FINITE SAMPLE GUARANTEES FOR LEARNING THE DYNAMICS OF SYSTEMS

by

Lei Xin

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Electrical and Computer Engineering

West Lafayette, Indiana

December 2023

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Shreyas Sundaram, Co-Chair

Elmore Family School of Electrical and Computer Engineering

Dr. George T.C. Chiu, Co-Chair

School of Mechanical Engineering

Dr. Jianghai Hu

Elmore Family School of Electrical and Computer Engineering

Dr. Philip E. Paré

Elmore Family School of Electrical and Computer Engineering

Approved by:

Dr. Milind Kulkarni

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Shreyas Sundaram, for his guidance and support throughout my Ph.D. journey. I would like to thank him for accepting me as a Ph.D. student when I was an undergrad. I was impressed by his knowledge, the way he tackles research problems, and the way he explains complicated concepts using simple words. I also learnt a lot of presentation skills from him. His encouragements help me to gain more confidence. I deeply appreciate him for devoting his time into helping me grow as a good researcher.

I would like to express my appreciation to Dr. George T.C. Chiu, who became my co-advisor in my second year of Ph.D. During our weekly research meetings, he provided countless valuable feedback and insights that allowed me to look at problems from different perspectives and helped me connect theory and practice. I am very fortunate to have two amazing advisors, and I could never imagine better Ph.D. advisors.

I would like to thank my committee members, Professors Jianghai Hu and Philip E. Paré. Their valuable feedback helped improve the quality of this thesis significantly. My first course on modern control theory was taught by Professor Jianghai Hu, and I can't overstate how much it helped me build a solid foundation in the field. I thoroughly enjoyed our collaboration on a research paper with Professor Philip E. Paré, and I will never forget the barbecue we had together with his group in the midst of a rainstorm.

Many thanks also go to my lab mates: Amritha, Aritra, Tong, Kananart, Jayanth, Jiajun, Mengxue, Lili, Dylan, Baike, Mustafa, Lintao, Hemant, Russel, Yijing, Maria, Nadia, Alvin, Delong, Carrig, Younggil, Michael, and Opdyke, for the inspiring discussions we had, and for the enjoyable moments we shared together. Your presence made this journey even more rewarding. I learnt a lot from each of you. I would like to thank my colleagues who worked on the WAMV project with me: Jianwen, Wenjian, Hyunsang. Our discussions on the project allow me to think about real world problems from different angles. I would also like to thank all of the friends I have made at Purdue. I am fortunate to have created countless cherished memories with each of you.

I would like to extend my deepest appreciation to my parents and all my family members for their steadfast love and support. Their boundless encouragement, both emotionally and academically, has been the cornerstone of my success, and I am profoundly grateful for their enduring faith in me. I could never have achieved such milestones without their unwavering assistance and belief in my abilities.

Finally, I would like to express my heartfelt gratitude to the entire Purdue University community for the dedication to maintaining an exceptional academic environment. Your unwavering commitment to fostering intellectual growth and creating a vibrant scholarly atmosphere has significantly enriched my educational journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The problem of system identification is to learn the system dynamics from data. While classical system identification theories focused primarily on achieving asymptotic consistency, recent efforts have sought to characterize the number of samples needed to achieve a desired level of accuracy in the learned model. This thesis focuses on finite sample analysis for identifying/learning dynamical systems.

In the first part of this thesis, we provide novel results on finite sample analysis for learning different linear systems. We first consider the system identification problem of a fully observed system (i.e., all states of the system can be perfectly measured), leveraging data generated from an auxiliary system that shares "similar" dynamics. We provide insights on the benefits of using the auxiliary data, and guidelines on selecting the weight parameter during the model training process. Subsequently, we consider the system identification problem for a partially observed autonomous linear system, where only a subset of states and multiple short trajectories of the system can be observed. We present a finite sample error bound and characterize the learning rate.

In the second part of this thesis, we explore the practical usage of finite sample analysis under several different scenarios. We first consider a parameter learning problem in a distributed setting, where a group of agents wishes to collaboratively learn the underlying model. We propose a distributed parameter estimation algorithm and provide finite time bounds on the estimation error. We show that our analysis allows us to determine a time at which the communication can be stopped (due to the costs associated with communications), while meeting a desired estimation accuracy. Subsequently, we consider the problem of online change point detection for a linear system, where the user observes data in an online manner, and the goal is to determine when the underlying system dynamics change. We provide an online change point detection algorithm, and a data-dependent threshold that allows one to achieve a pre-specified upper bound on the probability of making a false alarm. We further provide a finite-sample-based lower bound for the probability of detecting a change point with a certain delay.

Finally, we extend the results to linear model identification from non-linear systems. We provide a data acquisition algorithm followed by a regularized least squares algorithm, along with an associated finite sample error bound on the learned linearized dynamics. Our error bound demonstrates a trade-off between the error due to nonlinearity and the error due to noise, and shows that one can learn the linearized dynamics with arbitrarily small error given sufficiently many samples.

1. INTRODUCTION

Learning a predictive model from data is an important problem in many fields [1], including machine learning, economics and control theory. Classical control theories are typically model-based. When modeling from first principles is not possible, one can attempt to learn a predictive model from data. The problem of system identification is to learn the parameters of dynamical system, given the measurements of the inputs to and outputs from the system. While classical system identification theories focused primarily on achieving asymptotic consistency [2]–[4], these results may not directly translate into guarantees on the quality of learned model, given a finite number of samples. Having a finite sample bound for the error is not only of interest on its own, but also can be integrated with techniques like robust control to come up with overall performance guarantees for the closed loop system, e.g., [5], [6]. Consequently, we focus on studying the following problem:

Can we characterize the quality of the learned dynamics from a finite number of samples collected from the system?

In this thesis, we attempt to systematically address the above problem. We present novel non-asymptotic guarantees for learning the dynamics of systems under various settings. We summarize our contributions below. The details and comparisons over existing results are provided in the beginning of each chapter.

1.1 Overview and Contributions

1.1.1 Learning Dynamical Systems by Leveraging Data from Similar Systems

In Chapter 2, we study the problem of identifying the dynamics of a linear system when one has access to samples generated by a similar (but not identical) system, in addition to data from the true system. We use a weighted least squares approach, and provide a finite sample error bound of the learned model as a function of the number of samples and various system parameters from the two systems as well as the weight assigned to the auxiliary data. We show that the auxiliary data can help to reduce the intrinsic system identification error due to noise, at the price of adding a portion of error that is due to the differences between

the two system models. We further provide a data-dependent bound that is computable when some prior knowledge about the systems, such as upper bounds on noise levels and model difference, is available. This bound can also be used to determine the weight that should be assigned to the auxiliary data during the model training stage. Our analysis can be applied to a variety of important settings. For example, if the system dynamics change at some point in time (e.g., due to a fault), how should one leverage data from the prior system in order to learn the dynamics of the new system? As another example, if there is abundant data available from a simulated (but imperfect) model of the true system, how should one weight that data compared to the real data from the system? Our analysis provides insights into the answers to these questions.

1.1.2 Learning the Dynamics of Autonomous Linear Systems From Multiple Trajectories

In Chapter 3, we consider the problem of learning the dynamics of partially-observed autonomous linear systems (i.e., systems that are not affected by external control inputs) from observations of multiple trajectories of those systems, with finite sample guarantees. Existing results on learning rate and consistency of autonomous linear system identification rely on observations of steady state behaviors from a single long trajectory, and are not applicable to unstable systems. In contrast, we consider the scenario of learning system dynamics based on multiple short trajectories, where there are no easily observed steady state behaviors. We provide a finite sample analysis, which shows that the dynamics can be learned at a rate $\mathcal{O}(\frac{1}{\sqrt{N}})$ for both stable and unstable systems, where N is the number of trajectories, when the initial state of the system has zero mean (which is a common assumption in the existing literature). We further generalize our result to the case where the initial state has non-zero mean. We show that one can adjust the length of the trajectories to achieve a learning rate of $\mathcal{O}(\sqrt{\frac{\log N}{N}})$ for strictly stable systems and a learning rate of $\mathcal{O}(\frac{(\log N)^d}{\sqrt{N}})$ for marginally stable systems, where d is some constant.

1.1.3 Finite Sample Guarantees for Distributed Online Parameter Estimation with Communication Costs

In Chapter 4, we study the problem of distributed online parameter learning. Existing distributed online optimization algorithms typically provide bounds on regret, which may not be directly translated into bounds on error of the learned model. In this chapter, we propose a distributed online learning algorithm in a networked setting, which enables each agent to improve its learning accuracy by communicating with neighbors in the network. We provide non-asymptotic bounds on the learning error, leveraging the statistical properties of the underlying model. Our analysis demonstrates a trade-off between learning error and communication costs. Further, our analysis allows us to determine a time at which the communication can be stopped (due to the costs associated with communications), while meeting a desired estimation accuracy.

1.1.4 Online Change Points Detection for Linear Dynamical Systems with Finite Sample Guarantees

In Chapter 5, we study the problem of online change point detection, where the goal is to detect abrupt changes in properties of time series, ideally as soon as possible after those changes occur. Existing work on online change point detection either assumes i.i.d data, focuses on asymptotic analysis, does not present theoretical guarantees on the trade-off between detection accuracy and detection delay, or is only suitable for detecting single change points. In this chapter, we study the online change point detection problem for linear dynamical systems with unknown dynamics, where the data exhibits temporal correlations and the system could have multiple change points. We develop a data-dependent threshold that can be used in our test that allows one to achieve a pre-specified upper bound on the probability of making a false alarm. We further provide a finite-sample-based bound for the probability of detecting a change point. Our bound demonstrates how parameters used in our algorithm affect the detection probability and delay, and provides guidance on the minimum required time between changes to guarantee detection.

1.1.5 Learning Linearized Models from Nonlinear Systems with Finite Data

In Chapter 6, we consider the problem of identifying a linearized system model from a nonlinear system. This is different from many existing works on linear model identification, where the underlying assumption is that the system is indeed linear. We provide a multiple trajectories-based deterministic data acquisition algorithm followed by a regularized least squares algorithm, along with an associated finite sample error bound on the learned linearized dynamics. Our error bound demonstrates a trade-off between the error due to nonlinearity and the error due to noise, and shows that one can learn the linearized dynamics with arbitrarily small error given sufficiently many samples. We validate our results through experiments, where we also show the potential insufficiency of linear system identification using a single trajectory with i.i.d random inputs (which is a common setup in the literature), when nonlinearity does exist.

Finally, in Chapter 7, we conclude the thesis by summarizing our main results and discussing directions for future work.

1.2 Mathematical Notation and Terminology

Vectors are taken to be column vectors unless otherwise specified. Let $\mathbb{R}$ and $\mathbb{N}$ denote the sets of real numbers and natural numbers, respectively. Let $\sigma_n(\cdot)$ and $\sigma_{min}(\cdot)$ be the n-th largest and smallest singular value, respectively, of a symmetric matrix. The eigenvalues of a given matrix are ordered with nonincreasing magnitude, i.e., $|\lambda_1(\cdot)| \geq \cdots \geq |\lambda_{min}(\cdot)|$. The spectral radius of a given matrix is denoted as $\rho(\cdot)$. A square matrix A is called strictly stable if $\rho(A) < 1$, marginally stable if $\rho(A) \leq 1$, and unstable if $\rho(A) > 1$. For a given matrix A, we use $A(i,j)$ to denote the element in its i-th row and j-th column, A' to denote its transpose, $A^\dagger$ to denote its pseudoinverse, and $\text{vec}(A)$ to denote its vectorization (i.e., the vector obtained by stacking the columns of A from left). We use $\|A\|$, $\|A\|_1$ and $\|A\|_F$ to denote the spectral norm, 1-norm and Frobenius norm, respectively, of matrix A. A Gaussian distributed random vector is denoted as $u \sim \mathcal{N}(\mu, \Sigma)$, where μ is the mean and Σ is the covariance matrix. We use I to denote the identity matrix. We use $\mathbb{E}$ to denote the expectation. We use $\text{tr}(\cdot)$ to denote the trace of a given matrix. Let $\mathbf{1}_n$ denote a vector of dimension n with all of its elements equal to 1. The symbol $\prod_{t=i}^{j} A_t$ is used to denote the matrix product, $A_i A_{i+1} \cdots A_j$. We use the symbol mod to denote the modulo operator. The symbols $\cup$ and $\cap$ are used to denote the union and intersection of sets, respectively. The symbol $\sigma(\cdot)$ is used to denote the sigma field generated by the corresponding random vectors. We use $\mathcal{S}^{n-1}$ to denote the unit sphere in n-dimensional space. The open l_1 ball in d-dimensional space with center at x_0 and radius r is denoted by $\mathcal{B}_d(x_0, r) \triangleq \{x \in \mathbb{R}^d : \|x - x_0\|_1 < r\}$. We denote $\mathbf{e}_i^d$ as a d-dimensional vector with the i-th component equal to 1 and all other components equal to 0. The symbols $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are used to denote the floor and ceiling functions, respectively. We use $\mathbf{0}$ to denote a zero vector with dimension that is clear from the context.

2. LEARNING DYNAMICAL SYSTEMS BY LEVERAGING DATA FROM SIMILAR SYSTEMS

2.1 Introduction

The existing literature on finite sample analysis of system identification is typically either single trajectory-based or multiple trajectories-based. The single trajectory setup assumes that one has samples from a single trajectory from the system, which enables system identification in an online manner, i.e., there is no need to restart the system from an initial state. This setup has been studied extensively over the past few years and is still an ongoing research topic [7]–[12]. A key challenge in the analysis is addressing the dependencies of samples from the single trajectory. The derived sample complexity results typically show how the system identification error goes to zero by increasing the number of samples used in the single trajectory. For the multiple trajectories setup, it is typically assumed that one has access to data generated from multiple independent trajectories [13]–[17]. In practice, the multiple trajectories setup has the advantage of being able to handle unstable systems, and other cases where collecting a single long trajectory is infeasible. Technically, the assumption of independence of data usually allows for more direct use of standard concentration inequalities. Consequently, the derived results typically only show that the error goes to zero by increasing the number of trajectories. The recent paper [18] carefully addresses learning dynamical systems from a mix of both dependent and independent data, i.e., learning from multiple trajectories each with a non-trivial length. The paper [18] provides sharp bounds that hold in expectation, and shows that the error goes to zero at a rate that is determined by the product of the number of trajectories and the number of samples used from these trajectories.

We note that all of the above works make the assumption that the data used for system identification are generated from the true system model that one wants to learn. However, in many cases, collecting abundant data from the true system can be costly or even infeasible. In such cases, one may want to rely on data generated from other systems that share similar dynamics. For example, for non-engineered systems like animals, one may only have a limited amount of data from the true individual animal one wants to model, due to the challenge of

conducting experiments for such systems. On the other hand, it may be possible to collect data from other animals in the herd or from a reasonably good simulator. Furthermore, when the dynamics of a system changes (e.g., due to failures), one needs to decide whether to discard all of the previous data, or to leverage the old information in an appropriate way. In settings such as the ones described above, it is of great interest to understand how one can leverage the data generated from systems that share similar (but not identical) dynamics. This idea is similar to the notion of *transfer learning* in the machine learning community, where one wants to transfer knowledge from related tasks to a new task [19]. However, in contrast to system identification, most of the papers on transfer learning (in the context of estimation) consider learning a static mapping from a feature space to a label space [20]. The recent works [21], [22] study joint learning of multiple dynamical systems, assuming all systems are weighted equally in the training stage. However, an open question remains on how to effectively utilize samples from other systems to enhance the accuracy of the model for the true system of interest, especially when the number of samples from the true system is limited.

Our conference paper [23] provides finite sample analysis of system identification with the help of an auxiliary system, using a weighted least squares approach, under the assumption of having access to multiple trajectories from both the true system and the auxiliary system. The paper [23] decomposes the overall system identification error into the error due to noise and the error due to model difference, and shows that the auxiliary data can help to reduce the error due to noise by introducing a portion of constant error that is due to the difference in the models between the true and auxiliary systems. However, although the algorithm in [23] uses all samples from these trajectories (different from [13], where only two data points from each trajectory are used, assuming all samples are generated from the same system), the result is conservative in characterizing the effect of the trajectory length. In particular, the error due to noise can only go to zero by increasing the number of trajectories from the two systems.

In this chapter, we address the above problem. Our contributions are as follows.

- We provide finite sample data-independent bounds for learning dynamical systems by leveraging data from an auxiliary system, using a weighted least squares approach. Again, we decompose the error into a portion due to noise and a portion due to model difference. Different from [23], we show that the error due to noise can go to zero by increasing either the number of trajectories or the trajectory length from the two systems, or both.[1] Our analysis is general in that when the two systems have same system matrices (such that we only have the error due to noise), our result qualitatively matches the results in the recent paper [18], which characterizes how the expected error goes to zero with respect to the number of trajectories and the trajectory length, given samples from the same system. Importantly, our bounds provide insights on general guidelines for assigning weights to the auxiliary system, when there is not enough prior knowledge about the systems.

- We also provide a data-dependent bound that is computable when some prior knowledge about the systems, such as upper bounds on noise levels and model difference, is available (based on a regularized weighted least squares approach). The data-dependent bound can be used in a data-driven scheme for selecting a good weight parameter that provides better performance guarantees in practice.

To the best of our knowledge, our work in this chapter is the first to study finite sample analysis for weighted least squares-based system identification given different systems. Technically, we overcome the challenges of addressing the dependencies of samples from independent system trajectories in a less conservative way, compared to [13], [23]. We carefully analyze how different weights used in system identification and difference of system dynamics affect the finite sample error. We provide a new upper bound of the sample covariance matrix in the multiple trajectories setup for systems that have inputs. We also provide a new lower bound for the smallest eigenvalue of the sample covariance matrix for non-Gaussian time series in a more general context. This result could be of independent interest since it can be used in the analyses of many regression-based problems.

[1]↑We note that the paper [23] also allows the auxiliary system to be time-varying. However, the derived bound again degrades when the trajectory length becomes longer. In this chapter, we will assume the auxiliary system is time-invariant.

This chapter is organized as follows. Section 2.2 formulates the system identification problem and introduces the algorithm we use. In Section 2.3, we present our main results. We present various numerical examples capturing different scenarios in Section 2.4 to illustrate our results, and conclude in Section 2.5. All of the proofs are included in Section 2.6.

2.2 Problem formulation and algorithm

Consider the following discrete time linear time-invariant (LTI) system

$$\bar{x}_{t+1} = \bar{A}\bar{x}_t + \bar{B}\bar{u}_t + \bar{w}_t, \tag{2.1}$$

where $\bar{x}_t \in \mathbb{R}^n$, $\bar{u}_t \in \mathbb{R}^p$, $\bar{w}_t \in \mathbb{R}^n$, are the state, input, and process noise, respectively, and $\bar{A} \in \mathbb{R}^{n\times n}$ and $\bar{B} \in \mathbb{R}^{n\times p}$ are the system matrices we wish to learn from data. In this chapter, we also assume that both the input $\bar{u}_t$ and state $\bar{x}_t$ can be perfectly measured.

Suppose that we have access to N_r independent experiments of system (2.1), in which the system restarts from an initial state $\bar{x}_0$, and each experiment is of length T_r. We call the state-input pairs collected from each experiment a *rollout* (or trajectory), and denote the set of samples we have as $\{(\bar{x}_t^i, \bar{u}_t^i) : 1 \le i \le N_r, 0 \le t \le T_r\}$. Note that we use the superscript to denote the rollout index and the subscript to denote the time index.

Let $\bar{z}_t^i = \begin{bmatrix} \bar{x}_t^{i\prime} & \bar{u}_t^{i\prime} \end{bmatrix}' \in \mathbb{R}^{n+p}$ for $t \ge 0$. For each rollout i, define $\bar{X}^i = \begin{bmatrix} \bar{x}_1^i & \cdots & \bar{x}_{T_r}^i \end{bmatrix} \in \mathbb{R}^{n\times T_r}$, $\bar{Z}^i = \begin{bmatrix} \bar{z}_0^i & \cdots & \bar{z}_{T_r-1}^i \end{bmatrix} \in \mathbb{R}^{(n+p)\times T_r}$, $\bar{W}^i = \begin{bmatrix} \bar{w}_0^i & \cdots & \bar{w}_{T_r-1}^i \end{bmatrix} \in \mathbb{R}^{n\times T_r}$. Further, define the batch matrices $\bar{X} = \begin{bmatrix} \bar{X}^1 & \cdots & \bar{X}^{N_r} \end{bmatrix} \in \mathbb{R}^{n\times N_r T_r}$, $\bar{Z} = \begin{bmatrix} \bar{Z}^1 & \cdots & \bar{Z}^{N_r} \end{bmatrix} \in \mathbb{R}^{(n+p)\times N_r T_r}$, $\bar{W} = \begin{bmatrix} \bar{W}^1 & \cdots & \bar{W}^{N_r} \end{bmatrix} \in \mathbb{R}^{n\times N_r T_r}$. Denoting $\Theta \triangleq \begin{bmatrix} \bar{A} & \bar{B} \end{bmatrix} \in \mathbb{R}^{n\times(n+p)}$, we have

$$\bar{X} = \Theta\bar{Z} + \bar{W}.$$

Typically, one would like to solve the following optimization problem:

$$\min_{\tilde{\Theta}\in\mathbb{R}^{n\times(n+p)}} \|\bar{X} - \tilde{\Theta}\bar{Z}\|_F^2,$$

and obtain an estimate $\Theta_{LS} \triangleq \begin{bmatrix} \bar{A}_{LS} & \bar{B}_{LS} \end{bmatrix}$, of which the analytical form is

$$\Theta_{LS} = \bar{X}\bar{Z}'(\bar{Z}\bar{Z}')^{-1},$$

under the assumption that the matrix $\bar{Z}\bar{Z}'$ is invertible. The quality of the recovered estimate will depend on N_r and T_r; in particular, if both N_r and T_r are small, the obtained estimate could have large estimation error [7], [13].

Suppose that, in addition to samples from the true system, we also have access to samples generated from an auxiliary system that shares "similar" (but unknown) dynamics to system (2.1). In particular, consider an auxiliary discrete time linear time-invariant system

$$\hat{x}_{t+1} = \hat{A}\hat{x}_t + \hat{B}\hat{u}_t + \hat{w}_t, \tag{2.2}$$

where $\hat{x}_t \in \mathbb{R}^n$, $\hat{u}_t \in \mathbb{R}^p$, $\hat{w}_t \in \mathbb{R}^n$ are the state, input, and process noise, respectively, and $\hat{A} \in \mathbb{R}^{n\times n}$ and $\hat{B} \in \mathbb{R}^{n\times p}$ are system matrices.[2] The above dynamics can be rewritten as

$$\hat{x}_{t+1} = (\bar{A} + \delta_A)\hat{x}_t + (\bar{B} + \delta_B)\hat{u}_t + \hat{w}_t, \tag{2.3}$$

where $\delta_A = \hat{A} - \bar{A}, \delta_B = \hat{B} - \bar{B}$. Intuitively, the samples generated from the above system will be useful for identifying system (2.1) if both $\|\delta_A\|$ and $\|\delta_B\|$ are small. For example, suppose that we want to identify the dynamics of a vehicle, the auxiliary system could be another vehicle of the same type produced by the same manufacturer. We also provide a scenario involving a change in dynamics of a system where the true and the auxiliary systems have the same state representation in our experiment section later in the chapter.[3]

Thus, suppose that we also have access to N_p independent experiments of system (2.2), in which the system restarts from an initial state $\hat{x}_0$, and each experiment is of length T_p. Let $\{(\hat{x}_t^i, \hat{u}_t^i) : 1 \le i \le N_p, 0 \le t \le T_p\}$ denote the samples from these experiments. Let $\hat{z}_t^i = \begin{bmatrix} \hat{x}_t^{i\prime} & \hat{u}_t^{i\prime} \end{bmatrix}' \in \mathbb{R}^{n+p}$ for $t \ge 0$. The matrices $\hat{X}^i \in \mathbb{R}^{n\times T_p}$, $\hat{Z}^i \in \mathbb{R}^{(n+p)\times T_p}$, $\hat{W}^i \in \mathbb{R}^{n\times T_p}$, $\hat{X} \in$

[2]↑In the terminology of transfer learning, system (2.1) can be referred to as a target system, and system (2.2) can be referred to as a source system.

[3]↑In practice, the auxiliary system needs to share the same set of state variables to be considered "similar", although our results hold for general systems where the states are unrelated.

$\mathbb{R}^{n\times N_pT_p}$, $\hat{Z}\in\mathbb{R}^{(n+p)\times N_pT_p}$, $\hat{W}\in\mathbb{R}^{n\times N_pT_p}$ are defined similarly, using the signals $\hat{u}_t^i, \hat{x}_t^i, \hat{w}_t^i$ from system (2.2). Let $X=\begin{bmatrix}\bar{X} & \hat{X}\end{bmatrix}\in\mathbb{R}^{n\times(N_rT_r+N_pT_p)}$, $Z=\begin{bmatrix}\bar{Z} & \hat{Z}\end{bmatrix}\in\mathbb{R}^{(n+p)\times(N_rT_r+N_pT_p)}$, $W=\begin{bmatrix}\bar{W} & \hat{W}\end{bmatrix}\in\mathbb{R}^{n\times(N_rT_r+N_pT_p)}$ and $\delta_\Theta=\begin{bmatrix}\delta_A & \delta_B\end{bmatrix}\in\mathbb{R}^{n\times(n+p)}$. Defining

$$\Delta^i=\begin{bmatrix}\delta_\Theta\hat{z}_0^i & \cdots & \delta_\Theta\hat{z}_{T_p-1}^i\end{bmatrix}\in\mathbb{R}^{n\times T_p},$$

for all $i\in\{1,\ldots,N_p\}$, and denoting

$$\Delta=\begin{bmatrix}\mathbf{0} & \cdots & \mathbf{0} & \Delta^1 & \cdots & \Delta^{N_p}\end{bmatrix}\in\mathbb{R}^{n\times(N_rT_r+N_pT_p)},$$

where we use $\mathbf{0}$ to denote zero matrices with appropriate dimensions, we have the relationship

$$X=\Theta Z+W+\Delta. \tag{2.4}$$

Letting $q\in\mathbb{R}_{\geq 0}$ be a design parameter that specifies the relative weight assigned to samples generated from the auxiliary system (2.2), we can define $Q=\mathrm{diag}(I_{N_rT_r}, qI_{N_pT_p})\in\mathbb{R}^{(N_rT_r+N_pT_p)\times(N_rT_r+N_pT_p)}$. Setting the regularization parameter $\lambda\geq 0$, we are interested in the following (regularized-) weighted least squares problem:

$$\min_{\tilde{\Theta}\in\mathbb{R}^{n\times(n+p)}}\{\|XQ^{\frac{1}{2}}-\tilde{\Theta}ZQ^{\frac{1}{2}}\|_F^2+\lambda\|\tilde{\Theta}\|_F^2\}. \tag{2.5}$$

The well known (regularized-) weighted least squares estimate [24] is $\Theta_{WLS}\triangleq\begin{bmatrix}\bar{A}_{WLS} & \bar{B}_{WLS}\end{bmatrix}$, which has the form

$$\Theta_{WLS}=XQZ'(ZQZ'+\lambda I_{n+p})^{-1}, \tag{2.6}$$

when the matrix $ZQZ'+\lambda I_{n+p}$ is invertible. Using (2.4), the system identification error can be expressed as

$$\begin{aligned}\|\Theta_{WLS}-\Theta\|=\|&-\lambda\Theta(ZQZ'+\lambda I_{n+p})^{-1}\\&+WQZ'(ZQZ'+\lambda I_{n+p})^{-1}\\&+\Delta QZ'(ZQZ'+\lambda I_{n+p})^{-1}\|.\end{aligned} \tag{2.7}$$

In particular, when the regularization parameter is set to be $\lambda = 0$, we recover the standard weighted least squares estimate.

The above steps are encapsulated in Algorithm 1.

Algorithm 1 System Identification Using Auxiliary Data

1: Gather N_r rollouts of samples (each of length T_r) generated from the true system (2.1), each starting from $\bar{x}_0^i$ for all $1 \leq i \leq N_r$.

2: Gather N_p rollouts of samples (each of length T_p) generated from the auxiliary system (2.2), each starting from $\hat{x}_0^i$ for all $1 \leq i \leq N_p$.

3: Construct the matrices X, Q, Z. Compute $\Theta_{WLS} = XQZ'(ZQZ' + \lambda I_{n+p})^{-1}$.

4: Return the first n columns of Θ_{WLS} as an estimated $\bar{A}$, and the remaining columns of Θ_{WLS} as an estimated $\bar{B}$.

Remark 1. *The weight parameter q specifies how much we weight the data from the auxiliary system relative to the data from the true system, and can depend on the number of samples (N_r, T_r and N_p, T_p) from each of those systems or the data available to us. The specific choice of q will be discussed in detail later as we present our main results.*

Our results will leverage the following definition of sub-Gaussian random vectors [25].

Definition 2.2.1. *A real-valued random variable w is called sub-Gaussian with parameter R^2 if we have*

$$\forall \alpha \in \mathbb{R}, \mathbb{E}[\exp(\alpha w)] \leq \exp(\frac{\alpha^2 R^2}{2}).$$

A random vector $x \in \mathbb{R}^n$ is called R^2 sub-Gaussian if for all unit vectors $v \in \mathcal{S}^{n-1}$ the random variable $v'x$ is R^2 sub-Gaussian.

We make the following assumption. Recall that the inputs $\bar{u}_t, \hat{u}_t$ are known, while the noise terms $\bar{w}_t, \hat{w}_t$ are unknown.

Assumption 1. *The random vectors $\bar{w}_t, \bar{u}_t, \bar{x}_0, \hat{w}_t, \hat{u}_t, \hat{x}_0$ are independent sub-Gaussian with independent coordinates for all $t \geq 0$. Furthermore, they have positive definite covariance matrices and sub-Gaussian parameters $\sigma_{\bar{w}}^2, \sigma_{\bar{u}}^2, \sigma_{\bar{x}_0}^2, \sigma_{\hat{w}}^2, \sigma_{\hat{u}}^2, \sigma_{\hat{x}_0}^2$, respectively.*

We note that independent random inputs are commonly used in the context of system identification to provide excitation of the system dynamics [8], [13]. Studying the optimal input for system identification is an active area of research [26]. Further, random initial states can be easily obtained from deterministic initial states that are equal to zero. A simple way to achieve this is to apply zero input and treat the first state as the initial state for each trajectory (which is random due to noise). Our results could also be generalized to include bounded deterministic initial states.

To ease the notation, we make some definitions now. We denote $\bar{\sigma}_{max} = \max(\sigma_{\bar{w}}, \sigma_{\bar{u}}, \sigma_{\bar{x}_0})$. Letting $v(j)$ denote the j-th component of a vector v, we define

$$\bar{\sigma}_{min} \triangleq \min(\{\mathbf{E}[\bar{w}(i)^2], \mathbf{E}[\bar{u}(j)^2], \mathbf{E}[\bar{x}_0(i)^2]\}),$$
$$\bar{\sigma}_* \triangleq \max\left(\left\{\frac{\mathbf{E}[\bar{w}(i)^4]}{\mathbf{E}[\bar{w}(i)^2]^2}, \frac{\mathbf{E}[\bar{u}(j)^4]}{\mathbf{E}[\bar{u}(j)^2]^2}, \frac{\mathbf{E}[\bar{x}_0(i)^4]}{\mathbf{E}[\bar{x}_0(i)^2]^2}\right\}\right),$$

for all $t \geq 0, 1 \leq i \leq n, 1 \leq j \leq p$, where we omitted the time index t for the ease of exposition. Further, define the following matrix for $t \geq 0$:

$$\bar{G}_t \triangleq \sum_{i=0}^{t} \bar{A}^i \bar{A}^{i\prime} + \sum_{i=0}^{t-1} \bar{A}^i \bar{B} \bar{B}' \bar{A}^{i\prime}. \tag{2.8}$$

The terms $\hat{\sigma}_{max}, \hat{\sigma}_{min}, \hat{\sigma}_*$ and $\hat{G}_t$ are defined similarly for the auxiliary system (2.2).

In the next section, we provide data-independent bounds (assuming $\lambda = 0$) and a data-dependent bound (assuming $\lambda > 0$) of the system identification error in (2.7). We study the case when $\lambda = 0$ in the data-independent bounds to highlight our key insights (the benefits of the auxiliary data and the role of the weight parameter q), and the results for $\lambda > 0$ can be easily generalized. The data-independent finite sample upper bounds characterize the error as a function of $N_r, T_r, N_p, T_p, q, \|\delta_\Theta\|$ and other parameters from the true system and the auxiliary system. While the data-independent error bounds provide insights on the benefits of using the auxiliary samples, the derived results may not be used directly in practice, since they involve unknown system parameters. To address that, we also provide a data-dependent bound for the case when $\lambda > 0$. The non-zero regularization parameter λ not only helps us to derive the data-dependent result, but also provides the user with

more flexibility to tune the estimate in practice. One could set λ to be small to reduce the impact of regularization on the estimate. The derived data-dependent bound is computable, applicable to more general input and noise, and can be used in real-world applications to select the weight parameter q (and regularization parameter λ). More specifically, the bound characterizes the error as a function of $\sigma_{\bar{w}}, \sigma_{\hat{w}}, q, \|\delta_\Theta\|$, λ, $\|\Theta\|$, and the available data. Both our data-independent bounds and data-dependent bound will provide insights and guidance on selecting an appropriate weight parameter q. We will assume that system (2.1) and system (2.2) have the same stability in our discussions, i.e., both $\rho(\bar{A})$ and $\rho(\hat{A})$ are less than 1, or both $\rho(\bar{A})$ and $\rho(\hat{A})$ are equal to 1, or both $\rho(\bar{A})$ and $\rho(\hat{A})$ are greater than 1 (although $\rho(\bar{A})$ does not need to equal to $\rho(\hat{A})$), but similar insights can be extended even if they are different.

2.3 Finite Sample Guarantees of the System Identification Error

In this section, we provide data-independent bounds (assuming $\lambda = 0$), and a data-dependent bound (assuming $\lambda > 0$) on the system identification error in (2.7). The proof of the data-independent bounds follow by upper bounding the error terms $\|WQZ'(ZQZ')^{-\frac{1}{2}}\|$, $\|(ZQZ')^{-\frac{1}{2}}\|$, and $\|\Delta QZ'\|$ separately. The proof of the data-dependent bound follows by directly evaluating an upper bound of the term (2.7) from data, but with the replacement of the noise-dependent term $\|WQZ'(ZQZ' + \lambda I_{n+p})^{-\frac{1}{2}}\|$ by a high-probability upper bound. All of the proofs are presented in section 2.6.

2.3.1 Data-independent Bounds

Here, we present our first main result, a data-independent finite sample upper bound on the weighted least squares estimation error in (2.7) when $\lambda = 0$. In the following result, we let c, c_1 denote some positive constants.[4]

Theorem 2.3.1. *Fix $q \geq 0$, $\delta \in (0, \frac{2}{e})$, and let Assumption 1 hold. Denote $\bar{\zeta} = \frac{\bar{\sigma}_{min}}{c_1 \bar{\sigma}_*}$ and $\hat{\zeta} = \frac{\hat{\sigma}_{min}}{c_1 \hat{\sigma}_*}$. Suppose that $N_r T_r \geq \max\{8c_1^2\bar{\sigma}_*^2(\log\frac{2}{\delta} + (n+p)\log\frac{144\bar{g}(\frac{\delta}{2})}{\bar{\zeta}^2(N_rT_r-1)}) + 1, 33\}$, $N_pT_p \geq \max\{8c_1^2\hat{\sigma}_*^2(\log\frac{2}{\delta} + (n+p)\log\frac{144\hat{g}(\frac{\delta}{2})}{\hat{\zeta}^2(N_pT_p-1)}) + 1, 33\}$, $\bar{g}(\frac{\delta}{2}) \geq \frac{\bar{\zeta}^2(N_rT_r-1)}{16}$, and $\hat{g}(\frac{\delta}{2}) \geq \frac{\hat{\zeta}^2(N_pT_p-1)}{16}$.*

[4] ↑See Remark 6 and Remark 7 in section 2.6 for more discussions on the constants c, c_1.

Then with probability at least $1-5\delta$*, the weighted least squares estimate* Θ_{WLS} *from Algorithm 1 using* $\lambda = 0$ *satisfies*

$$\|\Theta_{WLS} - \Theta\| \leq \underbrace{\frac{20\max(\sigma_{\bar{w}}, \sqrt{q}\sigma_{\hat{w}})\sqrt{\log\frac{9^n}{\delta} + (n+p)\log(\phi)}}{\sqrt{N_r T_r \bar{\zeta}^2 + qN_p T_p \hat{\zeta}^2}}}_{\text{Error due to noise}} + \underbrace{q\|\delta_\Theta\|\frac{33\hat{g}(\delta)}{N_r T_r \bar{\zeta}^2 + qN_p T_p \hat{\zeta}^2}}_{\text{Error due to difference between true and auxiliary systems}}, \tag{2.9}$$

where

$$\phi = \phi(N_r, T_r, N_p, T_p, q) = \frac{33(\bar{g}(\delta) + q\hat{g}(\delta))}{N_r T_r \bar{\zeta}^2 + qN_p T_p \hat{\zeta}^2} + 1,$$

$$\bar{g}(\delta) = N_r \sum_{t=0}^{T_r - 1} (\mathrm{tr}(\bar{G}_t) + p)(\frac{1}{c}\log(\frac{2}{\delta}) + 1)\bar{\sigma}^2_{max},$$

$$\hat{g}(\delta) = N_p \sum_{t=0}^{T_p - 1} (\mathrm{tr}(\hat{G}_t) + p)(\frac{1}{c}\log(\frac{2}{\delta}) + 1)\hat{\sigma}^2_{max}.$$

Remark 2. ***Interpretation of Theorem 2.3.1.*** *Recall that* N_r *is the number of rollouts from the true system (2.1),* T_r *is the length of each rollout of the true system (2.1),* N_p *is the number of rollouts from the auxiliary system (2.2), and* T_p *is the length of each rollout of the auxiliary system (2.2). Consequently, the quantities* $N_r T_r$ *and* $N_p T_p$ *capture the total number of samples from the true system and the auxiliary system, respectively. Further, recall that* $\sigma_{\bar{w}}, \sigma_{\hat{w}}$ *capture the noise levels from the two systems, and* $\|\delta_\Theta\|$ *captures the difference between the two system models. For strictly stable systems (2.1)-(2.2),* $\bar{g}(\delta)$ *and* $\hat{g}(\delta)$ *grow at most linearly with respect to* T_r *and* T_p*. For marginally stable systems,* $\bar{g}(\delta)$ *and* $\hat{g}(\delta)$ *grow at most polynomially with respect to* T_r, T_p *(see Proposition 2.6.2 in section 2.6). The terms* $\bar{g}(\delta)$ *and* $\hat{g}(\delta)$ *grow at most linearly with respect to* N_r, N_p*, irrespective of the spectral radius of the systems. Consequently, the parameter* ϕ *remains bounded with respect to* T_r, T_p *for strictly stable systems, grows at most polynomially with respect to* T_r, T_p *for marginally stable systems, and remains bounded with respect to* N_r, N_p*, irrespective of the spectral radius of the systems. We discuss some further observations below.*

Error due to noise and error due to model difference: Theorem 2.3.1 decomposes the overall estimation error into the error due to noise (or the intrinsic error) and the error due to model difference. Suppose that $q = 1$, and both systems are strictly stable for now. The error due to noise depends on the noise levels from the true system and the auxiliary system, and can be reduced by increasing the number of samples from the true system and the auxiliary system (increase $N_r T_r$ or $N_p T_p$). Theorem 2.3.1 is an improvement over the result in [23], since Theorem 2.3.1 shows that one can reduce the error due to noise by increasing either the number of rollouts or the length of these rollouts (or both), whereas the result in [23] only shows the error due to noise can be reduced by increasing the number of rollouts. The dependence on $\sqrt{n+p}$ is due to the dimension of the system model we wish to learn. The error due to model difference depends on how similar the two systems are, and becomes smaller if the auxiliary system is more similar to the true system (smaller $\|\delta_\Theta\|$), or if there are more samples from the true system than auxiliary system (increase $N_r T_r$). Consequently, one can observe that increasing the number of samples from the auxiliary system helps to reduce the error due to noise, at the price of adding a portion of error due to model difference (note that the error due to model difference is always bounded when we increase N_p or T_p). In particular, when the two systems are exactly the same, i.e., $\|\delta_\Theta\| = 0$, Theorem 2.3.1 recovers the learning rate $\mathcal{O}(\frac{1}{\sqrt{N_r T_r + N_p T_p}})$, which qualitatively matches the learning rate as reported in [18], when all samples are generated from the same system.

When the two systems are both marginally stable, one can see that the error due to noise can still go to zero as T_r, T_p increase, since the term ϕ grows at most polynomially with respect to T_r, T_p. However, the error due to model difference may amplify as T_p increases. We provide a slightly refined bound for large $N_p T_p$ in Theorem 2.3.2 to capture this case.

The benefits of collecting multiple trajectories: The existing literature has shown that the multiple trajectories setup has the benefit of handling unstable systems (when all samples are collected from the true system), since restarting the system from an initial state prevents the system state from going to infinity over time [13]. This benefit is captured by our result. In particular, fixing T_r, T_p, q, one can observe that the error due to noise always goes to zero as we increase N_r or N_p, irrespective of the spectral radius of the two systems, since the parameter ϕ is bounded. Further, the error due to model difference always goes to zero as

we increase N_r, and remains constant as we increase N_p, again irrespective of the spectral radius of the two systems.[5]

The selection of weight parameter q: In practice, selecting a good weight parameter q based on Theorem 2.3.1 requires an oracle, since one has to know the specific values of the different parameters in Theorem 2.3.1. Further, due to the different realizations of random variables, the optimal weight might differ at each experiment. A commonly used approach for tuning parameters in the training process is to leverage a cross-validation process (see [27] for an overview). In section 2.3.2, we also provide a data-dependent bound, which is computable and can help one to select a good value of q based on data. However, general guidelines can be given based on the upper bound provided by Theorem 2.3.1 when N_p or T_p is large and $\|\delta_\Theta\|$ is small, supposing that the two systems are strictly stable (for simplicity):

- *When $N_r T_r$ is small, we can set q to be relatively large to make sure that the first term in the error bound is small (use the auxiliary data to reduce the error due to noise). Consequently, the error bound is essentially dominated by the second term, which is small if the two systems are similar. This corresponds to the case where we have little data from the true system, and thus there may be a large identification error due to using only that data. In this case, it is worth placing more weight on the data from the auxiliary system, up to the point that the reduction in estimation error due to the larger amount of data is balanced out by the differences between the systems.*

- *When $N_r T_r$ is large, we can decrease q to reduce the second term as well, since the first term is already made small enough. This corresponds to the case where we have a large amount of data from the true system, and only need the data from the auxiliary system to slightly improve our estimates. In this case, we place a lower weight on the auxiliary data in order to avoid excessive bias due to the difference in the dynamics of the two systems.*

Furthermore, Theorem 2.3.1 demonstrates how the weight parameter should scale. For example, it can be verified that one can set $q = \mathcal{O}(\frac{1}{\sqrt{N_r}})$ to ensure consistency, when N_p grows

[5]↑In fact, the literature has shown that unstable systems that satisfy a certain regularity condition can also be consistently estimated using a single trajectory [11]. It would be interesting to capture that in our setup in future work.

linearly with respect to N_r (when T_r, T_p are fixed). These ideas will also be illustrated experimentally in Section 2.4.

Finally, the following corollary of Theorem 2.3.1 provides a sufficient condition under which using the data from the auxiliary system (setting $q \neq 0$) leads to a smaller error bound compared to using data only from the true system (setting $q = 0$), when both the true system and the auxiliary system are strictly stable.

Corollary 1. *Suppose that both system (2.1) and system (2.2) are strictly stable, i.e., $\rho(\bar{A}) < 1$ and $\rho(\hat{A}) < 1$. Consider the estimation error bound provided in Theorem 2.3.1. Suppose that q satisfies the following inequality:*

$$\begin{aligned} &\frac{\sigma_{\bar{w}}\sqrt{\log\frac{9^n}{\delta} + (n+p)\log(\frac{33\bar{g}(\delta)}{N_r T_r \bar{\zeta}^2} + 1)}}{\sqrt{N_r T_r \bar{\zeta}^2}} \\ &> \frac{\max(\frac{\sigma_{\bar{w}}}{\sqrt{q}}, \sigma_{\hat{w}})\sqrt{\log\frac{9^n}{\delta} + (n+p)\log(\frac{\gamma}{\zeta^2} + 1)}}{\sqrt{N_p T_p \hat{\zeta}^2}} \\ &+ \|\delta_\Theta\|\frac{\gamma}{20\hat{\zeta}^2}, \end{aligned} \tag{2.10}$$

where $\zeta = \min(\bar{\zeta}, \hat{\zeta})$, and γ is any positive constant that satisfies

$$\begin{aligned} \max(33(\mathrm{tr}(\bar{G}_t) + p)(\frac{1}{c}\log(\frac{2}{\delta}) + 1)\bar{\sigma}^2_{max}, \\ 33(\mathrm{tr}(\hat{G}_t) + p)(\frac{1}{c}\log(\frac{2}{\delta}) + 1)\hat{\sigma}^2_{max}) \leq \gamma \end{aligned}$$

for all $t \geq 0$. Then the resulting error bound will be smaller than the error bound obtained using $q = 0$.

Remark 3. ***Interpretation of Corollary 1.*** *Note that γ always exists since $\mathrm{tr}(\bar{G}_t)$ and $\mathrm{tr}(\hat{G}_t)$ are bounded for strictly stable systems (see Proposition 2.6.2 in section 2.6). We also note that the above condition might be conservative, and may not be easily checked in practice since it involves unknown parameters. However, we describe the insights provided by this condition here. One may observe that condition (2.10) is more likely to hold if $\|\delta_\Theta\|$ is small (the true system and the auxiliary system shares "similar" dynamics), and $N_p T_p$ is*

large (one has abundant samples from the auxiliary system), as these conditions can make the right hand side of the inequality smaller. In other words, the auxiliary samples tend to be more informative in such cases. In contrast, condition (2.10) is less likely to hold if $N_r T_r$ is large, since it will make the left hand side of the inequality smaller, i.e., if we already have a lot of samples from the true system, then the auxiliary samples tend to be less informative. The effect of the noise can be quite subtle, since it shows up in various places. However, loosely speaking, having a smaller $\sigma_{\hat{w}}$ while assigning higher weight q may still help to make the right hand side of the inequality smaller by making the term $\max(\frac{\sigma_{\bar{w}}}{\sqrt{q}}, \sigma_{\hat{w}})$ smaller, when the terms $\hat{\zeta}$ and ζ are not affected too much. In other words, we might be able to benefit from the auxiliary system if the auxiliary system is not too noisy, and if we attach enough importance to the auxiliary samples.

The following result is a slightly refined bound for large $N_p T_p$.

Theorem 2.3.2. *Under the same conditions in Theorem 2.3.1, with probability at least $1 - 5\delta$, the weighted least squares estimate Θ_{WLS} from Algorithm 1 using $\lambda = 0$ satisfies*

$$\|\Theta_{WLS} - \Theta\| \leq \underbrace{\frac{20 \max(\sigma_{\bar{w}}, \sqrt{q}\sigma_{\hat{w}}) \sqrt{\log \frac{9^n}{\delta} + (n+p)\log(\phi)}}{\sqrt{N_r T_r \bar{\zeta}^2 + q N_p T_p \hat{\zeta}^2}}}_{\text{Error due to noise}} + \underbrace{\|\delta_\Theta\| (1 + \frac{33\bar{g}(\delta)}{N_r T_r \bar{\zeta}^2 + q N_p T_p \hat{\zeta}^2})}_{\text{Error due to difference between true and auxiliary systems}} . \tag{2.11}$$

Remark 4. *Theorem 2.3.2 yields a tighter bound when $N_p T_p$ is large. Fixing $q > 0$, it can be observed that the error bound converges exactly to $\|\delta_\Theta\|$ as T_p increases, when the auxiliary system is marginally stable or strictly stable ($\rho(\hat{A}) \leq 1$), since the ϕ term grows at most polynomially with respect to T_p, and the error due to model difference converges to $\|\delta_\Theta\|$. Further, the bound converges exactly to $\|\delta_\Theta\|$ as N_p increases, irrespective of the spectral radius of the system. In other words, one is essentially learning the dynamics of the auxiliary system when we use a lot of auxiliary samples.*

2.3.2 Data-dependent Bound

In this section, we provide a data-dependent upper bound of the system identification error using Algorithm 1, assuming $\lambda > 0$. The regularized solution with strictly positive λ helps us to establish the data-dependent bound, and provides the user with more flexibility to tune the estimate in practice. The bound is computable when some prior knowledge about the systems (as will be discussed below) is available, and applies to more general input and noise. One can also use it for the selection of weight parameter q and regularization parameter λ in practice (by selecting the weight parameter q and regularization parameter λ that give a smaller error bound).

Theorem 2.3.3. *Consider the systems (2.1)-(2.2), where the random vectors $\bar{w}_t, \bar{u}_t, \bar{x}_0, \hat{w}_t, \hat{u}_t, \hat{x}_0$ are independent, and $\bar{w}_t, \hat{w}_t$ are sub-Gaussian with parameters $\sigma_{\bar{w}}^2$ and $\sigma_{\hat{w}}^2$, respectively, for all $t \geq 0$. Fix $q \geq 0$, $\lambda > 0$, and $\delta > 0$. Let $V = \lambda I_{n+p}$ and $\bar{V} = (ZQZ' + V)V^{-1}$. With probability at least $1 - \delta$, the regularized weighted least squares estimate Θ_{WLS} obtained from running Algorithm 1 on the above systems satisfies*

$$\begin{aligned} \|\Theta_{WLS} - \Theta\| \leq & \frac{\max(\sigma_{\bar{w}}, \sqrt{q}\sigma_{\hat{w}})\sqrt{\frac{32}{9}(\log\frac{9^n}{\delta} + \frac{1}{2}\log\det(\bar{V}))}}{\sqrt{\lambda_{min}(ZQZ' + \lambda I_{n+p})}} \\ & + q\|\delta_\Theta\|\|\hat{Z}\hat{Z}'(ZQZ' + \lambda I_{n+p})^{-1}\| \\ & + \|\Theta\|\frac{\lambda}{\lambda_{min}(ZQZ' + \lambda I_{n+p})}. \end{aligned} \tag{2.12}$$

Remark 5. *Practically, one can compute the error bound in Theorem 2.3.3 using various different q and λ, and choose a value of q and λ that give the smallest error bound. We will illustrate the selection of q in Section 2.4. Note that the model difference term $\|\delta_\Theta\|$ in (2.12) can be replaced by an upper bound on the difference between the models (if that is available). In practice, an appropriate upper bound of the term $\|\delta_\Theta\|$ could be obtained using prior knowledge or previous estimates from data. For example, if one knows that an auxiliary system is different from the true system only in certain subsystems, where the entries are restricted to a fixed range, that information can be used to compute an upper bound of $\|\delta_\Theta\|$. Also, the difference between the subsystems of the true system and the auxiliary can*

be estimated from data using any existing techniques (e.g., least squares method). This may be helpful if doing experiments on the subsystems is easy. The bound on $\|\Theta\|$ can be obtained similarly using prior knowledge (e.g., using known range on entries). The noise distribution of the two systems and their corresponding sub-Gaussian parameters can also be estimated from data [28].

2.4 Numerical Experiments to Illustrate Various Scenarios for System Identification from Auxiliary Data

In order to validate our main results in Theorem 2.3.1 and Theorem 2.3.3 and gain more insights, we now provide numerical examples of the weighted least squares-based system identification algorithm (Algorithm 1). All of the numerical results are averaged over 10 independent experiments.

2.4.1 Predetermined weight

In this section, we provide numerical experiments using various predetermined weight parameters q. Such a situation may occur if we have a firm belief that the auxiliary system has similar dynamics to the true system, but upper bounds on $\|\delta_\Theta\|, \sigma_{\bar{w}}$ and $\sigma_{\hat{w}}$ are not available. Setting $\lambda = 0$, the experiments are performed using the following true system and auxiliary system:

$$\bar{A} = \begin{bmatrix} 0.6 & 0.5 & 0.4 \\ 0 & 0.5 & 0.4 \\ 0 & 0 & 0.4 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \\ 0.5 & 0.5 \end{bmatrix}, \tag{2.13}$$

$$\hat{A} = \begin{bmatrix} 0.7 & 0.5 & 0.4 \\ 0 & 0.5 & 0.4 \\ 0 & 0 & 0.4 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 1.1 & 0.5 \\ 0.5 & 1 \\ 0.5 & 0.5 \end{bmatrix}. \tag{2.14}$$

We set $\bar{x}_0, \hat{x}_0, \bar{u}_t, \hat{u}_t, \bar{w}_t, \hat{w}_t$ to be zero mean Gaussian random vectors with covariance matrices being identity matrices. The model difference of the above two systems is $\|\delta_\Theta\| \approx 0.1414$. The numbers of rollouts N_r and N_p are set to be 1. We provide experiments to illustrate various scenarios, including those we mentioned earlier in Section 2.1. Note that the experiments in

this section are conducted by varying the lengths of the trajectories from the two systems, while the experiments in [23] are performed by varying the number of trajectories from the two systems.

2.4.1.1 Scenario 1: Both T_r and T_p are increasing

In the first experiment, we set the length of the trajectory from the auxiliary system be $T_p = 3T_r$. In practice, one may encounter such a scenario when gathering data from the true system is time consuming or costly, whereas gathering data from an auxiliary system (such as a simulator) is faster or cheaper.

In Fig. 2.1, we plot the estimation error $\|\Theta - \Theta_{WLS}\|$ versus T_r using different weight parameters q. As expected, when one does not have enough data from the true system (T_r is small), setting $q > 0$ leads to a smaller estimation error of system matrices. However, the curve for $q = 1$ and $q = 10^{10}$ (corresponding to treating all samples equally and paying almost no attention to the samples from the true system, respectively) eventually plateau and incur more error than not using the the auxiliary data ($q = 0$). This phenomenon matches with the theoretical guarantee in Theorem 2.3.1. Specifically, when q is a nonzero constant and both T_p and T_r are increasing in a linear relationship, it can be verified that the upper bound in Theorem 2.3.1 will not go to zero as T_r increases. In other words, there is no need to attach high importance to the auxiliary data when one has enough data from the true system. In contrast, setting q to be diminishing with T_r could perform consistently better than $q = 0$ in this example, even when T_r becomes large. Indeed, one can choose $q = \mathcal{O}(\frac{1}{\sqrt{T_r}})$ in the upper bound given by (2.9) in Theorem 2.3.1, and show that the upper bound becomes $\mathcal{O}(\frac{1}{\sqrt{T_r}})$.

Key Takeaway: When T_p and T_r are both increasing linearly, having q diminish with T_r with a rate of $\mathcal{O}(\frac{1}{\sqrt{T_r}})$ helps to reduce the system identification error when T_r is small (by leveraging data from the auxiliary system), and avoids excessive bias from the auxiliary system when T_r is large.

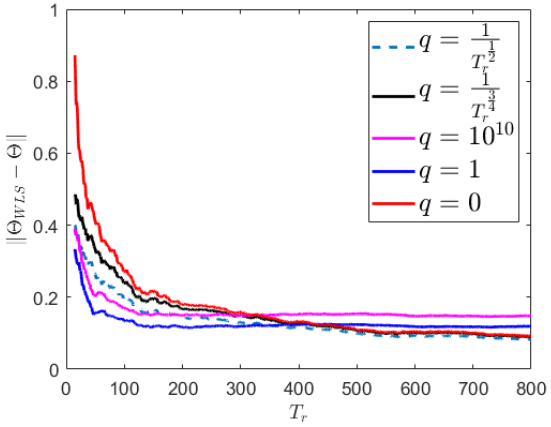

Figure 2.1. Scenario 1: Both T_r and T_p increase over time ($T_p = 3T_r$). Choosing $q = \mathcal{O}(\frac{1}{\sqrt{T_r}})$ strikes a good balance between reducing error when T_r is small and ensuring consistency when T_r is large

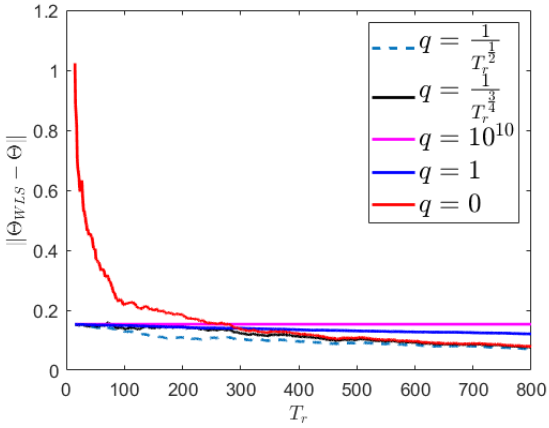

Figure 2.2. Scenario 2: T_p is fixed, and T_r increases over time. Having q diminish with T_r could reduce the error when T_r is small, and avoid unwanted bias from the auxiliary system when T_r is large

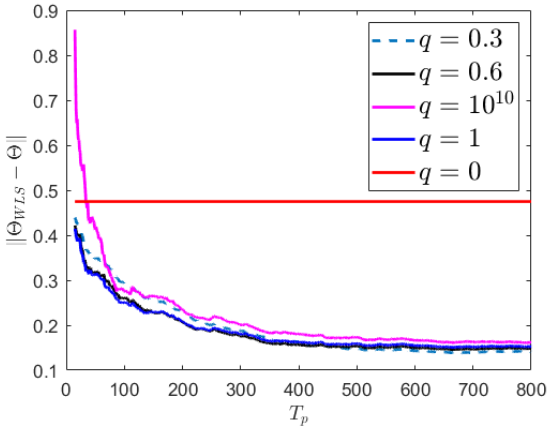

Figure 2.3. Scenario 3: T_r is fixed, and T_p increases over time. Setting q to be relatively balanced could make the error smaller than the extreme cases ($q = 0, 10^{10}$)

2.4.1.2 Scenario 2: T_p is fixed but T_r is increasing

For the second experiment, we fix the number of samples from the auxiliary system to be $T_p = 2400$, and look at what happens as the number of samples from the true system increases. In practice, one may encounter such a scenario when the system dynamics change at some point in time (e.g., due to faults). In this case, the true system we want to learn is the one after the fault, and the auxiliary system is the one prior to the fault. Consequently, while the data from the old (auxiliary) system may not accurately represent the new (true) system dynamics, leveraging the old data might be beneficial in this case.

In Fig. 2.2, we plot the estimation error versus T_r for different weight parameters q. As expected, setting $q > 0$ leads to a much smaller error during the initial phase when T_r is small. This can be confirmed by Theorem 2.3.1 since the overall estimation error is essentially the error due to the model difference. Namely, the auxiliary data helps to build a good initial estimate when T_r is small. When we set the weight to be $q = 10^{10}$, we are paying little attention to the samples from the true system, i.e., we are not gaining any new information as we collect more data from the true system. Consequently, the error is almost a flat line as T_r increases when $q = 10^{10}$. As can be observed from Theorem 2.3.1, when T_p is fixed, we can always make the error go to 0 as we increase T_r, using the weights we selected in this experiment. However, when q is set to be too large, it could make the error even larger due to the model difference (or bias) introduced by the auxiliary system. This is captured by Theorem 2.3.1 since when q is set to be too large (such that qT_p is large compared to T_r), even when T_r becomes larger, the second term in the error bound (2.9) (capturing model difference) is still large.

Key Takeaway: When T_p is fixed and large, and T_r increases over time, setting q to be nonzero builds a good initial estimate for the true system dynamics when we have little data from the true system. Again, having q diminish with T_r could reduce the system identification error when T_r is small, and avoid unwanted bias from the auxiliary system when T_r is large.

2.4.1.3 Scenario 3: T_r is fixed but T_p is increasing

In the last experiment, we fix the number of samples from the true system to be $T_r = 50$. As discussed earlier, one may encounter such a scenario when one has only a limited amount of time to gather data from the true system. Consequently, leveraging information from other "similar" systems (e.g., from a reasonably accurate simulator) could be helpful to augment the data. This is the most subtle case, since Theorem 2.3.1 does not ensure consistency when T_r is fixed.

In Fig. 2.3, we plot the the estimation error versus T_p using different weight parameters q. As it can be seen, setting $q = 0$ (not using the auxiliary samples) gives a flat line, which represents the error we can achieve purely based on $T_r = 50$ samples from the true system. When $q = 10^{10}$, we are paying little attention to the true system, and essentially learning the dynamics of the auxiliary system. In contrast, the results for $q = 1, 0.6, 0.3$ suggest that setting a relatively balanced weight q to the auxiliary data could make the error smaller than the two extreme cases ($q = 0, 10^{10}$) in this example. However, in practice, one may want to leverage a cross-validation process to tune the hyper-parameter q, when there is not enough prior knowledge about the dynamics of the true system and the auxiliary system.

Key Takeaway: Although consistency cannot be guaranteed when T_r is fixed and T_p increases over time, a relatively balanced q could make the error smaller than the extreme cases ($q = 0, 10^{10}$).

2.4.2 Selecting weight based on Theorem 2.3.3

In this section, we study selecting the weight parameter q using Theorem 2.3.3 using a fixed regularization parameter λ. We plot the true error $\|\Theta - \Theta_{WLS}\|$ and the theoretical data-dependent bound in Theorem 2.3.3 as a function of weight parameter q varying from 0 to 2, where the increment is set to be 0.01. This corresponds to the situation where some upper bounds on $\|\delta_\Theta\|, \sigma_{\bar{w}}$ and $\sigma_{\hat{w}}$ are available (as discussed in Remark 5). We set the confidence parameter to be $\delta = 0.01$, and all other parameters in Theorem 2.3.3 are assumed to be known exactly for simplicity. The system matrices of the true system and the auxiliary system are set to be

$$\bar{A} = \begin{bmatrix} 0.6 & 0.5 & 0.4 \\ 0 & 0.5 & 0.4 \\ 0 & 0 & 0.4 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \\ 0.5 & 0.5 \end{bmatrix}, \tag{2.15}$$

$$\hat{A} = \begin{bmatrix} 0.7 & 0.5 & 0.4 \\ 0 & 0.5 & 0.4 \\ 0 & 0 & 0.4 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 1+\mathbf{\Delta} & 0.5 \\ 0.5 & 1 \\ 0.5 & 0.5 \end{bmatrix}. \tag{2.16}$$

We set the regularization parameter to be $\lambda = 1$. We set $\bar{x}_0, \hat{x}_0, \bar{u}_t, \hat{u}_t, \bar{w}_t, \hat{w}_t$ to be zero mean Gaussian random vectors, where the covariance matrices of $\bar{x}_0, \hat{x}_0, \bar{u}_t, \hat{u}_t$ are set to be identity matrices. The trajectory lengths are set to be $T_r = 10, T_p = 50$, and the number of trajectories from the auxiliary system is set to be $N_p = 20$. The covariance matrices of $\bar{w}_t, \hat{w}_t$ are set to be $\sigma_{\bar{w}}^2 I_{n+p}, \sigma_{\hat{w}}^2 I_{n+p}$, and the values of $\mathbf{\Delta}, \sigma_{\bar{w}}, \sigma_{\hat{w}}, N_r$ are specified under the figures.

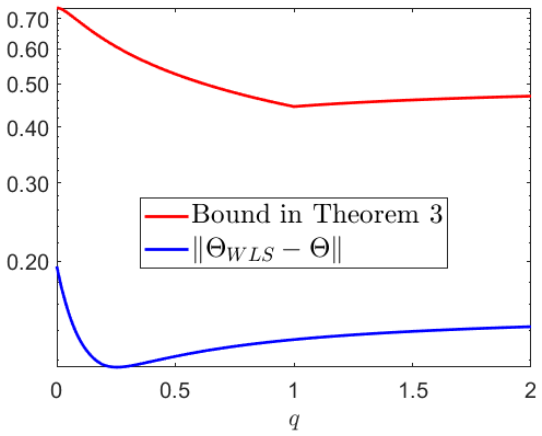

Figure 2.4. Baseline case 1: $\mathbf{\Delta} = 0.1, \sigma_{\bar{w}} = \sigma_{\hat{w}} = 1, N_r = 20$. An intermediate value of weighting parameter q is optimal

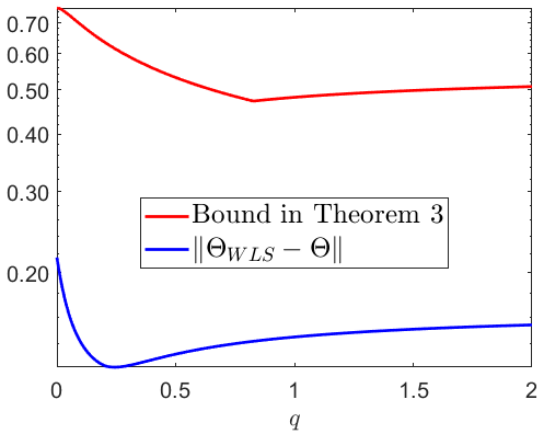

Figure 2.5. Baseline case 2: $\mathbf{\Delta} = 0.11, \sigma_{\bar{w}} = 1, \sigma_{\hat{w}} = 1.1, N_r = 19$. An intermediate value of weighting parameter q is optimal

As can be seen in Fig. 2.4 and Fig. 2.5, setting q to be non-zero could result in smaller error bounds, which show the benefits of leveraging the auxiliary data. However, for the values of the weight q we plotted, the optimal weights that obtain the smallest error bounds

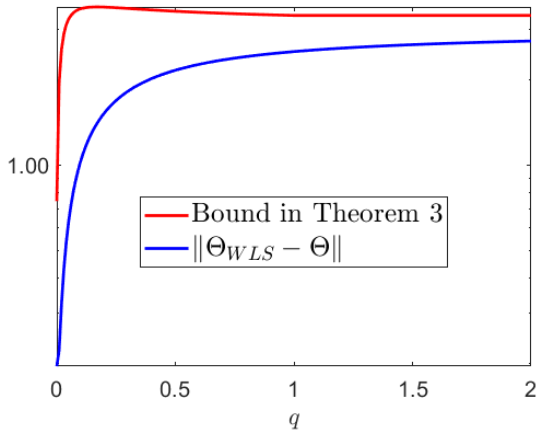

Figure 2.6. Large model difference: $\mathbf{\Delta} = 3, \sigma_{\bar{w}} = \sigma_{\hat{w}} = 1, N_r = 20$. In this case, it is optimal to not use data from the auxiliary system ($q = 0$)

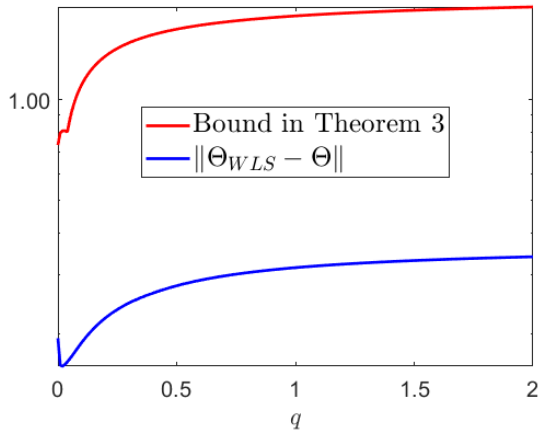

Figure 2.7. Noisy auxiliary system: $\mathbf{\Delta} = 0.1, \sigma_{\bar{w}} = 1, \sigma_{\hat{w}} = 5, N_r = 20$. In this case, it is optimal to not use data from the auxiliary system ($q = 0$)

in Theorem 2.3.3 do not align with the optimal weights q that minimize the true error $\|\Theta - \Theta_{WLS}\|$. Such mismatches could be due to the conservativeness of the bound. On the other hand, the optimal weight from the bound still captures how the true optimal weight should scale. As can be seen in Fig. 2.6, Fig. 2.7, and Fig. 2.9, the optimal weight for both the bound and the true error tend to be small when (1) the model difference is large; or (2) the auxiliary system becomes much more noisy; or (3) when one has a large number of samples from the true system. Such empirical results also match with our observations in Corollary 1 when $\lambda = 0$. In Fig. 2.8, both the optimal weight from our bound and the true optimal weight are greater than 1, since the true system is very noisy and hence the data from the true system tend to be less informative compared to the data from the auxiliary system. We further note that, in practice, selecting the exact optimal weight q that minimizes $\|\Theta_{WLS} - \Theta\|$ is very hard, and one would instead focus on selecting a relatively good weight q. The weight that results in a small error bound can be integrated with techniques like robust control to improve the overall system performance guarantee.

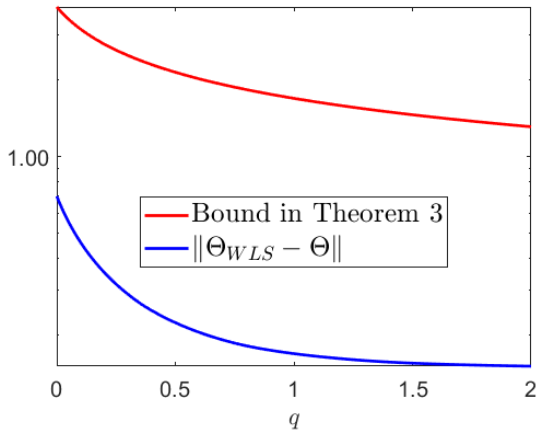

Figure 2.8. Noisy true system: $\mathbf{\Delta} = 0.1, \sigma_{\bar{w}} = 5, \sigma_{\hat{w}} = 1, N_r = 20$. In this case, it is optimal to assign higher weight to the auxiliary system ($q = 2$)

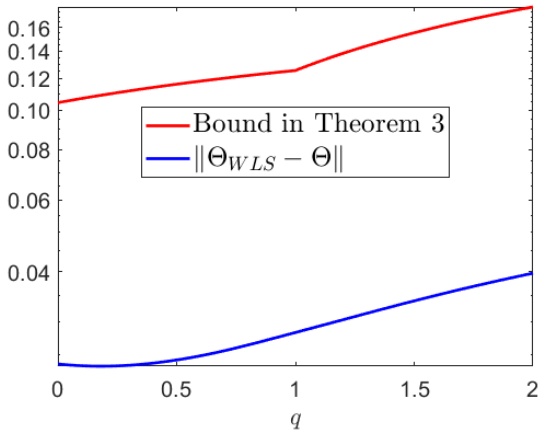

Figure 2.9. Large number of true samples: $\mathbf{\Delta} = 0.1, \sigma_{\bar{w}} = \sigma_{\hat{w}} = 1, N_r = 1200$. In this case, it is optimal to not use data from the auxiliary system ($q = 0$)

2.5 Chapter Summary

In this chapter, we provided finite sample analysis of system identification using a weighted least squares approach, when one has an auxiliary system that shares similar dynamics as the true system we want to learn. The analysis improves the result in [23] as we show the error due to noise can be reduced by increasing either the number of trajectories or the trajectory length of the true system and the auxiliary system, or both. Our analysis provides insights on the benefits of using the auxiliary system, and how to weight the data from the auxiliary system. We also provided a data-dependent bound that is computable when some prior knowledge about the systems is available, which is tighter and can be used to determine the appropriate weight parameter in the training process.

There are various directions for future research. First, as shown in [11], [12], the least squares estimator is consistent for certain types of unstable systems even if multiple trajectories of data are not available. It would be interesting to study how to capture that in our analysis. Second, it would be of interest to relax the conditions on the full rankness of the covariance matrix of noise/input in our setup such that one could handle systems with longer

memory. Another interesting direction is to develop lower bounds for such transfer learning-based system identification methods, which could potentially enable the development of an optimal estimator. Some possible approaches are to leverage assumptions like sparsity [20] or prior knowledge, e.g., known auxiliary system model. Finally, studying how to leverage the idea of learning from similar systems/transfer learning in control-related problems would also be a rich area for future research [29].

2.6 Proofs of Results

2.6.1 Intermediate Results

We will leverage the following Hanson-Wright inequality to upper bound the terms $\|\bar{Z}\bar{Z}'\|$ and $\|\hat{Z}\hat{Z}'\|$.

Lemma 1. *[30, Theorem 1.1] Let* $X = \begin{bmatrix} X_1 & \dots X_n \end{bmatrix}' \in \mathbb{R}^n$ *be a random vector with independent components* X_i *which satisfy* $\mathbb{E}[X_i] = 0$ *and* $\|X_i\|_{\psi_2} \leq K$*, where* $\|\cdot\|_{\psi_2}$ *denotes the sub-Gaussian norm, i.e.,* $\|X_i\|_{\psi_2} = \inf\{\zeta > 0 : \mathbb{E}[\exp(X_i^2/\zeta^2)] \leq 2\}$*. Let* A *be an* $n \times n$ *matrix. Then, for every* $t \geq 0$*, we have*

$$P(|X'AX - \mathbb{E}[X'AX]| > t) \leq 2\exp(-c_0 \min(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|})),$$

where c_0 *is some positive universal constant.*

We have the following result.

Lemma 2. *Let Assumption 1 hold. For any fixed* $\delta \in (0, \frac{2}{e})$*, each of the following inequalities holds with probability at least* $1 - \delta$*:*

$$\|\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t\bar{z}_t'\| \leq \bar{g}(\delta),$$

$$\|\sum_{i=1}^{N_r}\sum_{t=0}^{T_p-1} \hat{z}_t\hat{z}_t'\| \leq \hat{g}(\delta),$$

where

$$\bar{g}(\delta) = N_r \sum_{t=0}^{T_r-1} (\mathrm{tr}(\bar{G}_t) + p)(\frac{1}{c}\log(\frac{2}{\delta}) + 1)\bar{\sigma}_{max}^2,$$

$$\hat{g}(\delta) = N_p \sum_{t=0}^{T_p-1} (\mathrm{tr}(\hat{G}_t) + p)(\frac{1}{c}\log(\frac{2}{\delta}) + 1)\hat{\sigma}^2_{max},$$

and c is some positive constant.

Proof. We will only show the first inequality since the analysis for the second one is essentially the same. Let $\mathbf{Z}_{T_r}^{N_r} = \begin{bmatrix} \mathbf{Z}^{1'} & \dots & \mathbf{Z}^{N_r'} \end{bmatrix}' \in \mathbb{R}^{(n+p)N_rT_r}$, where $\mathbf{Z}^i = \begin{bmatrix} \bar{z}_0^{i'} & \dots & \bar{z}_{T_r-1}^{i'} \end{bmatrix}' \in \mathbb{R}^{(n+p)T_r}$ for $i = 1, \dots, N_r$. We have

$$\| \sum_{i=1}^{N_r} \sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'} \| \leq \sum_{i=1}^{N_r} \sum_{t=0}^{T_r-1} \bar{z}_t^{i'} \bar{z}_t^i = \mathbf{Z}_{T_r}^{N_r'} \mathbf{Z}_{T_r}^{N_r}. \tag{2.17}$$

Note that we have

$$\mathbf{Z}_{T_r}^{N_r'} \mathbf{Z}_{T_r}^{N_r} \leq \mathbb{E}[\mathbf{Z}_{T_r}^{N_r'} \mathbf{Z}_{T_r}^{N_r}] + |\mathbf{Z}_{T_r}^{N_r'} \mathbf{Z}_{T_r}^{N_r} - \mathbb{E}[\mathbf{Z}_{T_r}^{N_r'} \mathbf{Z}_{T_r}^{N_r}]|. \tag{2.18}$$

Now we will upper bound the two terms after the inequality in (2.18). We consider the term $\mathbb{E}[\mathbf{Z}_{T_r}^{N_r'} \mathbf{Z}_{T_r}^{N_r}]$ first. Let

$$H = \begin{bmatrix} H_1 & H_2 \end{bmatrix} \in \mathbb{R}^{(n+p)T_r \times (n+p)T_r},$$

where $H_1 \in \mathbb{R}^{(n+p)T_r \times nT_r}$ is defined as

$$H_1 = \begin{bmatrix} I_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \bar{A} & I_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \bar{A}^2 & \bar{A} & I_n & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{A}^{T_r-1} & \bar{A}^{T_r-2} & \bar{A}^{T_r-3} & \cdots & \bar{A} & I_n \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and $H_2 \in \mathbb{R}^{(n+p)T_r \times pT_r}$ is defined as

$$H_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ I_p & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \bar{B} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \bar{A}\bar{B} & \bar{B} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{A}^{T_r-2}\bar{B} & \bar{A}^{T_r-3}\bar{B} & \bar{A}^{T_r-4}\bar{B} & \cdots & \bar{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & I_p \end{bmatrix},$$

where we use $\mathbf{0}$ to denote zero matrices with appropriate dimensions. Further, let $\mathbf{H} = \operatorname{diag}(H, \cdots, H) \in \mathbb{R}^{(n+p)N_rT_r \times (n+p)N_rT_r}$ and $g = \begin{bmatrix} g^{1'} & g^{2'} & \cdots & g^{N_r'} \end{bmatrix}' \in \mathbb{R}^{(n+p)N_rT_r}$, where $g^i = \begin{bmatrix} \bar{x}_0^{i'} & \bar{w}_0^{i'} & \cdots & \bar{w}_{T_r-2}^{i'} & \bar{u}_0^{i'} & \cdots & \bar{u}_{T_r-1}^{i'} \end{bmatrix}' \in \mathbb{R}^{(n+p)T_r}$ for $i = 1, \ldots, N_r$. With these definitions, we have $\mathbf{H}g = \begin{bmatrix} \mathbf{Z}^{1'} & \ldots & \mathbf{Z}^{N_r'} \end{bmatrix}' = \mathbf{Z}_{T_r}^{N_r}$, and hence

$$\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r} = g'\mathbf{H}'\mathbf{H}g = \operatorname{tr}(g'\mathbf{H}'\mathbf{H}g) = \operatorname{tr}(gg'\mathbf{H}'\mathbf{H}). \tag{2.19}$$

Taking the expectation, and from the relationship $\operatorname{tr}(AB) \leq \lambda_{max}(A)\operatorname{tr}(B)$ for real symmetric A and real $B \succeq 0$ [31], we have

$$\begin{aligned} \mathbb{E}[\operatorname{tr}(gg'\mathbf{H}'\mathbf{H})] &= \operatorname{tr}(\mathbb{E}[gg']\mathbf{H}'\mathbf{H}) \leq \|\mathbb{E}[gg']\| \operatorname{tr}(\mathbf{H}'\mathbf{H}) \\ &= \bar{\sigma}_{max}^2 \operatorname{tr}(\mathbf{H}'\mathbf{H}) = \bar{\sigma}_{max}^2 N_r \operatorname{tr}(H'H) \\ &= \bar{\sigma}_{max}^2 N_r \Big(\sum_{t=0}^{T_r-1} \operatorname{tr}(\bar{G}_t) + \sum_{k=0}^{T_r-1} \operatorname{tr}(I_p)\Big) \\ &= \bar{\sigma}_{max}^2 N_r \sum_{t=0}^{T_r-1} (\operatorname{tr}(\bar{G}_t) + p), \end{aligned} \tag{2.20}$$

where $\bar{G}_t$ is defined in (2.8). Now we consider the term $|\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r} - \mathbb{E}[\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r}]|$ in (2.18). From (2.19), we have

$$|\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r} - \mathbb{E}[\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r}]| = |g'\mathbf{H}'\mathbf{H}g - \mathbb{E}[g'\mathbf{H}'\mathbf{H}g]|.$$

From [32], we have each component of g has sub-Gaussian norm that is upper bounded by $4\bar{\sigma}_{max}$. We can apply Lemma 1 to the above term with the replacement of c_0 by $\min(1, c_0)$ to obtain

$$\begin{aligned} &P(|g'\mathbf{H}'\mathbf{H}g - \mathbb{E}[g'\mathbf{H}'\mathbf{H}g]| > t) \\ &\leq 2\exp(-\min(\frac{c_0 t^2}{256\bar{\sigma}_{max}^4\|\mathbf{H}'\mathbf{H}\|_F^2}, \frac{c_0 t}{16\bar{\sigma}_{max}^2\|\mathbf{H}'\mathbf{H}\|})) \\ &\leq 2\exp(-\min(\frac{ct^2}{\bar{\sigma}_{max}^4\|\mathbf{H}'\mathbf{H}\|_F^2}, \frac{ct}{\bar{\sigma}_{max}^2\|\mathbf{H}'\mathbf{H}\|})), \end{aligned} \tag{2.21}$$

where $c \triangleq \frac{c_0}{256}$.

Fixing $\delta \in (0, \frac{2}{e})$ and setting $t = \frac{1}{c}\log(\frac{2}{\delta})\bar{\sigma}_{max}^2 \operatorname{tr}(\mathbf{H}'\mathbf{H})$, we have

$$\begin{aligned} \frac{ct^2}{\bar{\sigma}_{max}^4\|\mathbf{H}'\mathbf{H}\|_F^2} &= \frac{1}{c}(\frac{\operatorname{tr}(\mathbf{H}'\mathbf{H})}{\|\mathbf{H}'\mathbf{H}\|_F})^2(\log(\frac{2}{\delta}))^2 \\ &\geq (\log(\frac{2}{\delta}))^2 \geq \log(\frac{2}{\delta}), \end{aligned}$$

and

$$\frac{ct}{\bar{\sigma}_{max}^2\|\mathbf{H}'\mathbf{H}\|} = \log(\frac{2}{\delta})\frac{\operatorname{tr}(\mathbf{H}'\mathbf{H})}{\|\mathbf{H}'\mathbf{H}\|} \geq \log(\frac{2}{\delta}),$$

where we used the fact that $\|\mathbf{H}'\mathbf{H}\| \leq \|\mathbf{H}'\mathbf{H}\|_F \leq \|\mathbf{H}\|_F^2 = \operatorname{tr}(\mathbf{H}'\mathbf{H})$.

Combining the above inequalities with (2.21), we have with probability at least $1 - \delta$

$$\begin{aligned} |\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r} - \mathbb{E}[\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r}]| &= |g'\mathbf{H}'\mathbf{H}g - \mathbb{E}[g'\mathbf{H}'\mathbf{H}g]| \\ &\leq \frac{1}{c}\log(\frac{2}{\delta})\bar{\sigma}_{max}^2 \operatorname{tr}(\mathbf{H}'\mathbf{H}) \\ &= \frac{1}{c}\log(\frac{2}{\delta})\bar{\sigma}_{max}^2 N_r \sum_{t=0}^{T_r-1}(\operatorname{tr}(\bar{G}_t) + p). \end{aligned}$$

Consequently, considering the above inequality in conjunction with (2.20), and from (2.18), we have with probability at least $1-\delta$

$$\mathbf{Z}_{T_r}^{N_r'}\mathbf{Z}_{T_r}^{N_r} \leq N_r \sum_{t=0}^{T_r-1} (\mathrm{tr}(\bar{G}_t)+p)(\frac{1}{c}\log(\frac{2}{\delta})+1)\bar{\sigma}_{max}^2.$$

□

Remark 6. *The constant c in Lemma 2 depends on the constant c_0 in the Hanson-Wright inequality in Lemma 1. Attempts to explicitly characterize c_0 can be found in [33], [34]. One can also derive similar upper bounds using the Markov inequality to get rid of the constant c, but at the price of having linear dependence on δ in the denominators of the bounds.*

We will leverage the following definitions on ϵ-net and the block martingale small-ball conditions in [7].

Definition 2.6.1. *Let (T,d) be a metric space. Consider a subset $K \subset T$ and let $\epsilon > 0$. A subset $\mathbf{N} \subseteq K$ is called an ϵ-net of K if every point in K is within distance ϵ of some point of $\mathbf{N}$, i.e.,*

$$\forall x \in K \quad \exists x_0 \in \mathbf{N} : d(x.x_0) \leq \epsilon.$$

Definition 2.6.2. *[7, Definition 2.1]) Let $\{Z_t\}_{t\geq 1}$ be a $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}$. We say $\{Z_t\}_{t\geq 1}$ satisfies the (k,v,p)-block martingale small-ball (BMSB) condition if, for any $j \geq 0$, one has $\frac{1}{k}\sum_{i=1}^{k} P(|Z_{j+i}| \geq v|\mathcal{F}_j) > p$ almost surely.*

The following result establishes a lower bound of the smallest eigenvalue of the sample covariance matrix for general time series, leveraging the above definitions, a concentration inequality in [35], and the ideas in [36]. Note that we use $v(i)$ to denote the i-th component of a vector v.

Lemma 3. *Let $\{l_t\}_{t\geq 0}$ be a sequence of random vectors that is adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $l_t \in \mathbb{R}^d$. Let $\{\eta_t\}_{t\geq 1}$ be another sequence of random vectors such that η_t is $\mathcal{F}_t$-measurable, where $\eta_t \in \mathbb{R}^d$. Further, suppose $\eta_{t+1}|\mathcal{F}_t$ has zero mean and independent coordinates, where each coordinate has bounded fourth moment for all $t \geq 0$. Suppose that $\mathbb{E}[\eta_{t+1}\eta_{t+1}'|\mathcal{F}_t] \succeq \sigma_\eta^2 I_d$, and $\max_{1\leq i\leq d} \frac{\mathbb{E}[\eta_{t+1}(i)^4|\mathcal{F}_t]}{\mathbb{E}[\eta_{t+1}(i)^2|\mathcal{F}_t]^2} \leq c_\eta$ for all $t \geq 0$, where $\sigma_\eta, c_\eta \in \mathbb{R}_{>0}$.*

Let $\zeta = \frac{\sigma_\eta}{c_1 c_\eta}$, *where* c_1 *is a positive absolute constant. Define the sequence* $z_{t+1} = l_t + \eta_{t+1}$ *for* $t \geq 0$, *where* $z_0 \in \mathbb{R}^d$. *Fix* $\delta > 0$ *and a constant* $M \geq \frac{\zeta^2(T-1)}{16}$ *such that it holds* $\|\sum_{t=0}^{T-1} z_t z_t'\| \leq M$ *with probability at least* $1 - \frac{\delta}{2}$. *Then, if* $T \geq 8c_1^2 c_\eta^2 (\log \frac{2}{\delta} + d \log \frac{144M}{\zeta^2(T-1)}) + 1$, *we have with probability at least* $1 - \delta$,

$$\sum_{t=0}^{T-1} z_t z_t' \succeq \frac{\zeta^2(T-1)}{32} I_d.$$

Proof. Note that for any fixed $v \in \mathcal{S}^{d-1}$ and $t \geq 0$, we have

$$v' z_{t+1} | \mathcal{F}_t = v' l_t | \mathcal{F}_t + v' \eta_{t+1} | \mathcal{F}_t.$$

From Proposition 2.6.3, we have

$$P(|v' z_{t+1}| \geq \sqrt{\frac{\sigma_\eta^2}{2}} | \mathcal{F}_t) \geq \frac{1}{c_1 \times c_\eta}$$

almost surely. Since the scalar process $\{v' z_t\}_{t \geq 1}$ is adapted to the filtration $\{\mathcal{F}_t\}_{t \geq 1}$, the above inequality implies that $v' z_t$ satisfies the $(1, \frac{\sqrt{2}\sigma_\eta}{2}, \frac{1}{c_1 \times c_\eta})$ BMSB condition (see Definition 2.6.2). Denoting $m = \frac{\zeta^2(T-1)}{16}$, we can now apply Lemma 11 to obtain

$$\begin{aligned} P(\sum_{t=1}^{T-1} (v' z_t)^2 \leq m) &\leq \exp(\frac{-(T-1)}{8c_1^2 c_\eta^2}) \\ &\leq \exp(\log \frac{\delta}{2} + \log(\frac{\zeta^2(T-1)}{144M})^d) \\ &= \frac{\delta}{2} (\frac{m}{9M})^d, \end{aligned} \tag{2.22}$$

where the last inequality is due to our assumption that $T \geq 8c_1^2 c_\eta^2 (\log \frac{2}{\delta} + d \log \frac{144M}{\zeta^2(T-1)}) + 1$.

Since $M \geq m$, we can let $\mathbf{N}(\frac{m}{4M})$ be a $\frac{m}{4M}$- net of $\mathcal{S}^{d-1}$ with the smallest cardinality (see Definition 2.6.1). From Lemma 6, we know that there are at most $(\frac{9M}{m})^d$ elements in $\mathbf{N}(\frac{m}{4M})$. Applying a union bound to combine the events in (2.22) for all $v \in \mathbf{N}(\frac{m}{4M})$, we have with probability at least $1 - \frac{\delta}{2}$

$$\min_{v \in \mathbf{N}(\frac{m}{4M})} \sum_{t=1}^{T} (v' z_t)^2 \geq m. \tag{2.23}$$

Note that for any realization of the matrix $\sum_{t=0}^{T-1} z_t z_t'$, we can fix a $v^* \in \mathcal{S}^{n-1}$ such that $\lambda_{min}(\sum_{t=0}^{T-1} z_t z_t') = \sum_{t=0}^{T-1} v^{*'} z_t z_t' v^*$, and let $v_0 \in \mathbf{N}(\frac{m}{4M})$ be a vector such that $\|v^* - v_0\| \leq \frac{m}{4M}$, to obtain

$$\begin{aligned}
\lambda_{min}(\sum_{t=0}^{T-1} z_t z_t') &= \sum_{t=0}^{T-1} (v_0 + v^* - v_0)' z_t z_t' (v_0 + v^* - v_0) \\
&= \sum_{t=0}^{T-1} v_0' z_t z_t' v_0 + \sum_{t=0}^{T-1} v_0' z_t z_t' (v^* - v_0) \\
&\quad + \sum_{t=0}^{T-1} (v^* - v_0)' z_t z_t' v_0 + \sum_{t=0}^{T-1} (v^* - v_0)' z_t z_t' (v^* - v_0) \\
&\geq \sum_{t=0}^{T-1} (v_0' z_t)^2 - \frac{m}{2M} \| \sum_{t=0}^{T-1} z_t z_t \| \\
&\geq \sum_{t=1}^{T-1} (v_0' z_t)^2 - \frac{m}{2M} \| \sum_{t=0}^{T-1} z_t z_t \| \\
&\geq \min_{v \in \mathbf{N}(\frac{m}{4M})} \sum_{t=1}^{T-1} (v' z_t)^2 - \frac{m}{2M} \| \sum_{t=0}^{T-1} z_t z_t \|.
\end{aligned}$$

Applying a union bound to combine the event $\| \sum_{t=0}^{T-1} z_t z_t' \| \leq M$ and the event in (2.23), we have with probability at least $1 - \delta$,

$$\lambda_{min}(\sum_{t=0}^{T-1} z_t z_t') \geq \frac{m}{2} = \frac{\zeta^2 (T-1)}{32}.$$

□

We have the following result.

Proposition 2.6.1. *Let Assumption 1 hold. Denote* $\bar{\zeta} = \frac{\bar{\sigma}_{min}}{c_1 \bar{\sigma}_*}$ *and* $\hat{\zeta} = \frac{\hat{\sigma}_{min}}{c_1 \hat{\sigma}_*}$, *where* c_1 *is a positive absolute constant. Fixing* $\delta \in (0,1)$, *suppose that* $N_r T_r \geq 8 c_1^2 \bar{\sigma}_*^2 (\log \frac{2}{\delta} + (n+p) \log \frac{144 \bar{g}(\frac{\delta}{2})}{\bar{\zeta}^2 (N_r T_r - 1)}) + 1$, $N_p T_p \geq 8 c_1^2 \hat{\sigma}_*^2 (\log \frac{2}{\delta} + (n+p) \log \frac{144 \hat{g}(\frac{\delta}{2})}{\hat{\zeta}^2 (N_p T_p - 1)}) + 1$, $\bar{g}(\frac{\delta}{2}) \geq \frac{\bar{\zeta}^2 (N_r T_r - 1)}{16}$, *and* $\hat{g}(\frac{\delta}{2}) \geq \frac{\hat{\zeta}^2 (N_p T_p - 1)}{16}$, *where* $\bar{g}(\frac{\delta}{2}), \hat{g}(\frac{\delta}{2})$ *are defined in Lemma 2. Then, we have with probability at least* $(1-\delta)^2$

$$ZQZ' \succeq \frac{(N_r T_r - 1)\bar{\zeta}^2 + q(N_p T_p - 1)\hat{\zeta}^2}{32} I_{n+p}.$$

Proof. We have

$$ZQZ' = \bar{Z}\bar{Z}' + q\hat{Z}\hat{Z}' = \sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'} + q\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1} \hat{z}_k^j \hat{z}_k^{j'}. \tag{2.24}$$

We now focus on the first summation in (2.24) since the analysis for the second one is essentially the same. We can define the sequence $\{z_t\}_{t\geq 0}$ as

$$z_t = \begin{cases} \bar{z}_t^1 & \text{if} \quad 0 \leq t \leq T_r - 1 \\ \bar{z}_{t-T_r}^2 & \text{if} \quad T_r \leq t \leq 2T_r - 1 \\ \bar{z}_{t-2T_r}^3 & \text{if} \quad 2T_r \leq t \leq 3T_r - 1 \\ \vdots & \vdots, \end{cases}$$

where $\bar{z}_b^a$ for $a > N_r$ and $b = 0, \ldots, T_r - 1$ are generated using the same way as $\bar{z}_b^a$ for $a = 1, \ldots, N_r$ and $b = 0, \ldots, T_r - 1$. In words, z_t is the sequence formed by concatenating the sequence $\{\bar{z}_t^1\}_{t=0}^{T_r-1}, \{\bar{z}_t^2\}_{t=0}^{T_r-1}, \ldots$. The sequences $\{w_t\}_{t\geq 0}$ and $\{u_t\}_{t\geq 0}$ are defined similarly using the signals $\bar{w}_t^i$ and $\bar{u}_t^i$. Further, we define the sequence $\{x_t\}_{t\geq 0}$ as $x_t = \bar{x}_0^{t+1}$, where $\bar{x}_0^a$ for $a > N_r$ are generated using the same way as $\bar{x}_0^a$ for $a = 1, \ldots, N_r$. With these definitions, we have

$$\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'} = \sum_{t=0}^{N_r T_r - 1} z_t z_t'.$$

We now verify the conditions in Lemma 3. Note that for t satisfying $t \geq 0$ and $(t+1) \bmod T_r \neq 0$, we have

$$z_{t+1} = \begin{bmatrix} \Theta z_t \\ 0 \end{bmatrix} + \begin{bmatrix} w_t \\ u_{t+1} \end{bmatrix}.$$

For t satisfying $t \geq 0$ and $(t+1) \bmod T_r = 0$, we have

$$z_{t+1} = \begin{bmatrix} x_{\frac{t+1}{T_r}} \\ u_{t+1} \end{bmatrix}.$$

Let $l_t = \begin{bmatrix} \Theta z_t \\ 0 \end{bmatrix}$ and $\eta_{t+1} = \begin{bmatrix} w_t \\ u_{t+1} \end{bmatrix}$ for t satisfying $t \geq 0$ and $(t+1) \bmod T_r \neq 0$. Similarly, let $l_t = \mathbf{0}$ and $\eta_{t+1} = \begin{bmatrix} x_{\frac{t+1}{T_r}} \\ u_{t+1} \end{bmatrix}$ for t satisfying $t \geq 0$ and $(t+1) \bmod T_r = 0$. Define the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t = \sigma(\{l_i\}_{i=0}^t \cup \{\eta_j\}_{j=1}^t)$. We have l_t is $\mathcal{F}_t$-measurable for $t \geq 0$, η_t is $\mathcal{F}_t$-measurable for $t \geq 1$, $\mathbb{E}[\eta_{t+1}\eta_{t+1}'|\mathcal{F}_t] \succeq \bar{\sigma}_{min}^2 I_{n+p}$ and $\max_{1\leq i\leq (n+p)} \frac{\mathbb{E}[\eta_{t+1}(i)^4|\mathcal{F}_t]}{\mathbb{E}[\eta_{t+1}(i)^2|\mathcal{F}_t]^2} \leq \bar{\sigma}_*$ for $t \geq 0$ due to our assumption.

Fixing $\delta \in (0,1)$, from Lemma 2, we have $\| \sum_{t=0}^{N_r T_r - 1} z_t z_t' \| \leq \bar{g}(\frac{\delta}{2})$ with probability at least $1 - \frac{\delta}{2}$. Consequently, letting $N_r T_r \geq 8c_1^2 \bar{\sigma}_*^2 (\log \frac{2}{\delta} + (n+p) \log \frac{144 \bar{g}(\frac{\delta}{2})}{\bar{\zeta}^2 (N_r T_r - 1)}) + 1$, we can apply Lemma 3 to get

$$\sum_{i=1}^{N_r} \sum_{t=0}^{T_r - 1} \bar{z}_t^i \bar{z}_t^{i'} = \sum_{t=0}^{N_r T_r - 1} z_t z_t' \succeq \frac{(N_r T_r - 1)\bar{\zeta}^2}{32} I_{n+p}$$

with probability at least $1 - \delta$. Applying a similar procedure for the second summation in (2.24) and leveraging the independence of data, we have with probability at least $(1-\delta)^2$

$$\sum_{i=1}^{N_r} \sum_{t=0}^{T_r - 1} \bar{z}_t^i \bar{z}_t^{i'} + q \sum_{j=1}^{N_p} \sum_{k=0}^{T_p - 1} \hat{z}_k^j \hat{z}_k^{j'} \succeq \frac{(N_r T_r - 1)\bar{\zeta}^2 + q(N_p T_p - 1)\hat{\zeta}^2}{32} I_{n+p}.$$

□

We will use the following lemma, which provides an upper bound for self-normalized martingales.

Lemma 4. *([37, Theorem 1]) Let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration. Let $\{w_t\}_{t\geq 1}$ be a real valued stochastic process such that w_t is $\mathcal{F}_t$-measurable, and w_t is conditionally sub-Gaussian on $\mathcal{F}_{t-1}$ with parameter R^2. Let $\{z_t\}_{t\geq 1}$ be an $\mathbb{R}^m$-valued stochastic process such that z_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $m \times m$ dimensional positive definite matrix. For all $t \geq 0$, define*

$$\bar{V}_t = V + \sum_{s=1}^t z_s z_s', S_t = \sum_{s=1}^t w_s z_s.$$

Then, for any $\delta > 0$, and for all $t \geq 0$,

$$P(\|\bar{V}_t^{-\frac{1}{2}} S_t\|^2 \leq 2R^2 \log(\frac{\det(\bar{V}_t^{\frac{1}{2}}) \det(V^{-\frac{1}{2}})}{\delta})) \geq 1 - \delta.$$

The following lemma generalizes the above result to the case where w_t is multi-dimensional, and will be used to bound the error term $\|WQZ'(ZQZ')^{-\frac{1}{2}}\|$. The proof is similar to [36, Proposition V.4].

Lemma 5. *Let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration. Let $\{w_t\}_{t\geq 1}$ be a $\mathbb{R}^n$-valued stochastic process such that w_t is $\mathcal{F}_t$-measurable, and w_t is conditionally sub-Gaussian on $\mathcal{F}_{t-1}$ with parameter R^2. Let $\{z_t\}_{t\geq 1}$ be a $\mathbb{R}^m$-valued stochastic process such that z_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $m \times m$ dimensional positive definite matrix. For all $t \geq 0$, define*

$$\bar{V}_t = V + \sum_{s=1}^{t} z_s z_s', S_t = \sum_{s=1}^{t} z_s w_s'.$$

Then, for any $\delta > 0$, and for all $t \geq 0$,

$$P(\|\bar{V}_t^{-\frac{1}{2}} S_t\| \leq \sqrt{\frac{32}{9} R^2 (\log \frac{9^n}{\delta} + \frac{1}{2} \log \det(\bar{V}_t V^{-1}))} \geq 1 - \delta.$$

Proof. We have

$$\|\bar{V}_t^{-\frac{1}{2}} S_t\| = \|\bar{V}_t^{-\frac{1}{2}} \sum_{s=1}^{t} z_s w_s'\| = \sup_{v \in \mathcal{S}^{n-1}} \|\bar{V}_t^{-\frac{1}{2}} \sum_{s=1}^{t} z_s w_s' v\|.$$

Note that for any fixed unit vector $v \in \mathcal{S}^{n-1}$, the random variable $w_s' v$ is conditionally sub-Gaussian with parameter R^2. Let $\mathbf{N}(\frac{1}{4})$ be a $\frac{1}{4}$- net of $\mathcal{S}^{n-1}$ with the smallest cardinality (see Definition 2.6.1). From Lemma 6, we know that there are at most 9^n elements in $\mathbf{N}(\frac{1}{4})$. For any fixed $\delta \in (0, 1)$ and $v \in \mathbf{N}(\frac{1}{4})$, we can apply Lemma 4 to obtain with probability at least $1 - \frac{\delta}{9^n}$

$$\begin{aligned} \|\bar{V}_t^{-\frac{1}{2}} \sum_{s=1}^{t} z_s w_s' v\|^2 &\leq 2R^2 \log \frac{9^n \det(\bar{V}_t^{\frac{1}{2}}) \det(V^{-\frac{1}{2}})}{\delta} \\ &= 2R^2 (\log \frac{9^n}{\delta} + \frac{1}{2} \log \det(\bar{V}_t V^{-1})). \end{aligned}$$

Applying a union bound over all 9^n events, from Lemma 7, we have with probability at least $1-\delta$

$$\begin{aligned}\|\bar{V}_t^{-\frac{1}{2}}\sum_{s=1}^{t} z_s w_s'\| &\le \frac{4}{3}\sup_{v\in\mathbf{N}(\frac{1}{4})}\|\bar{V}_t^{-\frac{1}{2}}\sum_{s=1}^{t} z_s w_s' v\| \\ &\le \sqrt{\frac{32}{9}R^2(\log\frac{9^n}{\delta}+\frac{1}{2}\log\det(\bar{V}_t V^{-1})}.\end{aligned}$$

□

2.6.2 Proofs of Theorem 2.3.1

Proof. Recall that the system identification error in (2.7) (using $\lambda = 0$) satisfies

$$\begin{aligned}\|\Theta_{WLS}-\Theta\| \le \|(ZQZ')^{-\frac{1}{2}}ZQW'\|\|(ZQZ')^{-\frac{1}{2}}\| \\ + \|\Delta QZ'\|\|(ZQZ')^{-1}\|,\end{aligned} \tag{2.25}$$

under the invertibility assumption. Let N_rT_r, N_pT_p, δ satisfy the conditions in Theorem 2.3.1, and let $V = \frac{N_rT_r\bar{\zeta}^2+qN_pT_p\hat{\zeta}^2}{33}I_{n+p}$. From Proposition 2.6.1, we have with probability at least $1-2\delta$

$$ZQZ' \succeq V, \tag{2.26}$$

conditioning on which we have

$$\|(ZQZ')^{-\frac{1}{2}}\| \le \|V^{-\frac{1}{2}}\|, \tag{2.27}$$

where we used the relationship $\frac{(N_rT_r-1)\bar{\zeta}^2+q(N_pT_p-1)\hat{\zeta}^2}{32} \ge \frac{N_rT_r\bar{\zeta}^2+qN_pT_p\hat{\zeta}^2}{33}$ when $\min\{N_rT_r, N_pT_p\} \ge 33$ in (2.26), and Lemma 8 in conjunction with Lemma 9 for (2.27).

Further, conditioning on (2.26), we also have $ZQZ' \succeq V \Rightarrow 2ZQZ' \succeq ZQZ'+V \Rightarrow (ZQZ')^{-1} \preceq 2(ZQZ'+V)^{-1}$, where we used Lemma 8. Applying Lemma 10, we have

$$\|(ZQZ')^{-\frac{1}{2}}ZQW'\| \le \sqrt{2}\|(ZQZ'+V)^{-\frac{1}{2}}ZQW'\|. \tag{2.28}$$

Next, to use Lemma 2, we define a new pair of sequences $\{z_t\}_{t\geq 1}$ and $\{w_t\}_{t\geq 1}$ using the signals used in the terms $ZQZ' = \sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1}\bar{z}_t^i\bar{z}_t^{i'} + q\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1}\hat{z}_k^j\hat{z}_k^{j'}$ and $ZQW' = \sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1}\bar{z}_t^i\bar{w}_t^{i'} + q\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1}\hat{z}_k^j\hat{w}_k^{j'}$. That is,

$$z_t = \begin{cases} \bar{z}_{t-1}^1 & \text{if} \quad 1 \leq t \leq T_r \\ \bar{z}_{t-T_r-1}^2 & \text{if} \quad T_r+1 \leq t \leq 2T_r \\ \bar{z}_{t-2T_r-1}^3 & \text{if} \quad 2T_r+1 \leq t \leq 3T_r \\ \vdots & \vdots \\ \bar{z}_{t-(N_r-1)T_r-1}^{N_r} & \text{if} \quad (N_r-1)T_r+1 \leq t \leq N_rT_r \\ \sqrt{q}\hat{z}_{t-N_rT_r-1}^1 & \text{if} \quad N_rT_r+1 \leq t \leq N_rT_r+T_p \\ \sqrt{q}\hat{z}_{t-N_rT_r-T_p-1}^2 & \text{if} \quad N_rT_r+T_p+1 \leq t \leq N_rT_r+2T_p \\ \sqrt{q}\hat{z}_{t-N_rT_r-2T_p-1}^3 & \text{if} \quad N_rT_r+2T_p+1 \leq t \leq N_rT_r+3T_p \\ \vdots & \vdots, \end{cases}$$

and

$$w_t = \begin{cases} \bar{w}_{t-1}^1 & \text{if} \quad 1 \leq t \leq T_r \\ \bar{w}_{t-T_r-1}^2 & \text{if} \quad T_r+1 \leq t \leq 2T_r \\ \bar{w}_{t-2T_r-1}^3 & \text{if} \quad 2T_r+1 \leq t \leq 3T_r \\ \vdots & \vdots \\ \bar{w}_{t-(N_r-1)T_r-1}^{N_r} & \text{if} \quad (N_r-1)T_r+1 \leq t \leq N_rT_r \\ \sqrt{q}\hat{w}_{t-N_rT_r-1}^1 & \text{if} \quad N_rT_r+1 \leq t \leq N_rT_r+T_p \\ \sqrt{q}\hat{w}_{t-N_rT_r-T_p-1}^2 & \text{if} \quad N_rT_r+T_p+1 \leq t \leq N_rT_r+2T_p \\ \sqrt{q}\hat{w}_{t-N_rT_r-2T_p-1}^3 & \text{if} \quad N_rT_r+2T_p+1 \leq t \leq N_rT_r+3T_p \\ \vdots & \vdots, \end{cases}$$

where $\sqrt{q}\hat{z}_b^a, \sqrt{q}\hat{w}_b^a$ for $a > N_p$ and $b = 0, \ldots, T_p-1$ are generated using the same way as $\sqrt{q}\hat{z}_b^a, \sqrt{q}\hat{w}_b^a$ for $a = 1, \ldots, N_p$ and $b = 0, \ldots, T_p-1$.

Consequently, we have

$$ZQZ' = \sum_{t=1}^{N_rT_r+N_pT_p} z_t z_t',$$

and

$$ZQW' = \sum_{t=1}^{N_rT_r+N_pT_p} z_t w_t'.$$

Now define the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t = \sigma(\{z_{i+1}\}_{i=0}^t \cup \{w_j\}_{j=1}^t)$. With these definitions, we can see that the noise terms w_t are $\mathcal{F}_t$-measurable, and $w_t|\mathcal{F}_{t-1}$ are sub-Gaussian with parameter $\max(\sigma_{\bar{w}}^2, q\sigma_{\hat{w}}^2)$ for all $t \geq 1$. Consequently, we can apply Lemma 5 to obtain with probability at least $1-\delta$

$$\sqrt{2}\|(ZQZ'+V)^{-\frac{1}{2}}ZQW'\| \leq 3\max(\sigma_{\bar{w}}, \sqrt{q}\sigma_{\hat{w}})\sqrt{\log\frac{9^n}{\delta} + \log\det((ZQZ'+V)V^{-1})}. \tag{2.29}$$

Further, from Lemma 2, we have with probability at least $1-2\delta$

$$\begin{aligned}\det((ZQZ'+V)V^{-1}) &\leq \frac{\|ZQZ'+V\|^{n+p}}{\det(V)} \\ &\leq \frac{(\|\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1}\bar{z}_t^i\bar{z}_t^{i'}\| + q\|\sum_{i=1}^{N_p}\sum_{t=0}^{T_p-1}\hat{z}_t^i\hat{z}_t^{i'}\| + \|V\|)^{n+p}}{\det(V)} \\ &\leq (\frac{33(\bar{g}(\delta)+q\hat{g}(\delta))}{N_rT_r\bar{\zeta}^2 + qN_pT_p\hat{\zeta}^2} + 1)^{n+p} = \phi^{n+p}.\end{aligned} \tag{2.30}$$

Applying a union bound over the events in (2.28), (2.29), and (2.30), we have with probability at least $1-5\delta$

$$\|(ZQZ')^{-\frac{1}{2}}ZQW'\| \leq 3\max(\sigma_{\bar{w}}, \sqrt{q}\sigma_{\hat{w}})\sqrt{\log\frac{9^n}{\delta} + (n+p)\log(\phi)}. \tag{2.31}$$

Next, conditioning on the event in Lemma 2, notice that we also have

$$\begin{aligned}\|\Delta QZ'\| &= \|\sum_{i=1}^{N_p}\sum_{t=0}^{T_p-1} q\delta_\Theta\hat{z}_t^i\hat{z}_t^{i'}\| \leq q\|\delta_\Theta\|\|\sum_{i=1}^{N_p}\sum_{t=0}^{T_p-1}\hat{z}_t\hat{z}_t'\| \\ &\leq q\|\delta_\Theta\|\hat{g}(\delta).\end{aligned} \tag{2.32}$$

Finally, combining (2.27), (2.31) and (2.32), we have the desired result. □

2.6.3 Proof of Corollary 1

Proof. Setting $q = 0$, from Theorem 2.3.1, we have with probability at least $1 - \delta$

$$\|\Theta_{WLS} - \Theta\| \leq \frac{20\sigma_{\bar{w}}\sqrt{\log\frac{9^n}{\delta} + (n+p)\log(\frac{33\bar{g}(\delta)}{N_r T_r \bar{\zeta}^2} + 1)}}{\sqrt{N_r T_r \bar{\zeta}^2}}. \tag{2.33}$$

When $q \neq 0$, from Theorem 2.3.1, after some algebraic manipulations, we can show that with probability at least $1 - \delta$

$$\|\Theta_{WLS} - \Theta\| \leq \frac{20\max(\frac{\sigma_{\bar{w}}}{\sqrt{q}}, \sigma_{\hat{w}})\sqrt{\log\frac{9^n}{\delta} + (n+p)\log(\frac{\gamma}{\hat{\zeta}^2} + 1)}}{\sqrt{N_p T_p \hat{\zeta}^2}} + \|\delta_\Theta\|\frac{\gamma}{\hat{\zeta}^2}. \tag{2.34}$$

The proof follows by setting the upper bound in (2.33) to be greater than the one in (2.34). □

2.6.4 Proof of Theorem 2.3.2

Proof. The proof follows the same procedure as the proof of Theorem 2.3.1. However, instead of bounding the term $\|\Delta QZ'(ZQZ')^{-1}\|$ by $\|\Delta QZ'\|\|(ZQZ')^{-1}\|$, we use the following bound

$$\begin{aligned}
&\|\Delta QZ'(ZQZ')^{-1}\| = \\
&\|(\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1} q\delta_\Theta \hat{z}_k^j \hat{z}_k^{j'})(\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'} + q\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1} \hat{z}_k^j \hat{z}_k^{j'})^{-1}\| \\
&\leq \|\delta_\Theta\|\|I_{n+p} - (\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'})(\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'} + q\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1} \hat{z}_k^j \hat{z}_k^{j'})^{-1})\| \\
&\leq \|\delta_\Theta\|(1 + \|\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'}\|\|(\sum_{i=1}^{N_r}\sum_{t=0}^{T_r-1} \bar{z}_t^i \bar{z}_t^{i'} + q\sum_{j=1}^{N_p}\sum_{k=0}^{T_p-1} \hat{z}_k^j \hat{z}_k^{j'})^{-1})\|),
\end{aligned} \tag{2.35}$$

where we used the relationship $AB = (C + A)B - CB$ for real matrices A, B, C in the first inequality. When the events in Proposition 2.6.1 and Lemma 2 happen, we can upper bound the right hand side of the last inequality in (2.35) by $\|\delta_\Theta\|(1 + \frac{33\bar{g}(\delta)}{N_r T_r \bar{\zeta}^2 + q N_p T_p \hat{\zeta}^2})$.

□

2.6.5 Proof of Theorem 2.3.3

Proof. From (2.7), note that the system identification error satisfies

$$\begin{aligned}\|\Theta_{WLS} - \Theta\| &\leq \lambda\|\Theta\|\|(ZQZ' + \lambda I_{n+p})^{-1}\| \\ &+ \|(ZQZ' + \lambda I_{n+p})^{-\frac{1}{2}} ZQW'\|\|(ZQZ' + \lambda I_{n+p})^{-\frac{1}{2}}\| \\ &+ \|\Delta QZ'(ZQZ' + \lambda I_{n+p})^{-1}\| \\ &\leq \frac{\lambda\|\Theta\|}{\lambda_{min}(ZQZ' + \lambda I_{n+p})} \\ &+ \frac{\|(ZQZ' + \lambda I_{n+p})^{-\frac{1}{2}} ZQW'\|}{\sqrt{\lambda_{min}(ZQZ' + \lambda I_{n+p})}} \\ &+ q\|\delta_\Theta\|\|\hat{Z}\hat{Z}'(ZQZ' + \lambda I_{n+p})^{-1}\|.\end{aligned} \tag{2.36}$$

Note that all terms in the above inequality can be evaluated from data, except for the term $\|(ZQZ' + \lambda I_{n+p})^{-\frac{1}{2}} ZQW'\|$. We can follow a similar procedure to apply Lemma 5 as in the proof of Theorem 2.3.1. Fixing $\delta > 0$, from Lemma 5, we have with probability at least $1 - \delta$

$$\|(ZQZ' + \lambda I_{n+p})^{-\frac{1}{2}} ZQW'\| \leq \max(\sigma_{\bar{w}}, \sqrt{q}\sigma_{\hat{w}})\sqrt{\frac{32}{9}(\log\frac{9^n}{\delta} + \frac{1}{2}\log\det(\bar{V}))}.$$

The result then follows. □

2.6.6 Auxiliary Results

Lemma 6. *([38, Corollary 4.2.13]) Let $\epsilon > 0$, and let $\mathbf{N}(\mathcal{S}^{n-1}, \epsilon)$ be the smallest possible cardinality of an ϵ-net of the unit Euclidean sphere $\mathcal{S}^{n-1}$. We have the following inequality:*

$$\mathbf{N}(\mathcal{S}^{n-1}, \epsilon) \leq (\frac{2}{\epsilon} + 1)^n.$$

Lemma 7. *([38, Lemma 4.4.1]) Let A be an m by n matrix and $\epsilon \in [0,1)$. Then, for any ϵ-net $\mathbf{N}$ of the sphere $\mathcal{S}^{n-1}$, we have*

$$\|A\| \leq \frac{1}{1-\epsilon} \cdot \sup_{x \in \mathbf{N}} \|Ax\|.$$

Lemma 8. *([39, Lemma 3]) Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be positive definite matrices. If $A \preceq B$, then we have $A^{-1} \succeq B^{-1}$.*

Lemma 9. *([39, Theorem 2]) Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be positive semidefinite matrices. If $A \preceq B$, then we have $A^{\frac{1}{2}} \preceq B^{\frac{1}{2}}$.*

Lemma 10. *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be positive semidefinite matrices. Let $C \in \mathbb{R}^{n\times m}$. If $A \preceq B$, then we have*

$$\|A^{\frac{1}{2}}C\| \leq \|B^{\frac{1}{2}}C\|.$$

Proof. From $A \preceq B$, we have

$$C'AC \preceq C'BC,$$

which implies

$$\|A^{\frac{1}{2}}C\| = \sqrt{\lambda_{max}(C'AC)} \leq \sqrt{\lambda_{max}(C'BC)} = \|B^{\frac{1}{2}}C\|.$$

□

Proposition 2.6.2. *Assuming that $\rho(\bar{A}) < 1$ and $\rho(\hat{A}) < 1$, we have both $\mathrm{tr}(\bar{G}_t)$ and $\mathrm{tr}(\hat{G}_t)$ are $\mathcal{O}(1)$. If $\rho(\bar{A}) = 1$ and $\rho(\hat{A}) = 1$, we have $\mathrm{tr}(\bar{G}_t) = \mathcal{O}(t^{2\bar{\kappa}})$ and $\mathrm{tr}(\hat{G}_t) = \mathcal{O}(t^{2\hat{\kappa}})$, where $\bar{\kappa}$, $\hat{\kappa}$ are the largest Jordan blocks corresponding to the unit eigenvalues of $\bar{A}$ and $\hat{A}$, respectively.*

Proof. We only consider the term $\mathrm{tr}(\bar{G}_t)$ as the term $\mathrm{tr}(\hat{G}_t)$ is essentially the same. Defining $\bar{F}_t = \begin{bmatrix} I_n & \bar{A} & \cdots & \bar{A}^t & \bar{B} & \bar{A}\bar{B} & \bar{A}^2\bar{B} & \cdots & \bar{A}^{t-1}\bar{B} \end{bmatrix} \in \mathbb{R}^{n\times(tn+tp+n)}$ for $t \geq 0$, we have $\bar{G}_t = \bar{F}_t\bar{F}_t'$. Further, we have

$$\mathrm{tr}(\bar{G}_t) \leq n\|\bar{G}_t\| \leq n\|\bar{F}_t\|^2 \leq n(\sum_{i=0}^{t}\|\bar{A}^i\| + \sum_{i=0}^{t-1}\|\bar{A}^i\|\|\bar{B}\|)^2.$$

From [40, Lemma E.2.] (i.e., Lemma 16), we have $\sum_{i=0}^{t}\|\bar{A}^i\|$ is $\mathcal{O}(1)$ for strictly stable systems, and $\mathcal{O}(t^{\hat{\kappa}})$ for marginally stable systems. The result then follows. □

Lemma 11. *[7, Proposition 2.5]) Suppose that $\{Z_t\}_{t\geq 1}$ satisfies the (k, v, p)-BMSB condition. Denoting $\lfloor\cdot\rfloor$ as the floor function, we have*

$$P(\sum_{i=1}^{T} Z_i^2 \leq \frac{v^2p^2}{8}k\lfloor T/k\rfloor) \leq \exp(-\frac{\lfloor T/k\rfloor p^2}{8}).$$

Proposition 2.6.3. *[35, Proposition C.2.] Let $\mu \in \mathbb{R}$ and $M \in \mathbb{R}^{d\times d}$ be a full rank matrix. Let $w \in \mathbb{R}^d$ be a random vector such that each coordinate $w(i)$ has positive variance and finite fourth moment. Further, each coordinate $w(i)$ is independent and zero-mean. Then, for any fixed $v \in \mathcal{S}^{d-1}$,*

$$P_w(|\langle v,\ \mu + Mw\rangle| \geq \sqrt{\lambda_{min}(M\Sigma M')/2}) \geq \frac{1}{c_1 C_w},$$

where $\Sigma = \mathbb{E}_w[ww']$, c_1 is an absolute constant, and $C_w = \max_{1\leq i\leq d} \frac{\mathbb{E}[w(i)^4]}{\mathbb{E}[w(i)^2]^2}$.

Remark 7. *The constant c_1 is due to the application of the Rosenthal's inequality. As suggested in [35, Proposition C.2.], one can take $c_1 = 192$.*

3. LEARNING THE DYNAMICS OF AUTONOMOUS LINEAR SYSTEMS FROM MULTIPLE TRAJECTORIES

3.1 Introduction

[6] As mentioned in Chapter 2, non-asymptotic analysis of system identification based on a single trajectory has been studied extensively over the past few years [7], [9], [11], [12], [42], [43]. However, in practice, performing system identification using a single trajectory could be problematic for many applications. For example, having the system run for a long time could incur risks when the system is unstable. Furthermore, only historical snippets of data about the system may be available, without the ability to easily observe long-run behavior. Additionally, in settings where one has the ability to restart the system or have several copies of the system running in parallel, one may obtain *multiple* trajectories generated by the system dynamics [44]. The paper [13] studies the sample complexity of identifying a system whose state is fully measured using only the final data points from multiple trajectories. Using a similar setup, the paper [14] explores the benefits of adding an ℓ_1 regularizer. The paper [15] studies the sample complexity of partially-measured system identification by including nuclear norm regularization, again only using the final samples from each trajectory. For partially-measured systems, the paper [16] allows for more efficient use of data. As mentioned in [16], compared to the single trajectory setup, the multiple trajectories setup usually allows for more direct application of concentration inequalities due to the assumption of independence over multiple trajectories.

In addition to the potential lack of single long trajectories, in many settings we may not be able to actually apply inputs to the system in order to perform system identification; this could be due to the costs of applying inputs, or due to the fact that we are simply observing an autonomous system that we cannot control. The uncontrolled system may also be serving as a subsystem connected to the main system that one wants to control, and having a better model of the subsystem could be useful in controlling the main system. For partially-measured systems, the characterization of finite sample error of purely stochastic systems (systems that are entirely driven by unmeasurable noise) is more challenging as indicated in

[6]↑The material in this chapter was published at the 2022 American Control Conference [41]

[40]. There, the goal is to estimate the system matrices as well as the steady state Kalman filter gain of the corresponding system. The paper [40] shows that classical stochastic system identification algorithm can achieve a learning rate of $\mathcal{O}(\frac{1}{\sqrt{\bar{N}}})$ (up to logarithmic factors) for both strictly stable and marginally stable systems, where $\bar{N}$ denotes the number of samples in a single trajectory.

In this chapter, we are motivated by the challenge of system identification for partially-measured and autonomous stochastic linear systems (with no controlled inputs) as in [40], but for the case where a single long-run trajectory is unavailable. Existing results on consistency and learning rate of stochastic system identification algorithms (including [40]) typically convert the original system to a statistically equivalent form of the Kalman filter that is assumed to have reached steady state [2], [40], [45]. The analysis is then performed by assuming that the covariance matrix of the initial state of the system is the same as the steady state Kalman filter error covariance matrix, which simplifies the analysis. We note, however, that this assumption is invalid when one has no long run observation of the system trajectory, since it is in general unclear how long one should wait until the Kalman filter "converges" (even if it converges exponentially fast) for an unknown system. Furthermore, the available short trajectories may not be long enough to guarantee that the underlying filter has converged. Consequently, the single trajectory-based results cannot be directly applied to the multiple (short) trajectories case. Our goal in this chapter is to estimate the system matrices (up to similarity transformations) using only multiple trajectories of transient responses of a partially-measured system that is entirely driven by noise.

Our work is inspired by recent work on stochastic system identification (with a single long trajectory) [40], and system identification with multiple trajectories (but with controlled inputs) [16], and extends them in the following ways. First, we provide results on the sample complexity of learning the dynamics of an autonomous stochastic linear system using multiple trajectories, without assuming that the associated Kalman filter has reached steady state (i.e., the initial states can have arbitrary covariance matrix). Compared to [16] and [40], our results neither rely on controlled inputs, nor rely on observations of steady state behaviors of the system. Second, we provide the asymptotic learning rate of the multiple-trajectories-based stochastic system identification algorithm. If N is the number of trajectories, we prove

a learning rate of $\mathcal{O}(\frac{1}{\sqrt{N}})$ when the initial state in each trajectory has zero mean (which is a common assumption in the existing literature). This rate is consistent with [16] and [40] (up to logarithmic factors). We further generalize the result to the case when the initial state in each trajectory has non-zero mean. In such case, we show that we can adjust the length of the regressor to achieve a learning rate of $\mathcal{O}(\sqrt{\frac{\log N}{N}})$ for strictly stable systems and a learning rate of $\mathcal{O}(\frac{(\log N)^d}{\sqrt{N}})$ for marginally stable systems, where d is some constant.

3.2 Problem Formulation

Consider a discrete time linear time-invariant system with no user specified inputs:

$$x_{k+1} = Ax_k + w_k, \quad y_k = Cx_k + v_k, \tag{3.1}$$

where $x_k \in \mathbb{R}^n$, $y_k \in \mathbb{R}^m$, $w_k \in \mathbb{R}^n$, $v_k \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times n}$. The noise terms w_k and v_k are assumed to be i.i.d Gaussian, i.e., $w_k \sim \mathcal{N}(0, Q)$, $v_k \sim \mathcal{N}(0, R)$. The initial state is also assumed to be independent of w_k and v_k, and is distributed as $x_0 \sim \mathcal{N}(\mu, \Sigma_0)$. In addition, whether μ is zero or non-zero is assumed to be known. If μ is non-zero, the system matrix A is assumed to be strictly stable or marginally stable. The system order n is also assumed to be known. We refer to the above system as an autonomous stochastic linear system. We will make the following assumption.

Assumption 2. *The output covariance matrix R is positive definite. The pair (A, C) is observable and $(A, Q^{\frac{1}{2}})$ is controllable.*

Under the above assumption, the *Kalman Filter* associated with system (3.1) is a system of the form

$$\hat{x}_{k+1} = A\hat{x}_k + K_k e_k, \quad y_k = C\hat{x}_k + e_k, \tag{3.2}$$

where $\hat{x}_k$ is an estimate of state x_k, with the initial estimate being the mean of the initial state in system (3.1), i.e., $\hat{x}_0 = \mu$. The sequence of matrices $K_k \in \mathbb{R}^{n\times m}$ is called the Kalman gain, given by

$$K_k = AP_kC'(CP_kC' + R)^{-1}, \tag{3.3}$$

where the estimation error covariance $P_k \in \mathbb{R}^{n\times n}$ is updated based on the Riccati equation

$$P_{k+1} = AP_kA' + Q - AP_kC'(CP_kC' + R)^{-1}CP_kA',$$

with $P_0 = \Sigma_0$. Finally, $e_k = y_k - C\hat{x}_k$ are independent zero mean Gaussian innovations with covariance matrix given by

$$\bar{R}_k = CP_kC' + R. \tag{3.4}$$

Since the outputs of system (3.1) and system (3.2) have identical statistical properties [46], we will perform analysis on system (3.2). The subspace identification problem for stochastic systems that we tackle in this chapter is to identify the system matrices (A, C) up to a similarity transformation, given *multiple trajectories* of outputs of the system (3.1). As a byproduct, we will also simultaneously learn the Kalman filter gain K_k of the corresponding system, at some time step k. In particular, we are interested in the quality of the estimates of (A, C) given a finite number of samples.

3.3 Subspace Identification Technique

Here we describe a variant of classical subspace identification algorithm [46] to estimate (A, C) matrices (up to a similarity transformation). We will first establish some definitions.

Suppose that we have access to N independent output trajectories of system (3.1), each of some length $T \in \mathbb{N}$, and each obtained right after restarting the system from an initial state $x_0 \sim \mathcal{N}(\mu, \Sigma_0)$. We denote the data from these trajectories as $\{y_k^i : 1 \leq i \leq N, 0 \leq k \leq T-1\}$, where the superscript denotes the trajectory index and the subscript denotes the time index. Let $p + f = T$, where p, f are design parameters that satisfy $p, f > n$, where n is the order of the system. We split the output samples from each output trajectory i into

past and future outputs with respect to p, and denote the past output and future output vectors for trajectory i as:

$$\begin{aligned} Y_-^i &\triangleq \begin{bmatrix} y_0^{i'} & y_1^{i'} & \cdots & y_{p-1}^{i'} \end{bmatrix}', \\ Y_+^i &\triangleq \begin{bmatrix} y_p^{i'} & y_{p+1}^{i'} & \cdots & y_{p+f-1}^{i'} \end{bmatrix}', \end{aligned} \tag{3.5}$$

respectively. The batch past output and batch future output matrices are formed by stacking all N output trajectories:

$$Y_- \triangleq \begin{bmatrix} Y_-^1 & Y_-^2 & \cdots & Y_-^N \end{bmatrix}, \quad Y_+ \triangleq \begin{bmatrix} Y_+^1 & Y_+^2 & \cdots & Y_+^N \end{bmatrix}. \tag{3.6}$$

The past and future innovations E_-^i, E_+^i, E_-, E_+ are defined similarly, using the signals e_k^i rather than y_k^i.

Let the batch matrix of initial states be $\hat{X}_0 \triangleq \begin{bmatrix} \hat{x}_0^1 & \hat{x}_0^2 & \cdots & \hat{x}_0^N \end{bmatrix}$. Define the largest norm of innovation covariance matrices as $\bar{\mathcal{R}}_T \triangleq \max_{t\in 0,\ldots,T-1} \|\bar{R}_t\|$, where $\bar{R}_t$ is defined in (3.4). For any $l \geq 1$, the extended observability matrix $\mathcal{O}_l \in \mathbb{R}^{ml\times n}$ and the reversed extended controllability matrix $\mathcal{K}_p \in \mathbb{R}^{n\times mp}$ are defined as:

$$\begin{aligned} \mathcal{O}_l &\triangleq \begin{bmatrix} C' & (CA)' & \cdots & (CA^{l-1})' \end{bmatrix}', \\ \mathcal{K}_p &\triangleq \begin{bmatrix} ((A-K_{p-1}C)\cdots(A-K_1C)K_0)' \\ \vdots \\ ((A-K_{p-1}C)K_{p-2})' \\ K_{p-1}' \end{bmatrix}. \end{aligned}$$

Define

$$G \triangleq \mathcal{O}_f \mathcal{K}_p. \tag{3.7}$$

Let $K \in \mathbb{R}^{n\times m}$ be the steady state Kalman gain $K = APC'(CPC' + R)^{-1}$, where $P \in \mathbb{R}^{n\times n}$ is the solution to the Riccati equation, $P = APA' + Q - APC'(CPC' + R)^{-1}CPA'$. From Kalman filtering theory, the matrix $A - KC$ has spectral radius strictly less than 1

[47]. Denote the reversed extended controllability matrix formed by the steady state Kalman gain K as

$$\mathbf{K}_p \triangleq \begin{bmatrix} (A-KC)^{p-1}K & (A-KC)^{p-2}K & \cdots & K \end{bmatrix}.$$

We further make the following assumption.

Assumption 3. *We have* $\text{rank}(\mathcal{K}_p) = \text{rank}(\mathbf{K}_p) = n$.

The rank condition on $\mathbf{K}_p$ is standard, e.g., [4], [40]. The rank condition on $\mathcal{K}_p$ is needed to ensure that G has rank n, which can be satisfied in practice by choosing p to be relatively large if the rank condition on $\mathbf{K}_p$ is satisfied (see Proposition 3.6.4 in section 3.6).

Finally, for any integer $a \geq 0$ and $b \geq 2$, define the block-Toeplitz matrix $\mathcal{T}_b^a \in \mathbb{R}^{mb \times mb}$ as:

$$\mathcal{T}_b^a \triangleq \begin{bmatrix} I_m & 0 & \cdots & 0 \\ CK_a & I_m & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ CA^{b-2}K_a & CA^{b-3}K_{a+1} & \cdots & I_m \end{bmatrix}.$$

3.3.1 Linear Regression

The subspace identification technique first uses linear regression to estimate G from (3.7), which will subsequently form the basis for the recovery of the system parameters.

For any output trajectory $i \in \{1, \cdots, N\}$, by iterating (3.2), the future output matrix Y_+^i satisfies

$$Y_+^i = \mathcal{O}_f \hat{x}_p^i + \mathcal{T}_f^p E_+^i. \tag{3.8}$$

Note that at any time step k, the state $\hat{x}_k^i$ can be expressed from (3.2) as

$$\begin{aligned} \hat{x}_k^i &= A\hat{x}_{k-1}^i + K_{k-1}(y_{k-1}^i - C\hat{x}_{k-1}^i) \\ &= K_{k-1}y_{k-1}^i + (A - K_{k-1}C)\hat{x}_{k-1}^i. \end{aligned}$$

By expanding the above relationship recursively, we have

$$\begin{aligned}\hat{x}_p^i &= K_{p-1}y_{p-1}^i + \cdots + (A - K_{p-1}C)\cdots(A - K_1C)K_0y_0^i + (A - K_{p-1}C)\cdots(A - K_0C)\hat{x}_0^i \\ &= \mathcal{K}_pY_-^i + (A - K_{p-1}C)\cdots(A - K_0C)\hat{x}_0^i.\end{aligned}$$

By substituting the above equality into (3.8), the relationship between the batch output matrices is given by

$$Y_+ = GY_- + \mathcal{O}_f(A - K_{p-1}C)\cdots(A - K_0C)\hat{X}_0 + \mathcal{T}_f^pE_+.$$

An estimate of G (motivated by the least squares approach) is

$$\hat{G} = Y_+Y_-^{'}(Y_-Y_-^{'})^{-1}. \tag{3.9}$$

Consequently, the estimation error for matrix G can be expressed as

$$\hat{G} - G = \mathcal{T}_f^pE_+Y_-^{'}(Y_-Y_-^{'})^{-1} + \mathcal{O}_f(A - K_{p-1}C)\cdots(A - K_0C)\hat{X}_0Y_-^{'}(Y_-Y_-^{'})^{-1}, \tag{3.10}$$

where the second term can be dropped if $\|\hat{X}_0\| = 0$, i.e., the initial state of system (3.1) has zero mean. When $\|\hat{X}_0\|$ is known to be non-zero but the system is marginally stable ($\rho(A) \leq 1$), we can leverage the fact that the norm $\|(A - K_{p-1}C)\cdots(A - K_0C)\|$ converges to zero exponentially fast with p (see Proposition 3.6.2 in section 3.6) by setting $p = c\log N$ for some positive constant c, to make the second term go to zero asymptotically. The above steps are encapsulated in Algorithm 2.

Algorithm 2 Linear regression to calculate an estimate $\hat{G}$ of G

Input N output trajectories $\{y_k^i : 1 \leq i \leq N, 0 \leq k \leq T-1\}$, parameters p, f

1: For each output trajectory $i \in \{1, \cdots, N\}$, construct the past output and future output Y_-^i, Y_+^i as in (3.5).

2: Construct the batch past output and batch future output Y_-, Y_+ as in (3.6).

3: Return $\hat{G} = Y_+Y_-^{'}(Y_-Y_-^{'})^{-1}$.

Remark 8. *Note that the matrix G itself can be used to predict future outputs from past outputs. The value GY_-^i represents the Kalman prediction of the next f outputs using the past p measurements, assuming the initial state has zero mean. Its role is similar to the Markov parameters that map inputs to outputs in the case when one has measured inputs.*

3.3.2 Balanced Realization

The following balanced realization algorithm uses a standard Singular Value Decomposition to extract the estimated system matrices $(\hat{A}, \hat{C}, \hat{K}_{p-1})$ from the estimate $\hat{G}$.

Algorithm 3 Balanced realization to calculate estimates $(\hat{A}, \hat{C}, \hat{K}_{p-1})$ of (A, C, K_{p-1}) up to a similarity transformation

---

Input The estimate $\hat{G}$, parameters n, m, f

1: Perform the Singular Value Decomposition: $\hat{G} = \begin{bmatrix} \hat{U}_1 & \hat{U}_2 \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \hat{V}_1' \\ \hat{V}_2' \end{bmatrix}$, where $\Sigma_1 \in \mathbb{R}^{n \times n}$ contains the n-largest singular values of $\hat{G}$.

2: Compute the estimated observability matrix $\hat{\mathcal{O}}_f = \hat{U}_1 \hat{\Sigma}_1^{\frac{1}{2}}$, and let the top m rows of $\hat{\mathcal{O}}_f$ be $\hat{C}$.

3: Compute the estimated reversed extended controllability matrix $\hat{\mathcal{K}}_p = \hat{\Sigma}_1^{\frac{1}{2}} \hat{V}_1'$, and let the last m columns of $\hat{\mathcal{K}}_p$ be $\hat{K}_{p-1}$.

4: Compute $\hat{A} = (\hat{\mathcal{O}}_f^u)^\dagger \hat{\mathcal{O}}_f^l$, where $\hat{\mathcal{O}}_f^u$ is the submatrix formed by the top $m(f-1)$ rows of $\hat{\mathcal{O}}_f$, and $\hat{\mathcal{O}}_f^l$ is the submatrix formed by dropping the first m rows of $\hat{\mathcal{O}}_f$.

5: Return $(\hat{A}, \hat{C}, \hat{K}_{p-1})$.

3.4 Main Results

In this section, we will present our main results on bounding the estimation error $\|\hat{G} - G\|$ from (3.10). We will show that the term $\|(Y_- Y_-')^{-1}\|$ decreases with a rate of $\mathcal{O}(\frac{1}{N})$, and then upper bound the growth rate of other terms in (3.10) separately. Using recent results on the balanced realization algorithm with the adjustments to accommodate the non-steady state Kalman filter, we then show that the estimation error of the system matrices A, C, K_{p-1} will also be bounded when the error $\|\hat{G} - G\|$ is small enough.

First, denote the covariance matrix of the weighted past innovation $\mathcal{T}_p^0 E_-^i$, $1 \leq i \leq N$, as:

$$\Sigma_E \triangleq \mathbb{E}[\mathcal{T}_p^0 E_-^i E_-^{i'} \mathcal{T}_p^{0'}] = \mathcal{T}_p^0 \operatorname{diag}(\bar{R}_0, \cdots, \bar{R}_{p-1}) \mathcal{T}_p^{0'}.$$

Let $\sigma_E \triangleq \sigma_{min}(\Sigma_E)$. We first show that the weighted innovation covariance matrix Σ_E is strictly positive definite. The proof is similar to [40, Lemma 2].

Proposition 3.4.1. *Let $\sigma_E \triangleq \sigma_{min}(\Sigma_E)$. We have $\sigma_E \geq \sigma_{min}(R) > 0$.*

Proof. For any output trajectory i, its corresponding weighted past innovation $\mathcal{T}_p^0 E_-^i$ can be written as

$$\mathcal{T}_p^0 E_-^i = Y_-^i - \mathcal{O}_p \hat{x}_0^i = \mathcal{O}_p (x_0^i - \hat{x}_0^i) + \mathbf{T} W_-^i + V_-^i,$$

where W_-^i and V_-^i are the process and output noises respectively in system (3.1), and are defined similarly as Y_-^i. Matrix $\mathbf{T}$ is a block-Toeplitz matrix which accounts for the weight of the process noise in system (3.1), and its explicit form is omitted in the interest of space. Since $x_0^i - \hat{x}_0^i$, V_-^i and W_-^i are independent, we have

$$\Sigma_E = \mathbb{E}[\mathcal{T}_p^0 E_-^i E_-^{i'} \mathcal{T}_p^{0'}] \succeq \mathbb{E}[V_-^i V_-^{i'}] = \operatorname{diag}(R, \cdots, R).$$

Hence, we have $\sigma_E \geq \sigma_{min}(R) > 0$, where the second inequality comes from Assumption 2. □

Next we will show that the term $\|(Y_- Y_-^{'})^{-1}\|$ is decreasing with a rate of $\mathcal{O}(\frac{1}{N})$. Since $Y_- = \mathcal{O}_p \hat{X}_0 + \mathcal{T}_p^0 E_-$, the explicit form of $Y_- Y_-^{'}$ is

$$\begin{aligned} Y_- Y_-^{'} = {} & \mathcal{O}_p \hat{X}_0 \hat{X}_0^{'} \mathcal{O}_p^{'} + \mathcal{T}_p^0 E_- E_-^{'} \mathcal{T}_p^{0'} \\ & + \mathcal{O}_p \hat{X}_0 E_-^{'} \mathcal{T}_p^{0'} + \mathcal{T}_p^0 E_- \hat{X}_0^{'} \mathcal{O}_p^{'}, \end{aligned} \tag{3.11}$$

and we will bound these terms separately.

We will rely on the following lemma from [13, Lemma 2], which provides a non-asymptotic lower bound of a standard Wishart matrix.

Lemma 12. *Let $u_i \sim \mathcal{N}(0, I_{mp})$, $i = 1, \ldots, N$ be i.i.d random vectors. For any fixed $\delta > 0$, we have*

$$\sqrt{\lambda_{min}(\sum_{i=1}^{N} u_i u_i^*)} \geq \sqrt{N} - \sqrt{mp} - \sqrt{2 \log \frac{1}{\delta}}$$

with probability at least $1 - \delta$.

Proposition 3.4.2. *For any fixed $\delta > 0$, let $N \geq N_0 \triangleq 8mp + 16 \log \frac{2}{\delta}$. We have*

$$\mathcal{T}_p^0 E_- E_-^{'} \mathcal{T}_p^{0'} \succeq \frac{N}{4} \sigma_E I_{mp}$$

with probability at least $1 - \delta$.

Proof. For any output trajectory i, note that the past innovation E_-^i has the same distribution as a single Gaussian random vector, $E_-^i \sim \mathcal{N}(0, \text{diag}(\bar{R}_0, \cdots, \bar{R}_{p-1}))$, since the innovations $e_0^i, \cdots, e_{p-1}^i$ are independent zero-mean Gaussian random vectors with covariance matrices $\bar{R}_0, \cdots, \bar{R}_{p-1}$, respectively. We can rewrite E_-^i as

$$E_-^i = \text{diag}(\bar{R}_0^{\frac{1}{2}}, \cdots, \bar{R}_{p-1}^{\frac{1}{2}}) \mathbf{u}_i,$$

where $\mathbf{u}_i$ are i.i.d random vectors with $\mathbf{u}_i \sim \mathcal{N}(0, I_{mp})$. Let $U_- = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{bmatrix}$. Fixing $\delta > 0$ and applying Lemma 12, with probability of at least $1 - \delta$, we have

$$\sqrt{\lambda_{min}(U_- U_-^{'})} \geq \sqrt{N} - \sqrt{mp} - \sqrt{2 \log \frac{2}{\delta}}. \tag{3.12}$$

Considering the inequality $2(a^2 + b^2) \geq (a + b)^2$ and the assumption that $N \geq N_0 \triangleq 8mp + 16 \log \frac{2}{\delta}$, we have

$$2(mp + 2 \log \frac{2}{\delta}) \geq (\sqrt{mp} + \sqrt{2 \log \frac{2}{\delta}})^2,$$

$$\Rightarrow \frac{\sqrt{N}}{2} \geq \sqrt{mp} + \sqrt{2 \log \frac{2}{\delta}}.$$

In conjunction with (3.12), this yields $\sqrt{\lambda_{min}(U_-U_-')} \geq \frac{1}{2}\sqrt{N}$ with probability at least $1-\delta$. Hence, we have

$$U_-U_-' \succeq \frac{N}{4} I_{mp}$$

with probability of at least $1-\delta$.

Finally, multiplying by $\mathcal{T}_p^0 \operatorname{diag}(\bar{R}_0^{\frac{1}{2}}, \cdots, \bar{R}_{p-1}^{\frac{1}{2}})$ from the left and $\operatorname{diag}(\bar{R}_0^{\frac{1}{2}}, \cdots, \bar{R}_{p-1}^{\frac{1}{2}})\mathcal{T}_p^{0\prime}$ from the right gives $\mathcal{T}_p^0 E_-E_-'\mathcal{T}_p^{0\prime} \succeq \frac{N}{4}\Sigma_E \succeq \frac{N}{4}\sigma_E I_{mp}$ with probability at least $1-\delta$. □

To bound the cross terms due to the possibly non-zero batch matrix of initial states in (3.11), $\hat{X}_0 E_-'$, we will be using the following lemma from [42, Lemma A.1].

Lemma 13. *Let $M \in \mathbb{R}^{m\times n}$ be a matrix with $m \geq n$, and let $\eta \in \mathbb{R}$ be such that $\|M\| \leq \eta$. Let $Z \in \mathbb{R}^{m\times k}$ be a matrix with independent standard normal entries. Then, for all $t \geq 0$, with probability at least $1 - 2\exp(\frac{-t^2}{2})$,*

$$\|M'Z\| \leq \eta(\sqrt{2(n+k)} + t).$$

Proposition 3.4.3. *For any fixed $\delta > 0$, let $\gamma_p \triangleq \bar{\mathcal{R}}_T^{\frac{1}{2}}(\sqrt{2(n+mp)} + \sqrt{2\log\frac{2}{\delta}})$, and $\gamma_f \triangleq \bar{\mathcal{R}}_T^{\frac{1}{2}}(\sqrt{2(n+mf)} + \sqrt{2\log\frac{2}{\delta}})$. For $N \geq n$, each of these inequalities hold with probability at least $1-\delta$:*

$$\|\hat{X}_0 E_-'\| \leq \|\hat{X}_0\|\gamma_p,$$

$$\|\hat{X}_0 E_+'\| \leq \|\hat{X}_0\|\gamma_f.$$

Proof. We will only show the first inequality as the second one is almost identical. We can rewrite $E_- = \operatorname{diag}(\bar{R}_0^{\frac{1}{2}}, \cdots, \bar{R}_{p-1}^{\frac{1}{2}})U_-$, where $U_- = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N\end{bmatrix}$, where $\mathbf{u}_i$ are i.i.d random vectors with $\mathbf{u}_i \sim \mathcal{N}(0, I_{mp})$. Applying Lemma 13, we obtain

$$\begin{aligned}\|\hat{X}_0 E_-'\| &= \|\hat{X}_0 U_-' \operatorname{diag}(\bar{R}_0^{\frac{1}{2}}, \cdots, \bar{R}_{p-1}^{\frac{1}{2}})\| \\ &\leq \|\hat{X}_0\|\bar{\mathcal{R}}_T^{\frac{1}{2}}(\sqrt{2(n+mp)} + t)\end{aligned}$$

with probability at least $1-2\exp(\frac{-t^2}{2})$. Finally, setting $\delta = 2\exp(\frac{-t^2}{2})$, we have $t = \sqrt{2\log\frac{2}{\delta}}$. Plugging t into the above inequality, we get the desired form. □

Now we are ready to show that $\|(Y_-Y_-^{'})^{-1}\|$ is decreasing with a rate of $\mathcal{O}(\frac{1}{N})$.

Lemma 14. *Fix any $\delta > 0$ and let $N \geq \max\{N_0, N_1\}$, where $N_0 = 8mp + 16\log\frac{2}{\delta}$, and $N_1 \triangleq \frac{16\|\mathcal{T}_p^0\|\|\mathcal{O}_p\|\|\hat{X}_0\|\gamma_p}{\sigma_E}$. Define $\zeta \triangleq \mathcal{O}_p\mu\mu^{'}\mathcal{O}_p^{'}$. We have*

$$\|(Y_-Y_-^{'})^{-1}\| \leq \frac{8}{N\sigma_{min}(\sigma_E I_{mp} + 8\zeta)}$$

with probability at least $1-2\delta$.

Proof. Recall the explicit form of $Y_-Y_-^{'}$ in (3.11). Letting $u \in \mathbb{R}^{mp}$ be an arbitrary unit vector, we can write

$$\begin{aligned} u^{'}Y_-Y_-^{'}u &= u^{'}\mathcal{O}_p\hat{X}_0\hat{X}_0^{'}\mathcal{O}_p^{'}u + u^{'}\mathcal{T}_p^0E_-E_-^{'}\mathcal{T}_p^{0'}u + u^{'}\mathcal{O}_p\hat{X}_0E_-^{'}\mathcal{T}_p^{0'}u + u^{'}\mathcal{T}_p^0E_-\hat{X}_0^{'}\mathcal{O}_p^{'}u \\ &\geq u^{'}\mathcal{O}_p\hat{X}_0\hat{X}_0^{'}\mathcal{O}_p^{'}u + u^{'}\mathcal{T}_p^0E_-E_-^{'}\mathcal{T}_p^{0'}u - 2\|\hat{X}_0E_-^{'}\|\|\mathcal{T}_p^0\|\|\mathcal{O}_p\|, \end{aligned}$$

where we used the Cauchy–Schwarz inequality. Fixing $\delta > 0$, letting $N \geq N_0$, applying Proposition 3.4.2 and Proposition 3.4.3, and using a union bound, we have

$$u^{'}Y_-Y_-^{'}u \geq u^{'}\mathcal{O}_p\hat{X}_0\hat{X}_0^{'}\mathcal{O}_p^{'}u + \frac{N}{4}\sigma_E - 2\|\mathcal{T}_p^0\|\|\mathcal{O}_p\|\|\hat{X}_0\|\gamma_p$$

with probability at least $1-2\delta$.

Conditioning on the above event and letting $N \geq N_1 = \frac{16\|\mathcal{T}_p^0\|\|\mathcal{O}_p\|\|\hat{X}_0\|\gamma_p}{\sigma_E}$, we have

$$\begin{aligned} u^{'}Y_-Y_-^{'}u &\geq u^{'}\mathcal{O}_p\hat{X}_0\hat{X}_0^{'}\mathcal{O}_p^{'}u + \frac{N}{8}\sigma_E \\ &= u^{'}\mathcal{O}_pN\mu\mu^{'}\mathcal{O}_p^{'}u + u^{'}\frac{N}{8}\sigma_E I_{mp}u, \end{aligned}$$

where the equality is due to the fact that $\hat{x}_0^i = \mu$ for all i.

Consequently, we have

$$Y_-Y_-^{'} \succeq \mathcal{O}_pN\mu\mu^{'}\mathcal{O}_p^{'} + \frac{N}{8}\sigma_E I_{mp}.$$

Taking the inverse we get the desired result. □

To see that N will eventually be greater than N_1 even if $\|\hat{X}_0\|$ is non-zero, note that when the system is strictly stable or marginally stable, $\|\mathcal{T}_p^0\|$ and $\|\mathcal{O}_p\|$ will grow no faster than $\mathcal{O}(p^{\mathbf{d}})$ for some constant $\mathbf{d}$ (see Proposition 3.6.1 for $\|\mathcal{T}_p^0\|$, and [40, Corollary E.1] for $\|\mathcal{O}_p\|$). Further, γ_p is $\mathcal{O}(p^{\frac{1}{2}})$, and $\|\hat{X}_0\| = \sqrt{N}\|\mu\| = \mathcal{O}(\sqrt{N})$. Thus if $p = \mathcal{O}(\log N)$, N will eventually be greater than N_1 as N increases.

Now we will show that the term $\|E_+E_-'\|$ is $\mathcal{O}(\sqrt{N})$. We will leverage the following Lemma from[13, Lemma 1] to bound the product of the innovation terms.

Lemma 15. *Let $f_i \in \mathbb{R}^m$, $g_i \in \mathbb{R}^n$ be independent random vectors $f_i \sim \mathcal{N}(0, \Sigma_f)$ and $g_i \sim \mathcal{N}(0, \Sigma_g)$, for $i = 1, \cdots, N$. Let $N \geq 2(n+m)\log\frac{1}{\delta}$. For any fixed $\delta > 0$, we have*

$$\|\sum_{i=1}^{N} f_i g_i^*\| \leq 4\|\Sigma_f\|^{\frac{1}{2}}\|\Sigma_g\|^{\frac{1}{2}}\sqrt{N(m+n)\log\frac{9}{\delta}}.$$

with probability at least $1-\delta$.

Proposition 3.4.4. *For any fixed $\delta > 0$, let $N \geq N_2 \triangleq 2(mf+mp)\log\frac{1}{\delta}$. We have*

$$\|\mathcal{T}_f^p E_+ E_-' \mathcal{T}_p^{0'}\| \leq 4\|\mathcal{T}_f^p\|\|\mathcal{T}_p^0\|\bar{\mathcal{R}}_T\sqrt{N(mf+mp)\log\frac{9}{\delta}}$$

with probability at least $1-\delta$.

Proof. Note that the columns of E_+ are independent Gaussian random vectors, i.e., $E_+^i \sim \mathcal{N}(0, \mathrm{diag}(\bar{R}_p, \cdots, \bar{R}_{p+f-1}))$. Similarly, the columns of E_- are independent Gaussian random vectors, i.e., $E_-^i \sim \mathcal{N}(0, \mathrm{diag}(\bar{R}_0, \cdots, \bar{R}_{p-1}))$. Further, E_+^i and E_-^i are independent from classical results of Kalman filtering theory [47]. Applying Lemma 15, we get the desired result. □

With Proposition 3.4.4, we are now in place to prove the bound on the estimation error of G .

Theorem 3.4.1 (Bound on estimation error of G)**.** *Consider the Kalman filter form (3.2) of system (3.1) under Assumptions 2 and 3, and let G be defined as in (3.7). For any fixed $\delta > 0$,*

let $\hat{G}$ defined in (3.9) be the output of the linear regression described in Algorithm 2 given N trajectories of outputs, where $N \geq \max\{N_0, N_1, N_2\}$, where $N_0 = 8mp + 16\log\frac{2}{\delta}$, $N_1 = \frac{16\|\mathcal{T}_p^0\|\|\mathcal{O}_p\|\|\hat{X}_0\|\gamma_p}{\sigma_E}$, $N_2 = 2(mf+mp)\log\frac{1}{\delta}$. We have:

$$\|\hat{G} - G\| \leq \frac{\epsilon_1}{\sqrt{N}\sigma_{min}(\sigma_E I_{mp} + 8\zeta)} + \frac{\|\hat{X}_0\|\epsilon_2 + \|\hat{X}_0\|^2\epsilon_3}{N\sigma_{min}(\sigma_E I_{mp} + 8\zeta)}$$

with probability at least $1 - 4\delta$, where

$$\begin{aligned}
\epsilon_1 &= 32\|\mathcal{T}_f^p\|\|\mathcal{T}_p^0\|\bar{\mathcal{R}}_T\sqrt{(mf+mp)\log\frac{9}{\delta}} \\
\epsilon_2 &= 8\gamma_f\|\mathcal{T}_f^p\|\|\mathcal{O}_p\| + 8\|(A - K_{p-1}C)\cdots(A - K_0C)\|\gamma_p\|\mathcal{T}_p^0\|\|\mathcal{O}_f\|, \\
\epsilon_3 &= 8\|\mathcal{O}_f\|\|\mathcal{O}_p\|\|(A - K_{p-1}C)\cdots(A - K_0C)\|, \\
\gamma_p &= \bar{\mathcal{R}}_T^{\frac{1}{2}}(\sqrt{2(n+mp)} + \sqrt{2\log\frac{2}{\delta}}), \\
\gamma_f &= \bar{\mathcal{R}}_T^{\frac{1}{2}}(\sqrt{2(n+mf)} + \sqrt{2\log\frac{2}{\delta}}), \quad \zeta \triangleq \mathcal{O}_p\mu\mu^{'}\mathcal{O}_p^{'}.
\end{aligned}$$

Proof. Recall the expression of the error $\hat{G} - G$ in (3.10). We first bound the error term $\mathcal{T}_f^p E_+ Y_-^{'}(Y_-Y_-^{'})^{-1}$. Using $Y_- = \mathcal{O}_p\hat{X}_0 + \mathcal{T}_p^0 E_-$, we have

$$\begin{aligned}
\|\mathcal{T}_f^p E_+ Y_-^{'}(Y_-Y_-^{'})^{-1}\| \leq{}& \|\mathcal{T}_f^p E_+ \hat{X}_0^{'}\mathcal{O}_p^{'}\|\|(Y_-Y_-^{'})^{-1}\| \\
&+ \|\mathcal{T}_f^p E_+ E_-^{'}\mathcal{T}_p^{0'}\|\|(Y_-Y_-^{'})^{-1}\|.
\end{aligned}$$

Fix $\delta > 0$ and let $N \geq \max\{N_0, N_1, N_2\}$. Applying Proposition 3.4.3, Proposition 3.4.4 and Lemma 14 to the above inequality and using a union bound, we obtain

$$\|\mathcal{T}_f^p E_+ Y_-^{'}(Y_-Y_-^{'})^{-1}\| \leq \frac{8\|\hat{X}_0\|\gamma_f\|\mathcal{T}_f^p\|\|\mathcal{O}_p\|}{N\sigma_{min}(\sigma_E I_{mp} + 8\zeta)} + \frac{32\|\mathcal{T}_f^p\|\|\mathcal{T}_p^0\|\bar{\mathcal{R}}_T\sqrt{(mf+mp)\log\frac{9}{\delta}}}{\sqrt{N}\sigma_{min}(\sigma_E I_{mp} + 8\zeta)} \tag{3.13}$$

with probability at least $1 - 4\delta$.

Second, we bound the error term $\mathcal{O}_f(A - K_{p-1}C)\cdots(A - K_0C)\hat{X}_0 Y_-^{'}(Y_-Y_-^{'})^{-1}$. We have

$$\|\hat{X}_0 Y_-^{'}\| = \|\hat{X}_0\hat{X}_0^{'}\mathcal{O}_p^{'} + \hat{X}_0 E_-^{'}\mathcal{T}_p^{0'}\|.$$

Conditioning on the event $\|\hat{X}_0 E_-^{'}\| \leq \|\hat{X}_0\|\gamma_p$ from Proposition 3.4.3, we have

$$\|\hat{X}_0 Y_-^{'}\| \leq \|\hat{X}_0\|^2\|\mathcal{O}_p\| + \|\hat{X}_0\|\gamma_p\|\mathcal{T}_p^0\|.$$

Conditioning on the above event and the event $\|(Y_-Y_-^{'})^{-1}\| \leq \frac{8}{N\sigma_{min}(\sigma_E I_{mp}+8\zeta)}$ from Lemma 14, we have

$$\begin{aligned}
&\|\mathcal{O}_f(A-K_{p-1}C)\cdots(A-K_0C)\hat{X}_0Y_-^{'}(Y_-Y_-^{'})^{-1}\| \leq \\
&\|\mathcal{O}_f\|\|(A-K_{p-1}C)\cdots(A-K_0C)\|\|\hat{X}_0Y_-^{'}\|\|(Y_-Y_-^{'})^{-1}\| \\
&\leq \frac{8\|\mathcal{O}_f\|\|(A-K_{p-1}C)\cdots(A-K_0C)\|\|\hat{X}_0\|^2\|\mathcal{O}_p\|}{N\sigma_{min}(\sigma_E I_{mp}+8\zeta)} \\
&+ \frac{8\|\mathcal{O}_f\|\|(A-K_{p-1}C)\cdots(A-K_0C)\|\|\hat{X}_0\|\gamma_p\|\mathcal{T}_p^0\|}{N\sigma_{min}(\sigma_E I_{mp}+8\zeta)}.
\end{aligned} \tag{3.14}$$

Finally, we combine the two upper bounds from (3.13) and (3.14) to get the desired form. □

Below, we present some interpretations of Theorem 3.4.1.

Learning rate when $\|\hat{X}_0\|$ is zero, and the effects of trajectory length: When $\|\hat{X}_0\| = 0$, i.e., the initial state of the system (3.1) has zero mean, the upper bound of the error will not depend on ϵ_2, ϵ_3. Noting the dependencies on p, f in ϵ_1, setting p and f to be small will generally result in a smaller error bound of G, since we are estimating a smaller G. However, p, f should be greater than the order n (and p should also be large enough such that Assumption 3 is satisfied), so that Algorithm 3 can recover the system matrices from G. The estimator $\hat{G}$ can achieve a learning rate of $\mathcal{O}(\frac{1}{\sqrt{N}})$. This rate is faster than the single trajectory case reported in [40] in that there are no logarithmic factors, and it applies to both stable and unstable systems. This confirms the benefits of being able to collect multiple independent trajectories starting from $x_0 \sim \mathcal{N}(0, \Sigma_0)$.

Learning rate when $\|\hat{X}_0\|$ is nonzero, and the effects of trajectory length: When $\|\hat{X}_0\|$ is nonzero, the error bound will depend on ϵ_2, ϵ_3. Note that $\|\hat{X}_0\| = \sqrt{N}\|\mu\|$ when the initial state of each trajectory has mean μ. The term $\frac{\|\hat{X}_0\|^2\epsilon_3}{N\sigma_{min}(\sigma_E I_{mp}+8\zeta)}$ is $\mathcal{O}(1)$ when p is fixed. In such case, if the system is known to be marginally stable ($\rho(A) \leq 1$), we can leverage the fact that the norm $\|(A-K_{p-1}C)\cdots(A-K_0C)\|$ in ϵ_3 converges to zero exponentially fast

with p (see Proposition 3.6.2 in section 3.6), by setting $p = c \log N$ for some sufficiently large c, to force the term $\|(A - K_{p-1}C) \cdots (A - K_0 C)\|$ to go to zero no slower than $\mathcal{O}(\frac{1}{\sqrt{N}})$. The term $\bar{\mathcal{R}}_T$ is $\mathcal{O}(1)$ since the Kalman filter converges [47]. For the same reason, by fixing a small $f > n$, $\|\mathcal{T}_f^p\|$ is $\mathcal{O}(1)$. In addition, $\|\mathcal{T}_p^0\|$ and $\|\mathcal{O}_p\|$ are $\mathcal{O}(1)$ for stable systems, and $\mathcal{O}(p^{\mathbf{d}})$ for some constant $\mathbf{d}$ for marginally stable systems (see Proposition 3.6.1 in section 3.6 for $\|\mathcal{T}_p^0\|$, and [40, Corollary E.1] for $\|\mathcal{O}_p\|$). As a result, the error will decrease with a rate of $\mathcal{O}(\sqrt{\frac{\log N}{N}})$ for strictly stable systems, and $\mathcal{O}(\frac{(\log N)^d}{\sqrt{N}})$ for some constant d for marginally stable systems, even if $\|\hat{X}_0\|$ is non-zero.

The next step shows that the realization error of system matrix estimates $(\hat{A}, \hat{C}, \hat{K}_{p-1})$ provided by Algorithm 3 is bounded. Based on our assumption that $\mathcal{O}_f$ and $\mathcal{K}_p$ have rank n, the true G also has rank n. The proof of the following theorem entirely follows [40, Theorem 4], with the only difference being the replacement of steady state Kalman gain K by non-steady state Kalman gain K_{p-1}.

Theorem 3.4.2 (Bound on realizations of system matrices)**.** *Let G and $\hat{G}$ be defined in (3.7) and (3.9). Let the estimates based on $\hat{G}$ using Algorithm 3 be $\hat{\mathcal{O}}_f, \hat{\mathcal{K}}_p, \hat{A}, \hat{C}, \hat{K}_{p-1}$, and the corresponding matrices based on the true G using Algorithm 3 be $\tilde{\mathcal{O}}_f, \tilde{\mathcal{K}}_p, \tilde{A}, \tilde{C}, \tilde{K}_{p-1}$. If G has rank n and $\|\hat{G} - G\| \leq \frac{\sigma_n(G)}{4}$, then there exists an orthonormal matrix $\mathcal{T} \in \mathbb{R}^{n \times n}$ such that:*

$$\|\hat{\mathcal{O}}_f - \tilde{\mathcal{O}}_f \mathcal{T}\| \leq 2\sqrt{\frac{10n}{\sigma_n(G)}}\|\hat{G} - G\|,$$

$$\|\hat{C} - \tilde{C}\mathcal{T}\| \leq \|\hat{\mathcal{O}}_f - \tilde{\mathcal{O}}_f \mathcal{T}\|,$$

$$\|\hat{A} - \mathcal{T}'\tilde{A}\mathcal{T}\| \leq \frac{\sqrt{\|G\|} + \sigma_o}{\sigma_o^2}\|\hat{\mathcal{O}}_f - \tilde{\mathcal{O}}_f \mathcal{T}\|,$$

$$\|\hat{K}_{p-1} - \mathcal{T}'\tilde{K}_{p-1}\| \leq 2\sqrt{\frac{10n}{\sigma_n(G)}}\|\hat{G} - G\|,$$

where $\sigma_o \triangleq min(\sigma_n(\hat{\mathcal{O}}_f^u), \sigma_n(\tilde{\mathcal{O}}_f^u))$. Recall that the notation $\hat{\mathcal{O}}_f^u, \tilde{\mathcal{O}}_f^u$ refers to the submatrix formed by the top $m(f-1)$ rows of the respective matrix.

Remark 9. *Note that the matrices $\tilde{A}, \tilde{C}, \tilde{K}_{p-1}$ are equivalent to the original A, C, K_{p-1} matrices up to a similarity transformation. As p increases, $\|G\| = \|\mathcal{O}_f \mathcal{K}_p\|$ is $\mathcal{O}(1)$ since f is fixed, and $\|\mathcal{K}_p\|$ is also $\mathcal{O}(1)$ (see Proposition 3.6.3 in section 3.6). From Proposition*

3.6.4 in section 3.6, $\sigma_n(G)$ is lower bounded as p increases. As suggested in[40, Remark 3], the random term $\sigma_n(\hat{\mathcal{O}}_f^u)$ in σ_o can be replaced by a deterministic bound as

$$\sigma_n(\hat{\mathcal{O}}_f^u) \geq \sigma_n(\tilde{\mathcal{O}}_f^u) - \|\hat{\mathcal{O}}_f - \tilde{\mathcal{O}}_f \mathcal{T}\|.$$

Hence σ_o will be lower bounded by $\frac{\sigma_n(\tilde{\mathcal{O}}_f^u)}{2} > 0$ when the error $\|\hat{\mathcal{O}}_f - \tilde{\mathcal{O}}_f \mathcal{T}\|$ is small enough, where the inequality is due to the fact that we assumed the system is observable. Consequently, the term $\frac{\sqrt{\|G\|}+\sigma_o}{\sigma_o^2}$ is always $\mathcal{O}(1)$.

As a result, all estimation errors of system matrices depend linearly on $\|\hat{G} - G\|$, even if p is increasing. Hence, the realization error will decrease at least as fast as $\mathcal{O}(\frac{1}{\sqrt{N}})$ when $\|\hat{X}_0\| = 0$, and p is fixed. When $\|\hat{X}_0\|$ is non-zero, the error can decrease at a rate of $\mathcal{O}(\sqrt{\frac{\log N}{N}})$ for strictly stable systems, and at a rate of $\mathcal{O}(\frac{(\log N)^d}{\sqrt{N}})$ for some constant d for marginally stable systems by setting $p = c \log N$ for some positive constant c. Note that as p goes to infinity, the matrix $\hat{K}_{p-1}$ estimates the steady state Kalman gain $\mathcal{T}'\tilde{K}$.

On the other hand, the dependencies on $\sigma_n(G)$ and $\sigma_n(\hat{\mathcal{O}}_f^u)$ also show that the estimation error of system matrices depends on the "normalized estimation error" of G. Consequently, although our bound suggests that setting p, f to be small could potentially reduce the estimation error of G (when $\mu = 0$), it may not necessarily reduce the error of the system matrices. A similar issue also appears in the recovery of system matrices from Markov parameters [16]. It is of interest to study how trajectory length directly affects the realization error in future work.

3.5 Chapter Summary

In this chapter, we performed finite sample analysis of learning the dynamics of autonomous systems using multiple trajectories. Our results rely neither on controlled inputs, nor on observations of steady state behaviors of the system. We proved a learning rate that is consistent with [16] and [40] (up to logarithmic factors).

3.6 Proofs of Results

Lemma 16. *([40, Lemma E.2]). Consider the series $S_t = \sum_{i=0}^{t} \|A^i\|$. If the matrix A is strictly stable ($\rho(A) < 1$), then $S_t = \mathcal{O}(1)$; if the matrix A is marginally stable ($\rho(A) = 1$), then $S_t = \mathcal{O}(t^{\mathbf{d}})$, where $\mathbf{d}$ is the largest Jordan block of A corresponding to a unit circle eigenvalue $\|\lambda\| = 1$.*

Proposition 3.6.1. *The norm $\|\mathcal{T}_p^0\|$ is $\mathcal{O}(p^{\mathbf{d}})$ with p for some constant $\mathbf{d}$ when the system matrix A is marginally stable, and is $\mathcal{O}(1)$ when the system matrix A is strictly stable.*

Proof. Letting $K_{max}(p) = \max_{t\in 0,\dots,p-2} \|K_t\|$, where K_t is defined in (3.3). We have

$$\begin{aligned}\|\mathcal{T}_p^0\| &\le \|I_{mp}\| + \|C\|K_{max}(p) + \|C\|\|A\|K_{max}(p) + \cdots + \|C\|\|A^{p-2}\|K_{max}(p) \\ &= 1 + \|C\|K_{max}(p)\sum_{i=0}^{p-2}\|A^i\|.\end{aligned}$$

From Kalman filtering theory, the Kalman gain K_t converges to its steady state K under Assumption 2. Hence $K_{max}(p) = \mathcal{O}(1)$. From Lemma 16, we have the above sum is $\mathcal{O}(1)$ if A is strictly stable, and $\mathcal{O}(p^{\mathbf{d}})$ when A is marginally stable. □

Lemma 17. *([48, Theorem 6.6]). Let $U + A_1, U + A_2, \cdots$ be a sequence of $n \times n$ matrices. Given $\epsilon > 0$, there is a $\delta(\epsilon)$ such that if $\|A_k\| \le \delta(\epsilon)$ for all k, then*

$$\|(U + A_k)\cdots(U + A_1)\| \le \sigma(\rho(U) + \epsilon)^k$$

for some constant σ.

Proposition 3.6.2. *For any fixed integer k, where $p - 1 \ge k \ge 0$, we have*

$$\|(A - K_{p-1}C)\cdots(A - K_kC)\| = \mathcal{O}(e^{-c_0 p}),$$

for some positive constant c_0.

Proof. From Kalman filtering theory, the Kalman gain K_t converges to its steady state K, and the matrix $A - KC$ has spectral radius less than 1 under Assumption 2. Hence, we

can write $A - K_t C = A - KC + \eta_t$ for $t \geq 0$, where $\|\eta_t\|$ converges to 0. Pick ϵ such that $\epsilon + \rho(A - KC) < 1$. To apply Lemma 17, let $U = A - KC$, and let $k + t(\epsilon)$ be the smallest index such that $\|\eta_t\| \leq \delta(\epsilon)$ for all $t \geq k + t(\epsilon)$. Letting $p \geq k + t(\epsilon) + 1$, we have

$$\begin{aligned}
\|(A - K_{p-1}C) \cdots (A - K_k C)\| &\leq \| \prod_{t=1}^{p-k-t(\epsilon)} (U + \eta_{p-t})\| \| \prod_{t=p-k-t(\epsilon)+1}^{p-k} (U + \eta_{p-t})\| \\
&\leq \sigma(\rho(U) + \epsilon)^{p-k-t(\epsilon)} \| \prod_{t=p-k-t(\epsilon)+1}^{p-k} (U + \eta_{p-t})\| \\
&= \mathcal{O}(e^{-c_0 p}),
\end{aligned}$$

where the second inequality comes from Lemma 17. □

Proposition 3.6.3. *The norm $\|\mathcal{K}_p\|$ is $\mathcal{O}(1)$ with p.*

Proof. From Kalman filtering theory, the Kalman gain K_{p-1} converges to its steady state K, and the matrix $A - KC$ has spectral radius less than 1 under Assumption 2. Hence for any $t \geq 1$ we can write $A - K_t C = A - KC + \eta_t$, where $\|\eta_t\|$ converges to 0. Let $K_{max}(p) = \max_{t \in 0,\ldots,p-1} \|K_t\|$. We have $K_{max}(p) = \mathcal{O}(1)$ since K_t converges to K. Pick ϵ such that $\epsilon + \rho(A - KC) < 1$. To apply Lemma 17, let $U = A - KC$, and let $t(\epsilon)$ be the smallest index such that $\|\eta_t\| \leq \delta(\epsilon)$ for all $t \geq t(\epsilon)$. Letting $p \geq t(\epsilon) + 1$, we have

$$\begin{aligned}
\|\mathcal{K}_p\| &\leq K_{max}(p) + K_{max}(p) \sum_{t=2}^{p} \| \prod_{j=2}^{t} (U + \eta_{p-j+1})\| \\
&= K_{max}(p) + K_{max}(p) \sum_{t=p-t(\epsilon)+2}^{p} \| \prod_{j=2}^{t} (U + \eta_{p-j+1})\| + K_{max}(p) \sum_{t=2}^{p-t(\epsilon)+1} \| \prod_{j=2}^{t} (U + \eta_{p-j+1})\|.
\end{aligned}$$

From Proposition 3.6.2, we have

$$K_{max}(p) \sum_{t=p-t(\epsilon)+2}^{p} \| \prod_{j=2}^{t} (U + \eta_{p-j+1})\| = \mathcal{O}(e^{-c_0 p}). \tag{3.15}$$

From Lemma 17, we have

$$\begin{aligned} K_{max}(p) \sum_{t=2}^{p-t(\epsilon)+1} \| \prod_{j=2}^{t} (U + \eta_{p-j+1}) \| &\leq K_{max}(p) \sum_{t=2}^{p-t(\epsilon)+1} \sigma(\rho(U) + \epsilon)^{t-1} \\ &\leq \frac{K_{max}(p)\sigma}{1 - \rho(U) - \epsilon} = \mathcal{O}(1), \end{aligned} \tag{3.16}$$

where the second inequality comes from geometric series.

Finally, combining (3.15) and (3.16), we obtain $\|\mathcal{K}_p\| = \mathcal{O}(1)$.

□

Proposition 3.6.4. *Assume that* $\text{rank}(\mathcal{O}_f) = \text{rank}(\mathbf{K}_p) = n$*, where* n *is the order of the system. Fix any positive integer* $k, n \leq k < p$*. Let* $\mathbf{K}_{ss}$ *be the matrix formed by the last* k *block columns of the reversed extended controllability matrix* $\mathbf{K}_p$*. For sufficiently large* p*, we have the following inequalities:*

$$\begin{aligned} \sigma_n(G) &\geq \frac{\sigma_n(\mathcal{O}_f \mathbf{K_{ss}})}{2} > 0, \\ \sigma_n(\mathcal{K}_p) &\geq \frac{\sigma_n(\mathbf{K_{ss}})}{2} > 0. \end{aligned}$$

Proof. We will only show the first inequality as the second one is similar. Recall the definition of G in (3.7). We can rewrite $G = [M \quad \mathcal{O}_f \mathcal{K}_{tv}]$, where $\mathcal{K}_{tv}$ is the matrix formed by the last k block columns of $\mathcal{K}_p$, and M is some residual matrix. We have

$$GG' \succeq \mathcal{O}_f \mathcal{K}_{tv} \mathcal{K}_{tv}' \mathcal{O}_f'.$$

Hence, we have

$$\begin{aligned} \sigma_n(G) &\geq \sigma_n(\mathcal{O}_f \mathcal{K}_{tv}) = \sigma_n(\mathcal{O}_f \mathbf{K}_{ss} + \mathcal{O}_f \mathcal{K}_{tv} - \mathcal{O}_f \mathbf{K}_{ss}) \\ &\geq \sigma_n(\mathcal{O}_f \mathbf{K}_{ss}) - \|\mathcal{O}_f\| \|\mathcal{K}_{tv} - \mathbf{K}_{ss}\|, \end{aligned}$$

where the last inequality comes from the application of [49, Theorem 3.3.16(c)]. From Kalman filtering theory, the Kalman gain K_t converges to its steady state K under Assumption 2. Consequently, we have $\|\mathcal{K}_{tv} - \mathbf{K}_{ss}\|$ converges to zero as p increases. We see that

$\sigma_n(G)$ will eventually be lower bounded by $\frac{\sigma_n(\mathcal{O}_f \mathbf{K}_{ss})}{2} > 0$ as p increases, where the inequality comes from the assumption that $\mathcal{O}_f$ and $\mathbf{K}_p$ have full rank and the Cayley-Hamilton Theorem. $\square$

4. FINITE SAMPLE GUARANTEES FOR DISTRIBUTED ONLINE PARAMETER ESTIMATION WITH COMMUNICATION COSTS

4.1 Introduction

[7] In this chapter, we consider the system identification problem (or more generally, parameter estimation) in a distributed setting. In many cases, the available datasets are usually split among multiple agents/learners and come in a streaming manner, which require online processing. Coordination among the various agents to process their data also comes with a communication cost, and thus algorithms must be designed to balance the amount of communication with the speed and accuracy of learning.

The problem of distributed learning/optimization has been studied extensively over the last few decades, e.g.,[51], [52]. These papers typically provide theoretical guarantees on the convergence of local solutions to the optimizer of the sum of local functions over the network. When it comes to distributed online parameter estimation, the existing literature typically focuses on proving asymptotic convergence of the estimate to the true value, e.g., [53], [54]. There is another branch of research on distributed online learning that focuses on providing bounds on regret, which is defined as the difference between the costs generated by the sequence of local decisions and the true optimal costs obtained in hindsight, e.g., [55]–[57]. The bound on regret can be used as an appropriate metric to evaluate a proposed algorithm, as a sublinear regret implies that the algorithm performs as well as its centralized counterpart on average (over time). However, it is unclear how such bounds can be translated into the bounds on the accuracy of the learned model after a finite number of time-steps. The paper [58] studies distributed state estimation problem with finite time convergence guarantee with a fixed observation matrix, and under Byzantine faults. In contrast, we consider the problem where the observation/feature matrix is random, which is often encountered in general machine learning problems.

[7]↑The material in this chapter was published at the 2022 Conference on Decision and Control [50].

In this chapter, we propose a distributed online parameter estimation algorithm in a networked setting, which enables each agent to improve its estimation accuracy by communicating with neighbors in the network. Our algorithm can be viewed as an extension of the distributed least squares method in [59] to an online setting. In our algorithm, each agent stores two estimates of the true parameter: one computed purely based on local data and one computed after communicating with neighbors in the network. We provide finite time (or sample) upper bounds on the estimation errors of both of these two estimates, which highlight the role of communication. Our results demonstrate a trade-off between estimation error and communication costs. To balance such a trade-off, we discuss how we can leverage our finite time error bounds to determine a time at which the communication can be stopped (due to the costs associated with maintaining communications), while meeting a desired estimation accuracy. We also provide a numerical example to validate our results.

4.2 Problem Formulation

Consider a group of m agents $\mathcal{V}$ interconnected over an undirected and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. An edge $(i, j) \in \mathcal{E}$ is an unordered pair, which indicates a bidirectional communication link between agents i and j. Let $\mathcal{N}_i \triangleq \{j : (i, j) \in \mathcal{E}\}$ be the set of neighbors of agent i. The goal of these agents is to collaboratively estimate an unknown parameter $\Theta \in \mathbb{R}^{l\times n}$ with finite time guarantees, under a *finite number of communication steps*. At each time step $t = 1, 2, \cdots$, each agent $i \in \mathcal{V}$ gathers the data pair $(x_{i,t}, y_{i,t})$ generated by the following model

$$y_{i,t} = \Theta x_{i,t} + \eta_{i,t}, \tag{4.1}$$

where $y_{i,t} \in \mathbb{R}^l$ is the label vector, $x_{i,t} \in \mathbb{R}^n$ is the feature vector, and $\eta_{i,t} \in \mathbb{R}^l$ is the noise. We make the following assumption.

Assumption 4. *The feature vector $x_{i,t}$ and noise $\eta_{i,t}$ are Gaussian random vectors that are independent over time and agents, where $x_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma_x^2 I_n)$ and $\eta_{i,t} \sim \mathcal{N}(0, \sigma_\eta^2 I_l)$. The mean $\mu_{i,t} \in \mathbb{R}^n$ is deterministic with $\sup\{\|\mu_{i,t}\| : i \in \mathcal{V}, t \in \mathbb{Z}_{\geq 1}\} = \hat{u} \in \mathbb{R}_{\geq 0}$.*

The above model can be used to capture many problems. For example, it can be used to capture the problem of dynamical system identification via multiple independent trajectories (assuming zero initial condition and without process noise), where $y_{i,t}$ is the output of the system in each trajectory, $x_{i,t}$ is the input applied in each trajectory, and Θ is the Markov parameter matrix of the system, e.g., [15]. We note that the $x_{i,t}$ considered in our model allows for time-varying and agent-dependent mean $\mu_{i,t}$, and hence is more general than the analogous system identification problem, which typically considers zero-mean Gaussian inputs. We also note that our algorithm does not require any parameters of the model to be known in advance. However, we assume that there are known upper bounds on $\sigma_x, \sigma_\eta, \hat{\mu}, \|\Theta\|$, and there is a known non-zero lower bound on σ_x. These bounds will facilitate the design of certain user-specified parameters in our algorithm, which will become clear when we present our results.

Remark 10. *One may observe that a trivial solution to the above problem might be to not communicate at all, i.e., each agent only updates based on its local dataset. However, such a solution does not leverage the distributed nature of the problem, which provides each agent with the potential to speed up the learning by communicating with the other agents in the network. On the other hand, communications with the other agents should be carefully designed, as information from others might become less useful when each agent already has a good estimate based on the information it has so far. In the sequel, we study a distributed algorithm that leverages the communication network, which allows all agents to learn the model efficiently (when some upper/lower bounds on $\sigma_x, \sigma_\eta, \hat{\mu}, \|\Theta\|$ are available). More specifically, the algorithm allows every agent to hold an estimate with an estimation error comparable to that of the centralized solution throughout time, while saving communication costs.*

4.3 Algorithm

In this section, we describe a two-time-scale distributed algorithm. At each time step $t = 1, 2, \cdots$, based on its local dataset, each agent $i \in \mathcal{V}$ wishes to solve the following least squares problem:

$$\min_{\tilde{\Theta} \in \mathbb{R}^{l \times n}} \sum_{j=1}^{t} \|y_{i,j} - \tilde{\Theta} x_{i,j}\|^2. \tag{4.2}$$

The least squares local estimate for agent i, given its samples collected up to time step t, is

$$\hat{\Theta}_{i,t+1} = (\sum_{j=1}^{t} y_{i,j} x_{i,j}^{'})(\sum_{j=1}^{t} x_{i,j} x_{i,j}^{'})^{-1}, \tag{4.3}$$

assuming the matrix $\sum_{j=1}^{t} x_{i,j} x_{i,j}^{'}$ is invertible.

The above estimate can be updated iteratively with the arrival of new data pair $(x_{i,t}, y_{i,t})$, through

$$\begin{aligned} \alpha_{i,t+1} &= \alpha_{i,t} + y_{i,t} x_{i,t}^{'}, \\ \beta_{i,t+1} &= \beta_{i,t} + x_{i,t} x_{i,t}^{'}, \\ \hat{\Theta}_{i,t+1} &= \alpha_{i,t+1} \beta_{i,t+1}^{\dagger}, \end{aligned} \tag{4.4}$$

where $\alpha_{i,1} = 0, \beta_{i,1} = 0$. Note that $\beta_{i,t+1}^{\dagger} = \beta_{i,t+1}^{-1}$ once $\beta_{i,t+1}$ becomes invertible. Also, $\beta_{i,t+1}^{-1}$ can be updated iteratively using the Sherman-Morrison formula [60], which states $\beta_{i,t+1}^{-1} = \beta_{i,t}^{-1} - \frac{\beta_{i,t}^{-1} x_{i,t} x_{i,t}^{'} \beta_{i,t}^{-1}}{1 + x_{i,t}^{'} \beta_{i,t}^{-1} x_{i,t}}$.

The algorithm enters the communication phase when the conditions $t \bmod \zeta = 0$ and $t \leq S$ are satisfied, where $\zeta \in \mathbb{Z}_{\geq 1}$ and $S \in \mathbb{Z}_{\geq 0}$, i.e., when the current time step t is an integer multiple of the pre-specified communication period ζ and is less than the pre-specified stopping time S. Letting the superscript k denote the current communication time step, each agent $i \in \mathcal{V}$ sets $\alpha_{i,t+1}^{0} = \alpha_{i,t+1}, \beta_{i,t+1}^{0} = \beta_{i,t+1}$. At each communication time step k, each

agent $i \in \mathcal{V}$ broadcasts its current $\alpha_{i,t+1}^k$ and $\beta_{i,t+1}^k$ to its neighbors $j \in \mathcal{N}_i$, and receives $\alpha_{j,t+1}^k$ and $\beta_{j,t+1}^k$ from $j \in \mathcal{N}_i$. The update is given by

$$\begin{aligned}\alpha_{i,t+1}^{k+1} &= W(i,i)\alpha_{i,t+1}^k + \sum_{j\in\mathcal{N}_i} W(i,j)\alpha_{j,t+1}^k, \\ \beta_{i,t+1}^{k+1} &= W(i,i)\beta_{i,t+1}^k + \sum_{j\in\mathcal{N}_i} W(i,j)\beta_{j,t+1}^k,\end{aligned} \tag{4.5}$$

for $k = 0, 1, \ldots, T-1$, where $T \in \mathbb{Z}_{\geq 1}$ is the number of pre-specified total communication steps whenever the algorithm enters the communication phase, and $W \in \mathbb{R}^{m\times m}$ is the matrix where $W(i,j)$ is the weight agent $i \in \mathcal{V}$ assigns to agent $j \in \mathcal{V}$. We make the following assumption on W, which is commonly used, e.g., [61].

Assumption 5. *The weight matrix $W \in \mathbb{R}^{m\times m}$ associated with the communication graph $G = (\mathcal{V}, \mathcal{E})$ is assumed to satisfy: (1) $W(i,j) \in \mathbb{R}$ and $W(i,j) \geq 0$ for all $i, j \in \mathcal{V}$, and $W(i,j) = 0$ if $j \notin \mathcal{N}_i$ and $i \neq j$; (2)$W\mathbf{1}_m = \mathbf{1}_m$; (3) $W = W'$ and (4) $\rho(W) \triangleq \max\{\lambda_2(W), -\lambda_m(W)\} < 1$.*

The local estimate after communication is set to be $\bar{\Theta}_{i,t+1} = \alpha_{i,t+1}^T(\beta_{i,t+1}^T)^\dagger$. If there is no communication happened at the current time step t, agent i just keeps its estimate from the previous time-step, i.e., $\bar{\Theta}_{i,t+1} = \bar{\Theta}_{i,t}$.

The above steps are encapsulated in Algorithm 4.

Remark 11. *Note that Algorithm 4 has two time scales. In practice, this can capture the scenario where communication occurs at a much faster rate than obtaining samples. Further, note that both $\hat{\Theta}_{i,t+1}$ (without communication) and $\bar{\Theta}_{i,t+1}$ (after communication) are estimates of the true parameter Θ. In the next section, we will provide bounds on the finite time estimation errors $\|\hat{\Theta}_{i,t+1} - \Theta\|$ and $\|\bar{\Theta}_{i,t+1} - \Theta\|$. In practice, one could choose the estimate with smaller (estimated) error bound as the "true" output of the algorithm. In section 4.5, we will discuss how to choose the user-specified parameters ζ, S and T to enable efficient learning.*

Algorithm 4 Distributed Online Estimation Algorithm

Input Weight matrix W, stopping time S, communication period ζ, number of communication steps T

Each $v_i \in \mathcal{V}$ initializes $\alpha_{i,1} = 0, \beta_{i,1} = 0, \bar{\Theta}_{i,1} = 0$
for $t = 1, 2, 3, \ldots$ **do**
  for $v_i \in \mathcal{V}$ **do** ▷ Implement in parallel
    Gather the data pair $(x_{i,t}, y_{i,t})$, where $x_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma_x^2 I_n)$
    Update $\alpha_{i,t+1}, \beta_{i,t+1}, \hat{\Theta}_{i,t+1}$ as in (4.4)
    if $t \bmod \zeta = 0$ and $t \leq S$ **then**
      Set $\alpha_{i,t+1}^0 = \alpha_{i,t+1}, \beta_{i,t+1}^0 = \beta_{i,t+1}$
      for $k = 0, 1, \ldots, T-1$ **do**
        Broadcast $\alpha_{i,t+1}^k$ $\beta_{i,t+1}^k$ to $j \in \mathcal{N}_i$, and receive $\alpha_{j,t+1}^k$ $\beta_{j,t+1}^k$ from $j \in \mathcal{N}_i$
        Update $\alpha_{i,t+1}^{k+1}, \beta_{i,t+1}^{k+1}$ as in (4.5)
      end for
      $\bar{\Theta}_{i,t+1} = \alpha_{i,t+1}^T (\beta_{i,t+1}^T)^\dagger$
    else
      $\bar{\Theta}_{i,t+1} = \bar{\Theta}_{i,t}$
    end if
  end for
end for

4.4 Analysis of the Error

4.4.1 Local Estimation Error Without Communication

We will start with bounding the estimation error using only local samples. Note that for any agent $i \in \mathcal{V}$, we have

$$\begin{aligned}\|\hat{\Theta}_{i,t+1} - \Theta\| &= \|\alpha_{i,t+1}\beta_{i,t+1}^{-1} - \Theta\| \\ &= \|(\sum_{j=1}^{t} y_{i,j}x_{i,j}^{'})(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})^{-1} - \Theta\| \\ &= \|(\sum_{j=1}^{t} \eta_{i,j}x_{i,j}^{'})(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})^{-1}\| \\ &\leq \|\sum_{j=1}^{t} \eta_{i,j}x_{i,j}^{'}\|\|(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})^{-1}\|,\end{aligned} \tag{4.6}$$

assuming the the matrix $\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}$ is invertible. The proof follows by upper bounding the above terms separately.

Now we start with our first result.

Lemma 18. *Let Assumption 4 hold. Fix $\delta > 0$ and let $t \geq \max(t_1, t_2)$, where $t_1 = 8n + 16log\frac{2}{\delta}, t_2 = (\frac{16\hat{\mu}(\sqrt{4n}+\sqrt{2\log\frac{2}{\delta}})}{\sigma_x})^2$. For any $i \in \mathcal{V}$, letting $\bar{\mu}_{i,t} = \frac{4}{t\sigma_x^2}\sum_{j=1}^{t}\mu_{i,j}\mu_{i,j}^{'}$, with probability at least $1-2\delta$, we have both of the following inequalities:*

$$\begin{aligned}\|\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}\| &\leq t(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2), \\ \lambda_{min}(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}) &\geq \frac{t\sigma_x^2}{8}\lambda_{min}(I_n + \bar{\mu}_{i,t}).\end{aligned}$$

Proof. Fixing $i \in \mathcal{V}$, we can rewrite $x_{i,j} = \sigma_x u_{i,j} + \mu_{i,j}$, where $u_{i,j} \sim \mathcal{N}(0, I_n)$ for $j = 1, \ldots .t$. We have

$$\begin{aligned}\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'} &= \sum_{j=1}^{t}(\sigma_x u_{i,j} + \mu_{i,j})(\sigma_x u_{i,j}^{'} + \mu_{i,j}^{'}) \\ &= \sum_{j=1}^{t}\sigma_x^2 u_{i,j}u_{i,j}^{'} + \sum_{j=1}^{t}\mu_{i,j}\mu_{i,j}^{'} + \sum_{j=1}^{t}\sigma_x u_{i,j}\mu_{i,j}^{'} + \sum_{j=1}^{t}\sigma_x\mu_{i,j}u_{i,j}^{'}.\end{aligned} \tag{4.7}$$

To derive the upper bound in Lemma 18, we start with upper bounding the norm of the first term in the last equality of (4.7). Fixing $\delta > 0$ and applying Lemma 12, we have with probability at least $1 - \delta$,

$$\sqrt{\lambda_1(\sum_{j=1}^{t} u_{i,j}u_{i,j}^{'})} \leq \sqrt{t} + \sqrt{n} + \sqrt{2\log\frac{2}{\delta}}. \tag{4.8}$$

Further, we have

$$\begin{aligned} & \frac{1}{2}\sqrt{t} \geq \sqrt{n} + \sqrt{2\log\frac{2}{\delta}} \\ \iff \quad & \frac{t}{4} \geq (\sqrt{n} + \sqrt{2\log\frac{2}{\delta}})^2. \end{aligned}$$

Noting the inequality $2(a^2 + b^2) \geq (a+b)^2$, we can write $2(n + 2\log\frac{2}{\delta}) \geq (\sqrt{n} + \sqrt{2\log\frac{2}{\delta}})^2$. Letting $t \geq 8n + 16\log\frac{2}{\delta}$, one can then show that the following holds with probability at least $1 - \delta$:

$$\sqrt{\lambda_1(\sum_{i=1}^{t} u_{i,j}u_{i,j}^{'})} \leq \sqrt{t} + \sqrt{n} + \sqrt{2\log\frac{2}{\delta}} \leq \frac{3}{2}\sqrt{t}.$$

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

$$\|\sum_{i=1}^{t} \sigma_x^2 u_{i,j}u_{i,j}^{'}\| \leq \frac{9}{4}\sigma_x^2 t. \tag{4.9}$$

For the second term in the last equality of (4.7), from Assumption 4, we have

$$\|\sum_{i=1}^{t} \mu_{i,j}\mu_{i,j}^{'}\| \leq t\hat{\mu}^2. \tag{4.10}$$

For the last two terms in the last equality of (4.7), since $t \geq n$, we have

$$\begin{aligned} \|\sum_{j=1}^{t} \sigma_x u_{i,j}\mu_{i,j}^{'}\| &= \|\sum_{j=1}^{t} \sigma_x \mu_{i,j}u_{i,j}^{'}\| \\ &= \sigma_x \| \begin{bmatrix} \mu_{i,1} \cdots \mu_{i,t} \end{bmatrix} \begin{bmatrix} u_{i,1} \cdots u_{i,t} \end{bmatrix}^{'} \| \\ &\leq \sigma_x \sqrt{t}\hat{\mu}(\sqrt{4n} + \sqrt{2\log\frac{2}{\delta}}), \end{aligned} \tag{4.11}$$

with probability at least $1-\delta$, where the inequality comes from applying Lemma 13 and the fact that $\| \begin{bmatrix} \mu_{i,1} \cdots \mu_{i,t} \end{bmatrix} \| \leq \sqrt{t}\hat{\mu}$. Combining (4.9), (4.10) and (4.11) using a union bound, letting $t \geq \max\{8n + 16\log\frac{2}{\delta}, (\frac{16\hat{\mu}(\sqrt{4n}+\sqrt{2\log\frac{2}{\delta}})}{\sigma_x})^2\}$, we have with probability at least $1-2\delta$,

$$\begin{aligned}
\|\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}\| &\leq t(\frac{9}{4}\sigma_x^2 + \hat{\mu}^2) + 2\sigma_x\sqrt{t}\hat{\mu}(\sqrt{4n} + \sqrt{2\log\frac{2}{\delta}}) \\
&\leq t(\frac{9}{4}\sigma_x^2 + \hat{\mu}^2) + t\frac{\sigma_x^2}{8} \\
&= t(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2),
\end{aligned}$$

which is of the desired form.

Now we prove the lower bound in Lemma 18. We first lower bound the smallest eigenvalue of first term in the last equality of (4.7). Note that conditioning on the event in (4.8), we also have

$$\sqrt{\lambda_{min}(\sum_{j=1}^{t} u_{i,j}u_{i,j}^{'})} \geq \sqrt{t} - \sqrt{n} - \sqrt{2\log\frac{2}{\delta}}$$

from Lemma 12. Letting $t \geq 8n + 16\log\frac{2}{\delta}$, we have $\sqrt{\lambda_{min}(\sum_{j=1}^{t} u_{i,j}u_{i,j}^{'})} \geq \frac{1}{2}\sqrt{t}$, which implies

$$\sum_{i=1}^{t} \sigma_x^2 u_{i,j}u_{i,j}^{'} \succeq \frac{1}{4}\sigma_x^2 t I_n. \tag{4.12}$$

Further, from (4.7), note that

$$\lambda_{min}(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}) \geq \lambda_{min}(\sum_{j=1}^{t} \sigma_x^2 u_{i,j}u_{i,j}^{'} + \sum_{j=1}^{t} \mu_{i,j}\mu_{i,j}^{'}) - 2\|\sum_{j=1}^{t} \sigma_x u_{i,j}\mu_{i,j}^{'}\|,$$

where the inequality comes from [49, Theorem 3.3.16(c)]. Now, conditioning on the event in (4.11) and the event in (4.12), denoting $\bar{\mu}_{i,t} = \frac{4}{t\sigma_x^2}\sum_{j=1}^{t} \mu_{i,j}\mu_{i,j}^{'}$, we have

$$\begin{aligned}
&\lambda_{min}(\sum_{j=1}^{t} \sigma_x^2 u_{i,j}u_{i,j}^{'} + \sum_{j=1}^{t} \mu_{i,j}\mu_{i,j}^{'}) - 2\|\sum_{j=1}^{t} \sigma_x u_{i,j}\mu_{i,j}^{'}\| \\
&\geq \frac{t\sigma_x^2}{4}\lambda_{min}(I_n + \bar{\mu}_{i,t}) - 2\sigma_x\sqrt{t}\hat{\mu}(\sqrt{4n} + \sqrt{2\log\frac{2}{\delta}}).
\end{aligned}$$

Hence, when $t \geq (\frac{16\hat{\mu}(\sqrt{4n}+\sqrt{2\log\frac{2}{\delta}})}{\sigma_x})^2$, we have

$$\frac{t\sigma_x^2}{4}\lambda_{min}(I_n + \bar{\mu}_{i,t}) - 2\sigma_x\sqrt{t}\hat{\mu}(\sqrt{4n} + \sqrt{2\log\frac{2}{\delta}}) \geq \frac{t\sigma_x^2}{8}\lambda_{min}(I_n + \bar{\mu}_{i,t}).$$

□

Next, we will bound the error due to noise.

Proposition 4.4.1. *Let Assumption 4 hold. Fix $\delta > 0$ and let $t \geq t_3 = 2(n+l)\log\frac{1}{\delta}$. For any $i \in \mathcal{V}$, we have with probability at least $1 - 2\delta$,*

$$\|\sum_{j=1}^{t}\eta_{i,j}x_{i,j}^{'}\| \leq \sqrt{t}\sigma_\eta\left(4\sigma_x\sqrt{(n+l)\log\frac{9}{\delta}} + \hat{\mu}(\sqrt{2(l+n)} + \sqrt{2\log\frac{2}{\delta}})\right).$$

Proof. Fixing $i \in \mathcal{V}$, we can rewrite $\eta_{i,j} = \sigma_\eta f_{i,j}$ and $x_{i,j} = \sigma_x g_{i,j} + \mu_{i,j}$, where $f_{i,j}$, $g_{i,j}$ are independent Gaussian random vectors with $f_{i,j} \sim \mathcal{N}(0, I_l)$ and $g_{i,j} \sim \mathcal{N}(0, I_n)$, for $j = 1, \ldots, t$. We have

$$\begin{aligned}\|\sum_{j=1}^{t}\eta_{i,j}x_{i,j}^{'}\| &= \|\sum_{j=1}^{t}\sigma_\eta f_{i,j}(\sigma_x g_{i,j}^{'} + \mu_{i,j}^{'})\| \\ &\leq \|\sum_{j=1}^{t}\sigma_\eta\sigma_x f_{i,j}g_{i,j}^{'}\| + \|\sum_{j=1}^{t}\sigma_\eta f_{i,j}\mu_{i,j}^{'}\|.\end{aligned}$$

Fixing $\delta > 0$ and letting $t \geq 2(n+l)\log\frac{1}{\delta}$, applying Lemma 15, we have with probability at least $1 - \delta$

$$\|\sum_{j=1}^{t}\sigma_\eta\sigma_x f_{i,j}g_{i,j}^{'}\| \leq 4\sigma_x\sigma_\eta\cdot\sqrt{t(n+l)\log\frac{9}{\delta}}. \tag{4.13}$$

Next, notice that

$$\|\sum_{j=1}^{t}\sigma_\eta f_{i,j}\mu_{i,j}^{'}\| = \sigma_\eta\|\begin{bmatrix}\mu_{i,1}\cdots\mu_{i,t}\end{bmatrix}\begin{bmatrix}f_{i,1}\cdots f_{i,t}\end{bmatrix}^{'}\|.$$

Using the fact that $\| \begin{bmatrix} \mu_{i,1} \cdots \mu_{i,t} \end{bmatrix} \| \leq \sqrt{t}\hat{\mu}$ and applying Lemma 15, we have with probability at least $1-\delta$

$$\| \sum_{j=1}^{t} \sigma_\eta f_{i,j} \mu_{i,j}^{'} \| \leq \sigma_\eta \sqrt{t}\hat{\mu}(\sqrt{2(n+l)} + \sqrt{2\log\frac{2}{\delta}}). \tag{4.14}$$

Applying a union bound over the events in (4.13) and (4.14), we get the desired form. □

Theorem 4.4.1. *Let Assumption 4 hold. Fix $\delta > 0$ and let $t \geq \max(t_1, t_2, t_3)$, where $t_1 = 8n + 16log\frac{2}{\delta}, t_2 = (\frac{16\hat{\mu}(\sqrt{4n}+\sqrt{2\log\frac{2}{\delta}})}{\sigma_x})^2, t_3 = 2(n+l)\log\frac{1}{\delta}$. For any $i \in \mathcal{V}$, letting $\bar{\mu}_{i,t} = \frac{4}{t\sigma_x^2}\sum_{j=1}^{t} \mu_{i,j}\mu_{i,j}^{'}$, we have with probability at least $1-4\delta$,*

$$\|\hat{\Theta}_{i,t+1} - \Theta\| \leq \frac{C_1}{\sqrt{t}\sigma_x^2 \lambda_{min}(I_n + \bar{\mu}_{i,t})}, \tag{4.15}$$

where $C_1 = 8\sigma_\eta(4\sigma_x\sqrt{(n+l)\log\frac{9}{\delta}} + \hat{\mu}(\sqrt{2(l+n)} + 2\sqrt{\log\frac{2}{\delta}}))$.

Proof. Recall the expression of the estimation error in (4.6). Noting that $\|(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})^{-1}\| = \frac{1}{\lambda_{min}(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})}$, we can combine the second event in Lemma 18 and the event in Proposition 4.4.1 via a union bound to get the desired result. □

Remark 12. *Theorem 4.4.1 shows that the error is $\mathcal{O}(\frac{1}{\sqrt{t}})$. Note that when the mean $\mu_{i,j}$ is non-zero but invariant for all t, the bound could become more conservative when $n > 1$ (since the $\bar{\mu}_{i,t}$ term will not make the denominator larger). If this is known in advance, one could set $\hat{y}_{i,t} = y_{i,2t-1} - y_{i,2t}$ and $\hat{x}_{i,t} = x_{i,2t-1} - x_{i,2t}$. One then has $\hat{y}_{i,t} = \Theta\hat{x}_{i,t} + \hat{\eta}_{i,t}$, where $\hat{x}_{i,t} \sim \mathcal{N}(0, 2\sigma_x^2 I_n)$ and $\hat{\eta}_{i,t} \sim \mathcal{N}(0, 2\sigma_\eta^2 I_l)$. The same bound will still apply to the least squares solution using the transformed dataset, i.e., with the price of reducing the amount of samples by one-half, one could force the mean-dependent terms in Theorem 4.4.1 to go to zero. Such a transformation could result in a smaller bound when $\hat{\mu}$ is large enough.*

4.4.2 Global Estimation Error

Next, we look at the estimation error of the least squares estimate supposing that one has access to all samples across the network up to time step t. The global estimate and its associated estimation error are

$$\begin{aligned}\hat{\Theta}_{t+1} &\triangleq (\sum_{i=1}^{m}\sum_{j=1}^{t} y_{i,j}x_{i,j}^{'})(\sum_{i=1}^{m}\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})^{-1} \\ &= (\sum_{i=1}^{m}\alpha_{i,t+1})(\sum_{i=1}^{m}\beta_{i,t+1})^{-1},\end{aligned} \tag{4.16}$$

$$\|\hat{\Theta}_{t+1} - \Theta\| = \|(\sum_{i=1}^{m}\sum_{j=1}^{t} \eta_{i,j}x_{i,j}^{'})(\sum_{i=1}^{m}\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'})^{-1}\|,$$

assuming the matrix $\sum_{i=1}^{m}\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}$ is invertible. The proof of the following theorem entirely follows Theorem 4.4.1 due to Assumption 4, with slight adjustments to accommodate possibly different means of $x_{i,j}$ across the network.

Theorem 4.4.2. *Let Assumption 4 hold. Fix $\delta > 0$, and let $t \geq \frac{1}{m}\max(t_1, t_2, t_3)$, where $t_1 = 8n + 16log\frac{2}{\delta}, t_2 = (\frac{16\hat{\mu}(\sqrt{4n}+\sqrt{2\log\frac{2}{\delta}})}{\sigma_x})^2, t_3 = 2(n+l)\log\frac{1}{\delta}$. Letting $\bar{\mu}_t = \frac{4}{mt\sigma_x^2}\sum_{i=1}^{m}\sum_{j=1}^{t}\mu_{i,j}\mu_{i,j}^{'}$, we have with probability at least $1-4\delta$,*

$$\|\hat{\Theta}_{t+1} - \Theta\| \leq \frac{C_1}{\sqrt{mt}\sigma_x^2\lambda_{min}(I_n + \bar{\mu}_t)},$$

where C_1 is defined in Theorem 4.4.1.

Remark 13. *Theorem 4.4.2 indicates that the global estimation error bound is approximately $\frac{1}{\sqrt{m}}$ of the local estimation error bound for agent $i \in \mathcal{V}$ in Theorem 4.4.1 (when $\bar{\mu}_t$ in Theorem 4.4.2 is approximately equal to $\bar{\mu}_{i,t}$ in Theorem 4.4.1). Next, we will analyze the local estimation error after finite communication steps, which shows how communication could help agents benefit from the global dataset.*

4.4.3 Local Estimation Error After Communication

To derive the error bound of the local estimate after communication, we first define some quantities for notational simplicity. Recall the roles of the stopping time S and the communication period ζ in Algorithm 4. For t satisfying $t \bmod \zeta = 0$ and $t \leq S$ (note that we will only consider such t in this section), define $\bar{\alpha}_{t+1} \triangleq \frac{1}{m}\sum_{i=1}^{m} \alpha_{i,t+1}$ and $\bar{\beta}_{t+1} \triangleq \frac{1}{m}\sum_{i=1}^{m} \beta_{i,t+1}$. For any $i \in \mathcal{V}$, note that

$$\begin{aligned} \|\bar{\Theta}_{i,t+1} - \Theta\| &= \|\alpha_{i,t+1}^T(\beta_{i,t+1}^T)^{-1} - \Theta\| \\ &= \|\alpha_{i,t+1}^T(\beta_{i,t+1}^T)^{-1} - \bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1} + \bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1} - \Theta\| \\ &\leq \|\alpha_{i,t+1}^T(\beta_{i,t+1}^T)^{-1} - \bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1}\| + \|\bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1} - \Theta\|, \end{aligned} \tag{4.17}$$

under the invertibility assumption.

The second portion of the above inequality can be bounded using Theorem 4.4.2, since $\|\bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1} - \Theta\| = \|\hat{\Theta}_{t+1} - \Theta\|$. Now we will focus on bounding the first term, which corresponds to the error due to network convergence at time step t, using T steps of communication. For t satisfying $t \bmod \zeta = 0$ and $t \leq S$, fixing $i \in \mathcal{V}$ and defining $\epsilon_{\bar{\alpha}_{i,t+1}^T} \triangleq \alpha_{i,t+1}^T - \bar{\alpha}_{t+1}$ and $\epsilon_{(\beta_{i,t+1}^T)^{-1}} \triangleq (\beta_{i,t+1}^T)^{-1} - \bar{\beta}_{t+1}^{-1}$, we have

$$\begin{aligned} \|\alpha_{i,t+1}^T(\beta_{i,t+1}^T)^{-1} - \bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1}\| &= \|(\bar{\alpha}_{t+1} + \epsilon_{\bar{\alpha}_{i,t+1}^T})(\bar{\beta}_{t+1}^{-1} + \epsilon_{(\beta_{i,t+1}^T)^{-1}}) - \bar{\alpha}_{t+1}\bar{\beta}_{t+1}^{-1}\| \\ &\leq \|\bar{\alpha}_{t+1}\epsilon_{(\beta_{i,t+1}^T)^{-1}}\| + \|\epsilon_{\bar{\alpha}_{i,t+1}^T}\bar{\beta}_{t+1}^{-1}\| + \|\epsilon_{\bar{\alpha}_{i,t+1}^T}\epsilon_{(\beta_{i,t+1}^T)^{-1}}\| \\ &\leq \|\bar{\alpha}_{t+1}\|\|\epsilon_{(\beta_{i,t+1}^T)^{-1}}\| + \|\epsilon_{\bar{\alpha}_{i,t+1}^T}\|\|\bar{\beta}_{t+1}^{-1}\| + \|\epsilon_{\bar{\alpha}_{i,t+1}^T}\|\|\epsilon_{(\beta_{i,t+1}^T)^{-1}}\|, \end{aligned} \tag{4.18}$$

and we will bound the above terms separately.

Before we proceed, we will define some probabilistic events. Let t satisfy $t \bmod \zeta = 0$ and $t \leq S$. With the replacement of δ by $\hat{\delta}$, let $t \geq \max(t_1, t_2, t_3)$ defined in Lemma 18 and Proposition 4.4.1. Let E_1 be the event such that the event in Lemma 18 occurs for all $i \in \mathcal{V}$ at time step t, i.e., $E_1 \triangleq \{\{\|\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}\| \leq t(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2)\} \cap \{\lambda_{min}(\sum_{j=1}^{t} x_{i,j}x_{i,j}^{'}) \geq \frac{t\sigma_x^2}{8}\lambda_{min}(I_n + \bar{\mu}_{i,t})\}\}$ for all $i \in \mathcal{V}$, where $\bar{\mu}_{i,t} = \frac{4}{t\sigma_x^2}\sum_{j=1}^{t} \mu_{i,j}\mu_{i,j}^{'}$. Similarly, let E_2 be the event such that the event in Proposition 4.4.1 occurs for all $i \in \mathcal{V}$ at time step t, i.e.,

$E_2 \triangleq \{\|\sum_{j=1}^{t}\eta_{i,j}x_{i,j}^{'}\| \leq \sqrt{t}\sigma_\eta(4\sigma_x\sqrt{(n+l)\log\frac{9}{\hat{\delta}}} + \hat{\mu}(\sqrt{2(l+n)} + \sqrt{2\log\frac{2}{\hat{\delta}}}))\}$ for all $i \in \mathcal{V}$. Applying a union bound over all $i \in \mathcal{V}$, we have

$$E_3 \triangleq E_1 \cap E_2 \tag{4.19}$$

occurs with probability at least $1 - 4m\hat{\delta}$.

Proposition 4.4.2. *Conditioning on event E_3 in (4.19), we have*

$$\|\bar{\alpha}_{t+1}\| \leq tc_1 + \sqrt{t}c_2,$$

where $c_1 \triangleq \|\Theta\|(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2)$, $c_2 \triangleq \sigma_\eta(4\sigma_x\sqrt{(n+l)\log\frac{9}{\hat{\delta}}} + \hat{\mu}(\sqrt{2(l+n)} + \sqrt{2\log\frac{2}{\hat{\delta}}}))$.

Proof. We have

$$\begin{aligned}
\|\bar{\alpha}_{t+1}\| &= \|\frac{1}{m}\sum_{i=1}^{m}\alpha_{i,t+1}\| \\
&= \|\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{t}y_{i,j}x_{i,j}^{'}\| \\
&= \|\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{t}(\Theta x_{i,j} + \eta_{i,j})x_{i,j}^{'}\| \\
&\leq \frac{1}{m}(\|\Theta\|\|\sum_{i=1}^{m}\sum_{j=1}^{t}x_{i,j}x_{i,j}^{'}\| + \|\sum_{i=1}^{m}\sum_{j=1}^{t}\eta_{i,j}x_{i,j}^{'}\|) \\
&\leq t\|\Theta\|(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2) + \sqrt{t}\sigma_\eta(4\sigma_x\sqrt{(n+l)\log\frac{9}{\delta}} + \hat{\mu}(\sqrt{2(l+n)} + \sqrt{2\log\frac{2}{\delta}})),
\end{aligned}$$

where the last inequality is due to event E_3. □

We will use the following result from [62].

Lemma 19. *Consider a weight matrix $W \in \mathbb{R}^{m\times m}$ that satisfies Assumption 5. The following inequality holds:*

$$\max_{i\in\{1,\cdots,m\}}\sum_{j\in\{1,\cdots,m\}}\left|W^T(i,j) - \frac{1}{m}\right| \leq \sqrt{m}(\rho(W))^T,$$

where $\rho(W) = \max\{\lambda_2(W), -\lambda_m(W)\}$.

Proposition 4.4.3. *Let Assumption 5 hold. Conditioning on event E_3 in (4.19), for all $i \in \mathcal{V}$, we have*

$$\|\epsilon_{\bar{\alpha}_{i,t+1}^T}\| \leq m^{\frac{3}{2}}\sqrt{l}(\rho(W))^T(tc_1 + \sqrt{t}c_2),$$

where c_1 and c_2 are defined in Proposition 4.4.2.

Proof. Define $A_{t+1}^T = \left[\alpha_{1,t+1}^T, \cdots, \alpha_{m,t+1}^T\right]$ and $\bar{A}_{t+1} = \left[\bar{\alpha}_{t+1}, \cdots, \bar{\alpha}_{t+1}\right]$. For all $i \in \mathcal{V}$, we have

$$\|\epsilon_{\bar{\alpha}_{i,t+1}^T}\| \leq \|A_{t+1}^T - \bar{A}_{t+1}\|, \tag{4.20}$$

since $\epsilon_{\bar{\alpha}_{i,t+1}^T}$ is a submatrix of the matrix $A_{t+1}^T - \bar{A}_{t+1}$. Now, let $W \in \mathbb{R}^{m\times m}$ be the weight matrix associated with the communication graph, $\bar{W} \in \mathbb{R}^{m\times m}$ be a matrix with all components equal to $\frac{1}{m}$, and

$$\hat{\alpha}_{t+1}^{i,j} = \begin{bmatrix}\alpha_{1,t+1}(i,j) & \alpha_{2,t+1}(i,j) & \cdots & \alpha_{m,t+1}(i,j)\end{bmatrix}'.$$

We have

$$\begin{aligned}
\|A_{t+1}^T - \bar{A}_{t+1}\| &\leq \|A_{t+1}^T - \bar{A}_{t+1}\|_F \\
&= \|\operatorname{vec}(A_{t+1}^T - \bar{A}_{t+1})\| \\
&= \left\| \begin{bmatrix} W^T - \bar{W} & 0 & \cdots & 0 \\ 0 & W^T - \bar{W} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W^T - \bar{W} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_{t+1}^{1,1} \\ \hat{\alpha}_{t+1}^{1,2} \\ \vdots \\ \hat{\alpha}_{t+1}^{l,n} \end{bmatrix} \right\| \\
&\leq \|W^T - \bar{W}\| \left\| \begin{bmatrix}\alpha_{1,t+1} & \alpha_{2,t+1} & \cdots & \alpha_{m,t+1}\end{bmatrix} \right\|_F.
\end{aligned} \tag{4.21}$$

Now applying Lemma 19, we have

$$\begin{aligned}
\|W^T - \bar{W}\| = \|(W^T)' - \bar{W}'\| &\leq \sqrt{m}\|(W^T)' - \bar{W}'\|_1 \\
&\leq m(\rho(W))^T.
\end{aligned} \tag{4.22}$$

Further, since the rank of the matrix $\begin{bmatrix}\alpha_{1,t+1} & \alpha_{2,t+1} & \cdots & \alpha_{m,t+1}\end{bmatrix}$ is at most l, conditioning on event E_3, we have

$$\begin{aligned}\left\|\begin{bmatrix}\alpha_{1,t+1} & \alpha_{2,t+1} & \cdots & \alpha_{m,t+1}\end{bmatrix}\right\|_F &\leq \sqrt{l}\left\|\begin{bmatrix}\alpha_{1,t+1} & \alpha_{2,t+1} & \cdots & \alpha_{m,t+1}\end{bmatrix}\right\| \\ &\leq \sqrt{ml}\max_{i\in\{1,\cdots,m\}}(\|\alpha_{i,t+1}\|) \\ &\leq \sqrt{ml}\max_{i\in\{1,\cdots,m\}}(\|\Theta\|\|\sum_{j=1}^{t}x_{i,j}x_{i,j}^{'}\| + \|\sum_{j=1}^{t}\eta_{i,j}x_{i,j}^{'}\|) \\ &\leq \sqrt{ml}(t\|\Theta\|(\frac{19}{8}\sigma_x^2+\hat{\mu}^2) \\ &+ \sqrt{t}\sigma_\eta(4\sigma_x\sqrt{(n+l)\log\frac{9}{\hat{\delta}}} + \hat{\mu}(\sqrt{2(l+n)} + \sqrt{2\log\frac{2}{\hat{\delta}}})).\end{aligned} \tag{4.23}$$

The result follows by substituting (4.22) and (4.23) into (4.21). □

Proposition 4.4.4. *Conditioning on event E_3 in (4.19), we have*

$$\|\bar{\beta}_{t+1}^{-1}\| \leq \frac{8}{\sigma_x^2 t}.$$

Proof. Conditioning on event E_3 in (4.19), we have

$$\bar{\beta}_{t+1} = \frac{1}{m}\sum_{i=1}^{m}\beta_{i,t+1} \succeq \frac{\sigma_x^2}{8}tI_n.$$

Taking the inverse we get the desired result. □

Proposition 4.4.5. *Let Assumption 5 hold. Conditioning on event E_3 in (4.19), for all $i \in \mathcal{V}$, we have*

$$\|\epsilon_{(\bar{\beta}_{i,t+1}^T)^{-1}}\| \leq \rho(W))^T\frac{c_3}{t},$$

where $c_3 \triangleq \frac{152m^{\frac{3}{2}}\sqrt{5n}}{\sigma_x^2} + \frac{64m^{\frac{3}{2}}\sqrt{5n}}{\sigma_x^4}\hat{\mu}^2$.

Proof. Denote $\epsilon_{\beta^T_{i,t+1}} = \beta^T_{i,t+1} - \bar{\beta}_{t+1}$. Letting $W \in \mathbb{R}^{m\times m}$ be the weight matrix associated with the communication graph, $\bar{W} \in \mathbb{R}^{m\times m}$ be the matrix with all components equal to $\frac{1}{m}$. Following the same procedure as in the proof of Proposition 4.4.3, for all $i \in \mathcal{V}$, we have

$$\begin{aligned}\|\epsilon_{\bar{\beta}^T_{i,t+1}}\| &\leq \|W^T - \bar{W}\| \left\| \begin{bmatrix} \beta_{1,t+1} & \beta_{2,t+1} & \cdots & \beta_{m,t+1} \end{bmatrix} \right\|_F \\ &\leq m^{\frac{3}{2}} \sqrt{n} (\rho(W))^T \max_{i\in\{1,\cdots,m\}} (\|\beta_{i,t+1}\|) \\ &\leq m^{\frac{3}{2}} \sqrt{n} (\rho(W))^T t(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2),\end{aligned} \tag{4.24}$$

where the last inequality is due to event E_3.

Further, for all $i \in \mathcal{V}$, we have

$$\begin{aligned}\|\epsilon_{(\bar{\beta}^T_{i,t+1})^{-1}}\| &= \|(\beta^T_{i,t+1})^{-1} - \bar{\beta}^{-1}_{t+1}\| \\ &\leq \sqrt{5} \max\{\frac{1}{\sigma^2_{min}(\beta^T_{i,t+1})}, \frac{1}{\sigma^2_{min}(\bar{\beta}_{t+1})}\} \|\epsilon_{\bar{\beta}^T_{i,t+1}}\|,\end{aligned} \tag{4.25}$$

where the inequality comes from [63].

Conditioning on event E_3 and noting that $\beta^T_{i,t+1} = \sum_{i=1}^m q_i \beta_{i,t+1}$ for some weights $0 \leq q_i \leq 1$ and $\sum_{i=1}^m q_i = 1$, we have

$$\sigma_{min}(\beta^T_{i,t+1}) = \sigma_{min}(\sum_{i=1}^m q_i \beta_{i,t+1}) \geq \frac{\sigma_x^2}{8} t. \tag{4.26}$$

Consequently, substituting the above inequality and Proposition 4.4.4 into (4.25), and combining with (4.24), we obtain

$$\begin{aligned}\|\epsilon_{(\bar{\beta}^T_{i,t+1})^{-1}}\| &\leq \frac{64\sqrt{5}}{\sigma_x^4 t^2} \|\epsilon_{\bar{\beta}^T_{i,t+1}}\| \\ &\leq \frac{64\sqrt{5}}{\sigma_x^4 t^2} m^{\frac{3}{2}} \sqrt{n} (\rho(W))^T t(\frac{19}{8}\sigma_x^2 + \hat{\mu}^2).\end{aligned}$$

□

Now we are ready to bound the local estimation error after communication.

Theorem 4.4.3. *Let Assumptions 4 and 5 hold. Fix $\hat{\delta} > 0$ and let $t \geq \max(t_1, t_2, t_3)$, where $t_1 = 8n + 16log\frac{2}{\hat{\delta}}, t_2 = (\frac{16\hat{\mu}(\sqrt{4n}+\sqrt{2\log\frac{2}{\hat{\delta}}})}{\sigma_x})^2, t_3 = 2(n+l)\log\frac{1}{\hat{\delta}}$. For $t \bmod \zeta = 0$ and $t \leq S$, fix $\delta > 0$ and denote $\bar{\mu}_t = \frac{4}{mt\sigma_x^2}\sum_{i=1}^{m}\sum_{j=1}^{t}\mu_{i,j}\mu_{i,j}^{'}$. We have with probability at least $1-4m\hat{\delta}-4\delta$,*

$$\|\bar{\Theta}_{i,t+1} - \Theta\| \leq \underbrace{(\rho(W))^T C_0}_{\text{Error due to network convergence}} + \underbrace{\frac{C_1}{\sqrt{mt}\sigma_x^2\lambda_{min}(I_n + \bar{\mu}_t)}}_{\text{Error due to noise}}, \tag{4.27}$$

for all $i \in \mathcal{V}$, where $C_0 = c_3(c_1 + t^{-1/2}c_2) + \frac{8m^{\frac{3}{2}}\sqrt{l}(c_1+t^{-1/2}c_2)}{\sigma_x^2} + (\rho(W))^T m^{\frac{3}{2}}\sqrt{l}c_3(c_1 + t^{-1/2}c_2)$, and C_1, c_1, c_2, c_3 are defined in Theorem 4.4.1, Proposition 4.4.2 and Proposition 4.4.5.

Proof. Recall the decomposition of error from (4.17) and (4.18). Note that the event E_3 in (4.19) occurs with probability at least $1 - 4m\hat{\delta}$ when $t \geq \max(t_1, t_2, t_3)$. Combine event E_3 and the event in Theorem 4.4.2 using a union bound. Applying Propositions 4.4.2, 4.4.3, 4.4.4 and 4.4.5, we get the desired result. □

Remark 14. *Theorem 4.4.3 demonstrates a trade-off between estimation error and communication costs. By choosing $\hat{\delta}$ small, as T tends to infinity, the first term in the bound tends to zero, and agent $i \in \mathcal{V}$ can almost recover the same performance guarantee as if it had access to all samples across the network up to time step t (note that the second term in the error bound reduces the local estimation error bound in Theorem 4.4.1 by approximately $\frac{1}{\sqrt{m}}$). The speed the first term goes to zero depends on the network topology. Further, this result implies that by choosing T large such that the first term in the bound is small, communication becomes less important as t increases (i.e., as each agent keeps collecting samples), since the the second term goes to zero more slowly. Consequently, the improvements of the new local estimate after communication over the old estimates $\bar{\Theta}_{i,t+1}$ and $\hat{\Theta}_{i,t+1}$ will become smaller. In the next section, we discuss how to choose those user specified parameters to balance the trade-off between estimation error and communication costs, leveraging the above observation.*

4.5 Determining the communication period, the stopping time, and the number of communication steps

In short, the communication can be stopped when the minimum between the largest local error bound (over agents) in Theorem 4.4.1 and the error bound in Theorem 4.4.3 is less than some pre-specified threshold value $\epsilon \in \mathbb{R}_{>0}$. To achieve that, one needs to first specify the communication period ζ. Note that larger ζ corresponds to sparser communication. Further, one can specify how much error at most due to network convergence in Theorem 4.4.3 (first term in (4.27)) can be tolerated, denoted as $\epsilon_N \in \mathbb{R}_{>0}$. Smaller ϵ_N would require more communication steps. Based on that, one can compute the number of communication steps T that makes the error due to network convergence in Theorem 4.4.3 always less than ϵ_N. Consequently, one can then evaluate the bounds in Theorem 4.4.1 and Theorem 4.4.3 and determine the stopping time S. Note that the bounds in Theorem 4.4.1 and Theorem 4.4.3 involve parameters that may be unknown in practice. However, it suffices to replace $\bar{\mu}_t, \bar{\mu}_{i,t}$ by 0, and $\sigma_x, \sigma_\eta, \hat{\mu}, \|\Theta\|$ by their corresponding estimated upper/lower bounds.

Although communication could still help to reduce estimation error after $t > S$, even infinite communication steps can only allow each agent to recover the same estimation error bound as if it had access to the global dataset, under which the reduction of error could be negligible in practice when $\|\hat{\Theta}_{i,t+1} - \Theta\|$ or $\|\bar{\Theta}_{i,t+1} - \Theta\|$ is already small enough. Consequently, it might be preferable for these agents to start updating purely based on local data, considering the communication costs. We will illustrate this idea in the next section empirically.

4.6 Numerical Experiment

In this example, we consider a network of $m = 6$ agents trying to learn model (4.1), where

$$\Theta = \begin{bmatrix} 1.6 & 0.3 \\ 0.8 & 0.3 \end{bmatrix}, \sigma_x = 3, \sigma_\eta = 1,$$

and $\mu_{i,t} = 0$ for all i and t. The weight matrix associated with the communication graph is

$$W = \begin{bmatrix} 1/3 & 1/3 & 0 & 0 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \end{bmatrix}.$$

We set $\zeta = 20$ and assume that all parameters in Theorem 4.4.1 and Theorem 4.4.3 are known for simplicity. The number of communication steps is set to $T = 38$, which is computed based on the guidelines suggested in Section 4.5 such that the error due to network convergence in Theorem 4.4.3 is always less than 0.01 (using $\hat{\delta} = 0.001$). The communication is stopped when the smallest error bound between the one in Theorem 4.4.1 (using $\delta = 0.05$) and the one in Theorem 4.4.3 (using $\delta = 0.05, \hat{\delta} = 0.001$) is less than 0.5, which leads to $S = 1620$. We plot the average (over agents) local estimation error without communication $\|\hat{\Theta}_{i,t+1} - \Theta\|$, the average local estimation error after communication $\|\bar{\Theta}_{i,t+1} - \Theta\|$, and the global estimation error $\|\hat{\Theta}_{t+1} - \Theta\|$. All results are averaged over 10 independent runs.

As expected, the error $\|\bar{\Theta}_{i,t+1} - \Theta\|$ is almost the same as $\|\hat{\Theta}_{t+1} - \Theta\|$ when communication happens. Further, the error $\|\bar{\Theta}_{i,t+1} - \Theta\|$ decreases relatively rapidly, and is much smaller than $\|\hat{\Theta}_{i,t+1} - \Theta\|$ at the beginning. However, the error $\|\bar{\Theta}_{i,t+1} - \Theta\|$ decreases more slowly, and its improvement over $\|\hat{\Theta}_{i,t+1} - \Theta\|$ becomes less obvious, as each agent gathers more samples. Although the communication is stopped at $t = 1620$, leveraging the global dataset has only marginal improvements over the estimates $\hat{\Theta}_{i,t+1}, \bar{\Theta}_{i,t+1}$ after $t = 1620$, implying communication becomes less important, which confirms our observation in Theorem 4.4.3. On the other hand, although the minimum between $\|\hat{\Theta}_{i,t+1} - \Theta\|$ and $\|\bar{\Theta}_{i,t+1} - \Theta\|$ is less than 0.5 after $t = 1620$, the simulation also implies that our finite time bound is conservative. It is of interest to develop tighter bounds in future work.

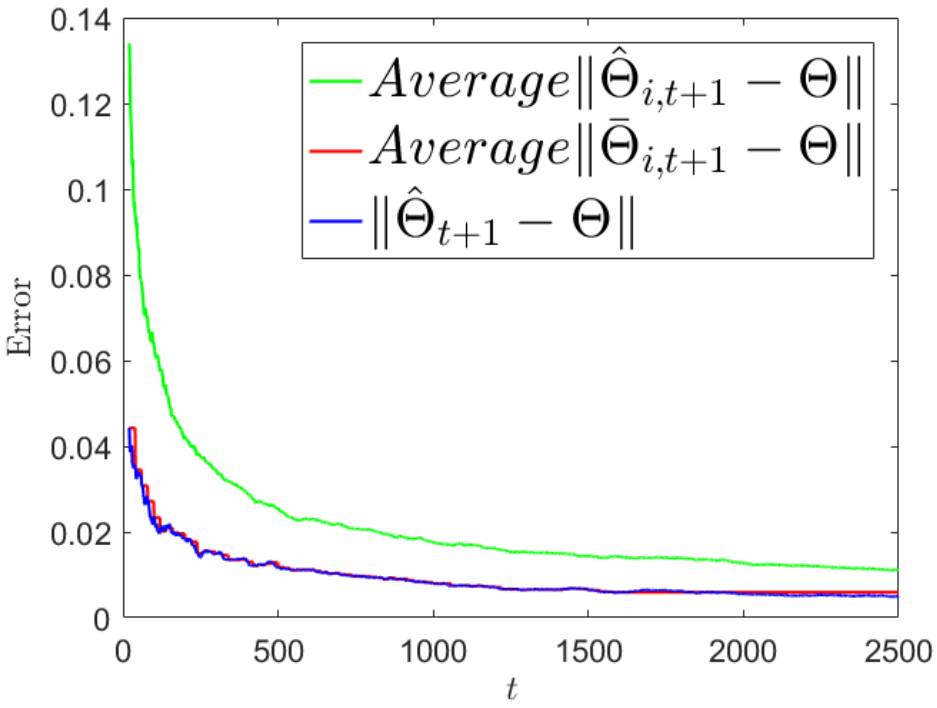

Figure 4.1. Average $\|\hat{\Theta}_{i,t+1} - \Theta\|$, average $\|\bar{\Theta}_{i,t+1} - \Theta\|$, and $\|\hat{\Theta}_{t+1} - \Theta\|$. The communication is stopped after $t = 1620$.

4.7 Chapter Summary

In this chapter, we proposed an online distributed parameter estimation algorithm with finite time performance guarantees. Our results demonstrate a trade-off between estimation error and communication costs, and we show that one can leverage the error bounds to determine a time at which the communication can be stopped.

5. ONLINE CHANGE POINTS DETECTION FOR LINEAR DYNAMICAL SYSTEMS WITH FINITE SAMPLE GUARANTEES

5.1 Introduction

The problem of change point detection (CPD) is to detect when abrupt changes in properties of time series occur. This problem has wide application in various fields, including monitoring of medical conditions, speech recognition, environmental surveillance, and image analysis [64]. In an offline setting, the goal is to determine the change points by looking at the entire dataset all at once [65]. In contrast, in an online (or sequential) setting, the goal is to detect when changes occur as soon as possible based on new data points arriving in a streaming manner [66].

Online CPD under the assumption of independence of samples over time has been studied extensively over the past few decades [67]–[70], and many approaches have been proposed; for example, [71], [72] leveraged Bayesian methods for online CPD, while [73], [74] explored the use of neural networks. The CPD problem is more challenging when the data exhibits correlations over time, and coming up with theoretical guarantees for such settings is an active area of research [75]. Indeed, time-correlations between data are commonly observed in practice. For example, data generated by dynamical systems is inherently correlated over time; such systems commonly occur in control theory, machine learning and economics [76]. As a practical example, when learning system dynamics from observed data, one may want to know if the system changes dynamics at some point of time to avoid using biased data [23]. There are several papers that focus on CPD for dynamical systems [77]–[79]. However, the existing works typically do not provide finite sample guarantees for the probabilities of making false and true alarms, i.e., it is unclear how well these algorithms perform given a finite number of data points.

A common approach used for online CPD is to compare a data-based statistic against a given threshold, with an alarm being raised if the statistic is larger than the threshold. There are some papers that provide threshold values that can achieve desired probabilities of false

and true alarms, based on Bayesian approaches [80], [81]. However, these works typically require a known probability distribution of the data prior to the change or do not have a theoretical analysis that demonstrates the relationship between detection accuracy and detection delay. Indeed, a rigorous characterization of the relationship between detection accuracy and detection delay is rarely provided in the literature. The paper [82] studies change point detection leveraging multivariate singular spectrum analysis, where the authors model the time series data as being generated by a spatio-temporal model and theoretically demonstrate the trade-off between detection delay and false alarm rate. However, the main focus of their analysis is on how a user-specified threshold value used in their test affects the performance of the algorithm in expectation. In contrast, our focus in this chapter is on finite sample guarantees for the probabilities of making false and true alarms, and how the number of samples used in our algorithm at each time step affects the detection accuracy and detection delay.

Lastly, we note that all of the above mentioned works do not theoretically demonstrate how the presence of multiple change points affects the performance of the CPD algorithms. This aspect is captured in our algorithm and its associated analysis.

In summary, our contributions are as follows.

- We provide an online change point detection algorithm for linear dynamical systems that is suitable for multiple change points. The algorithm is based on a least squares approach and can be easily implemented. We develop a data-dependent threshold that can be used in our test, which enables the user to achieve a pre-specified false alarm probability (assuming certain prior knowledge about the system is available). We note that the threshold does not require perfect knowledge of the system parameters at any given time. This is different from many existing works on CPD, e.g., [80], [83], [84], which usually require known distribution prior to the change (when there is only a single change point).

- We provide a finite-sample-based lower bound for the probability of making a true alarm after changes occur with a certain delay. Our result demonstrates the trade-off

between accurate detection and detection delay, as well as the ability to detect fast changes.

Our theoretical analysis and guarantees are demonstrated and validated via numerical examples at the end of this chapter.

5.2 Problem formulation and algorithm

Consider the discrete time linear dynamical system

$$x_{k+1} = A_k x_k + B_k u_k + w_k, \tag{5.1}$$

where $k \in \mathbb{N}$ is the time index, $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^p$ is the input, $w_k \in \mathbb{R}^n$ is the process noise, $A_k \in \mathbb{R}^{n\times n}$ and $B_k \in \mathbb{R}^{n\times p}$ are system matrices. The input and process noise are assumed to be i.i.d Gaussian, with $u_k \sim \mathcal{N}(0, \sigma_u^2 I_p)$ and $w_k \sim \mathcal{N}(0, \sigma_w^2 I_n)$. We define $\sigma_{min} \triangleq \min(\sigma_w, \sigma_u)$. The initial state x_0 is assumed to be independent of u_k and w_k. The system matrices A_k, B_k are deterministic but unknown. Let $\Theta_k = \begin{bmatrix} A_k & B_k \end{bmatrix}$. We further assume that there are known parameters b_{σ_w} and b_Θ that satisfy $\sigma_w \leq b_{\sigma_w}$ and $\|\Theta_k\| \leq b_\Theta$ for all $k \geq 0$. We call a time index $\hat{k} \in \mathbb{N}_{\geq 1}$ a *change point* if $\Delta_{\hat{k}} \triangleq \|\Theta_{\hat{k}-1} - \Theta_{\hat{k}}\| > 0$. Our goal is to determine the change points using observed data from system (5.1) in an online fashion. From the perspective of control theory, one can treat the above model as a switched system with unknown dynamics [85], where the goal is to detect when the system switches dynamics and learn a model for each mode of the system, so that one can design a better controller; one can also treat A_k in system (5.1) as a closed loop system under control, and the input u_k as a small exploratory input (which is a common strategy used in the adaptive control literature [5]).

Let $N \geq 2$ be a design parameter that corresponds to the length of the interval of previously seen data we would like to use at each step to detect change points (our analysis later will provide guidance on how to select N). Let $\{(x_i, u_j) : 0 \leq i \leq 2N-1, 0 \leq j \leq 2N-2\}$ denote the initial dataset. At each time step $k \geq 2N-1$, a sample (x_{k+1}, u_k)

generated by system (5.1) is observed. Let the label matrices for the *reference window* and the *test window* at time step k be

$$X_k^{ref} = \begin{bmatrix} x_{k-2N+3} & \cdots & x_{k-N+1} \end{bmatrix} \in \mathbb{R}^{n\times(N-1)}, \tag{5.2}$$

and

$$X_k^{test} = \begin{bmatrix} x_{k-N+3} & \cdots & x_{k+1} \end{bmatrix} \in \mathbb{R}^{n\times(N-1)}, \tag{5.3}$$

respectively. Let $z_k = \begin{bmatrix} x_k' & u_k' \end{bmatrix}' \in \mathbb{R}^{n+p}$ for $k \geq 0$, and define the regressor matrices for the reference window and the test window at time step k as

$$\begin{aligned} Z_k^{ref} &= \begin{bmatrix} z_{k-2N+2} & \cdots & z_{k-N} \end{bmatrix} \in \mathbb{R}^{(n+p)\times(N-1)}, \\ Z_k^{test} &= \begin{bmatrix} z_{k-N+2} & \cdots & z_k \end{bmatrix} \in \mathbb{R}^{(n+p)\times(N-1)}. \end{aligned} \tag{5.4}$$

Furthermore, we denote the noise matrices for the reference window and the test window at time step k as

$$\begin{aligned} W_k^{ref} &= \begin{bmatrix} w_{k-2N+2} & \cdots & w_{k-N} \end{bmatrix} \in \mathbb{R}^{n\times(N-1)}, \\ W_k^{test} &= \begin{bmatrix} w_{k-N+2} & \cdots & w_k \end{bmatrix} \in \mathbb{R}^{n\times(N-1)}. \end{aligned} \tag{5.5}$$

Based on these data windows, we consider the following intuitive approach. Let $\hat{\Theta}_k^{ref}$ and $\hat{\Theta}_k^{test}$ be the estimated system models using the data from the reference window and the test window, respectively. If the data from the reference window and the test window are generated by the same dynamics, then their estimated models should be similar; if the data from these windows are generated by different dynamics, i.e., if there is a change point, their estimated models should be quite different, i.e., one may flag a change point if the metric $\|\hat{\Theta}_k^{ref} - \hat{\Theta}_k^{test}\|$ is larger than some user-specified threshold value. To find $\hat{\Theta}_k^{ref}$ and $\hat{\Theta}_k^{test}$, we will solve the following regularized least squares problems at each time step k

$$\min_{\tilde{\Theta}_k^{ref}\in\mathbb{R}^{n\times(n+p)}} \{\|X_k^{ref} - \tilde{\Theta}_k^{ref} Z_k^{ref}\|_F^2 + \lambda\|\tilde{\Theta}_k^{ref}\|_F^2\},$$

and

$$\min_{\tilde{\Theta}_k^{test} \in \mathbb{R}^{n \times (n+p)}} \{\|X_k^{test} - \tilde{\Theta}_k^{test} Z_k^{test}\|_F^2 + \lambda \|\tilde{\Theta}_k^{test}\|_F^2\},$$

where $\lambda > 0$ is a regularization parameter. The closed-form solutions of the above problems are given by

$$\hat{\Theta}_k^{ref} = X_k^{ref}(Z_k^{ref})'(Z_k^{ref}(Z_k^{ref})' + \lambda I_{n+p})^{-1}, \tag{5.6}$$

and

$$\hat{\Theta}_k^{test} = X_k^{test}(Z_k^{test})'(Z_k^{test}(Z_k^{test})' + \lambda I_{n+p})^{-1}, \tag{5.7}$$

respectively. In this chapter, we will develop a data-dependent threshold $\gamma_k = \gamma_k(\delta)$ that has provable finite-sample guarantees, where $\delta \in (0, 1)$ is a user-specified upper bound of the false alarm probability. The specific expression of γ_k will be given later when we present our results (Lemma 20).

Our guarantees will apply to change points that are sufficiently separated (in time) from other change points. We make the following definition to make this formal.

Definition 5.2.1 (Sufficiently Separated Change Point)**.** *Suppose the system has $q \in \mathbb{N}_{\geq 1}$ change points. Let the sequence of change points be $0 < k_1 < \cdots < k_q$. We call a change point k_i, $1 \leq i \leq q-1$, sufficiently separated if $k_i - k_{i-1} \geq 4N - 1$ and $k_{i+1} - k_i \geq 4N - 1$, where $k_0 = 0$. The change point k_q is sufficiently separated if $k_q - k_{q-1} \geq 4N - 1$.*

If the system has infinitely many change points $0 < k_1 < k_2 < \cdots$, we call a change point k_i, $i \geq 1$, sufficiently separated if $k_i - k_{i-1} \geq 4N - 1$ and $k_{i+1} - k_i \geq 4N - 1$, where $k_0 = 0$.

We let the value S_k represent the most recent change point predicted by the algorithm at time step $k \geq 2N - 1$, where the initialization is given by $S_{2N-2} = 0$. If the current metric $\|\hat{\Theta}_k^{ref} - \hat{\Theta}_k^{test}\|$ is greater than the threshold γ_k, and $k - S_{k-1} > 2N - 2$, then the algorithm will predict a change point and set $S_k = k$; otherwise, no change point will be predicted and the algorithm will set $S_k = S_{k-1}$. The requirement of $k - S_{k-1} > 2N - 2$ is needed in addition to $\|\hat{\Theta}_k^{ref} - \hat{\Theta}_k^{test}\| \geq \gamma_k$ to deal with potentially multiple change points. Intuitively, if the metric is larger than the threshold for some consecutive time steps, we may not want to flag all of them as change points if all of the change points are sufficiently separated, so we wait for a period of time to make sure that the current dynamics have settled before making

the next prediction. In general, if there is no change point at time step k, i.e., $\|\Delta_k\| = 0$, we want the algorithm to output $S_k \neq k$; if there is a change point at time step k, we would like the algorithm to detect that as soon as possible, i.e., if $\|\Delta_k\| > 0$, we want the algorithm to output $S_t \geq k$ for some $t \geq k$, and the value $\min_{S_t \geq k}\{S_t\}$ should be small.

The above steps are encapsulated in Algorithm 5.

Algorithm 5 Online Change Point Detection

Input False alarm probability $\delta \in (0,1)$, window size $N \geq 2$, parameters $\lambda > 0, b_{\sigma_w}, b_\Theta$

Initialize $S_{2N-2} = 0$
for $k = 2N-1, 2N, 2N+1, \ldots$ **do**
    Gather the sample (x_{k+1}, u_k)
    Compute $\hat{\Theta}_k^{ref}$ and $\hat{\Theta}_k^{test}$ as in (5.6) and (5.7), respectively
    Compute γ_k as in (5.10)
    if $\|\hat{\Theta}_k^{ref} - \hat{\Theta}_k^{test}\| \geq \gamma_k$ and $k - S_{k-1} > 2N-2$ **then**
        Set $S_k = k$
        Flag k as a change point
    else
        Set $S_k = S_{k-1}$
    end if
end for

Remark 15. *One can update the least squares solution using the Sherman-Morrison formula [60], which provides an efficient way to update the matrix inverse if the changes are 'small' (in our case, the change at each time step is the replacement of the oldest sample by the most recent one). The threshold γ_k will be provided in the next section, and depends on the parameters b_{σ_w} and b_Θ (which depend on prior knowledge of the system, or can be estimated in practice). If those parameters are not available, one can replace γ_k with any other positive value, but at the cost of losing performance guarantees. In general, a smaller threshold would lead to a higher probability of both false and true alarms.*

In the next section, we will present our main results showing that Algorithm 5 ensures the false alarm probability will be less than δ. We further provide a finite-sample-based lower bound on the true alarm probability, which demonstrates the trade-off between detection accuracy and detection delay.

5.3 Finite Sample Analysis of Algorithm 5

In this section, we present theoretical guarantees for Algorithm 5. Some of the proofs are included in section 5.6. In Section 5.3.1, we present some intermediate results that are used later. Our main results (the finite-sample bounds on the probabilities of making false and true alarms) are presented in Section 5.3.2. We first make the following assumption on system (5.1).

Assumption 6. *There exists a constant $\beta > 0$ such that $\mathrm{tr}(\mathbf{E}[x_k x_k']) \leq \beta$ for all $k \geq 0$.*

If we have $\rho(A_k) < 1$ for all $k \geq 0$, i.e., all potential systems are strictly stable, Assumption 6 simply requires the dynamics to not change too frequently [86] (as a sufficient condition) or that there are finitely many change points.

5.3.1 Intermediate results

The lemma below provides a value for the threshold γ_k, along with an associated probability bound on the recovered system matrices $\hat{\Theta}_k^{ref}$ and $\hat{\Theta}_k^{test}$.

Lemma 20. *Consider any time step $k^* \geq 2N-1$. Let $\bar{V}_{k^*}^{ref} = (Z_{k^*}^{ref}(Z_{k^*}^{ref})' + V)V^{-1}$ and $\bar{V}_{k^*}^{test} = (Z_{k^*}^{test}(Z_{k^*}^{test})' + V)V^{-1}$, where $V = \lambda I_{n+p}$. Suppose*

$$X_{k^*}^{ref} = \Theta_{k^*-2N+2} Z_{k^*}^{ref} + W_{k^*}^{ref}, \tag{5.8}$$

and

$$X_{k^*}^{test} = \Theta_{k^*-N+2} Z_{k^*}^{test} + W_{k^*}^{test}, \tag{5.9}$$

i.e., the system matrices at the start of the reference interval and the test interval are the same throughout the interval, and that the threshold is chosen as

$$\begin{aligned}\gamma_{k^*} = &\frac{b_{\sigma_w}\sqrt{\frac{32}{9}(\log\frac{9^n 2}{\delta} + \frac{1}{2}\log\det(\bar{V}_{k^*}^{ref}))}}{\sqrt{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}} \\ &+ \frac{\lambda b_\Theta}{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})} \\ &+ \frac{b_{\sigma_w}\sqrt{\frac{32}{9}(\log\frac{9^n 2}{\delta} + \frac{1}{2}\log\det(\bar{V}_{k^*}^{test}))}}{\sqrt{\lambda_{min}(Z_{k^*}^{test}(Z_{k^*}^{test})' + \lambda I_{n+p})}} \\ &+ \frac{\lambda b_\Theta}{\lambda_{min}(Z_{k^*}^{test}(Z_{k^*}^{test})' + \lambda I_{n+p})}.\end{aligned} \tag{5.10}$$

Then we have

$$P(\|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*-2N+2}\| + \|\hat{\Theta}_{k^*}^{test} - \Theta_{k^*-N+2}\| \geq \gamma_{k^*}) \leq \delta.$$

Proof. We will focus on bounding the term $\|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*-2N+2}\|$ as the analysis for the term $\|\hat{\Theta}_{k^*}^{test} - \Theta_{k^*-N+2}\|$ is almost the same. Recalling the expression of $\hat{\Theta}_{k^*}^{ref}$ in (5.6), we have

$$\begin{aligned}&\|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*-2N+2}\| \\ &= \|W_{k^*}^{ref}(Z_{k^*}^{ref})'(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1} - \lambda\Theta_{k^*-2N+2}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1}\| \\ &\leq \|W_{k^*}^{ref}(Z_{k^*}^{ref})'(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1}\| + \|\lambda\Theta_{k^*-2N+2}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1}\| \\ &\leq \|W_{k^*}^{ref}(Z_{k^*}^{ref})'(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1}\| + \frac{\lambda b_\Theta}{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}.\end{aligned} \tag{5.11}$$

For the first term after the last inequality of (5.11), we have

$$\begin{aligned}&\|W_{k^*}^{ref}(Z_{k^*}^{ref})'(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1}\| \\ &\leq \|W_{k^*}^{ref}(Z_{k^*}^{ref})'(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1/2}\|\|(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1/2}\| \\ &= \frac{\|W_{k^*}^{ref}(Z_{k^*}^{ref})'(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})^{-1/2}\|}{\sqrt{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}} \\ &= \frac{\|(\lambda I_{n+p} + \sum_{t=k^*-2N+2}^{k^*-N} z_t z_t')^{-1/2}(\sum_{t=k^*-2N+2}^{k^*-N} z_t w_t')\|}{\sqrt{\lambda_{min}(Z_k^{ref}(Z_k^{ref})' + \lambda I_{n+p})}}.\end{aligned} \tag{5.12}$$

Now we will bound the numerator of the last equality of (5.12) using Lemma 5. Define the sequence pairs $\{\bar{z}_t\}_{t\geq 1}$ and $\{\bar{w}_t\}_{t\geq 1}$, where $\bar{z}_t = z_{k^*-2N+1+t}$ and $\bar{w}_t = w_{k^*-2N+1+t}$. Then we have $\|(\lambda I_{n+p} + \sum_{t=k^*-2N+2}^{k^*-N} z_t z_t')^{-1/2}(\sum_{t=k^*-2N+2}^{k^*-N} z_t w_t')\| = \|(\lambda I_{n+p} + \sum_{t=1}^{N-1} \bar{z}_t \bar{z}_t')^{-1/2}(\sum_{t=1}^{N-1} \bar{z}_t \bar{w}_t')\|$. Further, define the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t = \sigma(\{\bar{z}_i\}_{i=1}^{1+t} \cup \{\bar{w}_j\}_{j=1}^{t})$. With these definitions, we have the noise terms $\bar{w}_t$ are $\mathcal{F}_t$-measurable, and $\bar{w}_t|\mathcal{F}_{t-1}$ are sub-Gaussian with parameter σ_w^2 for all $t \geq 1$. Consequently, fixing $\delta > 0$, we can apply Lemma 5 to obtain with probability at least $1 - \frac{\delta}{2}$

$$\frac{\|(\lambda I_{n+p} + \sum_{t=k^*-2N+2}^{k^*-N} z_t z_t')^{-1/2}(\sum_{t=k^*-2N+2}^{k^*-N} z_t w_t')\|}{\sqrt{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}} \leq \frac{\sqrt{\frac{32}{9}\sigma_w^2(\log\frac{9^n 2}{\delta} + \frac{1}{2}\log\det(\bar{V}_{k^*}^{ref})}}{\sqrt{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}}. \tag{5.13}$$

Combining the above inequality with (5.11), using $\sigma_w \leq b_{\sigma_w}$, we have with probability at least $1 - \frac{\delta}{2}$

$$\|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*-2N+2}\| \leq \frac{b_{\sigma_w}\sqrt{\frac{32}{9}(\log\frac{9^n 2}{\delta} + \frac{1}{2}\log\det(\bar{V}_{k^*}^{ref}))}}{\sqrt{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}} + \frac{\lambda b_\Theta}{\lambda_{min}(Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p})}. \tag{5.14}$$

Following a similar procedure for the term $\|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*-N+2}\|$, and applying a union bound, we have the desired result. □

The following lemma bounds the probability of the metric $\|\hat{\Theta}_k^{ref} - \hat{\Theta}_k^{test}\|$ being larger than the threshold γ_k if there is not a change point in both the reference interval and the test interval.

Lemma 21. *Consider any time step $k^* \geq 2N-1$. If it holds that $\Theta_{k^*} = \Theta_{k^*-1} = \ldots = \Theta_{k^*-2N+1}$, then we have*

$$P(\|\hat{\Theta}_{k^*}^{ref} - \hat{\Theta}_{k^*}^{test}\| \geq \gamma_{k^*}) \leq \delta.$$

Proof. Since $\Theta_{k^*} = \Theta_{k^*-1} = \ldots = \Theta_{k^*-2N+1}$, we have

$$\begin{aligned}
\|\hat{\Theta}_{k^*}^{ref} - \hat{\Theta}_{k^*}^{test}\| &= \|(\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*}) - (\hat{\Theta}_{k^*}^{test} - \Theta_{k^*})\| \\
&\leq \|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*}\| + \|\hat{\Theta}_{k^*}^{test} - \Theta_{k^*}\| \\
&= \|\hat{\Theta}_{k^*}^{ref} - \Theta_{k^*-2N+2}\| + \|\hat{\Theta}_{k^*}^{test} - \Theta_{k^*-N+2}\|.
\end{aligned}$$

Further, note that we also have

$$X_{k^*}^{ref} = \Theta_{k^*-2N+2} Z_{k^*}^{ref} + W_{k^*}^{ref},$$

$$X_{k^*}^{test} = \Theta_{k^*-N+2} Z_{k^*}^{test} + W_{k^*}^{test}.$$

Applying Lemma 20, we get the desired result. □

The following two lemmas are used to establish our finite-sample bound on the true alarm probability, with proofs provided in section 5.6.

Lemma 22. *Let k^* be a time step such that $k^* \geq 2N-1$. For any fixed $\bar{\delta} > 0$, let $N \geq \max(42, 200(n+p)\log(\frac{13}{\bar{\delta}}))$. Then with probability at least $1-2\bar{\delta}$, the following inequalities hold simultaneously:*

$$Z_{k^*}^{ref}(Z_{k^*}^{ref})' + \lambda I_{n+p} \succeq \frac{N\sigma_{min}^2 + 42\lambda}{42} I_{n+p},$$

$$Z_{k^*}^{test}(Z_{k^*}^{test})' + \lambda I_{n+p} \succeq \frac{N\sigma_{min}^2 + 42\lambda}{42} I_{n+p}.$$

Lemma 23. *Let k^* be a time step such that $k^* \geq 2N-1$. For any fixed $\bar{\delta} > 0$, with probability at least $1-2\bar{\delta}$, the following inequalities hold simultaneously:*

$$\|Z_{k^*+N-2}^{ref}(Z_{k^*+N-2}^{ref})'\| \leq \frac{C_1}{\bar{\delta}},$$

$$\|Z_{k^*+N-2}^{test}(Z_{k^*+N-2}^{test})'\| \leq \frac{C_1}{\bar{\delta}},$$

where $C_1 = (N-1)(\beta + \sigma_u^2 p)$.

The following result bounds the probability of the threshold value γ_k being small compared to the magnitude of the change, and will be used later to lower bound the probability of making an accurate detection.

Lemma 24. *Let $k^* \geq 2N-1$ be a change point. Let $N_1 = 200(n+p)(\log(\frac{7\lambda}{C_1}) + \|\Delta_{k^*}\| \sqrt{\frac{N\sigma_{min}^2+42\lambda}{2500b_{\sigma_w^2}^2(n+p)}} - \frac{168\lambda b_\Theta}{\sqrt{2500b_{\sigma_w^2}(n+p)(N\sigma_{min}^2+42\lambda)}})$. Consider any N satisfying $N \geq \max(42, N_1, \frac{336\lambda b_\Theta}{\|\Delta_{k^*}\|\sigma_{min}^2} - \frac{42\lambda}{\sigma_{min}^2})$. Further let $\lambda \leq \frac{4C_1}{e\delta_e}$, where*

$$\delta_e = \frac{8C_1}{\lambda \exp(\|\Delta_{k^*}\|\sqrt{\frac{N\sigma_{min}^2+42\lambda}{10000b_{\sigma_w}^2(n+p)}} - \sqrt{\frac{\log(\frac{9^n 2}{\delta})}{n+p}})}.$$

Then we have

$$P(\{2\gamma_{k^*+N-2} \leq \|\Delta_{k^*}\|\}) \geq 1-\delta_e,$$

where γ_{k^*+N-2} is defined in Lemma 20, and C_1 is defined in Lemma 23.

Proof. Fix $\bar{\delta} > 0$. From Lemma 22, when $N \geq \max(42, 200(n+p)\log(\frac{13}{\bar{\delta}}))$, we have with probability at least $1-2\bar{\delta}$ the following inequalities

$$\lambda_{min}(Z_{k^*+N-2}^{ref}(Z_{k^*+N-2}^{ref})' + \lambda I_{n+p}) \geq \frac{N\sigma_{min}^2+42\lambda}{42}, \tag{5.15}$$

$$\lambda_{min}(Z_{k^*+N-2}^{test}(Z_{k^*+N-2}^{test})' + \lambda I_{n+p}) \geq \frac{N\sigma_{min}^2+42\lambda}{42}. \tag{5.16}$$

From Lemma 23, letting $\lambda \leq \frac{C_1}{e\bar{\delta}}$, we have with probability at least $1-2\bar{\delta}$

$$\begin{aligned} \det(\bar{V}_{k^*+N-2}^{ref}) &= \frac{\det(Z_{k^*+N-2}^{ref}(Z_{k^*+N-2}^{ref})' + V)}{\det(V)} \\ &\leq \frac{(\|Z_{k^*+N-2}^{ref}(Z_{k^*+N-2}^{ref})'\| + \lambda)^{n+p}}{\lambda^{n+p}} \\ &\leq (\frac{C_1}{\bar{\delta}\lambda}+1)^{n+p} \leq (\frac{2C_1}{\bar{\delta}\lambda})^{n+p}, \end{aligned} \tag{5.17}$$

and

$$\det(\bar{V}_{k^*+N-2}^{test}) \leq (\frac{2C_1}{\bar{\delta}\lambda})^{n+p}, \tag{5.18}$$

where $\bar{V}^{ref}_{k^*+N-2}$ and $\bar{V}^{test}_{k^*+N-2}$ are defined in Lemma 20. Applying a union bound, we have the events in (5.15)-(5.18) occur simultaneously with probability at least $1-4\bar{\delta}$, which implies

$$2\gamma_{k^*+N-2} \leq \frac{50b_{\sigma_w}\sqrt{\log(\frac{9^n 2}{\delta})} + 50b_{\sigma_w}\log(\frac{2C_1}{\bar{\delta}\lambda})\sqrt{n+p}}{\sqrt{N\sigma^2_{min}+42\lambda}} + \frac{168\lambda b_\Theta}{N\sigma^2_{min}+42\lambda},$$

where we used the relationship that $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for positive a, b, and that $\lambda \leq \frac{C_1}{e\bar{\delta}} \Rightarrow \frac{2C_1}{\bar{\delta}\lambda} \geq e \Rightarrow \sqrt{\log(\frac{2C_1}{\bar{\delta}\lambda})} \leq \log(\frac{2C_1}{\bar{\delta}\lambda})$.

Setting the right hand side of the above inequality to $\|\Delta_{k^*}\|$, after some algebraic manipulations, we have

$$\bar{\delta} = \frac{2C_1}{\lambda \exp(\frac{c}{50b_{\sigma_w}\sqrt{n+p}})}, \tag{5.19}$$

where $c = \|\Delta_{k^*}\|\sqrt{N\sigma^2_{min}+42\lambda} - \frac{168\lambda b_\Theta}{\sqrt{N\sigma^2_{min}+42\lambda}} - 50b_{\sigma_w}\sqrt{\log(\frac{9^n 2}{\delta})}$. When $N \geq \frac{336\lambda b_\Theta}{\|\Delta_{k^*}\|\sigma^2_{min}} - \frac{42\lambda}{\sigma^2_{min}}$, we have $c \geq \frac{\|\Delta_{k^*}\|\sqrt{N\sigma^2_{min}+42\lambda}}{2} - 50b_{\sigma_w}\sqrt{\log(\frac{9^n 2}{\delta})}$, which implies $\bar{\delta} \leq \frac{\delta_e}{4}$. Finally, recall that we also required the conditions $N \geq \max(42, 200(n+p)\log(\frac{13}{\bar{\delta}}))$ and $\lambda \leq \frac{C_1}{e\bar{\delta}}$. Substituting (5.19) into these conditions, using $\bar{\delta} \leq \frac{\delta_e}{4}$, and after some simplifications, we have the desired result. □

Remark 16. *In Theorem 5.3.2, we will show that a smaller δ_e corresponds to a higher probability of true alarm. Note that N will be larger than N_1 for sufficiently large N since the term C_1 grows linearly fast.*

The following result lower bounds the probability of the metric $\|\hat{\Theta}^{ref}_k - \hat{\Theta}^{test}_k\|$ being larger than the threshold γ_k with some delay, when there is a change point.

Lemma 25. *Consider any time step k^* that is a sufficiently separated change point. Suppose the conditions in Lemma 24 are satisfied. Then we have*

$$P(\bigcup_{t=k^*}^{k^*+N-2}\{\|\hat{\Theta}^{ref}_t - \hat{\Theta}^{test}_t\| \geq \gamma_t\}) \geq 1-\delta-\delta_e,$$

where δ_e is defined in Lemma 24.

Proof. Note that we have

$$P(\bigcup_{t=k^*}^{k^*+N-2}\{\|\hat{\Theta}_t^{ref}-\hat{\Theta}_t^{test}\|\geq\gamma_t\})\geq P(\{\|\hat{\Theta}_{k^*+N-2}^{ref}-\hat{\Theta}_{k^*+N-2}^{test}\|\geq\gamma_{k^*+N-2}\}).$$

We will lower bound the probability after the above inequality now. We have

$$\begin{aligned}
&\|\hat{\Theta}_{k^*+N-2}^{ref}-\hat{\Theta}_{k^*+N-2}^{test}\| \\
&=\|(\Theta_{k^*-1}+\hat{\Theta}_{k^*+N-2}^{ref}-\Theta_{k^*-1})-(\Theta_{k^*}+\hat{\Theta}_{k^*+N-2}^{test}-\Theta_{k^*})\| \\
&=\|\Delta_{k^*}+(\hat{\Theta}_{k^*+N-2}^{ref}-\Theta_{k^*-1})-(\hat{\Theta}_{k^*+N-2}^{test}-\Theta_{k^*})\| \\
&\geq\|\Delta_{k^*}\|-\|\hat{\Theta}_{k^*+N-2}^{ref}-\Theta_{k^*-1}\|-\|\hat{\Theta}_{k^*+N-2}^{test}-\Theta_{k^*}\|,
\end{aligned} \tag{5.20}$$

where the last inequality is due to the triangle inequality. Since k^* is a sufficiently separated change point, we have $\Theta_{k^*-1}=\Theta_{k^*-2}=\ldots=\Theta_{k^*-4N+1}$ and $\Theta_{k^*}=\Theta_{k^*+1}=\ldots=\Theta_{k^*+4N-2}$. Hence, we have

$$\|\hat{\Theta}_{k^*+N-2}^{ref}-\Theta_{k^*-1}\|=\|\hat{\Theta}_{k^*+N-2}^{ref}-\Theta_{k^*-N}\|. \tag{5.21}$$

For the same reason, we have

$$X_{k^*+N-2}^{ref}=\Theta_{k^*-N}Z_{k^*+N-2}^{ref}+W_{k^*+N-2}^{ref}, \tag{5.22}$$

and

$$X_{k^*+N-2}^{test}=\Theta_{k^*}Z_{k^*+N-2}^{test}+W_{k^*+N-2}^{test}. \tag{5.23}$$

Applying Lemma 20, we can obtain that with probability at least $1-\delta$

$$\|\hat{\Theta}_{k^*+N-2}^{ref}-\Theta_{k^*-N}\|+\|\hat{\Theta}_{k^*+N-2}^{test}-\Theta_{k^*}\|\leq\gamma_{k^*+N-2}, \tag{5.24}$$

which implies that

$$\|\hat{\Theta}_{k^*+N-2}^{ref}-\hat{\Theta}_{k^*+N-2}^{test}\|\geq\|\Delta_{k^*}\|-\gamma_{k^*+N-2} \tag{5.25}$$

with probability at least $1-\delta$ from (5.20) and (5.21). Furthermore, applying Lemma 24 and a union bound, we have with probability at least $1-\delta-\delta_e$

$$\begin{aligned}\|\hat{\Theta}^{ref}_{k^*+N-2} - \hat{\Theta}^{test}_{k^*+N-2}\| &\geq \|\Delta_{k^*}\| - \gamma_{k^*+N-2} \\ &\geq \gamma_{k^*+N-2}.\end{aligned}$$

□

5.3.2 Main results: Finite-sample probability bounds on making false and true alarms

Now, we state our first main theorem, which shows that the the probability of false alarm is upper bounded by δ (recall that setting $S_k = k$ is always associated with flagging time step k as a change point).

Theorem 5.3.1 (Probability of False Alarm)**.** *Consider any time step $k^* \geq 2N-1$. If it holds that $\Theta_{k^*} = \Theta_{k^*-1} = \ldots = \Theta_{k^*-2N+1}$, we have*

$$P(S_{k^*} = k^*) \leq \delta.$$

Proof. From Algorithm 5, we have $P(S_{k^*} = k^*) \leq P(\|\hat{\Theta}^{ref}_{k^*} - \hat{\Theta}^{test}_{k^*}\| \geq \gamma_{k^*})$. Applying Lemma 21, we have $P(S_{k^*} = k^*) \leq \delta$. □

Remark 17. *Note that δ is a user specified parameter that can be arbitrarily small. However, a small δ could reduce the true alarm probability, which will be discussed when we present Theorem 5.3.2. We also note that although we assumed that both the input u_k and the noise w_k are i.i.d Gaussian, this result actually holds for independent sub-Gaussian noise and arbitrary input that is independent of future noise. However, our finite-sample lower bound of the true alarm probability, i.e., if there is a change point, requires the assumption of i.i.d Gaussian input and i.i.d Gaussian noise. We leave the analysis of more general distributions of input and noise to future work.*

Next, we state our result that lower bounds the probability of making a true alarm (with some delay).

Theorem 5.3.2 (Probability of True Alarm)**.** *Consider any time step k^* that is a sufficiently separated change point. Suppose the conditions on N, λ in Lemma 24 are also satisfied. Recall that $C_1 = (N-1)(\beta + \sigma_u^2 p)$, and*

$$\delta_e = \frac{8C_1}{\lambda \exp(\|\Delta_{k^*}\|\sqrt{\frac{N\sigma_{min}^2 + 42\lambda}{10000 b_{\sigma_w}^2 (n+p)}} - \sqrt{\frac{\log(\frac{9n^2}{\delta})}{n+p}})}.$$

Then we have

$$P(\bigcup_{t=k^*}^{k^*+N-2} \{S_t = t\}) \geq 1 - (2N+1)\delta - \delta_e,$$

where the term $(2N+1)\delta$ captures the uncertainty due to settling from the previous change point, and the term δ_e captures the uncertainty due to noise.

Proof. Define the events $E_1 \triangleq \bigcup_{t=k^*}^{k^*+N-2} \{\|\hat{\Theta}_t^{ref} - \hat{\Theta}_t^{test}\| \geq \gamma_t\}$ and $E_2 \triangleq \bigcap_{t=k^*-2N}^{k^*-1} \{\|\hat{\Theta}_t^{ref} - \hat{\Theta}_t^{test}\| \leq \gamma_t\}$. We have $P(E_1) \geq 1 - \delta - \delta_e$ by applying Lemma 25. Since k^* is a sufficiently separated change point, we also have $\Theta_{k^*-1} = \Theta_{k^*-2} = \ldots = \Theta_{k^*-4N+1}$. Hence we can apply Lemma 21 and combine the $2N$ events in E_2 using a union bound to obtain $P(E_2) \geq 1 - 2N\delta$. Recall the conditions for setting $S_t = t$ in Algorithm 5. Conditioning on $E_1 \cap E_2$, let $\hat{t}$ be the smallest time step such that $\|\hat{\Theta}_t^{ref} - \hat{\Theta}_t^{test}\| \geq \gamma_t$ for $t = k^*, \ldots, k^* + N - 2$. Note that on the event $E_1 \cap E_2$ we also have

$$\begin{aligned} \hat{t} - S_{\hat{t}-1} &\geq k^* - S_{k^*-2N-1} \geq k^* - (k^* - 2N - 1) \\ &> 2N - 2, \end{aligned}$$

with probability 1, which implies $S_{\hat{t}} = \hat{t}$. Consequently, combining events E_1 and E_2 using a union bound, we have

$$P(\bigcup_{t=k^*}^{k^*+N-2} \{S_t = t\}) \geq P(E_1 \cap E_2) \geq 1 - (2N+1)\delta - \delta_e.$$

□

Remark 18. ***Interpretation of Theorem 5.3.2.*** *Recall that setting $S_t = t$ implies flagging time step t as a change point. Theorem 5.3.2 provides a high probability lower bound*

on true detection in the long run (i.e., predict the change point after k^*, but within any pre-specified large time steps), where the associated delay is bounded by $N-2$. Note that the term $-2N\delta$ is added to the probability lower bound in Lemma 25. This term is included due to the need to deal with potentially multiple change points, since the algorithm may conclude that the dynamics have settled by observing the event $E_2 = \bigcap_{t=k^*-2N}^{k^*-1}\{\|\hat{\Theta}_t^{ref} - \hat{\Theta}_t^{test}\| \leq \gamma_t\}$, i.e., we want to make sure that the reason the metric is greater than the threshold is that we have a new change point instead of that we are observing some residual effects from the previous change point. Below we further discuss the effects of some parameters.

***Discussions on the effects of** δ, N**:** *Suppose that the window size N is fixed for now. We see from Theorem 5.3.2 that a smaller false probability δ would lead to a smaller uncertainty due to the need to confirm the dynamics has settled from the previous change point. However, if δ is set to be too small, δ_e may become very large and make the overall lower bound small, which implies a potentially smaller probability of true detection in the long run. To make the true alarm probability lower bound large and maintain a small false alarm probability δ, we can use a larger window size N. More specifically, by setting $\delta = \frac{a}{\exp(\sqrt{N})}$ for some constant $a > 0$, we see that a larger N corresponds to a smaller false alarm probability, and ensures both a smaller uncertainty due to settling and a smaller uncertainty due to noise, since the term C_1 in the numerator of δ_e grows at most linearly with respect to N, and the denominator of δ_e grows exponentially fast with respect to the square root of N. The price, on the other hand, is that the guaranteed delay is larger due to a larger N. Consequently, this result demonstrates a trade-off between detection accuracy and detection delay, i.e., to maintain both a low false alarm probability and a high true alarm probability in the long run, one has to suffer from a potentially larger delay by using a larger N. Such a non-asymptotic characterization is different from existing approaches, where algorithm performance is typically measured in expectation, e.g., using Average Running Length [82]. Furthermore, as N becomes larger, there may be fewer change points that satisfy the sufficient separation condition, i.e., the speed of the change of the dynamics may appear to be too fast relative to the time interval we are monitoring. In other words, this result further implies that a smaller N is suitable for more frequent changes, but at the price of being less likely to predict the change accurately.*

Discussions on the effects of $\|\Delta_{k^*}\|$***:*** *Finally, note that the denominator of* δ_e *also grows exponentially fast with respect to the magnitude of change* $\|\Delta_{k^*}\|$ *(supposing that* N *is large enough). Hence, a larger change in dynamics could lead to a higher probability (lower bound) of detecting a change point (matching intuition).*

Theorem 5.3.1 and Theorem 5.3.2 cover all possible situations if all change points are sufficiently separated. More specifically, consider any time step $k^* \geq 2N-1$. If there is not a change point over the past $2N$ time steps, Theorem 5.3.1 ensures that there will not be a false alarm with high probability. If the current time step k^* is a change point, Theorem 5.3.2 ensures that Algorithm 5 will detect it, i.e., predict exactly one change point, within $N-1$ time steps, with high probability. Furthermore, the design of Algorithm 5 ensures that the Algorithm will not flag any change point at $k^*+N-1, k^*+N, \ldots, k^*+2N-2$, if a change point was predicted at $k^*, \ldots, k^*+N-2$ (note that the events $\{S_t = t\}_{t=k^*}^{k^*+2N-2}$ are mutually exclusive, since it is necessary to have $t - S_{t-1} > 2N-2$ to set $S_t = t$ from Algorithm 5). These properties imply that all time steps are covered by either Theorem 5.3.1 or Theorem 5.3.2.

5.4 Numerical experiment

In this section, we provide some numerical examples of the online change point detection algorithm. In general, a threshold that has non-asymptotic guarantees can be conservative in practice. However, as we will see, even if we use the exact theoretical threshold, Algorithm 5 can still achieve reasonably good performance, which indicates that the derived threshold is not overly conservative in practice (and it does not require perfect knowledge of the parameters at any time, unlike [80], [83], [84]).

The system we consider here is the linearized longitudinal dynamics of a UAV reported in [87], where we set the sampling rate to be 0.1 seconds using zero-order hold. The system has 5 states, representing inertial velocity components of the airframe projected onto a body

frame axis, the pitch angle, the pitch angular rate, and the altitude. The input is the elevator deflection. Assuming full state observation, the matrices A_k and B_k are given by

$$A_k = \begin{bmatrix} 0.9371 + \epsilon_k^A & 0.068 & -0.9507 & -0.0367 & 0 \\ -0.0085 & 0.2761 & -0.0207 & 0.411 & 0 \\ 0.0035 & -0.0164 & 0.9991 & 0.043 & 0 \\ 0.0548 & -0.1914 & -0.0253 & 0.0593 & 0 \\ -0.0086 & 0.0726 & -1.6984 & -0.0146 & 1 \end{bmatrix},$$

$$B_k = \begin{bmatrix} 0.361 + \epsilon_k^B & -4.8436 & -0.3888 & -5.6967 & 0.0492 \end{bmatrix}',$$

where ϵ_k^A and ϵ_k^B are perturbations we added to to the system, and we set $\epsilon_k^A = 0, -1, -1$ for $0 \leq k \leq 2499$, $2500 \leq k \leq 4999$, and $k \geq 5000$, respectively; $\epsilon_k^B = 0, 2, 0$ for $0 \leq k \leq 2499$, $2500 \leq k \leq 4999$, and $k \geq 5000$, respectively. In other words, $k_1 = 2500$ and $k_2 = 5000$ are two change points. We set $\lambda = \sigma_w = \sigma_u = 1$, and $x_0 = 0$. The parameters used in our threshold value γ_k are assumed to be tight bounds for simplicity, i.e., $b_\Theta = \max_{k \geq 0}(\|\Theta_k\|)$ and $b_{\sigma_w} = \sigma_w$. We performed experiments using $N = 50, 150, 250, 350, 450$, each with 10 independent runs, where the length of each experiment is set to 9000 steps. The bound on false alarm probability is set to $\delta = \frac{1000}{\exp\sqrt{N}}$ (as suggested in Remark 18).

In Figures 5.1-5.3, we plot the average $\|\hat{\Theta}_k^{ref} - \hat{\Theta}_k^{test}\|$ versus γ_k for $N = 50, 250, 450$. In Table 5.1, we report the performance of the CPD algorithm. Here, AD1 refers to the average detection time between $k = 2500$ and $k = 4999$, AD2 refers to the average detection time between $k = 5000$ and $k = 8999$, MD1 refers to the number of experiments where no detection was made between $k = 2500$ and $k = 4999$, and MD2 refers to the number of experiments where no detection was made between $k = 5000$ and $k = 8999$. For all experiments, none of them made a false alarm before $k = 2500$. Further, we can see that as N increases, the probability of true detection in the long run increases (captured by a

smaller misdetection rate). However, the corresponding delay also increases accordingly. These empirical observations are consistent with our theoretical findings in Theorem 5.3.2.

Table 5.1. Empirical Performance of Algorithm 5 Over 10 Independent Runs. The bound on false alarm probability is set to $\delta = \frac{1000}{\exp\sqrt{N}}$.

N	δ	AD1	MD1	AD2	MD2
50	0.85	2550	9	N/A	10
150	4.8×10^{-3}	2629.3	1	N/A	10
250	1.3×10^{-4}	2685.8	0	5218.9	1
350	7.5×10^{-6}	2701.2	0	5280.9	0
450	6.1×10^{-7}	2755	0	5321.8	0

5.5 Chapter Summary

In this chapter, we studied online change point detection for linear dynamical systems, where there are potentially multiple change points. Our analysis provides a data-dependent dynamic threshold that allows the user to specify a desired upper bound of the false alarm probability. We also provided a finite-sample lower bound on the probability of correctly identifying the change point with some delay. Our analysis demonstrates the trade-off between detection accuracy and detection delay, and characterizes how frequently changes can occur while still maintaining detection with a given probability. It is noted that our focus in this chapter is on fully-observed systems, i.e., all system states can be perfectly measured. It would be of interest to extend our analysis to partially-observed systems, where only a subset of system states can be measured. Other promising directions for future work would be to analyze different types of changes for dynamical systems, e.g., changes in noise distribution or changes in models that are possibly state-dependent, and to consider nonlinear system dynamics.

5.6 Proofs of Results

The following result is used to establish Lemma 22.

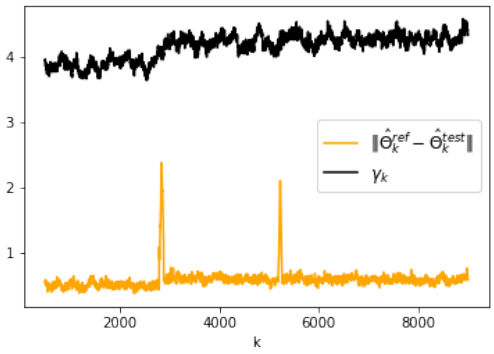

Figure 5.1. Online Change Point Detection with $N = 50$. The use of small N results in a threshold that is too high to flag change points, although we see spikes in the test statistics.

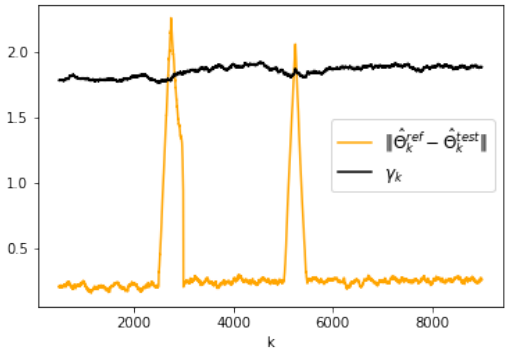

Figure 5.2. Online Change Point Detection with $N = 250$. The threshold successfully captures the two change points using a moderate N.

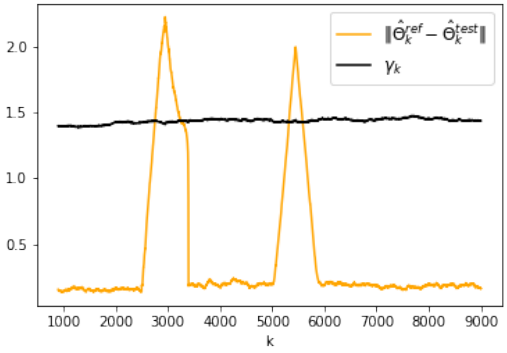

Figure 5.3. Online Change Point Detection with $N = 450$. The threshold successfully captures the two change points, but the use of a larger N incurs higher delay.

Lemma 26. *[88, Lemma 36] Let $\{z_t\}_{t\geq 0}$ be a sequence of random vectors that is adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $z_t \in \mathbb{R}^d$ for all $t \geq 0$. Suppose z_t is conditionally Gaussian on $\mathcal{F}_{t-1}$ with $\mathbb{E}[z_t z_t'|\mathcal{F}_{t-1}] \succeq \sigma_z^2 I_d$ for all $t \geq 1$, where $\sigma_z > 0$. Then, for any fixed $\delta > 0$ and any $T \geq 200d\log(\frac{12}{\delta})$, the following inequality holds with probability at least $1-\delta$:*

$$\sum_{t=0}^{T-1} z_t z_t' \succeq \frac{(T-1)\sigma_z^2}{40} I_d.$$

5.6.1 Proof of Lemma 22

Proof. We will only show the first inequality as the proof for the second one is almost the same. Note that

$$Z_{k^*}^{ref}(Z_{k^*}^{ref})' = \sum_{t=k^*-2N+2}^{k^*-N} z_t z_t'.$$

Rename the sequences $\bar{z}_t = z_{k^*-2N+2+t}$, $\bar{u}_t = u_{k^*-2N+2+t}$, $\bar{w}_t = w_{k^*-2N+2+t}$, and $\bar{\Theta}_t = \Theta_{k^*-2N+2+t}$ for $t \geq 0$. Define the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t = \sigma(\bar{z}_0, \bar{z}_1, \ldots, \bar{z}_t)$ for $t \geq 0$. Note that $z_t|\mathcal{F}_{t-1}$ is a Gaussian random vector for $t \geq 1$, since

$$\bar{z}_t|\mathcal{F}_{t-1} = \begin{bmatrix} \bar{\Theta}_{t-1}\bar{z}_{t-1}|\mathcal{F}_{t-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \bar{w}_{t-1}|\mathcal{F}_{t-1} \\ \bar{u}_t|\mathcal{F}_{t-1} \end{bmatrix}.$$

From the above equality, we also have

$$\mathbb{E}[\bar{z}_t\bar{z}_t'|\mathcal{F}_{t-1}] \succeq \begin{bmatrix} \sigma_w^2 I_n & 0 \\ 0 & \sigma_u^2 I_p \end{bmatrix} \succeq \sigma_{min}^2 I_{n+p}.$$

Consequently, after some algebraic manipulations, we can apply Lemma 26 to obtain with probability at least $1-\bar{\delta}$

$$\begin{aligned} Z_k^{ref}(Z_k^{ref})' + \lambda I_{n+p} &= \sum_{t=0}^{N-2} \bar{z}_t \bar{z}_t' + \lambda I_{n+p} \\ &\succeq (\frac{(N-2)\sigma_{min}^2}{40} + \lambda) I_{n+p} \\ &\succeq \frac{N\sigma_{min}^2 + 42\lambda}{42} I_{n+p}, \end{aligned}$$

when $N \geq \max(42, 200(n+p)\log(\frac{13}{\bar{\delta}}))$.

The result then follows by applying a union bound. □

5.6.2 Proof of Lemma 23

Proof. Recall that we have

$$\|Z_{k^*+N-2}^{ref}(Z_{k^*+N-2}^{ref})'\| = \|\sum_{t=k^*-N}^{k^*-2} z_t z_t'\|, \tag{5.26}$$

and

$$\|Z_{k^*+N-2}^{test}(Z_{k^*+N-2}^{test})'\| = \|\sum_{t=k^*}^{k^*+N-2} z_t z_t'\|. \tag{5.27}$$

We will only show the first inequality, as the second one is almost identical. From the Markov inequality, we have with probability at least $1-\bar{\delta}$

$$\|Z_{k^*+N-2}^{ref}(Z_{k^*+N-2}^{ref})'\| \leq \frac{\mathbf{E}[\|\sum_{t=k^*-N}^{k^*-2} z_t z_t'\|]}{\bar{\delta}}. \tag{5.28}$$

Now we bound the term $\mathbf{E}[\|\sum_{t=k^*-N}^{k^*-2} z_t z_t'\|]$. Since the term $z_t z_t'$ has unit rank, we have $\|z_t z_t'\| = \mathrm{tr}(z_t z_t')$, and hence

$$\mathbf{E}[\|\sum_{t=k^*-N}^{k^*-2} z_t z_t'\|] \leq \sum_{t=k^*-N}^{k^*-2} \mathbf{E}[\|z_t z_t'\|] = \sum_{t=k^*-N}^{k^*-2} \mathrm{tr}(\mathbf{E}[z_t z_t']). \tag{5.29}$$

Note that for all $t \geq 0$, from Assumption 6, we have

$$\mathrm{tr}(\mathbf{E}[z_t z_t']) = \mathrm{tr} \begin{bmatrix} \mathbf{E}[x_t x_t'] & 0 \\ 0 & \sigma_u^2 I_p \end{bmatrix} \leq \beta + \sigma_u^2 p. \tag{5.30}$$

Hence, we have with probability at least $1 - \bar{\delta}$

$$\| Z^{ref}_{k^*+N-2} (Z^{ref}_{k^*+N-2})' \| \leq \frac{(N-1)(\beta + \sigma_u^2 p)}{\bar{\delta}}. \tag{5.31}$$

Following a similar procedure for (5.27) and applying a union bound, we have the desired result. □

6. LEARNING LINEARIZED MODELS FROM NONLINEAR SYSTEMS WITH FINITE DATA

6.1 Introduction

[8] As indicated in above, system identification is an important problem in control theory since a good model can facilitate model-based control design [90]. Although physical systems are typically nonlinear, linear models are frequently used in practice due to their simplicity [91], and their ability to approximate nonlinear systems around a given reference point. Consequently, it is of interest to understand identification of appropriate linear models from data generated by nonlinear systems. On the other hand, we note that when it comes to finite sample analysis for linear system identification, almost all existing works that have finite sample guarantees (including those in the previous chapters of this thesis) assume that the underlying system is truly linear, except for [92]. Furthermore, independent Gaussian random inputs are typically applied to ensure persistent excitation.

The study on nonlinear system identification is less well-understood, in general, as compared to the case for linear system identification. Recent works on finite sample analysis for nonlinear system identification include [93]–[95]. It is worth noting that to obtain finite sample guarantees, the existing works on nonlinear system identification typically require that a certain model structure to be known in advance. However, when the specific model structure is unknown, a reasonable alternative goal is to learn a linearized model from the nonlinear system, due to the well-studied techniques on linear system control as discussed above.

There is a branch of recent research that focuses on learning a linear system representation that completely captures the behaviours of a nonlinear system using the Koopman Operator [96]. This approach typically requires carefully selected basis functions (e.g., leveraging neural networks [97]), and the analysis focuses on the noiseless setting. In contrast, our focus in this work is to learn a linearized system model, in the sense that the linear model captures the linear part of the nonlinear system after Taylor expansion, and to provide finite sample guarantees when the system has noise.

[8]↑The material in this chapter was published at the 2023 Conference on Decision and Control [89].

Most relevant to our work is the recent paper [92], which provides a finite sample error bound for learning linear models from systems that have unmodeled dynamics that could capture nonlinearities, using a single system trajectory. However, the method proposed in [92] assumes the system is "well-behaved" by requiring the unmodeled dynamics/nonlinear terms to be (globally) Lipschitz [98]. The method also requires the system to satisfy certain additional properties to ensure consistent estimation, supposing the inputs are carefully chosen. In contrast, we show in this work that one can learn a linearized system model from a nonlinear system with arbitrarily small error without the Lipschitzness assumption, given sufficiently many short trajectories, supposing that one has control over the initial conditions of the experiments.

In summary, our contributions in this chapter are as follows.

- We provide a deterministic, multiple trajectories-based data acquisition algorithm that ensures persistent excitation under the constraint of being close to the reference point. Using this algorithm followed by a regularized least squares estimation algorithm, we provide a finite sample error bound of the identified linearized dynamics of a nonlinear system.

- Our bound shows that one can learn the linearized dynamics with arbitrarily small error, given sufficiently many experiments in the multiple trajectories setup, and demonstrates a trade-off between the error due to noise and the error due to nonlinearity. The bound further characterizes the benefits of using regularization. When the system is perfectly linear, we show a learning rate that matches the existing results on learning perfectly linear systems using random inputs.

- We provide numerical experiments to validate our results and insights, and show the potential insufficiency of linear system identification using random inputs from a single trajectory when nonlinearity does exist.

This chapter is organized as follows. Section 6.2 introduces the system identification problem and the algorithms we use. In Section 6.3, we present our theoretical results. We

present numerical examples in in Section 6.4 to validate our results, and conclude in Section 6.5.

6.2 Problem Formulation and System Identification Algorithm

Consider the following discrete time nonlinear time invariant system

$$x_{k+1} = f(z_k) + w_k, \tag{6.1}$$

where $f : \mathbb{R}^{n+p} \to \mathbb{R}^n$, $z_k = \begin{bmatrix} x_k' & u_k' \end{bmatrix}' \in \mathbb{R}^{n+p}$, $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^p$, and $w_k \in \mathbb{R}^n$. Here, x_k, u_k and w_k are the state, input, and process noise, respectively. The noise terms w_k are assumed to be independent sub-Gaussian random vectors with parameter σ_w^2.

Note that sub-Gaussian distributions are commonly used to model noise processes [12]. In particular, every Gaussian random vector is sub-Gaussian.

Assume that for each component function of f, all second order partial derivatives exist and are continuous on $\mathbb{R}^{n+p}$. From Taylor's theorem [99], system (6.1) using reference point $z_k = \mathbf{0}$ can be rewritten as

$$x_{k+1} = Ax_k + Bu_k + o + w_k + r_k, \tag{6.2}$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times p}$, are system matrices that capture the linear part of $f(z_k)$, $o = f(\mathbf{0}) \in \mathbb{R}^n$, and $r_k = h(z_k) \in \mathbb{R}^n$ is a remainder vector that contains higher order terms that are state/input dependent, where $h : \mathbb{R}^{n+p} \to \mathbb{R}^n$. The above model is less studied in the literature on finite sample analysis for system identification, and we will consider this model in the sequel. Note that we assume o is possibly non-zero to capture scenarios where the equilibrium points of the system are unknown. When the system is perfectly linear, we have $o = r_k = \mathbf{0}$, which is the commonly used model in the literature. In this chapter, we assume that both the state x_k and input u_k can be perfectly measured. Suppose that we can restart the system multiple times from an arbitrary initial state x_0 using arbitrary input u_0, and obtain multiple length 1 trajectories (i.e., state-input pairs obtained by running the system for a single time step, as will be explained next). Using superscript to denote the

trajectory index, we denote the set of samples we have as $\{(x_1^i, x_0^i, u_0^i) : 1 \leq i \leq N\}$. Our goal is to learn the linear approximation system matrices $\Theta \triangleq \begin{bmatrix} A & B & o \end{bmatrix} \in \mathbb{R}^{n \times (n+p+1)}$ in system (6.2) from the set of samples available to us.

Our result will leverage the following mild assumption on the remainder vector $r_k = h(z_k)$ in system (6.2).

Assumption 7. *Let $r_{i,k}$ denote the i-th component of r_k. There exist $c > 0$ and $\beta = \beta(c)$ such that $|r_{i,k}| \leq \beta \|z_k\|_1^2$ for all $i \in \{1, \ldots, n\}$ and all $z_k \in \mathcal{B}_{n+p}(\boldsymbol{0}, c)$.*

Remark 19. *The above assumption is, in fact, a direct result of assuming that each component function of the original nonlinear dynamics f has all second order partial derivatives being continuous on $\mathbb{R}^{n+p}$, due to Taylor's theorem for multivariable functions from [100, Corollary 1]. Intuitively, this assumption says that the higher order terms are dominated by the second order terms, if the arguments of the function are sufficiently close to the origin. Note that it does not require the function h to be Lipschitz (which is the assumption used in [92]). As an example, consider a scalar system with the dynamics given by $f(z_k) = x_k + u_k + x_k^2 + x_k^3$. Here $r_k = x_k^2 + x_k^3$ satisfies Assumption 7 for $c = 1$ and $\beta = 2$ since $|x_k^2 + x_k^3| \leq |x_k^2| + |x_k^3| \leq 2|x_k|^2 \leq 2\|z_k\|_1^2$ for all $z_k \in \mathcal{B}_2(\boldsymbol{0}, 1)$, but the corresponding function h is not Lipschitz. In general, a larger c may lead to a larger β.*

Let $q > 0$ be a (small) design parameter that constrains the magnitude of the initial conditions z_0, and N be the number of experiments to perform. We deploy a data collection scheme specified in Algorithm 6.

Algorithm 6 Data Acquisition

Input Norm constraint parameter $q > 0$, number of experiments $N > 0$

Initialize $s_1 = 1$
for $i = 1, \ldots, N$ **do**
  if $i \bmod (n+p) \neq 0$ **then**
    Set $z_0^i = \begin{bmatrix} x_0^{i'} & u_0^{i'} \end{bmatrix}' = s_i \times q\mathbf{e}_{i \bmod (n+p)}^{n+p}$
    Collect x_1^i, where $x_1^i = Ax_0^i + Bu_0^i + w_0^i + o + r_0^i$
    Set $s_{i+1} = s_i$
  else
    Set $z_0^i = \begin{bmatrix} x_0^{i'} & u_0^{i'} \end{bmatrix}' = s_i \times q\mathbf{e}_{n+p}^{n+p}$
    Collect x_1^i, where $x_1^i = Ax_0^i + Bu_0^i + w_0^i + o + r_0^i$
    Set $s_{i+1} = -s_i$
  end if
end for
Output $\{(x_1^i, x_0^i, u_0^i) : 1 \leq i \leq N\}$

Remark 20. *Intuitively, we want the initial conditions to stay as close to the reference point (in this case, the origin) as possible, to avoid excessive bias from the higher order terms. Hence, the reason of using of multiple length 1 trajectories is to prevent the noise from driving the system too far from the reference point, and amplifying the effects from r_k. The key idea of Algorithm 6 is to ensure persistent excitation (i.e., the smallest eigenvalue of the sample covariance matrix becomes larger as one gets more data), subject to the constraint on bounded distance to the origin (specified by q). Later on in our theoretical result, we will demonstrate how q will affect the finite sample estimation error bound for learning Θ.*

We establish some definitions now. Define the batch matrices

$$\begin{aligned} X &= \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^N \end{bmatrix} \in \mathbb{R}^{n \times N} \\ W &= \begin{bmatrix} w_0^1 & w_0^2 & \cdots & w_0^N \end{bmatrix} \in \mathbb{R}^{n \times N} \\ R &= \begin{bmatrix} r_0^1 & r_0^2 & \cdots & r_0^N \end{bmatrix} \in \mathbb{R}^{n \times N}. \end{aligned} \tag{6.3}$$

Let $\hat{z}_0^i = \begin{bmatrix} z_0^{i'} & 1 \end{bmatrix}' \in \mathbb{R}^{n+p+1}$. Define the regressor matrix

$$Z = \begin{bmatrix} \hat{z}_0^1 & \hat{z}_0^2 & \cdots & \hat{z}_0^N \end{bmatrix} \in \mathbb{R}^{(n+p+1)\times N}. \tag{6.4}$$

We have the following relationship

$$X = \Theta Z + W + R. \tag{6.5}$$

To learn the linear model Θ, we would like to solve the following regularized least squares problem

$$\min_{\tilde{\Theta}\in\mathbb{R}^{n\times(n+p+1)}} \{\|X - \tilde{\Theta}Z\|_F^2 + \lambda\|\tilde{\Theta}\|_F^2\},$$

where $\lambda \geq 0$ is a regularization parameter. The closed-form solution of the above problem is given by

$$\hat{\Theta} = XZ'(ZZ' + \lambda I_{n+p+1})^{-1}, \tag{6.6}$$

under the invertibility assumption [24]. The estimation error is then given by

$$\begin{aligned} \|\hat{\Theta} - \Theta\| = \| &- \lambda\Theta(ZZ' + \lambda I_{n+p+1})^{-1} \\ &+ WZ'(ZZ' + \lambda I_{n+p+1})^{-1} \\ &+ RZ'(ZZ' + \lambda I_{n+p+1})^{-1}\|. \end{aligned} \tag{6.7}$$

For the ease of reference, the above steps are encapsulated in Algorithm 7.

Algorithm 7 System Identification Using Multiple Length 1 Trajectories

Input Dataset $\{(x_1^i, x_0^i, u_0^i) : 1 \leq i \leq N\}$, regularization parameter $\lambda \geq 0$

1: Construct the matrices X, Z. Compute $\hat{\Theta} = XZ'(ZZ' + \lambda I_{n+p+1})^{-1}$.

2: Extract the estimated system matrices A, B, o from the estimate $\hat{\Theta} = \begin{bmatrix} \hat{A} & \hat{B} & \hat{o} \end{bmatrix}$.

In the next section, we will provide a finite sample bound of the system identification error (6.7) using Algorithm 6 and Algorithm 7. The bound explicitly characterizes how the

error depends on N, q, σ_w, λ, and other system parameters, and will provide guidance on selecting q, λ.

6.3 Theoretical Analysis

To upper bound the system identification error in (6.7) with high probability, we bound the terms $\| - \lambda\Theta(ZZ' + \lambda I_{n+p+1})^{-1}\|$, $\|WZ'(ZZ' + \lambda I_{n+p+1})^{-1/2}\|$, $\|(ZZ' + \lambda I_{n+p+1})^{-1/2}\|$, $\|RZ'(ZZ' + \lambda I_{n+p+1})^{-1}\|$ separately. We provide some intermediate results first in Section 6.3.1. Our main result is presented in Section 6.3.2.

6.3.1 Intermediate results

The following result shows the persistent excitation property of Algorithm 6. Note that the requirement on $N \geq 4(n+p)$ below is mainly used for numerical simplification.

Lemma 27. *Suppose that Algorithm 6 is used to generate data. Let $N \geq 4(n+p)$. Then we have the following inequalities*

$$\lambda_{min}(ZZ') \geq N \min\{\frac{q^2}{2(n+p)}, \frac{1}{2}\},$$
$$\lambda_{max}(ZZ') \leq N \max\{\frac{2q^2}{n+p}, 2\}.$$

Proof. To ease the notation, we write $\mathbf{e}_i^{n+p}$ as $\mathbf{e}_i$ for $i = 1, \ldots, n+p$ in the sequel. We focus on the lower bound first. Denote $N_1 = \lfloor \frac{N}{2(n+p)} \rfloor \times 2(n+p)$. Since the assumption $N \geq 4(n+p)$ implies $N_1 \geq 4(n+p)$, we have

$$\begin{aligned} ZZ' &= \sum_{i=1}^{N} \hat{z}_0^i \hat{z}_0^{i'} \succeq \sum_{i=1}^{N_1} \hat{z}_0^i \hat{z}_0^{i'} \\ &= \left(\sum_{i=1,1+(n+p),1+2(n+p),\ldots}^{N_1-(n+p)+1} \begin{bmatrix} s_i \mathbf{e}_1 q \\ 1 \end{bmatrix} \begin{bmatrix} (s_i \mathbf{e}_1 q)' & 1 \end{bmatrix} \right) \\ &+ \left(\sum_{i=2,2+(n+p),2+2(n+p),\ldots}^{N_1-(n+p)+2} \begin{bmatrix} s_i \mathbf{e}_2 q \\ 1 \end{bmatrix} \begin{bmatrix} (s_i \mathbf{e}_2 q)' & 1 \end{bmatrix} \right) \\ &+ \cdots \\ &+ \left(\sum_{i=n+p,n+p+(n+p),\ldots}^{N_1} \begin{bmatrix} s_i \mathbf{e}_{n+p} q \\ 1 \end{bmatrix} \begin{bmatrix} (s_i \mathbf{e}_{n+p} q)' & 1 \end{bmatrix} \right) \\ &= \begin{bmatrix} M_1 & M_2 \\ M_2' & M_3 \end{bmatrix}, \end{aligned} \tag{6.8}$$

where $M_1 \in \mathbb{R}^{(n+p)\times(n+p)}$, $M_2 \in \mathbb{R}^{(n+p)\times 1}$, and $M_3 \in \mathbb{R}^{1\times 1}$.

For the submatrix M_1, we have

$$\begin{aligned} M_1 &= \sum_{j=1}^{n+p} \sum_{i=1}^{\frac{N_1}{n+p}} \mathbf{e}_j \mathbf{e}_j' q^2 = \sum_{j=1}^{n+p} \frac{N_1}{n+p} \mathbf{e}_j \mathbf{e}_j' q^2 \\ &= \operatorname{diag}(\frac{N_1}{n+p} q^2, \cdots, \frac{N_1}{n+p} q^2), \end{aligned} \tag{6.9}$$

where we used the property that $s_i^2 = 1$ for all i, and the fact that $N_1 \bmod 2(n+p) = 0$

For the submatrix M_2, we have

$$\begin{aligned} M_2 &= \left(\sum_{i=1,1+(n+p),1+2(n+p),\ldots}^{N_1-(n+p)+1} s_i \mathbf{e}_1 q \right) + \cdots + \left(\sum_{i=n+p,n+p+(n+p),\ldots}^{N_1} s_i \mathbf{e}_{n+p} q \right) \\ &= \mathbf{0} + \mathbf{0} + \ldots + \mathbf{0} = \mathbf{0}, \end{aligned} \tag{6.10}$$

where we used the property that $s_i = 1$ if $i \in \{j(n+p)+1, j(n+p)+2, \ldots, j(n+p)+(n+p) | j \text{ is even}\}$ and $s_i = -1$ if $i \in \{j(n+p)+1, j(n+p)+2, \ldots, j(n+p)+(n+p) | j \text{ is odd}\}$, and the fact that $N_1 \bmod 2(n+p) = 0$, i.e., the number of positive terms is exactly the same as the number of negative terms for each summation.

Lastly, for the scalar matrix M_3, we have

$$M_3 = \sum_{i=1}^{N_1} 1^2 = N_1. \tag{6.11}$$

Combining (6.9)-(6.11), we have

$$\begin{aligned} \lambda_{min}(ZZ') &\geq \lambda_{min}\left(\begin{bmatrix} M_1 & M_2 \\ M_2' & M_3 \end{bmatrix}\right) \\ &= \min\{\frac{N_1}{n+p}q^2, N_1\}. \end{aligned} \tag{6.12}$$

Using the property $\lfloor \frac{N}{c} \rfloor c \geq N - c$ for any $c > 0$, we have

$$N_1 = \lfloor \frac{N}{2(n+p)} \rfloor \times 2(n+p) \geq N - 2(n+p) \geq \frac{N}{2}, \tag{6.13}$$

where the second inequality is due to our assumption that $N \geq 4(n+p)$.

Hence, the above inequality in conjunction with (6.12) yields

$$\lambda_{min}(ZZ') \geq N \min\{\frac{q^2}{2(n+p)}, \frac{1}{2}\}, \tag{6.14}$$

which is of the desired form.

Next, we show the upper bound. Denoting $N_2 = \lceil \frac{N}{2(n+p)} \rceil \times 2(n+p)$, using $N \leq N_2$, we have

$$ZZ' = \sum_{i=1}^{N} \hat{z}_0^i \hat{z}_0^{i'} \preceq \sum_{i=1}^{N_2} \hat{z}_0^i \hat{z}_0^{i'}, \tag{6.15}$$

where $\hat{z}_0^1, \hat{z}_0^2, \ldots, \hat{z}_0^{N_2}$ are generated from Algorithm 6 with input parameter N_2. Since N_2 mod $2(n+p) = 0$, we can follow a similar procedure as in the proof of the lower bound to obtain

$$\begin{aligned}\lambda_{max}(ZZ') &\leq \max\{\frac{N_2}{n+p}q^2, N_2\} \\ &\leq \max\{\frac{N+2(n+p)}{n+p}q^2, N+2(n+p)\} \\ &\leq N\max\{\frac{2q^2}{n+p}, 2\},\end{aligned} \tag{6.16}$$

where the second inequality is due to the relationship $N_2 \leq N + 2(n+p)$, and the last inequality is due to the assumption that $N \geq 4(n+p)$. □

Leveraging Lemma 5, we have the following result that upper bounds the contribution from noise.

Lemma 28. *Suppose that Algorithm 6 is used to generate data. Let $N \geq 4(n+p)$ and $q \leq \sqrt{n+p}$. Then for any fixed $\delta \in (0,1)$, we have with probability at least $1-\delta$*

$$\|WZ'(ZZ' + \lambda I_{n+p+1})^{-1/2}\| \leq 3\sigma_w\sqrt{\log\frac{9^n}{\delta} + (n+p+1)\log(1+\frac{4(n+p)}{q^2+\zeta})},$$

where $\zeta = \frac{4\lambda(n+p)}{N}$.

Proof. Denoting $\bar{V}_N = \lambda I_{n+p+1} + ZZ'$, we have

$$\|WZ'(ZZ' + \lambda I_{n+p+1})^{-1/2}\| = \|\bar{V}_N^{-1/2}ZW'\|.$$

Let $\hat{V}_N = (\lambda + \frac{Nq^2}{2(n+p)})I_{n+p+1}$. When $N \geq 4(n+p)$ and $q \leq \sqrt{n+p}$, we can apply Lemma 27 to get $\bar{V}_N \succeq \hat{V}_N$. Since $\bar{V}_N \succeq \hat{V}_N \Rightarrow 2\bar{V}_N \succeq \bar{V}_N + \hat{V}_N \Rightarrow \bar{V}_N^{-1} \preceq 2(\bar{V}_N + \hat{V}_N)^{-1}$, we can write

$$\begin{aligned}\|\bar{V}_N^{-1/2}ZW'\| &\leq \sqrt{2}\|(\bar{V}_N + \hat{V}_N)^{-1/2}ZW'\| \\ &= \sqrt{2}\|(\hat{V}_N + \lambda I_{n+p+1}) + \sum_{i=1}^N \hat{z}_0^i \hat{z}_0^{i'})^{-1/2}(\sum_{i=1}^N \hat{z}_0^i w_0^{i'})\|,\end{aligned}$$

where the inequality is due to [101, Lemma 10].

Denote $V = \hat{V}_N + \lambda I_{n+p+1}$. Define the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t = \sigma(\{\hat{z}_0^{i+1}\}_{i=0}^t \cup \{w_0^j\}_{j=1}^t)$. Since the sequence of $\hat{z}_0^i$ generated by Algorithm 6 is deterministic, and the noise terms are independent, for any fixed $\delta \in (0,1)$, we can apply Lemma 5 to obtain with probability at least $1-\delta$

$$\sqrt{2}\|(\bar{V}_N + \hat{V}_N)^{-1/2} ZW'\| \leq 3\sigma_w \sqrt{\log \frac{9^n}{\delta} + \frac{1}{2}\log\det((V+ZZ')V^{-1})}.$$

When $q \leq \sqrt{n+p}$, we can apply the upper bound in Lemma 27 to obtain

$$\begin{aligned} det((V+ZZ')V^{-1}) &= \frac{det(V+ZZ')}{det(V)} \\ &\leq \frac{(2\lambda + \frac{Nq^2}{2(n+p)} + \|ZZ'\|)^{n+p+1}}{(2\lambda + \frac{Nq^2}{2(n+p)})^{n+p+1}} \\ &\leq (1 + \frac{2N}{2\lambda + \frac{Nq^2}{2(n+p)}})^{n+p+1} \\ &= (1 + \frac{4(n+p)}{q^2+\zeta})^{n+p+1}, \end{aligned}$$

where we used the fact that the determinant is the product of eigenvalues. The result then follows. □

Next, we bound the contribution from the higher order terms.

Lemma 29. *Suppose that Algorithm 6 is used to generate data. Let $N \geq 4(n+p)$ and $q \leq \sqrt{n+p}$. Fix constants c and β that satisfy Assumption 7, and denote $\gamma = \frac{\lambda(n+p)}{Nq^2}$. Then if $q < c$, we have*

$$\|RZ'(ZZ' + \lambda I_{n+p+1})^{-1}\| \leq \sqrt{\frac{2\beta^2(n^2+np)}{1+\gamma}}q + \frac{2(n+p)\sqrt{\lambda N n \beta^2 q^4}}{Nq^2 + 2\lambda(n+p)}. \tag{6.17}$$

Proof. Note that

$$\|RZ'(ZZ' + \lambda I_{n+p+1})^{-1}\| \leq \|R\|\|Z'(ZZ' + \lambda I_{n+p+1})^{-1}\|. \tag{6.18}$$

For the term $\|R\|$, using $R_{i,j}$ to denote its (i,j) entry, we have

$$\|R\| \leq \|R\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{N} R_{i,j}^2} \leq \sqrt{Nn\beta^2 q^4}, \tag{6.19}$$

where the second inequality is due to the fact that $\|z_0^i\|_1 = q$ for all $i = 1, \ldots, N$, the assumption that $q < c$, and Assumption 7.

For the term $\|Z'(ZZ' + \lambda I_{n+p+1})^{-1}\|$, we have

$$\|Z'(ZZ' + \lambda I_{n+p+1})^{-1}\| = \sqrt{\|(ZZ' + \lambda I_{n+p+1})^{-1}ZZ'(ZZ' + \lambda I_{n+p+1})^{-1}\|}.$$

Note that

$$\begin{aligned}
&\|(ZZ' + \lambda I_{n+p+1})^{-1}ZZ'(ZZ' + \lambda I_{n+p+1})^{-1}\| = \\
&\|(ZZ' + \lambda I_{n+p+1})^{-1}(ZZ' + \lambda I_{n+p+1})(ZZ' + \lambda I_{n+p+1})^{-1} \\
&\quad - \lambda(ZZ' + \lambda I_{n+p+1})^{-1}(ZZ' + \lambda I_{n+p+1})^{-1}\| \\
&\leq \|(ZZ' + \lambda I_{n+p+1})^{-1}\| + \lambda\|(ZZ' + \lambda I_{n+p+1})^{-1}\|^2.
\end{aligned} \tag{6.20}$$

From Weyl's inequality [102], we have

$$\|(ZZ' + \lambda I_{n+p+1})^{-1}\| = \frac{1}{\lambda_{min}(ZZ' + \lambda I_{n+p+1})} \leq \frac{1}{\lambda_{min}(ZZ') + \lambda}.$$

Using the above inequality and (6.20), since $N \geq 4(n+p)$ and $q \leq \sqrt{n+p}$, we can apply Lemma 27 to get

$$\|Z'(ZZ' + \lambda I_{n+p+1})^{-1}\| \leq \sqrt{\frac{2(n+p)}{Nq^2 + 2\lambda(n+p)}} + \frac{2\sqrt{\lambda}(n+p)}{Nq^2 + 2\lambda(n+p)},$$

where we used the relationship that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$.

Finally, combining the above inequality with (6.19), and after some algebraic manipulations, we have the desired result. □

6.3.2 Main Result

Now we present our main result, a finite sample upper bound of the system identification error (6.7).

Theorem 6.3.1. *Suppose that Algorithm 6 is used to generate data. Let $N \geq 4(n+p)$ and $q \leq \sqrt{n+p}$. Fix constants c, β that satisfy Assumption 7, and a confidence parameter $\delta \in (0,1)$. Then if $q < c$, with probability at least $1-\delta$, the estimation error of Algorithm 7 satisfies*

$$\begin{aligned}\|\hat{\Theta} - \Theta\| \leq & \underbrace{\frac{5\sigma_w\sqrt{\log\frac{9^n}{\delta} + (n+p+1)\log(1+\frac{4(n+p)}{q^2})}}{\sqrt{Nq^2/(n+p)+\lambda}}}_{\text{Error due to noise}} \\ & + \underbrace{\sqrt{\frac{2(n^2+np)}{1+\gamma}}\beta q}_{\text{Error due to nonlinearity}} \\ & + \underbrace{\frac{2(n+p)(\lambda\|\Theta\| + \sqrt{\lambda N n \beta^2 q^4})}{2\lambda(n+p)+Nq^2}}_{\text{Error due to regularization}},\end{aligned} \tag{6.21}$$

where $\gamma = \frac{\lambda(n+p)}{Nq^2}$.

Proof. Recall the estimation error in (6.7). We have

$$\begin{aligned}\|\hat{\Theta} - \Theta\| \leq \lambda\|\Theta\|\|(ZZ' + \lambda I_{n+p+1})^{-1}\| + \|RZ'(ZZ' + \lambda I_{n+p+1})^{-1}\| \\ + \|WZ'(ZZ' + \lambda I_{n+p+1})^{-1/2}\|\|(ZZ' + \lambda I_{n+p+1})^{-1/2}\|.\end{aligned} \tag{6.22}$$

Noting that

$$\begin{aligned}\|(ZZ' + \lambda I_{n+p+1})^{-1/2}\| &= \frac{1}{\sqrt{\lambda_{min}(ZZ' + \lambda I_{n+p+1})}} \\ &\leq \frac{1}{\sqrt{\lambda_{min}(ZZ') + \lambda}},\end{aligned} \tag{6.23}$$

from Weyl's inequality [102], the result directly follows from applying Lemma 27, Lemma 28, and Lemma 29 after some algebraic manipulations. □

Remark 21. ***Interpretation of Theorem 6.3.1.*** *Note that Theorem 6.3.1 holds irrespective of the spectral radius of the system matrix A, which captures a well known advantage*

of the multiple trajectories setup. Below we discuss other key insights provided by Theorem 6.3.1.

Trade-off between error due to noise and error due to nonlinearity: *Suppose that $\lambda = 0$ for now. When the system is perfectly linear, one has $\beta = 0$. Consequently, the upper bound in Theorem 6.3.1 only contains the error due to noise, which goes to zero with a rate of $\mathcal{O}(\frac{1}{\sqrt{N}})$. This implies a consistent estimator of which the convergence rate matches the results in the existing literature for learning perfectly linear system using random inputs [12], [13]. When there does exist nonlinearity, i.e., $\beta > 0$, one can observe that the error due to nonlinearity can be made arbitrarily small by choosing a smaller q used in Algorithm 6 (recall that q captures the magnitude of the initial conditions). On the other hand, a smaller q would also make the denominator of the term capturing error due to noise small, thus leading to a larger error due to noise. In other words, if one starts close enough to the reference point (by setting q to be small), one would have less bias due to nonlinearity, at the cost of having a smaller signal to noise ratio (thus a larger error due to noise). However, the error due to noise can always be made almost zero by increasing the number of experiments N. Consequently, if one can afford to generate a large amount of data, it is preferable to use a small q due to the low bias introduced by the nonlinear terms, and the small error introduced by the noise (which is due to the large amount of samples). These insights are different from system identification for truly linear systems, where it is commonly believed that a larger signal to noise ratio is always better. We will also illustrate these ideas empirically in Section 6.4.*

Role of regularization: *Suppose that N, q are fixed. As λ becomes larger, we can observe that both the error due to noise and the error due to nonlinearity goes to zero, and the error due to regularization converges to $\|\Theta\|$. This result implies that setting λ to be relatively large can be helpful if σ_w is large (system is very noisy) or β is large (system has strong nonlinearity), while $\|\Theta\|$ is small. However, the optimal λ can be hard to obtain if (some upper bounds of) the parameters in (6.21) are unknown in advance. In practice, cross validation techniques [27] are commonly used to select a good value of λ.*

6.4 Numerical Examples

In this section, we provide simulated numerical examples to validate the insights for system identification using Algorithm 6 and Algorithm 7. We also compare the results against the single trajectory setup, where the input is set to be independent zero mean Gaussian, with slight adjustments to deal with the offset o in our setup (6.2), i.e., by appending ones in the regressor matrix. More specifically, we still use Algorithm 7 in the single trajectory setup, but the dataset is generated without restarting the system, see [6], [12] for examples. Such comparisons are made since Gaussian inputs are commonly used in the literature on linear system identification [8], [13]. For simplicity, we set $\lambda = 0$ for all experiments. All results are averaged over 10 independent experiments.

6.4.1 System with mild nonlinearity

In the first example, we investigate the performance of the system identification algorithms under mild nonlinearity. The model we use here captures the dynamics of a nonlinear pendulum.[9] The system states are the pendulum angle and its velocity, and the input is the torque applied. We set the mass and length of the pendulum to be 1 kg and 1 meter, respectively. After discretization using Euler's method by setting the sampling time to be 0.05 seconds, the dynamics is given by

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \end{bmatrix} = \begin{bmatrix} x_{1,k} + 0.05 x_{2,k} \\ -0.49 \sin(x_{1,k}) + x_{2,k} + 0.05 u_k \end{bmatrix} + w_k, \tag{6.24}$$

where we set w_k to be independent Gaussian random vectors with zero mean and covariance matrix given by $0.25 I_2$. The linearized system matrices around the origin are given by

$$A = \begin{bmatrix} 1 & 0.05 \\ -0.49 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0.05 \end{bmatrix}, o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{6.25}$$

[9] https://courses.engr.illinois.edu/ece486/fa2019/handbook/lec02.html

We plot the system identification error using Algorithm 6 and Algorithm 7 versus the number of experiments N for $q = 1.2, 0.9, 0.6$ in Fig. 6.1. As can be observed, a smaller q could lead to a larger error when N is small, due to a smaller signal to noise ratio. However, a smaller q can eventually result in a smaller error when N is large enough due to less bias, which confirms our observations in Theorem 6.3.1.

In the single trajectory setup, we plot the error using i.i.d zero mean Gaussian inputs with different variance σ_u^2, where N here represents the number of samples used in the single trajectory. A common heuristic is that one should apply small inputs to learn a good linear approximation around a given reference point, i.e., the variance σ_u^2 should be small. However, as shown in Fig. 6.2, the error plateaus at around 0.6, even for small variance inputs. The key reason is that the random input and process noise can always drive the system states to undesired regions and excite the higher order terms, unless the input is carefully designed. In fact, the paper [92] shows that random inputs in the single trajectory setup could result in inconsistent estimation under certain conditions even for Lipschitz nonlinearity.

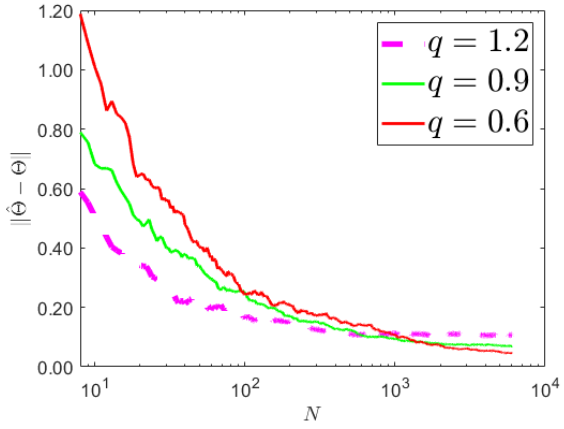

Figure 6.1. System identification error using Algorithms 6-7 with different q, mild nonlinearity

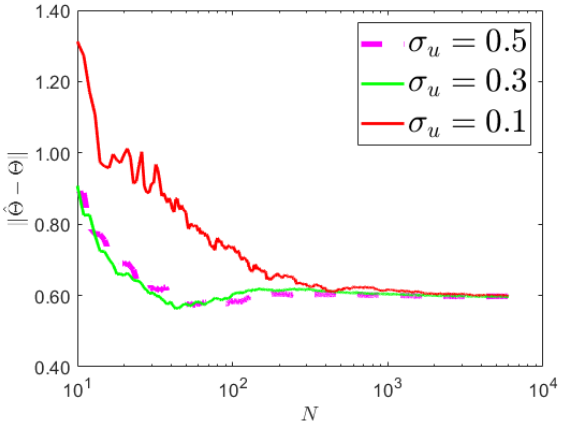

Figure 6.2. System identification error using a single trajectory with different σ_u, mild nonlinearity

6.4.2 System with strong nonlinearity

In the second example, we investigate the performance of the system identification algorithms under strong nonlinearity (where the assumption of lipschitzness used in [92] no longer holds). The virtual model we use here is given by

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \end{bmatrix} = \begin{bmatrix} 0.9 & 0.5 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} x_{1,k} \\ x_{2,k} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u_k + \begin{bmatrix} x_{1,k}^3 + x_{2,k}^5 \\ x_{1,k} x_{2,k} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} + w_k, \tag{6.26}$$

where we again set w_k to be independent Gaussian random vectors with zero mean and covariance matrix given by $0.25 I_2$.

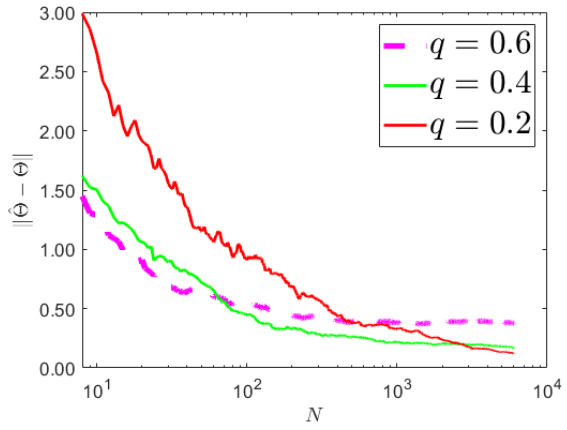

Figure 6.3. System identification error using Algorithms 6-7 with different q, strong nonlinearity

Again, we plot the system identification error using Algorithm 6 and Algorithm 7 versus the number of experiments N for $q = 0.6, 0.4, 0.2$. As can be observed, similar trends still hold, i.e., a smaller q results in a larger error when N is small, but is beneficial in the long run, even for system with relatively strong nonlinearity.

In contrast, in the single trajectory setup, we applied i.i.d zero mean Gaussian inputs with variance $\sigma_u^2 = 0.1^2, 0.01^2, 0.001^2$. However, all of them fail to converge and result in numerical issues since the noise and non-zero offset drive the system states to regions where nonlinearity dominates.

6.5 Conclusion and future work

In this chapter, we proposed system identification algorithms to learn the linearized model of a system. Unlike existing works, we assume that the underlying dynamics could be nonlinear. We presented a finite sample error bound of the algorithms, which shows that one can learn the linearized dynamics with arbitrarily small error given sufficiently many samples, and demonstrates a trade-off between the error due to noise and the error due to nonlinearity. Our bound further characterizes the benefits of using regularization. As shown in [103], initializing states at different locations might come at different costs. Consequently, future work would focus on studying how to optimize the data collection procedure under constraints on initial state/input.

7. SUMMARY AND FUTURE WORK

In this thesis, we studied finite sample performance guarantees for learning the dynamics of systems. In Chapter 2, we studied system identification of a fully observed system, when one has access to data generated from a "similar" system. Our result shows that one can leverage the auxiliary data to reduce the error due to the noise, at the cost of adding a bias that depends on the difference between the true and auxiliary systems. In Chapter 3, we studied the finite sample guarantees for learning a partially observed linear autonomous system, under multiple trajectories of transient response of the system. We proved a learning rate that is consistent with classical results, and extended the analysis to the case where the initial state has possibly non-zero mean. In Chapter 4, we proposed a distributed online parameter estimation algorithm in a networked setting. We provided finite time estimation error bounds, and showed that our results allow one to determine a time at which the communication can be stopped (due to the costs associated with maintaining communications), while meeting a desired estimation accuracy. In Chapter 5, we studied the online change points detection problem for linear dynamical systems. We proposed a detection algorithm and a data-dependent detection rule that allows the user to achieve a desired upper bound on the false alarm probability. We also provided a finite-sample-based lower bound for the probability of making a true alarm with a certain delay. Finally, in Chapter 6, we made a step toward learning the dynamics of a nonlinear system. We proposed a data-acquisition algorithm followed by a regularized least squares algorithm to enable the identification of the linearized model of the nonlinear system. We also provided a finite sample error bound, which demonstrates a trade-off between the error due to noise and the error due to nonlinearity.

Finally, we describe some directions for future work.

7.1 Developing Lower Bounds for Learning from Similar Systems

One interesting direction to explore is the establishment of lower bounds for the weighted least squares-based system identification technique discussed in Chapter 2. These results could help characterize the optimal performance achievable and potentially facilitate the

development of improved estimators. Several potential approaches involve leveraging assumptions, such as sparsity [20], or utilizing prior knowledge, such as a known auxiliary system model.

7.2 Learning Controllers from Similar Systems

Our work in Chapter 2 studies the system identification problem for similar systems. We are interested in exploring the possibility of integrating these techniques into control tasks. For example, can we leverage pre-trained controllers designed for similar systems to enhance the control of the target system?

7.3 Change Point Detection for Nonlinear Systems

Since many physical systems are nonlinear, developing change point detection algorithms for nonlinear systems will be of great importance. One approach is to use a similar sliding windows-based technique (as studied in Chapter 5), assuming that a certain model structure is known. Nevertheless, an open question remains regarding the extent of guarantees that can be provided in such cases.

7.4 Change Point Detection vs Sensor Fault Detection

Our techniques in Chapter 5 assume that all of the system states can be perfectly measured. However, in practice, it is typical that only a subset of the system states can be observed (and the measurements are corrupted by noise). Our ongoing work leverages a similar approach by estimating the Markov parameter matrix of the system in an online manner. We are also interested in distinguishing between changes in system dynamics and sensor faults.

7.5 Linear Control for Nonlinear Systems

An interesting and important question is to provide performance characterization of linear control laws for nonlinear systems. In particular, since linearization around an equilibrium

point is a commonly used technique in practice, we would like to understand how a controller designed from an imperfect linearized model performs, when an upper bound on the error of the learned model is available.

7.6 Federated Learning

Federated learning has been widely used in practice. We would like to explore if the techniques we developed in Chapter 4 can be extended to the federated learning setup, especially under heterogeneous data.

7.7 Real-World Problems

While our work is primarily theoretical, understanding how real world problems differ from theories is crucial. Typically, the discrepancy between theory and practice comes at the violation of assumptions. For example, the noise added to a system is not Gaussian, which could break the guarantees one could provide. Consequently, it is always important to study problems under general assumptions, e.g., sub-Gaussian noise instead of Gaussian noise.

REFERENCES

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*. AMLBook New York, 2012, vol. 4.

[2] D. Bauer, M. Deistler, and W. Scherrer, "Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs," *Automatica*, vol. 35, no. 7, pp. 1243–1254, 1999.

[3] M. Jansson and B. Wahlberg, "On consistency of subspace methods for system identification," *Automatica*, vol. 34, no. 12, pp. 1507–1519, 1998.

[4] T. Knudsen, "Consistency analysis of subspace identification methods based on a linear regression approach," *Automatica*, vol. 37, no. 1, pp. 81–89, 2001.

[5] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] L. Ye, H. Zhu, and V. Gupta, "On the sample complexity of decentralized linear quadratic regulator with partially nested information structure," *arXiv preprint arXiv:2110.07112*, 2021.

[7] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Proc. Conference On Learning Theory*, 2018, pp. 439–473.

[8] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *American control conference*, IEEE, 2019, pp. 5655–5661.

[9] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," in *Proc. Conference on Learning Theory*, 2019, pp. 2714–2802.

[10] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time lti system identification," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1186–1246, 2021.

[11] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[12] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *Proc. International Conference on Machine Learning*, 2019, pp. 5610–5618.

[13] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, pp. 1–47, 2019.

[14] S. Fattahi and S. Sojoudi, "Data-driven sparse system identification," in *Proc. Allerton Conference on Communication, Control, and Computing*, 2018, pp. 462–469.

[15] Y. Sun, S. Oymak, and M. Fazel, "Finite sample system identification: Optimal rates and the role of regularization," in *Proc. Learning for Dynamics and Control Conference*, 2020, pp. 16–25.

[16] Y. Zheng and N. Li, "Non-asymptotic identification of linear dynamical systems using multiple trajectories," *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1693–1698, 2020.

[17] L. Xin, G. Chiu, and S. Sundaram, "Learning the dynamics of autonomous linear systems from multiple trajectories," in *2022 American Control Conference (ACC)*, IEEE, 2022, pp. 3955–3960.

[18] S. Tu, R. Frostig, and M. Soltanolkotabi, "Learning from many trajectories," *arXiv preprint arXiv:2203.17193*, 2022.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[20] H. Bastani, "Predicting with proxies: Transfer learning in high dimension," *Management Science*, vol. 67, no. 5, pp. 2964–2984, 2021.

[21] A. Modi, M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Joint learning of linear time-invariant dynamical systems," *arXiv preprint arXiv:2112.10955*, 2021.

[22] M. K. S. Faradonbeh and A. Modi, "Joint learning-based stabilization of multiple unknown linear systems," *IFAC-PapersOnLine*, vol. 55, no. 12, pp. 723–728, 2022.

[23] L. Xin, L. Ye, G. Chiu, and S. Sundaram, "Identifying the dynamics of a system by leveraging data from similar systems," in *2022 American Control Conference (ACC)*, IEEE, 2022, pp. 818–824.

[24] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[25] O. Rivasplata, "Subgaussian random variables: An expository note," *Internet publication, PDF*, vol. 5, 2012.

[26] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Conference on Learning Theory*, PMLR, 2020, pp. 3487–3582.

[27] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation.," *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[28] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[29] L. Li, C. De Persis, P. Tesi, and N. Monshizadeh, "Data-based transfer stabilization in linear systems," *arXiv preprint arXiv:2211.05536*, 2022.

[30] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electronic Communications in Probability*, vol. 18, pp. 1–9, 2013.

[31] F. Zhang and Q. Zhang, "Eigenvalue inequalities for matrix product," *IEEE Transactions on Automatic Control*, vol. 51, no. 9, pp. 1506–1509, 2006.

[32] H. Zhang and S. X. Chen, "Concentration inequalities for statistical inference," *arXiv preprint arXiv:2011.02258*, 2020.

[33] D. Dadush, C. Guzmán, and N. Olver, "Fast, deterministic and sparse dimensionality reduction," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 2018, pp. 1330–1344.

[34] K. Moshksar, "On the absolute constant in hanson-wright inequality," *arXiv preprint arXiv:2111.00557*, 2021.

[35] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," in *2019 American Control Conference (ACC)*, IEEE, 2019, pp. 5582–5588.

[36] N. Matni and S. Tu, "A tutorial on concentration bounds for system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, IEEE, 2019, pp. 3741–3749.

[37] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, 2011.

[38] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[39] N. Chan and M. K. Kwong, "Hermitian matrix inequalities and a conjecture," *The American Mathematical Monthly*, vol. 92, no. 8, pp. 533–541, 1985.

[40] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *Conference on Decision and Control (CDC)*, arXiv:1903.09122, IEEE, 2019, pp. 3648–3654.

[41] L. Xin, G. Chiu, and S. Sundaram, "Learning the dynamics of autonomous linear systems from multiple trajectories," in *Proc. American Control Conference*, 2022.

[42] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," *arXiv preprint arXiv:1806.05722*, 2018.

[43] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite-time system identification for partially observed LTI systems of unknown order," *arXiv preprint arXiv:1902.01848*, 2019.

[44] Y. Xing, B. Gravell, X. He, K. H. Johansson, and T. Summers, "Linear system identification under multiplicative noise from multiple trajectory data," in *Proc. American Control Conference*, 2020, pp. 5157–5261.

[45] M. Deistler, K. Peternell, and W. Scherrer, "Consistency and relative efficiency of subspace methods," *Automatica*, vol. 31, no. 12, pp. 1865–1875, 1995.

[46] P. Van Overschee and B. De Moor, *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media, 2012.

[47] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[48] D. J. Hartfiel, *Nonhomogeneous matrix products*. World Scientific, 2002.

[49] R. A. Horn and C. R. Johnson, "Topics in matrix analysis, 1991," *Cambridge University Presss, Cambridge*, vol. 37, p. 39, 1991.

[50] L. Xin, G. Chiu, and S. Sundaram, "Finite sample guarantees for distributed online parameter estimation with communication costs," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, 2022, pp. 5980–5985.

[51] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[52] M. Zhu and S. Martinez, "On distributed convex optimization under inequality and equality constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2011.

[53] S. Kar, J. M. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012.

[54] Q. Zhang and J.-F. Zhang, "Distributed parameter estimation over unreliable networks with markovian switching topologies," *IEEE Transactions on Automatic Control*, vol. 57, no. 10, pp. 2545–2560, 2012.

[55] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed optimization via dual averaging," in *52nd IEEE Conference on Decision and Control*, IEEE, 2013, pp. 1484–1489.

[56] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed convex optimization on dynamic networks," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3545–3550, 2016.

[57] D. Yuan, A. Proutiere, and G. Shi, "Distributed online linear regressions," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 616–639, 2020.

[58] L. Su and S. Shahrampour, "Finite-time guarantees for byzantine-resilient distributed state estimation with noisy measurements," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3758–3771, 2019.

[59] R. Tron and R. Vidal, "Distributed computer vision algorithms through distributed averaging," in *CVPR 2011*, IEEE, 2011, pp. 57–63.

[60] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.

[61] L. Ye and S. Sundaram, "Distributed maximization of submodular and approximately submodular functions," in *2020 59th IEEE Conference on Decision and Control (CDC)*, IEEE, 2020, pp. 2979–2984.

[62] P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of markov chains," *The annals of applied probability*, pp. 36–61, 1991.

[63] G. W. Stewart, "On the perturbation of pseudo-inverses, projections and linear least squares problems," *SIAM review*, vol. 19, no. 4, pp. 634–662, 1977.

[64] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[65] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107 299, 2020.

[66] A. S. Polunchenko and A. G. Tartakovsky, "State-of-the-art in sequential change-point detection," *Methodology and computing in applied probability*, vol. 14, no. 3, pp. 649–684, 2012.

[67] A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov, "Efficient computer network anomaly detection by changepoint detection methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 4–11, 2012.

[68] T. Flynn and S. Yoo, "Change detection with the kernel cumulative sum algorithm," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, IEEE, 2019, pp. 6092–6099.

[69] W.-C. Chang, C.-L. Li, Y. Yang, and B. Póczos, "Kernel change-point detection with auxiliary deep generative models," *arXiv preprint arXiv:1901.06077*, 2019.

[70] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[71] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[72] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[73] M. K. Titsias, J. Sygnowski, and Y. Chen, "Sequential changepoint detection in neural networks with checkpoints," *Statistics and Computing*, vol. 32, no. 2, pp. 1–19, 2022.

[74] J. Lee, Y. Xie, and X. Cheng, "Training neural networks for sequential change-point detection," *arXiv preprint arXiv:2210.17312*, 2022.

[75] A. G. Tartakovsky, "On asymptotic optimality in sequential changepoint detection: Non-iid case," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3433–3450, 2017.

[76] L. Zadeh and C. Desoer, *Linear system theory: the state space approach*. Courier Dover Publications, 2008.

[77] O. Zoeter and T. Heskes, "Change point problems in linear dynamical systems," *Journal of Machine Learning Research*, vol. 6, pp. 1999–2026, 2005.

[78] Y. Kawahara, T. Yairi, and K. Machida, "Change-point detection in time-series data based on subspace identification," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, IEEE, 2007, pp. 559–564.

[79] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Nonparametric bayesian learning of switching linear dynamical systems," *Advances in neural information processing systems*, vol. 21, 2008.

[80] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[81] J. Han, K. Lee, A. Tong, and J. Choi, "Confirmatory bayesian online change point detection in the covariance structure of gaussian processes," *arXiv preprint arXiv:1905.13168*, 2019.

[82] A. Alanqary, A. Alomar, and D. Shah, "Change point detection via multivariate singular spectrum analysis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 218–23 230, 2021.

[83] J. Geng, B. Zhang, L. M. Huie, and L. Lai, "Online change-point detection of linear regression models," *IEEE Transactions on Signal Processing*, vol. 67, no. 12, pp. 3316–3329, 2019.

[84] T. Banerjee and V. V. Veeravalli, "Data-efficient minimax quickest change detection with composite post-change distribution," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5172–5184, 2015.

[85] Z. Sun and S. S. Ge, "Analysis and synthesis of switched linear control systems," *Automatica*, vol. 41, no. 2, pp. 181–195, 2005.

[86] G. Zhai, B. Hu, K. Yasuda, and A. N. Michel, "Stability analysis of switched systems with stable and unstable subsystems: An average dwell time approach," *International Journal of Systems Science*, vol. 32, no. 8, pp. 1055–1061, 2001.

[87] A. E. Ahmed, A. Hafez, A. Ouda, H. E. H. Ahmed, and H. M. Abd-Elkader, "Modeling of a small unmanned aerial vehicle," *Adv Robot Autom*, vol. 4, no. 126, p. 2, 2015.

[88] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1328–1337.

[89] L. Xin, G. Chiu, and S. Sundaram, "Learning linearized models from nonlinear systems with finite data," *arXiv preprint arXiv:2309.08805*, 2023.

[90] L. Ljung, "System identification," *Wiley encyclopedia of electrical and electronics engineering*, pp. 1–19, 1999.

[91] W. J. Rugh, *Linear system theory*. Prentice-Hall, Inc., 1996.

[92] A. Sarker, P. Fisher, J. E. Gaudio, and A. M. Annaswamy, "Accurate parameter estimation for safety-critical systems with unmodeled dynamics," *Artificial Intelligence*, p. 103 857, 2023.

[93] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems," *Journal of Machine Learning Research*, vol. 23, no. 140, pp. 1–49, 2022.

[94] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 1433–1462, 2022.

[95] D. Foster, T. Sarkar, and A. Rakhlin, "Learning nonlinear dynamical systems from a single trajectory," in *Learning for Dynamics and Control*, PMLR, 2020, pp. 851–861.

[96] A. Mauroy and J. Goncalves, "Linear identification of nonlinear systems: A lifting technique based on the koopman operator," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, IEEE, 2016, pp. 6500–6505.

[97] W. Hao, B. Huang, W. Pan, D. Wu, and S. Mou, "Deep koopman representation of nonlinear time varying systems," *arXiv preprint arXiv:2210.06272*, 2022.

[98] Ş. Cobzaş, R. Miculescu, A. Nicolae, *et al.*, *Lipschitz functions*. Springer, 2019.

[99] R. Courant, F. John, A. A. Blank, and A. Solomon, *Introduction to calculus and analysis*. Springer, 1965, vol. 1.

[100] G. B. Folland, "Higher-order derivatives and taylor's formula in several variables," *Preprint*, pp. 1–4, 2005.

[101] L. Xin, L. Ye, G. Chiu, and S. Sundaram, "Learning dynamical systems by leveraging data from similar systems," *arXiv preprint arXiv:2302.04344*, 2023.

[102] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[103] A. A. Ahmadi, A. Chaudhry, V. Sindhwani, and S. Tu, "Safely learning dynamical systems from short trajectories," in *Learning for Dynamics and Control*, PMLR, 2021, pp. 498–509.