UNIVERSITY OF CALIFORNIA
RIVERSIDE

The Sample Complexity of Learning Dynamical Systems

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Yahya Sattar

September 2023

Dissertation Committee:

Professor Samet Oymak, Chairperson
Professor Amit K. Roy-Chowdhury
Professor Fabio Pasqualetti

The Dissertation of Yahya Sattar is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

> "*Blessed is the One in Whose Hands rests all authority. And He is Most Capable of everything. He is the One Who created death and life in order to test which of you is best in deeds. And He is the Almighty, All-Forgiving. He is the One Who created seven heavens, one above the other. You will never see any imperfection in the creation of the Most Compassionate. So look again: do you see any flaws? Then look again and again—your sight will return frustrated and weary.*"
>
> —Al Quran, Surah Al Mulk

I express my deep gratitude to my advisor Prof. Samet Oymak for his support and guidance. Without his continuous assistance and supervision, this thesis would not have been possible. I have learned from him many useful proof techniques which have helped me solve some of the challenging research problems.

I am extremely thankful to my co-advisors Prof. Necmiye Ozay and Prof. Laura Balzano for their support and guidance during the last two years of my PhD. I feel very fortunate to have worked with them. Without their support and valuable feedback, this thesis would not have been possible. I have learned from them many important lessons about mentorship and professionalism. I thank them for being an inspiration to me and for being always available to answer my questions over Slack.

I am very thankful to my committee member Prof. Amit Roy-Chowdhury for believing in me during the hardest time of my PhD. I thank him for introducing me to my advisor Prof. Samet Oymak and his research direction. I thank him for the support and guidance I received from him when he was our Graduate Advisor. I am very thankful to my second committee member Prof. Fabio Pasqualetti for his wonderful course on control systems. This was my first course in Control Theory at UCR and it helped me understand many important concepts related to my research on dynamical systems.

I am extremely thankful to Prof. Zubair Khalid for the two wonderful years I spent working under his supervision at LUMS. Words cannot express the efforts he put in to train me as a researcher. I am very thankful to Prof. Momin Uppal for his support and guidance throughout. I truly admire his teaching methodology and I believe he inspired many students including me to pursue research in information theory. His dedication to teaching and mentoring is unparalleled.

Further, I want to express my deep gratitude to Prof. Waqas Majeed, Prof. Anastasios Mourikis, Prof. Hamed Mohsenian and Prof. Pervez Hoodbhoy for being outstanding instructors. Their lectures used to be captivating and interesting. I feel extremely fortunate to be able to learn from them.

I want to thank my amazing collaborators and lab-mates Zhe Du, Davoud Ataee tarzanagh, Mingchen Li, Xuechen Zhang, Yingcong Li, Xiangyu Chang and Emrullah Ildiz. It was wonderful to share the lab-space with the many amazing friends from Multi-Aspect Data Lab and Database Lab – Uday Singh Saini, Miguel Gutierrez, Rutuja Gurav, William Shiao, Dawon Ahn, Yunshu Wu, Shahrzad Shirazi, Mehnaz Tabassum Mahin and Tomal Majumder.

During my stay at Riverside, I was fortunate to have the company of Shahid Saleem, Awez Shaik, Harish Reddy, Malik Al Masri, Usman Javed, Muhammad Athwan, Aijaz Hashmi, Ameer Ahmed Siddique, Abdus Salam, Sadrul Ula, Sarder Ali Tariq, Akil Sevim, Rakib Hyder, Abed AlRahman Al Makdah, Akber Raza, Wajahat Abbas, Chaudhry Ghulam Sarwar, Umar Farooq, Abu Bakr Siddique, Hasnain Farooq, Syed Adeel Zaffar, Fahad Warraich, Haris Gilani, Sardar Atif and many more.

I am extremely thankful to my parents for their unwavering support and unconditional love. The hardworking nature of my father and the sacrifice of my mother have been a guiding light for me throughout my life. I am thankful to my sister Salma for her love and affection. I thank her for buying me my first laptop when I was studying at LUMS. My love and thanks also go to my brothers Zakariya, Shoaib and Muhammad. I am thankful to Allah for blessing me with such loving brothers.

Last, but certainly not the least, my most affectionate thought goes to my wife Faiza, who has stood by me through all my good and bad times. Words cannot express my appreciation for her love and support. It was her who took care of the entire family when I was busy with never-ending publication deadlines. I cannot thank her enough for the countless sacrifices she made to help me achieve my goals. My most affectionate love goes to my children Faris and Umar – the warmth of my heart. I am extremely thankful to Allah for blessing me beyond my imagination.

Yahya Sattar,
California,
September 2023.

To Amma and Abba for the sacrifices they made for my education.

ABSTRACT OF THE DISSERTATION

The Sample Complexity of Learning Dynamical Systems

by

Yahya Sattar

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, September 2023
Professor Samet Oymak, Chairperson

Machine learning has emerged as a leading force in revolutionizing technology, education, and almost every aspect of our lives. Reinforcement learning is a sub-field of machine learning that deals with the effects of dynamic feedback and systems that interact with the environment. In these settings, classic statistical and algorithmic guarantees often do not hold because of non i.i.d. data, dynamic feedback, and distribution shift.

We develop a framework for single trajectory learning of nonlinear dynamical systems using mixing arguments. Our main result studies the landscape of empirical risk minimization for learning nonlinear dynamical systems from a single trajectory, and provides uniform gradient convergence guarantee, which is combined with novel one-point convexity to facilitate the learning of nonlinear dynamical systems. Our proposed framework allows for non-convex loss landscape and our sample complexity and statistical error rates are optimal in terms of the trajectory length, dimensions of the system and input/noise strength.

Next, we study the problem of learning bilinear dynamical systems from a single trajectory of the system's states and inputs. Our main contribution is the application of martingale small-ball arguments to derive learning guarantees for non-mixing bilinear dynamical systems. We further extend our analysis to time varying dynamical systems by studying the problem of learning non-mixing Markov jump systems. Specifically, we learn the dynamics in each mode and the Markov transition matrix, underlying the evolution of the mode switches, from a single trajectory of the system's states, inputs, and modes. Our sample complexity and statistical error rates are optimal in terms of the trajectory length, the dimensions of the system and the input/noise strength.

Lastly, as a preliminary to the problem of finding the best LTI dynamical system that can minimize least-squares loss given a single trajectory of an unknown dynamical system, we study the simpler problem of finding the best linear model in high dimensions, given a dataset. Specifically, we analyze projected gradient descent algorithm to estimate the population minimizer in the finite sample regime. We show that the nonlinearity of the problem can be treated as uncorrelated noise and establish linear convergence rate and data-dependent estimation error bounds for the projected gradient descent algorithm.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Dynamical systems are fundamental for modeling a wide range of problems appearing in complex physical processes, cyber-physical systems, natural language processing, and machine learning. Classical optimal control literature heavily relies on modeling the underlying system as a linear time-invariant (LTI) dynamical system to synthesize control policies leading to elegant solutions such as PID controller and Kalman filter [1–3]. Kalman filters have been vital in the planning and control of vehicles, particularly aircraft, spacecraft and ships. They have been used in the implementation of navigation systems of spacecrafts [4], and in the trajectory estimation for the famous Apollo program [5]. Reinforcement learning (RL) is a sub-field of machine learning that studies how to use past data to enhance the future manipulation of a dynamical system. Modern RL algorithms have attained super-human level performance in playing Atari games from the pixels [6] to mastering the game of Go [7, 8]. They also find critical applications in many fields including robotics, self-driving cars, finance and smart grids [9, 10].

Further, modern approaches for processing sequential data in natural language processing (NLP), such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers have connections to nonlinear dynamical systems [11,12]. RNNs and LSTMs are integral components in language modeling and NLP applications, and they are inherently dynamical systems because they maintain hidden states. While transformers are not dynamical systems, they are sequence models that infer the auto-regressive nature of NLP tasks [13]. Thus, learning and decision-making with temporally-dependent data represent a common challenge that connect transformers and dynamical systems.

In many of the real world problems, we have to estimate or approximate the underlying dynamical system 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 dynamical system from the time series of its trajectories [14–18]. In this thesis, we aim to learn the dynamical systems which are governed by a general nonlinear state equation,

$$\boldsymbol{x}_{t+1} = \phi(\boldsymbol{x}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \qquad t = 0, \dots T-1, \tag{1.0.1}$$

where $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is the system dynamics, $\boldsymbol{x}_t \in \mathbb{R}^n$ 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 square-loss empirical risk minimization algorithms for learning the system dynamics from a single finite trajectory $(\boldsymbol{x}_t, \boldsymbol{u}_t)_{t=0}^{T-1}$.

1.1 Background

Learning dynamical systems has a long history, with major theoretical results being related to asymptotic properties [19] under strong assumptions on persistence of excitation [20]. More recently, the trend has been to move away from asymptotics [21]. There is a recent surge of interest toward understanding the sample complexity of learning standard linear time-invariant (LTI) dynamical systems from a single finite trajectory under mild assumptions, using statistical tools like martingales [22–25] or mixing-time arguments [26,27]. Specifically, [28] provides precise rates for the non-asymptotic identification of standard LTI dynamical systems using a single trajectory. Single trajectory learning of dynamical systems is challenging because of temporally correlated data [22,23,29]. Moreover, if the dynamical systems are governed by nonlinear, bilinear or time-varying state equations, then deriving non-asymptotic learning guarantees is even more challenging. The reason is that, it is not straightforward to extend the above mentioned statistical tools like martingales or mixing-time arguments to nonlinear dynamical systems.

There has been some recent works on non-asymptotic learning of certain classes of nonlinear dynamical systems. [30] 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. [31,32] study theoretical properties of nonlinear state equations with a goal towards understanding recurrent networks and nonlinear systems. [33] provides theoretical guarantees for the recovery of generalized linear dynamical systems and [34] provides the first offline algorithm that can learn generalized linear models without the mixing assumption. In the non-parametric setting, [29] analyzes the performance of the non-parametric least squares estimator (LSE) and shows that the non-parametric LSE converges to the ground truth regression function at the minimax optimal rate. In a follow up work, [35] provides a fast rate excess risk bound which shows that whenever a trajectory hypercontractivity condition holds, the risk of the LSE on dependent data matches the i.i.d. rate order-wise after a burn-in time.

1.2 Contributions and Thesis Outline

The main contributions of this thesis is to develop a framework for single trajectory learning of dynamical systems beyond standard LTI dynamical systems, such as nonlinear, bilinear and markov jump systems.

Chapter 2: In this chapter, we study the problem of learning nonlinear dynamical systems,

$$\boldsymbol{x}_{t+1} = \phi(\boldsymbol{x}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \qquad t = 0, \dots T-1, \tag{1.2.1}$$

where $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is the system dynamics, $\boldsymbol{x}_t \in \mathbb{R}^n$ 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. We assume the system is driven by inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{x}_t) + \boldsymbol{z}_t$, where $\boldsymbol{\pi}(\cdot)$ is a fixed control policy and $\boldsymbol{z}_t$ is excitation for exploration. With our choice of inputs, the state equation (1.2.1) becomes,

$$\boldsymbol{x}_{t+1} = \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \qquad t = 0, \dots T-1, \tag{1.2.2}$$

where $\tilde{\phi}$ denotes the closed-loop nonlinear system. Towards estimating $\boldsymbol{\theta}_\star$, we formulate an empirical risk minimization (ERM) problem over single finite trajectory as follows,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{\theta}), \qquad \hat{\mathcal{L}}(\boldsymbol{\theta}) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{x}_{t+1} - \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2, \tag{1.2.3}$$

where $L \geq 1$ is the approximate mixing-time of the nonlinear dynamical system. To solve (1.2.3), we investigate the properties of the gradient descent algorithm, given by the following iterate

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_{\tau} - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}_{\tau}), \tag{1.2.4}$$

Our goal in this chapter is to study the statistical and computational efficiency of the gradient descent algorithm (1.2.4) for learning $\boldsymbol{\theta}_\star$ from a single finite trajectory. For this purpose, first, we develop new statistical guarantees for the uniform convergence of the gradients of the empirical loss, which is combined with a novel one-point convexity and smoothness (OPCS) condition to estimate $\boldsymbol{\theta}_\star$ with an error rate of $\mathcal{O}(\sqrt{dL/T})$, where $L \geq 1$ is the mixing-time. While we focus on nonlinear state equations in this chapter, our technical ideas (e.g., combining mixing-time and optimization landscape arguments) have implications for richer class of dynamical systems. Finally, we specialize our main results to two special cases of interest: (a) Standard LTI dynamical systems $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_t$, and (b) Nonlinear state equation $\boldsymbol{x}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{x}_t) + \boldsymbol{u}_t + \boldsymbol{w}_t$. This chapter is based on the following publications:

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

[37] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *The Journal of Machine Learning Research*, 23(1):6248–6296, 2022.

Chapter 3: In this chapter, we study the problem of learning bilinear dynamical systems which are governed by the state equation,

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}_0\boldsymbol{x}_t + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k\boldsymbol{x}_t + \boldsymbol{w}_{t+1}, \qquad t = 0, \dots T-1, \tag{1.2.5}$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$ is the state, $\boldsymbol{u}_t \in \mathbb{R}^p$ is the input, and $\boldsymbol{w}_t \in \mathbb{R}^n$ is the process noise at time t. $\{\boldsymbol{A}_k\}_{k=0}^{p} \in \mathbb{R}^{n\times n}$ are the state matrices which govern the dynamics of the system.

Our goal in this chapter is to analyze the performance of the least-squares estimator for estimating the state matrices $\{\boldsymbol{A}_k\}_{k=0}^{p}$ given a single finite trajectory of (1.2.5). Specifically, we are able to show that the least-squares estimator converges to the true dynamics with an error rate of $\mathcal{O}(\sqrt{n(p+1)/T})$. We obtain this result by extending the martingale small-ball argument to bilinear dynamical systems. This chapter is based on the following publication:

[38] Yahya Sattar, Samet Oymak, and Necmiye Ozay. Finite sample identification of bilinear dynamical systems. *IEEE 61st Conference on Decision and Control (CDC)*, pages 6705–6711. IEEE, 2022.

Chapter 4: In this chapter, we study the problem of learning Markov jump linear dynamical systems (MJS) which are governed by the following state equation,

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{B}_{\omega(t)}\boldsymbol{u}_t + \boldsymbol{w}_t \qquad \omega(t) \sim \text{Markov Chain}(\boldsymbol{T}), \qquad t = 0, \dots T-1, \tag{1.2.6}$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$, $\boldsymbol{u}_t \in \mathbb{R}^p$ and $\boldsymbol{w}_t \in \mathbb{R}^n$ are the state, input, and process noise of the MJS at time t. There are s modes in total, and the dynamics of mode i is given by the state matrix $\boldsymbol{A}_i$ and input matrix $\boldsymbol{B}_i$. The active mode at time t is indexed by $\omega(t) \in [s]$.

In this chapter, we provide the first comprehensive non-asymptotic learning guarantees for MJS given a single finite trajectory of (1.2.6). We provide an algorithm (2) to estimate the MJS dynamics with an error rate of $\mathcal{O}(\sqrt{(n+p)/T})$. We obtain this result by extending the martingale small-ball argument to MJS. This chapter is based on the following publication:

[39] Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Identification and adaptive control of markov jump systems: Sample complexity and regret bounds. *IEEE Transactions on Automatic Control (TAC)*, under submission, 2023.

Chapter 5: System dynamics often exhibit sparse structure which can improve sample complexity. However we lack theory to estimate such sparse structures through nonlinearities. Towards this goal, this chapter studies the simpler problem of finding the best linear model that can minimize least-squares loss given the data $(\boldsymbol{x}_i, y_i)_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$. Specifically, we are interested in the high-dimensional regime where we have fewer samples than the parameter dimension, i.e., we assume $n \ll p$. In this case, the problem is ill-posed; however, if the population minimizer $\boldsymbol{\theta}_\star := \mathbb{E}[y\boldsymbol{x}]$ lies on a low-dimensional manifold, we can take advantage of this information to solve a constrained empirical risk minimization problem. Let $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}$ be the regularization function, that promotes the desired structure, such as sparsity or low-rank. Then, we are interested to solve the following constrained empirical risk minimization (ERM),

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{\theta}) \le R. \tag{1.2.7}$$

where $\boldsymbol{y} = [y_1 \ \dots \ y_n]^\top \in \mathbb{R}^n$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$ are the output labels and data matrix respectively. To solve (1.2.7), we investigate the properties of the projected gradient descent algorithm, given by the following iterate

$$\boldsymbol{\theta}_{\tau+1} = \mathcal{P}_{\mathcal{K}}(\boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_\tau, \mu_\tau)), \tag{1.2.8}$$

where $\mathcal{P}_{\mathcal{K}}$ projects onto the constraint set $\mathcal{K} := \{\boldsymbol{\theta} \in \mathbb{R}^p \mid \mathcal{R}(\boldsymbol{\theta}) \leq R\}$ and η is the step size.

The main contribution of this chapter is to study the statistical and computational efficiency of the projected gradient descent algorithm (1.2.8) for finding the best linear model $\boldsymbol{\theta}_\star$ given the data $(\boldsymbol{x}_i, y_i)_{i=1}^n$. Our results in this chapter hold for both sub-gaussian and sub-exponential data $(\boldsymbol{x}_i, y_i)_{i=1}^n$. This chapter is based on the following publication:

[40] Yahya Sattar and Samet Oymak. Quickly finding the best linear model in high dimensions via projected gradient descent. *IEEE Transactions on Signal Processing*, 68:818–829, 2020.

Chapter 6: In this chapter, we present numerical experiments to verify our theoretical findings from the previous chapters.

Chapter 7: In this chapter, we provide concluding remarks for the thesis and also discuss possible future directions.

Lastly, the supplementary proofs for the results in Chapters 2 and 5 are presented in **Appendices A, B and C**, respectively.

1.3 Work Outside the Scope of this Thesis

The following articles, while relevant to the theme of this thesis and completed during our doctoral studies, are not part of this thesis:

[41] Zhe Du, Yahya Sattar, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Data-driven control of markov jump systems: Sample complexity and regret bounds. *American Control Conference (ACC)*, pages 4901–4908. IEEE, 2022.

[42] Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Samet Oymak, Laura Balzano, and Necmiye Ozay. Certainty equivalent quadratic control for markov jump systems. *American Control Conference (ACC)*, pages 2871–2878. IEEE, 2022.

The following article, while completed during our doctoral studies, is unrelated to the theme of this thesis:

[43] Mingchen Li, Yahya Sattar, Christos Thrampoulidis, and Samet Oymak. Exploring weight importance and hessian bias in model pruning. *arXiv preprint arXiv:2006.10903*, 2020.

1.4 Notations

We use boldface uppercase (lowercase) letters to denote matrices (vectors). For a vector $\boldsymbol{v}$, we denote its Euclidean norm by $\|\boldsymbol{v}\|_{\ell_2}$, its ℓ_1 norm by $\|\boldsymbol{v}\|_{\ell_1}$, and its ℓ_∞ norm by $\|\boldsymbol{v}\|_{\ell_\infty}$, respectively. For a matrix $\boldsymbol{M}$, $\rho(\boldsymbol{M}), \|\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_F$ denote the spectral radius, spectral norm and Frobenius norm respectively. $\mathbf{vec}(\boldsymbol{M}) \in \mathbb{R}^{mn}$ denotes the vectorization of a matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, and $\mathbf{mtx}(\cdot)$ denotes its inverse, that is, $\mathbf{mtx}\big(\mathbf{vec}(\boldsymbol{M})\big) = \boldsymbol{M}$.

The Kronecker product of two matrices $\boldsymbol{M}$ and $\boldsymbol{N}$ is denoted as $\boldsymbol{M} \otimes \boldsymbol{N}$. We denote by $\boldsymbol{V}_{1:s}$ a set of s matrices $\{\boldsymbol{V}_i\}_{i=1}^s$ of same dimensions. We define $[s] := \{1, 2, \ldots, s\}$ and $\|\boldsymbol{V}_{1:s}\| := \max_{i\in[s]} \|\boldsymbol{V}_i\|$. The i-th row or column of a matrix $\boldsymbol{M}$ is denoted by $[\boldsymbol{M}]_{i,:}$ or $[\boldsymbol{M}]_{:,i}$ respectively. We use $\mathbf{1}$ to denote a vector of all ones.

$c, c_0, c_1, \ldots, C, C_0$ 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 $\mathbb{R}^d$. For ease of notation, $\mathcal{B}^d := \mathcal{B}^d(0, 1)$ denote the unit ball 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. Finally, orders of magnitude notation $\tilde{\mathcal{O}}(\cdot)$ hides $\log(1/\delta)$ or $\log^2(1/\delta)$ terms.

Given a set S, let $\text{cl}(S)$ and $\text{clconv}(S)$ be the minimal closed set and minimal closed-convex set containing S respectively. Let $\text{rad}(S)$ denote the set radius $\sup_{\boldsymbol{v}\in S} \|\boldsymbol{v}\|_{\ell_2}$. For closed sets, let $\mathcal{P}_S(\cdot)$ be the projection operator defined as $\mathcal{P}_S(\boldsymbol{a}) = \arg\min_{\boldsymbol{v}\in S} \|\boldsymbol{a} - \boldsymbol{v}\|_{\ell_2}$.

Chapter 2

Nonlinear System Identification

2.1 Introduction

Dynamical systems are fundamental for modeling a wide range of problems appearing in complex physical processes, cyber-physical systems and machine learning. Classical optimal control literature heavily relies on modeling the underlying system as a linear time-invariant (LTI) dynamical system to synthesize control policies leading to elegant solutions such as PID controller and Kalman filter [1–3]. 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 [44–48]. 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 [14–18].

We aim to learn the dynamics of nonlinear systems which are governed by following state equation,

$$\boldsymbol{x}_{t+1} = \phi(\boldsymbol{x}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \tag{2.1.1}$$

where $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is the system dynamics, $\boldsymbol{x}_t \in \mathbb{R}^n$ 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.

2.1.1 Relation to Prior 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 [21]. Recently [22–25, 49–61] explore linear system identification in great depth. [11] provides preliminary guarantees for recurrent networks (RNN) and [12] shows the role of stability in RNNs. There is also a substantial amount of work on model-free approaches [62–66] which avoid learning the dynamics and find the optimal control input by directly optimizing over policy space. In a different line of work, [67] 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, [68, 69] developed learning methods that exploit other control-theoretic priors. Nonetheless, none of these works characterize the sample complexity of the problem.

More recently, [30] 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. [70] 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. [71] 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, [72] proposes an online model learning predictive control framework to control unknown nonlinear dynamical systems, [73] 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 [34] provides the first offline algorithm that can learn generalized linear models without the mixing assumption.

Closer to our work, [31,32] 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

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

loss. The concurrent work of [33] provides related results for the recovery of generalized linear dynamical systems ($\boldsymbol{x}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{x}_t) + \boldsymbol{w}_t$) using complementary techniques. [33] uses martingale arguments and analyze GLMtron algorithm of [74], while we use mixing time arguments and analyze gradient descent.

Perhaps the most established technique in the statistics literature for dealing with non-independent, time-series data is the use of mixing-time arguments [75]. In the machine learning literature, mixing-time arguments have been used to develop generalization bounds [27, 76–78] 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 [79, 80]. 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.

2.1.2 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 2.1). 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 (2.1.1) via gradient descent. Unlike the related results on nonlinear systems by [31,32], 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 [79,80], our bounds properly capture the noise dependence and allow for learning the ground-truth dynamics with high accuracy and small sample complexity (see Section 2.3.1 for further discussion).

- **Applications:** We specialize our results by establishing theoretical guarantees for learning linear ($\boldsymbol{x}_{t+1} = \boldsymbol{A}_\star\boldsymbol{x}_t + \boldsymbol{B}_\star\boldsymbol{u}_t + \boldsymbol{w}_t$) as well as nonlinear ($\boldsymbol{x}_{t+1} = \phi(\boldsymbol{\Theta}_\star\boldsymbol{x}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t$)

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 $\boldsymbol{x}_t = \phi(\boldsymbol{A}_1\boldsymbol{x}_{t-1} + \boldsymbol{A}_2\boldsymbol{x}_{t-2} + \cdots + \boldsymbol{A}_m\boldsymbol{x}_{t-m}) + \boldsymbol{w}_{t-1}$ 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.

2.2 Preliminaries and Problem Setup

We assume the system is driven by inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{x}_t) + \boldsymbol{z}_t$, 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 (2.1.1) becomes,

$$\boldsymbol{x}_{t+1} = \phi(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t) + \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t := \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t, \tag{2.2.1}$$

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 $(\boldsymbol{x}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ generated by the nonlinear system (2.2.1). We also assume access to a stabilizing control policy $\boldsymbol{\pi}(\cdot)$. A special case of (2.2.1) is a linear state equation with $\boldsymbol{\theta}_\star = [\boldsymbol{A}_\star \;\; \boldsymbol{B}_\star]$, $\boldsymbol{\pi}(\boldsymbol{x}_t) = -\boldsymbol{K}\boldsymbol{x}_t$ and

$$\boldsymbol{x}_{t+1} = (\boldsymbol{A}_\star - \boldsymbol{B}_\star\boldsymbol{K})\boldsymbol{x}_t + \boldsymbol{B}_\star\boldsymbol{z}_t + \boldsymbol{w}_t, \tag{2.2.2}$$

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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{\theta}), \quad \text{subject to} \quad \hat{\mathcal{L}}(\boldsymbol{\theta}) := \frac{1}{2(T-L)}\sum_{t=L}^{T-1} \|\boldsymbol{x}_{t+1} - \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2, \tag{2.2.3}$$

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 (2.2.3), we investigate the properties of the gradient descent algorithm, given by the following iterate

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau), \tag{2.2.4}$$

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 Definition 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 $(\boldsymbol{x}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$. This will in turn allow us to leverage techniques developed for i.i.d. data to solve problems with sequential data.

2.2.1 Assumptions on the System and the Inputs

We assume that the closed-loop system $\tilde{\phi}$ is stable. Stability in the case of standard linear time-invariant dynamical systems is connected to the spectral radius of the closed-loop system [22,65]. The definition below provides a natural generalization of stability to nonlinear dynamical 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 (2.2.1) resulting from initial state $\boldsymbol{x}_0 = \boldsymbol{\alpha}$, $(\boldsymbol{z}_\tau)_{\tau=0}^{t-1}$ and $(\boldsymbol{w}_\tau)_{\tau=0}^{t-1}$ by $\boldsymbol{x}_t(\boldsymbol{\alpha})$. 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*

$$\|\boldsymbol{x}_t(\boldsymbol{\alpha}) - \boldsymbol{x}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2}. \tag{2.2.5}$$

Note that, for a stable LTI dynamical system ($\rho(\boldsymbol{A}_\star) < 1$), 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 $\boldsymbol{x}_t$ 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 [81] and sum-of-squares [82] 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 $\boldsymbol{x}_t(\boldsymbol{\alpha})$ and $\boldsymbol{x}_t(0)$ do not have to converge, rather their difference $\|\boldsymbol{x}_t(\boldsymbol{\alpha}) - \boldsymbol{x}_t(0)\|_{\ell_2}$ exponentially converges to zero (assuming $\|\boldsymbol{\alpha}\|_{\ell_2}$ is bounded).

To keep the exposition simple, we will also assume $\boldsymbol{x}_0 = 0$ 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.*

2.2.2 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 $\boldsymbol{x}_0 = 0$. 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,*

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) &= \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, (\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}))], \\ \textit{where} \quad \mathcal{L}(\boldsymbol{\theta}, (\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})) &:= \frac{1}{2}\|\boldsymbol{x}_L - \tilde{\phi}(\boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}; \boldsymbol{\theta})\|_{\ell_2}^2. \end{aligned} \tag{2.2.6}$$

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*

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}_\star, \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\rangle \geq \alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}^2, \tag{2.2.7}$$

$$\|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq \beta\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}. \tag{2.2.8}$$

We emphasize that, as opposed to traditional strong convexity and smoothness assumptions [83], 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 [84]. 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 $\boldsymbol{x}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{x}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t$, with γ-increasing activation (i.e., $\phi'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$) and Gaussian excitation/noise (see Lemma 65). 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 [85] 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 $\boldsymbol{x}_t = \phi(\boldsymbol{A}_1\boldsymbol{x}_{t-1} + \boldsymbol{A}_2\boldsymbol{x}_{t-2} + \cdots + \boldsymbol{A}_m\boldsymbol{x}_{t-m}) + \boldsymbol{w}_{t-1}$ when the joint feature vector $[\boldsymbol{x}_{L-1}^\top \ \boldsymbol{x}_{L-2}^\top \cdots \boldsymbol{x}_{L-m}^\top]^\top$ 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* (2.2.4) *to solve the ERM problem* (2.2.3)*. 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*

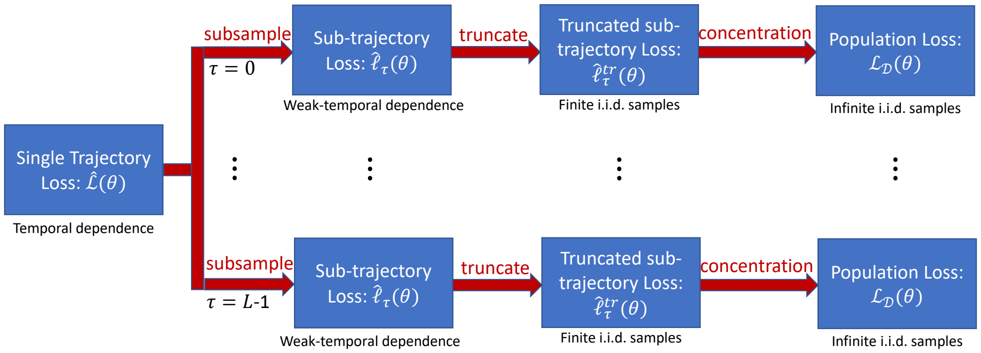
$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\sqrt{\frac{d}{T(1-\rho)}} \tag{2.2.9}$$

- ***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*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \lesssim (1 - \frac{\alpha^2}{128\beta^2})^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}. \tag{2.2.10}$$

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 [33, 71, 86]. 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 (2.2.10) 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 [84]. 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 (2.2.1).



Single Trajectory Loss: $\hat{\mathcal{L}}(\theta)$
Temporal dependence
subsample
$\tau = 0$
Sub-trajectory Loss: $\hat{\ell}_\tau(\theta)$
Weak-temporal dependence
truncate
Truncated sub-trajectory Loss: $\hat{\ell}^{tr}_\tau(\theta)$
Finite i.i.d. samples
concentration
Population Loss: $\mathcal{L}_\mathcal{D}(\theta)$
Infinite i.i.d. samples
subsample
$\tau = L$-1
Sub-trajectory Loss: $\hat{\ell}_\tau(\theta)$
Weak-temporal dependence
truncate
Truncated sub-trajectory Loss: $\hat{\ell}^{tr}_\tau(\theta)$
Finite i.i.d. samples
concentration
Population Loss: $\mathcal{L}_\mathcal{D}(\theta)$
Infinite i.i.d. samples

Figure 2.1: 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 $\|\nabla\hat{\ell}_\tau(\boldsymbol{\theta}) - \nabla\hat{\ell}^{\text{tr}}_\tau(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})C_\rho\rho^{L-1}$, where $\hat{\ell}^{\text{tr}}_\tau(\boldsymbol{\theta})$ is the loss constructed with finite i.i.d. samples. Next, we show the uniform convergence of the empirical gradient as $\|\nabla\hat{\ell}^{\text{tr}}_\tau(\boldsymbol{\theta}) - \nabla\mathcal{L}_\mathcal{D}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\sqrt{d/N}$, where $\mathcal{L}_\mathcal{D}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}^{\text{tr}}_\tau(\boldsymbol{\theta})]$ is the population loss. Finally, we combine these with the local one-point convexity of the population loss to get our main results.

2.3 Main Results

2.3.1 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 (2.2.1), the sample $\boldsymbol{x}$ corresponds to the variables $(\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})$ triple and the loss function $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})$ is given by (2.2.6). Define the empirical and population losses,

$$\hat{\mathcal{L}}_\mathcal{S}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{and} \quad \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})]. \tag{2.3.1}$$

Let $\boldsymbol{\theta}_\star$ denotes the population minimizer which we wish to estimate via gradient descent. Recent works by [79, 80] 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 (2.2.3).

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 $i = 1, \dots, N$. 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

$$\hat{\mathcal{L}}_Q(\boldsymbol{\theta}) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \phi(\boldsymbol{z}_i^\top \boldsymbol{\theta}))^2. \tag{2.3.2}$$

- **The role of noise:** Suppose ϕ is identity and the problem is purely linear regression. Then, gradient descent estimator will achieve statistical accuracy $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star\|_{\ell_2} \lesssim \sigma\sqrt{d/N}$. [79, 80] 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 $\mathcal{B}^d(\boldsymbol{\theta}_\star, r)$.

- **Activation ϕ:** Both [79, 80] can only handle bounded activation ϕ. [80] uses boundedness to control Rademacher complexity, whereas, [79] 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 [79], 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 (2.3.1) satisfy*

$$\max(\|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}')\|_{\ell_2}, \|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2}) \le L_{\mathcal{D}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\ell_2}. \tag{2.3.3}$$

Observe that, by definition, the Lipschitz constant obeys $L_{\mathcal{D}} \ge \beta$ where β is the one-point smoothness parameter in Assumption 3. However, $L_{\mathcal{D}}$ is allowed be much larger than β. Specifically, $L_{\mathcal{D}}$ will only appear logarithmically in our bounds, hence, we can tolerate very large values of $L_{\mathcal{D}}$. 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 (2.3.1) is upper bounded as a function of the noise level σ_0 and distance to the population minimizer via*

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}) - \mathbb{E}[\nabla\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})]\|_{\psi_1} \le \sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}, \tag{2.3.4}$$

where the subexponential norm of a random variable X is defined as $\|X\|_{\psi_1} := \sup_{k\ge1} \frac{(\mathbb{E}[|X|^k])^{1/k}}{k}$ 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 [79] 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 (2.3.2). 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 B, we show that in the case of nonlinear state equations $\boldsymbol{x}_{t+1}=\phi(\boldsymbol{\Theta}_\star\boldsymbol{x}_t)+\boldsymbol{z}_t+\boldsymbol{w}_t$, 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 $\boldsymbol{\Sigma}[\boldsymbol{x}_t]\le\beta_+^2\boldsymbol{I}_n$ (see the proof of Lemma 68), then Assumption 4 holds with $L_{\mathcal{D}}=c((1+\sigma)\beta_+^2n+\|\boldsymbol{\Theta}_\star\|_F\beta_+^3n^{3/2}\log^{3/2}(2T))$ 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*

$$\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta})-\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}\le c_0(\sigma_0+K\|\boldsymbol{\theta}-\boldsymbol{\theta}_\star\|_{\ell_2})C_{\log}\sqrt{\frac{d}{N}}. \tag{2.3.5}$$

Proof sketch: Our proof technique uses peeling argument [87] 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 [86] (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 [79,80] 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, [79] 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 (2.2.3), 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 a single finite trajectory.

2.3.2 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 [31]) *Consider the state equation (2.2.1). Suppose $\tilde{\phi}(0,0;\boldsymbol{\theta}) = 0$, $\boldsymbol{x}_0 = 0$. Given, $t \geq L > 0$, for each state $\boldsymbol{x}_t$, we define its fictional proxy $\boldsymbol{x}_{t,L}$ by resetting $\boldsymbol{x}_{t-L} = 0$ but preserving the excitation $\boldsymbol{z}_\tau$ and noise $\boldsymbol{w}_\tau$ from $t-L$ to $t-1$. Alternately, $\boldsymbol{x}_{t,L}$ is obtained by driving the system with excitations $\boldsymbol{z}'_\tau$ and additive noise $\boldsymbol{w}'_\tau$ until time $t-1$, where*

$$\boldsymbol{z}'_\tau = \begin{cases} 0 \ \textit{if}\, \tau < t-L \\ \\ \boldsymbol{z}_\tau \ \textit{else} \end{cases} \qquad \textit{and} \qquad \boldsymbol{w}'_\tau = \begin{cases} 0 \ \textit{if}\, \tau < t-L \\ \\ \boldsymbol{w}_\tau \ \textit{else} \end{cases}. \tag{2.3.6}$$

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

The L-truncated state vector $\boldsymbol{x}_{t,L}$ is identically distributed as $\boldsymbol{x}_L$. 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 [31, 32]. The difference between L-truncated and non-truncated state vectors is guaranteed to be bounded as

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t,L}\|_{\ell_2} \leq C_\rho \rho^L \|\boldsymbol{x}_{t-L}\|_{\ell_2}. \tag{2.3.7}$$

This directly follows from Definition 1. To tightly capture the effect of truncation, we also bound the Euclidean norm of states $\boldsymbol{x}_t$ as follows.

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

Following this and (2.3.7), we can obtain weakly dependent sub-trajectories by properly sub-sampling a single trajectory $(\boldsymbol{x}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$. For this purpose, we first define a sub-trajectory and its truncation as follows.

Definition 7 (Truncated sub-trajectories [31]) *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* $(\boldsymbol{x}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ *at points* $\tau+L, \tau+2L, \dots, \tau+NL$ *and truncate the states by* $L-1$ *to get the* τ_{th} *truncated sub-trajectory* $(\bar{\boldsymbol{x}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N$*, defined as*

$$(\bar{\boldsymbol{x}}^{(i)}, \boldsymbol{z}^{(i)}) := (\boldsymbol{x}_{\tau+iL,L-1}, \boldsymbol{z}_{\tau+iL}) \quad \textit{for} \quad i = 1, \dots, N \tag{2.3.8}$$

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 $(\bar{\boldsymbol{x}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N$ 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* $(\bar{\boldsymbol{x}}^{(i)})_{i=1}^N$ *are all independent and are identically distributed as* $\boldsymbol{x}_{L-1}$*. Moreover,* $(\bar{\boldsymbol{x}}^{(i)})_{i=1}^N, (\boldsymbol{z}^{(i)})_{i=1}^N, (\boldsymbol{w}^{(i)})_{i=1}^N$ *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* $(\bar{\boldsymbol{y}}^{(i)}, \bar{\boldsymbol{x}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^N := (\boldsymbol{x}_{\tau+iL+1,L}, \boldsymbol{x}_{\tau+iL,L-1}, \boldsymbol{z}_{\tau+iL})_{i=1}^N$ *as*

$$\hat{\ell}_\tau^{tr}(\boldsymbol{\theta}) := \frac{1}{2N}\sum_{i=1}^N \|\bar{\boldsymbol{y}}^{(i)} - \tilde{\phi}(\bar{\boldsymbol{x}}^{(i)}, \boldsymbol{z}^{(i)}; \boldsymbol{\theta})\|_{\ell_2}^2. \tag{2.3.9}$$

Observe that the triplets $(\bar{\boldsymbol{y}}^{(i)}, \bar{\boldsymbol{x}}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^{N}$ are independent and identically distributed as $(\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})$. 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 2.3.1 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 (2.2.3).

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

$$\hat{\mathcal{L}}^{tr}(\boldsymbol{\theta}) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{x}_{t+1,L} - \tilde{\phi}(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 = \frac{1}{L} \sum_{\tau=0}^{L-1} \hat{\ell}_{\tau}^{tr}(\boldsymbol{\theta}). \tag{2.3.10}$$

Let $\mathcal{X}$ be the convex hull of all states $\boldsymbol{x}_t$ 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 $(\boldsymbol{x}, \boldsymbol{z}) \subset \mathcal{X} \times \mathcal{Z}$. 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 given by (2.2.1). 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 $(\boldsymbol{x}, \boldsymbol{z}) \subset \mathcal{X} \times \mathcal{Z}$, we have that $\|\nabla_{\boldsymbol{x}} \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\| \leq B_{\tilde{\phi}}$, $\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} \leq C_{\tilde{\phi}}$ and $\|\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\| \leq D_{\tilde{\phi}}$ for some scalars $B_{\tilde{\phi}}, C_{\tilde{\phi}}, D_{\tilde{\phi}} > 0$ and $1 \leq k \leq n$. Let $\beta_+ > 0$ be as* *defined 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*

$$|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{tr}(\boldsymbol{\theta})| \le 2n\beta_+ C_\rho \rho^{L-1} B_{\tilde{\phi}}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}), \tag{2.3.11}$$

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\hat{\mathcal{L}}^{tr}(\boldsymbol{\theta})\|_{\ell_2} \le 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \tag{2.3.12}$$

Proof sketch: To prove Theorem 11, we use the Mean-value Theorem together with Assumptions 1 and 2. First, using (2.2.3) and (2.3.10), we obtain

$$\begin{aligned}|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta})| \le \frac{1}{2} \max_{L \le t \le (T-1)} |\|\tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t - \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 \\ - \|\tilde{\phi}(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t - \tilde{\phi}(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|.\end{aligned} \tag{2.3.13}$$

Suppose, the maximum is achieved at $(\boldsymbol{x}, \bar{\boldsymbol{x}}, \boldsymbol{z}, \boldsymbol{w})$ (where $\bar{\boldsymbol{x}}$ is the truncated state). Then, we use the identity $a^2 - b^2 = (a+b)(a-b)$ to upper bound the difference $|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta})|$ as a product of two terms $|a+b|$ and $|a-b|$ with $a := \|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}$ and $b := \|\tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}$. 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 $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\ell_2}$). Combining the two bounds gives us the statement (2.3.11) of the Theorem. A similar proof technique is used to upper bound the gradient distance $\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\hat{\mathcal{L}}^{\mathrm{tr}}(\boldsymbol{\theta})\|_{\ell_2}$. ■

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.

2.3.3 Non-asymptotic Identification of Nonlinear Dynamical 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 $(\boldsymbol{x}, \boldsymbol{z}) \subset \mathcal{X} \times \mathcal{Z}$, we have $\|\nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} \le C_{\tilde{\phi}}$ and $\|\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\| \le D_{\tilde{\phi}}$ 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*

$$L \ge L_0 \quad \textit{where} \quad L_0 = \left\lceil 1 + \frac{\log(C\rho K_{\tilde{\phi}} n \sqrt{N/d})}{1-\rho} \right\rceil. \tag{2.3.14}$$

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

$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le 2c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) C_{\log} \sqrt{\frac{d}{N}}. \tag{2.3.15}$$

Proof sketch: Theorem 12 can be proved by combining the results of Theorems 4 and 11. The idea is to split the truncated loss $\hat{\mathcal{L}}^{\text{tr}}$ (Definition 10) as an average of L truncated subtrajectory losses $\hat{\ell}_\tau^{\text{tr}}$ (Definition 9) as: $\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta})$. Recall that $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta})]$. Then, we use Theorem 4 with a union bound over all $0 \le \tau \le L-1$ to upper bound $\|\nabla \hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}$ which is used to show the uniform convergence of the truncated loss $\hat{\mathcal{L}}^{\text{tr}}$ as: $\|\nabla \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le \frac{1}{L}\sum_{\tau=0}^{L-1} \|\nabla \hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}$. Combining this with Theorem 11 and picking L via (2.3.14), 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 (2.2.2), 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*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \le \left(1 - \frac{\alpha^2}{128\beta^2}\right)^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{c\sigma_0}{\alpha} C_{\log}\sqrt{\frac{d}{N}}. \tag{2.3.16}$$

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 $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \le \left(1 - \frac{\alpha^2}{128\beta^2}\right)^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{5\nu}{\alpha}$. 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 Section 2.4. 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 (2.2.1) provides n equations. For example, we can accomplish better sample complexity for separable dynamical systems (see Section 2.3.4) 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 [31,33,71]. In contrast, it is well-known that this dependency (on $\rho(\boldsymbol{A}_\star)$) can be avoided for learning linear dynamical systems [22]. Recently, [34] showed, under a strong invertibility condition, that dependency on the mixing time can be avoided for the generalized linear models $\boldsymbol{x}_{t+1} = \phi(\boldsymbol{A}_\star \boldsymbol{x}_t) + \boldsymbol{w}_t$. This leaves open the question of whether learning without mixing is possible in situations beyond the generalized linear models.

2.3.4 Non-asymptotic Identification of Separable Dynamical Systems

Suppose now that the nonlinear dynamical system is separable, that is, the nonlinear state equation (2.2.1) can be split into n state updates via

$$\boldsymbol{x}_{t+1}[k] = \tilde{\phi}_k(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_k^\star) + \boldsymbol{w}_t[k], \quad \text{for } 1 \leq k \leq n, \tag{2.3.17}$$

where $\boldsymbol{x}_t[k]$ and $\boldsymbol{w}_t[k]$ denote the k_{th} entry of $\boldsymbol{x}_t$ 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 (2.2.3) is alternately given by,

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} (\boldsymbol{x}_{t+1}[k] - \tilde{\phi}_k(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta}_k))^2. \tag{2.3.18}$$

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 (2.2.3) 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 (2.2.3), 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,

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k \in \mathbb{R}^{\bar{d}}} \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k). \tag{2.3.19}$$

Similarly global iterations (2.2.4) follows the iterations of the subproblems, that is, the GD iterate (2.2.4) 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,

$$\begin{aligned} \hat{\mathcal{L}}_\mathcal{S}(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{x}_i), \\ \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) \quad \text{where} \quad \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) = \mathbb{E}[\mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{x})], \end{aligned} \tag{2.3.20}$$

where $\mathcal{L}_k(\cdot, \boldsymbol{x})$ is a loss function that admits a sample $\boldsymbol{x}$ and outputs the corresponding loss. When learning (2.3.17), the sample $\boldsymbol{x}$ corresponds to the variables $(\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})$ triple and the loss function $\mathcal{L}_k(\boldsymbol{\theta}, \boldsymbol{x})$ is given by

$$\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})) := \frac{1}{2}(\boldsymbol{x}_L[k] - \tilde{\phi}_k(\boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}; \boldsymbol{\theta}_k))^2. \tag{2.3.21}$$

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 (2.3.20) hold for all* $1 \le k \le n$*. Additionally, suppose the nonlinear dynamical system is separable, that is, the nonlinear state equation follows (2.3.17). Let* $K_{\tilde{\phi}}$ *be as in Theorem 12. Let* $N = \lfloor (T-L)/L \rfloor$*, where we pick* L *via*

$$L \ge L_0 \quad \textit{where} \quad L_0 = \lceil 1 + \frac{\log(C\rho K_{\tilde{\phi}} n\sqrt{N/\bar{d}})}{1-\rho} \rceil. \tag{2.3.22}$$

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*

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le \big(1 - \frac{\alpha^2}{128\beta^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma_0}{\alpha} C_{\log}\sqrt{\frac{\bar{d}}{N}} \quad \textit{for all} \quad 1 \le k \le n. \tag{2.3.23}$$

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 (2.3.20), 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 (2.3.20) 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 standard LTI 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.

2.4 Applications of Theorems 13 and 14

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

$$\boldsymbol{x}_{t+1} = \phi(\boldsymbol{A}_\star \boldsymbol{x}_t) + \boldsymbol{B}_\star \boldsymbol{z}_t + \boldsymbol{w}_t, \tag{2.4.1}$$

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 LTI dynamical systems ($\phi = \boldsymbol{I}_n$); and **(b)** Nonlinear state equations

$$\boldsymbol{x}_{t+1} = \phi(\boldsymbol{\Theta}_\star \boldsymbol{x}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t, \tag{2.4.2}$$

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 $\boldsymbol{\pi}(\boldsymbol{x}_t) = 0$. 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.

2.4.1 Linear Dynamical Systems

To simplify the notation, we define the following concatenated vector/matrix: $\boldsymbol{h}_t := [\boldsymbol{x}_t^\top \; \boldsymbol{z}_t^\top]^\top$ and $\boldsymbol{\Theta}_\star := [\boldsymbol{A}_\star \; \boldsymbol{B}_\star]$. Letting $\phi = \boldsymbol{I}_n$, the state update (2.4.1) is alternately given by: $\boldsymbol{x}_{t+1} = \boldsymbol{\Theta}_\star \boldsymbol{h}_t + \boldsymbol{w}_t$. 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 LTI dynamical system is separable as in (2.3.17). Therefore, given a finite trajectory $(\boldsymbol{x}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ of the linear dynamical system (2.4.1) ($\phi = \boldsymbol{I}_n$), we construct the empirical loss as follows,

$$\hat{\mathcal{L}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} (\boldsymbol{x}_{t+1}[k] - \boldsymbol{\theta}_k^\top \boldsymbol{h}_t)^2. \tag{2.4.3}$$

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,

$$\boldsymbol{G}_t := [\boldsymbol{A}_\star^{t-1}\boldsymbol{B}_\star \ \boldsymbol{A}_\star^{t-2}\boldsymbol{B}_\star \ \cdots \ \boldsymbol{B}_\star] \quad \text{and} \quad \boldsymbol{F}_t := [\boldsymbol{A}_\star^{t-1} \ \boldsymbol{A}_\star^{t-2} \ \cdots \ \boldsymbol{I}_n]. \tag{2.4.4}$$

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 $\boldsymbol{h}_t$ satisfies the following bounds (see Section A for detail)

$$(1 \wedge \lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p} \preceq \boldsymbol{\Sigma}[\boldsymbol{h}_t] \preceq (1 \vee \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_{n+p}). \tag{2.4.5}$$

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 (2.4.1) with $\phi = \boldsymbol{I}_n$. Suppose $\rho(\boldsymbol{A}_\star) < 1$. Let $C_\rho \ge 1$ 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 $\gamma_+ \ge \gamma_- > 0$ be as defined in (2.4.5) and set $\kappa = \gamma_+/\gamma_-$. Let $N = \lfloor (T-L)/L \rfloor$, where we pick L via*

$$L \ge L_0 \quad \text{where} \quad L_0 = \left\lceil 1 + \frac{\log(CC_\rho\beta_+N(n+p)/\gamma_+)}{1-\rho} \right\rceil. \tag{2.4.6}$$

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* $1 \le k \le n$, *all gradient descent iterates* $\boldsymbol{\Theta}_\tau = [\boldsymbol{\theta}_1^{(\tau)} \cdots \boldsymbol{\theta}_n^{(\tau)}]^\top$ *on* $\hat{\mathcal{L}}$ *satisfy*

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le \big(1 - \frac{\gamma_-^2}{128\gamma_+^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma\sqrt{\kappa}}{\sqrt{\gamma_-}}\log(6N+3)\sqrt{\frac{n+p}{N}}. \tag{2.4.7}$$

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 $\boldsymbol{\Sigma}[\boldsymbol{h}_{L-1}]$. 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 $\boldsymbol{\Sigma}[\boldsymbol{h}_{L-1}]$. 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.

2.4.2 Nonlinear State Equations

In this section, we apply Theorem 14 to learn the nonlinear state equation (2.4.2). Observe that the nonlinear system (2.4.2) 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 $(\boldsymbol{x}_t, \boldsymbol{z}_t)_{t=0}^{T-1}$ of (2.4.2), we construct the empirical loss as follows,

$$\hat{\mathcal{L}}(\boldsymbol{\Theta}) = \sum_{k=1}^n \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \quad \text{where} \quad \hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) \coloneqq \frac{1}{2(T-L)}\sum_{t=L}^{T-1}(\boldsymbol{x}_{t+1}[k] - \phi(\boldsymbol{\theta}_k^\top\boldsymbol{x}_t) - \boldsymbol{z}_t[k])^2. \tag{2.4.8}$$

The following corollary of Theorem 14 states our main result on the statistical and convergence guarantees of gradient descent for learning the nonlinear system (2.4.2).

Corollary 16 *Suppose the nonlinear system (2.4.2) satisfies (C_ρ, ρ)-stability according to Def. 1. Suppose ϕ is γ-increasing (i.e. $\phi'(x) \ge \gamma > 0$ for all $x \in \mathbb{R}$), has bounded first and second derivatives, that is, $|\phi'|, |\phi''| \le 1$, 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*

$$L \ge L_0 \quad \text{where} \quad L_0 = \Big\lceil 1 + \frac{\log(CC_\rho(1 + \|\boldsymbol{\Theta}_\star\|_F C_\rho(1+\sigma)/(1-\rho))Nn)}{1-\rho} \Big\rceil. \tag{2.4.9}$$

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

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le \big(1 - \frac{\gamma^4(1-\rho)^4}{512C_\rho^4 n^2}\big)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma}{\gamma^2(1-\rho)} C_\rho D_{\log} \sqrt{\frac{n}{N}}. \tag{2.4.10}$$

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 [85] 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 $\boldsymbol{\Theta}_\star \in \mathbb{R}^{n \times n}$ since each sample gives n equations. The sample complexity bound depends on the condition number of the covariance matrix $\boldsymbol{\Sigma}[\boldsymbol{x}_t]$, which can be shown to be bounded by $C_\rho^2/(1-\rho)^2$ (see Section B). 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 Section 2.5. 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 Section 2.5. 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.

2.5 Proofs of the Main Results

2.5.1 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 [86] (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*

$$\|\frac{1}{N}\sum_{i=1}^{n}\boldsymbol{v}_i - \boldsymbol{\mu}\|_{\ell_2} \leq CK\frac{\sqrt{d}+t}{\sqrt{N}}. \tag{2.5.1}$$

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

$$\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}_i - \boldsymbol{\mu}\|_{\ell_2} \leq CK(\tau + 1)\sqrt{d/N}. \tag{2.5.2}$$

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. *Uniform convergence with covering argument:* We will use a peeling argument [87]. 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

$$\max(\sigma_0/K, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) \leq r_i \leq \max(\sigma_0/K, \mathrm{e}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \tag{2.5.3}$$

Applying Lemma 18 together with a union bound over the $P + 1$ covers and elements of the covers, we guarantee the following: With probability at least $1 - \sum_{i=0}^{P}\exp(-100d\log(3r_i/\varepsilon_i))$, within all covers $\mathcal{N}_i$, gradient vector at all points $\boldsymbol{\theta} \in \mathcal{N}_i$ satisfies

$$\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma_0 + Kr_i)\log(3r_i/\varepsilon_i)\sqrt{d/N}. \tag{2.5.4}$$

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 $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\ell_2} \le \varepsilon$. This yields

$$\begin{aligned}
&\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \\
&\qquad\qquad \le \|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}')\|_{\ell_2} + \|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2} + \|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}') - \nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2}, \\
&\qquad\qquad \lesssim \varepsilon_i L_{\mathcal{D}} + (\sigma_0 + Kr_i)\log(3r_i/\varepsilon_i)\sqrt{d/N}. \qquad (2.5.5)
\end{aligned}$$

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

$$\begin{aligned}
\|\nabla\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} &\lesssim (\sigma_0 + Kr_i)\log(3(1 + L_{\mathcal{D}}N/K))\sqrt{d/N}, \\
&\lesssim (\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(1 + L_{\mathcal{D}}N/K))\sqrt{d/N}, \qquad (2.5.6)
\end{aligned}$$

where we used (2.5.3) to get the last inequality. Finally, observing that $\log(3r_i/\varepsilon_i) \ge 1$, the probability bound simplifies to

$$1 - p_0 - \sum_{i=0}^{P} \exp(-100d\log(3r_i/\varepsilon_i)) \ge 1 - p_0 - \log(\frac{Kr}{\sigma_0})\exp(-100d). \qquad (2.5.7)$$

This completes the proof. ■

2.5.2 Proof of Lemma 6

Proof. Suppose $\boldsymbol{x}_0 = 0$. We claim that $\|\boldsymbol{x}_t\|_{\ell_2} \le \beta_+\sqrt{n}(1 - \rho^t)$ 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 $\boldsymbol{x}_1$ satisfies the following bound and obeys the induction

$$\|\boldsymbol{x}_1\|_{\ell_2} \le B\sqrt{n} + c_w\sigma\sqrt{n} \le C_\rho\sqrt{n}(B + c_w\sigma) = \beta_+\sqrt{n}(1 - \rho^1). \qquad (2.5.8)$$

Suppose the bound holds until $t-1$, where $t \le T$, and let us apply induction. First observe that $\|\boldsymbol{x}_{t,L}\|_{\ell_2}$ obeys the same upper bound as $\|\boldsymbol{x}_L\|_{\ell_2}$ by construction. Recalling (2.3.7), we get the following by induction

$$\begin{aligned}
\|\boldsymbol{x}_t - \boldsymbol{x}_{t,t-1}\|_{\ell_2} \le C_\rho \rho^{t-1}\|\boldsymbol{x}_1\|_{\ell_2} \implies \|\boldsymbol{x}_t\|_{\ell_2} &\le C_\rho \rho^{t-1}\|\boldsymbol{x}_1\|_{\ell_2} + \|\boldsymbol{x}_{t,t-1}\|_{\ell_2}, \\
&\overset{(a)}{\le} C_\rho \rho^{t-1}\|\boldsymbol{x}_1\|_{\ell_2} + \beta_+\sqrt{n}(1-\rho^{t-1}), \\
&\overset{(b)}{\le} \sqrt{n}(C_\rho \rho^{t-1}(B + c_w\sigma) + \beta_+(1-\rho^{t-1})), \\
&\le \beta_+\sqrt{n}(1-\rho^t), \qquad (2.5.9)
\end{aligned}$$

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

2.5.3 Proof of Lemma 8

Proof. By construction $\bar{\boldsymbol{x}}^{(i)}$ 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, $\{\bar{\boldsymbol{x}}^{(i)}\}_{i=1}^N$ are all independent of each other. To show the independence of $\{\bar{\boldsymbol{x}}^{(i)}\}_{i=1}^N$ 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 $\{\bar{\boldsymbol{x}}^{(i)}\}_{i=1}^N$. Identical argument shows the independence of $\{\bar{\boldsymbol{x}}^{(i)}\}_{i=1}^N$ 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, $\{\bar{\boldsymbol{x}}^{(i)}\}_{i=1}^N, \{\boldsymbol{z}^{(i)}\}_{i=1}^N, \{\boldsymbol{w}^{(i)}\}_{i=1}^N$ are all independent of each other. This completes the proof. ■

2.5.4 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 (2.2.3) and (2.3.10) respectively. The distance between them can be bounded as follows.

$$
\begin{aligned}
&|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})| \\
&\qquad = |\frac{1}{2(T-L)}\sum_{t=L}^{T-1}\|\boldsymbol{x}_{t+1} - \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 - \frac{1}{2(T-L)}\sum_{t=L}^{T-1}\|\boldsymbol{x}_{t+1,L} - \tilde{\phi}(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\qquad \le \frac{1}{2(T-L)}\sum_{t=L}^{T-1}|\|\boldsymbol{x}_{t+1} - \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 - \|\boldsymbol{x}_{t+1,L} - \tilde{\phi}(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\qquad \le \frac{1}{2}\max_{L \le t \le (T-1)}|\|\boldsymbol{x}_{t+1} - \tilde{\phi}(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2 - \|\boldsymbol{x}_{t+1,L} - \tilde{\phi}(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\qquad \le \frac{1}{2}|\|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}^2 - \|\tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}^2|, \\
&\qquad \le \frac{1}{2}(|\|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} - \|\tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}|) \\
&\qquad\quad (|\|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} + \|\tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}|),
\end{aligned} \tag{2.5.10}
$$

where, $(\boldsymbol{x}, \bar{\boldsymbol{x}}, \boldsymbol{z}, \boldsymbol{w})$ corresponds to the maximum index and we used the identity $a^2 - b^2 = (a+b)(a-b)$. Denote the k_{th} element of $\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ by $\tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ and that of $\boldsymbol{w}$ by w_k for

$1 \leq k \leq n$. To proceed, using Mean-value Theorem, with probability at least $1 - p_0$, we have

$$
\begin{aligned}
|\tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) + w_k| &\leq c_w \sigma + \sup_{\tilde{\boldsymbol{\theta}} \in [\boldsymbol{\theta}, \boldsymbol{\theta}_\star]} \|\nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \tilde{\boldsymbol{\theta}})\|_{\ell_2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}, \\
&\leq c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2} \quad \text{for all} \quad 1 \leq k \leq n, \qquad (2.5.11) \\
\implies \|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} &\leq \sqrt{n} \max_{1 \leq k \leq n} |\tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}_k(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) + w_k|, \\
&\leq \sqrt{n}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \qquad (2.5.12)
\end{aligned}
$$

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

$$
\begin{aligned}
&\frac{1}{2} |\|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} + \|\tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \boldsymbol{w} - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}| \\
&\qquad \leq \sqrt{n}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}). \qquad (2.5.13)
\end{aligned}
$$

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

$$
\begin{aligned}
\Delta &\leq \|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) + \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2}, \\
&\leq \|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\|_{\ell_2} + \|\tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - \tilde{\phi}(\bar{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star)\|_{\ell_2}, \\
&\leq \sup_{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}]} \|\nabla_{\boldsymbol{x}} \tilde{\phi}(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})\| \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\ell_2} + \sup_{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}]} \|\nabla_{\boldsymbol{x}} \tilde{\phi}(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star)\| \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\ell_2}, \\
&\overset{(a)}{\leq} B_{\tilde{\phi}} C_\rho \rho^{L-1} \beta_+ \sqrt{n} + B_{\tilde{\phi}} C_\rho \rho^{L-1} \beta_+ \sqrt{n}, \\
&= 2 B_{\tilde{\phi}} C_\rho \rho^{L-1} \beta_+ \sqrt{n}, \qquad (2.5.14)
\end{aligned}
$$

with probability at least $1 - p_0$, where we get (a) from (2.3.7) and the initial assumption that $\|\nabla_{\boldsymbol{x}} \tilde{\phi}(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\| \leq B_{\tilde{\phi}}$. Multiplying this bound with (2.5.13) 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

$$\begin{aligned}
\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta})-\nabla\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})\|_{\ell_2} &\leq \frac{1}{T-L}\sum_{t=L}^{T-1}\|\nabla_{\boldsymbol{\theta}}\tilde{\phi}(\boldsymbol{x}_t,\boldsymbol{z}_t;\boldsymbol{\theta})^{\top}(\tilde{\phi}(\boldsymbol{x}_t,\boldsymbol{z}_t;\boldsymbol{\theta})-\boldsymbol{x}_{t+1}) \\
&\quad -\nabla_{\boldsymbol{\theta}}\tilde{\phi}(\boldsymbol{x}_{t,L-1},\boldsymbol{z}_t;\boldsymbol{\theta})^{\top}(\tilde{\phi}(\boldsymbol{x}_{t,L-1},\boldsymbol{z}_t;\boldsymbol{\theta})-\boldsymbol{x}_{t+1,L})\|_{\ell_2}, \\
&\leq \max_{L\leq t\leq(T-1)}\|\nabla_{\boldsymbol{\theta}}\tilde{\phi}(\boldsymbol{x}_t,\boldsymbol{z}_t;\boldsymbol{\theta})^{\top}(\tilde{\phi}(\boldsymbol{x}_t,\boldsymbol{z}_t;\boldsymbol{\theta})-\boldsymbol{x}_{t+1}) \\
&\quad -\nabla_{\boldsymbol{\theta}}\tilde{\phi}(\boldsymbol{x}_{t,L-1},\boldsymbol{z}_t;\boldsymbol{\theta})^{\top}(\tilde{\phi}(\boldsymbol{x}_{t,L-1},\boldsymbol{z}_t;\boldsymbol{\theta})-\boldsymbol{x}_{t+1,L})\|_{\ell_2}, \\
&\leq \|\nabla_{\boldsymbol{\theta}}\tilde{\phi}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})^{\top}(\tilde{\phi}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})-\tilde{\phi}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}_\star)-\boldsymbol{w}) \\
&\quad -\nabla_{\boldsymbol{\theta}}\tilde{\phi}(\bar{\boldsymbol{x}},\boldsymbol{z};\boldsymbol{\theta})^{\top}(\tilde{\phi}(\bar{\boldsymbol{x}},\boldsymbol{z};\boldsymbol{\theta})-\tilde{\phi}(\bar{\boldsymbol{x}},\boldsymbol{z};\boldsymbol{\theta}_\star)-\boldsymbol{w})\|_{\ell_2}\leq\sqrt{n}\Lambda, \qquad (2.5.15)
\end{aligned}$$

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

$$\begin{aligned}
\Lambda := \max_{1\leq k\leq n}&\|(\tilde{\phi}_k(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})-\tilde{\phi}_k(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}_\star)-w_k)\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&-(\tilde{\phi}_k(\bar{\boldsymbol{x}},\boldsymbol{z};\boldsymbol{\theta})-\tilde{\phi}_k(\bar{\boldsymbol{x}},\boldsymbol{z};\boldsymbol{\theta}_\star)-w_k)\nabla_{\boldsymbol{\theta}}\tilde{\phi}_k(\bar{\boldsymbol{x}},\boldsymbol{z};\boldsymbol{\theta})\|_{\ell_2}, \qquad (2.5.16)
\end{aligned}$$

where $\tilde{\phi}_k(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ denotes the k_{th} element of $\tilde{\phi}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$. Without losing generality, suppose k is the coordinate achieving maximum value and attaining Λ. Note that $\Lambda=\alpha(\boldsymbol{x})-\alpha(\bar{\boldsymbol{x}})$ for some function α. Using Mean-value Theorem, we bound $\Lambda\leq\sup_{\tilde{\boldsymbol{x}}\in[\boldsymbol{x},\bar{\boldsymbol{x}}]}\|\nabla_{\boldsymbol{x}}\alpha(\tilde{\boldsymbol{x}})\|\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|_{\ell_2}$,

$$
\begin{aligned}
\Lambda \leq \sup_{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}]} & \| (\tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - w_k) \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) \\
& + \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) (\nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta})^\top - \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star)^\top) \| \| \boldsymbol{x} - \bar{\boldsymbol{x}} \|_{\ell_2}, \\
\leq \sup_{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}]} & \big[|\tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - w_k| \| \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) \| \\
& + \| \nabla_{\boldsymbol{\theta}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) \|_{\ell_2} \| \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) - \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) \|_{\ell_2} \big] \| \boldsymbol{x} - \bar{\boldsymbol{x}} \|_{\ell_2}, \\
\overset{(a)}{\leq} \sup_{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}]} & \big[D_{\tilde{\phi}} |\tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) - \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) - w_k| \\
& + C_{\tilde{\phi}} \| \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) - \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) \|_{\ell_2} \big] \| \boldsymbol{x} - \bar{\boldsymbol{x}} \|_{\ell_2},
\end{aligned}
\tag{2.5.17}
$$

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

$$
\begin{aligned}
\sup_{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}]} \| \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}) - \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \boldsymbol{\theta}_\star) \|_{\ell_2} & \leq \sup_{\substack{\tilde{\boldsymbol{x}} \in [\boldsymbol{x}, \bar{\boldsymbol{x}}] \\ \tilde{\boldsymbol{\theta}} \in [\boldsymbol{\theta}, \boldsymbol{\theta}_\star]}} \| \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{x}} \tilde{\phi}_k(\tilde{\boldsymbol{x}}, \boldsymbol{z}; \tilde{\boldsymbol{\theta}}) \| \| \boldsymbol{\theta} - \boldsymbol{\theta}_\star \|_{\ell_2}, \\
& \leq D_{\tilde{\phi}} \| \boldsymbol{\theta} - \boldsymbol{\theta}_\star \|_{\ell_2}.
\end{aligned}
\tag{2.5.18}
$$

Finally, plugging the bounds from (2.5.11) and (2.5.18) into (2.5.17), with probability at least $1 - p_0$, we have

$$
\begin{aligned}
\| \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) \|_{\ell_2} \leq \sqrt{n} \Lambda & \leq \sqrt{n} (D_{\tilde{\phi}} (c_w \sigma + C_{\tilde{\phi}} \| \boldsymbol{\theta} - \boldsymbol{\theta}_\star \|_{\ell_2}) + C_{\tilde{\phi}} D_{\tilde{\phi}} \| \boldsymbol{\theta} - \boldsymbol{\theta}_\star \|_{\ell_2}) \| \boldsymbol{x} - \bar{\boldsymbol{x}} \|_{\ell_2}, \\
& \leq 2n \beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}} (c_w \sigma + C_{\tilde{\phi}} \| \boldsymbol{\theta} - \boldsymbol{\theta}_\star \|_{\ell_2}),
\end{aligned}
\tag{2.5.19}
$$

This completes the proof. ■

2.5.5 Proof of Theorem 12

Proof. Theorem 12 is a direct consequence of combining the results from Sections 2.3.1 and 2.3.2. To begin our proof, consider the truncated sub-trajectory loss $\hat{\ell}_\tau^{\text{tr}}$ from Definition 9 which also implies that $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta})]$. Hence, $\hat{\ell}_\tau^{\text{tr}}$ 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 $\hat{\ell}_\tau^{\text{tr}}$, 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

$$\|\nabla\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{d/N}, \tag{2.5.20}$$

for all $0 \leq \tau \leq L-1$, where we get the advertised probability by union bounding over all $0 \leq \tau \leq L-1$. Next, observe that the truncated loss $\hat{\mathcal{L}}^{\text{tr}}$ can be split into (average of) L sub-trajectory losses via $\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\tau=0}^{L-1}\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta})$. 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

$$\begin{aligned}
\|\nabla\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} &\leq \frac{1}{L}\sum_{\tau=0}^{L-1}\|\nabla\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}, \\
&\leq \max_{0\leq\tau\leq(L-1)}\|\nabla\hat{\ell}_\tau^{\text{tr}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2}, \\
&\leq c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{d/N}.
\end{aligned} \tag{2.5.21}$$

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

$$\begin{aligned}
\|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} &\leq \|\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} + \|\hat{\mathcal{L}}(\boldsymbol{\theta}) - \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta})\|_{\ell_2}, \\
&\leq c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{d/N} \\
&\quad + 2n\beta_+C_\rho\rho^{L-1}D_{\tilde{\phi}}(c_w\sigma + C_{\tilde{\phi}}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}).
\end{aligned} \tag{2.5.22}$$

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

$$\begin{aligned} & 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}} \leq c_0(\sigma_0/c_w\sigma \wedge K/C_{\tilde{\phi}}) \log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}, \\ \iff & \rho^{L-1} \leq (\sigma_0/c_w\sigma \wedge K/C_{\tilde{\phi}}) \frac{c_0 \log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}}{2n\beta_+ C_\rho D_{\tilde{\phi}}}, \\ \iff & L \geq 1 + \big[\log\big(\frac{2n\beta_+ C_\rho D_{\tilde{\phi}}\sqrt{N/d}}{c_0 \log(3(L_{\mathcal{D}}N/K+1))}\big) + \log(c_w\sigma/\sigma_0 \vee C_{\tilde{\phi}}/K)\big]/\log(\rho^{-1}), \\ \impliedby & L \geq \Big\lceil 1 + \frac{\log((2/c_0)n\beta_+ C_\rho D_{\tilde{\phi}}\sqrt{N/d}(c_w\sigma/\sigma_0 \vee C_{\tilde{\phi}}/K))}{1-\rho}\Big\rceil. \end{aligned} \tag{2.5.23}$$

Hence, picking L via (2.5.23), 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

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq 2c_0(\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}. \tag{2.5.24}$$

This completes the proof. ■

2.5.6 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: $\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq \nu + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$.

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 $\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq \nu + (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}$ and $r \geq 5\nu/\alpha$. 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*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \leq \big(1 - \frac{\alpha^2}{128\beta^2}\big)^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{5\nu}{\alpha}. \tag{2.5.25}$$

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

$$\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2}^2 = \|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - 2\eta\left\langle\boldsymbol{\delta}_\tau, \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\right\rangle + \eta^2\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2. \tag{2.5.26}$$

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

$$\begin{aligned}\left\langle\boldsymbol{\delta}_\tau, \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\right\rangle &\ge \langle\boldsymbol{\delta}_\tau, \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\rangle - |\left\langle\boldsymbol{\delta}_\tau, \nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\right\rangle|,\\ &\ge \alpha\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - (\nu + (\alpha/2)\|\boldsymbol{\delta}_\tau\|_{\ell_2})\|\boldsymbol{\delta}_\tau\|_{\ell_2} \ge (\alpha/2)\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 - \nu\|\boldsymbol{\delta}_\tau\|_{\ell_2}.\end{aligned} \tag{2.5.27}$$

Similarly,

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le \|\nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\|_{\ell_2} + \|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_\tau) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le (3/2)\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu. \tag{2.5.28}$$

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

$$\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2}^2 \le \|\boldsymbol{\delta}_\tau\|_{\ell_2}^2(1 - \eta\alpha/2 + 4\eta^2\beta^2) \le (1 - \frac{\alpha^2}{64\beta^2})\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2. \tag{2.5.29}$$

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.

$$\begin{aligned}\|\boldsymbol{\delta}_{\tau+1}\|_{\ell_2}^2 &\le \|\boldsymbol{\delta}_\tau\|_{\ell_2}^2(1 - \eta\alpha + (9/4)\eta^2\beta^2) + 2\eta\nu\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \eta^2(3\nu\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu^2),\\ &\le (1 - \frac{3\alpha^2}{64\beta^2})\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 + \frac{\alpha}{8\beta^2}\nu\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \frac{\alpha^2}{256\beta^4}(3\nu\beta\|\boldsymbol{\delta}_\tau\|_{\ell_2} + \nu^2),\\ &\le (\frac{16}{\alpha^2} + \frac{1}{2\beta^2} + \frac{3\alpha}{64\beta^3} + \frac{\alpha^2}{256\beta^4})\nu^2 \le \frac{25}{\alpha^2}\nu^2,\end{aligned} \tag{2.5.30}$$

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

$$\|\boldsymbol{\delta}_\tau\|_{\ell_2}^2 \le (1 - \frac{\alpha^2}{64\beta^2})^\tau\|\boldsymbol{\delta}_0\|_{\ell_2}^2 \implies \|\boldsymbol{\delta}_\tau\|_{\ell_2} \le (1 - \frac{\alpha^2}{128\beta^2})^\tau\|\boldsymbol{\delta}_0\|_{\ell_2}. \tag{2.5.31}$$

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

- 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 $\boldsymbol{\delta}_{\tau+1} \le 5\nu/\alpha$.

This completes the 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 $N \ge 16c_0^2K^2\log^2(3(L_{\mathcal{D}}N/K+1))d/\alpha^2$ in Theorem 12 to obtain

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \le (\alpha/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2} + 2c_0\sigma_0\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N}, \tag{2.5.32}$$

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 $N \ge 16c_0^2K^2\log^2(3(L_{\mathcal{D}}N/K+1))d/\alpha^2$, the statistical error rate of our non-asymptotic identification can be upper bounded as follows,

$$\frac{5\nu}{\alpha} = \frac{10c_0\sigma_0}{\alpha}\log(3(L_{\mathcal{D}}N/K+1))\sqrt{d/N} \lesssim \sigma_0/K. \tag{2.5.33}$$

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.

2.5.7 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 (2.3.20) satisfy

$$\max(\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2}, \|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2}) \le L_{\mathcal{D}}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}, \tag{2.5.34}$$

for all $1 \le k \le n$. 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

$$\|\nabla\mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{x}) - \mathbb{E}[\nabla\mathcal{L}_k(\boldsymbol{\theta}_k, \boldsymbol{x})\|_{\psi_1} \le \sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} \quad \text{for all} \quad 1 \le k \le n. \tag{2.5.35}$$

To proceed, using Theorem 4 with Assumptions 4 and 5 replaced by (2.5.34) and (2.5.35) 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 $1 \le k \le n$, we have

$$\|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \le c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}. \tag{2.5.36}$$

- **Small impact of truncation:** Next, we relate the gradients of the single trajectory loss $\hat{\mathcal{L}}_k$ in (2.3.18) and the multiple trajectory loss $\hat{\mathcal{L}}_k^{\text{tr}}$ (defined below). Similar to (2.3.18), the truncated loss for separable dynamical systems is alternately given by

$$\begin{aligned} \hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k), \\ \text{where} \quad \hat{\mathcal{L}}_k^{\text{tr}}(\boldsymbol{\theta}_k) &:= \frac{1}{2(T-L)}\sum_{t=L}^{T-1}(\boldsymbol{x}_{t+1,L}[k] - \tilde{\phi}_k(\boldsymbol{x}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta}_k))^2, \end{aligned} \tag{2.5.37}$$

where $\boldsymbol{x}_{t,L}[k]$ denotes the k_{th} element of the truncated vector $\boldsymbol{x}_{t,L}$. 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 $\hat{\mathcal{L}}^{\text{tr}}_k$ 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

$$\|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}^{\text{tr}}_k(\boldsymbol{\theta}_k)\|_{\ell_2} \le 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}}(c_w\sigma + C_{\tilde{\phi}}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star_k\|_{\ell_2}). \tag{2.5.38}$$

- **Combined result:** Next, we combine (2.5.36) and (2.5.38) 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 $\hat{\mathcal{L}}^{\text{tr}}_k$ 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

$$\|\nabla\hat{\mathcal{L}}^{\text{tr}}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \le c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star_k\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}. \tag{2.5.39}$$

Combining this with (2.5.38), 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

$$\begin{aligned}\|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &\le \|\nabla\hat{\mathcal{L}}^{\text{tr}}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} + \|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}^{\text{tr}}_k(\boldsymbol{\theta}_k)\|_{\ell_2},\\ &\le c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star_k\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}\\ &+ 2n\beta_+ C_\rho \rho^{L-1} D_{\tilde{\phi}}(c_w\sigma + C_{\tilde{\phi}}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star_k\|_{\ell_2}).\end{aligned} \tag{2.5.40}$$

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

$$L \ge \left\lceil 1 + \frac{\log((2/c_0)n\beta_+ C_\rho D_{\tilde{\phi}}\sqrt{N/\bar{d}}(c_w\sigma/\sigma_0 \vee C_{\tilde{\phi}}/K))}{1 - \rho}\right\rceil. \tag{2.5.41}$$

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

$$\begin{aligned}\|\nabla\hat{\mathcal{L}}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &\le 2c_0(\sigma_0 + K\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2})\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}, \\ &\overset{(a)}{\le} (\alpha/2)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + 2c_0\sigma_0\log(3(L_{\mathcal{D}}N/K + 1))\sqrt{\bar{d}/N}, \quad (2.5.42)\end{aligned}$$

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

- **One-point convexity & smoothness:** Lastly, Assumption 3 on the Auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ states that, there exist scalars $\beta \ge \alpha > 0$ such that, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$ and $1 \le k \le n$, the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$ of (2.3.20) satisfies

$$\langle\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\rangle \ge \alpha\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2, \quad (2.5.43)$$

$$\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \le \beta\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}. \quad (2.5.44)$$

- **Finalizing the proof:** We are now ready to use Theorem 19 with gradient concentration bound given by (2.5.42) and the OPCS Assumptions given by (2.5.43) and (2.5.44). 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 (2.5.43) and the one-point smoothness assumption (2.5.44) to get the statement of the theorem. This completes the proof. ■

Chapter 3

Bilinear System Identification

3.1 Introduction

Bilinear systems constitute an important class of nonlinear systems used in modeling systems in a variety of domains from engineering to biology [88]. They, together with state affine systems, also provide global approximators for more general nonlinear systems [89, 90], and have recently been invoked in the study of Koopman operators for systems with control inputs [91–93]. Due to the ubiquity of bilinear models, identification of such models from input-output data has also received interest in the literature both in continuous-time [94, 95] and discrete-time [96]. However, a theoretical understanding of learning a bilinear model from a finite noisy trajectory, and in particular, how the accuracy of the learned model depends on the trajectory length is lacking. We aim to answer this question for discrete-time bilinear models, learned from a single state-input trajectory using least squares.

3.1.1 Relation to Prior Work

There is a growing body of literature on non-asymptotic properties and sample complexity of learning dynamical systems. For linear systems, the recent results include [22–25, 28, 50–52, 54, 55, 60, 97–99] that establish that accuracy of the learned models improve at a rate $\mathcal{O}(1/\sqrt{T})$, where T is the trajectory length. These results are extended to certain classes of switched [39, 41, 100] and nonlinear systems [29, 31, 34, 101, 102], where, with the exception of [34], mixing-time arguments are used to ease the statistical analysis. One shortcoming of such arguments is that while, in general, as the contraction rate or "stability" of the system decreases, the signal to noise ratio increases and identification gets better due to stronger excitation, mixing-time based arguments capture the opposite dependence [51]. By adapting the martingale small-ball condition as in [51], we show this shortcoming can also be avoided for bilinear system identification.

3.1.2 Contributions

To summarize, we make the following contributions towards bilinear system identification: (i) For a bilinear system with state dimension n and input dimension p, the system dynamics involve $p+1$ matrices of size $n \times n$. We estimate these dynamics with an error rate $\mathcal{O}(\sqrt{n(p+1)/T})$. Our error rate is optimal in terms of the trajectory length T and the dimension of the unknown matrices. (ii) Recently, [29] asked an important question, *"Is learning without mixing possible in situations beyond generalized linear models?"* We provide a positive answer to this by extending martingale small-ball argument to bilinear systems. (iii) We correctly capture the dependence of random input and noise on the identification of marginally mean-square stable bilinear systems. Finally, we perform numerical experiments to support our theoretical results.

3.2 Preliminaries and Problem Setup

3.2.1 Bilinear Dynamical Systems

We consider the identification of bilinear dynamical systems which are governed by the state equation,

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}_0\boldsymbol{x}_t + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k\boldsymbol{x}_t + \boldsymbol{w}_{t+1}. \tag{3.2.1}$$

Here $\boldsymbol{x}_t \in \mathbb{R}^n$ is the state, $\boldsymbol{u}_t \in \mathbb{R}^p$ is the input, and $\boldsymbol{w}_t \in \mathbb{R}^n$ is the process noise at time t. $\{\boldsymbol{A}_k\}_{k=0}^{p} \in \mathbb{R}^{n\times n}$ are the state matrices which govern the dynamics of the system. Throughout, we assume that the input signal and noise are normally distributed.

Assumption 6 *We have* $\{\boldsymbol{u}_t\}_{t=0}^{\infty} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{u}}^2\boldsymbol{I}_p)$ *and* $\{\boldsymbol{w}_t\}_{t=1}^{\infty} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2\boldsymbol{I}_n)$*, where* $\sigma_{\boldsymbol{u}}, \sigma_{\boldsymbol{w}} > 0$.

Our primary goal in this chapter is to estimate the unknown state matrices $\{\boldsymbol{A}_k\}_{k=0}^{p}$ from finite samples obtained from a single trajectory of (3.2.1). For this purpose, we introduce the following concatenated matrix/vector notation,

$$\begin{aligned} \boldsymbol{A}_\star &:= \begin{bmatrix} \boldsymbol{A}_0 & \sigma_{\boldsymbol{u}}\boldsymbol{A}_1 & \cdots & \sigma_{\boldsymbol{u}}\boldsymbol{A}_p \end{bmatrix}, \\ \tilde{\boldsymbol{x}}_t &:= \begin{bmatrix} \boldsymbol{x}_t^\top & \sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u}_t[1]\boldsymbol{x}_t^\top & \cdots & \sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u}_t[p]\boldsymbol{x}_t^\top \end{bmatrix}^\top = \tilde{\boldsymbol{u}}_t \otimes \boldsymbol{x}_t, \end{aligned} \tag{3.2.2}$$

where $\boldsymbol{A}_\star \in \mathbb{R}^{n\times n(p+1)}$, $\tilde{\boldsymbol{x}}_t \in \mathbb{R}^{n(p+1)}$ and we define $\tilde{\boldsymbol{u}}_t := [1 \; \sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u}_t^\top]^\top$. With these definitions, the state update equation (3.2.1) can alternately be written as,

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}_\star \tilde{\boldsymbol{x}}_t + \boldsymbol{w}_{t+1}. \tag{3.2.3}$$

Suppose we have access to a single finite trajectory $\{(\boldsymbol{u}_t, \boldsymbol{x}_t, \boldsymbol{x}_{t+1})\}_{t=0}^T$ of the bilinear dynamical system (3.2.1). Then, to carry out finite sample identification of $\boldsymbol{A}_\star$ using the method of linear least squares, we define the following concatenated matrices,

$$\boldsymbol{Y}_T := \begin{bmatrix} \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_{T+1}^\top \end{bmatrix}, \; \tilde{\boldsymbol{X}}_T := \begin{bmatrix} \tilde{\boldsymbol{x}}_1^\top \\ \vdots \\ \tilde{\boldsymbol{x}}_T^\top \end{bmatrix}, \; \boldsymbol{W}_T := \begin{bmatrix} \boldsymbol{w}_2^\top \\ \vdots \\ \boldsymbol{w}_{T+1}^\top \end{bmatrix}. \tag{3.2.4}$$

To estimate the dynamics, we solve the following least-squares problem,

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}\in\mathbb{R}^{n\times n(p+1)}} \frac{1}{2T}\|\boldsymbol{Y}_T - \tilde{\boldsymbol{X}}_T\boldsymbol{A}^\top\|_F^2. \tag{3.2.5}$$

When the problem is over-determined, the solution to the least-squares problem (3.2.5) is given by $\hat{\boldsymbol{A}}^\top = (\tilde{\boldsymbol{X}}_T^\top\tilde{\boldsymbol{X}}_T)^{-1}\tilde{\boldsymbol{X}}_T^\top\boldsymbol{Y}_T$ and the associated estimation error is given by, $\hat{\boldsymbol{A}}^\top - \boldsymbol{A}_\star^\top = (\tilde{\boldsymbol{X}}_T^\top\tilde{\boldsymbol{X}}_T)^{-1}\tilde{\boldsymbol{X}}_T^\top\boldsymbol{W}_T$. This implies that the estimation error can be upper-bounded as follows,

$$\|\hat{\boldsymbol{A}} - \boldsymbol{A}_\star\| = \|(\tilde{\boldsymbol{X}}_T^\top\tilde{\boldsymbol{X}}_T)^{-1}\tilde{\boldsymbol{X}}_T^\top\boldsymbol{W}_T\| \le \|\tilde{\boldsymbol{X}}_T^\top\boldsymbol{W}_T\|/\lambda_{\min}\big(\tilde{\boldsymbol{X}}_T^\top\tilde{\boldsymbol{X}}_T\big). \tag{3.2.6}$$

To make the problem (3.2.5) well-conditioned, we also need a stability guarantee on the bilinear system (3.2.1). This will make sure that the design matrix $\tilde{\boldsymbol{X}}_T$ has smaller condition number to help better estimation. However, because of the randomness in $\boldsymbol{u}_t$, the dynamical behavior of the bilinear system (3.2.1) is also random. Therefore, it is common to define the stability of bilinear dynamical systems in the mean-square sense [103], which is the topic of our next subsection.

Algorithm 1 Bilinear System Identification

```
Input: Trajectory $\{(\boldsymbol{u}_t, \boldsymbol{x}_t, \boldsymbol{x}_{t+1})\}_{t=1}^T$ of bilinear dynamical system (3.2.1)
Estimate $\{\boldsymbol{A}_k\}_{k=0}^p$:
    Construct $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^T$ according to (3.2.2)
    Construct $\tilde{\boldsymbol{X}}_T, \boldsymbol{Y}_T$ according to (3.2.4)
    Find the least-squares estimator $\hat{\boldsymbol{A}} = \big((\tilde{\boldsymbol{X}}_T^\top \tilde{\boldsymbol{X}}_T)^{-1} \tilde{\boldsymbol{X}}_T^\top \boldsymbol{Y}_T\big)^\top$
    We have $\hat{\boldsymbol{A}}_0 = \hat{\boldsymbol{A}}[:, 1:n]$, and $\hat{\boldsymbol{A}}_k = \sigma_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{A}}[:, kn+1:(k+1)n]$ for $k = 1, \ldots, p$
Output: $\{\hat{\boldsymbol{A}}_k\}_{k=0}^m$
```

3.2.2 Mean-square stability of bilinear systems

Definition 20 ([103]) *The bilinear system in* (3.2.1) *is mean-square stable (MSS) if there exists $\boldsymbol{x}_\infty \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_\infty \in \mathbb{R}_+^{n \times n}$, such that for any initial state $\boldsymbol{x}_0$, as $t \to \infty$, we have*

$$\|\mathbb{E}[\boldsymbol{x}_t] - \boldsymbol{x}_\infty\|_{\ell_2} \to 0, \quad \|\,\mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top] - \boldsymbol{\Sigma}_\infty\| \to 0. \tag{3.2.7}$$

Here the expectation is over the input sequence $\{\boldsymbol{u}_t\}_{t=0}^\infty$, the noise process $\{\boldsymbol{w}_t\}_{t=1}^\infty$ and the initial state $\boldsymbol{x}_0$. In the noise free case $(\boldsymbol{w}_t = 0)$, we have $\boldsymbol{x}_\infty = 0$ and $\boldsymbol{\Sigma}_\infty = 0$.

The mean square stability of the bilinear system in (3.2.1) is related to the spectral radius of the following augmented state matrix [103],

$$\begin{aligned} \tilde{\boldsymbol{A}} &:= \boldsymbol{F} \otimes \boldsymbol{F} + \sum_{k=1}^p \sum_{\ell=1}^p \gamma_{k\ell} \boldsymbol{A}_\ell \otimes \boldsymbol{A}_k, \\ \text{where} \quad \boldsymbol{F} &:= \boldsymbol{A}_0 + \sum_{k=1}^p \mathbb{E}[\boldsymbol{u}_t[k]] \boldsymbol{A}_k, \\ \text{and} \quad \gamma_{k\ell} &:= \mathbb{E}[\boldsymbol{u}_t[k] \boldsymbol{u}_t[\ell]] - \mathbb{E}[\boldsymbol{u}_t[k]]\, \mathbb{E}[\boldsymbol{u}_t[\ell]]. \end{aligned} \tag{3.2.8}$$

Moreover, under Assumption 6, this further simplifies to,

$$\tilde{\boldsymbol{A}} = \boldsymbol{A}_0 \otimes \boldsymbol{A}_0 + \sigma_{\boldsymbol{u}}^2 \sum_{k=1}^p \boldsymbol{A}_k \otimes \boldsymbol{A}_k. \tag{3.2.9}$$

From Proposition 3 in [103], $\tilde{\boldsymbol{A}}$ can be viewed as a mapping from $\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]$ to $\mathbb{E}[\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top]$. Specifically, in the noise-free case, we have $\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top]) = \tilde{\boldsymbol{A}}\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top])$. Therefore, the bilinear system in (3.2.1) is MSS if and only if $\rho(\tilde{\boldsymbol{A}}) < 1$. This leads to our second assumption, which is stated as follows.

Assumption 7 *The bilinear system* (3.2.1) *is marginally mean-square stable, i.e.,* $\rho(\tilde{\boldsymbol{A}}) \leq 1$.

Using marginal mean-square stability, we can show that the second moment properties of the states $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ can be bounded as follows.

Lemma 21 *Consider the bilinear system in* (3.2.1). *Suppose Assumption* 6 *holds and let* $\tilde{\boldsymbol{A}}$ *be as in* (3.2.9). *Then, for all* $t \geq 0$, *we have*

$$\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]) = \tilde{\boldsymbol{A}}^t\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top]) + \sigma_{\boldsymbol{w}}^2\sum_{i=0}^{t-1}\tilde{\boldsymbol{A}}^i\mathbf{vec}(\boldsymbol{I}_n),$$

$$\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] \leq C_{\tilde{\boldsymbol{A}}}\rho(\tilde{\boldsymbol{A}})^t n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 n\sum_{i=0}^{t-1}C_{\tilde{\boldsymbol{A}}}\rho(\tilde{\boldsymbol{A}})^i.$$

Lemma 21 shows that if $\{\boldsymbol{w}_t\}_{t\geq1} = 0$ and $\rho(\tilde{\boldsymbol{A}}) < 1$, then starting from any initial state $\boldsymbol{x}_0$ with finite $\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2]$, the state $\boldsymbol{x}_t$ exponentially converges to 0. This implies, when $\rho(\tilde{\boldsymbol{A}}) < 1$, the process noise can assist learning by providing excitation and not allowing the trajectory to converge to 0.

3.3 Main Results

At the core of our analysis is showing that the random process $\{\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{u}}_t \otimes \boldsymbol{x}_t\}_{t\geq1}$ satisfies the martingale small-ball condition which is defined as follows.

Definition 22 (Martingale small-ball [22]) *Let* $\{\mathcal{F}_t\}_{t\geq1}$ *denotes a filtration and* $\{Z_t\}_{t\geq1}$ *be an* $\{\mathcal{F}_t\}_{t\geq1}$*-adapted random process taking values in* $\mathbb{R}$. *We say* $\{Z_t\}_{t\geq1}$ *satisfies the* (k, ν, q)*-*

block martingale small-ball (BMSB) condition if, for any $j \geq 0$, one has $\frac{1}{k}\sum_{i=1}^{k}\mathbb{P}\left(|Z_{j+i}| \geq \nu \mid \mathcal{F}_j\right) \geq q$ almost surely. Given a process $\{\boldsymbol{x}_t\}_{t\geq 1}$ taking values in $\mathbb{R}^d$, we say it satisfies the $(k, \boldsymbol{\Gamma}_{sb}, q)$-BMSB condition for $\boldsymbol{\Gamma}_{sb} \succ 0$ if, for any fixed $\boldsymbol{v} \in \mathcal{S}^{d-1}$, the process $Z_t = \langle \boldsymbol{v}, \boldsymbol{x}_t \rangle$ satisfies $(k, \sqrt{\boldsymbol{v}^\top \boldsymbol{\Gamma}_{sb} \boldsymbol{v}}, q)$-BMSB.

To show that $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 1}$ satisfies BMSB condition, let $\mathcal{F}_t := \sigma(\boldsymbol{x}_0, \dots, \boldsymbol{x}_t, \boldsymbol{u}_0, \dots, \boldsymbol{u}_t, \boldsymbol{w}_1, \dots, \boldsymbol{w}_t)$ denotes the filtration generated by the states, the input and the noise processes when $t \geq 1$. Furthermore, let $\mathcal{F}_0 := \sigma(\boldsymbol{x}_0, \boldsymbol{u}_0)$. Then, $\boldsymbol{x}_t, \boldsymbol{u}_t$ and $\boldsymbol{w}_t$ become $\mathcal{F}_t$-measurable and, recalling (3.2.2), $\tilde{\boldsymbol{x}}_t$ is also $\mathcal{F}_t$-measurable.

Theorem 23 (BMSB condition for $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 1}$) *Consider the bilinear dynamical system in (3.2.1). Suppose Assumption 6 holds and let $\tilde{\boldsymbol{x}}_t$ be as in (3.2.2). Then, the process $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 1}$ satisfies the $(k, c^2\sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_{n(p+1)}, q)$-martingale small-ball condition, with the constants $k = 1, c = 1/2$ and $q = 9/320$.*

The theorem above uses martingale small-ball with $k = 1$. We remark that using $k > 1$ is expected to help capture the role of additional excitation terms in the BMSB lower bound, specifically, the dependence on $\tilde{\boldsymbol{A}}$. However, this requires bounding higher order moments that involve cross-products of the input signal and noise terms and is left as future research.

We are now ready to state our main result to estimate the dynamics $\{\boldsymbol{A}_k\}_{k=0}^{p}$ from a single finite trajectory $\{(\boldsymbol{u}_t, \boldsymbol{x}_t, \boldsymbol{x}_{t+1})\}_{t=0}^{T}$ of the bilinear dynamical system (3.2.1).

Theorem 24 (Bilinear system identification) *Fix $\delta \in (0,1)$ and suppose we are given a single trajectory $\{(\boldsymbol{u}_t, \boldsymbol{x}_t, \boldsymbol{x}_{t+1})\}_{t=0}^{T}$ of the bilinear dynamical system in (3.2.1). Suppose*

Assumptions 6 and 7 hold, and the trajectory length T satisfies the following lower bound,

$$\begin{aligned} T &\gtrsim n(p+1) + \log(12\bar{\Gamma}/(\sigma_{\boldsymbol{w}}^2\delta)) + \log(3/\delta), \\ \textit{where, } \bar{\Gamma} &\coloneqq C_{\tilde{\boldsymbol{A}}}(n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 nT)(p+1). \end{aligned} \tag{3.3.1}$$

Then, with probability at least $1-\delta$, Algorithm 1 ensures

$$\max\left\{\|\hat{\boldsymbol{A}}_0 - \boldsymbol{A}_0\|, \{\sigma_u\|\hat{\boldsymbol{A}}_k - \boldsymbol{A}_k\|\}_{k=1}^p\right\} \lesssim \sqrt{\frac{n(p+1) + \log(12\bar{\Gamma}/(\sigma_{\boldsymbol{w}}^2\delta)) + \log(3/\delta)}{T}}. \tag{3.3.2}$$

In words, (3.3.2) ensures the estimation of all state matrices as soon as the sample size exceeds the effective degrees of freedom $n(p+1)$. The estimation of $\{\boldsymbol{A}_k\}_{k=1}^p$ naturally depends on the input strength, as $\boldsymbol{u}_t[k]$ is a multiplier of $\boldsymbol{A}_k$ in (3.2.1). Please note that Theorem 24 only holds under the condition that $\rho(\tilde{\boldsymbol{A}}) \leq 1$. This implies that we cannot increase $\sigma_{\boldsymbol{u}}$ arbitrarily to obtain better estimation. This is because, under Assumption 6, we have $\tilde{\boldsymbol{A}} = \boldsymbol{A}_0 \otimes \boldsymbol{A}_0 + \sigma_{\boldsymbol{u}}^2 \sum_{k=1}^p \boldsymbol{A}_k \otimes \boldsymbol{A}_k$. Therefore, the largest possible $\sigma_{\boldsymbol{u}}$ is given by $\sigma_{\boldsymbol{u},\max} \coloneqq \max\{\sigma_{\boldsymbol{u}} > 0 : \rho(\boldsymbol{A}_0 \otimes \boldsymbol{A}_0 + \sigma_{\boldsymbol{u}}^2 \sum_{k=1}^p \boldsymbol{A}_k \otimes \boldsymbol{A}_k) \leq 1\}$.

Our estimation error is independent of the noise variance $\sigma_{\boldsymbol{w}}^2$. This is because the size of the noise variance $\sigma_{\boldsymbol{w}}^2$ directly influences the size of the states leading to a cancellation in the signal-to-noise ratio. On the other hand the size of the input variance $\sigma_{\boldsymbol{u}}^2$ indirectly influences the size of the states by influencing the spectral radius of $\tilde{\boldsymbol{A}}$. As a result, increasing $\sigma_{\boldsymbol{u}}^2$ helps learning. These observations are further strengthened by numerical experiments in Section 6.2.

Unlike the existing results [29, 33, 71, 101] on finite time identification of nonlinear dynamical systems, the error bounds in Theorem 24 do not degrade with increasing instability. We emphasize that, our result guarantees identification even in the case of non-mixing bilinear systems. This shows learning without mixing is possible beyond generalized linear models.

3.4 Proofs of the Main Results

3.4.1 Proof of Theorem 23

Proof. In this subsection, we will show that the random process $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 1}$ satisfies $(1, c^2\sigma_{\boldsymbol{w}}^2\boldsymbol{I}_{n(p+1)}, q)$-BMSB condition, for some constants $c, q > 0$. For this purpose, we need to show that, for any fixed $\boldsymbol{v} \in \mathcal{S}^{n(p+1)-1}$, the random process $\{Z_t\}_{t\geq 1} \coloneqq \{\langle \boldsymbol{v}, \tilde{\boldsymbol{x}}_t\rangle\}_{t\geq 1}$ satisfies $(1, c\sigma_{\boldsymbol{w}}\|\boldsymbol{v}\|_{\ell_2}, q)$-BMSB condition, that is, for any $j \geq 0$, we need to show that $\mathbb{P}(|Z_{j+1}| \geq c\sigma_{\boldsymbol{w}}\|\boldsymbol{v}\|_{\ell_2} \mid \mathcal{F}_j) \geq q$ almost surely. To proceed, for any $j \geq 0$, consider the concatenated state vector,

$$\tilde{\boldsymbol{x}}_{j+1} = \begin{bmatrix} \boldsymbol{x}_{j+1} \\ \sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u}_{j+1}[1]\boldsymbol{x}_{j+1} \\ \vdots \\ \sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u}_{j+1}[p]\boldsymbol{x}_{j+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1} \\ \bar{\boldsymbol{u}}_{j+1}[1](\boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}) \\ \vdots \\ \bar{\boldsymbol{u}}_{j+1}[p](\boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}) \end{bmatrix}, \tag{3.4.1}$$

where we set $\bar{\boldsymbol{u}}_t = \sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u}_t$, so that $\{\bar{\boldsymbol{u}}_t\}_{t=0}^\infty \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$. To proceed, using (3.4.1), we have that

$$Z_{j+1} \coloneqq \langle \boldsymbol{v}, \tilde{\boldsymbol{x}}_{j+1}\rangle = \langle \boldsymbol{v}_0 + \bar{\boldsymbol{u}}_{j+1}[1]\boldsymbol{v}_1 + \cdots + \bar{\boldsymbol{u}}_{j+1}[p]\boldsymbol{v}_p, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle, \tag{3.4.2}$$

where we set $\boldsymbol{v} = [\boldsymbol{v}_0^\top \ \boldsymbol{v}_1^\top \ \cdots \ \boldsymbol{v}_p^\top]^\top$ such that $\boldsymbol{v}_i \coloneqq \boldsymbol{v}[ni+1 : n(i+1)]$. Next, we concatenate $\boldsymbol{v}_i$'s to form the matrix,

$$\boldsymbol{V} \coloneqq [\boldsymbol{v}_1 \cdots \boldsymbol{v}_p] \in \mathbb{R}^{n\times p}. \tag{3.4.3}$$

Combining this with (3.4.2), we have that $Z_{j+1} = \langle \boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle$. Therefore, we are interested in lower bounding the following probability,

$$= \mathbb{P}\big(|\langle \boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle| \geq c\sigma_{\boldsymbol{w}}\|\boldsymbol{v}\|_{\ell_2} \mid \mathcal{F}_j\big). \tag{3.4.4}$$

To lower bound the probability in (3.4.4), we define the following three events,

$$
\begin{aligned}
\mathcal{E}_z &:= \left\{ |\langle \boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle| \geq c\sigma_{\boldsymbol{w}}\|\boldsymbol{v}\|_{\ell_2} \;\middle|\; \mathcal{F}_j \right\}, \\
\mathcal{E}_w &:= \left\{ |\langle \boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle| \geq \sigma_{\boldsymbol{w}}\|\boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2} \;\middle|\; \mathcal{F}_j \right\}, \qquad (3.4.5) \\
\mathcal{E}_u &:= \left\{ \|\boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2} \geq c\|\boldsymbol{v}\|_{\ell_2} \;\middle|\; \mathcal{F}_j \right\}.
\end{aligned}
$$

Note that, $\mathcal{E}_w \cap \mathcal{E}_u \subset \mathcal{E}_z$. This implies that, we have, $\mathbb{P}(\mathcal{E}_z) \geq \mathbb{P}(\mathcal{E}_w \cap \mathcal{E}_u) = \mathbb{P}(\mathcal{E}_w \mid \mathcal{E}_u)\,\mathbb{P}(\mathcal{E}_u)$. Therefore, to lower bound the probability of the event $\mathcal{E}_z$, it suffices to lower bound the probability of these two events: $\mathcal{E}_w \mid \mathcal{E}_u$ and $\mathcal{E}_u$.

(a) $\mathbb{P}(\mathcal{E}_w \mid \mathcal{E}_u)$: Given that, we have $\boldsymbol{w}_{j+1} \sim \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$, for any fixed vector $\boldsymbol{q} \in \mathbb{R}^n$, $\langle \boldsymbol{q}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle \big| \mathcal{F}_j \sim \mathcal{N}\big(\langle \boldsymbol{q}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j\rangle, \sigma_{\boldsymbol{w}}^2\|\boldsymbol{q}\|_{\ell_2}^2\big)$. Therefore, integrating the probability density function of a standard Gaussian random variable, it can be shown that,

$$
\mathbb{P}\big(|\langle \boldsymbol{q}, \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}\rangle| \geq \sigma_{\boldsymbol{w}}\|\boldsymbol{q}\|_{\ell_2} \;\big|\; \mathcal{F}_j\big) \geq 3/10. \qquad (3.4.6)
$$

We obtain the above result by integrating the probability density function of a Gaussian random variable as follows,

$$
\begin{aligned}
\forall \alpha \in \mathbb{R} \quad \mathbb{P}_{Z\sim\mathcal{N}(0,\sigma^2)}(|\alpha + Z| \geq \sigma) &\geq \mathbb{P}_{Z\sim\mathcal{N}(0,\sigma^2)}(|Z| \geq \sigma) = \mathbb{P}_{Z'\sim\mathcal{N}(0,1)}(|Z'| \geq 1), \\
&= 1 - \mathbb{P}_{Z'\sim\mathcal{N}(0,1)}(|Z'| \leq 1) = 1 - 2\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-z'^2/2} dz', \\
&\geq 1 - 2(7/20) = 3/10. \qquad (3.4.7)
\end{aligned}
$$

To proceed, setting $\boldsymbol{q} = \boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}$ and $\boldsymbol{p} = \boldsymbol{A}_\star\tilde{\boldsymbol{x}}_j + \boldsymbol{w}_{j+1}$, let $f_{\boldsymbol{Q}}(\boldsymbol{q})$, $f_{\boldsymbol{P}}(\boldsymbol{p})$ denote the probability density functions of the random vectors $\boldsymbol{q} \mid \mathcal{F}_j$ and $\boldsymbol{p} \mid \mathcal{F}_j$, respectively, under the event $\mathcal{E}_u$. Observe that $\boldsymbol{q} \mid \mathcal{F}_j$ and $\boldsymbol{p} \mid \mathcal{F}_j$ are independent under $\mathcal{E}_u$. Therefore, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_w \mid \mathcal{E}_u) &= \int \int f_{\boldsymbol{Q}}(\boldsymbol{q}) f_{\boldsymbol{P}}(\boldsymbol{p}) \mathbf{1}_{(|\langle \boldsymbol{q},\boldsymbol{p}\rangle| \geq \sigma_{\boldsymbol{w}} \|\boldsymbol{q}\|_{\ell_2})} d\boldsymbol{p} d\boldsymbol{q}, \\ &= \int f_{\boldsymbol{Q}}(\boldsymbol{q}) \underbrace{\int f_{\boldsymbol{P}}(\boldsymbol{p}) \mathbf{1}_{(|\langle \boldsymbol{q},\boldsymbol{p}\rangle| \geq \sigma_{\boldsymbol{w}} \|\boldsymbol{q}\|_{\ell_2})} d\boldsymbol{p}}_{\mathbb{P}(|\langle \boldsymbol{q},\boldsymbol{p}\rangle| \geq \sigma_{\boldsymbol{w}} \|\boldsymbol{q}\|_{\ell_2}) \text{ for fixed } \boldsymbol{q} \in \mathbb{R}^n} d\boldsymbol{q} \overset{\text{(i)}}{\geq} (3/10) \int f_{\boldsymbol{Q}}(\boldsymbol{q}) d\boldsymbol{q} = 3/10, \end{aligned} \quad (3.4.8)$$

where $\mathbf{1}_{(\cdot)}$ denotes the indicator function, and we obtain (i) from (4.5.4). Hence, we showed that $\mathbb{P}(\mathcal{E}_w \mid \mathcal{E}_u) \geq 3/10$.

(b) $\mathbb{P}(\mathcal{E}_u)$*:* Next, to lower bound the probability of the event $\mathcal{E}_u$, we consider the following,

$$\begin{aligned}\|\boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 &= \|\boldsymbol{v}_0\|_{\ell_2}^2 + \|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 + 2\langle \boldsymbol{v}_0, \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\rangle, \\ &= \|\boldsymbol{v}_0\|_{\ell_2}^2 + \|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 + 2\left\langle \boldsymbol{V}^\top \boldsymbol{v}_0, \bar{\boldsymbol{u}}_{j+1}\right\rangle. \end{aligned} \quad (3.4.9)$$

Let $\mathcal{E}_\Xi = \{\|\boldsymbol{v}_0\|_{\ell_2}^2 + \|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq \Xi\}$ and $\mathcal{E}_+ = \{\langle \boldsymbol{V}^\top \boldsymbol{v}_0, \bar{\boldsymbol{u}}_{j+1}\rangle \geq 0\}$. Since $\bar{\boldsymbol{u}}_{j+1}$ is rotationally invariant and $\boldsymbol{V}^\top \boldsymbol{v}_0$ is a fixed vector $\mathbb{P}(\mathcal{E}_+) = 1/2$. More generally, $\mathcal{E}_\Xi$ and $\mathcal{E}_+$ are independent again due to rotational invariance (sign and magnitude of $\bar{\boldsymbol{u}}_{j+1}$ are independent). Combining this with (3.4.9), for any Ξ, we have

$$\mathbb{P}\left(\|\boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq \Xi\right) \geq \mathbb{P}(\mathcal{E}_\Xi \cap \mathcal{E}_+) = 0.5\,\mathbb{P}(\|\boldsymbol{v}_0\|_{\ell_2}^2 + \|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq \Xi). \quad (3.4.10)$$

Therefore, to lower bound the probability of event $\mathcal{E}_u$, it suffices to lower bound the probability of the event $\{\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq c\|\boldsymbol{V}\|_F^2\}$, for some constant $c > 0$. Let $\boldsymbol{V}$ have singular value decomposition $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{R}^\top$ with $\|\boldsymbol{V}\|_F^2 = \|\boldsymbol{\Sigma}\|_F^2 = \sum_{i=1}^p \sigma_i^2$. Furthermore, since $\bar{\boldsymbol{u}}_{j+1} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ and $\boldsymbol{Q}, \boldsymbol{R}$ are orthogonal matrices, we have $\boldsymbol{g} := \boldsymbol{R}^\top \bar{\boldsymbol{u}}_{j+1} \sim \mathcal{N}(0, \boldsymbol{I}_p)$. Therefore, we have

$$\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 = \|\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{R}^\top \bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 = \|\boldsymbol{\Sigma}\boldsymbol{R}^\top \bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 = \|\boldsymbol{\Sigma}\boldsymbol{g}\|_{\ell_2}^2 = \sum_{i=1}^p \sigma_i^2 \boldsymbol{g}[i]^2. \quad (3.4.11)$$

This further implies,

$$\mathbb{E}[\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2] = \mathbb{E}\big[\sum_{i=1}^p \sigma_i^2 \boldsymbol{g}[i]^2\big] = \sum_{i=1}^p \sigma_i^2\, \mathbb{E}[\boldsymbol{g}[i]^2] = \sum_{i=1}^p \sigma_i^2 = \|\boldsymbol{V}\|_F^2. \tag{3.4.12}$$

Similarly, we also have,

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^4] &= \mathbb{E}\big[(\sum_{i=1}^p \sigma_i^2 \boldsymbol{g}[i]^2)^2\big] = \mathbb{E}\big[\sum_{i=1}^p \sigma_i^4 \boldsymbol{g}[i]^4 + \sum_{i=1}^p \sum_{\substack{j=1\\ j\neq i}}^p \sigma_i^2\sigma_j^2 \boldsymbol{g}[i]^2\boldsymbol{g}[j]^2\big],\\ &= \sum_{i=1}^p \sigma_i^4\, \mathbb{E}[\boldsymbol{g}[i]^4] + \sum_{i=1}^p \sum_{\substack{j=1\\ j\neq i}}^p \sigma_i^2\sigma_j^2\, \mathbb{E}[\boldsymbol{g}[i]^2\boldsymbol{g}[j]^2] \overset{\text{(i)}}{=} 3\sum_{i=1}^p \sigma_i^4 + \sum_{i=1}^p \sum_{\substack{j=1\\ j\neq i}}^p \sigma_i^2\sigma_j^2, \qquad (3.4.13)\\ &\leq 3\big(\sum_{i=1}^p \sigma_i^2\big)^2 = 3\|\boldsymbol{V}\|_F^4,\end{aligned}$$

where we get (i) from $\mathbb{E}[\boldsymbol{g}[i]^4] = 3$ and the independence of $\boldsymbol{g}[i]$ and $\boldsymbol{g}[j]$ for all $i \neq j$. Combining (3.4.12) and (3.4.13) with the Paley-Zygmund inequality, for a fixed $\gamma \in (0,1)$, we have

$$\begin{aligned}&\mathbb{P}\big(\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq \gamma\, \mathbb{E}[\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2]\big) \geq (1-\gamma)^2 \frac{\mathbb{E}[\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2]^2}{\mathbb{E}[\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^4]},\\ &\implies \mathbb{P}\big(\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq \gamma \|\boldsymbol{V}\|_F^2\big) \geq (1-\gamma)^2 \frac{1}{3}, \qquad (3.4.14)\\ &\implies \mathbb{P}\big(\|\boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq (1/4)\|\boldsymbol{V}\|_F^2\big) \geq 3/16,\end{aligned}$$

where we obtain the last line by setting $\gamma = 1/4$. Finally, combining (3.4.10) and (3.4.14), we have

$$\mathbb{P}\big(\|\boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2}^2 \geq \|\boldsymbol{v}_0\|_{\ell_2}^2 + (1/4)\sum_{i=1}^p \|\boldsymbol{v}_i\|_{\ell_2}^2\big) \geq (1/2)(3/16) = 3/32. \tag{3.4.15}$$

Combining this with $\|\boldsymbol{v}\|_{\ell_2}^2 = \sum_{i=0}^p \|\boldsymbol{v}_i\|_{\ell_2}^2$, we obtain

$$\mathbb{P}\big(\|\boldsymbol{v}_0 + \boldsymbol{V}\bar{\boldsymbol{u}}_{j+1}\|_{\ell_2} \geq (1/2)\|\boldsymbol{v}\|_{\ell_2}\big) \geq 3/32. \tag{3.4.16}$$

Hence, setting $c = 1/2$, we found that $\mathbb{P}(\mathcal{E}_u) \geq 3/32$. Putting all together, we have $\mathbb{P}(\mathcal{E}_z) \geq \mathbb{P}(\mathcal{E}_w \mid \mathcal{E}_u)\mathbb{P}(\mathcal{E}_u) \geq 9/320$. This verifies our claim that the process $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 1}$ satisfies $(1, c^2\sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_{n(p+1)}, q)$-BMSB condition, with the constants $c = 1/2$ and $q = 9/320$. ■

3.4.2 Proof of Theorem 24

Proof. For the sake of completeness, before we present the proof of Theorem 24, we present a meta result from [22] which will be used to prove Theorem 24.

Theorem 25 (Meta-theorem [22]) *Fix $\delta \in (0,1)$, $T \in \mathbb{N}$ and $0 \prec \mathbf{\Gamma}_{sb} \prec \bar{\Gamma}$. Then if $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{t=1}^T \in (\mathbb{R}^d \times \mathbb{R}^n)^T$ is a random sequence such that (a) $\boldsymbol{y}_t = \boldsymbol{A}_\star \boldsymbol{x}_t + \boldsymbol{w}_t$, where $\boldsymbol{w}_t \mid \mathcal{F}_{t-1}$ is $\sigma_{\boldsymbol{w}}^2$-subgaussian and mean zero, (b) $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ satisfy the $(k, \mathbf{\Gamma}_{sb}, q)$-small ball condition, and (c) such that $\mathbb{P}\left(\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\top \npreceq T\bar{\mathbf{\Gamma}}\right) \le \delta$. Then if*

$$T \ge \frac{10k}{q^2}\left(\log(1/\delta) + 2d\log(10/q) + \log(\det(\bar{\mathbf{\Gamma}}\mathbf{\Gamma}_{sb}^{-1}))\right),$$

we have

$$\mathbb{P}\left(\|\hat{\boldsymbol{A}}(T) - \boldsymbol{A}_\star\| \ge \frac{90\sigma_{\boldsymbol{w}}}{q}\sqrt{\frac{n + d\log(10/q) + \log(\det(\bar{\mathbf{\Gamma}}\mathbf{\Gamma}_{sb}^{-1})) + \log(1/\delta)}{T\lambda_{\min}(\mathbf{\Gamma}_{sb})}}\right) \le 3\delta.$$

Our proof strategy is to verify that the conditions (a), (b), and (c) of Theorem 25 hold for the bilinear dynamical system in (3.2.1) and then apply Theorem 25 to estimate $\boldsymbol{A}_\star$.

(a) Sub-gaussian noise: Following the re-parameterization in (3.2.3), we have $\boldsymbol{x}_{t+1} = \boldsymbol{A}_\star \tilde{\boldsymbol{x}}_t + \boldsymbol{w}_{t+1}$. Moreover, under Assumption 6, the process noise $\boldsymbol{w}_t | \mathcal{F}_{t-1}$ is $\sigma_{\boldsymbol{w}}^2$-subgaussian and mean zero.

(b) BMSB condition: Theorem 23 proves that the random process $\{\tilde{\boldsymbol{x}}_t\}_{t\ge1}$ satisfies $(1, c^2\sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_{n(p+1)}, q)$-BMSB condition, with the constants $c = 1/2$ and $q = 9/320$.

(c) State correlation bound: Recall the definition of $\tilde{\boldsymbol{u}}_t$, $\tilde{\boldsymbol{x}}_t$ from (3.2.2) and $\tilde{\boldsymbol{X}}_T$ from (3.2.4). We have

$$\begin{aligned}\|\tilde{\boldsymbol{X}}_T^\top \tilde{\boldsymbol{X}}_T\| &= \|\sum_{t=1}^T (\tilde{\boldsymbol{u}}_t \otimes \boldsymbol{x}_t)(\tilde{\boldsymbol{u}}_t^\top \otimes \boldsymbol{x}_t^\top)\| = \|\sum_{t=1}^T (\tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top \otimes \boldsymbol{x}_t \boldsymbol{x}_t^\top)\|, \\ &\overset{\text{(i)}}{\le} \sum_{t=1}^T \|\tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top\| \|\boldsymbol{x}_t \boldsymbol{x}_t^\top\| \le \sum_{t=1}^T \|\tilde{\boldsymbol{u}}_t\|_{\ell_2}^2 \|\boldsymbol{x}_t\|_{\ell_2}^2,\end{aligned} \tag{3.4.17}$$

where we obtain (i) from the triangle inequality and the fact that $\|\boldsymbol{C} \otimes \boldsymbol{D}\| \le \|\boldsymbol{C}\|\|\boldsymbol{D}\|$. This further implies,

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\boldsymbol{X}}_T^\top \tilde{\boldsymbol{X}}_T\|\right] \le \sum_{t=1}^{T} \mathbb{E}[\|\tilde{\boldsymbol{u}}_t\|_{\ell_2}^2 \|\boldsymbol{x}_t\|_{\ell_2}^2] &\overset{\text{(ii)}}{\le} \sum_{t=1}^{T} (p+1) C_{\tilde{\boldsymbol{A}}}(n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 nt), \\ &\le T C_{\tilde{\boldsymbol{A}}}(n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 nT)(p+1),\end{aligned} \tag{3.4.18}$$

where we obtain (ii) from the independence of $\boldsymbol{u}_t$ and $\boldsymbol{x}_t$. Moreover, we have $\mathbb{E}[\|\tilde{\boldsymbol{u}}_t\|_{\ell_2}^2] = 1 + \sigma_{\boldsymbol{u}}^{-2}\,\mathbb{E}[\|\boldsymbol{u}_t\|_{\ell_2}^2] = 1 + p$, and we use Lemma 21 along with Assumption 7 to bound $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]$. Hence, setting

$$\bar{\Gamma} := C_{\tilde{\boldsymbol{A}}}(n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 nT)(m+1), \tag{3.4.19}$$

we have, $\mathbb{E}[\|\sum_{t=1}^{T} \tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_t^\top\|] = \mathbb{E}[\|\tilde{\boldsymbol{X}}_T^\top \tilde{\boldsymbol{X}}_T\|] \le T\bar{\Gamma}$. Next, we use Markov inequality to show that

$$\begin{aligned}\mathbb{P}\Big(\sum_{t=1}^{T} \tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_t^\top \npreceq (T\bar{\Gamma}/\delta)\boldsymbol{I}_{n(p+1)}\Big) &= \mathbb{P}\Big(\lambda_{\max}\big(\sum_{t=1}^{T} \tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_t^\top\big) \ge T\bar{\Gamma}/\delta\Big), \\ &\le \mathbb{E}\Big[\lambda_{\max}\big(\sum_{t=1}^{T} \tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_t^\top\big)\Big]\delta/(T\bar{\Gamma}) \le \delta.\end{aligned} \tag{3.4.20}$$

We are now ready to use Theorem 25 from [22] to obtain our final result.

(d) Finalizing the proof: In Theorem 25, we set $\bar{\boldsymbol{\Gamma}} = (1/\delta)C_{\tilde{\boldsymbol{A}}}(n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 nT)(p+1)\boldsymbol{I}_{n(p+1)}$, $\boldsymbol{\Gamma}_{sb} = (1/4)\sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_{n(p+1)}$, $k = 1$, $q = 9/320$, and $d = n(p+1)$. This gives,

$$\bar{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_{sb}^{-1} = 4\bar{\Gamma}/(\sigma_{\boldsymbol{w}}^2\delta)\boldsymbol{I}_{n(p+1)} = (4/\delta)C_{\tilde{\boldsymbol{A}}}(n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2]/\sigma_{\boldsymbol{w}}^2 + nT)(p+1)\boldsymbol{I}_{n(p+1)}. \tag{3.4.21}$$

Using this in Theorem 25, and replacing δ with $\delta/3$, when the trajectory length T satisfies,

$$T \gtrsim n(p+1) + \log(12\bar{\Gamma}/(\sigma_{\boldsymbol{w}}^2\delta)) + \log(3/\delta), \tag{3.4.22}$$

we have

$$\mathbb{P}\left(\|\hat{\boldsymbol{A}} - \boldsymbol{A}_\star\| \lesssim \sqrt{\frac{n(p+1) + \log(12\bar{\Gamma}/(\sigma_{\boldsymbol{w}}^2\delta)) + \log(3/\delta)}{T}}\right) \ge 1 - \delta. \tag{3.4.23}$$

Finally, using the fact that the spectral norm of a sub-matrix is upper bounded by that of the original matrix establishes the statement of the theorem. This completes the proof. ■

3.4.3 Proof of Lemma 21

Proof. To begin, consider the following

$$
\begin{aligned}
&\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^{\top}]) \\
&\quad = \mathbf{vec}\Bigg(\mathbb{E}\Bigg[\Big((\boldsymbol{A}_0 + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k)\boldsymbol{x}_t + \boldsymbol{w}_{t+1} \Big) \Big((\boldsymbol{A}_0 + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k)\boldsymbol{x}_t + \boldsymbol{w}_{t+1} \Big)^{\top} \Bigg] \Bigg) \\
&\quad \overset{\text{(i)}}{=} \mathbf{vec}\Bigg(\mathbb{E}\Bigg[(\boldsymbol{A}_0 + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k)\boldsymbol{x}_t\boldsymbol{x}_t^{\top} (\boldsymbol{A}_0 + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k)^{\top} \Bigg] + \mathbb{E}[\boldsymbol{w}_{t+1}\boldsymbol{w}_{t+1}^{\top}] \Bigg), \\
&\quad \overset{\text{(ii)}}{=} \mathbb{E}\Bigg[(\boldsymbol{A}_0 + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k) \otimes (\boldsymbol{A}_0 + \sum_{k=1}^{p} \boldsymbol{u}_t[k]\boldsymbol{A}_k)\mathbf{vec}(\boldsymbol{x}_t\boldsymbol{x}_t^{T}) \Bigg] + \mathbf{vec}(\sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n), \\
&\quad \overset{\text{(iii)}}{=} \Big(\boldsymbol{A}_0 \otimes \boldsymbol{A}_0 + \sigma_{\boldsymbol{u}}^2 \sum_{k=1}^{p} \boldsymbol{A}_k \otimes \boldsymbol{A}_k \Big) \mathbf{vec}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^{\top}]) + \sigma_{\boldsymbol{w}}^2 \mathbf{vec}(\boldsymbol{I}_n), \\
&\quad = \tilde{\boldsymbol{A}}\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^{\top}]) + \sigma_{\boldsymbol{w}}^2 \mathbf{vec}(\boldsymbol{I}_n),
\end{aligned}
\tag{3.4.24}
$$

where we get (i) from the independence of $\boldsymbol{u}_t$ and $\boldsymbol{x}_t$, (ii) from the linearity of $\mathbf{vec}(\cdot)$ operator, and (iii) from Assumption 6. Here we use the definition of $\tilde{\boldsymbol{A}}$ from (3.2.9). Repeating the recursion in (3.4.24) till $t = 0$, we have

$$
\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^{\top}]) = \tilde{\boldsymbol{A}}^t \mathbf{vec}(\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^{\top}]) + \sigma_{\boldsymbol{w}}^2 \sum_{i=0}^{t-1} \tilde{\boldsymbol{A}}^i \mathbf{vec}(\boldsymbol{I}_n). \tag{3.4.25}
$$

Next, using (3.4.25), we bound the expected squared Euclidean norm of the states $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ as follows,

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] = \mathbb{E}[\boldsymbol{x}_t^\top \boldsymbol{x}_t] = \mathbb{E}[\mathrm{trace}(\boldsymbol{x}_t\boldsymbol{x}_t^\top)] = \mathrm{trace}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]) &= \sum_{j=1}^{n} \lambda_j(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]), \\
&\le \sqrt{n\sum_{j=1}^{n}\lambda_j^2(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top])} = \sqrt{n}\|\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]\|_F = \sqrt{n}\|\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top])\|_{\ell_2}, \\
&= \sqrt{n}\|\tilde{\boldsymbol{A}}^t\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top]) + \sigma_{\boldsymbol{w}}^2\sum_{i=0}^{t-1}\tilde{\boldsymbol{A}}^i\mathbf{vec}(\boldsymbol{I}_n)\|_{\ell_2}, \\
&\le \sqrt{n}\|\tilde{\boldsymbol{A}}^t\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top])\|_{\ell_2} + \sqrt{n}\|\sigma_{\boldsymbol{w}}^2\sum_{i=0}^{t-1}\tilde{\boldsymbol{A}}^i\mathbf{vec}(\boldsymbol{I}_n)\|_{\ell_2}, \\
&\le \sqrt{n}\|\tilde{\boldsymbol{A}}^t\|\|\mathbf{vec}(\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top])\|_{\ell_2} + \sigma_{\boldsymbol{w}}^2\sqrt{n}\sum_{i=0}^{t-1}\|\tilde{\boldsymbol{A}}^i\|\|\mathbf{vec}(\boldsymbol{I}_n)\|_{\ell_2}, \\
&\le C_{\tilde{\boldsymbol{A}}}\rho(\tilde{\boldsymbol{A}})^t\sqrt{n}\|\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top]\|_F + \sigma_{\boldsymbol{w}}^2 n\sum_{i=0}^{t-1}C_{\tilde{\boldsymbol{A}}}\rho(\tilde{\boldsymbol{A}})^i \\
&\le C_{\tilde{\boldsymbol{A}}}\rho(\tilde{\boldsymbol{A}})^t n\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{w}}^2 n\sum_{i=0}^{t-1}C_{\tilde{\boldsymbol{A}}}\rho(\tilde{\boldsymbol{A}})^i,
\end{aligned}$$

where $\lambda_j(\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top])$ denotes the j-th eigenvalue of $\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]$, for $j = 1, \dots, n$. This completes the proof. ■

Chapter 4

Learning Markov Jump Systems

4.1 Introduction

A canonical problem at the intersection of machine learning and control is that of adaptive control of an unknown dynamical system. An intelligent autonomous system is likely to encounter such a task; from an observation of the inputs and outputs, it needs to both learn and effectively control the dynamics. A commonly used control paradigm is the Linear Quadratic Regulator (LQR), which is theoretically well understood when system dynamics are linear and known. LQR also provides an interesting benchmark, when system dynamics are unknown, for reinforcement learning (RL) with continuous state and action spaces and for adaptive control [104–109].

A generalization of linear dynamical systems called Markov jump linear systems (MJSs) models dynamics that switch between multiple linear systems, called modes, according to an underlying finite Markov chain. MJS allows for modeling a richer set of problems where the underlying dynamics can abruptly change over time. One can, similarly, generalize the

LQR paradigm to MJS by using mode-dependent cost matrices, which allow different control goals under different modes. For instance, a Mars rover optimally exploring an unknown heterogeneous terrain, optimal solar power generation on a cloudy day, or controlling investments in financial markets may be modeled as MJS-LQR problems with unknown system dynamics [110–114].

While the MJS-LQR problem is well understood when one has perfect knowledge of the system dynamics [115, 116], in practice, such knowledge is not always possible, and one may have to resort to adaptive control. Earlier works have aimed at analyzing the asymptotic properties (i.e., stability) of adaptive controllers for unknown MJSs both in continuous-time [117] and discrete-time [118] settings. However, despite the practical importance of MJSs, non-asymptotic sample complexity results and regret analysis for MJSs are lacking. When the Markovian modes switch in an i.i.d. fashion, and the Markov matrix is the only unknown, recent works study data-driven stability verification [119] and stabilization [120] with non-asymptotic guarantees. However, it is difficult to extend these works to more general MJSs with completely unknown dynamics. One major challenge brought by MJSs is that one needs to consider both the state/input in the continuous space and the Markovian mode switching sequence in the discrete space. Furthermore, the state data generated by the same mode are temporally separated with the mode switching, thus having time-varying statistical properties and posing difficulties to sample complexity analysis.

One advantage of MJSs is that, stability is only required in the mean-square sense, which relaxes the deterministic counterpart that is commonly needed for non-switched systems. This, however, brings new challenges to the analysis since unstable realization is

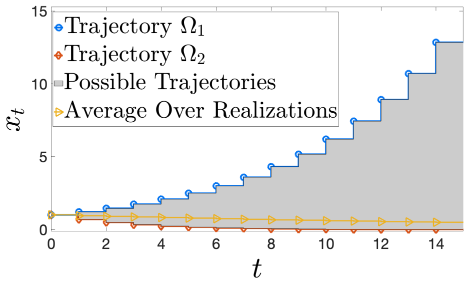


Figure 4.1: State trajectories for a two-modes MJS: Mode 1: $x_{t+1} = 1.2x_t$, Mode 2: $x_{t+1} = 0.7x_t$, Markov matrix $[[0.6, 0.4]^\intercal, [0.3, 0.7]^\intercal]^\intercal$, and $x_0 = 1$. Blue and red curves: mode switching sequences $\Omega_1 = \{1, 1, \dots\}$ and $\Omega_2 = \{2, 2, \dots\}$. Yellow curve: average over all realizations. Gray area: region for all possible trajectories.

possible with mean-square stability. Figure 4.1 shows an example (adapted from [116]) of an MJS that is stable in the mean-square sense despite having an unstable mode. Clearly, under an unfavorable mode switching sequence, the system trajectory can still blow up. Therefore, statistical tools such as high probability light-tail bounds are not applicable without strong assumptions on the joint spectral radius of the system (cf. [57]). Perhaps more surprisingly, there are examples of MJS with all modes individually stable, however due to switching, the system exhibits an unstable behavior on average, and the MJS is not mean-square stable [116, Example 3.17]. Therefore, finding controllers to individually stabilize the mode dynamics does not guarantee that the overall system will be stable when mode switches over time. This more relaxed notion of mean-square stability presents major challenges in learning, controlling, and statistical analysis.

4.1.1 Relation to Prior Work

Our work is related to several topics in model-based reinforcement learning, system identification, and adaptive control.

- **System Identification:** Learning dynamical models has a long history in the control community, with major theoretical results being related to asymptotic properties under strong assumptions on persistence of excitation [20]. The problem becomes harder for hybrid and switched systems where the initial focus was on computational complexity as opposed to sample complexity of learning [121, 122]. There are some recent results on asymptotic consistency [123] in the stochastic jump systems, a special case of MJSs where the modes switch in an i.i.d. manner. Identification of MJSs with hidden mode sequence has also attracted significant attention [124, 125].

- **Sample Complexity of System Identification:** There is a recent surge of interest toward understanding the sample complexity of learning linear dynamical systems from a single trajectory under mild assumptions [61], using statistical tools like martingales [22, 24, 25] or mixing time arguments [26, 27]. Recently, [28] provides precise rates for the finite-time identification of LTI (linear time-invariant) systems using a single trajectory. The literature gets scarcer for switched systems. In [126], a novel approach based on Lyapunov equation is proposed for systems with stochastic switches, yet theoretical guarantees are lacking. [57] is one of the early works – and it seems to be the only work not assuming persistence of excitation – to provide finite sample analysis for learning systems with stochastic switches, yet with additional strong assumptions like independent switches and small joint spectral radius. The proof techniques developed within our work aim to obviate such assumptions. We tackle the open problem of learning MJS from finite samples, obtained from a single trajectory, with theoretical guarantees under mild assumptions.

4.1.2 Contributions

We provide the first comprehensive system identification and regret guarantees for learning and controlling Markov jump linear systems using a single trajectory while assuming only marginal mean-square stability (see Definition 26). Specifically, our contributions are as follows[1]: We provide an algorithm (Algorithm. 2) to estimate the MJS dynamics with an error rate of $\mathcal{O}(\sqrt{(n+p)/T})$, where n and p are the state and input dimensions respectively, and T is the trajectory length. Our error rate is optimal in terms of the trajectory length T and the dimensions (n and p) of the unknown matrices.

4.2 Preliminaries and Problem Setup

We consider the identification and adaptive control of MJSs which are governed by the following state equation,

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{B}_{\omega(t)}\boldsymbol{u}_t + \boldsymbol{w}_t \quad \text{s.t.} \quad \omega(t) \sim \text{Markov Chain}(\boldsymbol{T}), \tag{4.2.1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$, $\boldsymbol{u}_t \in \mathbb{R}^p$ and $\boldsymbol{w}_t \in \mathbb{R}^n$ are the state, input, and process noise of the MJS at time t with $\{\boldsymbol{w}_t\}_{t=0}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$. There are s modes in total, and the dynamics of mode i is given by the state matrix $\boldsymbol{A}_i$ and input matrix $\boldsymbol{B}_i$. The active mode at time t is indexed by $\omega(t) \in [s]$. Throughout, we assume the state $\boldsymbol{x}_t$ and the mode $\omega(t)$ can be observed at time t. The mode switching sequence $\{\omega(t)\}_{t=0}^{\infty}$ follows a Markov chain with transition matrix $\boldsymbol{T} \in \mathbb{R}_{+}^{s \times s}$ such that for all $t \geq 0$, the ij-th element of $\boldsymbol{T}$ denotes the conditional probability $[\boldsymbol{T}]_{ij} := \mathbb{P}\big(\omega(t+1) = j \mid \omega(t) = i\big)$ for all $i, j \in [s]$. Throughout, we assume the initial state $\boldsymbol{x}_0$, the mode switching sequence $\{\omega(t)\}_{t=0}^{\infty}$, and the noise $\{\boldsymbol{w}_t\}_{t=0}^{\infty}$ are mutually independent.

[1] orders of magnitude here are up to polylogarithmic factors

We use MJS$(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}, \boldsymbol{T})$ to refer to an MJS with state equation (4.2.1), parameterized by the matrix tuple $(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}, \boldsymbol{T})$. We call a sequence of controllers $\boldsymbol{K}_{1:s} := \{\boldsymbol{K}_1, \dots, \boldsymbol{K}_s\}$ a mode-dependent state-feedback controller for the MJS if the input is given by $\boldsymbol{u}_t = \boldsymbol{K}_{\omega(t)} \boldsymbol{x}_t$. Under $\boldsymbol{K}_{1:s}$, the MJS becomes closed-loop with state matrices $\boldsymbol{L}_{1:s}$ where $\boldsymbol{L}_i := \boldsymbol{A}_i + \boldsymbol{B}_i \boldsymbol{K}_i$.

Due to the randomness in the mode sequence $\{\omega(t)\}_{t=0}^{\infty}$, it is common to consider the stability of MJS in the mean-square sense which is defined as follows.

Definition 26 (Mean-square stability [116]) *We say the MJS in* (4.2.1) *is mean-square stable (MSS) if when setting $\boldsymbol{u}_t = 0$, there exists $\boldsymbol{x}_\infty, \boldsymbol{\Sigma}_\infty$ such that for any initial state $\boldsymbol{x}_0$ and mode $\omega(0)$, as $t \to \infty$, we have*

$$\|\mathbb{E}[\boldsymbol{x}_t] - \boldsymbol{x}_\infty\|_{\ell_2} \to 0, \quad \|\mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top] - \boldsymbol{\Sigma}_\infty\| \to 0, \tag{4.2.2}$$

where the expectation is over the Markovian mode switching sequence $\{\omega(t)\}_{t=0}^{\infty}$, the noise $\{\boldsymbol{w}_t\}_{t=0}^{\infty}$ and the initial state $\boldsymbol{x}_0$. In the noise-free case (i.e., $\boldsymbol{w}_t = 0$), we have $\boldsymbol{x}_\infty = 0$, $\boldsymbol{\Sigma}_\infty = 0$. We say the MJS in (4.2.1) *is (mean-square) stabilizable if there exists mode-dependent controller $\boldsymbol{K}_{1:s}$ such that the closed-loop MJS $\boldsymbol{x}_{t+1} = (\boldsymbol{A}_{\omega(t)} + \boldsymbol{B}_{\omega(t)} \boldsymbol{K}_{\omega(t)}) \boldsymbol{x}_t$ is MSS. We call such $\boldsymbol{K}_{1:s}$ a stabilizing controller.*

Similarly to the Lyapunov stability of LTI systems, MJSs also have the spectral radius criterion to determine the MSS. For notation brevity, let $\boldsymbol{L}_{1:s}$ denote the MJS state matrices, where $\boldsymbol{L}_i = \boldsymbol{A}_i + \boldsymbol{B}_i \boldsymbol{K}_i$ for the closed-loop case and $\boldsymbol{L}_i = \boldsymbol{A}_i$ otherwise. Define the augmented state matrix $\tilde{\boldsymbol{L}} \in \mathbb{R}^{sn^2 \times sn^2}$ with the ij-th $n^2 \times n^2$ block given by $[\tilde{\boldsymbol{L}}]_{ij} := [\boldsymbol{T}]_{ji} \boldsymbol{L}_j \otimes \boldsymbol{L}_j$. Then, $\rho(\tilde{\boldsymbol{L}}) < 1$ if and only if the MJS is MSS [116, Theorem 3.9]. This follows from the fact that the matrix $\tilde{\boldsymbol{L}}$ maps $\mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top]$ to $\mathbb{E}[\boldsymbol{x}_{t+1} \boldsymbol{x}_{t+1}^\top]$ (see (4.4.5) in the appendix). Particularly, when the MJS has $\rho(\tilde{\boldsymbol{L}}) \le 1$, we refer to it as *marginally* MSS. The notions of marginally (mean-square) stabilizability and marginally (mean-square) stabilizing controller follow similarly.

• **System Identification:** System identification problems seek to estimate unknown system dynamics from a single (or multiple) trajectory(ies) of the system's states, inputs and mode observations. In the MJS setting, our goal is to estimate the state/input matrices $\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}$ and the Markov transition matrix $\boldsymbol{T}$ from a single trajectory of the system's states, inputs and mode observations $\{\boldsymbol{x}_t, \boldsymbol{u}_t, \omega(t)\}_{t=0}^{T}$, and provide finite sample estimation guarantees. In this work, the main assumption for the MJS to be identified is as follows.

Assumption 8 *The MJS in* (4.2.1) *has ergodic Markov chain and is marginally mean-square stabilizable.*

Ergodicity guarantees that the distribution of the mode sequence $\omega(t)$ converges to a unique strictly positive stationary distribution [127, Theorem 4.3.5]. Throughout, we let $\boldsymbol{\pi}_\infty \in \mathbb{R}_+^s$ denote the stationary distribution of $\boldsymbol{T}$ such that $\boldsymbol{\pi}_\infty^\top = \boldsymbol{\pi}_\infty^\top \boldsymbol{T}$, and define $\pi_{\min} \coloneqq \min_{i\in[s]} \boldsymbol{\pi}_\infty(i)$, $\pi_{\max} \coloneqq \max_{i\in[s]} \boldsymbol{\pi}_\infty(i)$. Ergodicity ensures that the MJS could have enough "visits" to every mode $i \in [s]$, thus providing enough number of samples to learn $[\boldsymbol{T}]_{i,:}$, $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ for all $i \in [s]$. We further define the mixing time [128] that describes how fast a Markov chain converges to its stationary distribution.

Definition 27 (Markov chain mixing time) *Consider an ergodic Markov matrix $\boldsymbol{T} \in \mathbb{R}_+^{s\times s}$ with stationary distribution $\boldsymbol{\pi}_\infty \in \mathbb{R}_+^s$. For $\epsilon \geq 0$, define the mixing time as*

$$t_{MC}(\epsilon) \coloneqq \min\left\{t \in \mathbb{N} : \max_{i\in[s]} \frac{1}{2}\|([\boldsymbol{T}^t]_{i,:})^\top - \boldsymbol{\pi}_\infty\|_{\ell_1} \leq \epsilon\right\}. \tag{4.2.3}$$

Particularly, when the parameter ϵ is omitted, $t_{MC} \coloneqq t_{MC}(\frac{1}{4})$.

```
Algorithm 2 MJS-SYSID
Input: A marginally mean-square stabilizing controller $\boldsymbol{K}_{1:s}$; variances $\sigma_{\boldsymbol{w}}^2$ and $\sigma_{\boldsymbol{z}}^2$; MJS trajectory $\{\boldsymbol{x}_t, \boldsymbol{z}_t, \omega(t)\}_{t=0}^T$, generated using inputs $\boldsymbol{u}_t = \boldsymbol{K}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{z}_t$; exploration niose $\{\boldsymbol{z}_t\}_{t=0}^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{z}}^2 \boldsymbol{I}_p)$
Estimate $\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}$: for all modes $i \in [s]$ do
    $S_i = \{t \mid \omega(t) = i\}$
    $\hat{\boldsymbol{\Theta}}_{1,i}, \hat{\boldsymbol{\Theta}}_{2,i} = \underset{\boldsymbol{\Theta}_1 \in \mathbb{R}^{n \times n}, \boldsymbol{\Theta}_2 \in \mathbb{R}^{n \times p}}{\arg\min} \frac{1}{2|S_i|} \sum_{t \in S_i} \|\boldsymbol{x}_{t+1} - \boldsymbol{\Theta}_1 \boldsymbol{x}_t/\sigma_{\boldsymbol{w}} - \boldsymbol{\Theta}_2 \boldsymbol{z}_t/\sigma_{\boldsymbol{z}}\|_{\ell_2}^2$
    $\hat{\boldsymbol{B}}_i = \hat{\boldsymbol{\Theta}}_{2,i}/\sigma_{\boldsymbol{z}}$ and $\hat{\boldsymbol{A}}_i = \hat{\boldsymbol{\Theta}}_{1,i}/\sigma_{\boldsymbol{w}} - \hat{\boldsymbol{B}}_i \boldsymbol{K}_i$
Estimate $\boldsymbol{T}$: $[\hat{\boldsymbol{T}}]_{ij} = \sum_{t=1}^T \mathbf{1}_{(\omega(t-1)=i, \omega(t)=j)} / \sum_{t=1}^T \mathbf{1}_{(\omega(t-1)=i)}$
Output: $\hat{\boldsymbol{A}}_{1:s}, \hat{\boldsymbol{B}}_{1:s}, \hat{\boldsymbol{T}}$
```

As mentioned earlier, MJS presents unique statistical analysis challenges due to Markovian jumps and MSS. In the following, Section 4.3 presents our system identification procedures together with theoretical guarantees overcoming these challenges.

4.3 Main Results

Our MJS identification procedure is given in Algorithm 2. We assume one has access to an stabilizing controller $\boldsymbol{K}_{1:s}$ to start the identification, which has been a standard assumption in data-driven control of LTI systems [50, 129–132]. Note that, if the open-loop MJS is already marginally MSS, then one can simply set $\boldsymbol{K}_{1:s} = 0$ and carry out MJS identification. Given an MJS trajectory $\{\boldsymbol{x}_t, \boldsymbol{z}_t, \omega(t)\}_{t=0}^T$, generated using the input $\boldsymbol{u}_t = \boldsymbol{K}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{z}_t$, we solve s least-squares regression problems to estimate $\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}$. Moreover, using the empirical frequency of observed modes, we estimate $\boldsymbol{T}$.

The following theorem gives our main results on learning the dynamics of an unknown MJS from finite samples obtained from a single trajectory. One can refer to Theorems 32 and 36 in Section 4.4 for the detailed theorem statements and proofs.

Theorem 28 (Identification of MJS) *Suppose we run Algorithm 2 with the trajectory length* $T \geq \max\{2T_0, \hat{\mathcal{O}}(\frac{(n+p)\log(T)}{\pi_{\min}(1-\gamma)})\}$, *where* $T_0 \coloneqq t_{MC}(\pi_{\min}/2)$ *and* $\gamma \coloneqq \hat{\mathcal{O}}(\frac{1}{\pi_{\min}}\sqrt{\frac{\pi_{\max}T_0}{T}})$. *Suppose,* $\{\boldsymbol{z}_t\}_{t=0}^T \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{z}}^2 \boldsymbol{I}_p)$ *and* $\{\boldsymbol{w}_t\}_{t=0}^T \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$. *Then, under Assumption 8, with probability at least* $1-\delta$, *for all* $i \in [s]$, *we have*

$$\begin{aligned}\max\{\frac{\sigma_{\boldsymbol{z}}}{\sigma_{\boldsymbol{w}} + \sigma_{\boldsymbol{z}}}\|\hat{\boldsymbol{A}}_i - \boldsymbol{A}_i\|, \frac{\sigma_{\boldsymbol{z}}}{\sigma_{\boldsymbol{w}}}\|\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i\|\} &\leq \hat{\mathcal{O}}\left(\sqrt{\frac{(n+p)\log(T)}{\pi_{\min}(1-\gamma)T}}\right), \\ \textit{and} \quad \|\hat{\boldsymbol{T}} - \boldsymbol{T}\| &\leq \hat{\mathcal{O}}\left(\frac{1}{\pi_{\min}}\sqrt{\frac{\log(T)}{T}}\right).\end{aligned} \tag{4.3.1}$$

Proof sketch: Let $\boldsymbol{h}_t \coloneqq [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}} \ \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top$ and $\boldsymbol{\Theta}_i^\star \coloneqq [\sigma_{\boldsymbol{w}}(\boldsymbol{A}_i + \boldsymbol{B}_i\boldsymbol{K}_i) \ \sigma_{\boldsymbol{z}}\boldsymbol{B}_i]$ for all $i \in [s]$. Then the output of each sample in $\{(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{z}_t, \omega(t))\}_{t\in S_i}$ can be related to the inputs as follows,

$$\boldsymbol{x}_{t_k+1} = \boldsymbol{\Theta}_i^\star \boldsymbol{h}_{t_k} + \boldsymbol{w}_{t_k} \quad \text{for} \quad k = 1, 2, \ldots, |S_i|, \tag{4.3.2}$$

where we set $S_i \coloneqq \{t \mid \omega(t) = i\} \equiv \{t_1, t_2, \cdots, t_{|S_i|}\}$. This shows that, for each $i \in [s]$, the problem of estimating $(\boldsymbol{A}_i, \boldsymbol{B}_i)$ is equivalent to the problem of estimating $\boldsymbol{\Theta}_i^\star$ from the sequence of covariate-response pairs $(\boldsymbol{h}_{t_k}, \boldsymbol{x}_{t_k+1})_{k\geq 1}$. Specifically, following Algorithm 2, we solve a regression problem. For this purpose, we define the following concatenated matrices: $\boldsymbol{Y}_i$ has $\{\boldsymbol{x}_{t+1}^\top\}_{t\in S_i}$ on its rows, $\boldsymbol{H}_i$ has $\{\boldsymbol{h}_t^\top\}_{t\in S_i}$ on its rows and $\boldsymbol{W}_i$ has $\{\boldsymbol{w}_t^\top\}_{t\in S_i}$ on its rows. Observe that, we have $\boldsymbol{Y}_i = \boldsymbol{H}_i\boldsymbol{\Theta}_i^{\star\top} + \boldsymbol{W}_i$ and the regression problem in Algorithm 2 becomes,

$$\hat{\boldsymbol{\Theta}}_i^\top = \underset{\boldsymbol{\Theta}_i \in \mathbb{R}^{n\times(n+p)}}{\arg\min} \frac{1}{2|S_i|}\|\boldsymbol{Y}_i - \boldsymbol{H}_i\boldsymbol{\Theta}_i^\top\|_F^2. \tag{4.3.3}$$

When the problem is over-determined, the solution to the least-squares problem (4.3.3) is given by $\hat{\boldsymbol{\Theta}}_i^\top = \boldsymbol{H}_i^\dagger \boldsymbol{Y}_i = (\boldsymbol{H}_i^\top \boldsymbol{H}_i)^{-1} \boldsymbol{H}_i^\top \boldsymbol{Y}_i$ and the associated estimation error is given by, $\hat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_i^\star = \big((\boldsymbol{H}_i^\top \boldsymbol{H}_i)^{-1} \boldsymbol{H}_i^\top \boldsymbol{W}_i\big)^\top$. This implies that the estimation error can be upper-bounded as follows,

$$\|\hat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_i^\star\| = \|(\boldsymbol{H}_i^\top \boldsymbol{H}_i)^{-1} \boldsymbol{H}_i^\top \boldsymbol{W}_i\| \le \frac{\|\boldsymbol{H}_i^\top \boldsymbol{W}_i\|}{\lambda_{\min}\big(\boldsymbol{H}_i^\top \boldsymbol{H}_i\big)}, \tag{4.3.4}$$

We upper bound the estimation error in (4.3.4) as follows: (a) First, we prove that the covariates process $\{\boldsymbol{h}_{t_k}\}_{k=1}^{|S_i|}$ satisfies $(k, \boldsymbol{I}_{n+p}, q)$-block Martingale small-ball condition, with the constants $k = 1$ and $q = 3/10$. (b) Next, we use Assumption 8 and Markov inequality to show that $\mathbb{P}\big(\boldsymbol{H}_i^\top \boldsymbol{H}_i \npreceq (|S_i|\bar{\Gamma}/\delta)\boldsymbol{I}_{n+p}\big) \le \delta$, for some $\bar{\Gamma} = \mathcal{O}(T)$. (c) Next, we use Assumption 8 (ergodicity) and Freedman's inequality to show that, using $T \ge 2T_0$, where $T_0 := t_{\text{MC}}(\pi_{\min}/2)$, we have $\mathbb{P}\big(\cap_{i=1}^s \{|S_i| \ge \pi_{\min}(1-\gamma)T\}\big) \ge 1-\delta$, where $\gamma := \hat{\mathcal{O}}\big(\frac{1}{\pi_{\min}}\sqrt{\frac{\pi_{\max} T_0}{T}}\big)$. Finally, we combine (a), (b) and (c) with Theorem 2.4 from [22] to obtain our main result on single trajectory learning of $\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}$. ■

Our system identification result achieves near-optimal ($\hat{\mathcal{O}}(\sqrt{(n+p)/T})$) dependence on the trajectory length T. Note that the overall sample complexity grows as $T \gtrsim (n+p)/\pi_{\min}$. A degrees-of-freedom counting argument would show that the dependency of $T \gtrsim (n+p)/\pi_{\min}$ is optimal. The reason is that, each vector state equation we fit has n scalar equations. The total degrees of freedom for each dynamics pair $(\boldsymbol{A}_i, \boldsymbol{B}_i)$ is $n \times (n+p)$. Additionally, for the least-frequent mode, in steady-state, we should observe $\pi_{\min} T$ equations. Putting these together, we would minimally need $n \times \pi_{\min} T \ge n \times (n+p)$, which means we need $T \ge (n+p)/\pi_{\min}$ samples to estimate the MJS dynamics $(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s})$. Note that, our sample complexity is not effected directly by the number of MJS modes s. However, s indirectly effects sample complexity via $\pi_{\min}$, which is the probability of least-frequent mode in the steady state.

It is well known that the least squares problem has a unique solution when the regressor matrix has full rank. For the least squares problem in Algorithm 2, the unknown input matrix $\boldsymbol{B}_i$ has regressor given by the random exploration noise $\{\boldsymbol{z}_t\}_{t \in S_i}$, which can be guaranteed to be full-rank when $\boldsymbol{z}_t$ has non-degenerate covariance. This ensures one can uniquely recover $\boldsymbol{B}_i$ thus is the reason we apply the additional $\boldsymbol{z}_t$ to the input $\boldsymbol{u}_t$. On the other hand, the regressor $\{\boldsymbol{x}_t\}_{t \in S_i}$ associated with the state matrix $\boldsymbol{A}_i$ is automatically guaranteed to be full-rank due to the presence of the additive process noise $\boldsymbol{w}_t$ in the MJS dynamics (4.2.1). This implies that, when $\boldsymbol{B}_{1:s}$ are known a priori, the exploration noise $\boldsymbol{z}_t$ is no longer needed, and one is still able to learn the remaining $\boldsymbol{A}_{1:s}$. The sample complexity guarantee for this case is provided in Corollary below.

Corollary 29 (Identification with known $\boldsymbol{B}_{1:s}$) *Consider the same setting of Theorem 28. Additionally, suppose $\boldsymbol{B}_{1:s}$ are known. Then, setting $\sigma_{\boldsymbol{z}} = 0$ and solving only for the state matrices, with probability at least $1 - \delta$, for all $i \in [s]$, we have $\|\hat{\boldsymbol{A}}_i - \boldsymbol{A}_i\| \leq \tilde{\mathcal{O}}\big(\sqrt{\frac{(n+p)\log(T)}{\pi_{\min}(1-\gamma)T}}\big)$.*

In Corollary 29, we show that, when $\boldsymbol{B}_{1:s}$ is assumed to be known, $\boldsymbol{A}_{1:s}$ can be estimated regardless of the exploration strength $\sigma_{\boldsymbol{z}}$. This is because the excitation for the state matrix arises from noise $\boldsymbol{w}_t$.

4.4 Proofs of the Main Results

4.4.1 MJS Covariance Dynamics Under MSS

Consider MJS$(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}, \boldsymbol{T})$ with process noise $\boldsymbol{w}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ and input $\boldsymbol{u}_t = \boldsymbol{K}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{z}_t$ under a stabilizing controller $\boldsymbol{K}_{1:s}$ and excitation for exploration $\boldsymbol{z}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{z}})$. Let $\boldsymbol{L}_i := \boldsymbol{A}_i + \boldsymbol{B}_i\boldsymbol{K}_i$ be the closed-loop state matrix. Let $\tilde{\boldsymbol{L}} \in \mathbb{R}^{sn^2 \times sn^2}$ be the augmented closed-loop state matrix with ij-th $n^2 \times n^2$ block given by $[\tilde{\boldsymbol{L}}]_{ij} = [\boldsymbol{T}]_{ji}\boldsymbol{L}_j \otimes \boldsymbol{L}_j$. Let $\tau_{\tilde{\boldsymbol{L}}} > 0$ and $\rho_{\tilde{\boldsymbol{L}}} \in [0,1]$ be two constants such that $\|\tilde{\boldsymbol{L}}^k\| \le \tau_{\tilde{\boldsymbol{L}}}\rho_{\tilde{\boldsymbol{L}}}^k$. By definitions of $\tau(\tilde{\boldsymbol{L}})$ and $\rho(\tilde{\boldsymbol{L}})$, one can choose them for $\tau_{\tilde{\boldsymbol{L}}}$ and $\rho_{\tilde{\boldsymbol{L}}}$ respectively. Let $\boldsymbol{\Sigma}_i(t) := \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top \mathbf{1}_{(\omega(t)=i)}]$, $\boldsymbol{\Sigma}(t) := \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]$,

$$\boldsymbol{s}_t := \begin{bmatrix} \mathbf{vec}(\boldsymbol{\Sigma}_1(t)) \\ \vdots \\ \mathbf{vec}(\boldsymbol{\Sigma}_s(t)) \end{bmatrix}, \quad \tilde{\boldsymbol{B}}_t := \begin{bmatrix} \sum_{j=1}^s \boldsymbol{\pi}_{t-1}(j)\boldsymbol{T}_{j1}(\boldsymbol{B}_j \otimes \boldsymbol{B}_j) \\ \vdots \\ \sum_{j=1}^s \boldsymbol{\pi}_{t-1}(j)\boldsymbol{T}_{js}(\boldsymbol{B}_j \otimes \boldsymbol{B}_j) \end{bmatrix}, \quad \text{and} \quad \tilde{\boldsymbol{\Pi}}_t := \boldsymbol{\pi}_t \otimes \boldsymbol{I}_{n^2}. \tag{4.4.1}$$

The following lemma shows how $\boldsymbol{s}_t$ depends on $\boldsymbol{s}_0$, $\boldsymbol{\Sigma}_{\boldsymbol{z}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{w}}$, which will be used to upper bound $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]$ in Lemma 31.

Lemma 30 *The vectorized covariance $\boldsymbol{s}_t$ has the following dynamics,*

$$\boldsymbol{s}_t = \tilde{\boldsymbol{L}}^t\boldsymbol{s}_0 + (\tilde{\boldsymbol{B}}_t + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{B}}_{t-1} + \cdots + \tilde{\boldsymbol{L}}^{t-1}\tilde{\boldsymbol{B}}_1)\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}}) + (\tilde{\boldsymbol{\Pi}}_t + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Pi}}_{t-1} + \cdots + \tilde{\boldsymbol{L}}^{t-1}\tilde{\boldsymbol{\Pi}}_1)\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

Proof. To begin, we evaluate $\boldsymbol{\Sigma}_i(t)$, from the equivalent MJS dynamics $\boldsymbol{x}_{t+1} = \boldsymbol{L}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{B}_{\omega(t)}\boldsymbol{z}_t + \boldsymbol{w}_t$, as follows,

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \mathbf{1}_{(\omega(t+1)=i)}] &= \sum_{j=1}^s \mathbb{E}[\boldsymbol{L}_j\boldsymbol{x}_t\boldsymbol{x}_t^\top \boldsymbol{L}_j \mathbf{1}_{(\omega(t+1)=i,\omega(t)=j)}] \\ &+ \sum_{j=1}^s \mathbb{E}[\boldsymbol{B}_j\boldsymbol{z}_t\boldsymbol{z}_t^\top \boldsymbol{B}_j^\top \mathbf{1}_{(\omega(t+1)=i,\omega(t)=j)}] + \mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t^\top \mathbf{1}_{(\omega(t+1)=i)}]. \end{aligned} \tag{4.4.2}$$

Since $\boldsymbol{w}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ and $\boldsymbol{z}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{z}})$, we get

$$\boldsymbol{\Sigma}_i(t+1) = \sum_{j=1}^{s} \boldsymbol{T}_{ji} \boldsymbol{L}_j \boldsymbol{\Sigma}_j(t) \boldsymbol{L}_j^\top + \sum_{j=1}^{s} \boldsymbol{\pi}_t(j) \boldsymbol{T}_{ji} \boldsymbol{B}_j \boldsymbol{\Sigma}_{\boldsymbol{z}} \boldsymbol{B}_j^\top + \boldsymbol{\pi}_{t+1}(i) \boldsymbol{\Sigma}_{\boldsymbol{w}}. \tag{4.4.3}$$

Vectorizing both sides of the above equation, we have

$$\begin{aligned}\mathbf{vec}(\boldsymbol{\Sigma}_i(t+1)) = &\sum_{j=1}^{s} \boldsymbol{T}_{ji} (\boldsymbol{L}_j \otimes \boldsymbol{L}_j) \mathbf{vec}(\boldsymbol{\Sigma}_j(t)) \\ &+ \sum_{j=1}^{s} \boldsymbol{\pi}_t(j) \boldsymbol{T}_{ji} (\boldsymbol{B}_j \otimes \boldsymbol{B}_j) \mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}}) + \boldsymbol{\pi}_{t+1}(i) \mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}}).\end{aligned} \tag{4.4.4}$$

Stacking this for every $i \in [s]$, we obtain

$$\begin{bmatrix} \mathbf{vec}(\boldsymbol{\Sigma}_1(t+1)) \\ \vdots \\ \mathbf{vec}(\boldsymbol{\Sigma}_s(t+1)) \end{bmatrix} = \tilde{\boldsymbol{L}} \begin{bmatrix} \mathbf{vec}(\boldsymbol{\Sigma}_1(t)) \\ \vdots \\ \mathbf{vec}(\boldsymbol{\Sigma}_s(t)) \end{bmatrix} + \tilde{\boldsymbol{B}}_{t+1} \mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}}) + \tilde{\boldsymbol{\Pi}}_{t+1} \mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}}). \tag{4.4.5}$$

Propagating this dynamics from t to 0 gives the desired result. ■

We next provide a key lemma that upper bounds $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]$ and $\|\boldsymbol{\Sigma}(t)\|_F$, which are later used extensively in system identification analysis.

Lemma 31 *For $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]$ and $\|\boldsymbol{\Sigma}(t)\|_F$, under MSS given in Definition 26, we have*

$$\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] \le \sqrt{ns}\tau_{\tilde{\boldsymbol{L}}} \big(\rho_{\tilde{\boldsymbol{L}}}^t \mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sqrt{n} \|\boldsymbol{B}_{1:s}\|^2 \|\boldsymbol{\Sigma}_{\boldsymbol{z}}\| \sum_{t'=1}^{t} \rho_{\tilde{\boldsymbol{L}}}^{t-t'} + \sqrt{n} \|\boldsymbol{\Sigma}_{\boldsymbol{w}}\| \sum_{t'=1}^{t} \rho_{\tilde{\boldsymbol{L}}}^{t-t'} \big), \tag{4.4.6}$$

$$\|\boldsymbol{\Sigma}(t)\|_F \le \sqrt{s}\tau_{\tilde{\boldsymbol{L}}} \big(\rho_{\tilde{\boldsymbol{L}}}^t \mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sqrt{n} \|\boldsymbol{B}_{1:s}\|^2 \|\boldsymbol{\Sigma}_{\boldsymbol{z}}\| \sum_{t'=1}^{t} \rho_{\tilde{\boldsymbol{L}}}^{t-t'} + \sqrt{n} \|\boldsymbol{\Sigma}_{\boldsymbol{w}}\| \sum_{t'=1}^{t} \rho_{\tilde{\boldsymbol{L}}}^{t-t'} \big). \tag{4.4.7}$$

Proof. First we derive an upper bound for $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]$. The upper bound for $\|\boldsymbol{\Sigma}(t)\|_F$ follows similarly. For state $\boldsymbol{x}_t$, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] &= \sum_{i=1}^{s} \mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2 \mathbf{1}_{(\omega(t)=i)}] = \sum_{i=1}^{s} \text{trace } (\mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top \mathbf{1}_{(\omega(t)=i)}]) = \sum_{i=1}^{s} \text{trace}(\boldsymbol{\Sigma}_i(t)), \\ &= \sum_{i=1}^{s} \sum_{j=1}^{n} \lambda_j(\boldsymbol{\Sigma}_i(t)) \le \sqrt{ns \sum_{i=1}^{s} \sum_{j=1}^{n} \lambda_j^2(\boldsymbol{\Sigma}_i(t))} \le \sqrt{ns \sum_{i=1}^{s} \|\boldsymbol{\Sigma}_i(t)\|_F^2}.\end{aligned} \tag{4.4.8}$$

Then, by definition of $\boldsymbol{s}_t$ in (4.4.1), we have

$$\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] \leq \sqrt{ns}\|\boldsymbol{s}_t\|_{\ell_2}. \tag{4.4.9}$$

Now, applying the dynamics of $\boldsymbol{s}_t$ from Lemma 30, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] &\leq \sqrt{ns}\Big(\|\tilde{\boldsymbol{L}}^t\|\|\boldsymbol{s}_0\|_{\ell_2} + \sum_{t'=1}^{t}\|\tilde{\boldsymbol{L}}^{t-t'}\|\|\tilde{\boldsymbol{B}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2} + \sum_{t'=1}^{t}\|\tilde{\boldsymbol{L}}^{t-t'}\|\|\tilde{\boldsymbol{\Pi}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}})\|_{\ell_2}\Big)\\ &\leq \sqrt{ns}\tau_{\tilde{\boldsymbol{L}}}\Big(\rho_{\tilde{\boldsymbol{L}}}^t\|\boldsymbol{s}_0\|_{\ell_2} + \sum_{t'=1}^{t}\rho_{\tilde{\boldsymbol{L}}}^{t-t'}\|\tilde{\boldsymbol{B}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2} + \sum_{t'=1}^{t}\rho_{\tilde{\boldsymbol{L}}}^{t-t'}\|\tilde{\boldsymbol{\Pi}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}})\|_{\ell_2}\Big),\end{aligned} \tag{4.4.10}$$

where the second line follows from $\|\tilde{\boldsymbol{L}}^t\| \leq \tau_{\tilde{\boldsymbol{L}}}\rho_{\tilde{\boldsymbol{L}}}^t$.

Now, we evaluate $\|\boldsymbol{s}_0\|_{\ell_2}$, $\|\tilde{\boldsymbol{B}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2}$, and $\|\tilde{\boldsymbol{\Pi}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}})\|_{\ell_2}$ separately. For the first term, we have

$$\|\boldsymbol{s}_0\|_{\ell_2} = \sqrt{\sum_{i=1}^{s}\|\boldsymbol{\Sigma}_i(0)\|_F^2} = \sqrt{\sum_{i=1}^{s}\boldsymbol{\pi}_0(i)^2\|\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top]\|_F^2} \leq \|\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top]\|_F \leq \mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2]. \tag{4.4.11}$$

Let $[\tilde{\boldsymbol{B}}_{t'}]_i$ denote the ith block of $\tilde{\boldsymbol{B}}_{t'}$, i.e., $[\tilde{\boldsymbol{B}}_{t'}]_i = \sum_{j=1}^{s}\boldsymbol{\pi}_{t-1}(j)\boldsymbol{T}_{ji}(\boldsymbol{B}_j \otimes \boldsymbol{B}_j)$, then

$$\begin{aligned}\|\tilde{\boldsymbol{B}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2} = \sqrt{\sum_{i=1}^{s}\|[\tilde{\boldsymbol{B}}_{t'}]_i\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2}^2} &\leq \sum_{i=1}^{s}\|[\tilde{\boldsymbol{B}}_{t'}]_i\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2}\\ &= \sum_{i=1}^{s}\|\sum_{j=1}^{s}\boldsymbol{\pi}_{t'-1}(j)\boldsymbol{T}_{ji}(\boldsymbol{B}_j \otimes \boldsymbol{B}_j)\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{z}})\|_{\ell_2}\\ &= \sum_{i=1}^{s}\|\sum_{j=1}^{s}\boldsymbol{\pi}_{t'-1}(j)\boldsymbol{T}_{ji}(\boldsymbol{B}_j\boldsymbol{\Sigma}_{\boldsymbol{z}}\boldsymbol{B}_j^\top)\|_F\\ &\leq \|\boldsymbol{B}_{1:s}\|^2\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\| \cdot \sum_{i=1}^{s}\|\sum_{j=1}^{s}\boldsymbol{\pi}_{t'-1}(j)\boldsymbol{T}_{ji}\boldsymbol{I}_n\|_F\\ &= \|\boldsymbol{B}_{1:s}\|^2\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\| \cdot \sum_{i=1}^{s}\|\boldsymbol{\pi}_{t'}(i)\boldsymbol{I}_n\|_F\\ &\leq \sqrt{n}\|\boldsymbol{B}_{1:s}\|^2\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|.\end{aligned} \tag{4.4.12}$$

Lastly, we have

$$\|\tilde{\boldsymbol{\Pi}}_{t'}\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}})\|_{\ell_2} = \sqrt{\sum_{i=1}^{s}\|\boldsymbol{\pi}_{t'}(i)\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}})\|_{\ell_2}^2} \le \|\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{w}})\|_{\ell_2} = \|\boldsymbol{\Sigma}_{\boldsymbol{w}}\|_F = \sqrt{n}\|\boldsymbol{\Sigma}_{\boldsymbol{w}}\|. \tag{4.4.13}$$

Plugging (4.4.11)–(4.4.13) into (4.4.10), we obtain

$$\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] \le \sqrt{ns}\tau_{\tilde{\boldsymbol{L}}}\big(\rho_{\tilde{\boldsymbol{L}}}^t\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sqrt{n}\|\boldsymbol{B}_{1:s}\|^2\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|\sum_{t'=1}^{t}\rho_{\tilde{\boldsymbol{L}}}^{t-t'} + \sqrt{n}\|\boldsymbol{\Sigma}_{\boldsymbol{w}}\|\sum_{t'=1}^{t}\rho_{\tilde{\boldsymbol{L}}}^{t-t'}\big), \tag{4.4.14}$$

which gives the bound for $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]$ in (4.4.6). To obtain the bound for $\|\boldsymbol{\Sigma}(t)\|_F$ in (4.4.7), note that $\|\boldsymbol{\Sigma}(t)\|_F = \|\sum_{i=1}^s \boldsymbol{\Sigma}_i(t)\|_F \le \sqrt{s\sum_{i=1}^s\|\boldsymbol{\Sigma}_i(t)\|_F^2} \le \sqrt{s}\|\boldsymbol{s}_t\|_{\ell_2}$. We then follow a similar line of reasoning as above to get the statement of the lemma. This completes the proof. ■

4.4.2 Estimation of $\boldsymbol{T}$

The following theorem adapted from [133, Lemma 7] provides the sample complexity result for estimating Markov matrix $\boldsymbol{T}$, which is a corresponds to the sample complexity on $\|\hat{\boldsymbol{T}} - \boldsymbol{T}\|$ in Theorem 28.

Theorem 32 *Suppose we have an ergodic Markov chain $\boldsymbol{T} \in \mathbb{R}^{s\times s}$ with mixing time t_{MC} and stationary distribution $\boldsymbol{\pi}_\infty \in \mathbb{R}^s$. Let $\pi_{\max} := \max_{i\in[s]}\boldsymbol{\pi}_\infty(i)$ and $\pi_{\min} := \min_{i\in[s]}\boldsymbol{\pi}_\infty(i)$. Given a state sequence $\omega(0), \omega(1), \dots, \omega(T)$ of the Markov chain, define the empirical estimator $\hat{\boldsymbol{T}}$ of the Markov matrix as follows,*

$$[\hat{\boldsymbol{T}}]_{ij} = \frac{\sum_{t=1}^{T-1}\mathbf{1}_{(\omega(t)=i,\omega(t+1)=j)}}{\sum_{t=1}^{T-1}\mathbf{1}_{(\omega(t)=i)}}, \tag{4.4.15}$$

Assume for some $\delta > 0$, $T \ge \underline{T}_{MC,1}(C_{MC}, \frac{\delta}{4}) := \{68C_{MC}\pi_{\max}\pi_{\min}^{-2}\log(\frac{4s}{\delta})\}^2$, where $C_{MC} := t_{\mathrm{MC}}\cdot\max\{3, 3-3\log(\pi_{\max}\log(s))\}$. Then, we have with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{T}} - \boldsymbol{T}\| \le \frac{4\|\boldsymbol{T}\|}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}C_{MC}\log(T)\log(4sC_{MC}\log(T)/\delta)}{T}}. \tag{4.4.16}$$

Proof. We first consider estimators computed using a sub-trajectory of $\omega(0), \omega(1), \ldots, \omega(T)$, then combine them together to show the error bound for $\hat{\boldsymbol{T}}$ in the claim. For $C_{MC} := t_{\text{MC}} \cdot \max\{3, 3 - 3\log(\pi_{\max}\log(s))\}$, let $L = C_{MC}\log(T)$. Then, for $\ell = 0, 1, \ldots, L-1$, define $\hat{\boldsymbol{T}}^{(\ell)} \in \mathbb{R}^{s\boldsymbol{x}s}$ such that $[\hat{\boldsymbol{T}}^{(\ell)}]_{ij} = \frac{\sum_{k=1}^{\lfloor T/L \rfloor} \mathbf{1}_{(\omega(kL+\ell)=i, \omega(kL+1+\ell)=j)}}{\sum_{k=1}^{\lfloor T/L \rfloor} \mathbf{1}_{(\omega(kL+\ell)=i)}}$. In other words, $\hat{\boldsymbol{T}}^{(\ell)}$ is the estimator computed using data with sub-sampling period L. Following the proof of [133, Lemma 7], we know for any $\epsilon < \pi_{\min}/2$, suppose $L \geq 6t_{MC}\log(\epsilon^{-1})$.

$$\mathbb{P}\{\|\hat{\boldsymbol{T}}^{(\ell)} - \boldsymbol{T}\| \leq 4\pi_{\min}^{-1}\|\boldsymbol{T}\|\epsilon\} \geq 1 - 4s\exp\{-\frac{T\epsilon^2}{17\pi_{\max}L}\}. \tag{4.4.17}$$

By setting $\delta = 4s\exp\{-\frac{T\epsilon^2}{17\pi_{\max}L}\}$, one can also interpret the above result as: for all $\delta > 0$, suppose

$$L \geq 3t_{MC}\log\left(\frac{T}{17\pi_{\max}L\log(\frac{4s}{\delta})}\right), \tag{4.4.18}$$

then when

$$T \geq 68L\pi_{\max}\pi_{\min}^{-2}\log(\frac{4s}{\delta}), \tag{4.4.19}$$

we have with probability at least $1 - \delta$

$$\|\hat{\boldsymbol{T}}^{(\ell)} - \boldsymbol{T}\| \leq \frac{4\|\boldsymbol{T}\|}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}C_{MC}\log(T)\log(4s/\delta)}{T}}. \tag{4.4.20}$$

One can verify (4.4.18) holds by plugging in $L = C_{MC}\log(T)$ and using definition $C_{MC} := t_{\text{MC}} \cdot \max\{3, 3 - 3\log(\pi_{\max}\log(s))\}$; (4.4.19) holds under the premise condition $T \geq \underline{T}_{MC,1}(C_{MC}, \frac{\delta}{4}) := \{68C_{MC}\pi_{\max}\pi_{\min}^{-2}\log(\frac{4s}{\delta})\}^2$.

Note that by definition, $\hat{\boldsymbol{T}}$ can be viewed as a convex combination of $\hat{\boldsymbol{T}}^{(\ell)}$ for all $\ell = 0, 1, \ldots, L$, thus by triangle inequality and union bound, we have with probability $1 - L\delta$,

$$\|\hat{\boldsymbol{T}} - \boldsymbol{T}\| \leq \frac{4\|\boldsymbol{T}\|}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}C_{MC}\log(T)\log(4s/\delta)}{T}}. \tag{4.4.21}$$

Finally, by replacing $L\delta$ with δ, we could show (4.4.16) and conclude the proof. ■

4.4.3 Estimation of $\boldsymbol{A}_{1:s}$ and $\boldsymbol{B}_{1:s}$

In this section, we estimate the unknown MJS dynamics $\boldsymbol{A}_{1:s}$ and $\boldsymbol{B}_{1:s}$ from finite samples obtained from a single trajectory of (4.2.1). Given a stabilizing controller $\boldsymbol{K}_{1:s}$, under the input $\boldsymbol{u}_t = \boldsymbol{K}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{z}_t$, the MJS state equation (4.2.1) becomes,

$$\boldsymbol{x}_{t+1} = \boldsymbol{L}_{\omega(t)}\boldsymbol{x}_t + \boldsymbol{B}_{\omega(t)}\boldsymbol{z}_t + \boldsymbol{w}_t, \quad \text{s.t.} \quad \omega(t) \sim \text{Markov Chain}(\boldsymbol{T}), \tag{4.4.22}$$

where $\boldsymbol{L}_{\omega(t)} := \boldsymbol{A}_{\omega(t)} + \boldsymbol{B}_{\omega(t)}\boldsymbol{K}_{\omega(t)}$ denotes the closed-loop state matrix, and $\{\boldsymbol{z}_t\}_{t=0}^{T} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{z}}^2 \boldsymbol{I}_p)$ is the i.i.d. excitation for exploration. To estimate the unknown system dynamics $(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s})$, we run the closed-loop MJS (4.4.22) for T time-steps and collect the trajectory $(\boldsymbol{x}_t, \boldsymbol{z}_t, \omega(t))_{t=0}^{T}$. Then, we run Algorithm 2 on the collected trajectory to obtain the estimates $(\hat{\boldsymbol{A}}_{1:s}, \hat{\boldsymbol{B}}_{1:s})$. To proceed, let $\boldsymbol{h}_t := [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}} \;\; \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top$ and $\boldsymbol{\Theta}_i^\star := [\sigma_{\boldsymbol{w}}\boldsymbol{L}_i \;\; \sigma_{\boldsymbol{z}}\boldsymbol{B}_i]$ for all $i \in [s]$. Then the output of each sample in $\{(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{z}_t, \omega(t))\}_{t\in S_i}$ can be related to the inputs as follows,

$$\boldsymbol{x}_{t_k+1} = \boldsymbol{\Theta}_i^\star \boldsymbol{h}_{t_k} + \boldsymbol{w}_{t_k} \quad \text{for} \quad k = 1, 2, \dots, |S_i|, \tag{4.4.23}$$

where we set $S_i := \{t \mid \omega(t) = i\} \equiv \{t_1, t_2, \cdots, t_{|S_i|}\}$. This shows that, for each $i \in [s]$, the problem of estimating $(\boldsymbol{A}_i, \boldsymbol{B}_i)$ is equivalent to the problem of estimating $\boldsymbol{\Theta}_i^\star$ from the sequence of covariate-response pairs $(\boldsymbol{h}_{t_k}, \boldsymbol{x}_{t_k+1})_{k\geq 1}$. Specifically, following Algorithm 2, we solve a regression problem. For this purpose, we define the following concatenated matrices,

$$\boldsymbol{Y}_i = \begin{bmatrix} \boldsymbol{x}_{t_1+1}^\top \\ \boldsymbol{x}_{t_2+1}^\top \\ \vdots \\ \boldsymbol{x}_{t_{|S_i|}+1}^\top \end{bmatrix}, \quad \boldsymbol{H}_i = \begin{bmatrix} \boldsymbol{h}_{t_1}^\top \\ \boldsymbol{h}_{t_2}^\top \\ \vdots \\ \boldsymbol{h}_{t_{|S_i|}}^\top \end{bmatrix}, \quad \boldsymbol{W}_i = \begin{bmatrix} \boldsymbol{w}_{t_1}^\top \\ \boldsymbol{w}_{t_2}^\top \\ \vdots \\ \boldsymbol{w}_{t_{|S_i|}}^\top \end{bmatrix}, \tag{4.4.24}$$

that is, $\boldsymbol{Y}_i$ has $\{\boldsymbol{x}_{t+1}^\top\}_{t\in S_i}$ on its rows, $\boldsymbol{H}_i$ has $\{\boldsymbol{h}_t^\top\}_{t\in S_i}$ on its rows and $\boldsymbol{W}_i$ has $\{\boldsymbol{w}_t^\top\}_{t\in S_i}$ on its rows. Observe that, we have $\boldsymbol{Y}_i = \boldsymbol{H}_i\boldsymbol{\Theta}_i^{\star\top} + \boldsymbol{W}_i$ and the regression problem in Algorithm 2 becomes,

$$\hat{\boldsymbol{\Theta}}_i^\top = \underset{\boldsymbol{\Theta}_i\in\mathbb{R}^{n\times(n+p)}}{\arg\min}\ \frac{1}{2|S_i|}\|\boldsymbol{Y}_i - \boldsymbol{H}_i\boldsymbol{\Theta}_i^\top\|_F^2. \tag{4.4.25}$$

When the problem is over-determined, the solution to the least-squares problem (4.4.25) is given by $\hat{\boldsymbol{\Theta}}_i^\top = \boldsymbol{H}_i^\dagger\boldsymbol{Y}_i = (\boldsymbol{H}_i^\top\boldsymbol{H}_i)^{-1}\boldsymbol{H}_i^\top\boldsymbol{Y}_i$ and the associated estimation error is given by, $\hat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_i^\star = \big((\boldsymbol{H}_i^\top\boldsymbol{H}_i)^{-1}\boldsymbol{H}_i^\top\boldsymbol{W}_i\big)^\top$. This implies that the estimation error can be upper-bounded as follows,

$$\|\hat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_i^\star\| = \|(\boldsymbol{H}_i^\top\boldsymbol{H}_i)^{-1}\boldsymbol{H}_i^\top\boldsymbol{W}_i\| \le \frac{\|\boldsymbol{H}_i^\top\boldsymbol{W}_i\|}{\lambda_{\min}\big(\boldsymbol{H}_i^\top\boldsymbol{H}_i\big)}, \tag{4.4.26}$$

To make the problem (4.4.25) well-conditioned, we also need a stability guarantee on the closed-loop MJS (4.4.22). This will make sure that the design matrix $\boldsymbol{H}_i$ has smaller condition number to help better estimation. Specifically, we will use the notion of mean-square stability introduced by Definition 26 to achieve this.

At the core of our analysis is showing that the process $\{\boldsymbol{h}_t := [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}}\ \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top\}_{t\in S_i}$ satisfies the martingale small-ball condition (for each $i\in[s]$), which is defined as follows.

Definition 33 (Martingale small-ball [22]) *Let $\{\mathcal{F}_t\}_{t\ge1}$ denotes a filtration and $\{Z_t\}_{t\ge1}$ be an $\{\mathcal{F}_t\}_{t\ge1}$-adapted random process taking values in $\mathbb{R}$. We say $\{Z_t\}_{t\ge1}$ satisfies the (k,ν,q)-block martingale small-ball (BMSB) condition if, for any $j\ge0$, one has $\frac{1}{k}\sum_{i=1}^k\mathbb{P}\big(|Z_{j+i}|\ge\nu \mid \mathcal{F}_j\big)\ge q$ almost surely. Given a process $\{\boldsymbol{x}_t\}_{t\ge1}$ taking values in $\mathbb{R}^d$, we say it satisfies the $(k,\boldsymbol{\Gamma}_{\text{sb}},q)$-BMSB condition for $\boldsymbol{\Gamma}_{\text{sb}}\succ0$ if, for any fixed $\boldsymbol{v}\in\mathcal{S}^{d-1}$, the process $Z_t=\langle\boldsymbol{v},\boldsymbol{x}_t\rangle$ satisfies $(k,\sqrt{\boldsymbol{v}^\top\boldsymbol{\Gamma}_{\text{sb}}\boldsymbol{v}},q)$-BMSB.*

To show that the random process $\{\boldsymbol{h}_t := [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}}\ \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top\}_{t\in S_i}$ satisfies BMSB condition, let $\mathcal{F}_t := \sigma(\boldsymbol{x}_0,\dots,\boldsymbol{x}_t,\boldsymbol{z}_0,\dots,\boldsymbol{z}_t,\boldsymbol{w}_0,\dots,\boldsymbol{w}_{t-1},\omega(1),\dots,\omega(t))$ denotes the filtration generated by the states, the excitation and the noise processes, and the mode switching sequence when $t \geq 1$. Furthermore, let $\mathcal{F}_0 := \sigma(\boldsymbol{x}_0,\boldsymbol{z}_0,\omega(0))$. Then, $\boldsymbol{x}_t, \boldsymbol{z}_t$ and $\omega(t)$ become $\mathcal{F}_t$-measurable and $\boldsymbol{w}_t$ is $\mathcal{F}_{t+1}$-measurable.

Theorem 34 (BMSB condition for $\{\boldsymbol{h}_t\}_{t\geq 1}$) *Consider closed-loop MJS (4.4.22). Suppose $\{\boldsymbol{z}_t\}_{t=0}^\infty \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{z}}^2\boldsymbol{I}_p)$ and $\{\boldsymbol{w}_t\}_{t=0}^\infty \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{w}}^2\boldsymbol{I}_n)$. Then, the covariate process $\{\boldsymbol{h}_t = [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}}\ \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top\}_{t\geq 1}$ satisfies the $(k,\boldsymbol{I}_{n+p},q)$-martingale small-ball condition, with the constants $k=1$ and $q=3/10$.*

The theorem above uses martingale small-ball with $k=1$. We remark that using $k>1$ is expected to help capture the role of additional excitation terms in the BMSB lower bound, specifically, the dependence on $\tilde{\boldsymbol{L}}$. However, this requires bounding higher order moments that involve cross-products of the input signal and noise terms and is left as future research.

Next, under the ergodicity of Markov chain (Assumption 8), we establish a high probability lower bound on the cardinality of the set $S_i := \{t \mid \omega(t) = i\} \equiv \{t_1, t_2, \cdots, t_{|S_i|}\}$. Our result is stated in the following theorem, which plays a critical role in establishing finite sample learning guarantees for the unknown MJS state and input matrices $\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}$.

Theorem 35 (Lower bound on $|S_i|$) *Let $\{\omega(t)\}_{t=0}^\infty$ be an ergodic Markov chain with the transition matrix $\boldsymbol{T} \in \mathbb{R}_+^{s\times s}$. Let $t_{\mathrm{MC}}(\epsilon)$ be as in Definition 27, and define $T_0 := t_{\mathrm{MC}}(\pi_{\min}/2)$. Let S_i be as in Algorithm 2. Fix $\delta \in (0,1)$, such that $\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}} \leq \pi_{\min}/2$. Then,* *choosing $T \geq 2T_0$, we have*

$$\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{|S_i| \geq \frac{\pi_{\min}T}{4}\left(1 - \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}}\right)\right\}\right) \geq 1-\delta. \tag{4.4.27}$$

The theorem above states that, choosing $T \geq 2t_{\mathrm{MC}}(\pi_{\min}/2)$, an ergodic Markov chain is guaranteed to visit each mode $i \in [s]$, at least $\hat{\mathcal{O}}(\pi_{\min}T)$ times. We remark that, our estimate is consistent with the asymptotic case when $T \to \infty$. Note that, the term $\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}}$ in (4.4.27) can be made arbitrary small by choosing sufficiently large trajectory length T. Finally, we combine Theorems 34 and 35 with Theorem 2.4 from [22] to obtain our main result on single trajectory learning of $\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s}$.

Theorem 36 (Identification of MJS) *Fix $\delta \in (0,1)$, such that,*

$$\gamma := \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}T_0\log(2sT_0/\delta)}{T-2T_0}} \leq \frac{1}{2}. \tag{4.4.28}$$

Suppose we run Algorithm 2 with the trajectory length T satisfying the following lower bound,

$$T \gtrsim \max\left\{2T_0, \frac{(n+p)+\log(6s\bar{\Gamma}/\delta)+\log(6s/\delta)}{\pi_{\min}(1-\gamma)}\right\}, \tag{4.4.29}$$

where $T_0 := t_{\mathrm{MC}}(\pi_{\min}/2)$ and $\bar{\Gamma} := \sqrt{ns}\tau_{\tilde{\boldsymbol{L}}}\big(\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2]/\sigma_{\boldsymbol{w}}^2 + (\sigma_{\boldsymbol{z}}^2/\sigma_{\boldsymbol{w}}^2)\sqrt{n}\|\boldsymbol{B}_{1:s}\|^2T + \sqrt{n}T\big) + p$. Suppose $\{\boldsymbol{z}_t\}_{t=0}^{T} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{z}}^2\boldsymbol{I}_p)$, $\{\boldsymbol{w}_t\}_{t=0}^{T} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{w}}^2\boldsymbol{I}_n)$. Let $C_K := \max_{i\in[s]}\|\boldsymbol{K}\|$. Then, under Assumption 8, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{\|\hat{\boldsymbol{A}}_i - \boldsymbol{A}_i\| \lesssim \frac{(C_K\sigma_{\boldsymbol{w}}+\sigma_{\boldsymbol{z}})}{\sigma_{\boldsymbol{z}}}\sqrt{\frac{(n+p)+\log(6s\bar{\Gamma}/\delta)+\log(6s/\delta)}{\pi_{\min}(1-\gamma)T}}\right\}\right) \geq 1-\delta,$$
$$\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{\|\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i\| \lesssim \frac{\sigma_{\boldsymbol{w}}}{\sigma_{\boldsymbol{z}}}\sqrt{\frac{(n+p)+\log(6s\bar{\Gamma}/\delta)+\log(6s/\delta)}{\pi_{\min}(1-\gamma)T}}\right\}\right) \geq 1-\delta. \tag{4.4.30}$$

Here, a few remarks are in place. First, the result appears to be convoluted however most of the dependencies are logarithmic (specifically, the dependency on the failure probability δ

and $\log(T)$ terms). Besides these, the dominant term, when estimating $\boldsymbol{A}_{1:s}$, $\boldsymbol{B}_{1:s}$ reduces to

$$\frac{(C_K\sigma_{\boldsymbol{w}} + \sigma_{\boldsymbol{z}})}{\sigma_{\boldsymbol{z}}}\sqrt{\frac{n+p}{\pi_{\min}T}} \quad \text{and} \quad \frac{\sigma_{\boldsymbol{w}}}{\sigma_{\boldsymbol{z}}}\sqrt{\frac{n+p}{\pi_{\min}T}},$$

respectively, which is identical to our statement in Theorem 28. Note that the overall sample complexity grows as $T \gtrsim (n+p)/\pi_{\min}$. A degrees-of-freedom counting argument would show that the dependency of $T \gtrsim (n+p)/\pi_{\min}$ is optimal. The reason is that, each vector state equation we fit has n scalar equations. The total degrees of freedom for each dynamics pair $(\boldsymbol{A}_i, \boldsymbol{B}_i)$ is $n \times (n+p)$. Additionally, for the least-frequent mode, in steady-state, we should observe $\pi_{\min}T$ equations. Putting these together, we would minimally need $n \times \pi_{\min}T \geq n \times (n+p)$, which means we need $T \geq (n+p)/\pi_{\min}$ samples to estimate the MJS dynamics $(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s})$. Note that, our sample complexity is not effected directly by the number of MJS modes s. However, s indirectly effects sample complexity via $\pi_{\min}$, which is the probability of least-frequent mode in the steady state.

4.5 Proofs of Intermediate Theorems and Lemmas

4.5.1 Proof of Theorem 34

Proof. In this subsection, we will show that the process $\{\boldsymbol{h}_t = [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}} \; \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top\}_{t\geq 1}$ satisfies $(1, \boldsymbol{I}_{n+p}, q)$-BMSB condition, for some constant $q > 0$. For this purpose, we need to show that, for any fixed $\boldsymbol{v} \in \mathcal{S}^{n+p-1}$, the random process $\{Z_t\}_{t\geq 1} \coloneqq \{\langle \boldsymbol{v}, \boldsymbol{h}_t\rangle\}_{t\geq 1}$ satisfies $(1, \|\boldsymbol{v}\|_{\ell_2}, q)$-BMSB condition, that is, for any $j \geq 0$, we need to show that $\mathbb{P}(|Z_{j+1}| \geq \|\boldsymbol{v}\|_{\ell_2} \mid \mathcal{F}_j) \geq q$ almost surely. For any $j \geq 0$, consider the concatenated state vector,

$$\boldsymbol{h}_{j+1} = \begin{bmatrix} \boldsymbol{x}_{j+1}/\sigma_{\boldsymbol{w}} \\ \boldsymbol{z}_{j+1}/\sigma_{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{L}_{\omega(j)}\boldsymbol{x}_j + \boldsymbol{B}_{\omega(j)}\boldsymbol{z}_j + \boldsymbol{w}_j)/\sigma_{\boldsymbol{w}} \\ \boldsymbol{z}_{j+1}/\sigma_{\boldsymbol{z}} \end{bmatrix}. \tag{4.5.1}$$

For any fixed $\boldsymbol{v} \in \mathcal{S}^{n+p-1}$, let $\boldsymbol{v}_1 \in \mathbb{R}^n$ and $\boldsymbol{v}_2 \in \mathbb{R}^p$ such that $\boldsymbol{v} = [\boldsymbol{v}_1^\top \ \boldsymbol{v}_2^\top]^\top$. Combining this with (4.5.1), we get

$$Z_{j+1} := \langle \boldsymbol{v}, \boldsymbol{h}_{j+1} \rangle = \sigma_{\boldsymbol{w}}^{-1} \left\langle \boldsymbol{v}_1, \boldsymbol{L}_{\omega(j)}\boldsymbol{x}_j + \boldsymbol{B}_{\omega(j)}\boldsymbol{z}_j + \boldsymbol{w}_j \right\rangle + \sigma_{\boldsymbol{z}}^{-1} \langle \boldsymbol{v}_2, \boldsymbol{z}_{j+1} \rangle . \tag{4.5.2}$$

To proceed, let $\{\mathcal{F}_t\}_{t\geq 1}$ denotes the filtration as defined before Theorem 34. Then, it is easy to see that $\sigma_{\boldsymbol{w}}^{-1} \left\langle \boldsymbol{v}_1, \boldsymbol{L}_{\omega(j)}\boldsymbol{x}_j + \boldsymbol{B}_{\omega(j)}\boldsymbol{z}_j + \boldsymbol{w}_j \right\rangle \bigm| \mathcal{F}_j \sim \mathcal{N}(\sigma_{\boldsymbol{w}}^{-1} \left\langle \boldsymbol{v}_1, \boldsymbol{L}_{\omega(j)}\boldsymbol{x}_j + \boldsymbol{B}_{\omega(j)}\boldsymbol{z}_j \right\rangle, \|\boldsymbol{v}_1\|_{\ell_2}^2)$. This is because $\boldsymbol{x}_j, \boldsymbol{z}_j$ and $\omega(j)$ are $\mathcal{F}_j$-measurable, whereas, $\boldsymbol{w}_j$ is $\mathcal{F}_{j+1}$-measurable. Similarly, $\sigma_{\boldsymbol{z}}^{-1} \langle \boldsymbol{v}_2, \boldsymbol{z}_{j+1} \rangle \bigm| \mathcal{F}_j \sim \mathcal{N}(0, \|\boldsymbol{v}_2\|_{\ell_2}^2)$. Furthermore, since $\boldsymbol{w}_j$ and $\boldsymbol{z}_{j+1}$ are independent, $Z_{j+1} \bigm| \mathcal{F}_j$ has the following distribution,

$$Z_{j+1} \bigm| \mathcal{F}_j \sim \mathcal{N}(\sigma_{\boldsymbol{w}}^{-1} \left\langle \boldsymbol{v}_1, \boldsymbol{L}_{\omega(j)}\boldsymbol{x}_j + \boldsymbol{B}_{\omega(j)}\boldsymbol{z}_j \right\rangle, \|\boldsymbol{v}_1\|_{\ell_2}^2 + \|\boldsymbol{v}_2\|_{\ell_2}^2). \tag{4.5.3}$$

Therefore, integrating the probability density function of a standard Gaussian random variable, it can be shown that,

$$\mathbb{P}\left(|\langle \boldsymbol{v}, \boldsymbol{h}_{j+1} \rangle| \geq \|\boldsymbol{v}\|_{\ell_2} \bigm| \mathcal{F}_j \right) \geq 3/10, \tag{4.5.4}$$

where we obtain the above result by integrating the probability density function of a Gaussian random variable as follows,

$$\begin{aligned} \forall \alpha \in \mathbb{R} \quad \mathbb{P}_{Z\sim\mathcal{N}(0,\sigma^2)}(|\alpha + Z| \geq \sigma) \geq \mathbb{P}_{Z\sim\mathcal{N}(0,\sigma^2)}(|Z| \geq \sigma) &= \mathbb{P}_{Z'\sim\mathcal{N}(0,1)}(|Z'| \geq 1), \\ &= 1 - \mathbb{P}_{Z'\sim\mathcal{N}(0,1)}(|Z'| \leq 1), \\ &= 1 - 2\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-z'^2/2} dz', \\ &\geq 1 - 2(7/20) = 3/10. \end{aligned} \tag{4.5.5}$$

This verifies our claim that the process $\{\boldsymbol{h}_t = [\boldsymbol{x}_t^\top/\sigma_{\boldsymbol{w}} \ \boldsymbol{z}_t^\top/\sigma_{\boldsymbol{z}}]^\top\}_{t\geq 1}$ satisfies $(1, \boldsymbol{I}_{n+p}, 3/10)$-BMSB condition. This completes the proof. ■

4.5.2 Proof of Theorem 35

Proof. To begin, recall that $t_{\mathrm{MC}}(\epsilon) := \min\{t \in \mathbb{N} : \max_{j\in[s]} \frac{1}{2}\|([\boldsymbol{T}^t]_{j,:})^\top - \boldsymbol{\pi}_\infty\|_{\ell_1} \leq \epsilon\}$, and $([\boldsymbol{T}^L]_{i,:})\mathbf{1} = \boldsymbol{\pi}_\infty^\top \mathbf{1} = 1$, for all $i \in [s]$. Therefore, choosing $L \geq t_{\mathrm{MC}}(\pi_{\min}/2)$, we have

$$\max_{j\in[s]} \|([\boldsymbol{T}^L]_{j,:})^\top - \boldsymbol{\pi}_\infty\|_{\ell_\infty} \leq \frac{\pi_{\min}}{2}. \tag{4.5.6}$$

To proceed, let $\mathbb{Z}^+ := \{1, 2, 3, \ldots\}$ denotes the set of positive integers. Then, to lower bound $|S_i|$ in Algorithm 2, we split the set $S_i := \{t \mid \omega(t) = i\}$ into $L \geq 1$ subsets via $S_i = \cup_{\ell=0}^{L-1} S_i^{(\ell)}$, such that

$$S_i^{(\ell)} := \left\{t \mid \omega(t) = i, \ (t-\ell)/L \in \mathbb{Z}^+ \rfloor\right\}, \tag{4.5.7}$$

where $0 \leq \ell \leq L-1$ is a fixed offset. Let $\{\mathcal{F}_t\}_{t\geq 1}$ denotes the filtration as defined before Theorem 34. To ease the notation, we let $\tilde{\omega}(k) := \omega(\ell + kL)$, and $\tilde{\mathcal{F}}_k := \mathcal{F}_{\ell+kL}$, for all $k \in \mathbb{Z}^+$. Then, one can see that $\tilde{\omega}(k)$ is $\tilde{\mathcal{F}}_k$-measurable. To proceed, define $\boldsymbol{\delta}_k, \boldsymbol{\Delta}_k \in \mathbb{R}^s$ such that

$$\begin{aligned} \boldsymbol{\delta}_k(i) &:= \mathbf{1}_{(\tilde{\omega}(k)=i)} - \mathbb{E}[\mathbf{1}_{(\tilde{\omega}(k)=i)} \mid \tilde{\mathcal{F}}_{k-1}], \\ \boldsymbol{\Delta}_k(i) &:= \sum_{j=1}^k \boldsymbol{\delta}_j(i). \end{aligned} \tag{4.5.8}$$

Note that for all $i \in [s]$, the random process $\{\boldsymbol{\Delta}_k(i)\}_{k\in\mathbb{Z}^+}$, adapted to the filtration $\{\tilde{\mathcal{F}}_k\}_{k\in\mathbb{Z}^+}$, forms a martingale, that is, we have

$$\begin{aligned} \mathbb{E}[\boldsymbol{\Delta}_{k+1}(i) \mid \tilde{\mathcal{F}}_k] &= \mathbb{E}[\sum_{j=1}^{k+1} \boldsymbol{\delta}_j(i) \mid \tilde{\mathcal{F}}_k] \\ &= \sum_{j=1}^k \boldsymbol{\delta}_j(i) + \mathbb{E}[\mathbf{1}_{(\tilde{\omega}(k+1)=i)} - \mathbb{E}[\mathbf{1}_{(\tilde{\omega}(k+1)=i)} \mid \tilde{\mathcal{F}}_k] \mid \tilde{\mathcal{F}}_k] \\ &= \sum_{j=1}^k \boldsymbol{\delta}_j(i) = \boldsymbol{\Delta}_k(i). \end{aligned} \tag{4.5.9}$$

Therefore, $\boldsymbol{\delta}_k(i) = \boldsymbol{\Delta}_k(i) - \boldsymbol{\Delta}_{k-1}(i)$ can be viewed as the martingale difference sequence. Since $\mathbb{E}[\boldsymbol{\delta}_k(i) \mid \tilde{\mathcal{F}}_{k-1}] = 0$, we have $\mathbb{E}[\boldsymbol{\delta}_k(i)^2 \mid \tilde{\mathcal{F}}_{k-1}] = \mathrm{Var}(\boldsymbol{\delta}_k(i) \mid \tilde{\mathcal{F}}_{k-1}) = \mathrm{Var}(\mathbf{1}_{(\tilde{\omega}(k)=i)} \mid \tilde{\mathcal{F}}_{k-1}) \le \mathbb{E}[\mathbf{1}^2_{(\tilde{\omega}(k)=i)} \mid \tilde{\mathcal{F}}_{k-1}] \le \mathbb{E}[\mathbf{1}_{(\tilde{\omega}(k)=i)} \mid \tilde{\mathcal{F}}_{k-1}] = \mathbb{P}(\tilde{\omega}(k) = i \mid \tilde{\omega}(k-1)) = [\boldsymbol{T}^L]_{\tilde{\omega}(k-1),i}$. When $L \ge t_{\mathrm{MC}}(\pi_{\min}/2)$, then using (4.5.6), we get $[\boldsymbol{T}^L]_{\tilde{\omega}(k-1),i} \le \boldsymbol{\pi}_\infty(i) + \max_{j\in[s]} \|([\boldsymbol{T}^L]_{j,:})^\top - \boldsymbol{\pi}_\infty\|_{\ell_\infty} \le 2\pi_{\max}$. Therefore,

$$\sum_{k=1}^{\tilde{T}} \mathbb{E}[\boldsymbol{\delta}_k(i)^2 \mid \tilde{\mathcal{F}}_{k-1}] \le 2\pi_{\max}\tilde{T}, \tag{4.5.10}$$

where we use the definition $\tilde{T} := \lfloor \frac{T-\ell}{L} \rfloor$. Combining this with the observation that $|\boldsymbol{\delta}_k(i)| < 1$, we have

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{k=1}^{\tilde{T}} \mathbf{1}_{(\tilde{\omega}(k)=i)} - \sum_{k=1}^{\tilde{T}} \mathbb{E}[\mathbf{1}_{(\tilde{\omega}(k)=i)} \mid \tilde{\mathcal{F}}_{k-1}]\right| \ge \frac{\epsilon}{2}\tilde{T}\right) &\overset{\text{(i)}}{=} \mathbb{P}(|\boldsymbol{\Delta}_{\tilde{T}}(i)| \ge \frac{\epsilon}{2}\tilde{T}), \\ &\overset{\text{(ii)}}{\le} \exp(-\frac{\tilde{T}\epsilon^2/8}{2\pi_{\max} + \epsilon/6}), \\ &\overset{\text{(iii)}}{\le} \exp(-\frac{\tilde{T}\epsilon^2}{17\pi_{\max}}),\end{aligned} \tag{4.5.11}$$

where (i) follows from the definition of $\boldsymbol{\Delta}_{\tilde{T}}(i)$, (ii) follows from Freedman's inequality [134], and (iii) follows from picking $\epsilon \le \pi_{\min}/2$. Moreover, when $L \ge t_{\mathrm{MC}}(\pi_{\min}/2)$, we also have

$$\begin{aligned}\left|\sum_{k=1}^{\tilde{T}} \mathbb{E}[\mathbf{1}_{(\tilde{\omega}(k)=i)} \mid \tilde{\mathcal{F}}_{k-1}] - \boldsymbol{\pi}_\infty(i)\tilde{T}\right| &= \left|\sum_{k=1}^{\tilde{T}} \mathbb{P}(\tilde{\omega}(k) = i \mid \tilde{\omega}(k-1)) - \boldsymbol{\pi}_\infty(i)\tilde{T}\right|, \\ &\le \sum_{k=1}^{\tilde{T}} \left|[\boldsymbol{T}^L]_{\tilde{\omega}(k-1),i} - \boldsymbol{\pi}_\infty(i)\right|, \\ &\le \tilde{T} \max_{j\in[s]} \|([\boldsymbol{T}^L]_{j,:})^\top - \boldsymbol{\pi}_\infty\|_\infty, \\ &\le \frac{\pi_{\min}}{2}\tilde{T}.\end{aligned} \tag{4.5.12}$$

Combining (4.5.12) with (4.5.11), and union bounding over $0 \le \ell \le L-1$, we obtain

$$\mathbb{P}\left(\bigcap_{\ell=0}^{L-1}\left\{|S_i^{(\ell)}| \ge \boldsymbol{\pi}_\infty(i)\tilde{T} - \frac{\pi_{\min}}{2}\tilde{T} - \frac{\epsilon}{2}\tilde{T}\right\}\right) \ge 1 - \sum_{\ell=0}^{L-1} \exp(-\frac{\tilde{T}\epsilon^2}{17\pi_{\max}}). \tag{4.5.13}$$

To proceed, define the events $\mathcal{E}_1 := \bigcap_{\ell=0}^{L-1} \left\{|S_i^{(\ell)}| \ge (\pi_{\min}/2 - \epsilon/2)\tilde{T}\right\}$ and $\mathcal{E}_2 := \left\{|S_i| \ge (\pi_{\min}/2 - \epsilon/2)(T-L)\right\}$. Note that $\mathcal{E}_1 \subset \mathcal{E}_2$ because, $|S_i| = \sum_{\ell=0}^{L-1} |S_i^{(\ell)}|$ and $\sum_{\ell=0}^{L-1} \tilde{T} = \sum_{\ell=0}^{L-1} \lfloor \frac{T-\ell}{L} \rfloor = T - L$.

This implies that $\mathbb{P}(\mathcal{E}_2) \geq \mathbb{P}(\mathcal{E}_1)$. Combing this with (4.5.13), and union bounding over all $i \in [s]$, we have

$$\begin{aligned} &\mathbb{P}\bigg(\bigcap_{i=1}^{s}\Big\{|S_i| \geq (\pi_{\min}/2 - \epsilon/2)(T-L)\Big\}\bigg) \geq 1 - sL\exp\Big(-\frac{(T/L-2)\epsilon^2}{17\pi_{\max}}\Big), \\ \implies &\mathbb{P}\bigg(\bigcap_{i=1}^{s}\Big\{|S_i| \geq (\pi_{\min}/4 - \epsilon/4)T\Big\}\bigg) \overset{\text{(i)}}{\geq} 1 - sL\exp\Big(-\frac{(T/L-2)\epsilon^2}{17\pi_{\max}}\Big), \end{aligned} \tag{4.5.14}$$

where (i) follows from choosing $T \geq 2L$. Finally, setting $\delta = sL\exp(-\frac{(T/L-2)\epsilon^2}{17\pi_{\max}})$ and replacing ϵ with $\sqrt{\frac{17\pi_{\max}L\log(sL/\delta)}{T-2L}}$, we obtain the statement of the theorem,

$$\mathbb{P}\bigg(\bigcap_{i=1}^{s}\bigg\{|S_i| \geq \frac{\pi_{\min}T}{4}\bigg(1 - \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}L\log(sL/\delta)}{T-2L}}\bigg)\bigg\}\bigg) \geq 1 - \delta. \tag{4.5.15}$$

This completes the proof. ■

4.5.3 Proof of Theorem 36

Proof. For the sake of completeness, before we present the proof of Theorem 36, we present a meta result from [22] which will be used to prove Theorem 36.

Theorem 37 (Meta-theorem [22]) *Fix $\delta \in (0,1)$, $T \in \mathbb{N}$ and $0 \prec \boldsymbol{\Gamma}_{\text{sb}} \prec \bar{\Gamma}$. Then if $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{t=1}^T \in (\mathbb{R}^d \times \mathbb{R}^n)^T$ is a random sequence such that (a) $\boldsymbol{y}_t = \boldsymbol{A}_\star \boldsymbol{x}_t + \boldsymbol{w}_t$, where $\boldsymbol{w}_t \mid \mathcal{F}_t$ is $\sigma_{\boldsymbol{w}}^2$-subgaussian and mean zero, (b) $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ satisfy the $(k, \boldsymbol{\Gamma}_{\text{sb}}, q)$-small ball condition, and (c) such that $\mathbb{P}\big(\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t^\top \npreceq T\bar{\boldsymbol{\Gamma}}\big) \leq \delta$. Then if*

$$T \geq \frac{10k}{q^2}\big(\log(1/\delta) + 2d\log(10/q) + \log(\det(\bar{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_{\text{sb}}^{-1}))\big),$$

we have

$$\mathbb{P}\bigg(\|\hat{\boldsymbol{A}} - \boldsymbol{A}_\star\| \geq \frac{90\sigma_{\boldsymbol{w}}}{q}\sqrt{\frac{n + d\log(10/q) + \log(\det(\bar{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_{\text{sb}}^{-1})) + \log(1/\delta)}{T\lambda_{\min}(\boldsymbol{\Gamma}_{\text{sb}})}}\bigg) \leq 3\delta.$$

Our proof strategy is to verify that the conditions (a), (b), and (c) of Theorem 37 hold for the MJS in (4.4.22) and then apply Theorem 37 to estimate $(\boldsymbol{A}_{1:s}, \boldsymbol{B}_{1:s})$. Before that, let S_i be as defined in Algorithm 2, that is, $S_i := \{t \mid \omega(t) = i\}$. Then, the samples $\{(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{z}_t, \omega(t))\}_{t \in S_i}$ used to estimate $(\boldsymbol{A}_i, \boldsymbol{B}_i)$ are related as follows,

$$\boldsymbol{x}_{t_k+1} = \boldsymbol{\Theta}_i^\star \boldsymbol{h}_{t_k} + \boldsymbol{w}_{t_k} \quad \text{for} \quad k = 1, 2, \dots, |S_i|, \tag{4.5.16}$$

where we set $S_i := \{t \mid \omega(t) = i\} \equiv \{t_1, t_2, \cdots, t_{|S_i|}\}$, $\boldsymbol{h}_{t_k} := [\boldsymbol{x}_{t_k}^\top / \sigma_{\boldsymbol{w}} \ \ \boldsymbol{z}_{t_k}^\top / \sigma_{\boldsymbol{z}}]^\top$ and $\boldsymbol{\Theta}_i^\star := [\sigma_{\boldsymbol{w}} \boldsymbol{L}_i \ \ \sigma_{\boldsymbol{z}} \boldsymbol{B}_i]$. This shows that, for each $i \in [s]$, the problem of estimating $(\boldsymbol{A}_i, \boldsymbol{B}_i)$ is equivalent to the problem of estimating $\boldsymbol{\Theta}_i^\star$ from the sequence of covariate-response pairs $(\boldsymbol{h}_{t_k}, \boldsymbol{x}_{t_k+1})_{k \geq 1}$. Moreover, let $\{\mathcal{F}_t\}_{t \geq 1}$ denotes the filtration as defined before Theorem 34.

(a) Sub-Gaussian noise: Following the re-parameterization in (4.5.16), the covariate-response pairs $(\boldsymbol{h}_{t_k}, \boldsymbol{x}_{t_k+1})_{k \geq 1}$ are generated from a linear response time series $\boldsymbol{x}_{t_k+1} = \boldsymbol{\Theta}_i^\star \boldsymbol{h}_{t_k} + \boldsymbol{w}_{t_k}$ for $k = 1, 2, \dots, |S_i|$. Moreover, under the Assumption that $\{\boldsymbol{w}_t\}_{t=0}^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$ and that $\boldsymbol{w}_{t_k}$ is $\mathcal{F}_{t_k+1}$-measureable, $\boldsymbol{w}_{t_k} \mid \mathcal{F}_{t_k} \sim \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$.

(b) BMSB condition: Theorem 34 proves that the covariates process $\{\boldsymbol{h}_{t_k}\}_{k=1}^{|S_i|}$ satisfies $(k, \boldsymbol{I}_{n+p}, q)$-BMSB condition, with the constants $k = 1$ and $q = 3/10$.

(c) Covariates correlation bound: Recalling the definition of $\boldsymbol{h}_{t_k}$ and $\boldsymbol{H}_i$ from (4.4.24), we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{H}_i^\top \boldsymbol{H}_i\|] &= \mathbb{E}[\|\sum_{k=1}^{|S_i|} \boldsymbol{h}_{t_k} \boldsymbol{h}_{t_k}^\top\|] \leq \sum_{k=1}^{|S_i|} \mathbb{E}[\|\boldsymbol{h}_{t_k} \boldsymbol{h}_{t_k}^\top\|] \leq \sum_{k=1}^{|S_i|} \mathbb{E}[\|\boldsymbol{h}_{t_k}\|_{\ell_2}^2], \\
&= \sum_{k=1}^{|S_i|} (\mathbb{E}[\|\boldsymbol{x}_{t_k}\|_{\ell_2}^2] / \sigma_{\boldsymbol{w}}^2 + \mathbb{E}[\|\boldsymbol{z}_{t_k}\|_{\ell_2}^2] / \sigma_{\boldsymbol{z}}^2), \\
&\overset{\text{(i)}}{\leq} \sum_{k=1}^{|S_i|} \sqrt{ns} \tau_{\tilde{\boldsymbol{L}}} \big(\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] + \sigma_{\boldsymbol{z}}^2 t_k \sqrt{n} \|\boldsymbol{B}_{1:s}\|^2 + \sigma_{\boldsymbol{w}}^2 t_k \sqrt{n} \big) / \sigma_{\boldsymbol{w}}^2 + \sum_{k=1}^{|S_i|} p, \\
&\leq \sqrt{ns} \tau_{\tilde{\boldsymbol{L}}} \big(\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2] / \sigma_{\boldsymbol{w}}^2 + (\sigma_{\boldsymbol{z}}^2 / \sigma_{\boldsymbol{w}}^2) \sqrt{n} \|\boldsymbol{B}_{1:s}\|^2 T + \sqrt{n} T \big) |S_i| + p |S_i|
\end{aligned} \tag{4.5.17}$$

where we obtain (i) from combining Lemma 31 with Assumption 8 (which says $\rho(\tilde{\boldsymbol{L}}) \le 1$). Hence, setting

$$\bar{\Gamma} := \sqrt{ns}\,\tau_{\tilde{\boldsymbol{L}}}\big(\,\mathbb{E}[\|\boldsymbol{x}_0\|_{\ell_2}^2]/\sigma_{\boldsymbol{w}}^2 + (\sigma_{\boldsymbol{z}}^2/\sigma_{\boldsymbol{w}}^2)\sqrt{n}\|\boldsymbol{B}_{1:s}\|^2 T + \sqrt{n}T\big) + p, \tag{4.5.18}$$

we have, $\mathbb{E}[\|\sum_{k=1}^{|S_i|}\boldsymbol{h}_{t_k}\boldsymbol{h}_{t_k}^\top\|] = \mathbb{E}[\|\boldsymbol{H}_i^\top\boldsymbol{H}_i\|] \le |S_i|\bar{\Gamma}$. Next, we use Markov inequality to show that

$$\begin{aligned}\mathbb{P}\big(\sum_{k=1}^{|S_i|}\boldsymbol{h}_{t_k}\boldsymbol{h}_{t_k}^\top \not\preceq (|S_i|\bar{\Gamma}/\delta)\boldsymbol{I}_{n+p}\big) &= \mathbb{P}\big(\lambda_{\max}(\sum_{k=1}^{|S_i|}\boldsymbol{h}_{t_k}\boldsymbol{h}_{t_k}^\top) \ge |S_i|\bar{\Gamma}/\delta\big),\\ &\le \mathbb{E}\big[\lambda_{\max}(\sum_{k=1}^{|S_i|}\boldsymbol{h}_{t_k}\boldsymbol{h}_{t_k}^\top)\big]\delta/(|S_i|\bar{\Gamma}) \le \delta.\end{aligned} \tag{4.5.19}$$

We are now ready to use Theorem 2.4 from [22] to obtain our final result.

(d) Finalizing the proof: We use Theorem 37, with $\bar{\boldsymbol{\Gamma}} = (\bar{\Gamma}/\delta)\boldsymbol{I}_{n+p}$, $\boldsymbol{\Gamma}_{\text{sb}} = \boldsymbol{I}_{n+p}$, $k = 1$, $q = 3/10$, and $d = n + p$ to upper bound the estimation error (4.4.26) with high probability. Suppose the cardinality of the set $S_i = \{t \mid \omega(t) = i\}$ satisfies,

$$|S_i| \gtrsim (n+p) + \log(3s\bar{\Gamma}/\delta) + \log(3s/\delta), \tag{4.5.20}$$

for each $i \in [s]$. Then, using Theorem 37, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{\|\hat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_i^\star\| \lesssim \sigma_{\boldsymbol{w}}\sqrt{\frac{(n+p) + \log(3s\bar{\Gamma}/\delta) + \log(3s/\delta)}{|S_i|}}\right\}\right) \ge 1 - \delta. \tag{4.5.21}$$

Combining (4.5.21) with Theorem 35, we fix $\delta \in (0,1)$, such that $\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}} \le \pi_{\min}/2$, and choose the trajectory length T satisfying

$$T \gtrsim \max\left\{2T_0, \frac{(n+p) + \log(3s\bar{\Gamma}/\delta) + \log(3s/\delta)}{\pi_{\min}\big(1 - \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}}\big)}\right\}, \tag{4.5.22}$$

then, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{\|\hat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_i^\star\| \lesssim \sigma_{\boldsymbol{w}}\sqrt{\frac{(n+p) + \log(3s\bar{\Gamma}/\delta) + \log(3s/\delta)}{T\pi_{\min}\big(1 - \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}}\big)}}\right\}\right) \ge 1 - 2\delta. \tag{4.5.23}$$

To proceed, using standard result from linear algebra that the spectral norm of a sub-matrix is upper bounded by the spectral norm of the original matrix, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{\|\hat{\boldsymbol{A}}_i - \boldsymbol{A}_i\| \lesssim \frac{(C_K\sigma_{\boldsymbol{w}} + \sigma_{\boldsymbol{z}})}{\sigma_{\boldsymbol{z}}}\sqrt{\frac{(n+p) + \log(3s\bar{\Gamma}/\delta) + \log(3s/\delta)}{T\pi_{\min}\big(1 - \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}}\big)}}\right\}\right) \geq 1 - 2\delta,\\
&\mathbb{P}\left(\bigcap_{i=1}^{s}\left\{\|\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i\| \lesssim \frac{\sigma_{\boldsymbol{w}}}{\sigma_{\boldsymbol{z}}}\sqrt{\frac{(n+p) + \log(3s\bar{\Gamma}/\delta) + \log(3s/\delta)}{T\pi_{\min}\big(1 - \frac{1}{\pi_{\min}}\sqrt{\frac{17\pi_{\max}T_0\log(sT_0/\delta)}{T-2T_0}}\big)}}\right\}\right) \geq 1 - 2\delta,
\end{aligned} \tag{4.5.24}$$

where we used the relation $\|\hat{\boldsymbol{A}}_i - \boldsymbol{A}_i\| \leq \|\hat{\boldsymbol{L}}_i - \boldsymbol{L}_i\| + \|\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i\|\|\boldsymbol{K}_i\|$ and $\|\boldsymbol{K}_i\| \leq C_K$ to upper bound the estimation error of the state matrices $\{\boldsymbol{A}_i\}_{i=1}^s$. Finally, replacing δ with $\delta/2$, we get the statement of the theorem. This completes the proof. ■

Chapter 5

Finding Best Linear Model

5.1 Introduction

Supervised learning is concerned with finding a relation between the input-output pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$. The simplest relations are linear functions where the output y_i is estimated by a linear function of the input, that is, $\hat{y}_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle$. Using quadratic loss, we can find the optimal $\boldsymbol{\theta}$ with a simple linear regression which minimizes

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)^2. \tag{5.1.1}$$

If the samples are i.i.d. and input has identity covariance, the minimizer of the population loss ($n \to \infty$) is simply given by

$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[\mathcal{L}(\boldsymbol{\theta})] = \mathbb{E}[y\boldsymbol{x}]. \tag{5.1.2}$$

where $(\boldsymbol{x}, y)$ is drawn from same distribution as data. We will refer to this population minimizer as the best linear model (BLM). In many applications, we operate in the high-dimensional regime where we have fewer samples than the parameter dimension i.e. $n \ll p$. In this case, the problem is ill-posed; however, if $\boldsymbol{\theta}_\star$ lies on a low-dimensional manifold, we can take advantage of this information to solve the problem. We assume $\boldsymbol{\theta}_\star$ is structured-sparse, for instance, it can be a signal that is sparse in a dictionary or it can be a low-rank matrix. If $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}$ is a regularization function, that promotes this structure, we can solve the constrained empirical risk minimization (ERM)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{\theta}) \le R. \tag{5.1.3}$$

where $\boldsymbol{y} = [y_1 \ \dots \ y_n]^\top \in \mathbb{R}^n$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$ are the output labels and data matrix respectively. This problem is well-studied in the statistics and compressed sensing (CS) literature. However, much of the theory literature is concerned with the scenario where the problem is realizable i.e. the outputs are explicitly generated with respect to some ground truth vector $\boldsymbol{a}$. In the simplest scenario, input/output relation can be $y = \langle \boldsymbol{x}, \boldsymbol{a}\rangle + z$ where z is an independent zero-mean noise variable. In this case, one simply has $\boldsymbol{\theta}_\star = \boldsymbol{a}$. Realizability also appears in the literature on semi-parametric single-index models [135–137] where the conditional expectation satisfies $\mathbb{E}[y \mid \boldsymbol{x}] = g(\boldsymbol{x}^\top \boldsymbol{a})$ for some $\boldsymbol{a}$. Interestingly, as discussed in (5.3.3), these works often assume problem setups to ensure BLM coincides with the ground truth parameter $\boldsymbol{a}$. We remark that the realizability assumption is typically more suitable for signal processing applications where the task is reconstructing a ground truth signal or image. In contrast, machine learning (ML) aims to find a model minimizing the test error however exact model parameters are not the primary concern. Our work is closer to ML and analyzes constrained ERM problem (5.1.3) while circumventing realizability issue.

While linear models find ubiquitous use due to their simplicity and interpretability, their performance might be non-ideal if the input/output relation is highly nonlinear. In these instances, linear models can be used as building blocks for more complex machine learning models by employing boosting [138]. Essentially, after learning the BLM, one can fit a more complex model on the residual to further capture nonlinearity. This approach has the potential to improve the model interpretability [139] and it can also reduce the sample complexity required for fitting the more complex model thanks to the reduced residual variance [140–142]. Indeed, residual learning is very popular in deep learning applications thanks to the success of residual networks [143]. We also remark that problem might be fully nonlinear and (5.1.3) might be non-informative. A classic example is quadratic dependence (e.g. phase retrieval problem) where $y = |\boldsymbol{a}^\top \boldsymbol{x}|$ so that $\mathbb{E}[y\boldsymbol{x}] = 0$ for normally distributed inputs. Finally, we remark that even if BLM estimator may fail to achieve small population loss (test error) single-handedly, it can be used for determining useful input features which is critical for interpretability. In the small sample regime, this is facilitated by using ℓ_1 or sparsity constraints.

Bias in the data can negatively affect the estimation quality. Assuming input is zero-mean, instead of solving (5.1.3) we can solve a modified problem which accounts for the mean of the output as well. Again, denoting the regularization function by $\mathcal{R}$, we shall consider the intercept-enabled problem

$$\hat{\boldsymbol{\theta}}, \hat{\mu} = \arg\min_{\boldsymbol{\theta},\mu} \mathcal{L}(\boldsymbol{\theta}, \mu) \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{\theta}) \le R. \tag{5.1.4}$$

where the loss is given by $\mathcal{L}(\boldsymbol{\theta}, \mu) = \frac{1}{2}\left\|\boldsymbol{y} - [\boldsymbol{X}\ \ \boldsymbol{1}]\begin{bmatrix}\boldsymbol{\theta}\\ \mu\end{bmatrix}\right\|_{\ell_2}^2$. We will show that solving problem (5.1.4) is essentially equivalent to solving (5.1.3) with debiased output hence it will result in more accurate estimation. The goal of this chapter is studying problem (5.1.4) under a general algorithmic framework, establishing finite-sample statistical and algorithmic convergence, and addressing practical considerations on the data distribution. In particular, we are interested in how well one can estimate the best linear model (BLM) given by the pair $(\boldsymbol{\theta}_\star = \mathbb{E}[y\boldsymbol{x}], \mu_\star = \mathbb{E}[y])$. For estimation, we will utilize the projected gradient descent algorithm given by the iterates

$$\begin{aligned}\boldsymbol{\theta}_{\tau+1} &= \mathcal{P}_{\mathcal{K}}(\boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_\tau, \mu_\tau)),\\ \mu_{\tau+1} &= \mu_\tau - \eta\nabla\mathcal{L}_\mu(\boldsymbol{\theta}_\tau, \mu_\tau),\end{aligned} \tag{5.1.5}$$

where $\mathcal{P}_{\mathcal{K}}$ projects onto the constraint set $\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^p \mid \mathcal{R}(\boldsymbol{\theta}) \le R\}$ and η is the step size.

5.1.1 Relation to Prior Work

There is a significant amount of literature on nonlinear (or one-bit) CS [136,144–153]. [145, 154–157] study algorithmic and statistical convergence rates for first order methods such as projected/proximal gradient descent. For nonlinear CS, [145, 146, 148, 158] provide statistical analysis of single index estimation with a focus on Gaussian data. Recently, one-bit CS techniques have been extended to subgaussian distributions using dithering trick which adds noise before quantization [137, 159–161]. Dithering is introduced to guarantee consistent estimation of the ground-truth parameter. The papers [162–166] address non-gaussianity by utilizing Stein identity which requires access to the distribution of the input samples. Closer to us [167] studies the constrained empirical risk minimization with linear functions and squared loss with a focus on convex problems. In comparison our analysis applies to a broader class of distributions and focus on first order algorithms. Much of our analysis focuses on addressing subexponential samples, which requires tools from high-dimensional probability [86, 168, 169]. [170] similarly studies high-dimensional estimation with subexponential design matrix for a planted linear model where $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{w}$. In contrast to [170], we consider the more general setup of (5.1.3) which allows for arbitrary input/output relations and explore the properties of gradient descent rather than convex programs.

Our results apply to general regularizers and borrow ideas from [145–148]. Similar to these, we view the nonlinearity between input and output as an additive noise. The convergence analysis of projected gradient descent is a rather well-understood topic and we utilize insights from [154–157] for our analysis.

5.1.2 Contributions

At a high-level our work has three distinguishing features:

• **Projected gradient descent to find BLM:** Nonlinear CS literature is typically concerned with a ground-truth vector to be recovered. For instance, one-bit CS aims to learn $\boldsymbol{\theta}$ from samples of type $y = \text{sgn}(\boldsymbol{a}^\top \boldsymbol{x})$. Unlike these, our approach applies to arbitrary input / outputs with subexponential tails, hence the results apply under much weaker assumptions. For instance, closely related work [145] analyzes PGD for nonlinear compressed sensing however their results are only valid for normally distributed inputs.

• **Subexponential samples:** Most nonlinear CS results apply to Gaussian or subgaussian data when dithering trick is utilized [137, 159–161]. We take advantage of the recent techniques for subexponential distributions to provide statistical/computational guarantees for heavier-tailed distributions.

• **Analyzing the intercept-enabled design matrix:** Intercept term is commonly used in regression analysis to estimate the output bias [171]. We analyze the intercept-enabled problem (5.1.4) by studying the statistical properties of the concatenated design matrix $[\boldsymbol{X}\ \mathbf{1}]$. Empirically this modification leads to a substantial performance improvement when labels are not zero-mean.

5.2 Preliminaries and Problem Setup

In this section we introduce statistical quantities which are utilized to characterize the benefits of the regularization $\mathcal{R}$.

Suppose we are given n i.i.d. samples $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim (\boldsymbol{x}, y)$. To keep the exposition clean, we assume that $\boldsymbol{x}$ is whitened, that is, it has zero-mean and identity covariance. Our goal will be finding a linear relation between the modified input-output pairs $([\boldsymbol{x}_i^\top\ 1]^\top, y_i)_{i=1}^n$. In the population limit, optimal model parameters are given by

$$\boldsymbol{\theta}_\star, \mu_\star = \arg\min_{\boldsymbol{\theta},\mu} \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \mu)] = \mathbb{E}[y\boldsymbol{x}], \mathbb{E}[y].$$

Thus, in the limiting case, $\mu_\star$ captures the mean of the output and $\boldsymbol{\theta}_\star$ is the population minimizer of $\mathcal{L}(\boldsymbol{\theta})$. Our goal is estimating the population minimizers $\boldsymbol{\theta}_\star, \mu_\star$ using finite samples $(\boldsymbol{x}_i, y_i)_{i=1}^n$. As discussed in Section 5.1, assuming $\boldsymbol{\theta}_\star$ is structured sparse, we consider a non-asymptotic estimation of $\boldsymbol{\theta}_\star, \mu_\star$ via problem (5.1.4). To proceed with analysis, we assume $\mathcal{R}$ is a proper function (i.e. closed sub-level sets) and set

$$\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^p \mid \mathcal{R}(\boldsymbol{\theta}) \le R\}, \tag{5.2.1}$$

$$\mathcal{K}_{\text{ext}} = \{[\boldsymbol{\theta}^\top\ \mu]^\top \in \mathbb{R}^{p+1} \mid \mathcal{R}(\boldsymbol{\theta}) \le R\}. \tag{5.2.2}$$

We investigate the PGD algorithm (5.1.5) which can be written as

$$\begin{bmatrix}\boldsymbol{\theta}_{\tau+1}\\ \mu_{\tau+1}\end{bmatrix} = \mathcal{P}_{\mathcal{K}_{\text{ext}}}\left(\begin{bmatrix}\boldsymbol{\theta}_{\tau}\\ \mu_{\tau}\end{bmatrix} + \eta[\boldsymbol{X}\ \mathbf{1}]^{\top}\left(\boldsymbol{y} - [\boldsymbol{X}\ \mathbf{1}]\begin{bmatrix}\boldsymbol{\theta}_{\tau}\\ \mu_{\tau}\end{bmatrix}\right)\right), \tag{5.2.3}$$

where η is a fixed learning rate and $[\boldsymbol{X}\ \mathbf{1}] \in \mathbb{R}^{n\times(p+1)}$ is the intercept-enabled design matrix constructed as follows

$$[\boldsymbol{X}\ \mathbf{1}] = \begin{bmatrix}\boldsymbol{x}_1^{\top}\ 1\\ \vdots\\ \boldsymbol{x}_n^{\top}\ 1\end{bmatrix}. \tag{5.2.4}$$

Following [145, 172] PGD analysis can be related to the *tangent ball* around the population parameter $\boldsymbol{\theta}_\star$ which is given by

$$\mathcal{C} = \text{cl}(\{\alpha\boldsymbol{v} \mid \boldsymbol{v} + \boldsymbol{\theta}_\star \in \mathcal{K},\ \alpha \ge 0\}) \bigcap \mathcal{B}^p. \tag{5.2.5}$$

Similarly, we define the *extended tangent ball* as follows

$$\mathcal{C}_{\text{ext}} = \left\{\begin{bmatrix}\alpha\boldsymbol{v}\\ \gamma\end{bmatrix} \mid \alpha \ge 0,\ \boldsymbol{v} \in \mathcal{C},\ \gamma \in \mathbb{R}\right\} \bigcap \mathcal{B}^{p+1}. \tag{5.2.6}$$

We remark that our (extended) tangent ball definition is the intersection of the (extended) tangent cone with the unit Euclidian ball. While related literature mostly uses tangent cone [172, 173], we introduce the tangent ball for notational convenience.

The two definitions above ($\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$) are closely related. For any vector $\boldsymbol{v} \in \mathcal{C}$ and scalar $|\gamma| \le 1$, we have that $[\sqrt{1-\gamma^2}\boldsymbol{v}^{\top}\ \gamma]^{\top} \in \mathcal{C}_{\text{ext}}$. In the following we will express the convergence rates and residual errors of the PGD algorithm (5.1.5) in terms of the statistical properties of the tangent balls.

- **Technical approach:** To keep the discussion focused, throughout we assume that R is correctly specified i.e., $R = \mathcal{R}(\boldsymbol{\theta}_\star)$. Denoting the parameter estimation error in (5.2.3) by $\boldsymbol{h}_\tau = [\boldsymbol{\theta}_\tau{}^\top\ \mu_\tau]^\top - [\boldsymbol{\theta}_\star{}^\top\ \mu_\star]^\top$, and the effective noise by $\boldsymbol{w} = \boldsymbol{y} - [\boldsymbol{X}\ \mathbf{1}][\boldsymbol{\theta}_\star{}^\top\ \mu_\star]^\top$, the PGD update can be shown to obey [155] (see Eq. (VI.10))

$$\|\boldsymbol{h}_{\tau+1}\|_{\ell_2} \le \kappa\left(\|\boldsymbol{h}_\tau\|_{\ell_2}\rho(\mathcal{C}) + \eta\nu(\mathcal{C})\right) \tag{5.2.7}$$

where κ is a numerical constant which is equal to 1 for convex regularizer $\mathcal{R}$ and 2 for arbitrary $\mathcal{R}$ and

$$\rho(\mathcal{C}) = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}_{\text{ext}}} |\boldsymbol{u}^\top(\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\boldsymbol{v}|, \tag{5.2.8}$$

$$\nu(\mathcal{C}) = \sup_{\boldsymbol{v}\in\mathcal{C}_{\text{ext}}} |\boldsymbol{v}^\top[\boldsymbol{X}\ \mathbf{1}]^\top\boldsymbol{w}|. \tag{5.2.9}$$

Here ρ captures the algorithmic convergence and ν captures the statistical accuracy in terms of regularization. To achieve statistical learning bounds, we need to characterize the quantities above in finite sample. Existing literature provides a fairly good understanding of the related terms when $\boldsymbol{X}$ has subgaussian rows or $\boldsymbol{w}$ is independent of $\boldsymbol{X}$. The technical contributions of this work are i) extending these results to subexponential samples, ii) allowing for nonlinear dependencies between the noise and data, and iii) addressing the bias term by studying the concatenated matrix $[\boldsymbol{X}\ \mathbf{1}]$. To proceed with statistical analysis, we introduce Gaussian width.

Definition 38 ((Perturbed) Gaussian width [86]) *The Gaussian width of a set $T \subset \mathcal{B}^p$ is defined as*

$$\omega(T) = \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_p)}[\sup_{\boldsymbol{v}\in T} \boldsymbol{v}^\top\boldsymbol{g}]. \tag{5.2.10}$$

Let $C \geq 1$ be an absolute constant. Given an integer $n \geq 1$, the perturbed Gaussian width $\omega_n(T)$ of $T \subset \mathcal{B}^d$ is defined as

$$\omega_n(T) = \min_{\substack{clconv(S) \supseteq T \\ rad(S) \leq C}} \omega(S) + \frac{\gamma_1(S)}{\sqrt{n}} \tag{5.2.11}$$

where $\gamma_1(S)$ is Talagrand's γ_1-functional (see [168]) with ℓ_2-metric. Note that one can always choose $S = T$.

Gaussian width helps to quantify the complexity of the regularized problem and determines the sample complexity of the linear inverse problems i.e. high-dimensional problems become manageable in the regime $n \gtrsim \omega^2(\mathcal{C})$ [172, 173]. Perturbed width is introduced more recently in [86] to address subexponential samples. [86] shows that, for standard regularizers such as ℓ_0, ℓ_1, subspace, and rank constraints, one has

$$\omega^2(\mathcal{C}) \sim \omega_n^2(\mathcal{C}) \tag{5.2.12}$$

in the interesting regime $n \geq \omega^2(\mathcal{C})$. For these regularizes, perturbed width leads to a similar statistical accuracy as Gaussian width but also applies to subexponential samples. For general sets $\mathcal{C}$, the ratio $\omega_n(\mathcal{C})/\omega(\mathcal{C})$ may be large however it can be upper bounded by using $\gamma_1(\mathcal{C}) \lesssim \omega(\mathcal{C})\sqrt{p \log p}$.

As illustrated in Table 5.1, square of the Gaussian width captures the degrees of freedom for practical regularizers [172, 173]. Table 5.1 is obtained by setting $R = \mathcal{R}(\boldsymbol{\theta}_\star)$ in (5.2.1). In practice, a good choice for R can be found by using cross validation or based on the characteristics of data (e.g. [174]). We remark that setting $R > \mathcal{R}(\boldsymbol{\theta}_\star)$, leads to a large tangent ball, specifically $\mathcal{C} = \mathcal{B}^p$. This can be addressed by using the fact that PGD output is robust to the choice of R around $\mathcal{R}(\boldsymbol{\theta}_\star)$ (see Theorem 2.6 of [155]). Alternatively, one can utilize the proximal variation which solves the regularized problem $\min_{\boldsymbol{\theta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2+\lambda\mathcal{R}(\boldsymbol{\theta})$. To keep our discussion focused, we (implicitly) assume R is correctly specified throughout. The next statistical quantity required in our analysis is the Orlicz norm defined as follows.

Definition 39 (Orlicz norms) *For a scalar random variable Orlicz-a norm is defined as*

$$\|X\|_{\psi_a}=\sup_{p\geq 1} p^{-1/a}(\mathbb{E}[|X|^p])^{1/p}$$

Orlicz-a norm of a vector $\boldsymbol{x}\in\mathbb{R}^d$ *is defined as* $\|\boldsymbol{x}\|_{\psi_a}=\sup_{\boldsymbol{v}\in\mathcal{B}^d}\|\boldsymbol{v}^\top\boldsymbol{x}\|_{\psi_a}$. *Subexponential and subgaussian norms are special cases of Orlicz-a norm given by* $\|\cdot\|_{\psi_1}$ *and* $\|\cdot\|_{\psi_2}$ *respectively.*

Based on perturbed Gaussian width definition, we will show that one can upper bound the critical quantities (5.2.8) and (5.2.9). In return, this will reveal the statistical and computational performance of the PGD algorithm. This is the topic of the next section which states our main results.

Constraint	**Parameter vector model**	$\boldsymbol{\omega^2(\mathcal{C})}$
None	$\boldsymbol{\theta}_\star\in\mathbb{R}^p$	p
Sparsity $\|\cdot\|_{\ell_0}$	s non-zero entries	$s\log(6p/s)$
ℓ_1 norm $\|\cdot\|_{\ell_1}$	s non-zero entries	$s\log(6p/s)$
Subspace	$\boldsymbol{\theta}_\star\in\mathcal{S},\ \dim(\mathcal{S})=k$	k
Matrix rank	$\text{rank}(\text{mtx}((\boldsymbol{\theta}_\star)))\leq r$	$rp^{1/2}$

Table 5.1: List of low-dimensional models and corresponding Gaussian widths (up to a constant factor) for the constraint sets $\mathcal{K}=\{\boldsymbol{\theta}\mid\mathcal{R}(\boldsymbol{\theta})\leq\mathcal{R}(\boldsymbol{\theta}_\star)\}$. If constraint is set membership such as subspace, $\mathcal{R}(\boldsymbol{\theta})=0$ inside the set and ∞ outside. Furthermore, we represent the vector $\boldsymbol{\theta}_\star\in\mathbb{R}^p$ in matrix form as $\text{mtx}(\boldsymbol{\theta}_\star)\in\mathbb{R}^{\sqrt{p}\times\sqrt{p}}$.

5.3 Main Results

In this section we estimate the convergence rate and the statistical accuracy of the PGD algorithm as a function of sample size, complexity of the parameter (e.g. sparsity level), and the distribution of the data (whether subgaussian or subexponential). Our main theorem establishes a linear convergence rate of PGD and shows that PGD achieves statistically efficient error rates. We first describe the data model.

Definition 40 (Isotropic vector) *$\boldsymbol{x} \in \mathbb{R}^p$ is called an isotropic Orlicz-a vector if it is zero-mean with identity covariance and if its Orlicz-a norm $\|\boldsymbol{x}\|_{\psi_a}$ is bounded by an absolute constant.*

Definition 41 (σ-noisy datasets) *We call a dataset $\mathcal{S}$ with i.i.d. samples $(\boldsymbol{x}_i, y_i)_{i=1}^n$ σ-Orlicz-a if for all pairs $(\boldsymbol{x}, y) \in \mathcal{S}$, the input $\boldsymbol{x}$ is isotropic Orlicz-a vectors and the residual at the BLM obeys $\|y - \boldsymbol{x}^\top \boldsymbol{\theta}_\star - \mu_\star\|_{\psi_a} \le \sigma$.*

We call σ-Orlicz-1 dataset σ-subexponential and σ-Orlicz-2 dataset σ-subgaussian.

Note that the residual at the BLM corresponds to the noise in our problem which may may depend on the input in a nonlinear fashion. If we solve (5.1.3) rather than (5.1.4), the noise term σ will essentially grow to $\sigma + \mu_\star$ since zero-mean input features $\boldsymbol{x}$ cannot explain the label mean. This highlights the advantage of (5.1.4). Our main results capture the PGD performance for different dataset models described below.

Theorem 42 (Subgaussian) *Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subgaussian dataset. Assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$ and set learning rate $\eta = 1/n$. Let $\mathcal{R}$ be an arbitrary regularizer. Starting from an initial estimate $[\boldsymbol{\theta}_0^\top \; \mu_0]^\top$ obeying $\mathcal{R}(\boldsymbol{\theta}_0) \le R$, with probability at least $1 - 6\exp(-c_0 t^2/2) -$*

$4n^{-100}$, *all PGD iterates* (5.2.3) *obey*

$$\left\|\begin{bmatrix}\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\star\\ \\ \mu_\tau-\mu_\star\end{bmatrix}\right\|_{\ell_2}\leq\ (c\frac{\omega(\mathcal{C})+t}{\sqrt{n}})^\tau\left\|\begin{bmatrix}\boldsymbol{\theta}_0-\boldsymbol{\theta}_\star\\ \\ \mu_0-\mu_\star\end{bmatrix}\right\|_{\ell_2}+C\sigma\frac{(\omega(\mathcal{C})+t)\sqrt{\log(n)}}{\sqrt{n}}. \tag{5.3.1}$$

We remark that $\mathcal{R}(\boldsymbol{\theta}_0)\leq R$ is not a major assumption since one can first project $\boldsymbol{\theta}_0$ to the constrained set before starting PGD. For subexponential samples, we have the following theorem which applies to convex regularizers.

Theorem 43 (Subexponential) *Suppose* $(\boldsymbol{x}_i,y_i)_{i=1}^n$ *is a* σ*-subexponential dataset. Set* $q=(n+p)\log^3(n+p)$*. Set learning rate* $\eta=c_0/q$*, suppose* $\mathcal{R}$ *is convex and* $n\gtrsim(\omega_n(\mathcal{C})+t)^2$*. Starting from initialization* $[\boldsymbol{\theta}_0^\top\ \mu_0]^\top$ *satisfying* $\mathcal{R}(\boldsymbol{\theta}_0)\leq R$*, with probability at least* $1-9\exp(-c_0\min(n,t\sqrt{n},t^2))-5(n+p)^{-100}$*, all PGD iterates* (5.2.3) *obey*

$$\left\|\begin{bmatrix}\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\star\\ \\ \mu_\tau-\mu_\star\end{bmatrix}\right\|_{\ell_2}\leq\left(1-\frac{cn}{q}\right)^\tau\left\|\begin{bmatrix}\boldsymbol{\theta}_0-\boldsymbol{\theta}_\star\\ \\ \mu_0-\mu_\star\end{bmatrix}\right\|_{\ell_2}+C\sigma\frac{(\omega_n(\mathcal{C})+t)\log(n)}{\sqrt{n}}. \tag{5.3.2}$$

Both of these results show that PGD iterates converge to population parameters $\boldsymbol{\theta}_\star$, $\mu_\star$ at a linear rate. The subexponential theorem requires a more conservative choice of learning rate. The statistical estimation error grows as $\omega(\mathcal{C})/\sqrt{n}$ for subgaussian and $\omega_n(\mathcal{C})/\sqrt{n}$ for subexponential. Since our results apply in the regime $n\gtrsim\omega^2(\mathcal{C})$, following (5.2.12), statistical errors associated with subgaussian and subexponential are same up to a constant for typical regularizers.

Our main results follow from Theorems 44 and 45 which are the topics of the following sections.

5.3.1 Controlling the Convergence Rate of PGD

In this section, we study the convergence rate characterized by the $\rho(\mathcal{C})$ term. The challenges we address are (i) characterizing the restricted singular values of the subexponential data matrices and (ii) addressing the concatenated all ones vector.

Theorem 44 (Convergence rate) *Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subgaussian dataset and $[\boldsymbol{X}\ \mathbf{1}]$ is the intercept-enabled design matrix, where $\mathbf{1}$ is a vector of all ones. Let $\mathcal{C}$ and $\mathcal{C}_{ext}$ be the tangent balls as defined in (5.2.5) and (5.2.6) respectively. Assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$. Setting $\eta = 1/n$, with probability at least $1 - 4e^{-t^2}$ we have*

$$\rho(\mathcal{C}) \lesssim \frac{\omega(\mathcal{C}) + t}{\sqrt{n}}. \tag{5.3.3}$$

If the dataset is σ-subexponential, then setting $\eta = c_0/(n+p)\log^3(n+p)$ and assuming $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$, with probability $1 - 5\exp(-c\min(n, t\sqrt{n}, t^2)) - 3(n+p)^{-100}$, we have

$$\rho(\mathcal{C}) \leq 1 - C_0 \eta n. \tag{5.3.4}$$

Note that, subexponential requires a smaller choice of learning rate which results in slower convergence.

5.3.2 Bounding the Error due to Nonlinearity

Next, we provide a bound on the effective noise level $\nu(\mathcal{C})$; which is crucial for assessing statistical accuracy. This term arises from the nonlinearity and noise associated with the relation between input and output. For example, for single-index models, we have $\mathbb{E}[y \mid \boldsymbol{x}] = \phi(\boldsymbol{x}^\top \boldsymbol{\theta}_{\text{GT}})$ for some link function ϕ and ground truth $\boldsymbol{\theta}_{\text{GT}}$, and ϕ becomes the source of the nonlinearity. Our approach is similar to [145–148, 167] and treats the nonlinearity as a noise. The finite sample noise is captured by the residual vector

$$\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\star - \mathbf{1}\mu_\star. \tag{5.3.5}$$

Following the $\nu(\mathcal{C})$ term in (5.2.9), The contribution of the residual $\boldsymbol{w}$ to the estimated parameter is captured by the vector

$$\mathrm{e} = [\boldsymbol{X}\ \mathbf{1}]^\top \boldsymbol{w} = \sum_{i=1}^{n}(y_i - \mu_\star - \boldsymbol{x}_i^\top \boldsymbol{\theta}_\star)\begin{bmatrix}\boldsymbol{x}_i \\ 1\end{bmatrix}. \tag{5.3.6}$$

Our key observation is that the properties of e can be characterized under fairly general assumptions compared to the existing literature; which is mostly restricted to zero-mean subgaussian samples.

Theorem 45 (Statistical error) *Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim (\boldsymbol{x}, y)$ is a σ-subgaussian dataset. Let the tangent balls $\mathcal{C}$ and $\mathcal{C}_{ext}$ be as defined in (5.2.5) and (5.2.6) respectively. Assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$. Then, with probability at least $1 - 2\exp(-t^2/2) - 4n^{-100}$, we have*

$$\frac{\nu(\mathcal{C})}{n} \lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}}. \tag{5.3.7}$$

where $\nu(\mathcal{C})$ is the effective noise given by (5.2.9). If $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subexponential dataset and $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$, with probability at least $1 - 4\exp(-c\ \min(t\sqrt{n}, t^2)) - 2n^{-100}$, we have

$$\frac{\nu(\mathcal{C})}{n} \lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}}. \tag{5.3.8}$$

This theorem establishes the crucial finite sample upper bounds on $\nu(\mathcal{C})$ for both subgaussian and subexponential data as a function of Gaussian width of the tangent ball. Combining our bounds on $\rho(\mathcal{C})$ and $\nu(\mathcal{C})$ and utilizing the recursion (5.2.7), we can obtain the PGD convergence characteristics and prove the main theorems.

5.3.3 Discussion on Realizability

Consider a single-index model where our dataset satisfies $y = \phi(\boldsymbol{a}^\top \boldsymbol{x})$. For simplicity assume $\|\boldsymbol{a}\|_{\ell_2} = 1$. As mentioned in the introduction, if we wish to recover $\boldsymbol{a}$, it would be ideal to ensure BLM $\boldsymbol{\theta}_\star$ corresponds to $\boldsymbol{a}$. Below we highlight the two established ways of achieving this [135–137, 146, 159–161].

- **Gaussianity assumption:** Suppose $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_p)$. In this case, we use the independence of orthogonal projections of $\boldsymbol{x}$. Specifically, $g := \boldsymbol{a}^\top \boldsymbol{x} \sim \mathcal{N}(0,1)$ is independent of $\boldsymbol{x} - \boldsymbol{a}\boldsymbol{a}^\top \boldsymbol{x}$. This yields

$$\mathbb{E}[y\boldsymbol{x}] = \mathbb{E}[\phi(\boldsymbol{a}^\top \boldsymbol{x})\boldsymbol{x}] = \boldsymbol{a}\,\mathbb{E}[\phi(\boldsymbol{a}^\top \boldsymbol{x})\boldsymbol{a}^\top \boldsymbol{x}] = \boldsymbol{a}\,\mathbb{E}[g\phi(g)] \tag{5.3.9}$$

  Hence $\boldsymbol{\theta}_\star$ is related to $\boldsymbol{a}$ by a simple scaling of $\mathbb{E}[g\phi(g)]$.

- **Dithering** can be used in quantization to prevent the bias in the quantization error. Suppose the quantization function ϕ rounds its input to the nearest discrete level $(c\delta)_{c=-\infty}^{\infty}$. We can apply dithering during signal acquisition via $y = \phi(\boldsymbol{a}^\top \boldsymbol{x} + w)$ where w is independent and uniformly distributed over $[-\delta/2, \delta/2]$. The application of dithering guarantees that BLM is an unbiased estimate of $\boldsymbol{a}$ by noticing $\mathbb{E}_w[\phi(c+w)] = c$. This gives

$$\mathbb{E}[y\boldsymbol{x}] = \mathbb{E}_{w,\boldsymbol{x}}[\phi(\boldsymbol{a}^\top \boldsymbol{x} + w)\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{a}\boldsymbol{x}^\top \boldsymbol{x}] = \boldsymbol{a}. \tag{5.3.10}$$

5.4 Proofs of the Main Results

This section proves our main results and outlines the proofs of Theorems 42, 43, 44 and 45. Throughout, we use the same notation as described in Section 5.2.

5.4.1 Proof of Theorem 43

Proof. We provide our analysis for subexponential samples. The extension to subgaussian samples is accomplished in an identical fashion. Set the estimation error at iteration τ to be $\boldsymbol{h}_\tau = [\boldsymbol{\theta}_\tau{}^\top \ \mu_\tau]^\top - [\boldsymbol{\theta}_\star{}^\top \ \mu_\star]^\top$. Note that, when $\rho(\mathcal{C}) < 1$ and $\mathcal{R}$ is a convex regularizer, then the recursion (5.2.7) can be iteratively expanded as

$$\begin{aligned}\|\boldsymbol{h}_\tau\|_{\ell_2} &\leq \|\boldsymbol{h}_0\|_{\ell_2}\rho(\mathcal{C})^\tau + \eta\nu(\mathcal{C})\sum_{k=0}^{\tau-1}\rho(\mathcal{C})^k \\ &\leq \|\boldsymbol{h}_0\|_{\ell_2}\rho(\mathcal{C})^\tau + \frac{\eta\nu(\mathcal{C})}{1-\rho(\mathcal{C})}. \end{aligned} \tag{5.4.1}$$

With the advertised probability, subexponential statements of Theorems 44 and 45 hold. Hence, for some constants, we have that $\rho(\mathcal{C}) \leq 1 - c_0\eta n$, $\nu(\mathcal{C}) \leq C\sqrt{n}\sigma(\omega_n(\mathcal{C}) + t)\log(n)$ and $\eta = c/q$ with $q = (n+p)\log^3(n+p)$. Plugging these in (5.4.1), we find the following upper bound on the right hand side,

$$\begin{aligned}\|\boldsymbol{h}_{\tau+1}\|_{\ell_2} &\leq (1-c_0\eta n)^\tau\|\boldsymbol{h}_0\|_{\ell_2} + \frac{\eta}{c_0\eta n}C\sqrt{n}\sigma(\omega_n(\mathcal{C}) + t)\log(n) \\ &= (1-\frac{c_0cn}{q})^\tau\|\boldsymbol{h}_0\|_{\ell_2} + \sigma\frac{C}{c_0}\frac{(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}}, \end{aligned} \tag{5.4.2}$$

which is the desired bound. The case of subgaussian samples is again a corollary of Theorems 44 and 45. This concludes the proof of our main result. ■

5.4.2 Proof of Theorem 44 for subgaussian samples

Proof. We start our proof with the following lemma.

Lemma 46 *Let $(\boldsymbol{x}_i)_{i=1}^n \sim \boldsymbol{x} \in \mathbb{R}^p$ be i.i.d. isotropic subgaussian samples. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be concatenated data and $[\boldsymbol{X} \ \mathbf{1}]$ is the intercept-enabled design matrix, where $\mathbf{1}$ is a vector of* *all ones. Let $\mathcal{T}$ be a closed set with Euclidian radius bounded by a constant and*

$$\mathcal{T}_{ext} = \{\tilde{\boldsymbol{v}} \mid \tilde{\boldsymbol{v}} = [\beta \boldsymbol{v}^\top \ \gamma]^\top \quad \textit{where} \quad \boldsymbol{v} \in \mathcal{T},\ |\beta| \leq C_1,\ |\gamma| \leq C_2\}. \tag{5.4.3}$$

for some positive constants C_1, C_2. Assume $n \gtrsim (\omega(\mathcal{T}) + t)^2$. Then, with probability at least $1 - 2e^{-t^2}$ we have

$$\sup_{\tilde{\boldsymbol{v}} \in \mathcal{T}_{ext}} |\tilde{\boldsymbol{v}}^\top (\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \mathbf{1}]^\top [\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}}. \tag{5.4.4}$$

The proof of Lemma 46 is deferred to Section C.1. To proceed, we apply the result of Lemma 46 over the sets $\mathcal{T}_{\text{ext}} = \mathcal{C}_{\text{ext}} - \mathcal{C}_{\text{ext}}$ and $\mathcal{T}_{\text{ext}} = \mathcal{C}_{\text{ext}} + \mathcal{C}_{\text{ext}}$ to control $\boldsymbol{u} - \boldsymbol{v}$ and $\boldsymbol{u} + \boldsymbol{v}$ vectors. Controlling these helps us bound the cross-product $|\boldsymbol{u}^\top (\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \mathbf{1}]^\top [\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}|$. In Section C.2, we use this argument to show that with the desired probability we have

$$\sup_{\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}} |\tilde{\boldsymbol{u}}^\top (\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \mathbf{1}]^\top [\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \lesssim \frac{\omega(\mathcal{C}) + t}{\sqrt{n}}. \tag{5.4.5}$$

Using the fact that left-hand side is the rate of convergence $\rho(\mathcal{C})$ concludes the proof for subgaussian samples. ■

5.4.3 Proof of Theorem 44 for subexponential samples

Proof. Let $(\boldsymbol{x}_i)_{i=1}^n \sim \boldsymbol{x} \in \mathbb{R}^p$ be i.i.d. isotropic subexponential vectors and $\boldsymbol{X}$ be the associated design matrix as previously. Let $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in (5.2.5) and (5.2.6) respectively. Assume $n \gtrsim \omega_n^2(\mathcal{C})$. Our proof strategy is based on the observation that, we can bound the (restricted) singular values of $[\boldsymbol{X}\ \mathbf{1}]^\top [\boldsymbol{X}\ \mathbf{1}]$ with high probability for subexponential data as follows.

- **Upper bounding the singular values:** In this section we will upper bound the largest eigenvalue of the matrix $[\boldsymbol{X}\ \mathbf{1}]^\top [\boldsymbol{X}\ \mathbf{1}]$ with high probability. Towards this goal, we utilize Matrix Chernoff bound from [175]. For the sake of completeness, we present the Matrix Chernoff Theorem in the following.

Theorem 47 (Matrix Chernoff [175]) *Consider a finite sequence $\{\boldsymbol{X}_i\}_{i=1}^n$ of independent, random, positive semidefinite matrices with common dimension d. Assume that*

$$\|\boldsymbol{X}_i\| \leq L \quad \textit{for } i = 1, \dots, n. \tag{5.4.6}$$

Define the sum $\boldsymbol{M} = \sum_{i=1}^n \boldsymbol{X}_i$ and let $\zeta_{\max}$ be an upper bound on the spectral norm of the expectation $\mathbb{E}[\boldsymbol{M}]$ i.e. $\zeta_{\max} \geq \|\mathbb{E}[\boldsymbol{M}]\| = \|\sum_{i=1}^n \mathbb{E}[\boldsymbol{X}_i]\|$. We have that

$$\mathbb{P}\left(\|\boldsymbol{M}\| \geq (1+\epsilon)\zeta_{\max}\right) \leq d\left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\frac{\zeta_{\max}}{L}}, \epsilon \geq 0. \tag{5.4.7}$$

We will use Theorem 47 to bound the largest eigenvalue of $[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}]$. Observe that

$$[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}] = \sum_{i=1}^n \begin{bmatrix} \boldsymbol{x}_i \\ 1 \end{bmatrix} [\boldsymbol{x}_i^\top\ 1]. \tag{5.4.8}$$

Clearly this matrix is positive semidefinite. To bound $\|[\boldsymbol{x}_i^\top\ 1]^\top[\boldsymbol{x}_i^\top\ 1]\|$, we use the following lemma.

Lemma 48 (Spectral norm bound) *Let $(\boldsymbol{x}_i)_{i=1}^n$ be i.i.d. isotropic subexponential samples in $\mathbb{R}^p$. Then, with probability at least $1 - 2(n+p)^{-100}$ the spectral norm of all $\boldsymbol{x}_i\boldsymbol{x}_i^\top$ matrices can be bounded as*

$$\|\boldsymbol{x}_i\boldsymbol{x}_i^\top\| \leq \|\boldsymbol{x}_i\|_{\ell_2}^2 \leq cp\log^2(n+p). \tag{5.4.9}$$

The proof of lemma 48 is deferred to Section C.3. Lemma 48 guarantees that $\|[\boldsymbol{x}_i^\top\ 1]^\top[\boldsymbol{x}_i^\top\ 1]\| \leq \|[\boldsymbol{x}_i^\top\ 1]^\top\|_{\ell_2}^2 = \|\boldsymbol{x}_i\|_{\ell_2}^2 + 1 \leq Cp\log^2(n+p)$. Hence, we do satisfy the conditions required by Theorem 47. Before using Theorem 47 we will upper bound the spectral norm of the expectation $\mathbb{E}[[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}]]$ as follows.

Lemma 49 (Spectral norm bound of expectation) *Let $\boldsymbol{x} \in \mathbb{R}^p$ be an isotropic subexponential vector, $\tilde{\boldsymbol{x}} = [\boldsymbol{x}^\top \ 1]^\top$ and let $B = Cp\log^2(n+p)$ for sufficiently large constant $C > 0$. Then we have*

$$\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \ \middle| \ \|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B\right] \le 2\boldsymbol{I}_p. \tag{5.4.10}$$

The proof of Lemma 49 is deferred to Section C.4. Thus, applying Lemma 49 on the set of all $[\boldsymbol{x}_i^\top \ 1]^\top$ satisfying $\|[\boldsymbol{x}_i^\top \ 1]^\top[\boldsymbol{x}_i^\top \ 1]\| \le Cp\log^2(n+p)$, we find that with probability $1 - 2(n+p)^{-100}$ the following holds

$$\|\,\mathbb{E}[[\boldsymbol{X}\ \boldsymbol{1}]^\top[\boldsymbol{X}\ \boldsymbol{1}]]\| = \|\,\mathbb{E}[\sum_{i=1}^n \begin{bmatrix} \boldsymbol{x}_i \\ \\ 1 \end{bmatrix} [\boldsymbol{x}_i^\top \ 1]]\| \le \|\sum_{i=1}^n 2\boldsymbol{I}_p\| = 2n. \tag{5.4.11}$$

Hence, we can pick $\zeta_{\max} \ge 2n$ to upper bound the largest eigenvalue of $\mathbb{E}[[\boldsymbol{X}\ \boldsymbol{1}]^\top[\boldsymbol{X}\ \boldsymbol{1}]]$. Now, using Theorem 47 with $\zeta_{\max} = C_0C(n+p)\log^3(n+p), L = Cp\log^2(n+p)$ and $\epsilon = e - 1$ we get

$$\begin{aligned}\mathbb{P}\left(\|[\boldsymbol{X}\ \boldsymbol{1}]^\top[\boldsymbol{X}\ \boldsymbol{1}]\| \ge eC_0C(n+p)\log^3(n+p)\right) &\le p\left[\frac{e^{e-1}}{e^e}\right]^{C_0\frac{n+p}{p}\log(n+p)} \\ &= pe^{-C_0\frac{n+p}{p}\log(n+p)} \le (n+p)^{-100}. \end{aligned} \tag{5.4.12}$$

Union bounding, with probability at least $1 - 3(n+p)^{-100}$,

$$\|[\boldsymbol{X}\ \boldsymbol{1}]^\top[\boldsymbol{X}\ \boldsymbol{1}]\| \lesssim (n+p)\log^3(n+p). \tag{5.4.13}$$

- **Lower bounding the singular values:** In this section we will lower bound the gain of $[\boldsymbol{X}\ \boldsymbol{1}]$ restricted to the tangent ball $\mathcal{C}_{\text{ext}}$. We will utilize the notion of restricted singular value (RSV) to proceed.

Definition 50 (Restricted singular value) *Given a matrix $\boldsymbol{M}$ and a closed set $\mathcal{C}$, the RSV of $\boldsymbol{M}$ at $\mathcal{C}$ is defined as*

$$\sigma(\boldsymbol{M},\mathcal{C}) = \min_{\boldsymbol{v}\in\mathcal{C}} \frac{\|\boldsymbol{M}\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_2}}. \tag{5.4.14}$$

In the following, we will lower bound $\min_{\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}},\|\tilde{\boldsymbol{v}}\|_{\ell_2}=1} \|[\boldsymbol{X}\ \mathbf{1}]\tilde{\boldsymbol{v}}\|_{\ell_2}$ which is the RSV of $[\boldsymbol{X}\ \mathbf{1}]$ at $\mathcal{C}_{\text{ext}}$. Observe that any $\tilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}$ with unit Euclidian norm obeys $\tilde{\boldsymbol{v}} = [\sqrt{1-\gamma^2}\boldsymbol{v}^\top\ \gamma]^\top$ for $|\gamma| \le 1$ and $\boldsymbol{v} \in \mathcal{C}$, $\|\boldsymbol{v}\|_{\ell_2} = 1$. Consequently

$$\begin{aligned}\|[\boldsymbol{X}\ \mathbf{1}]\tilde{\boldsymbol{v}}\|_{\ell_2}^2 &= \|\sqrt{1-\gamma^2}\boldsymbol{X}\boldsymbol{v} + \gamma\mathbf{1}\|_{\ell_2}^2 \\ &= (1-\gamma^2)\ \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2\mathbf{1}^\top\mathbf{1} + 2\gamma\sqrt{1-\gamma^2}\mathbf{1}^\top\boldsymbol{X}\boldsymbol{v} \\ &\ge (1-\gamma^2)\ \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2 n + 2\gamma\sqrt{1-\gamma^2}\boldsymbol{v}^\top\sum_{i=1}^n \boldsymbol{x}_i. \end{aligned} \tag{5.4.15}$$

Setting $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ and minimizing both sides over $\tilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}, \|\tilde{\boldsymbol{v}}\|_{\ell_2} = 1$, we get

$$\begin{aligned}\min_{\substack{\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}\\ \|\tilde{\boldsymbol{v}}\|_{\ell_2}=1}} \|[\boldsymbol{X}\ \mathbf{1}]\tilde{\boldsymbol{v}}\|_{\ell_2}^2 &\ge \min_{|\gamma|\le 1}\Big((1-\gamma^2)\ \min_{\substack{\boldsymbol{v}\in\mathcal{C}\\ \|\boldsymbol{v}\|_{\ell_2}=1}} \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2 n\Big) - 2n \sup_{\substack{\boldsymbol{v}\in\mathcal{C}\\ \|\boldsymbol{v}\|_{\ell_2}=1}} |\boldsymbol{v}^\top\bar{\boldsymbol{x}}| \\ &\ge \min\Big(\min_{\substack{\boldsymbol{v}\in\mathcal{C}\\ \|\boldsymbol{v}\|_{\ell_2}=1}} \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2, n\Big) - 2n\sup_{\boldsymbol{v}\in\mathcal{C}}|\boldsymbol{v}^\top\bar{\boldsymbol{x}}|. \end{aligned} \tag{5.4.16}$$

In essence, (5.4.16) bounds RSV of $[\boldsymbol{X}\ \mathbf{1}]$ in terms of the RSV of $\boldsymbol{X}$ and some simpler terms. The following theorem from [86] (Theorem D.11) gives a lower lower bound on the RSV of a matrix $\boldsymbol{X}$ with i.i.d. subexponential rows.

Theorem 51 (Bounding RSV [86]) *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. isotropic subexponential rows. Let $\mathcal{C}$ be a tangent ball as in (5.2.5) and suppose the sample size obeys $n \gtrsim (\omega_n(\mathcal{C}) + t)$. Then with probability at least $1 - 3\exp(-c\min(n, t\sqrt{n}, t^2))$, we have that*

$$\min_{\substack{\boldsymbol{v}\in\mathcal{C}\\ \|\boldsymbol{v}\|_{\ell_2}=1}} \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 \ge c_0 n. \tag{5.4.17}$$

Next, we shall state a lemma from [86] (Lemma D.7) to upper bound the term involving the sample average $\bar{\boldsymbol{x}}$.

Lemma 52 (Bounding empirical width [86]) *Suppose $\mathcal{C}$ is a subset of the unit Euclidian ball and $(\boldsymbol{x}_i)_{i=1}^n$ are i.i.d. zero-mean vectors with bounded subexponential norm. Define the empirical average vector $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_i \boldsymbol{x}_i$. We have that*

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{C}} |\boldsymbol{u}^\top \bar{\boldsymbol{x}}| \le C\frac{(\omega_n(\mathcal{C})+t)}{\sqrt{n}}\right) \ge 1 - 2\exp(-c\cdot\min(t\sqrt{n}, t^2)). \tag{5.4.18}$$

Plugging the bounds of Theorem 51 and Lemma 52 into (5.4.16) we find that, there exist constants $c, c_0, C_0 > 0$ such that with probability at least $1 - 5\exp(-c\min(n, t\sqrt{n}, t^2))$, we can lower bound the RSV of $[\boldsymbol{X}\ \mathbf{1}]$ as,

$$\min_{\substack{\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}\\ \|\tilde{\boldsymbol{v}}\|_{\ell_2}=1}} \|[\boldsymbol{X}\ \mathbf{1}]\tilde{\boldsymbol{v}}\|_{\ell_2}^2 \ge c_0 n - C_0 n\frac{\omega_n(\mathcal{C})+t}{\sqrt{n}} \ge c_0 n/2, \tag{5.4.19}$$

where the last line follows from the assumption that $n \gtrsim (\omega_n(\mathcal{C})+t)^2$.

- **Upper bounding the convergence rate:** Union bounding the events (5.4.13) and (5.4.19), we obtain upper and lower bounds on the singular values of $[\boldsymbol{X}\ \mathbf{1}]$ with the desired probability. Hence, we can bound the convergence rate of PGD as follows. Setting $q = (n+p)\log^3(n+p)$, we have (5.4.13) $\|[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}]\| \le Cq$. Therefore, choosing learning rate $\eta = 1/Cq$, the matrix $\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}]$ is positive semidefinite (PSD). Hence, applying the generalized Cauchy-Schwarz inequality for PSD matrices, we find

$$\begin{aligned}
\rho(\mathcal{C}) &= \sup_{\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} |\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \\
&= \sup_{\substack{\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}\\ \|\tilde{\boldsymbol{u}}\|_{\ell_2}=\|\tilde{\boldsymbol{v}}\|_{\ell_2}=1}} |\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \\
&\le \sup_{\substack{\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}\\ \|\tilde{\boldsymbol{u}}\|_{\ell_2}=\|\tilde{\boldsymbol{v}}\|_{\ell_2}=1}} [|\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{u}}|^{1/2}|\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}|^{1/2}] \\
&= \sup_{\substack{\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}\\ \|\tilde{\boldsymbol{v}}\|_{\ell_2}=1}} |\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\eta[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \\
&= 1-\eta \min_{\substack{\tilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}\\ \|\tilde{\boldsymbol{v}}\|_{\ell_2}=1}} \|[\boldsymbol{X}\ \mathbf{1}]\tilde{\boldsymbol{v}}\|_{\ell_2}^2 \\
&\le 1-c_0\eta n/2. \qquad (5.4.20)
\end{aligned}$$

Here the last inequality follows from (5.4.19). This completes the proof for subexponential samples. ■

5.4.4 Proof of Theorem 45 for subgaussian samples

Proof. Suppose the dataset $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim (\boldsymbol{x}, y)$ is σ-subgaussian. Let $\boldsymbol{X}, [\boldsymbol{X}\ \mathbf{1}], \mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in Section 5.2, recall $\boldsymbol{w}$ from (5.3.5) and assume $n \gtrsim (\omega(\mathcal{C})+t)^2$.

Representing a vector $\tilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}$ as $\tilde{\boldsymbol{v}} = [\sqrt{1-\gamma^2}\boldsymbol{v}^\top \;\gamma]^\top$ for $\boldsymbol{v} \in \mathcal{C}$ and $|\gamma| \le 1$, we have

$$\begin{aligned}
\nu(\mathcal{C}) &= \sup_{\tilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}} |\tilde{\boldsymbol{v}}^\top [\boldsymbol{X}\;\mathbf{1}]^\top \boldsymbol{w}| \\
&= \sup_{\substack{\boldsymbol{v}\in\mathcal{C}\\ |\gamma|\le 1}} |\sqrt{1-\gamma^2}\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w} + \gamma \mathbf{1}^\top \boldsymbol{w}| \\
&\le \sup_{\substack{\boldsymbol{v}\in\mathcal{C}\\ |\gamma|\le 1}} |\sqrt{1-\gamma^2}\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w}| + \sup_{|\gamma|\le 1} |\gamma \mathbf{1}^\top \boldsymbol{w}| \\
&\le \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w}| + |\mathbf{1}^\top \boldsymbol{w}|. \qquad (5.4.21)
\end{aligned}$$

In the following we will upper bound the terms $\sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w}|$ and $|\mathbf{1}^\top \boldsymbol{w}|$ separately and will combine them to get an upper bound on the residual error.

- **Upper bounding the first term in (5.4.21):** In order to upper bound the first term in (5.4.21), define the clipping function

$$\text{clip}(a, B) = \begin{cases} a \text{ if } |a| \le B \\ \\ \text{sign}(a)B \text{ else} \end{cases}. \qquad (5.4.22)$$

The following lemma immediately follows from union bounding the large deviations of subgaussian and subexponential variables X and shows that $X = \text{clip}(X, B)$ with high probability.

Lemma 53 *Let $(w_i)_{i=1}^n$ be i.i.d. subgaussian random variables with $\|w_i\|_{\psi_2} \le \sigma$. There exists a constant $C > 0$ such that picking $B = C\sqrt{\log(n)}$, with probability $1 - 2n^{-100}$ for all i, we have*

$$w_i = \textit{clip}(w_i, \sigma B). \qquad (5.4.23)$$

If instead $(w_i)_{i=1}^n$ are i.i.d. subexponential with $\|w_i\|_{\psi_1} \le \sigma$, then picking $B = C\log(n)$ leads to the same result.

Using Lemma 53, $\|\boldsymbol{w}\|_{\ell_\infty} \leq \sigma B$ with probability $1 - 2n^{-100}$. Conditioned on this event, we have

$$\sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w}| = \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^n \text{clip}(w_i, \sigma B)\boldsymbol{x_i}|. \tag{5.4.24}$$

Setting $\boldsymbol{z}_i = \text{clip}(w_i, \sigma B)\boldsymbol{x_i} = w_i \boldsymbol{x}_i$, (5.4.24) can be re-written as

$$\begin{aligned}\sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w}| &= \frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^n \boldsymbol{z}_i| \\ &\leq \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| + \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^n \mathbb{E}[\boldsymbol{z}_i]| \\ &\leq \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| + n\|\mathbb{E}[\boldsymbol{z}_1]\|_{\ell_2}. \end{aligned} \tag{5.4.25}$$

Note that $\boldsymbol{z}_i = w_i \boldsymbol{x}_i$ is subgaussian since w_i is bounded. The subgaussian norm obeys

$$\|\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i]\|_{\psi_2} \lesssim \|\boldsymbol{z}_i\|_{\psi_2} \lesssim \sigma\sqrt{\log(n)}\|\boldsymbol{x}_i\|_{\psi_2} \lesssim \sigma\sqrt{\log(n)}. \tag{5.4.26}$$

Define the average vector $\bar{\boldsymbol{z}} = n^{-1/2} \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])$ which is still subgaussian with same norm (up to a constant). Standard results from functional analysis (specifically generic chaining) [168] guarantee

$$\begin{aligned}\frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| &= \frac{1}{\sqrt{n}} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^\top \bar{\boldsymbol{z}}| \\ &\lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}}. \end{aligned} \tag{5.4.27}$$

with probability at least $1 - 2e^{-t^2/2}$. This bounds the first term of (5.4.25). Next, we address the expectation term $\|\mathbb{E}[\boldsymbol{z}_1]\|_{\ell_2}$ via following lemma.

Lemma 54 *Suppose $\boldsymbol{x}$ is an isotropic Orlicz-a vector and $\|w\|_{\psi_a} \leq \sigma$ where $w = y - \boldsymbol{x}^\top \boldsymbol{\theta}_\star - \mu_\star$. Let $B = C\log^{1/a}(n)$ for sufficiently large constant $C > 0$. For $a = 1, 2$, we have that*

$$\|\mathbb{E}[w\boldsymbol{x} \mid |w| \leq \sigma B]\|_{\ell_2} \lesssim \sigma p^2 n^{-201}. \tag{5.4.28}$$

The proof of Lemma 54 is deferred to Section C.6. Combining (5.4.27) and Lemma 54 into (5.4.25), with probability at least $1-2e^{-t^2/2}-2n^{-100}$, we find that,

$$\begin{aligned}\frac{1}{n}\sup_{\boldsymbol{v}\in\mathcal{C}}|\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{w}| &\lesssim \frac{\sigma(\omega(\mathcal{C})+t)\sqrt{\log(n)}}{\sqrt{n}}+\sigma p^2 n^{-200}\\ &\lesssim \frac{\sigma(\omega(\mathcal{C})+t)\sqrt{\log(n)}}{\sqrt{n}} \qquad (5.4.29)\end{aligned}$$

which is the desired bound for the first term in (5.4.21).

- **Upper bounding the second term in (5.4.21):** The vector $\boldsymbol{w}$ is zero-mean with $\|\boldsymbol{w}\|_{\psi_2}\le\sigma$. Hence, $\|\mathbf{1}^\top\boldsymbol{w}\|_{\psi_2}\le\sigma\sqrt{n}$ which implies that with probability $1-2n^{-100}$,

$$|\mathbf{1}^\top\boldsymbol{w}|\lesssim\sigma\sqrt{n\log n}. \qquad (5.4.30)$$

Combining the bound above with (5.4.29), we get the advertised bound on the residual, namely

$$\frac{1}{n}\nu(\mathcal{C})\lesssim\frac{\sigma(\omega(\mathcal{C})+t)\sqrt{\log(n)}}{\sqrt{n}}, \qquad (5.4.31)$$

with probability at least $1-2\exp(-t^2/2)-4n^{-100}$. This completes the proof for σ-subgaussian data. ■

5.4.5 Proof of Theorem 45 for subexponential samples

Proof. Suppose the dataset $(\boldsymbol{x}_i,y_i)_{i=1}^n\sim(\boldsymbol{x},y)$ is σ-subexponential. Let $\boldsymbol{X},[\boldsymbol{X}\ \mathbf{1}],\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in Section 5.2, recall $\boldsymbol{w}$ from (5.3.5) and assume $n\gtrsim(\omega_n(\mathcal{C})+t)^2$. Similar to the subgaussian case, we split the residual into two terms via (5.4.21) and bound each term separately to get a final bound.

- **Upper bounding the first term in (5.4.21):** Let $\boldsymbol{z}_i=w_i\boldsymbol{x}_i$. With probability $1-2n^{-100}$, we have that $\|\boldsymbol{w}\|_{\ell_\infty}\lesssim\sigma\log n$. We continue the analysis conditioned on this event.

With bounded w_i, $\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i]$ is subexponential via

$$\|\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i]\|_{\psi_1} \lesssim \|\boldsymbol{z}_i\|_{\psi_1} \lesssim \sigma \log n \|\boldsymbol{x}_i\|_{\psi_1} \lesssim \sigma \log n. \tag{5.4.32}$$

Combining this with Lemma 52, guarantees that

$$\frac{1}{n} \sup_{\boldsymbol{v} \in \mathcal{C}} |\boldsymbol{v}^\top \sum_{i=1}^{n} (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| \lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t) \log(n)}{\sqrt{n}} \tag{5.4.33}$$

with probability at least $1 - 2\exp(-\mathcal{O}(\min(t\sqrt{n}, t^2)))$. Next, using Lemma 54, we also upper bound $\|\mathbb{E}[\boldsymbol{z}_1]\|_{\ell_2}$ by $C\sigma p^2 n^{-201}$. Combining this with (5.4.33) and substituting into (the deterministic inequality) (5.4.25), with probability at least $1 - 2\exp(-\mathcal{O}(\min(t\sqrt{n}, t^2))) - 2n^{-100}$ we have,

$$\frac{1}{n} \sup_{\boldsymbol{v} \in \mathcal{C}} |\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{w}| \lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t) \log(n)}{\sqrt{n}}. \tag{5.4.34}$$

- **Upper bounding the second term in (5.4.21):** Using $\|w_i\|_{\psi_1} \lesssim \sigma$ and applying Lemma 52 (over one-dimensional $\mathbb{R}$), we find that $|\mathbf{1}^\top \boldsymbol{w}| \lesssim \sigma(1+t)\sqrt{n}$ with probability $1 - 2\exp(-c\min(t\sqrt{n}, t^2))$. Combining this with (5.4.34) and plugging into (5.4.21), we get the advertised upper bound

$$\begin{aligned} \frac{1}{n}\nu(\mathcal{C}) &\lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t) \log(n)}{\sqrt{n}} + \frac{(1+t)\sigma}{\sqrt{n}} \\ &\lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t) \log(n)}{\sqrt{n}} \end{aligned} \tag{5.4.35}$$

which holds with probability at least $1 - 4\exp(-c\min(t\sqrt{n}, t^2)) - 2n^{-100}$. This completes the proof for σ-subexponential data. ■

Chapter 6

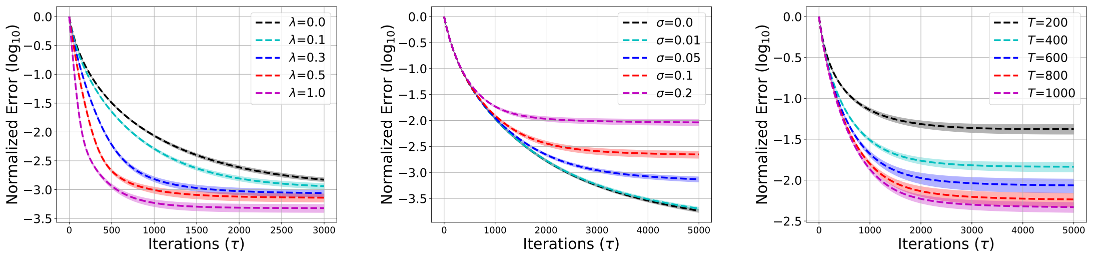
Numerical Experiments

6.1 Nonlinear System Identification

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 6.1: 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.

For our experiments, we choose unstable nonlinear dynamical systems ($\rho(\boldsymbol{A}) > 1$) governed by nonlinear state equation $\boldsymbol{x}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t) + \boldsymbol{w}_t$ 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



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

Figure 6.1: We run gradient descent to learn nonlinear dynamical system governed by state equation $\boldsymbol{x}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t) + \boldsymbol{w}_t$. 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.

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 $\boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{x}_t + \boldsymbol{z}_t$. 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 6.2, 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 nonlinear state equation $\boldsymbol{x}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t) + \boldsymbol{w}_t$.

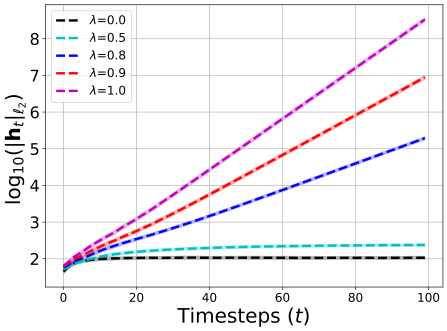


Figure 6.2: For a properly chosen random unstable system the state vectors diverge for LDS while they stay bounded for leaky ReLU systems with small leakage.

- **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 6.1a, 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 6.1b, 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.

- **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 6.1c, 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 6.2 is generated by evolving the state through 100 timesteps and recording the Euclidean norm of $\boldsymbol{x}_t$ at each timestep. This is repeated 500 times with $\rho(\boldsymbol{A}) > 1$ and using leaky-ReLU activations. In Figure 6.2, we plot the mean and one standard deviation of the Euclidean norm of the states $\boldsymbol{x}_t$ 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 $\boldsymbol{x}_0$ has all entries positive. Then, using leaky-ReLU will not help the trajectory to converge.

6.2 Bilinear System Identification

For our experiments, we choose a bilinear dynamical system (3.2.1) with state dimension $n = 8$ and input dimension $p = 4$. $\boldsymbol{A}_0$ is generated with $\mathcal{N}(0,1)$ entries and scaled to have its largest eigenvalues equal to 0.6. Similarly, $\{\boldsymbol{A}_k\}_{k=1}^p$ are generated with $\mathcal{N}(0,1)$ entries and scaled to have their largest eigenvalue equal to $1/p$. Using $\boldsymbol{x}_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\boldsymbol{I}_n)$, $\{\boldsymbol{u}_t\}_{t=0}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{u}}^2\boldsymbol{I}_p)$ and $\{\boldsymbol{w}_t\}_{t=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{w}}^2\boldsymbol{I}_n)$, we generate a single finite trajectory $\{(\boldsymbol{u}_t,\boldsymbol{x}_t,\boldsymbol{x}_{t+1})\}_{t=0}^T$ of the bilinear dynamical system (3.2.1), which is given as an input to Algorithm 1.

We plot, (i) the normalized estimation error of $\boldsymbol{A}_0$ given by $\|\hat{\boldsymbol{A}}_0 - \boldsymbol{A}_0\|/\|\boldsymbol{A}_0\|$, and (ii) the average normalized estimation error of $\{\boldsymbol{A}_k\}_{k=1}^p$ given by $(1/p)\sum_{k=1}^p \|\hat{\boldsymbol{A}}_k - \boldsymbol{A}_k\|/\|\boldsymbol{A}_k\|$.

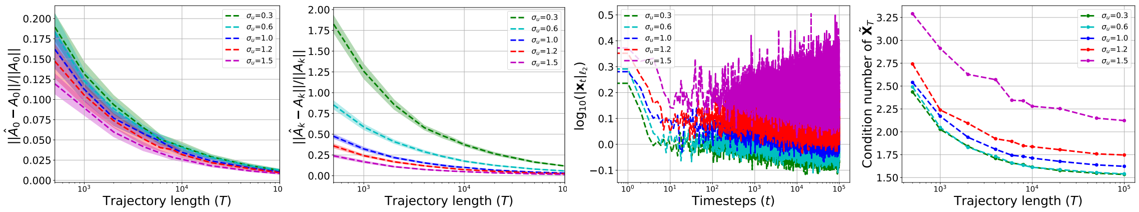


Figure 6.3: Identification with varying input variance $\sigma_{\boldsymbol{u}}^2$

Each experiment is repeated 20 times and we plot the mean and one standard deviation. We also plot, (iii) the Euclidean norm of the states $\{\|\boldsymbol{x}_t\|_{\ell_2}\}_{t=0}^T$, and (iv) the condition number of the design matrix $\tilde{\boldsymbol{X}}_T$. To verify our theoretical results from Section 3.3, We perform the following two different types of experiments.

- **Input strength:** In this experiment, we run Algorithm 1 with different values of $\sigma_{\boldsymbol{u}}$ and T, while setting the values of $n, p, \rho(\boldsymbol{A})$ and $\rho(\boldsymbol{A}_k)$ as described above. We also set $\sigma_{\boldsymbol{w}} = 0.3$. The results of this experiment are plotted in Figure 6.3. As predicted by our theory, the estimation errors of $\{\boldsymbol{A}_k\}_{k=0}^p$ converge to 0 with the increasing trajectory length. Another important observation is that the estimation errors also decrease with increasing $\sigma_{\boldsymbol{u}}$. This is more prominent in the case of $\{\boldsymbol{A}_k\}_{k=1}^p$, which is consistent with the message of Theorem 24. Furthermore, Table 6.2 shows that increasing $\sigma_{\boldsymbol{u}}$ results in an increase in the spectral radius of the augmented state matrix $\tilde{\boldsymbol{A}}$. This also implies that we cannot increase $\sigma_{\boldsymbol{u}}$ above a certain threshold. Otherwise, the bilinear system might become unstable and we might not be able to learn the dynamics $\{\boldsymbol{A}_k\}_{k=0}^p$.

- **Noise level:** In this experiment, we run Algorithm 1 with different values of $\sigma_{\boldsymbol{w}}$ and T, while setting the values of $n, p, \rho(\boldsymbol{A})$ and $\rho(\boldsymbol{A}_k)$ as described above. We also set $\sigma_{\boldsymbol{u}} = 1.5$. The results of this experiment are plotted in Figure 6.4. Larger trajectory

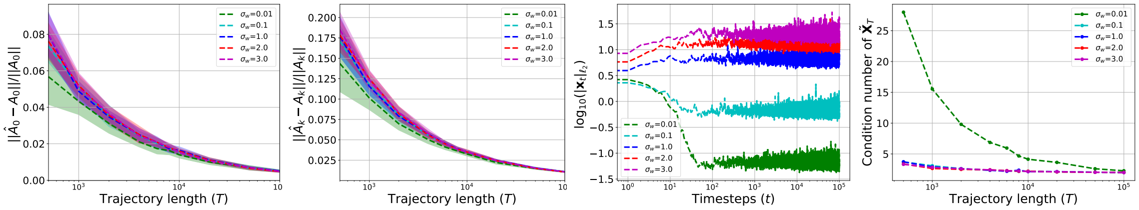


Figure 6.4: Identification with varying noise variance $\sigma_{\boldsymbol{w}}^2$

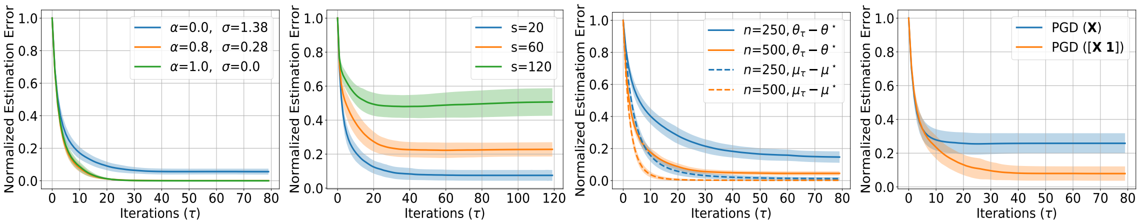
length helps here as well. Interestingly, the estimation errors are independent of the noise strength $\sigma_{\boldsymbol{w}}$. This is as predicted by Theorem 24. From Figure 6.4, we also see that, when the trajectory length is sufficiently large, the condition number of $\tilde{\boldsymbol{X}}_T$ is similar for different noise levels. When the trajectory length and the noise level are very small, $\tilde{\boldsymbol{X}}_T$ has larger condition number because of the random initialization of $\boldsymbol{x}_0$ and the decrease in Euclidean norm of $\boldsymbol{x}_t$ with time (see Figure 6.4). If the noise is 0 and the unknown bilinear system has $\rho(\tilde{\boldsymbol{A}}) < 1$, then as shown in Lemma 21, the states will converge to 0 exponentially fast. Therefore, most of the samples in the collected trajectory $\{(\boldsymbol{u}_t, \boldsymbol{x}_t, \boldsymbol{x}_{t+1})\}_{t=0}^T$ will be zero.

$\sigma_{\boldsymbol{u}}$	0.3	0.6	1.0	1.2	1.5
$\rho(\tilde{\boldsymbol{A}})$	0.369	0.402	0.509	0.595	0.764

Table 6.2: $\rho(\tilde{\boldsymbol{A}})$ increases with increasing $\sigma_{\boldsymbol{u}}$

6.3 Finding Best Linear Model in High Dimensions

We consider a standard single-index model where for some ground truth vector $\boldsymbol{a}$ and link function ϕ, the input/output relation is given by $y_i = \phi(\boldsymbol{a}^T \boldsymbol{x}_i)$. We pick $\boldsymbol{a}$



(a) Nonlinearity (b) Sparsity (c) Sample size (d) Bias estimation

Figure 6.5: We run PGD with leaky-ReLU activation ($\max(x, \alpha x)$) using subexponential design matrix $[\boldsymbol{X}\ \mathbf{1}]$. We plot the normalized estimation error $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^\star\|_{\ell_2}^2 / \|\boldsymbol{\theta}^\star\|_{\ell_2}^2$ with varying degree of (a) nonlinearity, (b) sparsity and (c) sample size. (c) also shows the bias estimation error $\|\boldsymbol{\mu}_\tau - \boldsymbol{\mu}^\star\|_{\ell_2}^2 / \|\boldsymbol{\mu}^\star\|_{\ell_2}^2$. The estimation error decays quickly however the eventual error varies as a function of the nonlinearity, sparsity and sample size as predicted by Theorem 43. (d) PGD with bias estimation outperforms vanilla PGD (using $\boldsymbol{X}$ alone).

to be an s sparse vector with i.i.d. $\mathcal{N}(0,1)$ nonzero entries and set the dimension to be $p = 800$. Based on the sparsity prior, we run PGD as iterative hard thresholding where $\boldsymbol{\theta}_\tau$ is projected to be s-sparse after every iteration. As a link function, we considered leaky-ReLU (i.e. $\max(x, \alpha x)$ where $0 \leq \alpha \leq 1$); which is of interest for deep learning. We generate $\boldsymbol{x}_i$'s with i.i.d. exponentially distributed entries (with parameter $\lambda = 1$) and then remove the mean and normalize the covariance to identity. We pick a learning rate of $\eta = 0.5/(n+p)$ in all experiments, where n is the sample size and p is the dimension of parameter. The shaded areas in the plots correspond to one standard deviation.

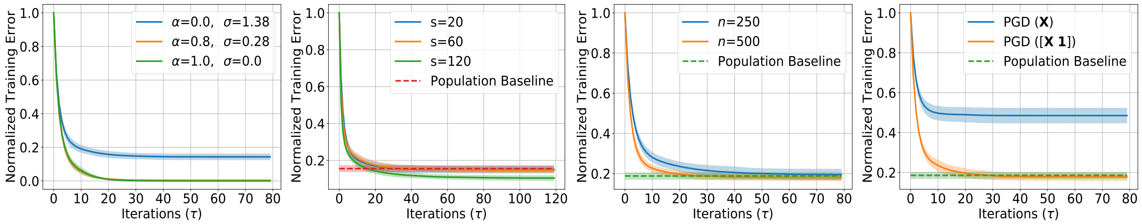
To assess the performance of PGD, we use the following metrics at a gradient iterate $\boldsymbol{\theta}_\tau$:

- **normalized estimation error:** $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^\star\|_{\ell_2}^2 / \|\boldsymbol{\theta}^\star\|_{\ell_2}^2$,
- **normalized training error:** $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau - \mu_\tau \mathbf{1}\|_{\ell_2}^2 / \|\boldsymbol{y}\|_{\ell_2}^2$.
- **normalized test error:** similarly defined but evaluated on a fresh dataset of size n.

The population BLM $\boldsymbol{\theta}^\star$ is estimated using $100,000$ samples by solving a linear regression. To verify our theoretical results, we study the effect of the nonlinearity, sparsity, and sample size on the quality of the PGD estimate. Figures 6.5, 6.6, and 6.7 plot the estimation error, training error, and test error for the same set of configurations described below. For Figures 6.6 and 6.7, the PGD errors are compared to the BLM baseline $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}^\star-\mu^\star\mathbf{1}\|_{\ell_2}^2/\|\boldsymbol{y}\|_{\ell_2}^2$ which is the error at the BLM $\boldsymbol{\theta}^\star,\mu^\star$ (highlighted as the dashed green line).

- **The degree of nonlinearity:** Figures 6.5a, 6.6a, 6.7a plot the errors over different degrees of nonlinearity (quantified by the parameter α) with $s=20$ and $n=500$. The estimation error grows with the increase in the degree of nonlinearity and we almost perfectly recover $\boldsymbol{\theta}^\star$ for the linear case. We also state the effective noise level $\sigma=\mathbb{E}[(y-\langle\boldsymbol{\theta}^\star,\boldsymbol{x}\rangle-\mu^\star)^2]^{1/2}$. σ is zero for the linear case ($\alpha=1$) and it increases with decreasing the value of α, resulting in larger estimation error as predicted by our theory. Note that if nonlinearity is mild, BLM can achieve good test accuracy (Fig. 6.7a). At $\alpha=0.8$ normalized test error is around 0.0043.

- **Sparsity:** Figures 6.5b, 6.6b, 6.7b plot the errors over different levels of sparsity (s) while setting $n=500$ and using ReLU function ($\alpha=0$). The estimation improves with increasing sparsity (smaller s) which is consistent with Table 5.1 where larger s leads to larger (perturbed) Gaussian width. Hence, these figures are consistent with Theorem 43, which states that the statistical estimation error grows as $\omega_n(\mathcal{C})/\sqrt{n}$. Sparsity $s=20$ achieves same training error as the BLM baseline and has good test performance. As sparsity grows ($s=60,120$), PGD achieves lower training error than the baseline. This implies that PGD is overfitting as verified in Figure 6.7b.

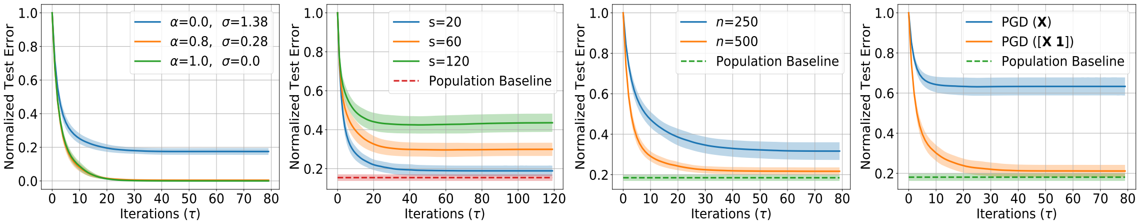


(a) Nonlinearity (b) Sparsity (c) Sample size (d) Bias estimation

Figure 6.6: Normalized training errors $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau - \mu_\tau \mathbf{1}\|_{\ell_2}^2 / \|\boldsymbol{y}\|_{\ell_2}^2$ associated with the same set of experiments in Figure 6.5. Errors are plotted by varying (a) the degree of nonlinearity, (b) sparsity, (c) sample size, and (d) the design matrix.

- **Sample size:** Figures 6.5c, 6.6c, 6.7c plot the errors over different sample sizes n while setting $s = 20$ and using ReLU function ($\alpha = 0$). Figure 6.5c also plots the normalized estimation error of the output mean defined as $|\mu_\tau - \mu^\star|^2 / |\mu^\star|^2$. The estimation improves with increasing sample size. This is again consistent with Theorem 43. Note that both $n = 250, n = 500$ achieve similar training errors as the BLM baseline however $n = 250$ has noticeably larger test error which can be anticipated from Figure 6.5c.

- **Effect of debiasing:** Figures 6.5d, 6.6d, 6.7d compare the errors over the design matrices $\boldsymbol{X}$ and $[\boldsymbol{X}\ \mathbf{1}]$, setting $s = 20$ and $n = 500$ and using ReLU function ($\alpha = 0$). We observe that the design matrix $[\boldsymbol{X}\ \mathbf{1}]$ yields much better performance compared to the original matrix $\boldsymbol{X}$. While in theory the design matrix $[\boldsymbol{X}\ \mathbf{1}]$ has a similar convergence guarantee to $\boldsymbol{X}$, in practice it improves the estimation significantly thanks to addressing the output mean and reducing the output variance. Finally, perhaps not surprisingly, we remark that $\boldsymbol{\theta}^\star$ is not equal to the ground truth parameter $\boldsymbol{a}$ and it is not perfectly s sparse. Measuring the correlation coefficient $\rho = \frac{\langle \boldsymbol{a}, \boldsymbol{\theta}^\star \rangle}{\|\boldsymbol{a}\|_{\ell_2} \|\boldsymbol{\theta}^\star\|_{\ell_2}}$ for varying sparsity s reveals that $\rho \approx 0.969$ for $s = 20$, $\rho \approx 0.982$ for $s = 60$ and $\rho \approx 0.991$ for $s = 200$. Here ρ is obtained by averaging over 20 realizations of random $\boldsymbol{a}$ and $\boldsymbol{\theta}^\star$ is empirically found from 10^5 samples.



(a) Nonlinearity (b) Sparsity (c) Sample size (d) Bias estimation

Figure 6.7: Normalized test errors $\|\boldsymbol{y}_{\text{test}} - \boldsymbol{X}_{\text{test}}\boldsymbol{\theta}_\tau - \mu_\tau\mathbf{1}\|_{\ell_2}^2/\|\boldsymbol{y}_{\text{test}}\|_{\ell_2}^2$ associated with the same set of experiments in Figure 6.5. This is same as Figure 6.6 however evaluations are done on a test dataset $(\boldsymbol{y}_{\text{test}}, \boldsymbol{X}_{\text{test}})$

Increased correlation with larger s is previously pointed out by the interesting work [176]. Per discussion in Section 5.3.3, Gaussian samples are guaranteed to be consistent and achieve correlation of 1. Indeed, repeating the same experiments with Gaussian data results in $\rho > 0.995$ for all choices of s where 0.995 (rather than 1) is due to estimating BLM with finite samples.

Chapter 7

Conclusions and Future Directions

In this chapter, we present the conclusions of this thesis and discuss some possible future directions. Specifically, we get the following conclusions:

• **Nonlinear system identification:** We can learn nonlinear dynamical systems by utilizing stability and mixing-time arguments. We show that, under reasonable assumptions, one can learn the dynamics of a nonlinear stabilized system from a single finite trajectory. We find that, we can combine stability with one-point convexity and smoothness condition to learn the nonlinear dynamical system (2.2.1) with an error rate $\mathcal{O}(\sqrt{dL/T})$, which is optimal in terms of trajectory length T and the dimension of the unknown dynamics. Our general approach can treat important dynamical systems, such as standard LTI dynamical systems and the setups of [31–33] as special cases. We provide both sample size and estimation error guarantees on standard LTI dynamical systems and certain nonlinear state equations.

• **Bilinear system identification:** We provide finite sample analysis for learning discrete-time bilinear systems. We find that: (i) under marginal mean-square stability, we can estimate the bilinear systems of the form (3.2.1) with an error rate $\mathcal{O}(\sqrt{n(p+1)/T})$, which is optimal in terms of trajectory length T and the dimension of the unknown matrices, and (ii) the estimation gets better with increasing input variance $\sigma_{\boldsymbol{u}}^2$, whereas, it is independent of the noise variance $\sigma_{\boldsymbol{w}}^2$.

- **Markov jump system identification:** Markov jump systems are fundamental to a rich class of control problems where the underlying dynamics are changing with time. Despite its importance, statistical understanding (sample complexity and error rates) of MJS have been lacking due to the technicalities such as Markovian transitions and weaker notion of mean-square stability. At a high-level, this thesis overcomes these challenges to provide finite sample system identification guarantees for MJS. Notably, the resulting estimation error $\mathcal{O}(\sqrt{(n+p)/T})$ is optimal in the trajectory length and coincides with the standard LTI system identification up to polylogarithmic factors.

- **Finding best linear model:** We study the problem of finding the best linear model high dimensions. We analyze the projected gradient descent algorithm and show its fast convergence as well as statistical accuracy in a data-dependent fashion. Our results hold for sub-exponential data as well which is heavier tailed compared to well-studied sub-gaussian. In both cases, we prove that nonlinearity of the problem can be treated as uncorrelated noise, and we establish favorable statistical guarantees to estimate the best linear model. Our bounds have a similar flavor to guarantees known for regularized linear regression with independent noise.

7.1 Future Directions

In this section, we discuss some possible future directions. Specifically, our technical tools and ideas from Chapters 2, 3,4 and 5 can be extended or improved to solve the following research problems:

• **Learning nonlinear ARX models:** While we focus on nonlinear state equations, our technical ideas (e.g., combining mixing-time and optimization landscape arguments) have implications for richer class of systems. For instance, nonlinear ARX form $\boldsymbol{x}_t = f_\star(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \cdots, \boldsymbol{x}_{t-m}) + \boldsymbol{w}_{t-1}$ is a powerful generalization of the state equations that we investigate. Koopman lifting provides another class of nonlinear problems. It would be interesting to extend our framework (i.e., merging one-point convexity and smoothness with mixing-time arguments to enable success of gradient descent) to provide non-asymptotic learning guarantees for these systems.

• **Learning non-mixing nonlinear models:** It will be interesting to explore alternative approaches to mixing-time arguments for learning nonlinear dynamical systems. Martingale based arguments have the potential to provide tighter statistical guarantees and mitigate dependence on the spectral radius [22]. It will be interesting to extend these arguments to provide non-asymptotic learning guarantees for nonlinear dynamical systems.

• **Adaptive control for nonlinear systems:** It will be interesting to extend our nonlinear system identification framework to nonlinear adaptive control. One can analyze the nonlinear control problem in both model-free and model-based settings. In these settings, one can analyze the Koopman operator theoretic framework as well as the model predictive control framework to provide end-to-end guarantees. Combining learning and control in the nonlinear setting is challenging. A major challenge is the distribution shift induced by deploying the learned policy. Mitigating the distribution shift in nonlinear adaptive control is itself an interesting research direction.

- **Learning bilinear systems with control:** Our analysis from Chapter 3 can be extended to estimate a more general bilinear system $\boldsymbol{x}_{t+1} = \boldsymbol{A}_0\boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k]\boldsymbol{A}_k\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_{t+1}$. In this case, because of the additional $\boldsymbol{B}\boldsymbol{u}_t$ term, the estimation gets better with increasing input variance $\sigma_{\boldsymbol{u}}^2$ or decreasing the noise variance $\sigma_{\boldsymbol{w}}^2$. In the future, we would like to apply these results for learning more general nonlinear systems by learning a bilinear or state-affine approximation in a higher dimensional space using Koopman operator-like techniques, with the main challenge being the need to jointly learn a lifting and the dynamics in the lifted space.

- **Finding best LTI model in high dimensions:** It would be interesting to extend our results from Chapter 5 to finding the best LTI dynamical system that can minimize least-squares loss given a single trajectory of an unknown dynamical system.

Bibliography

[1] 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.

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

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

[4] David Gaylor and E Glenn Lightsey. Gps/ins kalman filter design for spacecraft operating in the proximity of international space station. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*, page 5445, 2003.

[5] Leonard A McGee. *Discovery of the Kalman filter as a practical tool for aerospace and industry*, volume 86847. National Aeronautics and Space Administration, 1985.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[7] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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

[9] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[10] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354, 2018.

[11] 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.

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

[13] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

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

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

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

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

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

[19] Michel Verhaegen. Subspace model identification part 3. analysis of the ordinary output-error state-space model identification algorithm. *International Journal of control*, 58(3):555–586, 1993.

[20] Lennart Ljung. System identification. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–19, 1999.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1025–1032, 2008.

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

[28] Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001, 2020.

[29] Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *conference on Learning Theory*, pages 3333–3364. PMLR, 2022.

[30] 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.

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

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

[33] 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.

[34] 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.

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

[36] 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 (CAMSAP)*, pages 116–120. IEEE, 2019.

[37] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *The Journal of Machine Learning Research*, 23(1):6248–6296, 2022.

[38] Yahya Sattar, Samet Oymak, and Necmiye Ozay. Finite sample identification of bilinear dynamical systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6705–6711. IEEE, 2022.

[39] Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Identification and adaptive control of markov jump systems: Sample complexity and regret bounds. *arXiv preprint arXiv:2111.07018*, 2021.

[40] Yahya Sattar and Samet Oymak. Quickly finding the best linear model in high dimensions via projected gradient descent. *IEEE Transactions on Signal Processing*, 68:818–829, 2020.

[41] Zhe Du, Yahya Sattar, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Data-driven control of markov jump systems: Sample complexity and regret bounds. In *2022 American Control Conference (ACC)*, pages 4901–4908. IEEE, 2022.

[42] Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Samet Oymak, Laura Balzano, and Necmiye Ozay. Certainty equivalent quadratic control for markov jump systems. In *2022 American Control Conference (ACC)*, pages 2871–2878. IEEE, 2022.

[43] Mingchen Li, Yahya Sattar, Christos Thrampoulidis, and Samet Oymak. Exploring weight importance and hessian bias in model pruning. *arXiv preprint arXiv:2006.10903*, 2020.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

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

[61] Samet Oymak and Necmiye Ozay. Revisiting Ho–Kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

[69] 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, 2020.

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

[71] 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.

[72] 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 (CDC)*, pages 757–762. IEEE, 2021.

[73] 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. *Advances in Neural Information Processing Systems*, 33:14532–14543, 2020.

[74] 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.

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

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

[77] 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.

[78] 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.

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

[80] 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.

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

[82] 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.

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

[84] 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.

[85] 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.

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

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

[88] Ronald R. Mohler. *Bilinear Control Processes: With Applications to Engineering, Ecology, and Medicine.* Elsevier, 1973.

[89] Spyros Svoronos, George Stephanopoulos, and Rutherford Aris. Bilinear approximation of general non-linear dynamic systems with linear inputs. *International Journal of Control*, 31(1):109–126, 1980.

[90] James Ting-Ho Lo. Global bilinearization of systems with control appearing linearly. *SIAM Journal on Control*, 13(4):879–885, 1975.

[91] Krzysztof Kowalski and W-H Steeb. *Nonlinear dynamical systems and Carleman linearization*. World Scientific, 1991.

[92] Debdipta Goswami and Derek A Paley. Global bilinearization and controllability of control-affine nonlinear systems: A koopman spectral approach. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 6107–6112. IEEE, 2017.

[93] Daniel Bruder, Xun Fu, and Ram Vasudevan. Advantages of bilinear koopman realizations for the modeling and control of systems with unknown dynamics. *IEEE Robotics and Automation Letters*, 6(3):4369–4376, 2021.

[94] Jer-Nan Juang. Continuous-time bilinear system identification. *Nonlinear Dynamics*, 39(1):79–94, 2005.

[95] Eduardo D Sontag, Yuan Wang, and Alexandre Megretski. Input classes for identifiability of bilinear systems. *IEEE Transactions on Automatic Control*, 54(2):195–207, 2009.

[96] N Berk Hizir, Minh Q Phan, Raimondo Betti, and Richard W Longman. Identification of discrete-time bilinear systems through equivalent linear models. *Nonlinear Dynamics*, 69(4):2065–2078, 2012.

[97] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. *American Control Conference*, 2019.

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

[99] Boualem Djehiche and Othmane Mazhar. Efficient learning of hidden state lti state space models of unknown order. *arXiv preprint arXiv:2202.01625*, 2022.

[100] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Data driven estimation of stochastic switched linear systems of unknown order. *arXiv preprint arXiv:1909.04617*, 2019.

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

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

[103] C Kubrusly and O Costa. Mean square stability conditions for discrete stochastic bilinear systems. *IEEE Transactions on Automatic Control*, 30(11):1082–1087, 1985.

[104] Marco C Campi and PR Kumar. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM J. Control Optim.*, 36(6):1890–1907, 1998.

[105] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proc. of COLT*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[106] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *FOCM*, pages 1–47, 2019.

[107] 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.

[108] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators. *arXiv preprint arXiv:2007.12291*, 2020.

[109] Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *ICML*, pages 23–31. PMLR, 2020.

[110] KA Loparo and F Abdel-Malek. A probabilistic approach to dynamic power system security. *IEEE transactions on circuits and systems*, 37(6):787–798, 1990.

[111] DO Cajueiro. *Stochastic optimal control of jumping Markov parameter processes with applications to finance.* PhD thesis, PhD thesis, 2002, Instituto Tecnológico de Aeronáutica-ITA, Brazil, 2002.

[112] Valery Ugrinovskii* and Hemanshu R Pota. Decentralized control of power systems via robust control of uncertain markov jump parameter systems. *International Journal of Control*, 78(9):662–677, 2005.

[113] Lars Blackmore, Stanislav Funiak, and Brian C Williams. Combining stochastic and greedy search in hybrid estimation. In *AAAI*, pages 282–287, 2005.

[114] Lars EO Svensson, Noah Williams, et al. Optimal monetary policy under uncertainty: a markov jump-linear-quadratic approach. *Federal Reserve Bank of St. Louis Review*, 90(4):275–293, 2008.

[115] Howard J Chizeck, Alan S Willsky, and D Castanon. Discrete-time markovian-jump linear quadratic optimal control. *International Journal of Control*, 43(1):213–231, 1986.

[116] Oswaldo Luiz Valle Costa, Marcelo Dutra Fragoso, and Ricardo Paulino Marques. *Discrete-time Markov jump linear systems.* Springer, 2006.

[117] Peter E Caines and Ji-Feng Zhang. On the adaptive control of jump parameter systems via nonlinear filtering. *SIAM J. Control Optim.*, 33(6):1758–1777, 1995.

[118] F Xue and L Guo. Necessary and sufficient conditions for adaptive stablizability of jump linear systems. *Communications in Information and Systems*, 1(2):205–224, 2001.

[119] Konstantinos Gatsis and George J Pappas. Statistical learning for analysis of networked control systems over unknown channels. *Automatica*, 125:109386, 2021.

[120] Mathijs Schuurmans, Pantelis Sopasakis, and Panagiotis Patrinos. Safe learning-based control of stochastic jump linear systems: a distributionally robust approach. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6498–6503. IEEE, 2019.

[121] Necmiye Ozay, Mario Sznaier, Constantino M Lagoa, and Octavia I Camps. A sparsification approach to set membership identification of switched affine systems. *IEEE Transactions on Automatic Control*, 57(3):634–648, 2011.

[122] F Lauer and G Bloch. Hybrid system identification: Theory and algorithms for learning switching models, vol. 478. *Cham, Switzerland: Springer*, 2018.

[123] Pedro Hespanhol and Anil Aswani. Statistical consistency of set-membership estimator for linear systems. *IEEE Control Systems Letters*, 4(3):668–673, 2020.

[124] Jitendra Tugnait. Adaptive estimation and identification for discrete systems with markov jump parameters. *IEEE Transactions on Automatic control*, 27(5):1054–1065, 1982.

[125] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Bayesian nonparametric methods for learning markov switching processes. *IEEE Signal Processing Magazine*, 27(6):43–54, 2010.

[126] Sahin Lale, Oguzhan Teke, Babak Hassibi, and Anima Anandkumar. Stability and identification of random asynchronous linear time-invariant systems. In *Learning for Dynamics and Control*, pages 651–663. PMLR, 2021.

[127] Robert G Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[128] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[129] Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9. PMLR, 2018.

[130] 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.

[131] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *NeurIPS*, pages 2645–2653, 2012.

[132] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

[133] Anru Zhang and Mengdi Wang. Spectral state compression of markov processes. *IEEE transactions on information theory*, 66(5):3202–3231, 2019.

[134] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[135] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2016.

[136] Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 16–21. IEEE, 2008.

[137] Christos Thrampoulidis and Ankit Singh Rawat. The generalized lasso for sub-gaussian measurements with dithered quantization. *arXiv preprint arXiv:1807.06976*, 2018.

[138] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[139] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[140] Mario Srouji, Jian Zhang, and Ruslan Salakhutdinov. Structured control nets for deep reinforcement learning. In *International Conference on Machine Learning*, pages 4742–4751. PMLR, 2018.

[141] Ohad Shamir. Are resnets provably better than linear predictors? In *Advances in neural information processing systems*, pages 507–516, 2018.

[142] Tapani Raiko, Harri Valpola, and Yann LeCun. Deep learning made easier by linear transformations in perceptrons. In *Artificial intelligence and statistics*, pages 924–932, 2012.

[143] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[144] Ravi Ganti, Nikhil Rao, Rebecca M Willett, and Robert Nowak. Learning single index models in high dimensions. *arXiv preprint arXiv:1506.08910*, 2015.

[145] Samet Oymak and Mahdi Soltanolkotabi. Fast and reliable parameter estimation from nonlinear observations. *SIAM Journal on Optimization*, 27(4):2276–2300, 2017.

[146] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

[147] Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

[148] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

[149] Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

[150] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling theory, a renaissance*, pages 3–66. Springer, 2015.

[151] Sjoerd Dirksen, Hans Christian Jung, and Holger Rauhut. One-bit compressed sensing with partial gaussian circulant matrices. *arXiv preprint arXiv:1710.03287*, 2017.

[152] Sjoerd Dirksen and Shahar Mendelson. Robust one-bit compressed sensing with partial circulant matrices. *arXiv preprint arXiv:1812.06719*, 2018.

[153] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

[154] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[155] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[156] Raja Giryes, Yonina C Eldar, Alex M Bronstein, and Guillermo Sapiro. Tradeoffs between convergence speed and reconstruction accuracy in inverse problems. *IEEE Transactions on Signal Processing*, 66(7):1676–1690, 2018.

[157] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[158] Martin Genzel. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *IEEE Transactions on Information Theory*, 63(3):1601–1619, 2017.

[159] Sjoerd Dirksen and Shahar Mendelson. Robust one-bit compressed sensing with non-gaussian measurements. *arXiv preprint arXiv:1805.09409*, 2018.

[160] Laurent Jacques and Valerio Cambareri. Time for dithering: fast and quantized random embeddings via the restricted isometry property. *Information and Inference: A Journal of the IMA*, 6(4):441–476, 2017.

[161] Chunlei Xu and Laurent Jacques. Quantized compressive sensing with rip matrices: The benefit of dithering. *arXiv preprint arXiv:1801.05870*, 2018.

[162] Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina Eldar, and Tong Zhang. Sparse nonlinear regression: Parameter estimation under nonconvexity. In *International Conference on Machine Learning*, pages 2472–2481, 2016.

[163] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pages 3851–3860, 2017.

[164] Zhuoran Yang, Krishna Balasubramanian, Zhaoran Wang, and Han Liu. Learning non-gaussian multi-index model via second-order stein's method. *Advances in Neural Information Processing Systems*, 2017.

[165] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. On stein's identity and near-optimal estimation in high-dimensional index models. *arXiv preprint arXiv:1709.08795*, 2017.

[166] H. L. Yap, M. B. Wakin, and C. J. Rozell. Stable manifold embeddings with structured random matrices. *IEEE Journal on Selected Topics in Signal Processing,*, 7(4):720–730, 2013.

[167] Martin Genzel and Gitta Kutyniok. The mismatch principle: Statistical learning under large model uncertainties. *arXiv preprint arXiv:1808.06329*, 2018.

[168] Michel Talagrand. Gaussian processes and the generic chaining. In *Upper and Lower Bounds for Stochastic Processes*, pages 13–73. Springer, 2014.

[169] Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

[170] Vidyashankar Sivakumar, Arindam Banerjee, and Pradeep K Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *Advances in neural information processing systems*, pages 2206–2214, 2015.

[171] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[172] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[173] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 2014.

[174] Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.

[175] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[176] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

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

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

Appendix A

Proof of Corollary 15

A.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 55 **((C_ρ, ρ)-stability)** *Fix excitations $(\boldsymbol{z}_t)_{t=0}^\infty$ and noise $(\boldsymbol{w}_t)_{t=0}^\infty$. Denote the state sequence (2.4.1) ($\phi = \boldsymbol{I}_n$) resulting from initial state $\boldsymbol{x}_0 = \boldsymbol{\alpha}$, $(\boldsymbol{z}_\tau)_{\tau=0}^t$ and $(\boldsymbol{w}_\tau)_{\tau=0}^t$ by $\boldsymbol{x}_t(\boldsymbol{\alpha})$. Suppose $\rho(\boldsymbol{A}_\star) < 1$. Then, there exists $C_\rho \geq 1$ and $\rho \in (\rho(\boldsymbol{A}_\star), 1)$ such that $\|\boldsymbol{x}_t(\boldsymbol{\alpha}) - \boldsymbol{x}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2}$.*

Proof. To begin, consider the difference,

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

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

$$\|\boldsymbol{x}_t(\boldsymbol{\alpha}) - \boldsymbol{x}_t(0)\|_{\ell_2} = \|\boldsymbol{A}_\star^t(\boldsymbol{\alpha} - 0)\|_{\ell_2} \leq \|\boldsymbol{A}_\star^t\| \|\boldsymbol{\alpha}\|_{\ell_2}. \tag{A.1.1}$$

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, $\|\boldsymbol{x}_t(\boldsymbol{\alpha}) - \boldsymbol{x}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2}$. This completes the proof. ■

A.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 [31] that bounds the Euclidean norm of a subgaussian vector.

Lemma 56 *Let $\boldsymbol{a} \in \mathbb{R}^n$ be a zero-mean subgaussian random vector with $\|\boldsymbol{a}\|_{\psi_2} \leq L$. Then for any $m \geq n$, there exists $C > 0$ such that*

$$\mathbb{P}(\|\boldsymbol{a}\|_{\ell_2} \leq CL\sqrt{m}) \geq 1 - 2\exp(-100m). \tag{A.2.1}$$

To apply Lemma 56, we require the subgaussian norm of the state vector $\boldsymbol{x}_t$ and the concatenated vector $\boldsymbol{x}_t$. We will do that by first bounding the corresponding covariance matrices as follows.

Theorem 57 (Covariance bounds) *Consider the LDS in (2.4.1) 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 (2.4.4). Then, the covariance matrix of the vectors $\boldsymbol{x}_t$ and $\boldsymbol{h}_t := [\boldsymbol{x}_t^\top \ \boldsymbol{z}_t^\top]^\top$ satisfies*

$$\lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_n \preceq \boldsymbol{\Sigma}[\boldsymbol{x}_t] \preceq \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top)\boldsymbol{I}_n, \tag{A.2.2}$$

$$(1 \wedge \lambda_{\min}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p} \preceq \boldsymbol{\Sigma}[\boldsymbol{h}_t] \preceq (1 \vee \lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t\boldsymbol{F}_t^\top))\boldsymbol{I}_{n+p}, \tag{A.2.3}$$

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

$$\boldsymbol{x}_t = \underbrace{\sum_{i=0}^{t-1} \boldsymbol{A}_\star^{t-1-i} \boldsymbol{B}_\star \boldsymbol{z}_i}_{\boldsymbol{g}_t} + \underbrace{\sum_{i=0}^{t-1} \boldsymbol{A}_\star^{t-1-i} \boldsymbol{w}_i}_{\boldsymbol{\omega}_t}. \tag{A.2.4}$$

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

$$\begin{aligned}
\boldsymbol{\Sigma}[\boldsymbol{x}_t] &= \boldsymbol{\Sigma}[\boldsymbol{g}_t + \boldsymbol{\omega}_t] = \boldsymbol{\Sigma}[\boldsymbol{g}_t] + \boldsymbol{\Sigma}[\boldsymbol{\omega}_t] = \mathbb{E}[\boldsymbol{g}_t \boldsymbol{g}_t^\top] + \mathbb{E}[\boldsymbol{\omega}_t \boldsymbol{\omega}_t^\top] \\
&= \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} (\boldsymbol{A}_\star^i) \boldsymbol{B}_\star \, \mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_j^\top] \boldsymbol{B}_\star^\top (\boldsymbol{A}_\star^j)^\top + \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} (\boldsymbol{A}_\star^i) \, \mathbb{E}[\boldsymbol{w}_i \boldsymbol{w}_j^\top] (\boldsymbol{A}_\star^j)^\top \\
&\overset{(a)}{=} \sum_{i=0}^{t-1} (\boldsymbol{A}_\star^i) \boldsymbol{B}_\star \boldsymbol{B}_\star^\top (\boldsymbol{A}_\star^i)^\top + \sigma^2 \sum_{i=0}^{t-1} (\boldsymbol{A}_\star^i)(\boldsymbol{A}_\star^i)^\top,
\end{aligned} \tag{A.2.5}$$

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 $\boldsymbol{x}_t$ and the concatenated vector $\boldsymbol{h}_t := [\boldsymbol{x}_t^\top \ \boldsymbol{z}_t^\top]^\top$.

$$\lambda_{\min}(\boldsymbol{G}_t \boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t \boldsymbol{F}_t^\top) \boldsymbol{I}_n \preceq \boldsymbol{\Sigma}[\boldsymbol{x}_t] \preceq \lambda_{\max}(\boldsymbol{G}_t \boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t \boldsymbol{F}_t^\top) \boldsymbol{I}_n, \tag{A.2.6}$$

$$(1 \wedge \lambda_{\min}(\boldsymbol{G}_t \boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t \boldsymbol{F}_t^\top)) \boldsymbol{I}_{n+p} \preceq \boldsymbol{\Sigma}[\boldsymbol{h}_t] \preceq (1 \vee \lambda_{\max}(\boldsymbol{G}_t \boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t \boldsymbol{F}_t^\top)) \boldsymbol{I}_{n+p}, \tag{A.2.7}$$

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 $\|\boldsymbol{x}_t\|_{\psi_2} \lesssim \sqrt{\|\boldsymbol{\Sigma}[\boldsymbol{x}_t]\|} \le \sqrt{\lambda_{\max}(\boldsymbol{G}_t \boldsymbol{G}_t^\top + \sigma^2 \boldsymbol{F}_t \boldsymbol{F}_t^\top)}$ and

$\|\boldsymbol{h}_t\|_{\psi_2} \lesssim \sqrt{\boldsymbol{\Sigma}[\boldsymbol{h}_t]} \leq 1 \vee \sqrt{\lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)}$. Combining these with Lemma 56, we find that, with probability at least $1 - 4T\exp(-100n)$, for all $1 \leq t \leq T$, we have $\|\boldsymbol{x}_t\|_{\ell_2} \leq c\sqrt{\beta_+ n}$ and $\|\boldsymbol{h}_t\|_{\ell_2} \leq c_0\sqrt{\beta_+(n+p)}$, where we set $\beta_+ = 1 \vee \max_{1\leq t\leq T}\lambda_{\max}(\boldsymbol{G}_t\boldsymbol{G}_t^\top + \sigma^2\boldsymbol{F}_t\boldsymbol{F}_t^\top)$. This verifies Lemma 6 and consequently Assumption 2.

A.3 Verification of Assumption 3

Recall that, we define the following concatenated vector/matrix for linear dynamical systems: $\boldsymbol{h}_t := [\boldsymbol{x}_t^\top \ \boldsymbol{z}_t^\top]^\top$ 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,

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k), \quad \text{where} \quad \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) := \frac{1}{2}\mathbb{E}[(\boldsymbol{x}_L[k] - \boldsymbol{\theta}_k^\top\boldsymbol{h}_{L-1})^2]. \tag{A.3.1}$$

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 58 (One-point convexity & smoothness) *Consider the setup of Theorem 57 and the auxiliary loss given by (A.3.1). 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 $1 \leq k \leq n$, the gradient $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$ satisfies,*

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

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

Proof. To begin, we take the gradient of the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ (A.3.1) to get $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) = \mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) - \boldsymbol{h}_{L-1}\boldsymbol{w}_{L-1}[k]]$. Note that, $\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{w}_{L-1}[k]] = 0$ for linear dynamical systems because $\boldsymbol{w}_{L-1}$ and $\boldsymbol{h}_{L-1}$ are independent and we have $\mathbb{E}[\boldsymbol{w}_{L-1}] = \mathbb{E}[\boldsymbol{h}_{L-1}] = 0$.

Therefore, using Theorem 57 with $t = L - 1$, we get the following one point convexity bound,

$$\begin{aligned}\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\right\rangle &= \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top](\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)\right\rangle, \\ &\geq \gamma_- \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2, \qquad \text{(A.3.2)}\end{aligned}$$

where $\gamma_- := 1 \wedge \lambda_{\min}(\boldsymbol{\Gamma}_{L-1})$. Similarly, setting $\gamma_+ := 1 \vee \lambda_{\max}(\boldsymbol{\Gamma}_{L-1})$, we also have

$$\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \leq \|\,\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top]\|\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} \leq \gamma_+\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}. \qquad \text{(A.3.3)}$$

This completes the proof. ■

A.4 Verification of Assumption 4

Let $\mathcal{S} := (\boldsymbol{x}_L^{(i)}, \boldsymbol{x}_{L-1}^{(i)}, \boldsymbol{z}_{L-1}^{(i)})_{i=1}^N$ be N i.i.d. copies of $(\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})$ generated from N i.i.d. trajectories of the system (2.4.1) with $\phi = \boldsymbol{I}_n$. Let $\boldsymbol{h}_{L-1}^{(i)} := [\boldsymbol{x}_{L-1}^{(i)\top}\ \boldsymbol{z}_{L-1}^{(i)\top}]^\top$ 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

$$\hat{\mathcal{L}}_\mathcal{S}(\boldsymbol{\Theta}) = \sum_{k=1}^n \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k), \quad \text{where} \quad \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) := \frac{1}{2N}\sum_{i=1}^N (\boldsymbol{x}_L^{(i)}[k] - \boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1}^{(i)})^2. \qquad \text{(A.4.1)}$$

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 59 (Lipschitz gradient) *Consider the same setup of Theorem 57. Consider the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ and its finite sample approximation $\hat{\mathcal{L}}_{k,\mathcal{S}}$ from (A.3.1) and (A.4.1) respectively. Let $\gamma_+ > 0$ be as in Lemma 58. 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 $1 \leq k \leq n$, we have*

$$\max(\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) - \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2}, \|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) - \nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2}) \leq 2\gamma_+\|\boldsymbol{\theta}_k - \boldsymbol{\theta}'_k\|_{\ell_2}. \qquad \text{(A.4.2)}$$

Proof. To begin, recall the auxiliary loss from (A.3.1). We have that

$$\begin{aligned}\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)-\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}'_k)\|_{\ell_2} &= \|\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top](\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^\star)-\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top](\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k^\star)\|_{\ell_2},\\ &\le \|\,\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top]\|\|\boldsymbol{\theta}_k-\boldsymbol{\theta}'_k\|_{\ell_2},\\ &\le \gamma_+\|\boldsymbol{\theta}_k-\boldsymbol{\theta}'_k\|_{\ell_2}. \qquad \text{(A.4.3)}\end{aligned}$$

To obtain a similar result for the finite sample loss $\hat{\mathcal{L}}_{k,\mathcal{S}}$, we use Corollary 5.50 from [177] 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 $\boldsymbol{h}_{L-1}^{(i)}$ with $t=10,\varepsilon=1$ shows that, for $N\gtrsim n+p$, with probability at least $1-2\exp(-100(n+p))$, we have

$$\|\frac{1}{N}\sum_{i=1}^N\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top-\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top]\|\le\gamma_+. \qquad \text{(A.4.4)}$$

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

$$\begin{aligned}&\|\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k)-\nabla\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}'_k)\|_{\ell_2}\\ &\qquad\le\|\frac{1}{N}\sum_{i=1}^N\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^\star)-\frac{1}{N}\sum_{i=1}^N\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top(\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k^\star)\|_{\ell_2},\\ &\qquad\le\|\frac{1}{N}\sum_{i=1}^N\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top\|\|\boldsymbol{\theta}_k-\boldsymbol{\theta}'_k\|_{\ell_2},\\ &\qquad\le\left[\|\,\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top]\|+\|\frac{1}{N}\sum_{i=1}^N\boldsymbol{h}_{L-1}^{(i)}(\boldsymbol{h}_{L-1}^{(i)})^\top-\mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top]\|\right]\|\boldsymbol{\theta}_k-\boldsymbol{\theta}'_k\|_{\ell_2},\\ &\qquad\le 2\gamma_+\|\boldsymbol{\theta}_k-\boldsymbol{\theta}'_k\|_{\ell_2}, \qquad \text{(A.4.5)}\end{aligned}$$

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

A.5 Verification of Assumption 5

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

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}, (\boldsymbol{x}_L, \boldsymbol{h}_{L-1})) &= \sum_{k=1}^{n} \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{h}_{L-1})), \\ \text{where} \quad \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{h}_{L-1})) &:= \frac{1}{2}(\boldsymbol{x}_L[k] - \boldsymbol{\theta}_k^\top \boldsymbol{h}_{L-1})^2. \end{aligned} \tag{A.5.1}$$

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

Lemma 60 (Subexponential gradient) *Consider the same setup of Theorem 57. Let $\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{h}_{L-1}))$ be as defined in (A.5.1) and $\gamma_+ > 0$ be as in lemma 58. Then, at any point $\boldsymbol{\Theta}$, for all $1 \le k \le n$, we have*

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

Proof. Using standard bounds on the subgaussian norm of a random vector, we find that $\|\boldsymbol{h}_{L-1}\|_{\psi_2} \lesssim \sqrt{\boldsymbol{\Sigma}[\boldsymbol{h}_{L-1}]} \le \sqrt{\gamma_+}$, where $\gamma_+ > 0$ is as defined in Lemma 58. Combining this with $\|\boldsymbol{w}_{L-1}[k]\|_{\psi_2} \le \sigma$, we get the following subexponential norm bound,

$$\begin{aligned}&\|\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{h}_{L-1})) - \mathbb{E}[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{h}_{L-1}))]\|_{\psi_1} \\ &\qquad = \|(\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top - \mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top])(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star) - \boldsymbol{h}_{L-1}\boldsymbol{w}_{L-1}[k]\|_{\psi_1}, \\ &\qquad \le \|(\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top - \mathbb{E}[\boldsymbol{h}_{L-1}\boldsymbol{h}_{L-1}^\top])(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)\|_{\psi_1} + \|\boldsymbol{h}_{L-1}\boldsymbol{w}_{L-1}[k]\|_{\psi_1}, \\ &\qquad \lesssim \gamma_+ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma\sqrt{\gamma_+}, \end{aligned} \tag{A.5.2}$$

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.6 Finalizing the 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 61 *Consider the same setup of Theorem 57. 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:*

$$\|\nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t)\|_{\ell_2} = \|\boldsymbol{h}_t\|_{\ell_2} \le c_0\sqrt{\beta_+(n+p)} =: C_{\tilde{\phi}}, \tag{A.6.1}$$

$$\|\nabla_{\boldsymbol{h}_t}\nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^\top \boldsymbol{h}_t)\| = \|\boldsymbol{I}_{n+p}\| \le 1 =: D_{\tilde{\phi}}, \tag{A.6.2}$$

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,

$$N \gtrsim \kappa^2\log^2(3(2\gamma_+)N/\gamma_+ + 3)(n+p) \Leftrightarrow N \gtrsim \kappa^2\log^2(6N+3)(n+p), \tag{A.6.3}$$

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

$$\begin{aligned} & L \ge 1 + \big[\log(c_0(n+p)\sqrt{\beta_+}C_\rho\sqrt{N/(n+p)}) + \log(c/\sqrt{\gamma_+} \vee c\sqrt{\beta_+(n+p)}/\gamma_+)\big]/\log(\rho^{-1}) \\ \Longleftarrow \quad & L \ge \Big\lceil 1 + \frac{\log(CC_\rho\beta_+N(n+p)/\gamma_+)}{1-\rho}\Big\rceil, \end{aligned} \tag{A.6.4}$$

where, $C > 0$ is a constant. Finally, given the trajectory length $T \gtrsim L(N+1)$, where N and L are given by (A.6.3) and (A.6.4) 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 $1 \le k \le n$, all gradient descent iterates $\mathbf{\Theta}^{(\tau)}$ on $\hat{\mathcal{L}}$ satisfy

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \le (1 - \frac{\gamma_-^2}{128\gamma_+^2})^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{5c}{\gamma_-}\sigma\sqrt{\gamma_+}\log(6N+3)\sqrt{\frac{n+p}{N}}. \tag{A.6.5}$$

We remark that, choosing $N \gtrsim \kappa^2\log^2(6N+3)(n+p)$, the residual term in (A.6.5) 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. ■

Appendix B

Proof of Corollary 16

B.1 Verification of Assumption 2

Lemma 62 *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,

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

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

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

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

■

Lemma 63 (Bounded states) *Suppose, the nonlinear system* (2.4.2) *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* $\boldsymbol{x}_0 = 0$, *for all* $0 \leq t \leq T$, *we have:*

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

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

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

Proof. (a) Given $\|\boldsymbol{z}_t\|_{\psi_2} \leq 1$ and $\|\boldsymbol{w}_t\|_{\psi_2} \leq \sigma$, we use Lemma 56 to obtain $\mathbb{P}(\|\boldsymbol{z}_t\|_{\ell_2} \lesssim \sqrt{n}) \geq 1-2T\exp(-100n)$ and $\mathbb{P}(\|\boldsymbol{w}_t\|_{\ell_2} \lesssim \sigma\sqrt{n}) \geq 1-2T\exp(-100n)$ for all $0 \leq t \leq T-1$. 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 $\boldsymbol{x}_t$.

(b) Recall that $\boldsymbol{x}_0 = 0$. We claim that $\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] \leq \beta_+^2 n(1-\rho^t)^2$, 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 $\mathbb{E}[\|\boldsymbol{z}_t\|_{\ell_2}] \leq \sqrt{n}$ and $\mathbb{E}[\|\boldsymbol{w}_t\|_{\ell_2}] \leq \sigma\sqrt{n}$. Using this results, we show that $\boldsymbol{x}_1$ satisfies the following bound and obeys the induction

$$\mathbb{E}[\|\boldsymbol{x}_1\|_{\ell_2}^2] = \mathbb{E}[\|\phi(0) + \boldsymbol{z}_0 + \boldsymbol{w}_0\|_{\ell_2}^2] \le (1+\sigma^2)n \le C_\rho^2(1+\sigma)^2 n = \beta_+^2 n(1-\rho^1)^2.$$

This implies $\mathbb{E}[\|\boldsymbol{x}_1\|_{\ell_2}] \le \beta_+\sqrt{n}(1-\rho^1)$ as well. Suppose the bound holds until $t-1$, that is, $\mathbb{E}[\|\boldsymbol{x}_{t-1}\|_{\ell_2}^2] \le \beta_+^2 n(1-\rho^{t-1})^2$ (which also means $\mathbb{E}[\|\boldsymbol{x}_{t-1}\|_{\ell_2}] \le \beta_+\sqrt{n}(1-\rho^{t-1})$). We now apply the induction as follows: First observe that $\mathbb{E}[\|\boldsymbol{x}_{t,L}\|_{\ell_2}]$ obeys the same upper bound as $\mathbb{E}[\|\boldsymbol{x}_L\|_{\ell_2}]$ by construction. To proceed, recalling (2.3.7), we get the following by induction

$$\begin{aligned}
&\|\boldsymbol{x}_t - \boldsymbol{x}_{t,t-1}\|_{\ell_2} \le C_\rho \rho^{t-1}\|\boldsymbol{x}_1\|_{\ell_2} \\
\implies \quad & \|\boldsymbol{x}_t\|_{\ell_2} \le C_\rho\rho^{t-1}\|\boldsymbol{x}_1\|_{\ell_2} + \|\boldsymbol{x}_{t,t-1}\|_{\ell_2}, \\
\implies \quad & \|\boldsymbol{x}_t\|_{\ell_2}^2 \le (C_\rho\rho^{t-1}\|\boldsymbol{x}_1\|_{\ell_2} + \|\boldsymbol{x}_{t,t-1}\|_{\ell_2})^2, \\
\implies & \mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2] \le C_\rho^2\rho^{2(t-1)}\,\mathbb{E}[\|\boldsymbol{x}_1\|_{\ell_2}^2] + \mathbb{E}[\|\boldsymbol{x}_{t-1}\|_{\ell_2}^2] + 2C_\rho\rho^{t-1}\,\mathbb{E}[\|\boldsymbol{x}_1\|_{\ell_2}]\,\mathbb{E}[\|\boldsymbol{x}_{t-1}\|_{\ell_2}], \\
&\overset{\text{(a)}}{\le} C_\rho^2\rho^{2(t-1)}(1+\sigma)^2 n + \beta_+^2 n(1-\rho^{t-1})^2 + 2nC_\rho\rho^{t-1}(1+\sigma)\beta_+(1-\rho^{t-1}), \\
&\overset{\text{(b)}}{\le} \beta_+^2 n(\rho^{2(t-1)}(1-\rho^1)^2 + (1-\rho^{t-1})^2 + 2\rho^{t-1}(1-\rho^{t-1})(1-\rho^1)), \\
&= \beta_+^2 n[\rho^{2t-2}(1+\rho^2-2\rho) + 1 + \rho^{2t-2} - 2\rho^{t-1} + (2\rho^{t-1} - 2\rho^{2t-2})(1-\rho)], \\
&= \beta_+^2 n(1+\rho^{2t} - 2\rho^t), \\
&= \beta_+^2 n(1-\rho^t)^2, \qquad \text{(B.1.1)}
\end{aligned}$$

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

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

$$\implies \mathbb{P}(\|\boldsymbol{z}_t\|_{\ell_2} \geq \sqrt{2cn}(1+t)) \leq \exp(-cnt^2), \tag{B.1.2}$$

$$\text{and} \quad \mathbb{P}(\|\boldsymbol{w}_t\|_{\ell_2} \geq \sigma\sqrt{2cn}(1+t)) \leq \exp(-cnt^2). \tag{B.1.3}$$

To proceed, let $X = \|\boldsymbol{x}_t\|_{\ell_2}$ and $Y = \sum_{\tau=0}^{t-1} C_\rho \rho^\tau (\|\boldsymbol{z}_t\|_{\ell_2} + \|\boldsymbol{w}_t\|_{\ell_2})$ and note that $X \leq Y$. Now, using (B.1.2), (B.1.3) and union bounding over all $0 \leq t \leq T-1$, we get the following high probability upper bound on Y, that is,

$$\mathbb{P}(Y \geq \sum_{\tau=0}^{t-1} C_\rho \rho^\tau \sqrt{2cn}(1+\sigma)(1+t)) \leq 2T\exp(-cnt^2),$$

$$\implies \quad \mathbb{P}(Y \geq C_\rho\sqrt{10n\log(2T)}(1+t)(1+\sigma)/(1-\rho)) \leq \exp(-5nt^2), \tag{B.1.4}$$

where we choose $c = 5\log(2T)$ to get the final concentration bound of Y. Finally using this bound in Lemma 62, we get

$$\mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^3] \leq 32\beta_+^3(\log(2T)n)^{3/2}, \tag{B.1.5}$$

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

B.2 Verification of Assumption 3

Theorem 64 *Suppose the nonlinear system (2.4.2) 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 63. Then, the matrix $\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]$ satisfies*

$$(1+\sigma^2)\boldsymbol{I}_n \preceq \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top] \preceq \beta_+^2 n\boldsymbol{I}_n. \tag{B.2.1}$$

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

$$\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top] \preceq \mathbb{E}[\|\boldsymbol{x}_t\boldsymbol{x}_t^\top\|]\boldsymbol{I}_n \preceq \mathbb{E}[\|\boldsymbol{x}_t\|_{\ell_2}^2]\boldsymbol{I}_n \preceq \beta_+^2 n\boldsymbol{I}_n, \tag{B.2.2}$$

where we get the last inequality by applying Lemma 63. To get a lower bound, note that $\boldsymbol{\Sigma}[\boldsymbol{x}_t] = \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top] - \mathbb{E}[\boldsymbol{x}_t]\mathbb{E}[\boldsymbol{x}_t]^\top$. Since, all of these matrices are positive semi-definite, we get the following lower bound,

$$\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top] \succeq \boldsymbol{\Sigma}[\boldsymbol{x}_t] = \boldsymbol{\Sigma}[\phi(\boldsymbol{\Theta}_\star\boldsymbol{x}_{t-1}) + \boldsymbol{z}_{t-1} + \boldsymbol{w}_{t-1}] \succeq \boldsymbol{\Sigma}[\boldsymbol{z}_{t-1} + \boldsymbol{w}_{t-1}] = (1+\sigma^2)\boldsymbol{I}_n. \tag{B.2.3}$$

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

■

To verify Assumption 3 for the nonlinear system (2.4.2), denoting the k_{th} row of $\boldsymbol{\Theta}$ by $\boldsymbol{\theta}_k^\top$, the auxiliary loss for the nonlinear system (2.4.2) is given by,

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\Theta}) = \sum_{k=1}^n \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) \quad \text{where} \quad \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) := \frac{1}{2}\mathbb{E}[(\boldsymbol{x}_L[k] - \phi(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}) - \boldsymbol{z}_{L-1}[k])^2]. \tag{B.2.4}$$

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

Lemma 65 (One-point convexity & smoothness) *Consider the setup of Theorem 64 and the auxiliary loss given by (B.2.4). 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 63. Then, for all $1 \leq k \leq n$, the gradients $\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)$ satisfy,*

$$\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\rangle \geq \gamma^2(1+\sigma^2)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2,$$

$$\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} \leq \beta_+^2 n\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}.$$

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 (B.2.4), we have

$$\begin{aligned}
\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) &= \mathbb{E}[(\phi(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1}) - \boldsymbol{w}_{L-1}[k])\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}],\\
&= \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}]\\
&- \mathbb{E}[\boldsymbol{w}_{L-1}[k]\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}],\\
&= \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)], \qquad \text{(B.2.5)}
\end{aligned}$$

where $\mathbb{E}[\boldsymbol{w}_{L-1}[k]\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}] = 0$ because $\boldsymbol{x}_{L-1}$ 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,

$$\begin{aligned}
\left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\right\rangle &= \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)]\right\rangle,\\
&\geq \gamma^2 \left\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star, \mathbb{E}[\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top](\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)\right\rangle,\\
&\geq \gamma^2(1+\sigma^2)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}^2. \qquad \text{(B.2.6)}
\end{aligned}$$

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

$$\begin{aligned}
\|\nabla\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\|_{\ell_2} &= \|\mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star)]\|_{\ell_2},\\
&\leq \mathbb{E}[\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top\|]\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2},\\
&\leq \mathbb{E}[\|\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top\|]\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}.\\
&\leq \beta_+^2 n\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}, \qquad \text{(B.2.7)}
\end{aligned}$$

where β_+ is as defined in Lemma 63. This completes the proof. ■

B.3 Verification of Assumption 4

Let $\mathcal{S} = (\boldsymbol{x}_L^{(i)}, \boldsymbol{x}_{L-1}^{(i)}, \boldsymbol{z}_{L-1}^{(i)})_{i=1}^N$ be N i.i.d. copies of $(\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})$ generated from N i.i.d. trajectories of the system (2.4.2). Then, the finite sample approximation of the auxiliary loss $\mathcal{L}_{\mathcal{D}}$ is given by,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\Theta}) = \sum_{k=1}^{n} \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) \text{ where } \hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k) := \frac{1}{2N} \sum_{i=1}^{N} (\boldsymbol{x}_L^{(i)}[k] - \phi(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1}^{(i)}) - \boldsymbol{z}_{L-1}^{(i)}[k])^2. \quad \text{(B.3.1)}$$

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 66 (Lipschitz gradient) *Consider the same setup of Theorem 64. Consider the auxiliary loss $\mathcal{L}_{k,\mathcal{D}}$ and its finite sample approximation $\hat{\mathcal{L}}_{k,\mathcal{S}}$ from (B.2.4) and (B.3.1) respectively. Suppose, ϕ has bounded first and second derivatives, that is, $|\phi'|, |\phi''| \le 1$. Let β_+ be as in Lemma 63. 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*

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

Proof. To begin recall that, $\nabla \mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k) = \mathbb{E}[(\phi(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1}) - \phi(\boldsymbol{\theta}_k^{\star\top} \boldsymbol{x}_{L-1}))\phi'(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}]$. 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,

$$\begin{aligned}
\|\nabla^2\mathcal{L}_{k,\mathcal{D}}(\boldsymbol{\theta}_k)\| &= \|\,\mathbb{E}[(\phi(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1}))\phi''(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top] \\
&\quad + \mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top]\|, \\
&\le \mathbb{E}[\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}, \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi''(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top\|] \\
&\quad + \mathbb{E}[\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top\|], \\
&\le \mathbb{E}[\|(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top\|] + \mathbb{E}[\|\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top\|], \\
&\le \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}\,\mathbb{E}[\|\boldsymbol{x}_{L-1}\|_{\ell_2}^3] + \mathbb{E}[\|\boldsymbol{x}_{L-1}\|_{\ell_2}^2], \\
&\lesssim \beta_+^3(\log(2T)n)^{3/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \beta_+^2 n, \qquad \text{(B.3.2)}
\end{aligned}$$

where we get the last inequality by applying Lemma 63. Similarly, to bound the Lipschitz constant of the empirical gradient

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

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

$$\begin{aligned}
\|\nabla^2\hat{\mathcal{L}}_{k,\mathcal{S}}(\boldsymbol{\theta}_k)\| &\le \frac{1}{N}\sum_{i=1}^{N}\|(\phi(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}^{(i)}) - \phi(\boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1}^{(i)}) - \boldsymbol{w}_{L-1}^{(i)}[k])\phi''(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}^{(i)})\boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top\| \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}^{(i)})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}^{(i)})\boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top\|, \\
&\overset{\text{(a)}}{\le} \frac{1}{N}\sum_{i=1}^{N}[\|(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1}^{(i)} - \boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1}^{(i)})\boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top\| + (1 + |\boldsymbol{w}_{L-1}^{(i)}[k]|)\|\boldsymbol{x}_{L-1}^{(i)}(\boldsymbol{x}_{L-1}^{(i)})^\top\|], \\
&\le \frac{1}{N}\sum_{i=1}^{N}[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2}\|\boldsymbol{x}_{L-1}^{(i)}\|_{\ell_2}^3 + (1 + |\boldsymbol{w}_{L-1}^{(i)}[k]|)\|\boldsymbol{x}_{L-1}^{(i)}\|_{\ell_2}^2], \\
&\lesssim \beta_+^3 n^{3/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + (1+\sigma)\beta_+^2 n, \qquad \text{(B.3.3)}
\end{aligned}$$

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 63.

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

■

B.4 Verification of Assumption 5

Given a single sample $(\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})$ from the trajectory of the nonlinear system (2.4.2), the single sample loss is given by,

$$\mathcal{L}(\boldsymbol{\Theta}, (\boldsymbol{x}_L, \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1})) = \sum_{k=1}^{n} \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}[k])),$$

$$\text{where} \quad \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}[k])) := \frac{1}{2}(\boldsymbol{x}_L[k] - \phi(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{L-1}) - \boldsymbol{z}_{L-1}[k])^2. \tag{B.4.1}$$

Before stating a lemma on bounding the subexponential norm of the gradient of the single sample loss (B.4.1), we will state an intermediate lemma to prove the Lipschitzness of the state vector.

Lemma 67 (Lipschitzness of the state vector) *Suppose the nonlinear system* (2.4.2) *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* $\boldsymbol{x}_0 = 0$*. Fixing all* $\{\boldsymbol{v}_i\}_{i \neq \tau}$ *(i.e., all except* $\boldsymbol{v}_\tau$*),* $\boldsymbol{x}_{t+1}$ *is* $C_\rho \rho^{t-\tau}(1+\sigma^2)^{1/2}$ *Lipschitz function of* $\boldsymbol{v}_\tau$ *for* $0 \le \tau \le t$*.*

Proof. To begin, observe that $\boldsymbol{x}_{t+1}$ is deterministic function of the sequence $\{\boldsymbol{v}_\tau\}_{\tau=0}^{t}$. Fixing all $\{\boldsymbol{v}_i\}_{i \neq \tau}$, we denote $\boldsymbol{x}_{t+1}$ as a function of $\boldsymbol{v}_\tau$ by $\boldsymbol{x}_{t+1}(\boldsymbol{v}_\tau)$. Given a pair of vectors $(\boldsymbol{v}_\tau, \hat{\boldsymbol{v}}_\tau)$, using (C_ρ, ρ)-stability of the nonlinear system (2.4.2), for any $t \geq \tau$, we have

$$\begin{aligned}
\|\boldsymbol{x}_{t+1}(\boldsymbol{v}_\tau) - \boldsymbol{x}_{t+1}(\hat{\boldsymbol{v}}_\tau)\|_{\ell_2} &\leq C_\rho \rho^{t-\tau} \|\boldsymbol{x}_{\tau+1}(\boldsymbol{v}_\tau) - \boldsymbol{x}_{\tau+1}(\hat{\boldsymbol{v}}_\tau)\|_{\ell_2}, \\
&\leq C_\rho \rho^{t-\tau} \|\phi(\boldsymbol{\Theta}_\star \boldsymbol{x}_\tau) + \boldsymbol{z}_\tau + \boldsymbol{w}_\tau - \phi(\boldsymbol{\Theta}_\star \boldsymbol{x}_\tau) - \hat{\boldsymbol{z}}_\tau - \hat{\boldsymbol{w}}_\tau\|_{\ell_2}, \\
&\leq C_\rho \rho^{t-\tau} (\|\boldsymbol{z}_\tau - \hat{\boldsymbol{z}}_\tau\|_{\ell_2} + \sigma \|1/\sigma \boldsymbol{w}_\tau - 1/\sigma \hat{\boldsymbol{w}}_\tau\|_{\ell_2}), \\
&\overset{\text{(a)}}{\leq} C_\rho \rho^{t-\tau} (1+\sigma^2)^{1/2} (\|\boldsymbol{z}_\tau - \hat{\boldsymbol{z}}_\tau\|_{\ell_2}^2 + 1/\sigma^2 \|\boldsymbol{w}_\tau - \hat{\boldsymbol{w}}_\tau\|_{\ell_2}^2)^{1/2}, \\
&\leq C_\rho \rho^{t-\tau} (1+\sigma^2)^{1/2} \|\boldsymbol{v}_\tau - \hat{\boldsymbol{v}}_\tau\|_{\ell_2}, \qquad \text{(B.4.2)}
\end{aligned}$$

where we get (a) by using Cauchy-Schwarz inequality. This implies $\boldsymbol{x}_{t+1}$ is $C_\rho \rho^{t-\tau}(1+\sigma^2)^{1/2}$ Lipschitz function of $\boldsymbol{v}_\tau$ for $0 \leq \tau \leq t$ 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 (B.4.1).

Lemma 68 (Subexponential gradient) *Consider the same setup of Lemma 67. Let $\mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}[k])$ be as in (B.4.1) and $\beta_+ := C_\rho(1+\sigma)/(1-\rho)$. Suppose $|\phi'(x)| \leq 1$ for all $x \in \mathbb{R}$. Then, at any point $\boldsymbol{\Theta}$, for all $1 \leq k \leq n$, we have*

$$\begin{aligned}
&\|\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}[k])) - \mathbb{E}[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k, (\boldsymbol{x}_L[k], \boldsymbol{x}_{L-1}, \boldsymbol{z}_{L-1}[k]))]\|_{\psi_1} \\
&\qquad \lesssim \beta_+^2 \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \sigma \beta_+. \qquad \text{(B.4.3)}
\end{aligned}$$

Proof. We first bound the subgaussian norm of the state vector $\boldsymbol{x}_t$ following [31] 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 $\boldsymbol{x}_t$ is a deterministic function of $\boldsymbol{q}_t$, that is, $\boldsymbol{x}_t = f(\boldsymbol{q}_t)$ 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

$$\|f(\boldsymbol{q}_t) - f(\hat{\boldsymbol{q}}_t)\|_{\ell_2} \leq L_f \|\boldsymbol{q}_t - \hat{\boldsymbol{q}}_t\|_{\ell_2}. \qquad \text{(B.4.4)}$$

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,

$$\|f(\boldsymbol{q}_t) - f(\hat{\boldsymbol{q}}_t)\|_{\ell_2} \le \sum_{i=0}^{t-1} \|f(\boldsymbol{b}_{i+1}) - f(\boldsymbol{b}_i)\|_{\ell_2}. \tag{B.4.5}$$

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 $\boldsymbol{x}_t$ as a function of $\boldsymbol{v}_i$ and using the result of Lemma 67, we have

$$\begin{aligned}
\|f(\boldsymbol{q}_t) - f(\hat{\boldsymbol{q}}_t)\|_{\ell_2} &\le \sum_{i=0}^{t-1} C_\rho \rho^{t-1-i}(1+\sigma^2)^{1/2} \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i\|_{\ell_2}, \\
&\overset{(a)}{\le} C_\rho (1+\sigma^2)^{1/2} \Big(\sum_{i=0}^{t-1} \rho^{2(t-1-i)}\Big)^{1/2} \underbrace{\Big(\sum_{i=0}^{t-1} \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i\|_{\ell_2}^2\Big)^{1/2}}_{\|\boldsymbol{q}_t - \hat{\boldsymbol{q}}_t\|_{\ell_2}}, \\
&\overset{(b)}{\le} \frac{C_\rho (1+\sigma^2)^{1/2}}{(1-\rho^2)^{1/2}} \|\boldsymbol{q}_t - \hat{\boldsymbol{q}}_t\|_{\ell_2},
\end{aligned} \tag{B.4.6}$$

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 $\boldsymbol{x}_t$ 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, $\boldsymbol{x}_t$ is β_K-Lipschitz function of $\boldsymbol{q}_t$, for any fixed unit length vector $\boldsymbol{a}$, $\boldsymbol{a}^\top \boldsymbol{x}_t$ is still β_K-Lipschitz function of $\boldsymbol{q}_t$. This implies $\|\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t]\|_{\psi_2} \lesssim \beta_K$. Secondly, β_K-Lipschitz function of a Gaussian vector obeys the variance inequality $\textbf{var}[\boldsymbol{a}^\top \boldsymbol{x}_t] \le \beta_K^2$ (page 49 of [178]), which implies the covariance bound $\boldsymbol{\Sigma}[\boldsymbol{x}_t] \preceq \beta_K^2 \boldsymbol{I}_n$. Combining these results with $\|\boldsymbol{w}_t[k]\|_{\psi_2} \le \sigma$, we get the following subexponential norm bound,

$$
\begin{aligned}
&\|\nabla\mathcal{L}_k(\boldsymbol{\theta}_k,(\boldsymbol{x}_L[k],\boldsymbol{x}_{L-1},\boldsymbol{z}_{L-1}[k]))-\mathbb{E}[\nabla\mathcal{L}_k(\boldsymbol{\theta}_k,(\boldsymbol{x}_L[k],\boldsymbol{x}_{L-1},\boldsymbol{z}_{L-1}[k]))]\|_{\psi_1}\\
&\qquad\leq\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1},\boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^\star)\\
&\qquad-\mathbb{E}[\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1},\boldsymbol{\theta}_k^{\star\top}\boldsymbol{x}_{L-1})\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{x}_{L-1}\boldsymbol{x}_{L-1}^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^\star)]\|_{\psi_1}\\
&\qquad+\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_{L-1})\boldsymbol{w}_{L-1}[k]\boldsymbol{x}_{L-1}\|_{\psi_1},\\
&\qquad\lesssim\beta_K^2\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^\star\|_{\ell_2}+\sigma\beta_K,\\
&\qquad\lesssim\beta_+^2\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^\star\|_{\ell_2}+\sigma\beta_+, \qquad\qquad (\text{B.4.7})
\end{aligned}
$$

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. ■

B.5 Finalizing the Proof of Corollary 16

Proof. We have verified Assumptions 2, 3, 4 and 5 for the nonlinear system 2.4.2. Hence, we are ready to use Theorem 14 to learn the dynamics $\boldsymbol{\Theta}_\star$ of the nonlinear system (2.4.2) . Before that, we find the values of the system related constants to be used in Theorem 14 as follows.

Remark 69 *Consider the same setup of Lemma 67. Let $\beta_+\geq\beta_K>0$ be as defined in Lemmas 63 and 68 respectively. Then, with probability at least $1-4T\exp(-100n)$, for all $1\leq t\leq T$, $\boldsymbol{\Theta}\in\mathcal{B}^{n\times n}(\boldsymbol{\Theta}_\star,r)$ and $1\leq k\leq n$, the scalars C_ϕ, D_ϕ take the following values.*

$\|\nabla_{\boldsymbol{\theta}_k}\phi(\boldsymbol{\theta}_k^\top\boldsymbol{x}_t)\|_{\ell_2}=\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_t)\boldsymbol{x}_t\|_{\ell_2}\leq\|\boldsymbol{x}_t\|_{\ell_2}\lesssim\beta_+\sqrt{n}=:C_\phi,$

$\|\nabla_{\boldsymbol{x}_t}\nabla_{\boldsymbol{\theta}_k}\phi(\boldsymbol{\theta}_k^\top\boldsymbol{x}_t)\|=\|\phi'(\boldsymbol{\theta}_k^\top\boldsymbol{x}_t)\boldsymbol{I}_n+\phi''(\boldsymbol{\theta}_k^\top\boldsymbol{x}_t)\boldsymbol{x}_t\boldsymbol{\theta}_k^\top\|\lesssim 1+\beta_+\sqrt{n}\|\boldsymbol{\theta}_k\|_{\ell_2}\lesssim 1+\|\boldsymbol{\Theta}_\star\|_F\beta_+\sqrt{n}=:D_\phi$

where without loss of generality we choose $\boldsymbol{\Theta}^{(0)} = 0$ *and* $r = \|\boldsymbol{\Theta}_\star\|_F$. *Furthermore, the Lipschitz constant and the gradient noise coefficients take the following values:* $L_{\mathcal{D}} = c((1+\sigma)\beta_+^2 n + \|\boldsymbol{\Theta}_\star\|_F \beta_+^3 n^{3/2}\log^{3/2}(2T))$, $K = c\beta_+^2$ *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 (2.4.2) via gradient descent,

$$\begin{aligned} & N \gtrsim \frac{\beta_+^4}{\gamma^4(1+\sigma^2)^2}\log^2(3((1+\sigma)\beta_+^2 n + \|\boldsymbol{\Theta}_\star\|_F\beta_+^3 n^{3/2}\log^{3/2}(2T))N/\beta_+^2 + 3)n, \\ \Longrightarrow \quad & N \gtrsim \frac{C_\rho^4}{\gamma^4(1-\rho)^4}\log^2(3(1+\sigma)n + 3\|\boldsymbol{\Theta}_\star\|_F\beta_+ n^{3/2}\log^{3/2}(2T)N + 3)n, \end{aligned} \tag{B.5.1}$$

where $\frac{\beta_+^2}{1+\sigma^2} \le \frac{C_\rho^2(1+\sigma)^2/(1-\rho)^2}{(1+\sigma)^2/2} = \frac{2C_\rho^2}{(1-\rho)^2}$ is an upper bound on the condition number of the covariance matrix $\boldsymbol{\Sigma}[\boldsymbol{x}_t]$. Similarly, the approximate mixing time of the nonlinear system (2.4.2) is given by,

$$\begin{aligned} & L \ge 1 + \big[\log(c_0 C_\rho\beta_+(1+\|\boldsymbol{\Theta}_\star\|_F\beta_+\sqrt{n})n\sqrt{N/n}) + \log(c/\beta_+ \vee c\sqrt{n}/\beta_+)\big]/\log(\rho^{-1}), \\ \Longleftarrow \quad & L \ge \Big\lceil 1 + \frac{\log(CC_\rho(1+\|\boldsymbol{\Theta}_\star\|_F\beta_+)Nn)}{1-\rho}\Big\rceil, \end{aligned} \tag{B.5.2}$$

where $C > 0$ is a constant. Finally, given the trajectory length $T \gtrsim L(N+1)$, where N and L are as given by (B.5.1) and (B.5.2) respectively, starting from $\boldsymbol{\Theta}^{(0)} = 0$ and using the learning rate $\eta = \frac{\gamma^2(1+\sigma^2)}{16\beta_+^4 n^2} \ge \frac{\gamma^2(1-\rho)^4}{32C_\rho^4(1+\sigma)^2 n^2}$, with probability at least $1 - Ln\big(4T + \log(\frac{\|\boldsymbol{\Theta}_\star\|_F C_\rho(1+\sigma)}{\sigma(1-\rho)})\big)\exp(-100n)$ for all $1 \le k \le n$, all gradient descent iterates $\boldsymbol{\Theta}^{(\tau)}$ on $\hat{\mathcal{L}}$

satisfy

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

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

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

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

Appendix C

Remaining Proofs from Chapter 5

C.1 Proof of Lemma 46

Proof. We start by expanding the convergence term by substituting $\tilde{\boldsymbol{v}} = [\beta\boldsymbol{v}^\top \ \gamma]^\top$ as follows,

$$\begin{aligned}
|\tilde{\boldsymbol{v}}^\top(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X} \ \boldsymbol{1}]^\top[\boldsymbol{X} \ \boldsymbol{1}])\tilde{\boldsymbol{v}}| &= |\frac{1}{n}\|[\boldsymbol{X} \ \boldsymbol{1}]\tilde{\boldsymbol{v}}\|_{\ell_2}^2 - \|\tilde{\boldsymbol{v}}\|_{\ell_2}^2| \\
&= |\frac{1}{n}\|\beta\boldsymbol{X}\boldsymbol{v} + \gamma\boldsymbol{1}\|_{\ell_2}^2 - \|[\beta\boldsymbol{v}^\top \ \gamma]^\top\|_{\ell_2}^2| \\
&= |\frac{1}{n}(\beta^2\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2\boldsymbol{1}^\top\boldsymbol{1} + 2\beta\gamma\boldsymbol{1}^\top\boldsymbol{X}\boldsymbol{v}) - \beta^2\|\boldsymbol{v}\|_{\ell_2}^2 - \gamma^2| \\
&= |\frac{1}{n}\beta^2\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 - \beta^2\|\boldsymbol{v}\|_{\ell_2}^2 + \frac{1}{n}\gamma^2 n - \gamma^2 + 2\beta\gamma\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^\top\boldsymbol{x}_i \\
&\le \beta^2|\frac{1}{n}\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 - \|\boldsymbol{v}\|_{\ell_2}^2| + |2\beta\gamma||\boldsymbol{v}^\top\frac{\sum_{i=1}^n \boldsymbol{x}_i}{n}| \\
&\lesssim |\boldsymbol{v}^\top(\boldsymbol{I} - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{v}| + |\boldsymbol{v}^\top\bar{\boldsymbol{x}}|, \qquad \text{(C.1.1)}
\end{aligned}$$

where, $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ is the empirical average vector of i.i.d. subgaussian rows $(\boldsymbol{x}_i)_{i=1}^n$. Thus, using (C.1.1), we can write

$$\sup_{\tilde{\boldsymbol{v}}\in\mathcal{T}_{\text{ext}}} |\tilde{\boldsymbol{v}}^\top(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \lesssim \sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^\top(\boldsymbol{I} - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{v}| + \sup_{\boldsymbol{v}\in\mathcal{T}}|\boldsymbol{v}^\top\bar{\boldsymbol{x}}|. \tag{C.1.2}$$

Given $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ is isotropic subgaussian, Lemma 6.14 in [155] guarantees

$$\sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^\top(\boldsymbol{I} - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{v}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}}, \tag{C.1.3}$$

with probability at least $1 - e^{-t^2}$. Furthermore, since $(\boldsymbol{x}_i)_{i=1}^n$'s have bounded subgaussian norm, $\bar{\boldsymbol{x}}$ is also bounded and standard generic chaining bounds guarantee that [168]

$$\sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^\top \frac{\sum_{i=1}^n \boldsymbol{x}_i}{n}| = \sup_{\boldsymbol{v}\in\mathcal{T}}|\boldsymbol{v}^\top\bar{\boldsymbol{x}}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}}, \tag{C.1.4}$$

with probability at least $1 - e^{-t^2}$. Combining the results (C.1.3) and (C.1.4) into (C.1.2), we find that

$$\sup_{\tilde{\boldsymbol{v}}\in\mathcal{T}_{\text{ext}}} |\tilde{\boldsymbol{v}}^\top(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \mathbf{1}]^\top[\boldsymbol{X}\ \mathbf{1}])\tilde{\boldsymbol{v}}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}} \tag{C.1.5}$$

holds with probability at least $1 - 2e^{-t^2}$. This completes the proof of Lemma 46. ■

C.2 Proof of Equation 5.4.5

Proof. Let the tangent balls $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in (5.2.5) and (5.2.6) respectively. Define the sets

$$\mathcal{T}_- = \mathcal{C}_{\text{ext}} - \mathcal{C}_{\text{ext}} \quad \text{and} \quad \mathcal{T}_+ = \mathcal{C}_{\text{ext}} + \mathcal{C}_{\text{ext}} \tag{C.2.1}$$

and note that

$$\omega(\mathcal{C} - \mathcal{C}) = \mathbb{E}[\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}} \boldsymbol{g}^\top(\boldsymbol{u} - \boldsymbol{v})] \le \mathbb{E}[\sup_{\boldsymbol{u}\in\mathcal{C}} \boldsymbol{g}^\top\boldsymbol{u} + \sup_{\boldsymbol{v}\in-\mathcal{C}} \boldsymbol{g}^\top\boldsymbol{v}] = 2\omega(\mathcal{C}). \tag{C.2.2}$$

Similarly, $\omega(\mathcal{C}+\mathcal{C}) \le 2\omega(\mathcal{C})$. Applying Lemma 46 on $\mathcal{T}_+$ and $\mathcal{T}_-$, with advertised probability of $1-4\exp(-t^2)$, we have

$$\sup_{\boldsymbol{a}\in\mathcal{T}_+\cup\mathcal{T}_-} |\Lambda(\boldsymbol{a},\boldsymbol{a})| \lesssim \frac{\omega(\mathcal{C})+t}{\sqrt{n}} \tag{C.2.3}$$

where $\Lambda(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{a}^\top(\boldsymbol{I}-\frac{1}{n}[\boldsymbol{X}\ \ \mathbf{1}]^\top[\boldsymbol{X}\ \ \mathbf{1}])\boldsymbol{b}$. Now, for any $\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}_{\text{ext}}$, picking $\boldsymbol{u}+\boldsymbol{v}\in\mathcal{T}_+$ and $\boldsymbol{u}-\boldsymbol{v}\in\mathcal{T}_-$, we have

$$|\boldsymbol{\Lambda}(\boldsymbol{u}+\boldsymbol{v},\boldsymbol{u}+\boldsymbol{v})|, |\boldsymbol{\Lambda}(\boldsymbol{u}-\boldsymbol{v},\boldsymbol{u}-\boldsymbol{v})| \lesssim \frac{\omega(\mathcal{C})+t}{\sqrt{n}}. \tag{C.2.4}$$

To proceed, note that

$$\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{v}) = \frac{\boldsymbol{\Lambda}(\boldsymbol{u}+\boldsymbol{v},\boldsymbol{u}+\boldsymbol{v})-\boldsymbol{\Lambda}(\boldsymbol{u}-\boldsymbol{v},\boldsymbol{u}-\boldsymbol{v})}{4}. \tag{C.2.5}$$

Hence, $|\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{v})| = |\boldsymbol{u}^\top(\boldsymbol{I}-\frac{1}{n}[\boldsymbol{X}\ \ \mathbf{1}]^\top[\boldsymbol{X}\ \ \mathbf{1}])\boldsymbol{v}| \lesssim (\omega(\mathcal{C})+t)/\sqrt{n}$ holds with the advertised probability. ■

C.3 Proof of Lemma 48

Proof. Let $(\boldsymbol{x}_i)_{i=1}^n \sim \boldsymbol{x}\in\mathbb{R}^p$ be i.i.d. isotropic subexponential samples and $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ is the corresponding design matrix. Let x_{ij} denotes the ij^{th} element of the matrix $\boldsymbol{X}$. Since each x_{ij} has subexponential norm bounded by a constant, there exists a constant $C>0$ such that $|x_{ij}|\le C\log(n+p)$ holds with probability at least $1-2(n+p)^{-102}$ using subexponential tail bound. Union bounding over all entries of $\boldsymbol{X}$ yields that $|x_{ij}|\le C\log(n+p)$ holds for all i,j with probability at least $1-2(n+p)^{-100}$. Hence, we can bound each row $\boldsymbol{x}_i$ of $\boldsymbol{X}$ with probability at least $1-2(n+p)^{-100}$ via

$$\|\boldsymbol{x}_i\|_{\ell_2} \le C\sqrt{p}\log(n+p), \tag{C.3.1}$$

or equivalently, we have

$$\|\boldsymbol{x}_i\boldsymbol{x}_i^\top\| \le \|\boldsymbol{x}_i\|_{\ell_2}^2 \le cp\log^2(n+p). \tag{C.3.2}$$

This completes the proof of Lemma 48. ■

C.4 Proof of Lemma 49

Proof. Recall that $(\boldsymbol{x}_i)_{i=1}^n \sim \boldsymbol{x} \in \mathbb{R}^p$ are i.i.d. isotropic subexponential vectors and $\tilde{\boldsymbol{x}} = [\boldsymbol{x}^\top\ 1]^\top$. We can estimate the covariance matrix of $\tilde{\boldsymbol{x}}$ given $\|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B$ using law of total probability as follows

$$\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right] = \mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \;\middle|\; \|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B\right]\mathbb{P}\left(\|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B\right) + \mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \;\middle|\; \|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 > B\right]\mathbb{P}\left(\|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 > B\right). \tag{C.4.1}$$

Since a covariance matrix is positive-semidefinite, each term in (C.4.1) is individually positive semidefinite. Hence, we will drop the second term in (C.4.1) to get the following lower bound on the covariance matrix

$$\mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top] \succeq \mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \;\big|\; \|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B]\,\mathbb{P}(\|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B) \tag{C.4.2}$$

Using Lemma 48, it follows that $\|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 = \|[\boldsymbol{x}^\top\ 1]^\top\|_{\ell_2}^2 \le Cp\log^2(n+p) = B$ holds with probability at least $1 - 2(n+p)^{-100}$. Hence, following (C.4.2), we get

$$\begin{aligned}\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top \;\middle|\; \|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 \le B\right] &\preceq \frac{\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]}{\mathbb{P}\left(\|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right)} \\ &\preceq \frac{1}{1-2(n+p)^{-100}}\boldsymbol{I}_p \preceq 2\boldsymbol{I}_p. \end{aligned} \tag{C.4.3}$$

This completes the proof of Lemma 49. ■

C.5 Proof of Lemma 53

Proof. Subgaussian case: Using subgaussian tail, for large enough constant $C > 0$, for each i, we have $|w_i| \le C\sigma\sqrt{\log(n)} = \sigma B$ with probability at least $1 - 2n^{-101}$. This implies $\mathrm{clip}(w_i, \sigma B) = w_i$. Union bounding over all entries of $\boldsymbol{w}$, we find the result which holds with probability at least $1 - 2n^{-100}$.

Subexponential case: follows similarly with $B = C\log(n)$. ■

C.6 Proof of Lemma 54

Proof. We prove the result for subexponential samples. Subgaussian case follows similarly. Without loss of generality, let $\sigma = 1$ as everything can be scaled accordingly. Defining clip function as previously, set $\boldsymbol{z} = \mathrm{clip}(w, B)\boldsymbol{x}$. Furthermore, let w_{tail} denote the tail of $|w|$, such that,

$$w_{\text{tail}} = \begin{cases} |w| & \text{if } |w| > B \\ 0 & \text{otherwise} \end{cases}. \tag{C.6.1}$$

w_{tail} is an upper bound on the error due to clipping, that is,

$$|w - \mathrm{clip}(w, B)| \le w_{\text{tail}}. \tag{C.6.2}$$

We proceed by upper bounding $\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2}$ in terms of w_{tail}, using subadditive property of ℓ_2-norm and the orthogonality of w and $\boldsymbol{x}$ (i.e., $\mathbb{E}[w\boldsymbol{x}] = \mathbb{E}[\boldsymbol{x}(y - \boldsymbol{x}^\top\boldsymbol{\theta}_\star - \mu_\star)] = \mathbb{E}[y\boldsymbol{x}] -$

$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{\theta}_\star - \mu_\star \mathbb{E}[\boldsymbol{x}] = 0)$, as follows

$$\begin{aligned} \|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} &= \|\mathbb{E}[\text{clip}(w, B)\boldsymbol{x}]\|_{\ell_2} \\ &= \|\mathbb{E}[(w - \text{clip}(w, B))\boldsymbol{x}]\|_{\ell_2} \\ &\le \mathbb{E}[|w - \text{clip}(w, B)|\|\boldsymbol{x}\|_{\ell_2}] \\ &\le \mathbb{E}[w_{\text{tail}} \max(\|\boldsymbol{x}\|_{\ell_2}, \sqrt{p}B)]. \end{aligned} \tag{C.6.3}$$

Using subexponentiality, for some constant $c > 0$, we have that, $\mathbb{P}(w_{\text{tail}} > \sqrt{c}t) \le 2e^{-t}$ and $\mathbb{P}(\|\boldsymbol{x}\|_{\ell_2} > \sqrt{cp}t) \le 2pe^{-t}$, where, the latter follows from union bounding over all entries of $\boldsymbol{x}$. Union bounding these two events, we get the following tail bound for their product,

$$\mathbb{P}(w_{\text{tail}}\|\boldsymbol{x}\|_{\ell_2} > c\sqrt{p}t^2) \le 4pe^{-t}. \tag{C.6.4}$$

For notational convenience, set

$$g = w_{\text{tail}} \max(\|\boldsymbol{x}\|_{\ell_2}, \sqrt{p}B), \tag{C.6.5}$$

and note that g satisfies the following property due to (C.6.1)

$$\begin{cases} \text{either} & g > \sqrt{p}B^2 \\ \text{or} & g = 0 \end{cases}. \tag{C.6.6}$$

Furthermore, from (C.6.4) we get the following tail distribution

$$Q_g(t) = \mathbb{P}(g > t) \le 4pe^{-[\frac{t}{c\sqrt{p}}]^{1/2}}. \tag{C.6.7}$$

for $t \geq \alpha := \sqrt{p}B^2$. Combining (C.6.5), (C.6.6) and (C.6.7) into (C.6.3) and denoting probability density function of g by f_g, we get

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \leq \mathbb{E}[g] = \int_\alpha^\infty t f_g(t)dt = -\int_\alpha^\infty t dQ_g(t) \\
&= -tQ_g(t)\Big|_\alpha^\infty + \int_\alpha^\infty Q_g(t)dt \\
&= \sqrt{p}B^2 Q_g(\sqrt{p}B^2) + \int_\alpha^\infty Q_g(t)dt \\
&\overset{(a)}{\leq} 4p^2B^2e^{-B/\sqrt{c}} + 4p\int_{\sqrt{p}B^2}^\infty e^{-[\frac{t}{c\sqrt{p}}]^{1/2}}dt. \qquad \text{(C.6.8)}
\end{aligned}$$

where, (a) follows from (C.6.7). To bound the term on the right hand side, we do a change of variable in (C.6.8) by setting $\tau = [t/(c\sqrt{p})]^{1/2}$ to get,

$$\begin{aligned}
4p\int_{\sqrt{p}B^2}^\infty e^{-[\frac{t}{c\sqrt{p}}]^{1/2}}dt &\leq 8cp^2\int_{\frac{B}{\sqrt{c}}}^\infty \tau e^{-\tau}d\tau \\
&\leq 8cp^2\big[-\tau e^{-\tau}\big|_{\frac{B}{\sqrt{c}}}^\infty + \int_{\frac{B}{\sqrt{c}}}^\infty e^{-\tau}d\tau\big] \\
&= 8cp^2\big[\frac{B}{\sqrt{c}}e^{-\frac{B}{\sqrt{c}}} + e^{-\frac{B}{\sqrt{c}}}\big] \\
&\leq 8cp^2(\frac{B}{\sqrt{c}}+1)e^{-\frac{B}{\sqrt{c}}}. \qquad \text{(C.6.9)}
\end{aligned}$$

Combining this with (C.6.8), we get

$$\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \leq 4p^2(B^2 + 2c(B/\sqrt{c}+1))e^{-B/\sqrt{c}} \overset{(a)}{\leq} C_0p^2n^{-201}, \qquad \text{(C.6.10)}$$

where, we get (a) by picking $B = C\log(n)$ with sufficiently large $C > 0$. Finally, note that conditioned on $|w| \leq B$, $\boldsymbol{z} = w\boldsymbol{x}$ and

$$\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \geq \|\mathbb{E}[w\boldsymbol{x} \mid |w| \leq B]\|_{\ell_2}\mathbb{P}(|w| \leq B). \qquad \text{(C.6.11)}$$

Since $\mathbb{P}(|w| \le B) > 1/2$, this yields $\|\mathbb{E}[w\boldsymbol{x} \mid |w| \le B]\|_{\ell_2} \lesssim p^2 n^{-201}$ which is the advertised result with $\sigma = 1$. Similarly for subgaussian samples, one can show that

$$\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \lesssim p^2 B^2 e^{-B^2/c}. \tag{C.6.12}$$

Picking $B = C\sqrt{\log(n)}$ with sufficiently large $C > 0$, we get the same result, concluding the proof of Lemma 54. ■