Proceedings of Machine Learning Research vol 144:1–36, 2021

On Uninformative Optimal Policies in Adaptive LQR with Unknown B-Matrix

Ingvar Ziemann ZIEMANN@KTH.SE
Henrik Sandberg HSAN@KTH.SE
KTH Royal Institute of Technology

Abstract

This paper presents local asymptotic minimax regret lower bounds for adaptive Linear Quadratic Regulators (LQR). We consider affinely parametrized B-matrices and known A-matrices and aim to understand when logarithmic regret is impossible even in the presence of structural side information. After defining the intrinsic notion of an *uninformative optimal policy* in terms of a singularity condition for Fisher information we obtain local minimax regret lower bounds for such uninformative instances of LQR by appealing to van Trees' inequality (Bayesian Cramér-Rao) and a representation of regret in terms of a quadratic form (Bellman error). It is shown that if the parametrization induces an uninformative optimal policy, logarithmic regret is impossible and the rate is at least order square root in the time horizon. We explicitly characterize the notion of an *uninformative optimal policy* in terms of the nullspaces of system-theoretic quantities and the particular instance parametrization.

Keywords: Linear Quadratic Regulator, Adaptive Control, Regret, Fundamental Limitations, Fisher Information.

1. Introduction

Possibly given some structural side information, what is the asymptotic order of magnitude of regret for a fixed unknown instance of the linear quadratic regular (LQR)? We introduce a framework for thinking about this question in terms of the Fisher information matrix about an underlying parameter, θ, generated by the optimal policy. This information quantity captures how much we learn about the underlying parameter by playing optimally in each round. It is shown that when this information matrix, depending only on the optimal policy, satisfies a certain degeneracy property, logarithmic regret is impossible and the order of magnitude must be square-root in the observation horizon. Our argument relies on information comparison; we observe that a low regret policy must yield information about θ comparable to the optimal policy. That is, we use Fisher information to capture the *exploration-exploitation* trade-off in adaptive LQR in terms of regret lower bounds.

Moreover, this reliance on Fisher information allows us to give conditions for when structural side-information is insufficient for logarithmic regret. If accurate structural models are available to describe certain systems, one asks what the impact of such structure – side information – may be on the efficiency of learning algorithms. For instance, if we are controlling a networked system, we may want to impose graph structure. One can also imagine physical constraints imposing some symmetry or relation between certain coefficients.

Indeed, depending on the problem structure, it has been observed that there is a sharp regret phase-transition in learning linear quadratic regulators in its asymptotic scaling with the time horizon, T. There are several upper bound results in the $\sqrt{T}$ regime of regret when the entire system is unknown (Abbasi-Yadkori and Szepesvári, 2011; Mania et al., 2019; Cohen et al., 2019), with Simchowitz and Foster (2020) establishing also an up-to-logs matching lower bound. However, in the presence of certain side-information or structure, logarithmic rates are attainable (Faradonbeh et al., 2020a; Cassel et al., 2020). The question of precisely specifying what structure makes logarithmic regret possible however remains open. We attempt to adress this by showing when it is not.

Contribution. We provide a natural framework for capturing the phase transition in regret depending on the structure of a nominal instance of LQR and possible side information encoded by an affine map. When the Fisher information of the trajectory corresponding to optimal (stationary) policy for this instance satisfies a certain singularity condition, we say that the instance is *uninformative* (see Section 4.1). Assuming that the A-matrix is known and that $\mathsf{vec}\, B = L\theta + B_0$, we establish that uninformativeness is a sufficient condition for logarithmic regret to be impossible. Indeed, our Theorem 5 shows that for all uninformative instances, regret R_T (see (4)), satisfies for some constant $C(\theta, \varepsilon) > 0$

$$\limsup_{T\to\infty} \sup_{\theta' \in B(\theta, \varepsilon T^{-1/4})} \frac{R_T^{\pi}(\theta')}{\sqrt{T}} \geq C(\theta, \varepsilon). \tag{1}$$

To establish this result, a number of intermediate observations are made. First, we extend the exact (modulo terms $O(1)$) representation of regret as the cumulative Bellman error to the LQR setting in Lemma 11. We then interpret this Bellman error, which is a sum of quadratic forms, as a sequence of estimation variances. The minimax regret, the left hand side of (1), is then lower bounded by placing a suitable family of priors over shrinking subsets of $B(\theta, \varepsilon)$, to which we then apply Van Trees' inequality (Bayesian Cramér-Rao). This interpretation is formalized in Lemma 6, which lower bounds (1) by inverse Fisher information. The final observation is that Fisher information itself has a relationship to regret in LQR. Specifically, Lemma 9 establishes a principle of information comparison by spectral perturbation (Davis and Kahan, 1970; Wedin, 1972; Cai et al., 2018). We show that any policy which has low regret automatically has Fisher information comparable to the Fisher information corresponding to use of the optimal policy (which is singular if the instance is uninformative). Roughly speaking, for any policy π, this idea states that

$$\text{Fisher Information}(\pi) = \text{Fisher Information}(\pi^*) + O(\text{Regret})$$

where π^* is the optimal policy. Fisher information thus allows us to make precise the exploration-exploitation trade-off in adaptive LQR. In other words, the learner has to combat the singularity in Fisher information by adding excitation, however, since regret bounds information, this extra excitation necessarily has non-trivial contribution to regret. We believe this gives an appealing framework for understanding regret in online control.

1.1. Related Work

The problem of adaptively controlling an unknown instance of a linear quadratic system has a rather long history and dates back to at least Åström and Wittenmark (1973). Early works (Goodwin et al., 1981; Lai et al., 1982; Campi and Kumar, 1998) only asked that the adaptive algorithm be asymptotically optimal on average, which in our modern language can be rephrased as having sublinear regret. Historically, in this setting the emphasis on regret minimization as a means to study the performance of an adaptive algorithm appears first in the works by Lai (1986) and Lai and Wei (1986). See also Guo (1995). Noteably, in these last three works regret is shown to scale logarithmically with time, albeit subject to rather strong structural conditions.

These logarithmic rates are of course in stark contrast with the current trend, where the emphasis has been on establishing bounds in the regime $\sqrt{T}$ (and with high probability). The present incarnation of this problem, the adaptive LQR model, was essentially popularized by Abbasi-Yadkori and Szepesvári (2011) in which the authors produced an algorithm with $\tilde{O}(\sqrt{T})$ regret. A line of work following that publication focuses on improving and providing more computationally tractable algorithms in this setting (Ouyang et al., 2017; Dean et al., 2018; Abeille and Lazaric, 2018; Abbasi-Yadkori et al., 2019; Mania et al., 2019; Cohen et al., 2019; Faradonbeh et al., 2020b; Abeille and Lazaric, 2020). The emphasis of these works is entirely on providing upper bounds. Recently, some effort has been made to understand the complexity of the problem in terms of lower bounds. In particular, Simchowitz and Foster (2020) provides matching (modulo constants and log factors) upper and lower bounds scaling correctly with the dimensional dependence given that the entire set of parameters (A, B) are unknown. The works of Cassel et al. (2020) and Ziemann and Sandberg (2020a) are also interesting in that they, for certain specific cases, provide $\sqrt{T}$ lower bounds that take the structure of the problem into account. Further, Cassel et al. (2020) shows that when the A-matrix is known and the optimal policy satisfies a certain non-degeneracy condition logarithmic rates are in fact achievable and a similar observation is made by Ziemann and Sandberg (2020a) in the context of memoryless systems.

Our lower bound shares some features with that in Simchowitz and Foster (2020), such as dimensional dependence, which also considers a local minimax version of regret. However, their bounds do not apply to the situation in which the learner is presented with structural side-information. Our proof approach is different and relies on Van Trees' inequality (van Trees, 2004; Bobrovsky et al., 1987). This necessarily involves the Fisher information, which, quite naturally, allows for taking problem structure into account by considering different parametrizations of the problem dynamics. Cramér-Rao type bounds have previously been used to establish regret lower bounds in adaptive LQR (Ziemann and Sandberg, 2020a,b). However, these papers consider either restricted (memoryless) structure or do not take into account the possible singularity of Fisher information, which very much drives our result. We also note that the idea to bound a minimax complexity by a suitable family of Bayesian problems is well-known in the statistics literature (Gill et al., 1995). See also van der Vaart (2000); Tsybakov (2008); Ibragimov and Has'minskii (2013) and the references therein.

Indeed, the adaptive control problem is intimitely connected to parameter estimation (Polderman, 1986). From the outset algorithm design has to a large extent been based on certainty equivalence; that is, estimating the parameters and plugging these estimates into an optimality equation, as if they were the ground truth (Åström and Wittenmark, 1973; Mania et al., 2019). Our lower bound condition, *uninformativeness*, is related to identifiability and inspired by a similar phenomenon in point estimation, which may become arbitrarily hard when Fisher information is singular (Rothenberg, 1971; Goodrich and Caines, 1979;

Stoica and Söderström, 1982; Stoica and Marzetta, 2001). We also note that non-singularity of Fisher information is strongly related to the size of the smallest singular value of the covariates matrix in linear system identification (Faradonbeh et al., 2018; Simchowitz et al., 2018; Sarkar and Rakhlin, 2019; Jedra and Proutiere, 2020), which actually quantifies the corresponding rate of convergence (Jedra and Proutiere, 2019).

Low regret linear quadratic control also fits into a wider context of online decision-making. Lai and Robbins (1985); Burnetas and Katehakis (1996) solve asymptotically a set of problems known as bandits. Similar to LQR, depending on the problem structure and regret definition, the complexity of these problems also switches between $\log T$ and $\sqrt{T}$ (Shamir, 2013). Bandits are in some sense memoryless Markov Decision Processes (MDP). For finite state and action spaces, Burnetas and Katehakis (1997), characterizes instance specific regret for MDPs. See also Ok et al. (2018) and the referencs therein. Returning to the control context, there have also been recent advances in the partially observed adaptive Linear Quadratic Gaussian (LQG) setting, (Lale et al., 2020) and unknown cost LQR with adversarial disturbances (Hazan et al., 2020). We note that logarithmic rates reappear in the partially observed case due to certain excitation properties of the output sequence. Moreover, there is an interesting parallel line of work which emphasizes robustness in adaptive LQR (Dean et al., 2019; Umenberger et al., 2019). See also Recht (2019); Matni et al. (2019) and the references therein for an overview of the relationship between control and learning.

1.2. Outline

We first define the problem in Section 2. Section 3 relates regret to sub-optimal solutions of the Bellman equation. Section 4 reviews Fisher information and studies *uninformativeness*, which is the key notion used in Section 5 to derive our regret lower bound. Finally Section 6 concludes. All proofs can be found in the arXiv version of this paper (Ziemann and Sandberg, 2020c), which includes our appendices.

Notation. We use $\succeq$ (and $\succ$) for (strict) inequality in the matrix positive definite partial order. By $\|\cdot\|$ we denote the standard 2-norm by $\|\cdot\|_\infty$ the matrix operator norm (induced $l^2 \to l^2$) and $\rho(\cdot)$ denotes the spectral radius. Moreover, $\otimes$, vec and $\dagger$ are used to denote the Kronecker product, vectorization (mapping a matrix into a column vector by stacking its columns), and the Moore-Penrose pseudoinverse, respectively. For a sequence of vectors $\{v_t\}_{t=1}^n, v_t \in \mathbb{R}^d$ we use $v^n = (v_1, \dots, v_n)$ defined on the n-fold product $\mathbb{R}^{d\times n}$. The set of k-times continuously differentiable functions on $\mathbb{R}^d$ is denoted $C^k(\mathbb{R}^d)$. We use D for Jacobian, d for differential and ∇ for the gradient. That is, for a scalar function, ∇f denotes a column vector of its first derivatives. We write $\mathbf{E}$ for the expectation operator, with superscripts indicating policy, and subscripts indicating parameters.

2. Problem Formulation

Fix an unknown parameter $\theta \in \Theta$ where Θ is an open subset of $\mathbb{R}^{d_\theta}$. Let $\{x_t\}$ be a controlled process on $\mathbb{R}^{d_x}$ with dynamics

$$x_{t+1} = Ax_t + B(\theta)u_t + w_t, \qquad x_0 = 0, \qquad t = 0, 1, 2, \dots \tag{2}$$

with control process $\{u_t\}$ on $\mathbb{R}^{d_u}$, additive noise process $\{w_t\}$ on $\mathbb{R}^{d_x}$ so that $A \in \mathbb{R}^{d_x \times d_x}$ and $B = B(\theta) \in \mathbb{R}^{d_x \times d_u}$. It is assumed that B depends affinely on θ, to be made precise momentarily. We consider the stage cost function $c(x, u) = x^\top Qx + u^\top Ru$, with $Q \in \mathbb{R}^{d_x \times d_x}, R \in \mathbb{R}^{d_u \times d_u}$. We further denote the σ-field generated by $x_1, \dots, x_t$ and possible auxilliary randomization by $\mathcal{F}_t$. The adaptive control objective is to design a policy π that minimizes the cumulative cost

$$V_T^\pi(\theta) = \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi c(x_t, u_t) = \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \left[x_t^\top Q x_t + u_t^\top R u_t \right] \tag{3}$$

without a priori knowledge of the parameter θ. This paper studies the fundamental limitations to this problem. We make the following standing assumptions about (2)-(3).

A1. $\{w_t\}$ is iid $p(\cdot)$ with $\mathbf{E} w_t w_t^\top = \Sigma_w$ and $p(\cdot) \in C^1(\mathbb{R}^{d_x})$ with finite Fisher information.

A2. The cost function is strongly convex. More precisely, $Q \succ 0$ and $R \succ 0$.

A3. The dynamics (2) are stabilizable at θ; there exists $K(\theta)$ with $\rho(A - B(\theta)K(\theta)) < 1$.

A4. $B(\theta)$ is affine in θ; $\operatorname{\mathsf{vec}} B(\theta) = L\theta + \operatorname{\mathsf{vec}} B_0$ for matrices $L \in \mathbb{R}^{d_x d_u \times d_\theta}, B_0 \in \mathbb{R}^{d_x \times d_u}$.

We say that a tuple $(\theta, A, B(\cdot), Q, R, p)$ satisfying the above assumptions is a parametrized instance of LQR. Our goal will be to devise regret lower bounds in terms of the nominal instance, θ, and the structure of the parametrization, i.e, in terms of the matrix L. To motivate why this flexibility in L is beneficial, consider the following example.

Example 1 *Consider the system (2) and suppose that it is known that B has network strucutre, say, it is the Laplacian of some known graph* `G` *with unknown weights. For simpicity let us assume that $d_x = d_u = 3$ and that the graph adjacency matrix is given by*

$$\texttt{Adj} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

so that nodes 1 and 2 and 1 and 3 are connected by an edge. We assume that this topological information is available to the user. The Laplacian is thus of the form

$$B = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_2 & \theta_4 & 0 \\ \theta_3 & 0 & \theta_5 \end{bmatrix} \text{ with } \operatorname{\mathsf{vec}} B = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_2 & \theta_4 & 0 & \theta_3 & 0 & \theta_5 \end{bmatrix}^\top.$$

Clearly, $\operatorname{\mathsf{vec}} B = L\theta$, where $\theta \in \mathbb{R}^5$ for some matrix $L \in \mathbb{R}^{9 \times 5}$. From an identification perspective, one thus suspects that the side-information of knowing the network topology reduces the hardness of the adaptive control task. Namely, the number of parameters to be identified is only $d_\theta = 5$ instead of $9 = \text{number of entries}(B)$.

Optimality and Regret. The objective (3) is equivalent to minimizing the regret

$$R_T^\pi(\theta) = V_T^\pi(\theta) - V_T^*(\theta) \tag{4}$$

where $V_T^*(\theta) = V_T^{\pi^*}(\theta)$ is the cumulative cost corresponding to the optimal policy $\pi^*(\theta)$, defined implicitly by the Riccati recursion (Bertsekas et al., 1995).

3. Regret and Cumulative Bellman Error

One expects that $V_n^*(\theta) \approx V_n^{\pi^\infty}(\theta)$ where π^∞ instead of relying on the Riccati recursion relies on its stationary limit. The stationary policy π^∞ is then characterized by minimization of the quadratic form

$$\phi(x,u,\theta) = x^\top Qx + u^\top Ru + [A(\theta)x + B(\theta)u]P(\theta)[A(\theta)x + B(\theta)u)] - x^\top P(\theta)x.$$

If $P(\theta)$ solves the discrete algebraic Riccati equation $\min_{u \in \mathbb{R}^{d_u}} \phi(x,u,\theta) = 0$ is equivalent to the average cost Bellman equation for this problem (Bertsekas et al., 1995). Our first theorem exploits the quadratic nature of ϕ to transform the problem essentially into one of sequential estimation.

Theorem 1 *Under assumptions A1-A4 it holds that*

$$\begin{aligned} R_T^\pi(\theta) &= \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \left(\phi(x_t,u_t,\theta)\right) + O(1) \\ &= \sum_{t=0}^{T-1} \mathbf{E}_\theta^\pi \Big((u_t - K(\theta)x_t)^\top \big[R + B^\top(\theta)P(\theta)B(\theta)\big](u_t - K(\theta)x_t)\Big) + O(1) \qquad (5) \end{aligned}$$

where the term $O(1)$ is uniformly bounded in the stability region of the optimal stationary policy $K(\theta)$, i.e. on each nonempty set $\{\theta' \in \mathbb{R}^{d_\theta} | \rho(A - B(\theta')K(\theta)) \leq 1-\zeta\}$, $\zeta \in (0,1)$.

See Appendix B for the proof. Note that such a neighborhood always exists for some ζ by upper semi-continuity of the spectral radius and stability of $A - B(\theta)K(\theta)$. The proof idea behind Theorem 1 is due to Burnetas and Katehakis (1997). Similar expressions are also obtained in e.g. Faradonbeh et al. (2020a); Simchowitz and Foster (2020).

Notice that when the choice of u_t is made x_t is known. Hence (5) informs us that regret minimization is essentially equivalent to minimizing a cumulative weighted estimation error for the sequence of estimands $K(\theta)x_t$. To be clear, our perspective is that we wish to estimate the function value of $\theta \mapsto K(\theta)x_t$ where the function $K(\theta)x_t$ is revealed at time t by virtue of observations of the x_t. Moving to a Bayesian setting, the entire trajectory (x^{T+1}, u^T) is then interepreted as a noisy observation of the underlying parameter θ. A natural approach for variance lower bounds is to rely on Fisher information and use Cramér-Rao type bounds.

4. Fisher Information Theory

Let us recall the definition of Fisher information. For a parametrized family of probability densities $\{q_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^d$, Fisher information $\mathtt{I}_p(\theta) \in \mathbb{R}^{d \times d}$ is

$$\mathtt{I}_q(\theta) = \int \nabla_\theta \log q_\theta(x) \left[\nabla_\theta \log q_\theta(x)\right]^\top q_\theta(x) dx \qquad (6)$$

whenever the integral exists. For a density λ, we also define the location integral

$$\mathtt{J}(\lambda) = \int \nabla_\theta \log \lambda(\theta) \left[\nabla_\theta \log \lambda(\theta)\right]^\top \lambda(\theta) d\theta. \qquad (7)$$

Again, provided of course that the integral exists. See Ibragimov and Has'minskii (2013) for details about these integrals and their existence.

We now study Fisher information where q in (6) is the joint density of x^{T+1}.

Lemma 2 *Under Assumption 1, Fisher information about θ given observations of (x^{T+1}, u^T) in the model (2) is given by*

$$\mathtt{I}^T(\theta; u^T) = \mathbf{E}\sum_{t=0}^{T}[\mathsf{D}_\theta[B(\theta)u_t]]^\top\, \mathsf{J}(p)\, \mathsf{D}_\theta[B(\theta)u_t]. \tag{8}$$

See appendix C for the proof. We now turn to investigating Fisher information corresponding to observations generated by the optimal policy $u_t = K(\theta)x_t$. It will be especially interesting to study when (8) becomes singular.

4.1. Uninformative Optimal Policies

Naively, the perspective discussed in Section 3 viewing (5) as a cumulative estimation error suggests a lower bound on the scale $\log T$, since one might think that the errors variances should decay as $1/t$. However, when Fisher information is singular, this reasoning might fail.

We say that an instance $(\theta, A, B(\cdot), Q, R, p)$ is *uninformative* if

1. $\mathtt{I}_*^T(\theta) = \mathbf{E}\sum_{t=0}^{T}[\mathsf{D}_\theta[B(\theta)K(\theta)x_t]]^\top\, \mathsf{J}(p)\, \mathsf{D}_\theta[B(\theta)K(\theta)x_t]$ is singular for all T under closed loop dynamics

$$x_{t+1} = (A - B(\theta)K(\theta))x_t + w_t; \text{ and}$$

2. There exists a vector $\tilde{v}$ in the nullspace of $\mathtt{I}^T(\theta)$ such that $\mathsf{D}_\theta\, \mathsf{vec}\, K(\theta)\tilde{v} \neq 0$.

In other words, uninformativeness stipulates that observation of an optimally regulated example is not sufficient to (locally) identify the optimal policy.

Algebraic Charaterization of Uninformativeness. The above description of what constitutes an uninformative optimal policy is somewhat indirect. We can characterize uninformativeness directly in terms of the instance parameters.

Proposition 3 *The instance $(\theta, A, B(\cdot), Q, R, p)$ is uninformative if and only if there exists a vector $\tilde{v}$ such that*

$$\begin{cases} \tilde{v} & \in \ker L^\top[KK^\top \otimes \mathsf{J}(p)]L \\ \tilde{v} & \notin \ker \mathsf{D}_\theta\, \mathsf{vec}\, K(\theta) \end{cases} \tag{9}$$

where L is such that $\mathsf{vec}\, B = L\theta + B_0$.

The proof is given in Appendix E.2. Any subspace $\mathtt{U}$ of maximal dimension for which all nonzero $\tilde{v} \in \mathtt{U}$ satisfy (9) is called an *information singular subspace*. We will later see that the dimension of any such subspace gives the dimensional dependence in our lower bound.

As Cassel et al. (2020) show that logarithmic regret rates for known A-matrix and unknown B-matrix are attainable if $KK^\top$ is nonsingular, it is interesting to note that our notion of uninformativeness immediately rules out the possibility of nonsingular $KK^\top$.

Corollary 4 *If the instance $(\theta, A, B(\cdot), Q, R, p)$ is uninformative with $\mathsf{vec}\, B = \theta$ so that $L = I_{d_\theta}$, then $KK^\top$ is singular.*

The following examples illustrate the concept in terms of certain simple model structures.

Example 2 *Consider a "scalar" LQR, with nonzero $A = a \in \mathbb{R}$ known, and $B = \theta \in \mathbb{R}$ unknown. Since the optimal linear feedback law is 0 if and only if $\theta = 0$, it follows that scalar LQR is uninformative if and only if the input matrix is $B = \theta = 0$. Notice that scalar $B \approx 0$ is precisely the construction used in the lower-bound proof of Cassel et al. (2020).*

In the next example, we illustrate that one can explicitly compute the dimension of $\mathtt{U}$, the number of parameters not excited by the optimal policy.

Example 3 *Consider a "memoryless" linear quadratic regulator (c.f. Ziemann and Sandberg (2020a))*

$$\begin{bmatrix} r_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} G & 0 \\ I & 0 \end{bmatrix} \begin{bmatrix} r_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0 \\ -F \end{bmatrix} u_t + \begin{bmatrix} n_t \\ v_t \end{bmatrix}, \tag{10}$$

with $R = I_{d_u}, Q = I_{d_x}$. Due to the memoryless property in the second x-coordinate y_t, the optimal regulation of this instance can be reduced to

$$\begin{cases} y_{t+1} = r_t - Fu_t + v_t \\ \min_{\{u_t\}} \mathbf{E} \sum_{t=1}^{T} \|r_t - Fu_t\|^2 + \|u_t\|^2 \end{cases} \tag{11}$$

which has optimal policy in feedforward form $u_t = (F^\top F + I_{d_u})^{-1} F^\top r_t$. Suppose that $\mathsf{vec}\, F = \theta$, so that the structure (10) is known. It can be shown this instance is uninformative if and only if $KK^\top$ is singular. Moreover, $\dim \mathtt{U} = d_y \times \dim \ker KK^\top$. See the appendix for the proof.

Note that in this case, the condition agrees exactly with that in Cassel et al. (2020).

5. Main Result

Our main result is a *local* asymptotic minimax regret lower bound. It considers a shrinking neighborhood in parameter space around θ and states that if the nominal instance θ is *uninformative* any policy π suffers regret on the order of magnitude $\sqrt{T}$ on at least one instance of this neighborhood.

Theorem 5 *Assume that A1-A4 hold and suppose that the nominal instance $(\theta, A, B(\cdot), Q, R, p)$ is uninformative. Then for any $\Gamma \succ 0$ and any $\varepsilon > 0$ such that $B(\theta, \varepsilon) \subset \{\theta' \in \mathbb{R}^{d_\theta} | \rho(A - B(\theta)K(\theta')) < 1\}$, any policy π satisfies*

$$\limsup_{T \to \infty} \sup_{\theta' \in B(\theta, \varepsilon T^{-1/4})} \frac{R_T^\pi(\theta')}{\sqrt{T}} \geq C(\theta, \varepsilon, \Gamma) \tag{12}$$

where

$$C(\theta,\varepsilon,\Gamma) = \inf_{C'>0} \max\left\{ p^2 \operatorname{tr}\left(\frac{\Gamma}{F(\theta,\varepsilon,C')} \otimes \left[R + B^\top(\theta)P(\theta)B(\theta)\right] \times \mathsf{D}_\theta \operatorname{\mathsf{vec}} K(\theta)\tilde{W}_0\tilde{W}_0^\top [\mathsf{D}_\theta \operatorname{\mathsf{vec}} K(\theta)]^\top \right), C' \right\} \tag{13}$$

where $p = \liminf \mathbf{P}^\pi_{\theta'\sim\lambda}\left(\left\{\sum_{t=k\lceil\sqrt{T}\rceil}^{(k+1)\lfloor\sqrt{T}\rfloor} x_t x_t^\top \succeq \sqrt{T}\Gamma\right\}\right)$*, and where* $\tilde{W}_0$ *is an orthonormal matrix with columns spanning an eigenspace of dimension* $\lceil \dim \mathtt{U}/2\rceil$ *of* $\mathtt{U}$*, and further*

$$F(\theta,\varepsilon,C') = \|L\|_\infty^2 \|Q^{-1}\|_\infty g^{\pi^*}(\theta) \frac{\operatorname{tr}\mathsf{J}(p)}{\lceil \dim \mathtt{U}/2\rceil} \|\,\mathsf{D}_\theta^2 \operatorname{tr} K(\theta)K^\top(\theta)\|_\infty \times \frac{\varepsilon^2}{2} + 2\|L\|_\infty^2 \|[R + B^\top(\theta)P(\theta)B(\theta)]^{-1}\|_\infty \frac{\operatorname{tr}\mathsf{J}(p)}{\lceil \dim \mathtt{U}/2\rceil} C' + \|\,\mathsf{J}(\lambda)\|_\infty$$

where the infimum with respect to π *is taken subject to* $\limsup_{T\to\infty} \frac{R_T^\pi(\theta)}{\sqrt{T}} \leq C'$ *and* F *is defined for any* $\lambda \in C_c^\infty[B(\theta,\varepsilon)]$*.*

The bound can be optimized in terms of the state covariance proxy Γ. For instance, the choice $\Gamma = \Sigma_w$ results in $p = 1$ in (13). The quotient $\operatorname{tr}\Gamma / F(\theta,\varepsilon,C_\pi)$ is approximately the trace of the state (observation) variance divided by an upper bound for inverse normalized Fisher information and as such can be interpreted as a signal-to-noise ratio (SNR). The optimization problem (13) can thus be understood as to balance high SNR with good control performance. Indeed, this balancing constitutes one of the key observations leading to the proof of Theorem 5 (Appendix A); if regret is $O(C_\pi\sqrt{T})$, the dominant component of $F(\theta,\varepsilon,C_\pi)$ is $O(C_\pi)$.

We also note in passing that together with the scalar case, Example 2, we immediately see that the global minimax regret, also with unknown A-matrix, is at least of order $\sqrt{T}$.

Dimensional Depedence. Observe that the optimization problem (13) has value asymptotically proportional to

$$\sqrt{\operatorname{tr}\left(\underbrace{\left[\Gamma \otimes \left[R + B^\top(\theta)P(\theta)B(\theta)\right]\right]}_{\succ 0} \mathsf{D}_\theta \operatorname{\mathsf{vec}} K(\theta)\tilde{W}_0\tilde{W}_0^\top [\mathsf{D}_\theta \operatorname{\mathsf{vec}} K(\theta)]^\top\right)} \Bigg/ \sqrt{\frac{\operatorname{tr}\mathsf{J}(p)}{\dim \mathtt{U}}}$$
$$\propto \sqrt{\operatorname{tr}\left(\mathsf{D}_\theta \operatorname{\mathsf{vec}} K(\theta)\tilde{W}_0\tilde{W}_0^\top [\mathsf{D}_\theta \operatorname{\mathsf{vec}} K(\theta)]^\top\right)} \Bigg/ \sqrt{\frac{\operatorname{tr}\mathsf{J}(p)}{\dim \mathtt{U}}}$$
$$\propto \sqrt{\frac{(\dim \mathtt{U})^2}{d_x}} \tag{14}$$

where $\mathtt{U}$ is the information singular subspace (9). That is, the dimensional depedence is proportional to the size of the parameter subspace not excited by the optimal policy relevant for identification of $K(\theta)$. Notice further that if the entire matrix B is unknown, $\mathtt{U}$ is a subspace of $\mathbb{R}^{d_x d_u}$ which further relates to the dimensions of the problem.

Outline of the Proof of Theorem 5 We provide a brief outline of the proof, of which the details are given in Appendix A. The following steps are key:

- We relate regret to a sequential estimation problem via cumulative Bellman error. This is the content of Theorem 1.
- We consider the sup of regret over a shrinking neighborhood. This sup is relaxed to a Bayesian estimation problem for an increasingly concentrated sequence of priors to which we apply van Trees' inequality.
- The role of these increasingly concentrated priors is to allow us to consider the local properties of Fisher information on the domain of the prior.
- A truncation argument is used to decouple conditional Fisher information from the state trajectory.
- We establish a principle of information comparsion; small regret implies that the Fisher information of that policy is near singular under the hypothesis that the optimal policy is uninformative. This implies large estimation error and thus large regret. The proof of this principle relies on the Davas-Kahan $\sin\theta$ Theorem (Davis and Kahan, 1970).

6. Discussion and Conclusion

We have established a framework which is able to both qualitatively and quantitatively explain why $\sqrt{T}$ regret occurs in certain instances of adaptive linear quadratic control. Essentially, if the optimal policy does not sufficiently excite all directions of the underlying parameter space necessary for identification of that same optimal policy, logarithmic regret is impossible. We make this explicit by a singularity condition on that optimal policy's Fisher information which takes problem structure into account. Indeed, singular information has deep connections to classical notions of identifiability (Rothenberg, 1971; Goodrich and Caines, 1979; Stoica and Söderström, 1982; Stoica and Marzetta, 2001), and we believe that further investigation of this relation would be an interesting line of future work. An extension to unknown A-matrix is of course also highly relevant.

Moreover, the tightness of our notion remains an open question. Some preliminary evidence is provided by the results of Cassel et al. (2020) which establish logarithmic regret under the nonsingularity condition $KK^\top \succ 0$ in the case of known A-matrix. Our Corollary 4 is consistent with their observations. This opens an exciting problem: to prove or disprove whether logarithmic regret is attainable if and only if the instance is not uninformative.

Acknowledgments

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 express their gratitude to Yishao Zhou for valuable input.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117, 2019.

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. *Proceedings of Machine Learning Research*, 80, 2018.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. *arXiv preprint arXiv:2007.06482*, 2020.

Brian DO Anderson and John B Moore. *Optimal filtering*. Courier Corporation, 2012.

Karl Johan Åström and Björn Wittenmark. On self tuning regulators. *Automatica*, 9(2): 185–199, 1973.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Ben-Zion Bobrovsky, E Mayer-Wolf, and M Zakai. Some classes of global cramér-rao bounds. *The Annals of Statistics*, pages 1421–1438, 1987.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

T Tony Cai, Anru Zhang, et al. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

Marco C Campi and PR Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 2020.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear–quadratic regulators. *Automatica*, 117:108982, 2020a.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive control and learning. *Automatica*, 117:108950, 2020b.

Richard D Gill, Boris Y Levit, et al. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, 1(1-2):59–79, 1995.

R Goodrich and P Caines. Necessary and sufficient conditions for local second-order identifiability. *IEEE Transactions on Automatic Control*, 24(1):125–127, 1979.

Graham C Goodwin, Peter J Ramadge, and Peter E Caines. Discrete time stochastic adaptive control. *SIAM Journal on Control and Optimization*, 19(6):829–853, 1981.

Lei Guo. Convergence and logarithm laws of self-tuning regulators. *Automatica*, 31(3): 435–450, 1995.

Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pages 408–421. PMLR, 2020.

Il'dar Abdulovich Ibragimov and Rafail Zalmanovich Has'minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2676–2681. IEEE, 2019.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems: Optimality of the least-squares estimator. *arXiv preprint arXiv:2003.07937*, 2020.

TL Lai. Asymptotically efficient adaptive control in stochastic regression models. *Advances in Applied Mathematics*, 7(1):23–45, 1986.

Tze Lai and Ching-Zong Wei. Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control*, 31 (10):898–906, 1986.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *arXiv preprint arXiv:2003.11227*, 2020.

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3724–3740. IEEE, 2019.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8874–8882, 2018.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

Jan Willem Polderman. On the necessity of identifying the true parameter in adaptive lq control. *Systems & control letters*, 8(2):87–91, 1986.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

Thomas J Rothenberg. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591, 1971.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018.

Petre Stoica and Thomas L Marzetta. Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing*, 49(1):87–90, 2001.

Petre Stoica and Torsten Söderström. On non-singular information matrices and local identifiability. *International Journal of Control*, 36(2):323–329, 1982.

Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008.

Jack Umenberger, Mina Ferizbegovic, Thomas B Schön, and Håkan Hjalmarsson. Robust exploration in linear quadratic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 15336–15346, 2019.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Harry L van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory.* John Wiley & Sons, 2004.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Ingvar Ziemann and Henrik Sandberg. On a phase transition of regret in linear quadratic control: The memoryless case. *IEEE Control Systems Letters*, 5(2):695–700, 2020a.

Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for unbiased adaptive control of linear quadratic regulators. *IEEE Control Systems Letters*, 4(3):785–790, 2020b.

Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive lqr with unknown b-matrix. *https://arxiv.org/abs/2011.09288*, 2020c.