Proceedings of Machine Learning Research vol 120:1–11, 2020 2nd Annual Conference on Learning for Dynamics and Control

Learning nonlinear dynamical systems from a single trajectory

Dylan J. Foster DYLANF@MIT.EDU
Alexander Rakhlin RAKHLIN@MIT.EDU
Tuhin Sarkar TSARKAR@MIT.EDU
Massachusetts Institute of Technology

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M.Zeilinger

Abstract

We introduce algorithms for learning nonlinear dynamical systems of the form $x_{t+1} = \sigma(\Theta^\star x_t) + \varepsilon_t$, where $\Theta^\star$ is a weight matrix, σ is a nonlinear link function, and ε_t is a mean-zero noise process. We give an algorithm that recovers the weight matrix $\Theta^\star$ from a single trajectory with optimal sample complexity and linear running time. The algorithm succeeds under weaker statistical assumptions than in previous work, and in particular i) does not require a bound on the spectral norm of the weight matrix $\Theta^\star$ (rather, it depends on a generalization of the spectral radius) and ii) enjoys guarantees for non-strictly-increasing link functions such as the ReLU. Our analysis has two key components: i) we give a general recipe whereby global stability for nonlinear dynamical systems can be used to certify that the state-vector covariance is well-conditioned, and ii) using these tools, we extend well-known algorithms for efficiently learning generalized linear models to the dependent setting.[1]

1. Introduction

We consider nonlinear dynamical systems of the form

$$x_{i+1} = f^\star(x_i) + \varepsilon_i, \tag{1}$$

where $f^\star : \mathbb{R}^d \to \mathbb{R}^d$ is an unknown function, $\{\varepsilon_i\}_{i=0}^n$ is an independent, mean-zero noise process in $\mathbb{R}^d$ and $x_0 = \mathbf{0}$. Dynamical systems are ubiquitous in applied mathematics, engineering, and computer science, with applications including control systems, time series analysis, econometrics, and natural language processing. The recent success of deep reinforcement learning (Mnih et al., 2015; Silver et al., 2017; Lillicrap et al., 2016) has led to renewed interest in developing efficient algorithms for learning complex nonlinear systems such as (1) from data.

In this paper, we focus on the task of estimating the dynamics $f^\star$ given a single trajectory $\{x_i\}_{i=1}^{n+1}$, where $f^\star$ belongs to a known function class $\mathcal{F}$. We focus on the following questions:

- What is the sample complexity of recovering the dynamics $f^\star$? How is it determined by $\mathcal{F}$?
- What algorithmic principles enable computationally efficient recovery of the dynamics?

1. The full version of this paper is available at `https://arxiv.org/abs/2004.14681`.

For linear dynamical systems where $f^\star(x) = \Theta^\star x$, subroutines for efficiently estimating dynamics from data form a core building block of *certainty-equivalent* control, which enjoys optimal sample complexity guarantees for this simple setting (Mania et al., 2019; Simchowitz and Foster, 2020). While linear dynamical systems have been the subject of intense recent interest (Dean et al., 2019; Hazan et al., 2017; Tu and Recht, 2018; Hazan et al., 2018; Simchowitz et al., 2018; Sarkar and Rakhlin, 2019; Simchowitz et al., 2019; Mania et al., 2019; Sarkar et al., 2019), nonlinear dynamical systems are comparatively poorly understood.

1.1. On the performance of least squares

On the algorithmic side, a natural starting point for learning the system (1) is the least squares estimator

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \|f(x_i) - x_{i+1}\|^2, \tag{2}$$

where $\|\cdot\|$ denotes the entrywise ℓ_2 norm. A basic observation is that the in-sample prediction error (or, denoising error) of this estimator is bounded by the so-called *offset Rademacher complexity* introduced by Rakhlin and Sridharan (2014); Liang et al. (2015):

Proposition 1 *The least-squares estimator (2) guarantees*

$$\mathbb{E}_\varepsilon\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{f}_n(x_i) - f^\star(x_i)\right\|^2\right] \le \mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}}\left[\frac{1}{n}\sum_{i=1}^{n} 4\langle \varepsilon_i, g(x_i)\rangle - \|g(x_i)\|^2\right] =: \mathfrak{R}_n^\circ(\mathcal{G}), \tag{3}$$

where $\mathcal{G} = \mathcal{F} - f^\star$ and $\mathbb{E}_\varepsilon$ denotes expectation with respect to $\{\varepsilon_i\}_{i=1}^n$.[2]

The proof is a simple consequence of the basic inequality for least squares (van de Geer, 2000). The offset Rademacher process captures the notion of localization/self-normalization (Bartlett et al., 2005; Koltchinskii, 2006; de la Peña et al., 2008): The negative quadratic term penalizes fluctuations from the term involving the random variables $\{\varepsilon_i\}_{i=1}^n$, leading to fast rates for prediction error. In particular, if ε_i has subgaussian parameter τ^2 and $f^\star(x) = \Theta^\star x$ is a familiar linear dynamical sytem, we have $\mathfrak{R}_n^\circ(\mathcal{G}^{\text{linear}}) \lesssim \tau^2 \cdot \frac{d^2}{n}$. The utility of this approach, however, lies in the fact that it easily extends beyond the linear setting. For example, if $\mathcal{G}$ consists of a class of *generalized linear* dynamical systems of the form

$$x_{i+1} = \sigma(\Theta^\star x_i) + \varepsilon_i, \tag{4}$$

where $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is a 1-Lipschitz link function, we enjoy a similar guarantee: $\mathfrak{R}_n^\circ(\mathcal{G}^{\text{glm}}) \lesssim \tau^2 \cdot \frac{d^2}{n}$. More generally, even though (3) has a complex dependent structure (the variables $\varepsilon_1, \dots, \varepsilon_n$ determine the evolution of $x_1, \dots, x_n$ via (1)), it is possible to bound the value for general function classes $\mathcal{F}$ such as neural networks, kernels, decision trees using sequential covering numbers and chaining techniques introduced in Rakhlin et al. (2014); Rakhlin and Sridharan (2014). However, there are number of important questions that remain if one wishes to use this type of learning guarantee for real-world control applications.

2. Note that x_i is measurable with respect to the σ-algebra $\sigma(\varepsilon_1, \dots, \varepsilon_{i-1})$.

- *Efficient algorithms.* Even for simple nonlinear systems such as the generalized linear model (4), computing the least-squares estimator (2) may be computationally intractable in the worst case. For what classes of interest can we obtain algorithms that are both computationally efficient *and* sample-efficient?

- *Out-of-sample performance.* The prediction error guarantee (3) only concerns performance on the realized sequence $\{x_t\}_{t=1}^n$. For control applications such as certainty-equivalent control, it is essential to bound the performance of the estimator $\widehat{f}_n$ on counterfactual sequences in which the data generating process is $x_{t+1} = \widehat{f}_n(x_t) + \varepsilon_t$ (i.e., error in simulation). For linear and generalized linear systems, a sufficient condition for such a guarantee is to recover the weight matrix $\Theta^\star$ in parameter norm. Under what conditions on the data generating process can we obtain such guarantees?

1.2. Contributions.

We provide a new efficient algorithm for recovery of *generalized linear systems* (4). Our algorithm runs in nearly-linear time and obtains optimal $O\left(\sqrt{d^2/n}\right)$ sample complexity for recovery in Frobenius norm. Conceptually, our key technical observations are as follows:

- We provide a general recipe based on Lyapunov functions for proving that data remains well-conditioned/nearly isometric for stable dynamical systems, without assuming linearity.

- We show that efficient algorithms for learning generalized linear models in the i.i.d. setting (Kalai and Sastry, 2009; Kakade et al., 2011) cleanly port to the dependent setting. Here the key insight is that the empirical counterparts of simple non-convex losses arising from generalized linear models remain well-behaved even under dependent data.

Our algorithm improves prior work on two fronts: First, we do not require a bound on the spectral norm of $\Theta^\star$, and instead require a bound on a parameter that generalizes the notion of the spectral radius to the nonlinear setting. Second, we can recover $\Theta^\star$ even when the link function σ is the ReLU, eschewing invertibility assumptions from previous results.

Related work. Our results are closely related to recent work of Oymak (2019); Bahmani and Romberg (2019); Sattar and Oymak (2020). We refer the reader to the full version of the paper for a detailed comparison and further discussion of related work.

Notation. Throughout this paper we use $c > 0$, $c' > 0$, and $c'' > 0$ to denote absolute numerical constants whose value may vary depending on context. We use non-asymptotic big-oh notation, writing $f = \mathcal{O}(g)$ if there exists a numerical constant such that $f(x) \le c \cdot g(x)$ and $f = \tilde{\mathcal{O}}(g)$ if $f \le c \cdot g \max\{\log g, 1\}$. We say a random vector $v \in \mathbb{R}^d$ is subgaussian with variance proxy τ^2 if $\sup_{\|\theta\|=1} \sup_{p \ge 1} \left\{ p^{-1/2} \left(\mathbb{E}[|\langle v, \theta \rangle|^p] \right)^{1/p} \right\} = \tau$ and $\mathbb{E}[v] = \mathbf{0}$, and we denote this by $v \sim \mathsf{subG}(\tau^2)$. We let $\|\cdot\|_{\mathsf{op}}$ denote the spectral norm and $\|\cdot\|_F$ denote the Frobenius norm. For a convex set X, we let $\mathrm{Proj}_X(\cdot)$ denote euclidean projection onto the set. Unless otherwise stated, all dynamical systems considered in this paper are assumed to start from $x_0 = \mathbf{0}$.

2. Stability, Lower Isometry, and Recovery

Well-conditioned data plays a fundamental role in statistical estimation. For linear regression, it is well-known that the minimax rates for parameter recovery are governed by the spectrum of empirical design matrix $X_n \in \mathbb{R}^{n\times d}$ formed by stacking $x_1, \ldots, x_n$ as rows (Hastie et al., 2015; Wainwright, 2019)). In particular, letting $\widehat{\Sigma}_n = X_n^\top X_n$ denote the empirical covariance, a sufficient condition for recovery is the *lower isometry* property $\widehat{\Sigma}_n \succeq \frac{1}{2} I$; See Lecué and Mendelson (2018) for a contemporary discussion. In this section, we develop tools for proving lower isometry guarantees for nonlinear dynamical systems such as (1). To begin, we make a mild assumption on the noise process.

Assumption 1 *The noise variables $\{\varepsilon_t\}_{t=1}^n$ are independent. Each increment is isotropic (zero-mean, with $\mathbb{E}\left[\varepsilon_i \varepsilon_i^\top\right] = I$) and satisfies $\varepsilon_t \sim \mathsf{subG}(\tau^2)$.*[3]

While Assumption 1 ensures that each increment ε_i is well-behaved, it is not clear a-priori whether the empirical design matrix should enjoy favorable conditioning—indeed, the observations $\{x_i\}_{i=1}^{n+1}$ evolve from the noise process in a complex dependent fashion. In general, the behavior of the empirical design matrix will heavily depend on the system $f^\star$. Here we show that classical results in control theory on *exponential stability* of the system $f^\star$ provide sufficient conditions for both upper and lower control of the spectrum of the empirical design matrix. While our guarantees apply to the noisy system (1), our assumptions depend on the behavior of the system in absence of noise:

$$x_{t+1} = f(x_t), \tag{5}$$

where $x_0 = \mathbf{0}$.

Definition 2 (Global exponential stability) *A noiseless system (5) given by map $f : \mathbb{R}^d \to \mathbb{R}^d$ is globally exponentially stable (g.e.s.) with respect to a norm $\|\cdot\|_\square$ if there exist constants $C_f > 0$ and $\rho_f < 1$ depending only on f such that for all $k \geq 1$,*

$$\|f^k\|_{\mathsf{op}(\square)} \leq C_f \rho_f^k, \tag{6}$$

where $\|f\|_{\mathsf{op}(\square)} := \sup_{x\in\mathbb{R}^d} \frac{\|f(x)\|_\square}{\|x\|_\square}$.[4]

In this paper, we focus on systems where $f^\star$ satisfies the g.e.s. property, and where this is certified by a quadratic Lyapunov function.

Definition 3 *A map $f : \mathbb{R}^d \to \mathbb{R}^d$ is (K, ρ)-g.e.s. if there exists a matrix $K \succ 0$ and constant $0 \leq \rho < 1$ such that for all $x \in \mathbb{R}^d$,*

$$\|f(x)\|_K^2 \leq \rho \cdot \|x\|_K^2, \tag{7}$$

where $\|x\|_K := \sqrt{\langle x, Kx\rangle}$.

3. Our results extend to general covariance Σ under the standard assumption that $\Sigma^{-1}\varepsilon_t$ is subgaussian.
4. When $f : \mathbb{R}^d \to \mathbb{R}^d$, we let f^k denote the k-times composition of f, i.e. $f^k = \underbrace{f \circ \cdots \circ f}_{k \text{ times}}$.

Any (K,ρ)-g.e.s. map satisfies (6) with $\|\cdot\|_\square = \|\cdot\|_K$, $C_f = 1$, and $\rho_f = \rho^{1/2}$. The equation (7) is homogeneous under rescaling, and consequently we will assume without loss of generality that $K \succeq I$ for the remainder of the paper.

In general, finding certificates of stability for nonlinear dynamical systems is a difficult problem. Providing necessary and sufficient conditions for stability for rich classes of nonlinear dynamical systems remains an active area of research, with most development proceeding on a fairly case-by-case basis. We develop a general reduction from lower and upper isometry to (K,ρ)-stability, which allows us to leverage developments in control in a black-box fashion as opposed to having to prove concentration results case-by-case. Our main result here is Theorem 4, which shows that any (K,ρ)-g.e.s. system enjoys both upper and lower isometry.

Theorem 4 *Consider the noisy system (1), and let noise process satisfy Assumption 1. Suppose the map $f^\star$ satisfies the (K,ρ)-g.e.s. property Definition 3 in the absence of noise. Then for any $\delta > 0$, once $n \geq cd \cdot \frac{\tau^4}{(1-\rho)^2} \log(R_{K,\rho}/\delta + 1)$, with probability at least $1-\delta$ the iterates $\{x_i\}_{i=1}^n$ of the noisy system satisfy*

$$\frac{1}{4} \cdot I \preceq \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \preceq 4R_{K,\rho} \cdot I, \tag{8}$$

where $R_{K,\rho} := \frac{\mathsf{tr}(K)}{1-\rho}$ is the effective radius of the system and $c > 0$ is an absolute constant.

The key feature of Theorem 4 is that we only need to assume the (K,ρ)-g.e.s. property on the map $f^\star$ in the absence of noise, yet the theorem gives a guarantee on the trajectory generated by the noisy system (1) as long as Assumption 1 is satisfied.

The proof has three parts, each of which relies on the machinery of self-normalization. We first use the structure of the dynamics (1) to show that the lower isometry in (8) holds as soon as we have a weak *upper* bound on the covariance of the form $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \preceq \frac{B}{\delta} \cdot I$, where δ is the failure probability. We then show that in (K,ρ)-g.e.s. systems, this condition is satisfied with $B = R_{K,\rho}$. Finally, the strong upper bound in (8) is attained by using a self-normalized inequality to boost the weak upper bound and remove the $1/\delta$ factor.

2.1. Lower isometry for generalized linear systems

We now provide sufficient conditions under which the generalized linear systems that are the focus of our main learning results satisfy the g.e.s. property. We make the following mild regularity assumption on the link function.

Assumption 2 *The link function $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ has the form $\sigma(x) := (\sigma_1(x_1), \cdots, \sigma_d(x_d))$, where each coordinate function $\sigma_i : \mathbb{R} \to \mathbb{R}$ is non-decreasing, 1-Lipschitz, and satisfies $\sigma_i(0) = 0$.*

With this assumption, the following constrained Lyapunov equation provides a sufficient condition under which the generalized linear system satisfies the g.e.s. property.

Proposition 5 *Suppose there exists a* diagonal *matrix $K \succ 0$ and scalar $\rho < 1$ such that*

$$\Theta^\top K \Theta \preceq \rho \cdot K. \tag{9}$$

Then the map $f = \sigma \circ \Theta$ is (K,ρ)-g.e.s. whenever σ satisfies Assumption 2.

Proposition 5 can be used to invoke Theorem 4 for any generalized linear system of the form $f^\star = \sigma \circ \Theta$. Thus, we can ensure lower and upper isometry hold for generalized linear systems whenever their stability is certified the Lyapunov condition (9).

The equation (9) strengthens the usual Lyapunov condition for linear systems by adding the additional constraint that K is diagonal. This condition is stronger than the classical spectral radius condition that $\rho(\Theta) < 1$, but it can easily be seen that some type of strengthening is necessary, as the classical condition is not sufficient for nonlinear systems. For example, the matrix $\Theta = \left(\begin{smallmatrix} 1 & 1 \\ -1 & -1 \end{smallmatrix}\right)$ has $\rho(\Theta) = 0$, but the map $x \mapsto \mathsf{relu}(\Theta x)$ is not g.e.s.—indeed, we have $\mathsf{relu}(\Theta e_1) = e_1$, where e_1 is the first standard basis vector. A sufficient condition for Proposition 5 is that Θ has spectral norm bounded by unity, but the condition (9) is a strictly weaker than this assumption. Further sufficient conditions include: 1) $\rho(\Theta) < 1$ and Θ has non-negative entries (Rantzer, 2011, Proposition 2), and 2) $\rho(|\Theta|) < 1$, where $|\Theta|$ denotes element-wise absolute value operator. We refer to the full version of this paper for additional discussion.

3. Algorithms for generalized linear dynamical systems

We now leverage the isometry results of Section 2 to develop efficient algorithms with parameter recovery guarantees for generalized linear systems. Following Section 2, we make the following assumption on the generalized linear system.

Assumption 3 *The system (1) is generalized linear ($f^\star = \sigma \circ \Theta^\star$) and is (K, ρ)-g.e.s. in the sense of Proposition 5. Furthermore, $\|\Theta^\star\|_F \le W$, where W is known to the learner.*[5]

Background: Learning generalized linear models Our algorithm for learning generalized linear dynamical systems builds on developments for learning generalized linear models in statistical learning. Consider the simpler setting where we receive $\{(x_i, y_i)\}_{i=1}^n$ i.i.d., where $y = \sigma(\langle \theta^\star, x \rangle) + \varepsilon$ and $\mathbb{E}[\varepsilon \mid x] = 0$. For this setting the population loss $L(\theta) := \mathbb{E}_{x,y}\big[(\sigma(\langle \theta, x \rangle) - y)^2\big]$ is not convex. However, if the link function σ is strictly increasing the and the population covariance $\mathbb{E}[xx^\top]$ is well-conditioned, the loss satisfies a gradient-dominance type property, and gradient descent on the empirical loss will converge to $\theta^\star$ given sufficiently many samples (Mei et al., 2018). To provide guarantees even when σ is not strictly increasing, we opt to use a variant of the GLMtron algorithm introduced by Kakade et al. (2011). The GLMtron algorithm performs gradient descent using a "pseudogradient" for the empirical loss in which the derivative of σ is simply dropped: $\theta^{(t+1)} = \theta^{(t)} - \frac{1}{n}\sum_{i=1}^n \big(\sigma(\langle \theta^{(t)}, x_i \rangle) - y_i\big) x_i$. Following the pseudogradient allows the algorithm to efficiently provide prediction error guarantees even when σ is not strictly increasing, and this is the starting point for our approach.

3.1. Algorithm and guarantees

Algorithm 1 is a natural extension of GLMtron to handle the vector-valued target variables and matrix-valued parameters that arise in our dynamical system setting.

Algorithm 1 is closely related to projected gradient descent on the empirical square loss $\widehat{L}(\Theta, X_{n+1}) := \frac{1}{n}\sum_{t=1}^n \|\sigma(\Theta x_t) - x_{t+1}\|^2$, but rather than following the gradient, the algorithm

5. There is no restriction on the range of the parameter W, but some of our sample complexity guarantees depend on it polynomially.

Algorithm 1 Parameter estimation for generalized linear systems

1: **input**: Single trajectory: $X_{n+1} = \{x_i\}_{i=1}^{n+1}$, Learning rate schedule: η_t.
2: **initialize**: $\widehat{\Theta}^{(1)} = \mathbf{0}_{d\times d}$.
3: Define $\mathcal{M} = \left\{\Theta \in \mathbb{R}^{d\times d} \mid \|\Theta\|_F \le W\right\}$.
4: **for** $t = 1, \dots, m$ **do**
5: $\quad \widehat{\Theta}^{(t+1)} = \mathrm{Proj}_{\mathcal{M}}(\widehat{\Theta}^{(t)} - \eta_t \widehat{G}(\widehat{\Theta}^{(t)}, X_{n+1}))$,
$\quad\quad$ where $\widehat{G}(\Theta^{(t)}, X_{n+1}) := \frac{1}{n}\sum_{i=1}^{n}(\sigma(\Theta^{(t)}x_i) - x_{i+1})x_i^\top$.
6: **return**: $\widehat{\Theta} = \widehat{\Theta}^{(m)}$ (Option I), or $\widehat{\Theta} = \widehat{\Theta}^{(t)}$ with $t \in [m]$ uniform (Option II).

follows the pseudogradient $\widehat{G}$ attained by dropping the link derivative σ' from the gradient (Line 5). This modification allows for prediction guarantees without assuming a lower bound on the link function derivative, and allows for weaker dependence on the derivative lower bound for parameter recovery guarantees. In particular, we show that the algorithm obtains the best of both worlds in a certain sense. First, with only the assumption that the link function is Lipschitz, the algorithm ensures that iterates have low prediction error on average. Consequently, if we select an iterate uniformly at random (Option II in Algorithm 1), the iterate will have low prediction error (a "slow rate" of type $1/\sqrt{n}$) in expectation. On the other hand, suppose the following assumption holds.

Assumption 4 *There exists a constant $\zeta > 0$ such that for all i, $|\sigma_i(x) - \sigma_i(y)| \ge \zeta|x - y|$ for all $x, y \in \mathbb{R}$.*

In this case, the algorithm enjoys linear convergence, and taking the last iterate (Option I in Algorithm 1) leads to a "fast" $1/n$-type rate for prediction error, as well as a parameter recovery guarantee.

To state the performance guarantee, we let $\mathcal{E}(\Theta) := \frac{1}{n}\sum_{i=1}^{n}\|(\sigma(\Theta x_i) - \sigma(\Theta^\star x_i))\|^2$ denote the in-sample prediction error, and let $\mathbb{E}_{\mathcal{A}}$ denote expectation with respect to the algorithm's internal randomness (uniform selection of the iterate returned in Line 6 under Option II).

Theorem 6 *Let $\delta > 0$ be fixed and let Assumptions 1-3 hold. Whenever $n \ge \frac{c\tau^4 d}{1-\rho}\log(R_{K,\rho}/\delta + 1)$, Algorithm 1 enjoys the following guarantees:*

1. ***Slow rate.*** *If $\eta_t = \frac{1}{16R_{K,\rho}}$ and $m \ge C_0 \cdot \sqrt{n}$, then with probability at least $1 - \delta$, Algorithm 1 with Option II has*

$$\mathbb{E}_{\mathcal{A}}\left[\mathcal{E}(\widehat{\Theta})\right] \le C_1 \cdot \sqrt{\frac{d^2}{n}\log(4R_{K,\rho}/\delta + 1)}, \tag{10}$$

where $C_0 \le c \cdot WB(\tau^2 d R_{K,\rho}\log(4R_{K,\rho}/\delta + 1))^{-1/2}$ and $C_1 \le c\tau W\sqrt{\frac{\sigma_{\max}(K)}{1-\rho}}$.

2. ***Fast rate.*** *Suppose that Assumption 4 holds in addition to Assumptions 1-3. If $\eta_t = \frac{\zeta^2}{(16R_{K,\rho})^2}$ and $m \ge C_2 \cdot \log\left(1 + \frac{nW^2B^2}{\tau^2 R_{K,\rho}}\right)$, then with probability at least $1-\delta$, Algorithm 1 with Option I has*

$$\mathcal{E}(\widehat{\Theta}) \le C_3 \cdot \frac{d^2}{n}\log\left(4R_{K,\rho}/\delta + 1\right), \quad \text{and} \quad \|\widehat{\Theta} - \Theta^\star\|_F^2 \le C_4 \cdot \frac{d^2}{n}\log\left(4R_{K,\rho}/\delta + 1\right),$$

where $C_2 \le cB^2\zeta^{-4}$, $C_3 \le c\tau^2 B^2\zeta^{-6} \cdot \frac{\sigma_{\max}(K)}{(1-\rho)}$ and $C_4 \le c\tau^2\zeta^{-4} \cdot \frac{\sigma_{\max}(K)}{(1-\rho)}$.

Assumption 4 is satisfied for the so-called "leaky ReLU" $\mathsf{relu}_\beta(x) := \max\{x, \beta x\}$ with $\beta > 0$, but not for the ReLU. Our next theorem shows that under stronger assumptions on the noise process, the algorithm succeeds at parameter recovery for the ReLU as well. We make the following assumption.

Assumption 5 *The link function σ is the ReLU ($\sigma_i(x_i) = \mathsf{relu}(x_i) := \max\{x_i, 0\}$) and the noise process is Gaussian, with $\varepsilon_i \sim \mathcal{N}(0, I)$.*

The gaussian assumption ensures for any pair of parameters, sufficiently large probability mass lies in the region where the ReLU is active. In particular, we use that for any pair $u, v \in \mathbb{R}^d$, $\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}[(\mathsf{relu}(\langle u, \varepsilon\rangle) - \mathsf{relu}(\langle v, \varepsilon\rangle))^2] \propto \|u - v\|^2$ (Lemma 18). Similar guarantees can be established for log-concave distributions using arguments in Balcan and Long (2013), but we consider only the gaussian case for simplicity.

Theorem 7 (Parameter recovery for the ReLU) *Suppose assumptions Assumptions 1-3 and Assumption 5 hold. Let $\delta > 0$ be fixed and suppose $n \geq c\frac{\tau^4 d^3}{1-\rho}\log(R_{K,\rho}/\delta + 1)$. Then when $\eta_t = (16R_{K,\rho})^{-2}e^{-4\rho R_{K,\rho}}$ and $m \geq C_0 \cdot \log n$, Algorithm 1 with Option I guarantees that with probability at least $1 - \delta$, $\|\widehat{\Theta} - \Theta^\star\|_F^2 \leq C_1 \cdot \frac{d^2}{n} \cdot R_{K,\rho}^2 \log^2(2R_{K,\rho}n/\delta + 1)$, where $C_0 \leq cB^2 e^{8\rho R_{K,\rho}}$ and $C_1 \leq \frac{c\tau W^2}{(1-\rho)^2} \cdot e^{8\rho R_{K,\rho}}$.*

Let us discuss some key features of Theorem 6 and Theorem 7. First, in the fast rate regime where Assumption 4 holds, Theorem 6 attains the usual parametric rate $\mathbb{E}\|\widehat{\Theta} - \Theta^\star\|_F \lesssim \sqrt{\frac{d^2}{n}}$, which is optimal for this setting (Tsybakov, 2008). The algorithm is also linearly convergent in this regime, and so the runtime to attain parameter recovery is nearly linear. On the other hand, the dependence on problem-dependent parameters such as K and ρ in the results can almost certainly be improved for all of the results. For example, while there are certainly systems for which $\|x_t\|$ grows as $\frac{1}{1-\rho}$, it is not clear whether exponential dependence on this parameter in Theorem 7 is required for parameter recovery with the ReLU. More generally, the factor $e^{R_{K,\rho}}$ in Theorem 7 can be replaced with $\max_i e^{\|\mu_i\|^2}$, where μ_i is an upper bound on the (conditional) expected value of x_i at time i. Our analysis simply bounds $\|\mu_i\|^2$ by $R_{K,\rho}$, and any improvements to this norm bound for systems of interest will immediately lead to improved rates.

4. Discussion

We have shown that the exponential stability, in conjunction with Lyapunov arguments, offers a simple approach to establishing isometry guarantees for data generated by nonlinear dynamical systems, and we have provided efficient algorithms for learning and parameter recovery in generalized linear systems. We hope that the analysis techniques introduced here will find use beyond the generalized linear setting, as well as for end-to-end control.

Going forward, it will be interesting to draw further connections and build stronger bridges between Lyapunov theory and empirical process theory for dependent data. For example, what properties of data generated by dynamical systems can we use Lyapunov functions to certify, going beyond lower and upper isometry?

Acknowledgements We thank Adam Klivans, Alexandre Megretski, and Karthik Sridharan for helpful discussions. We acknowledge the support of ONR award #N00014-20-1-2336 and NSF TRIPODS award #1740751.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Sohail Bahmani and Justin Romberg. Convex programming for estimation in nonlinear recurrent models. *arXiv preprint arXiv:1908.09915*, 2019.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

Victor H de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 6702–6712, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, pages 1260–1285, 2015.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2016.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of LQR is efficient. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. *Conference on Learning Theory (COLT)*, 2019.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression. In *Conference on Learning Theory (COLT)*, 2014.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 2014. To appear.

Anders Rantzer. Distributed control of positive systems. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 6608–6611. IEEE, 2011.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. Finite-time system identification for partially observed LTI systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

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, 2017.

Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online LQR. *arXiv preprint arXiv:2001.09576*, 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, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802, 2019.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5005–5014, 2018.

Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.