Journal of Machine Learning Research 23 (2022) 1-49 Submitted 6/20; Revised 5/22; Published 5/22

Non-asymptotic and Accurate Learning of Nonlinear Dynamical Systems

Yahya Sattar YSATT001@UCR.EDU
Department of Electrical and Computer Engineering
University of California Riverside, CA 92521, USA

Samet Oymak OYMAK@ECE.UCR.EDU
Department of Electrical and Computer Engineering
University of California Riverside, CA 92521, USA

Editor: Moritz Hardt

Abstract

We consider the problem of learning a nonlinear dynamical system governed by a nonlinear state equation $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{h}_t, \boldsymbol{u}_t; \boldsymbol{\theta}) + \boldsymbol{w}_t$. Here $\boldsymbol{\theta}$ is the unknown system dynamics, $\boldsymbol{h}_t$ is the state, $\boldsymbol{u}_t$ is the input and $\boldsymbol{w}_t$ is the additive noise vector. We study gradient based algorithms to learn the system dynamics $\boldsymbol{\theta}$ from samples obtained from a single finite trajectory. If the system is run by a stabilizing input policy, then using a mixing-time argument we show that temporally-dependent samples can be approximated by i.i.d. samples. We then develop new guarantees for the uniform convergence of the gradient of the empirical loss induced by these i.i.d. samples. Unlike existing works, our bounds are noise sensitive which allows for learning the ground-truth dynamics with high accuracy and small sample complexity. When combined, our results facilitate efficient learning of a broader class of nonlinear dynamical systems as compared to the prior works. We specialize our guarantees to entrywise nonlinear activations and verify our theory in various numerical experiments.

Keywords: nonlinear dynamical systems, stability, uniform convergence, learning from single trajectory

1. Introduction

Dynamical systems are fundamental for modeling a wide range of problems appearing in complex physical processes, cyber-physical systems and machine learning. Contemporary neural network models for processing sequential data, such as recurrent networks and LSTMs, can be interpreted as nonlinear dynamical systems and establish state-of-the-art performance in machine translation and speech recognition (Bahdanau et al., 2015; Graves et al., 2013; Li et al., 2013; Mikolov et al., 2010; Sak et al., 2014). Classical optimal control literature heavily relies on modeling the underlying system as a linear dynamical system (LDS) to synthesize control policies leading to elegant solutions such as PID controller and Kalman filter (Åström and Hägglund, 1995; Ho and Kálmán, 1966; Welch and Bishop, 1995). In many of these problems, we have to estimate or approximate the system dynamics from data, either because the system is initially unknown or because it is time-varying. This is alternatively known as the system identification problem which is the task of learning an unknown system from the time series of its trajectories (Åström and Eykhoff, 1971; Chen et al., 1990; Hochreiter and Schmidhuber, 1997; Ljung, 1998; Pintelon and Schoukens, 2012).

In this paper, we aim to learn the dynamics of nonlinear systems which are governed by following state equation,

$$\boldsymbol{h}_{t+1} = \phi(\boldsymbol{h}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \tag{1.1}$$

where $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is the system dynamics, $\boldsymbol{h}_t \in \mathbb{R}^n$ is the state vector, $\boldsymbol{u}_t \in \mathbb{R}^p$ is the input and $\boldsymbol{w}_t \in \mathbb{R}^n$ is the additive noise at time t. Our goal is understanding the statistical and computational efficiency of gradient based algorithms for learning the system dynamics from a single finite trajectory.

Contributions: Although system identification is classically well-studied, obtaining non-asymptotic sample complexity bounds is challenging especially when it comes to nonlinear systems. We address this challenge by connecting the system identification problem (which has temporally dependent samples) to classical statistical learning setup where data is independent and identically distributed (see Figure 1). We leverage this connection to show that gradient descent achieves stellar computational and statistical guarantees for nonlinear system identification. We establish this under a novel *one-point convexity and smoothness (OPCS)* condition (see Assumption 3) which allows for non-convex optimization landscape. Thus, our central contribution is providing an analysis framework for system identification through first-order methods with finite sample estimation guarantees. Specifically, we make the following contributions.

• **Learning nonlinear systems via gradient descent:** We work with (properly defined) stable nonlinear systems and use stability in conjunction with mixing-time arguments to address the problem of learning the system dynamics from a single finite trajectory. Under proper and intuitive assumptions, this leads to sample complexity and convergence guarantees for learning nonlinear dynamical systems (1.1) via gradient descent. Unlike the related results on nonlinear systems by Bahmani and Romberg (2020); Oymak (2019), our analysis accounts for the noise, achieves optimal statistical error rates in terms of the dimension d and the sample size N, and applies to a broader class of nonlinear systems.

• **Accurate statistical learning:** Of independent interest, we develop new statistical guarantees for the uniform convergence of the gradients of the empirical loss. Improving over earlier works of Foster et al. (2018); Mei et al. (2018), our bounds properly capture the noise dependence and allow for learning the ground-truth dynamics with high accuracy and small sample complexity (see §3 for further discussion).

• **Applications:** We specialize our results by establishing theoretical guarantees for learning linear ($\boldsymbol{h}_{t+1} = \boldsymbol{A}_\star \boldsymbol{h}_t + \boldsymbol{B}_\star \boldsymbol{u}_t + \boldsymbol{w}_t$) as well as nonlinear ($\boldsymbol{h}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{h}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t$) dynamical systems via gradient descent which highlight the optimality of our guarantees. We verify our theoretical results through various numerical experiments with nonlinear activations.

• **Broader implications:** Finally, while we focus on nonlinear state equations, our technical ideas (e.g., combining mixing-time and optimization landscape arguments, see Assumptions 1 and 3) have implications for richer class of systems. For instance, nonlinear ARX form $\boldsymbol{h}_t = \phi(\boldsymbol{A}_1 \boldsymbol{h}_{t-1} + \boldsymbol{A}_2 \boldsymbol{h}_{t-2} + \cdots + \boldsymbol{A}_m \boldsymbol{h}_{t-m}) + \boldsymbol{w}_{t-1}$ is a powerful generalization of the state equations that we investigate. Koopman lifting provides another class of nonlinear problems. We anticipate that our framework (i.e., merging one-point convexity and smoothness with mixing-time arguments to enable success of gradient descent) will also find applications for these systems.

Organization: We introduce the problem under consideration in §2 and provide uniform convergence guarantees for empirical gradients in §3. We relate the gradients of single trajectory loss and multiple trajectory loss in §4. Our main results on learning nonlinear systems are presented in §5 and applied to two special cases in §6. §7 provides numerical experiments to corroborate our theoretical results. §8 discusses the related works and §9 concludes the paper. Lastly, §10 presents the proofs of our main results.

Notations: We use boldface uppercase (lowercase) letters to denote matrices (vectors). For a vector $\boldsymbol{v}$, we denote its Euclidean norm by $\|\boldsymbol{v}\|_{\ell_2}$. For a matrix $\boldsymbol{M}$, $\rho(\boldsymbol{M}), \|\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_F$ denote the spectral radius, spectral norm and Frobenius norm respectively. $c, c_0, c_1, \ldots, C, C_0$ denote positive absolute constants. $\mathcal{S}^{d-1}$ denotes the unit sphere while $\mathcal{B}^d(\boldsymbol{a}, r)$ denotes the Euclidean ball of radius r, centered at $\boldsymbol{a}$, in $\mathbb{R}^d$. The normal distribution is denoted by $\mathcal{N}(\mu, \sigma^2)$. For a random vector $\boldsymbol{v}$, we denote its covariance matrix by $\boldsymbol{\Sigma}[\boldsymbol{v}]$. We use $\gtrsim$ and $\lesssim$ for inequalities that hold up to a constant factor. We denote by $a \vee b$, the maximum of two scalars a and b. Similarly, $a \wedge b$ denotes the minimum of the two scalars. Given a number a, $\lfloor a \rfloor$ denotes the largest integer less than or equal to a, whereas, $\lceil a \rceil$ denotes the smallest integer greater than or equal to a.

2. Problem Setup

We assume the system is driven by inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{h}_t) + \boldsymbol{z}_t$, where $\boldsymbol{\pi}(\cdot)$ is a fixed control policy and $\boldsymbol{z}_t$ is excitation for exploration. For statistical analysis, we assume the excitation and noise are random, that is, $(\boldsymbol{z}_t)_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_z$ and $(\boldsymbol{w}_t)_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_w$ for some distributions $\mathcal{D}_z$ and $\mathcal{D}_w$. With our choice of inputs, the state equation (1.1) becomes,

$$\boldsymbol{h}_{t+1} = \phi(\boldsymbol{h}_t, \boldsymbol{\pi}(\boldsymbol{h}_t) + \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t := \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \tag{2.1}$$

where $\tilde{\phi}$ denotes the closed-loop nonlinear system. Throughout, we assume the nonlinear functions $\phi(\cdot, \cdot; \boldsymbol{\theta})$ and $\tilde{\phi}(\cdot, \cdot; \boldsymbol{\theta})$ are differentiable in $\boldsymbol{\theta}$. For clarity of exposition, we will not explicitly state this assumption when it is clear from the context. To estimate $\boldsymbol{\theta}_\star$ in a non-asymptotic setting, we assume access to a finite trajectory $(\boldsymbol{h}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ generated by the nonlinear system (2.1). We also assume access to a stabilizing control policy $\boldsymbol{\pi}(\cdot)$. A special case of (2.1) is a linear state equation with $\boldsymbol{\theta}_\star = [\boldsymbol{A}_\star \ \boldsymbol{B}_\star]$, $\boldsymbol{\pi}(\boldsymbol{h}_t) = -\boldsymbol{K}\boldsymbol{h}_t$ and

$$\boldsymbol{h}_{t+1} = (\boldsymbol{A}_\star - \boldsymbol{B}_\star \boldsymbol{K})\boldsymbol{h}_t + \boldsymbol{B}_\star \boldsymbol{z}_t + \boldsymbol{w}_t, \tag{2.2}$$

Towards estimating $\boldsymbol{\theta}_\star$, we formulate an empirical risk minimization (ERM) problem over single finite trajectory as follows,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{\theta}), \quad \text{subject to} \quad \hat{\mathcal{L}}(\boldsymbol{\theta}) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{h}_{t+1} - \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2, \tag{2.3}$$

where $L \geq 1$ is a churn period which is useful for simplifying the notation later on, as L will also stand for the approximate mixing-time of the system. To solve (2.3), we investigate the properties of the gradient descent algorithm, given by the following iterate

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}_\tau), \tag{2.4}$$

where $\eta > 0$ is the fixed learning rate. ERM with i.i.d. samples is a fairly well-understood topic in classical machine learning. However, samples obtained from a single trajectory of a dynamical system are temporally dependent. For stable systems (see Def. 1), it can be shown that this dependence decays exponentially over the time. Capitalizing on this, we show that one can obtain almost i.i.d. samples from a given trajectory $(\boldsymbol{h}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$. This will in turn allow us to leverage techniques developed for i.i.d. data to solve problems with sequential data.

2.1 Assumptions on the System and the Inputs

We assume that the closed-loop system $\tilde{\phi}$ is stable. Stability in linear dynamical systems is connected to the spectral radius of the closed-loop system (Krauth et al., 2019; Simchowitz et al., 2018). The definition below provides a natural generalization of stability to nonlinear systems.

Definition 1 ((C_ρ, ρ)-stability) *Given excitation $(\boldsymbol{z}_t)_{t\geq 0}$ and noise $(\boldsymbol{w}_t)_{t\geq 0}$, denote the state sequence (2.1) resulting from initial state $\boldsymbol{h}_0 = \boldsymbol{\alpha}$, $(\boldsymbol{z}_\tau)_{\tau=0}^{t-1}$ and $(\boldsymbol{w}_\tau)_{\tau=0}^{t-1}$ by $\boldsymbol{h}_t(\boldsymbol{\alpha})$. Let $C_\rho \geq 1$ and $\rho \in (0,1)$ be system related constants. We say that the closed loop system $\tilde{\phi}$ is (C_ρ, ρ)-stable if, for all $\boldsymbol{\alpha}$, $(\boldsymbol{z}_t)_{t\geq 0}$ and $(\boldsymbol{w}_t)_{t\geq 0}$ triplets, we have*

$$\|\boldsymbol{h}_t(\boldsymbol{\alpha}) - \boldsymbol{h}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2}. \tag{2.5}$$

Note that, for a stable LDS ($\rho(\boldsymbol{A}_\star) < 1$), as a consequence of Gelfand's formula, there exists $C_\rho \geq 1$ and $\rho \in (\rho(\boldsymbol{A}_\star), 1)$ such that (C_ρ, ρ)-stability holds. A concrete example of nonlinear stable system is a contractive system where $\tilde{\phi}$ is ρ-Lipschitz function of $\boldsymbol{h}_t$ for some $\rho < 1$. We remark that, our interest in this work is not verifying the stability of a nonlinear system, but using stability of the closed-loop nonlinear system as an ingredient of the learning process. Verifying stability of the nonlinear systems can be very challenging, however, system analysis frameworks such as integral quadratic constraints (Megretski and Rantzer, 1997) and sum-of-squares (Prajna et al., 2002) may provide informative bounds.

Assumption 1 (Stability) *The closed-loop system $\tilde{\phi}$ is (C_ρ, ρ)-stable for some $\rho < 1$.*

Assumption 1 implies that the closed-loop system forgets a past state exponentially fast. This is different from the usual notion of "exponential Lyapunov stability" which requires the exponential convergence to a point in the state space. On the other hand, in the case of (C_ρ, ρ)-stability, the trajectories $\boldsymbol{h}_t(\boldsymbol{\alpha})$ and $\boldsymbol{h}_t(0)$ do not have to converge, rather their difference $\|\boldsymbol{h}_t(\boldsymbol{\alpha}) - \boldsymbol{h}_t(0)\|_{\ell_2}$ exponentially converges to zero (assuming $\|\boldsymbol{\alpha}\|_{\ell_2}$ is bounded). To keep the exposition simple, we will also assume $\boldsymbol{h}_0 = 0$ throughout. For data driven guarantees, we will make use of the following independence and boundedness assumptions on excitation and noise.

Assumption 2 (Boundedness) *There exist scalars $B, c_w, \sigma > 0$, such that $(\boldsymbol{z}_t)_{t\geq 0} \overset{i.i.d.}{\sim} \mathcal{D}_z$ and $(\boldsymbol{w}_t)_{t\geq 0} \overset{i.i.d.}{\sim} \mathcal{D}_w$ obey $\|\tilde{\phi}(0, \boldsymbol{z}_t; \boldsymbol{\theta}_\star)\|_{\ell_2} \leq B\sqrt{n}$ and $\|\boldsymbol{w}_t\|_{\ell_\infty} \leq c_w \sigma$ for $0 \leq t \leq T-1$ with probability at least $1 - p_0$ over the generation of data.*

2.2 Optimization Machinery

To concretely show how stability helps, we define the following loss function, obtained from i.i.d. samples at time $L-1$ and can be used as a proxy for $\mathbb{E}[\hat{\mathcal{L}}]$.

Definition 2 (Auxiliary Loss) *Suppose $\boldsymbol{h}_0 = 0$. Let $(\boldsymbol{z}_t)_{t\geq 0} \overset{i.i.d.}{\sim} \mathcal{D}_z$ and $(\boldsymbol{w}_t)_{t\geq 0} \overset{i.i.d.}{\sim} \mathcal{D}_w$. The auxiliary loss is defined as the expected loss at timestamp $L-1$, that is,*

$$\begin{aligned} &\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, (\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}))], \\ \text{where} \quad &\mathcal{L}(\boldsymbol{\theta}, (\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})) := \frac{1}{2}\|\boldsymbol{h}_L - \tilde{\phi}(\boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}; \boldsymbol{\theta})\|_{\ell_2}^2. \end{aligned} \tag{2.6}$$

Our generic system identification results via gradient descent will utilize the one-point convexity hypothesis. This is a special case of Polyak-Łojasiewicz inequality and provides a generalization of strong convexity to nonconvex functions.

Assumption 3 (One-point convexity & smoothness (OPCS)) *There exist scalars $\beta \geq \alpha > 0, r > 0$ such that, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, the auxiliary loss $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ of Definition 2 satisfies*

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}_\star, \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \rangle \geq \alpha \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}^2, \tag{2.7}$$

$$\|\nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}. \tag{2.8}$$

We emphasize that, as opposed to traditional strong convexity and smoothness assumptions (Nesterov, 2003), Assumption 3 is fairly mild, as it only assumes strong convexity and smoothness with respect to $\boldsymbol{\theta}_\star$. One-point convexity (OPC) is also known as restricted secant inequality and implies Polyak-Lojasiewicz condition (Karimi et al., 2016). To our knowledge, ours is the first work that use OPC with one-point smoothness (rather than global smoothness). A concrete example of a nonlinear system satisfying OPCS is the nonlinear state equation $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{h}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t$, with γ-increasing activation (i.e. $\phi'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$) and Gaussian excitation/noise (see Lemma 30). We expect many activations including ReLU to work as well. The main challenge is verifying OPCS of the population loss. For ReLU, Lemma 6.1 of Kalan et al. (2019) shows this property for i.i.d. Gaussian features. Extending this to subgaussian features would yield the ReLU result. The OPCS assumption can also be verified for nonlinear ARX $\boldsymbol{h}_t = \phi(\boldsymbol{A}_1 \boldsymbol{h}_{t-1} + \boldsymbol{A}_2 \boldsymbol{h}_{t-2} + \cdots + \boldsymbol{A}_m \boldsymbol{h}_{t-m}) + \boldsymbol{w}_{t-1}$ when the joint feature vector $[\boldsymbol{h}_{L-1}^\top \ \boldsymbol{h}_{L-2}^\top \ \cdots \ \boldsymbol{h}_{L-m}^\top]^\top$ has favorable covariance properties (e.g., positive definiteness) and ϕ is γ-increasing.

To proceed, if the gradient of $\hat{\mathcal{L}}(\boldsymbol{\theta})$ is close to that of $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ and Assumption 3 holds, gradient descent converges to the population minimum up to a statistical error governed by the noise level. The following statement summarizes our main results in Theorems 12 and 13. Below $\lesssim$ subsumes the logarithmic factors involving the problem variables.

Theorem 3 (Main result – informal) *Suppose we run gradient descent algorithm (2.4) to solve the ERM problem (2.3). Suppose Assumptions 1 - 5 hold. Suppose $r \gtrsim \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}$ and $T \gtrsim \frac{d}{\alpha^2(1-\rho)}$. The following statements hold with high probability over the trajectory.*

- ***Uniform convergence of gradient:*** *For all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, $\nabla \hat{\mathcal{L}}(\boldsymbol{\theta})$ satisfies*

$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) \sqrt{\frac{d}{T(1-\rho)}} \tag{2.9}$$

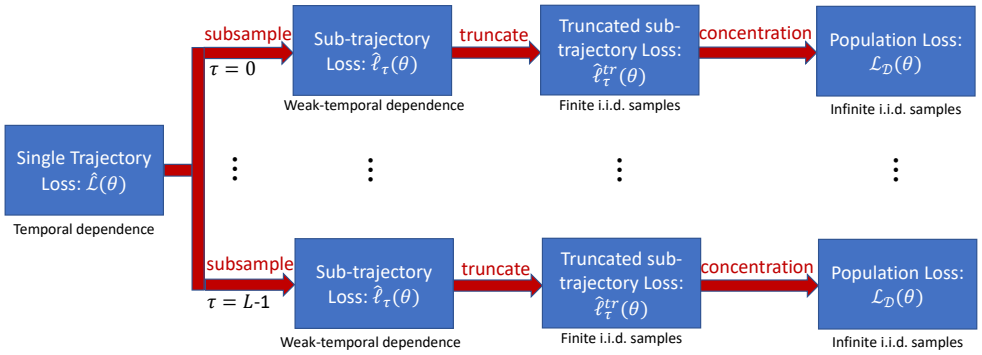


Figure 1: We learn nonlinear dynamical systems from a single trajectory by minimizing the empirical loss $\hat{\mathcal{L}}(\boldsymbol{\theta})$. The idea is to split $\hat{\mathcal{L}}(\boldsymbol{\theta})$ as an average of L sub-trajectory losses as $\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\tau=0}^{L-1}\hat{\ell}_\tau(\boldsymbol{\theta})$, through shifting and sub-sampling. Observing that each sub-trajectory has weakly dependent samples because of stability, we use a mixing time argument to show that $\|\nabla\hat{\ell}_\tau(\boldsymbol{\theta}) - \nabla\hat{\ell}_\tau^{\rm tr}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2})C_\rho\rho^{L-1}$, where $\hat{\ell}_\tau^{\rm tr}(\boldsymbol{\theta})$ is the loss constructed with finite i.i.d. samples (§4). Next, we show the uniform convergence of the empirical gradient as $\|\nabla\hat{\ell}_\tau^{\rm tr}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2})\sqrt{d/N}$, where $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}_\tau^{\rm tr}(\boldsymbol{\theta})]$ is the population loss (§3). Finally, we combine these with the local one-point convexity of the population loss to get our main results (§5).

- ***Convergence of gradient descent:*** *Set the learning rate* $\eta = \alpha/(16\beta^2)$ *and fix* $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. *All gradient descent iterates* $\boldsymbol{\theta}_\tau$ *on* $\hat{\mathcal{L}}(\boldsymbol{\theta})$ *satisfy*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \lesssim (1 - \frac{\alpha^2}{128\beta^2})^\tau\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}. \tag{2.10}$$

Observe that, our bounds exhibit optimal scaling in terms of the dimension d, the noise level σ and the trajectory length T. However, they degrade when stability parameter ρ approaches to one. Also note that this behavior is common in stability/mixing-based learning of dynamical systems (Boffi et al., 2021; Foster et al., 2020; Oymak, 2018). We remark that finite time identification of nonlinear dynamical systems without using stability arguments or establishing milder ρ-dependence is an exciting direction. Finally, observe that the computational convergence rate of (2.10) is $1 - \frac{\alpha^2}{128\beta^2}$. This rate can be strenghtened to $1 - \mathcal{O}(\alpha/\beta)$ if one assumes the stronger condition of global β-smoothness of $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ through existing arguments (Karimi et al., 2016). In contrast, we enforce weaker local one-point smoothness at the expense of β/α (condition number) times more computation.

In the following sections, we provide our formal results on the uniform convergence of gradient of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{\theta})$ and the identification of nonlinear dynamical systems (2.1).

3. Accurate Statistical Learning with Gradient Descent

To provide finite sample guarantees, we need to characterize the properties of the empirical loss and its gradients. Towards this goal, this section establishes new gradient based statistical

learning guarantees. Let $\mathcal{S} = (\boldsymbol{x}_i)_{i=1}^N$ be N i.i.d. samples from a distribution $\mathcal{D}$ and $\mathcal{L}(\cdot, \boldsymbol{x})$ be a loss function that admits a sample $\boldsymbol{x}$ and outputs the corresponding loss. When learning the nonlinear system (2.1), the sample $\boldsymbol{x}$ corresponds to the variables $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$ triple and the loss function $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})$ is given by (2.6). Define the empirical and population losses,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})]. \tag{3.1}$$

Let $\boldsymbol{\theta}_\star$ denotes the population minimizer which we wish to estimate via gradient descent. Recent works by Foster et al. (2018); Mei et al. (2018) provide finite sample learning guarantees via uniform convergence of the empirical gradient over a local ball $\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. However these works suffer from two drawbacks which we address here. To contrast the results, let us consider the following toy regression problem which is a simplification of our original task (2.3).

Generalized linear model: Suppose labels y_i are generated as, $y_i = \phi(\boldsymbol{z}_i^\top \boldsymbol{\theta}_\star) + w_i$ for some activation $\phi : \mathbb{R} \to \mathbb{R}$ where $\boldsymbol{z}_i \in \mathbb{R}^d$ is the input, w_i is the noise and $i = 1, \ldots, N$. Assume $N \gtrsim d$, $\boldsymbol{z}_i$ is zero-mean subgaussian vector with identity covariance and w_i has variance σ^2. Consider the quadratic loss

$$\hat{\mathcal{L}}_Q(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{i=1}^{N} (y_i - \phi(\boldsymbol{z}_i^\top \boldsymbol{\theta}))^2. \tag{3.2}$$

- **The role of noise:** Suppose ϕ is identity and the problem is purely linear regression. Then, gradient descent estimator will achieve statistical accuracy $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star\|_{\ell_2} \lesssim \sigma\sqrt{d/N}$. Foster et al. (2018); Mei et al. (2018) yield the coarser bound $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star\|_{\ell_2} \lesssim (\sigma + rC)\sqrt{d/N}$ for some scalars $r, C > 0$ coming from the uniform convergence of the empirical gradient over a local ball $\mathcal{B}(\boldsymbol{\theta}_\star, r)$.
- **Activation ϕ:** Both Foster et al. (2018); Mei et al. (2018) can only handle bounded activation ϕ. Foster et al. (2018) uses boundedness to control Rademacher complexity, whereas, Mei et al. (2018) requires bounded activation to make sure that the gradient of the loss is subgaussian. On the other hand, even for pure linear regression, gradients are subexponential rather than subgaussian (as it involves $\boldsymbol{z}_i \boldsymbol{z}_i^\top$).

Below we address both of these issues. We restrict our attention to low-dimensional setup, however we expect the results to extend to sparsity/ℓ_1 constraints in a straightforward fashion by adjusting covering numbers. In a similar spirit to Mei et al. (2018), we study the loss landscape over a local ball $\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. We first determine the conditions under which empirical and population gradients are close.

Assumption 4 (Lipschitz gradients) *There exist numbers $L_{\mathcal{D}}, p_0 > 0$ such that with probability at least $1 - p_0$ over the generation of data, for all pairs $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, the gradients of empirical and population losses in (3.1) satisfy*

$$\max(\|\nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}')\|_{\ell_2}, \|\nabla \hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2}) \le L_{\mathcal{D}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\ell_2}. \tag{3.3}$$

Observe that, by definition, the Lipschitz constant obeys $\mathcal{L}_{\mathcal{D}} \ge \beta$ where β is the one-point smoothness parameter in Assumption 3. However, $\mathcal{L}_{\mathcal{D}}$ is allowed be much larger than β.

Specifically, $\mathcal{L}_{\mathcal{D}}$ will only appear logarithmically in our bounds, hence, we can tolerate very large values of $\mathcal{L}_{\mathcal{D}}$. On the other hand β controls the convergence rate of gradient descent, hence, it must not be very large, compared to α, to guarantee fast linear convergence.

Assumption 5 (Subexponential gradient noise) *There exist scalars $K, \sigma_0 > 0$ such that, given $\boldsymbol{x} \sim \mathcal{D}$, at any point $\boldsymbol{\theta}$, the subexponential norm of the gradient of single sample loss $\mathcal{L}$ in (3.1) is upper bounded as a function of the noise level σ_0 and distance to the population minimizer via*

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}) - \mathbb{E}[\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})]\|_{\psi_1} \le \sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}, \tag{3.4}$$

where the subexponential norm of a random variable X is defined as $\|X\|_{\psi_1} := \sup_{k \ge 1} \frac{(\mathbb{E}[|X|^k])^{1/k}}{k}$ and that of a random vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as $\|\boldsymbol{x}\|_{\psi_1} := \sup_{\boldsymbol{v} \in \mathcal{S}^{n-1}} \|\boldsymbol{v}^\top \boldsymbol{x}\|_{\psi_1}$.

This assumption is an improvement over the work of Mei et al. (2018) and will help us distinguish the gradient noise due to optimization ($K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$) and due to noise σ_0 at the population minima.

As an example, consider the quadratic loss in (3.2). In the case of linear regression ($\phi(x) = x$), it is easy to show that Assumption 4 holds with $L_{\mathcal{D}} = 2$ and $p_0 = 2\exp(-100d)$, whereas, Assumption 5 holds with $K = c$ and $\sigma_0 = c_0\sigma$ for some scalars $c, c_0 > 0$. Moreover, in Appendix A.2, we show that in the case of nonlinear state equations $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{h}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t$, Assumptions 4 and 5 hold as long as ϕ has bounded first and second derivatives, that is, $|\phi'(x)|, |\phi''(x)| \le 1$ for all $x \in \mathbb{R}$. Specifically, using $\boldsymbol{z}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ and $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, if we bound the state covariance as $\boldsymbol{\Sigma}[\boldsymbol{h}_t] \preceq \beta_+^2 \boldsymbol{I}_n$ (see the proof of Lemma 33), then Assumption 4 holds with $L_{\mathcal{D}} = c((1+\sigma)\beta_+^2 n + \|\boldsymbol{\Theta}_\star\|_F \beta_+^3 n^{3/2} \log^{3/2}(2T))$ and $p_0 = 4T\exp(-100n)$, whereas, Assumption 5 holds with $K = c\beta_+^2$ and $\sigma_0 = c\sigma\beta_+$.

The next theorem establishes uniform concentration of the gradient as a function of the noise level and the distance from the population minima. To keep the exposition clean, from here on we set $C_{\log} = \log(3(L_{\mathcal{D}} N/K + 1))$.

Theorem 4 (Uniform convergence of gradient) *Suppose the gradients of $\mathcal{L}_{\mathcal{D}}$ and $\hat{\mathcal{L}}_{\mathcal{S}}$ obey Assumptions 4 and 5. Then, there exists $c_0 > 0$ such that, with probability at least $1 - p_0 - \log(\frac{Kr}{\sigma_0})\exp(-100d)$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have*

$$\|\nabla \hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) C_{\log} \sqrt{\frac{d}{N}}. \tag{3.5}$$

Proof sketch: Our proof technique uses peeling argument (Geer et al., 2000) to split the Euclidean ball $\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ into $P+1$ sets $\{\mathcal{S}_i\}_{i=0}^P$. Given a set $\mathcal{S}_i \subset \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and the associated radius r_i, we pick an ϵ_i-covering of the set $\mathcal{S}_i$. We then apply Lemma D.7 of Oymak (2018) (by specializing it to unit ball) together with a union bound over the elements of $P+1$ covers, to guarantee uniform convergence of the empirical gradient over the elements of $P+1$ covers. Combining this with Assumption 4, we guarantee a uniform convergence of the empirical gradient to its population counterpart over all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. ■

Theorem 4 provides a refined control over the gradient quality in terms of the distance $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$. The reason why Foster et al. (2018); Mei et al. (2018) are getting coarser

dependence on the noise level as compared to ours is their assumption that the gradient of the loss is subgaussian over all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ with subgaussian norm bounded by $\sigma + rC$, that is, there is a universal upper bound on the subgaussian norm of the gradient of the loss function over all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$.

To show the uniform convergence of the empirical gradient, Mei et al. (2018) requires the following assumptions on the gradient and the Hessian of the loss over all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$: (i) the gradient of the loss is subgaussian, (ii) the Hessian of the loss, evaluated on a unit vector, is subexponential, and (iii) the Hessian of the population loss is bounded at one point. Comparing (i) with Assumption 5, we observe that Assumption 5 is milder and is satisfied by a broader class of loss functions as compared to (i). For example, even for pure linear regression, the gradients are subexponential rather than subgaussian (as it involves $\boldsymbol{z}_i \boldsymbol{z}_i^\top$). On the other hand, our uniform convergence result requires Assumption 4 which might look restrictive. However, observe that the Lipschitz constant only appears logarithmically in our bounds, hence, Assumption 4 is fairly mild.

Going back to the original problem (2.3), observe that Theorem 4 bounds the impact of finite samples. In the next section, we provide bounds on the impact of learning from a single trajectory. Combining them relates the gradients of the auxiliary loss $\mathcal{L}_{\mathcal{D}}$ and the finite trajectory loss $\hat{\mathcal{L}}$ which will help learning $\boldsymbol{\theta}_\star$ from finite data obtained from a single trajectory.

4. Learning from a Single Trajectory

In this section we bound the impact of dependence in the data obtained from a single trajectory. For this purpose we use perturbation-based techniques to relate the gradients of the single trajectory loss $\hat{\mathcal{L}}$ and the multiple trajectory loss $\hat{\mathcal{L}}^{\text{tr}}$ (defined below). Before that, we introduce a few more concepts and definitions.

Definition 5 (Truncated state vector (Oymak, 2019)) *Consider the state equation (2.1). Suppose $\tilde{\phi}(0, 0; \boldsymbol{\theta}) = 0$, $\boldsymbol{h}_0 = 0$. Given, $t \geq L > 0$, for each state $\boldsymbol{h}_t$, we define its fictional proxy $\boldsymbol{h}_{t,L}$ by resetting $\boldsymbol{h}_{t-L} = 0$ but preserving the excitation $\boldsymbol{z}_\tau$ and noise $\boldsymbol{w}_\tau$ from $t - L$ to $t - 1$. Alternately, $\boldsymbol{h}_{t,L}$ is obtained by driving the system with excitations $\boldsymbol{z}'_\tau$ and additive noise $\boldsymbol{w}'_\tau$ until time $t - 1$, where*

$$\boldsymbol{z}'_\tau = \begin{cases} 0 \text{ if } \tau < t - L \\ \boldsymbol{z}_\tau \text{ else} \end{cases}, \quad \text{and} \quad \boldsymbol{w}'_\tau = \begin{cases} 0 \text{ if } \tau < t - L \\ \boldsymbol{w}_\tau \text{ else} \end{cases}. \tag{4.1}$$

We call the obtained state $\boldsymbol{h}_{t,L}$ as the L-truncated (or simply truncated) state at time t.

The L-truncated state vector $\boldsymbol{h}_{t,L}$ is identically distributed as $\boldsymbol{h}_L$. Hence, using truncation argument we can obtain i.i.d. samples from a single trajectory which will be used to bound the impact of dependence in the data. At its core our analysis uses a mixing time argument based on contraction and is used in related works by Bahmani and Romberg (2020); Oymak (2019). The difference between L-truncated and non-truncated state vectors is guaranteed to be bounded as

$$\|\boldsymbol{h}_t - \boldsymbol{h}_{t,L}\|_{\ell_2} \leq C_\rho \rho^L \|\boldsymbol{h}_{t-L}\|_{\ell_2}. \tag{4.2}$$

This directly follows from Definition 1 and asserts that the effect of past states decreases exponentially with truncation length L. To tightly capture the effect of truncation, we also bound the Euclidean norm of states $\boldsymbol{h}_t$ as follows.

Lemma 6 (Bounded states) *Suppose Assumptions 1 and 2 hold. Then, with probability at least $1-p_0$, we have $\|\boldsymbol{h}_t\|_{\ell_2} \le \beta_+\sqrt{n}$ for all $0 \le t \le T$, where $\beta_+ := C_\rho(c_w\sigma + B)/(1-\rho)$.*

Following this and (4.2), we can obtain weakly dependent sub-trajectories by properly sub-sampling a single trajectory $(\boldsymbol{h}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$. For this purpose, we first define a sub-trajectory and its truncation as follows.

Definition 7 (Truncated sub-trajectories (Oymak, 2019)) *Let sampling period $L \ge 1$ be an integer. Set the sub-trajectory length $N = \lfloor \frac{T-L}{L} \rfloor$. We sub-sample the trajectory $(\boldsymbol{h}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ at points $\tau+L, \tau+2L, \ldots, \tau+NL$ and truncate the states by $L-1$ to get the τ_{th} truncated sub-trajectory $(\bar{\boldsymbol{h}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N$, defined as*

$$(\bar{\boldsymbol{h}}^{(i)}, \boldsymbol{z}^{(i)}) := (\boldsymbol{h}_{\tau+iL,L-1}, \boldsymbol{z}_{\tau+iL}) \quad \text{for} \quad i = 1, \ldots, N \tag{4.3}$$

where $0 \le \tau \le L-1$ is a fixed offset.

For notational convenience, we also denote the noise at time $\tau + iL$ by $\boldsymbol{w}^{(i)}$. The following lemma states that the τ_{th} truncated sub-trajectory $(\bar{\boldsymbol{h}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N$ has independent samples.

Lemma 8 (Independence) *Suppose $(\boldsymbol{z}_t)_{t=0}^\infty \overset{i.i.d.}{\sim} \mathcal{D}_z$ and $(\boldsymbol{w}_t)_{t=0}^\infty \overset{i.i.d.}{\sim} \mathcal{D}_w$. Then, the τ_{th} truncated states $(\bar{\boldsymbol{h}}^{(i)})_{i=1}^N$ are all independent and are identically distributed as $\boldsymbol{h}_{L-1}$. Moreover, $(\bar{\boldsymbol{h}}^{(i)})_{i=1}^N, (\boldsymbol{z}^{(i)})_{i=1}^N, (\boldsymbol{w}^{(i)})_{i=1}^N$ are all independent of each other.*

For the purpose of analysis, we will define the loss restricted to a sub-trajectory and show that each sub-trajectory can have favorable properties that facilitate learning.

Definition 9 (Truncated sub-trajectory loss) *We define the truncated loss in terms of truncated (sub-sampled) triplets $(\bar{\boldsymbol{y}}^{(i)}, \bar{\boldsymbol{h}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N := (\boldsymbol{h}_{\tau+iL+1,L}, \boldsymbol{h}_{\tau+iL,L-1}, \boldsymbol{z}_{\tau+iL})_{i=1}^N$ as*

$$\hat{\ell}_\tau^{tr}(\boldsymbol{\theta}) := \frac{1}{2N}\sum_{i=1}^N \|\bar{\boldsymbol{y}}^{(i)} - \tilde{\phi}(\bar{\boldsymbol{h}}^{(i)}, \boldsymbol{z}^{(i)}; \boldsymbol{\theta})\|_{\ell_2}^2. \tag{4.4}$$

Observe that the triplets $(\bar{\boldsymbol{y}}^{(i)}, \bar{\boldsymbol{h}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N$ are independent and identically distributed as $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$. Therefore, we have $\mathcal{L}_\mathcal{D}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta})]$, that is, $\hat{\ell}_\tau^{\text{tr}}$ is a finite sample approximation of $\mathcal{L}_\mathcal{D}$ and we will use results from Section 3 to bound the Euclidean distance between them. Before, stating our results on uniform convergence of empirical losses, we want to demonstrate the core idea regarding stability. For this purpose, we define the truncated loss which is truncated version of the empirical loss (2.3).

Definition 10 (Truncated loss) *Let $\boldsymbol{h}_{t+1,L} = \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t$. We define the truncated (empirical) risk as*

$$\hat{\mathcal{L}}^{tr}(\boldsymbol{\theta}) := \frac{1}{2(T-L)}\sum_{t=L}^{T-1} \|\boldsymbol{h}_{t+1,L} - \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{\ell}_\tau^{tr}(\boldsymbol{\theta}). \tag{4.5}$$

Let $\mathcal{H}$ be the convex hull of all states $\boldsymbol{h}_t$ and $\mathcal{Z}$ be the convex hull of all the inputs $\boldsymbol{z}_t$ such that Assumptions 1 and 2 are valid. As a regularity condition, we require the problem to behave nicely over state-excitation pairs $(\boldsymbol{h}, \boldsymbol{z}) \subset \mathcal{H} \times \mathcal{Z}$. Throughout, $\tilde{\phi}_k$ denotes the scalar function associated to the k_{th} entry of $\tilde{\phi}$.

The following theorem states that, in the neighborhood of $\boldsymbol{\theta}_\star$, the empirical risk $\hat{\mathcal{L}}$ behaves like the truncated risk $\hat{\mathcal{L}}^{\text{tr}}$, when the approximate mixing-time L is chosen sufficiently large.

Theorem 11 (Small impact of truncation) *Consider the state equation (2.1). Suppose Assumptions 1 and 2 hold. Suppose there exists $r > 0$ such that, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and for all $(\boldsymbol{h}, \boldsymbol{z}) \subset \mathcal{H} \times \mathcal{Z}$, we have that $\|\nabla_{\boldsymbol{h}} \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\| \leq B_{\tilde{\phi}}$, $\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} \leq C_{\tilde{\phi}}$ and $\|\nabla_{\boldsymbol{h}} \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\| \leq D_{\tilde{\phi}}$ for some scalars $B_{\tilde{\phi}}, C_{\tilde{\phi}}, D_{\tilde{\phi}} > 0$ and $1 \leq k \leq n$. Let $\beta_+ > 0$ be as in Lemma 6. Then, with probability at least $1 - p_0$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have*

$$|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{tr}(\boldsymbol{\theta})| \leq 2n\beta_+ C_\rho \rho^{L-1} B_{\tilde{\phi}} (c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}), \tag{4.6}$$

$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}^{tr}(\boldsymbol{\theta})\|_{\ell_2} \leq 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}} (c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \tag{4.7}$$

Proof sketch: To prove Theorem 11, we use the Mean-value Theorem together with Assumptions 1 and 2. First, using (2.3) and (4.5), we obtain

$$\begin{aligned} |\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})| \leq \frac{1}{2} \max_{L \leq t \leq (T-1)} | &\|\tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t - \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 \\ &- \|\tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t - \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 |. \end{aligned} \tag{4.8}$$

Suppose, the maximum is achieved at $(\boldsymbol{h}, \bar{\boldsymbol{h}}, \boldsymbol{z}, \boldsymbol{w})$ (where $\bar{\boldsymbol{h}}$ is the truncated state). Then, we use the identity $a^2 - b^2 = (a+b)(a-b)$ to upper bound the difference $|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})|$ as a product of two terms $|a+b|$ and $|a-b|$ with $a := \|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}$ and $b := \|\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}$. We upper bound the term $|a+b|$ by bounding each quantity a and b using the Mean-value Theorem together with Assumption 2. Similarly, the term $|a-b|$ is upper bounded by first applying triangle inequality and then using the Mean-value Theorem together with Assumptions 1 and 2 (to bound the difference $\|\boldsymbol{h} - \bar{\boldsymbol{h}}\|_{\ell_2}$). Combining the two bounds gives us the statement (4.6) of the Theorem. A similar proof technique is used to upper bound the gradient distance $\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})\|_{\ell_2}$. ■

Combining Theorems 4 and 11 allows us to upper bound the Euclidean distance between the gradients of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{\theta})$ and the auxiliary loss $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ which is the topic of the next section.

5. Main Results

5.1 Non-asymptotic Identification of Nonlinear Systems

In this section, we provide our main results on statistical and convergence guarantees of gradient descent for learning nonlinear dynamical systems, using finite samples generated from a single trajectory. Before stating our main result on non-asymptotic identification of nonlinear systems, we state a theorem to bound the Euclidean distance between the gradients the empirical loss $\hat{\mathcal{L}}(\boldsymbol{\theta})$ and the auxiliary loss $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$.

Theorem 12 (Uniform convergence of gradient) *Fix $r > 0$. Suppose Assumptions 1 and 2 on the system and Assumptions 4 and 5 on the Auxiliary Loss hold. Also suppose for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $(\boldsymbol{h}, \boldsymbol{z}) \subset \mathcal{H} \times \mathcal{Z}$, we have $\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} \le C_{\tilde{\phi}}$ and $\|\nabla_{\boldsymbol{h}} \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\| \le D_{\tilde{\phi}}$ for all $1 \le k \le n$ for some scalars $C_{\tilde{\phi}}, D_{\tilde{\phi}} > 0$. Define $K_{\tilde{\phi}} := (2/c_0)\beta_+ D_{\tilde{\phi}}(c_w \sigma/\sigma_0 \vee C_{\tilde{\phi}}/K)$. Let $\beta_+ > 0$ be as in Lemma 6 and $N = \lfloor (T-L)/L \rfloor$, where we pick L via*

$$L \ge L_0 \quad \text{where} \quad L_0 = \Big\lceil 1 + \frac{\log(C\rho K_{\tilde{\phi}} n \sqrt{N/d})}{1-\rho} \Big\rceil. \tag{5.1}$$

Then, with probability at least $1 - 2Lp_0 - L\log(\frac{Kr}{\sigma_0})\exp(-100d)$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have

$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le 2c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) C_{\log} \sqrt{\frac{d}{N}}. \tag{5.2}$$

Proof sketch: Theorem 12 can be proved by combining the results of Theorems 4 and 11. The idea is to split the truncated loss $\hat{\mathcal{L}}^{\mathrm{tr}}$ (Def. 10) as an average of L truncated subtrajectory losses $\hat{\ell}_\tau^{\mathrm{tr}}$ (Def. 9) as: $\hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta})$. Recall that $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta})]$. Then, we use Theorem 4 with a union bound over all $0 \le \tau \le L-1$ to upper bound $\|\nabla \hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}$ which is used to show the uniform convergence of the truncated loss $\hat{\mathcal{L}}^{\mathrm{tr}}$ as: $\|\nabla \hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le \frac{1}{L}\sum_{\tau=0}^{L-1} \|\nabla \hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}$. Combining this with Theorem 11 and picking L via (5.1), we get the statement of the theorem. ■

Observe that $K_{\tilde{\phi}}$ depends on the system related constants and the noise level. For example, for a linear dynamical system (2.2), we can show that $K_{\tilde{\phi}} = c\sqrt{n+p}$. Note that, if we choose $N \gtrsim K^2 C_{\log}^2 d/\alpha^2$ in Theorem 12, we get $\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim \sigma_0 C_{\log}\sqrt{d/N} + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$. Combining this result with Assumption 3 gives our final result on non-asymptotic identification of nonlinear dynamical systems from a single trajectory.

Theorem 13 (Non-asymptotic identification) *Consider the setup of Theorem 12. Also suppose the Auxiliary loss satisfies Assumption 3. Let $N = \lfloor (T-L)/L \rfloor$, where we pick L as in Theorem 12. Suppose $N \gtrsim K^2 C_{\log}^2 d/\alpha^2$. Given $r > 0$, set learning rate $\eta = \alpha/(16\beta^2)$ and pick $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. Assuming $\sigma_0 \lesssim rK$, with probability at least $1 - 2Lp_0 - L\log(\frac{Kr}{\sigma_0})\exp(-100d)$, all gradient descent iterates $\boldsymbol{\theta}_\tau$ on $\hat{\mathcal{L}}$ satisfy*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \le \Big(1 - \frac{\alpha^2}{128\beta^2}\Big)^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{c\sigma_0}{\alpha} C_{\log}\sqrt{\frac{d}{N}}. \tag{5.3}$$

Proof sketch: To prove Theorem 13, we first show that, when (i) the auxiliary loss $\mathcal{L}_{\mathcal{D}}$ satisfies one-point convexity and smoothness (Assumption 3), (ii) for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, $\nabla \hat{\mathcal{L}}$ satisfies $\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le \nu + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$, and (iii) $r \ge 5\nu/\alpha$; then, setting learning rate $\eta = \alpha/(16\beta^2)$ and fixing $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, all gradient descent iterates $\boldsymbol{\theta}_\tau$ on $\hat{\mathcal{L}}$ satisfy $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \le \big(1 - \frac{\alpha^2}{128\beta^2}\big)^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{5\nu}{\alpha}$. Combining this with Theorem 12, we get the desired result. Specifically, we use Theorem 12 with $N \gtrsim K^2 C_{\log}^2 d/\alpha^2$, to get the gradient convergence in the form of (ii) with $\nu = c\sigma_0 C_{\log}\sqrt{\frac{d}{N}}$. Plugging this back to the gradient descent convergence result established above, we get the statement of the theorem. ■

Observe that, Theorem 13 requires $\mathcal{O}(d)$ samples to learn the dynamics $\boldsymbol{\theta}_\star \in \mathbb{R}^d$, hence, our sample complexity captures the correct dependence on the dimension of unknown system dynamics. Furthermore, it achieves $\sigma\sqrt{d/N}$ error rate, which is optimal in both d and N. Recall that the gradient noise σ_0 is a function of the process noise σ, and role of σ will be more clear in § 6. We remark that while this theorem provides strong dependence, the results can be further refined when the number of states n is large since each sample in (2.1) provides n equations. For example, we can accomplish better sample complexity for separable dynamical systems (see §5.2) which is the topic of next section.

Lastly, observe that L is proportional to $1/(1-\rho)$. As a result, our sample complexity bound degrades with stability. In the extreme case, when $\rho = 1$, the approximate mixing time L goes to infinity, and our analysis does not hold. This has been previously observed in stability/mixing-based learning of nonlinear dynamical systems (Boffi et al., 2021; Foster et al., 2020; Oymak, 2019). In contrast, it is well-known that this dependency (on $\rho(\boldsymbol{A}_\star)$) can be avoided for learning linear dynamical systems (Simchowitz et al., 2018). Recently, Jain et al. (2021) showed, under a strong invertibility condition, that dependency on the mixing time can be avoided for the generalized linear models $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}_\star \boldsymbol{h}_t) + \boldsymbol{w}_t$. This leaves open the question of whether learning without mixing is possible in situations beyond the generalized linear models.

5.2 Separable Dynamical Systems

Suppose now that the nonlinear dynamical system is separable, that is, the nonlinear state equation (2.1) can be split into n state updates via

$$\boldsymbol{h}_{t+1}[k] = \tilde{\phi}_k(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_k^\star) + \boldsymbol{w}_t[k], \quad \text{for } 1 \le k \le n, \tag{5.4}$$

where $\boldsymbol{h}_t[k]$ and $\boldsymbol{w}_t[k]$ denote the k_{th} entry of $\boldsymbol{h}_t$ and $\boldsymbol{w}_t$ respectively while $\tilde{\phi}_k$ denotes the scalar function associated to the k_{th} entry of $\tilde{\phi}$. The overall system is given by the concatenation $\boldsymbol{\theta}_\star = [\boldsymbol{\theta}_1^{\star\top} \cdots \boldsymbol{\theta}_n^{\star\top}]^\top$. For simplicity, let us assume $\boldsymbol{\theta}_k^\star \in \mathbb{R}^{\bar{d}}$, where $\bar{d} = d/n$. In the case of separable dynamical systems, the empirical loss in (2.3) is alternately given by,

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} (\boldsymbol{h}_{t+1}[k] - \tilde{\phi}_k(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_k))^2. \tag{5.5}$$

As before, we aim to learn the system dynamics $\boldsymbol{\theta}_\star$ via gradient descent. The gradient of the empirical loss simplifies to $\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) = [\nabla\hat{\mathcal{L}}_1(\boldsymbol{\theta}_1)^\top \cdots \nabla\hat{\mathcal{L}}_n(\boldsymbol{\theta}_n)^\top]^\top$. From this, we observe that learning $\boldsymbol{\theta}_\star$ via (2.3) is equivalent to learning each of its components $\boldsymbol{\theta}_k^\star$ by solving n separate ERM problems in $\mathbb{R}^{\bar{d}}$. Denoting $\hat{\boldsymbol{\theta}}$ to be the solution of the ERM problem (2.3), we have the following equivalence: $\hat{\boldsymbol{\theta}} \equiv [\hat{\boldsymbol{\theta}}_1^\top \cdots \hat{\boldsymbol{\theta}}_n^\top]^\top$, where $\hat{\boldsymbol{\theta}}_k \in \mathbb{R}^{\bar{d}}$ is the solution to the following minimization problem,

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k \in \mathbb{R}^{\bar{d}}} \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k). \tag{5.6}$$

Similarly global iterations (2.4) follows the iterations of the subproblems, that is, the GD iterate (2.4) implies $\boldsymbol{\theta}_k^{(\tau+1)} = \boldsymbol{\theta}_k^{(\tau)} - \eta\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k^{(\tau)})$. Before, stating our main result on learning

separable nonlinear dynamical systems, we will show how the Auxiliary loss $\mathcal{L}_{\mathcal{D}}$ and its finite sample approximation $\hat{\mathcal{L}}_{\mathcal{S}}$ can be split into the sum of n losses as follows,

$$\begin{aligned}\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{x}_i), \\ \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) \quad \text{where} \quad \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) = \mathbb{E}[\mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{x})],\end{aligned} \tag{5.7}$$

where $\mathcal{L}_k(\cdot, \boldsymbol{x})$ is a loss function that admits a sample $\boldsymbol{x}$ and outputs the corresponding loss. When learning (5.4), the sample $\boldsymbol{x}$ corresponds to the variables $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$ triple and the loss function $\mathcal{L}_k(\boldsymbol{\theta}, \boldsymbol{x})$ is given by

$$\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})) := \frac{1}{2}(\boldsymbol{h}_L[k] - \tilde{\phi}_k(\boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}; \boldsymbol{\theta}_k))^2. \tag{5.8}$$

The following theorem gives refined sample complexity for learning the dynamics of separable nonlinear dynamical systems.

Theorem 14 (Refined complexity) *Suppose Assumptions 1 and 2 on the system and Assumptions 3, 4 and 5 on the Auxiliary Loss (5.7) hold for all $1 \le k \le n$. Additionally, suppose the nonlinear dynamical system is separable, that is, the nonlinear state equation follows (5.4). Let $K_{\tilde{\phi}}$ be as in Theorem 12. Let $N = \lfloor (T-L)/L \rfloor$, where we pick L via*

$$L \ge L_0 \quad \textit{where} \quad L_0 = \Big\lceil 1 + \frac{\log(C\rho K_{\tilde{\phi}} n\sqrt{N/\bar{d}})}{1-\rho} \Big\rceil. \tag{5.9}$$

Suppose $N \gtrsim K^2 C_{\log}^2 \bar{d}/\alpha^2$. Given $r > 0$, set the learning rate $\eta = \alpha/(16\beta^2)$ and pick $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. Assuming $\sigma_0 \lesssim rK$, with probability at least $1 - 2Lnp_0 - Ln\log(\frac{Kr}{\sigma_0})\exp(-100\bar{d})$, all gradient descent iterates $\boldsymbol{\theta}_\tau = [\boldsymbol{\theta}_1^{(\tau)\top} \cdots \boldsymbol{\theta}_n^{(\tau)\top}]^\top$ on $\hat{\mathcal{L}}$ satisfy

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le \big(1 - \frac{\alpha^2}{128\beta^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma_0}{\alpha} C_{\log}\sqrt{\frac{\bar{d}}{N}} \quad \textit{for all} \quad 1 \le k \le n. \tag{5.10}$$

Proof sketch: The proof technique for Theorem 14 is similar to that of Theorem 13. First, using Assumptions 4 and 5 on the Auxiliary loss (5.7), we get an upper bound on $\|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2}$ for all $1 \le k \le n$. Next, using Assumption 1 and 2 on the system, we upper bound $\|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k)\|_{\ell_2}$ for all $1 \le k \le n$. Combining these two bounds, we get an upper bound on the gradient distance $\|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2}$ for all $1 \le k \le n$. After picking N and L in the same way as we we did in Theorem 13, we use Theorem 3 with Assumption 3 on the Auxiliary loss (5.7) and the derived bound on $\|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2}$ to get the statement of the theorem. ■

Observe that, in the case of separable dynamical systems we require $\mathcal{O}(\bar{d})$ samples to learn the dynamics $\boldsymbol{\theta}_\star \in \mathbb{R}^d$. We achieve refined sample complexity because each sample provides n equations and $\bar{d} = d/n$. Common dynamical systems like linear dynamical systems and nonlinear state equations are very structured and have separable state equations. Hence, applying Theorem 14 to these systems results in accurate sample complexity and error rates which is the topic of the next section.

6. Applications

In this section, we apply our results from the previous section to learn two different dynamical systems of the following form,

$$\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}_\star \boldsymbol{h}_t) + \boldsymbol{B}_\star \boldsymbol{z}_t + \boldsymbol{w}_t, \tag{6.1}$$

where $\boldsymbol{A}_\star \in \mathbb{R}^{n\times n}$, $\boldsymbol{B}_\star \in \mathbb{R}^{n\times p}$ are the unknown system dynamics, $\boldsymbol{z}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ and $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. Specifically we learn the dynamics of the following dynamical systems: **(a)** Standard linear dynamical systems ($\phi = \boldsymbol{I}_n$); and **(b)** Nonlinear state equations

$$\boldsymbol{h}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{h}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t, \tag{6.2}$$

where the nonlinear function $\phi : \mathbb{R} \to \mathbb{R}$ applies entry-wise on vector inputs. For the clarity of exposition, we focus on stable systems and set the feedback policy $\boldsymbol{\pi}(\boldsymbol{h}_t) = 0$. For linear dynamical systems, this is equivalent to assuming $\rho(\boldsymbol{A}_\star) < 1$. For nonlinear state equation, we assume (C_ρ, ρ)-stability holds according to Definition 1.

6.1 Linear Dynamical Systems

To simplify the notation, we define the following concatenated vector/matrix: $\boldsymbol{x}_t := [\boldsymbol{h}_t^\top \ \boldsymbol{z}_t^\top]^\top$ and $\boldsymbol{\Theta}_\star := [\boldsymbol{A}_\star \ \boldsymbol{B}_\star]$. Letting $\phi = \boldsymbol{I}_n$, the state update (6.1) is alternately given by: $\boldsymbol{h}_{t+1} = \boldsymbol{\Theta}_\star \boldsymbol{x}_t + \boldsymbol{w}_t$. To proceed, let $\boldsymbol{\theta}_k^{\star\top}$ denotes the k_{th} row of $\boldsymbol{\Theta}_\star$, then $\boldsymbol{\Theta}_\star \equiv [\boldsymbol{\theta}_1^\star \cdots \boldsymbol{\theta}_n^\star]^\top$. Observe that the standard linear dynamical system is separable as in (5.4). Therefore, given a finite trajectory $(\boldsymbol{h}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ of the linear dynamical system (6.1) ($\phi = \boldsymbol{I}_n$), we construct the empirical loss as follows,

$$\hat{\mathcal{L}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} (\boldsymbol{h}_{t+1}[k] - \boldsymbol{\theta}_k^\top \boldsymbol{x}_t)^2. \tag{6.3}$$

Before stating our main result, we introduce a few more concepts to capture the properties of gradient descent for learning the dynamics $\boldsymbol{\theta}_k^\star$. Define the matrices,

$$\boldsymbol{G}_t := [\boldsymbol{A}_\star^{t-1}\boldsymbol{B}_\star \ \boldsymbol{A}_\star^{t-2}\boldsymbol{B}_\star \cdots \boldsymbol{B}_\star] \quad \text{and} \quad \boldsymbol{F}_t := [\boldsymbol{A}_\star^{t-1} \ \boldsymbol{A}_\star^{t-2} \cdots \boldsymbol{I}_n]. \tag{6.4}$$

Then, the matrices $\boldsymbol{G}_t\boldsymbol{G}_t^\top$ and $\boldsymbol{F}_t\boldsymbol{F}_t^\top$ are the finite time controllability Gramians for the control and noise inputs, respectively. It is straightforward to see that the covariance matrix of the concatenated vector $\boldsymbol{x}_t$ satisfies the following bounds (see § A.1 for detail)

$$(1 \wedge \lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p} \preceq \boldsymbol{\Sigma}[\boldsymbol{x}_t] \preceq (1 \vee \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_{n+p}). \tag{6.5}$$

Define, $\gamma_- := 1 \wedge \lambda_{\min}(\boldsymbol{G}_{L-1}\boldsymbol{G}_{L-1}^\top + \sigma^2 \boldsymbol{F}_{L-1}\boldsymbol{F}_{L-1}^\top)$, $\gamma_+ := 1 \vee \lambda_{\max}(\boldsymbol{G}_{L-1}\boldsymbol{G}_{L-1}^\top + \sigma^2 \boldsymbol{F}_{L-1}\boldsymbol{F}_{L-1}^\top)$ and $\beta_+ = 1 \vee \max_{1\le t\le T} \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top)$. The following corollary of Theorem 14 states our main result on the statistical and convergence guarantees of gradient descent for learning the dynamics of linear dynamical systems.

Corollary 15 *Consider the system (6.1) with $\phi = \boldsymbol{I}_n$. Suppose $\rho(\boldsymbol{A}_\star) < 1$. Let $C_\rho \geq 1$ and $\rho \in (\rho(\boldsymbol{A}_\star), 1)$ be scalars. Suppose $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ and $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. Let $\gamma_+ \geq \gamma_- > 0$ be as defined in (6.5) and set $\kappa = \gamma_+/\gamma_-$. Let $N = \lfloor (T-L)/L \rfloor$, where we pick L via*

$$L \geq L_0 \quad \text{where} \quad L_0 = \Big\lceil 1 + \frac{\log(CC_\rho \beta_+ N(n+p)/\gamma_+)}{1-\rho} \Big\rceil. \tag{6.6}$$

Suppose $N \gtrsim \kappa^2 \log^2(6N+3)(n+p)$. Set the learning rate $\eta = \gamma_-/(16\gamma_+^2)$ and the initialization $\boldsymbol{\Theta}_0 = 0$. Assuming $\sigma \lesssim \|\boldsymbol{\Theta}_\star\|_F \sqrt{\gamma_+}$, with probability at least $1 - 4T\exp(-100n) - Ln\big(4 + \log(\frac{\|\boldsymbol{\Theta}_\star\|_F \sqrt{\gamma_+}}{\sigma})\big)\exp(-100(n+p))$, for all $1 \leq k \leq n$, all gradient descent iterates $\boldsymbol{\Theta}_\tau = [\boldsymbol{\theta}_1^{(\tau)} \cdots \boldsymbol{\theta}_n^{(\tau)}]^\top$ on $\hat{\mathcal{L}}$ satisfy

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \leq \big(1 - \frac{\gamma_-^2}{128\gamma_+^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma\sqrt{\kappa}}{\sqrt{\gamma_-}} \log(6N+3)\sqrt{\frac{n+p}{N}}. \tag{6.7}$$

Observe that Corollary 15 requires $\mathcal{O}(n+p)$ samples to learn the dynamics $\boldsymbol{A}_\star \in \mathbb{R}^{n \times n}$ and $\boldsymbol{B}_\star \in \mathbb{R}^{n \times p}$. The sample complexity captures the correct dependence on the dimension of unknown system dynamics, because each sample provides n equations and there are $n(n+p)$ unknown parameters. Our sample complexity bound correctly depends on the condition number κ of the covariance matrix $\boldsymbol{\Sigma}[\boldsymbol{x}_{L-1}]$. Moreover, $\gamma_- = 1 \wedge \lambda_{\min}(\boldsymbol{G}_{L-1}\boldsymbol{G}_{L-1}^\top + \sigma^2 \boldsymbol{F}_{L-1}\boldsymbol{F}_{L-1}^\top)$ is a non-decreasing function of the mixing time L. The intuition for this is that larger L takes into account more long-term excitations to lower bound the size of covariance matrix $\boldsymbol{\Sigma}[\boldsymbol{x}_{L-1}]$. When the condition number of $\boldsymbol{\Sigma}[\boldsymbol{x}_t]$ is close to 1, the sample complexity of the problem is lower and vice versa. Lastly, our statistical error rate $\sigma\sqrt{(n+p)/N}$ is optimal in the dimension $(n+p)$ and sample size N. The logarithmic dependence on $\|\boldsymbol{\Theta}_\star\|_F$ is an artifact of our general framework. We believe it can be possibly removed with a more refined concentration analysis.

6.2 Nonlinear State Equations

In this section, we apply Theorem 14 to learn the nonlinear state equation (6.2). Observe that the nonlinear system (6.2) is separable because we assume that the nonlinear function $\phi : \mathbb{R} \to \mathbb{R}$ applies entry-wise on vector inputs. Let $\boldsymbol{\theta}_k^{\star\top}$ denotes the k_{th} row of $\boldsymbol{\Theta}_\star$. Given a finite trajectory $(\boldsymbol{h}_{t+1}, \boldsymbol{h}_t)_{t=0}^{T-1}$ of (6.2), we construct the empirical loss as follows,

$$\hat{\mathcal{L}}(\boldsymbol{\Theta}) = \sum_{k=1}^n \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} (\boldsymbol{h}_{t+1}[k] - \phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t))^2. \tag{6.8}$$

The following corollary of Theorem 14 states our main result on the statistical and convergence guarantees of gradient descent for learning the nonlinear system (6.2).

Corollary 16 *Suppose the nonlinear system (6.2) satisfies (C_ρ, ρ)-stability according to Def. 1. Suppose ϕ is γ-increasing (i.e. $\phi'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$), has bounded first and second derivatives, that is, $|\phi'|, |\phi''| \leq 1$, and $\phi(0) = 0$. Suppose $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$ and $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. Let $N = \lfloor (T-L)/L \rfloor$, where we pick L via*

$$L \geq L_0 \quad \text{where} \quad L_0 = \Big\lceil 1 + \frac{\log(CC_\rho(1 + \|\boldsymbol{\Theta}_\star\|_F C_\rho(1+\sigma)/(1-\rho))Nn)}{1-\rho} \Big\rceil. \tag{6.9}$$

Setting $D_{\log} = \log(3(1+\sigma)n + 3C_\rho(1+\sigma)\|\mathbf{\Theta}_\star\|_F n^{3/2}\log^{3/2}(2T)N/(1-\rho)+3)$, *suppose* $N \gtrsim \frac{C_\rho^4}{\gamma^4(1-\rho)^4} D_{\log}^2 n$. *Set the learning rate* $\eta = \frac{\gamma^2(1-\rho)^4}{32C_\rho^4(1+\sigma)^2n^2}$ *and pick the initialization* $\mathbf{\Theta}_0 = 0$. *Assuming* $\sigma \lesssim \|\mathbf{\Theta}_\star\|_F$, *with probability at least* $1 - Ln\big(4T + \log(\frac{\|\mathbf{\Theta}_\star\|_F C_\rho(1+\sigma)}{\sigma(1-\rho)})\big)\exp(-100n)$, *for all* $1 \le k \le n$, *all gradient descent iterates* $\mathbf{\Theta}_\tau = [\boldsymbol{\theta}_1^{(\tau)} \cdots \boldsymbol{\theta}_n^{(\tau)}]^\top$ *on* $\hat{\mathcal{L}}$ *satisfy*

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le \big(1 - \frac{\gamma^4(1-\rho)^4}{512C_\rho^4 n^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma}{\gamma^2(1-\rho)} C_\rho D_{\log}\sqrt{\frac{n}{N}}. \tag{6.10}$$

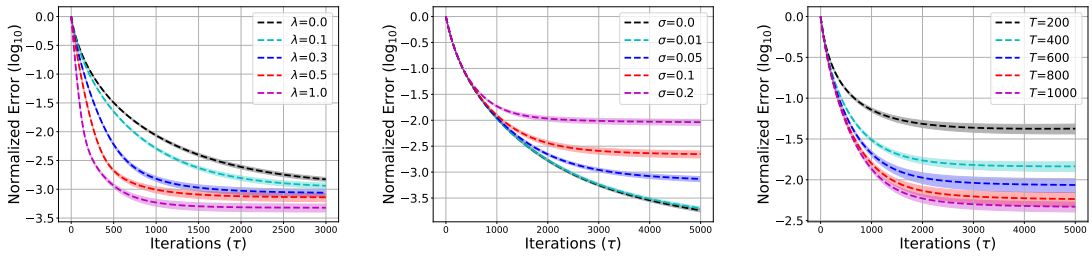
We believe that the condition of γ-increasing ϕ can be relaxed and we expect many nonlinear activations including ReLU to work. The main challenge is verifying one-point convexity of the population loss when ϕ is ReLU. Lemma 6.1 of Kalan et al. (2019) shows this property for i.i.d. Gaussian features. Extending this to subgaussian features, would yield the ReLU result. Theorem 16 requires $\mathcal{O}(n)$ samples to learn the dynamics $\mathbf{\Theta}_\star \in \mathbb{R}^{n\times n}$ since each sample gives n equations. The sample complexity bound depends on the condition number of the covariance matrix $\mathbf{\Sigma}[\boldsymbol{h}_t]$, which can be shown to be bounded by $C_\rho^2/(1-\rho)^2$ (see Section A.2). Lastly, similar to the linear case, our statistical error rate $\sigma\sqrt{n/N}$ is optimal in the dimension n and sample size N.

Remark 17 (Probability of success) *For our main results, instead of achieving* $1-\delta$ *probability of success with variable* $\delta \in (0,1)$, *we are content with achieving* $1 - K_{\log}\exp(-cd)$ *probability of success for an absolute constant* $c > 0$, *where* $K_{\log}$ *is a fixed constant which depends either logarithmically or linearly on the values of* n, L, T, N, σ_0, K *etc. Please note that, the probability of success in Theorems 12, 13 and 14 is coming from an application of Lemma 18 in §10. We simply apply this lemma using a fixed choice of* $t = c_0\sqrt{d}$. *This gives the error bound* $\tilde{\mathcal{O}}(\sigma_0\sqrt{d/N})$ *and the probability of success* $1 - K_{\log}\exp(-cd)$. *One can also obtain* $1-\delta$ *probability of success by setting* $t = c_0\sqrt{\log(K_{\log}/\delta)}$ *(instead of* $t = c_0\sqrt{d}$*), when applying Lemma 18 in §10. This gives the error bound* $\tilde{\mathcal{O}}(\sigma_0\sqrt{\frac{d\log(K_{\log}/\delta)}{N}})$. *In this case, one can easily see the trade-off between the probability of success and the error bound.*

7. Numerical Experiments

Leakage	$\|\boldsymbol{A}_\star\|$	$\|\boldsymbol{A}'_\star\|$	$\rho(\boldsymbol{A}_\star)$	$\rho(\boldsymbol{A}'_\star)$	$\sup_{\|\boldsymbol{x}\|_{\ell_2}=1}\|\phi(\boldsymbol{A}_\star\boldsymbol{x})\|_{\ell_2}$	$\sup_{\|\boldsymbol{x}\|_{\ell_2}=1}\|\phi(\boldsymbol{A}'_\star\boldsymbol{x})\|_{\ell_2}$
$\lambda = 0.00$	2.07	1.85	1.12	0.65	1.79	1.56
$\lambda = 0.50$	2.07	1.85	1.12	0.65	1.84	1.60
$\lambda = 0.80$	2.07	1.85	1.12	0.65	1.92	1.70
$\lambda = 1.00$	2.07	1.85	1.12	0.65	2.07	1.85

Table 1: This table lists the core properties of the (random) state matrix in our experiments. The values are averaged over 1000 random trials. For linear systems, the state matrix $\boldsymbol{A}_\star$ is unstable however the closed-loop matrix $\boldsymbol{A}'_\star$ is stable. We also list the nonlinear spectral norms (i.e. $\sup_{\|\boldsymbol{x}\|_{\ell_2}=1}\|\phi(\boldsymbol{A}_\star\boldsymbol{x})\|_{\ell_2}$) associated with $\boldsymbol{A}_\star$ and $\boldsymbol{A}'_\star$, as a function of different leakage levels of leaky-ReLUs, which are all larger than 1. Despite this, experiments show nonlinear systems are stable with $\boldsymbol{A}'_\star$ (some even with $\boldsymbol{A}_\star$). This indicates that Definition 1 is indeed applicable to a broad range of systems.



(a) Nonlinearity (b) Noise level (c) Trajectory length

Figure 2: We run gradient descent to learn nonlinear dynamical system governed by state equation $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{h}_t + \boldsymbol{B}\boldsymbol{u}_t) + \boldsymbol{w}_t$. We study the effect of nonlinearity, noise variance and trajectory length on the convergence of gradient descent. The empirical results verify what is predicted by our theory.

For our experiments, we choose unstable nonlinear dynamical systems ($\rho(\boldsymbol{A}) > 1$) governed by nonlinear state equation $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{h}_t + \boldsymbol{B}\boldsymbol{u}_t) + \boldsymbol{w}_t$ with state dimension $n = 80$ and input dimension $p = 50$. $\boldsymbol{A}$ is generated with $\mathcal{N}(0,1)$ entries and scaled to have its largest 10 eigenvalues greater than 1. $\boldsymbol{B}$ is generated with i.i.d. $\mathcal{N}(0,1/n)$ entries. For nonlinearity, we use either softplus ($\phi(x) = \ln(1+e^x)$) or leaky-ReLU ($\max(x, \lambda x)$, with leakage $0 \le \lambda \le 1$) activations. We run gradient descent with fixed learning rate $\eta = 0.1/T$, where T denotes the trajectory length. We choose a noisy stabilizing policy $\boldsymbol{K}$ for the linear system (ignoring ϕ) and set $\boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{h}_t + \boldsymbol{z}_t$. Here $\boldsymbol{K}$ is obtained by solving a discrete-time Riccati equation (by setting rewards $\boldsymbol{Q}, \boldsymbol{R}$ to identity) and adding random Gaussian noise with zero mean and variance 0.001 to each entry of the Riccati solution. We want to emphasize that any stabilizing policy will work here. For some nonlinear activations, as shown in Figure 3, one can learn the system dynamics using a policy which is unstable for the linear system but remains stable for the nonlinear system. Lastly, $\boldsymbol{z}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ and $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$.

We plot the normalized estimation error of $\boldsymbol{A}$ and $\boldsymbol{B}$ given by the formula $\|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_F^2 / \|\boldsymbol{A}\|_F^2$ (same for $\boldsymbol{B}$). Each experiment is repeated 20 times and we plot the mean and one standard deviation. To verify our theoretical results, we study the effect of the following on the convergence of gradient descent for learning the system dynamics.

• **Nonlinearity:** This experiment studies the effect of nonlinearity on the convergence of gradient descent for learning nonlinear dynamical system with leaky-ReLU activation. We run gradient descent over different values of λ (leakage). The trajectory length is set to $T = 2000$ and the noise variance is set to $\sigma^2 = 0.01$. In Figure 2a, we plot the normalized estimation error of $\boldsymbol{A}$ over different values of λ. We observe that, decreasing nonlinearity leads to faster convergence of gradient descent.

• **Noise level:** This experiment studies the effect of noise variance on the convergence of gradient descent for learning nonlinear dynamical system with softplus activation. The trajectory length is set to $T = 2000$. In Figure 2b, we plot the normalized estimation error of $\boldsymbol{A}$ over different values of noise variance. We observe that, the gradient descent linearly converges to the ground truth plus some residual which is proportional to the noise variance as predicted by our theory.

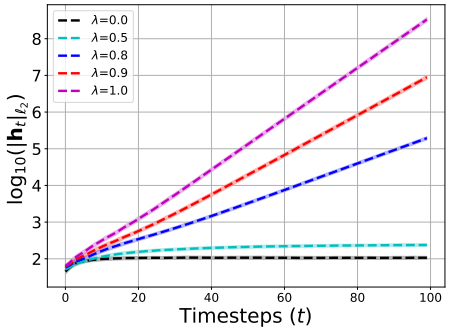


Figure 3: For a properly chosen random unstable system the state vectors diverge for LDS while they stay bounded for leaky ReLU systems with small leakage.

• **Trajectory length:** This experiment studies the effect of trajectory length on the statistical accuracy of learning system dynamics via gradient descent. We use softplus activation and the noise variance is set to $\sigma^2 = 0.01$. In Figure 2c, we plot the normalized estimation error of $\boldsymbol{A}$ over different values of T. We observe that, by increasing the trajectory length (number of samples), the estimation gets better, verifying our theoretical results.

We remark that, we get similar plots for the input matrix $\boldsymbol{B}$. Lastly, Figure 3 is generated by evolving the state through 100 timesteps and recording the Euclidean norm of $\boldsymbol{h}_t$ at each timestep. This is repeated 500 times with $\rho(\boldsymbol{A}) > 1$ and using leaky-ReLU activations. In Figure 3, we plot the mean and one standard deviation of the Euclidean norm of the states $\boldsymbol{h}_t$ over different values of λ (leakage). The states are bounded when we use leaky-ReLU with $\lambda \leq 0.5$ even when the corresponding LDS is unstable. This shows that the nonlinearity can help the states converge to a point in state space. However, this is not always true. For example, when $\boldsymbol{A} = 2\boldsymbol{I}$ and $\boldsymbol{h}_0$ has all entries positive. Then, using leaky-ReLU will not help the trajectory to converge.

8. Related Work

Nonlinear dynamical systems relate to the literature in control theory, reinforcement learning, and recurrent neural networks. We study nonlinear dynamical systems from optimization and learning perspective rather than control. While such problems are known to be challenging (especially under nonlinearity), there is a growing interest in understanding system identification and associated optimal control problems (e.g. LQR) in a non-asymptotic and data-dependent fashion (Recht, 2019). Recently Dean et al. (2018); Faradonbeh et al. (2018, 2020); Fattahi et al. (2019); Hardt et al. (2018); Hazan et al. (2017, 2018); Oymak and Ozay (2019); Sarkar and Rakhlin (2019); Sarkar et al. (2019, 2021); Simchowitz et al. (2018, 2019); Tsiamis and Pappas (2019); Tsiamis et al. (2020); Wagenmaker and Jamieson (2020) explore linear system identification in great depth. Allen-Zhu et al. (2019) provides preliminary guarantees for recurrent networks (RNN) and Miller and Hardt (2019) shows the role of stability in RNNs. There is also a substantial amount of work on model-free approaches (see e.g., Dann and Brunskill, 2015; Fazel et al., 2018; Krauth et al., 2019; Malik et al., 2019; Zou et al., 2019) which avoid learning the dynamics and find the optimal

control input by directly optimizing over policy space. In a different line of work, Singh et al. (2021) proposed a learning framework for trajectory planning from learned dynamics. They propose a regularizer of dynamics that promotes stabilizability of the learned model, which allows tracking reference trajectories based on estimated dynamics. Also, Khosravi and Smith (2020a,b) developed learning methods that exploit other control-theoretic priors. Nonetheless, none of these works characterize the sample complexity of the problem.

More recently, Mania et al. (2022) proposes an active learning approach for non-asymptotic identification of nonlinear dynamical systems whose state transitions depend linearly on a known feature embedding of state-action pairs. Kakade et al. (2020) extends this to an online nonlinear control problem, and provides the lower confidence-based continuous control algorithm, which enjoys $\mathcal{O}(\sqrt{T})$ regret bound. (Boffi et al., 2021) studies the problem of adaptive control of a known discrete-time nonlinear system subject to unmodeled disturbances, and uses online least squares algorithms to estimate the unknown parameter. In a similar line of work, Lale et al. (2021) proposes an online model learning predictive control framework to control unknown nonlinear dynamical systems, Mhammedi et al. (2020) proposes a learning-theoretic framework for continuous control in which the environment is summarized by a low-dimensional continuous latent state with linear dynamics and quadratic costs, but the agent operates on high-dimensional, nonlinear observations, and Jain et al. (2021) provides the first offline algorithm that can learn generalized linear models without the mixing assumption.

Closer to our work, Bahmani and Romberg (2020); Oymak (2019) study theoretical properties of nonlinear state equations with a goal towards understanding recurrent networks and nonlinear systems. While some high-level ideas, such as mixing-time arguments, are shared, our results (a) apply to a broader class of nonlinear systems (e.g. mild assumptions on nonlinearity), (b) utilize a variation of the spectral radius for nonlinear systems[1], (c) account for process noise, and (d) develop new statistical guarantees for the uniform convergence of the gradient of the empirical loss. The concurrent work of Foster et al. (2020) provides related results for the recovery of generalized linear dynamical systems ($\boldsymbol{h}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{h}_t) + \boldsymbol{w}_t$) using complementary techniques. Foster et al. (2020) uses martingale arguments and analyze GLMtron algorithm of Kakade et al. (2011), while we use mixing time arguments and analyze gradient descent.

A very preliminary version of this work has appeared in a workshop paper (Sattar and Oymak, 2019) where we provide preliminary guarantees for the identification nonlinear dynamical systems. In contrast to this work, Sattar and Oymak (2019) does not provide sample complexity and statistical error bounds and learns a simple noiseless system by assuming the one-point convexity of the empirical loss (with i.i.d. samples). On the other hand, this work provides new guarantees for non-asymptotic identification of nonlinear dynamical systems under process noise. It develops new statistical guarantees for the uniform convergence of the gradients of the empirical loss and applies the developed framework to learn nonlinear state equations $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{h}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t$. Lastly, it also provides the necessary technical framework and the associated proofs.

Perhaps the most established technique in the statistics literature for dealing with non-independent, time-series data is the use of mixing-time arguments (Yu, 1994). In the

[1] Rather than enforcing contraction (i.e. 1-Lipschitzness)-based stability which corresponds to using spectral norm rather than spectral radius.

machine learning literature, mixing-time arguments have been used to develop generalization bounds (Kuznetsov and Mohri, 2017; McDonald et al., 2017; Mohri and Rostamizadeh, 2007, 2008) which are analogous to the classical generalization bounds for i.i.d. data. We utilize mixing-time for nonlinear stabilizable systems to connect our temporally-dependent problem to standard supervised learning task with a focus on establishing statistical guarantees for gradient descent.

Finite sample convergence of the gradients of the empirical loss (to the population gradient) is studied by Foster et al. (2018); Mei et al. (2018). These guarantees are not sufficient for our analysis as they only apply to problems with bounded nonlinearities and do not accurately capture the noise dependence. We address this by establishing stronger uniform convergence guarantees for empirical gradients and translate our bounds to the system identification via mixing-time/stability arguments.

9. Conclusions

We proposed a general approach for learning nonlinear dynamical systems by utilizing stabizability and mixing-time arguments. We showed that, under reasonable assumptions, one can learn the dynamics of a nonlinear stabilized systems from a single finite trajectory. Our general approach can treat important dynamical systems, such as LDS and the setups of Bahmani and Romberg (2020); Foster et al. (2020); Oymak (2019) as special cases. We provided both sample size and estimation error guarantees on LDS and certain nonlinear state equations. Finally, the numerical experiments verify our theoretical findings on statistical and computational efficiency of gradient descent for learning nonlinear systems.

There are many interesting future avenues. One direction is exploring alternative approaches to mixing-time arguments. Martingale based arguments have the potential to provide tighter statistical guarantees and mitigate dependence on the spectral radius (Simchowitz et al., 2018). Another important direction is learning better control policies by optimizing the policy function $\boldsymbol{\pi}$ in a data driven fashion. This topic has attracted significant attention for linear systems (Dean et al., 2018; Recht, 2019) and led to strong regret guarantees (Cohen et al., 2019; Mania et al., 2019), however, nonlinearity presents significant challenges. Our framework is more suitable for model based approaches (as it learns system dynamics $\boldsymbol{\theta}_\star$), however, model-free guarantees would be similarly intriguing.

Acknowledgments

This work was partially supported by an NSF grant CNS-1932254, an NSF CAREER award CCF-2046816, and a Research Scholar award from Google.

10. Proofs of the Main Results

In this section, we present the proofs of our main results.

10.1 Proof of Theorem 4

Before we begin our proof, we state a lemma to bound the Euclidean norm of a sum of i.i.d. subexponential random vectors. The following lemma is a restatement of Lemma D.7 of Oymak (2018) (by specializing it to unit ball) and it follows from an application of generic chaining tools.

Lemma 18 *Let $C > 0$ be a universal constant. Suppose $N \geq d$. Let $(\boldsymbol{v}_i)_{i=1}^N \in \mathbb{R}^d$ be i.i.d. vectors obeying $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{v}_i]$ and subexponential norm $\|\boldsymbol{v}_i - \boldsymbol{\mu}\|_{\psi_1} \leq K$. With probability at least $1 - 2\exp(-c\min(t\sqrt{N}, t^2))$, we have that*

$$\|\frac{1}{N}\sum_{i=1}^{n} \boldsymbol{v}_i - \boldsymbol{\mu}\|_{\ell_2} \leq CK\frac{\sqrt{d}+t}{\sqrt{N}}. \tag{10.1}$$

Alternatively, setting $t = \tau\sqrt{d}$ for $\tau \geq 1$, with probability at least $1 - 2\exp(-c\tau d)$, we have

$$\|\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}_i - \boldsymbol{\mu}\|_{\ell_2} \leq CK(\tau+1)\sqrt{d/N}. \tag{10.2}$$

Throughout the proof of Theorem 4. we pick the constraint set $\mathcal{C} = \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, however, these ideas are general and would apply to any set with small covering numbers (such as sparsity, ℓ_1, rank constraints).

Proof of uniform convergence with covering argument: We will use a peeling argument (Geer et al., 2000). Split the ball $\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ into $P+1 = \lceil \log(Kr/\sigma_0)\rceil + 1$ sets via following arguments,

$$\mathcal{B}^d(\boldsymbol{\theta}_\star, r) = \cup_{i=0}^{P}\mathcal{S}_i \quad \text{where} \quad \mathcal{S}_i = \begin{cases} \mathcal{B}^d(\boldsymbol{\theta}_\star, \sigma_0/K) & \text{if} \quad i = 0, \\ \mathcal{B}^d(\boldsymbol{\theta}_\star, \min(r, \mathrm{e}^i\sigma_0/K)) - \mathcal{B}^d(\boldsymbol{\theta}_\star, \mathrm{e}^{i-1}\sigma_0/K) & \text{else.} \end{cases}$$

By Assumption 4, with probability at least $1-p_0$, $\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta})$, $\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ are $L_{\mathcal{D}}$-Lipschitz. Given a set $\mathcal{S}_i$ and the associated radius $r_i = \min(r, \mathrm{e}^i\sigma_0/K)$, pick an $\varepsilon_i \leq r_i \leq r$ covering $\mathcal{N}_i$ of the set $\mathcal{S}_i \subset \mathcal{B}^d(\boldsymbol{\theta}_\star, r_i)$ such that $\log|\mathcal{N}_i| \leq d\log(3r_i/\varepsilon_i)$. Observe that over $\mathcal{S}_i$, by construction, we have

$$\max(\sigma_0/K, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) \leq r_i \leq \max(\sigma_0/K, \mathrm{e}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \tag{10.3}$$

Applying Lemma 18 together with a union bound over the $P+1$ covers and elements of the covers, we guarantee the following: Within all covers $\mathcal{N}_i$, gradient vector at all points $\boldsymbol{\theta} \in \mathcal{N}_i$ satisfies

$$\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma_0 + Kr_i)\log(3r_i/\varepsilon_i)\sqrt{d/N}, \tag{10.4}$$

with probability at least $1 - \sum_{i=0}^{P}\exp(-100d\log(3r_i/\varepsilon_i))$. Given both events hold with probability at least $1 - p_0 - \sum_{i=0}^{P}\exp(-100d\log(3r_i/\varepsilon_i))$, for any $\boldsymbol{\theta} \in \mathcal{S}_i$, pick $\boldsymbol{\theta}' \in \mathcal{N}_i$ so that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\ell_2} \leq \varepsilon$. This yields

$$\begin{aligned} &\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \\ &\qquad \leq \|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}')\|_{\ell_2} + \|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2} + \|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}') - \nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2}, \\ &\qquad \lesssim \varepsilon_i L_{\mathcal{D}} + (\sigma_0 + Kr_i)\log(3r_i/\varepsilon_i)\sqrt{d/N}. \end{aligned} \tag{10.5}$$

Setting $\varepsilon_i = \min(1, \frac{K}{L_{\mathcal{D}}}\sqrt{d/N})r_i$ for $0 \le i \le P$, for any $\boldsymbol{\theta} \in \mathcal{S}_i$ (and thus for any $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$), we have

$$\begin{aligned}\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} &\lesssim (\sigma_0 + Kr_i)\log(3(1+L_{\mathcal{D}}N/K))\sqrt{d/N}, \\ &\lesssim (\sigma_0 + K\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2})\log(3(1+L_{\mathcal{D}}N/K))\sqrt{d/N}, \end{aligned} \tag{10.6}$$

where we used (10.3) to get the last inequality. Finally, observing that $\log(3r_i/\varepsilon_i) \ge 1$, the probability bound simplifies to

$$1 - p_0 - \sum_{i=0}^{P}\exp(-100d\log(3r_i/\varepsilon_i)) \ge 1 - p_0 - \log(\frac{Kr}{\sigma_0})\exp(-100d). \tag{10.7}$$

This completes the proof. ■

10.2 Proof of Lemma 6

Proof Suppose $\boldsymbol{h}_0 = 0$. We claim that $\|\boldsymbol{h}_t\|_{\ell_2} \le \beta_+\sqrt{n}(1-\rho^t)$ with probability at least $1-p_0$, where $\beta_+ := C_\rho(c_w\sigma + B)/(1-\rho)$. Note that, using the bounds on $\boldsymbol{z}_t, \boldsymbol{w}_t$, the state vector $\boldsymbol{h}_1$ satisfies the following bound and obeys the induction

$$\|\boldsymbol{h}_1\|_{\ell_2} \le B\sqrt{n} + c_w\sigma\sqrt{n} \le C_\rho\sqrt{n}(B + c_w\sigma) = \beta_+\sqrt{n}(1-\rho^1). \tag{10.8}$$

Suppose the bound holds until $t-1$, where $t \le T$, and let us apply induction. First observe that $\|\boldsymbol{h}_{t,L}\|_{\ell_2}$ obeys the same upper bound as $\|\boldsymbol{h}_L\|_{\ell_2}$ by construction. Recalling (4.2), we get the following by induction

$$\begin{aligned}\|\boldsymbol{h}_t - \boldsymbol{h}_{t,t-1}\|_{\ell_2} \le C_\rho\rho^{t-1}\|\boldsymbol{h}_1\|_{\ell_2} \implies \|\boldsymbol{h}_t\|_{\ell_2} &\le C_\rho\rho^{t-1}\|\boldsymbol{h}_1\|_{\ell_2} + \|\boldsymbol{h}_{t,t-1}\|_{\ell_2}, \\ \|\boldsymbol{h}_t\|_{\ell_2} &\overset{(a)}{\le} C_\rho\rho^{t-1}\|\boldsymbol{h}_1\|_{\ell_2} + \beta_+\sqrt{n}(1-\rho^{t-1}), \\ &\overset{(b)}{\le} \sqrt{n}(C_\rho\rho^{t-1}(B + c_w\sigma) + \beta_+(1-\rho^{t-1})), \\ &\le \beta_+\sqrt{n}(1-\rho^t), \end{aligned} \tag{10.9}$$

where, we get (a) from the induction hypothesis and (b) from the bound on $\boldsymbol{h}_1$. This bound also implies $\|\boldsymbol{h}_t\|_{\ell_2} \le \beta_+\sqrt{n}$ with probability at least $1-p_0$, for all $0 \le t \le T$, and completes the proof. ■

10.3 Proof of Lemma 8

Proof By construction $\bar{\boldsymbol{h}}^{(i)}$ only depends on the vectors $\{\boldsymbol{z}_t, \boldsymbol{w}_t\}_{t=\tau+(i-1)L+1}^{\tau+iL-1}$. Note that the dependence ranges $[\tau + (i-1)L + 1, \tau + iL - 1]$ are disjoint intervals for each $i's$. Hence, $\{\bar{\boldsymbol{h}}^{(i)}\}_{i=1}^N$ are all independent of each other. To show the independence of $\{\bar{\boldsymbol{h}}^{(i)}\}_{i=1}^N$ and $\{\boldsymbol{z}^{(i)}\}_{i=1}^N$, observe that the inputs $\boldsymbol{z}^{(i)} = \boldsymbol{z}_{\tau+iL}$ have timestamps $\tau + iL$; which is not covered by $[\tau + (i-1)L + 1, \tau + iL - 1]$ - the dependence ranges of $\{\bar{\boldsymbol{h}}^{(i)}\}_{i=1}^N$. Identical argument shows the independence of $\{\bar{\boldsymbol{h}}^{(i)}\}_{i=1}^N$ and $\{\boldsymbol{w}^{(i)}\}_{i=1}^N$. Lastly, $\{\boldsymbol{z}^{(i)}\}_{i=1}^N$ and $\{\boldsymbol{w}^{(i)}\}_{i=1}^N$ are independent

of each other by definition. Hence, $\{\bar{\boldsymbol{h}}^{(i)}\}_{i=1}^N, \{\boldsymbol{z}^{(i)}\}_{i=1}^N, \{\boldsymbol{w}^{(i)}\}_{i=1}^N$ are all independent of each other. This completes the proof. ■

10.4 Proof of Theorem 11

Proof Our proof consists of two parts. The first part bounds the Euclidean distance between the truncated and non-truncated losses while the second part bounds the Euclidean distance between their gradients.

• **Convergence of loss:** To start, recall $\hat{\mathcal{L}}(\boldsymbol{\theta})$ and $\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})$ from (2.3) and (4.5) respectively. The distance between them can be bounded as follows.

$$
\begin{aligned}
&|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})| \\
&\quad = |\frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{h}_{t+1} - \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 - \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{h}_{t+1,L} - \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\quad \le \frac{1}{2(T-L)} \sum_{t=L}^{T-1} |\|\boldsymbol{h}_{t+1} - \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 - \|\boldsymbol{h}_{t+1,L} - \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\quad \le \frac{1}{2} \max_{L \le t \le (T-1)} |\|\boldsymbol{h}_{t+1} - \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 - \|\boldsymbol{h}_{t+1,L} - \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\quad \le \frac{1}{2} \|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}^2 - \|\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\quad \le \frac{1}{2} (|\|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} - \|\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}|) \\
&\qquad (|\|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} + \|\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}|), \qquad (10.10)
\end{aligned}
$$

where, $(\boldsymbol{h}, \bar{\boldsymbol{h}}, \boldsymbol{z}, \boldsymbol{w})$ corresponds to the maximum index ($\bar{\boldsymbol{h}}$ be the truncated state) and we used the identity $a^2 - b^2 = (a+b)(a-b)$. Denote the k_{th} element of $\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})$ by $\tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})$ and that of $\boldsymbol{w}$ by w_k for $1 \le k \le n$. To proceed, using Mean-value Theorem, with probability at least $1 - p_0$, we have

$$
\begin{aligned}
|\tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) + w_k| &\le c_w \sigma + \sup_{\tilde{\boldsymbol{\theta}} \in [\boldsymbol{\theta}, \boldsymbol{\theta}_\star]} \|\nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \tilde{\boldsymbol{\theta}})\|_{\ell_2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}, \\
&\le c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2} \quad \text{for all} \quad 1 \le k \le n, \qquad (10.11) \\
\implies \|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} &\le \sqrt{n} \max_{1 \le k \le n} |\tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) + w_k|, \\
&\le \sqrt{n}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \qquad (10.12)
\end{aligned}
$$

This further implies that, with probability at least $1 - p_0$, we have

$$
\begin{aligned}
&\frac{1}{2} |\|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} + \|\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}| \\
&\qquad \le \sqrt{n}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \qquad (10.13)
\end{aligned}
$$

To conclude, applying triangle inequality and using Mean-value Theorem, the difference term $\Delta := |\|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} - \|\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}|$ is bounded as follows,

$$\begin{aligned}
\Delta &\le \|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}, \\
&\le \|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} + \|\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star)\|_{\ell_2}, \\
&\le \sup_{\tilde{\boldsymbol{h}} \in [\boldsymbol{h}, \bar{\boldsymbol{h}}]} \|\nabla_{\boldsymbol{h}} \tilde{\phi}(\tilde{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\| \|\boldsymbol{h} - \bar{\boldsymbol{h}}\|_{\ell_2} + \sup_{\tilde{\boldsymbol{h}} \in [\boldsymbol{h}, \bar{\boldsymbol{h}}]} \|\nabla_{\boldsymbol{h}} \tilde{\phi}(\tilde{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star)\| \|\boldsymbol{h} - \bar{\boldsymbol{h}}\|_{\ell_2}, \\
&\overset{\text{(a)}}{\le} B_{\tilde{\phi}} C_\rho \rho^{L-1} \beta_+ \sqrt{n} + B_{\tilde{\phi}} C_\rho \rho^{L-1} \beta_+ \sqrt{n}, \\
&= 2 B_{\tilde{\phi}} C_\rho \rho^{L-1} \beta_+ \sqrt{n}, \qquad (10.14)
\end{aligned}$$

with probability at least $1 - p_0$, where we get (a) from (4.2) and the initial assumption that $\|\nabla_{\boldsymbol{h}} \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})\| \le B_{\tilde{\phi}}$. Multiplying this bound with (10.13) yields the advertised bound on the loss difference.

• **Convergence of gradients:** Next, we take the gradients of $\hat{\mathcal{L}}(\boldsymbol{\theta})$ and $\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})$ to bound Euclidean distance between them. We begin with

$$\begin{aligned}
\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})\|_{\ell_2} \le \frac{1}{T-L} \sum_{t=L}^{T-1} &\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})^\top (\tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta}) - \boldsymbol{h}_{t+1}) \\
&- \nabla_{\boldsymbol{\theta}} \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})^\top (\tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}) - \boldsymbol{h}_{t+1,L})\|_{\ell_2}, \\
\le \max_{L \le t \le (T-1)} &\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta})^\top (\tilde{\phi}(\boldsymbol{h}_t, \boldsymbol{z}_t; \boldsymbol{\theta}) - \boldsymbol{h}_{t+1}) \\
&- \nabla_{\boldsymbol{\theta}} \tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})^\top (\tilde{\phi}(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}) - \boldsymbol{h}_{t+1,L})\|_{\ell_2}, \\
\le &\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})^\top (\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \boldsymbol{w}) \\
&- \nabla_{\boldsymbol{\theta}} \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})^\top (\tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \boldsymbol{w})\|_{\ell_2}, \\
\le &\sqrt{n} \Lambda, \qquad (10.15)
\end{aligned}$$

where $(\boldsymbol{h}, \bar{\boldsymbol{h}}, \boldsymbol{z}, \boldsymbol{w})$ corresponds to the maximum index ($\bar{\boldsymbol{h}}$ be the truncated state) and we define Λ to be the entry-wise maximum

$$\begin{aligned}
\Lambda := \max_{1 \le k \le n} &\|(\tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - w_k) \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta}) \\
&- (\tilde{\phi}_k(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}_k(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - w_k) \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\bar{\boldsymbol{h}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}, \qquad (10.16)
\end{aligned}$$

where $\tilde{\phi}_k(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})$ denotes the k_{th} element of $\tilde{\phi}(\boldsymbol{h}, \boldsymbol{z}; \boldsymbol{\theta})$. Without losing generality, suppose k is the coordinate achieving maximum value and attaining Λ. Note that $\Lambda = \alpha(\boldsymbol{h}) - \alpha(\bar{\boldsymbol{h}})$ for some function α, hence, using Mean-value Theorem as previously, we bound $\Lambda \le$

$\sup_{\tilde{\boldsymbol{h}}\in[\boldsymbol{h},\bar{\boldsymbol{h}}]} \|\nabla_{\boldsymbol{h}}\alpha(\tilde{\boldsymbol{h}})\|\|\boldsymbol{h}-\bar{\boldsymbol{h}}\|_{\ell_2}$ as follows,

$$\begin{aligned}
\Lambda &\le \sup_{\tilde{\boldsymbol{h}}\in[\boldsymbol{h},\bar{\boldsymbol{h}}]} \|(\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) - \tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star) - w_k)\nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) \\
&\qquad + \nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta})(\nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta})^\top - \nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star)^\top)\|\|\boldsymbol{h}-\bar{\boldsymbol{h}}\|_{\ell_2}, \\
&\le \sup_{\tilde{\boldsymbol{h}}\in[\boldsymbol{h},\bar{\boldsymbol{h}}]} \big[|\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) - \tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star) - w_k|\|\nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta})\| \\
&\qquad + \|\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta})\|_{\ell_2}\|\nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) - \nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star)\|_{\ell_2}\big]\|\boldsymbol{h}-\bar{\boldsymbol{h}}\|_{\ell_2}, \\
&\overset{(a)}{\le} \sup_{\tilde{\boldsymbol{h}}\in[\boldsymbol{h},\bar{\boldsymbol{h}}]} \big[D_{\tilde{\phi}}|\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) - \tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star) - w_k| \\
&\qquad + C_{\tilde{\phi}}\|\nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) - \nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star)\|_{\ell_2}\big]\|\boldsymbol{h}-\bar{\boldsymbol{h}}\|_{\ell_2}, \qquad (10.17)
\end{aligned}$$

where we get (a) from the initial assumptions $\|\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\boldsymbol{h},\boldsymbol{z};\boldsymbol{\theta})\|_{\ell_2} \le C_{\tilde{\phi}}$ and $\|\nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\boldsymbol{h},\boldsymbol{z};\boldsymbol{\theta})\| \le D_{\tilde{\phi}}$. To proceed, again using Mean-value Theorem, we obtain

$$\begin{aligned}
\sup_{\tilde{\boldsymbol{h}}\in[\boldsymbol{h},\bar{\boldsymbol{h}}]} \|\nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}) - \nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\boldsymbol{\theta}_\star)\|_{\ell_2} &\le \sup_{\substack{\tilde{\boldsymbol{h}}\in[\boldsymbol{h},\bar{\boldsymbol{h}}]\\ \tilde{\boldsymbol{\theta}}\in[\boldsymbol{\theta},\boldsymbol{\theta}_\star]}} \|\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{h}}\tilde{\phi}_k(\tilde{\boldsymbol{h}},\boldsymbol{z};\tilde{\boldsymbol{\theta}})\|\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2}, \\
&\le D_{\tilde{\phi}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2}. \qquad (10.18)
\end{aligned}$$

Finally, plugging the bounds from (10.11) and (10.18) into (10.17), with probability at least $1-p_0$, we have

$$\begin{aligned}
\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta})\|_{\ell_2} &\le \sqrt{n}\Lambda, \\
&\le \sqrt{n}(D_{\tilde{\phi}}(c_w\sigma + C_{\tilde{\phi}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2}) + C_{\tilde{\phi}}D_{\tilde{\phi}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2})\|\boldsymbol{h}-\bar{\boldsymbol{h}}\|_{\ell_2}, \\
&\le 2n\beta_+C_\rho\rho^{L-1}D_{\tilde{\phi}}(c_w\sigma + C_{\tilde{\phi}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2}), \qquad (10.19)
\end{aligned}$$

This completes the proof. ■

10.5 Proof of Theorem 12

Proof Theorem 12 is a direct consequence of combining the results from Sections 3 and 4. To begin our proof, consider the truncated sub-trajectory loss $\hat{\ell}_\tau^{\mathrm{tr}}$ from Definition 9 which also implies that $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta})]$. Hence, $\hat{\ell}_\tau^{\mathrm{tr}}$ it is a finite sample approximation of the Auxiliary loss $\mathcal{L}_{\mathcal{D}}$. To proceed, using Theorem 4 with Assumptions 4 and 5 on the Auxiliary loss $\mathcal{L}_{\mathcal{D}}$ and its finite sample approximation $\hat{\ell}_\tau^{\mathrm{tr}}$, with probability at least $1 - Lp_0 - L\log(\frac{Kr}{\sigma_0})\exp(-100d)$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have

$$\|\nabla\hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le c_0(\sigma_0 + K\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}, \qquad (10.20)$$

for all $0 \le \tau \le L-1$, where we get the advertised probability by union bounding over all $0 \le \tau \le L-1$. Next, observe that the truncated loss $\hat{\mathcal{L}}^{\mathrm{tr}}$ can be split into (average of)

L sub-trajectory losses via $\hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\tau=0}^{L-1}\hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta})$. This implies that, with probability at least $1 - Lp_0 - L\log(\frac{Kr}{\sigma_0})\exp(-100d)$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have

$$\begin{aligned}\|\nabla\hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} &\le \frac{1}{L}\sum_{\tau=0}^{L-1}\|\nabla\hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2},\\ &\le \max_{0\le\tau\le(L-1)}\|\nabla\hat{\ell}_\tau^{\mathrm{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2},\\ &\le c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}. \end{aligned} \tag{10.21}$$

Combining this with Theorem 11, with the advertised probability, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have

$$\begin{aligned}&\|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}\\ &\quad\le \|\hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} + \|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta})\|_{\ell_2},\\ &\quad\le c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N} + 2n\beta_+C_\rho\rho^{L-1}D_{\tilde{\phi}}(c_w\sigma + C_{\tilde{\phi}}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}).\end{aligned}$$

To simplify the result further, we pick L to be large enough so that the second term in the above inequality becomes smaller than or equal to the first one. This is possible when

$$\begin{aligned}&2n\beta_+C_\rho\rho^{L-1}D_{\tilde{\phi}} \le c_0(\sigma_0/c_w\sigma \wedge K/C_{\tilde{\phi}})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N},\\ \iff\quad &\rho^{L-1} \le (\sigma_0/c_w\sigma \wedge K/C_{\tilde{\phi}})\frac{c_0\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}}{2n\beta_+C_\rho D_{\tilde{\phi}}},\\ \iff\quad &L \ge 1 + \Big[\log\Big(\frac{2n\beta_+C_\rho D_{\tilde{\phi}}\sqrt{N/d}}{c_0\log(3(L_{\mathcal{D}}N/K+1))}\Big) + \log(c_w\sigma/\sigma_0 \vee C_{\tilde{\phi}}/K)\Big]/\log(\rho^{-1}),\\ \impliedby\quad &L \ge \Big\lceil 1 + \frac{\log((2/c_0)n\beta_+C_\rho D_{\tilde{\phi}}\sqrt{N/d}(c_w\sigma/\sigma_0 \vee C_{\tilde{\phi}}/K))}{1-\rho}\Big\rceil.\end{aligned} \tag{10.22}$$

Hence, picking L via (10.22), with probability at least $1 - 2Lp_0 - L\log(\frac{Kr}{\sigma_0})\exp(-100d)$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le 2c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}. \tag{10.23}$$

This completes the proof. ■

10.6 Proof of Theorem 13

Before we begin the proof, we state a theorem to show the linear convergence of gradient descent for minimizing an empirical loss $\hat{\mathcal{L}}$ when the population loss $\mathcal{L}_{\mathcal{D}}$ satisfies one-point convexity and the Euclidean distance between the gradients of the two losses is upper bounded as follows: $\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le \nu + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$.

Theorem 19 (OPCS convergence) *Suppose Assumption 3 holds. Assume for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, $\nabla\hat{\mathcal{L}}$ satisfies $\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le \nu + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$ and $r \ge 5\nu/\alpha$. Set learning rate $\eta = \alpha/(16\beta^2)$ and pick $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. All gradient descent iterates $\boldsymbol{\theta}_\tau$ on $\hat{\mathcal{L}}$ satisfy*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \le \big(1 - \frac{\alpha^2}{128\beta^2}\big)^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{5\nu}{\alpha}. \tag{10.24}$$

Proof Set $\boldsymbol{\delta}_\tau = \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star$. At a given iteration τ we have that $\boldsymbol{\delta}_{\tau+1} = \boldsymbol{\delta}_\tau - \eta\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)$ which implies

$$\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2}^2 = \|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - 2\eta\left\langle \boldsymbol{\delta}_\tau, \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\right\rangle + \eta^2\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2. \tag{10.25}$$

Using Assumptions 3 and $\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le \nu + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$, we have that

$$\begin{aligned}\left\langle \boldsymbol{\delta}_\tau, \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\right\rangle &\ge \langle \boldsymbol{\delta}_\tau, \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\rangle - |\left\langle \boldsymbol{\delta}_\tau, \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\right\rangle|,\\ &\ge \alpha\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - (\nu + (\alpha/2)\|\boldsymbol{\delta}_\tau\|_{\ell_2})\|\boldsymbol{\delta}_\tau\|_{\ell_2} \ge (\alpha/2)\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - \nu\|\boldsymbol{\delta}_\tau\|_{\ell_2}.\end{aligned} \tag{10.26}$$

Similarly,

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le \|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\|_{\ell_2} + \|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le (3/2)\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu. \tag{10.27}$$

Suppose $\|\boldsymbol{\delta}_\tau\|_{\ell_2} \ge 4\nu/\alpha$. Then, $(\alpha/2)\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - \nu\|\boldsymbol{\delta}_\tau\|_{\ell_2} \ge (\alpha/4)\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2$ and $(3/2)\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu \le 2\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2}$. Hence, using the learning rate $\eta = \frac{\alpha}{16\beta^2}$, we obtain

$$\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2}^2 \le \|\boldsymbol{\delta}_\tau\|_{\ell_2}^2(1 - \eta\alpha/2 + 4\eta^2\beta^2) \le (1 - \frac{\alpha^2}{64\beta^2})\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2.$$

Now, imagine the scenario $\|\boldsymbol{\delta}_\tau\|_{\ell_2} \le 4\nu/\alpha$. We would like to prove that $\boldsymbol{\delta}_{\tau+1}$ satisfies a similar bound namely $\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2} \le 5\nu/\alpha$. This is shown as follows.

$$\begin{aligned}\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2}^2 &\le \|\boldsymbol{\delta}_\tau\|_{\ell_2}^2(1 - \eta\alpha + (9/4)\eta^2\beta^2) + 2\eta\nu\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \eta^2(3\nu\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu^2),\\ &\le (1 - \frac{3\alpha^2}{64\beta^2})\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 + \frac{\alpha}{8\beta^2}\nu\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \frac{\alpha^2}{256\beta^4}(3\nu\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu^2),\\ &\le (\frac{16}{\alpha^2} + \frac{1}{2\beta^2} + \frac{3\alpha}{64\beta^3} + \frac{\alpha^2}{256\beta^4})\nu^2 \le \frac{25}{\alpha^2}\nu^2,\end{aligned}$$

which implies $\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2} \le 5\nu/\alpha$. To get the final result observe that during initial iterations, as long as $\|\boldsymbol{\delta}_\tau\|_{\ell_2} \ge 4\nu/\alpha$, we have

$$\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 \le (1 - \frac{\alpha^2}{64\beta^2})^\tau\|\boldsymbol{\delta}_0\|_{\ell_2}^2 \implies \|\boldsymbol{\delta}_\tau\|_{\ell_2} \le (1 - \frac{\alpha^2}{128\beta^2})^\tau\|\boldsymbol{\delta}_0\|_{\ell_2}.$$

After the first instance $\|\boldsymbol{\delta}_\tau\|_{\ell_2} < 4\nu/\alpha$, iterations will never violate $\|\boldsymbol{\delta}_\tau\|_{\ell_2} \le 5\nu/\alpha$. The reason is

- If $\|\boldsymbol{\delta}_\tau\|_{\ell_2} < 4\nu/\alpha$: we can only go up to $5\nu/\alpha$ and $\boldsymbol{\delta}_{\tau+1} \le 5\nu/\alpha$.
- If $4\nu/\alpha \le \|\boldsymbol{\delta}_\tau\|_{\ell_2} \le 5\nu/\alpha$: we have to go down hence $\boldsymbol{\delta}_{\tau+1} \le 5\nu/\alpha$.

■

Proof The proof of Theorem 13 readily follows from combining our gradient convergence result (i.e., Theorem 12) with Theorem 19. We begin by picking $N \geq 16c_0^2K^2\log^2(3(L_{\mathcal{D}}N/K+1))d/\alpha^2$ in Theorem 12 to obtain

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq (\alpha/2)\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2} + 2c_0\sigma_0\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}, \tag{10.28}$$

with probability at least $1-2Lp_0 - L\log(\frac{Kr}{\sigma_0})\exp(-100d)$ for all $\boldsymbol{\theta}\in\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. We then use Theorem 19 with $\nu = 2c_0\sigma_0\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}$ and set $c = 10c_0$ to get the statement of the theorem. Lastly, observe that by choosing $N \geq 16c_0^2K^2\log^2(3(L_{\mathcal{D}}N/K+1))d/\alpha^2$, the statistical error rate of our non-asymptotic identification can be upper bounded as follows,

$$\frac{5\nu}{\alpha} = \frac{10c_0\sigma_0}{\alpha}\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N} \lesssim \sigma_0/K. \tag{10.29}$$

Therefore, to ensure that Theorem 19 is applicable, we assume that the noise is small enough, so that $\sigma_0 \lesssim rK$. This completes the proof. ■

10.7 Proof of Theorem 14

Proof Our proof strategy is similar to that of Theorem 13, that is, we first show the gradient convergence result for each component $\hat{\mathcal{L}}_k$ of the empirical loss $\hat{\mathcal{L}}$. We then use Theorem 19 to learn the dynamics of separable dynamical systems using finite samples obtained from a single trajectory.

• **Uniform gradient convergence:** In the case of separable dynamical systems, Assumption 4 states that, there exist numbers $L_{\mathcal{D}}, p_0 > 0$ such that with probability at least $1-p_0$ over the generation of data, for all pairs $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, the gradients of empirical and population losses in (5.7) satisfy

$$\max(\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2}, \|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2}) \leq L_{\mathcal{D}}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}, \tag{10.30}$$

for all $1\leq k\leq n$. Similarly, Assumption 5 states that, there exist scalars $K, \sigma_0 > 0$ such that, given $\boldsymbol{x}\sim\mathcal{D}$, at any point $\boldsymbol{\theta}$, the subexponential norm of the gradient is upper bounded as a function of the noise level σ_0 and distance to the population minimizer via

$$\|\nabla\mathcal{L}_k(\boldsymbol{\theta}_k,\boldsymbol{x}) - \mathbb{E}[\nabla\mathcal{L}_k(\boldsymbol{\theta}_k,\boldsymbol{x})\|_{\psi_1} \leq \sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} \quad \text{for all} \quad 1\leq k\leq n. \tag{10.31}$$

To proceed, using Theorem 4 with Assumptions 4 and 5 replaced by (10.30) and (10.31) respectively, with probability at least $1-np_0 - n\log(\frac{Kr}{\sigma_0})\exp(-100\bar{d})$, for all $\boldsymbol{\theta}\in\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1\leq k\leq n$, we have

$$\|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \leq c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{\bar{d}/N}. \tag{10.32}$$

• **Small impact of truncation:** Next, we relate the gradients of the single trajectory loss $\hat{\mathcal{L}}_k$ in (5.5) and the multiple trajectory loss $\hat{\mathcal{L}}_k^{\text{tr}}$ (defined below). Similar to (5.5), the truncated loss for separable dynamical systems is alternately given by

$$\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k), \text{where } \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} (\boldsymbol{h}_{t+1,L}[k] - \tilde{\phi}_k(\boldsymbol{h}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}_k))^2, \quad (10.33)$$

where $\boldsymbol{h}_{t,L}[k]$ denotes the k_{th} element of the truncated vector $\boldsymbol{h}_{t,L}$. We remark that Assumptions 1 and 2 are same for both non-separable and separable dynamical systems. Therefore, repeating the same proof strategy of Theorem 11, with $\hat{\mathcal{L}}^{\text{tr}}$ and $\hat{\mathcal{L}}$ replaced by $\hat{\mathcal{L}}_k^{\text{tr}}$ and $\hat{\mathcal{L}}_k$ respectively, with probability at least $1 - np_0$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1 \le k \le n$, we have

$$\|\nabla \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k)\|_{\ell_2} \le 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}). \quad (10.34)$$

• **Combined result:** Next, we combine (10.32) and (10.34) to obtain a uniform convergence result for the gradient of the empirical loss $\hat{\mathcal{L}}_k$. Observe that, similar to $\hat{\mathcal{L}}^{\text{tr}}$, the truncated loss $\hat{\mathcal{L}}_k^{\text{tr}}$ can also be split into L truncated sub-trajectory losses (see the proof of Theorem 12). Each of these truncated sub-trajectory loss is identically distributed as $\hat{\mathcal{L}}_{k,\mathcal{S}}$. Therefore, using a similar line of reasoning as we did in the proof of Theorem 12, with probability at least $1 - Lnp_0 - Ln\log(\frac{Kr}{\sigma_0})\exp(-100\bar{d})$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1 \le k \le n$, we have

$$\|\nabla \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k) - \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \le c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}) \log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}. \quad (10.35)$$

Combining this with (10.34), with probability at least $1 - Lnp_0 - Ln\log(\frac{Kr}{\sigma_0})\exp(-100\bar{d})$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1 \le k \le n$, we have

$$\begin{aligned}
&\|\nabla \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \\
&\le \|\nabla \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k) - \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} + \|\nabla \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k)\|_{\ell_2}, \\
&\le c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}) \log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N} + 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}).
\end{aligned}$$

To simplify the result further, we pick L to be large enough so that the second term in the above inequality becomes smaller than or equal to the first one. This is possible when

$$L \ge \Big\lceil 1 + \frac{\log((2/c_0)n\beta_+ C_\rho D_{\tilde{\phi}} \sqrt{N/\bar{d}}(c_w\sigma/\sigma_0 \vee C_{\tilde{\phi}}/K))}{1-\rho} \Big\rceil. \quad (10.36)$$

Hence, picking L as above, with probability at least $1 - 2Lnp_0 - Ln\log(\frac{Kr}{\sigma_0})\exp(-100\bar{d})$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1 \le k \le n$, we have

$$\begin{aligned}
\|\nabla \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &\le 2c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}) \log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}, \\
&\overset{(a)}{\le} (\alpha/2)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + 2c_0\sigma_0 \log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}, \quad (10.37)
\end{aligned}$$

where we get (a) by choosing $N \geq 16c_0^2K^2\log^2(3(L_{\mathcal{D}}N/K+1))\bar{d}/\alpha^2$.

• **One-point convexity & smoothness:** Lastly, Assumption 3 on the Auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ states that, there exist scalars $\beta \geq \alpha > 0$ such that, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1 \leq k \leq n$, the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$ of (5.7) satisfies

$$\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\right\rangle \geq \alpha\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2, \tag{10.38}$$

$$\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \leq \beta\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}. \tag{10.39}$$

• **Finalizing the proof:** We are now ready to use Theorem 19 with gradient concentration bound given by (10.37) and the OPCS Assumptions given by (10.38) and (10.39). Specifically, we use Theorem 19 with $\nu = 2c_0\sigma_0\log(3(L_{\mathcal{D}}N/K+1))\sqrt{\bar{d}/N}$, the one-point convexity assumption (10.38) and the one-point smoothness assumption (10.39) to get the statement of the theorem. This completes the proof. ■

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 6676–6688, 2019.

Karl Johan Åström and Peter Eykhoff. System identification—a survey. *Automatica*, 7(2): 123–162, 1971.

Karl Johan Åström and Tore Hägglund. *PID controllers: theory, design, and tuning*, volume 2. Instrument Society of America Research Triangle Park, NC, 1995.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Sohail Bahmani and Justin Romberg. Convex programming for estimation in nonlinear recurrent models. *Journal of Machine Learning Research*, 21(235):1–20, 2020.

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.

Sheng Chen, SA Billings, and PM Grant. Non-linear system identification using neural networks. *International Journal of Control*, 51(6):1191–1214, 1990.

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

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

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

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 66(4):1802–1808, 2020.

Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control*, pages 2682–2689. IEEE, 2019.

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*, volume 80, pages 1467–1476. PMLR, 2018.

Dylan Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8745–8756, 2018.

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.

Sara A Geer, Sara van de Geer, and D Williams. *Empirical processes in M-estimation*, volume 6. Cambridge university press, 2000.

Alex Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE, 2013.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, volume 30, 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*, volume 31, pages 4639–4648, 2018.

BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Prateek Jain, Suhas S Kowshik, Dheeraj Nagaraj, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 34, pages 8518–8531, 2021.

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

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*, volume 24, pages 927–935, 2011.

Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A Salman Avestimehr. Fitting relus via sgd and quantized sgd. In *2019 IEEE International Symposium on Information Theory*, pages 2469–2473, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Mohammad Khosravi and Roy S Smith. Convex nonparametric formulation for identification of gradient flows. *IEEE Control Systems Letters*, 5(3):1097–1102, 2020a.

Mohammad Khosravi and Roy S Smith. Nonlinear system identification with prior knowledge on the region of attraction. *IEEE Control Systems Letters*, 5(3):1091–1096, 2020b.

Karl Krauth, Stephen Tu, and Benjamin Recht. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, volume 32, pages 8514–8524, 2019.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control*, pages 757–762. IEEE, 2021.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Shuai Li, Sanfeng Chen, and Bo Liu. Accelerating a recurrent neural network to finite-time convergence for solving time-varying sylvester equation by using a sign-bi-power activation function. *Neural processing Letters*, 37(2):189–205, 2013.

Lennart Ljung. System identification. In *Signal Analysis and Prediction*, pages 163–173. Springer, 1998.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925, 2019.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, volume 32, pages 10154–10164, 2019.

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.

Daniel J McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Nonparametric risk bounds for time-series forecasting. *The Journal of Machine Learning Research*, 18(1):1044–1083, 2017.

Alexandre Megretski and Anders Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997.

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

Zakaria Mhammedi, Dylan J Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, and John Langford. Learning the linear quadratic regulator from nonlinear observations. In *Advances in Neural Information Processing Systems*, volume 33, pages 14532–14543, 2020.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations*, 2019.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, volume 20, pages 1025–1032, 2007.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, volume 21, pages 1097–1104, 2008.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Samet Oymak. Learning compact neural networks with regularization. In *International Conference on Machine Learning*, pages 3966–3975. PMLR, 2018.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. In *Conference on Learning Theory*, pages 2551–2579. PMLR, 2019.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference*, pages 5655–5661. IEEE, 2019.

Rik Pintelon and Johan Schoukens. *System identification: a frequency domain approach*. John Wiley & Sons, 2012.

Stephen Prajna, Antonis Papachristodoulou, and Pablo A Parrilo. Introducing sostools: A general purpose sum of squares programming solver. In *2002 41st IEEE Conference on Decision and Control*, volume 1, pages 741–746. IEEE, 2002.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

Tuhin Sarkar, Alexander Rakhlin, and Munther Dahleh. Nonparametric system identification of stochastic switched linear systems. In *2019 IEEE 58th Conference on Decision and Control*, pages 3623–3628. IEEE, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time lti system identification. *Journal of Machine Learning Research*, 22:1–61, 2021.

Yahya Sattar and Samet Oymak. A simple framework for learning stabilizable systems. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 116–120. IEEE, 2019.

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.

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

Sumeet Singh, Spencer M Richards, Vikas Sindhwani, Jean-Jacques E Slotine, and Marco Pavone. Learning stabilizable nonlinear dynamics with contraction-based regularization. *The International Journal of Robotics Research*, 40(10-11):1123–1150, 2021.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control*, pages 3648–3654. IEEE, 2019.

Anastasios Tsiamis, Nikolai Matni, and George Pappas. Sample complexity of kalman filtering for unknown systems. In *Learning for Dynamics and Control*, pages 435–444. PMLR, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, page 210–268. Cambridge University Press, 2012.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Greg Welch and Gary Bishop. *An Introduction to the Kalman Filter*. University of North Carolina at Chapel Hill, USA, 1995.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 32, pages 8668–8678, 2019.

Appendix A. Proof of Corollaries 15 and 16

A.1 Application to Linear Dynamical Systems

A.1.1 Verification of Assumption 1

The following lemma states that a linear dynamical system satisfies (C_ρ, ρ)-stability if the spectral radius $\rho(\boldsymbol{A}_\star) < 1$.

Lemma 20 ((C_ρ, ρ)-stability) *Fix excitations $(\boldsymbol{z}_t)_{t=0}^{\infty}$ and noise $(\boldsymbol{w}_t)_{t=0}^{\infty}$. Denote the state sequence (6.1) ($\phi = \boldsymbol{I}_n$) resulting from initial state $\boldsymbol{h}_0 = \boldsymbol{\alpha}$, $(\boldsymbol{z}_\tau)_{\tau=0}^{t}$ and $(\boldsymbol{w}_\tau)_{\tau=0}^{t}$ by $\boldsymbol{h}_t(\boldsymbol{\alpha})$. Suppose $\rho(\boldsymbol{A}_\star) < 1$. Then, there exists $C_\rho \geq 1$ and $\rho \in (\rho(\boldsymbol{A}_\star), 1)$ such that $\|\boldsymbol{h}_t(\boldsymbol{\alpha}) - \boldsymbol{h}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2}$.*

Proof To begin, consider the difference,

$$\boldsymbol{h}_t(\boldsymbol{\alpha}) - \boldsymbol{h}_t(0) = \boldsymbol{A}_\star \boldsymbol{h}_{t-1}(\boldsymbol{\alpha}) + \boldsymbol{B}_\star \boldsymbol{z}_{t-1} - \boldsymbol{A}_\star \boldsymbol{h}_{t-1}(0) - \boldsymbol{B}_\star \boldsymbol{z}_{t-1} = \boldsymbol{A}_\star (\boldsymbol{h}_{t-1}(\boldsymbol{\alpha}) - \boldsymbol{h}_{t-1}(0)).$$

Repeating this recursion till $t = 0$ and taking the norm, we get

$$\|\boldsymbol{h}_t(\boldsymbol{\alpha}) - \boldsymbol{h}_t(0)\|_{\ell_2} = \|\boldsymbol{A}_\star^t(\boldsymbol{\alpha} - 0)\|_{\ell_2} \leq \|\boldsymbol{A}_\star^t\| \|\boldsymbol{\alpha}\|_{\ell_2}. \tag{A.1}$$

Given $\rho(\boldsymbol{A}_\star) < 1$, as a consequence of Gelfand's formula, there exists $C_\rho \geq 1$ and $\rho \in (\rho(\boldsymbol{A}_\star), 1)$ such that, $\|\boldsymbol{A}_\star^t\| \leq C_\rho \rho^t$, for all $t \geq 0$. Hence, $\|\boldsymbol{h}_t(\boldsymbol{\alpha}) - \boldsymbol{h}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2}$. This completes the proof. ■

A.1.2 Verification of Assumption 2

To show that the states of a stable linear dynamical system are bounded with high probability, we state a standard Lemma from Oymak (2019) that bounds the Euclidean norm of a subgaussian vector.

Lemma 21 *Let $\boldsymbol{a} \in \mathbb{R}^n$ be a zero-mean subgaussian random vector with $\|\boldsymbol{a}\|_{\psi_2} \le L$. Then for any $m \ge n$, there exists $C > 0$ such that*

$$\mathbb{P}(\|\boldsymbol{a}\|_{\ell_2} \le CL\sqrt{m}) \ge 1 - 2\exp(-100m). \tag{A.2}$$

To apply Lemma 21, we require the subgaussian norm of the state vector $\boldsymbol{h}_t$ and the concatenated vector $\boldsymbol{x}_t$. We will do that by first bounding the corresponding covariance matrices as follows.

Theorem 22 (Covariance bounds) *Consider the LDS in (6.1) with $\phi = \boldsymbol{I}_n$. Suppose $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ and $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. Let $\boldsymbol{G}_t$ and $\boldsymbol{F}_t$ be as in (6.4). Then, the covariance matrix of the vectors $\boldsymbol{h}_t$ and $\boldsymbol{x}_t = [\boldsymbol{h}_t^\top \ \boldsymbol{z}_t^\top]^\top$ satisfies*

$$\lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_n \preceq \boldsymbol{\Sigma}[\boldsymbol{h}_t] \preceq \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_n, \tag{A.3}$$

$$(1 \wedge \lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p} \preceq \boldsymbol{\Sigma}[\boldsymbol{x}_t] \preceq (1 \vee \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p}, \tag{A.4}$$

Proof We first expand the state vector $\boldsymbol{h}_t$ as a sum of two independent components $\boldsymbol{g}_t$ and $\boldsymbol{\omega}_t$ as follows,

$$\boldsymbol{h}_t = \underbrace{\sum_{i=0}^{t-1} \boldsymbol{A}_\star^{t-1-i}\boldsymbol{B}_\star\boldsymbol{z}_i}_{\boldsymbol{g}_t} + \underbrace{\sum_{i=0}^{t-1} \boldsymbol{A}_\star^{t-1-i}\boldsymbol{w}_i}_{\boldsymbol{\omega}_t}. \tag{A.5}$$

Observe that, $\boldsymbol{g}_t$ denotes the state evolution due to control input and $\boldsymbol{\omega}_t$ denotes the state evolution due to noise. Furthermore, $\boldsymbol{g}_t$ and $\boldsymbol{\omega}_t$ are both independent and zero-mean. Therefore, we have

$$\begin{aligned}
\boldsymbol{\Sigma}[\boldsymbol{h}_t] &= \boldsymbol{\Sigma}[\boldsymbol{g}_t + \boldsymbol{\omega}_t] = \boldsymbol{\Sigma}[\boldsymbol{g}_t] + \boldsymbol{\Sigma}[\boldsymbol{\omega}_t] = \mathbb{E}[\boldsymbol{g}_t\boldsymbol{g}_t^\top] + \mathbb{E}[\boldsymbol{\omega}_t\boldsymbol{\omega}_t^\top] \\
&= \sum_{i=0}^{t-1}\sum_{j=0}^{t-1}(\boldsymbol{A}_\star^i)\boldsymbol{B}_\star\,\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_j^\top]\boldsymbol{B}_\star^\top(\boldsymbol{A}_\star^j)^\top + \sum_{i=0}^{t-1}\sum_{j=0}^{t-1}(\boldsymbol{A}_\star^i)\,\mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_j^\top](\boldsymbol{A}_\star^j)^\top \\
&\overset{(a)}{=} \sum_{i=0}^{t-1}(\boldsymbol{A}_\star^i)\boldsymbol{B}_\star\boldsymbol{B}_\star^\top(\boldsymbol{A}_\star^i)^\top + \sigma^2\sum_{i=0}^{t-1}(\boldsymbol{A}_\star^i)(\boldsymbol{A}_\star^i)^\top,
\end{aligned} \tag{A.6}$$

where we get (a) from the fact that $\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_j^\top] = \boldsymbol{I}_p$ and $\mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_j^\top] = \sigma^2\boldsymbol{I}_n$ when $i = j$, and zero otherwise. To proceed, let $\boldsymbol{G}_t := [\boldsymbol{A}_\star^{t-1}\boldsymbol{B}_\star \ \boldsymbol{A}_\star^{t-2}\boldsymbol{B}_\star \cdots \boldsymbol{B}_\star]$ and $\boldsymbol{F}_t := [\boldsymbol{A}_\star^{t-1} \ \boldsymbol{A}_\star^{t-2} \cdots \boldsymbol{I}_n]$. Observing $\boldsymbol{G}_t\boldsymbol{G}_t^\top = \sum_{i=0}^{t-1}(\boldsymbol{A}_\star^i)\boldsymbol{B}_\star\boldsymbol{B}_\star^\top(\boldsymbol{A}_\star^i)^\top$ and $\boldsymbol{F}_t\boldsymbol{F}_t^\top = \sum_{i=0}^{t-1}(\boldsymbol{A}_\star^i)(\boldsymbol{A}_\star^i)^\top$, we obtain the following bounds on the covariance matrix of the state vector $\boldsymbol{h}_t$ and the concatenated vector $\boldsymbol{x}_t = [\boldsymbol{h}_t^\top \ \boldsymbol{z}_t^\top]^\top$.

$$\lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_n \preceq \boldsymbol{\Sigma}[\boldsymbol{h}_t] \preceq \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_n, \tag{A.7}$$

$$(1 \wedge \lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p} \preceq \boldsymbol{\Sigma}[\boldsymbol{x}_t] \preceq (1 \vee \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p}, \tag{A.8}$$

where to get the second relation, we use the fact that $\boldsymbol{\Sigma}[\boldsymbol{z}_t] = \boldsymbol{I}_p$. This completes the proof. ■

Once we bound the covariance matrices, using standard bounds on the subgaussian norm

of a random vector, we find that $\|\boldsymbol{h}_t\|_{\psi_2} \lesssim \sqrt{\boldsymbol{\Sigma}[\boldsymbol{h}_t]} \le \sqrt{\lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)}$ and $\|\boldsymbol{x}_t\|_{\psi_2} \lesssim \sqrt{\boldsymbol{\Sigma}[\boldsymbol{x}_t]} \le 1 \vee \sqrt{\lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)}$. Combining these with Lemma 21, we find that, with probability at least $1 - 4T\exp(-100n)$, for all $1 \le t \le T$, we have $\|\boldsymbol{h}_t\|_{\ell_2} \le c\sqrt{\beta_+ n}$ and $\|\boldsymbol{x}_t\|_{\ell_2} \le c_0\sqrt{\beta_+(n+p)}$, where we set $\beta_+ = 1 \vee \max_{1\le t\le T} \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)$. This verifies Lemma 6 and consequently Assumption 2.

A.1.3 Verification of Assumption 3

Recall that, we define the following concatenated vector/matrix for linear dynamical systems: $\boldsymbol{x}_t = [\boldsymbol{h}_t^\top \ \boldsymbol{z}_t^\top]^\top$ and $\boldsymbol{\Theta}_\star = [\boldsymbol{A}_\star \ \boldsymbol{B}_\star]$. Let $\boldsymbol{\theta}_k^{\star\top}$ denotes the k_{th} row of $\boldsymbol{\Theta}_\star$. Then, the auxiliary loss for linear dynamical system is defined as follows,

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k), \quad \text{where} \quad \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) := \frac{1}{2}\mathbb{E}[(\boldsymbol{h}_L[k] - \boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1})^2]. \tag{A.9}$$

Using the derived bounds on the covariance matrix, it is straightforward to show that the auxiliary loss satisfies the following one-point convexity and smoothness conditions.

Lemma 23 (One-point convexity & smoothness) *Consider the setup of Theorem 22 and the auxiliary loss given by (A.9). Define $\boldsymbol{\Gamma}_t := \boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top$. Let $\gamma_- := 1 \wedge \lambda_{\min}(\boldsymbol{\Gamma}_{L-1})$ and $\gamma_+ := 1 \vee \lambda_{\max}(\boldsymbol{\Gamma}_{L-1})$. For all $1 \le k \le n$, the gradient $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$ satisfies,*

$$\begin{aligned}\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\right\rangle &\ge \gamma_- \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2,\\ \|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &\le \gamma_+ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}.\end{aligned}$$

Proof To begin, we take the gradient of the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ (A.9) to get $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) = \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) - \boldsymbol{x}_{L-1}\boldsymbol{w}_{L-1}[k]]$. Note that, $\mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{w}_{L-1}[k]] = 0$ for linear dynamical systems because $\boldsymbol{w}_{L-1}$ and $\boldsymbol{x}_{L-1}$ are independent and we have $\mathbb{E}[\boldsymbol{w}_{L-1}] = \mathbb{E}[\boldsymbol{x}_{L-1}] = 0$. Therefore, using Theorem 22 with $t = L-1$, we get the following one point convexity bound,

$$\begin{aligned}\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\right\rangle &= \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top](\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)\right\rangle,\\ &\ge \gamma_- \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2.\end{aligned} \tag{A.10}$$

Similarly, we also have

$$\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \le \|\mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top]\|\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le \gamma_+\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}. \tag{A.11}$$

This completes the proof. ■

A.1.4 Verification of Assumption 4

Let $\mathcal{S} := (\boldsymbol{h}_L^{(i)}, \boldsymbol{h}_{L-1}^{(i)}, \boldsymbol{z}_{L-1}^{(i)})_{i=1}^N$ be N i.i.d. copies of $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$ generated from N i.i.d. trajectories of the system (6.1) with $\phi = \boldsymbol{I}_n$. Let $\boldsymbol{x}_{L-1}^{(i)} := [\boldsymbol{h}_{L-1}^{(i)\top} \ \boldsymbol{z}_{L-1}^{(i)\top}]^\top$ and $\boldsymbol{\Theta} := [\boldsymbol{A} \ \boldsymbol{B}]$ be the concatenated vector/matrix. Then, the finite sample approximation of the auxiliary loss $\mathcal{L}_{\mathcal{D}}$ is given by

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k), \quad \text{where} \quad \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) := \frac{1}{2N}\sum_{i=1}^{N}(\boldsymbol{h}_L^{(i)}[k] - \boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1}^{(i)})^2. \tag{A.12}$$

The following lemma states that both $\nabla\mathcal{L}_{k,\mathcal{D}}$ and $\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}$ are Lipschitz with high probability.

Lemma 24 (Lipschitz gradient) *Consider the same setup of Theorem 22. Consider the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ and its finite sample approximation $\hat{\mathcal{L}}_{k,\mathcal{S}}$ from (A.9) and (A.12) respectively. Let $\gamma_+ > 0$ be as in Lemma 23. For $N \gtrsim n + p$, with probability at least $1 - 2\exp(-100(n+p))$, for all pairs $\boldsymbol{\Theta}, \boldsymbol{\Theta}'$ and for all $1 \le k \le n$, we have*

$$\max(\|\nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) - \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2}, \|\nabla \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2}) \le 2\gamma_+ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}. \quad \text{(A.13)}$$

Proof To begin, recall the auxiliary loss from (A.9). We have that

$$\begin{aligned} \|\nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) - \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2} &= \|\mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top](\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) - \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top](\boldsymbol{\theta}'_k - \boldsymbol{\theta}_k^\star)\|_{\ell_2}, \\ &\le \|\,\mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top]\|\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}, \\ &\le \gamma_+ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}. \end{aligned} \quad \text{(A.14)}$$

To obtain a similar result for the finite sample loss $\hat{\mathcal{L}}_{k,\mathcal{S}}$, we use Corollary 5.50 from Vershynin (2012) which bounds the concentration of empirical covariance around its population when the sample size is sufficiently large. Specifically, applying this corollary on the empirical covariance of $\boldsymbol{x}_{L-1}^{(i)}$ with $t = 10, \varepsilon = 1$ shows that, for $N \gtrsim n + p$, with probability at least $1 - 2\exp(-100(n+p))$, we have

$$\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top - \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top]\| \le \gamma_+. \quad \text{(A.15)}$$

Thus, the gradient $\nabla \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k)$ also satisfies the Lipschitz property, that is, for $N \gtrsim n + p$, with probability at least $1 - 2\exp(-100(n+p))$, we have

$$\begin{aligned} &\|\nabla \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2} \\ &\quad\le \|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) - \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top(\boldsymbol{\theta}'_k - \boldsymbol{\theta}_k^\star)\|_{\ell_2}, \\ &\quad\le \|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top\|\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}, \\ &\quad\le \Big[\|\,\mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top]\| + \|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top - \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top]\|\Big]\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}, \\ &\quad\le 2\gamma_+\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}, \end{aligned} \quad \text{(A.16)}$$

for all $1 \le k \le n$. Combining the two results, we get the statement of the lemma. This completes the proof. ■

A.1.5 Verification of Assumption 5

Given a single sample $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$ from the trajectory of a linear dynamical system, setting $\boldsymbol{x}_{L-1} = [\boldsymbol{h}_{L-1}^\top \ \boldsymbol{z}_{L-1}^\top]^\top$, the single sample loss is given by,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta}, (\boldsymbol{h}_L, \boldsymbol{x}_{L-1})) &= \sum_{k=1}^n \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1})), \\ \text{where} \quad \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1})) &:= \frac{1}{2}(\boldsymbol{h}_L[k] - \boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1})^2. \end{aligned} \quad \text{(A.17)}$$

The following lemma shows that the gradient of the above loss is subexponential.

Lemma 25 (Subexponential gradient) *Consider the same setup of Theorem 22. Let* $\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1}))$ *be as defined in (A.17) and* $\gamma_+ > 0$ *be as in lemma 23. Then, at any point* $\boldsymbol{\Theta}$*, for all* $1 \le k \le n$*, we have*

$$\|\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1})) - \mathbb{E}[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1}))]\|_{\psi_1} \lesssim \gamma_+ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma\sqrt{\gamma_+}.$$

Proof Using standard bounds on the subgaussian norm of a random vector, we find that $\|\boldsymbol{x}_{L-1}\|_{\psi_2} \lesssim \sqrt{\boldsymbol{\Sigma}[\boldsymbol{x}_{L-1}]} \le \sqrt{\gamma_+}$, where $\gamma_+ > 0$ is as defined in Lemma 23. Combining this with $\|\boldsymbol{w}_{L-1}[k]\|_{\psi_2} \le \sigma$, we get the following subexponential norm bound,

$$\begin{aligned}
&\|\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1})) - \mathbb{E}[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{x}_{L-1}))]\|_{\psi_1} \\
&\qquad = \|(\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top - \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top])(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) - \boldsymbol{x}_{L-1}\boldsymbol{w}_{L-1}[k]\|_{\psi_1}, \\
&\qquad \le \|(\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top - \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top])(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)\|_{\psi_1} + \|\boldsymbol{x}_{L-1}\boldsymbol{w}_{L-1}[k]\|_{\psi_1}, \\
&\qquad \lesssim \gamma_+ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma\sqrt{\gamma_+},
\end{aligned}$$

where we get the last inequality from the fact that, the product of two subgaussian random variables results in a subexponential random variable with its subexponential norm bounded by the product of the two subgaussian norms. ■

A.1.6 Proof of Corollary 15

Proof Our proof strategy is based on verifying Assumptions 1, 2, 3, 4 and 5 for a stable linear dynamical system and then applying Theorem 14. Since, we already verified all the assumptions, we are ready to use Theorem 14. Before that, we find the values of the system related constants to be used in Theorem 14 as follows.

Remark 26 *Consider the same setup of Theorem 22. For a stable linear dynamical system, with probability at least* $1 - 4T\exp(-100n)$*, for all* $1 \le t \le T$*, the scalars* $C_{\tilde{\phi}}, D_{\tilde{\phi}}$ *take the following values:*

$$\|\nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_t)\|_{\ell_2} = \|\boldsymbol{x}_t\|_{\ell_2} \le c_0\sqrt{\beta_+(n+p)} = C_{\tilde{\phi}}, \tag{A.18}$$

$$\|\nabla_{\boldsymbol{x}_t}\nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_t)\| = \|\boldsymbol{I}_{n+p}\| \le 1 = D_{\tilde{\phi}}, \tag{A.19}$$

where $\beta_+ = 1 \vee \max_{1\le t\le T} \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top)$*. Furthermore, the Lipschitz constant and the gradient noise coefficients take the following values:* $L_{\mathcal{D}} = 2\gamma_+$*,* $K = c\gamma_+$ *and* $\sigma_0 = c\sigma\sqrt{\gamma_+}$*. Lastly, we also have* $p_0 = 2\exp(-100(n+p))$*.*

Using these values, we get the following sample complexity bound for learning linear dynamical system via gradient descent,

$$N \gtrsim \kappa^2 \log^2(3(2\gamma_+)N/\gamma_+ + 3)(n+p) \Leftrightarrow N \gtrsim \kappa^2 \log^2(6N+3)(n+p), \tag{A.20}$$

where $\kappa = \gamma_+/\gamma_-$ is an upper bound on the condition number of the covariance matrix $\mathbf{\Sigma}[\boldsymbol{x}_t]$. Similarly, the approximate mixing time for the linear dynamical system is given by,

$$L \geq 1 + \Big[\log(c_0(n+p)\sqrt{\beta_+}C_\rho\sqrt{N/(n+p)}) + \log(c/\sqrt{\gamma_+} \vee c\sqrt{\beta_+(n+p)}/\gamma_+)\Big]/\log(\rho^{-1})$$
$$\Longleftarrow \quad L \geq \Big\lceil 1 + \frac{\log(CC_\rho\beta_+N(n+p)/\gamma_+)}{1-\rho}\Big\rceil, \tag{A.21}$$

where, $C > 0$ is a constant. Finally, given the trajectory length $T \gtrsim L(N+1)$, where N and L are given by (A.20) and (A.21) respectively, starting from $\mathbf{\Theta}^{(0)} = 0$ and using learning rate $\eta = \gamma_-/(16\gamma_+^2)$ (in Theorem 14), with probability at least $1 - 4T\exp(-100n) - Ln\big(4 + \log(\frac{\|\mathbf{\Theta}_\star\|_F\sqrt{\gamma_+}}{\sigma})\big)\exp(-100(n+p))$ for all $1 \leq k \leq n$, all gradient descent iterates $\mathbf{\Theta}^{(\tau)}$ on $\hat{\mathcal{L}}$ satisfy

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \leq (1 - \frac{\gamma_-^2}{128\gamma_+^2})^\tau\|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{5c}{\gamma_-}\sigma\sqrt{\gamma_+}\log(6N+3)\sqrt{\frac{n+p}{N}}. \tag{A.22}$$

We remark that, choosing $N \gtrsim \kappa^2\log^2(6N+3)(n+p)$, the residual term in (A.22) can be bounded as follows,

$$\frac{5c}{\gamma_-}\sigma\sqrt{\gamma_+}\log(6N+3)\sqrt{\frac{n+p}{N}} \lesssim \sigma/\sqrt{\gamma_+}.$$

Therefore, to ensure that Theorem 14 is applicable, we assume that the noise is small enough, so that $\sigma \lesssim \sqrt{\gamma_+}\|\mathbf{\Theta}_\star\|_F$ (we choose $\mathbf{\Theta}^{(0)} = 0$ and $r = \|\mathbf{\Theta}_\star\|_F$). This completes the proof. ■

A.2 Application to Nonlinear State Equations

Lemma 27 *Let X be a non-negative random variable upper bounded by another random variable Y. Fix an integer $k > 0$. Fix a constant $C > 1 + k\log 3$ and suppose for some $B > 0$ we have that $\mathbb{P}(Y \geq B(1+t)) \leq \exp(-Ct^2)$ for all $t > 0$. Then, the following bound holds,*

$$\mathbb{E}[X^k] \leq (2^k + 2)B^k.$$

Proof Split the real line into regions $\mathcal{R}_i = \{x \mid Bi \leq x \leq B(i+1)\}$. Observe that $\mathbb{P}(Y \in \mathcal{R}_0) + \mathbb{P}(Y \in \mathcal{R}_1) \leq 1$ and $\mathbb{P}(Y \in \mathcal{R}_{i+1}) \leq \exp(-Ci^2)$ for $i \geq 1$. Then,

$$\mathbb{E}[Y^k] \leq \sum_{i=0}^{\infty}(B(i+1))^k\mathbb{P}(Y \in \mathcal{R}_i),$$
$$\leq (2^k+1)B^k + \sum_{i=1}^{\infty}(i+2)^kB^k\exp(-Ci^2).$$

Next, we pick $C > 0$ sufficiently large to satisfy $\exp(-Ci^2)(i+2)^k \leq \exp(-i^2) \leq \exp(-i)$. This can be guaranteed by picking C to satisfy, for all i

$$\exp((C-1)i^2) \geq (i+2)^k \iff (C-1)i^2 \geq k\log(i+2),$$
$$\iff C \geq 1 + \sup_{i\geq 1}\frac{k\log(i+2)}{i^2},$$
$$\iff C \geq 1 + k\log 3.$$

Following this, we obtain $\sum_{i=1}^{\infty}(i+2)^k B^k \exp(-Ci^2) \le B^k$. Thus, we find $\mathbb{E}[Y^k] \le (2^k+2)B^k$. ■

A.2.1 Verification of Assumption 2

Lemma 28 (Bounded states) *Suppose, the nonlinear system (6.2) is (C_ρ, ρ)-stable and $\phi(0) = 0$. Suppose, $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$, $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ and let $\beta_+ := C_\rho(1+\sigma)/(1-\rho)$. Then, starting from $\boldsymbol{h}_0 = 0$, for all $0 \le t \le T$, we have:*

(a) $\mathbb{P}(\|\boldsymbol{h}_t\|_{\ell_2} \le c\beta_+\sqrt{n}) \ge 1 - 4T\exp(-100n)$.

(b) $\mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^2] \le \beta_+^2 n$.

(c) $\mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^3] \le C\beta_+^3(\log(2T)n)^{3/2}$.

Proof **(a)** Given $\|\boldsymbol{z}_t\|_{\psi_2} \le 1$ and $\|\boldsymbol{w}_t\|_{\psi_2} \le \sigma$, we use Lemma 21 to obtain $\mathbb{P}(\|\boldsymbol{z}_t\|_{\ell_2} \lesssim \sqrt{n}) \ge 1 - 2T\exp(-100n)$ and $\mathbb{P}(\|\boldsymbol{w}_t\|_{\ell_2} \lesssim \sigma\sqrt{n}) \ge 1 - 2T\exp(-100n)$ for all $0 \le t \le T-1$. Using these results along-with (C_ρ, ρ)-stability in Lemma 6, we get the desired bound on the Euclidean norm of the state vector $\boldsymbol{h}_t$.

(b) Recall that $\boldsymbol{h}_0 = 0$. We claim that $\mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^2] \le \beta_+^2 n(1-\rho^t)^2$, where $\beta_+ := C_\rho(1+\sigma)/(1-\rho)$. Note that, using standard results on the distribution of squared Euclidean norm of a Gaussian vector, we have $\mathbb{E}[\|\boldsymbol{z}_t\|_{\ell_2}^2] = n$ and $\mathbb{E}[\|\boldsymbol{w}_t\|_{\ell_2}^2] = \sigma^2 n$, which implies $\mathbb{E}[\|\boldsymbol{z}_t\|_{\ell_2}] \le \sqrt{n}$ and $\mathbb{E}[\|\boldsymbol{w}_t\|_{\ell_2}] \le \sigma\sqrt{n}$. Using this results, we show that $\boldsymbol{h}_1$ satisfies the following bound and obeys the induction

$$\mathbb{E}[\|\boldsymbol{h}_1\|_{\ell_2}^2] = \mathbb{E}[\|\phi(0) + \boldsymbol{z}_t + \boldsymbol{w}_t\|_{\ell_2}^2] \le (1+\sigma^2)n \le C_\rho^2(1+\sigma)^2 n = \beta_+^2 n(1-\rho^1)^2.$$

This implies $\mathbb{E}[\|\boldsymbol{h}_1\|_{\ell_2}] \le \beta_+\sqrt{n}(1-\rho^1)$ as well. Suppose the bound holds until $t-1$, that is, $\mathbb{E}[\|\boldsymbol{h}_{t-1}\|_{\ell_2}^2] \le \beta_+^2 n(1-\rho^{t-1})^2$ (which also means $\mathbb{E}[\|\boldsymbol{h}_{t-1}\|_{\ell_2}] \le \beta_+\sqrt{n}(1-\rho^{t-1})$). We now apply the induction as follows: First observe that $\mathbb{E}[\|\boldsymbol{h}_{t,L}\|_{\ell_2}]$ obeys the same upper bound as $\mathbb{E}[\|\boldsymbol{h}_L\|_{\ell_2}]$ by construction. To proceed, recalling (4.2), we get the following by induction

$$\begin{aligned}
&\|\boldsymbol{h}_t - \boldsymbol{h}_{t,t-1}\|_{\ell_2} \le C_\rho\rho^{t-1}\|\boldsymbol{h}_1\|_{\ell_2} \\
\implies &\|\boldsymbol{h}_t\|_{\ell_2} \le C_\rho\rho^{t-1}\|\boldsymbol{h}_1\|_{\ell_2} + \|\boldsymbol{h}_{t,t-1}\|_{\ell_2}, \\
\implies &\|\boldsymbol{h}_t\|_{\ell_2}^2 \le (C_\rho\rho^{t-1}\|\boldsymbol{h}_1\|_{\ell_2} + \|\boldsymbol{h}_{t,t-1}\|_{\ell_2})^2, \\
\implies &\mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^2] \le C_\rho^2\rho^{2(t-1)}\,\mathbb{E}[\|\boldsymbol{h}_1\|_{\ell_2}^2] + \mathbb{E}[\|\boldsymbol{h}_{t-1}\|_{\ell_2}^2] + 2C_\rho\rho^{t-1}\,\mathbb{E}[\|\boldsymbol{h}_1\|_{\ell_2}]\,\mathbb{E}[\|\boldsymbol{h}_{t-1}\|_{\ell_2}], \\
&\qquad \overset{(a)}{\le} C_\rho^2\rho^{2(t-1)}(1+\sigma)^2 n + \beta_+^2 n(1-\rho^{t-1})^2 + 2nC_\rho\rho^{t-1}(1+\sigma)\beta_+(1-\rho^{t-1}), \\
&\qquad \overset{(b)}{\le} \beta_+^2 n(\rho^{2(t-1)}(1-\rho^1)^2 + (1-\rho^{t-1})^2 + 2\rho^{t-1}(1-\rho^{t-1})(1-\rho^1)), \\
&\qquad = \beta_+^2 n[\rho^{2t-2}(1+\rho^2-2\rho) + 1 + \rho^{2t-2} - 2\rho^{t-1} + (2\rho^{t-1} - 2\rho^{2t-2})(1-\rho)], \\
&\qquad = \beta_+^2 n(1+\rho^{2t} - 2\rho^t), \\
&\qquad = \beta_+^2 n(1-\rho^t)^2, \qquad \text{(A.23)}
\end{aligned}$$

where we get (a) from the induction hypothesis and (b) from the bound on $\boldsymbol{h}_1$. This bound also implies $\mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^2] \le \beta_+^2 n$ and completes the proof.

(c) Recall that, we have $\|\boldsymbol{z}_t\|_{\psi_2} \le 1$, $\|\boldsymbol{w}_t\|_{\psi_2} \le \sigma$, $\mathbb{E}[\|\boldsymbol{z}_t\|_{\ell_2}] \le \sqrt{n}$ and $\mathbb{E}[\|\boldsymbol{w}_t\|_{\ell_2}] \le \sigma\sqrt{n}$. Combining these bounds with standard concentration inequalities of a Guassian random vector, we have

$$\mathbb{P}(\|\boldsymbol{z}_t\|_{\ell_2} \ge \mathbb{E}[\|\boldsymbol{z}_t\|_{\ell_2}] + t) \le \exp(-t^2/2) \quad \text{and} \quad \mathbb{P}(\|\boldsymbol{w}_t\|_{\ell_2} \ge \mathbb{E}[\|\boldsymbol{w}_t\|_{\ell_2}] + t) \le \exp(-t^2/(2\sigma^2)),$$

$$\implies \mathbb{P}(\|\boldsymbol{z}_t\|_{\ell_2} \ge \sqrt{2cn}(1+t)) \le \exp(-cnt^2), \tag{A.24}$$

$$\text{and} \quad \mathbb{P}(\|\boldsymbol{w}_t\|_{\ell_2} \ge \sigma\sqrt{2cn}(1+t)) \le \exp(-cnt^2). \tag{A.25}$$

To proceed, let $X = \|\boldsymbol{h}_t\|_{\ell_2}$ and $Y = \sum_{\tau=0}^{t-1} C_\rho \rho^\tau (\|\boldsymbol{z}_t\|_{\ell_2} + \|\boldsymbol{w}_t\|_{\ell_2})$ and note that $X \le Y$. Now, using (A.24), (A.25) and union bounding over all $0 \le t \le T-1$, we get the following high probability upper bound on Y, that is,

$$\begin{aligned} & \mathbb{P}(Y \ge \sum_{\tau=0}^{t-1} C_\rho \rho^\tau \sqrt{2cn}(1+\sigma)(1+t)) \le 2T\exp(-cnt^2), \\ \implies \quad & \mathbb{P}(Y \ge C_\rho \sqrt{10n\log(2T)}(1+t)(1+\sigma)/(1-\rho)) \le \exp(-5nt^2), \end{aligned}$$

where we choose $c = 5\log(2T)$ to get the final concentration bound of Y. Finally using this bound in Lemma 27, we get

$$\mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^3] \le 32\beta_+^3(\log(2T)n)^{3/2}, \tag{A.26}$$

where $\beta_+ = C_\rho(1+\sigma)/(1-\rho)$, as defined earlier. This completes the proof. ■

A.2.2 Verification of Assumption 3

Theorem 29 *Suppose the nonlinear system (6.2) satisfies (C_ρ, ρ)-stability. Suppose $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$ and $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$. Let β_+ be as in Lemma 28. Then, the matrix $\mathbb{E}[\boldsymbol{h}_t\boldsymbol{h}_t^\top]$ satisfies*

$$(1+\sigma^2)\boldsymbol{I}_n \le \mathbb{E}[\boldsymbol{h}_t\boldsymbol{h}_t^\top] \le \beta_+^2 n\boldsymbol{I}_n. \tag{A.27}$$

Proof We first upper bound the matrix $\mathbb{E}[\boldsymbol{h}_t\boldsymbol{h}_t^\top]$ by bounding its largest singular value as follows,

$$\mathbb{E}[\boldsymbol{h}_t\boldsymbol{h}_t^\top] \le \mathbb{E}[\|\boldsymbol{h}_t\boldsymbol{h}_t^\top\|]\boldsymbol{I}_n \le \mathbb{E}[\|\boldsymbol{h}_t\|_{\ell_2}^2]\boldsymbol{I}_n \le \beta_+^2 n\boldsymbol{I}_n, \tag{A.28}$$

where we get the last inequality by applying Lemma 28. To get a lower bound, note that $\boldsymbol{\Sigma}[\boldsymbol{h}_t] = \mathbb{E}[\boldsymbol{h}_t\boldsymbol{h}_t^\top] - \mathbb{E}[\boldsymbol{h}_t]\mathbb{E}[\boldsymbol{h}_t]^\top$. Since, all of these matrices are positive semi-definite, we get the following lower bound,

$$\mathbb{E}[\boldsymbol{h}_t\boldsymbol{h}_t^\top] \ge \boldsymbol{\Sigma}[\boldsymbol{h}_t] = \boldsymbol{\Sigma}[\phi(\boldsymbol{\Theta}_\star\boldsymbol{h}_{t-1}) + \boldsymbol{z}_t + \boldsymbol{w}_t] \ge \boldsymbol{\Sigma}[\boldsymbol{z}_t + \boldsymbol{w}_t] = (1+\sigma^2)\boldsymbol{I}_n. \tag{A.29}$$

Combining the two bounds gives us the statement of the lemma. This completes the proof. ■

To verify Assumption 3 for the nonlinear system (6.2), denoting the k_{th} row of $\boldsymbol{\Theta}$ by $\boldsymbol{\theta}_k^\top$, the auxiliary loss for the nonlinear system (6.2) is given by,

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) \quad \text{where} \quad \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) := \frac{1}{2}\mathbb{E}[(\boldsymbol{h}_L[k] - \phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) - \boldsymbol{z}_{L-1}[k])^2]. \tag{A.30}$$

Using the derived bounds on the matrix $\mathbb{E}[\boldsymbol{h}_t \boldsymbol{h}_t^\top]$, it is straightforward to show that the auxiliary loss satisfies the following one-point convexity and smoothness conditions.

Lemma 30 (One-point convexity & smoothness) *Consider the setup of Theorem 29 and the auxiliary loss given by (A.30). Suppose, ϕ is γ-increasing (i.e. $\phi'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$) and 1-Lipschitz. Let β_+ be as in Lemma 28. Then, for all $1 \leq k \leq n$, the gradients $\nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$ satisfy,*

$$\begin{aligned} \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) \right\rangle &\geq \gamma^2(1+\sigma^2)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2, \\ \|\nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &\leq \beta_+^2 n \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}. \end{aligned}$$

Proof Given two distinct scalars a, b we define $\phi'(a,b) := \frac{\phi(a)-\phi(b)}{a-b}$. Observe that $0 < \gamma \leq \phi'(a,b) \leq 1$ because of the assumption that ϕ is 1-Lipschitz and γ-increasing. Now, recalling the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ from (A.30), we have

$$\begin{aligned} \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) &= \mathbb{E}[(\phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) - \phi(\boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1}) - \boldsymbol{w}_{L-1}[k])\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}], \\ &= \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1} - \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}] \\ &\quad - \mathbb{E}[\boldsymbol{w}_{L-1}[k]\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}], \\ &= \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)], \end{aligned} \tag{A.31}$$

where $\mathbb{E}[\boldsymbol{w}_{L-1}[k]\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}] = 0$ because $\boldsymbol{h}_{L-1}$ and $\boldsymbol{w}_{L-1}$ are independent and we have $\mathbb{E}[\boldsymbol{w}_{L-1}] = 0$. Next, using γ-increasing property of ϕ, we get the following one-point convexity bound,

$$\begin{aligned} \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) \right\rangle &= \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)] \right\rangle, \\ &\geq \gamma^2 \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top](\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) \right\rangle, \\ &\geq \gamma^2(1+\sigma^2)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2. \end{aligned} \tag{A.32}$$

Similarly, using 1-Lipschitzness of ϕ, we get the following smoothness bound,

$$\begin{aligned} \|\nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &= \|\mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)]\|_{\ell_2}, \\ &\leq \mathbb{E}[\|\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top\|]\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}, \\ &\leq \mathbb{E}[\|\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top\|]\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}. \\ &\leq \beta_+^2 n \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}, \end{aligned} \tag{A.33}$$

where β_+ is as defined in Lemma 28. This completes the proof. ■

A.2.3 Verification of Assumption 4

Let $\mathcal{S} = (\boldsymbol{h}_L^{(i)}, \boldsymbol{h}_{L-1}^{(i)}, \boldsymbol{z}_{L-1}^{(i)})_{i=1}^N$ be N i.i.d. copies of $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$ generated from N i.i.d. trajectories of the system (6.2). Then, the finite sample approximation of the auxiliary loss $\mathcal{L}_{\mathcal{D}}$ is given by,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) \;\text{ where }\; \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) := \frac{1}{2N}\sum_{i=1}^{N}(\boldsymbol{h}_L^{(i)}[k] - \phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)}) - \boldsymbol{z}_{L-1}^{(i)}[k])^2. \quad \text{(A.34)}$$

The following lemma states that both $\nabla\mathcal{L}_{k,\mathcal{D}}$ and $\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}$ are Lipschitz with high probability.

Lemma 31 (Lipschitz gradient) *Consider the same setup of Theorem 29. Consider the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ and its finite sample approximation $\hat{\mathcal{L}}_{k,\mathcal{S}}$ from (A.30) and (A.34) respectively. Suppose, ϕ has bounded first and second derivatives, that is, $|\phi'|, |\phi''| \le 1$. Let β_+ be as in Lemma 28. Then, with probability at least $1 - 4T\exp(-100n)$, for all pairs $\boldsymbol{\Theta}, \boldsymbol{\Theta}' \in \mathcal{B}^{n\times n}(\boldsymbol{\Theta}_\star, r)$ and for $1 \le k \le n$, we have*

$$\begin{aligned}\max(\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2}, \|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2}) \\ \lesssim ((1+\sigma)\beta_+^2 n + r\beta_+^3 n^{3/2}\log^{3/2}(2T))\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}.\end{aligned}$$

Proof To begin recall that, $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) = \mathbb{E}[(\phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1}))\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}]$. To bound the Lipschitz constant of the gradient $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$, we will upper bound the spectral norm of the Hessian as follows,

$$\begin{aligned}\|\nabla^2\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\| &= \|\,\mathbb{E}[(\phi(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1}))\phi''(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top] \\ &\quad + \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top]\|, \\ &\le \mathbb{E}[\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1})(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1})\phi''(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top\|] \\ &\quad + \mathbb{E}[\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top\|], \\ &\le \mathbb{E}[\|(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1})\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top\|] + \mathbb{E}[\|\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top\|], \\ &\le \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}\,\mathbb{E}[\|\boldsymbol{h}_{L-1}\|_{\ell_2}^3] + \mathbb{E}[\|\boldsymbol{h}_{L-1}\|_{\ell_2}^2], \\ &\lesssim \beta_+^3(\log(2T)n)^{3/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \beta_+^2 n, \end{aligned} \quad \text{(A.35)}$$

where we get the last inequality by applying Lemma 28. Similarly, to bound the Lipschitz constant of the empirical gradient

$$\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) = 1/N\sum_{i=1}^{N}(\phi(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1}^{(i)}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1}^{(i)}) - \boldsymbol{w}_{L-1}^{(i)}[k])\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1}^{(i)})\boldsymbol{h}_{L-1}^{(i)},$$

we bound the spectral norm of the Hessian of the empirical loss $\hat{\mathcal{L}}_{k,\mathcal{S}}$ as follows,

$$\begin{aligned}
\|\nabla^2\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k)\| \leq & \frac{1}{N}\sum_{i=1}^{N}\|(\phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1}^{(i)}) - \boldsymbol{w}_{L-1}^{(i)}[k])\phi''(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)})\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top\| \\
& + \frac{1}{N}\sum_{i=1}^{N}\|\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)})\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)})\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top\|, \\
\overset{(a)}{\leq} & \frac{1}{N}\sum_{i=1}^{N}[\|(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{h}_{L-1}^{(i)})\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top\| + (1+|\boldsymbol{w}_{L-1}^{(i)}[k]|)\|\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top\|], \\
\leq & \frac{1}{N}\sum_{i=1}^{N}[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}\|\boldsymbol{h}_{L-1}^{(i)}\|_{\ell_2}^3 + (1+|\boldsymbol{w}_{L-1}^{(i)}[k]|)\|\boldsymbol{h}_{L-1}^{(i)}\|_{\ell_2}^2], \\
\lesssim & \beta_+^3 n^{3/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + (1+\sigma)\beta_+^2 n, \qquad \text{(A.36)}
\end{aligned}$$

with probability at least $1 - 4T\exp(-100n)$, where we get (a) by using a similar argument as we used in the case of auxiliary loss while the last inequality comes from Lemma 28. Combining the two bounds, gives us the statement of the lemma. This completes the proof. ■

A.2.4 Verification of Assumption 5

Given a single sample $(\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})$ from the trajectory of the nonlinear system (6.2), the single sample loss is given by,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\Theta}, (\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1})) &= \sum_{k=1}^{n}\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k])), \\
\text{where} \quad \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k])) &:= \frac{1}{2}(\boldsymbol{h}_L[k] - \phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) - \boldsymbol{z}_{L-1}[k])^2. \qquad \text{(A.37)}
\end{aligned}$$

Before stating a lemma on bounding the subexponential norm of the gradient of the single sample loss (A.37), we will state an intermediate lemma to prove the Lipschitzness of the state vector.

Lemma 32 (Lipschitzness of the state vector) *Suppose the nonlinear system (6.2) is (C_ρ, ρ)-stable, $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$ and $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$. Let $\boldsymbol{v}_t := [\boldsymbol{z}_t^\top \ 1/\sigma\boldsymbol{w}_t^\top]^\top$ and $\boldsymbol{h}_0 = 0$. Fixing all $\{\boldsymbol{v}_i\}_{i\neq\tau}$ (i.e., all except $\boldsymbol{v}_\tau$), $\boldsymbol{h}_{t+1}$ is $C_\rho\rho^{t-\tau}(1+\sigma^2)^{1/2}$ Lipschitz function of $\boldsymbol{v}_\tau$ for $0 \leq \tau \leq t$.*

Proof To begin, observe that $\boldsymbol{h}_{t+1}$ is deterministic function of the sequence $\{\boldsymbol{v}_\tau\}_{\tau=0}^t$. Fixing all $\{\boldsymbol{v}_i\}_{i\neq\tau}$, we denote $\boldsymbol{h}_{t+1}$ as a function of $\boldsymbol{v}_\tau$ by $\boldsymbol{h}_{t+1}(\boldsymbol{v}_\tau)$. Given a pair of vectors $(\boldsymbol{v}_\tau, \hat{\boldsymbol{v}}_\tau)$, using (C_ρ, ρ)-stability of the nonlinear system (6.2), for any $t \geq \tau$, we have

$$\begin{aligned}
\|\boldsymbol{h}_{t+1}(\boldsymbol{v}_\tau) - \boldsymbol{h}_{t+1}(\hat{\boldsymbol{v}}_\tau)\|_{\ell_2} &\leq C_\rho\rho^{t-\tau}\|\boldsymbol{h}_{\tau+1}(\boldsymbol{v}_\tau) - \boldsymbol{h}_{\tau+1}(\hat{\boldsymbol{v}}_\tau)\|_{\ell_2}, \\
&\leq C_\rho\rho^{t-\tau}\|\phi(\boldsymbol{\Theta}_\star\boldsymbol{h}_\tau) + \boldsymbol{z}_\tau + \boldsymbol{w}_\tau - \phi(\boldsymbol{\Theta}_\star\boldsymbol{h}_\tau) - \hat{\boldsymbol{z}}_\tau - \hat{\boldsymbol{w}}_\tau\|_{\ell_2}, \\
&\leq C_\rho\rho^{t-\tau}(\|\boldsymbol{z}_\tau - \hat{\boldsymbol{z}}_\tau\|_{\ell_2} + \sigma\|1/\sigma\boldsymbol{w}_\tau - 1/\sigma\hat{\boldsymbol{w}}_\tau\|_{\ell_2}), \\
&\overset{(a)}{\leq} C_\rho\rho^{t-\tau}(1+\sigma^2)^{1/2}(\|\boldsymbol{z}_\tau - \hat{\boldsymbol{z}}_\tau\|_{\ell_2}^2 + 1/\sigma^2\|\boldsymbol{w}_\tau - \hat{\boldsymbol{w}}_\tau\|_{\ell_2}^2)^{1/2}, \\
&\leq C_\rho\rho^{t-\tau}(1+\sigma^2)^{1/2}\|\boldsymbol{v}_\tau - \hat{\boldsymbol{v}}_\tau\|_{\ell_2}, \qquad \text{(A.38)}
\end{aligned}$$

where we get (a) by using Cauchy-Schwarz inequality. This implies $\boldsymbol{h}_{t+1}$ is $C_\rho \rho^{t-\tau}(1+\sigma^2)^{1/2}$ Lipschitz function of $\boldsymbol{v}_\tau$ for $0 \le \tau \le t$ and completes the proof. ■

We are now ready to state a lemma to bound the subexponential norm of the gradient of the single sample loss (A.37).

Lemma 33 (Subexponential gradient) *Consider the same setup of Lemma 32. Let* $\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k]))$ *be as in (A.37) and* $\beta_+ := C_\rho(1+\sigma)/(1-\rho)$. *Suppose* $|\phi'(x)| \le 1$ *for all* $x \in \mathbb{R}$. *Then, at any point* $\boldsymbol{\Theta}$, *for all* $1 \le k \le n$, *we have*

$$\|\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k])) - \mathbb{E}[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k]))]\|_{\psi_1} \lesssim \beta_+^2 \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma\beta_+.$$

Proof We first bound the subgaussian norm of the state vector $\boldsymbol{h}_t$ following Oymak (2019) as follows: Setting $\boldsymbol{v}_t = [\boldsymbol{z}_t^\top \ 1/\sigma \boldsymbol{w}_t^\top]^\top$, define the vectors $\boldsymbol{q}_t := [\boldsymbol{v}_0^\top \cdots \boldsymbol{v}_{t-1}^\top]^\top \in \mathbb{R}^{2nt}$ and $\hat{\boldsymbol{q}}_t := [\hat{\boldsymbol{v}}_0^\top \cdots \hat{\boldsymbol{v}}_{t-1}^\top]^\top \in \mathbb{R}^{2nt}$. Observe that $\boldsymbol{h}_t$ is a deterministic function of $\boldsymbol{q}_t$, that is, $\boldsymbol{h}_t = f(\boldsymbol{q}_t)$ for some function f. To bound the Lipschitz constant of f, for all (deterministic) vector pairs $\boldsymbol{q}_t$ and $\hat{\boldsymbol{q}}_t$, we find the scalar L_f satisfying

$$\|f(\boldsymbol{q}_t) - f(\hat{\boldsymbol{q}}_t)\|_{\ell_2} \le L_f \|\boldsymbol{q}_t - \hat{\boldsymbol{q}}_t\|_{\ell_2}. \tag{A.39}$$

For this purpose, we define the vectors $\{\boldsymbol{b}_i\}_{i=0}^t$ as follows: $\boldsymbol{b}_i = [\hat{\boldsymbol{v}}_0^\top \cdots \hat{\boldsymbol{v}}_{i-1}^\top \ \boldsymbol{v}_i^\top \cdots \boldsymbol{v}_{t-1}]^\top$. Observing that $\boldsymbol{b}_0 = \boldsymbol{q}_t$ and $\boldsymbol{b}_t = \hat{\boldsymbol{q}}_t$, we write the telescopic sum,

$$\|f(\boldsymbol{q}_t) - f(\hat{\boldsymbol{q}}_t)\|_{\ell_2} \le \sum_{i=0}^{t-1} \|f(\boldsymbol{b}_{i+1}) - f(\boldsymbol{b}_i)\|_{\ell_2}. \tag{A.40}$$

Observe that $f(\boldsymbol{b}_{i+1})$ and $f(\boldsymbol{b}_i)$ differs only in $\boldsymbol{v}_i, \hat{\boldsymbol{v}}_i$ terms in the argument. Hence, viewing $\boldsymbol{h}_t$ as a function of $\boldsymbol{w}_i$ and using the result of Lemma 32, we have

$$\begin{aligned}
\|f(\boldsymbol{q}_t) - f(\hat{\boldsymbol{q}}_t)\|_{\ell_2} &\le \sum_{i=0}^{t-1} C_\rho \rho^{t-1-i}(1+\sigma^2)^{1/2} \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i\|_{\ell_2}, \\
&\overset{(a)}{\le} C_\rho (1+\sigma^2)^{1/2} \Big(\sum_{i=0}^{t-1} \rho^{2(t-1-i)} \Big)^{1/2} \underbrace{\Big(\sum_{i=0}^{t-1} \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i\|_{\ell_2}^2 \Big)^{1/2}}_{\|\boldsymbol{q}_t - \hat{\boldsymbol{q}}_t\|_{\ell_2}}, \\
&\overset{(b)}{\le} \frac{C_\rho (1+\sigma^2)^{1/2}}{(1-\rho^2)^{1/2}} \|\boldsymbol{q}_t - \hat{\boldsymbol{q}}_t\|_{\ell_2},
\end{aligned} \tag{A.41}$$

where we get (a) by applying the Cauchy-Schwarz inequality and (b) follows from $\rho < 1$. Setting $\beta_K = C_\rho(1+\sigma^2)^{1/2}/(1-\rho^2)^{1/2}$, we found that $\boldsymbol{h}_t$ is β_K-Lipschitz function of $\boldsymbol{q}_t$. Since $\boldsymbol{v}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_{2n})$, the vector $\boldsymbol{q}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_{2nt})$. Since, $\boldsymbol{h}_t$ is β_K-Lipschitz function of $\boldsymbol{q}_t$, for any fixed unit length vector $\boldsymbol{a}$, $\boldsymbol{a}^\top \boldsymbol{h}_t$ is still β_K-Lipschitz function of $\boldsymbol{q}_t$. This implies $\|\boldsymbol{h}_t - \mathbb{E}[\boldsymbol{h}_t]\|_{\psi_2} \lesssim \beta_K$. Secondly, β_K-Lipschitz function of a Gaussian vector obeys the variance inequality $\mathbf{var}[\boldsymbol{a}^\top \boldsymbol{h}_t] \le \beta_K^2$ (page 49 of Ledoux (2001)), which implies the covariance

bound $\mathbf{\Sigma}[\boldsymbol{h}_t] \preceq \beta_K^2 \boldsymbol{I}_n$. Combining these results with $\|\boldsymbol{w}_t[k]\|_{\psi_2} \leq \sigma$, we get the following subexponential norm bound,

$$\begin{aligned}
\|\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k])) - \mathbb{E}[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{h}_L[k], \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}[k]))]\|_{\psi_1} \\
&\leq \|\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1}) \phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) \boldsymbol{h}_{L-1} \boldsymbol{h}_{L-1}^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) \\
&\quad - \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}, \boldsymbol{\theta}_k^{\star\top} \boldsymbol{h}_{L-1}) \phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) \boldsymbol{h}_{L-1} \boldsymbol{h}_{L-1}^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)]\|_{\psi_1} \\
&\quad + \|\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}) \boldsymbol{w}_{L-1}[k] \boldsymbol{h}_{L-1}\|_{\psi_1}, \\
&\lesssim \beta_K^2 \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma \beta_K, \\
&\lesssim \beta_+^2 \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma \beta_+,
\end{aligned} \tag{A.42}$$

where we get the last two inequalities from the fact that the product of a bounded function (ϕ is 1-Lipschitz because $|\phi'(x)| \leq 1$ for all $x \in \mathbb{R}$) with a subgaussian/subexponential random vector is still a subgaussian/subexponential random vector. This completes the proof. ■

A.2.5 Proof of Corollary 16

Proof We have verified Assumptions 2, 3, 4 and 5 for the nonlinear system 6.2. Hence, we are ready to use Theorem 14 to learn the dynamics $\mathbf{\Theta}_\star$ of the nonlinear system (6.2) . Before that, we find the values of the system related constants to be used in Theorem 14 as follows.

Remark 34 *Consider the same setup of Lemma 32. Let $\beta_+ \geq \beta_K > 0$ be as defined in Lemmas 28 and 33 respectively. Then, with probability at least $1 - 4T\exp(-100n)$, for all $1 \leq t \leq T$, $\mathbf{\Theta} \in \mathcal{B}^{n\times n}(\mathbf{\Theta}_\star, r)$ and $1 \leq k \leq n$, the scalars C_ϕ, D_ϕ take the following values.*

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}_k} \phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t)\|_{\ell_2} = \|\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t) \boldsymbol{h}_t\|_{\ell_2} \leq \|\boldsymbol{h}_t\|_{\ell_2} \lesssim \beta_+ \sqrt{n} = C_\phi, \\
&\|\nabla_{\boldsymbol{h}_t} \nabla_{\boldsymbol{\theta}_k} \phi(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t)\| = \|\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t) \boldsymbol{I}_n + \phi''(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t) \boldsymbol{h}_t \boldsymbol{\theta}_k^\top\| \lesssim 1 + \beta_+ \sqrt{n} \|\boldsymbol{\theta}_k\|_{\ell_2} \lesssim 1 + \|\mathbf{\Theta}_\star\|_F \beta_+ \sqrt{n} = D_\phi
\end{aligned}$$

where without loss of generality we choose $\mathbf{\Theta}^{(0)} = 0$ and $r = \|\mathbf{\Theta}_\star\|_F$. Furthermore, the Lipschitz constant and the gradient noise coefficients take the following values: $L_{\mathcal{D}} = c((1+\sigma)\beta_+^2 n + \|\mathbf{\Theta}_\star\|_F \beta_+^3 n^{3/2} \log^{3/2}(2T))$, $K = c\beta_+^2$ and $\sigma_0 = c\sigma\beta_+$. Lastly, we also have $p_0 = 4T\exp(-100n)$.

Using these values, we get the following sample complexity bound for learning nonlinear system (6.2) via gradient descent,

$$\begin{aligned}
&N \gtrsim \frac{\beta_+^4}{\gamma^4 (1+\sigma^2)^2} \log^2(3((1+\sigma)\beta_+^2 n + \|\mathbf{\Theta}_\star\|_F \beta_+^3 n^{3/2} \log^{3/2}(2T)) N/\beta_+^2 + 3) n, \\
\implies \quad &N \gtrsim \frac{C_\rho^4}{\gamma^4 (1-\rho)^4} \log^2(3(1+\sigma) n + 3\|\mathbf{\Theta}_\star\|_F \beta_+ n^{3/2} \log^{3/2}(2T) N + 3) n,
\end{aligned} \tag{A.43}$$

where $\frac{\beta_+^2}{1+\sigma^2} \leq \frac{C_\rho^2 (1+\sigma)^2/(1-\rho)^2}{(1+\sigma)^2/2} = \frac{2C_\rho^2}{(1-\rho)^2}$ is an upper bound on the condition number of the covariance matrix $\mathbf{\Sigma}[\boldsymbol{h}_t]$. Similarly, the approximate mixing time of the nonlinear system (6.2)

is given by,

$$L \geq 1 + \big\lceil \log(c_0 C_\rho \beta_+ (1 + \|\mathbf{\Theta}_\star\|_F \beta_+ \sqrt{n}) n \sqrt{N/n}) + \log(c/\beta_+ \vee c\sqrt{n}/\beta_+) \big\rceil / \log(\rho^{-1}),$$
$$\Longleftarrow \quad L \geq \Big\lceil 1 + \frac{\log(C C_\rho (1 + \|\mathbf{\Theta}_\star\|_F \beta_+) N n)}{1 - \rho} \Big\rceil, \tag{A.44}$$

where $C > 0$ is a constant. Finally, given the trajectory length $T \gtrsim L(N+1)$, where N and L are as given by (A.43) and (A.44) respectively, starting from $\mathbf{\Theta}^{(0)} = 0$ and using the learning rate $\eta = \frac{\gamma^2(1+\sigma^2)}{16\beta_+^4 n^2} \geq \frac{\gamma^2(1-\rho)^4}{32 C_\rho^4 (1+\sigma)^2 n^2}$, with probability at least $1 - Ln\big(4T + \log(\frac{\|\mathbf{\Theta}_\star\|_F C_\rho (1+\sigma)}{\sigma(1-\rho)})\big)\exp(-100n)$ for all $1 \leq k \leq n$, all gradient descent iterates $\mathbf{\Theta}^{(\tau)}$ on $\hat{\mathcal{L}}$ satisfy

$$\begin{aligned}
\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} &\leq \big(1 - \frac{\gamma^4(1+\sigma^2)^2}{128\beta_+^4 n^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \\
&\quad + \frac{5c}{\gamma^2(1+\sigma^2)} \sigma \beta_+ \log(3(1+\sigma)n + 3\|\mathbf{\Theta}_\star\|_F \beta_+ n^{3/2} \log^{3/2}(2T) N + 3) \sqrt{\frac{n}{N}}. \\
&\leq \big(1 - \frac{\gamma^4(1-\rho)^4}{512 C_\rho^4 n^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \\
&\quad + \frac{10 c C_\rho}{\gamma^2(1-\rho)} \sigma \log(3(1+\sigma)n + 3C_\rho(1+\sigma)\|\mathbf{\Theta}_\star\|_F n^{3/2} \log^{3/2}(2T) N/(1-\rho) + 3) \sqrt{\frac{n}{N}},
\end{aligned}$$

where we get the last inequality by plugging in the value of $\beta_+ = C_\rho \sigma/(1-\rho)$ and using the inequality $(1+\sigma^2) \geq \frac{(1+\sigma)^2}{2}$. We remark that, choosing $N \gtrsim \frac{C_\rho^4}{\gamma^4(1-\rho)^4} \log^2(3(1+\sigma)n + 3C_\rho(1+\sigma)\|\mathbf{\Theta}_\star\|_F n^{3/2} \log^{3/2}(2T) N/(1-\rho) + 3) n$, the residual term in the last inequality can be bounded as,

$$\frac{10 c C_\rho}{\gamma^2(1-\rho)} \log(3(1+\sigma)n + 3C_\rho(1+\sigma)\|\mathbf{\Theta}_\star\|_F n^{3/2} \log^{3/2}(2T) N/(1-\rho) + 3) \sqrt{\frac{n}{N}} \lesssim \sigma.$$

Therefore, to ensure that Theorem 14 is applicable, we assume that $\sigma \lesssim \|\mathbf{\Theta}_\star\|_F$ (where we choose $\mathbf{\Theta}^{(0)} = 0$ and $r = \|\mathbf{\Theta}_\star\|_F$). This completes the proof. ■