On a Phase Transition of Regret in Linear Quadratic Control: The Memoryless Case

Ingvar Ziemann, Henrik Sandberg

On a Phase Transition of Regret in Linear Quadratic Control: The Memoryless Case

Ingvar Ziemann, Henrik Sandberg

***Abstract*— We consider an idealized version of adaptive control of a multiple input multiple output (MIMO) system without state. We demonstrate how rank deficient Fisher information in this simple memoryless problem leads to the impossibility of logarithmic rates of regret. Our analysis rests on a version of the Cramér-Rao inequality that takes into account possible ill-conditioning of Fisher information and a pertubation result on the corresponding singular subspaces. This is used to define a sufficient condition, which we term *uniformativeness*, for regret to be at least order square root in the samples.**

I. Introduction

Recently, there has been a revitilization of interest in the adaptive linear quadratic regulator (LQR) as it serves as good theoretically tractable example of reinforcement learning in continuous state and action spaces, [1]-[2]. Much progress has been made toward analyzing the statistical convergence rate, the *regret* incurred, of adaptive algorithms. Several works over the past decade, [3], [4] and [5], have been able to prove *upper bounds* on the regret at a rate of approximately $\sqrt{T}$ in the time horizon. However, in some special cases, [6], [7] [8] and [9], the authors have actually been able to prove regret to scale at a rate of $\log T$, ensuring considerably faster convergence. In particular, [7] and [8] show that, suitably modified, the Åström-Wittenmark self-tuning regulator [10] for SISO tracking problems converges at the rate $\log T$. Given these two very different rates, it is thus natural to ask whether regret undergoes a phase transition in its asymptotic scaling. Here, we consider a simplified, and memoryless, version of the linear quadratic problem to verify that such a phenomenon indeed occurs. The point of such an analysis, as presented here, is to isolate the essence of this phase transition. With this in mind, Our goal is to identify and give conditions for when the *lower bound changes* from order $\log T$ to $\sqrt{T}$.

a) Contribution: As hinted above, the main contribution of this note is to establish a sufficient condition, *uninformativeness*, for regret to necessarily scale on the order $\sqrt{T}$, see Definition 4.4 and Theorem 5.2. We will see that this phase transition depends both on the rank of the optimal linear feedforward matrix and on the excitation of the reference signal. In the uninformative regime, there is an asymptotically non-neglible trade-off between exploration and exploitation. These results partially answer an unresolved question in the literature [11], [12], as to precisely when logarithmic rates are attainable for linear quadratic problems by showing when they are not. On the theoretical side, the method of analysis to provide lower bounds here is novel and rests on comparing the Fisher information of any policy to the Fisher information of the optimal policy having knowledge of the system's parameters using singular subspace perturbation theory. This clarifies some connections between regret and parameter estimation ([11], [13]) via the spectral properties of the Fisher information. We believe this proof strategy to potentially be of general interest. Moreover, we are among the first to record a $\sqrt{T}$ lower bound for linear quadratic adaptive control problems.

b) Related Work: After the initial submission of this manuscript the authors became aware of the results by [14] and [15] which independently arrive at $\sqrt{T}$ lower bounds for similar problems. It is particularly interesting to compare our work to [14] which, as our work, contains an instance specific bound. There, it is shown that regret for an unknown LQR scales as $\dim(\text{state}) \times \dim(\text{inputs})^2 \times \sqrt{T}$. By contrast, we consider a memoryless tracking problem where $\dim(\text{state}) = 0$ and the above results are not applicable. In particular, when there is a reference signal to track, we show that regret can be of order $\sqrt{T}$ even if the dimension of the state is zero. We also remark that our proof is very different from those in [14] and [15] and has the appealing property of generating lower bounds directly dependent on an information quantity, the Fisher information.

c) Notation: We use $\succeq$ (and $\succ$) for (strict) inequality in the matrix positive definite partial order. By $\|\cdot\|$ we denote the standard 2-norm and by $\|\cdot\|_\infty$ the matrix operator norm. Moreover, $\otimes$, vec and $\dagger$ are used to denote the Kronecker product, vectorization (mapping a matrix into a column vector), and the Moore-Penrose pseudoinverse, respectively. We use ∇ for gradient or Jacobian. All vectors, including gradients, are in this paper represented as column vectors. For two functions f, g, $\limsup |f(t)/g(t)| = 0$, for some norm $|\cdot|$, is written as $f = o(g)$. If instead $\limsup |f(t)/g(t)| \leq C, C > 0$, we write $f = O(g)$. Asymptotic lower bounds are written as $f = \Omega(g)$ which means that $\liminf f(t)/g(t) \geq c, c > 0$. In general, these limits will be for large times, usually indexed by t or T. For stochastic quantities, we also use the notation o, O, in which case it corresponds to convergence in probability. We write $\mathbf{E}$ for the expectation operator.

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 three anonymous referees for their helpful comments.

II. Problem Formulation

We consider the memoryless adaptive control problem

$$\begin{cases} \min_{(u_t)} & \sum_{t=1}^{T} \mathbf{E}\|r_t - y_t\|^2 + \lambda \mathbf{E}\|u_t\|^2, \\ \text{s. t.} & y_t = Bu_t + w_t, \end{cases} \tag{1}$$

where $y_t, r_t, w_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$ and $\lambda \geq 0$. $B \in \mathbb{R}^{n\times m}$ is assumed unknown in advance. Moreover, we assume that a unique solution to (1) of the form $u_t = Kr_t$ exists, in which case $K = (B^\top B + \lambda I)^\dagger B^\top \in \mathbb{R}^{m\times n}$. Our goal is to investigate whether, depending on the dimensionality of B, there may be phase transition in the regret – learning-based performance – any algorithm can attain. We are able to prove that there are two regimes for regret (defined below): one in which regret scales like $\log T$ and one in which it scales like $\sqrt{T}$. To this end, our Theorem 5.2 gives regret lower bounds for (1) and a sufficient condition for the $\sqrt{T}$-scaling limit to occur. This is then contrasted with Theorem 5.4 which gives a logarithmic lower bound valid in both regimes.

To qualify this, some further assumptions are necessary. We suppose B has rank at least 1. We write $K = K(B, \lambda) \in \mathbb{R}^{m\times n}$ for the optimal linear law and its Jacobian is $G = \nabla_B \operatorname{vec} K(B, \lambda)$. The reference signal, r_t, is assumed to be known in advance and we will make a standard persistence of excitation assumption, namely that $\sum_{k=1}^{t} r_k r_k^\top \succ tcI + o(t)$ and that $\|r_t\| > c'$ for some $c, c' > 0$ and sufficiently large t. The noise $w_t \in \mathbb{R}^n$ is assumed to be mean zero, indepedent and identically distributed, and have density $q(w)$, which admits Fisher information[1]. The control $u_t \in \mathbb{R}^m$ is constrained to depend on only past inputs and outputs and is in particular oblivious of the parameter B – it is adaptive. To compare adaptive laws to the optimal law, one introduces the regret.

Definition 2.1: The regret of u_t, $R_T = R_T(\{u_t\}; B)$ is

$$R_T = \sum_{t=1}^{T} \mathbf{E}\|r_t - Bu_t\|^2 - \sum_{t=1}^{T} \mathbf{E}\|r_t - BKr_t\|^2 + \lambda \mathbf{E}\sum_{t=1}^{T} \|u_t\|^2 - \lambda \mathbf{E}\sum_{t=1}^{T} \|Kr_t\|^2. \tag{2}$$

This measures the cumulative difference between the cost incurred by the adaptive law (u_t) and the optimal law Kr_t which uses knowledge of B.

Example 2.2: Consider a scalar system $y_t = bu_t + w_t$ with variance 1 of w_t. Suppose that r_t is sufficiently excited, say $r_t = 1$ for all time and that $\lambda = 0$. This case is then covered by [7] and [8] (by setting all lag parameters of y to zero), where it is shown that

$$R_T = O(\log T)$$

for a policy based on least squares and certainty equivalence.

[1]The density of w_t needs to satisfy certain absolute continuity and mean square differentiability conditions. We prefer not to go into these details and simply assume existence, see Definition 4.1 and consult [16] for details.

III. Regret Decomposition

The following result is key, as it directly relates regret to an estimation error.

Lemma 3.1: For any linear policy $u_t = \hat{K}_t r_t$, we have

$$R_T = \sum_{t=1}^{T} \mathbf{E} \operatorname{tr}\left[\left((I\otimes r_t r_t^\top) + \lambda (I\otimes B)^\top (I\otimes r_t r_t^\top)(I\otimes B)\right) \times \operatorname{vec}(\hat{K}_t - K)(\operatorname{vec}(\hat{K}_t - K))^\top\right]. \tag{3}$$

Proof: Setting $u_t = \hat{K}_t r_t$, we see that K also is for each t the minimizer of

$$\|r_t - B\hat{K}_t r_t\|^2 - \|r_t - BKr_t\|^2 + \lambda\left(\|\hat{K}_t r_t\|^2 - \|Kr_t\|^2\right).$$

Vectorizing, we observe that this is a quadratic expression in $\operatorname{vec}\hat{K}_t$, minimized at $\operatorname{vec} K$, where its value is zero. A straightforward computation shows that the Hessian in vectorized variables is

$$2(I\otimes r_t r_t^\top) + 2\lambda (I\otimes B)^\top (I\otimes r_t r_t^\top)(I\otimes B)$$

Since there are no higher order terms, Taylor expansion around the minimum K gives

$$\begin{aligned} &\|r_t - B\hat{K}_t r_t\|^2 - \|r_t - BKr_t\|^2 + \lambda\left(\|\hat{K}_t r_t\|^2 - \|Kr_t\|^2\right) \\ &= \operatorname{tr}\left[\left((I\otimes r_t r_t^\top) + \lambda (I\otimes B)^\top (I\otimes r_t r_t^\top)(I\otimes B)\right) \times \operatorname{vec}(\hat{K}_t - K)(\operatorname{vec}(\hat{K}_t - K))^\top\right]. \end{aligned}$$

The result follows by summation and expectation. ∎

This shows that regret is linear in the estimation error, $\operatorname{vec}(\hat{K}_t - K)(\operatorname{vec}(\hat{K}_t - K))^\top$. To relate this to any particular policy, we make the following definitions.

Definition 3.2: The control sequence (u_t) is α-fast convergent if for all B, $u_t = Kr_t + v_t$ with $\sqrt{\mathbf{E}\operatorname{tr} v_t v_t^\top} = o(t^{-\alpha})$.

Remark 3.3: Since $\sqrt{\mathbf{E}\operatorname{tr} v_t v_t^\top} = o(t^{-\alpha})$, Chebyshev's inequality implies that $v_t = o(t^{-\alpha})$ in probability.

Definition 3.4: An α-fast convergent policy $u_t = Kr_t + v_t$ is called β-unbiased if $\mathbf{E}v_t = o(t^{-\beta})$.

In particular, α-fast convergence prohibits constant strategies such as selecting K which is optimal for one parametrization but sub-optimal for others. The reason for introducing β-unbiasedness is similar. Observe also that trivially an α-fast convergent policy is α-unbiased.

Lemma 3.5: Any α-fast convergent policy can be written as $u_t = \hat{K}_t r_t$, for some sequence of matrices $\hat{K}_t = K + o(t^{-\alpha})$.

Proof: Since by assumption $\|r_t\| > c'$ uniformly in time for large t and some constant c', there exists a linear transformation V_t such that $v_t = V_t r_t$ and $\mathbf{E}\|V_t\| = o(t^{-\alpha})/\|r_t\| = o(t^{-\alpha})$. Take $\hat{K}_t = K + V_t$. ∎

Lemma 3.5 shows that for α-fast convergent policies, it suffices to consider linear representations. In the subsequent analysis, we will also need some asymptotic control of the gradient of these linear representations, with respect to the parameter, B.

Definition 3.6: An α-fast convergent policy, $u_t = \hat{K}_t r_t$ is regular if $\nabla_B \mathbf{E}\operatorname{vec}\hat{K}_t = \nabla_B \operatorname{vec} K + o(1)$.

IV. Information

As indicated by the regret decomposition, Lemma 3.1, our regret analysis will essentially be estimation-theoretic. The following notion is key in the Cramér-Rao bound we will use.

Definition 4.1: For a parametrized family of probability densities $\{p_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^d$, Fisher information $I_\theta \in \mathbb{R}^{d\times d}$ is

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

whenever the integral exists.

Since the density, $q(w)$, of the noise in (1) was assumed sufficiently regular, for the parameter $\theta = B$, the parametrized family of densities induced by $y_k = Bu_k + w_k$, $k = 1, \dots, t$ admits Fisher information, $I_{t,B} \in \mathbb{R}^{mn\times mn}$. Let $J = \int q(w)[\nabla_w q(w)][\nabla_w q(w)]^\top dw$, then one has that

$$\begin{aligned} I_{t,B} &= \sum_{k=1}^{t} \int q(y - Bu_t) \\ &\quad \times \nabla_B \log q(y - Bu_k) \left[\nabla_B \log q(y - Bu_k)\right]^\top dy \\ &= \sum_{k=1}^{t} u_k \otimes J \otimes u_k^\top, \quad (4) \end{aligned}$$

by the chain rule for Fisher information and change of variables. We note in passing that J is often called Fisher information about the location parameter of q. If in addition $u_t = Kr_t + v_t$ is α-fast convergent (4) becomes

$$\begin{aligned} I_{t,B} &= \underbrace{\sum_{k=1}^{t} \mathbf{E}\left[Kr_k \otimes J \otimes r_k^\top K^\top\right]}_{I^*_{t,B}} + \sum_{k=1}^{t} \mathbf{E} v_k \otimes J \otimes v_k^\top \\ &= \sum_{k=1}^{t} \mathbf{E}\left[v_k \otimes J \otimes r_k^\top K^\top + Kr_k \otimes J \otimes v_k^\top\right]. \quad (5) \end{aligned}$$

where $v_k = o(k^{-\alpha})$ due to α-fast convergence. Above, one recognizes $I^*_{t,B}$ as the Fisher information generated by the optimal trajectory, where $v_k \equiv 0$ and the optimal law is always applied. Observe that unless K has rank n, $I^*_{t,B}$, is degenerate for all t. A degenerate Fisher information has bleak implications for model identifiability. For instance, if $I_{t,K}$ is degenerate, the analysis of [17] shows that no finite variance estimator with a given bias exists except for under very special circumstances (see also Theorem 4.2). Fortunately, the remainder term in (5) may be chosen to complete rank-deficiency. However, all is not won, since the requirement that u_t is α-fast convergence entails that this term is small.

a) A Multi-Scale Cramér-Rao Bound: We now state the general Cramér-Rao analysis of [17].

Theorem 4.2 (cf. [17]): Let $\{p_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, be a family of densities with Fisher information I_θ. Suppose that $\alpha = \alpha(\theta) \in \mathbb{R}^{d'}$ is a vector-valued function of θ and let $\hat{\alpha}$ be any estimate of α. Let I_θ have singular value decomposition

$$I_\theta = ZSZ^\top = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} Z_1^\top \\ Z_2^\top \end{bmatrix}.$$

Suppose $A = \nabla_\theta \mathbf{E}\hat{\alpha}$ exists and set $\begin{bmatrix} A_1 & A_2 \end{bmatrix} = A \begin{bmatrix} Z_1 & Z_2 \end{bmatrix}$. Then one has that

$$\mathbf{E}[(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^\top] \succeq A_1 S_1^\dagger A_1^\top + A_2 S_2^\dagger A_2^\top. \quad (6)$$

Moreover, if $A \neq AI_\theta I_\theta^\dagger$, the covariance is infinite in the sense that $\operatorname{tr} \mathbf{E}[(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^\top] \geq \sigma, \forall \sigma \in \mathbb{R}$.

Proof: The proof of this result is essentially the same as in [17], and is thus omitted for brevity. ∎

This is the Cramér-Rao inequality split into two blocks. The idea in the sequel will be to use that the second block of an α-fast convergent policy is very near rank deficient whenever $I^*_{t,B}$ in (5) loses rank. This also relates to the reason we prefer to state the bound in terms of pseudo-inverses; we cannot a priori guarantee that the second block is non-zero.

For our model (1), the following consequence is immediate and is just the above Theorem restated for our problem.

Corollary 4.3: Let $\hat{K}_t$ be any estimator of the optimal policy K of (1) such that $H = \nabla_B \mathbf{E} \operatorname{vec} \hat{K}_t(B, \lambda) \in \mathbb{R}^{nm\times nm}$ exists. Decompose $I_{t,B}$ (5) spectrally and define H_1, H_2 via

$$I_{t,B} = U\Lambda U^\top = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} U_1^\top \\ U_2^\top \end{bmatrix} \quad (7)$$

$$\begin{bmatrix} H_1 & H_2 \end{bmatrix} = H \begin{bmatrix} U_1 & U_2 \end{bmatrix}. \quad (8)$$

Then

$$\begin{aligned} &\mathbf{E} \operatorname{vec}(\hat{K}_t - K)(\operatorname{vec}(\hat{K}_t - K))^\top \\ &\qquad \succeq H_1 \Lambda_1^\dagger H_1^\top + H_2 \Lambda_2^\dagger H_2^\top. \quad (9) \end{aligned}$$

We let the block sizes in (7) correspond to those in (11) and Λ_2 contains the smallest singular values. The notation (7)-(8) is fixed throughout the rest of the paper.

b) Uninformative Optimal Policies: The regret decomposition (3) shows that regret depends fundamentally on

$$\operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) \mathbf{E} \operatorname{vec}(\hat{K}_t - K)(\operatorname{vec}(\hat{K}_t - K))^\top\right).$$

Since the estimation error is lower-bounded by (9), degeneracy of Fisher information should be a factor in regret if for the small block, say Λ_2,

$$\operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) H_2 \Lambda_2^\dagger H_2^\top\right) \neq 0. \quad (10)$$

Now, the Fisher-information, $I_{t,B}$, of an α-fast convergent policy one imagines is close to that of the optimal policy, $I^*_{t,B}$. To that end, apply the spectral decomposition

$$I^*_{t,B} = O\Lambda' O^\top = \begin{bmatrix} O_1 & O_2 \end{bmatrix} \begin{bmatrix} \Lambda_1' & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} O_1^\top \\ O_2^\top \end{bmatrix} \quad (11)$$

where the columns of O_2 span $\ker I^*_{t,B}$. This relates to (7) where U_2 spans the smallest singular space of $\dim \ker I^*_{t,B}$. We expect regret to scale differently in the following regime.

Definition 4.4: The pair (r_t, B) is said to be γ-uninformative if $G \neq GI^*_{t,B}(I^*_{t,B})^\dagger$ and

$$\operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) G_2 G_2^\top\right) \geq \gamma \quad (12)$$

for $\gamma > 0$, $G_2 = GO_2$ where $G = \nabla_B \operatorname{vec} K(B, \lambda)$, $I^*_{t,B}$ is given in (5), and O_2 is defined through (11).

This definition is meaningful when $I^*_{t,B}$ of (5) does not have full rank (for otherwise $G = GI^*_{t,B}(I^*_{t,B})^\dagger$) and is

best understood as a quantitative observability-type condition; the kernel of the optimal policy's information matrix should be visible through the cost (regret). Note also the implicit dependence on B, via G_2 which depends both on the subspace $\ker I^*_{t,B}$ and on the optimal policy K. Geometrically, γ-uninformativeness means that $\ker I^*_{t,B}$ (5), which is spanned by O_2 and $G_2 = GO_2$, is γ-separated from being perpendicular to the reference signal's power, $r_t r_t^\top$, in the geometry of symmetric matrices. A rank drop in $I^*_{t,B}$ of (5) can occur when $\operatorname{rank} K = \operatorname{rank} B < m$. Note here also that [6] establishes logarithmic rates precisely when B is invertible, i.e. of full rank. We now give an example illustrating γ-uninformativeness.

Example 4.5: Suppose $y_t, r_t, w_t \in \mathbb{R}$ and $u \in \mathbb{R}^m$ so that y_t is given by $y_t = b^\top u_t + w_t$ with $b \in \mathbb{R}^m$ and b given by the first standard Euclidean basis vector

$$b^\top = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}.$$

If furthermore $\lambda = 0$, the optimal policy can be expressed as $u_t = br_t$. Thus in the notation of (11)-(12), $G = \nabla_b b = I$.

Let us also assume that w_t is Gaussian with variance 1 and suppose $r_t = 1$ for all t. Substitution into (5) shows that Fisher information of the optimal policy is

$$I^*_{t,b^\top} = \sum_{k=1}^{t} bb^\top$$

which has nullspace $\ker b^\top$ and clearly $G \neq GI^*_{t,B}(I^*_{t,B})^\dagger$. This also means that $O_2^\top$ of (11)-(12) is given by

$$O_2 = \begin{bmatrix} 0_{1\times m-1} \\ I_{m-1\times m-1} \end{bmatrix}.$$

Thus, in this case, since $G_2 = \nabla_b b O_2 = IO_2 = O_2$, wherefore we have that

$$\operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) G_2 G_2^\top\right) = \operatorname{tr} O_2 O_2^\top = m - 1.$$

That is, the pair $(1, b^\top)$ is $(m-1)$-uninformative.

c) Spectral Information Comparison: The main reason for using G_2 (11)-(12) instead of H_2 (7)-(8), in Definition 4.4 is that we wish to study how the parametrization of (1) impacts regret. If we had defined uninformativeness in terms of the H_2, which depends on the choice of algorithm through $H = \nabla_B \operatorname{vec} \hat{K}_t(B, \lambda)$, it would be hard to claim that a phase transition occurs through the parametrization and our lower bound would not be algorithm indepedent. Of course, this entails that we need to control the gap between the system quantity (12) and the algorithm-dependent quantity (10). To this end, we perform a perturbation analysis of the subspace spanned by U_2 (spans $\ker I^*_{t,B}$), to relate it to O_2 (spans the subspace of the $\dim \ker I^*_{t,B}$ smallest singular values of $I_{t,B}$).

Lemma 4.6: Let H_2 be as in (8) and G_2 be as in (11)-(12). Then for any regular and α-fast convergent policy $\hat{K}_t$, $\alpha > 0$, one has that

$$H_2 P_t = G_2 + o(1)$$

for some sequence of orthonormal matrices (P_t).

Proof: The proof relies on Wedin's $\sin\Theta$ Theorem [18] (quoted in the appendix) which describes the perturbation theory of range and nullspace in the singular value decomposition. By (5) it follows that

$$\frac{1}{t} I_{t,B} = \frac{1}{t} I^*_{t,B} + o(t^{-2\alpha} \log t)$$

using α-fast convergence[2]. Moreover, it is clear that rescaling Fisher informations does not change the singular value decomposition except for rescaled singular values. This gives control of the residuals (1.8) in [18] and for sufficiently large t the separation conditions there are satisfied since $(1/t)\Lambda_2 = o(1)$. We recall also from Lemma 1 of [19] that the $\sin\Theta$ Theorem provides an upper bound on the distance $d(U_2, O_2) = \min_P \|U_2 P - O_2\|_\infty$, where P is optimized over the orthogonal group. Apply now Wedin's Theorem in combination with Lemma 1 of [19] to conclude that

$$U_2 P_t = O_2 + o(t^{-2\alpha})$$

where P_t optimizes $d(U_2, O_2)$. Finally, by regularity $H = G + o(1)$, and therefore $H_2 P_t \to G_2$. ∎

Remark 4.7: The matrix P_t is necessary to account for the possibly arbitrary ordering of the singular vectors corresponding to $\ker I^*_{t,B}$. It exists by continuity and compactness of the orthogonal group in the standard matrix topology.

V. Fundamental Limitations

We need one more lemma which relates regret to the spectral properties of $I_{t,B}$ and $I^*_{t,B}$.

Lemma 5.1: For any γ-uninformative pair (r_t, B), and regular and α-fast convergent policy $\hat{K}_t$, $\alpha > 0$, with Λ_2 as in (7) and H_2 as in (8), one has that

$$\operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) H_2 \Lambda_2^\dagger H_2^\top\right) \geq \gamma \sigma_{\min}(\Lambda_2^\dagger) \times (1 + o(1)).$$

Proof: The trace cyclic property and Lemma 4.6 yields

$$\begin{aligned} &\operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) H_2 H_2^\top\right) \\ &\quad = \operatorname{tr}\left(\left(I \otimes r_t r_t^\top\right) H_2 P_t P_t^\top H_2^\top\right) \\ &\quad = \operatorname{tr}\left(P_t^\top H_2^\top \left(I \otimes r_t r_t^\top\right) H_2 P_t\right) \\ &\qquad = \operatorname{tr}\left(G_2^\top \left(I \otimes r_t r_t^\top\right) G_2\right) + o(1) \end{aligned}$$

using that P_t from Lemma 4.6 is orthonormal and where G_2 is as in (11)-(12). Multiplication by $\Lambda_2^\dagger \succeq 0$ rescales the bound by $\sigma_{\min}(\Lambda_2^\dagger)$. ∎

Lemma 5.1 is used together with the second part of the Cramér-Rao bound, Theorem 4.2. Either $\sigma_{\min}(\Lambda_2^\dagger)$ is nonzero, or the policy has infinite variance. In either case, we will be able to establish our regret bound. We are now in position to state our main result.

Theorem 5.2: Consider the model (1) and suppose that (r_t, B) is γ-uninformative for each t in a sequence of subsets $\tau_T \subset \mathbb{N}, \tau_T \subset \{1, \dots, T\}$ with $|\tau_T| > cT$, for some $\gamma > 0$ and $c \in (0, 1)$. Then any regular α-fast convergent policy, with $\alpha > 0$, which is $(1/2)$-unbiased satisfies the following regret lower bound

$$R_T = \Omega(\sqrt{T}). \tag{13}$$

[2]The appearance of the logarithmic factor is due to the case $\alpha = 1/2$, since $\int t^{-1} \sim \log t \neq O(1)$.

More generally, any α-fast convergent policy (not necessarily $(1/2)$-unbiased) satisfies

$$R_T = \Omega(\max(T^{\alpha}, T^{1-2\alpha}) = \Omega(T^{1/3})).$$

Remark 5.3: In our proof, the idea behind $(1/2)$-unbiasedness is to prevent super-efficiency of the policy; following our analysis α-fast convergence is not sufficient to guarantee that the policy $\hat{K}_t$ does not perform better at certain points B in parameter space. The second part of the Theorem shows that this can be relaxed somewhat (but at some cost).

Proof: According to Lemma 3.5 we may restrict our attention to linear policies. Let Λ_2 be as in (7) and to emphasize its time-dependence, we now write $\Lambda_2 = \Lambda_2(t)$. Let also H_2 be as in (8). Combining the results that we have established this far offers

$$\begin{aligned} R_T &\geq \sum_{t=1}^{T} \mathbf{E}\,\mathrm{tr}\left[\left(I \otimes r_t r_t^{\top}\right)\mathrm{vec}(\hat{K}_t - K)(\mathrm{vec}(\hat{K}_t - K))^{\top}\right] \\ &\geq \sum_{t=1}^{T}\left[\mathrm{tr}\left(\left(I \otimes r_t r_t^{\top}\right) H_2 \Lambda_2^{\dagger}(t) H_2^{\top}\right)\right] \\ &\geq \sum_{j=1}^{\lfloor cT \rfloor}\left[\gamma \sigma_{\min}(\Lambda_2^{\dagger}(t_j))(1+o(1))\right]. \quad (14) \end{aligned}$$

This merits some explanation: The first inequality in (14) above is obtained by dropping the part pertaining to B, in (3). The second inequality is Corollary 4.3 to the Cramér-Rao bound, Theorem 4.2. In the third inequality, we use the assumption that the model is uninformative for a sequence of sets τ_T and thus also for a subsequence $\{t_j\}$ of time $\{t\}$. We then apply Lemma 5.1 throwing away all terms which are not in some set τ_T. Fourth, we observe that the length of this subsequence by assumption is proportional to T, i.e. cardinality greater than $\lfloor cT \rfloor$ for some $c \in (0,1)$.

Now, we may assume that $\sigma_{\min}(\Lambda_2(t_j)) > 0$ for all t_j of this subsequence, for otherwise we may use the condition $G \neq G I_{t,B}^*(I_{t,B}^*)^{\dagger}$ of Definition 4.4. Since $H = G + o(1)$ and $I_{t,B}/t = I_{t,B}^*/t + o(1)$ the second part of Theorem 4.2 would in this case imply infinite variance, which contradicts α-fast convergence.

The idea of the rest of the proof is to balance the convergence rate of the policy with the necessary exploration for large Fisher information. Fix $\alpha, \alpha' > 0$. If the policy is not α-fast convergent (so that it only is α'-fast convergent, for $\alpha' < \alpha$), one has that

$$\begin{aligned} \sum_{t=1}^{T}\left[\mathrm{tr}\left((I \otimes r_t r_t^{\top})\mathbf{E}\left[\underbrace{\mathrm{vec}(\hat{K}_t - K)(\mathrm{vec}(\hat{K}_t - K))^{\top}}_{\Omega(t^{-2\alpha})}\right]\right)\right] \\ = \Omega(T^{1-2\alpha}). \quad (15) \end{aligned}$$

Hence, unless the policy is $(1/4)$-fast convergent (i.e. $\alpha \geq 1/4$), we have $R(T) = \Omega(\sqrt{T})$.

We will now balance this term with smallest singular value of Fisher information. Using the fact that the policy is assumed $(1/2)$-unbiased, it clear from (5) that $\sigma_{\max}(\Lambda_2(t)) = o(t^{1-2\alpha}) + o(t^{-1/2})$. To see this, note that the mixed term (mixed in v_t, Kr_t) of (5) has contribution at most $o(t^{-1/2})$ by the unbiasedness assumption and that the term quadratic in v_t has contribution $o(t^{-2\alpha})$. Hence $\sigma_{\min}(\Lambda_2^{\dagger}(t)) = \Omega(\min(t^{2\alpha-1}, t^{-1/2}))$. From this, we gather that

$$\sum_{j=1}^{\lceil cT \rceil}\left[\gamma \sigma_{\min}(\Lambda_2^{\dagger}(t_j))(1+o(1))\right] = \Omega(\min(\sqrt{T}, T^{2\alpha})).$$

Balancing this with $\Omega(T^{1-2\alpha})$ in (15) by setting $\alpha = 1/4$ finishes the proof of the first part of the theorem.

If we drop the unbiasedness assumption, we may make the same analysis as above, but the mixed term in (5) may dominate and instead be on the order of magnitude $o(t^{-\alpha})$ (instead of $o(t^{-1/2})$). Mutatis mutandis, one arrives at a lower bound $\Omega(T^{\alpha})$, which can be balanced with $\Omega(T^{1-2\alpha})$ at $\alpha = 1/3$, yielding the result. ∎

Let us now put Theorem 5.2 into perspective by comparing with what kind of lower bound we can prove if Fisher information has full rank.

Theorem 5.4: Suppose that $\sigma_{\min}(I_{t,B}^*) \geq \delta t + o(t), \delta > 0$ and assume the additional regularity condition $\sigma_{\min} H = \sigma_{\min} \nabla_B \mathbf{E}\,\mathrm{vec}\,\hat{K}_t(B) = \Omega(1)$. Then for any $\alpha > 0$ and any regular α-fast convergent policy, (1) satisfies the regret lower bound

$$R_T = \Omega(\log T).$$

In this regime, there is only very little trade-off between exploration and exploitation. Essentially, an optimal policy which has full rank, and thus full rank Fisher information, will already excite all directions of B and so there is little to be gained by adding extra excitation.

Proof: Write, using (3), and the Cramér-Rao bound, Theorem 4.2, with the second block of dimension zero,

$$\begin{aligned} R_T &\geq \sum_{t=1}^{T} \mathbf{E}\,\mathrm{tr}\left[\left((I \otimes r_t r_t^{\top})\right)\mathrm{vec}(\hat{K}_t - K)(\mathrm{vec}(\hat{K}_t - K))^{\top}\right] \\ &\geq \sum_{t=1}^{T} \mathbf{E}\,\mathrm{tr}\left[\left(I \otimes r_t r_t^{\top}\right)\underbrace{H I_{t,B}^{-1} H^{\top}}_{=\Omega(\delta/(t+o(t)))}\right]. \end{aligned}$$

The result is now immediate since $\sum_{t=1}^{T} \delta/(t+o(t))$ scales like $\log T$. ∎

Theorems 5.2 and 5.4 together provide strong evidence that a phase transition occurs. We now return to Example 2.2 to understand why logarithmic rates are feasible in [8].

Example 5.5: Let us revisit the scalar system $y_t = bu_t + w_t$. Here Fisher information is given by $I_{t,b}^* = \sum_{k=1}^{t} r_k^2 b^2 = tb^2$, for $r_t = 1$, which, in particular, has no nullspace. Hence the optimal policy is not uninformative for any constant and the singular value condition of Theorem 5.4 is satisfied. Indeed, the lower bound $\Omega(\log T)$ presented above matches (in order) the upper bound due to [8] discussed in Example 2.2.

VI. Discussion

Our analysis here is similar spirit to [13]. The largest difference is that we take a closer look at the Cramér-Rao bound when the information matrix is nearly degenerate. In

this regime learning becomes difficult. Similarly, the proof strategy is here also based on comparing the information "collected" by any algorithm and that of the optimal algorithm. However, degeneracy makes this comparison more difficult and we need to resort to singular space perturbation theory, [18]. It is this difference that allows us to demonstrate the $\sqrt{T}$-rate.

Using this, we show that if the optimal policy to (1) gives degenerate information in a certain sense, then regret must be super-logarithmic. In this regime, one is forced to introduce supplementary excitation beyond the randomness already present in the algorithm. It is also interesting to note that the lower bound strongly suggests that a certainty equivalent controller perturbed by noise with full rank covariance of order $1/\sqrt{t}$ is a good idea since this corresponds closely to the case for which the lower bound is optimized to $\Omega(\sqrt{T})$. Indeed, this was the strategy pursued by [11] attaining regret of order $\sqrt{T}$ for the full LQR.

We also wish to mention that the concept of α-fast convergence is much inspired by the notion of uniformly fast convergence for the related bandit problems, see [20] and [21]. Indeed, what we have considered here can also be seen as a stochastic contextual bandit with side information y_t and context decided by r_t, see [22] for an overview of bandits.

Directions for Future Work: It would be very desirable to extend the present analysis to dynamic models such as LQR. On the one hand, such extension could provide a Fisher information point of view on the LQR lower bound of [14]. On the other hand, this could help describe necessary and sufficient side information for logarithmic regret, [11]. Such situations can be covered by our method by adapting the parametrization of the problem which would result in a different (constrained) Fisher information. To this end, [12] shows that under an identifiability condition, logarithmic regret is attainable for LQR and it would be very interesting to compare this notion to uninformativeness. It would also be interesting to relax or replace the assumption of $(1/2)$-unbiasedness in Theorem 5.2 for the full $\Omega(\sqrt{T})$-lower bound. Another potential direction is to take a more directly information-theoretic route toward lower bounds as in [23] and as is traditional for bandits [20]. This was recently done for system identification in [24].

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Optimism-based adaptive regulation of linear-quadratic systems," *arXiv preprint arXiv:1711.07230*, 2017.

[5] H. Mania, S. Tu, and B. Recht, "Certainty equivalent control of LQR is efficient," *arXiv preprint arXiv:1902.07826*, 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] T. Lai and C.-Z. Wei, "Extended least squares and their applications to adaptive control and prediction in linear systems," *IEEE Transactions on Automatic Control*, vol. 31, no. 10, pp. 898–906, 1986.

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

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

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

[11] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive linear–quadratic regulators," *Automatica*, vol. 117, 2020.

[12] ——, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, 2020.

[13] I. Ziemann and H. Sandberg, "Regret lower bounds for unbiased adaptive control of linear quadratic regulators," *IEEE Control Systems Letters*, vol. 4, no. 3, pp. 785–790, 2020.

[14] M. Simchowitz and D. J. Foster, "Naive exploration is optimal for online lqr," *arXiv preprint arXiv:2001.09576*, 2020.

[15] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," *arXiv preprint arXiv:2002.08095*, 2020.

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

[17] P. Stoica and T. L. Marzetta, "Parameter estimation problems with singular information matrices," *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 87–90, 2001.

[18] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[19] T. T. Cai, A. Zhang, *et al.*, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *The Annals of Statistics*, vol. 46, no. 1, pp. 60–89, 2018.

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

[21] 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.

[22] T. Lattimore and C. Szepesvári, "Bandit algorithms," *preprint*, 2018.

[23] M. Raginsky, "Divergence-based characterization of fundamental limitations of adaptive dynamical systems," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2010, pp. 107–114.

[24] Y. Jedra and A. Proutiere, "Sample complexity lower bounds for linear system identification," *2019 IEEE Conference on Decision and Control (CDC)*, 2019.

APPENDIX

We require the following version of Wedin's $\sin\Theta$ Theorem:

Consider matrices M and $\tilde{M} = M + T$ for some perturbation T, with singular value decompositions $M = V_1\Gamma_1 W_1^\top + V_2\Gamma_2 W_2^\top$, $\tilde{M} = \tilde{V}_1\tilde{\Gamma}_1\tilde{W}_1^\top + \tilde{V}_2\tilde{\Gamma}_2\tilde{W}_2^\top$. If for $\delta > 0, \eta \geq 0$, $\sigma_{\min}(\tilde{\Gamma}_1) \geq \eta + \delta$ and $\sigma_{\max}(\Gamma_2) \leq \eta$ then

$$\max_P \|\tilde{V}_2 P - V_2\|_\infty = O(\|T\|_\infty)$$

where the maximization is over the orthogonal group.

The result is due to Wedin [18]. The formulation of the $\sin\Theta$ distance as an optimization over the norm $\|\cdot\|_\infty$ appears for instance in [19].