Nonasymptotic Regret Analysis of Adaptive Linear Quadratic Control with Model Misspecification

Bruce D. Lee[1] BRUCELE@SEAS.UPENN.EDU
Anders Rantzer[2] ANDERS.RANTZER@CONTROL.LTH.SE
Nikolai Matni[1] NMATNI@SEAS.UPENN.EDU
[1] *Department of Electrical and Systems Engineering, University of Pennsylvania*
[2] *Department of Automatic Control, Lund University*

Editors: A. Abate, K. Margellos, A. Papachristodoulou

Abstract

The strategy of pre-training a large model on a diverse dataset, then fine-tuning for a particular application has yielded impressive results in computer vision, natural language processing, and robotic control. This strategy has vast potential in adaptive control, where it is necessary to rapidly adapt to changing conditions with limited data. Toward concretely understanding the benefit of pre-training for adaptive control, we study the adaptive linear quadratic control problem in the setting where the learner has prior knowledge of a collection of basis matrices for the dynamics. This basis is misspecified in the sense that it cannot perfectly represent the dynamics of the underlying data generating process. We propose an algorithm that uses this prior knowledge, and prove upper bounds on the expected regret after T interactions with the system. In the regime where T is small, the upper bounds are dominated by a term that scales with either $\texttt{poly}(\log T)$ or $\sqrt{T}$, depending on the prior knowledge available to the learner. When T is large, the regret is dominated by a term that grows with δT, where δ quantifies the level of misspecification. This linear term arises due to the inability to perfectly estimate the underlying dynamics using the misspecified basis, and is therefore unavoidable unless the basis matrices are also adapted online. However, it only dominates for large T, after the sublinear terms arising due to the error in estimating the weights for the basis matrices become negligible. We provide simulations that validate our analysis. Our simulations also show that offline data from a collection of related systems can be used as part of a pre-training stage to estimate a misspecified dynamics basis, which is in turn used by our adaptive controller.

1. Introduction

Transfer learning, whereby a model is pre-trained on a large dataset, and then finetuned for a specific application, has exhibited great success in computer vision (Dosovitskiy et al., 2020) and natural language processing (Devlin et al., 2018). Efforts to apply these methods to control have shown exciting preliminary results, particularly in robotics (Dasari et al., 2019). The principle underpinning the success of transfer learning is to use diverse datasets to extract compressed, broadly useful features, which can be used in conjunction with comparatively simple models for downstream objectives. These simple models can be finetuned with relatively little data from the downstream task. However, errors in the pre-training stage may cause this two-step strategy to underperform learning from scratch when ample task-specific data is available. This tradeoff may be acceptable in settings such as adaptive control, where the learner must rapidly adapt to changes with limited data.

Driven by the potential of pre-training in adaptive control, we study the adaptive linear quadratic regulator (LQR) in a setting where imperfect prior information about the system is available. The

adaptive LQR problem consists of a learner interacting with an unknown linear system

$$x_{t+1} = A^\star x_t + B^\star u_t + w_t, \tag{1}$$

with state x_t, input u_t, and noise w_t assuming values in $\mathbb{R}^{d_{\mathsf{X}}}$, $\mathbb{R}^{d_{\mathsf{U}}}$, and $\mathbb{R}^{d_{\mathsf{X}}}$, respectively. The learner is evaluated by its ability to minimize its *regret*, which compares the cost incurred by playing the learner for T time steps with the cost attained by the optimal LQR controller. Prior work has studied the adaptive LQR problem in the settings where $A^\star$, $B^\star$, or both $A^\star$ and $B^\star$ are fully unknown to the learner and with either stochastic or bounded adversarial noise processes (Abbasi-Yadkori and Szepesvári, 2011; Hazan et al., 2020; Cassel et al., 2020). In this work, we analyze a setting where the learner has a set of basis matrices for the system dynamics; however, $\begin{bmatrix} A^\star & B^\star \end{bmatrix}$ is not in the subspace spanned by this basis, leading to misspecification.

1.1. Related Work

Adaptive Control Originating from autopilot development for high-performance aircraft in the 1950s (Gregory, 1959), adaptive control addressed the need for controllers to adapt to varying altitude, speed, and flight configuration (Stein, 1980). Interest in adaptive control theory grew over the subsequent decades, notably advanced by the landmark paper by Åström and Wittenmark (1973), who studied self tuning regulators. The history of adaptive control is documented in numerous texts (Åström and Wittenmark, 2013; Ioannou and Sun, 1996; Narendra and Annaswamy, 2012).

Nonasymptotic Adaptive LQR The non-asymptotic study of the adaptive LQR problem was pioneered by Abbasi-Yadkori and Szepesvári (2011). Subsequent works (Dean et al., 2018; Cohen et al., 2019; Mania et al., 2019) developed computationally efficient algorithms with upper bounds on the regret scaling with $\sqrt{T}$. Simchowitz and Foster (2020) provide lower bounds which show that the rate $\sqrt{d_{\mathsf{U}}^2 d_{\mathsf{X}} T}$ is optimal when the system is entirely unknown. The dependence of the lower bounds on system theoretic constants is refined by Ziemann and Sandberg (2022), enabling Tsiamis et al. (2022) to show the regret may depend exponentially on d_{X} for specific classes of systems. If either $A^\star$ or $B^\star$ are known, then the regret bounds may be improved to $\mathtt{poly}(\log T)$ (Cassel et al., 2020; Jedra and Proutiere, 2022). Alternative formulations of the adaptive LQR problem consider bounded non-stochastic disturbances (Hazan et al., 2020; Simchowitz et al., 2020) and minimax settings (Rantzer, 2021; Cederberg et al., 2022; Renganathan et al., 2023). In contrast to existing work studying the adaptive LQR problem from a nonasymptotic perspective, we consider bounded misspecification between a representation estimate for the dynamics and the data generating process.

Multi-task Representation Learning The source of mispecification we consider is inspired by theoretical work studying multi-task representation learning in the linear regression setting (Du et al., 2020; Tripuraneni et al., 2020). These papers have been followed by a collection of work studying the use of multi-task representation learning in the presence of data correlated across time, which arises in system identification (Modi et al., 2021; Zhang et al., 2023b) and imitation learning (Zhang et al., 2023a). All of these works show that by pre-training a shared representation on a set of source tasks, the sample complexity of learning the target task may be reduced.

1.2. Contributions

We introduce a notion of misspecification between an estimate for the basis of the dynamics and the data generating process. We then propose an adaptive control algorithm that uses this misspecified basis, and subsequently analyze the regret of this algorithm. This leads to the following insights:

- Our results generalize the understanding of when logarithmic regret is possible in adaptive LQR beyond the cases of known $A^\star$ or $B^\star$ studied by Cassel et al. (2020); Jedra and Proutiere (2022).

- When misspecification is present, the regret incurs a term that is linear in T. The coefficient for this term decays gracefully as the level of misspecification diminishes. As a result, small misspecification means the regret is dominated by sublinear terms in the low data regime of interest for adaptive control. These terms are favorable to those possible in the absence of prior knowledge.

We validate our theory with numerical experiments, and show the benefit using a dynamics representation determined by pre-training on offline data from related systems for adaptive control.

1.3. Notation

The Euclidean norm of a vector x is denoted $\|x\|$. For a matrix A, the spectral norm is denoted $\|A\|$, and the Frobenius norm is denoted $\|A\|_F$. We use $\dagger$ to denote the Moore-Penrose pseudo-inverse. The spectral radius of a square matrix is denoted $\rho(A)$. The minimum eigenvalue of a symmetric, positive definite matrix A is denoted $\lambda_{\min}(A)$For $f, g : D \to \mathbb{R}$, we write $f \lesssim g$ if for some $c > 0$, $f(x) \leq cg(x)\, \forall x \in D$. We denote the solutions to the discrete Lyapunov equation by $\texttt{dlyap}(A, Q)$ and the discrete algebraic Riccati equation by $\texttt{DARE}(A, B, Q, R)$.

2. Problem Formulation

2.1. System model

We consider the system (1) where the noise w_t has independent identically distributed elements that are mean zero and σ^2-sub-Gaussian for some $\sigma^2 \in \mathbb{R}$ with $\sigma^2 \geq 1$. We additionally assume that the noise has identity covariance: $\mathbf{E}\left[w_t w_t^\top\right] = I$.[1] We suppose the dynamics admit the decomposition

$$\begin{bmatrix} A^\star & B^\star \end{bmatrix} = \mathsf{vec}^{-1}(\Phi^\star \theta^\star), \tag{2}$$

where $\Phi^\star \in \mathbb{R}^{d_{\mathsf{X}}(d_{\mathsf{X}}+d_{\mathsf{U}}) \times d_\theta}$ specifies the model structure, and has orthonormal columns. Meanwhile, $\theta^\star \in \mathbb{R}^{d_\theta}$ specifies the parameters. The operator vec^{-1} maps a vector in $\mathbb{R}^{d_{\mathsf{X}}(d_{\mathsf{X}}+d_{\mathsf{U}})}$ into a matrix in $\mathbb{R}^{d_{\mathsf{X}} \times (d_{\mathsf{X}}+d_{\mathsf{U}})}$ by stacking length d_{X} blocks of the the vector into columns of the matrix, working top to bottom and left to right. We can write this as a linear combination of basis matrices:

$$\begin{bmatrix} A^\star & B^\star \end{bmatrix} = \sum_{i=1}^{d_\theta} \theta_i^\star \begin{bmatrix} \Phi_i^{A,\star} & \Phi_i^{B,\star} \end{bmatrix}, \text{ where } \begin{bmatrix} \Phi_i^{A,\star} & \Phi_i^{B,\star} \end{bmatrix} = \mathsf{vec}^{-1}\, \Phi_i^\star,$$

and $\Phi_i^\star$ is the i^{th} column of $\Phi^\star$. This decomposition of the data generating process is a natural extension of the low-dimensional linear representations considered in Du et al. (2020) to the setting of multiple related dynamical systems with shared structure determined by $\Phi^\star$. It captures many practically relevant settings, such as when $A^\star$ and $B^\star$ depend on a few physical parameters $\theta^\star$, and $\Phi^\star$ describes the structure through which these physical parameters enter the dynamics.

We assume that both $\Phi^\star$ and $\theta^\star$ are unknown; however, we have an estimate $\hat{\Phi} \in \mathbb{R}^{d_{\mathsf{X}}(d_{\mathsf{X}}+d_{\mathsf{U}}) \times d_\theta}$, also with orthonormal columns, for $\Phi^\star$. Such an estimate may be obtained by performing a pre-training step on offline data from a collection of systems related to (1) by the shared matrix $\Phi^\star$ in (2). Due to the noise present in the offline data, this estimate will be imperfect, resulting in misspecification.[2] To quantify the level of misspecification between this estimate and the underlying data generating process, we use the following subspace distance metric.

Definition 1 ((Stewart and Sun, 1990)) *Let $\Phi_\perp^\star$ complete the basis of $\Phi^\star$ such that $\begin{bmatrix} \Phi^\star & \Phi_\perp^\star \end{bmatrix}$ is an orthogonal matrix. Then the* subspace distance *between $\Phi^\star$ and $\hat{\Phi}$ is $d(\Phi^\star, \hat{\Phi}) \triangleq \left\| \hat{\Phi}^\top \Phi_\perp^\star \right\|$.*

1. Noise that enters the process through a non-singular matrix H can be addressed by rescaling the dynamics by H^{-1}.
2. Sample complexity bounds for learning $\hat{\Phi}$ are provided by Zhang et al. (2023b), so this step is not studied here.

Note that the above distance is small when the range of the matrices $\hat{\Phi}$ and $\Phi^\star$ is similar.[3]

As long as $d(\Phi^\star, \hat{\Phi})$ is sufficiently small and the dimension $d_\theta < d_{\mathsf{X}}(d_{\mathsf{X}} + d_{\mathsf{U}})$, the estimate $\hat{\Phi}$ allows the learner to fit a model with less data than would be required to estimate $\begin{bmatrix} A^\star & B^\star \end{bmatrix}$ from scratch. This benefit comes at the cost of a bias in the learner's model that grows with $d(\Phi^\star, \hat{\Phi})$.

2.2. Learning objective

The goal of the learner is to interact with system (1) while keeping the cumulative cost small, where the cumulative cost is defined for matrices $Q \succeq I$ and $R = I$ as[4]

$$C_T \triangleq \sum_{t=1}^{T} c_t, \text{ where } c_t \triangleq x_t^\top Q x_t + u_t^\top R u_t. \tag{3}$$

To define an algorithm that keeps the cost small, we first introduce the infinite horizon LQR cost:

$$\mathcal{J}(K) \triangleq \limsup_{T\to\infty} \frac{1}{T} \mathbf{E}^K C_T, \tag{4}$$

where the superscript of K on the expectation denotes that the cost is evaluated under the state feedback controller $u_t = Kx_t$. To ensure that there exists a controller such that (4) has finite cost, we assume $(A^\star, B^\star)$ is stabilizable. Under this assumption, (4) is minimized by the LQR controller $K_\infty(A^\star, B^\star)$, where $K_\infty(A, B) \triangleq -(B^\top P_\infty(A,B)B + R)^{-1} B^\top P_\infty(A,B)A$, and $P_\infty(A,B) \triangleq \mathtt{DARE}(A, B, Q, R)$. We define the shorthands $P^\star \triangleq P_\infty(A^\star, B^\star)$ and $K^\star \triangleq K_\infty(A^\star, B^\star)$. To characterize the infinite horizon LQR cost of an arbitrary stabilizing controller K, we additionally define the solution P_K to the lyapunov equation for the closed loop system under an arbitrary K such that $\rho(A^\star + B^\star K) < 1$: $P_K \triangleq \mathtt{dlyap}(A^\star + B^\star K, Q + K^\top R K)$. For a controller K satisfying $\rho(A^\star + B^\star K) < 1$, $\mathcal{J}(K) = \mathrm{tr}(P_K)$. We have that $P_{K^\star} = P^\star$.

The infinite horizon LQR controller provides a baseline level of performance that our learner cannot surpass in the limit as $T \to \infty$. Borrowing the notion of regret from online learning, as in Abbasi-Yadkori and Szepesvári (2011), we quantify the performance of our learning algorithm by comparing the cumulative cost C_T to the scaled infinite horizon cost attained by the LQR controller if the system matrices $\begin{bmatrix} A^\star & B^\star \end{bmatrix}$ were known:

$$\mathbf{R}_T \triangleq C_T - T\mathcal{J}(K^\star). \tag{5}$$

In light of the above reformulation, the goal of the learner is to interact with the system (1) to maximize the information about the relevant parameters for control while simultaneously regulating the system to minimize $\mathbf{R}_T$. A reasonable strategy to do so is for the learner to use its history of interaction with the system to construct a model for the dynamics, e.g. by determining estimates $\hat{A}$ and $\hat{B}$. It may then use these estimates as part of a *certainty equivalent* (CE) design by synthesizing controllers $\hat{K} = K_\infty(\hat{A}, \hat{B})$. Prior work has shown that if the model estimate is sufficiently close to the true dynamics, then the cost of playing the controller $\hat{K}$ exceeds the cost of playing $K^\star$ by a quantity that is quadratic in the estimation error (Mania et al., 2019; Simchowitz and Foster, 2020).

Lemma 1 (Theorem 3 of Simchowitz and Foster (2020)) *Define* $\varepsilon \triangleq \frac{1}{2916\|P^\star\|^{10}}$. *If* $\left\|\begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} - \begin{bmatrix} A^\star & B^\star \end{bmatrix}\right\|_F^2 \le \varepsilon$, *then* $\mathcal{J}(\hat{K}) - \mathcal{J}(K^\star) \le 142\,\|P^\star\|^8 \left\|\begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} - \begin{bmatrix} A^\star & B^\star \end{bmatrix}\right\|_F^2$.

3. This distance may be small when $\left\|\hat{\Phi} - \Phi^\star\right\|$ is not. However, small $\left\|\hat{\Phi} - \Phi_\star\right\|$ implies small subspace distance.

4. Generalizing to arbitrary $Q \succ 0$ and $R \succ 0$ can be performed by scaling the cost and changing the input basis.

Algorithm 1 Certainty Equivalent Control with Continual Exploration

Input: Stabilizing controller K_0, initial epoch length τ_1, number of epochs k_{fin}, exploration sequence $\sigma_1^2, \sigma_2^2, \sigma_3^2, \dots \sigma_{k_{\mathsf{fin}}}^2$, state bound x_b, controller bound K_b
Initialize: $\hat{K}_1 \leftarrow K_0$, $\tau_0 \leftarrow 0$, $T \leftarrow \tau_1 2^{k_{\mathsf{fin}}-1}$.
for $k = 1, 2, \dots, k_{\mathsf{fin}}$ **do**
  for $t = \tau_{k-1}+1, \dots, \tau_k$ **do**
    if $\|x_t\|^2 \geq x_b^2 \log T$ or $\left\|\hat{K}_k\right\| \geq K_b$ **then** Abort and play K_0 forever
    Play $u_t = \hat{K}_k x_t + \sigma_k g_t$, where $g_t \sim \mathcal{N}(0, I)$
  $\hat{\theta}_k \leftarrow \mathtt{LS}(\hat{\Phi}, x_{\tau_{k-1}+1:\tau_k+1}, u_{\tau_{k-1}+1:\tau_k})$ ▷ Algorithm 2
  $\begin{bmatrix}\hat{A}_k & \hat{B}_k\end{bmatrix} \leftarrow \mathsf{vec}^{-1}\left(\hat{\Phi}\hat{\theta}_k\right)$
  $\hat{K}_{k+1} \leftarrow K_\infty(\hat{A}_k, \hat{B}_k)$
  $\tau_{k+1} \leftarrow 2\tau_k$

2.3. Algorithm description

Our proposed algorithm, Algorithm 1, is a CE algorithm akin to that proposed in Cassel et al. (2020). The algorithm takes a stabilizing controller K_0 as an input. Starting from this controller, Algorithm 1 follows a doubling epochs strategy. At the end of each epoch, it uses the data collected during the epoch along with the estimate for $\hat{\Phi}$ to obtain an estimate for $\hat{\theta}$ by solving a least squares problem (as detailed in Algorithm 2). The estimated parameters $\hat{\theta}$ are combined with the estimate $\hat{\Phi}$ to obtain the dynamics estimate $\begin{bmatrix}\hat{A} & \hat{B}\end{bmatrix}$. This estimate is used to synthesize a CE controller $\hat{K} = K_\infty(\hat{A}, \hat{B})$. In the next epoch, the learner plays the resultant controller with exploratory noise added. Before playing each input, the algorithm checks whether the state or the controller exceed bounds determined by algorithm inputs x_b and K_b. If they do, it aborts the certainty equivalent scheme, and plays the initial stabilizing controller for all time. Doing so enables bounding of the regret during unlikely events where the CE controller fails. The key difference from the CE algorithms proposed in prior work is that the system identification step solves a least squares problem defined in terms of the estimate $\hat{\Phi}$ to estimate the unknown parameters.

Algorithm 2 Least squares: $\mathtt{LS}(\hat{\Phi}, x_{1:t+1}, u_{1:t+1})$

Input: Model structure estimate $\hat{\Phi}$, state data $x_{1:t+1}$, input data $u_{1:t}$
Return: $\hat{\theta} = \Lambda^\dagger \left(\sum_{s=1}^t \hat{\Phi}^\top \left(\begin{bmatrix}x_s \\ u_s\end{bmatrix} \otimes I_{d_{\mathsf{X}}}\right) x_{s+1}\right)$, where $\Lambda = \sum_{s=1}^t \hat{\Phi}^\top \left(\begin{bmatrix}x_s \\ u_s\end{bmatrix}\begin{bmatrix}x_s \\ u_s\end{bmatrix}^\top \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}$.

3. Regret Bounds

We now present our bounds on the expected[5] regret incurred by Algorithm 1. Further discussion and complete proofs may be found in Lee et al. (2023).

Consider running Algorithm 1 for $T = \tau_1 2^{k_{\mathsf{fin}}-1}$ timesteps, where τ_1 is the initial epoch length and k_{fin} is the number of epochs. To bound the regret incurred by this algorithm, we decompose the regret into that achieved by the algorithm under a high probability success event, and that incurred

5. In contrast to high probability regret bounds, expected regret provides an understanding of what happens in the unlikely events where controller performs poorly.

during a failure event under which the state or controller bound in line 5 of Algorithm 1 are violated. To ensure the failure event occurs with a small probability, we make the following assumption on the state and controller bounds, which uses the shorthand $\Psi_{B^\star} \triangleq \max\{1, \|B^\star\|\}$.

Assumption 1 *We assume that* $x_b \geq 400 \|P_{K_0}\|^2 \Psi_{B^\star} \sigma \sqrt{d_{\mathsf{X}} + d_{\mathsf{U}}}$ *and* $K_b \geq \sqrt{2\|P_{K_0}\|}$.

We make additional assumptions about the remaining arguments supplied as inputs to the algorithm in two cases: one where no additional assumptions about the dynamics are made (Section 3.1), and one where we assume the system structure estimate is such that the initial controller and the optimal controller provide sufficient excitation to identify the unknown parameters (Section 3.2).

3.1. Certainty equivalent control with continual exploration

To ensure an estimate satisfying the condition in Lemma 1 is attainable, the gap between the model structure estimate and the ground truth cannot be too large, leading to the following assumption.

Assumption 2 *Let* ε *be as in Lemma 1 and* K_0 *be an initial stabilizing controller. Define* $\beta_1 \triangleq C_{\mathsf{bias},1} \sigma^4 \|P_{K_0}\|^{12} \Psi_{B^\star}^8 \|\theta^\star\|^2 (d_{\mathsf{X}} + d_{\mathsf{U}}) \sqrt{\frac{d_\theta}{d_{\mathsf{U}}}}$ *for a sufficiently large universal constant* $C_{\mathsf{bias},1}$. *We assume our representation error satisfies* $d(\hat{\Phi}, \Phi^\star) \leq \frac{\varepsilon^2}{4\beta_1^2}$.

The requirement above arises from the way that misspecification enters our bounds on the estimation error $\left\| \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} - \begin{bmatrix} A^\star & B^\star \end{bmatrix} \right\|_F^2$. See the definition of $\mathcal{E}_{\mathsf{est},1}$ in Section 3.3.

In this setting, we run Algorithm 1 with exploratory inputs injected to ensure identifiability of the unknown parameters. Doing so provides the regret guarantees in the following theorem.

Theorem 2 *Consider applying Algorithm 1 with initial stabilizing controller* K_0 *for* $T = \tau_1 2^{k_{\mathsf{fin}}} - 1$ *timesteps for some positive integers* k_{fin}, *and* τ_1. *Suppose that for some* $\gamma \geq 1$, *the exploration sequence is given by* $\sigma_k^2 = \max\left\{ \frac{\sqrt{d_{\mathsf{U}}/d_\theta}}{\sqrt{\tau_1 2^{k-1}}}, \gamma d(\hat{\Phi}, \Phi^\star)^{1/2} \right\} \forall k \geq 1$.[6] *Suppose the state bound* x_b *and the controller bound* K_b *satisfy Assumption 1 and that* $\hat{\Phi}$ *satisfies Assumption 2. Let* ε *be as in Lemma 1. There exists a universal positive constant* $C_{\mathsf{warm\ up}}$ *such that if* $\tau_1 = \tau_{\mathsf{warm\ up}} \log^2 T$ *for*

$$\tau_{\mathsf{warm\ up}} \geq C_{\mathsf{warm\ up}} \sigma^4 \|P_{K_0}\|^3 \max\left\{ \Psi_{B^\star}^2 (d_{\mathsf{X}} + d_{\mathsf{U}}), x_b^2, -\log_{\left(1 - \frac{1}{\|P^\star\|}\right)} \|P^\star\|, \left(\sqrt{d_\theta d_{\mathsf{U}}/\varepsilon}\right)^2 \right\},$$

then the expected regret satisfies

$$\boldsymbol{E}[\mathbf{R}_T] \leq c_0 \log^2(T) + c_1 \sqrt{d_\theta d_{\mathsf{U}}} \sqrt{T} \log T + c_2 \sqrt{d(\hat{\Phi}, \Phi^\star)} T,$$

where $c_0 = \texttt{poly}(d_{\mathsf{X}}, d_{\mathsf{U}}, \|P_{K_0}\|, \Psi_{B^\star}, \tau_{\mathsf{warm\ up}}, x_b, K_b)$, $c_1 = \texttt{poly}(\|P_{K_0}\|, \Psi_{B^\star}, \sigma)$ *and* $c_2 = \texttt{poly}(d_{\mathsf{U}}, d_{\mathsf{X}}, d_\theta, \|P_{K_0}\|, \Psi_{B^\star}, \|\theta^\star\|, \sigma, \gamma)$.

The constants c_0 and c_2 in the above bound depend on system dimensions, system-theoretic quantities, and algorithm parameters including the state and controller bounds, the initial epoch length, and the initial controller. In contrast, the constant c_1, does not depend on system dimension. It is presented as such to emphasize that the dimensional dependence of the order $\sqrt{T}$ term is $\sqrt{d_\theta d_{\mathsf{U}}}$. This elucidates the dependence on the system and parameter dimensions in the regime

6. The γ allows the sequence to be defined with a bound on the level of misspecification, rather than precise knowledge.

where the $\sqrt{T}$ term is dominant. Consider the result in the absence of misspecification: $d(\hat{\Phi}, \Phi^\star) = 0$. In this case, the dominant term grows with $\sqrt{d_\theta d_{\mathsf{U}}}\sqrt{T}\log T$. As long as $d_\theta \leq d_{\mathsf{X}} d_{\mathsf{U}}$, this is smaller than the dependence of $\sqrt{d_{\mathsf{U}}^2 d_{\mathsf{X}}}$ which appears in the lower bounds for the regret of learning to control a system with entirely unknown $A^\star$ and $B^\star$ (Simchowitz and Foster, 2020). If the misspecification is nonzero, then the regret bound incurs an additional term that grows linearly with T. However, as long as $d(\hat{\Phi}, \Phi^\star)$ is sufficiently small, there exists a regime of T for which the $\sqrt{T}$ term dominates, and using the misspecified basis provides a benefit over learning from scratch.

3.2. Certainty equivalent control without additional exploration

In this section, we analyze the regret attained under the additional assumption that the process noise fully excites the relevant modes of the system under $K^\star$ and K_0. This may be guaranteed as follows.

Assumption 3 *Let $\hat{\Phi}$ be the estimate for dynamics representation, and let α be a number satisfying $\alpha \geq \frac{1}{3\|P^\star\|^{3/2}}$. We assume that $\lambda_{\min}\left(\hat{\Phi}^\top\left(\begin{bmatrix} I \\ K \end{bmatrix}\begin{bmatrix} I \\ K \end{bmatrix}^\top \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}\right) \geq \alpha^2$ for $K = K_0, K^\star$.*

The above assumption captures a setting where playing either the initial controller K_0 or the optimal controller $K^\star$ provides persistence of excitation without any exploratory input. This can be seen by noting that the matrix $\hat{\Phi}^\top\left(\begin{bmatrix} I \\ K \end{bmatrix}\begin{bmatrix} I \\ K \end{bmatrix}^\top \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}$ is a lower bound (in Loewner order) for the covariance matrix formed by taking the expectation of Λ/t in Algorithm 2 when $u_s = Kx_s$.

Under the above assumption, we may run Algorithm 1 without an additional exploratory input injected. As in Section 3.1, we require that the representation error is small enough to guarantee the closeness condition in Lemma 1 may be satisfied with our estimated model.

Assumption 4 *Let ε be as in Lemma 1, K_0 be a stabilizing controller, and α be a positive number such that Assumption 3 holds. We assume our representation error satisfies $d(\hat{\Phi}, \Phi^\star) \leq \sqrt{\frac{\varepsilon}{2\beta_2}}$, where $\beta_2 \triangleq C_{\mathsf{bias},2}\frac{\varepsilon\|P_{K_0}\|^9\Psi_{B^\star}^8\|\theta^\star\|^2(d_{\mathsf{X}}+d_{\mathsf{U}})}{d_\theta\min\{\alpha^2,\alpha^4\}}$ and $C_{\mathsf{bias},2}$ is a sufficiently large universal constant .*

This requirement again arises from dependence of the estimation error bounds on the misspecification. See the definition of $\mathcal{E}_{\mathsf{est},2}$ in Section 3.3. Under these assumptions, our regret bound may be improved to that in the following theorem.

Theorem 3 *Consider applying Algorithm 1 with initial stabilizing controller K_0 for $T = \tau_1 2^{k_{\mathsf{fin}}}$ timeteps for some positive integers k_{fin}, and τ_1. Additionally suppose the exploration sequence is zero for all time: $\sigma_k^2 = 0$ for $k = 1, \ldots, k_{\mathsf{fin}}$. Suppose the state bound x_b and the controller bound K_b satisfy Assumption 1, and that $\hat{\Phi}$ satisfies Assumption 3 and Assumption 4. Let ε be as in Lemma 1. There exists a positive universal constant $C_{\mathsf{warm\,up}}$ such that if $\tau_1 = \tau_{\mathsf{warm\,up}}\log T$, for*

$$\tau_{\mathsf{warm\,up}} \geq C_{\mathsf{warm\,up}}\sigma^4\|P_{K_0}\|^3\Psi_{B^\star}^2\max\left\{(d_{\mathsf{X}}+d_{\mathsf{U}}), x_b^2, -\log_{\left(1-\frac{1}{2\|P^\star\|}\right)}\|P^\star\|, d_\theta/(2\varepsilon\alpha^2)\right\},$$

then the expected regret satisfies

$$\boldsymbol{E}[\mathbf{R}_T] \leq c_1\log^2(T) + c_2 d(\hat{\Phi}, \Phi^\star)^2 T,$$

where $c_1 = \mathtt{poly}(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, \|P_{K_0}\|, \Psi_{B^\star}, \|\theta^\star\|, \sigma, \alpha^{-1}, \tau_{\mathsf{warm\,up}}, K_b, x_b)$ and $c_2 = \mathtt{poly}(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, \|P_{K_0}\|, \Psi_{B^\star}, \|\theta^\star\|, \alpha^{-1})$.

When the misspecification is zero, the expected regret grows with $\log^2 T$. Prior work (Cassel et al., 2020; Jedra and Proutiere, 2022) has shown that such rates are possible if either $A^\star$ or $B^\star$ are known to the learner. See Lee et al. (2023) for details about how to obtain logarithmic regret in these settings using Theorem 3. The above result expands on prior work by generalizing conditions for prior knowledge that are sufficient to achieve logarithmic regret.

With misspecification present, the above theorem has a term growing linearly T. In contrast to Theorem 2, the coefficient for this term is proportional to the level of misspecification squared, which is smaller than the square root dependence in Theorem 2. This result shows that if some coarse system knowledge depending on a few unknown parameters is available in advance and the unknown parameters are easily identifiable in the sense of Assumption 3, then there exists a substantial regime of T for which the regret incurred is much smaller than that attained by learning to control the system from scratch with fully unknown $A^\star$ and $B^\star$ (where the regret scales as $\sqrt{T}$).

3.3. Proof sketch

Our main result proceeds by first defining a success events for which the certainty equivalent control scheme never aborts, and generates dynamics estimates $\begin{bmatrix}\hat{A}_k & \hat{B}_k\end{bmatrix}$ which are sufficiently close to the true dynamics $\begin{bmatrix}A^\star & B^\star\end{bmatrix}$ at all times.The success events for Section 3.1 and Section 3.2 are $\mathcal{E}_{\mathsf{success},1} = \mathcal{E}_{\mathsf{bound}} \cap \mathcal{E}_{\mathsf{est},1}$ and $\mathcal{E}_{\mathsf{success},2} = \mathcal{E}_{\mathsf{bound}} \cap \mathcal{E}_{\mathsf{est},2}$ respectively, where

$$\mathcal{E}_{\mathsf{bound}} = \left\{\|x_t\|^2 \le x_b^2 \log T \quad \forall t = 1, \dots, T\right\} \cap \left\{\left\|\hat{K}_k\right\| \le K_b\,, \forall k = 1, \dots, k_{\mathsf{fin}}\right\},$$

$$\mathcal{E}_{\mathsf{est},1} = \left\{\left\|\begin{bmatrix}\hat{A}_k & \hat{B}_k\end{bmatrix} - \begin{bmatrix}A^\star & B^\star\end{bmatrix}\right\|_F^2 \le C_{\mathsf{est},1}\frac{\sigma^2\sqrt{d_\theta d_{\mathsf{U}}}\,\|P_{K_0}\|}{\sqrt{\tau_k}}\log T + \beta_1\sqrt{d(\hat{\Phi}, \Phi^\star)}\right\},$$

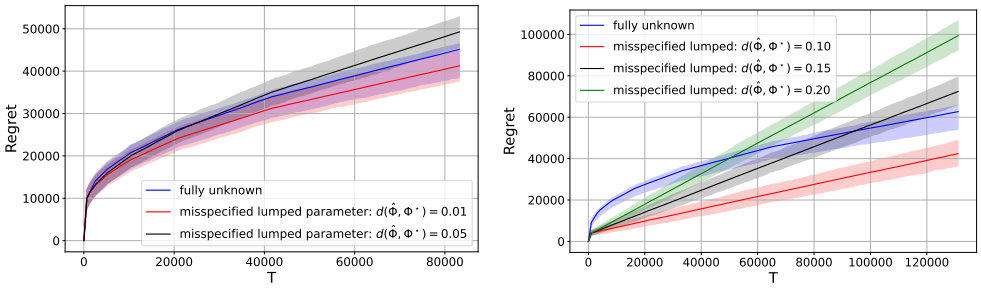
$$\mathcal{E}_{\mathsf{est},2} = \left\{\left\|\begin{bmatrix}\hat{A}_k & \hat{B}_k\end{bmatrix} - \begin{bmatrix}A^\star & B^\star\end{bmatrix}\right\|_F^2 \le C_{\mathsf{est},2}\frac{\sigma^2 d_\theta}{\tau_k \alpha^2}\log T + \beta_2 d(\hat{\Phi}, \Phi^\star)^2\right\},$$

and $C_{\mathsf{est},1}$ and $C_{\mathsf{est},2}$ are positive universal constants.

We use the success events to decompose the expected regret as in Cassel et al. (2020): $\mathbf{E}[\mathbf{R}_T] = R_1 + R_2 + R_3 - T\mathcal{J}(K^\star)$, where for $\mathcal{E}_{\mathsf{success}} = \mathcal{E}_{\mathsf{success},1}$ or $\mathcal{E}_{\mathsf{success}} = \mathcal{E}_{\mathsf{success},2}$,

$$R_1 = \mathbf{E}\left[\mathbf{1}(\mathcal{E}_{\mathsf{success}})\sum_{k=2}^{k_{\mathsf{fin}}} J_k\right], \quad R_2 = \mathbf{E}\left[\mathbf{1}(\mathcal{E}_{\mathsf{success}}^c)\sum_{t=\tau_1+1}^{T} c_t\right], \quad \text{and} \quad R_3 = \mathbf{E}\left[\sum_{t=1}^{\tau_1} c_t\right], \tag{6}$$

are the costs due to success, failure, and the first epoch respectively. Here, J_k are the epoch costs defined as $J_k = \sum_{t=\tau_k}^{\tau_{k+1}} c_t$. In the settings of both Section 3.1 and Section 3.2, R_3 is given by $\tau_1 \operatorname{tr}\big(P_{K_0}(I + \sigma_1^2 B^\star (B^\star)^\top)\big)$, while R_2 is controlled using the upper bounds on the state and controller to obtain a bound on the cost, which is then multiplied by the small probability of the failure event. To control R_1, we show that under the success event, the closeness condition Lemma 1 is satisfied. As a result, the cost of each epoch J_k is $(\tau_k - \tau_{k-1})\mathcal{J}(K^\star)$ in addition to a term proportional to $(\tau_k - \tau_{k-1})\Big(\left\|\begin{bmatrix}\hat{A}_k & \hat{B}_k\end{bmatrix} - \begin{bmatrix}A^\star & B^\star\end{bmatrix}\right\|_F^2 + \sigma_k^2\Big)$. Using the estimation error bounds of events $\mathcal{E}_{\mathsf{est},1}$ and $\mathcal{E}_{\mathsf{est},2}$ along with the choices for σ_k^2 in the two settings, we find that the quantity $R_1 - T\mathcal{J}(K^\star)$ is proportional to $\sqrt{T}\log T + \sqrt{d(\hat{\Phi}, \Phi^\star)T}$ in the setting of Section 3.1, and $\log^2 T + d(\hat{\Phi}, \Phi^\star)^2 T$ in the setting of Section 3.2. Combining terms provides the expected regret bounds in Theorems 2 and 3.



(*a*) Continual exploration (*b*) No exploration

Figure 2: We plot the regret of Algorithm 1 with $\hat{\Phi}$ describing a lumped parameter model (right), or a lumped parameter model and extended such that the condition Assumption 3 is violated (left). In both settings the representations are perturbed, resulting in a misspecification between the true representation $\Phi^\star$ and the representation estimate $\hat{\Phi}$. The regret is compared to that incurred by running Algorithm 1 with a fully unknown $A^\star$ and $B^\star$.

4. Numerical Example

To validate the trends predicted by our bounds, we run Algorithm 1 on the system (1) where $A^\star$ and $B^\star$ are obtained by linearizing and discretizing the cartpole dynamics defined by the equations $(M+m)\ddot{x} + m\ell(\ddot{\theta}\cos(\theta) - \dot{\theta}^2\sin(\theta)) = u$, and $m(\ddot{x}\cos(\theta) + \ell\ddot{\theta} - g\sin(\theta)) = 0$, for cart mass $M = 1$, pole mass $m = 1$, pole length $\ell = 1$, and gravitiy $g = 1$. Discretization uses Euler's approach with stepsize 0.25. The disturbance signal is generated as $w_t \sim \mathcal{N}(0, 0.01I)$.

We consider various inputs for the representation estimate $\hat{\Phi}$ and the exploration sequence $\sigma_1^2, \ldots, \sigma_{k_{\text{fin}}}^2$. The remaining parameters are discussed in Lee et al. (2023).

No Misspecification: In Figure 1, we plot the regret of Algorithm 1 in the absence of misspecification, i.e. $d(\hat{\Phi}, \Phi^\star) = 0$. We consider several instances for the representation: one for fully unknown $A^\star$ and $B^\star$, one which encodes a setting where the $A^\star$ matrix is known up to an unknown scaling, and one which captures a lumped parameter model where the discretized and linearized cartpole structure is known up to scale, but the values of the entries which vary with cart mass, pole mass, and pole length are unknown. We additionally consider extending the lumped parameter representation by adding a basis vector that ensures the condition in Assumption 3 is violated. For the representation capturing fully unknown $A^\star$ and $B^\star$, and the extended lumped parameter representation, the condition in Assumption 3 is not satisfied, so we run Algorithm 1 with exploration noise scaling as $\sigma_k^2 \propto \frac{1}{\sqrt{2^k}}$, and incur $\sqrt{T}$ regret, as predicted by Theorem 2. The extended lumped parameter representation incurs regret at a slower rate than the setting when the system is fully unknown. This is predicted by Theorem 2 due to the fact that the extended lumped parameter model has $d_\theta = 6 < 20 = d_{\mathsf{X}}(d_{\mathsf{X}} + d_{\mathsf{U}})$, so the coefficient on the $\sqrt{T}$ term is smaller. For the remaining settings, Assumption 3 is satisfied, so we run the algorithm with no additional exploration and incur logarithimic regret, as predicted in Theorem 3.

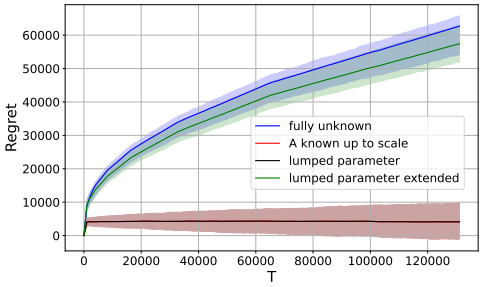


Figure 1: Regret of Algorithm 1 with various choices for $\hat{\Phi}$.

Artificial Misspecification: In Figure 2, we compare the regret from the fully unknown setting to the regret with a misspecified lumped parameter representation. Figure 2(*a*) considers the lumped parameter representation that is extended such that Assumption 3 is violated. Therefore the learner must continually inject noise to the system in order to explore. We artificially create misspecification by adding small perturbations to the true representation such that $d(\hat{\Phi}, \Phi^\star) > 0$. We see that in the low data regime, the regret incurred is less than that incurred when the model is fully unknown. When the misspecification level is 0.05, the bias in the identification due to the misspecification causes the regret to rapidly overtake the regret from the fully unkown setting. When the misspecification is small, the regret remains less than that from the fully unknown setting for the entire horizon of T values that are plotted. Figure 2(*b*) considers the lumped parameter setting without the extension, for which Assumption 3 is satisfied. As in Figure 2(*a*), we add a perturbation to the representation to create misspecification. In this setting, we consider much larger perturbations, such that $d(\hat{\Phi}, \Phi^\star)$ is 0.1, 0.15, or 0.20. For all three such situations, the regret begins much smaller than that of the fully unknown model, but overtakes it as T becomes large.

Learned Representation: In Figure 3, we consider a setting motivated by multi-task learning, in which a representation is learned using offline data from several systems related to the system of interest. In particular, we collect trajectories of length 1200 from five discretized and linearized cartpole systems generated with various values of the parameters (M, m, ℓ). The resulting data is in turn used to fit a a representation $\hat{\Phi}$.[7] Once the representation is obtained, we run Algorithm 1 in the absence of exploratory input. We see that the regret incurred is much lower than the setting in which the dynamics are fully unknown for the small data regime, but overtakes it as T becomes large. This aligns with the results from the artificial misspecification experiment. By computing the distance between the learned and true lumped parameter representations, we find that $d(\hat{\Phi}, \Phi^\star) = 0.2041$.

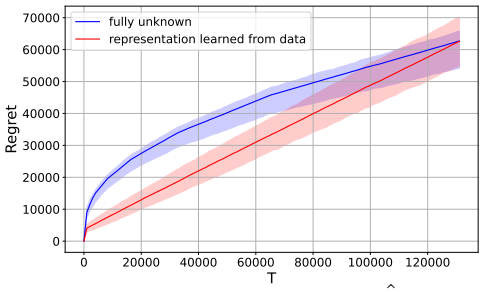


Figure 3: Regret of Alg. 1 with $\hat{\Phi}$ learned offline from related systems.

5. Conclusion

We studied adaptive LQR in the presence of misspecification between the learner's simple model structure for the dynamics, and the true dynamics. Our proposed algorithm performs well in both experiments and theory as long as this misspecification is sufficiently small. Our analysis shows a phase shift in the problem that depends on whether the learner's prior knowledge enables identification of the unknown parameters without additional exploration, thus allowing regret that is logarithmic in T. There are many interesting avenues for future work, including extending the analysis to settings where a collection of nonlinear basis functions for the dynamics are misspecified, e.g., by modifying the model studied in Kakade et al. (2020). Another direction is to analyze the setting where the learner's simple model is updated online by sharing data from a collection of related systems, as in recent work on federated learning in dynamical systems (Wang et al., 2023).

7. See Lee et al. (2023) for details.

Acknowledgments

We thank Leo Toso, Olle Kjellqvist, Thomas Zhang, and Ingvar Ziemann for helpful comments and discussions.

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. JMLR Workshop and Conference Proceedings, 2011.

Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

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

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*, pages 1328–1337. PMLR, 2020.

Daniel Cederberg, Anders Hansson, and Anders Rantzer. Synthesis of minimax adaptive controller for a finite set of linear systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 1380–1384. IEEE, 2022.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{t}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

PC Gregory. *Proceedings of the Self Adaptive Flight Control Systems Symposium*, volume 59. Wright Air Development Center, Air Research and Development Command, United . . . , 1959.

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

Petros A Ioannou and Jing Sun. *Robust adaptive control*, volume 1. PTR Prentice-Hall Upper Saddle River, NJ, 1996.

Yassir Jedra and Alexandre Proutiere. Minimal expected regret in linear quadratic control. In *International Conference on Artificial Intelligence and Statistics*, pages 10234–10321. PMLR, 2022.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Bruce D Lee, Anders Rantzer, and Nikolai Matni. Nonasymptotic regret analysis of adaptive linear quadratic control with model misspecification. *arXiv preprint arXiv:2401.00073*, 2023.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Aditya Modi, Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Joint learning of linear time-invariant dynamical systems. *arXiv preprint arXiv:2112.10955*, 2021.

Kumpati S Narendra and Anuradha M Annaswamy. *Stable adaptive systems*. Courier Corporation, 2012.

Anders Rantzer. Minimax adaptive control for a finite set of linear systems. In *Learning for Dynamics and Control*, pages 893–904. PMLR, 2021.

Venkatraman Renganathan, Andrea Iannelli, and Anders Rantzer. An online learning analysis of minimax adaptive control. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1034–1039. IEEE, 2023.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

Gunter Stein. Adaptive flight control: A pragmatic view. In *Applications of Adaptive Control*, pages 291–312. Elsevier, 1980.

Gilbert W Stewart and Ji-guang Sun. Matrix perturbation theory. *(No Title)*, 1990.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Anastasios Tsiamis, Ingvar M Ziemann, Manfred Morari, Nikolai Matni, and George J Pappas. Learning to control linear systems can be hard. In *Conference on Learning Theory*, pages 3820–3857. PMLR, 2022.

Han Wang, Leonardo F Toso, Aritra Mitra, and James Anderson. Model-free learning with heterogeneous dynamical systems: A federated lqr approach. *arXiv preprint arXiv:2308.11743*, 2023.

Thomas T Zhang, Katie Kang, Bruce D Lee, Claire Tomlin, Sergey Levine, Stephen Tu, and Nikolai Matni. Multi-task imitation learning for linear dynamical systems. In *Learning for Dynamics and Control Conference*, pages 586–599. PMLR, 2023a.

Thomas TCK Zhang, Leonardo Felipe Toso, James Anderson, and Nikolai Matni. Sample-efficient linear representation learning from non-iid non-isotropic data. In *The Twelfth International Conference on Learning Representations*, 2023b.

Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *arXiv preprint arXiv:2201.01680*, 2022.