2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

2024 IEEE 63rd Conference on Decision and Control (CDC) | 979-8-3503-1633-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/CDC56724.2024.10885804

Active Learning for Control-Oriented Identification of Nonlinear Systems

Bruce D. Lee, Ingvar Ziemann, George J. Pappas, Nikolai Matni

***Abstract*— Model-based reinforcement learning is an effective approach for controlling an unknown system. It is based on a longstanding pipeline familiar to the control community in which one performs experiments on the environment to collect a dataset, uses the resulting dataset to identify a model of the system, and finally performs control synthesis using the identified model. As interacting with the system may be costly and time consuming, targeted exploration is crucial for developing an effective control-oriented model with minimal experimentation. Motivated by this challenge, recent work has begun to study finite sample data requirements and sample efficient algorithms for the problem of optimal exploration in model-based reinforcement learning. However, existing theory and algorithms are limited to model classes which are linear in the parameters. Our work instead focuses on models with nonlinear parameter dependencies, and presents the first finite sample analysis of an active learning algorithm suitable for a general class of nonlinear dynamics. In certain settings, the excess control cost of our algorithm achieves the optimal rate, up to logarithmic factors. We validate our approach in simulation, showcasing the advantage of active, control-oriented exploration for controlling nonlinear systems.**

I. Introduction

In recent years, model-based reinforcement learning has been successfully applied to various application domains including robotics, healthcare, and autonomous driving [1, 2]. These approaches often proceed by performing experiments on a system to collect data, and then using the data to fit models for the dynamics. In the specified application domains, performing experiments requires interaction with the physical world, which can be both costly and time-consuming. It is therefore important to design the experimentation and identification procedures to efficiently extract the most information relevant to control. In particular, experiments must be designed with the downstream control objective in mind. This fact is well-established in classical controls and identification literature [3–6]. While these works provide some guidance for experiment design, they mostly focus on linear systems, and supply only asymptotic guarantees.

Driven by the empirical success of machine and deep learning in solving classes of complex control problems, the learning and control communities have recently begun revisiting the classical pipeline of identification to control, proposing new algorithms, and analyzing them from a non-asymptotic viewpoint. Early efforts focused on end-to-end control guarantees for unknown linear system under naive exploration (injecting white noise inputs) [7, 8]. These methods have also been refined by using active learning to collect better data for control synthesis [9]. This approach has been extended to nonlinear systems with a linear dependence on the unknown parameters [10]. Other works studying model-based control of nonlinear systems also assume linear dependence on the unknown parameters, or consider related simplifying assumptions in settings including tabular or low-rank Markov Decision Processes [11, 12]. Model-based reinforcement learning for a general class of nonlinear systems has also been considered [13]. However, their guarantees focus on the worst case uncertainty of any control policy rather than end-to-end control costs for a particular objective.

There is a significant gap in that there are no algorithms with strong guarantees (achieving the optimal rates) for model-based reinforcement learning of general nonlinear dynamical systems. We leverage recently developed machinery for non-asymptotic analysis of nonlinear system identification to tackle this problem [14].

A. Contribution

We introduce and analyze the Active Learning for Control-Oriented Identification (`ALCOI`) algorithm. This algorithm extends an approach for model-based reinforcement learning proposed by Wagenmaker et al. [10] for dynamical systems with a linear dependence on the unknown parameter to general nonlinear dynamics that satisfy some smoothness assumptions. The algorithm is inspired by a reduction of the excess control cost to the system identification error, which may then be controlled using novel finite sample system identification error bounds for smooth nonlinear systems.

Leveraging the aforementioned reduction of the excess control cost and system identification error bounds, we derive finite sample bounds for the excess cost of our algorithm.

Theorem I.1 (Main Result, Informal)**.** *Let the* `ALCOI` *algorithm interact with an unknown nonlinear dynamical system for some number of exploration rounds before proposing a control policy designed to optimize some objective. The excess cost of the proposed policy on the objective satisfies*

$$\text{excess cost} \leq \frac{\text{hardness of control} \times \text{hardness of identification}}{\#\,\text{exploration rounds}}.$$

The "hardness of control" captures how the error in estimation of the dynamics translates to error in control, while the "hardness of identification" captures how challenging it is to identify the parameters under the best possible exploration policy. Moreover, our analysis reveals how the respective hardness quantities interact. Wagenmaker et al. [10] provide upper and lower bounds for this problem in a setting where

The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania. Emails: {`brucele, ingvarz, pappasg, nmatni`}`@seas.upenn.edu`.

the dynamics model is linear in the unknown parameters. Our upper bound is tight up to logarithmic factors in this setting, and we conjecture that it is also tight up to logarithmic factors in the setting where the model is nonlinear in the parameters.

The non-asymptotic system identification result may be of independent interest. It derives from invoking recently developed machinery for the analysis of nonlinear system identification along with the delta method, a classical approach from statistics. These bounds provide rates that match the asymptotic limit up to logarithmic factors.

B. Related Work

a) Additional Work Analyzing Identification & Control: Finite sample guarantees for active exploration of pure system identification have been studied in linear [15], and nonlinear (with linear dependence on the unknown parameters) settings [16]. Lower bounds complementing the upper bounds for the end-to-end control are also present [9, 10], and have been specialized to the linear-quadratic regulator setting to characterize systems which are hard to learn to control [17]. Recent literature considers gradient-based approaches for experiment design in linear-quadratic control [18]. For more details on finite sample analysis of learning to control, see the survey by Tsiamis et al. [19]. The aforementioned results do not focus on general nonlinear systems. Such analysis exists for identification; however, in the absence of end-to-end control error bounds [14, 20]. In contrast, we achieve end-to-end control error bounds for active learning applied for learning to control general nonlinear systems.

b) Dual Control: A related paradigm to the "identify then control" scheme studied in this work is that of *dual control*, in which the learner must interact with an unknown system while simultaneously optimizing a control objective [21]. Åström and Wittenmark [22] study a version of this problem known as the self-tuning regulator, providing asymptotic guarantees of convergence. Non-asymptotic guarantees for the self tuning regulator have been studied more recently from the online learning perspective of regret [23]. Subsequent work provides matching upper and lower bounds for the regret of the self-tuning regulator problem [24]. Lower bounds refining the dependence on system-theoretic constants have also been established [25]. The regret of learning to control nonlinear dynamical systems (with linear dependence on the unknown parameter) has also been studied [26, 27]. As in the "identify then control" setting, prior work in dual control has not provided finite sample analysis of the end-to-end control error for systems with nonlinear dependence on the unknown parameters.

Notation: Expectation (respectively probability) with respect to all the randomness of the underlying probability space is denoted by $\mathbb{E}$ (respectively $\mathbb{P}$). 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 $\|A\|_F$. A symmetric, positive semi-definite matrix $A = A^\top$ is denoted $A \succeq 0$. $A \succeq B$ denotes that $A - B$ is positive semi-definite. Similarly, a symmetric, positive definite matrix A is denoted $A \succ 0$. The minimum eigenvalue of a symmetric, positive semi-definite matrix A is denoted $\lambda_{\min}(A)$. For a positive definite matrix A, we define the A-norm as $\|x\|_A^2 = x^\top A x$. The gradient of a scalar valued function $f : \mathbb{R}^n \to \mathbb{R}$ is denoted ∇f, and the Hessian is denoted $\nabla^2 f$. The Jacobian of a vector-valued function $g : \mathbb{R}^n \to \mathbb{R}^m$ is denoted Dg, and follows the convention for any $x \in \mathbb{R}^n$, the rows of $Dg(x)$ are the transposed gradients of $g_i(x)$. The p^{th} order derivative of g is denoted by $D^p g$. Note that for $p \geq 2$, $D^p g(x)$ is a tensor for any $x \in \mathbb{R}^n$. The operator norm of such a tensor is denoted by $\|D^p g(x)\|_{\mathsf{op}}$. For a function $f : \mathsf{X} \to \mathbb{R}^{d_\mathsf{Y}}$, we define $\|f\|_\infty \triangleq \sup_{x \in \mathsf{X}} \|f(x)\|$. A Euclidean norm ball of radius r centered at x is denoted $\mathcal{B}(x, r)$.

II. Problem Formulation

We consider a nonlinear dynamical system evolving according to

$$X_{t+1} = f(X_t, U_t; \phi^\star) + W_t, \quad t = 1, \dots T, \tag{1}$$

with state X_t assuming values in $\mathbb{R}^{d_\mathsf{X}}$, input U_t assuming values in $\mathbb{R}^{d_\mathsf{U}}$, and d_X-dimensional noise $W_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2 I)$ for some $\sigma_w > 0$. For simplicity, we assume $X_1 = 0$. Here, f is the dynamics function, which depends on an unknown parameter $\phi^\star \in \mathbb{R}^{d_\phi}$. We assume that there exists some positive B such that $\|\phi^\star\| \leq B - 1$ and $\|f(\cdot, \cdot, \phi)\|_\infty \leq B$ for all $\phi \in \mathbb{R}^{d_\phi}$ satisfying $\|\phi\| \leq B$.

We study a learner whose objective is to determine a policy $\pi = \{\pi_t\}_{t=1}^T$ from a policy class $\Pi^\star$ to minimize the cost $\mathcal{J}(\pi, \phi^\star)$, where

$$\mathcal{J}(\pi, \phi) = \mathbb{E}_\pi^\phi \left[\sum_{t=1}^T c_t(X_t, U_t) + c_{T+1}(X_{T+1}) \right]. \tag{2}$$

The functions c_t are known stage costs. The superscript on the expectation denotes that the dynamics (1) are rolled out under parameter ϕ, while the subscript denotes that the system is played in closed-loop under the feedback control policy $U_t = \pi_t(X_1, U_1, \dots, X_{t-1}, U_{t-1}, X_t)$ for $t = 1, \dots, T$. The learner follows a two step interaction protocol with an exploration phase, and an evaluation phase. In the exploration phase, the learner interacts with the system for a total of N episodes, each consisting of T timesteps, by playing exploration policies $\pi \in \Pi_{\mathrm{exp}}$. The policy class Π_{exp} is an exploration policy class, described in more detail below. The learner does not incur any cost during the exploration episodes, and seeks only to gain information about the system. After the N interaction episodes, it uses the collected data to propose a policy $\hat{\pi} \in \Pi^\star$. The learner is then evaluated on the expected cost of the proposed policy on a new evaluation episode. In particular, it incurs cost $\mathcal{J}(\hat{\pi}, \phi^\star)$.

The policy classes $\Pi^\star$ and Π_{exp} are known; $\Pi^\star$ consists of deterministic policies, but Π_{exp} may be random. We do not assume $\Pi^\star = \Pi_{\mathsf{exp}}$. We assume that the policy class $\Pi^\star$ has the parametric form:

$$\Pi^\star = \left\{ \pi^\theta | \theta \in \mathbb{R}^{d_\theta} \right\}.$$

Fig. 1. Identfication to control pipeline.

No such parametric assumption is made on the exploration class Π_{exp}. It instead consists of whatever experimental procedures are available. For instance, it may be the class of policies with power or energy bounded inputs. Policies belonging to Π_{exp} must be history dependent (causal). We assume that the learner is allowed to randomly select choices of policies in Π_{exp}.[1] Given these policy classes, the learning procedure should seek to identify the best exploitation policy belonging to $\Pi^\star$ by playing the most informative exploration policy in the class Π_{exp}.

A. Certainty Equivalent Control

We focus on learners which follow a model-based approach to synthesize a control policy from interaction with the system, outlined in Figure 1. In this section, we discuss the learner's procedure for the last two steps: system identification and control synthesis. In Section III, we return to the question of which experiments the learner should perform.

Given the data collected during the experimentation phase, the learner finds an estimate for the dynamics by solving a nonlinear least squares problem. In particular, suppose that during the N experimentation episodes of length T, the learner collects data $\left\{(U_t^n, X_t^n, X_{t+1}^n)\right\}_{n,t=1}^{N,T+1}$. The subscript denotes the time index within each episode, while the superscript denotes the episode index. Using this dataset, the learner may identify the dynamics of the system by solving

$$\hat{\phi} \in \operatorname*{argmin}_{\phi \in \mathbb{R}^{d_\phi}, \|\phi\| \le B} \sum_{n=1}^{N} \sum_{t=1}^{T} \left\| X_{t+1}^n - f(X_t^n, U_t^n; \phi) \right\|^2. \quad (3)$$

Solving this problem provides a parameter estimate which is an effective predictor under the distribution of states and inputs seen during the experimentation. This notion can be captured via the prediction error.

Definition II.1. *We define* $\mathbf{Err}_\pi^{\phi^\star}(\phi)$ *as the prediction error for a parameter* ϕ *under policy* π*:*

$$\mathbf{Err}_\pi^{\phi^\star}(\phi) = \mathbb{E}_\pi^{\phi^\star}\left[\frac{1}{T}\sum_{t=1}^{T} \left\| f(X_t, U_t; \phi) - f(X_t, U_t; \phi^\star) \right\|^2\right].$$

Once the learner estimates $\hat{\phi}$, the controller parameters are determined from the dynamics parameters by solving the policy optimization problem as

$$\theta^\star(\hat{\phi}) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_\theta}} J(\pi^\theta, \hat{\phi}). \quad (4)$$

The certainty equivalent policy may then be expressed as a function of the estimated dynamics parameters $\hat{\phi}$ as

$$\pi^\star(\hat{\phi}) \triangleq \pi^{\theta^\star(\hat{\phi})}. \quad (5)$$

[1] i.e. for any $\pi^1, \pi^2 \in \Pi_{\mathrm{exp}}$ and any $b \in [0,1]$, the policy π_{mix} which at the start of a new episode plays π^1 for the duration of the episode with probability b and π^2 for the duration of the episode with probability $1-b$ also belongs to Π_{exp}.

We note that both the nonlinear least squares problem (3) and the certainty equivalent control synthesis procedure of (5) may be computationally challenging. The focus of this work is to understand the statistical complexity of the problem rather than the computational complexity. In the episodic setting we consider, both of these problems are solved offline. Therefore, given sufficient time and compute, it is often possible to determine good approximations to the optimal solutions using non-convex optimization solvers and approaches for policy optimization from the model-based reinforcement learning literature [1].

B. Assumptions

By (5), the optimal policy for the objective (2) under the true parameter $\phi^\star$ defining the dynamics (1) is thus given by $\pi^\star(\phi^\star)$, and the corresponding objective value is $\mathcal{J}(\pi^\star(\phi^\star), \phi^\star)$. Meanwhile, the objective value attained under an estimate $\hat{\phi}$ is $\mathcal{J}(\pi^\star(\hat{\phi}), \phi^\star)$. We abuse notation and define the shorthand

$$\mathcal{J}_{\tilde{\phi}}(\phi) \triangleq \mathcal{J}(\pi^\star(\phi), \tilde{\phi}) \quad (6)$$

to describe the control cost of applying a certainty equivalence policy synthesized using parameter ϕ on a system with dynamics described by $\tilde{\phi}$. It has been shown by Wagenmaker et al. [9, 10] that for models which are linear in the parameters, the gap $\mathcal{J}_{\phi^\star}(\phi) - \mathcal{J}_{\phi^\star}(\phi^\star)$ is characterized by the squared parameter error weighted by the *model-task Hessian*, defined below.

Definition II.2. *The model-task Hessian for objective* (2) *and dynamics* (1) *is given by*

$$\mathcal{H}(\tilde{\phi}) = \nabla_\phi^2 \mathcal{J}_{\tilde{\phi}}(\phi)|_{\phi=\tilde{\phi}},$$

where $\mathcal{J}_{\tilde{\phi}}$ *is defined in* (6)*.*

To express the excess cost achieved by a certainty equivalent controller synthesized using the estimated model parameters $\hat{\phi}$, we operate under the following smoothness assumption on the dynamics.

Assumption II.1. *(Smooth Dynamics) The dynamics are four times differentiable with respect to* u *and* ϕ*. Furthermore, for all* $(x,u) \in \mathbb{R}^{d_{\mathsf{x}}} \times \mathbb{R}^{d_{\mathsf{u}}}$*, all* $\phi \in \mathbb{R}^{d_\phi}$*, and* $i,j \in \{0,1,2,3\}$ *such that* $1 \le i+j \le 4$*, the derivatives of* f *satisfy*

$$\left\| D_\phi^{(i)} D_u^{(j)} f(x,u;\phi) \right\|_{\mathsf{op}} \le L_f.$$

The above assumption is satisfied for, e.g., control-affine dynamics which depend smoothly on ϕ: $f(x_t, u_t; \phi) = g_1(x_t;\phi) + g_2(u_t;\phi)u$, with g_1 and g_2 each three time differentiable with respect to ϕ. In this example, differentiability with respect to u is immediate from the affine dependence.

We also require that the policy class $\Pi^\star$ is smooth.

Assumption II.2. *(Smooth Policy Class) For* $t = 1, \dots, T$*,* $x \in \mathcal{X}$*, and any policy* $\pi \in \Pi^\star$*, the function* $\pi_t^\theta(x)$ *is four-times differentiable in* θ*. Furthermore,* $\left\| D_\theta^{(i)} \pi_t^\theta(x) \right\|_{\mathsf{op}} \le L_\theta$ *for* $i = 1, \dots, 4$*,* $\theta \in \mathbb{R}^{d_\theta}$*,* $x \in \mathcal{X}$*, and* $t = 1, \dots, T$*.*

Note that such smoothness conditions are not imposed for the exploration policy class Π_{exp}. The exploration policy class could, for instance, consist of model predictive controllers with constraints on the injected input energy, which do not satisfy such smoothness assumptions.

We additionally require that the costs are bounded for policies in the class $\Pi_\star$ and all dynamics parameters in a neighborhood of the true parameter.

Assumption II.3. *(Regular costs) The stage costs c_t are three times differentiable and $\left\|D_u^{(i)} c_t(x,u)\right\|_{\text{op}} \leq L_{\mathsf{cost}}$ for $i = 1,\dots,3$, $\theta \in \mathbb{R}^{d_\theta}$, $(x,u) \in \mathbb{R}^{d_{\mathsf{X}}+d_{\mathsf{U}}}$. There exists some $r_{\mathsf{cost}}(\phi^\star) > 0$ such that for all $\phi \in \mathcal{B}(\phi^\star, r_{\mathsf{cost}}(\phi^\star))$, and all $\pi \in \Pi^\star$, we have $\mathbb{E}_\pi^\phi\left[\left(\sum_{t=1}^T c_t(X_t,U_t) + C_{T+1}(X_T)\right)^2\right] \leq L_{\mathsf{cost}}$.*

We additionally assume that the certainty equivalent policy is a smooth function of the dynamics parameter in a neighborhood around the optimal parameter.

Assumption II.4. *There exists some $r_\theta(\phi^\star) > 0$ such that for all $\phi \in \mathcal{B}(\phi^\star, r_\theta(\phi^\star))$,*

- $\nabla_\theta \mathcal{J}(\pi^\theta, \phi)|_{\theta=\theta^\star(\phi)} = 0$
- *$\theta^\star(\phi)$ is three times differentiable in ϕ and $\left\|D_\phi^i \theta^\star(\phi)\right\|_{\text{op}} \leq L_{\pi^\star}$ for some $L_{\pi^\star} > 0$ and $i \in \{1,2,3\}$.*

It is shown in Proposition 6 of Wagenmaker et al. [10] that the above condition holds if $\nabla_\theta^2 J(\pi^\theta, \phi^\star) \succ 0$.[2] The above assumption also holds in the setting of linear-quadratic regulation, as may be verified using the LQR derivative expressions in Simchowitz and Foster [24].

In order to bound the parameter recovery error in terms of the prediction error, additional identifiability conditions are needed. The following definition of a Lojasiewicz exploration policy is determined from a Lojasiewicz condition that arises in the optimization literature that measures the sharpness of an objective near its optimizer [29]. In our setting, it quantifies the degree of identifiability from using a particular exploration policy. It does so by bounding the growth of identification error as a polynomial of prediction error.

Definition II.3 (Lojasiewicz condition, Roulet and d'Aspremont [29])**.** *For positive numbers C_{Loja} and α, we say that a policy $\pi \in \Pi_{\text{exp}}$ is $(C_{\mathsf{Loja}}, \alpha)$-Lojasiewicz if*

$$\left\|\hat{\phi} - \phi^\star\right\| \leq C_{\mathsf{Loja}} \mathbf{Err}_\pi^{\phi^\star}(\hat{\phi})^\alpha.$$

To ensure parameter recovery is possible for the learner, we make the following assumption regarding identifiability.

Assumption II.5. *Fix some positive constant C_{Loja} and $\alpha \in (\frac{1}{4}, \frac{1}{2}]$. The learner has access to a policy $\pi^0 \in \Pi_{\text{exp}}$ which is $(C_{\mathsf{Loja}}, \alpha)$-Lojasiewicz.*

While the Lojasiewicz assumption ensures that the data collected via the exploration policy π^0 is sufficient to identify the parameters, the rate of recovery may be slow. To bypass this limitation, we assume that some policy in the exploration class satisfies a persistence of excitation condition. This condition can be expressed by first defining the Fisher information matrix for a parameter ϕ and a policy π as

$$\mathbf{FI}^\pi(\phi) \triangleq \frac{\mathbb{E}_\pi^\phi\left[\sum_{t=1}^T Df(X_t,U_t,\phi)^\top Df(X_t,U_t,\phi)\right]}{\sigma_w^2}, \quad (7)$$

where $Df(X_t,U_t,\phi)$ is the Jacobian of f with respect to ϕ. The Fisher information measures the signal-to-noise ratio of the data collected from an episode of interaction with the system under exploration policy π. With this definition, persistance of excitation is equivalent to the positive definiteness of the matrix $\mathbf{FI}^\pi(\phi^\star)$.

Assumption II.6. *There exists a policy $\pi \in \Pi_{\text{exp}}$ for which*

$$\mathbf{FI}^\pi(\phi^\star) \succeq \mu I \succ 0.$$

III. Proposed Algorithm and Main Result

The above smoothness assumptions allow us to characterize the excess control cost of a policy synthesized via certainty equivalence applied to a parameter estimate $\hat{\phi}$, (5). In particular, we extend a result from Wagenmaker et al. [10] from the linear in parameters setting to the smooth nonlinear setting.

Lemma III.1 (Thm. 1 of Wagenmaker et al. [10])**.** *Suppose Assumptions II.1-II.4 hold. Let $r_\theta(\phi^\star)$ be as defined in Assumption II.4. Then for $\hat{\phi} \in \mathcal{B}(\phi^\star, \min\{r_{\mathsf{cost}}(\phi^\star), r_\theta(\phi^\star)\})$,*

$$\mathcal{J}_{\phi^\star}(\hat{\phi}) - \mathcal{J}_{\phi^\star}(\phi^\star) \leq \left\|\hat{\phi} - \phi^\star\right\|_{\mathcal{H}(\phi^\star)}^2 + C_{\mathsf{cost}} \left\|\hat{\phi} - \phi^\star\right\|^3, \quad (8)$$

where $\mathcal{J}_{\phi^\star}(\phi)$ is as defined in (6) and

$$C_{\mathsf{cost}} = \mathsf{poly}(L_{\pi^\star}, L_f, L_\theta, L_{\mathsf{cost}}, \sigma_w^{-1}, T, d_{\mathsf{X}}).$$

Lemma III.1 informs us that the leading term of the excess cost is given by the parameter estimation error weighted by the model-task Hessian, $\left\|\hat{\phi} - \phi^\star\right\|_{\mathcal{H}(\phi^\star)}^2$.

Asymptotically, the distribution of the parameter estimation error is normally distributed with mean zero, and covariance given by the inverse Fisher information matrix under the data collection policy π evaluated at the true parameter value (cf. Theorem 1 of Ljung and Caines [30]):

$$\lim_{N\to\infty} \sqrt{N}\mathbf{FI}^\pi(\phi^\star)^{1/2}(\hat{\phi} - \phi_\star) \sim \mathcal{N}(0, I).$$

We provide a novel non-asymptotic result which characterizes the H-norm of the parameter error for a positive definite matrix H in terms of the Fisher information matrix.

Theorem III.1. *Suppose Assumption II.1 holds. Consider the least squares estimate $\hat{\phi}$ determined from (3) using data collected from N episodes via an exploration policy π which is $(C_{\mathsf{Loja}}, \alpha)$-Lojasiewicz for some $\alpha \in (\frac{1}{4}, \frac{1}{2}]$, and satisfies $\lambda_{\min}(\mathbf{FI}^\pi(\phi^\star)) > 0$, with $\mathbf{FI}^\pi(\phi^\star)$ as defined in (7). Let H be a positive definite matrix, β a positive number satisyfing $\beta \leq \sigma_w^2 \frac{\lambda_{\min}(\mathbf{FI}^\pi(\phi^\star))}{4}$, and $\delta \in (0, \frac{1}{4}]$. Then there exists a*

[2] Wagenmaker et al. [10] show this result for the linear in the parameters setting; however, it extends easily to the smooth nonlinear setting. See Appendix C of [28].

polynomial poly_α *depending on* α *such that the following condition holds. With probability at least* $1-\delta$,

$$\left\|\hat{\phi}-\phi^\star\right\|_H^2 \le 2(1+\xi) \times \left(\frac{\mathrm{tr}(H\mathbf{FI}^\pi(\phi^\star)^{-1})}{N} + 2\frac{\left\|H\mathbf{FI}^\pi(\phi^\star)^{-1}\right\|}{N}\log\frac{4}{\delta}\right), \tag{9}$$

where $\xi = 4\beta\left(\frac{1}{\sigma_w^2\lambda_{\min}(\mathbf{FI}^\pi(\phi^\star))}+d_\phi\right)$ *as long as*

$$N \ge \mathsf{poly}_\alpha\left(T, L_f, d_\phi, d_{\mathsf{X}}, \sigma_w, \log N, \log\frac{1}{\delta}, \log\frac{B}{\sigma_w}, C_{\mathsf{Loja}}, \frac{1}{\beta}\right).$$

By substituting the inequality (9) into the leading term of (8) and bounding $\left\|\mathcal{H}(\phi^\star)\mathbf{FI}^\pi(\phi^\star)^{-1}\right\| \le \mathrm{tr}(\mathcal{H}(\phi^\star)\mathbf{FI}^\pi(\phi^\star)^{-1})$, one would expect that the the excess cost of deploying the certainty equivalent policy synthesized on a least squares estimate determined from data collected using policy π is characterized by

$$\mathrm{tr}(\mathcal{H}(\phi^\star)\mathbf{FI}^\pi(\phi^\star)^{-1}). \tag{10}$$

In light of this, we would like to choose the exploration policy π which minimizes this upper bound:

$$\pi = \operatorname*{argmin}_{\tilde{\pi}\in\Pi_{\mathrm{exp}}} \mathrm{tr}(\mathcal{H}(\phi^\star)\mathbf{FI}^{\tilde{\pi}}(\phi^\star)^{-1}). \tag{11}$$

Our main result shows that the above intuition can be made rigorous through careful analysis and design of the exploration policy. It then proceeds to show that we can find an exploration policy approximately solving (11), even though the parameter $\phi^\star$ defining the exploration objective is unknown prior to experimentation. It is also necessary to address the fact that the model-task Hessian may not be positive definite. Thus optimizing the above objective could result in exploration policies which are not persistently exciting.

To circumvent the issue of the unknown parameter $\phi^\star$, we consider a two step approach in which we first obtain a crude parameter estimate, and then refine it by playing a targeted exploration policy. Denote the crude estimate by $\hat{\phi}^-$. This parameter can be used to search for a policy that approximately solves the optimization problem in (11). A straightforward approach to do so is to solve the problem under the estimated parameter:

$$\pi = \operatorname*{argmin}_{\tilde{\pi}\in\Pi_{\mathrm{exp}}} \mathrm{tr}(H(\hat{\phi}^-)\mathbf{FI}^{\tilde{\pi}}(\hat{\phi}^-)^{-1}). \tag{12}$$

To address the issue of a model-task Hessian which is not positive definite, we introduce regularization into the exploration design. In particular, we set the exploration policy π as

$$\pi = \operatorname*{argmin}_{\tilde{\pi}\in\Pi_{\mathrm{exp}}} \mathrm{tr}((H(\hat{\phi}^-)+\nu I)\mathbf{FI}^{\tilde{\pi}}(\hat{\phi}^-)^{-1}), \tag{13}$$

for an appropriately chosen regularization parameter ν.

The above discussion motivates Algorithm 1, named Active Learning for Control-Oriented Identification (ALCOI). The algorithm takes as input an initial policy satisfying the Lojasiewicz condition (Assumption II.5), the exploration policy class, the target policy, the number of exploration rounds, a parameter $\gamma \in (0,1)$ which controls the ratio of the exploration budget that the initial loja policy is played, the level of regularization ν, and the precision of the optimization for the exploration policy ε. Given these components, the algorithm proceeds in three stages. The first stage begins in Line 2 by playing the initial policy for a portion of the exploration budget controlled by γ. In Line 3, it uses the collected data to derive a coarse estimate $\hat{\phi}^-$ for the unknown parameters by solving a least squares problem. Next, the estimate $\hat{\phi}^-$ is used to construct the model-task Hessian as $\hat{\phi}^-$ as $\mathcal{H}(\hat{\phi}^-)$ and define an exploration objective $\mathrm{tr}\left(\left(\mathcal{H}(\hat{\phi}^-)+\lambda I\right)\mathbf{FI}^{\tilde{\pi}}(\hat{\phi}^-)^{-1}\right)$. Optimizing this objective to precision ε over the class of exploration policies provides the policy π_{exp}. This policy is run to collect data from the system, and obtain a fresh estimate $\hat{\phi}^+$ for $\phi^\star$. Finally, the estimate is used to synthesize the certainty equivalent policy as in (5).

Algorithm 1 $\mathtt{ALCOI}(\pi^0, \Pi_{\mathrm{exp}}, \Pi^\star, N, \gamma, \nu, \varepsilon)$

1: **Input:** Initial policy π^0, exploration policy class Π_{exp}, target policy class $\Pi^\star$, initial policy ratio γ, regularization parameter ν, optimization precision ε.
2: Play π^0 for $\lfloor \gamma N \rfloor$ episodes to collect $\left\{X_t^n, U_t^n, X_{t+1}^n\right\}_{t,k=1}^{T,\lfloor\gamma N\rfloor}$.
3: Fit $\hat{\phi}^-$ from the collected data by solving (3).
4: Determine exploration policy as

$$\pi_{\mathrm{exp}} \in \left\{\pi\in\Pi_{\mathrm{exp}} \middle| \mathrm{tr}\left(\mathcal{H}(\hat{\phi}^-)\mathbf{FI}^\pi(\hat{\phi}^-)^{-1}\right) \le (1+\varepsilon)\inf_{\tilde{\pi}} \mathrm{tr}\left(\mathcal{H}(\hat{\phi}^-)\mathbf{FI}^{\tilde{\pi}}(\hat{\phi}^-)^{-1}\right)\right\}.$$

.
5: Define π_{mix} which at the start of each episode plays π^0 with probability γ, and π_{exp} with probability $1-\gamma$.
6: Play π_{mix} for $\lfloor(1-\gamma)N\rfloor$ episodes, collecting data $\left\{X_t^n, U_t^n, X_{t+1}^n\right\}_{t,k=1}^{T,\lfloor(1-\gamma)N\rfloor}$.
7: Fit $\hat{\phi}^+$ by solving (3) with the data $\left\{X_t^n, U_t^n, X_{t+1}^n\right\}_{t,k=1}^{T,\lfloor(1-\gamma)N\rfloor}$.
8: **Return:** certainty equivalent policy $\hat{\pi} = \pi^\star(\hat{\phi}^+)$.

Our main result is a finite sample bound characterizing the excess cost of the policy return by Algorithm 1.

Theorem III.2 (Main Result)**.** *Suppose* f, π^0, Π_{exp}, $\Pi^\star$ *satisfy Assumptions II.1-II.6. Let* ν *be a non-negative regularization parameter such that* $\lambda_{\min}(\mathcal{H}(\phi^\star))+\nu > 0$. *Let the optimization tolerance* $\varepsilon \in (0, 1/2)$ *and the initial policy ratio* $\gamma \in (0, 1/2)$. *Consider running Algorithm 1 to generate a control policy* $\hat{\pi}$ *as* $\hat{\pi} = \mathtt{ALCOI}(\pi^0, \Pi_{\mathrm{exp}}, \Pi^\star, N, \gamma, \nu, \varepsilon)$.

Let $\delta \in \left(0, \frac{1}{4}\right]$ *be the failure probability, and* $\beta \in \left(0, \frac{\mu(\lambda_{\min}(H(\phi^\star))+\nu)}{512 d_\phi(\|H(\phi^\star)\|+\nu)}\right)$[3] *be a free parameter in the bound.*

[3]Recall that μ is the persistance of excitation parameter defined in Assumption II.6.

There exists a polynomial function $\texttt{poly}_\alpha$ *depending on the Lojasiewicz parameter* α *such that the following holds true. With probability at least* $1-\delta$*, it holds that*

$$\mathcal{J}(\hat{\pi},\phi^\star)-\mathcal{J}_{\phi^\star}(\phi^\star)\leq(1+4\gamma)(1+\varepsilon)(1+\xi)\left(2+4\log\frac{4}{\delta}\right)\times\frac{\inf_{\tilde{\pi}\in\Pi_{\mathrm{exp}}}\mathrm{tr}\left((\mathcal{H}(\phi^\star)+\nu I)\,\mathbf{FI}^{\tilde{\pi}}(\phi^\star)^{-1}\right)}{N},$$

where $\xi=\beta\left(3+16\left(\frac{128d_\phi(\|\mathcal{H}(\phi^\star)\|+\nu)}{\mu(\lambda_{\min}(\mathcal{H}(\phi^\star))+\nu)}+d_\phi\right)\right)$ *as long as*

$$N\geq\mathsf{poly}_\alpha\Bigg(T,L_f,L_{\mathsf{cost}},L_{\pi^\star},d_\phi,d_{\mathsf{X}},\frac{1}{\mu},r_{\mathsf{cost}}(\phi^\star),r_\theta(\phi^\star)\frac{1}{\lambda_{\min}(\mathcal{H}(\phi^\star))+\nu},\|\mathcal{H}(\phi^\star)\|,\nu,\sigma_w,\sigma_w^{-1},\log N,\log\frac{1}{\delta},\log B,C_{\mathsf{Loja}},\frac{1}{\gamma},\frac{1}{\beta}\Bigg).$$

The above result characterizes the excess control cost in terms of three key quantities. First, the term $\mathcal{H}(\phi^\star)$ is the model-task Hessian, which describes how error in identification of the dynamics model parameters translates to the control cost. Second is the inverse Fisher information of the optimal exploration policy term, which measures a signal-to-noise ratio quantifying the hardness of parameter identification. Finally, the number of exploration episodes N on the denominator captures the rate of decay from increasing the experimental budget.

In the setting where the dynamics model has linear dependence on the parameters, Wagenmaker et al. [10] present a lower bound on the excess control cost achieved by any learner following the model based interaction protocol described in Section II. If we choose the free parameter in the upper bound as $\beta\leq\left(\frac{128d_\phi(\|\mathcal{H}(\phi^\star)\|+\nu)}{\mu(\lambda_{\min}(\mathcal{H}(\phi^\star))+\nu)}+d_\phi\right)^{-1}$, then the upper bound of Theorem III.2 matches this lower bound up to universal constants, and the term $\log(4/\delta)$. Future work will pursue general lower bounds that hold for dynamics models with a nonlinear dependence on the unknown parameter.

The burn-in time is currently polynomial in the relevant system parameters; however, we do not pursue tight burn-in times in this work. It may be possible to improve the dependence of the burn-in on various system quantities, e.g. by leveraging stability or reachability to obtain optimal dependence of the burn-in on T. We additionally draw attention to the utility of the parameter β and the algorithm hyperparameters ν and γ for navigating the tradeoff between a good burn-in time, and optimal rates. One can take β, ν, γ and ε arbitrarily close to zero, meaning that the coefficient characterizing the excess cost can become arbitrarily close to $2+4\log\frac{8}{\delta}$. The cost of doing so is an increase in the burn-in time. A notable exception is the situation where $\mathcal{H}(\phi^\star)\succ 0$, i.e. the setting where all parameters are necessary for control. In this case, one can take $\nu=0$, while the burn-in time must exceed a polynomial in $\frac{1}{\lambda_{\min}(\mathcal{H}(\phi^\star))}$.

IV. Proof Sketch

Full proof details may be found in the appendix of Lee et al. [28]. Here, we present a sketch. Our main result proceeds by demonstrating the following sub-steps. In these sub-steps, let C be a polynomial of the problem parameters and log of the reciprocal of the failure probability, δ, as in the burn-in requirement of Theorem III.2.

1) With high probability, the coarse parameter estimation error decays gracefully with the total amount of data:
$$\left\|\hat{\phi}^- - \phi^\star\right\|\leq\frac{C}{(TN)^\alpha},\tag{14}$$
as long as N exceeds some polynomial burn-in time. This result is derived from recent results characterizing non-asymptotic bounds for identification [14], and takes the place of the estimator consistency requirements in classical asymptotic identification literature [30]. By making the number of episodes N sufficiently large, we can make this error arbitrarily small. It thus characterizes a type of "consistency burn-in".
2) As long as the coarse estimation error of (14) is sufficiently small, the ideal optimal exploration objective of (11) is well-approximated by the objective (12). In particular, for any exploration policy $\pi\in\Pi_{\mathrm{exp}}$,
$$\left|\mathrm{tr}\left(\mathcal{H}(\hat{\phi}^-)\mathbf{FI}^\pi(\hat{\phi}^-)^{-1}\right)-\mathrm{tr}(\mathcal{H}(\phi^\star)\mathbf{FI}^\pi(\phi^\star)^{-1})\right|\leq C\left\|\hat{\phi}^- - \phi^\star\right\|.\tag{15}$$
3) For N sufficiently large, we may use the consistency guarantee (14) to prove Theorem III.1. The proof of this fact follows by revisiting the delta method [31] through the lens of concentration inequalities. Doing so results in the near sharp[4] rates we obtain.

Using the above results, our argument proceeds according to the following series of inequalities applied to the excess cost. With high probability,

$$\begin{aligned}\mathcal{J}_{\phi^\star}(\hat{\phi}^+)-\mathcal{J}_{\phi^\star}(\phi^\star)&\leq\left\|\hat{\phi}^+-\phi^\star\right\|^2_{\mathcal{H}(\phi^\star)}+C_{\mathsf{cost}}\left\|\hat{\phi}^+-\phi^\star\right\|^3\\&\leq\xi(\delta)\frac{\mathrm{tr}((\mathcal{H}(\phi^\star)+\nu I)\,\mathbf{FI}^{\pi_{\mathsf{mix}}}(\phi^\star)^{-1})}{N}+\frac{C}{N^{3/2}},\end{aligned}\tag{16}$$

where the first inequality follows by Lemma III.1, and the second inequality follows by applying Theorem III.1 with $H=\mathcal{H}(\phi^\star)+\nu I$ for the first term, and $H=I$ for the second term. The quantity $\xi(\delta)$ is a trades off the burn-in time and the final bound. In our analysis, it can become as small as $2+4\log\frac{1}{\delta}$. Next, it follows that

$$\begin{aligned}&\mathrm{tr}((\mathcal{H}(\phi^\star)+\nu I)\,\mathbf{FI}^{\pi_{\mathsf{mix}}}(\phi^\star)^{-1})\\&\overset{(i)}{\leq}\frac{C}{N^\alpha}+\mathrm{tr}\left(\left(H(\hat{\phi}^-)+\nu I\right)\mathbf{FI}^{\pi_{\mathsf{mix}}}(\hat{\phi}^-)-1\right)\\&\overset{(ii)}{\leq}2\frac{C}{N^\alpha}+\frac{1}{1-\gamma}\inf_{\tilde{\pi}\in\Pi_{\mathrm{exp}}}\mathrm{tr}\left(\left(H(\hat{\phi}^-)+\nu I\right)\mathbf{FI}^{\tilde{\pi}}(\hat{\phi}^-)^{-1}\right)\\&\overset{(iii)}{\leq}3\frac{C}{N^\alpha}+\frac{1}{1-\gamma}\inf_{\tilde{\pi}\in\Pi_{\mathrm{exp}}}\mathrm{tr}\left((H(\phi^\star)+\nu I)\,\mathbf{FI}^{\tilde{\pi}}(\phi^\star)^{-1}\right),\end{aligned}$$

[4]Rates matching the asymptotic limit up to logarithmic factors.

where inequality (i) follows from (15) and (14), inequality (ii) follows from the definition of the policy π_{mix} and inequality (iii) follows from (15) and (14). The main result then follows by substituting the above bound into (16), and taking N to exceed a polynomial burn-in time so the higher order terms become negligible.

V. Numerical Validation

We deploy ALCOI on an illustrative example to illustrate the benefits of active control-oriented exploration. For more experiments, and further details, see Appendix C. Consider the two dimensional system

$$X_{t+1} = X_t + U_t + W_t + \sum_{i=1}^{4} \psi(X_t - \phi_\star^{(i)})$$

with X_t, U_t, W_t and $\phi_\star^{(i)}$ assuming values in $\mathbb{R}^2$. Here $\psi : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by $\psi(x) = 5\frac{x}{\|x\|}\exp(-x^2)$. The noise is distributed according to a standard normal distribution. The parameters $\phi_\star^{(1)}$, $\phi_\star^{(2)}$, $\phi_\star^{(3)}$, $\phi_\star^{(4)}$ are set as $\begin{bmatrix}5\\0\end{bmatrix}$, $\begin{bmatrix}-5\\0\end{bmatrix}$, $\begin{bmatrix}0\\5\end{bmatrix}$ and $\begin{bmatrix}0\\-5\end{bmatrix}$, respectively.

We consider model-based reinforcement learning with a horizon $T = 10$ and quadratic cost functions: for all $t \in [T]$,

$$c_t(x,u) = \left\| x - \begin{bmatrix}5.5\\0\end{bmatrix} \right\|^2, \; c_{T+1}(x) = \left\| x - \begin{bmatrix}5.5\\0\end{bmatrix} \right\|^2.$$

The policy class $\Pi^\star$ consists of feedback linearization controllers defined by parameters $\theta = (K, \hat{\phi}^{(1)}, \ldots, \hat{\phi}^{(4)})$, with $K \in \mathbb{R}^{2\times 2}$ and $\hat{\phi}^{(i)} \in \mathbb{R}^2$ for $i = 1, \ldots, 4$:

$$\pi^\theta(X_t) = K\left(X_t - \begin{bmatrix}-5.5\\0\end{bmatrix}\right) - \sum_{i=1}^{4} \psi(X_t - \hat{\phi}^{(i)}).$$

The exploration class Π_{exp} consists of policies with input energy bounded by T: $\sum_{t=1}^{T} \|U_t\|^2 \le T$.

We compare ALCOI with random exploration and approximate A-optimal experiment design. For random exploration, the learner injects isotropic Gaussian noise which is normalized such that $\sum_{t=1}^{T} \|U_t\|^2 = T$. For approximate A-optimal experiment design, the learner runs the ALCOI, but with the model-task Hessian estimate, $\mathcal{H}(\hat{\phi}^-)$, replaced by I.

Figure 2 illustrates that ALCOI achieves a lower excess control cost than the alternatives at all iterations. To understand why this is the case, note that in order to regulate the system to the position $X_t = \begin{bmatrix}5.5 & 0\end{bmatrix}^\top$, the parameter $\phi_\star^{(1)}$ must be identified accurately. However, due to the Gaussian kernel, accurately estimating $\phi_\star^{(1)}$ requires that the experiment data consists of trajectories where the state is near $\phi_\star^{(1)}$. Random exploration clearly fails to collect such trajectories. Approximate A-optimal experiment design does collect such trajectories; however, it also collects trajectories steering the state to $\begin{bmatrix}-5 & 0\end{bmatrix}^\top$, $\begin{bmatrix}0 & 5\end{bmatrix}^\top$, and $\begin{bmatrix}0 & -5\end{bmatrix}^\top$ in order to identify the parameters $\phi_\star^{(2)}, \phi_\star^{(3)}$ and $\phi_\star^{(3)}$. ALCOI, in contrast, designs experiments that are effective for identifying the parameters most relevant for control. For the chosen objective, this means that the algorithm invests the most exploration energy in collecting data in the neighborhood of $\phi_\star^{(1)}$. This illustrative example hints at the practical benefit of the proposed approach.

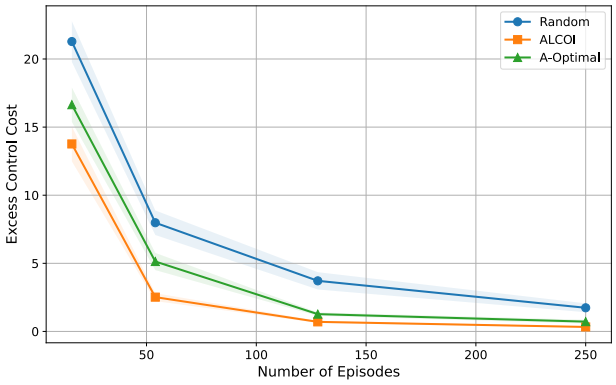

Fig. 2. Comparison of the proposed control-oriented identification procedure with approximate A-optimal design, and random experiment design. The mean over 100 runs is shown, with the standard error shaded.

VI. Conclusions

We have introduced and analyzed the Active Learning for Control-Oriented Identification (ALCOI) algorithm, marking a significant step towards understanding active exploration in model-based reinforcement learning for a general class of nonlinear dynamical systems. We provide finite sample bounds on the excess control cost achieved by the algorithm which offer insight into the interaction between the hardness of control and identification. Our bounds are known to be sharp up to logarithmic factors in the setting of nonlinear dynamical systems with linear dependence on the parameters, and we conjecture that they are sharp in general. Future work will attempt to verify that this is the case. It would also be interesting for future work to consider learning partially observed dynamics using general prediction error methods, rather than assuming a noiseless state observation.

Acknowledgment

BL and NM are supported by NSF Award SLES-2331880, NSF CAREER award ECCS-2045834 and AFOSR Award FA9550-24-1-0102. IZ is supported by a Swedish Research Council international postdoc grant. GP is supported in part by NSF Award SLES-2331880.

References

[1] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[2] T. M. Moerland, J. Broekens, A. Plaat, C. M. Jonker *et al.*, "Model-based reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 1, pp. 1–118, 2023.

[3] L. Ljung, "System identification," in *Signal analysis and prediction*. Springer, 1998, pp. 163–173.

[4] M. Gevers, "Towards a joint design of identification and control?" in *Essays on Control: Perspectives in the*

Theory and its Applications. Springer, 1993, pp. 111–151.

[5] H. Hjalmarsson, M. Gevers, and F. De Bruyne, "For model-based control design, closed-loop identification gives better performance," *Automatica*, vol. 32, no. 12, pp. 1659–1673, 1996.

[6] F. Pukelsheim, *Optimal design of experiments*. SIAM, 2006.

[7] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[8] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] A. J. Wagenmaker, M. Simchowitz, and K. Jamieson, "Task-optimal exploration in linear dynamical systems," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 641–10 652.

[10] A. Wagenmaker, G. Shi, and K. Jamieson, "Optimal exploration for model-based rl in nonlinear systems," *arXiv preprint arXiv:2306.09210*, 2023.

[11] M. Uehara and W. Sun, "Pessimistic model-based offline reinforcement learning under partial coverage," *arXiv preprint arXiv:2107.06226*, 2021.

[12] Y. Song and W. Sun, "Pc-mlp: Model-based reinforcement learning with policy cover guided exploration," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9801–9811.

[13] B. Sukhija, L. Treven, C. Sancaktar, S. Blaes, S. Coros, and A. Krause, "Optimistic active exploration of dynamical systems," *arXiv preprint arXiv:2306.12371*, 2023.

[14] I. Ziemann and S. Tu, "Learning with little mixing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4626–4637, 2022.

[15] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Conference on Learning Theory*. PMLR, 2020, pp. 3487–3582.

[16] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *arXiv preprint arXiv:2006.10277*, 2020.

[17] B. D. Lee, I. Ziemann, A. Tsiamis, H. Sandberg, and N. Matni, "The fundamental limitations of learning linear-quadratic regulators," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 4053–4060.

[18] S. Anderson and J. P. Hespanha, "Control-oriented identification for the linear quadratic regulator: Technical report," *arXiv preprint arXiv:2403.05455*, 2024.

[19] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite-sample perspective," *IEEE Control Systems Magazine*, vol. 43, no. 6, pp. 67–97, 2023.

[20] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems," *Journal of Machine Learning Research*, vol. 23, no. 140, pp. 1–49, 2022.

[21] A. A. Feldbaum, "Dual control theory. i," *Avtomatika i Telemekhanika*, vol. 21, no. 9, pp. 1240–1249, 1960.

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

[23] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, 2011.

[24] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

[25] I. Ziemann and H. Sandberg, "Regret lower bounds for learning linear quadratic gaussian systems," *arXiv preprint arXiv:2201.01680*, 2022.

[26] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun, "Information theoretic regret bounds for online nonlinear control," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 312–15 325, 2020.

[27] N. M. Boffi, S. Tu, and J.-J. E. Slotine, "Regret bounds for adaptive nonlinear control," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 471–483.

[28] B. Lee, I. Ziemann, G. Pappas, and N. Matni, "Active learning for control-oriented identification of nonlinear systems," 2024. [Online]. Available: https://sites.google.com/seas.upenn.edu/alcoi/home

[29] V. Roulet and A. d'Aspremont, "Sharpness, restart and acceleration," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] L. Ljung and P. E. Caines, "Asymptotic normality of prediction error estimators for approximate system models," *Stochastics*, vol. 3, no. 1-4, pp. 29–46, 1980.

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

[32] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 8921–8939.

[33] S. Mendelson, "Improving the sample complexity using global data," *IEEE transactions on Information Theory*, vol. 48, no. 7, pp. 1977–1991, 2002.

[34] I. Ziemann, S. Tu, G. J. Pappas, and N. Matni, "Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss," *arXiv preprint arXiv:2402.05928*, 2024.

[35] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[36] R. Tedrake, "Underactuated robotics: Learning, planning, and control for efficient and agile machines course notes for mit 6.832," *Working draft edition*, vol. 3, no. 4, p. 2, 2009.