Proceedings of Machine Learning Research vol 144:1–13, 2021

Regret Bounds for Adaptive Nonlinear Control

Nicholas M. Boffi BOFFI@G.HARVARD.EDU
Harvard University

Stephen Tu STEPHENTU@GOOGLE.COM
Google Brain Robotics

Jean-Jacques E. Slotine JJS@MIT.EDU
Massachusetts Institute of Technology

Abstract

We study the problem of adaptively controlling a known discrete-time nonlinear system subject to unmodeled disturbances. We prove the first finite-time regret bounds for adaptive nonlinear control with matched uncertainty in the stochastic setting, showing that the regret suffered by certainty equivalence adaptive control, compared to an oracle controller with perfect knowledge of the unmodeled disturbances, is upper bounded by $\widetilde{O}(\sqrt{T})$ in expectation. Furthermore, we show that when the input is subject to a k timestep delay, the regret degrades to $\widetilde{O}(k\sqrt{T})$. Our analysis draws connections between classical stability notions in nonlinear control theory (Lyapunov stability and contraction theory) and modern regret analysis from online convex optimization. The use of stability theory allows us to analyze the challenging infinite-horizon single trajectory setting.

Keywords: Adaptive control, online convex optimization, matched uncertainty.

1. Introduction

The goal of adaptive nonlinear control (Slotine and Li, 1991; Ioannou and Sun, 1996; Fradkov et al., 1999) is to control a continuous-time dynamical system in the presence of unknown dynamics; it is the study of concurrent learning and control of dynamical systems. There is a rich body of literature analyzing the stability and convergence properties of classical adaptive control algorithms. Under suitable assumptions (e.g., Lyapunov stability of the known part of the system), typical results guarantee asymptotic convergence of the unknown system to a fixed point or desired trajectory.

On the other hand, due to recent successes of reinforcement learning (RL) in the control of physical systems (Yang et al., 2019; OpenAI et al., 2019; Hwangbo et al., 2019; Williams et al., 2017; Levine et al., 2016), there has been a flurry of research in online RL algorithms for continuous control. In contrast to the classical setting of adaptive nonlinear control, online RL algorithms operate in discrete-time, and often come with finite-time regret bounds (Wang et al., 2019; Kakade et al., 2020; Jin et al., 2020; Cao and Krishnamurthy, 2020; Cai et al., 2020; Agarwal et al., 2020). These bounds provide a quantitative rate at which the control performance of the online algorithm approaches the performance of an oracle equipped with hindsight knowledge of the uncertainty.

In this work, we revisit the analysis of adaptive nonlinear control algorithms through the lens of modern reinforcement learning. Specifically, we show how to systematically port matched uncertainty adaptive control algorithms to discrete-time, and we use the machinery of online convex optimization (Hazan, 2016) to prove finite-time regret bounds. Our analysis uses the notions of contraction and incremental stability (Lohmiller and Slotine, 1998; Angeli, 2002) to draw a connection

between control regret, the quantity we are interested in, and function prediction regret, the quantity online convex optimization enables us to bound.

We present two main sets of results. First, we provide a discrete-time analysis of *velocity gradient* adaptation (Fradkov et al., 1999), a broad framework which encompasses e.g., classic adaptive sliding control (Slotine and Coetsee, 1986). We prove that in the deterministic setting, if a Lyapunov function describing the nominal system is strongly convex in the state, then the corresponding velocity gradient algorithm achieves constant regret with respect to a baseline controller having full knowledge of the system. Our second line of results considers the use of online least-squares *gradient based optimization* for the parameters. Under an incremental input-to-state stability assumption, we prove $\widetilde{O}(\sqrt{T})$ regret bounds in the presence of stochastic process noise. We further show that when the input is delayed by k timesteps, the regret degrades to $\widetilde{O}(k\sqrt{T})$. Importantly, our bounds hold for the challenging single trajectory infinite horizon setting, rather than the finite-horizon episodic setting more frequently studied in reinforcement learning. We conclude with simulations showing the efficacy of our proposed discrete-time algorithms in quickly adapting to unmodeled disturbances. Proofs and more details can be found in the full version of the paper (Boffi et al., 2020).

2. Related Work

There has been a renewed focus on the continuous state and action space setting in the reinforcement learning (RL) literature. The most well-studied problem for continuous control in RL is the Linear Quadratic Regulator (LQR) problem with unknown dynamics. For LQR, both upper and lower bounds achieving $\sqrt{T}$ regret are available (Abbasi-Yadkori and Szepesvári, 2011; Agarwal et al., 2019a; Mania et al., 2019; Cohen et al., 2019; Simchowitz and Foster, 2020; Hazan et al., 2020), for stochastic and adversarial noise processes. Furthermore, in certain settings it is even possible to obtain logarithmic regret (Agarwal et al., 2019b; Cassel et al., 2020; Foster and Simchowitz, 2020).

Results that extend beyond the classic LQR problem are less complete, but are rapidly growing. Recently, Kakade et al. (2020) showed $\sqrt{T}$ regret bounds in the finite horizon episodic setting for dynamics of the form $x_{t+1} = A\phi(x_t, u_t) + w_t$ where A is an unknown operator and ϕ is a known feature map, though their algorithm is generally not tractable to implement. Mania et al. (2020) show how to actively recover the parameter matrix A using trajectory optimization. Azizzadenesheli et al. (2018); Jin et al. (2020); Yang and Wang (2020); Zanette et al. (2020) show $\sqrt{T}$ regret bounds for *linear MDPs*, which implies that the associated Q-function is linear after a known feature transformation. Wang et al. (2019) extend this model to allow for generalized linear model Q-functions. Unlike the stability notions considered in this work, we are unaware of any algorithmic method of verifying the linear MDP assumption. Furthermore, the aforementioned regret bounds are for the finite-horizon episodic setting; we study the infinite-horizon single trajectory setting without resets.

Very few results categorizing regret bounds for adaptive nonlinear control exist; one recent example is Gaudio et al. (2019), who highlight that simple model reference adaptive controllers obtain constant regret in the continuous-time deterministic setting. In contrast, our work simultaneously tackles the issues of more general models, discrete-time systems, and stochastic noise. We note that several authors have ported various adaptive controllers into discrete-time (Pieper, 1996; Bartolini et al., 1995; Loukianov et al., 2018; Muñoz and Sbarbaro, 2000; Kanellakopoulos, 1994; Ordóñez et al., 2006). These results, however, are mostly concerned with asymptotic stability of the closed-loop system, as opposed to finite-time regret bounds.

3. Problem Statement

In this work, we focus on the following discrete-time[1], time-varying, and nonlinear dynamical system with linearly parameterized unknown in the matched uncertainty setting:

$$x_{t+1} = f(x_t, t) + B(x_t, t)(u_t - Y(x_t, t)\alpha) + w_t \,. \tag{3.1}$$

Here $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^d$, $f : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^n$ is a known nominal dynamics model, $B : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^{n \times d}$ is a known input matrix, $Y : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^{d \times p}$ is a matrix of known basis functions, and $\alpha \in \mathbb{R}^p$ is a vector of unknown parameters. The sequence of noise vectors $\{w_t\} \subseteq \mathbb{R}^n$ is assumed to satisfy the distributional requirements $\mathbb{E}[w_t] = 0$, $\|w_t\| \leqslant W$ almost surely, and that w_s is independent of w_t for all $s \neq t$. We further assume that $\alpha \in \mathcal{C} := \{\alpha \in \mathbb{R}^p : \|\alpha\| \leqslant D\}$, and that an upper bound for D is known. Without loss of generality, we set the origin to be a fixed-point of the nominal dynamics, so that $f(0, t) = 0$ for all t. Because the nominal dynamics is time-varying, this formalism captures the classic setting of nonlinear adaptive control, which considers the problem of tracking a time-varying desired trajectory x_t^d[2].

We study *certainty equivalence* controllers. In particular, we maintain a parameter estimate $\hat{\alpha}_t \in \mathcal{C}$ and play the input $u_t = Y(x_t, t)\hat{\alpha}_t$. Our goal is to design a learning algorithm that updates $\hat{\alpha}_t$ to cancel the unknown and which provides a guarantee of fast convergence to the performance of an ideal *comparator*. The comparator that we will study is an oracle that plays the ideal control $u_t = Y(x_t, t)\alpha$ at every timestep, leading to the dynamics $x_{t+1} = f(x_t, t) + w_t$. To measure the rate of convergence to this comparator, we study the following notion of *control regret*:

$$\mathsf{Regret}(T) := \mathbb{E}_{\{w_t\}} \left[\sum_{t=0}^{T-1} \|x_t^a\|^2 - \|x_t^c\|^2 \right] . \tag{3.2}$$

Here, the trajectory $\{x_t^a\}$ is generated by an adaptive control algorithm, while the trajectory $\{x_t^c\}$ is generated by the oracle with access to the true parameters α. Our notation for x_t^a and x_t^c suppresses the dependence of the trajectory on the noise sequence $\{w_t\}$. Our goal will be to design algorithms that exhibit *sub-linear* regret, i.e., $\mathsf{Regret}(T) = o(T)$, which ensures that the time-averaged regret asymptotically converges to zero. For ease of exposition, in the sequel we define $Y_t := Y(x_t^a, t)$ and $B_t := B(x_t^a, t)$, and we use the symbol $\tilde{\alpha}_t$ to denote the parameter estimation error $\hat{\alpha}_t - \alpha$.

3.1. Parameter Update Algorithms

We study two primary classes of parameter update algorithms inspired by online convex optimization (Hazan, 2016). The first is the family of *velocity gradient algorithms* (Fradkov et al., 1999), which perform online gradient-based optimization on a Lyapunov function for the nominal system. The second obviates the need for a known Lyapunov function, and directly performs online optimization on the least-squares prediction error. Here we discuss the discrete-time formulation, but a self-contained introduction to these algorithms in continuous-time can be found in the full paper.

1. Discrete-time systems may arise as a modeling decision, or due to finite sampling rates for the input, e.g., a continuous-time controller implemented on a computer. In the full paper, we study the latter situation, giving bounds on the rate for which a continuous-time controller must be sampled such that discrete-time closed-loop stability holds.
2. To see this, consider a system $y_{t+1} = g(y_t, t) + B(y_t, t)\,(u_t - Y(y_t, t)\alpha)$ and a desired trajectory y_t^d satisfying $y_{t+1}^d = g(y_t^d, t)$. Define the new variable $x_t := y_t - y_t^d$. Then $x_{t+1} = g(x_t + y_t^d, t) - g(y_t^d, t) + B(x_t + y_t^d, t)\,\big(u_t - Y(x_t + y_t^d, t)\alpha\big)$, so that the nominal dynamics $f(x_t, t) = g(x_t + y_t^d, t) - g(y_t^d, t)$ satisfies $f(0, t) = 0$ for all t. If the original y_t system is non-autonomous, the time-dependent desired trajectory will introduce a time-dependent nominal dynamics in the x_t system.

3.1.1. VELOCITY GRADIENT ALGORITHMS

Velocity gradient algorithms exploit access to a known Lyapunov function for the nominal dynamics. Specifically, assume the existence of a non-negative function $Q(x,t) : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}_{\geqslant 0}$, which is differentiable in its first argument, and a constant $\rho \in (0,1)$ such that for all x, t:

$$Q(f(x,t),t+1) \leqslant Q(x,t) - \rho\|x\|^2 . \tag{3.3}$$

Given such a $Q(x,t)$, velocity gradient methods update the parameters according to the iteration

$$\hat{\alpha}_{t+1} = \Pi_{\mathcal{C}}[\hat{\alpha}_t - \eta_t Y(x_t,t)^\mathsf{T} B(x_t,t)^\mathsf{T} \nabla Q(x_{t+1},t+1)] , \;\; \Pi_{\mathcal{C}}[x] := \arg\min_{y\in\mathcal{C}} \|x-y\| , \tag{3.4}$$

which can alternatively be viewed as projected gradient descent with respect to the parameters after noting that $Y(x_t,t)^\mathsf{T} B(x_t,t)^\mathsf{T} \nabla Q(x_{t+1},t+1) = \nabla_{\hat{\alpha}_t} Q(x_{t+1},t+1)$. As we will demonstrate, the use of $\nabla Q(x_{t+1},t+1)$ instead of $\nabla Q(x_t,t)$ in (3.4) is key to unlocking a sublinear regret bound.

3.1.2. ONLINE LEAST-SQUARES

Online least-squares algorithms are motivated by minimizing the approximation error directly rather than through stability considerations. For each time t, define the prediction error loss function

$$\ell_t(\hat{\alpha}) := \frac{1}{2}\|B(x_t,t)Y(x_t,t)(\hat{\alpha}-\alpha) + w_t\|^2 . \tag{3.5}$$

Unlike in the usual optimization setting, the loss at time t is unknown to the controller, due to its dependence on the unknown parameters α. However, its gradient $\nabla\ell_t(\hat{\alpha}_t)$ can be implemented after observing x_{t+1} through a discrete-time analogue of Luenberger's well-known approach for reduced-order observer design (Luenberger, 1979):

$$\nabla\ell_t(\hat{\alpha}_t) = Y(x_t,t)^\mathsf{T} B(x_t,t)^\mathsf{T}(x_{t+1} - f(x_t,t)) . \tag{3.6}$$

The simplest update rule that uses the gradient $\nabla f_t(\hat{\alpha}_t)$ is online gradient descent:

$$\hat{\alpha}_{t+1} = \Pi_{\mathcal{C}}[\hat{\alpha}_t - \eta_t \nabla f_t(\hat{\alpha}_t)] , \tag{3.7}$$

while a more sophisticated update rule is the online Newton method:

$$\hat{\alpha}_{t+1} = \Pi_{\mathcal{C},t}[\hat{\alpha}_t - \eta A_t^{-1}\nabla f_t(\hat{\alpha}_t)] , \;\; A_t = \lambda I + \sum_{s=0}^{t} M_s^\mathsf{T} M_s , \;\; M_s = B(x_s,s)Y(x_s,s) . \tag{3.8}$$

Above, the operator $\Pi_{\mathcal{C},t}[\cdot]$ denotes projection w.r.t. the A_t-norm: $\Pi_{\mathcal{C},t}[x] := \arg\min_{y\in\mathcal{C}}\|x-y\|_{A_t}$.

4. Regret Bounds for Velocity Gradient Algorithms

In this section, we provide a regret analysis for the velocity gradient algorithm. Here, we will assume a deterministic system, so that $w_t \equiv 0$. Unrolling the Lyapunov stability assumption (3.3) and using the non-negativity of $Q(x,t)$ yields $\sum_{t=0}^{T-1}\|x_t^c\|^2 \leqslant \frac{Q(x_0,0)}{\rho}$, which shows that the contribution of $\sum_{t=0}^{T-1}\|x_t^c\|^2$ to the regret is $O(1)$. Therefore, it suffices to bound $\sum_{t=0}^{T-1}\|x_t^a\|^2$ directly. The key assumption that enables application of the velocity gradient method in discrete-time is strong convexity of the Lyapunov function $Q(x,t)$ with respect to x. Recall that a C^1 function $h(x)$ is μ-strongly convex if for all x and y, $h(y) \geqslant h(x) + \langle\nabla h(x), y-x\rangle + \frac{\mu}{2}\|y-x\|^2$. Our first result is a data-dependent regret bound for the velocity gradient algorithm.

Theorem 1 *Fix a $\lambda > 0$. Consider the velocity gradient update (3.4) with $\hat{\alpha}_0 \in \mathcal{C}$ and learning rate $\eta_t = \frac{D}{\sqrt{\lambda + \sum_{i=0}^{t} \|Y_i^\mathsf{T} B_i^\mathsf{T} \nabla Q(x_{i+1}^a, i+1)\|^2}}$. Assume that the Lyapunov stability condition (3.3) is verified, and that for every t, the map $x \mapsto Q(x,t)$ is μ-strongly convex. Then for any $T \geqslant 1$:*

$$\sum_{t=0}^{T-1} \|x_t^a\|^2 + \frac{\mu}{2\rho} \sum_{t=0}^{T-1} \|B_t Y_t \tilde{\alpha}_t\|^2 \leqslant \frac{Q(x_0, 0)}{\rho} + \frac{5\sqrt{\lambda} D}{\rho} + \frac{3D}{\rho} \sqrt{\sum_{t=0}^{T-1} \|Y_t^\mathsf{T} B_t^\mathsf{T} \nabla Q(x_{t+1}^a, t+1)\|^2} \, .$$

By Theorem 1, a bound on $\sum_{t=0}^{T-1} \|Y_t^\mathsf{T} B_t^\mathsf{T} \nabla Q(x_{t+1}^a, t+1)\|^2$ ensures a bound on the control regret. One way to obtain a bound is to assume that $\|Y_t^\mathsf{T} B_t^\mathsf{T} \nabla Q(x_{t+1}^a, t+1)\| \leqslant G$ for all t, in which case Theorem 1 yields the sublinear guarantee $\mathsf{Regret}(T) \leqslant O(\sqrt{T})$. However, this can be strengthened by assuming that both $\nabla Q(x,t)$ and $f(x,t)$ are Lipschitz continuous.

Theorem 2 *Suppose that for every x and t, $\|\nabla Q(x,t)\| \leqslant L_Q \|x\|$ and $\|f(x,t)\| \leqslant L_f \|x\|$. Further assume that $\sup_{x,t} \|B(x,t)\| \leqslant M$ and $\sup_{x,t} \|Y(x,t)\| \leqslant M$. Then, under the hypotheses of Theorem 1, for any $T \geqslant 1$:*

$$\sum_{t=0}^{T-1} \|x_t^a\|^2 + \frac{\mu}{2\rho} \sum_{t=0}^{T-1} \|B_t Y_t \tilde{\alpha}_t\|^2 \leqslant \frac{3}{2} \left(\frac{Q(x_0, 0)}{\rho} + \frac{5\sqrt{\lambda} D}{\rho} \right) + \frac{27 D^2}{\rho^2} M^4 L_Q^2 \max \left\{ L_f^2, \frac{2\rho}{\mu} \right\} \, .$$

Theorem 2 yields the constant bound $\mathsf{Regret}(T) \leqslant O(1)$, which mirrors an earlier result in the continuous-time deterministic setting due to Gaudio et al. (2019).

5. Regret Bounds for Online Least-Squares Algorithms

In this section we study the use of online least-squares algorithms for adaptive control in the stochastic setting. A core challenge in this setting is that neither $\mathbb{E} \sum_{t=0}^{T-1} \|x_t^a\|^2$ nor $\mathbb{E} \sum_{t=0}^{T-1} \|x_t^c\|^2$ converges to a constant, but rather each grows as $\Omega(T)$. Our approach couples the trajectories $\{x_t^a\}$ and $\{x_t^c\}$ together using the same noise realization $\{w_t\}$, and then utilizes incremental stability to compare trajectories of the comparator and the adaptation algorithm. We first provide a brief introduction to contraction and incremental stability, and then we discuss our results.

5.1. Contraction and Incremental Stability

To prove regret bounds for our least-squares algorithms, we use the following generalization of input-to-state stability, which allows for a direct comparison between two trajectories of the system in terms of the strength of past inputs.

Definition 3 (cf. Angeli (2002)) *Let constants β, γ be positive and $\rho \in (0,1)$. The discrete-time dynamical system $f(x,t)$ is called (β, ρ, γ)-*exponentially-incrementally-input-to-state-stable *(E-δISS) for a pair of initial conditions (x_0, y_0) and signal u_t (which is possibly adapted to the history $\{x_s\}_{s \leqslant t}$) if the trajectories $x_{t+1} = f(x_t, t) + u_t$ and $y_{t+1} = f(y_t, t)$ satisfy for all $t \geqslant 0$:*

$$\|x_t - y_t\| \leqslant \beta \rho^t \|x_0 - y_0\| + \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|u_k\| \, . \tag{5.1}$$

A system is (β, ρ, γ)-E-δISS if it is (β, ρ, γ)-E-δISS for all initial conditions (x_0, y_0) and signals u_t.

Definition 3 can be verified by checking if the system $f(x,t)$ is *contracting*.

Definition 4 (cf. Lohmiller and Slotine (1998)) *The discrete-time dynamical system $f(x,t)$ is contracting with rate $\gamma \in (0,1)$ in the metric $M(x,t)$ if for all x and t:*

$$\frac{\partial f}{\partial x}(x,t)^\mathsf{T} M(f(x,t),t+1)\frac{\partial f}{\partial x}(x,t) \preccurlyeq \gamma M(x,t)\ .$$

We note that contraction, much like Lyapunov stability, can be verified for a particular system using e.g., sum-of-squares programming (Aylward et al., 2008).

Proposition 5 *Let $f(x,t)$ be contracting with rate $\gamma \in (0,1)$ in the metric $M(x,t)$. Assume that for all x,t we have $0 \prec \mu I \preccurlyeq M(x,t) \preccurlyeq LI$. Then $f(x,t)$ is $(\sqrt{L/\mu}, \sqrt{\gamma}, \sqrt{L/\mu})$-E-$\delta$ISS.*

Furthermore, contraction is robust to small perturbations – if the dynamics $f(x,t)$ are contracting, so are the dynamics $f(x,t)+w_t$ for small enough w_t.

Proposition 6 *Let $\{w_t\}$ be a fixed sequence satisfying $\sup_{t\geqslant 0}\|w_t\| \leqslant W$. Suppose that $f(x,t)$ is contracting with rate γ in the metric $M(x,t)$ with $M(x,t) \succcurlyeq \mu I$. Define the perturbed dynamics $g(x,t) := f(x,t)+w_t$. Suppose that for all t, the function $x \mapsto M(x,t)$ is L_M-Lipschitz. Furthermore, suppose that $\sup_{x,t}\|\frac{\partial f}{\partial x}(x,t)\| \leqslant L_f$. Then as long as $W \leqslant \frac{\mu(1-\gamma)}{L_f^2 L_M}$, we have that $g(x,t)$ is contracting with rate $\gamma + \frac{L_f^2 L_M W}{\mu}$ in the metric $M(x,t)$.*

Note that if the metric is state independent (i.e., $M(x,t) = M(t)$), then we can take $L_M = 0$ and hence the perturbed system $g(x,t)$ is contracting at rate γ for all realizations $\{w_t\}$.

5.2. Main Results

Our analysis proceeds by assuming that for almost all noise realizations $\{w_t\}$, the perturbed nominal system $f(x,t)+w_t$ is incrementally stable (E-δISS). We apply incremental stability to bound the control regret directly in terms of the prediction regret, $\mathsf{Regret}(T) \leqslant O(\sqrt{T}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}\|B_t Y_t \tilde{\alpha}_t\|^2})$. Because online convex optimization methods provide explicit guarantees on the prediction regret, we can apply existing results from the online optimization literature to generate a bound on the control regret. To see this, recall that the sequence of prediction error functions $\{\ell_t\}$ from (3.5) has the form $\ell_t(\hat{\alpha}) = \frac{1}{2}\|B_t Y_t(\hat{\alpha}-\alpha)+w_t\|^2$. Hence:

$$\frac{1}{2}\mathbb{E}\sum_{t=0}^{T-1}\|B_t Y_t \tilde{\alpha}_t\|^2 = \mathbb{E}\left[\sum_{t=0}^{T-1}\ell_t(\hat{\alpha}_t)-\ell_t(\alpha)\right] \leqslant \mathbb{E}\left[\sup_{\alpha\in\mathcal{C}}\sum_{t=0}^{T-1}\ell_t(\hat{\alpha}_t)-\ell_t(\alpha)\right]\ .$$

In this section, we make the following assumption regarding the dynamics.

Assumption 7 *The perturbed system $g(x_t,t) := f(x_t,t)+w_t$ is (β,ρ,γ)-E-δISS for all realizations $\{w_t\}$ satisfying $\sup_t\|w_t\| \leqslant W$. Also $\sup_{x,t}\|B(x,t)\| \leqslant M$ and $\sup_{x,t}\|Y(x,t)\| \leqslant M$.*

We define the constant $B_x := \beta\|x_0\| + \frac{\gamma(2DM^2+W)}{1-\rho}$ and $G := M^2(2DM^2+W)$. A key result, which relates control regret to prediction regret, is given in the following theorem.

Theorem 8 *Consider any adaptive update rule $\{\hat{\alpha}_t\}$. Under Assumption 7, for all $T \geqslant 1$:*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|x_t^a\|^2 - \|x_t^c\|^2\right] \leqslant \frac{2B_x\gamma}{1-\rho}\sqrt{T}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}\|B_t Y_t \tilde{\alpha}_t\|^2}\,.$$

We can immediately specialize Theorem 8 to both online gradient descent and online Newton. Both corollaries are a direct consequence of applying well-known regret bounds in online convex optimization to Theorem 8 (cf. Hazan (2016)). Our first corollary shows that online gradient descent achieves a $O(T^{3/4})$ control regret bound.

Corollary 9 *Suppose we use online gradient descent (3.7) to update the parameters, setting the learning rate $\eta_t = \frac{D}{G\sqrt{t+1}}$. Under Assumption 7, for all $T \geqslant 1$:*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|x_t^a\|^2 - \|x_t^c\|^2\right] \leqslant 2\sqrt{6}B_x\frac{\gamma}{1-\rho}\sqrt{GD}T^{3/4}\,.$$

This result immediately generalizes to the case of mirror descent, where dimension-dependence implicit in G and D can be reduced, and where recent implicit regularization results apply (Boffi and Slotine, 2020). Next, the regret can be improved to $O(\sqrt{T\log T})$ by using online Newton.

Corollary 10 *Suppose we use the online Newton method (3.8) to update the parameters, setting $\eta = 1$. Suppose furthermore that $M \geqslant 1$. Under Assumption 7, for all $T \geqslant 1$:*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|x_t^a\|^2 - \|x_t^c\|^2\right] \leqslant \frac{2B_x\gamma}{1-\rho}\sqrt{T}\sqrt{4D^2(\lambda+M^4) + pG^2\log(1+M^4T/\lambda)}\,.$$

We also note that in the deterministic setting, online gradient descent to update the parameters achieves $O(1)$ prediction and control regret, which is consistent with the results in Section 4 and with the results in Gaudio et al. (2019). We give a self-contained proof of this in the full paper.

5.3. Input Delay Results

Motivated by *extended matching* conditions commonly considered in continuous-time adaptive control (Krstić et al., 1995), we now extend our previous results to a setting where the input is time-delayed by k steps. Specifically, we consider the modified system:

$$x_{t+1} = f(x_t, t) + B(x_t, t)(\xi_t - Y(t)\alpha) + w_t\,,\;\; \xi_t = u_{t-k}\,. \tag{5.2}$$

Here, we simplify part of the model (3.1) by assuming that the matrix $Y(t)$ is state-independent. With this simplification, the certainty equivalence controller is given by $u_t = Y(t+k)\hat{\alpha}_t$. The baseline we compare to in the definition of regret is the nominal system $x_{t+1}^c = f(x_t^c, t) + w_t$, which is equivalent to playing the input $u_t = Y(t+k)\alpha$. Note that the gradient $\nabla\ell_t(\hat{\alpha}_t)$ can be implemented by the controller as $\nabla\ell_t(\hat{\alpha}_t) = Y_t^\mathsf{T} B_t^\mathsf{T}(x_{t+1} - f(x_t, t) - B_t(\xi_t - Y_t\hat{\alpha}_t))$.

Folk wisdom and basic intuition suggest that nonlinear adaptive control algorithms for the extended matching setting will perform worse than their matched counterparts; however, standard asymptotic guarantees do not distinguish between the performance of these two classes of algorithms. Here we show that the control regret rigorously captures this gap in performance. We begin with online gradient descent, which provides a regret bound of $O(T^{3/4} + k\sqrt{T})$.

Theorem 11 *Consider the online gradient descent update (3.7) for the k-step delayed system (5.2) with step size $\eta_t = \frac{D}{G\sqrt{t+1}}$. Under Assumption 7 and with state-independent Y_t, for all $T \geqslant k$:*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|x_t^a\|^2 - \|x_t^c\|^2\right] \leqslant kB_x^2 + \frac{2B_x M^2 D\gamma}{(1-\rho)^2} + \frac{2\sqrt{6}B_x\gamma\sqrt{GD}}{1-\rho}T^{3/4} + \frac{4B_x\gamma M^2 D}{1-\rho}k\sqrt{T}\,.$$

Furthermore, the regret improves to $O(k\sqrt{T\log T})$ when we use the online Newton method.

Theorem 12 *Consider the online Newton update (3.8) for the k-step delayed system (5.2) with $\eta = 1$. Suppose $M \geqslant 1$. Under Assumption 7 and with state-independent Y_t, for all $T \geqslant k$:*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|x_t^a\|^2 - \|x_t^c\|^2\right] \leqslant kB_x^2 + \frac{2B_x M^2 D\gamma}{(1-\rho)^2} + \frac{2B_x\gamma Gk}{1-\rho}\sqrt{\frac{pT}{\lambda}\log(1+M^2T/\lambda)}$$
$$+ \frac{2B_x\gamma}{1-\rho}\sqrt{T}\sqrt{4D^2(\lambda+M^4) + pG^2\log(1+M^4T/\lambda)}\,.$$

5.4. Is Incremental Stability Necessary?

The results in this section have crucially relied on incremental input-to-state stability (Definition 3). A natural question to ask is if it possible to relax this assumption to input-to-state stability (Sontag, 2008), while still retaining regret guarantees. In the full paper, we provide a partial answer to this question. Inspired by Rüffer et al. (2013), we show that if a system is exponentially input-to-state stable (which we define similarly to Definition 3, but in reference to a single trajectory), then it is E-δISS on a compact set of initial conditions, but only for certain *admissible* inputs. Next, we prove that under a persistence of excitation condition, the disturbances $\{B_t Y_t \tilde{\alpha}_t\}$ due to parameter mismatch yield an admissible sequence of inputs with high probability. Combining these results, we show a $\sqrt{T}\log T$ regret bound that holds with *constant* probability. We are currently unable to recover a high probability regret bound since the (β, ρ, γ) constants for our E-δISS reduction depend *exponentially* on the original problem constants and the size of the compact set. We leave resolving this issue, in addition to removing the persistence of excitation condition, to future work.

6. Simulations

6.1. Velocity Gradient Adaptation

We consider the cartpole stabilization problem, where we assume the true parameters are unknown. Let q be the cart position, θ the pole angle, and u the force applied to the cart. The dynamics are:

$$\ddot{q} = \frac{u + m_p s_\theta(\ell\dot{\theta}^2 + gc_\theta)}{m_c + m_p s_\theta^2}\,, \ddot{\theta} \quad = \frac{1}{\ell(m_c + m_p s_\theta^2)}\left(-uc_\theta - m_p\ell\dot{\theta}^2 c_\theta s_\theta - (m_c+m_p)gs_\theta\right)\,.$$

Here, $c_\theta = \cos\theta$ and $s_\theta = \sin\theta$. We discretize the dynamics via RK4 with timestep $\Delta t = .01$. The true (unknown) parameters are the cart mass $m_c = 1$g, the pole mass $m_p = 1$g, and pole length $\ell = 1$m. Let the state $x = (q, \dot{q}, \theta, \dot{\theta})$. We solve a discrete-time infinite-horizon LQR problem (with $Q = I_4$ and $R = .5$) for the linearization at $x_{\text{eq}} := (0, 0, \pi, 0)$, using the *wrong parameters* $m_c = .45$g, $m_p = .45$g, $\ell = .8$m. This represents a simplified model of uncertainty in the system or a simulation-to-reality gap. The solution to the discrete-time LQR problem yields a Lyapunov

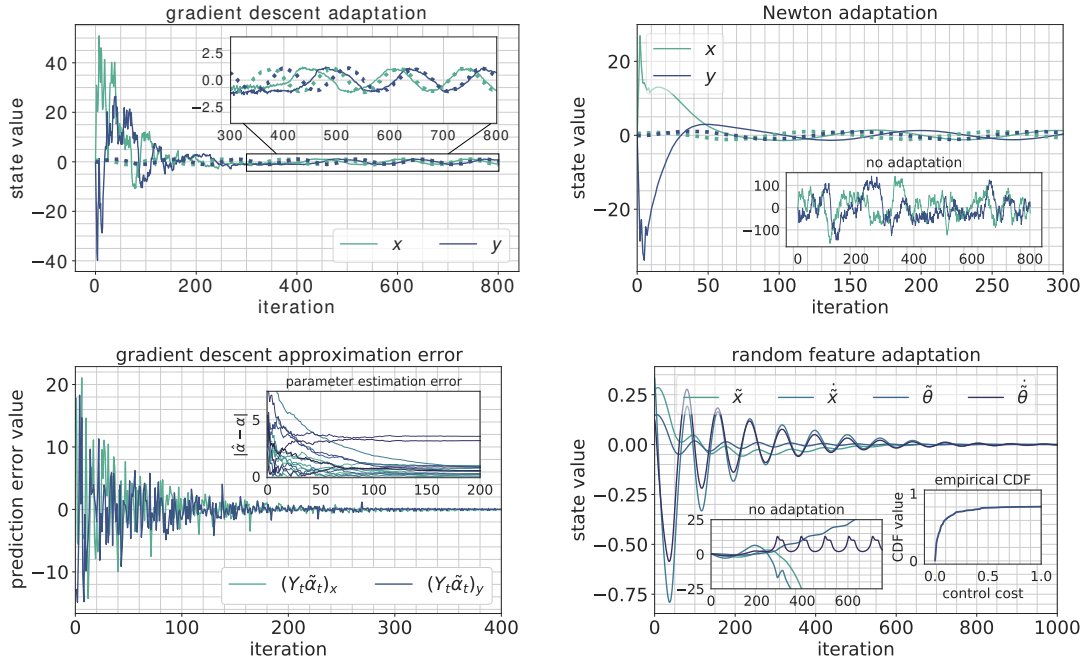

Figure 1: (Top left) Sample trajectory for online gradient descent (solid) and the comparator (dotted). Inset shows a close-up view near convergence. **(Top right)** Sample trajectory for online gradient descent (solid) and the comparator (dotted). Inset shows poor performance of the system *without* adaptation. **(Bottom left)** Prediction error for gradient descent (main figure) and parameter estimation error (inset). The parameters do not converge due to a lack of persistent excitation, but the prediction error still tends to zero. **(Bottom right)** LQR experiment with random features. Main figure shows the performance of one trajectory with adaptation. The right inset shows the empirical CDF of average control performance with adaptation. The left inset shows divergent behavior of one trajectory without adaptation.

function $Q(x) = \frac{1}{2}(x - x_{\mathrm{eq}})^{\mathsf{T}} P (x - x_{\mathrm{eq}})$, and a control law $u_t = -K(x_t - x_{\mathrm{eq}})$ that would locally stabilize the system around x_{eq} if the parameters were correct.

We use adaptive control to bootstrap our control policy computed with incorrect parameters to a stabilizing law for the true system. Specifically, we run the velocity gradient adaptive law (3.4) on the LQR Lyapunov function $Q(x)$ with basis functions $Y(x,t) \in \mathbb{R}^{1\times 400}$ given by random Gaussian features $\cos(\omega^{\mathsf{T}} x + b)$ with $\omega \sim N(0,1)$ and $b \sim \mathsf{Unif}(0, 2\pi)$ (cf. Rahimi and Recht (2007)). Note that $Q(x)$ is only an approximation to the true Lyapunov function (due to model-misspecification). We rollout 500 trajectories initialized uniformly at random in an ℓ_∞ ball of radius $\frac{1}{2}$ around x_{eq}, and measure the performance of the system both with and without adaptation through the average control regret $\frac{1}{T}\sum_{t=1}^{T} \|x_t - x_{\mathrm{eq}}\|^2$. The results are shown in the bottom-right pane of Figure 1. Without adaptation, every trajectory diverges, and an example is shown in the left inset. On the other hand, adaptation is often able to successfully stabilize the system. One example trajectory with adaptation is shown in the body of the pane. The right inset shows the empirical CDF of the average control cost with adaptation, indicating that $\sim 60\%$ of trajectories with adaptation have an average control regret less than 0.1, and $\sim 80\%$ less than 1. More generally, our approach of improving the quality of a controller through online adaptation with expressive, unstructured basis functions could be used as an additional layer on top of existing adaptive control algorithms to correct for errors in the structured, physical basis functions originating from the dynamics model.

6.2. Online Convex Optimization Adaptation

To demonstrate the applicability of our OCO-inspired discrete-time adaptation laws, we study the following discrete-time nonlinear system

$$\begin{aligned} x_{t+1} &= x_t + \tau\left(-y_t + \frac{x_t}{\sqrt{x_t^2+y_t^2}} - x_t + Y_x(x_t,t)^\mathsf{T}\tilde{\alpha}_t\right) + \sqrt{\tau}\sigma w_{t,1}\,, \\ y_{t+1} &= y_t + \tau\left(x_t + \frac{y_t}{\sqrt{x_t^2+y_t^2}} - y_t + Y_y(y_t,t)^\mathsf{T}\tilde{\alpha}_t\right) + \sqrt{\tau}\sigma w_{t,2} \end{aligned} \tag{6.1}$$

for $\tau = 0.05$, $\sigma = 0.1$, and $w_{t,i} \sim N(0,1)$. The nominal system for (6.1) is a forward-Euler discretization of the continuous-time system $\dot{x} = -y + \frac{x}{\sqrt{x^2+y^2}} - x$, $\dot{y} = x + \frac{y}{\sqrt{x^2+y^2}} - y$. In polar coordinates, the nominal system reads $\dot{r} = -(r-1)$, $\dot{\theta} = 1$, which is contracting in the Euclidean metric towards the limit cycle $\dot{\theta} = 1$ on the unit circle. This shows that the system in Euclidean coordinates is contracting in the radial direction in the metric $M(x,y) = \frac{\partial g}{\partial x}(x,y)^\mathsf{T}\frac{\partial g}{\partial x}(x,y)$, where g is the nonlinear mapping $(x,y) \mapsto (r,\theta)$. The basis functions are taken to be $Y_z(z_t,t)^\mathsf{T} = \sin(\omega(z_t + \sin(t)))$ where $z \in \{x,y\}$, the outer sin is taken element-wise, and $\omega \in \mathbb{R}^p$ is a vector of frequencies sampled uniformly between 0 and 2π. The estimated parameters $\hat{\alpha}_t$ are updated according to the OCO-inspired adaptive laws (3.7) or (3.8) analyzed in Section 5.2.

Results are shown in Figure 1. In the top-left pane, convergence of a sample trajectory towards the limit cycle is shown for gradient descent in solid, with the limit cycle itself plotted in dots. The inset displays a close-up view of convergence. In the top-right pane, convergence is shown for the online Newton method, which converges significantly faster and has a smoother trajectory than gradient descent. The inset displays a failure to converge without adaptation, demonstrating improved performance of the two adaptation algorithms in comparison to the system without adaptation. The bottom-left pane shows convergence of the two components of the prediction error $Y_t\tilde{\alpha}_t$ for gradient descent in the main figure, and shows parameter error trajectories in the inset. Note that the parameters do not converge to the true values due to a lack of persistent excitation.

7. Conclusion and Future Work

We present the first finite-time regret bounds for nonlinear adaptive control in discrete-time. Our work opens up many future directions of research. One direction is the possibility of logarithmic regret in our setting, given that it is achievable in various LQR problems (Agarwal et al., 2019b; Cassel et al., 2020; Foster and Simchowitz, 2020). A second question is handling state-dependent $Y(x,t)$ matrices in the k timestep delay setting, or more broadly, studying the extended matching conditions of Kanellakopoulos et al. (1989); Krstić et al. (1995) for which timestep delays are a special case. Another direction concerns proving regret bounds for the velocity gradient algorithm in a stochastic setting. Furthermore, in the spirit of Agarwal et al. (2019a); Hazan et al. (2020), an extension of our analysis to handle more general cost functions and adversarial noise sequences would be quite impactful. Finally, understanding if sublinear regret guarantees are possible for a non-exponentially incrementally stable system would be interesting.

Acknowledgments

The authors thank Naman Agarwal, Vikas Sindhwani, and Sumeet Singh for helpful feedback.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Conference on Learning Theory*, 2011.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Neural Information Processing Systems*, 2019b.

Naman Agarwal, Nataly Brukhim, Elad Hazan, and Zhou Lu. Boosting for control of dynamical systems. In *International Conference on Machine Learning*, 2020.

David Angeli. A lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3):410–421, 2002.

Erin M. Aylward, Pablo A. Parrilo, and Jean-Jacques E. Slotine. Stability and robustness analysis of nonlinear systems via contraction metrics and sos programming. *Automatica*, 44(8):2163–2170, 2008.

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, 2018.

Giorgio Bartolini, Antonella Ferrara, and Vadim I. Utkin. Adaptive sliding mode control in discrete-time systems. *Automatica*, 31(5):769–773, 1995.

Nicholas M. Boffi and Jean-Jacques E. Slotine. Implicit regularization and momentum algorithms in nonlinear adaptive control and prediction. *arXiv:1912.13154*, 2020.

Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. *arXiv:2011.13101*, 2020.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.

Tongyi Cao and Akshay Krishnamurthy. Provably adaptive reinforcement learning in metric spaces. *arXiv:2006.10875*, 2020.

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

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

Dylan J. Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, 2020.

Alexander L. Fradkov, Iliya V. Miroshnik, and Vladimir O. Nikiforov. *Nonlinear and Adaptive Control of Complex Systems*. 1999.

Joseph E. Gaudio, Travis E. Gibson, Anuradha M. Annaswamy, Michael A. Bolender, and Eugene Lavretsky. Connections between adaptive control and optimization in machine learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan, Sham M. Kakade, and Karan Singh. The nonstochastic control problem. In *31st International Conference on Algorithmic Learning Theory*, 2020.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

Petros A. Ioannou and Jing Sun. *Robust Adaptive Control*. 1996.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Neural Information Processing Systems*, 2020.

Ioannis Kanellakopoulos. A discrete-time adaptive nonlinear system. *IEEE Transactions on Automatic Control*, 39(11):2362–2365, 1994.

Ioannis Kanellakopoulos, Petar V. Kokotovic, and Riccardo Marino. Robustness of adaptive nonlinear control under an extended matching condition. *IFAC Proceedings Volumes*, 22(3):245–250, 1989.

Miroslav Krstić, Ioannis Kanellakopoulos, and Petar Kokotović. *Nonlinear and Adaptive Control Design*. 1995.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Winfried Lohmiller and Jean-Jacques E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

Alexander G. Loukianov, Antonio Navarrete-Guzmán, and Jorge Rivera. Adaptive discrete time sliding mode control for a class of nonlinear systems. In *2018 15th International Workshop on Variable Structure Systems (VSS)*, 2018.

David G. Luenberger. *Introduction to Dynamic Systems*. 1979.

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

Horia Mania, Michael I. Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv:2006.10277*, 2020.

David Muñoz and Daniel Sbarbaro. An adaptive sliding-mode controller for discrete nonlinear systems. *IEEE Transactions on Industrial Electronics*, 47(3):574–581, 2000.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand. *arXiv:1910.07113*, 2019.

Raúl Ordóñez, Jeffrey T. Spooner, and Kevin M. Passino. Experimental studies in nonlinear discrete-time adaptive prediction and control. *IEEE Transactions on Fuzzy Systems*, 14(2):275–286, 2006.

Jeff K. Pieper. A discrete time adaptive sliding mode controller. *IFAC Proceedings Volumes*, 29(1): 5227–5231, 1996.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machine. In *Neural Information Processing Systems*, 2007.

Björn S. Rüffer, Nathan van de Wouw, and Markus Mueller. Convergent systems vs. incremental stability. *Systems & Control Letters*, 62(3):277–285, 2013.

Max Simchowitz and Dylan J. Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, 2020.

Jean-Jacques E. Slotine and J. A. Coetsee. Adaptive sliding controller synthesis for non-linear systems. *International Journal of Control*, 43(6):1631–1651, 1986.

Jean-Jacques E. Slotine and Weiping Li. *Applied Nonlinear Control*. 1991.

Eduardo D. Sontag. *Input to State Stability: Basic Concepts and Results*, pages 163–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Yining Wang, Ruosong Wang, Simon S. Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv:1912.04136*, 2019.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M. Rehg, Byron Boots, and Evangelos A. Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

Lin F. Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 2020.

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*, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.