view from continuous control,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[6] T. Lai, “Asymptotically efficient adaptive control in stochastic regression models,” *Advances in Applied Mathematics*, vol. 7, no. 1, pp. 23–45, 1986.

[7] L. Guo, “Convergence and logarithm laws of self-tuning regulators,” *Automatica*, vol. 31, no. 3, pp. 435–450, 1995.

[8] A. Rantzer, “Concentration bounds for single parameter adaptive control,” in *2018 Annual American Control Conference (ACC)*, IEEE, 2018, pp. 1862–1866.

[9] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[10] Y. Abbasi-Yadkori and C. Szepesvári, “Regret bounds for the adaptive control of linear quadratic systems,” in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[11] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, “On optimality of adaptive linear-quadratic regulators,” *arXiv preprint arXiv:1806.10749*, 2018.

[12] H. Mania, S. Tu, and B. Recht, “Certainty equivalent control of LQR is efficient,” *arXiv preprint arXiv:1902.07826*, 2019.

[13] T. L. Graves and T. L. Lai, “Asymptotically efficient adaptive choice of control laws in controlled markov chains,” *SIAM journal on control and optimization*, vol. 35, no. 3, pp. 715–743, 1997.

[14] A. Garivier, P. Ménard, and G. Stoltz, “Explore first, exploit next: The true shape of regret in bandit problems,” *Mathematics of Operations Research*, vol. 44, no. 2, pp. 377–399, 2018.

[15] A. Garivier and O. Cappé, “The kl-ucb algorithm for bounded stochastic bandits and beyond,” in *Proceedings of the 24th annual conference on learning theory*, 2011, pp. 359–376.

[16] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015, vol. 75.

[17] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1.

[18] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[19] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[20] O. Kallenberg, *Foundations of modern probability*. Springer Science & Business Media, 2006.

[21] D. R. Cox and E. J. Snell, “A general definition of residuals,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 248–265, 1968.

[22] B. Efron and C. Stein, “The jackknife estimate of variance,” *The Annals of Statistics*, pp. 586–596, 1981.

[23] L. Ljung, “System identification: Theory for the user,” 1999.

[24] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.