Logarithmic Regret for Nonlinear Control

James Wang JWANG541@SEAS.UPENN.EDU
Bruce D. Lee BRUCELE@SEAS.UPENN.EDU
Ingvar Ziemann INGVARZ@SEAS.UPENN.EDU
Nikolai Matni NMATNI@SEAS.UPENN.EDU
All authors are with the Department of Electrical and System Engineering at the University of Pennsylvania.

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

We address the problem of learning to control an unknown nonlinear dynamical system through sequential interactions. Motivated by high-stakes applications in which mistakes can be catastrophic, such as robotics and healthcare, we study situations where it is possible for fast sequential learning to occur. Fast sequential learning is characterized by the ability of the learning agent to incur logarithmic regret relative to a fully-informed baseline. We demonstrate that fast sequential learning is achievable in a diverse class of continuous control problems where the system dynamics depend smoothly on unknown parameters, provided the optimal control policy is persistently exciting. Additionally, we derive a regret bound which grows with the square root of the number of interactions for cases where the optimal policy is not persistently exciting. Our results provide the first regret bounds for controlling nonlinear dynamical systems depending nonlinearly on unknown parameters. We validate the trends our theory predicts in simulation on a simple dynamical system.

Keywords: system identification, reinforcement learning, optimization

1. Introduction

Controlling an unknown nonlinear system through repeated sequential interaction is a fundamental problem in controls and reinforcement learning. Recent years have seen considerable impact of this paradigm in application areas ranging from walking robots (Yang et al., 2020), mastering games such as go and StarCraft (Silver et al., 2017) and even fine-tuning large language models (Ouyang et al., 2022). Problems of this form are often analyzed through the lens of Markov Decision Processes (MDP). Indeed, there is a wealth of literature on analyzing interactive sequential decision making in tabular MDPs (Burnetas and Katehakis, 1997; Dann and Brunskill, 2015). Extensions to this framework, typically motivated by studying large state and action spaces together with function approximation, are also abundant in the literature (Jiang et al., 2017; Zhou et al., 2021).

However, many problems, including certain robotics and healthcare tasks, are more naturally cast through the framework of continuous control. Such problems can be converted to tabular MDPs through discretization of the state and action spaces; however, doing so often results in intractable reinforcement learning problems. Conversely, the continuous control problem can be solved efficiently in special cases, such as the linear quadratic regulator (LQR) (Dean et al., 2020). Of the above motivating examples, robotic tasks in particular are plagued by costly data-collection (Kober et al., 2013). A similar situation arises in healthcare: giving the wrong treatment doses of a medicine repeatedly can have dire consequences. Consequently in these applications one would hope to find

fast learning algorithms that require as few interactions as possible with the unknown system to meet the desired performance criteria.

In the sequel, we measure the performance of an interactive sequential decision-maker by its regret—its performance as compared to the best policy (in a certain class), in hindsight. A fast learning algorithm in such sequential decision making tasks is characterized as one that attains *regret scaling logarithmically in the number of interactions with the unknown environment*. There has been a wealth of literature in characterizing when such rates are achievable in the setting of bandits (Lai and Robbins, 1985; Garivier et al., 2019) and analogs for tabular reinforcement learning (Burnetas and Katehakis, 1997; Ok et al., 2018; Xu et al., 2021). However, to date there has been no general characterization of when this is achievable in continuous control for nonlinear systems with nonlinear dependence on the unknown parameters. We thus ask: **are there conditions under which such fast learning algorithms exist for continuous control of nonlinear systems with nonlinear parameter dependencies?**

1.1. Contribution

Our main result answers the question of achievability of logarithmic regret in the affirmative.

Theorem 1 (Informal version of the main result) *If the optimal policy solving a given continuous control task is identifiable from an experiment running the optimal policy, polylogarithmic regret is attained by our Algorithm 1.*

The crux of our contribution is thus to establish a natural condition for logarithmic regret in nonlinear control problems and to provide a novel algorithm leveraging this condition which achieves logarithmic regret. To the best of our knowledge, this is the first algorithm achieving (poly-)logarithmic regret in general nonlinear control problems.

The intuition behind our result is as follows. If the data collected by running the optimal policy is sufficiently informative about the unknown parameters, then it is unnecessary to inject exploratory noise to perform online control. In particular, a policy which is near optimal will enjoy similarly informative data collection, allowing the learner to gradually approach the optimal policy by playing certainty equivalent controllers synthesized with estimates of the dynamics parameters. We formalize this intuition with a persistence of excitation condition, asking that the Fisher information matrix of the optimal policy is positive definite.

Finally, for completeness, we also provide an algorithm attaining sublinear regret in the absence of our identifiability condition. This result can be found in our full manuscript (https://arxiv.org/abs/2501.10261) along with all proofs and additional experiments.

1.2. Related Work

Logarithmic Regret in Bandits and RL The question of whether logarithmic regret is attainable or not is intimately connected with the exploration exploitation trade-off. Beginning with Lai and Robbins (1985) in the tabular bandit setting, *gap-dependent* regret bounds have been established showing that logarithmic regret is possible whenever there is a strict separation between the reward of the optimal action and that of a second best, or worse, action. Similar gap sufficient conditions for logarithmic regret also exist in tabular reinforcement learning (Burnetas and Katehakis, 1997; Ok et al., 2018; Xu et al., 2021). In the worst case, or for instance in linear bandits where there is

no gap, logarithmic regret is impossible and instead regret scales with the square root of the number of interactions with unknown environment (cf. e.g., Garivier et al., 2019).

Closed-Loop Identifiability and Adaptive Control Within the system identification community, the exploration-exploitation trade-off is often referred to as the *dual nature of control* (Feldbaum, 1960a,b) and is related to issues of *closed-loop identifiability* (Ljung, 1999). Roughly speaking, closed-loop identifiability issues arise because a fixed control law might not sufficiently excite the system under consideration in the necessary directions in state space (or feature space more generally). Indeed, in the Linear Quadratic Regulator (LQR) setting, Polderman (1986) gives an elegant geometric argument showing that the true parameters need to be identified in order to ascertain the optimal control law. It is also interesting to note that, precisely because the minimum variance controller is closed-loop identifiable (Lin et al., 1985) (in contrast to the more general LQR controller), logarithmic regret can be achieved in this setting (Lai, 1986). Reiterating the point above: the reason for the impossibility of pure exploitation is precisely a lack of closed-loop identifiability. This insight is leveraged in Simchowitz and Foster (2020) and Ziemann and Sandberg (2024) to show logarithmic regret is impossible in general in the linear quadratic Gaussian control problem. However, given some prior information about the system (e.g. if the way the input impacts the state transitions is known), then closed-loop identifiability may hold, making logarithmic regret achievable for LQR (Cassel et al., 2020; Jedra and Proutiere, 2022; Lee et al., 2024a). Alternatively, if the policy choice is restricted to a set in which all possible candidate provide closed-loop identifiability of the system parameters, then Lale et al. (2020) demonstrate logarithmic regret for the Linear Quadratic Gaussian (LQG).

Closed-loop identifiability issues similarly hinder the achievability of logarithmic regret in the online control of nonlinear systems. In the setting of nonlinear dynamical systems which depend linearly on some unknown parameters, Kakade et al. (2020); Boffi et al. (2021) propose algorithms that achieve regret scaling with the square root of the number of interactions. Lale et al. (2024) consider linear function approximators for smooth systems, and provide an algorithm achieving regret scaling with the square root of the number of interactions in general, and logarithmic regret if the system is sufficiently smooth. Critically, as with Lale et al. (2020), Lale et al. (2024) assume that all policies in the policy class provide closed-loop identifiability of the parameters. By contrast, we do not assume a priori access to a policy yielding such identifiability; we show that it suffices that the *unknown* optimal policy yields easy identification and our algorithm then adapts to this property. Moreover, we consider dynamical systems which depend nonlinearly on an unknown parameter, and propose an algorithm that incurs logarithmic regret as long as the optimal policy enables closed-loop identification.

Learning in Dynamical Systems Our contribution also draws on a recent line of work on learning in dynamical systems beginning with Simchowitz et al. (2018); Faradonbeh et al. (2018). The authors therein show that non-asymptotic parameter recovery from a single trajectory is possible in certain marginally stable, or unstable, linear dynamical systems. Mania et al. (2022) leverage the parameter recovery bounds to enable efficient exploration. Non-asymptotic identification of more general nonlinear systems is studied by Sattar and Oymak (2022); Foster et al. (2020); Ziemann and Tu (2022). Treven et al. (2023); Wagenmaker et al. (2024); Lee et al. (2024b) study control-oriented experiment design in an episodic setting for nonlinear systems.

1.3. Notation

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 $\left\|D^{(p)}g(x)\right\|_{\mathsf{op}}$. For a function $f : \mathsf{X} \to \mathbb{R}^{d_y}$, we define $\|f\|_\infty \triangleq \sup_{x\in\mathsf{X}} \|f(x)\|$. A Euclidean norm ball of radius r centered at x is denoted $\mathcal{B}(x, r)$.

2. Problem Formulation

We consider a nonlinear dynamical system given by the dynamics

$$x_{t+1} = f(x_t, u_t, \phi^*) + w_t, \quad t = 1, ..., T-1 \tag{1}$$

where the state $x_t \in \mathbb{R}^{d_x}$; the input $u_t \in \mathbb{R}^{d_u}$; and the additive noise $w_t \in \mathbb{R}^{d_x}$, with $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I)$. Let $x_1 \in \mathbb{R}^{d_x}$ be arbitrary. Here, f is the dynamics function and depends on a parameter $\phi^* \in \mathbb{R}^{d_\phi}$. We assume that there exists some positive B such that $\|\phi^*\| \leq B$ and $\|f(\cdot, \cdot, \phi)\|_\infty \leq B$ for all $\phi \in \mathbb{R}^{d_\phi}$ satisfying $\|\phi\| \leq B$.

We study an online learning problem under these dynamics. We consider a learner who has knowledge of the dynamics f, but not the parameter ϕ^*. In each episode $n = 1, \ldots, N$, the learner executes a policy π_n from the set of policies $\{\pi_0\} \cup \Pi$, where π_0 is an initial (possibly randomized) exploration policy, while Π is a class of deterministic controllers which take as input a point $x \in \mathbb{R}^{d_x}$ and return a control input $u \in \mathbb{R}^{d_u}$. Then, the learner observes a trajectory $(x_1, u_1), \ldots (x_T, u_T)$ (generated by unrolling (1) with $u_t \sim \pi_n(x_t)$); and incurs the cost $J(\pi_n, \phi^*)$, where

$$J(\pi, \phi) := \mathbb{E}_\pi^\phi \left[\sum_{t=1}^{T} c_t(x_t, u_t)\right] \tag{2}$$

for some cost functions $\{c_t\}_{t=1,\ldots,T}$ which are fixed across episodes. The subscript on the expectation denotes that the policy π is played, while the superscript denotes that the dynamics (1) are rolled out under ϕ. The expectation is taken over the noise w_t and the policy π_n. We suppose that the policy class Π is parametric: $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^{d_\theta}\}$.

The learner's objective is to achieve a low sum of costs over episodes. A natural metric is therefore to minimize the regret, defined as

$$\mathrm{Regret}(N) := \left(\sum_{n=1}^{N} J(\pi_n, \phi^*)\right) - N \min_{\pi\in\Pi} J(\pi, \phi^*). \tag{3}$$

We will explore no-regret learners for this setting, for which $\mathrm{Regret}(N)/N \to 0$ as $N \to \infty$.

2.1. Certainty Equivalent Control

Since the learner does not have access to the true dynamics ϕ^*, it cannot directly solve a policy optimization problem under the system $f(x, u, \phi^*)$ (Fazel et al., 2018) for the optimal controller. Instead, we leverage the principle of certainty equivalence. In particular, the learner uses the data

collected from its interactions to pose an estimate $\hat{\phi}$ for the parameter $\phi^\star$. Using this estimate, the learner solves the policy optimization problem,

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

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

$$\pi^*(\hat{\phi}) \triangleq \pi_{\theta^*(\hat{\phi})}. \tag{5}$$

Under some additional assumptions (which we will lay out in the following section), we can measure the performance of the policy $\pi^*(\hat{\phi})$ in terms of the quality of the estimate $\hat{\phi}$. Here, we define the prediction error of $\hat{\phi}$, on trajectories collected using a policy π (which is not necessarily $\pi^*(\hat{\phi})$), as

$$\mathrm{Err}_\pi^{\phi^*}(\phi) \triangleq \mathbb{E}_\pi^{\phi^*}\left[\frac{1}{T}\sum_{t=1}^{T} \|f(x_t, u_t, \phi) - x_{t+1}\|^2\right]. \tag{6}$$

2.2. Assumptions

In order to relate the excess cost achieved by a certainty equivalent controller synthesized under a dynamics estimate ϕ to the quality of the estimate $\hat{\phi}$, we impose some smoothness assumptions on the dynamics and policy class.

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

Assumption 2 *(Smooth exploitation policy class). For all policies $\pi \in \Pi$ and $x \in \mathcal{X}$, the function $\pi_\theta(x)$ is four-times differentiable in θ. Furthermore $\left\|D_\theta^{(i)} \pi_\theta(x)\right\|_{\mathsf{op}} \le L_\Pi$ for all $i = 1, ..., 4$, all $\theta \in \mathbb{R}^{d_\theta}$, and all $x \in \mathcal{X}$.*

We additionally require that the costs are bounded for policies in the class $\{\pi_0\} \cup \Pi$ and all dynamics parameters in a neighborhood of the true parameter. Intuitively, this allows our learning algorithm to occasionally play bad policies without incurring too much excess cost.

Assumption 3 *(Bounded costs). There exists $r_{\mathsf{cost}} > 0$ such that for all $\phi \in \mathcal{B}(\phi^*, r_{\mathsf{cost}})$, and all $\pi \in \{\pi_0\} \cup \Pi$, we have $\mathbb{E}_\pi^\phi\left[\left(\sum_{t=1}^T c_t(x_t, u_t)\right)^2\right] \le T^2 L_{\mathsf{cost}}^2$. (Together with Jensen's inequality, this immediately implies that $\mathbb{E}_\pi^\phi\left[\sum_{t=1}^T c_t(x_t, u_t)\right] \le T L_{\mathsf{cost}}$ for such ϕ and π.)*

As the task is episodic, the above assumption holds if the stage costs are uniformly bounded for all $x \in \mathbb{R}^{d_x}$ and $u \in \mathbb{R}^{d_u}$. Alternatively, if the stage costs are smooth, the above condition holds if the states and inputs are bounded with high probability. This is satisfied for Π by the smoothness of the dynamics (Assumption 1) and exploitation policy class (Assumption 2). A mild assumption that the initial policy π_0 plays bounded inputs suffices to guarantee the above condition also holds for π_0.

We additionally suppose that the certainty equivalent controller parameters, as a function of the estimated dynamics ϕ, are locally smooth near the true dynamics ϕ^*.

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

- $\nabla_\theta J(\pi_\theta, \phi) \mid_{\theta=\theta^*(\phi)} = 0$,
- *$\theta^*(\phi)$ is three times differentiable and $\left\| D_\phi^{(i)} \theta^*(\phi) \right\|_{\mathsf{op}} \le L_{\mathsf{ce}}$ for some $L_{\mathsf{ce}} > 0$ and $i \in \{1, 2, 3\}$.*

It is shown in Proposition 6 of Wagenmaker et al. (2024) that this condition holds if the minimizer of $J(\pi_\theta, \phi_\star)$ is unique, and $\nabla_\theta^2 J(\pi_\theta, \phi_\star) \succ 0$.

In order to bound the parameter recovery error in terms of the prediction error, additional identifiability conditions are needed. Ziemann et al. (2024) show that a rather minimal Lojasiewicz condition (cf. Roulet and d'Aspremont, 2017) relating the sharpness of an objective to its manifold of minimizers is sufficient for learning from dependent data. The following definition of a Lojasiewicz policy is taken from Lee et al. (2024b) and extends the corresponding definition from Ziemann et al. (2024) to decision-making. In the setting of Lee et al. (2024b), the following definition of a Lojasiewicz policy bounds the estimation error $\|\phi - \phi^*\|$ as a function of the prediction error $\mathrm{Err}_\pi^{\phi^*}(\phi)$ for all dynamics parameters ϕ.

Definition 2 *For positive numbers C and α, say that a policy $\pi \in \Pi$ is (C, α)-Lojasiewicz if*

$$\left\| \hat{\phi} - \phi^* \right\| \le C \, \mathrm{Err}_\pi^{\phi^*}(\hat{\phi})^\alpha \quad \textit{for all } \hat{\phi} \in \mathbb{R}^{d_\phi}.$$

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

Assumption 5 *(Initial Lojasiewicz policy). Fix some positive constant C_{Loja} and $\alpha \in (1/4, 1/2]$. The learner has access to a policy π_0 which is $(C_{\mathsf{Loja}}, \alpha)$-Lojasiewicz (here, we do not require that $\pi_0 \in \Pi$; furthermore, we allow π_0 to be randomized).*

This is satisfied in linear systems with $\alpha = 1/2$ if the initial controller π_0 plays Gaussian noise as input, and both the controller noise and process noise have positive definite covariance matrices. More generally, Theorem 2 of Musavi et al. (2024) implies that playing bounded i.i.d. random inputs suffices to satisfy this condition with $\alpha = 1/2$ in a broad class of analytic nonlinear systems (although the constant C_{Loja} may be very high).

While Assumption 5 ensures that the learner can identify the true dynamics ϕ^* using only data collected under π_0, the rate of recovery may be slow under only the assumptions listed previously. In order to obtain polylogarithmic regret bounds, we require the assumption that the optimal controller, defined by $\theta^* \triangleq \mathrm{argmin}_\theta J(\pi_\theta, \phi_\star)$, is persistently exciting. Persistence of excitation for a nonlinear dynamical system involves the positive definiteness of the matrix

$$\Sigma^\pi \triangleq \mathbb{E}_\pi^{\phi^*} \left[\frac{1}{T} \sum_{t=1}^{T} Df(x_t, u_t, \phi^*)^\top Df(x_t, u_t, \phi^*) \right] = D_\phi^{(2)} \left(\mathrm{Err}_\pi^{\phi^*}(\phi) \right) \Big|_{\phi=\phi^*}$$

where $Df(x_t, u_t, \phi^*)$ denotes the Jacobian of f with respect to ϕ evaluated at ϕ^*. It can be shown that Σ^π is a positive scalar multiple of the Fisher Information matrix (when the system evolves according to ϕ^* and π) and hence this condition is equivalent to requiring the positive definiteness of this Fisher Information matrix.

Assumption 6 *(Persistency of excitation for the optimal controller). The optimal policy under the true dynamics ϕ^*, denoted $\pi_{\theta^*} \triangleq \pi_{\theta^*(\phi^*)}$, is persistently exciting, i.e. for some $\mu > 0$,*

$$\mathbb{E}^{\phi^*}_{\pi_{\theta^*}}\left[\frac{1}{T}\sum_{t=1}^{T} Df(x_t,u_t,\phi^*)^\top Df(x_t,u_t,\phi^*)\right] \succeq \mu I_{d_\phi}.$$

Note that the above assumption is not satisfied in LQR in general when both the A^* and B^* matrices are unknown. However, Lee et al. (2024a) show that a sufficient condition for Assumption 6 to hold in linear systems is that either 1) the A^* matrix is known and the optimal controller K^* has full row rank or 2) the B^* matrix is known.

Finally, we reiterate that in the event that Assumption 6 does not hold, we can obtain slower, but still sublinear regret rates under very general conditions. See Appendix A of the extended manuscript for details.

3. Fast Learning

Under Assumptions 1, 2, 3, 4, 5, and 6, we give an algorithm (Algorithm 1) based on the aforementioned certainty equivalence principle which achieves polylogarithmic regret in our online nonlinear control setting. Given an initial Lojasiewicz policy π_0, the exploitation policy class Π, the number of episodes N, the number of initial phase episodes $N_{\mathsf{phase\,1}}$ (where $0 \le N_{\mathsf{phase\,1}} \le N$), and a confidence radius r_Φ, the algorithm proceeds in two phases.

In the first phase, the learner collects a dataset $\{(x_{n,t},u_{n,t},x_{n,t+1})\}_{t=1,\dots,T}^{n=1,\dots,N_{\mathsf{phase\,1}}}$ using π_0, and finds an confidence ball Φ with radius r_Φ such that $\phi^* \in \Phi$ with high probability. The confidence ball is centered at ϕ_0, which is the solution to a nonlinear least squares problem,

$$\phi_0 \in \operatorname*{argmin}_{\phi\in\mathbb{R}^{d_\phi},\|\phi\|\le B} \sum_{n=1}^{N_{\mathsf{phase\,1}}}\sum_{t=1}^{T} \|x_{n,t+1} - f(x_{n,t},u_{n,t},\phi)\|^2. \tag{7}$$

With a sufficiently small r_Φ, and conditioned on the event $\phi^* \in \Phi$, we show that policies synthesized using estimates that fall within this set enjoy a positive definite Fisher Information; equivalently, the prediction error $\mathrm{Err}^{\phi^*}_\pi(\phi)$ is strongly convex on Φ for all controllers π synthesized with dynamics estimates $\hat\phi \in \Phi$. This motivates an online convex optimization procedure in the second phase.

In the second phase, the learner interacts with the system by playing policies synthesized using parameter estimates from Φ. More specifically, the learner uses the certainty equivalent policy π corresponding to its current estimate of ϕ^* to collect a single trajectory $\mathcal{D} = \{(x_t,u_t,x_{t+1})\}_{t=1,\dots,T}$. The mean-squared-error of a dynamics estimate ϕ on the dataset $\mathcal{D}$ is

$$l_{\mathcal{D}}(\phi) \triangleq \frac{1}{T}\sum_{t=1}^{T} \|f(x_t,u_t,\phi) - x_{t+1}\|^2 \tag{8}$$

and the learner updates its estimate of ϕ^* using the gradient $\nabla l_{\mathcal{D}}(\phi)$, and repeats this process.

In general, the nonlinear least squares problem (7) and policy optimization problem (4) may be computationally challenging. The focus of this work is to understand the statistical complexity of the problem rather than the computational complexity. However, it is worth noting that the online optimization procedure is computationally efficient and therefore the learner may efficiently execute

Algorithm 1 Continuous Refinement

Require: Exploration policy π_0, exploitation policy class Π, number of episodes N, number of initial phase episodes $N_{\mathsf{phase\,1}}$, confidence radius r_Φ, hyperparameter μ

Play π_0 for $N_{\mathsf{phase\,1}}$ episodes to collect the dataset $\mathcal{D}_0 := \{(x_{n,t}^{(0)}, u_{n,t}^{(0)}, x_{n,t+1}^{(0)})\}_{t=1,\dots,T}^{n=1,\dots,N_{\mathsf{phase\,1}}}$ ▷ First phase

Set ϕ_0 via least squares (7) using $\mathcal{D}_0$

$\Phi \leftarrow \mathcal{B}(\phi_0, r_\Phi)$

for $i = 0, 1, \dots, N - N_{\mathsf{phase\,1}} - 1$ **do** ▷ Second phase

  Set $\pi_{i+1} \leftarrow \pi^*(\phi_i)$

  Play π_{i+1} to collect a single trajectory $\mathcal{D}_{i+1} := \{(x_t^{(i+1)}, u_t^{(i+1)}, x_{t+1}^{(i+1)})\}_{t=1,\dots,T}$

  $\psi_{i+1} \leftarrow \phi_i - \frac{8}{\mu \cdot (i+1)} \nabla l_{\mathcal{D}_{i+1}}(\phi_i)$, with $l_{\mathcal{D}}$ in (8)

  $\phi_{i+1} \leftarrow \operatorname{argmin}_{\phi \in \Phi} \|\phi - \psi_{i+1}\|$

the second phase of the dynamics estimation procedure online. Additionally, for particular systems (1) and objectives (2), the policy optimization problem (4) may be efficient. This is the case, for instance, if the optimal solution to the policy optimization problem can be achieved via feedback linearization (Charlet et al., 1989) by choosing the input to cancel out some portion of the dynamics. We consider such an example in Section 4.

Our main result bounds the regret incurred by Algorithm 1 in terms of N and $N_{\mathsf{phase\,1}}$ under the aforementioned smoothness and identifiability conditions.

Theorem 3 *Consider applying Algorithm 1 to the system (1) with initial policy π_0 satisfying Assumption 5, policy class Π satisfying Assumption 2, number of iterations N, number of initial phase episodes $N_{\mathsf{phase\,1}}$ and confidence radius r_Φ. Additionally suppose that the dynamics satisfy Assumption 1 and that the costs satisfy Assumption 3. Furthermore, suppose that the dynamics, objective, and policy class satisfy Assumption 4. Finally, suppose that the true optimal controller π_{θ^*} and the hyperparameter μ satisfy Assumption 6. Then,*

$$\mathbb{E}[\mathrm{Regret}(N)] \le \mathsf{poly}_\alpha(d_x, \sigma, L_f, L_{\mathsf{cost}}, \mu^{-1}) T \log N + T N_{\mathsf{phase\,1}} L_{\mathsf{cost}}$$

as long as the following both hold:

- $r_\Phi \le \mathsf{poly}\Big(r_{\mathsf{ce}}, r_{\mathsf{cost}}, d_x^{-1}, \sigma^{-1}, L_f^{-1}, L_\Pi^{-1}, L_{\mathsf{ce}}^{-1}, \mu\Big) T^{-1/2}$,
- $N_{\mathsf{phase\,1}} \ge \mathsf{poly}_\alpha\big(\log N, \log T, d_x, d_\phi, \sigma, C_{\mathsf{Loja}}, r_{\mathsf{ce}}^{-1}, r_{\mathsf{cost}}^{-1}, L_f, L_\Pi, L_{\mathsf{ce}}, \mu^{-1}, \log B\big) T^{1/(2\alpha)-1}$. *Here, the subscript α indicates that the degree of the polynomial depends on α.*

Theorem 3 states that if r_Φ is chosen small enough and the number of initial phase episodes $N_{\mathsf{phase\,1}}$ exceeds some burn-in which is polylogarithmic in N and polynomial in all other relevant system parameters, then the regret of Algorithm 1 grows at most linearly with $\log N$ and $N_{\mathsf{phase\,1}}$. Plugging in specific choices for r_Φ and $N_{\mathsf{phase\,1}}$ yields a polylogarithmic regret bound for Algorithm 1.

Corollary 4 *Suppose we apply Algorithm 1 in the setting of Theorem 3 with the parameters:*

- $r_\Phi = \mathsf{poly}\Big(r_{\mathsf{ce}}, r_{\mathsf{cost}}, d_x^{-1}, \sigma^{-1}, L_f^{-1}, L_\Pi^{-1}, L_{\mathsf{ce}}^{-1}, \mu\Big) T^{-1/2}$,

- $N_{\mathsf{phase\,1}} = \mathsf{poly}_\alpha\big(\log N, \log T, d_x, d_\phi, \sigma, C_{\mathsf{Loja}}, r_{\mathsf{ce}}^{-1}, r_{\mathsf{cost}}^{-1}, L_f, L_\Pi, L_{\mathsf{ce}}, \mu^{-1}, \log B\big) T^{1/(2\alpha)-1}$.

Then, Algorithm 1 achives regret depending polylogarithmically on the number of episodes N, i.e.,

$$\mathbb{E}[\mathrm{Regret}(N)] \le \mathsf{poly}_\alpha\big(\log N, \log T, d_x, d_\phi, \sigma, C_{\mathsf{Loja}}, r_{\mathsf{ce}}^{-1}, r_{\mathsf{cost}}^{-1}, L_f, L_\Pi, L_{\mathsf{ce}}, L_{\mathsf{cost}}, \mu^{-1}, \log B\big) T^{1/(2\alpha)}.$$

The full proof of Theorem 3 may be found in the full manuscript; we provide a brief sketch below.

Proof [Proof Sketch] For $N_{\mathsf{phase\,1}}$ satisfying the given bound, the system identification results of Ziemann and Tu (2022); Lee et al. (2024b) ensure that the confidence set Φ is constructed such that $\phi_\star \in \Phi$ with probability at least $1 - 1/N$. The regret is decomposed into three parts: that of the initial exploration phase, that of the second phase under the failure event where $\phi_\star \notin \Phi$, and that of the second phase under the success event, where $\phi_\star \in \Phi$. Using the bound on the episode costs, the regret incurred from the first phase is bounded by $TL_{\mathsf{cost}}N_{\mathsf{phase\,1}}$ and the regret incurred during the second phase under the failure event $\phi_\star \notin \Phi$ is bounded by $TL_{\mathsf{cost}}(N - N_{\mathsf{phase\,1}})\mathbb{P}[\phi_\star \notin \Phi] \le L_{\mathsf{cost}}$. The condition on the radius of the confidence set ensures that the prediction error is strongly convex when the learner plays a certainty equivalent controller synthesized using any system estimate $\phi \in \Phi$. This leads to an analysis similar to that of stochastic gradient descent (Robbins and Monro, 1951) on a strongly convex objective to bound the regret incurred during the second phase. Summing the contributions of the three components leads to the regret bound in Theorem 3. ■

Before proceeding, we note that while the regret of Algorithm 1 depends *polylogarithmically* on the number of episodes N, it depends *polynomially (quasilinearly when $\alpha = 1/2$)* on the episode length T. Intuitively, one might expect a sublinear dependence on T since increasing T increases the number of interactions the learner has with the system. The polynomial dependence on T arises because we consider an episodic setting without mixing assumptions within episodes. As a result, the sensitivity of the episode cost to parameter estimation scales quadratically in the episode horizon. Imposing mixing assumptions can reduce this to linear scaling by modifying the proof of Lemma D.1 of Wagenmaker et al. (2024). Therefore, by imposing stronger assumptions which lead to mixing, such as stability of the initial and optimal policies, one can likely achieve a logarithmic dependence on T. We leave formalizing this to future work.

4. Numerical Validation

4.1. Toy Experiment

We provide an simple example to illustrate the fast regret rates attained by Algorithm 1. For more experiments, see our full manuscript. Consider the two-dimensional nonlinear system

$$x_{t+1} = x_t + 5\exp\Big(-\|x_t - \phi^*\|^2\Big)\frac{x_t - \phi^*}{\|x_t - \phi^*\|} + u_t + w_t \tag{9}$$

where $x_t, u_t, w_t, \phi^* \in \mathbb{R}^2$, and with $x_1 = \begin{bmatrix}0 & 0\end{bmatrix}^\top$. The noise w_t has a standard normal distribution. We choose the unknown parameter $\phi^* = \begin{bmatrix}0.25 & 0.25\end{bmatrix}^\top$.

In this experiment, we use the horizon $T = 10$ and the number of episodes $N = 3000$. We will consider the quadratic cost functions

$$c_t(x, u) = \|x\|^2 \quad \text{for } t = 1, \dots, T.$$

The policy class Π consists of controllers parameterized by the dynamics estimate $\hat{\phi}$, with

$$\pi_{\hat{\phi}}(x) = -\left(x + 5 \exp\left(-\left\| x_t - \hat{\phi} \right\|^2 \right) \frac{x_t - \hat{\phi}}{\left\| x_t - \hat{\phi} \right\|} \right). \tag{10}$$

It can be shown that the dynamics (9) and policy class (10) satisfy Assumption 6. Our initial policy π_0 plays the controller π_ϕ corresponding to $\phi = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$, which can be shown to satisfy Assumption 5. In place of choosing $N_{\mathsf{phase\,1}}$ or r_Φ according to Theorem 1, we heuristically set $N_{\mathsf{phase\,1}} = 100$ and $r_\Phi = 0.2$. We note that the dynamics are not uniformly bounded globally, however they are uniformly bounded with high probability.

Under this choice of cost function and policy class, the learner's objective is to keep the system near the origin. Figure 1 illustrates the performance (measured in terms of regret) of Algorithm 1 on the toy dynamical system. The first plot shows that, after the initial $N_{\mathsf{phase\,1}}$-episode initial phase, the excess cost incurred per round begins to decay quickly, leading to the regret growing polylogarithmically with N. The second plot is included to better illustrate the regret attained by Algorithm 1; after the initial phase, the average regret appears to grow as a polynomial of the logarithm of the iteration. This toy example highlights the fast regret rates attained by Algorithm 1.

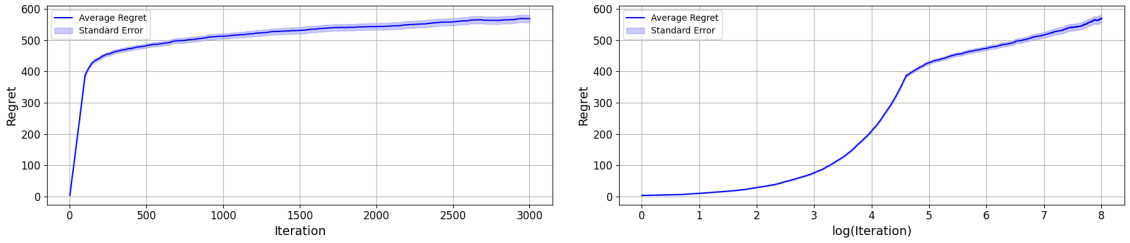

Figure 1: Average regret incurred by Algorithm 1 on the toy dynamical system (9), versus iterations and $\log(\text{iterations})$, respectively. The mean over 30 runs is shown, with the standard error shaded.

5. Conclusion

We have introduced Algorithm 1 for online learning in a broad class of nonlinear dynamical systems. We have also proven a general sufficient condition for polylogarithmic regret under a natural curvature condition — when the Fisher information matrix at the optimal policy is positive definite (detailed in our Assumption 6) — and show that polylogarithmic regret is achieved by our Algorithm 1. Finally, we have verified the performance of Algorithm 1 on a toy dynamical system and show that it achieves a fast regret rate in practice. Future work could extend these results to the single-trajectory setting. In particular, it could be interesting to extend the $\log^2 N$ regret rates of Cassel et al. (2020) and Lee et al. (2024a) in the single-trajectory partially known linear setting to the setting with nonlinear dynamics. Another exciting avenue for future work is to design an online learning algorithm which deploys optimal experiment design techniques (Wagenmaker et al., 2024) to optimally balance exploration and exploitation. Doing so may result in algorithms which automatically determine whether Assumption 6 is satisfied. Such an algorithm could achieve logarithmic regret if possible, and otherwise achieve $\sqrt{N}$ regret. Additionally, it may be possible to show improved dependence on the system-theoretic constants by using this approach.

Acknowledgements

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.

References

Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pages 471–483. PMLR, 2021.

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

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.

B Charlet, J Levine, and R Marino. On dynamic feedback linearization. *Systems & Control Letters*, 13(2):143–151, 1989.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

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

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

AA Feldbaum. Dual Control Theory. I. *Avtomatika i Telemekhanika*, 21(9):1240–1249, 1960a.

AA Feldbaum. Dual Control Theory. II. *Avtomatika i Telemekhanika*, 21(11):1453–1464, 1960b.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

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.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jiang17c.html.

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.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

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

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

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

Sahin Lale, Peter I Renn, Kamyar Azizzadenesheli, Babak Hassibi, Morteza Gharib, and Anima Anandkumar. FALCON: Fourier adaptive learning and control for disturbance rejection under extreme turbulence. *npj Robot*, 2(1), September 2024.

Bruce Lee, Anders Rantzer, and Nikolai Matni. Nonasymptotic regret analysis of adaptive linear quadratic control with model misspecification. In *6th Annual Learning for Dynamics & Control Conference*, pages 980–992. PMLR, 2024a.

Bruce D Lee, Ingvar Ziemann, George J Pappas, and Nikolai Matni. Active learning for control-oriented identification of nonlinear systems. *arXiv preprint arXiv:2404.09030*, 2024b.

Woei Lin, PR Kumar, and TI Seidman. Will the self-tuning approach work for general cost criteria? *Systems & control letters*, 6(2):77–85, 1985.

Lennart Ljung. *System identification: theory for the user*. PTR Prentice Hall, Upper Saddle River, NJ, 1999.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23(32):1–30, 2022.

Negin Musavi, Ziyao Guo, Geir Dullerud, and Yingying Li. Identification of analytic nonlinear dynamical systems with non-asymptotic guarantees. *arXiv preprint arXiv:2411.00656*, 2024.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

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, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. *Advances in Neural Information Processing Systems*, 36:38122–38153, 2023.

Andrew Wagenmaker, Guanya Shi, and Kevin G Jamieson. Optimal exploration for model-based rl in nonlinear systems. *Advances in Neural Information Processing Systems*, 36, 2024.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472. PMLR, 2021.

Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Tingnan Zhang, Jie Tan, and Vikas Sindhwani. Data efficient reinforcement learning for legged robots. In *Conference on Robot Learning*, pages 1–10. PMLR, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *IEEE Transactions on Automatic Control*, 2024.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35:4626–4637, 2022.

Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. In *Forty-first International Conference on Machine Learning*, 2024.