# 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 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. This to some extent resolves an open issue concerning the attainability of logarithmic regret rates in linear quadratic adaptive control. 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. Much progress has been made toward analyzing the statistical convergence rate, the *regret* incurred, of adaptive algorithms. Several works over the past decade, [1], [2] and [3], have been able to prove upper bounds on the regret at a rate of approximately $\sqrt{T}$ in the time horizon. There is a strong feeling that this should be order optimal even though no matching lower bound exist. However, in some special cases, [4], [5] and [6], the authors have actually been able to prove regret to scale at a rate of $\log T$, ensuring considerably faster convergence. In particular, [5] showed that the Åström-Wittenmark self-tuning regulator [7] for SISO tracking problems converges at the rate $\log T$.

Under an unbiasedness assumption, we gave a lower bound in [8] applicable to the general LQR and matching the special cases of [4] and [5] up to constants in the first order. However, as mentioned, no lower bound in the $\sqrt{T}$ regime is known. See also [9] for an interesting discussion of this apparent dichotomy. 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.

The contribution of this note is thus to provide lower bounds on regret which capture this phase transition. In particular, we shall prove that when a certain informativeness condition is not met by the optimal policy, logarithmic rates of regret are impossible for a certain class of policies we call $\alpha$-fast convergent. We will see that this phase transition depends both on the rank of the optimal linear feedback matrix and on the excitation of the reference signal. In this regime, there is an asymptotically non-neglible trade-off between exploration and exploitation. Our results partially answer an unresolved question in the literature [9], as to whether logarithmic rates are attainable in general or not for linear quadratic problems. As such, this work constitutes the first super-logarithmic lower bound available in the literature for linear quadratic stochastic adaptive control problems.

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).

### A. 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 Kroenecker product, vectorization, and Moore-penrose pseudoinverse, respectively. 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 $f = O(g)$. For asymptotic lower bounds we write $f = \Omega(g)$ 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 $T$. We also write $\mathbf{E}$ for the expectation operator and $\mathbf{V}$ for the covariance operator. We use $\nabla$ for gradient or Jacobian.

## 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. Our goal is to investigate whether depending on the parametrization, 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 and that that a unique optimal control policy to (1) exists (either $B$ is invertible or $\lambda > 0$) and is linear in $r_t$. We write $K = K(B, \lambda) \in \mathbb{R}^{m\times n}$ for

this optimal linear feedback 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 admit Fisher information[1] (Definition 2.6). 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. We will pose some restrictions on the possible rates of convergence an adaptive controller can have.

*Definition 2.1:* The control sequence $(u_t)$ is $\alpha$-fast convergent if for all $B$, $u_t = Kr_t + v_t$ with $v_t = o(t^{-\alpha})$ in $L^2$.
In particular, this prohibits constant strategies such as selecting $K$ which is optimal for one parametrization but sub-optimal for others. Note also that this definition relies implicitly on the uniqueness of the solution to (1). A weaker condition than $\alpha$-fast convergence, key to us, is introduced below.

*Definition 2.2:* The control sequence $(u_t)$ is $\beta$-unbiased if for all $B$ $u_t = F_t r_t + v_t$ with $v_t = o(t^{-\beta})$ and $\mathbf{E}F_t = K$ where $F_t$ depends only on past observations and inputs.
This assumption is weaker than $u_t$ being $\alpha$-fast convergent since then $u_t = Kr_t + o(t^{-\alpha})$ and the latter condition is verified with $F_t = K$ and $\beta = \alpha$. We also remark that for $\beta$-unbiased $u_t$, it is without loss of generality to write $u_t = (F_t + L_t)r_t$ since $\|r_t\| > c' > 0$ implies that there exists a transformation $L_t \in \mathbb{R}^{m\times n}$ such that $v_t = L_t r_t$ with $L_t = o(t^{-\beta})$. The quantity $L_t$ is referred to as the bias of the policy. We will thus, without loss of generality, write $u_t = \hat{K}_t r_t$ with $\hat{K}_t = F_t + L_t$ for $\beta$-unbiased policies. In the subsequent analysis, we will also need some (relatively little) control of the gradient of $L_t$.

*Definition 2.3:* A $\beta$-unbiased policy with bias $L_t$ is called regular if $\nabla_B L_t = o(1)$.

To compare adaptive laws to the optimal law, one introduces the regret.

*Definition 2.4:* The regret $R_T$ of a control law $u_t$ is

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

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 2.5:* Consider a scalar system $y_t = bu_t + w_t$ of with variance 1 of $w_t$. Suppose that $r_t$ is sufficiently rich, say $r_t = 1$ for all time and that $\lambda = 0$. This case is then covered by [5] (by setting all lag parameters of $y$ to zero), where it is shown that

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

[1]The density of $w_t$ needs to satisfy certain absolute continuity and mean square differentiability conditions. However, we prefer not to go into these details and simply assume existence.

for a policy based on least squares and certainty equivalence.

We will be more concerned with finding lower bounds. Our regret analysis will essentially be estimation-theoretic, and a key quantity that features there is the Fisher information.

*Definition 2.6:* For a parametrized family probability densities $\{p_\theta, \theta \in \Theta\}$ the Fisher information $I_\theta$ 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.
The quantity $\Delta = \nabla_\theta \log p_\theta(X), X \sim p_\theta$ above is referred to as the score vector. For a family of densities of the form $p_\theta(w) = p(w+\theta)$, we write $J = J_\theta(p)$ for Fisher information. This is called Fisher information about location parameter. This has special significance to us, since for one observation $y_t$ under the model (1) with known input $u_t$, one has that

$$I_B(y_t|u_t) = \int [\nabla_B \log p_B(y - Bu_t)]^2 p(y - Bu_t) dy = u_t u_t^\top \otimes J \quad (3)$$

by change of variables and where we write $[\nabla_B \log p_B(y - Bu_t)]^2 = \nabla_B \log p_B(y - Bu_t) \left[\nabla_B \log p_B(y - Bu_t)\right]^\top$. Above $p$ and $J$ denote the density and Fisher information about location parameter for $w_t$ respectively.

*Example 2.7:* If $w_t \sim N(0,\Sigma)$ then for $\Sigma \succ 0$, $J = \Sigma^{-1}$; for Gaussians, information is inversely proportional to noise.

## III. Regret Decomposition

*Lemma 3.1:* For any $\beta$-unbiased policy $u_t = \hat{K}_t r_t$, we have

$$R_T = \sum_{t=1}^{T} \operatorname{tr}\left(B^\top r_t r_t^\top B \mathbf{E}(K - \hat{K}_t)(K - \hat{K}_t)^\top\right) + \lambda \sum_{t=1}^{T} \operatorname{tr}\left(r_t r_t^\top \mathbf{E}(K - \hat{K}_t)(K - \hat{K}_t)^\top\right) + o(T^{1-\beta}) \quad (4)$$

where for $\beta = 1$, $o(T^{1-\beta})$ is replaced with $o(\log T)$.

*Proof:* We analyze the first line of (2) and expand the squares as inner products to find

$$\begin{aligned}\mathbf{E}\|r_t - Bu_t\|^2 &= \mathbf{E}\left[\langle r_t - B\hat{K}_t r_t, r_t - B\hat{K}_t r_t\rangle\right] \\ &= \mathbf{E}\|r_t\|^2 - \mathbf{E}\langle B\hat{K}_t r_t, r_t\rangle \\ &\quad - \mathbf{E}\langle r_t, B\hat{K}_t r_t\rangle + \mathbf{E}\|B\hat{K}_t r_t\|^2 \quad (5)\end{aligned}$$

and

$$\begin{aligned}&\mathbf{E}\|r_t - BKr_t\|^2 = \mathbf{E}\left[\langle r_t - BKr_t, r_t - BKr_t\rangle\right] \\ &= \mathbf{E}\|r_t\|^2 - \mathbf{E}\langle BKr_t, r_t\rangle - \mathbf{E}\langle r_t, BKr_t\rangle + \mathbf{E}\|BKr_t\|^2. \quad (6)\end{aligned}$$

Moreover note that

$$\begin{aligned}&\operatorname{tr}\left(r_t r_t^\top B \mathbf{E}\left[(K - \hat{K}_t)(K - \hat{K}_t)^\top\right] B^\top\right) \\ &= \operatorname{tr}\left(r_t r_t^\top B \mathbf{E}\left[KK^\top + \hat{K}_t\hat{K}_t^\top - K\hat{K}_t^\top - \hat{K}_t K^\top\right] B^\top\right) \\ &= \operatorname{tr}\left(r_t r_t^\top B \mathbf{E}\left[\hat{K}_t\hat{K}_t^\top - KK^\top\right] B^\top\right) + o(t^{-\beta}) \\ &= \mathbf{E}\|B\hat{K}_t r_t\|^2 - \mathbf{E}\|BKr_t\|^2 + o(t^{-\beta}).\end{aligned}$$

Subtracting (6) from (5) thus yields

$$\begin{aligned}&\mathbf{E}\|r_t - Bu_t\|^2 - \mathbf{E}\|r_t - BKr_t\|^2\\ &\quad= \mathbf{E}\|B\hat{K}_t r_t\|^2 - \|BKr_t\|^2 + o(t^{-\beta})\\ &= \operatorname{tr}\left(r_t r_t^\top B\mathbf{E}\left[(K-\hat{K}_t)(K-\hat{K}_t)^\top\right]B^\top\right) + o(t^{-\beta}).\end{aligned}$$

The first term is thus found after summation and application of the trace cyclic property. The analysis of the second term is similar. ∎

Since $\alpha$-fast convergent policies are in particular $\alpha$-unbiased, the critical problem dependent quantity is thus $\mathbf{E}\left[(K-\hat{K}_t)(K-\hat{K}_t)^\top\right]$.

## IV. Information

The significance of the regret analysis above is of course that if $\hat{K}_t$ is an unbiased estimate of $K$, then by Cramér-Rao

$$\mathbf{E}\left[\operatorname{vec}(K-\hat{K}_t)\operatorname{vec}(K-\hat{K}_t)^\top\right] \succeq I_{t,K}^\dagger \tag{7}$$

where $I_{t,K}$ is the information about $K$ after $t$ samples. Since $\hat{K}_t$ is not exactly unbiased, (7) is only asymptotically true, but this will be sufficient to carry out our analysis. For the Gaussian location model induced by $y_k = Bu_k + w_k, k = 1,\dots,t$, Fisher information becomes by (3) and the chain rule

$$\begin{aligned}I_{t,B} &= \sum_{k=1}^t \mathbf{E}u_k u_k^\top \otimes J\\ I_{t,K}^\dagger &:= [\nabla_B \operatorname{vec}(\hat{K}_t)] I_{t,B}^\dagger [\nabla_B \operatorname{vec}(\hat{K}_t)]^\top\end{aligned}$$

If in addition $u_t$ is $\alpha$-fast convergent one has that

$$\begin{aligned}I_{t,B} &= \sum_{k=1}^t \mathbf{E}\left(Kr_k r_k^\top K^\top + v_k v_k^\top\right)\otimes J\\ &= \underbrace{\sum_{k=1}^t \mathbf{E}\left[Kr_k r_k^\top K^\top \otimes J\right]}_{I_{t,B}^*} + \sum_{k=1}^t \mathbf{E}v_k v_k^\top \otimes J. \end{aligned}\tag{8}$$

where $v_k = o(k^{-\alpha})$ due to $\alpha$-fast convergence. Above, one recognizes $I_{t,B}^*$ as the Fisher information generated by the optimal trajectory, where $v_k \equiv 0$.

Observe that unless $K$ has full rank, $I_{t,B}^*$, is degenerate for all $t$. A degenerate Fisher information has bleak implications for model identifiability. Following the analysis of [10], no unbiased estimator exists except for very special circumstances for $K$ if $I_{t,K}$ is degenerate. Fortunately, the term $\sum_{k=1}^t \mathbf{E}v_k v_k^\top \otimes J$ may be chosen to complete rank-deficiency. However, all is not won, since the requirement that $u_t$ is $\alpha$-fast convergence entails that this term is small.

### A. A Multi-Scale Cramér-Rao Bound

We now repeat the analysis of [10] with the modification that we partition the information matrix into a positive definite block, and one positive semi-definite block[2].

We now fix the following notation which holds throughout the rest of the paper. Denote $\vec{K} = \operatorname{vec} K, \vec{K}_t = \operatorname{vec}\hat{K}_t$, $C = \mathbf{V}\vec{K}_t$ and $H = \nabla_B \operatorname{vec}\mathbf{E}\hat{K}_t$. Further, we spectrally decompose Fisher information as

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

and for an arbitrary appropriately dimensioned matrix $W$

$$\begin{bmatrix}W_1\\ W_2\end{bmatrix} = \begin{bmatrix}U_1^\top\\ U_2^\top\end{bmatrix}W, \qquad \begin{bmatrix}H_1 & H_2\end{bmatrix} = H\begin{bmatrix}U_1 & U_2\end{bmatrix}.$$

*Remark 4.1:* Of course, the quantities above all depend on time, $t$. However, for the moment, we supress this dependence to simplify notation.

*Theorem 4.2 (cf. [10]):* With notation as just introduced

$$C \succeq H_1\Lambda_1^{-1}H_1^\top + H_2\Lambda_2^\dagger H_2^\top.$$

*Proof:* Let $W$ be an arbitrary matrix of dimension $mn \times mn$. We have that

$$\begin{aligned}&\mathbf{E}\left[\left(\vec{K}_t - \vec{K}\right) - W^\top\Delta\right)\left(\vec{K}_t - \vec{K}\right) - W^\top\Delta\right)^\top\Bigg]\\ &\qquad= C - HW - W^\top H + W^\top I_{t,B}W \succeq 0\end{aligned}\tag{9}$$

since the score has mean zero. We now partition $I_{t,B}, W$ and $H$ as assumed. Substituting this decomposition into (9) yields

$$\begin{aligned}C &\succeq HW + W^\top H - W^\top I_{t,B}W\\ &= (H_1W_1 + H_2W_2) + (H_1W_1 + H_2W_2)^\top\\ &\quad - (W_1^\top\Lambda_1 W_1) - (W_2^\top\Lambda_2 W_2)\\ &= H_1\Lambda_1^{-1}H_1 - (W_1 - \Lambda_1^{-1}H_1^\top)^\top\Lambda_1(W_1 - \Lambda_1^{-1}H_1^\top)\\ &\quad + H_2\Lambda_2^\dagger H_2 - (W_2 - \Lambda_2^\dagger H_2^\top)^\top\Lambda_2(W_2 - \Lambda_2^\dagger H_2^\top).\end{aligned}\tag{10}$$

Choosing $W_i = \Lambda_i^\dagger H_i^\top, i = 1,2$ gives

$$C \succeq H_1\Lambda_1^{-1}H_1^\top + H_2\Lambda_2^\dagger H_2^\top$$

which was our claim. ∎

This is the Cramér-Rao inequality with an extra term $H_2\Lambda_2^\dagger H_2$ yielding extra variance for ill-conditioned Fisher information.

### B. Uninformative Optimal Policies

Observe that for any $\alpha$-fast convergent policy, (4) shows that regret depends fundamentally on

$$\operatorname{tr}\left(r_t r_t^\top \mathbf{E}\left[(K-\hat{K}_t)(K-\hat{K}_t)^\top\right]\right) = \operatorname{tr}\left(\left(I\otimes r_t r_t^\top\right)C\right)$$

with $C$, the vectorized covariance, as defined above. Since

$$C \succeq H_1\Lambda_1^{-1}H_1^\top + H_2\Lambda_2^\dagger H_2^\top,$$

the degeneracy of Fisher information should be a factor in regret if for the small block $\Lambda_2$

$$\operatorname{tr}\left(\left(I\otimes r_t r_t^\top\right)H_2\Lambda_2^\dagger H_2^\top\right) \neq 0. \tag{11}$$

Now, the Fisher-information, $I_{t,B}$, of an $\alpha$-fast convergent policy one imagines is close to that of the optimal policy, $I_{t,B}^*$. Apply now instead the Spectral theorem to $I_{t,B}^*$ and assume $B$ does not have full column rank. That is, let

$$I_{t,B}^* = O\Lambda' O^\top = \begin{bmatrix}O_1 & O_2\end{bmatrix}\begin{bmatrix}\Lambda_1' & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}O_1^\top\\ O_2^\top\end{bmatrix}.$$

[2] In their article, positive semi-definite corresponds to 0, however their analysis easily extends to our case of interest.

Recall that $G = \nabla_B \operatorname{vec} K(B,\lambda)$, and define $G_2 = GO_2$. We expect regret to scale differently in the following regime.

*Definition 4.3:* The pair $(r_t, B)$ is said to be $\gamma$-uninformative if $B$ does not have full column rank and

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

for $\gamma > 0$.

This 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 $K(B)$. Geometrically, $\gamma$-uninformativeness means that $\ker I^*_{t,B}$ (which is spanned by $O_2$ and recall $G_2 = GO_2$) is $\gamma$-separated from being perpendicular to the reference signal's power, $r_t r_t^\top$, in a weighted (by $G = \nabla K$) geometry of symmetric matrices. We now give an example of when this condition becomes relevant.

*Example 4.4:* Suppose that $y_t$ is a scalar given by $y_t = b^\top u_t + w_t$ with $b, u_t \in \mathbb{R}^m$ and $b$ given by the first standard Euclidean basis vector

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

In this case the optimal policy is given by $br_t$ and so differentiating $\mathbf{E}\hat{K}_t = b + o(1)$ with respect to $b$ yields $H = I + o(1)$.

Let us also assume that the variance of $w_t$ is 1 and choose $r_t$ such that

$$\sum_{k=1}^{t} r_k r_k^\top = tI + o(1),$$

This happens if $r_k$ is a sequence of standard Euclidean basis vector $e_i$ cyclically repeating themselves, so that $r_1 = e_1, \dots r_m = e_m, r_{m+1} = e_1, \dots$ and so on. In this case

$$I^*_{t,B} = t \begin{bmatrix} 1 & 0_{1\times m-1} \\ 0_{m-1\times 1} & 0_{m-1\times m-1} \end{bmatrix} + o(1)$$

and for large $t \notin \{t : r_t = e_1\}$ we have, since $G = I$, that

$$\begin{aligned}\operatorname{tr}(r_t r_t^\top G_2 G_2^\top) &= \operatorname{tr}\left(r_t r_t^\top \begin{bmatrix} 0 & 0_{1\times m-1} \\ 0_{m-1\times 1} & I_{m-1\times m-1} \end{bmatrix}\right) + o(1) \\ &= 1 + o(1).\end{aligned}$$

### C. Spectral Information Comparison

The main reason for replacing the matrix $H_2$ with $G_2$ 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 the gradient of the bias, 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 (11). To this end, we perform a perturbation analysis of the subspace spanned by $G_2$.

*Lemma 4.5:* Suppose that $(r_t, B)$ is $\gamma$-uninformative. Then for any regular and $\alpha$-fast convergent policy $\hat{K}_t$, one has that

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

for some orthonormal matrix $P_t$.

*Proof:* The proof relies on Wedin's $\sin\Theta$ Theorem [11] (quoted in the appendix) which describes the perturbation theory of range and nullspace in the singular value decomposition. By (8) it follows that

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

using $\alpha$-fast convergence[3]. 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 [11] 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 [12] 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 [12] 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 4.6:* 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

Before proving our main result, 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 $\gamma$-uninformative pair $(r_t, B)$, and regular and $\alpha$-fast convergent policy $\hat{K}_t$, 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 4.5 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 4.5 is orthonormal. Multiplication by $\Lambda_2^\dagger \succeq 0$ rescales the bound by $\sigma_{\min}(\Lambda_2^\dagger)$. ∎

*Theorem 5.2:* Suppose that $(r_t, B)$ is $\gamma$-uninformative for at least a constant fraction of time for some $\gamma > 0$ and that $\lambda > 0$. Then any regular $\alpha$-fast convergent policy with $\alpha \geq 1/3$, (1) has regret asymptotically lower bounded as

$$R_T = \Omega(T^{2/3}). \tag{13}$$

Moreover, under the additonal hypothesis of $\beta$-unbiasedness with $\beta \geq \min(1 - 2\alpha, 1/2)$, then for $\alpha < 1/3$ it holds that

$$R_T = \Omega(\sqrt{T}). \tag{14}$$

Before proceeding with the proof, some remarks are in order. We see that the result splits into two cases, $\alpha \geq 1/3$ and $\alpha < 1/3$. In the first regime, ill-conditioning of Fisher information translates into bad estimation; trying to converge

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

too fast convergence implies too much exploitation and too lite exploration. Indeed, the analysis shows that it is *impossible* for any adaptive control law $\hat{K}_t$ to converge at a rate faster than $t^{-1/3}$ in the class of uninformative instances. In the second regime, convergence is slower and there is a trade-off between exploration and exploitation, but as $\alpha \to 0$ the input signal simply becomes too different from the optimal input to be efficient.

Moreover, at $\alpha = 1/3$, there is a noise barrier in our analysis if we do not provide further control on the bias. It is also clear that the lower bound (14) holds uniformly for all $\alpha$ if one assumes $\beta \geq 1/2$. This assumption is not particularly strong, since for instance maximum likelihood estimation typically[4] has a bias of of order $1/T$, [13], which is thus covered. Note that this restriction is also in some sense optimal, since otherwise one could interpolate between some algorithm and random guessing of one particular parameter and the interpolating algorithm would achieve locally, at that parameter, better performance (and the lower bound does not depend on the parameter other than through uninformativeness).

*Proof:* To emphasize that $\Lambda_2^\dagger$ depends on $t$, we now write $\Lambda_2^\dagger(t)$. Combining the regret decomposition, Lemma 3.1, with Theorem 4.2 and Lemma 5.1 offers

$$\begin{aligned} &R_T \\ &\geq \lambda \sum_{t=1}^{T} \left[ \operatorname{tr}\left( r_t r_t^\top \mathbf{E}\left[ (K-\hat{K}_t)(K-\hat{K}_t)^\top \right] \right) \right] + o(T^{1-\beta}) \\ &\quad \geq \lambda \sum_{t=1}^{T} \left[ \operatorname{tr}\left( \left( I \otimes r_t r_t^\top \right) H_2 \Lambda_2^\dagger(t) H_2^\top \right) \right] + o(T^{1-\beta}) \\ &\quad\quad \geq \lambda \sum_{t=1}^{T} \left[ \varepsilon \sigma_{\min}(\Lambda_2^\dagger(t))(1+o(1)) \right] + o(T^{1-\beta}), \quad (15) \end{aligned}$$

where we simply discarded the first nonnegative term of Lemma 3.1 and where automatically $\beta \geq \alpha$ (the bias decays at least as fast as the overall convergence rate).

Now, since the policy is $\alpha$-fast convergent, it clear from (8) that $\sigma_{\max}(\Lambda_2(t)) = o(t^{1-2\alpha})$ and hence $\sigma_{\max}(\Lambda_2^\dagger(t)) = \Omega(t^{2\alpha-1})$. From this, we gather that

$$\sum_{t=1}^{T} \left[ \varepsilon \sigma_{\min}(\Lambda_2^\dagger(t))(1+o(1)) \right] = \Omega(T^{2\alpha}).$$

The result now follows by optimizing the trade-off between $\Omega(T^{2\alpha})$ and $o(T^{1-\beta}), \beta \geq \alpha$. In particular, if $\alpha \geq 1/3$ we always have at least $R_T = \Omega(T^{2/3})$.

Suppose instead that $\alpha < 1/3$. Just above we may then conclude that $R_T = \Omega(T^{2\alpha})$ which can be made small. However, note that if the policy is not $\alpha'$-fast convergent, one has that

$$\lambda \sum_{t=1}^{T} \left[ \operatorname{tr}\left( r_t r_t^\top \mathbf{E}\left[ \underbrace{(K-\hat{K}_t)(K-\hat{K}_t)^\top}_{\Omega(t^{-2\alpha'})} \right] \right) \right] = \Omega(T^{1-2\alpha'}). \quad (16)$$

[4]Assume for instance that the inputs are bounded away from zero asymptotically.

Further optimization thus yields for $\alpha \in [0, 1/3)$

$$R_T = \Omega(\min(T^{2\alpha}, T^{1-2\alpha})) = \Omega(\sqrt{T}),$$

which is attained at $\alpha = 1/4$. If $\beta \geq \min(1-2\alpha, 1/2)$ this term dominates the term $o(T^{1-\beta})$. ∎

*Remark 5.3:* If (12) is streghthened to

$$\operatorname{tr}\left( \left( I \otimes B r_t r_t^\top B^\top \right) G_2 G_2^\top \right) > \gamma \quad (17)$$

the first term of the regret decomposition in Lemma 3.1 can also be lower-bounded by the same arguments and the theorem extends to the case $\lambda = 0$ in this way.

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 that $\lambda > 0$. Then for any regular $\alpha$-fast convergent policy, which is also $\beta$-unbiased $\beta \geq 1$, (1) satisfies the regret lower bound

$$R_T = \Omega(\log T). \quad (18)$$

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:* Applying Cramér-Rao and comparing the lower bound $I_{t,K}^\dagger$ to a suitable average gives

$$\begin{aligned} &\mathbf{E}\left[ \operatorname{vec}(K-\hat{K}_t) \operatorname{vec}(K-\hat{K}_t)^\top \right] \succeq I_{t,K}^\dagger \\ &\qquad = [\nabla_B \operatorname{vec}(\hat{K}_t)] I_{t,B}^\dagger [\nabla_B \operatorname{vec}(\hat{K}_t)]^\top \\ &= [\nabla_B \operatorname{vec} K + o(1)](I_{t,B}^* + o(t^{1-2\alpha}))[\nabla_B \operatorname{vec} K + o(1)]^\top. \end{aligned}$$

We may assume $\alpha \geq 1/3$ for otherwise (16) still holds. Hence, for $C = \mathbf{E}\left[ \operatorname{vec}(K-\hat{K}_t) \operatorname{vec}(K-\hat{K}_t)^\top \right]$, and $h = [\nabla_B \operatorname{vec} K + o(1)]$, we have that

$$\begin{aligned} R_T &\geq \lambda \sum_{t=1}^{T} \left[ \operatorname{tr}\left( \left( I \otimes r_t r_t^\top \right) C \right) \right] + o(\log T) \\ &\geq \lambda \sum_{t=1}^{T} \left[ \operatorname{tr}\left( \left( I \otimes r_t r_t^\top \right) h \left( \delta t I + o(t) \right)^\dagger \right) h^\top \right] + o(\log T) \\ &= \Omega(\log T) \end{aligned}$$

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 2.5 to understand why logarithmic rates are feasible in [5].

*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 [5] discussed in Example 2.5.

## VI. Discussion

Our analysis here is similar spirit to [8]. The largest difference is that we take a closer look at the Cramér-Rao bound when the information matrix is near 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, [11]. 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 [9] attaining regret of order $\sqrt{T}$ for the full LQR. This should be contrasted with the SISO setting in Guo [5] elegantly proving the attainability of logarithmic rates in that case.

We also wish to mention that the concept of $\alpha$-fast convergence is much inspired by the notion of uniformly fast convergence for the related bandit problems, see [14] and [15]. 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 also [16] for an overview of bandits.

### A. Future Work

It would be very desirable to extend the present results to the full dynamic LQR and in particular, there, find conditions which ensure a similar phase transition as observed here. This would fully resolve the attainability issue concerning $\log T$ versus $\sqrt{T}$ regret, which we here have only begun to investigate. Moreover, we remark that the $(1-2\alpha)$-unbiasedness condition needed to prove the final part of Theorem 5.2 is not the most elegant condition. However unlikely it might seem that a policy which does not converge "fast" – say faster than $\alpha \geq 1/3$ – has any chance of attaining logarithmic regret in general, it would be satisfying to tighten the regime $\alpha < 1/3$. An interesting direction suggested in [8] is to make use of concepts from asymptotic statistics [17], such as local asymptotic normality theory. Another potential direction is to take a more information-theoretic route toward lower bounds as in [18] and as is traditional for bandits [14]. This was recently done for system identification in [19]. Finally, we note that our work opens up an interesting line of research: Is it possible to construct an algorithm which adaptively and without prior knowledge, attains a rate of $\sqrt{T}$ in the uninformative case and $\log T$ else?

## References


[1] 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.

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

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

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

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

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

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

[8] I. Ziemann and H. Sandberg, "Regret lower bounds for unbiased adaptive control of linear quadratic regulators," *IEEE Control Systems Letters, forthcoming*, 2020.

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

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

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

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

[18] 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.

[19] 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 [11]. The formulation of the $\sin\Theta$ distance as an optimization over the norm $\|\cdot\|_\infty$ appears for instance in [12].