THE UNIVERSITY OF CHICAGO

ON THE OPTIMAL ESTIMATION, CONTROL, AND MODELING OF DYNAMICAL SYSTEMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
WANTING XU

CHICAGO, ILLINOIS
AUGUST 2017

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to give sincere thanks to my advisors, Professor Mihai Anitescu and Professor Michael Stein, for their continuous guidance and encouragement throughout the years. Both of them have been extremely insightful on work and generous about their ideas and time – always available to discuss and provide feedbacks. I am grateful to Professor Mihai Anitescu for introducing me to interesting problems, holding countless inspiring discussions, and encouraging me to think creatively. I am also indebted to him for supporting me through the summers of 2014 and 2015 at Argonne National Laboratory and a conference travel, which expand my horizon and help me grow in academics. I give my gratitude to Professor Michael Stein for his meticulous guidance on my research, tremendous patience and especially his insights and rigor about work that motivate me to think harder and improve.

I am thankful to Professor Jonathan Weare for being my committee member and providing great feedbacks on my research. My gratitude also goes to all the faculty in the Department of Statistics at University of Chicago for educating me and exposing me to a variety of interesting fields.

It has been a pleasure to meet all the inspiring friends and colleagues in the Department: Likai, Soudeep, Kushal, Wooseok, Mohammad, Sayar, Vivak, Ulises, Si, Yunfan and Ken. I feel grateful to their friendship and companionship which make my life at Chicago truly memorable.

Many other people in the Department have helped me in diverse ways. I would like thank the staff members, especially Dr. Edward Friedman, John Zekos, Laura Rigazzi and Kathryn Kraynik for their indispensable supports on various issues.

Most importantly, I would like to give my heartfelt thanks to my family for their constant encouragement and support during ups and downs. Their love and care have been a continuous source of strength and bravery for me during my research and many other life pursuits of mine.

ABSTRACT

Five chapters are included in this thesis. The first chapter gives an overview of the remaining chapters, and abstracts for Chapters 2 – 5 are given in the following four paragraphs.

In Chapter 2, we consider a limited-memory multiple shooting method for weakly constrained variational data assimilation. Maximum-likelihood-based state estimation for dynamical systems with model error raises computational challenges in memory usage due to the much larger number of free variables when compared to the perfect model case. To address this challenge, we present a limited-memory method for maximum-likelihood-based estimation of state space models. We reduce the memory storage requirements by expressing the optimal states as a function of checkpoints bounding a shooting interval. All states can then be recomputed as needed from a recursion stemming from the optimality conditions. The matching of states at checkpoints are imposed, in a multiple shooting fashion, as constraints on the optimization problem, which is solved with an augmented Lagrangian method. We prove that for nonlinear systems under certain assumptions the condition number of the Hessian matrix of the augmented Lagrangian function is bounded above with respect to the number of shooting intervals. Hence the method is stable for increasing time horizon. The assumptions include satisfying the observability conditions of the linearized system on a shooting interval. We also propose a recursion-based gradient evaluation algorithm for computing the gradient, which in turn allows the algorithm to proceed by storing at any time only the checkpoints and the states on a shooting interval. We demonstrate our findings with simulations in different regimes for Burgers' equation.

In Chapter 3, we investigate a temporal decomposition approach to long-horizon dynamic optimization problems. The problems are discrete-time, linear, time-dependent and with box constraints on the control variables. We prove that an overlapping domains temporal decomposition, while inexact, approaches the solution of the long-horizon dynamic optimization problem exponentially fast in the size of the overlap. The resulting subproblems share no solution information and thus can be computed independently in parallel. Our findings are

demonstrated with a small, synthetic production cost model with real demand data.

In Chapter 4, we investigate the behavior of maximum likelihood estimators (MLE) of parameters of the squared exponential covariance commonly used in modeling the outputs of deterministic computer experiments. We consider the behavior of maximum likelihood estimators (MLE) of parameters of a Gaussian process with squared exponential covariance function when the computer model has some simple deterministic form. We prove that for regularly spaced observations on the line, the MLE of the scale parameter converges to zero if the computer model is a constant function and diverges to infinity for linear functions. When observing successive derivatives of a pth order monomial at zero, we find the asymptotic orders of the MLE of the scale parameter for all $p \geq 0$. For some commonly used test functions, we compare MLE with cross validation in a prediction problem and explore the joint estimation of range and scale parameters. The correlation matrix is nearly numerically singular even when the sample size is moderate. To overcome numerical difficulties, we perform exact computation by making use of exact results for the correlation matrix and restricting to parameter values and test functions that yield rational correlations and function values at the observation locations. We also consider the common approach of including a nugget effect to deal with the numerical difficulties, and explore its consequences on model fitting and prediction.

In Chapter 5, we consider modeling and predicting observations generated from a nonlinear circuit. Analyzing chaotic observations generated from some unknown nonlinear dynamics presents significant challenges for modeling the process and predicting future evolutions. We consider time series data which is measured from an electrical circuit and exhibits chaotic behavior. We investigate the performances of Gaussian process and neural network models in short-term prediction and capturing the long-term dynamics. One of the major difficulties in modeling observations generated by some physical process is the characterization of the model and observational errors. We explore the effects of different types of model and observational errors on the likelihood function of the initial state using simulated data qualitatively similar to our observations.

CHAPTER 1

INTRODUCTION

This thesis includes four projects generally on computational and statistical methods for optimal estimation, control, and modeling of dynamical systems. Optimal estimation and control are two fundamental problems for dynamical systems and have a wide range of applications [20, 33, 55]. Given a system, there often exists a mathematical model describing some aspects of the behavior of the system. The model may be developed from physical insights, fundamental laws or empirical testing, and establishes an interaction between the inputs and outputs of the system [89]. In addition, measurement devices are constructed to make observations from the system. However, uncertainty exists both in the dynamical model and the observations. For example, model developed based on Newtonian physics are only approximations to the macro structure of the system, but various parameters with micro resolutions are not determined absolutely [89]. Moreover, measurement devices do not provide perfect or even complete data due to device limitations and measurement noise. Consequently, optimal estimation aims to reconstruct, in the presence of noise both in the dynamical model describing the physical system and in the measurements, the underlying states of the system. The optimal state estimate is the basis on which the predictions of future observations can be made. A large number of methods are developed for optimal estimation, including Kalman filter, extended Kalman filter and particle methods [39, 69].

In Chapter 2, we consider one such state estimation problem. Our problem is motivated by applications in atmospheric sciences, where we have non-negligible model errors and huge dimensions of the state space. Both factors result in a large number of free variables, and lead to computational challenges in memory storage. The use of existing methods in this situation is limited by their reliance on linearity, memory usage, or slow convergence [68, 64, 17]. To address this challenge, we developed a maximum-likelihood based state estimation approach that reduces memory requirement and is shown, both theoretically and numerically, to be stable under increasing estimation horizon.

In addition to the states and observations, another important component of a dynamical system is the control. The objective of controlling a dynamical system is to modify its behavior by feedbacks so that its outputs approach a desired reference trajectory. The task is often carried out by a negative feedback control system where the difference between the output and reference is minimized to dynamically change the output. As a result, optimal control seeks the control laws for some system so that a given cost objective can be minimized. We consider optimal control in the context of dynamic programming (DP), which finds the decisions/controls minimizing, subject to some dynamical evolution of the decisions and the system's states, a cost function that is additive in time. A key aspect of the DP technique is to find decisions/controls that balance between a low present cost and a high future cost [20].

In Chapter 3, we consider a dynamic programming/optimal control problem that features long horizons and constrained controls. Our problem arises from production cost modeling (PCM) in the electrical power industry, which simulates the least-cost solution to generate sufficient energy to meet demand over a long time period. The problem can be mathematically formulated as a DP with additional constraints on the states and controls other than the dynamical evolution. The long horizon feature of such problems poses significant computational challenges. Sequential solutions, such as receding horizon control [72, 88], exist, but are too time-consuming in this situation. To tackle the problem, we considered a temporal decomposition approach, which decomposes the full horizon and solves the problems on a slightly larger embedding region for each decomposition interval, and proved that it approximates the solutions with an approximation error decaying exponentially in the length of the embedding regions. Consequently, when implemented in parallel for each decomposition interval, the method can significantly reduce computation time with little loss of accuracy.

The first two chapters consider the case where the formulation of the involved dynamical system is known. However, sometimes in practice, only the outputs of the system are observed, and hence the task is to model the observations and infer the underlying dynamics. Therefore, the last two chapters concern some statistical aspects in modeling computer experiments and chaotic dynamical systems. Computer experiments have been used extensively in studying complex scientific phenomena. It is a powerful tool, when the full scale physical experiments are impossible, to investigate the relationship between the response and the inputs [107]. Often, such computer experiments are computationally expensive, so a common alternative to running the computer code at all input values of interest is to run the code at a few input values and make cheap predictions at the others. It was proposed in [104, 105] to use Gaussian processes (GPs) to model the responses of the deterministic computer experiments, and the use of stochastic models provides a statistical basis for experimental design, parameter estimation, interpolation and uncertainty calibration.

In practice, smooth test functions comprised of elementary functions are usually used as test cases for studying how well GPs model the experiments. In Chapter 4, we investigate the asymptotic properties of the maximum likelihood estimator (MLE) of the squared exponential covariance when the computer experiment is some smooth, simple deterministic function. Our work points to some possible issues and consequences when using simple smooth test functions to study the effectiveness of GPs in modeling the computer experiments.

The study of nonlinear dynamics and chaos has been traditionally focused on characterizing and classifying the asymptotic behavior of the iterates [35, 56, 110, 111]. While such studies lead to important mathematical invariants that reveal properties of the system once its mathematical descriptions are known, challenges still remain for the *inverse problem* [27], i.e., to construct models and make predictions for observations generated from some unknown chaotic dynamics. The difficulty of the inverse problem partly stems from the characterization of and distinguishing between the model and observational errors. Both types of errors can be stochastic and correlated with complex structures, posing great challenges for the modeling task.

In an effort to investigate some aspects of this inverse problem, in Chapter 5, we analyzed voltage measurement data generated from a laboratory-built electrical circuit that exhibits chaotic behavior. We investigated the tradeoff between short-term prediction and long-term tracking for fitted Gaussian process and neural network models, and studied the effects of observational and model error structures on the likelihood function of the initial state with simulations.

This thesis contains material from two published papers by the author [120, 121]. In particular, Chapter 2 is based on [120], coauthored with Mihai Anitescu; and Chapter 4 uses contents from [121], coauthored with Michael L. Stein. Some material from each of these papers has also been incorporated into this introductory chapter. In the following, we provide more detailed overviews for each chapter.

1.1 Overview of Chapter 2

Chapter 2 considers weakly constrained variational data assimilation. Data assimilation is the process of estimating the underlying states of a physical system based on reconciliation of observations and physical laws governing its evolution [36, 38, 69]. The setup is most commonly described by a state space model with stochastic normal model error and measurement noise [69], for $j = 0, \dots, N$.

$$\begin{gathered} x_0 = x_B + \eta_B, \;\; x_{j+1} = M_j(x_j) + \eta_j, \;\; y_j = H_j(x_j) + \varepsilon_j, \\ \eta_B \sim \mathcal{N}(\mathbf{0}_J, Q_B), \;\; \eta_j \sim \mathcal{N}(\mathbf{0}_J, Q_j) \;\; \varepsilon_j \sim \mathcal{N}(\mathbf{0}_L, R_j). \end{gathered}$$

We are interested in the state estimation problem: given the observations y_j, the distributions of the background term η_B, model error η_j and observation error ε_j, determine the state x_j trajectory that best explains the observations. The problem is named data assimilation or 4DVar [36, 38, 69, 93] in atmospheric sciences applications. In this work we focus on *variational methods*: methods that aim to express the minus loglikelihood of the state space model and then minimize it with deterministic methods.

In the limiting case of zero model error, the system is called *strongly constrained* in the sense that every state is determined by the previous one and all states are functions of only the initial state. However, many sources (e.g., missing physics, discretization errors and semi-empirical/parametrized process descriptions) can contribute to model errors that have non-negligible effects [37, 114, 115]. The explicit inclusion of model error leads to the so-called *weakly constrained* [126] model. In the presence of model error, it is no longer possible to constrain every state using the model propagation $M_j(\cdot)$ as in the strongly constrained case, and hence the storage is $N+1$ fold larger since all states $x_0, x_1, \ldots, x_N$ are free variables. In the case of a large J or N, which we are increasingly approaching in atmospheric sciences as more refined physics models are coming online, the sheer amount of storage makes applications to real systems with higher resolution out of practical reach.

To this end, it is recently [7] proposed to reduce memory by using the constraints of the optimality conditions themselves. Enforcing the optimality conditions gives a recursion for computing each state based on the preceding two states, and hence reduces each state to a function of just the initial state. This method results in significant memory savings when N is large. However, the recursive nature of the method opens the door for instability when the estimation horizon increases or under certain model parameters, as discussed in [7]. That is, the recursion may exhibit rapid exponential increase of the solution, resulting in numerical overflow. To control the instability while maintaining the limited-memory property, we propose a multiple shooting approach which improves stability at the cost of modestly increased storage. We prove that for nonlinear systems within a certain regime, the condition number of the multiple shooting method is bounded above with respect to the number of shooting intervals, and hence the method is stable for increasing time horizon. We demonstrate our method using Burgers' equation under different parameter settings including large model error and sparse observations.

1.2 Overview of Chapter 3

In Chapter 3, we consider a temporal decomposition approach for long-horizon dynamic optimization. Long-horizon dynamic optimization problems appear in several application areas [11, 13, 24, 28, 40, 54, 63] and pose significant computational challenges because of the increase in the number of variables in proportion to the number of time periods considered. One such problem is in optimal planning of generation and transmission of electrical power. Such planning task involves a production cost model (PCM) which simulates the operation of generation and transmission systems by finding the least-cost solution to generating sufficient energy to meet demand over hundreds of thousands of periods. The large number of time period and constraints on the states and controls make the problem difficult to solve.

Researchers have therefore sought to identify approaches for long-horizon dynamic optimization that result in efficient temporal parallelism to address this complexity by bringing to bear more computing power. A recent approach for PCMs is to partition the simulation horizon and embed the annual problem into multiple overlapping weekly/monthly problems [11] that compute the contribution of an *inner time interval only* to the overall objective and then add up all these contributions. It is shown empirically that the approximation error decreases rapidly with the increase of the buffer region (the overlapping area) surrounding the inner time interval.

While used to great effect in [11], such temporal decomposition approaches were, up to our work, a heuristic with no theoretical basis. The main aim of this project is to investigate the theoretical properties of this temporal decomposition approach. We consider a long-horizon dynamic optimization problem with quadratic objective, linear dynamics and box constraints on the controls. We prove that, under certain assumptions, the approximation error of the decomposition approach decays exponentially in the size of the buffer region. The exponential decay rate enables one to choose embedding regions much shorter than the length of the full horizon; and since problems on each embedding region can be solved independently, the time to solution is significantly reduced when the approach is implemented in parallel. We demonstrate the theoretical results using a synthetic production cost model with real demand data.

1.3 Overview of Chapter 4

Chapter 4 considers the asymptotic properties of MLE of parameters of the squared exponential covariance function commonly used in modeling the output of the computer experiments. Computer experiments have been used extensively in investigating complex scientific phenomena. Each run of the computer experiments is deterministic in the sense that re-running the same code gives the same outputs. Often, each run of the code is computationally expensive, so a common alternative to running the code at all input values of interest is to run the code at some inputs and make cheaper predictions at others. [104] and [105] propose to model the deterministic computer experiment outputs as a realization of a Gaussian random field with covariance

$$\mathrm{Cov}\left(f(x), f(y)\right) = \theta_0 \prod_{u=1}^{d} e^{-\frac{|x_u - y_u|^{\gamma}}{\theta_u}},$$

where x_u, $y_u \in [0,1]$, $u = 1, \ldots, d$, $\theta_0 > 0$ is the scale parameter and $\theta_u > 0$ are range parameters. When $\gamma = 2$, the Gaussian process with covariance function is infinitely mean square differentiable and thus is an attractive choice when the output surface is known to be smooth [48, 95, 96, 109, 113]. The use of stochastic models provides a statistical basis for experimental design, parameter estimation, interpolation and uncertainty calibration.

Smooth test functions composed of elementary functions (e.g. polynomials, trigonometric functions and exponential functions) are often used as test cases for studying the effectiveness of Gaussian processes in modeling computer experiments. However, little is known about properties of the MLEs when observations are generated by these test functions.

In this project, we consider the asymptotic properties of the MLE when more and more observations are taken on a fixed domain, and the test function takes some simple deterministic form. We prove the asymptotic order of MLE of the scale parameter when the range parameter is fixed, and the test function is a pth order monomial $f(x) = x^p$. We explore the joint estimation of scale and range parameters through exact computation with three commonly used two-dimensional test functions in the computer experiments literature. We compare the MLE with the cross validation estimates in a prediction problem, and investigate the effect of the common approach of including a small nugget to overcome numerical difficulties in computing with the squared exponential covariance. The implications of modeling smooth deterministic outputs using the squared exponential covariance function are discussed based on the theoretical and numerical results.

1.4 Overview of Chapter 5

In Chapter 5, we analyze chaotic observations generated from some unknown dynamical system. The problems considered in the study of observed chaotic data from physical processes include, but are not limited to, calculating geometric and dynamical invariants of an underlying strange attractor [25, 106, 118], modeling the deterministic portion of the dynamical evolution from the observations [34, 66], and constructing a predictive model directly from the observations [27, 43]. In this work, we focus on the aspects of modeling the observed dynamical system and predicting its future evolution. We aim to provide some insights into possible issues in modeling the underlying dynamics and charactering the effects of the stochastic model and observational errors.

We consider analyzing and modeling voltage measurement data generated by a laboratory-built electrical circuit [84]. The observations are relatively smooth, concentrate on a low-dimensional attractor, and exhibit sensitive dependence on initial conditions. The nominal model describing the dynamics of the measurements produces systematic deficiencies when fitted to the data. We hence investigate modeling and prediction using Gaussian process and neural network models, both trained for predicting one-step ahead based on the preceding observations. To investigate the capacities of the predictors in capturing the dynamics, we investigate the tradeoff between one-step prediction and long-term tracking.

One of our goals for analyzing observations generated from some unknown dynamics is to investigate the effects and characterizations of the model and observational errors. We consider this aspect by performing simulations with data generated by our fitted models. The fitted models capture the chaotic feature of the observations, and the simulated data is qualitatively similar to the observations. We explore the effects of different types of model and observational errors on the likelihood function and the identifiability of the initial state.

CHAPTER 2

A LIMITED-MEMORY MULTIPLE SHOOTING METHOD FOR WEAKLY CONSTRAINED VARIATIONAL DATA ASSIMILATION

2.1 Introduction

Data assimilation is the process of estimating the underlying states of a physical system based on reconciliation of observations and physical laws governing its evolution [36, 38, 69]. The setup is most commonly described by a state space model with stochastic normal model error and measurement noise [69],

$$x_0 = x_B + \eta_B, \;\; x_{j+1} = M_j(x_j) + \eta_j, \;\; y_j = H_j(x_j) + \varepsilon_j, \tag{2.1.1}$$

$$\eta_B \sim \mathcal{N}(\mathbf{0}_J, Q_B), \;\; \eta_j \sim \mathcal{N}(\mathbf{0}_J, Q_j) \;\; \varepsilon_j \sim \mathcal{N}(\mathbf{0}_L, R_j). \tag{2.1.2}$$

where $x_j \in \mathbb{R}^J, y_j \in \mathbb{R}^L$. The mapping $M_j(\cdot) : \mathbb{R}^J \to \mathbb{R}^J$ models the physical law governing the evolution of the system dynamics, typically discretizations of partial differential equations. We assume $M_j(\cdot)$ is at least twice continuously differentiable. The random variable η_j models the stochastic model error and has a normal distribution with mean $\mathbf{0}_J$ and covariance $Q_j \in \mathbb{R}^{J\times J}$. The random variable η_B models the initial state as a normal distribution with mean x_B and covariance $Q_B \in \mathbb{R}^{J\times J}$. The function $H_j(\cdot) : \mathbb{R}^J \to \mathbb{R}^L$ maps the states into observed quantities, whereas ε_j models measurement error that has mean $\mathbf{0}_L$ and covariance $R_j \in \mathbb{R}^{L\times L}$. We also assume all covariance matrices to be positive definite.

With these definitions, we are interested in the *state estimation problem* [108]: We are given the background mean state x_B; evolution function $M_j(\cdot)$; measurement operator $H_j(\cdot)$; measured quantities y_j; and covariance matrices for background error, Q_B, model error, Q_j for $j = 0, 1, \ldots, N-1$, and measurement error, R_j for $j = 0, 1, \ldots, N$ at $N+1$ equally spaced time points. We want to determine the state trajectory $x_0, x_1, \ldots, x_N$ that best explains the data y_j under these assumptions. The problem is also named *data assimilation* or *4DVar* [36, 38, 69, 93] in atmospheric sciences applications, when $M_j(\cdot)$ is obtained from the discretization of 3D dynamics. In particular, we will focus on the circumstance where we are memory-limited, and thus we may be unwilling to simultaneously store the entire trajectory vector because of the $O(JN)$ memory requirements.

In the limiting case of $Q_j = \mathbf{0}_{J\times J}$, and thus $\eta_j = \mathbf{0}_J$, the system is called *strongly constrained* in the sense that every state is determined by the previous one and all states are functions of only the initial state x_0. However, many sources (e.g., missing physics, discretization errors and semi-empirical/parametrized process descriptions) can contribute to model errors that have non-negligible effects [37, 114, 115]. The explicit inclusion of the model error term in the physical evolution [47, 97, 98] leads precisely to (2.1.1)–(2.1.2). In atmospheric sciences, such models are called *weakly constrained* [126]. We note that, although the mean-zero, Gaussian, temporally uncorrelated model error, as described in (2.1.2), is commonly used in a typical operational setting [73, 114], the actual model error can exhibit non-zero mean, non-Gaussianity and temporal correlation. The mean term can be added to the dynamics to reduce the problem back to the form in (2.1.1)–(2.1.2). If the model error exhibits temporal correlation, then this can be accommodated by means of a shaping filter whereby the dynamics of the noise itself is modeled with an autoregressive-type approach and adjoined to the system dynamics [33, 126]. This situation can be again represented with our formulation (2.1.1)–(2.1.2) by using a larger system. The matrix Q_j can be any positive definite matrix. It thus can model a rich set of spatial correlations. On the other hand, to not affect the storage considerations of this project, we need to be able to apply it and its inverse using no more than the storage of a few state vectors. This is certainly the case if Q_j is sparse in a natural basis, such as the canonical or spectral (Fourier) basis. The latter case is one of the most frequently posited proposals for model error in atmospheric sciences [31, 85, 126]. When it comes to non-Gaussian noise, however, the extension of our formulation is not trivial. In our estimation it is likely that, if the distribution of the noise can be written explicitly and depends only on the state x_k, our recursive multiple shooting approach could apply as well. The specific form of the recursions and the numerical properties, however, would be quite different, so that direction would at least require a new analysis. Nevertheless, we note that the vast majority of current proposals for model error are done in terms of Gaussian distributions [31, 80, 85, 87, 115, 116, 126, 127]. We thus conclude that the formulation (2.1.1)–(2.1.2) can accommodate several cases and extensions of interest.

The paradigm (2.1.1)–(2.1.2) is called a *state space model*, and it is one of the most studied state estimation paradigms [69]. It has generated a large number of methods to solve it, including Kalman filters, extended Kalman filters, and particle methods [39, 69]. However, such methods may not be suitable to the kind of problems described here because of reliance on linearity of $M_j(\cdot)$ (Kalman filters) [68]; memory that increases superlinearly with the dimension of x (extended Kalman filters) [64]; and slow convergence, particularly when interested primarily in best estimates (particle methods) [17].

In this work we focus on *variational methods*: methods that aim to express the minus loglikelihood of the state space model (2.1.1)–(2.1.2) and then minimize it with deterministic methods, such as limited-memory BFGS [94]. The objective function of that minimization is the following weakly constrained function [80, 87, 115, 116, 127]:

$$\begin{aligned}
\Gamma(x_{0:N}) &= \frac{1}{N}\left(\sum_{j=0}^{N-1}\left(\gamma_j(x_j)+\phi_j\left(x_j,x_{j+1}\right)\right)+\gamma_N(x_N)\right), && (2.1.3)\\
\text{where } \phi_j(x_j,x_{j+1}) &= \left(x_{j+1}-M_j(x_j)\right)^T Q_j^{-1}\left(x_{j+1}-M_j(x_j)\right)/2, \qquad 0\le j\le N-1\\
\gamma_j(x_j) &= \left(y_j-H_j(x_j)\right)^T R_j^{-1}\left(y_j-H_j(x_j)\right)/2, \qquad 1\le j\le N\\
\gamma_0(x_0) &= (x_0-x_B)^T Q_B^{-1}(x_0-x_B)/2\\
&+ \left(y_0-H_0(x_0)\right)^T R_0^{-1}\left(y_0-H_0(x_0)\right)/2.
\end{aligned}$$

The best estimation of the states $x_0, x_1, \ldots, x_N$ then amounts to minimizing (2.1.3), which is equivalent to maximizing the likelihood of the state space model. In the strongly constrained case, only x_0 is a free variable. Using adjoint approaches for the minimization of (2.1.3) in that limiting case with a checkpointing strategy results in storage requirements of about $O(J \log(N))$ with a recomputation effort that is relatively bounded with J and N [51]. In the presence of model error, however, it is no longer possible to constrain the states by using model propagation, and hence the storage is $N+1$ fold larger since all states $x_0, x_1, \ldots, x_N$ are free variables. In the case of a large J or N, which we are increasingly approaching in atmospheric sciences as more refined physics models are coming online, the sheer amount of storage makes applications to real systems with higher resolution out of practical reach.

To this end, we recently [7] proposed to reduce memory by using the constraints of the optimality conditions themselves.

$$\begin{aligned} 0 &= \nabla_{x_0}\phi_0(x_0, x_1) + \nabla_{x_0}\gamma_0(x_0) && (2.1.4) \\ 0 &= \nabla_{x_j}\phi_j(x_j, x_{j+1}) + \nabla_{x_j}\phi_{j-1}(x_{j-1}, x_j) + \nabla_{x_j}\gamma_j(x_j), \qquad 1 \le j \le N && (2.1.5) \\ 0 &= \nabla_{x_N}\phi_{N-1}(x_{N-1}, x_N) + \nabla_{x_N}\gamma_N(x_N) && (2.1.6) \end{aligned}$$

Enforcing optimality conditions (2.1.4) and (2.1.5) gives a recursion for computing x_1 in terms of x_0 and x_{i+1} in terms of x_i and x_{i-1} for $1 \le i \le N-1$. Hence each state effectively is reduced to a function of just the initial state by using the optimality conditions as constraints; we call this recursively computable function $\lambda_i(x_0)$, $i = 1, 2, \ldots, N$. The objective function then becomes

$$\hat{\Gamma}(x_0) = \frac{1}{N}\left(\sum_{i=0}^{N-1} \gamma_i\left(\lambda_i(x_0)\right) + \phi_i\left(\lambda_i(x_0), \lambda_{i+1}(x_0)\right) + \gamma_N\left(\lambda_N(x_0)\right)\right). \tag{2.1.7}$$

When minimizing (2.1.7) only x_0 is a free variable. The evaluation of the components of $\hat{\Gamma}$ can be carried out by recursion. This results in significant memory savings when N is large.

Quasi-Newton methods such as L-BFGS can be used to minimize (2.1.7).

The recursive nature of the method opens the door for instability when the time horizon increases or under certain model parameters, as also discussed in [7]. That is, the recursion may exhibit rapid exponential increase of the solution, resulting in numerical overflow. Numerical experiments show that in the presence of large model error, large observation gap, large time step, or increased time horizon, the method may encounter such stability issues and fail to progress. The method that minimizes (2.1.7) in [7] uses essentially a single shooting idea. Each initial state x_0 determines the whole trajectory through $\lambda_i(x_0)$, and the optimality is found by satisfying the optimality condition at the end $\nabla_{x_N}\phi_{N-1} + \nabla_{x_N}\gamma_N = 0$. To control the instability that is induced by this recursion, we propose a multiple shooting approach for which multiple restart points across the whole horizon are used. We call such restart points *checkpoints*, given their identical functionality in adjoint calculations [51]. Each checkpoint sequence determines a "shooting" segment of the trajectory, and optimality is achieved by both minimizing the resulting function and matching at each checkpoint. To compute the function and its gradients on a shooting interval, we use a recursion like (2.1.5) restarted at the last checkpoint: a "shooting" approach. Employing checkpoints increases memory usage and introduces constraints to the optimization problem. However, at the cost of modestly increased storage, we expect the method to improve stability by reducing the length of recursion on each segment.

The rest of this chapter is organized as follows. Section 2 describes the low-memory multiple shooting method and proves the consistency of the solution with the full-memory data assimilation method. In Section 3, we show that for nonlinear systems within a certain regime, the condition number of the multiple shooting method is bounded above with respect to the number of shooting intervals. Section 4 describes a recursive limited-memory algorithm to evaluate the descent direction of the resulting optimization problem in preparation for numerical experiments. Section 5 presents numerical experiments that implement the multiple shooting method for Burgers' equation under different parameter settings. Improvements and limitations are discussed in the conclusion.

2.2 Multiple shooting approach

We note that the recursion defining x_{j+1} through (2.1.5) is a two-term recursion; therefore a checkpointing approach here would need two consecutive states. In the following, d pairs of checkpoints $\{x_{P_1-1}, x_{P_1}, \ldots, x_{P_d-1}, x_{P_d}\} \in \mathbb{R}^{2dJ}$ are equally spaced across the entire state. To simplify the discussion, we assume that the number of states on each shooting interval is constant; we let $k = N/(d+1)$ be that number. We also denote $P_0 = 0$ and $P_{d+1} = N$. For each shooting interval $[x_{P_i}, x_{P_{i+1}}]$ we define by $\hat{\Gamma}_i$ the component of the objective function (2.1.3) attached to that interval:

$$\begin{aligned}
\hat{\Gamma}_0(x_0) &= \frac{1}{N}\left(\sum_{j=0}^{P_1-1} \gamma_j(\widetilde{x}_j(x_0)) + \phi_j(\widetilde{x}_j(x_0), \widetilde{x}_{j+1}(x_0))\right), && \text{(2.2.1a)}\\
\hat{\Gamma}_i(x_{P_i-1}, x_{P_i}) &= \frac{1}{N}\left(\sum_{j=P_i}^{P_{i+1}-1} \gamma_j(\widetilde{x}_j(x_{P_i-1}, x_{P_i}))\right. && \text{(2.2.1b)}\\
&\qquad \left.+\phi_j(\widetilde{x}_j(x_{P_i-1}, x_{P_i}), \widetilde{x}_{j+1}(x_{P_i-1}, x_{P_i}))\right), \qquad 1 \le i \le d-1 &&\\
\hat{\Gamma}_d(x_{P_d-1}, x_{P_d}) &= \frac{1}{N}\left(\sum_{j=P_d}^{N-1} \gamma_k(\widetilde{x}_j(x_{P_i-1}, x_{P_i}))\right. && \text{(2.2.1c)}\\
&\qquad \left.+\phi_j(\widetilde{x}_j(x_{P_i-1}, x_{P_i}), \widetilde{x}_{j+1}(x_{P_i-1}, x_{P_i})) + \gamma_N(\widetilde{x}_N(x_{P_d}))\right). &&
\end{aligned}$$

The mappings $\widetilde{x}_j(x_{P_i-1}, x_{P_i})$ are defined implicitly from the optimality conditions (2.1.4) and (2.1.5). This step is possible as soon as $\nabla_{x_j}\phi(x_j, x_{j+1}) = \nabla_{x_j} M_j(x_j) Q_j^{-1}(x_{j+1} - M_j(x_j))$ is invertible in x_{j+1}. This is equivalent to requiring that $\nabla_{x_j} M_j(x_j) Q_j^{-1}$ be an invertible matrix. Since $M_j(\cdot)$ are propagating operators, they can be assumed to be invertible from properties of dynamical systems (see also the discussion at the beginning of [7, §3]). Since the covariance matrix Q_j is assumed to be positive definite, it immediately follows that the recursion (2.1.5) is uniquely solvable in x_{j+1}.

At points immediately following the checkpoints, the mappings $\widetilde{x}_{P_i+1}(x_{P_i-1}, x_{P_i})$ are the solution of the optimality conditions (2.1.4) and (2.1.5) at checkpoint P_i:

$$\begin{aligned} 0 &= \nabla_{x_0}\gamma_{x_0}(x_0) + \nabla_{x_0}\phi_{P_0}(x_0, \widetilde{x}_1) && (2.2.2) \\ 0 &= \nabla_{x_{P_i}}\phi_{P_{i-1}}(x_{P_i-1}, x_{P_i}) + \nabla_{x_{P_i}}\gamma_{P_i}(x_{P_i}) + \nabla_{x_{P_i}}\phi_{P_i}(x_{P_i}, \widetilde{x}_{P_i+1}), && (2.2.3) \end{aligned}$$

for $i = 1, \ldots, d$. At all other points, $\widetilde{x}_j(x_{P_i-1}, x_{P_i})$ is defined recursively from $\widetilde{x}_{j-1}(x_{P_i-1}, x_{P_i})$ and $\widetilde{x}_{j-2}(x_{P_i-1}, x_{P_i})$ by using the optimality conditions (2.1.5) as follows:

$$0 = \nabla_{x_j}\phi_{j-1}(\widetilde{x}_{j-1}, \widetilde{x}_j) + \nabla_{x_j}\gamma_j(\widetilde{x}_j) + \nabla_{x_j}\phi_j(\widetilde{x}_j, \widetilde{x}_{j+1}), \qquad (2.2.4)$$

for $P_i < j \leq P_{i+1} - 1$, $i = 0, \ldots, d$. Under model (2.1.1), the recursions (2.2.2)–(2.2.4) can be written at points immediately following checkpoints as

$$\begin{aligned} \widetilde{x}_1(x_0) &= M_0(x_0) + Q_0\nabla^{-T}M_0(x_0)Q_B^{-1}(x_0 - x_B) && (2.2.5) \\ &\quad -Q_0\nabla^{-T}M_0(x_0)\nabla^T H_0(x_0)R_0^{-1}(y_0 - H_0(x_0)), \\ \widetilde{x}_{P_i+1}(x_{P_i}, x_{P_i-1}) &= M_{P_i}(x_{P_i}) && (2.2.6) \\ &\quad +Q_{P_i}\nabla^{-T}M_{P_i}(x_{P_i})Q_{P_i-1}^{-1}(x_{P_i} - M_{P_i-1}(x_{P_i-1})) \\ &\quad -Q_{P_i}\nabla^{-T}M_{P_i}(x_{P_i})\nabla^T H_{P_i}(x_{P_i})R_{P_i}^{-1}(y_{P_i} - H_{P_i}(x_{P_i})), \end{aligned}$$

for $i = 1, 2, \ldots, d$. At all other points between checkpoints we obtain

$$\begin{aligned} \widetilde{x}_{j+1}(\widetilde{x}_j, \widetilde{x}_{j-1}) &= M_j(\widetilde{x}_j) + Q_j\nabla^{-T}M_j(\widetilde{x}_j)Q_{j-1}^{-1}(\widetilde{x}_j - M_{j-1}(\widetilde{x}_{j-1})) && (2.2.7) \\ &- Q_j\nabla^{-T}M_j(\widetilde{x}_j)\nabla^T H_j(\widetilde{x}_j)R_j^{-1}(y_j - H_j(\widetilde{x}_j)). \end{aligned}$$

Repeated use of (2.2.7) together with (2.2.5) and (2.2.6) results in computing all mappings $\widetilde{x}_j(x_{P_i-1}, x_{P_i})$

Then, by gathering the objective function components (2.2.1) and by imposing matching constraints at the checkpoint pairs, we obtain the following multiple shooting optimization problem.

$$
\begin{aligned}
\min \quad & \widetilde{\Gamma}(x_0, x_{P_1-1}, x_{P_1}, \ldots, x_{P_d-1}, x_{P_d}) \overset{\Delta}{=} \hat{\Gamma}_0(x_0) + \sum_{i=1}^{d} \hat{\Gamma}_i(x_{P_i-1}, x_{P_i}) && \text{(2.2.8a)} \\
\text{s.t.} \quad & c_1(x) = x_{P_1} - \widetilde{x}_{P_1}(x_0) = 0 && \text{(2.2.8b)} \\
& g_1(x) = x_{P_1-1} - \widetilde{x}_{P_1-1}(x_0) = 0 && \text{(2.2.8c)} \\
& c_{i+1}(x) = x_{P_{i+1}} - \widetilde{x}_{P_{i+1}}(x_{P_i-1}, x_{P_i}) = 0, \qquad 1 \leq i \leq d-1 && \text{(2.2.8d)} \\
& g_{i+1}(x) = x_{P_{i+1}-1} - \widetilde{x}_{P_{i+1}-1}(x_{P_i-1}, x_{P_i}) = 0, \qquad 1 \leq i \leq d-1 && \text{(2.2.8e)}
\end{aligned}
$$

The Lagrangian associated with the constraint problem (2.2.8) is

$$
L(x, \lambda, \psi) = \widetilde{\Gamma}(x) - \sum_{i=1}^{d} \lambda_i^T c_i(x) - \sum_{i=1}^{d} \psi_i^T g_i(x), \tag{2.2.9}
$$

where $x = (x_0, x_{P_1-1}, x_{P_1}, \ldots, x_{P_d-1}, x_{P_d})$ and $\lambda_i \in \mathbb{R}^J$, $\psi_i \in \mathbb{R}^J$ are Lagrange multipliers for the equality constraints $c_i(x) = 0$ and $g_i(x) = 0$, $i = 1, 2, \ldots, d$.

We also define the full memory form of the objective functions for each shooting interval as follows:

$$
\begin{aligned}
\Gamma_i(x_{P_i:P_{i+1}}) &= \frac{1}{N} \left(\sum_{j=P_i}^{P_{i+1}-1} \gamma_j(x_k) + \phi_j(x_j, x_{j+1}) \right), \qquad 0 \leq i \leq d, \\
\Gamma_d(x_{P_d:N}) &= \frac{1}{N} \left(\sum_{j=P_d}^{N-1} \gamma_j(x_j) + \phi_j(x_j, x_{j+1}) + \gamma_N(x_N) \right).
\end{aligned} \tag{2.2.10}
$$

We now define a list of symbols frequently used in the rest of the article.

Definition 2.2.1. *For $1 \leq i \leq d$ and $0 \leq j \leq N$, define*

(a)

$$\begin{aligned}
\beta_j(x_j, x_{j+1}) &= \nabla_{x_j}\gamma_j(x_j) + \nabla_{x_j}\phi_j(x_j, x_{j+1}), \qquad 0 \le j \le N-1 \\
\alpha_j(x_{j-1}, x_j) &= \nabla_{x_j}\phi_{j-1}(x_{j-1}, x_j), \qquad 1 \le j \le N \\
\theta_j(x_{j-1}, x_j, x_{j+1}) &= \alpha_j(x_{j-1}, x_j) + \beta_j(x_j, x_{j+1}), \qquad 1 \le j \le N-1 \\
\theta_0(x_0, x_1) &= \beta_0(x_0, x_1); \;\; \theta_N(x_{N-1}, x_N) = \alpha_N(x_{N-1}, x_N) + \nabla_{x_N}\gamma_N(x_N)
\end{aligned}$$

Note that for Γ_i *defined in* (2.2.10), *we have*

$$\begin{aligned}
\left(\frac{\partial \Gamma_i}{\partial(x_{P_i:P_{i+1}})}\right)^T &= \left[\beta_{P_i}^T, \theta_{P_i+1}^T, \ldots, \theta_{P_{i+1}-1}^T, \alpha_{P_{i+1}}^T\right], \qquad 0 \le i \le d-1 \\
\left(\frac{\partial \Gamma_d}{\partial(x_{P_d:N})}\right)^T &= \left[\beta_{P_d}^T, \theta_{P_d+1}^T, \ldots, \theta_{N-1}^T, \theta_N^T\right].
\end{aligned}$$

(b)

$$\begin{aligned}
L_j^{(0)}(x_0) &= \nabla_{x_0}\widetilde{x}_j(x_0), \qquad 0 \le j \\
L_j^{(P_i-1)}(x_{P_i-1}, x_{P_i}) &= \nabla_{x_{P_i-1}}\widetilde{x}_j(x_{P_i-1}, x_{P_i}), \qquad P_i - 1 \le j \\
L_j^{(P_i)}(x_{P_i-1}, x_{P_i}) &= \nabla_{x_{P_i}}\widetilde{x}_j(x_{P_i-1}, x_{P_i}), \qquad P_i - 1 \le j
\end{aligned}$$

(c) Let $\Lambda_i(x_{P_i-1}, x_{P_i})$ *be* $(k+1)J \times 2J$ *dimensional, and let* $\Lambda_0(x_0)$ *be* $(k+1)J \times J$ *dimensional matrices so that*

$$\begin{aligned}
\Lambda_i(x_{P_i-1}, x_{P_i}) &= \frac{\partial(\widetilde{x}_{P_i:P_{i+1}})}{\partial(x_{P_i-1}, x_{P_i})} = \begin{bmatrix} L_{P_i}^{(P_i-1)}(x_{P_i-1}, x_{P_i}) & L_{P_i}^{(P_i)}(x_{P_i-1}, x_{P_i}) \\ L_{P_i+1}^{(P_i-1)}(x_{P_i-1}, x_{P_i}) & L_{P_i+1}^{(P_i)}(x_{P_i-1}, x_{P_i}) \\ \vdots & \vdots \\ L_{P_{i+1}}^{(P_i-1)}(x_{P_i-1}, x_{P_i}) & L_{P_{i+1}}^{(P_i)}(x_{P_i-1}, x_{P_i}) \end{bmatrix}, \\
\Lambda_0(x_0) &= \frac{\partial(\widetilde{x}_{0:P_1})}{\partial(x_0)} = \left[L_0^{(0)}(x_0)^T, L_1^{(0)}(x_0)^T, \ldots, L_{P_1}^{(0)}(x_0)^T\right]^T.
\end{aligned}$$

Note that the first block row of Λ_i is $[0, I_J]$ and the first block row of Λ_0 is I_J. Let $L_0(x_0)$ and $L_i(x_{P_i-1}, x_{P_i})$ be the last two block rows respectively of $\Lambda_0(x_0)$ and $\Lambda_i(x_{P_i-1}, x_{P_i})$ so that

$$\begin{aligned} L_0(x_0) &= \begin{bmatrix} L^{(0)}_{P_1-1}(x_0) \\ L^{(0)}_{P_1}(x_0) \end{bmatrix} \\ L_i(x_{P_i-1}, x_{P_i}) &= \begin{bmatrix} L^{(P_i-1)}_{P_{i+1}-1}(x_{P_i-1}, x_{P_i}) & L^{(P_i)}_{P_{i+1}-1}(x_{P_i-1}, x_{P_i}) \\ L^{(P_i-1)}_{P_{i+1}}(x_{P_i-1}, x_{P_i}) & L^{(P_i)}_{P_{i+1}}(x_{P_i-1}, x_{P_i}) \end{bmatrix}. \end{aligned}$$

(d) *Let $J_i(x_{P_i-1}, x_{P_i})$ and $J_0(x_0)$ be $J(k+1) \times J(k+1)$ dimensional symmetric block tridiagonal matrices defined as follows (with the arguments of $\beta.$,$\theta.$,$\alpha.$ dropped for brevity).*

$$J_i = \begin{bmatrix} \nabla_{x_{P_i}} \beta_{P_i} & \nabla_{x_{P_i+1}} \theta_{P_i} & & \mathbf{0} \\ \nabla_{x_{P_i}} \theta_{P_i+1} & \nabla_{x_{P_i+1}} \theta_{P_i+1} & \ddots & \\ & \ddots & \ddots & \\ & \ddots & \nabla_{x_{P_{i+1}-1}} \theta_{P_{i+1}-1} & \nabla_{x_{P_{i+1}}} \theta_{P_{i+1}-1} \\ \mathbf{0} & & \nabla_{x_{P_{i+1}-1}} \theta_{P_{i+1}} & \nabla_{x_{P_{i+1}}} \alpha_{P_{i+1}} \end{bmatrix}.$$

Note that $J_i = \nabla^2 \Gamma_i$ for $0 \leq i \leq d-1$, and $\nabla^2 \Gamma_d$ differs from J_d by only the last diagonal block element so that $(J_d)_{(k,k)} + \nabla^2_{x_N} \gamma_N = (\nabla^2 \Gamma_d)_{(k,k)}$.

We now illustrate the relationship between the solution of the multiple shooting constrained optimization problem (2.2.8) and the solution of the full-memory data assimilation problem (2.1.3)

Theorem 2.2.2. *Let $x^*_{0:N}$ be a local minimizer of $\Gamma(x_{0:N})$ (2.1.3) that satisfies the first- and second-order sufficient conditions. Let $x^* = (x^*_0, x^*_{P_1-1}, x^*_{P_1}, \ldots, x^*_{P_d-1}, x^*_{P_d})$. Then*

(a) *x^* satisfies the KKT conditions of (2.2.8) with Lagrangian multipliers $\lambda^*_i = -\nabla_{x_{P_i}} \phi_{P_i-1}(x^*_{P_i-1}, x^*_{P_i})$, $\psi^*_i = 0$ for $1 \leq i \leq d$.*

(b) The Hessian matrix of the Lagrangian at optimality satisfies

$$
\begin{aligned}
w^T \nabla_x^2 L(x^*, \lambda^*, \psi^*) w &= \sum_{i=0}^{d} \hat{w}_i^T \Lambda_i^T J_i \Lambda_i \hat{w}_i \\
&+ \left(L_N^{(P_d-1)} w_{2d} + L_N^{(P_d)} w_{2d+1} \right)^T \nabla_{x_N}^2 \gamma_N \left(L_N^{(P_d-1)} w_{2d} + L_N^{(P_d)} w_{2d+1} \right),
\end{aligned}
$$

for $w = (w_1, \ldots, w_{2d+1}) \in \mathbb{R}^{(2d+1)J}$, $\hat{w}_i = (w_{2i}, w_{2i+1})$, $1 \leq i \leq d$, *and* $\hat{w}_0 = w_1$.

(c) x^* *satisfies the second-order sufficient conditions of (2.2.8).*

Proof. The optimality conditions (2.2.2)–(2.2.4) uniquely determine the recursion of $\widetilde{x}_j$, $0 \leq j \leq N$ (Theorem 1 of [7]). Therefore the solution $x_{0:N}^*$ of (2.1.3) coincides with the state propagated starting from the checkpoints by using the recursions (2.2.2)–(2.2.4), namely, $\widetilde{x}_j = x_j^*$ for $0 \leq j \leq N$. In the rest of the proof, the dependence of the symbols defined in Definition 2.2.1 on the checkpoints is suppressed for brevity.

First, we aim to verify part (a), that is, check the KKT conditions with Lagrangian multipliers $\lambda_i^* = -\nabla_{x_{P_i}} \phi_{P_i-1}(x_{P_i-1}^*, x_{P_i}^*)$, $\psi_i^* = 0$ for $1 \leq i \leq d$. Note that from the definitions of α_{P_i}, β_{P_i} (Definition 2.2.1(a)) and optimality conditions (2.1.4) and (2.1.5), we have that for $1 \leq i \leq d$,

$$
\begin{aligned}
\alpha_{P_i}(x_{P_i-1}^*, x_{P_i}^*) + \lambda_i^* &= 0, \qquad (2.2.11a) \\
\beta_{P_i}(x_{P_i}^*, x_{P_i+1}^*) - \lambda_i^* &= 0. \qquad (2.2.11b)
\end{aligned}
$$

By the chain rule and from the definition of the constraints (2.2.8b) and (2.2.8c) and Definitions 2.2.1(a) and (c), the first-order derivatives are

$$
\begin{aligned}
\nabla_{x_0} L(x^*, \lambda^*, \psi^*) &= \nabla_{x_0} \hat{\Gamma}_0(x_0^*) - \nabla_{x_0} c_1(x^*) \lambda_1^* - \nabla_{x_0} g_1(x^*) \psi_1^* \qquad (2.2.12) \\
&= \left(\frac{\partial(\widetilde{x}_{0:P_1})}{\partial(x_0)} \right)^T \frac{\partial \Gamma_0}{\partial(x_{0:P_1})} + L_{P_1}^{(0)T} \lambda_1^* + L_{P_1-1}^{(0)}{}^T \psi_1^* \\
&= \Lambda_0^T V_0 + L_{P_1-1}^{(0)}{}^T \psi_1^*, \text{ where}
\end{aligned}
$$

$$V_0 \;:=\; \begin{bmatrix} \theta_0(x_0^*, \widetilde{x}_1) \\ \theta_1(\widetilde{x}_0, \widetilde{x}_1, \widetilde{x}_2) \\ \vdots \\ \theta_{P_1-1}(\widetilde{x}_{P_1-2}, \widetilde{x}_{P_1-1}, \widetilde{x}_{P_1}) \\ \alpha_{P_1}(\widetilde{x}_{P_1-1}, \widetilde{x}_{P_1}) + \lambda_i^* \end{bmatrix}. \tag{2.2.13}$$

Optimality conditions (2.1.4), (2.1.5) and (2.2.11a) imply $V_0 = \mathbf{0}$, and hence we have $\nabla_{x_0} L(x^*, \lambda^*, \psi^*) = \mathbf{0}$.

For $1 \leq i \leq d-1$, from the definition of the constraints (2.2.8d) and (2.2.8e) and Definitions 2.2.1(a) and (c), we obtain that

$$\begin{aligned} \nabla_{(x_{P_i-1}, x_{P_i})} L(x^*, \lambda^*, \psi^*) &= \nabla_{(x_{P_i-1}, x_{P_i})} \hat{\Gamma}_i(x^*_{P_i-1}, x^*_{P_i}) \\ &- \begin{bmatrix} \nabla_{x_{P_i-1}} g_i(x^*)\psi_i^* + \nabla_{x_{P_i-1}} c_{i+1}(x^*)\lambda_{i+1}^* + \nabla_{x_{P_i-1}} g_{i+1}(x^*)\psi_{i+1}^* \\ \nabla_{x_{P_i}} c_i(x^*)\lambda_i^* + \nabla_{x_{P_i}} c_{i+1}(x^*)\lambda_{i+1}^* + \nabla_{x_{P_i}} g_{i+1}(x^*)\psi_{i+1}^* \end{bmatrix} \\ &= \left(\frac{\partial(\widetilde{x}_{P_i:P_{i+1}})}{\partial(x_{P_i-1}, x_{P_i})}\right)^T \frac{\partial \Gamma_i}{\partial(x_{P_i:P_{i+1}})} - \begin{bmatrix} \psi_i^* - L_{P_{i+1}-1}^{(P_i-1)\,T} \psi_{i+1}^* - L_{P_{i+1}}^{(P_i-1)T} \lambda_{i+1}^* \\ \lambda_i^* - L_{P_{i+1}-1}^{(P_i)\;\;T} \psi_{i+1}^* - L_{P_{i+1}}^{(P_i)\;T} \lambda_{i+1}^* \end{bmatrix} \\ &= \Lambda_i^T V_i - \begin{bmatrix} \psi_i^* - L_{P_{i+1}-1}^{(P_i-1)\,T} \psi_{i+1}^* \\ -L_{P_{i+1}-1}^{(P_i)\;\;T} \psi_{i+1}^* \end{bmatrix}, \text{ where} \end{aligned} \tag{2.2.14}$$

$$V_i \;:=\; \begin{bmatrix} \beta_{P_i}(x^*_{P_i}, \widetilde{x}_{P_i+1}) - \lambda_i^* \\ \theta_{P_i+1}(\widetilde{x}_{P_i}, \widetilde{x}_{P_i+1}, \widetilde{x}_{P_i+2}) \\ \vdots \\ \theta_{P_{i+1}-1}(\widetilde{x}_{P_{i+1}-2}, \widetilde{x}_{P_{i+1}-1}, \widetilde{x}_{P_{i+1}}) \\ \alpha_{P_{i+1}}(\widetilde{x}_{P_{i+1}-1}, \widetilde{x}_{P_{i+1}}) + \lambda_{i+1}^* \end{bmatrix}. \tag{2.2.15}$$

Optimality conditions (2.1.4) and (2.1.5) and (2.2.11a) and (2.2.11b) imply that $V_i = \mathbf{0}$, and hence we have $\nabla_{(x_{P_i-1}, x_{P_i})} L(x^*, \lambda^*, \psi^*) = \mathbf{0}$.

For the last shooting interval, from the definition of the constraints (2.2.8d) and (2.2.8e) and Definitions 2.2.1(a) and (c), we obtain that

$$\nabla_{(x_{P_d-1},x_{P_d})} L(x^*,\lambda^*,\psi^*) = \nabla_{(x_{P_d-1},x_{P_d})} \hat{\Gamma}_d(x^*_{P_d-1},x^*_{P_d}) \tag{2.2.16}$$

$$- \begin{bmatrix} \nabla_{x_{P_d-1}} g_d(x^*)\psi^*_d \\ \nabla_{x_{P_d}} c_d(x^*)\lambda^*_d \end{bmatrix} = \Lambda_d^T V_d - \begin{bmatrix} \psi^*_d \\ \mathbf{0} \end{bmatrix}, \text{ where}$$

$$V_d := \begin{bmatrix} \beta_{P_d}(x^*_{P_d}, \widetilde{x}_{P_d+1}) - \lambda^*_d \\ \theta_{P_d+1}(\widetilde{x}_{P_d}, \widetilde{x}_{P_d+1}, \widetilde{x}_{P_d+2}) \\ \vdots \\ \theta_{N-1}(\widetilde{x}_{N-2}, \widetilde{x}_{N-1}, \widetilde{x}_N) \\ \theta_N(\widetilde{x}_{N-1}, \widetilde{x}_N) \end{bmatrix}. \tag{2.2.17}$$

Optimality conditions (2.1.5) and (2.1.6) and (2.2.11b) imply that $V_d = \mathbf{0}$, and hence we have $\nabla_{(x_{P_d-1},x_{P_d})} L(x^*,\lambda^*,\psi^*) = \mathbf{0}$. This completes the proof of part (a).

We now derive the Hessian matrix. For $1 \le i \le d$, directly applying the chain rule to (2.2.12) and (2.2.14), we note that $V_i = \mathbf{0}$ for $0 \le i \le d-1$ give that $\nabla^2_{x_0} L(x^*,\lambda^*,\psi^*) = \Lambda_0^T J_0 \Lambda_0$ and that $\nabla^2_{(x_{P_{i}-1},x_{P_i})} L(x^*,\lambda^*,\psi^*) = \Lambda_i^T J_i \Lambda_i$ for $1 \le i \le d-1$.

For the last shooting interval, applying the chain rule to (2.2.16) and from Definitions 2.2.1(a) and (d) and the fact that $V_d = \mathbf{0}$, we obtain that

$$\nabla^2_{(x_{P_d-1},x_{P_d})} L(x^*,\lambda^*,\psi^*) \;=\; \Lambda_d^T J_d \Lambda_d + \begin{bmatrix} L_N^{(P_d-1)^T} \\ L_N^{(P_d)^T} \end{bmatrix} \nabla^2_{x_N}\gamma_N \begin{bmatrix} L_N^{(P_d-1)} L_N^{(P_d)} \end{bmatrix}.$$

Since the constraints are separable, there are no cross terms in the Hessian matrix.

For $w = (w_1,\ldots,w_{2d+1}) \in \mathbb{R}^{(2d+1)J}$, we define $\hat{w}_i = (w_{2i}, w_{2i+1})$ for $1 \le i \le d$ and

$\hat{w}_0 = w_1$. Then we have that

$$\begin{aligned} w^T\nabla_x^2 L(x^*,\lambda^*,\mu)w &= \sum_{i=0}^{d} \hat{w}_i^T \Lambda_i^T J_i \Lambda_i \hat{w}_i \\ &+ \left(L_N^{(P_d-1)} w_{2d} + L_N^{(P_d)} w_{2d+1}\right)^T \nabla_{x_N}^2 \gamma_N \left(L_N^{(P_d-1)} w_{2d} + L_N^{(P_d)} w_{2d+1}\right). \end{aligned} \tag{2.2.18}$$

This completes the proof of part (b).

The critical cone at optimality, from Definition 2.2.1(d) and (2.2.8d) and (2.2.8e), is

$$\begin{aligned} C(x^*,\lambda^*,\psi^*) &= \{w \in \mathbb{R}^{(2d+1)J} : \nabla c_i(x^*)w = 0, \nabla g_i(x^*)w = 0, 1 \le i \le d\} \\ &= \{\hat{w} \in \mathbb{R}^{(2d+1)J} : \hat{w}_i = L_{i-1}\hat{w}_{i-1}, 1 \le i \le d\}. \end{aligned} \tag{2.2.19}$$

We define the vector $u \in \mathbb{R}^{(N+1)J}$ by

$$u_j = \begin{cases} L_j^{(0)} w_1, & 0 \le j \le P_1 \\ L_j^{(P_i-1)} w_{2i} + L_j^{(P_i)} w_{2i+1}, & P_i + 1 \le j \le P_{i+1}, 1 \le i \le d \end{cases}$$

so that for $0 \le i \le d$,

$$\Lambda_i \hat{w}_i = \left[w_{2i+1}^T, u_{P_i+1}^T, \ldots u_{P_{i+1}}^T\right]^T. \tag{2.2.20}$$

From Definition 2.2.1(c) the first block row of Λ_i is $[0, I_J]$ for $1 \le i \le d$, and I_J for $i = 0$. Now we consider $w \in C(x^*,\lambda^*,\psi^*)$ and $w \ne \mathbf{0}$. This implies that $w_1 \ne \mathbf{0}$; and since $u_0 = w_1 \ne \mathbf{0}$, we have that $u \ne \mathbf{0}$. Note that since $w \in C(x^*,\lambda^*,\psi^*)$, $L_{P_i}^{(P_i-1)} = \mathbf{0}$, and $L_{P_i}^{(P_i)} = \mathbf{I}_J$, we have from (2.2.19) that $u_{P_i} = w_{2i+1}$, for $1 \le i \le d$. Substituting this equation in (2.2.18), using (2.2.20), using the expression of J_i from Definition 2.2.1(d), and using the fact that from Definition 2.2.1(a) we have that $\nabla_{x_{P_i}}\beta_{P_i} + \nabla_{x_{P_i}}\alpha_{P_i} = \nabla_{x_{P_i}}\theta_{P_i}$ for $1 \le i \le d$, we obtain that

$$w^T\nabla_x^2 L(x^*,\lambda^*,\psi^*)w = u_0^T\nabla_{x_0}\theta_0 u_0 + u_0^T\nabla_{x_1}\theta_0 u_1$$

$$
\begin{aligned}
&+ \sum_{j=1}^{N-1} (u_j^T \nabla_{x_{j-1}} \theta_j u_{j-1} + u_j^T \nabla_{x_j} \theta_j u_j + u_j^T \nabla_{x_{j+1}} \theta_j u_{j+1}) \\
&+ u_N^T \nabla_{x_{N-1}} \theta_N u_{N-1} + u_N^T \nabla_{x_N} \theta_N u_N = u^T \left(\nabla^2_{x_{0:N}} \Gamma(x^*_{0:N}) \right) u > 0.
\end{aligned}
$$

This completes the proof of part (c). □

2.3 Stability analysis

The constrained optimization problem (2.2.8) is now solved with an augmented Lagrangian method. From the Lagrangian function (2.2.9) and using the notations of (2.2.8), we define the augmented Lagrangian function.

$$
L_A(x, \lambda, \psi, \mu) = \widetilde{\Gamma}(x) - \sum_{i=1}^{d} \lambda_i^T c_i(x) - \sum_{i=1}^{d} \psi_i^T g_i(x) + \frac{\mu}{2} \sum_{i=1}^{d} \left(c_i(x)^T c_i(x) + g_i(x)^T g_i(x) \right) \tag{2.3.1}
$$

Here $\mu > 0$ is the penalty parameter that helps enforce feasibility. Augmented Lagrangian theory [94, Theorem 17.5] implies that, under the conditions stated in Theorem 2.2.2, there exists $\bar{\mu}$ so that for all $\mu \geq \bar{\mu}$, when λ^* and ψ^* are the Lagrange multipliers of (2.2.8), the solution x^* of (2.2.8) is a local minimizer of (2.3.1). Hence the convergence is assured without increasing μ indefinitely. The initial choice of μ is in practice a matter of some experimentation (as for example a starting value of $\mu = 0$ may result in divergence for non-convex problems), but it is a standard issue in nonlinear programming theory and practice [94]. Simple algorithms exist to update the value of μ on the way to convergence by increasing it when too-large infeasiblity is detected [94]. In the rest of this section, we assume μ is fixed at some value $\mu \geq \bar{\mu}$.

In this section we investigate the condition number of the Hessian matrix for L_A with respect to the number of shooting intervals. In ideal circumstances, the condition number would be bounded above by a constant and thus would prevent exponential growth of the solution in time, which is the signature of instability discussed in §2.1. Our aim is thus to identify under what circumstances this favorable situation can occur.

For this analysis we use several simplifications to our approach. While our investigations have indicated that similar results can be obtained without making the simplifications, leaving them out would significantly complicate and extend the analysis. We thus keep the number of time points in each shooting interval fixed at k, and we use for all d shooting intervals a fixed time step Δt. Since k is fixed, d grows linearly with N. We consider a constant covariance matrix for model error Q and observation error R for all time steps. The observation mapping is time-dependent linear; that is, $H_i(x_i) = B_i x_i$ for all $0 \le i \le N$ and some $B_i \in \mathbb{R}^{L \times J}$. Note that we allow observation gaps in time, which can be modeled by setting some B_i and the respective observations to $\mathbf{0}$. Theorem 2.2.2(b), definitions of the constraints (2.2.8b)–(2.2.8e) and of the critical cone (2.2.19), and Definition 2.2.1(c) imply that the Hessian for L_A at optimality satisfies

$$\begin{aligned} w^T \nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu) w &= \sum_{i=0}^{d} \hat{w}_i^T \Lambda_i^T J_i \Lambda_i \hat{w}_i + \mu \sum_{i=1}^{d} \|\hat{w}_i - L_{i-1}\hat{w}_{i-1}\|^2 \quad (2.3.2) \\ + \ \ & \left(L_N^{(P_d-1)} w_{2d} + L_N^{(P_d)} w_{2d+1}\right)^T B_N^T R^{-1} B_N \left(L_N^{(P_d-1)} w_{2d} + L_N^{(P_d)} w_{2d+1}\right) \end{aligned}$$

for any $w = (w_1, \ldots, w_{2d+1}) \in \mathbb{R}^{(2d+1)J}$, where we denote $\hat{w}_0 = w_1$, $\hat{w}_i = (w_{2i}, w_{2i+1})$ for $1 \le i \le d$.

We now introduce the definition of the observability matrix for each shooting interval, which is based on the standard one for the linearized system on a given system trajectory [67].

Definition 2.3.1. *For each $0 \le i \le d$, $P_i \le j$, denote*

$$\prod_{l=P_i}^{j} \nabla M_l(x_l) = \nabla M_j(x_j) \nabla M_{j-1}(x_{j-1}) \ldots \nabla M_{P_i}(x_{P_i}).$$

Define

$$C_i^T(x) = \left[B_{P_i}^T, \left(B_{P_i+1}\nabla M_{x_{P_i}}(x_{P_i})\right)^T, \ldots, \left(B_{P_i+k-2}\prod_{l=P_i}^{P_i+k-3}\nabla M_l(x_l)\right)^T \right]$$

as the observability matrix for the (i+1)th shooting interval.

For our work, the importance of the observability condition is that it will ensure that the objective function of (2.1.3) when applied to the linearized system is positive definite on one shooting interval.

Lemma 2.3.2. *$C_i(x)$ being full rank is equivalent to for any $0 \neq \boldsymbol{w} \in \mathbb{R}^{kJ}$ and $0 \leq i \leq d$,*

$$\begin{aligned} Q(\boldsymbol{w}) &:= \sum_{j=P_i}^{P_i+k-2} \left(\left(w_{j+1} - \nabla M_j(x_j)w_j\right)^T Q^{-1}\left(w_{j+1} - \nabla M_j(x_j)w_j\right) + w_j^T B_j^T R^{-1} B_j w_j \right) \\ &> 0. \end{aligned}$$

Proof. Suppose there exists $\mathbf{0} \neq s_0 \in \mathbb{R}^J$ such that $C_i s_0 = \mathbf{0}$. Then we define $\mathbf{s} = (s_{P_i}, \ldots, s_{P_i+k-1}) \in \mathbb{R}^{kJ}$ such that $s_{P_i} = s_0$, $s_{P_i+j} = \prod_{l=P_i}^{P_i+j-1}\nabla M_l(x_l)s_0$ for $1 \leq j \leq k-1$. Note that the assumption $C_i s_0 = \mathbf{0}$ and the definition of $\mathbf{s}$ imply that

$$\mathbf{0} = B_{P_i}s_0 = B_{P_i}s_{P_i}, \;\; \mathbf{0} = B_{P_i+j}\prod_{l=P_i}^{P_i+j-1}\nabla M_l(x_l)s_0 = B_{P_i+j}s_{P_i+j}, \; \forall 1 \leq j \leq k-2. \quad (2.3.3)$$

Then, (2.3.3) and the definition of $\mathbf{s}$ give that $Q(\mathbf{s}) = 0$. Note that $\mathbf{s} \neq \mathbf{0}$ since $s_0 \neq \mathbf{0}$.

On the other hand, suppose $Q(\mathbf{s}) = 0$ for some $\mathbf{0} \neq \mathbf{s} = (s_{P_i}, \ldots, s_{P_i+k-1}) \in \mathbb{R}^{kJ}$. Then $B_j s_j = \mathbf{0}$ and $s_{j+1} = \nabla M_j(x_j)s_j$ for $P_i \leq j \leq P_i + k - 2$. Then we have

$$\mathbf{0} = B_{P_i}s_{P_i}, \;\; \mathbf{0} = B_{P_i+j}\prod_{l=P_i}^{P_i+j-1}\nabla M_l(x_l)s_{P_i}, \; \forall 1 \leq j \leq k-2. \qquad (2.3.4)$$

Then, (2.3.4) implies that $C_i s_{P_i} = \mathbf{0}$. Note that $s_{P_i} \neq \mathbf{0}$ because otherwise $\mathbf{s} = \mathbf{0}$. □

A full-rank result holds for the Jacobian matrix of the recursion.

Lemma 2.3.3. *$\Lambda_i(x_{P_i-1}, x_{P_i})$ is full rank for $1 \le i \le d$,*

Proof. Adapting optimality recursion (2.2.6) to our simplified model gives

$$\begin{aligned} \widetilde{x}_{P_i+1} &= M_{P_i}(x_{P_i}) + Q\nabla^{-T} M_{P_i}(x_{P_i}) B_{P_i}^T R^{-1} \left(B_{P_i} x_{P_i} - y_{P_i} \right) \\ &+ Q\nabla^{-T} M_{P_i}(x_{P_i}) Q^{-1} \left(x_{P_i} - M_{P_i-1}(x_{P_i-1}) \right). \end{aligned}$$

and it implies $L_{P_i+1}^{(P_i-1)} = \frac{\partial \widetilde{x}_{P_i+1}}{\partial x_{P_i-1}} = -Q\nabla^{-T} M_{P_i}(x_{P_i}) Q^{-1} \nabla M_{P_i-1}(x_{P_i-1})$, which is invertible. Since the first block row of $\Lambda_i(x_{P_i-1}, x_{P_i})$ is $(\mathbf{0}, I)$ and $L_{P_i+1}^{(P_i-1)}$ is the (2,1)th block, $\Lambda_i(x_{P_i-1}, x_{P_i})$ is full rank. □

In addition to observability on one shooting interval, we will make slightly stronger assumptions than the ones implied by Lemmas 2.3.2 and 2.3.3. That is, we will assume that those bounds hold uniformly with the shooting interval index i.

Assumption 2.3.4. *There exist $\gamma_k > 0$ and $\rho_k > 0$ dependent on k but not on i, or d, such that for any $N > 0$, we have the following.*

(a) The observability matrices $C_i(x^*)$ are full rank for $0 \le i \le d$.

(b) Under (a), for all $0 \le i \le d$, $w = (w_{P_i}, \ldots, w_{P_i+k-1}) \in \mathbb{R}^{kJ}$,

$$\begin{aligned} \sum_{j=P_i}^{P_i+k-2} &\left(\left(w_{j+1} - \nabla M_j(x_j^*) w_j \right)^T Q^{-1} \left(w_{j+1} - \nabla M_j(x_j^*) w_j \right) + w_j^T B_j^T R^{-1} B_j w_j \right) \\ &\ge \gamma_k \|w\|^2, \end{aligned}$$

(c) $\lambda_{min}(\Lambda_i(x^*_{P_i-1}, x^*_{P_i})^T \Lambda_i(x^*_{P_i-1}, x^*_{P_i})) \ge \rho_k$ for all $1 \le i \le d$.

The second set of assumptions characterizes the system, states, and observations as follows.

Assumption 2.3.5. *For any $N > 0$,*

(a) $\max_{0\le j\le N}\left(\|x_j^*\|, \|x_B\|\right) \le C_1$ *and* $\max_{0\le j\le N}\|y_j\| \le C_2$ *for some constant* $C_1 > 0$ *and* $C_2 > 0$*;*

(b) $\max_{0\le j\le N}\|B_j\|_F \le b_0$ *for some constant* $b_0 > 0$*;*

(c) $\max_{0\le j\le N}\left(\|\nabla M_j(x_j^*)\|_F, \|\nabla^{-1}M_j(x_j^*)\|_F\right) \le A$ *for some constant* $A > 0$*;*

(d) $\max_{0\le j\le N}\left(\|M_j(x_j^*)\|_F\right) \le m_0$ *for some constant* $m_0 > 0$*;*

(e) $\max_{0\le j\le N}\|\nabla_{x_j} vec\left(\nabla^T M_j(x_j^*)\right)\|_F \le A_1$ *for some constant* $A_1 > 0$*.*

In fact, Assumptions 2.3.5(d) and (e) are consequences of (a) and the fact that M_j is at least twice continuously differentiable. We nonetheless state them as assumptions so that the bounds we will use in the proof will have convenient references.

We now make a small nonlinearity assumption. It is shown in [7] that for $s \times s$ matrix S and $s \times 1$ vector u and x, we have

$$\nabla_x(Su) = (u^T \otimes I_s)\nabla_x \mathrm{vec}(S) + S\nabla_x u. \tag{2.3.5}$$

Here we define $M_j^{(2)}(u) := (u^T \otimes I_J)\nabla_{x_j}\mathrm{vec}\left(\nabla^T M_j(x_j^*)\right)$. If u is not a function of x_j, then $M_j^{(2)}(u) = \nabla_{x_j}\left(\nabla^T M_j(x_j^*)u\right)$. Moreover, if the system is linear, then $M_j^{(2)}(u) = \mathbf{0}$; therefore bounds on $M_j^{(2)}(u)$ are bounds limiting nonlinearity. Note that under Assumption 2.3.5(e), denoting $C_0 = A_1\sqrt{J}$, we have for any $N > 0$ that

$$\max_{0\le j\le N}\|M_j^{(2)}(u)\|_F \le A_1\|u^T \otimes I_J\|_F \le C_0\|u\|. \tag{2.3.6}$$

For our proof, however, we need an even sharper restriction for the nonlinearity described below.

Assumption 2.3.6. *There exists $0 \leq b_k < \gamma_k$ such that for any $N > 0$,*

$$\max_{0 \leq j \leq N} \|M_j^{(2)} \left(Q^{-1} \left(x_{j+1}^* - M_j(x_j^*) \right) \right) \|_F \leq b_k,$$

where γ_k is as defined in Assumption 2.3.4.

Other than the observability assumption on each shooting interval, Assumptions 2.3.4 and 2.3.5 are primarily stating uniformity, and thus are only marginally stronger than the existing assumptions. Assumption 2.3.6 on the other hand, puts a relatively hard bound on how much nonlinearity we can tolerate in our analysis. At the end of this section we will discuss the effect of this assumption and its significance.

With these definitions and assumptions, we now proceed to the main results of our work. That is, we now prove that for the nonlinear system satisfying Assumptions 2.3.4, 2.3.5, and 2.3.6, the condition number of the Hessian matrix for the augmented Lagrangian is bounded above. First, we derive a lower bound.

Proposition 2.3.7. *Under Assumptions 2.3.4 and 2.3.6, for any $w \in \mathbb{R}^{kJ}$ and $\|w\| = 1$, we have that $w^T J_i(x_{P_i}^*) w \geq \gamma_k - b_k$ for $0 \leq i \leq d$.*

Proof. Referring back to Definition 2.2.1(a), we have that

$$\begin{aligned}
\nabla_{x_0}\beta_0 &= \nabla^T M_0(x_0^*) Q^{-1} \nabla M_0(x_0^*) + B_0^T R^{-1} B_0 - M_0^{(2)} \left(Q^{-1}(x_1^* - M_0(x_0^*)) \right) + Q_B^{-1}, \\
\nabla_{x_j}\beta_j &= \nabla^T M_j(x_j^*) Q^{-1} \nabla M_j(x_j^*) + B_j^T R^{-1} B_j - M_j^{(2)} \left(Q^{-1}(x_{j+1}^* - M_j(x_j^*)) \right), \\
& \qquad 0 < j \leq N-1 \\
\nabla_{x_j}\alpha_j &= Q^{-1}, \ \ 1 \leq j \leq N, \quad \nabla_{x_{j-1}}\theta_j = -Q^{-1}\nabla^T M_{j-1}(x_{j-1}^*), \ \ 1 \leq j \leq N \\
\nabla_{x_j}\theta_j &= \nabla_{x_j}\alpha_j + \nabla_{x_j}\beta_j, \ \ 0 < j < N \quad \nabla_{x_{j+1}}\theta_j = -\nabla^T M_j(x_j^*) Q^{-1}, \ \ 0 \leq j \leq N-1.
\end{aligned} \tag{2.3.7}$$

So for $\|w\| = 1$, referring to Definition 2.2.1 (d), we have

$$\begin{aligned} w^T J_i(x^*_{P_i})w \;\; \geq \;\; & \sum_{j=P_i}^{P_i+k-2} \Bigg(\left(w_{j+1} - \nabla M_j(x^*_j)w_j\right)^T Q^{-1} \left(w_{j+1} - \nabla M_j(x^*_j)w_j\right) \\ & \qquad + w_j^T B_j^T R^{-1} B_j w_j \Bigg) - \sum_{j=P_i}^{P_i+k-2} w_j^T M_j^{(2)} \left(Q^{-1}\left(x^*_{j+1} - M_j(x^*_j)\right)\right) w_j, \end{aligned}$$

for which equality holds for $1 \leq i \leq d$. For $i = 0$, the difference between the two sides is $w_0^T Q_B^{-1} w_0$, which is non-negative. By Assumption 2.3.4(b) we have that

$$\sum_{j=P_i}^{P_i+k-2} \left(\left(w_{j+1} - \nabla M_j(x^*_j)w_j\right)^T Q^{-1} \left(w_{j+1} - \nabla M_j(x^*_j)w_j\right) + w_j^T B_j^T R^{-1} B_j w_j \right) \geq \gamma_k,$$

and by Assumption 2.3.6 we have that $\left|\sum_{j=P_i}^{P_i+k-2} w_j^T M_j^{(2)} \left(Q^{-1}\left(x^*_{j+1} - M_j(x^*_j)\right)\right) w_j\right| \leq b_k$. Thus Proposition 2.3.7 follows. □

We now derive upper bounds in a series of lemmas.

Lemma 2.3.8. *Under Assumption 2.3.5, for each* $1 \leq i \leq d$, $P_i + 1 \leq j \leq P_i + k$, *and* $p = P_i - 1, P_i$, *we have that* $\|L_j^{(p)}(x^*_{P_i-1}, x^*_{P_i})\|_F \leq C_p^{(j-P_i+1)}$ *and* $\|L_j^{(0)}(x^*_0)\|_F \leq C_p^j$, *where* $C_p > 1$ *is a constant independent of* d.

Proof. For $0 \leq i \leq d$ and $P_i \leq j \leq P_{i+1} - 1$, define

$$F_{ij} \;\; = \;\; \nabla M_j(x^*_j) - Q\nabla_{x_j}\left(\nabla^{-T} M_j(x^*_j) B_j^T R^{-1}(y_j - B_j x^*_j)\right),$$

and for $0 \leq i \leq d$ and $P_i + 1 \leq j \leq P_{i+1} - 1$, define

$$\begin{aligned} G_{ij} \;\; &= \;\; Q\nabla_{x_j}\left(\nabla^{-T} M_j(x^*_j) Q^{-1}\left(x^*_j - M_{j-1}(x^*_{j-1})\right)\right) \\ K_{ij} \;\; &= \;\; -Q\nabla^{-T} M_j(x^*_j) Q^{-1} \nabla M_{j-1}(x^*_{j-1}). \end{aligned}$$

Also define

$$G_{10} \;\; = \;\; Q\nabla x_0 \left(\nabla^{-T} M_0(x_0^*) Q_B^{-1} \left(x_0^* - x_B\right)\right).$$

Then for any $1 \leq i \leq d$ and $P_i + 1 \leq j \leq P_i + k$, from optimality recursions (2.2.7) and the chain rule, the recursion of $L_j^{(P_i)}$ and $L_j^{(P_i-1)}$ can be written as

$$\begin{bmatrix} L_j^{(p)} \\ L_{j-1}^{(p)} \end{bmatrix} = \begin{bmatrix} F_{i,j-1} + G_{i,j-1} & K_{i,j-1} \\ I_J & \mathbf{0} \end{bmatrix} \begin{bmatrix} L_{j-1}^{(p)} \\ L_{j-2}^{(p)} \end{bmatrix}, \tag{2.3.8}$$

where $p = P_i, P_i - 1$. For the initial shooting interval, the recursion runs through $2 \leq j \leq P_1$ and $p = 0$. From (2.2.5), the initialization of the recursion for the initial shooting interval is

$$\begin{bmatrix} L_1^{(0)} \\ L_0^{(0)} \end{bmatrix} = \begin{bmatrix} F_{10} + G_{10} \\ I_J. \end{bmatrix}. \tag{2.3.9}$$

For the other shooting intervals $1 \leq i \leq d$, from (2.2.6), the recursion is initialized by

$$\begin{bmatrix} L_{P_i}^{(P_i-1)} \\ L_{P_i-1}^{(P_i-1)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ I_J \end{bmatrix}, \qquad \begin{bmatrix} L_{P_i}^{(P_i)} \\ L_{P_i-1}^{(P_i)} \end{bmatrix} = \begin{bmatrix} I_J \\ \mathbf{0} \end{bmatrix}. \tag{2.3.10}$$

Now we give upper bounds for the propagation matrices. For some $J \times 1$ vector $v(x_j^*)$, by differentiating both sides of $v(x_j^*) = \nabla^T M_j(x_j^*) \nabla^{-T} M_j(x_j^*) v(x_j^*)$ and using equation (2.3.5), we have that

$$\begin{aligned} \nabla_{x_j} \left(\nabla^{-T} M_j(x_j^*) v(x_j^*)\right) = & -\nabla^{-T} M_j(x_j^*) M_j^{(2)} \left(\nabla^{-T} M_j(x_j^*) v(x_j^*)\right) \\ & + \nabla^{-T} M_j(x_j^*) \nabla v(x_j^*). \end{aligned} \tag{2.3.11}$$

Now we can give bounds to each part involved in the propagation. By equation (2.3.11), Assumption 2.3.5, and equation (2.3.6), we have that

$$\begin{aligned}
& \|\nabla_{x_j}\left(\nabla^{-T} M_j(x_j^*) B_j^T R^{-1}(y_j - B_j x_j^*)\right)\|_F & \text{(2.3.12a)}\\
\leq\ & \|\nabla^{-T} M_j(x_j^*) M_j^{(2)}\left(\nabla^{-T} M_j(x_j^*) B_j^T R^{-1}(y_j - B_j x_j^*)\right)\|_F & \\
+\ & \|\nabla^{-T} M_j(x_j^*) B_j^T R^{-1} B_j\|_F & \text{(2.3.12b)}\\
\leq\ & C_0 A^2 b_0 \|R^{-1}\|_F (C_2 + b_0 C_1) + A b_0^2 \|R^{-1}\|_F, & \\
& \|\nabla_{x_0}\left(\nabla^{-T} M_0(x_0^*) Q_B^{-1}\left(x_0^* - x_B\right)\right)\|_F & \text{(2.3.12c)}\\
\leq\ & \|\nabla^{-T} M_0(x_0^*) M_0^{(2)}\left(\nabla^{-T} M_0(x_0^*) Q_B^{-1}\left(x_0^* - x_B\right)\right)\|_F & \\
+\ & \|\nabla^{-T} M_0(x_0^*) Q_B^{-1}\|_F \leq 2C_0 A^2 \|Q_B^{-1}\|_F C_1 + A\|Q_B^{-1}\|_F, & \\
& \|\nabla_{x_j}\left(\nabla^{-T} M_j(x_j^*) Q^{-1}\left(x_j^* - M_{j-1}(x_{j-1}^*)\right)\right)\|_F & \text{(2.3.12d)}\\
\leq\ & \|\nabla^{-T} M_j(x_j^*) M_j^{(2)}\left(\nabla^{-T} M_j(x_j^*) Q^{-1}\left(x_j^* - M_{j-1}(x_{j-1}^*)\right)\right)\|_F & \\
+\ & \|\nabla^{-T} M_j(x_j^*) Q^{-1}\|_F \leq C_0 A^2 \|Q^{-1}\|_F (C_1 + m_0) + A\|Q^{-1}\|_F. &
\end{aligned}$$

We then have that

$$\begin{aligned}
\|F_{ij}\|_F &\overset{\text{(2.3.12a)}}{\leq} A + \|Q\|_F\left(C_0 A^2 b_0 \|R^{-1}\|_F (C_2 + b_0 C_1) + A b_0^2 \|R^{-1}\|_F\right) := F\\
\|G_{ij}\|_F &\overset{\text{(2.3.12d)}}{\leq} \|Q\|_F\left(C_0 A^2 \|Q^{-1}\|_F (C_1 + m_0) + A\|Q^{-1}\|_F\right) := G_1\\
\|K_{ij}\|_F &\leq A^2 \|Q\|_F \|Q^{-1}\|_F := K\\
\|G_{10}\|_F &\overset{\text{(2.3.12c)}}{\leq} \|Q\|_F\left(2C_0 A^2 \|Q_B^{-1}\|_F C_1 + A\|Q_B^{-1}\|_F\right) := G_0.
\end{aligned}$$

Let $G = \max(G_1, G_0)$. Then, bounding each term in the propagation relations (2.3.8), (2.3.9), and (2.3.10) by its Forbenius norm, we have for $1 \leq i \leq d$, $P_i + 1 \leq j \leq P_i + k$, and $p = P_i - 1, P_i$ that

$$\|L_j^{(p)}\|_F \leq \left\|\begin{bmatrix} L_j^{(p)} \\ L_{j-1}^{(p)} \end{bmatrix}\right\|_F \leq \left\|\begin{bmatrix} F_{i,j-1} + G_{i,j-1} & K_{i,j-1} \\ I_J & \mathbf{0} \end{bmatrix}\right\|_F \left\|\begin{bmatrix} L_{j-1}^{(p)} \\ L_{j-2}^{(p)} \end{bmatrix}\right\|_F$$

$$
\begin{aligned}
&\leq \left(\sqrt{J+K^2+(F+G)^2}\right)^{j-P_i}\sqrt{J} \\
&\leq \left(\sqrt{J+K^2+(F+G)^2}\right)^{j-P_i+1} := C_p^{j-P_i+1},
\end{aligned}
$$

and for the initial shooting interval, similarly we have for $1 \leq j \leq P_1$ that

$$
\begin{aligned}
\|L_j^{(0)}\|_F &\leq \left(\sqrt{J+K^2+(F+G)^2}\right)^{j-1}\sqrt{J+(F+G)^2} \\
&\leq \left(\sqrt{J+K^2+(F+G)^2}\right)^{j} = C_p^j.
\end{aligned}
$$

□

Lemma 2.3.9. *Under Assumptions 2.3.5 and 2.3.6 and using notations in Definition 2.2.1(d), for each* $1 \leq i \leq d$ *we have that* $\|J_0(x_0^*)\|_F, \|J_i(x_{P_i-1}^*, x_{P_i}^*)\|_F \leq C_J$ *for some* $C_J > 0$ *independent of* d.

Proof. Because of the block tridiagonal structure of J_i for $0 \leq i \leq d$, we have that

$$
\begin{aligned}
\|J_i\|_F \quad &\leq \quad \sum_{j=P_i+1}^{P_{i+1}-1}\Big(\|\nabla_{x_{j-1}}\theta_j\|_F + \|\nabla_{x_j}\theta_j\|_F + \|\nabla_{x_{j+1}}\theta_j\|_F\Big) \\
&\quad\quad +\|\nabla_{x_{P_i}}\beta_{P_i}\|_F + \|\nabla_{x_{P_i+1}}\theta_{P_i}\|_F + \|\nabla_{x_{P_{i+1}-1}}\theta_{P_{i+1}}\|_F + \|\nabla_{x_{P_{i+1}}}\alpha_{P_{i+1}}\|_F \\
&\overset{(2.3.7)}{\leq} \quad \sum_{j=P_i+1}^{P_{i+1}-1}\Big(2A\|Q^{-1}\|_F + \|Q^{-1}\|_F + A^2\|Q^{-1}\|_F + b_0^2\|R^{-1}\|_F + b_k\Big) \\
&\quad\quad +2A\|Q^{-1}\|_F + \|Q^{-1}\|_F + A^2\|Q^{-1}\|_F + b_0^2\|R^{-1}\|_F + b_k \\
&\leq \quad k\Big(2A\|Q^{-1}\|_F + \|Q^{-1}\|_F + A^2\|Q^{-1}\|_F + b_0^2\|R^{-1}\|_F + b_k\Big) \\
&:= \quad C_J.
\end{aligned}
$$

□

Proposition 2.3.10. *For any* $w \in \mathbb{R}^{(2d+1)J}$ *and* $\|w\| = 1$, *we have that* $w^T\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)w \leq U_k$ *for some* $U_k > 0$ *independent of* d.

Proof. For $0 \leq i \leq d$, using Lemmas 2.3.8 and 2.3.9 and referring to Definition 2.2.1, we have that $\|J_i\|_F\|\Lambda_i\|_F^2 \leq 2(k+1)C_J C_p^{2k}$. Then, from (2.3.2), it follows that

$$\begin{aligned} w^T\nabla_x^2 L_A(x^*,\lambda^*,\psi^*,\mu)w &= \sum_{i=0}^{d}\hat{w}_i^T\Lambda_i^T J_i\Lambda_i\hat{w}_i + \mu\sum_{i=1}^{d}\|\hat{w}_i - L_{i-1}\hat{w}_{i-1}\|^2 \\ + &\left(L_N^{(P_d-1)}w_{2d} + L_N^{(P_d)}w_{2d+1}\right)^T B_N^T R^{-1} B_N \left(L_N^{(P_d-1)}w_{2d} + L_N^{(P_d)}w_{2d+1}\right) \\ \leq &\ 2(k+1)C_J C_p^{2k} + b_0^2\|R^{-1}\|_F C_p^{2k} + \mu(1+2C_p^k)^2. \end{aligned}$$

Defining U_k to be the last quantity above completes the proof. □

We are now in a position to state and prove our main result.

Theorem 2.3.11. *Under Assumptions 2.3.4, 2.3.5, and 2.3.6, the condition number of the Hessian matrix for the augmented Lagrangian is bounded above independent of the number of shooting intervals, d. That is,*

$$\kappa\Big(\nabla_x^2 L_A(x^*,\lambda^*,\psi^*,\mu)\Big) \leq \frac{U_k}{(\gamma_k - b_k)\min(\rho_k,1)}.$$

Proof. For any $w \in \mathbb{R}^{(2d+1)J}$ and $\|w\| = 1$, using Proposition 2.3.7 and Assumption 2.3.4(c), we have that

$$\begin{aligned} w^T\nabla_x^2 L_A(x^*,\lambda^*,\psi^*,\mu)w &= \sum_{i=0}^{d}\hat{w}_i^T\Lambda_i^T J_i\Lambda_i\hat{w}_i + \mu\sum_{i=1}^{d}\|\hat{w}_i - L_{i-1}\hat{w}_{i-1}\|^2 \\ &+\left(L_N^{(P_d-1)}w_{2d} + L_N^{(P_d)}w_{2d+1}\right)^T B_N^T R^{-1} B_N \left(L_N^{(P_d-1)}w_{2d} + L_N^{(P_d)}w_{2d+1}\right) \\ \geq &\quad \sum_{i=0}^{d}\hat{w}_i^T\Lambda_i^T J_i\Lambda_i\hat{w}_i \geq (\gamma_k - b_k)\sum_{i=0}^{d}\|\Lambda_i\hat{w}_i\|^2 \\ \overset{\textit{Assumption } 2.3.4}{\geq} &\quad (\gamma_k - b_k)\left(\rho_k\sum_{i=1}^{d}\|\hat{w}_i\|^2 + \|\hat{w}_0\|^2\right) \geq (\gamma_k - b_k)\min(\rho_k,1). \end{aligned}$$

Combining with Proposition 2.3.10, we obtain

$$\kappa\Big(\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)\Big) = \frac{\lambda_{max}\left(\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)\right)}{\lambda_{min}\left(\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)\right)} \leq \frac{U_k}{(\gamma_k - b_k)\min(\rho_k, 1)},$$

which completes the proof. □

Discussion. An interpretation of Theorem 2.3.11 is that, under observability Assumption 2.3.4 and small nonlinearity Assumption 2.3.6, the condition number of the multiple shooting problem is bounded above with the number of multiple shooting intervals d. This prevents the exponential increase of the solution, which we define as instability, and thus makes the multiple shooting problem computable. We note that the upper bounds of the lemmas preceding Theorem 2.3.11 allow for exponential increase *within the shooting interval*; but as long as observability holds, this increase stops at the end of a shooting interval. As for Assumptions 2.3.6, we note that the amount of nonlinearity needs to be upper bounded by the lower bound γ_k that is related to observability by Lemma 2.3.2. This points out that the bound on nonlinearity in Assumption 2.3.6 is not absolute; it only needs to be small compared with how much information can be found in the observations. That is, increasing the measurement space would increase the lower eigenvalue of $\sum B_i^T R^{-1} B_i$ and thus γ_k, which in turn would increase the prospects for Assumption 2.3.6 to hold.

Another important question is whether these assumptions are necessary. While an if and only if statement between observability and the bounded condition number of the multiple shooting Lagrangian probably does not hold, some of the assumptions are necessary in the following way. As we can see from Appendix A.1, multiple shooting without observations still results in exponential increase of the condition number and thus of the solution. Therefore some amount of observability, or, otherwise said, state space coverage by data, is necessary. As we can see from Appendix A.2, without multiple shooting the condition number of the Hessian matrix for the single shooting function (2.1.7) also increases exponentially and thus is unstable. We conclude that some form of observability and multiple shooting are necessary to obtain a stability result as Theorem 2.3.11.

2.4 Recursive gradient evaluation

When implementing minimization of the augmented Lagrangian function (2.3.1), gradient evaluation is required. In this section, we describe a recursive method for computing the gradient of (2.3.1) that fits into our memory-saving framework.

First we derive the gradients of the augmented Lagrangian function. Note that $\theta_j(\widetilde{x}_{j-1}, \widetilde{x}_j, \widetilde{x}_{j+1}) = \mathbf{0}$ for all $P_i + 1 \leq j \leq P_{i+1} - 1$, $0 \leq i \leq d$, and $\theta_0(x_0, \widetilde{x}_1) = \mathbf{0}$. For the first interval we obtain that

$$\begin{aligned} \nabla_{x_0} L_A(x, \lambda, \psi, \mu) &= {L_{P_1}^{(0)}}^T \left(\nabla_{x_{P_1}} \phi_{P_1-1}(\widetilde{x}_{P_1-1}, \widetilde{x}_{P_1}) + \lambda_1 - \mu c_1(x) \right) \\ +{L_{P_1-1}^{(0)}}^T (\psi_1 - \mu g_1(x)) \quad &+ {L_{P_1}^{(0)}}^T \left(\cancel{\sum_{j=1}^{P_1-1} \theta_j(\widetilde{x}_{j-1}, \widetilde{x}_j, \widetilde{x}_{j+1})} \right) + \cancel{\theta_0(x_0, \widetilde{x}_1)}. \end{aligned} \tag{2.4.1}$$

For $1 \leq i \leq d-1$, we obtain that

$$\begin{aligned} & \nabla_{x_{P_i-1}} L_A(x, \lambda, \psi, \mu) \\ = \;& {L_{P_{i+1}}^{(P_i-1)}}^T \left(\cancel{\sum_{j=P_i+1}^{P_{i+1}-1} \theta_j(\widetilde{x}_{j-1}, \widetilde{x}_j, \widetilde{x}_{j+1})} + \nabla_{x_{P_{i+1}}} \phi_{P_{i+1}-1}(\widetilde{x}_{P_{i+1}-1}, \widetilde{x}_{P_{i+1}} \right) \\ + \;& {L_{P_{i+1}}^{(P_i-1)}}^T (\lambda_{i+1} - \mu c_{i+1}(x)) + {L_{P_{i+1}-1}^{(P_i-1)}}^T (\psi_{i+1} - \mu g_{i+1}(x)) + \mu g_i(x) - \psi_i. \\ & \nabla_{x_{P_i}} L_A(x, \lambda, \psi, \mu) \\ = \;& {L_{P_{i+1}}^{(P_i)}}^T \left(\cancel{\sum_{j=P_i+1}^{P_{i+1}-1} \theta_j(\widetilde{x}_{j-1}, \widetilde{x}_j, \widetilde{x}_{j+1})} + \nabla_{x_{P_{i+1}}} \phi_{P_{i+1}-1}(\widetilde{x}_{P_{i+1}-1}, \widetilde{x}_{P_{i+1}}) \right) \\ + \;& {L_{P_{i+1}}^{(P_i)}}^T (\lambda_{i+1} - \mu c_{i+1}(x)) + {L_{P_{i+1}-1}^{(P_i)}}^T (\psi_{i+1} - \mu g_{i+1}(x)) \\ + \;& \mu c_i(x) - \lambda_i + \beta_{P_i}(x_{P_i}, \widetilde{x}_{P_i+1}). \end{aligned}$$

For the last shooting interval, we obtain that

$$\begin{aligned}
\nabla_{x_{P_d-1}} L_A(x,\lambda,\psi,\mu) &= L_N^{(P_d-1)^T}\left(\sum_{j=P_d+1}^{N-1}\theta_j(\widetilde{x}_{j-1},\widetilde{x}_j,\widetilde{x}_{j+1})+\theta_N(\widetilde{x}_{N-1},\widetilde{x}_N)\right)\\
&- \psi_d+\mu g_d(x),\\
\nabla_{x_{P_d}} L_A(x,\lambda,\psi,\mu) &= L_N^{(P_d)^T}\left(\sum_{j=P_d+1}^{N-1}\theta_j(\widetilde{x}_{j-1},\widetilde{x}_j,\widetilde{x}_{j+1})+\theta_N(\widetilde{x}_{N-1},\widetilde{x}_N)\right)\\
&+ \beta_{P_d}(x_{P_d},\widetilde{x}_{P_d+1})-\lambda_d.
\end{aligned}$$

Note that the derivatives are composed of a matrix-vector product for which the vector can be computed through one forward recursion similar to the one for the states. The Jacobian matrix $L_{P_{i+1}}^{(P_i)}$, however, needs to also be computed by forward recursion, and it turns out to be dense. The computation thus would require $O(J^2)$ storage and inhibit the low-memory advantage of our approach. Instead, we compute the matrix-vector product using a backward recursion separately on each multiple shooting interval, as follows. Since the evaluation procedure is the same for each interval, we illustrate our method with the first interval (assuming it has length N').

The target of our algorithm is to compute $v^T L_{N'}^{(0)}$ for some constant vector v. This algorithm can then be used to compute the gradient components defined in the beginning of this section. For example, for computing the first component (2.4.1) we note that we have two such matrix-vector products, where N' is, succesively P_1 and P_1-1 and v is succesively $\left(\nabla_{x_{P_1}}\phi_{P_1-1}(\widetilde{x}_{P_1-1},\widetilde{x}_{P_1})+\lambda_1-\mu c_1(x)\right)$ and $(\psi_1-\mu g_1(x))$. Similar embeddings hold for all other gradient components.

The computation of $v^T L_{N'}^{(0)}$ proceeds as follows. The optimality recursion states that $\theta_j(\widetilde{x}_{j-1}(x_0),\widetilde{x}_j(x_0),\widetilde{x}_{j+1}(x_0))=0$ for $1\le j\le N'-1$. Differentiating with respect to x_0 gives

$$L_{j+1}^{(0)} = -(\nabla_{x_{j+1}}\theta_j)^{-1}\left((\nabla_{x_{j-1}}\theta_j)L_{j-1}^{(0)}+(\nabla_{x_j}\theta_j)L_j^{(0)}\right). \tag{2.4.2}$$

Now we write the recursion ansatz and substitute (2.4.2) to obtain

$$\begin{aligned}
v^T L^{(0)}_{N'-l+1} &= c_l^T L^{(0)}_{N'-l} + b_l^T L^{(0)}_{N'-l-1} \\
&= -c_l^T (\nabla_{x_{N'-l}} \theta_{N'-l-1})^{-1} (\nabla_{x_{N'-l-2}} \theta_{N'-l-1}) L^{(0)}_{N'-l-2} \\
&+ \left(b_l^T - c_l^T (\nabla_{x_{N'-l}} \theta_{N'-l-1})^{-1} (\nabla_{x_{N'-l-1}} \theta_{N'-l-1}) \right) L^{(0)}_{N'-l-1} \\
&:= c_{l+1}^T L^{(0)}_{N'-l-1} + b_{l+1}^T L^{(0)}_{N'-l-2}
\end{aligned}$$

for $1 \leq l \leq N' - 2$, where c_{l+1} and b_{l+1} for $l = 2, \ldots, N' - 2$ are defined by sought-after recursions

$$c_{l+1}^T = b_l^T - c_l^T (\nabla_{x_{N'-l}} \theta_{N'-l-1})^{-1} (\nabla_{x_{N'-l-1}} \theta_{N'-l-1}), \tag{2.4.3}$$

$$b_{l+1}^T = -c_l^T (\nabla_{x_{N'-l}} \theta_{N'-l-1})^{-1} (\nabla_{x_{N'-l-2}} \theta_{N'-l-1}). \tag{2.4.4}$$

Then the matrix-vector product of interest can be expressed as $v^T L^{(0)}_{N'} = c^T_{N'-1} L^{(0)}_1 + b^T_{N'-1} L^{(0)}_0$, where $c_{N'-1}$ and $b_{N'-1}$ are obtained through recursions (2.4.3) and (2.4.4). It is a backward recursion with respect to the usage of state information x_j. The initial values for the recursion are

$$c_1^T = -v^T (\nabla_{x'_N} \theta_{N'-1})^{-1} (\nabla_{x_{N'-1}} \theta_{N'-1}), \quad b_1^T = -v^T (\nabla_{x'_N} \theta_{N'-1})^{-1} (\nabla_{x_{N'-2}} \theta_{N'-1}),$$

obtained by total differentiation of $\theta_{N'-1}(\widetilde{x}_{N'-2}(x_0), \widetilde{x}_{N'-1}(x_0), \widetilde{x}_{N'}(x_0)) = 0$.

Since the recursion can be computed separately on each shooting interval, the total storage does not exceed the number of multiple shooting checkpoints plus the length of an interval, which adds up to $2d+1+N/(d+1)$. We can use checkpointing within the shooting interval to reduce the storage even further, but we do not pursue that avenue here.

2.5 Numerical results

In this section, we apply our multiple shooting method to Burgers' equation in order to verify some of our theoretical findings. This is a one-spatial-dimension, time-dependent, partial differential equation that exhibits both diffusion and nonlinear advection. Since implementation of new ideas in an operational environment is a development-intensive process, in many research references discussing new state estimation methods Burgers' equation is considered an important first test of a method [8, 69, 78, 116].

The partial differential equation describing it is the following:

$$\frac{\partial x}{\partial t} + \frac{1}{2}\frac{\partial (x^2)}{\partial \chi} = \nu \frac{\partial^2 x}{\partial \chi^2}; \quad x(0,t) = x(1,t) = 0; \quad x(\chi, 0) = x_0(\chi), \tag{2.5.1}$$

where $\nu = 0.01$ is viscosity coefficient and $(\chi, t) \in (0,1) \times (0,T)$.

We denote by x_j^m the unknown value at grid coordinates $(j\Delta\chi, m\Delta t)$ and $\Delta\chi = 1/J$. We use a centered finite-difference discretization [8]:

$$\frac{x_j^{m+1} - x_j^m}{\Delta t} + \frac{(x_{j+1}^m)^2 - (x_{j-1}^m)^2}{4\Delta\chi} - \frac{\nu}{(\Delta\chi)^2}(x_{j+1}^{m+1} - 2x_j^{m+1} + x_{j-1}^{m+1}) = 0. \tag{2.5.2}$$

To demonstrate the benefits of multiple shooting, we choose parameters for which the single shooting method in [7] exhibits instability. To make the problem closer to intended application target, we also experiment with larger model error and sparser observations, which are known to be more difficult [7]. We compare the solution of the multiple shooting method with that obtained from directly minimizing the full-memory function (2.1.3) in our examples. Note that the full-memory problem itself is not without difficulties: it cannot be solved to high accuracy by LBFGS in any of our examples within 2,000 iterations. The norm of gradient of (2.1.3) decreases slowly approaching the end and never gets below 10^{-6}. In this section, we refer to the approach of minimizing the full-memory function as 4D-Var for brevity, although our example is (1+1)D.

2.5.1 Results for Burgers' equation

We choose $\Delta\chi = 1/500$, $\Delta t = \Delta\chi/500$, background state $x_B = \sin(\pi\chi)$, and background covariance $Q_B = 0.01I$. We generate the initial state x_0 by sampling from the background distribution, namely, $x_0 \sim \mathcal{N}(x_B, Q_B)$. The rest of the states are generated by model propagation plus a model error term, namely, $x_{j+1} = M_j(x_j) + \eta_j$ for $0 \leq j < N$, where $\eta_j \sim \mathcal{N}(0, Q)$ and $Q = (\Delta t)^2\text{diag}(2, 1, \ldots, 1, 2)$ is the covariance of model error. This scaling results in a standard deviation of about 10^{-3} for the model error for x. We note that, in our examples, the largest absolute value of x is around 1, so this makes the smallest relative error to be around 0.1%. In a subsequent example in §2.5.3, we take this standard deviation to be 10^{-4}. These are small values, but they were chosen to be comparable to the corresponding examples in [7] so that we can compare the performance of the method in this work to the one in [7]. In that initial work, instability was a significant issue which led to choosing such small values, and indeed, that algorithm blows up even for these examples. In §2.5.2 we will present simulations for much larger model errors with standard deviations of 3.2×10^{-2}, that is, at least a few percents of the solution. The observations are generated by applying the observation operator $H(x_j) = \sin(x_j)$ to the underlying states $U = \{x_0, \ldots, x_N\}$ plus a mean zero normal observation error term to mimic the action of a noisy nonlinear operator. The operator $H(\cdot) = \sin(\cdot)$ reflects linear response around zero and saturation away from zero (assuming a range for x_j of no more than π, which is true of the solution to the target problem under our assumptions), which are characteristics of many sensors. The shape of $H(\cdot)$ is not connected to the one of the initial condition, which is chosen to also be of the $sin(\cdot)$ type in order to be simple and consistent with the boundary conditions. At the end of this subsection, we give one example of a different choice of nonlinear observation operator but with similar linearity/saturation features, and show that the results obtained are similar. We note that for analytic simplicity our theoretical results consider only the linear observation operator case; but we expect the nonlinear one to be even harder, so we could use the results to validate the outcome of multiple shooting. The covariance of observation error is chosen as $R = 0.01I$. The observations are made with a gap of 10 steps in time and space.

Our aim is to minimize the augmented Lagrangian function (2.3.1). For achieving our limited-memory purpose, we use LBFGS [94] with $p = 6$ stored vectors. To obtain an initial point for minimization, we first perturb the underlying state U by the error of the background distribution. This mimics the situation where the estimation does not start cold; in other words, initial estimates of the states do exist from previous runs of the algorithm. On each shooting interval, we run the 4D-Var minimization of (2.2.10) with LBFGS for 200 iterations to get a "warm start" state $\{w_0, \dots, w_N\}$. Note that 4DVar is run only in the beginning on each interval separately on which p trajectories are stored. The largest amount of memory required is then $\max\{2d+1+\frac{N}{d+1}, (p+1)\frac{N}{d+1}\}$ state vectors. We also note that applying LBFGS to the 4DVar problem on the entire horizon requires $(p+1)N$ state vectors, which is most times d times larger. We add that in this and in the other numerical sections, it proved difficult to find another starting strategy that will reliably produce a point from which the multiple shooting algorithm will converge. On the other hand, this strategy does work and does not alter the storage reduction benefits of our approach.

The checkpoints of the warm start state $\{w_0, w_{P_1-1}, w_{P_1}, \dots, w_{P_d-1}, w_{P_d}\}$ are then used as the initial point for minimizing (2.3.1). The Lagrangian multiplier and penalty parameters are initially chosen as $\lambda_i^{(0)} = \mathbf{0}$, $\psi_i^{(0)} = \mathbf{0}$, $\mu^{(0)} = 10$ and are subject to the usual Lagrange multiplier and penalty parameter updates [94, Framework 17.3].

N	800	1000	1200	1400	1600	2400
d	12	14	15	17	19	36
storage	434	469	525	546	560	455
$\frac{\text{storage}}{(p+1)N}$	7.8%	6.7%	6.3%	5.6%	5.0%	2.7%

Table 2.1: Number of checkpoint pairs d and maximal storage for $\Delta t = \Delta\chi/500$.

In Table 2.1 we tabulate the number of checkpoint pairs d, number of stored vectors, and percentage of storage over full-memory storage for each of the examples in this section. For $N = 800$, $d = 12$ is the smallest number of checkpoint pairs to make the computation stable. For each $800 \leq N \leq 1600$, the corresponding d is chosen so that $d/\sqrt{N} = 12/\sqrt{800}$. For $N = 2400$, d is chosen to satisfy $d/N = 12/800$. We choose $d \propto N$ for $N = 800$ and 2400 to demonstrate that the method is stable for increasing N and hence to verify Theorem 2.3.11. For $1000 \leq N \leq 1600$, we choose another relation $d \propto \sqrt{N}$ to demonstrate empirically the consequences of a more aggressive checkpointing schedule.

Figure 2.1 compares the function value reduction of (2.3.1) at each iteration of LBFGS for increasing time horizon. For $800 \leq N \leq 1600$, the rate of the initial descent (before iteration 50) becomes smaller as N increases, which indicates slower convergence for increasing N. This means that a more aggressive checkpoint schedule (e.g., $d \propto \sqrt{N}$) can lead to slower convergence. In contrast, the rate of descent for $N = 2400$ is closer to that of $N = 800$ and much larger than those of $1000 \leq N \leq 1600$. It indicates that the method not only is stable but converges with similar speed for increasing N if d is allowed to increase linearly in N. Figure 2.2 shows the norm of gradient at each iteration. Figure 2.3 shows the Frobenius norm of constraints c_i, g_i, $1 \leq i \leq d$ at each iteration. Figure 2.4 plots the Euclidean distance scaled by $\Delta\chi$ of each iteration to the checkpoints of the full-memory 4D-Var solution. Note that the distance is not scaled by the number of states and is expected to increase with d.

In this experiment, we see significant reduction (by 8–9 orders of magnitude) for both the function value and the norm of gradient, even if the gradient did not decrease to a point that triggered the Lagrange multiplier update. Figure 2.5 plots the solution surface of multiple shooting and 4D-Var when $N = 2400$. Both of them approach a perturbed version of the noise-free solution. The inviscid form ($\nu = 0$) of Burgers' equation exhibits development of shocks. In our example, we employ a larger viscosity coefficient $\nu = 0.01$ which smooths out the waves and acts against the steepening effect of nonlinearity. Hence the solution surface of Burgers' equation in our setup (top left of Figure 2.5) does not exhibit such nonlinear effects. Figure 2.6 compares multiple shooting and 4D-Var solutions at fixed

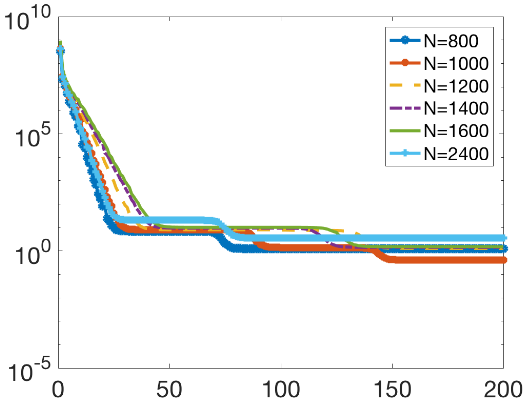

Figure 2.1: Function value of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/500$ and $N = 800, 1000, 1200, 1400, 1600, 2400$.

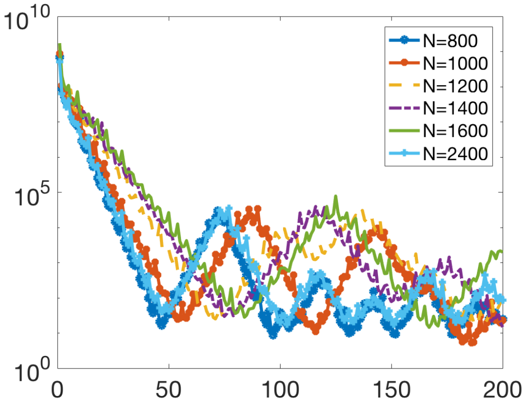

Figure 2.2: Gradient norm of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/500$ and $N = 800, 1000, 1200, 1400, 1600, 2400$.

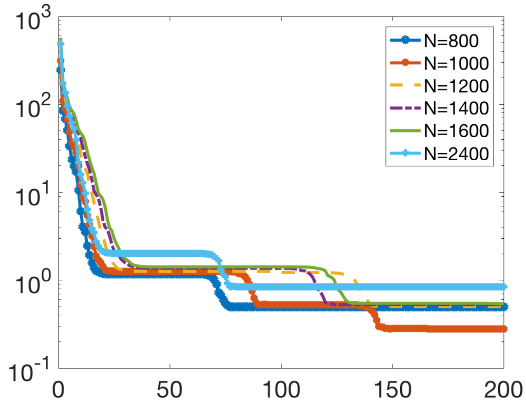

Figure 2.3: Norm of constraint at each iteration of LBFGS in minimizing (2.3.1) for $\Delta t = \Delta\chi/500$ and $N = 800, 1000, 1200, 1400, 1600, 2400$.

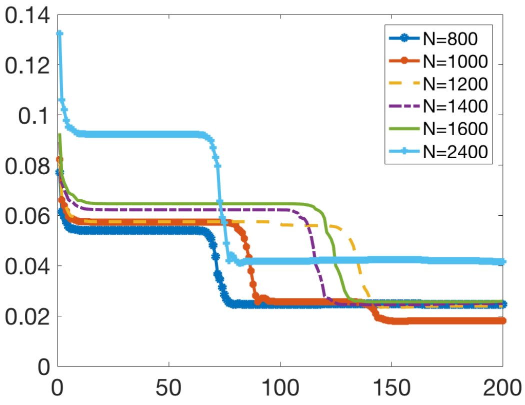

Figure 2.4: Distance to 4D-Var solution at each iteration of LBFGS in minimizing (2.3.1) for $\Delta t = \Delta\chi/500$ and $N = 800, 1000, 1200, 1400, 1600, 2400$.

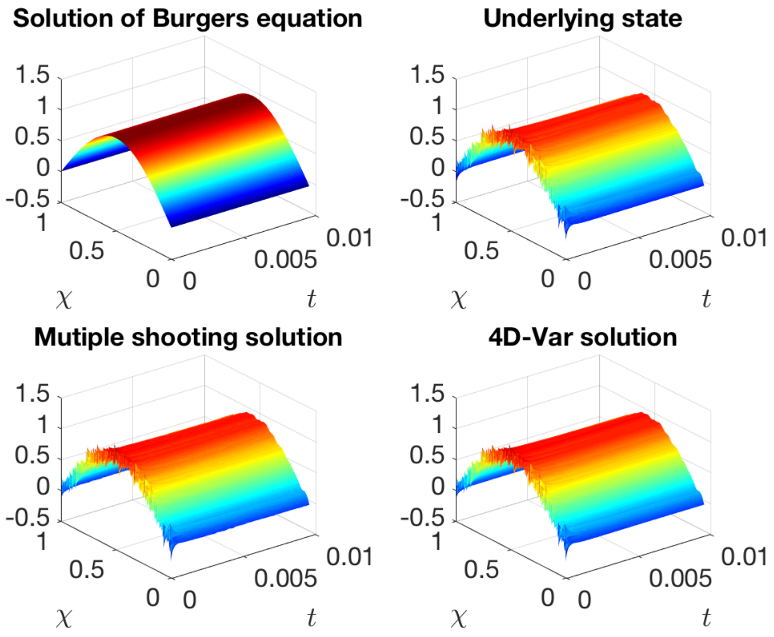

Figure 2.5: Exact solution of Burgers equation (top left), underlying state (top right) and states estimated with multiple shooting and 4D-Var for $\Delta t = \Delta\chi/500$ and $N = 2400$.

time and space nodes. Note that the two solutions are both close to the underlying state so that the trajectories overlap for most of the part. Although the problem is not solved to high accuracy as suggested by the norm of the gradient and norm of the constraint, we conclude that it does approach the 4D-Var solution.

From the simulations we see that keeping N/d fixed (at its lowest value) results in faster convergence compared with the alternatives. We thus conclude that the statement of Theorem 2.3.11 is satisfied, although its conditions are stronger than the case tested here (we did not enforce small nonlinearity and linearity of the observation operator). However, for the case of smaller N (e.g., 800 to 1600), even increasing d slower than linear in N (e.g., $\sqrt{N}$) would give stable results and thus even more memory savings at a cost of somewhat slower convergence.

To illustrate the effect of the observation mapping, we apply a different nonlinear operator $\widetilde{H}_j(x_j) = 1/\left(1 + e^{-x_j}\right) - 1/2$ which also exhibits linear response around zero and saturation away from zero. For $N = 800$, $d = 12$ and all the other parameters staying the same,

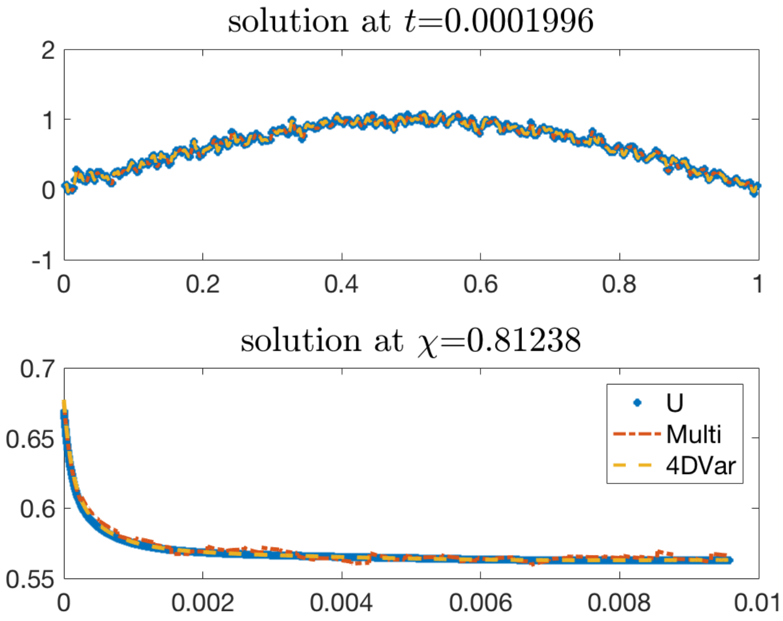

Figure 2.6: Underlying state, multiple shooting solution and 4D-Var solution at fixed time and space node for $\Delta t = \Delta\chi/500$ and $N = 2400$.

Figures 2.7 and 2.8 show the function value reduction and the norm of gradient of (2.3.1) at each iteration of LBFGS. The performance is similar to applying $H_j(x_j) = \sin(x_j)$ which is shown in Figures 2.1 and 2.2. Hence we conclude that the effect of the choice of observation mapping appears to be small.

2.5.2 Larger model error

In this section, we experiment with increased model error. We choose $\Delta\chi = 1/500$, $\Delta t = \Delta\chi/1000$, and a background covariance matrix $Q_B = 0.01I$. The covariance for the model error and observation error are chosen to be $10^{-3}I$. Observations are reduced to every 10 steps in time and every 100 steps in space. To initialize the minimization of (2.3.1), we run the 4D-Var minimization on one interval, and for the next interval we run 4D-Var constrained at the checkpoint by the solution from the previous interval.

Figure 2.9 shows the augmented Lagrangian function value decrease for $N = 500$ and number of checkpoint pairs $d = 38$. Figure 2.10 shows the norm of the gradient. Figure 2.11

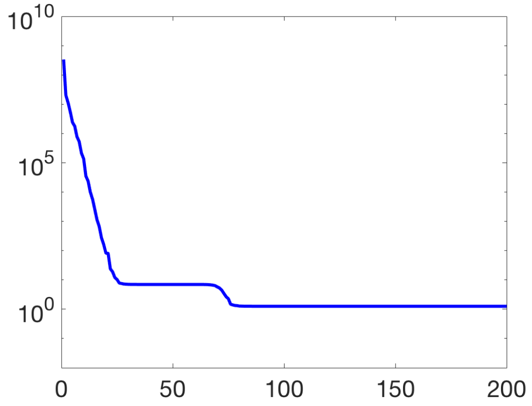

Figure 2.7: Function value of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/500$ and $N = 800$ under observation mapping $\widetilde{H}(\cdot)$.

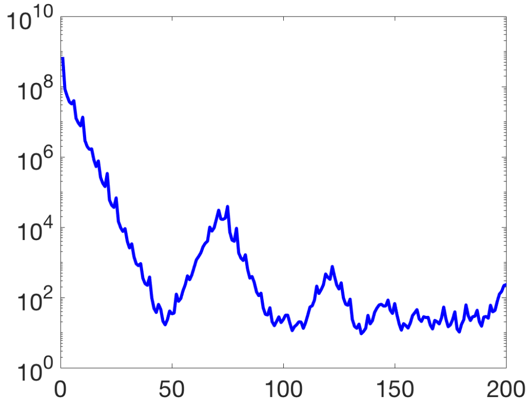

Figure 2.8: Gradient norm of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/500$ and $N = 800$ under observation mapping $\widetilde{H}(\cdot)$.

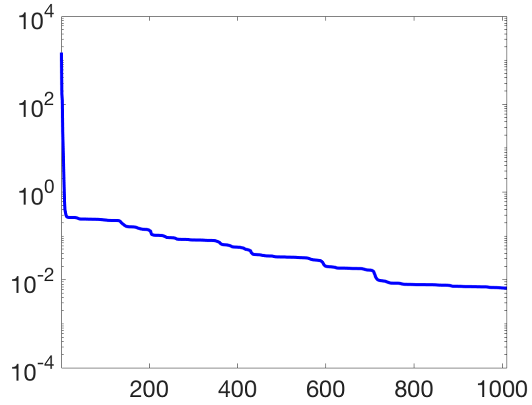

Figure 2.9: Function value of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/1000$, $Q = 10^{-3}I$, $N = 500$ and $d = 38$.

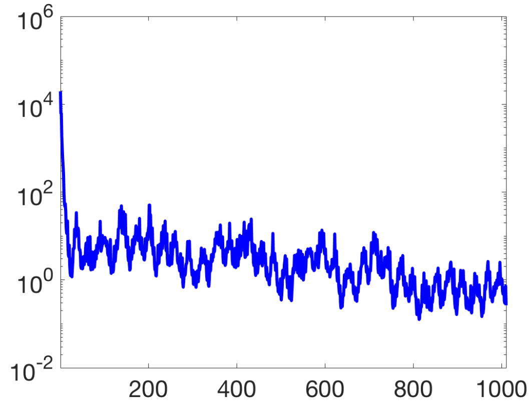

Figure 2.10: Gradient norm of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/1000$, $Q = 10^{-3}I$, $N = 500$ and $d = 38$.

compares the full-memory 4D-Var solution with that of multiple shooting. Increased model error results in the rough surface of the underlying states plot in Figure 2.11. Figure 2.12 compares the 4DVar and multiple shooting solutions at fixed time and space nodes. Note that the two solutions are close to each other so that their trajectories overlap.

Both the function value and the norm of the gradient converge slower after some significant initial progress. Since the norm of the gradient stalls and fails to progress below 0.1, we do not observe either Lagrangian multiplier or penalty parameter update during the experiments. However, both the function value and the norm of the gradient achieve 4 to 6

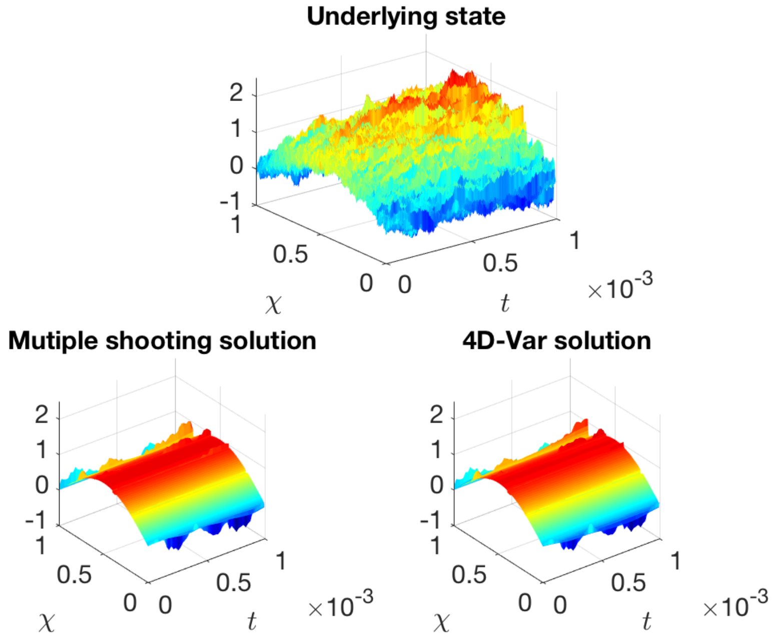

Figure 2.11: Underlying states, solution surface of multiples shooting and 4D-Var for $\Delta t = \Delta\chi/1000$, $Q = 10^{-3}I$, $N = 500$ and $d = 38$.

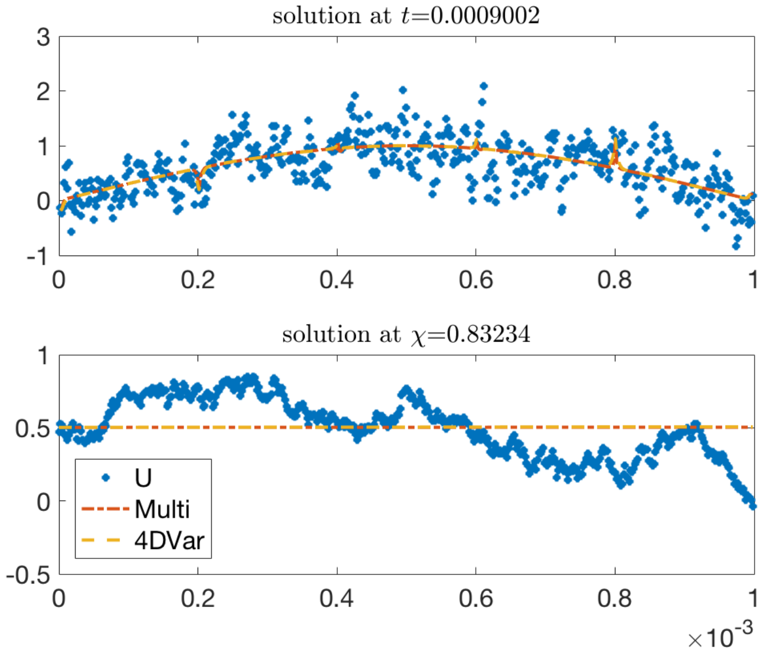

Figure 2.12: Underlying state, multiple shooting solution and 4D-Var solution at fixed time and space nodes for $\Delta t = \Delta\chi/1000$, $Q = 10^{-3}I$, $N = 500$ and $d = 38$.

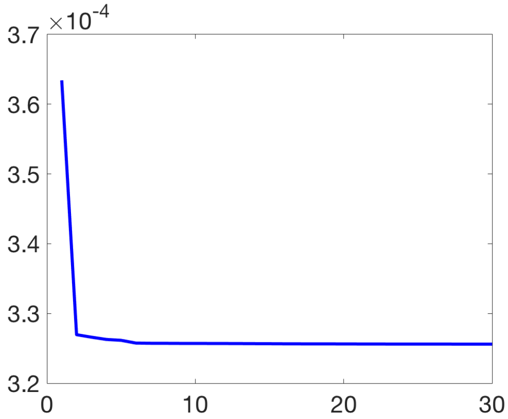

Figure 2.13: Function value of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/34$, $N = 300$ and $d = 30$. Observation gaps are 30 steps temporally and 200 steps spatially.

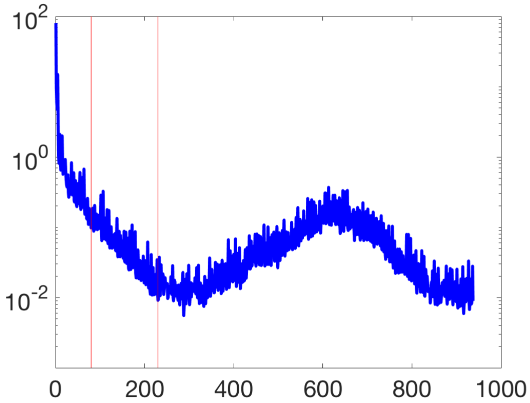

Figure 2.14: Gradient norm of (2.3.1) at each iteration of LBFGS for $\Delta t = \Delta\chi/34$, $N = 300$ and $d = 30$. Observation gaps are 30 steps temporally and 200 steps spatially. Reference line indicates Lagrangian multiplier updates.

orders of magnitude decrease, and the multiple shooting solution approaches reasonably well the full-memory 4D-Var solution. Clearly the problem has too much noise for the estimates to be close to the underlying state. However, the approach does show that multiple shooting has a performance comparable to that of 4DVar, with much less memory, and that is the goal of this project.

With the same parameters as those in [7, §5.2.5] but with a much longer horizon, $N = 500$ as opposed to $N = 110$, our method is able to produce iterations of moderate size, make nontrivial progress through minimization, and result in solutions comparable to that of full-memory method for a longer time horizon. Counting the storage during warm start, gradient evaluation, and stored vectors of LBFGS, the maximal number of states stored at any time of the algorithm is 91 and is about 18.2% of the total number of states N. The storage used by multiple shooting is 2.6% of the memory used by full-memory minimization using LBFGS with 6 vectors.

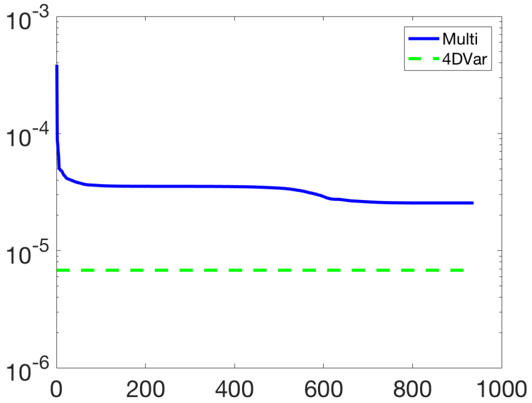

Figure 2.15: Norm of constraint at each iteration of LBFGS in minimizing (2.3.1) for $\Delta t = \Delta\chi/34$, $N = 300$ and $d = 30$. Observation gaps are 30 steps temporally and 200 steps spatially.

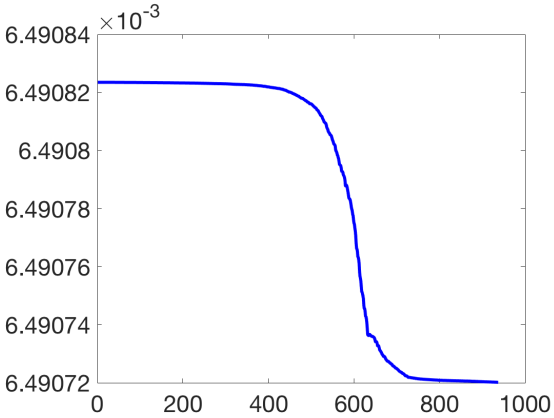

Figure 2.16: Distance to 4D-Var solution at each iteration of LBFGS in minimizing (2.3.1) for $\Delta t = \Delta\chi/34$, $N = 300$ and $d = 30$. Observation gaps are 30 steps temporally and 200 steps spatially.

2.5.3 Sparser observations

In this section, we consider the case where observations are sparser in both time and space. As in [7, §5.2.5], we choose $\Delta\chi = 1/700$, $\Delta t = \Delta\chi/34$, and background covariance as $Q_B = 10^{-3}I$. In particular, the much larger time step tests the ability of the approach to cope with increased instability. The covariance matrix for the model error and observation error are $Q = 10^{-8}I$ and $0.01I$, respectively. Observations are made every 30 steps in time and every 200 steps in space. The initial point for multiple shooting is the same warm-start point described in the precedent section. The parameters are the same as those in [7, §5.2.5] but with a longer horizon. We take $N = 300$ as opposed to $N = 32$ in [7], and we take number of checkpoint pairs $d = 30$. We note that this setup is significantly far from satisfying the observability condition. Indeed, the rank of the observability matrix in Definition 2.3.1 cannot be larger than 8, whereas Theorem 2.3.11 required a full rank, that is, 701.

For this experiment, Figure 2.13 shows the decrease of function value (2.3.1). Only the first 30 iterations are plotted since the function value stalls afterward. Lagrangian multipliers are updated at iteration 80 and 230, as shown by the vertical reference line in Figure 2.14. Figure 2.15 shows the norm of constraints c_i and g_i at each iteration. The horizontal reference line plotted is the norm of constraint for the 4D-Var solution. Figure 2.16 shows the Euclidean distance of each iteration to the 4D-Var solution scaled by $\Delta\chi$. The decrease in the norm of the gradient is significant (3–4 orders of magnitude), and the norm of the constraint is reduced by about 1 order of magnitude. The distance to the 4D-Var solution shows little progress compared with the initial guess obtained by running 4D-Var on each shooting interval, but Figures 2.15 and 2.16 suggest the reason is primarily that our warm-starting using 4D-Var on each shooting interval produces an initial point for multiple shooting close to the 4D-Var solution itself. On the other hand, even if in the distance to the 4D-Var solution there is not much progress beyond the warm start, the gradient is significantly reduced, and we can evaluate the convergence properties of the method, running LBGFS to detect whether we see an improvement, while needing less memory than 4D-Var with LBFGS (only 3.4% of the latter's). Therefore the multiple shooting method provides an improvement over 4D-Var with LBFGS in terms of memory and over single shooting in terms of stability even in this case, which is significantly outside the applicability of Theorem 2.3.11.

2.6 Conclusions

Determining the best state estimation for dynamical systems with model error raises new challenges in developing algorithms that reduce storage while maintaining stability. The reason is that, as opposed to the strongly constrained setups where only the initial state is free, all the states of a trajectory contribute to the number of degrees of freedom.

We present an approach where the number of degrees of freedom is reduced by the optimality conditions, as we previously introduced in [7], but now coupled with a multiple shooting approach in an augmented Lagrangian framework to improve stability. The multiple shooting approach can use a reverse recursion scheme on each shooting interval to ensure that the memory requirements for computing one gradient of the augmented Lagrangian never exceed $2d + 1 + \frac{N}{d+1}$ state vectors, where $d + 1$ is the number of shooting intervals and N is the length of the horizon. The full-memory data assimilation method, on the other hand, needs to store $N + 1$ state vectors when evaluating its gradient. We prove in Theorem 2.3.11 that under an observability assumption and when the nonlinearity is small relative to the parameter characterizing the observability, the condition number of the augmented Lagrangian matrix is bounded above, irrespective of the number of shooting intervals. This ensures that the multiple shooting approach is stable: the method does not exhibit exponentially increasing error for an increasing size of the assimilation interval. This is a feature not shared by the single shooting approach derived from [7]. Moreover, as pointed out in the Discussion following Theorem 2.3.11 and Appendix A.1, multiple shooting without observations still results in exponential increase of the condition number and thus of the solution. Therefore both multiple shooting and sufficiently informative observations appear to be necessary for stability to occur.

Our numerical simulations on cases described in [7] validate these points. First, for all of them the single shooting method showed an exponential increase of the solution and ran into overflow. For both small model error and larger model error setups, the multiple shooting approach converges to a solution close to that of the full-memory method while using only a fraction of the memory needed by the latter, never more than 8%. To achieve convergence, we needed to use the full-memory approach but only on the smaller, shooting intervals to create a good initial point for our multiple shooting approach. In the case of sparse observations, this initialization strategy was responsible for much of the improvement of the method in terms of distance to the full-memory 4D-Var solution, while using only 3.4% of the memory of the latter. But with that initialization strategy, which does not alter our maximum memory count, we reliably obtained reductions in the augmented Lagrangian gradients and solutions close to the ones of the full-memory approach. We are not aware of another optimization-based approach to reduce the memory requirements of weakly constrained data assimilation approaches. From the numerical experiments and the theory, we conclude that, particularly in the data-rich case, the multiple shooting method appears promising at reducing memory and producing a point of a quality comparable to that of the full-memory case without the instability of the previous single shooting approach.

We plan to explore new initialization strategies that empirically appear to be important for the robustness of the overall method. The method also has good potential for paralellism, although in that case the memory saving is less of a benefit. An interesting question would be to tie the stability of multiple shooting to a condition requiring enough information in the observations but weaker than observability on one shooting interval. We have observed the good behavior of the multiple shooting aproach in several such instances, but it is unclear how such a condition might be expressive enough and practical.

CHAPTER 3
EXPONENTIALLY ACCURATE TEMPORAL DECOMPOSITION FOR LONG-HORIZON LINEAR-QUADRATIC DYNAMIC OPTIMIZATION

3.1 Introduction

Long-horizon dynamic optimization problems appear in several application areas [11, 13, 24, 28, 40, 54, 63] and pose significant computational challenges because of the increase in the number of variables in proportion to the number of time periods considered. One very long horizon instance is optimal planning in the electrical power industry for transmission or generation expansion [11], which we now describe in some detail.

Such a planning analysis involves a production cost model (PCM). A PCM simulates the operation of generation and transmission systems by finding, during each time interval, the least-cost solution to generating sufficient energy to meet demand. As an abstraction, it is an optimal control problem, which can have nonlinear dynamics, control and state constraints. Most studies require running a PCM on an hourly scale for 1–20 years under different scenarios in order to address the operation and reliability aspects of the proposed transmission or expansion plan [101]. Doing so can result in a very large number of periods. For example, if a PCM is run for 12 years with an hourly scale, the number of time periods would exceed 100,000. Added to this are the tens of thousands of degrees of freedom at one time point, which are characteristic for planning at the interconnect level, making the problem a daunting one to solve. As a result, many planning studies, which involve investments of billions of dollars, are done with multiple approximations to make them fit the computing resources [91].

Researchers have therefore sought to identify approaches for long-horizon dynamic optimization that result in efficient temporal parallelism to address this complexity by bringing to bear more computing power. Approaches have included temporal decomposition strategies using Lagrangian decomposition [13, 54, 63] and a two-level optimization formulation with the lower level derived from a decomposition approach [28]. These ideas create the opportunity for faster computation using parallelism. For instance, a heuristic decomposition algorithm is presented in [40] for scheduling a batch chemical plant. The problem is decomposed into more tractable subproblems that are solved to optimality. Empirical evidence suggests largely reduced computational efforts and reasonable accuracy. Strengths and weaknesses of a number of temporal decomposition methods are investigated in [13]. A multiperiod nonlinear programming model is developed in [63] for production planning and distribution. Temporal decomposition is used for the solution and is shown to generate faster computation and good accuracy of the optimal solutions.

A recent approach for PCMs is to partition the simulation horizon and turn the annual problem into multiple overlapping weekly/monthly problems [11] that compute the contribution of an *inner time interval only* to the overall objective and then add up all these contributions. While such an approach cannot be an exact decomposition, it can be computed in parallel without information exchange between the problems on each decomposition interval, and therefore the computation can be sped up. Moreover, researchers have shown empirically in [11] that the error in the approach drops rapidly with the increase of the buffer region (the overlapping area) surrounding the inner time interval.

Our aim here is to provide theoretical support for approximate temporal decomposition of dynamic optimization problems with long horizons using overlapping intervals such as the work in [11]. A particular focus is on characterizing the error made by using such approximations.

For this initial foray, we will use a considerably simpler model than the PCMs in [11] or other planning models [63]. That is, our formulation is the following optimization problem:

$$\min_{x_k,u_k} \quad \sum_{k=n_1}^{n_2-1} \Big(u_k^T R_k u_k + (x_k - d_k)^T Q_k (x_k - d_k)\Big) \tag{3.1.1a}$$

$$+(x_{n_2} - d_{n_2})^T Q_{n_2}(x_{n_2} - d_{n_2}) \tag{3.1.1b}$$

$$\text{s.t.} \qquad x_{k+1} = A_k x_k + B_k u_k, \;\; x_{n_1} = x^0_{n_1}, \tag{3.1.1c}$$

$$l_k \leq u_k \leq b_k, \qquad n_1 \leq k \leq n_2 - 1, \tag{3.1.1d}$$

for some initial value $x^0_{n_1}$ given. We call such a problem linear-quadratic dynamic optimization problem. Such problems are known under various other names such as linear-quadratic (model predictive) control [46], or dynamic programming [20]. We choose the name *dynamic optimization* for problem (3.1.1) [26, 42] as we are interested in finding the solution of the optimization problem rather than computing the control rule or policy functions themselves. We will, however, use the terms control and dynamic programming as well when referring to the existing results and their interpretations. In (3.1.1) $[n_1, n_2]$ is the entire time horizon under consideration, and x_{n_1} is known. We refer to x_k, u_k, and d_k respectively as the supply or generation, control, and reference trajectory (also known as demand in PCM contexts). Problem (3.1.1) has a few simplifications and changes compared with [11, 63]: our objective is quadratic and not linear, and we do not allow for integer variables. We note that these different features are used in the target areas. Quadratic objectives are sometimes used instead of linear for the one-period cost function [71]. Economic dispatch, that is, a version of PCM where the scheduling decisions are all known in advance and thus no integer variables are present, is used in planning studies [19]. A more important approximation is that we do not allow for hard path constraints. For example, supply and demand mismatch is penalized in the objective but not enforced to be zero. Approaches exist to accommodate supply equaling demand, at least in some circumstances, as will be done in our numerical example in Section 3.4; we do not claim, however, that this can be done in general. Given the complexity of the analysis with even this simplified formulation, extensions that obtain results like ours under circumstances closer to [11, 63] will be investigated in future research.

Our approach, however, retains two important features from planning models that allow us to investigate approximate temporal decomposition: intertemporal constraints (3.1.1c)

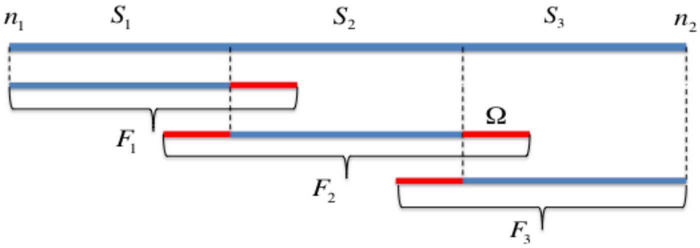

Figure 3.1: Illustration of the temporal decomposition scheme with three decomposition intervals. The entire horizon $[n_1, n_2]$ is decomposed into subintervals $S_{1:3}$, which are embedded in regions $F_{1:3}$ correspondingly. The red areas are buffer regions, each of length Ω.

and box constraints on the control (3.1.1d). In particular, it allows us to substantiate a key insight that makes the temporal decomposition approach work efficiently. That is, when the system (3.1.1c) is controllable, the closed loop control law attached to the optimal active set results in an asymptotically stable policy [20]. In turn, the effect of perturbations of the parameters d_k and initial state x_{n_1} on the solution decreases exponentially with the distance in time between the perturbation moment and the index of the state. Hence, the system can forget its past and ignore its future exponentially fast with the distance from both.

This observation suggests the following temporal decomposition approach. Given a fixed time period $S_i \subset [n_1, n_2]$, we are interested in finding a shorter embedding interval F_i with $S_i \subset F_i \subset [n_1, n_2]$, so that the solution on S_i obtained by solving problem (3.1.1) on F_i is close to the one obtained by solving problem (3.1.1) on $[n_1, n_2]$. As a result, the entire horizon can be decomposed approximately, but with little error, into pieces like S_i, and the optimal solutions on each piece can be computed in parallel by solving problem (3.1.1) on F_i. Figure 3.1 illustrates this decomposition scheme. The temporal decomposition approach then consists of approximating the optimal value of problem (3.1.1) on $[n_1, n_2]$ by the sum of the optimal values on S_i obtained from solving (3.1.1) on F_i, over all i. A formal definition of this decomposition approach is presented in Section 3.3.

Our work is to estimate the error of this decomposition approach. In our proofs we will use several results from optimal control theory, which were done in the case of $d_k = \mathbf{0}$ and in the absence of the bound constraints (3.1.1d), with respect to the notations in (3.1.1). In that case, the solution of (3.1.1) is provided by the linear-quadratic regulator (LQR), a feedback control law to achieve minimal cost. A derivation of the finite-horizon, discrete-time LQR based on the dynamic programming principle can be found in [20], which also shows that the resulting optimal trajectory tracks zero exponentially fast for time-independent linear systems. In this work, one particular control feature we will characterize and use is the rate of stabilization of the optimal trajectory for discrete time, time-varying linear-quadratic dynamic optimization problems. To this end, Zhang et al. [124] derive some important properties for the finite-horizon and infinite-horizon value function of the switched system discrete-time linear-quadratic dynamic optimization. The authors show that under some mild assumptions, the optimal trajectory stabilizes exponentially, and they give a workable estimate of that rate that we will use here. Some algorithms based on those theoretical results are also shown in [123] and [125]. A similar result to our Theorem 3.3.11 that upper bounds the approximation error of the optimal cost for the temporal decomposition approach is given in [70]. The authors show that, for a class of constrained discrete-time systems, the infinite-horizon cost associated with the moving-horizon feedback law converges to the optimal infinite-horizon cost as the moving horizon is extended. Our work inherits similar temporal decomposition features as that in [70]. However, in this chapter we additionally prove that, with a long but finite horizon, the *solutions* on the decomposition intervals converge as the embedding regions increase. Moreover, we characterize the convergence rate for the solutions and optimal cost as exponentially fast, which is crucial for the approach to be practical.

The rest of the chapter is organized as follows. Section 3.2 proves results about the box constrained control linear-quadratic problem. Section 3.3 describes the temporal decomposition approach and proves that, based on the results derived in Section 3.2, the error of the temporal decomposition method decays exponentially in the size of the embedding interval. In Section 3.4, we illustrate the theoretical findings by applying the temporal decomposition approach to a production cost model using real demand data. Proofs of results that are not central to the development of the main ideas are presented in Appendix B.1.

3.2 Box constrained control linear-quadratic problem

In this Section, we derive results for a sub-problem of the following box constrained control linear-quadratic problem:

$$\begin{aligned}
\min \quad & \Gamma_{n_1:n_2}(u_{n_1:n_2-1}, x_{n_1+1:n_2}) && \text{(3.2.1a)}\\
& \triangleq \sum_{k=n_1}^{n_2-1} \Big(u_k^T R_k u_k + (x_k - d_k)^T Q_k (x_k - d_k) \Big) && \text{(3.2.1b)}\\
& + (x_{n_2} - d_{n_2})^T Q_{n_2} (x_{n_2} - d_{n_2}) && \text{(3.2.1c)}\\
\text{s.t.} \quad & x_{k+1} = A_k x_k + B_k u_k, \qquad x_{n_1} = x_{n_1}^0, && \text{(3.2.1d)}\\
& l_k \le u_k \le b_k, \qquad n_1 \le k \le n_2 - 1, && \text{(3.2.1e)}
\end{aligned}$$

where the initial state $x_{n_1}^0$ is given. Throughout the article, we have that $A_k \in \mathbb{R}^{n \times n}$, $B_k \in \mathbb{R}^{n \times m}$, and R_k, Q_k are positive definite matrices. We make the following uniform boundedness assumption about the system.

Assumption 3.2.1. *For any n_1, n_2, $n_1 \le q \le n_2$, we have the following:*

(a) $\|A_q\|_2 \le C_A$, $\|B_q\|_2 \le C_B$, $\|Q_q\|_2 \le C_Q$, $\|R_q\|_2 \le C_R$ for some C_A, C_B, C_Q, $C_R > 0$.

(b) $\lambda_{min}(Q_q) \ge \lambda_Q > 0$, $\lambda_{min}(R_q) \ge \lambda_R > 0$.

(c) $\|b_q\|$, $\|l_q\| \le U$ for some $U > 0$.

The sub-problem of (3.2.1) we consider is an equality constrained problem obtained by considering some active subset of the box control constraints (3.2.1e).

3.2.1 An equality constrained sub-problem

To define the equality constrained sub-problem, we let $I_k \subset \{1, \dots, m\}$ be some index set for the elements of u_k attaining either the upper or lower bound of (3.2.1e), and denote $N_k = I_k^{\mathsf{c}}$. Let e_i be the ith standard basis vector. We associate I_k with a selection matrix C_k and a vector $\bar{b}_k$ defined as follows:

$$C_k(i,:) = \begin{cases} e_{j_i}^T, & u_k(j_i) = l_k(j_i) \\ -e_{j_i}^T, & u_k(j_i) = b_k(j_i) \end{cases}, \ \bar{b}_k(i) = \begin{cases} l_k(j_i), & u_k(j_i) = l_k(j_i) \\ -b_k(j_i), & u_k(j_i) = b_k(j_i) \end{cases}, \tag{3.2.2}$$

where $i = 1, \dots, |I_k|$, and j_i is the ith element in I_k. With these definitions, the equality constraints corresponding to I_k can be expressed as $C_k u_k = \bar{b}_k$, and the equality constrained sub-problem is defined as

$$\begin{aligned} \min_{x_k,u_k} \quad & \sum_{k=n_1}^{n_2-1} \Big(u_k^T R_k u_k + (x_k - d_k)^T Q_k (x_k - d_k) \Big) + (x_{n_2} - d_{n_2})^T Q_{n_2} (x_{n_2} - d_{n_2}) \\ \text{s.t.} \quad & x_{k+1} = A_k x_k + B_k u_k, \qquad x_{n_1} = x_{n_1}^0, \\ & C_k u_k = \bar{b}_k, \qquad n_1 \le k \le n_2 - 1. \end{aligned} \tag{3.2.3}$$

Note that when I_k is the active set of problem (3.2.1) at optimality, problems (3.2.3) and (3.2.1) have the same solutions.

Problem (3.2.3) is the primary topic we consider in this Section, and will appear later in Section 3.3 in a sensitivity analysis needed to prove temporal decomposition. In the rest of this Subsection, we focus on deriving properties for the solution of problem (3.2.3) for some index set I_k. In particular, we will show the exponential decay property of the dependence of the solutions of problem (3.2.3) on the initial state and terminal reference under certain conditions. This is crucial to establish the main temporal decomposition results in Section 3.3. To start with, we note that a reduced problem can be obtained by eliminating the equality constraints of (3.2.3). We partition u_k, B_k and R_k into blocks corresponding to I_k

and N_k. Denote

$$\widetilde{u}_k = [u_k(i)]_{i \in I_k}, \;\; \widetilde{B}_k = [B_k(:, i)]_{i \in I_k}$$

to be the elements (or columns) of u_k (or B_k) corresponding to the equality index set I_k. Similarly, for N_k, denote correspondingly

$$\hat{u}_k = [u_k(i)]_{i \in N_k}, \;\; \hat{B}_k = [B_k(:, i)]_{i \in N_k}.$$

Also, R_k can be partitioned into blocks corresponding to the index sets as follows:

$$\hat{R}_k = [R_k(i,j)]_{i \in N_k, j \in N_k}, \;\; \widetilde{R}_k = [R_k(i,j)]_{i \in I_k, j \in I_k}, \;\; \bar{R}_k = [R_k(i,j)]_{i \in I_k, j \in N_k}.$$

Then we have that

$$\begin{aligned} B_k u_k &= \hat{B}_k \hat{u}_k + \widetilde{B}_k \widetilde{u}_k, \\ u_k^T R_k u_k &= \hat{u}_k^T \hat{R}_k \hat{u}_k + 2\widetilde{u}_k^T \bar{R}_k \hat{u}_k + \widetilde{u}_k^T \widetilde{R}_k \widetilde{u}_k, \end{aligned}$$

and that the equality constraint $C_k u_k = \bar{b}_k$ is equivalent to $\widetilde{u}_k = \widetilde{b}_k$, where the ith element of $\widetilde{b}_k$ is $l_k(j_i)$ (or $b_k(j_i)$) if $u_k(j_i)$ attains the lower bound $l_k(j_i)$ (or the upper bound $b_k(j_i)$) for $i \in \{1, \ldots, |I_k|\}$, $j_i \in I_k$. Define a change of variable as follows:

$$\begin{aligned} v_k &= \hat{u}_k + \hat{R}_k^{-1} \bar{R}_k \widetilde{b}_k, \\ f_k &= \widetilde{B}_k \widetilde{b}_k - \hat{B}_k \hat{R}_k^{-1} \bar{R}_k \widetilde{b}_k. \end{aligned} \tag{3.2.4}$$

Note that $\hat{R}_k$ is invertible since R_k is positive definite. Then (u_k^*, x_k^*) is the solution of problem (3.2.3) if and only if (v_k^*, x_k^*), defined by (3.2.4), is the solution of the following problem:

$$
\begin{aligned}
\min_{x_k, v_k} \quad & \sum_{k=n_1}^{n_2-1} \left(v_k^T \hat{R}_k v_k + (x_k - d_k)^T Q_k (x_k - d_k) \right) && \text{(3.2.5a)} \\
& \quad +(x_{n_2} - d_{n_2})^T Q_{n_2} (x_{n_2} - d_{n_2}) && \text{(3.2.5b)} \\
\text{s.t.} \quad & x_{k+1} = A_k x_k + \hat{B}_k v_k + f_k, \;\; n_1 \le k \le n_2 - 1, \;\; x_{n_1} = x_{n_1}^0. && \text{(3.2.5c)}
\end{aligned}
$$

One can easily verify that (3.2.4) defines a one-to-one correspondence between the feasible sets of problems (3.2.3) and (3.2.5) and that the objective functions differ by a constant for the corresponding elements in the feasible sets. Note that the optimal values of problems (3.2.3) and (3.2.5) differ by a constant. However, since we are only interested in the *solutions* of problem (3.2.3) with which the solutions of problem (3.2.5) have a one-to-one relationship (3.2.4), we thus solve problem (3.2.5) in order to investigate properties for the solutions of (3.2.3).

Problem (3.2.5) is a linear-quadratic optimal control problem for which we need a notion of controllability for the sequence pair $\{A_k, \hat{B}_k\}_{k=n_1:n_2}$. Note that $\hat{B}_k$ is uniquely determined by the index set I_k under consideration, and hence the choice of the index sets $\mathcal{I} \overset{\Delta}{=} \{I_k\}$ will affect the controllability of the resulting $\{A_k, \hat{B}_k\}$. We make the following definition of controllability.

Definition 3.2.2. *For some index sets $\mathcal{I} = \{I_k\}_{k=n_1:n_2}$, let $\hat{B}_k = [B_k(:, i)]_{i \in I_k^{\mathsf{c}}}$. With some $0 < t < n_2 - n_1$, $\lambda_C > 0$,*

(a) define the controllability matrix associated with time steps $[q, q+t-1]$ as

$$
C_{q,t}(\mathcal{I}) = \left[\hat{B}_{q+t-1} \quad A_{q+t-1}\hat{B}_{q+t-2} \quad \ldots, \left(\prod_{l=1}^{t-1} A_{q+l} \right) \hat{B}_q \right];
$$

(b) the index set $\mathcal{I}$ is uniformly completely controllable with parameter λ_C, denoted as UCC(λ_C), if the sequence pair $\{A_k, \hat{B}_k\}$ is uniformly completely controllable with pa-

rameter λ_C [70, Definition 3.1], i.e., for any $n_1 \leq q \leq n_2$,

$$\lambda_{min}\left(C_{q,t}(\mathcal{I})C_{q,t}^T(\mathcal{I})\right) \geq \lambda_C > 0.$$

Now we derive the optimal control law and optimal states for problem (3.2.5) using a dynamic programming approach. When $d_k \equiv 0, \forall k \in n_1 : n_2$ and $f_k \equiv 0, \forall k \in n_1 : (n_2 - 1)$, the solution to problem (3.2.5) is well known from classical dynamic programming references. For our temporal decomposition, however, the dependence on d_k is crucial, whereas $f_k \neq 0$ is needed as an artifact of the box constraints. To simplify our notations, we use a *reverse product* notation as follows.

Definition 3.2.3. *We define*

$$\prod_{i=m}^{n} A_i = \begin{cases} A_n \dots A_m, & n \geq m \\ I, & n < m. \end{cases}$$

Proposition 3.2.4. *For $n_1 \leq k \leq n_2 - 1$, the optimal control laws for problem* (3.2.5) *are*

$$\begin{aligned} v_k^*(x_k) = L_k x_k + W_k^{-1} \sum_{i=k+1}^{n_2} \hat{B}_k^T \left(M_i^{k+1}\right)^T d_i \\ + W_k^{-1} \sum_{i=k+1}^{n_2-1} \hat{B}_k^T \left(S_i^{k+1}\right)^T f_i - W_k^{-1} \hat{B}_k K_{k+1} f_k, \end{aligned} \tag{3.2.6}$$

where

$$K_{n_2} = Q_{n_2}, \tag{3.2.7a}$$

$$K_k = A_k^T(K_{k+1} - K_{k+1}\hat{B}_k W_k^{-1} \hat{B}_k^T K_{k+1})A_k + Q_k, \quad n_1 \leq k \leq n_2 - 1, \tag{3.2.7b}$$

$$W_k = \hat{R}_k + \hat{B}_k^T K_{k+1} \hat{B}_k, \qquad n_1 \leq k \leq n_2 - 1, \tag{3.2.7c}$$

$$L_k = -W_k^{-1} \hat{B}_k^T K_{k+1} A_k, \qquad n_1 \leq k \leq n_2 - 1, \tag{3.2.7d}$$

$$
\begin{aligned}
D_k &= A_k + \hat{B}_k L_k, \qquad n_1 \le k \le n_2 - 1, && (3.2.7e)\\
M_i^k &= Q_i \prod_{l=k}^{i-1} D_l, \qquad i \ge k, \quad n_1 \le k \le n_2, && (3.2.7f)\\
S_i^k &= -K_{i+1} \prod_{l=k}^{i} D_l, \qquad i \ge k, \quad n_1 \le k \le n_2 - 1. && (3.2.7g)
\end{aligned}
$$

Proof. See Appendix B.1.1. □

We note that (3.2.7b)–(3.2.7e), and the expression of v_k^* when $d_k \equiv 0$, $f_k \equiv 0$ are the results of classical LQ control.

Definition 3.2.5. *For $n_1 \le k \le n_2 - 1$, define*

$$
E_k = \hat{B}_k^T W_k^{-1} \hat{B}_k,
$$

where W_k is defined in (3.2.7c)*.*

Proposition 3.2.6. *Let $x^*_{n_1+1:n_2}$ be the optimal states of* (3.2.5)*. Then we have that*

$$
x_k^* = \left(\prod_{i=n_1}^{k-1} D_i \right) x_{n_1} + \sum_{i=n_1+1}^{n_2} C_i^k d_i + \sum_{i=n_1}^{n_2-1} F_i^k f_i, \tag{3.2.8}
$$

where

$$
\begin{aligned}
C_i^k &= \sum_{s=n_1}^{\min(i,k)-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(M_i^{s+1} \right)^T, \\
F_i^k &= \sum_{s=n_1}^{\min(i,k)-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(S_i^{s+1} \right)^T + \left(\prod_{l=i+1}^{k-1} D_l \right) \left(I - E_i K_{i+1} \right) \mathbf{1}_{(k \ge i+1)}.
\end{aligned} \tag{3.2.9}
$$

Proof. See Appendix B.1.2 □

Next we investigate the properties of K_k defined by the Riccati recursion (3.2.7b) and the closed-loop matrices D_k defined in (3.2.7e). In the following, we only consider the index sets that are UCC(λ_C) according to Definition 3.2.2.

Proposition 3.2.7. *Under Assumption 3.2.1, if the index set $\mathcal{I}$ is UCC(λ_C), then for any $n_1 \leq q \leq n_2$, we have that $\|K_q\|_2 \leq \beta$ for some $\beta > 0$ independent of n_1, n_2, and the particular choice of $\mathcal{I}$.*

Proof. We note from the definition (3.2.7b) of matrix K_q that, while it is a function of the quantities in Definition 3.2.2, it does not depend on the reference d_k, or the shift f_k. We will thus reason about it on the system for which d_k and f_k are 0. That is, for any $x_q \in \mathbb{R}^n$, consider the problem

$$\min_{u_{q:n_2-1}} \quad \sum_{k=q}^{n_2-1} u_k^T \hat{R}_k u_k + x_k^T Q_k x_k + x_{n_2}^T Q_{n_2} x_{n_2} \tag{3.2.10a}$$

$$\text{s.t.} \quad x_{k+1} = A_k x_k + \hat{B}_k u_k, \qquad q \leq k \leq n_2 - 1. \tag{3.2.10b}$$

For $k \geq q$, successively applying $x_{k+1} = A_k x_k + \hat{B}_k u_k$ gives that for $j \geq 0$

$$x_{q+j} - \left(\prod_{l=0}^{j-1} A_{q+l} \right) x_q = \left[\hat{B}_{q+j-1} \quad A_{q+j-1}\hat{B}_{q+j-2} \quad \ldots \left(\prod_{l=1}^{j-1} A_{q+l} \right) \hat{B}_q \right] \begin{bmatrix} u_{q+j-1} \\ \vdots \\ u_q \end{bmatrix}, \tag{3.2.11}$$

and for $j = t$, (3.2.11) reduces to

$$x_{q+t} - \left(\prod_{l=0}^{t-1} A_{q+l} \right) x_q = C_{q,t} \begin{bmatrix} u_{q+t-1} \\ \vdots \\ u_q \end{bmatrix}.$$

$\mathcal{I}$ being UCC(λ_C) implies $C_{q,t}$ is uniformly completely controllable, and in particular that

$C_{q,t}$ is full rank. Then there exists $\hat{u} = (\hat{u}_q^T, \ldots, \hat{u}_{q+t-1}^T)^T$ so that

$$-\left(\prod_{l=0}^{t-1} A_l\right) x_q = C_{q,t} \begin{bmatrix} \hat{u}_{q+t-1} \\ \vdots \\ \hat{u}_q \end{bmatrix}. \tag{3.2.12}$$

Several $\hat{u}$ satisfy this relationship; we consider the one defined by

$$\hat{u} = -C_{q,t}^T (C_{q,t} C_{q,t}^T)^{-1} \left(\prod_{l=0}^{t-1} A_{q+l}\right) x_q.$$

Denote the corresponding states generated with $\hat{u}_{q:q+t-1}$ as $\hat{x}_{q:q+t}$, then $\hat{x}_{q+t} = \mathbf{0}$ by (3.2.12).

Assumption 3.2.1 implies that

$$\begin{aligned}
&\max_{1\le j\le t} \left\| \left[\hat{B}_{q+j-1} \quad A_{q+j-1}\hat{B}_{q+j-2} \quad \cdots \left(\prod_{l=1}^{j-1} A_{q+l}\right) \hat{B}_q \right] \right\|_2 \\
&\le \max_{1\le j\le t} \left(C_B + C_A C_B + \cdots + C_A^{j-1} C_B \right) \\
&\le \frac{C_B\left(1 - C_A^t\right)}{1 - C_A} \triangleq M.
\end{aligned}$$

Then from Assumption 3.2.1 and Definition 3.2.2, we have that

$$\|\hat{u}\| \le \frac{M}{\lambda_C} C_A^t \|x_q\|. \tag{3.2.13}$$

From (3.2.11), we have, for $1 \le j \le t-1$, that

$$\|\hat{x}_{q+j}\| \le C_A^j \|x_q\| + M\|\hat{u}\| \le \left(C_A^j + \frac{M^2}{\lambda_C} C_A^t \right) \|x_q\|. \tag{3.2.14}$$

Now we let $\hat{u}_k = \mathbf{0}$ for $k \ge q+t$. Then it follows that $\hat{x}_k = \mathbf{0}$ for $k \ge q+t$. Also note that (3.2.10) is a standard linear-quadratic regulator problem, and the optimal value is given by

$x_q^T K_q x_q$ [20]. As a result, we have that

$$\begin{aligned}
x_q^T K_q x_q &= \min_{u_k} \sum_{k=q}^{n_2-1} x_k^T Q_k x_k + u_k^T \hat{R}_k u_k + x_{n_2}^T Q_{n_2} x_{n_2} \\
&\leq \sum_{k=q}^{n_2-1} \hat{x}_k^T Q_k \hat{x}_k + \hat{u}_k^T \hat{R}_k \hat{u}_k + \hat{x}_{n_2}^T Q_{n_2} \hat{x}_{n_2} \\
&\leq \sum_{k=q}^{q+t-1} \hat{x}_k^T Q_k \hat{x}_k + \hat{u}_k^T \hat{R}_k \hat{u}_k \\
&\leq C_Q \sum_{k=q}^{q+t-1} \|\hat{x}_k\|^2 + C_R \sum_{k=q}^{q+t-1} \|\hat{u}_k\|^2 \\
&\overset{(3.2.13),(3.2.14)}{\leq} C_Q \left(1 + \sum_{i=1}^{t-1} \left(C_A^i + \frac{M^2}{\lambda_C} C_A^t \right)^2 \right) \|x_q\|^2 + C_R \frac{M^2 C_A^{2t}}{\lambda_C^2} \|x_q\|^2.
\end{aligned}$$

Letting

$$\beta = C_Q \left(1 + \sum_{i=1}^{t-1} \left(C_A^i + \frac{M^2}{\lambda_C} C_A^t \right)^2 \right) + C_R \frac{M^2 C_A^{2t}}{\lambda_C^2}$$

completes the proof. Note that β only depends on the quantities in Definition 3.2.2 and Assumption 3.2.1, which are independent of n_1, n_2, and the particular choice of $\mathcal{I}$ given it is UCC(λ_C). □

In the following, we prove that the closed-loop system is asymptotically stable with an exponential decay rate. While the asymptotic result is well known, we need bounds on the decay rate at any time index; this is what we prove below. The proof is motivated by [124].

Proposition 3.2.8. *Under Assumption 3.2.1, if the index set $\mathcal{I}$ is UCC(λ_C), then for any $q \leq j \leq n_2 - 1$, we have that*

$$\left\| \prod_{l=q}^{j} D_l \right\|_2 \leq C_1 \rho^{j-q+1},$$

where $C_1 = \sqrt{\beta/\lambda_Q}$, $\rho = 1/\sqrt{1 + (\lambda_Q/\beta)}$ and C_1, ρ are independent of n_1, n_2, and the *particular choice of $\mathcal{I}$.*

Proof. It is shown in [20] that the recursion (3.2.7b) is equivalent to

$$K_k = {D_k}^T K_{k+1} D_k + Q_k + {L_k}^T \hat{R}_k L_k. \tag{3.2.15}$$

For $q \leq j \leq n_2 - 1$, define $x_{j+1} = D_j x_j$. Note that in this proof, x_j is a synthetic sequence, and not the solution of the problems (3.2.1) or (3.2.3). Therefore the properties of x_j defined here do not necessarily reflect those of the solution sequence. Then (3.2.15) and Proposition 3.2.7 imply that

$$\begin{aligned} x_j^T K_j x_j &\geq x_{j+1}^T K_{j+1} x_{j+1} + x_j^T Q_j x_j \\ &\geq x_{j+1}^T K_{j+1} x_{j+1} + \frac{\lambda_Q}{\beta} x_j^T K_j x_j \\ &\geq \left(1 + \frac{\lambda_Q}{\beta}\right) x_{j+1}^T K_{j+1} x_{j+1}. \end{aligned} \tag{3.2.16}$$

Here we used the bounds from Assumption 3.2.1 and the fact that $x_j^T K_j x_j \geq x_{j+1}^T K_{j+1} x_{j+1}$, as implied by (3.2.15) and the positive definiteness of Q_k, $\hat{R}_k$. Also we have that

$$x_j^T K_j x_j \geq x_j^T Q_j x_j \geq \lambda_Q \|x_j\|^2. \tag{3.2.17}$$

As a result, for $n_2 - 1 \geq j \geq q$, we have the following:

$$\begin{aligned} \left\| \prod_{l=q}^{j} D_l x_q \right\|^2 = \|x_{j+1}\|^2 &\overset{(3.2.17)}{\leq} \frac{1}{\lambda_Q} x_{j+1}^T K_{j+1} x_{j+1} \\ &\overset{(3.2.16)}{\leq} \frac{1}{\lambda_Q (1 + \lambda_Q/\beta)} x_j^T K_j x_j \\ &\overset{(3.2.16)}{\leq} \frac{1}{\lambda_Q} \left(\frac{1}{1 + \lambda_Q/\beta} \right)^{j-q+1} x_q^T K_q x_q \\ &\overset{\text{Prop 3.2.7}}{\leq} \frac{\beta}{\lambda_Q} \left(\frac{1}{1 + \lambda_Q/\beta} \right)^{j-q+1} \|x_q\|^2, \end{aligned}$$

where the third inequality is obtained by repeatedly applying (3.2.16). □

We have the following uniform boundedness result of matrices frequently used in the rest of this Section.

Lemma 3.2.9. *Under Assumption 3.2.1, if the index set $\mathcal{I}$ is UCC(λ_C), then for any $n_1 \leq k \leq n_2 - 1$, we have that*

$$\|E_k\|_2 \leq C_E, \qquad \|L_k\|_2 \leq C_L, \qquad \|f_k\|_2 \leq l_0,$$

for some C_E, C_L and l_0 independent of n_1, n_2, and the particular choice of $\mathcal{I}$. Here E_k is defined in Definition 3.2.5, L_k in (3.2.7d)*, and f_k in* (3.2.4)*.*

Proof. See Appendix B.1.3. □

Next, we investigate properties of the optimal states x_k^* and controls u_k^* for problem (3.2.3). Due to the one-to-one correspondence between solutions of problems (3.2.3) and (3.2.5), we first consider the optimal states of (3.2.5) obtained in Proposition 3.2.6. We have the following lemma characterizing the dependence of x_k^* on d_i and f_i.

Lemma 3.2.10. *Let C_i^k and F_i^k be defined as in Proposition 3.2.6. Under Assumption 3.2.1, if the index set $\mathcal{I}$ is UCC(λ_C), we have that*

$$\|F_i^k\|_2,\ \|C_i^k\|_2 \leq C_2 \rho^{|i-k|},$$

for some $C_2 > 0$ independent of n_1, n_2, and the particular choice of $\mathcal{I}$. Here $\rho = 1/\sqrt{1 + (\lambda_Q/\beta)}$ as in Proposition 3.2.8.

Proof. See Appendix B.1.4. □

Proposition 3.2.8 and Lemma 3.2.10 establish the exponential decay properties with respect to $|k - n_1|$ and $|i - k|$ for matrices $\prod_{i=n_1}^{k-1} D_i$ and C_i^k, which encode the dependencies of the optimal states x_k^* of problem (3.2.5) on the initial value and the reference d_i, respectively, by Proposition 3.2.6. This property is the key to prove the following main result of this Section. Proposition 3.2.11 bounds the dependence of solutions x_k^*, u_k^* on the initial value x_{n_1} and terminal reference d_{n_2} with an exponential term. The importance of this result is shown in Section 3.3 when we investigate the sensitivity of problem (3.2.1) to the initial value and terminal reference.

Proposition 3.2.11. *Let x_k^* and u_k^* be the optimal states and controls of problem* (3.2.3). *Under Assumption 3.2.1, if the index set $\mathcal{I}$ is UCC(λ_C), then we have that*

$$\begin{aligned}
&\|\nabla_{x_{n_1}} x_k^*\|_2 \le Z_1 \rho^{k-n_1},\ \|\nabla_{d_{n_2}} x_k^*\|_2 \le Z_2 \rho^{n_2-k},\ n_1+1 \le k \le n_2,\\
&\|\nabla_{x_{n_1}} u_k^*\|_2 \le Z_1 \rho^{k-n_1},\ \|\nabla_{d_{n_2}} u_k^*\|_2 \le Z_2 \rho^{n_2-k},\ n_1 \le k \le n_2-1,
\end{aligned}$$

for some $Z_1, Z_2 > 0$ independent of n_1, n_2, and the particular choice of $\mathcal{I}$.

Proof. Due to the change of variable (3.2.4), the optimal states of problems (3.2.3) and (3.2.5) are the same, and the unconstrained parts of the optimal controls differ by a constant sequence. As a result, Proposition 3.2.6 and the optimal control law (3.2.6) give the following:

$$\begin{aligned}
\left\|\nabla_{x_{n_1}} x_k^*\right\|_2 &= \left\|\prod_{i=n_1}^{k-1} D_i\right\|_2 \overset{\text{Prop 3.2.8}}{\le} C_1 \rho^{k-n_1},\\
\left\|\nabla_{d_{n_2}} x_k^*\right\|_2 &= \left\|C_{n_2}^k\right\|_2 \overset{\text{Lemma 3.2.10}}{\le} C_2 \rho^{n_2-k},\\
\left\|\nabla_{x_{n_1}} u_k^*\right\|_2 &= \left\|\nabla_{x_{n_1}} v_k^*\right\|_2 \overset{(3.2.6)}{=} \|L_k \nabla_{x_{n_1}} x_k^*\|_2 \overset{\text{Lemma 3.2.9}}{\le} C_L \left\|\nabla_{x_{n_1}} x_k^*\right\|_2 \le C_L C_1 \rho^{k-n_1},\\
\left\|\nabla_{d_{n_2}} u_k^*\right\|_2 &= \left\|\nabla_{d_{n_2}} v_k^*\right\|_2 \overset{(3.2.6)}{\le} C_L \left\|\nabla_{d_{n_2}} x_k^*\right\|_2 + \left\|W_k^{-1} \hat{B}_k^T \left(\prod_{l=k+1}^{n_2-1} D_l\right)^T Q_{n_2}\right\|_2\\
&\overset{\text{Prop 3.2.8}}{\le} C_L \left\|\nabla_{d_{n_2}} x_k^*\right\|_2 + \frac{C_Q C_B}{\lambda_R} C_1 \rho^{n_2-k-1}\\
&\le C_L C_2 \rho^{n_2-k} + \frac{C_Q C_B}{\lambda_R} C_1 \rho^{n_2-k-1}.
\end{aligned}$$

Denoting $Z_1 = \max\left(C_1, C_1 C_L\right)$, and $Z_2 = \max\big(C_2, C_L C_2 + C_Q C_B C_1/\lambda_R \rho\big)$ completes the proof. □

The next result gives an uniform upper bound for the solutions of problem (3.2.3) whose index set is UCC(λ_C). First, we make the following assumptions about the size of the initial value $x^0_{n_1}$ and the reference trajectory.

Assumption 3.2.12. *For any n_1, n_2 and $n_1 \leq q \leq n_2$, we have that*

(a) $\|x^0_{n_1}\|_2 \leq u_0$ for some $u_0 > 0$,

(b) $\|d_q\|_2 \leq m_0$ for some $m_0 > 0$.

Since the initial state is part of input to the system, we can reasonably assume that the values are taken in some compact set. Note that the reference trajectory models the demand in a PCM which is our target application area. If we were to analyze asymptotics of our problem as $n_2 \to \infty$, uniformly bounded demand would be a tenuous assumption (though, with peak population scenarios currently considered, not impossible). The results here can be extended to polynomial increase of demand (as it will be compensated by exponential decays with rate ρ). To simplify the algebra, at this time we use Assumption 3.2.12(b), where the demand/reference trajectory is uniformly bounded over time.

Lemma 3.2.13. *Let x^*_k and u^*_k be the optimal states and controls of problem* (3.2.3). *Under Assumptions 3.2.1 and 3.2.12, if the index set $\mathcal{I}$ is UCC(λ_C), we have that,*

$$\|x^*_k\|_2 \leq C_g,\ n_1 + 1 \leq k \leq n_2; \ \ \|u^*_k\|_2 \leq C_u,\ n_1 \leq k \leq n_2 - 1$$

for some C_g, $C_u > 0$ independent of n_1, n_2, and the particular choice of $\mathcal{I}$.

Proof. Note again that, for problems (3.2.3) and (3.2.5), the optimal states are identical, and the unconstrained parts of the optimal controls satisfy the relation $v^*_k = \hat{u}^*_k + \hat{R}_k^{-1}\bar{R}_k\widetilde{b}_k$ by

(3.2.4). Consequently, Proposition 3.2.6, Lemma 3.2.9, and Lemma 3.2.10 give the following.

$$\begin{aligned}\|x_k^*\|_2 &\leq C_1\rho^{k-n_1}u_0 + \sum_{i=n_1+1}^{n_2} C_2\rho^{|k-i|}m_0 + \sum_{i=n_1}^{n_2-1} C_2\rho^{|k-i|}l_0 \\ &\leq C_1u_0 + 2m_0C_2\sum_{s=0}^{\infty}\rho^s + 2l_0C_2\sum_{s=0}^{\infty}\rho^s \\ &= C_1u_0 + \frac{2(m_0+l_0)C_2}{1-\rho} \triangleq C_g,\end{aligned}$$

where m_0 and u_0 are the bounds on the reference trajectory and initial state defined in Assumption 3.2.12. Note that (3.2.7c) gives that $\|W_k^{-1}\|_2 \leq 1/\lambda_R$. The optimal control law (3.2.6) and Proposition 3.2.8 give the following.

$$\begin{aligned}\|u_k^*\|_2 &\leq \|\widetilde{u}_k\|_2 + \|\hat{u}_k^*\|_2 = \|\widetilde{b}_k\|_2 + \|v_k^* - \hat{R}_k^{-1}\bar{R}_k\widetilde{b}_k\|_2 \leq \|v_k^*\|_2 + \|\widetilde{b}_k\|_2 + \frac{C_R}{\lambda_R}\|b_k\|_2 \\ &\leq C_LC_g + \frac{C_QC_B}{\lambda_R}\sum_{i=k+1}^{n_2}\rho^{i-k-1}m_0 + \frac{\beta C_B}{\lambda_R}\sum_{i=k+1}^{n_2-1}\rho^{i-k}l_0 + \frac{\beta C_B}{\lambda_R}l_0 + \left(2+\frac{C_R}{\lambda_R}\right)U \\ &\leq C_LC_g + \frac{m_0C_QC_B}{\lambda_R(1-\rho)} + \frac{l_0\beta C_B}{\lambda_R(1-\rho)} + \left(2+\frac{C_R}{\lambda_R}\right)U \triangleq C_u.\end{aligned}$$

This completes the proof. □

3.2.2 Box constrained control inequality problem

For the rest of this Section, we return to the inequality constrained problem (3.2.1), and investigate properties of its solutions and Lagrange multipliers using the results derived for problem (3.2.3). We make the following assumption about the active set $\mathcal{A}$ of problem (3.2.1).

Assumption 3.2.14. *The active set $\mathcal{A}$ of problem* (3.2.1) *is UCC(λ_C) as defined in Definition 3.2.2 (b).*

Corollary 3.2.15. *Let x_k^* and u_k^* be the optimal states and controls of problem* (3.2.1).

Under Assumptions 3.2.1, 3.2.12, and 3.2.14, we have that,

$$\|x_k^*\|_2 \le C_g,\ n_1 + 1 \le k \le n_2;\ \ \|u_k^*\|_2 \le C_u,\ n_1 \le k \le n_2 - 1$$

for C_g, $C_u > 0$ as those in Lemma 3.2.13.

Proof. Note that when the index set defining the equality constrained problem (3.2.3) is the active set $\mathcal{A}$ of problem (3.2.1), problems (3.2.1) and (3.2.3) have the same solution. Since $\mathcal{A}$ is UCC(λ_C), Lemma 3.2.13 gives the conclusion. □

In the following, for problem (3.2.1), we investigate the adjoint variables which are the Lagrange multipliers associated with the constraints $x_{k+1} = A_k x_k + B_k u_k$.

Proposition 3.2.16. *Let x_k^*, u_k^* be the solutions, and ϕ_k^* be the optimal adjoint variables for problem* (3.2.1). *For $n_1 \le k \le n_2 - 1$, we have that*

$$\phi_k^* = 2K_{k+1}x_{k+1}^* - 2\sum_{i=k+1}^{n_2}\left(M_i^{k+1}\right)^T d_i - 2\sum_{i=k+1}^{n_2-1}\left(S_i^{k+1}\right)^T f_i, \tag{3.2.18}$$

where K_k, M_i^k, S_i^k and f_i are defined with respect to the active constraints $C_k u_k^* = \bar{b}_k$ of problem (3.2.1) *at optimality.*

Proof. See Appendix B.1.5. □

Lemma 3.2.17. *Let ϕ_k^* be the optimal adjoint variables for problem* (3.2.1). *Then under Assumptions 3.2.1, 3.2.12, and 3.2.14, for $n_1 \le k \le n_2 - 1$, we have that*

$$\|\phi_k^*\| \le C_\phi$$

for some $C_\phi > 0$ independent of n_1 and n_2.

Proof. See Appendix B.1.6. □

3.3 A temporal decomposition approach

In this Section, we define a temporal decomposition approach to approximate the solutions and optimal values of problem (3.2.1). To partition the entire horizon, we decompose $[n_1, n_2]$ into n_0 subintervals of the same length $p = (n_2 - n_1)/n_0$. Denote the subintervals as

$$S_i = [n_1 + (i-1)p, n_1 + ip], \qquad i = 1, \dots, n_0. \tag{3.3.1}$$

For some buffer size $0 < \Omega < p$, define an embedding region F_i for each S_i as $F_i = \left[n_1'(i), n_2'(i)\right]$, where

$$n_1'(i) = \begin{cases} n_1, & i = 1, \\ n_1 + (i-1)p - \Omega, & i = 2, \dots, n_0, \end{cases} \qquad n_2'(i) = \begin{cases} n_1 + ip + \Omega, & i = 1, \dots, n_0 - 1, \\ n_2, & i = n_0. \end{cases} \tag{3.3.2}$$

Note that $S_i \subset F_i$ for $i = 1, \dots, n_0$. Figure 3.1 shows an illustration of such a decomposition scheme when $n_0 = 3$. We define the following parametrized problem.

Definition 3.3.1. *For $i = 1, \dots, n_0$, let $\theta = (\theta^{(h)}, \theta^{(d)})$. We define the parametrized problem P_θ^i as follows:*

$$\min_{w_k, h_k} \quad \sum_{k=n_1'(i)}^{n_2'(i)-1} \Big(w_k^T R_k w_k + (h_k - d_k)^T Q_k (h_k - d_k) \Big) \tag{3.3.3a}$$

$$+ (h_{n_2'(i)} - d_{n_2'(i)})^T Q_{n_2'(i)} (h_{n_2'(i)} - d_{n_2'(i)}) \tag{3.3.3b}$$

$$\text{s.t.} \quad h_{k+1} = A_k h_k + B_k w_k, \qquad n_1'(i) \le k \le n_2'(i) - 1 \tag{3.3.3c}$$

$$l_k \le w_k \le b_k, \qquad n_1'(i) \le k \le n_2'(i) - 1 \tag{3.3.3d}$$

$$h_{n_1'(i)} = \theta^{(h)}, \;\; d_{n_2'(i)} = \theta^{(d)}, \tag{3.3.3e}$$

where $d_{n_1'(i):n_2'(i)-1}$ are the same as those in problem (3.2.1).

Note that problem P_θ^i is problem (3.2.1) defined on a shorter interval F_i, but with a possibly different terminal reference vector $\theta^{(d)}$ and initial state $\theta^{(h)} = h^0_{n_1'(i)}$. For the latter, we invoke an assumption similar to Assumption 3.2.12.

Assumption 3.3.2. *For $i = 1, \ldots, n_0$, let $h^0_{n_1'(i)}$ be the initial value of problem P_θ^i, then $\|h^0_{n_1'(i)}\|_2 \le u_0$, where u_0 is the same as that in Assumption 3.2.12.*

Let $\theta_0(i) = (h^0_{n_1'(i)}, d_{n_2'(i)})$, where $h^0_{n_1'(i)}$ is any initial value satisfying Assumption 3.3.2 and $h^0_{n_1'(1)} = x^0_{n_1}$, and where $d_{n_2'(i)}$ is the reference in problem (3.2.1). Let $x^*_{n_1+1:n_2}$, $u^*_{n_1:n_2-1}$ be the optimal states and controls of problem (3.2.1) respectively, and let $h^*_{F_i}$ and $w^*_{F_i}$ be the optimal states and controls of problem $P^i_{\theta_0(i)}$. Denote

$$J\left([m_1, m_2], [n_1, n_2], x^0_{n_1}\right) \triangleq \sum_{k=m_1}^{m_2-1} {u^*_k}^T R_k u^*_k + (x^*_k - d_k)^T Q_k (x^*_k - d_k) \\ + (x^*_{n_2} - d_{n_2})^T Q_{n_2} (x^*_{n_2} - d_{n_2}) \mathbf{1}_{(m_2 = n_2)} \tag{3.3.4}$$

where $x^*_{m_1:m_2-1}$ and $u^*_{m_1:m_2-1}$ are respectively the optimal states and controls of problem (3.2.1) on $[n_1, n_2]$ with initial value $x^0_{n_1}$ restricted to $[m_1, m_2] \subset [n_1, n_2]$. Then the temporal decomposition approach consists of the approximation

$$J\left([n_1, n_2], [n_1, n_2], x^0_{n_1}\right) = \Gamma_{n_1:n_2}\left(u^*_{n_1:n_2-1}, x^*_{n_1+1:n_2}\right)$$

by $\sum_{i=1}^{n_0} J\left(S_i, F_i, h^0_{n_1'(i)}\right)$.

In other words, the optimal value of problem (3.2.1) is approximated by solving problem $P^i_{\theta_0(i)}$ on each embedding region F_i and summing over the solutions restricted on the subintervals $S_i \subset F_i$. On the target intervals S_i, solving $P^i_{\theta_0(i)}$ results in the states $h^*_{S_i}$ and controls $w^*_{S_i}$. To bound the error of this approximation, we need to relate the solution of $P^i_{\theta_0(i)}$ to the solution of (3.2.1) when solved on the full horizon.

To this end we define a *modified problem* on the embedding horizon F_i, whose solution vector is the same as the restriction to F_i of the solution of (3.2.1) for the full horizon $[n_1, n_2]$.

The modified problem is also an instance of (3.3.3), but its solution vector will be precisely the solution of (3.2.1) restricted to F_i. We have the following result.

Proposition 3.3.3. *Let* $(u^*_{n_1:n_2-1}, x^*_{n_1+1:n_2})$ *be the solutions and* ϕ^*_k *be the adjoint variables of problem* (3.2.1). *For* $i = 1, \ldots, n_0$, *define*

$$\begin{aligned}
\hat{h}_{n'_1(i)} &= \begin{cases} x^0_{n_1}, & i = 1 \\ x^*_{n'_1(i)}, & i = 2, \ldots, n_0, \end{cases} \\
\hat{d}_{n'_2(i)} &= \begin{cases} -Q^{-1}_{n'_2(i)}\phi^*_{n'_2(i)-1}/2 + x^*_{n'_2(i)}, & i = 1, \ldots, n_0 - 1 \\ d_{n_2}, & i = n_0. \end{cases}
\end{aligned} \tag{3.3.5}$$

Then $(u^*_{n'_1(i):n'_2(i)-1}, x^*_{n'_1(i)+1:n'_2(i)})$ *satisfies the KKT conditions and the second-order sufficient conditions of problem* $P^i_{\theta_1(i)}$ *with* $\theta_1(i) = (\hat{h}_{n'_1(i)}, \hat{d}_{n'_2(i)})$.

Proof. For $k = n_1, \ldots, n_2 - 1$, let $C_k u^*_k = \bar{b}_k$ be the active box constraints for problem (3.2.1) at optimality, and let λ^*_k be the associated optimal Lagrange multipliers. The KKT conditions for problem (3.2.1) are

$$2R_k u^*_k - C^T_k \lambda^*_k + B^T_k \phi^*_k = 0, \qquad n_1 \le k \le n_2 - 1 \tag{3.3.6a}$$

$$2Q_k(x^*_k - d_k) + A^T_k \phi^*_k - \phi^*_{k-1} = 0, \qquad n_1 + 1 \le k \le n_2 - 1 \tag{3.3.6b}$$

$$2Q_{n_2}(x^*_{n_2} - d_{n_2}) - \phi^*_{n_2-1} = 0, \tag{3.3.6c}$$

$$x^*_{k+1} = A_k x^*_k + B_k u^*_k, \qquad n_1 \le k \le n_2 - 1, \qquad x_{n_1} = x^0_{n_1}, \tag{3.3.6d}$$

$$l_k \le u^*_k \le b_k, \qquad n_1 \le k \le n_2 - 1 \tag{3.3.6e}$$

$$\lambda^*_k \ge 0, \qquad n_1 \le k \le n_2 - 1. \tag{3.3.6f}$$

Then for problem $P^i_{\theta_1(i)}$ with parameters $\hat{h}_{n'_1(i)}$ and $\hat{d}_{n'_2(i)}$ defined in (3.3.5), the KKT conditions are satisfied by the same solutions $(u^*_{n'_1(i):n'_2(i)-1}, x^*_{n'_1(i)+1:n'_2(i)})$ with the same

Lagrange multipliers λ_k^*, ϕ_k^* as follows:

$$2R_k u_k^* - C_k^T \lambda_k^* + B_k^T \phi_k^* = 0, \qquad n_1'(i) \le k \le n_2'(i) - 1 \tag{3.3.7a}$$

$$2Q_k(x_k^* - d_k) + A_k^T \phi_k^* - \phi_{k-1}^* = 0, \qquad n_1'(i) + 1 \le k \le n_2'(i) - 1 \tag{3.3.7b}$$

$$2Q_{n_2'(i)}(x_{n_2'(i)}^* - \hat{d}_{n_2'(i)}) - \phi_{n_2'(i)-1}^* = 0, \tag{3.3.7c}$$

$$x_{k+1}^* = A_k x_k^* + B_k u_k^*, \qquad n_1'(i) \le k \le n_2'(i) - 1, \qquad x_{n_1'(i)} = \hat{h}_{n_1'(i)}, \tag{3.3.7d}$$

$$l_k \le u_k^* \le b_k, \qquad n_1'(i) \le k \le n_2'(i) - 1 \tag{3.3.7e}$$

$$\lambda_k^* \ge 0, \qquad n_1'(i) \le k \le n_2'(i) - 1, \tag{3.3.7f}$$

where (3.3.7a)–(3.3.7b) and (3.3.7e)–(3.3.7f) directly follow from (3.3.6a)–(3.3.6b) and (3.3.6e)–(3.3.6f), respectively. Equations (3.3.7c) and (3.3.7d) follow from definitions of $\hat{h}_{n_1'(i)}$ and $\hat{d}_{n_2'(i)}$, respectively. The second-order condition is satisfied by virtue of the strong convexity of the problem. □

Proposition 3.3.3 indicates that, in order for problem P_θ^i to have the same solutions as problem (3.2.1) on F_i, the modified parameters (3.3.5) need to incorporate information from (3.2.1) about the adjoint variables $\phi_{n_2'(i)-1}^*$, and about the states $x_{n_1'(i)}^*$, $x_{n_2'(i)}^*$. We note that problem $P_{\theta_1(i)}^i$ defined in Proposition 3.3.3 is notional. It cannot be set up without having solved the full horizon $[n_1, n_2]$ problem, but its solution vector is *identical* to that of the full horizon problem restricted to F_i. In the following, we will prove that the solution of problem $P_{\theta_1(i)}^i$, on the subinterval S_i, is sufficiently close to that of $P_{\theta_0(i)}^i$. The latter problem is computable by using the reference trajectory corresponding only to the short interval F_i. Note that problem $P_{\theta_1(i)}^i$ can be viewed as the result of perturbing the parameter of problem $P_{\theta_0(i)}^i$. Therefore, to prove the relationship between the solutions of $P_{\theta_0(i)}^i$ and $P_{\theta_{1(i)}}^i$, we use the parametric sensitivity results derived from [22]. We note that our base problem (3.2.1) is a quadratic program, for which several results concerning the Lipschitz continuity with respect to parameters exist [22, 58]. Our aim, however, concerns more specific elements of the solution and seeks stronger results than directly using [22, 58] would allow. We aim to show that the entries corresponding to *a subset* of the solution vector components (the ones corresponding to the subintervals S_i in Figure 3.1) is Lipschitz continuous with respect to the initial state and terminal reference on the embedding regions F_i, *but with a Lipschitz constant L that decays exponentially in the buffer size Ω.* To achieve such an objective, we compute the directional derivative of the target components with respect to the perturbations, using results from [22], and then show that its value can be upper bounded using results such as Proposition 3.2.11. In turn, this gives the sought-after exponential decay result.

Definition 3.3.4. *For $\theta \in \mathbb{R}^q$, define the one-sided directional derivative of $y(\theta)$ along a direction $p \in \mathbb{R}^q$ at θ_0 as*

$$D_p y(\theta_0) = \lim_{t \downarrow 0} \frac{y(\theta_0 + tp) - y(\theta_0)}{t},$$

given that the limit exists.

Lemma 3.3.5. *Consider the following parametrized quadratic programming problem*

$$\begin{aligned} \min \quad & f(y,\theta) \overset{\Delta}{=} y^T G y/2 + y^T c(\theta) + \theta^T F \theta + y^T c_1 + \theta^T c_2 + C \\ \text{s.t.} \quad & Ay - r \leq 0 \\ & By - d(\theta) = 0, \end{aligned} \tag{3.3.8}$$

where G, F are positive definite, $\theta \in \mathbb{R}^q$ and $A^T = \begin{bmatrix} a_1, \dots, a_m \end{bmatrix} \in \mathbb{R}^{n \times m}$. Denote the solution of problem (3.3.8) *as $y(\theta)$. When $\theta = \theta_0$, let $y_0 = y(\theta_0)$ and the Lagrange multiplier corresponding to y_0 be $\bar{\lambda}$. Denote $I(y_0, \theta_0) = \{i : a_i^T y_0 = r_i, i = 1, \dots, m\}$ be the set of active inequality constraints, $I_+(y_0, \theta_0, \bar{\lambda}) = \{i \in I(y_0, \theta_0) : \bar{\lambda}_i > 0\}$ and $I_0(y_0, \theta_0, \bar{\lambda}) = \{i \in I(y_0, \theta_0) : \bar{\lambda}_i = 0\}$. If the linear independence constraint qualification (LICQ) holds at $y(\theta_0)$, then for any $p \in \mathbb{R}^q$, we have that*

$$D_p y(\theta_0) \quad = \quad \left(\frac{d y^*_{I'(\theta_0)}(\theta)}{d\theta} \bigg|_{\theta = \theta_0} \right) p,$$

where $y^*_{I'(\theta_0)}(\theta)$ *is the solution of the problem*

$$\begin{aligned} \min \quad & f(y,\theta) = y^T Gy/2 + y^T c(\theta) + \theta^T F\theta + y^T c_1 + \theta^T c_2 + C \\ \text{s.t.} \quad & A_{I'(\theta_0)} y - r' = 0 \\ & By - d(\theta) = 0, \end{aligned} \tag{3.3.9}$$

and where $I'(\theta_0) = I_+(y_0, \theta_0, \bar{\lambda}) \cup I_1$ *for some* $I_1 \subset I_0(y_0, \theta_0, \bar{\lambda})$*, and* $A_{I'(\theta_0)} = [a_i^T]_{i \in I'(\theta_0)}$*,* $r' = [r_i]_{i \in I'(\theta_0)}$*.*

Proof. See Appendix B.1.7. □

With Lemma 3.3.5, we are now ready to investigate the effect on solutions of perturbing the parameters of problem P^i_θ. Since the proof for each subinterval is the same, for notational simplicity we suppress the dependence of $n'_1(i)$, $n'_2(i)$ and P^i_θ on i whenever the index of the subinterval under consideration is clear.

Proposition 3.3.6. *Denote* $\theta_0 = (h^0_{n'_1}, d_{n'_2})$ *and* $\theta_1 = (\hat{h}_{n'_1}, \hat{d}_{n'_2})$ *as defined in* (3.3.5)*. For* $\theta = (\theta^{(h)}, \theta^{(d)})$*, let* $y(\theta)$ *be the solution of problem* P_θ*. We then have, for* $s \in [0,1]$

$$D_{\theta_1 - \theta_0} y\left(\theta_0 + s(\theta_1 - \theta_0)\right) = \left(\frac{dy^*_s(\theta)}{d\theta}\Big|_{\theta = \theta_0 + s(\theta_1 - \theta_0)} \right) (\theta_1 - \theta_0),$$

and $y^*_s(\theta)$ *is the solution of the following equality constrained problem,*

$$\begin{aligned} \min \quad & \sum_{k=n'_1}^{n'_2 - 1} w_k^T R_k w_k + (h_k - d_k)^T Q_k (h_k - d_k) + (h_{n'_2} - \theta^{(d)}_s)^T Q_{n'_2} (h_{n'_2} - \theta^{(d)}_s) \\ \text{s.t.} \quad & h_{k+1} = A_k h_k + B_k w_k, \qquad n'_1 \le k \le n'_2 - 1, \qquad h_{n'_1} = \theta^{(h)}_s, \\ & C_k(s) w_k = \bar{b}_k(s), \qquad n'_1 \le k \le n'_2 - 1, \end{aligned} \tag{3.3.10}$$

where rows of $C_k(s)$ *and* $\bar{b}_k(s)$ *are, respectively, subsets of rows of* $C'_k(s)$ *and* $\bar{b}'_k(s)$ *which are the selection matrix and bound vector defined by* (3.2.2) *corresponding to the active set of*

P_{θ_s} at optimality, and $\theta_s = \theta_0 + s(\theta_1 - \theta_0)$. In other words, the equations $C_k(s)w_k = \bar{b}_k(s)$ represent a subset of the active constraints of P_{θ_s} at optimality.

Proof. For any $\theta \in \mathbb{R}^{2n}$, problem P_θ is an instance of problem (3.3.8) with the following parameters:

$$G = \begin{bmatrix} 2R_{n_1'} & & & & & \\ & \ddots & & & & \\ & & 2R_{n_2'-1} & & & \\ & & & 2Q_{n_1'+1} & & \\ & & & & \ddots & \\ & & & & & 2Q_{n_2'} \end{bmatrix}, \quad F = \begin{bmatrix} Q_{n_1'} & \\ & Q_{n_2'} \end{bmatrix}, \quad r = \begin{bmatrix} b_{n_1'} \\ \vdots \\ b_{n_2'-1} \\ -l_{n_1'} \\ \vdots \\ -l_{n_2'-1} \end{bmatrix},$$

$$A = \left[\begin{array}{c|c} I_{(n_2'-n_1')m} & \\ & \mathbf{0}_{2(n_2'-n_1')m \times (n_2'-n_1')n} \\ -I_{(n_2'-n_1')m} & \end{array}\right], \quad c(\theta) = \begin{bmatrix} \mathbf{0}_{(n_2'-n_1')m+(n_2'-n_1'-1)n} \\ -2Q_{n_2'}\theta^{(d)} \end{bmatrix},$$

$$B = \left[\begin{array}{ccc|cccc} -B_{n_1'} & & & I & & & \\ & \ddots & & -A_{n_1'+1} & I & & \\ & & & & & \ddots & \\ & & -B_{n_2'-1} & & & -A_{n_2'-1} & I \end{array}\right], \quad d(\theta) = \begin{bmatrix} A_{n_1'}\theta^{(h)} \\ \mathbf{0}_{(n_2'-n_1'-1)n} \end{bmatrix}.$$

Here $Ax \leq r$ and $Bx = d(\theta)$ correspond respectively to the box constraints (3.3.3d) and the system dynamics (3.3.3c). Note that A and B have the same number of columns. G

and F are positive definite from Assumption 3.2.1. Let $\bar{A}$ be the matrix whose rows are subsets of rows of A corresponding to the active constraints of problem P_θ. Since an active constraint can achieve either lower or upper bound, but not both, the rows of $\bar{A}$ are linearly independent. Also, B has full row rank, and the rows of $\bar{A}$ are linearly independent of rows of B. As a result, LICQ holds for any $\theta \in \mathbb{R}^{2n}$ at optimality. For $s \in [0,1]$, directly applying Lemma 3.3.5 to problem P_{θ_s} gives the conclusion. □

Proposition 3.3.6 relates the directional derivative of the solution of problem P_{θ_s} with respect to the parameters θ to the solution of an equality constrained problem (3.3.10). Note that problem (3.3.10) has the same form as problem (3.2.3) for which we derive the exponential decay result Proposition 3.2.11 under some controllability conditions. Now we make similar controllability assumptions for the problems P_{θ_s}.

Assumption 3.3.7. *For $i = 1, \dots, n_0$ and $s \in [0,1]$, let $\theta_0(i) = (h^0_{n_1'(i)}, d_{n_2'(i)})$, $\theta_1(i) = (\hat{h}_{n_1'(i)}, \hat{d}_{n_2'(i)})$ as defined in* (3.3.5)*, and $\theta_s(i) = \theta_0(i) + s(\theta_1(i) - \theta_0(i))$, then the active sets of problems $P^i_{\theta_s(i)}$ at optimality are UCC(λ_C).*

Note that Assumption 3.3.7 assumes UCC(λ_C) for the active sets of the continuously indexed family of problems $P^i_{\theta_s(i)}$ on each embedding region F_i, which is stronger than Assumption 3.2.14 for problem (3.2.1). We note, however, that Assumption 3.3.7 is only made for the active sets at optimality.

Lemma 3.3.8. *Under Assumption 3.3.7, the index set for problem* (3.3.10) *is UCC(λ_C) for any $s \in [0,1]$ and $i = 1, \dots, n_0$.*

Proof. Since the proof for each $i = 1, \dots, n_0$ is the same, we suppress the dependence on i in the proof. By definition of problem (3.3.10), the index set $\mathcal{I}_s$ for (3.3.10), the problem that we use to compute the directional derivative of the solution with respect to the parameter θ, is a subset of the active set $\mathcal{A}_s$ for problem P_{θ_s}. As a result, the columns of the controllability matrix $C_{q,t}(\mathcal{A}_s)$ are contained in those of $C_{q,t}(\mathcal{I}_s)$. Since $\lambda_{min}(C_{q,t}(\mathcal{A}_s)C^T_{q,t}(\mathcal{A}_s)) \geq \lambda_C$ by

Assumption 3.3.7, we have that

$$\lambda_{min}(C_{q,t}(\mathcal{I}_s)C_{q,t}^T(\mathcal{I}_s)) \geq \lambda_{min}(C_{q,t}(\mathcal{A}_s)C_{q,t}^T(\mathcal{A}_s)) \geq \lambda_C,$$

and hence $\mathcal{I}_s$ is also UCC(λ_C). □

Together with Assumption 3.2.1, Lemma 3.3.8 justifies the application of the exponential decay result Proposition 3.2.11 to problem (3.3.10), and hence combined with Proposition 3.3.6, it bounds the distance between solutions of P_{θ_0} and P_{θ_1} as follows.

Proposition 3.3.9. *Let $y(\theta_0) = (w^*_{n_1':n_2'-1}, h^*_{n_1'+1:n_2'})$ be the solution of problem P_{θ_0}, and let $y(\theta_1) = (u^*_{n_1':n_2'-1}, x^*_{n_1'+1:n_2'})$ be the solution of problem P_{θ_1}. From Proposition 3.3.3, $y(\theta_1)$ is also the solution of problem* (3.2.1) *restricted to the embedding region F_i. Under Assumptions 3.2.1, 3.2.12, 3.2.14, 3.3.2 and 3.3.7, for $i = 1, \ldots, n_0$ and $k \in S_i$, we have that*

$$\|x_k^* - h_k^*\|_2, \;\; \|u_k^* - w_k^*\|_2 \leq Y\rho^{\Omega},$$

for some $Y > 0$ independent of n_1 and n_2, where ρ is defined in Proposition 3.2.8 and Ω is the buffer size as in (3.3.2).

Proof. From Leibniz-Newton, we have that

$$y(\theta_1) - y(\theta_0) = \int_0^1 D_{\theta_1-\theta_0} y\left(\theta_0 + s(\theta_1 - \theta_0)\right) ds,$$

which gives that

$$x_k^* - h_k^* = \int_0^1 D_{\theta_1-\theta_0}\widetilde{p}_k^*(\theta_s)\, ds, \;\; u_k^* - w_k^* = \int_0^1 D_{\theta_1-\theta_0}\widetilde{s}_k^*(\theta_s)\, ds, \tag{3.3.11}$$

where $(\widetilde{s}^*_{n_1':n_2'-1}(\theta_s), \widetilde{p}^*_{n_1'+1:n_2'}(\theta_s))$ is the solution of problem P_{θ_s}. Proposition 3.3.6 implies that

$$
\begin{aligned}
D_{\theta_1-\theta_0}\widetilde{p}_k^*(\theta_s) &= \begin{bmatrix} \nabla_{h_{n_1'}} p_k^*(\theta_s) & \nabla_{d_{n_2'}} p_k^*(\theta_s) \end{bmatrix} \begin{bmatrix} \hat{h}_{n_1'} - h_{n_1'}^0 \\ \hat{d}_{n_2'} - d_{n_2'} \end{bmatrix}, \\
D_{\theta_1-\theta_0}\widetilde{s}_k^*(\theta_s) &= \begin{bmatrix} \nabla_{h_{n_1'}} s_k^*(\theta_s) & \nabla_{d_{n_2'}} s_k^*(\theta_s) \end{bmatrix} \begin{bmatrix} \hat{h}_{n_1'} - h_{n_1'}^0 \\ \hat{d}_{n_2'} - d_{n_2'} \end{bmatrix},
\end{aligned}
$$

where $(s^*_{n_1':n_2'-1}(\theta_s), p^*_{n_1'+1:n_2'}(\theta_s))$ is the solution of the equality constrained problem (3.3.10). Note that each P_{θ_s} may have a different active set, which may also be different from that of problem (3.2.1). However, under Assumption 3.3.7, the active set of every P_{θ_s} is UCC(λ_C), and Lemma 3.3.8 implies that the index set for the corresponding problem (3.3.10) is UCC(λ_C) as well. In addition, the system parameters (e.g., R_k, Q_k, A_k, B_k) of problem (3.3.10) are bounded above by the corresponding quantities under Assumption 3.2.1. As a result, problem (3.3.10) satisfies all the conditions of Proposition 3.2.11, which can be applied to give that

$$
\begin{aligned}
&\|\nabla_{h_{n_1'}} p_k^*(\theta_s)\|_2, \ \|\nabla_{h_{n_1'}} s_k^*(\theta_s)\|_2 \le Z_1 \rho^{k-n_1'}, \\
&\|\nabla_{d_{n_2'}} p_k^*(\theta_s)\|_2, \ \|\nabla_{d_{n_2'}} s_k^*(\theta_s)\|_2 \le Z_2 \rho^{n_2'-k}.
\end{aligned} \tag{3.3.12}
$$

Note that as given in Proposition 3.2.11, Z_1, Z_2 and ρ are independent of the problem interval, and the particular choice of the index set.

Assumptions 3.2.12, 3.3.2 and Propositions 3.2.13, 3.2.17 give that

$$
\|\hat{h}_{n_1'} - h_{n_1'}^0\|_2 \le C_g + u_0, \ \ \|\hat{d}_{n_2'} - d_{n_2'}\|_2 \le \frac{C_\phi}{2\lambda_Q} + C_g + m_0, \tag{3.3.13}
$$

where u_0 and m_0 are defined in Assumptions 3.2.12 and 3.3.2. Combining (3.3.12) and

(3.3.13), we have, for $k \in S_i$,

$$\begin{aligned}
& \left\| D_{\theta_1-\theta_0} \widetilde{p}_k^*(\theta_s) \right\|_2, \ \left\| D_{\theta_1-\theta_0} \widetilde{s}_k^*(\theta_s) \right\|_2 \\
\leq \ & Z_1 \left(C_g + u_0 \right) \rho^{k-n_1'} + Z_2 \left(\frac{C_\phi}{2\lambda_Q} + C_g + m_0 \right) \rho^{n_2'-k} \\
\leq \ & \left(Z_1 \left(C_g + u_0 \right) + Z_2 \left(\frac{C_\phi}{2\lambda_Q} + C_g + m_0 \right) \right) \rho^{\Omega}.
\end{aligned}$$

Letting $Y = Z_1 \left(C_g + u_0 \right) + Z_2 \left(\frac{C_\phi}{2\lambda_Q} + C_g + m_0 \right)$ and combining with (3.3.11) complete the proof. □

Proposition 3.3.9 is our key result. It proves the main hypothesis of this work that solutions restricted to the subinterval S_i of (3.2.1) formulated over the long horizon $[n_1, n_2]$ are exponentially close to the solutions restricted to the interval S_i of the problem $P^i_{\theta_0(i)}$, which is set up and solved only on the embedding region F_i. The exponent is proportional to Ω, the buffer size. Now we derive the following error bound of the optimal values on each decomposition subinterval.

Proposition 3.3.10. *Under Assumptions 3.2.1, 3.2.12, 3.2.14, 3.3.2 and 3.3.7, we have*

$$\left| J(S_i, [n_1, n_2], x^0_{n_1}) - J(S_i, F_i, h^0_{n_1'(i)}) \right| \leq \frac{n_2 - n_1}{n_0} X \rho^{\Omega}$$

for some $X > 0$ independent of n_1 and n_2.

Proof. Let $(w^*_{n_1':n_2'-1}, h^*_{n_1'+1:n_2'})$ be the solution of problem $P^i_{\theta_0(i)}$, and let $(u^*_{n_1':n_2'-1}, x^*_{n_1'+1:n_2'})$ be the solution of $P^i_{\theta_1(i)}$ which, by Proposition 3.3.3, is also the solution of problem (3.2.1) on F_i. Since the active set of problem $P^i_{\theta_0(i)}$ at optimality is UCC(λ_C) by Assumption 3.3.7, and the initial state is bounded by u_0 from Assumption

3.3.2, Lemma 3.2.13 gives that for $j \in S_i$, $\|h_j^*\|_2 \leq C_g$, $\|w_j^*\|_2 \leq C_u$. Then we have

$$\begin{aligned}
&\left|(x_j^* - d_j)^T Q_j (x_j^* - d_j) - (h_j^* - d_j)^T Q_j (h_j^* - d_j)\right| \\
&\leq \left|(x_j^* - d_j)^T Q_j (x_j^* - h_j^*)\right| + \left|(x_j^* - h_j^*)^T Q_j (h_j^* - d_j)\right| \\
&\leq 2C_Q(C_g + m_0)\|x_j^* - h_j^*\|_2
\end{aligned} \tag{3.3.14}$$

and

$$\begin{aligned}
&\left|u_j^{*T} R_j u_j^* - w_j^{*T} R_j w_j^*\right| \\
&\leq \left|(u_j^* - w_j^*)^T R_j u_j^*\right| + \left|w_j^{*T} R_j (u_j^* - w_j^*)\right| \\
&\leq 2C_R C_u \|u_j^* - w_j^*\|_2.
\end{aligned} \tag{3.3.15}$$

Combining with Proposition 3.3.9, we have that

$$\begin{aligned}
&\left|J(S_i, [n_1, n_2], x_{n_1}^0) - J(S_i, F_i, h_{n_1'(i)}^0)\right| \\
&\leq \sum_{j \in S_i} \left(\left|(x_j^* - d_j)^T Q_j (x_j^* - d_j) - (h_j^* - d_j)^T Q_j (h_j^* - d_j)\right| + \left|u_j^{*T} R_j u_j^* - w_j^{*T} R_j w_j^*\right| \right) \\
&\leq 2\frac{n_2 - n_1}{n_0} \left(C_Q(C_g + m_0) + C_R C_u\right) Y \rho^\Omega.
\end{aligned}$$

Denoting $X = 2\left(C_Q(C_g + m_0) + C_R C_u\right) Y$ completes the proof. □

Now we bound the total error of optimal values generated by the decomposition approach.

Theorem 3.3.11. *Under Assumptions 3.2.1, 3.2.12, 3.2.14, 3.3.2 and 3.3.7, we have that*

$$\left|J([n_1, n_2], [n_1, n_2], x_{n_1}^0) - \sum_{i=1}^{n_0} J(S_i, F_i, h_{n_1'(i)}^0)\right| \leq (n_2 - n_1) X \rho^\Omega,$$

where $X > 0$ is the same as in Proposition 3.3.10.

Proof. Using Proposition 3.3.10, we have that

$$\begin{aligned}
&\left| J([n_1,n_2],[n_1,n_2],x^0_{n_1}) - \sum_{i=1}^{n_0} J(S_i,F_i,h^0_{n'_1(i)}) \right| \\
&\leq \sum_{i=1}^{n_0} \left| J(S_i,[n_1,n_2],x^0_{n_1}) - J(S_i,F_i,h^0_{n'_1(i)}) \right| \\
&\leq n_0 \frac{n_2-n_1}{n_0} X\rho^{\Omega} \\
&= (n_2-n_1)X\rho^{\Omega}.
\end{aligned}$$

□

Theorem 3.3.11 upper bounds the total error induced by the decomposition approach by the product of an exponential term ρ^{Ω} and the length of the horizon $n_2 - n_1$. The rate of decay is eventually dominated by the exponential term. The exponential decay rate in the buffer size Ω enables the buffer regions to be chosen significantly shorter than the entire horizon while producing reasonable approximations under increasing horizon. Hence, when this approach is implemented in parallel, the computation time can be significantly reduced with little compromise of accuracy.

We also note that the techniques developed in Section 3.2, particularly Proposition 3.2.11, and in the first part of Section 3.3 can be used beyond proving our main result, Theorem 3.3.11. We believe they can be useful in other contexts, and in particular, in model predictive control. For example, it appears that one can show with similar techniques that the trajectory obtained from a receding horizon control approach converges exponentially to the solution of the full horizon problem (3.1.1). For instance, Proposition 3.3.9 can be applied to show that, if the short horizon problem has length Ω, then the first optimal control vector $u_{n'_1}$ and the second state vector $x_{n'_1+1}$ are exponentially close in Ω to the corresponding elements of the solution of the full horizon problem (3.1.1). Due to the space limit, we aim to develop this observation in future research.

3.4 Numerical results

In this section, we apply the temporal decomposition approach to a simplified production cost model in order to verify some of our theoretical findings. We employ the estimated hourly demand data in the northern Illinois region from year 2011 to 2015 provided by PJM Interconnection [62]. The model we are considering is the following:

$$\min \quad \textstyle\sum_{k=1}^{N} c_1(x_k - d_k)^2 + c_2 x_k^2 + u_k^2 \tag{3.4.1a}$$

$$\text{s.t.} \qquad x_{k+1} = x_k + u_k \tag{3.4.1b}$$

$$-U \leq u_k \leq U. \tag{3.4.1c}$$

Here d_k is the electricity demand to be satisfied on hour k, described by the data from [62]. We assume this can be done by two fictitious generators: one with high quadratic cost, with parameter $c_1 = 10$, and one with low quadratic cost $c_2 = 5$. The cheaper generator has limited ability to change its output x_k, which is modeled by the box constraints (3.4.1c) (also called the ramp rate constraints [11]) combined with the dynamics (3.4.1b). The more expensive generator is fast and can thus serve all remaining load $d_k - x_k$. This situation models, for example, the situation where one has a cheap but slow coal plant and a fast but expensive gas plant. Here the control is u_k, the amount of change at hour k of the generation level of the cheaper generator. We note that the formulation has the form from (3.1.1).

To define the temporal decomposition approach described in Section 3.3, we partition the hourly scale five-year horizon into weeks, resulting in $n_0 = 261$ subintervals $S_1, \ldots, S_{n_0}$. With buffer size Ω, we define the embedding regions $F_1, \ldots, F_{n_0}$ as in (3.3.2). In order to apply our decomposition approach, we need to specify the initial state for the short horizon problems other than the earliest one, which uses the initial value of the full horizon problem. In general, finding a good initial state is difficult. In the case of production cost models that motivated this research, however, the demand, while random, is fairly stable [11] for the same time of the day in the week. Moreover, optimal generation levels tend to be stable too with a similar pattern [11]. Therefore for production cost models, a good initial guess is readily available. For the general dynamic optimization problem, such a good guess may not be available. On the other hand, for a given policy of choosing it – for example, choosing the analytical center of the feasible set – Proposition 3.3.9 can be used on test problems to determine a good choice of the buffer size Ω that allows for the effect of the initial condition policy to be small enough for the tolerance sought. The fact that Proposition 3.3.9 establishes exponential decay of the error with respect to the buffer size Ω allows such trade-offs to be carried out. For our production cost model example, as the demand pattern is relatively predictable as also indicated in [11], a good guess does exist at a given time of the day and week. As a consequence, we use as initial state the average demand at each hour of a day for all of 2011. At the initial time point $n_1'(i)$ of F_i, we thus set the initial value $x^0_{n_1'(i)}$ to the average demand for that hour. Denote the optimal objective function value of problem (3.4.1) as J^*, namely,

$$J^* \triangleq J([1, N], [1, N], x^0_1)$$

as defined in (3.3.4), and denote

$$J_i^* \triangleq J(S_i, F_i, x^0_{n_1'(i)})$$

for $i = 1, \ldots, n_0$. We solve both long and short horizon versions of (3.4.1) using the Ipopt software [21]. The model was defined by using the Julia/JuMP interface [83].

We now analyze how well the sum of the computation on the short intervals, J_i^*, approximates the long horizon problem, J^*. Figure 3.2 shows the relative approximation error $\left|J^* - \sum_{i=1}^{n_0} J_i^*\right| / J^*$ as the function of the buffer size (measured in hours) for an increasing value of the ramping constraint bound U in (3.4.1c). For each U, the largest buffer size we experiment with is the smallest value that results in a relative approximation error less than

10^{-5}. The cost of such large-scale planning projects is usually on the order of billions of dollars. A relative error on the order of 10^{-5} corresponds to a discrepancy of less than a hundred thousand dollars, which is already well within the tolerance level of planning. We observe from Figure 3.2 that, for all values of U the relative error decreases exponentially with Ω, which is the conclusion of our main result, Theorem 3.3.11. We conclude that Figure 3.2 validates the exponential decay of the approximation error of temporal decomposition with respect to the buffer size as proved in Theorem 3.3.11. We note that the target accuracy is achieved by buffer regions of less than 24 hours for all bounds U, although for different and larger PCM the results could be different. The order of magnitude of the buffer for which such accuracy is achieved is, however, of the same order – days – as in [11].

U	100	200	300	400	500	600	700	800	900	1000
t	91	62	48	34	14	12	10	8	7	5

Table 3.1: Longest period t (hour) for which the optimal controls of problem (3.4.1) achieve the bound.

Figure 3.2 also shows that the decay rate of the approximation error increases with increasing bound U of controls. Note that the error bound in Theorem 3.3.11 depends on the controllability of problem (3.4.1) at optimality. We thus investigate numerically the longest period t for which the problem (3.4.1) is not controllable; t as used here carries the same meaning as in Definition 3.2.2. Since problem (3.4.1) is one dimensional, from Definition 3.2.2, it follows that t is simply the longest contiguous period for which the optimal controls are on the bound. Table 3.1 shows t in hours for different choices of the bound U. The longest period of uncontrollability decreases as the bound becomes larger. Figure 3.3 shows the proportions of optimal controls of problem (3.4.1) that are on the bound for increasing U. When $U = 100$, more than 85 % of the optimal controls attain the bound, which approaches the controllability limit of the problem. Even for this tightest bound $U = 100$, however the value of t is 91, which is about three to four days. This bound is certainly covered by our weekly partitioned subintervals and thus ensures that the controllability Assumption 3.2.14

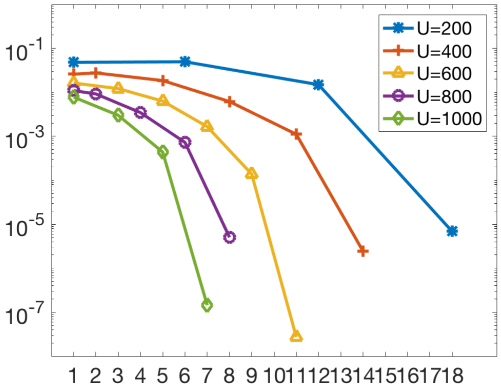

Figure 3.2: Relative error in approximation $\left|J^* - \sum_{i=1}^{n_0} J_i^*\right| / J^*$ at each buffer size (hour) for $U = 200, 400, 600, 800, 1000$.

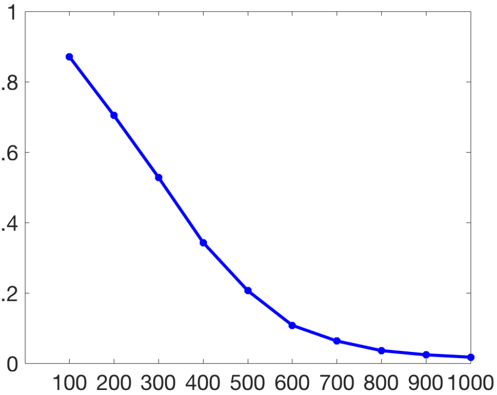

Figure 3.3: Proportions of optimal controls of problem (3.4.1) that are on the bound for $U = 100, \ldots, 1000$.

holds on each embedding region F_i. Therefore the conditions of our main result Theorem 3.3.11 are satisfied.

3.5 Conclusions

Temporal decompositions are useful techniques for exposing parallelism in dynamic optimization problems. Such approaches are particularly useful for production cost simulations in electricity planning problems, where the calculations can have hundreds of thousands of time periods. The version of temporal decomposition discussed in this work approximates the solution over the entire horizon by the one obtained by patching the solution from multiple dynamic optimization problems with much shorter, overlapping, horizons initialized at some guess of the state. In turn, this transforms a sequential problem into one that is immediately amenable to parallel computing, thus massively reducing the time to solution. While used to great effect in [11], such temporal decomposition approaches were, up to our work, a heuristic with no theoretical basis for its good approximating behavior.

In this work we prove that for the class of linear-quadratic dynamic optimization problems the temporal decomposition with overlap approaches the solution of the original problem exponentially fast in the size of the overlap. This approach partitions the entire horizon into subintervals and embeds each subinterval into the interval of interest plus a buffer region. The objective cost and, respectively, the solution on the entire horizon are then approximated by the the sum of the costs on each subinterval and, respectively, the solutions obtained from solving the corresponding problem on its embedding region. We prove that under some boundedness and controllability assumptions, the approximation error in both the solution and objective function decreases exponentially as a function of the buffer size. The exponential decay rate enables one to choose embedding regions much shorter than the length of the horizon; and since problems on each buffer region can be solved independently the time to solution is significantly reduced when the approach is implemented in parallel.

We validate our theoretical findings by using a numerical experiment that mimics a production cost evaluation over a five-year interval with hourly time periods and real data but with a simple, two-generator model. For all the cases, the relative error in the approximation of the objective function decreases exponentially with the buffer size. The decay rate decreases as more optimal controls attain the bound. In other words, the decay is slower when the system stays uncontrollable for longer periods. For this small experiment, even with the tightest bound on the controls and more than 85% of the optimal controls attaining the bounds, the buffer size needed for the relative error to be less than 10^{-5} is less than 24 periods – one day. Since the decomposed horizons have length only slightly more than one week, little extra effort has been added to solving problems when compared with the problem for the useful interval only, the one-week inner temporal region.

The class of dynamic optimization problems considered here is simplified when compared with [11] in that it is a linear-quadratic dynamic optimization problem with box control constraints. While our problem class does not include the complicating features such as combinations of linear objective, integer variables, and path constraints, it includes the intertemporal constraints that make the analysis of error difficult. Consequently, our approach gives analytical support for the rapid convergence of the temporal decomposition with overlapping intervals. Future work will address extending the results for these complicating features as well as applying the techniques of this work to model predictive control.

CHAPTER 4

MAXIMUM LIKELIHOOD ESTIMATION FOR A SMOOTH GAUSSIAN RANDOM FIELD MODEL

4.1 Introduction

Computer experiments have been used extensively in investigating complex scientific phenomena. The responses of many computer experiments are deterministic, in the sense that rerunning the same code with the same inputs will give identical outputs. Often, each run of the code is computationally expensive, so a common alternative to running the code at all input values of interest is to run the code at some inputs and make cheaper predictions at others. [104] and [105] propose to model the deterministic computer experiment outputs as a realization of a Gaussian random field with covariance

$$\mathrm{Cov}\left(f(x), f(y)\right) = \theta_0 \prod_{u=1}^{d} e^{-\frac{|x_u - y_u|^\gamma}{\theta_u}}, \tag{4.1.1}$$

where x_u, $y_u \in [0,1]$, $u = 1, \ldots, d$, $\theta_0 > 0$ is the scale parameter and $\theta_u > 0$ are range parameters. The use of stochastic models provides a statistical basis for experimental design, parameter estimation, interpolation and uncertainty calibration.

When $\gamma = 2$, the Gaussian process with covariance function (4.1.1) is infinitely mean square differentiable and thus is an attractive choice when the output surface is known to be smooth [48, 95, 96, 109, 113]. This covariance function is sometimes called "Gaussian" because of its functional form, but we prefer the name "squared exponential" to avoid confusion with a Gaussian process. In fact, smooth test functions composed of elementary functions (e.g. polynomials, trigonometric functions and exponential functions) are often used as test cases for studying the effectiveness of Gaussian processes in modeling computer experiments. However, little is known about properties of maximum likelihood estimators (MLEs) when observations are generated by these test functions. In this article, we are interested in the asymptotic properties of the MLE when more and more observations are taken on a fixed domain (fixed domain asymptotics, see [112]) for the Gaussian process when the computer model is some simple deterministic function. We aim to understand the implications of modeling smooth deterministic functions using the squared exponential covariance function.

We consider a mean zero Gaussian random field with covariance function (4.1.1) and $\gamma = 2$. For $d = 1$, we prove some asymptotic properties of the MLE for the scale parameter θ_0 when the range parameter θ_1 is fixed and the computer experiment response is a pth order monomial $f(x) = x^p$. We consider two situations for the observations. In the first case, observations $\mathbf{z}$ are taken on a regular grid on $[0, 1]$ so that $\mathbf{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \ldots, f(1))^T$. In the second case, the observations are successive derivatives of the response function at zero, namely, $\mathbf{z} = (f(0), f^{(1)}(0), \ldots, f^{(n-1)}(0))^T$. Automatic differentiation (AD) techniques [53] can be used to obtain derivatives of computer model output and there are certain problems for which higher order derivatives are needed [29, 52, 117]. Therefore considering what happens when one observes successive derivatives at a single location may be of some practical interest.

The rest of the chapter is organized as follows. Section 2 deals with regularly spaced observations on $[0, 1]$. The key finding is that the asymptotic order of the MLE $\hat{\theta}_0$ is $n^{-1/2}$ when $p = 0$ and at least $n^{1/2}$ when $p = 1$. In particular, $\hat{\theta}_0 \to 0$ when $p = 0$ and $\hat{\theta}_0 \to \infty$ when $p = 1$. Section 3 deals with the case where observations are derivatives at zero. An exact expression for the inverse Cholesky factor for the correlation matrix is obtained. For all $p \geq 0$, we prove that $\lim_{n\to\infty} n^{1/2-p}\hat{\theta}_0$ exists and is positive so that $\hat{\theta}_0 \to 0$ for $p = 0$ and $\hat{\theta}_0 \to \infty$ for all $p \geq 1$. Section 4 demonstrates the theoretical findings in Section 2 and 3, and explores numerically three commonly used two-dimensional test functions in the computer experiments literature. For estimating the scale parameter θ_0 and the two range parameters θ_1 and θ_2, we compare maximum likelihood method with leave-one-out cross validation in a prediction problem. We find that the likelihood method and cross validation perform differently for different test functions in terms of magnitude and calibration of prediction errors. We also explore MLE of the range parameter for pth order monomials when treating both scale and range parameters as unknown, and investigate its implications for practical test functions. In the numerical experiments, to deal with numerical singularity of the correlation matrix, we choose parameters such that the correlation matrix and observations are both rational, and do symbolic computation with *Mathematica* [119] to obtain exact results. Since a common approach to overcome the near singularity is to include a nugget effect, we investigate the effect of adding a nugget on the likelihood and prediction. We found that the likelihood generally decreases substantially with even a very small nugget, but prediction error can sometimes decrease a bit at first as the nugget size increases. All proofs of the theoretical results are presented in the Appendix.

4.2 Regularly spaced observations

In this section, we consider the observations are outputs of the model function $f(x) = x^p$ regularly spaced on $[0, 1]$. Fixing the range parameter θ_1, we prove that $\hat{\theta}_0 \to 0$ when the model function is constant and $\hat{\theta}_0 \to \infty$ when it is linear. Though some intermediate steps apply to all $p \geq 0$, we are only able to derive the asymptotic order and a lower bound on the asymptotic order for $p = 0$ and $p = 1$ respectively.

The observations are outputs of the model function $f(x) = x^p$ taken on a regular grid on $[0, 1]$ so that $\boldsymbol{z} = (\left(\frac{1}{n}\right)^p, \left(\frac{2}{n}\right)^p, \ldots, 1)^T$. The covariance matrix can be written as

$$\Sigma(\theta_0, \theta_1, n) = \theta_0 R(\theta_1, n), \tag{4.2.1}$$

where the (i, j)th element of $R(\theta_1, n)$ is

$$R(\theta_1, n)_{ij} = w^{(i-j)^2}, \qquad w = e^{-1/(\theta_1 n^2)}. \tag{4.2.2}$$

[81] gives the exact form of the inverse of the Cholesky factor for $R(\theta_1, n)$. Letting $R(\theta_1, n) = LL^T$, where L is the lower triangular Cholesky factor with positive diagonal elements, then

$$(L^{-1})_{ij} = \begin{cases} \frac{(-w)^{i-j}\left[{i-1 \atop j-1}\right]_{w^2}}{\prod_{k=1}^{i-1}(1-w^{2k})^{1/2}}, & i \geq j \\ 0, & i < j \end{cases} \tag{4.2.3}$$

where $\left[{k \atop m}\right]_q$ is the q-binomial coefficient defined by

$$\begin{bmatrix} k \\ m \end{bmatrix}_q = \frac{(1-q^{k-m+1})(1-q^{k-m+2})\dots(1-q^k)}{(1-q)(1-q^2)\dots(1-q^m)}$$

if $0 \leq m \leq k$ and 0 otherwise.

The log-likelihood function of θ_0 is

$$2l(\theta_0) = -n\log 2\pi - n\log\theta_0 - \log|R(\theta_1, n)| - \frac{1}{\theta_0}\boldsymbol{z}^T R(\theta_1, n)^{-1}\boldsymbol{z}$$

and the MLE of θ_0 is

$$\hat{\theta}_0 = \frac{1}{n}\boldsymbol{z}^T R(\theta_1, n)^{-1}\boldsymbol{z}.$$

With the form of L^{-1} in (4.2.3), the exact form of $\hat{\theta}_0$ can be written as

$$\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\sum_{j=1}^{i}(-w)^{i-j}\left[{i-1 \atop j-1}\right]_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})}.$$

For convenience we make the following notation for the rest of this article:

$$a_{ip}(w) := \frac{\left(\sum_{j=1}^{i}(-w)^{i-j}\left[{i-1 \atop j-1}\right]_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})} \tag{4.2.4}$$

where $p \geq 0$ and $i \geq 1$. Note that $\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} a_{ip}(w)$ for $w = e^{-1/(\theta_1 n^2)}$.

By considering the limit of the summand of $\hat{\theta}_0$, we obtain the following proposition.

Proposition 4.2.1. *Denote*

$$l_{ip} := \lim_{n\to\infty} a_{ip}(w) = \lim_{n\to\infty} \frac{\left(\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})}$$

then

$$l_{ip} = \begin{cases} \frac{(i-1)!\theta_1^p}{2^{i-1}\left(\frac{i-p-1}{2}!\right)^2}, & i-p \qquad odd \\ \\ 0, & i-p \qquad even \end{cases} \tag{4.2.5}$$

where $w = e^{-1/(\theta_1 n^2)}$, $p \geq 0$ *and* $i > p$.

Proof. See Section C.1.1. □

Lemma 4.2.2. $\frac{1}{n}\sum_{i=p+1}^{n} l_{ip} \sim \frac{n^{p-\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})}$ *as* $n \to \infty$.

Proof. See Section C.1.2. □

The previous results deal with the general case where $p \geq 0$. The following results concentrate on $p = 0$ and $p = 1$. We now derive the asymptotic order for $\hat{\theta}_0$ when $p = 0$ and a lower bound on the asymptotic order for $\hat{\theta}_0$ when $p = 1$.

Theorem 4.2.3. *If* $p = 0$, $\hat{\theta}_0 \sim \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}$ *as* $n \to \infty$.

Proof. See Section C.1.3. □

Theorem 4.2.4. *If* $p = 1$,

$$\liminf_{n\to\infty} \frac{\hat{\theta}_0}{\sqrt{n}} \geq \frac{\theta_1}{3\sqrt{2\pi}}.$$

Proof. See Section C.1.4. □

The difficulty of the proof lies in the fact that the dimension and elements of the correlation matrix $R(\theta_1, n)$ change with n. In particular, we can not simply apply an elementwise limit theorem to $\hat{\theta}_0 = \boldsymbol{z}^T R(\theta_1, n)^{-1} \boldsymbol{z}/n$. Proposition 4.2.1 proves the limits of $a_{ip}(w)$ for fixed i as $n \to \infty$ and Lemma 4.2.2 proves the asymptotic order of the average of those limits. However, to derive the asymptotic order of $\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} a_{ip}(w)$ we would need to have some results on the uniform convergence of $a_{ip}(w)$ for $1 \leq i \leq n$, which we have been unable to obtain.

We now state a conjecture about the asymptotic order of $\hat{\theta}_0$ for general $p \geq 0$. The case for $p = 0$ is proved in Theorem 4.2.3 with $C(0) = \sqrt{2/\pi}$, and Theorem 4.2.4 is a weaker version of the conjecture when $p = 1$.

Conjecture For all $p \geq 0$, $\lim_{n\to\infty} n^{1/2-p}\hat{\theta}_0 = C(p)$ where $C(p) = \theta_1^p/\sqrt{2\pi}2^p(p+1/2)$.

It might also be of interest to consider functions that are continuous but have some form of singularities. These functions are not smooth and hence not the main focus of our work. However, we present one example here to illustrate what can happen.

Proposition 4.2.5. *If*

$$f(x) = \begin{cases} 0, & x \leq 1/2 \\ g(x-1/2), & x > 1/2 \end{cases}$$

for some continuous function $g(x)$ satisfying $g(0) = 0$ and $c := \lim_{x\to 0} \frac{g(x)}{x^p} > 0$ for some $p \geq 1$, then

$$\liminf_{n\to\infty} \frac{\hat{\theta}_0}{n} > 0$$

as $n \to \infty$. In particular, $\hat{\theta}_0 \to \infty$ as $n \to \infty$.

Proof. See Section C.1.5. □

We recognize that fixing the range parameter as we have done here is rather artificial, but the mathematical difficulties of analyzing even this problem are formidable and we believe that the resulting asymptotic theory is interesting and informative despite its limitations.

As shown in [112, pp. 120-121], two non-identical squared exponential covariance functions for a Gaussian process on a finite interval correspond to orthogonal measures, suggesting that, unlike the case for Matérn covariance functions [122], it might be possible to estimate both the scale and range parameters consistently based on fixed domain asymptotics if in fact the Gaussian process model is correct (see for example [3]). In the present setting when the process is just a simple deterministic function, it is not at all clear what should happen, so we investigate the properties of joint estimates of scale and range parameters through numerical experiments in Section 4.

We do not have an intuitive explanation for the quantitative aspects of our asymptotic results, even for $p = 0$. Comparing to two settings for which asymptotic calculations can be easily done provides us with a clue to the qualitative behavior of $\hat{\theta}_0$ as p increases. If $\Sigma = \theta_0 I_n$, where I_n is the $n \times n$ identity matrix, and f is a continuous function on $[0, 1]$, then $\hat{\theta}_0 \sim \int_0^1 f(x)^2\, dx$ as $n \to \infty$, so $\hat{\theta}_0$ tends to a nonzero constant for any nontrivial f. For the exponential covariance function ($\gamma = 1$ and $d = 1$ in (4.1.1)), if f has a bounded second derivative on $[0, 1]$ then

$$\hat{\theta}_0 \sim \frac{1}{n} f(0)^2 + \frac{1}{2\theta_1 n} \int_0^1 \{f(x) + \theta_1 f'(x)\}^2 dx \tag{4.2.6}$$

as $n \to \infty$ (see Section C.1.6), so that $n\hat{\theta}_0$ tends to a positive finite constant as $n \to \infty$ when $f(x) = x^p$ for any nonnegative integer p. These results are in stark contrast to what we have proven and conjectured here for the squared exponential covariance function, that $n^{1/2-p}\hat{\theta}_0$ tends to a positive, finite constant. Thus, there must be something about the squared exponential model that makes us think θ_0, the variance of the process, is large when p is large. A possible intuitive explanation for this result is that if we think the underlying function is very smooth (which is the case when we use the squared exponential model) and we observe that the function just happens to to equal x^p at n densely spaced points, then we will conclude that this function must at least very nearly equal x^p over some broad interval, so that the larger p is, the more we think the function varies over this broad interval and the larger we think θ_0 is.

4.3 Derivatives at zero

We obtain the asymptotic order of $\hat{\theta}_0$ when the observations are the first $n-1$ derivatives at 0 for the response function $f(x) = x^p$, $p \geq 0$. Denote $\boldsymbol{z} = (f(0), f'(0), \ldots, f^{(n-1)}(0))^T$, and the covariance matrix of the observations $\boldsymbol{z}$ as $\Sigma_1(\theta_0, \theta_1, n)$.

Now we introduce a notation that is frequently used in the rest of this section.

Definition 4.3.1. *For a positive integer m, define the double factorial of m as*

$$m!! = \begin{cases} \prod_{k=1}^{m/2}(2k) = m(m-2)\ldots 2, & m \quad \textit{even} \\ \prod_{k=1}^{(m+1)/2}(2k-1) = m(m-2)\ldots 1, & m \quad \textit{odd} \end{cases}$$

and set $0!! = 1$. This double factorial notation is commonly used in combinatorics [50]. Note that $m!!$ is the product of all positive integers no larger than and having the same parity as m, which is different from the successive factorial $(m!)!$.

As before, f is modeled as a stationary Gaussian random field with covariance function $K(u) = \theta_0 e^{-u^2/\theta_1}$. Then the (i,j)th element of $\Sigma_1(\theta_0, \theta_1, n)$ can be computed as

$$\begin{aligned} \Sigma_1(\theta_0, \theta_1, n)_{ij} &= \frac{\partial^{i+j-2}}{\partial x^{i-1} \partial y^{j-1}} K(x-y)\Big|_{x=y=0} \\ &= (-1)^{j-1}\theta_0 \frac{d^{i+j-2}}{du^{i+j-2}} e^{-u^2/\theta_1}\Big|_{u=0} \\ &= \theta_0 \theta_1^{-\frac{i+j-2}{2}} (-1)^{i-1} H_{i+j-2} \end{aligned}$$

where

$$H_m = \begin{cases} 0, & m \text{ odd} \\ (-2)^{\frac{m}{2}}(m-1)!!, & m \text{ even} \end{cases}$$

is mth order Hermite polynomial at 0. The mth order Hermite polynomial is defined as

$$H_m(x) = (-1)^m e^{x^2} \frac{d^m}{dx^m} e^{-x^2}.$$

Defining $R_1(\theta_1, n)$ so that $\Sigma_1(\theta_0, \theta_1, n) = \theta_0 R_1(\theta_1, n)$, the MLE of θ_0 is $\hat{\theta}_0 = \frac{1}{n}\boldsymbol{z}^T R_1(\theta_1, n)^{-1}\boldsymbol{z}$.

The following proposition gives an exact form of the reverse Cholesky factorization [77] of $R_1(\theta_1, n)^{-1}$.

Proposition 4.3.2. *Let $D(\theta_1, n)$ be the lower triangular matrix with positive diagonal elements such that $R_1(\theta_1, n)^{-1} = D(\theta_1, n)^T D(\theta_1, n)$. Then for all $1 \leq i, j \leq n$, the (i, j)th element $D(\theta_1, n)_{ij} = d_{ij}$, where*

$$d_{ij} = \begin{cases} \dfrac{\theta_1^{\frac{j-1}{2}} \sqrt{(i-1)!}}{2^{\frac{j-1}{2}}(j-1)!(i-j)!!}, & \textit{if} \quad i \geq j, \quad \textit{and} \quad i+j \quad \textit{is} \quad \textit{even} \\ 0, & \textit{otherwise.} \end{cases} \tag{4.3.1}$$

Proof. See Section C.1.7. □

Note that d_{ij} depends only on i and j but not n. So the matrices $D(\theta_1, n)$ are nested as n increases. In fact in this case $R_1(\theta_1, n)$ is nested and in general, the reverse Cholesky factors of inverses of a sequence of nested matrices are nested (see Section C.1.8). This feature simplifies the proof of the asymptotic order of $\hat{\theta}_0$ for all $p \geq 0$ and is not shared by the setting considered in Section 2.

Theorem 4.3.3. *Suppose $f(x) = x^p$, $p \geq 0$, then*

$$\hat{\theta}_0 \sim \frac{n^{p-\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})}$$

as $n \to \infty$.

Proof. See Section C.1.9. □

Automatic differentiation (AD) can be used to obtain derivatives of functions coded as computer programs. AD exploits the fact that function evaluation can be broken down into elementary operations (e.g., addition, multiplication, $\exp(\cdot)$) and applies the chain rule. A comprehensive reference for AD can be found in [53]. Although in practice, only lower order derivatives are usually found, there are efforts to obtain higher order Taylor coefficients in one direction for certain problems [29, 52, 117].

For smooth functions, estimates based on multiple derivatives at zero should be closely related to estimates based on observing more frequently on a fixed domain, since more observations enable the calculation of higher order finite differences that approximate derivatives as the spacing gets small. More specifically, consider observing at k inputs $\boldsymbol{z}_k = (\left(\frac{1}{n}\right)^p, \ldots, \left(\frac{k}{n}\right)^p)^T$ for some $k \leq n$. For this case, MLE $\hat{\theta}_0(k) = \frac{1}{k}\sum_{i=1}^{k} a_{ip}(w)$ where $a_{ip}(w)$ is defined in (4.2.4) and $w = e^{-1/(\theta_1 n^2)}$, is the same as the MLE $\hat{\theta}_0'(k)$ when observing the first k finite differences $\boldsymbol{z}_k'$ at zero, since these first k finite differences at zero are a linear transformation of the first k observations, namely, $\boldsymbol{z}_k' = C\boldsymbol{z}_k$ for some $C \in \mathbb{R}^{k\times k}$. It follows that $\hat{\theta}_0'(k) = \frac{1}{n}\boldsymbol{z}_k'^T \left(CRC^T\right)^{-1} \boldsymbol{z}_k' = \frac{1}{n}\boldsymbol{z}_k R^{-1}\boldsymbol{z}_k = \hat{\theta}_0(k)$. Fixing k and letting $n \to \infty$ gives that finite differences converge to derivatives and $\hat{\theta}_0(k) \to \frac{1}{k}\sum_{i=1}^{k} l_{ip}$. The asymptotic order of the limit $\frac{1}{k}\sum_{i=1}^{k} l_{ip}$ as $k \to \infty$ is given by Lemma 4.2.2 and is exactly the same as what is obtained in Theorem 4.3.3, which is the asymptotic order of MLE when observations are derivatives at zero. However, this heuristic argument does not directly imply $\hat{\theta}_0$ has the same asymptotics for the two situations of observations. In particular, taking $k \to \infty$ at the same time as $n \to \infty$ is a different and harder problem.

4.4 Numerical results

In this section, first we illustrate our theoretical findings in Sections 2 and 3 numerically. Then we compare maximum likelihood estimators (MLE) with leave-one-out cross validation (CV) in a prediction problem for two commonly used test functions. We also show that for the Branin function, the MLE for the range parameter along one coordinate does not appear to exist. The parameters in the numerical experiments are chosen to make both the correlation matrix and the observations rational so that the matrix calculations can be done exactly with symbolic computations. A common approach to overcome the near singularity of the correlation matrix is to include a small nugget effect in the hope of improving conditioning at the same time introducing minimal modification to the matrix. We also investigate the effect of this approach on the likelihood and prediction.

4.4.1 Asymptotic behavior of MLE for scale parameter

As in Sections 2 and 3, we consider the test function as a pth order monomial $f(x) = x^p$ and two situations for the observations. In the first case, the observations $\boldsymbol{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \ldots, f(1))^T$ are taken on a regular grid on $[0,1]$, and in the second case,the observations $\boldsymbol{z'} = (f(0), f^{(1)}(0), \ldots, f^{(n-1)}(0))^T$ are the first $n-1$ derivatives of the test function at zero. Denote the MLE of the scale parameter for the two situations of observations as $\hat{\theta}_0$ and $\hat{\theta}'_0$ respectively. We consider $n = 2^k$, $k = 3, \ldots, 9$, and the range parameter $\theta_1 \approx 0.95$ is chosen to make the correlation matrix rational for all choices of n so that exact computations can be done. Note that exact computation is needed here to prevent numerical overflow even if the exact form of the Cholesky factor (4.2.3) is used.

Theorems 4.2.3 and 4.2.4 state that

$$\lim_{n\to\infty} \hat{\theta}_0 n^{1/2-p} = \sqrt{2/\pi},\ p = 0;\ \liminf_{n\to\infty} \hat{\theta}_0 n^{1/2-p} \geq \theta_1/3\sqrt{2\pi},\ p = 1.$$

The Conjecture in Section 2 states that $\hat{\theta}_0 n^{1/2-p}$ converges to some limit $C(p)$ as $n \to \infty$ for all $p \geq 0$. Figure 4.1 shows $\hat{\theta}_0$ for increasing n when $p = 0, 1, 2, 3$ on log scale. Theorem 4.2.3 and the Conjecture imply that $(\log n, \log \hat{\theta}_0)$ will be close to the reference line $y = (p-1/2)x + \log C(p)$ for n large. The numerical results show clear agreement with the theoretical results in Theorem 4.2.3 and the Conjecture. When observations are the first

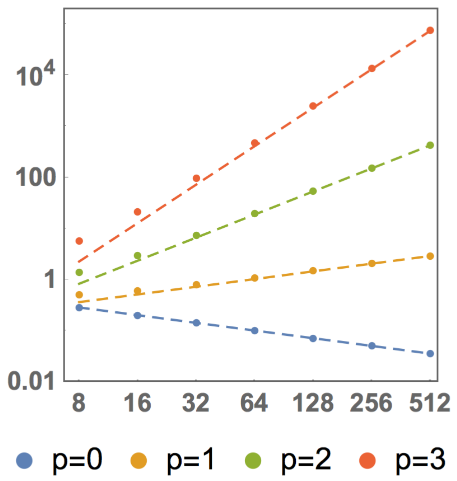

Figure 4.1: $\hat{\theta}_0$ when $n = 2^k$, $k = 3, \dots, 9$ for $\boldsymbol{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \dots, f(1))^T$. The slopes of the reference lines are the asymptotic orders $p - 1/2$ when $p = 0, 1, 2, 3$. Both axes are on log scale.

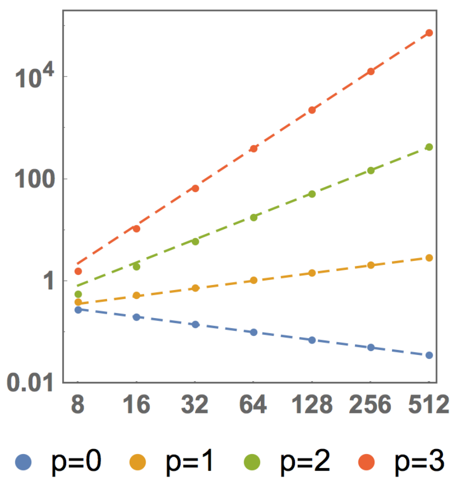

Figure 4.2: $\hat{\theta}'_0$ when $n = 2^k$, $k = 3, \dots, 9$ for $\boldsymbol{z}' = (f(0), f^{(1)}(0), \dots, f^{(n-1)}(0))^T$. The slopes of the reference lines are the asymptotic orders $p-1/2$ when $p = 0, 1, 2, 3$. Both axes are on log scale.

$n - 1$ derivatives at zero, Theorem 4.3.3 states that

$$\lim_{n \to \infty} \hat{\theta}'_0 n^{1/2-p} = \theta_1^p / \sqrt{2\pi} 2^p (p + 1/2) = C(p), \qquad p \geq 0.$$

Figure 4.2 shows $\hat{\theta}'_0$ for increasing n when $p = 0, 1, 2, 3$ on log scale with the same reference lines as those in Figure 4.1. For all four cases shown here, the agreement between the numerical and asymptotic results is good, even for $n = 8$.

4.4.2 Comparing MLE and CV in a prediction problem

We consider the first two functions on a 23×23 regular grid on $[0, 1] \times [0, 1]$ and let $\delta = 1/23$ be the spacing between neighboring points. The observations are taken on a 12×12 regular sub-grid, and the remaining 385 points are predictands. An illustration of the setup is shown in Figure 4.4. Observations are taken at every other location along each dimension to facilitate the use of the inverse Cholesky factor (4.2.3), so that exact computations can be done with rational correlations. The first test function we experiment with is a mixture of Gaussians [45, 57],

$$f(x_1, x_2) = c_1 e^{-s_1\left((x_1/\delta-\mu_1)^2+(x_2/\delta-\mu_2)^2\right)} + c_2 e^{-s_2\left((x_1/\delta-\widetilde{\mu}_1)^2+(x_2/\delta-\widetilde{\mu}_2)^2\right)}. \qquad (4.4.1)$$

We choose $e^{-s_1} = 399/400$ and $e^{-s_2} = 99/100$ to be rationals and $\mu_1 = \mu_2 = 8$, $\widetilde{\mu}_1 = \widetilde{\mu}_2 = 17$ to be integers so that all the observations are rational. We set $c_1 = 1$ and $c_2 = -1/2$ so that the function consists of a peak and a small dip. The second function we consider is a product of trigonometric and exponential functions [30],

$$f(x_1, x_2) = \cos(c_1 x_1 + c_2 x_2) e^{c_0 x_1 x_2}. \qquad (4.4.2)$$

We choose $c_1 = c_2$ such that $\cos(c_1\delta) = \cos(c_2\delta) = 24/25$. With this choice, $\sin(c_1\delta) = \sin(c_2\delta) = 7/25$ and $\cos(c_1 x_1 + c_2 x_2)$ is rational for all grid points (x_1, x_2) by trigonometric identities. c_0 is chosen to satisfy $e^{c_0\delta^2} = 500/499$ and with this choice $c_0 \approx 1.06$, which approximates $c_0 = 1$ used in [30]. Both the mixture of Gaussians (4.4.1) and trig-exponential function (4.4.2) are symmetric about the diagonal. The third test function we consider is the Branin function [23, 99] on a 27×27 regular grid on $[-5, 10] \times [0, 15]$,

$$f(x_1, x_2) = e(x_2 - f x_1^2 + g x_1 - r)^2 + s(1 - t)\cos(c_0 x_1) + s. \qquad (4.4.3)$$

The spacing between neighboring points is $\delta_b = 5/9$. We choose parameters as $e = 1$, $f = 5/36$, $g = 5/3$, $r = 6$, $s = 10$, $t = 1/24$ and $\cos(c_0\delta_b) = 4/5$. The observations are taken on the 14×14 regular sub-grid. The different setting of the grid for the Branin function is to make the observations rational at all grid points. The three test functions with the aforementioned parameters are shown in Figure 4.3. The parameters for the three test functions are rounded from the commonly used values to ensure rationality. For example, the recommended parameter values for the Branin function that are different from our choices are $f = 5.1/4\pi^2$, $g = 5/\pi$, $t = 1/8\pi$ [44, 99].

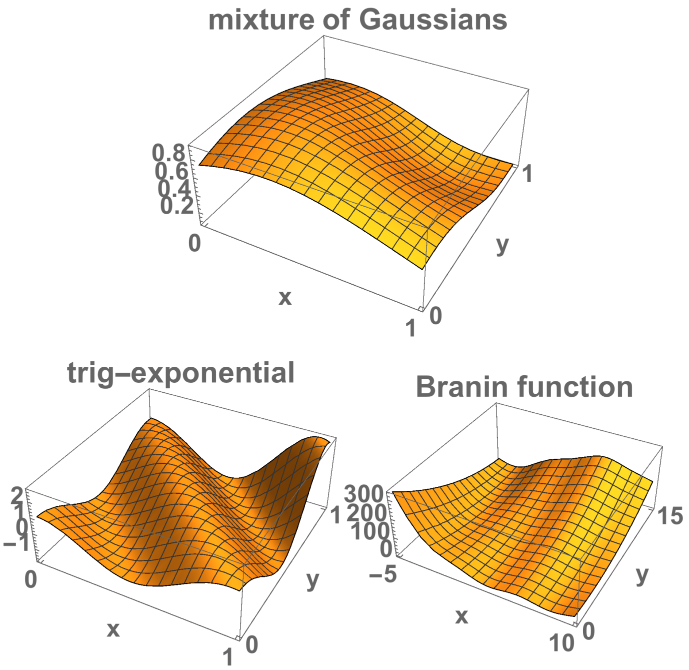

Figure 4.3: Surface for test functions (4.4.1), (4.4.2) and (4.4.3).

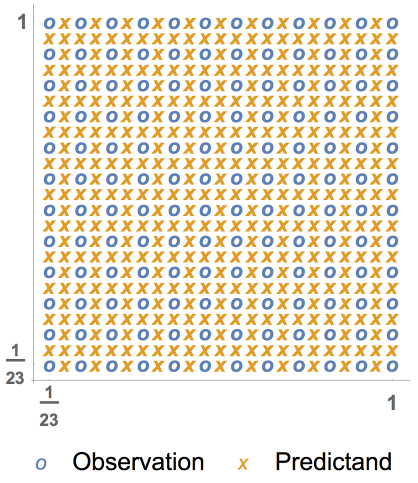

Figure 4.4: Locations of observations and predictands.

4.4.2.1 Exact computation results

We compare MLE and CV for mixture of Gaussians (4.4.1) and the trig-exponential function (4.4.2) in predicting at the 385 points not used to fit the model. Denote the observations as $\mathbf{z}$ and the log-likelihood function as $\mathcal{L}(\theta_0, \theta_1, \theta_2)$. The profile log-likelihood function of (θ_1, θ_2) is defined as

$$l_n(\theta_1, \theta_2) = \mathcal{L}\left(\hat{\theta}_0(\theta_1, \theta_2), \theta_1, \theta_2\right),$$

where $\hat{\theta}_0(\theta_1, \theta_2) = \text{argmax}_{\theta_0} \mathcal{L}(\theta_0, \theta_1, \theta_2)$. The profile log-likelihood function satisfies

$$\begin{aligned} 2l_n(\theta_1, \theta_2) = & -n \log 2\pi - n \log \hat{\theta}_0 - \log |R(\theta_1, m) \otimes R(\theta_2, m)| \\ & - \frac{1}{\hat{\theta}_0} \boldsymbol{z}^T \left(R(\theta_1, m) \otimes R(\theta_2, m)\right)^{-1} \boldsymbol{z} \end{aligned} \tag{4.4.4}$$

where $n = m^2$, $m = 12$, R is as defined in (4.2.2), and the MLE for the scale parameter is

$$\hat{\theta}_0(\theta_1, \theta_2) = \frac{1}{n} \boldsymbol{z}^T \left(R(\theta_1, m) \otimes R(\theta_2, m)\right)^{-1} \boldsymbol{z}. \tag{4.4.5}$$

Leave-one-out cross validation error is

$$p_n(\theta_1, \theta_2) = \sum_{i=1}^{n} \left(z_i - \hat{z}_{-i}(\theta_1, \theta_2)\right)^2 \tag{4.4.6}$$

where $\hat{z}_{-i}(\theta_1, \theta_2)$ is the best linear predictor (BLP) of z_i given z_j, $1 \leq j \leq m$ and $j \neq i$ under the Gaussian process model. The functions l_n and p_n are respectively maximized and minimized to obtain estimates of (θ_1, θ_2). Though both l_n and p_n are continuous functions, only certain values of (θ_1, θ_2) corresponds to rational correlations. We search over grids consisting of values that allow exact computation to optimize the corresponding functions.

We perform symbolic computations because the correlation matrix is very nearly singular. For example, if we take the set of observations as in Figure 4.4 with $\theta_1 = \theta_2$ chosen so that the correlation between neighboring points is 0.99, when doing double precision computations, the resulting correlation matrix is found to be not positive definite, nor is its inverse even when using the exact formula for the inverse [81].

Since the correlations between neighboring grid points $w_1 = e^{-\delta^2/\theta_1}$ and $w_2 = e^{-\delta^2/\theta_2}$ are uniquely identifiable with θ_1 and θ_2, we carry out the optimization in terms of (w_1, w_2). Denote by $C(w_1, w_2)$ the function to be optimized, either $\exp(l_n)$ or p_n. Throughout the optimization algorithm, we only consider rational w_1 and w_2 to allow exact computations. Successive grids with shrinking sizes are defined on $[0, 1] \times [0, 1]$ over which $C(w_1, w_2)$ is optimized. Once an optimizer (w_1^*, w_2^*) is found in the interior of a grid, we compare the log ratio of function values of the current iterate and the previous iterate with a convergence tolerance. Moreover, we define the 3×3 sub-grid with (w_1^*, w_2^*) at the center as $S(w_1^*, w_2^*)$, and compare the log ratio of maximal and minimal function values over $S(w_1^*, w_2^*)$ with the convergence tolerance. This comparison is done to help ensure the grid points are taken densely enough in a neighborhood of (w_1^*, w_2^*) so that a local optimum is obtained. We iterate until convergence. Details are provided in Algorithm 1.

We take $\varepsilon = 10^{-7}$ in all experiments. The initial grid search in Step 1 of Algorithm 1 led to $(0.99, 0.99)$ as the optimizer in Step 3 for both functions considered here. To check for multiple optima, we also started the algorithm with different initial grids. In addition, we search over smaller and denser grids inside $(0.01, 0.99) \times (0.01, 0.99)$ and see if an optimum could be obtained in the interior. For neither method did we find evidence for multiple local optima up to symmetry, in the sense that (w_1^*, w_2^*) generates the same cross validation error as (w_2^*, w_1^*) because of the symmetry in the observations and functions. When choosing the minimizer of the cross validation error for a grid in Step 3 of Algorithm 1, we select (w_1^*, w_2^*) with the convention that $w_1^* \leq w_2^*$.

Denote the true predictand value as $\mathbf{p} \in \mathbb{R}^{n_1}$ with $n_1 = 385$, and covariance matrices as $\Sigma_{zz} = \text{Cov}(\boldsymbol{z}, \boldsymbol{z}^T)$, $\Sigma_{pp} = \text{Cov}(\boldsymbol{p}, \boldsymbol{p}^T)$ and $\Sigma_{zp} = \text{Cov}(\boldsymbol{z}, \boldsymbol{p}^T)$. The predictions $\hat{\mathbf{p}}$ are obtained using the empirical best linear predictor (EBLP) and calibrated with empirical

Algorithm 1 Grid Search

```
Require: Convergence tolerance $\varepsilon$.
 1: Initialize grid $l_1 = l_2 = 0.01$, $r_1 = r_2 = 0.99$.
 2: Define $m_1 \times m_2$ regular grid $G = \{w_1^1, \ldots, w_{m_1}^1\} \times \{w_1^2, \ldots, w_{m_2}^2\} \subset [l_1, r_1] \times [l_2, r_2]$.
 3: Obtain optimizer $(w_1^*, w_2^*) \in G$ and corresponding function value $C^{(i)}(w_1^*, w_2^*)$ of $i$th iteration.
 4: (Test for convergence)
 5: if $\left|\log\left(\frac{C^{(i)}}{C^{(i-1)}}\right)\right| \leq \varepsilon$, $\left|\log\left(\frac{\max\{c^{(i)}(w_1,w_2),(w_1,w_2)\in S(w_1^*,w_2^*)\}}{\min\{c^{(i)}(w_1,w_2),(w_1,w_2)\in S(w_1^*,w_2^*)\}}\right)\right| \leq \varepsilon$ and $(w_1^*, w_2^*) \in \mathrm{int}(G)$ then
 6:     Return with $(w_1^*, w_2^*)$
 7: else
 8:     (Update searching grid)
 9:     for $k = 1, 2$ do
10:         if $w_k^* = w_{m_k}^k$ then        (optimized at right boundary)
11:             $dr_k \leftarrow 1 - w_{m_k}^k$
12:             $dl_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
13:         else if $w_k^* = w_1^k$ then        (optimized at left boundary)
14:             $dr_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
15:             $dl_k \leftarrow w_1^k$
16:         else        (optimized in interior)
17:             $dr_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
18:             $dl_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
19:         end if
20:         $l_k \leftarrow w_k^* - dl_k$
21:         $r_k \leftarrow w_k^* + dr_k$
22:     end for
23:     $i \leftarrow i + 1$
24:     Repeat steps 2 - 5.
25: end if
```

	MLE	CV
$(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$	$(0.024, 0.39, 0.39)$	$(3.29, 0.19, 0.66)$
$\mathrm{sd}\left(\frac{\hat{p}_i - p_i}{\sqrt{\mathrm{EMSE}(\hat{p}_i)}}\right)$	1.56	9.42
$\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(\hat{p}_i - p_i)^2}$	3.90×10^{-8}	3.99×10^{-7}

Table 4.1: Estimates of parameters (first row), standard deviations of standardized prediction errors (second row), and root mean squared prediction errors (last row) for mixture of Gaussians (4.4.1).

	MLE	CV
$(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$	$(1611.13, 0.42, 0.42)$	$(1.09 \times 10^{13}, 1.87, 1.87)$
$\mathrm{sd}\left(\frac{\hat{p}_i - p_i}{\sqrt{\mathrm{EMSE}(\hat{p}_i)}}\right)$	0.17	7.12×10^{-3}
$\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(\hat{p}_i - p_i)^2}$	7.94×10^{-7}	4.75×10^{-7}

Table 4.2: Estimates of parameters (first row), standard deviations of standardized prediction errors (second row), and root mean squared prediction errors (last row) for trig-exponential function (4.4.2).

mean squared error (EMSE). EBLP is BLP with θ replaced by its estimate $\hat{\theta}$ and is given by

$$\hat{\boldsymbol{p}} \quad = \quad \Sigma_{zp}^T(\hat{\theta})\Sigma_{zz}^{-1}(\hat{\theta})\boldsymbol{z}, \qquad (4.4.7)$$

and EMSE is the mean squared error (MSE) with θ replaced by its estimate $\hat{\theta}$ and is given by

$$\mathrm{EMSE}(\hat{\boldsymbol{p}}) \quad = \quad \Sigma_{pp}(\hat{\theta}) - \Sigma_{zp}^T(\hat{\theta})\Sigma_{zz}^{-1}(\hat{\theta})\Sigma_{zp}(\hat{\theta}).$$

For CV, we estimate the scale parameter $\widetilde{\theta}_0$ by $\hat{\theta}_0(\widetilde{\theta}_1, \widetilde{\theta}_2)$ using (4.4.5) as suggested by [86], where $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ are CV estimates for the range parameters. For the two functions, Tables 4.1 and 4.2 show the estimates, the standard deviations of standardized prediction errors, and the root mean squared errors.

Note that the CV estimates of the two range parameters for the mixture of Gaussians are not equal. Switching the two range parameter estimates (namely, $(0.66, 0.19)$) generates the same cross validation error, and if we were to employ the convention that $w_1^* \geq w_2^*$ in Step 3 of Algorithm 1, we would end up with the CV estimates for this case being $(0.66, 0.19)$. We find the unequal estimated range parameters somewhat surprising, so we did a further careful search along the diagonal $w_1 = w_2$ and could not find any points on this diagonal with smaller cross validation error than that produced by $(0.19, 0.66)$. We also experimented with estimating the mean of the Gaussian process. For mixture of Gaussians, the MLEs are $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\mu}) = (0.018, 0.38, 0.38, 0.20)$ when treating the mean as unknown and is estimated. The standard deviation of the standardized prediction errors and the root mean square error of predictions are 0.94 and 2.15×10^{-8} respectively, which are roughly similar to the results in Table 4.1 when fixing the mean at 0.

We now consider how MLE and CV estimates of the parameters perform when used for prediction. One of the important features of Gaussian processes is that they provide uncertainty estimates for the predictions, so we will look at both the quality of the point predictions and whether the standardized prediction errors (i.e., the errors divided by their estimated standard deviations) have standard deviation close to 1. If the Gaussian process model under consideration were correct, we should expect ML to do better than CV, but since the deterministic functions we consider are obviously not realizations of a Gaussian process, it is unclear which method will perform better. In terms of root mean square error, MLE is much better (by an order of magnitude, see Table 4.1) than CV for mixture of Gaussians and is moderately worse (67% larger, see Table 4.2) for trig-exponential. For mixture of Gaussians, the standardized prediction errors under MLE are reasonably well calibrated with a standard deviation of 1.56, whereas for CV, their standard deviation is 9.42, so that CV badly underestimates the variability of the prediction errors. For trig-exponential, both MLE and CV seriously overestimate the variability of the prediction errors, but CV much more so (Table 4.2).

Figure 4.5 show histograms of the standardized prediction errors for both functions and both estimates. Ideally, we might hope these histograms will approximate a standard normal distribution, but even if the truth were a Gaussian process, we should not be surprised to see something that does not look approximately normal because of possible strong dependencies between prediction errors at different locations. We see that in all four cases, the standardized prediction errors follow a vaguely symmetric distribution about 0. The most noteworthy feature in these plots occurs for the mixture of Gaussians based on CV, which was the case where the estimated range parameters were not equal. In this plot, we see that the standard deviation (sd) of the standardized prediction errors is much larger than 1 for predictands in odd columns (sd=16.10) and much smaller than 1 for predictands in even columns (sd=0.15). The tensor product form of the squared exponential covariance function implies that the EBLP of a predictand in an odd column only depends on observations in that column (see Section C.1.10), so that the form of the EBLP is entirely determined by the range parameter along columns. Similarly, EBLP in an odd row is only a function of observations in that row, whereas EBLPs in an even row and even column depend on all of the observations. Note that the estimated range parameter is large along the columns, so the fitted model thinks observations are much more strongly correlated in this direction. Since the EBLPs within odd columns only depend on within column correlations, it is then perhaps not surprising that the model is overoptimistic about the quality of interpolations in the direction with strong estimated correlations.

Figure 4.6 shows the (unstandardized) prediction root mean squared error averaged over each row and column of the large grid. For each function, we focus on the method yielding smaller prediction error. Since the estimates of the two range parameters are equal for both of these cases and the observations and function are symmetric about the diagonal, the prediction errors are also symmetric. Hence, averaging over rows and columns gives identical results, so we only present the root mean squared error averaged over the rows. Figures 4.7 and 4.9 show the prediction errors $\hat{p}_i - p_i$ at the corresponding locations for the two functions. Note that the magnitudes of the errors are only comparable for each function;

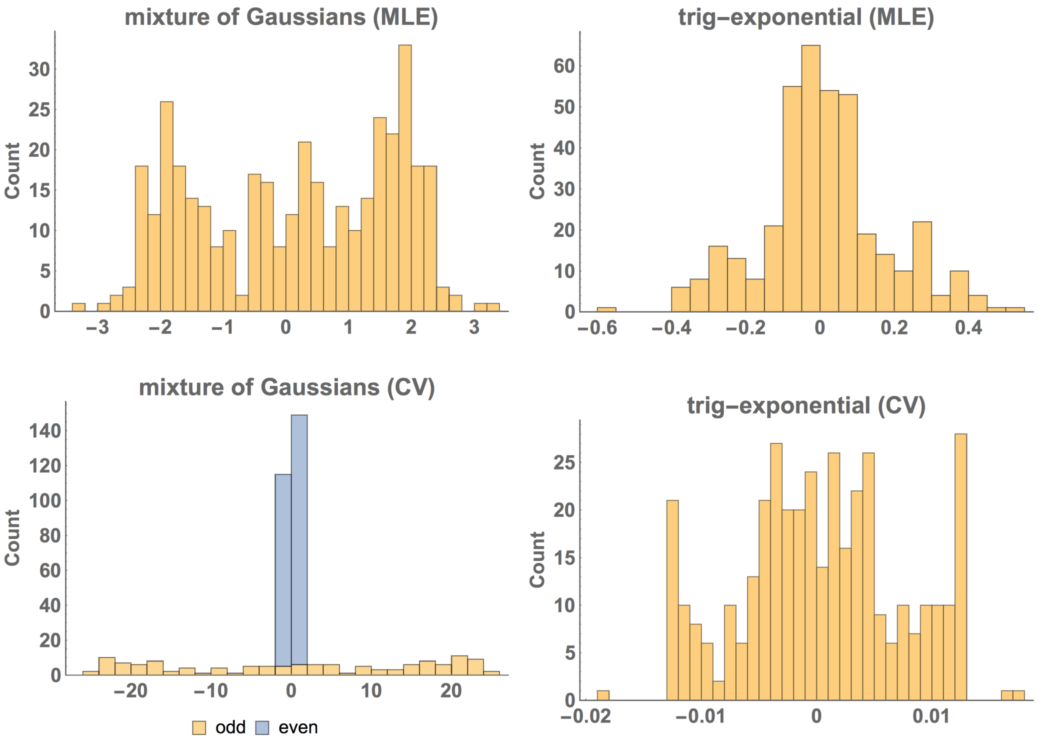

Figure 4.5: From left to right: histograms of standardized prediction errors generated by MLE for mixture of Gaussians, CV for mixture of Gaussians, MLE for trig-exponential function, and CV for trig-exponential function. The standardized prediction errors generated by CV for mixture of Gaussians are grouped into those in odd and even numbered columns. The histograms for the two groups are stacked.

they are not normalized across functions.

Figures 4.6–4.9 show that, for both test functions, the prediction errors are largest for the predictions on the second and second to last rows and columns (i.e., rows and columns 2 and 22 out of 23). We should generally expect prediction errors to be larger near a border of an observation domain than in the interior, but it is interesting to note that, at least for these functions, the errors tend to be larger, for example, in row 2 than row 1, even when comparing a predictand in row 2 and an odd column to one in row 1 and an even column, so that the distance from the predictand to the nearest observation equals $\frac{1}{23}$ in both cases.

Next, we show numerically that the MLE for Branin function (4.4.3) does not appear to exist. First of all, let us consider the estimation of the range parameter when $f(x) = x^p$

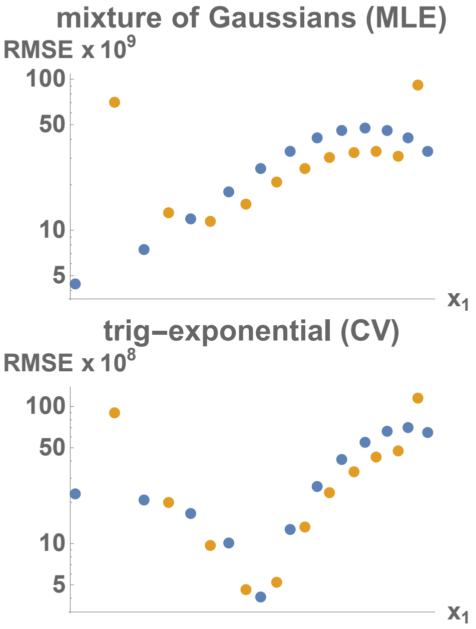

Figure 4.6: Root mean squared errors for each column and row of the 23 × 23 grid. Blue: odd; Yellow: even.

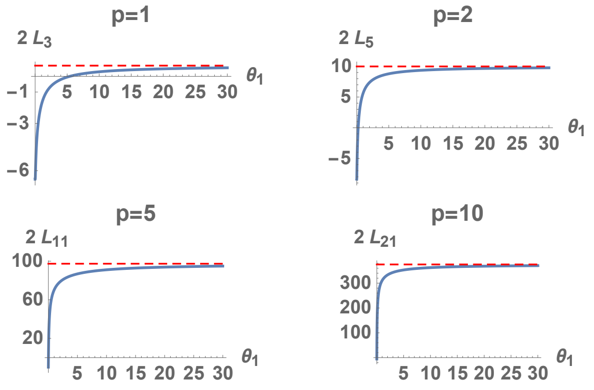

Figure 4.8: Profile log-likelihood $2L_{n^*(p)}(\theta_1)$ for $f(x) = x^p$, $n^*(p) = 2p + 1$ and $p = 1, 2, 5, 10$.

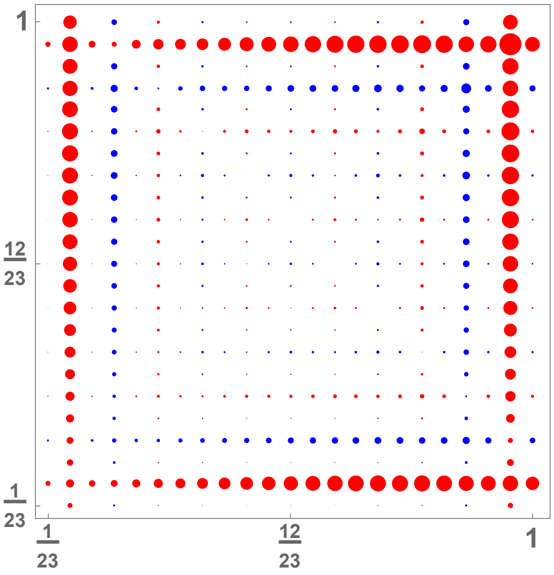

Figure 4.7: Prediction errors of MLE for mixture of Gaussians. Red: $\hat{p}_i > p_i$; Blue: $\hat{p}_i \leq p_i$. Area of disk is proportional to $|\hat{p}_i - p_i|$.

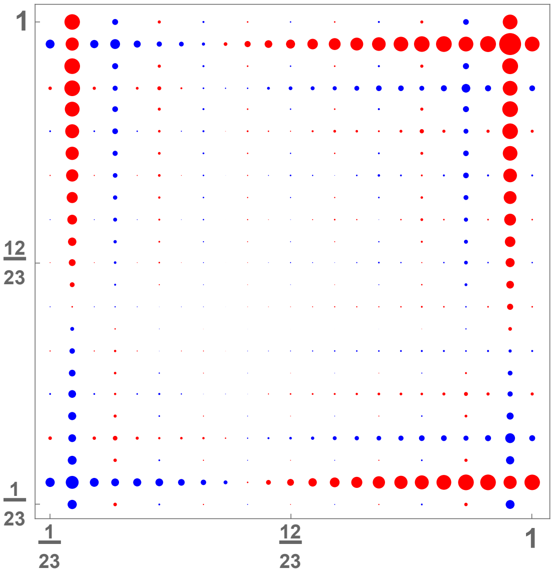

Figure 4.9: Prediction errors of CV for trig-exponential function. Red: $\hat{p}_i > p_i$; Blue: $\hat{p}_i \leq p_i$. Area of disk is proportional to $|\hat{p}_i - p_i|$.

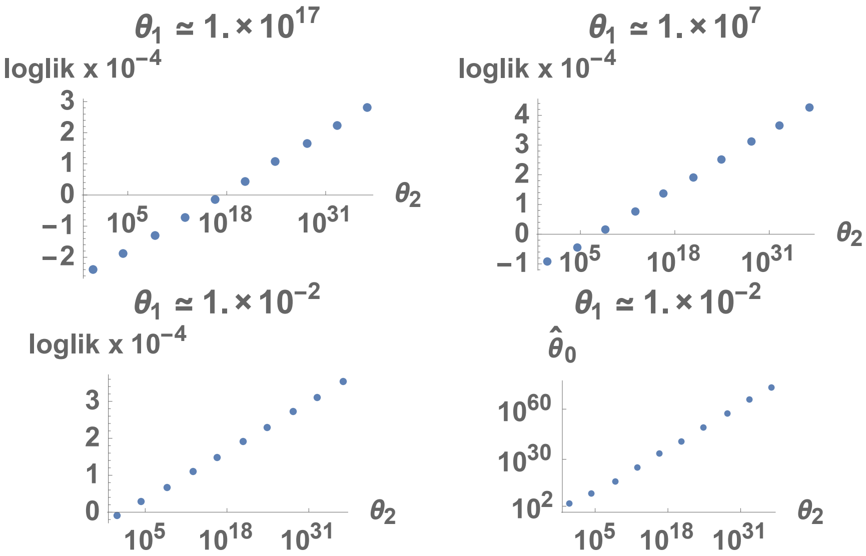

Figure 4.10: Profile log-likelihood l_{14} (4.4.4) and estimated scale parameter $\hat{\theta}_0$ for Branin function (4.4.3).

with $p \geq 1$. The profile log-likelihood $L_n(\theta_1)$ satisfying

$$2L_n(\theta_1) = -n\log(\hat{\theta}_0(\theta_1)) - \log|R(\theta_1, n)| - n\log 2\pi - n$$

is maximized to obtain MLE $\hat{\theta}_1$.

Our empirical evidence suggests that for each p and $n^*(p) = 2p+1$, when $n \geq n^*(p)$, $L_n(\theta_1)$ monotonically increases for $\theta_1 \in (0, \infty)$ so that the MLE for θ_1 does not exist. Moreover, $L_n(\theta_1)$ is bounded when $n = n^*(p)$ and increases to ∞ when $n > n^*(p)$. As noted in [79], this finding also appears to be documented in an unpublished thesis [61]. Figure 4.8 shows that at the critical value $n^*(p)$, $2L_{n^*(p)}(\theta_1)$ monotonically increases with a finite asymptote for some choices of p. Since the Branin function is a quadratic polynomial along its second dimension, we expect that the MLE for the second range parameter $\hat{\theta}_2$ does not exist. In fact, Figure 4.10 shows that for different choices of θ_1, the profile log-likelihood (4.4.4) increases for increasing θ_2. Also, the estimated scale parameter appears to be unbounded above as θ_2 increases.

4.4.2.2 Experiments with nugget effect

A common approach to overcome the numerical difficulties in computing with the covariance function (4.1.1) is to include a small nugget effect to stabilize the computation of the covariance matrix inversion [6, 103]. In the following, we add a small nugget effect δ_0 so that the covariance matrix of the observations has the form

$$\theta_0\big\{R(\theta_1, m) \otimes R(\theta_2, m) + \delta_0 I_{m^2}\big\}, \tag{4.4.8}$$

where $m = 12$ is the number of observations along each dimension. In the above formulation, we treat the nugget size δ_0 fixed when fitting the model. For the two test functions (4.4.1) and (4.4.2), we investigate the effect of including a nugget on model fitting and prediction.

With the alternative model (4.4.8), we evaluate the log-likelihood and prediction errors of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ obtained with exact computations in Section 4.4.2.1 for 11 values of δ_0 equally spaced on log scale between 10^{-14} and 10^{-12}. The value 14 is the largest integer k for which a nugget of 10^{-k} generally yields a covariance matrix that is found to be numerically nonsingular by *Mathematica*'s **CholeskyDecomposition** routine. For each δ_0, the scale parameter θ_0 is refitted with the model (4.4.8) but the range parameters are not changed. Figure 4.11 shows the log-likelihood and root mean squared errors of $(\hat{\theta}_1, \hat{\theta}_2)$ for the model (4.4.8).

For both test functions, the likelihood is substantially reduced when including even a nugget of 10^{-14} and decreases for increasing nugget size. To help see why the log-likelihood changes so much, note that log-likelihood can also be obtained with the conditional distributions of successive ordered observations so that

$$l(\theta|\mathbf{z}) \quad = \quad -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \log\left(\mathrm{sd}(z_i|z_1, \ldots, z_{i-1})\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{z_i - E(z_i|z_1, \ldots, z_{i-1})}{\mathrm{sd}(z_i|z_1, \ldots, z_{i-1})}\right)^2,$$

where $\mathrm{sd}(z_i|z_1, \ldots, z_{i-1})$ denotes conditional standard deviation and $E(z_i|z_1, \ldots, z_{i-1})$ de-

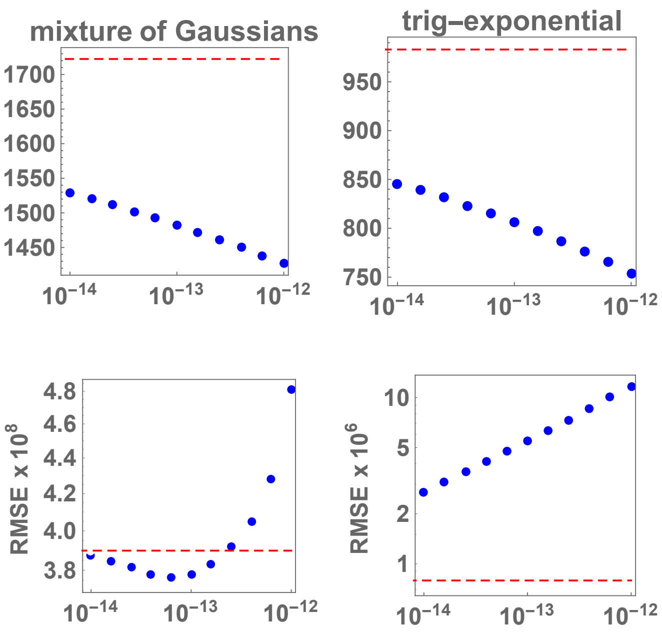

Figure 4.11: Log-likelihoods (top) and root mean squared prediction errors (bottom) of $(\hat{\theta}_1, \hat{\theta}_2)$ obtained with exact computations for model (4.4.8). Horizontal axes indicate nugget effect δ_0. Reference line indicates nugget-free case.

notes conditional mean. The log-likelihood can hence be expressed by conditional standardized errors and conditional standard deviations. We order the observations lexicographically so that $f(\frac{i_1}{m}, \frac{i_2}{m})$ precedes $f(\frac{j_1}{m}, \frac{j_2}{m})$ if and only if $i_1 < j_1$ or whenever $i_1 = j_1$, $i_2 < j_2$. The top panels of Figure 4.12 show, for each observation, the standardized errors and standard deviations conditional on previous observations. For each test function, we compare the case with $\delta_0 = 0$ and the case with the nugget size $\delta_0^* > 0$ yielding the smallest prediction error among the 11 positive values for δ_0 we considered, which for the mixture of Gaussians, is $\delta_0^* = 10^{-66/5} \approx 6.3 \times 10^{-14}$ and for the trig-exponential function, is $\delta_0^* = 10^{-14}$. For both test functions, the conditional standardized errors are similarly calibrated for $\delta_0 = 0$ and $\delta_0 = \delta_0^*$. However, some of the conditional standard deviations are much smaller when there is no nugget. Successive predictions of the ordered observations are more accurate at some locations when the model does not have a nugget.

For the trig-exponential function, Figure 4.11 shows root mean squared prediction error increases for increasing nugget size. However, better successive predictions at the test obser-

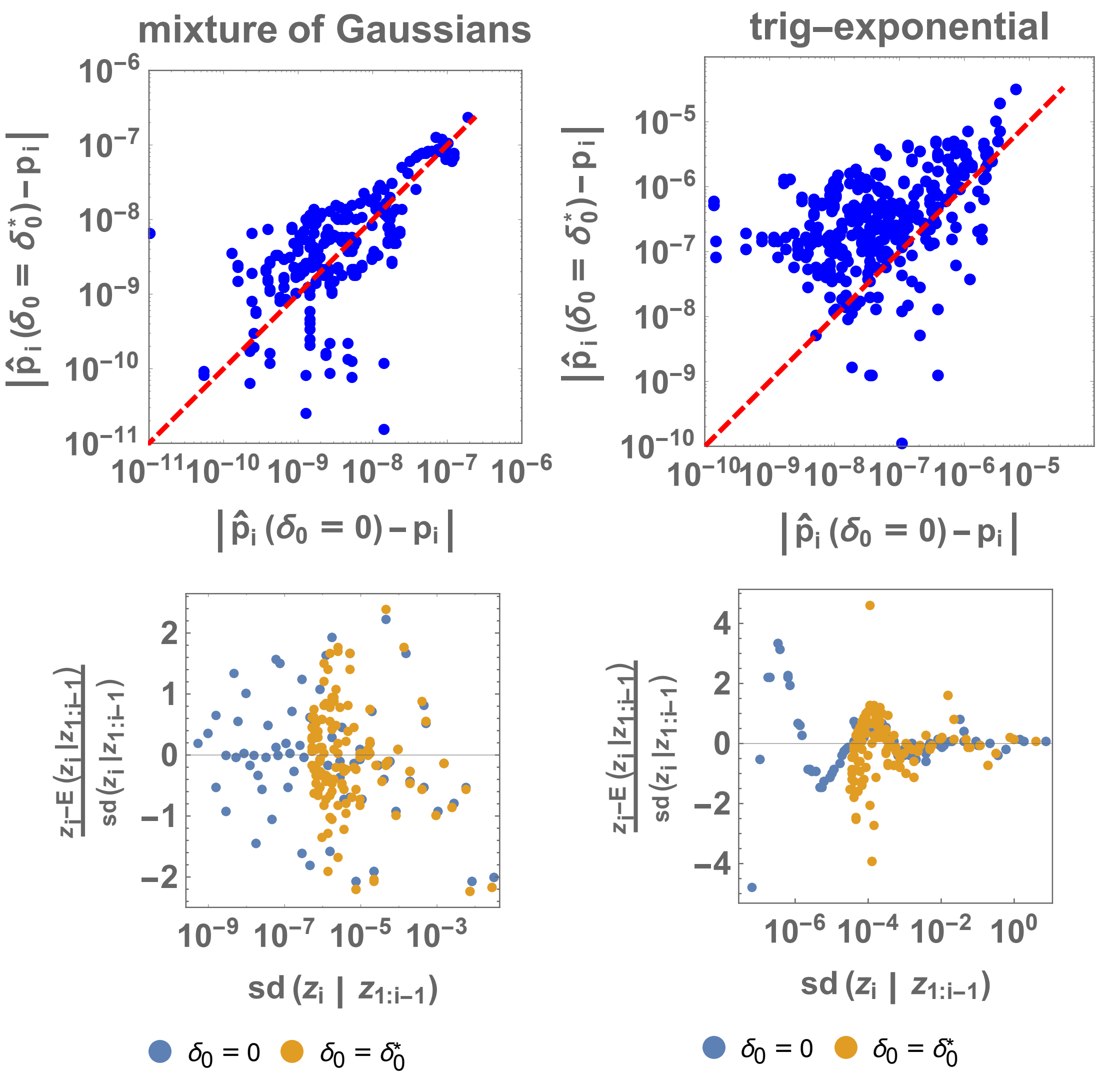

Figure 4.12: Absolute prediction error for $\delta_0 = 0$ and $\delta_0 = \delta_0^*$ yielding the smallest RMSE for each function (top); Successive conditional standardized errors for ordered observations (bottom).

vations does not necessarily imply better predictions at other locations: we do see that for the mixture of Gaussians, the prediction error is slightly smaller for $\delta_0 = 10^{-14}$ compared with the nugget-free case and further decreases as δ_0 increases before eventually increasing. The bottom panels of Figure 4.12 show the absolute prediction errors for each predictand when $\delta_0 = 0$ and $\delta_0 = \delta_0^*$. For the mixture of Gaussians, there is no obvious dominance for either $\delta_0 = 0$ and $\delta_0 = \delta_0^*$. In contrast, for the trig-exponential function, it is evident that the nugget-free model predicts better at most locations, perhaps especially for locations with the largest error, which tend to be near the boundaries of the observation region.

4.5 Conclusions

Since the approach was proposed by [104, 105], it has become quite common practice to model the deterministic output of a computer experiment as a realization of a Gaussian process. The Gaussian process with squared exponential covariance function is infinitely differentiable and thus is attractive if the computer model output is known to be smooth. In this article, we investigated the asymptotics for the maximum likelihood estimator of the scale parameter for this covariance when the computer response is a pth order monomial. Using exact computation, we investigated and compared MLE and CV estimates in a prediction problem.

Using the exact expression for the Cholesky factor and its inverse of the correlation matrix derived in [81], we proved that for regularly spaced observations, when the test function is a pth order monomial and the range parameter is fixed, the MLE of the scale parameter $\hat{\theta}_0 \to 0$ when $p = 0$ and $\hat{\theta}_0 \to \infty$ when $p = 1$ as the number of observations $n \to \infty$. When the observations are derivatives of the model function at zero, we derived the exact expression of the inverse Cholesky factor of the correlation matrix and proved asymptotic orders of $\hat{\theta}_0$ for all $p \geq 0$. We are unable to prove an asymptotic order for general $p > 1$ for regularly spaced observations. However, we conjecture that the asymptotic order is the same as that of the derivative case with a possibly different constant.

Though both MLE and CV are used in the computer experiment literature, it is not clear under what circumstances each method will yield smaller prediction errors with more calibrated standardized errors. When a model is misspecified, CV can sometimes be a good way of choosing parameters for prediction, since the cross validation criterion is based on prediction. For deterministic computer experiments, we know that in fact the outputs are not realizations of some Gaussian process model. Nevertheless, our experiments show that CV does not always improve upon the likelihood method. For example, CV estimates produce much larger prediction errors and poorer calibration for the mixture of Gaussians and modestly better predictions but far worse calibration for the trig-exponential function. This finding is consistent with the findings in [9] which shows CV appears to be less robust than MLE to model misspecification under regular grid design.

These numerical experiments used exact arithmetic, which will not generally be possible. Adding a small nugget is a common approach to alleviate the numerical instabilities in decomposing nearly singular covariance matrices. Our experiments suggest model fitting (as measured by the likelihood) deteriorates substantially by adding even a nugget that barely makes the covariance matrix numerically positive definite, whereas prediction can sometimes be slightly improved by adding a small nugget. Another interesting finding of our work is that, for the Branin test function, the MLEs do not appear to exist. This result can be viewed as an implication of the numerical finding that when the model function is a pth order monomial, the MLE of the range parameter does not exist when the number of observations exceeds a critical value (see also [61, 79]). For test functions that are a polynomial in one of its dimensions, the use of the likelihood method is not expected to produce meaningful estimates and inference. Though our results regarding the estimation of range parameters is empirical and based on limited numerical experiments, we believe that the examples shown give an indication of possible issues and consequences when using simple smooth test functions to study how well Gaussian process models work for deterministic computer models. Whether similar results might hold for, say, the numerical solution of a complex system of differential equations deserves further study.

CHAPTER 5

MODELING AND PREDICTING CHAOTIC CIRCUIT DATA

5.1 Introduction

The study of nonlinear dynamics and chaos has been traditionally focused on developing and investigating mathematical invariants that describe and classify the asymptotic behavior of the iterates [e.g., 56, 59]. This leads to important invariants such as Lyapunov exponent and fractal dimension with which knowledge about the future evolution of the dynamical system can be learned, once its mathematical description is known. However, challenges still remain to analyze observations generated by a dynamical system whose mathematical formulation is at least partially unknown. These challenges include, but are not limited to, calculating geometric and dynamical invariants of an underlying strange attractor [e.g., 25, 106, 118], modeling the deterministic portion of the dynamical evolution from the observations [e.g., 34, 66], and constructing a predictive model directly from the observations [e.g., 27, 43]. A comprehensive discussion of the problems arising from analyzing the observed chaotic data can be found in [2] and [1]. In this work, we focus on the aspects of modeling the observed dynamical system and predicting its future evolution. The modeling and prediction task is made difficult in particular by the characterization of possible stochastic model error of the underlying dynamics, observational error during measurement, and the separation of the two. The goal of this work is to provide some insights into modeling and estimation issues through a combination of real data analysis and simulation studies.

The modeling and prediction of observations generated from a system with nonlinear and chaotic dynamics has revolved around local techniques, global techniques and the combination of the two [2]. A common local prediction technique discussed in various instances [e.g., 27, 34, 43] is to construct the k-step ahead predictor at a time point in the form of a polynomial whose inputs are past observations and output is the prediction of the observation k steps ahead. The parameters of the polynomial are fitted using the near neighbors of the input and their corresponding k-step ahead observations. For example, the simplest nonlinear method of local forecasting proposed by [82] is to find the nearest neighbor s_t to the current observation s_n, and use the value s_{t+1} as the prediction for s_{n+1}. This prediction-by-analogue method is essentially fitting a constant – a zeroth order polynomial. Fitting with more than one near neighbor leads to prediction by averaging their outputs [100] in this case. Fitting a first order polynomial is considered in [27] and [84]; [43] investigates higher order polynomials. The use of higher order polynomials increases the model complexity and is expected to produce better predictions. However, since the number of parameters increases exponentially in the order of the polynomial, more computational efforts are also encountered. In addition, the local models are discontinuous, which is undesirable if the goal is to obtain a description of the underlying continuous dynamics.

Global models, on the other hand, describe the whole set of observations by representing the model mapping as an expansion in some basis functions, e.g., as a polynomial or ratio of polynomials, and fitting the parameters using the entire data set [e.g., 2, 14, 27]. The method is also subject to computational difficulties when the model is of high complexity. Modeling using radial basis functions [102] is an example of combining features from both the local and global techniques. The model is constructed and interpreted globally but also maintains good local properties through the locally centered radial basis functions [27]. In our work, we construct global models trained for one-step ahead prediction using Gaussian process (GP) and neural network (NN) models. GP provides a statistical basis for interpolation, model diagnostic and uncertainty calibration, while NN has proven effective in modeling a range of nonlinear problems [e.g., 49, 76].

We consider analyzing and modeling voltage measurement data generated by a laboratory-built electrical circuit [84]. The observations are relatively smooth, concentrate on a low-dimensional attractor, and exhibit sensitive dependence on the initial condition. A nominal differential equations model, based on a simplistic description of the circuit components, has systematic deficiencies when fitted to the data. We hence investigate modeling and prediction using GP and NN models. For both methods, we train a one-step ahead predictor based on the input-output pairs of m preceding observations and the current one for some embedding length $m > 0$. To investigate the capacities of the predictors in capturing the dynamics, we investigate the tradeoff between one-step prediction and long-term tracking. We find that both models perform similarly in one-step prediction, and the prediction error decreases as m increases. In contrast, the ability of the model propagations to track the observations improves at first but then degrades as m becomes larger for both models, which suggests a moderate value of m produces better balance between one-step prediction and long-term tracking.

One of our goals for analyzing observations generated from some unknown dynamics is to investigate the effects and characterizations of the model and observational errors. We consider this aspect by performing simulations with data generated by our fitted models. The fitted models capture the chaotic character of the observations, and the simulated data is qualitatively similar to the observations. We explore the effects of model and observational errors on the likelihood function and the identifiability of the initial state. We find that with independent and identically distributed (i.i.d.) observational error and no model error, the likelihood ratio between the true initial state and neighboring points increases exponentially in the number of observations. However, with even a tiny stochastic model error but otherwise correct dynamics, the true initial state no longer maximizes the likelihood function, and there does not seem to be an initial state able to track the observations for an indefinitely long time. A temporally correlated observational error with no model error, on the other hand, preserves the identifiability of the true initial state as the maximizer of the likelihood function. The information provided by the likelihood about the true initial state also grows exponentially with more observations but at a lower rate than that for the i.i.d. observational error.

The rest of the chapter is organized as follows. In Section 5.2, we introduce the data set and discuss the deficiencies of the nominal model in capturing the dynamics present in the data. In Section 5.3, we compare the prediction performances of GP and NN models and investigate their capacities in describing the long-term dynamics. Section 5.4 considers the effects of model and observational errors using simulation data qualitatively similar to the observations, and compares the fitted models in Section 5.3 with simulations under model errors estimated from a reconstructed physical circuit. Section 5.5 discusses ultimately unsuccessful efforts to model the system using the circuit simulator SPICE [41, 92].

5.2 The circuit data

In this work, we consider a time series consisting of voltage measurements of a laboratory-built electrical circuit [84]. The circuit was built on a breadboard using capacitors, resistors, operational amplifiers and multipliers. A diagram of the circuit is shown in Figure 5.1, where V_1, V_2 and V_3 are the nodes at which voltages are measured. In the rest of the section, we refer to the three voltages measurements V_1, V_2 and V_3 as coordinates x, y and z, so that the observation at time step n is $s_n = (x_n, y_n, z_n)$. The circuit was allowed to run for several minutes before data collection, and the measurements were taken at a frequency of 10kHz for about 1.5 minutes, resulting in three time series each with length 1 million. There were nine sets of measurements taken under different conditions in [84, Chapter 2], and the data we consider here belong to “set7”. Figure 5.2 shows the observations at the initial 100,000 time points, and Figure 5.3 shows the trajectories of the initial 1000 observations. The data points loop around and fall densely on a two-dimensional manifold as shown in Figure 5.2. Moreover, Figure 5.3 shows the observed trajectory is fairly smooth, indicating relatively low observational noise. Figure 5.4 shows the difference in Euclidean norm at each subsequent time point between two trajectories that start close by. Specifically, we select from all time points t, other than the initial one, the t^* that minimizes $\|s_0 - s_t\|$, and consider the differences $\|s_n - s_{t^*+n}\|$ for $n = 0, 1, \ldots, 500$. The exponentially growing divergence of the two trajectories suggests the observations exhibit sensitive dependence on the initial conditions.

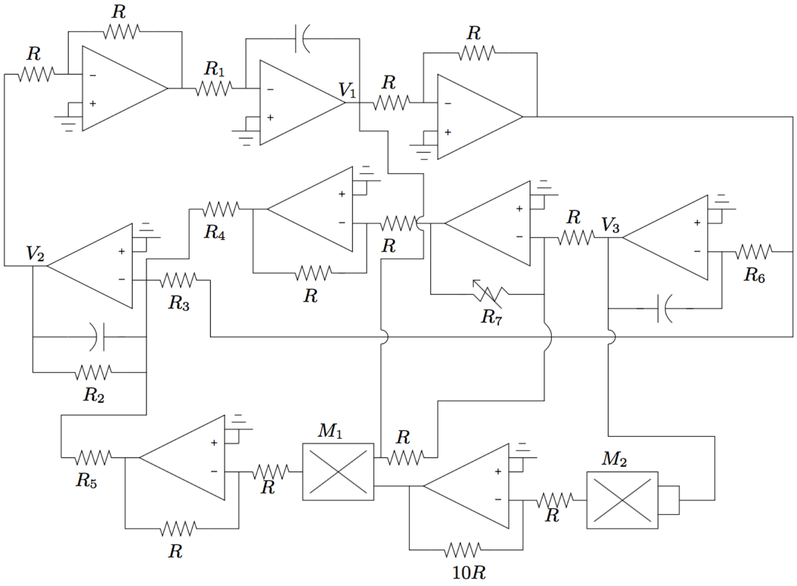

Figure 5.1: Circuit diagram [84, Figure 2.5]. V_1, V_2 and V_3 indicate the nodes at which voltages are measured.

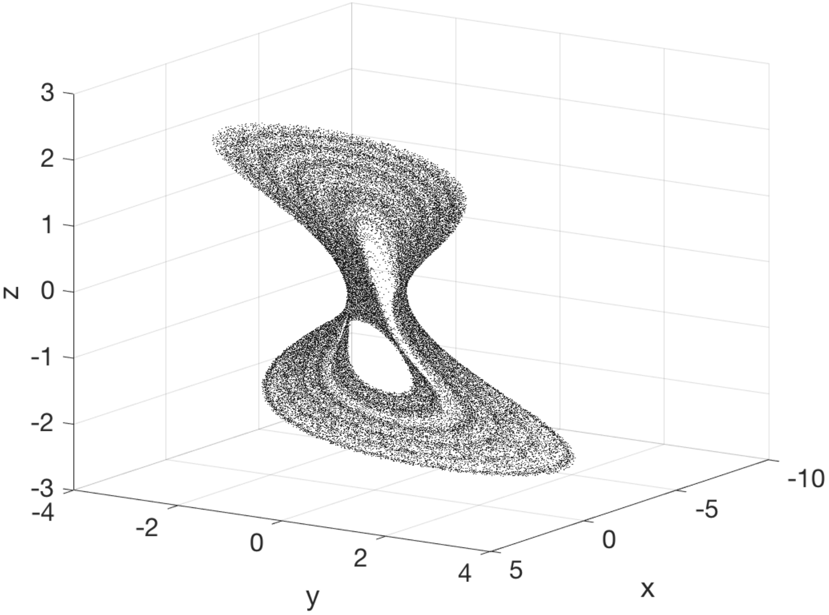

Figure 5.2: The initial 100,000 observations. x, y and z correspond to the voltage measurements V_1, V_2 and V_3 accordingly. The observations in the 3D space fall on a two-dimensional attractor.

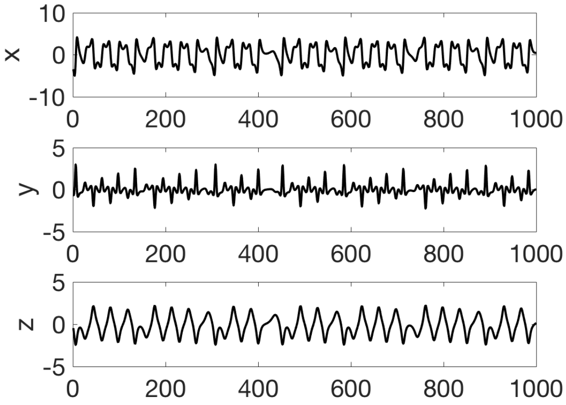

Figure 5.3: Trajectories of the initial 1000 observations.

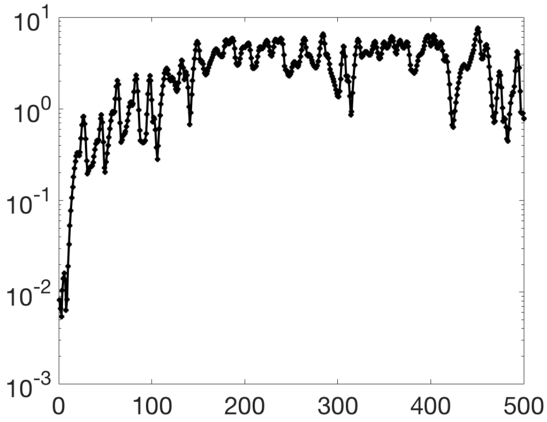

Figure 5.4: The trajectory differences $\|s_n - s_{t^*+n}\|$ for $n = 0, 1, \ldots, 500$. s_{t^*} is the closest observation to the initial observation s_0.

Because of the simple structure of the circuit, applying Kirchhoff's law to the idealized behavior of the circuit components yields the following nominal model:

$$\begin{aligned} \frac{dx}{dt} &= \theta_1 y \\ \frac{dy}{dt} &= -\theta_2 y + \theta_3 x - \theta_4(x+z) - \theta_5 x z^2 \\ \frac{dz}{dt} &= \theta_6 x. \end{aligned} \tag{5.2.1}$$

This set of ODEs (5.2.1) is isomorphic to the Moore-Spiegel system [90], which is a nonlinear thermodynamical oscillator that has its physical origin in fluid dynamics. It models the displacement $z(t)$ of a small mass element attached at a fixed point to an elastic spring oscillating in a temperature stratified fluid. The element exchanges heat with the ambient fluid and its buoyancy depends on the temperature. The Moore-Spiegel system, like the Lorenz attractor, is one of the classical low-order dynamical systems that exhibit chaotic behavior for certain choices of the parameter values [10, 90]. [84] points out that significant discrepancies exist between the observations and the nominal model (5.2.1). For example, with the parameter values used in the circuit, (5.2.1) settles down to a periodic orbit whereas the observations manifest chaotic behavior. In the following, we investigate the deficiency of the nominal model with parameters selected optimally by some criteria.

We estimate the parameters of (5.2.1) by minimizing the one-step ahead prediction error as follows:

$$\min \quad \frac{1}{2}\sum_{n=1}^{N-1} \left\| s_{n+1} - F^{(k)}\left(s_n, \theta, h/k\right) \right\|^2 \tag{5.2.2a}$$

$$\text{s.t.} \quad \theta > 0, \tag{5.2.2b}$$

where $\theta = (\theta_1, \ldots, \theta_6)$, $h = 10^{-4}$ is the time distance between observations, and $F(s, \theta, h)$ is obtained by integrating (5.2.1) with the Runge-Kutta 4th order method. To reduce the numerical error in the model propagation, we iterate k times the mapping F from one time point to the next, and $k = 6$ is chosen from $k = 1, \ldots, 10$ by minimizing the resulting prediction error. We solve (5.2.2) with $N = 2000$. Figure 5.5 shows the prediction errors of the fitted model (5.2.1) for the initial 100 time points. The trajectories for the prediction errors are fairly smooth and systematic, suggesting possible dynamics not captured by model (5.2.1). Figure 5.6 shows the observations and predictions projected to the x-z plane. Systematic departures of predictions from observations are also noticeable. For example, on the lower left portion of the plane where $x < 0$ and $z < 0$, the predictions tend to be ahead of the observations; while on the lower right portion where $x > 0$ and $z < 0$, the predictions tend to lag behind the observations. The systematic patterns in the prediction errors suggest the inadequacy of model (5.2.1) in describing the dynamics present in the observations.

In reality, the circuit components do not behave as simply as the nominal model (5.2.1) suggests. For example, the nominal model does not take into consideration any parasitic elements of the circuit, which are unavoidable and include stray inductance, capacitance and resistance [e.g., 4, 60]. These parasitics can alter the behavior of the circuit depending on the frequency of the signal, which in turn depends on the circuit components through (5.2.1). This intricate interaction is not captured by the nominal model, which assumes constant values for the circuit components at all frequencies. In addition, (5.2.1) models the behavior

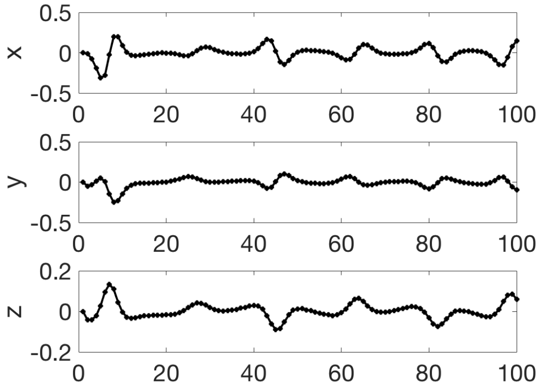

Figure 5.5: The difference between observations and one-step ahead predictions at the initial 100 time points with parameters estimated by solving (5.2.2). $N = 2000$, $h = 10^{-4}$, $k = 6$.

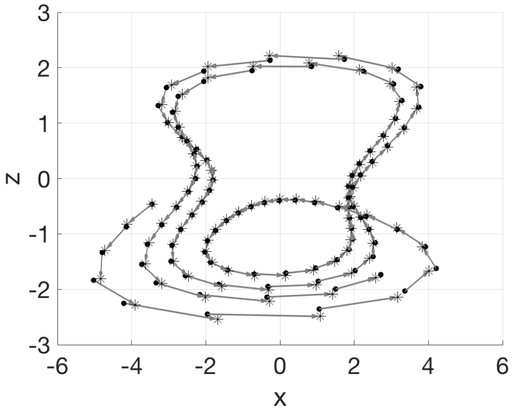

Figure 5.6: The observations (dot) and the predictions (asterisk) for the initial 100 time points projected to the x-z plane, with parameters estimated by solving (5.2.2). $N = 2000$, $h = 10^{-4}$, $k = 6$.

of the circuit on a macro level, but does not take into account effects with micro resolutions such as thermal noise, which is stochastic in nature. In the next section, we will explore alternative models for the same one-step ahead prediction problem, and investigate their capacities in capturing the long-term dynamics.

5.3 Gaussian process and neural network models

In this section, we consider a Gaussian process (GP) and a neural network (NN) model for prediction based on m previous observations with $m \geq 1$. Specifically, denote the observation at time step k as $s_k = (s_k(1), s_k(2), s_k(3))$, $k = 1, \ldots, N$. For each component $\nu \in \{1, 2, 3\}$, we model the outputs $s_k(\nu)$ as a GP (or NN) with inputs $(s_{k-m}, \ldots, s_{k-1})$ for $k = m + 1, \ldots, N$. Note that each input is $3m$-dimensional.

For the GP model, we consider the following squared-exponential covariance function:

$$\mathrm{Cov}\left(f(q), f(r)\right) = \sigma^2 e^{-\sum_{i=1}^{\beta}\left|\frac{q_i - r_i}{\theta_i}\right|^2} \tag{5.3.1}$$

for some $q, r \in \mathbb{R}^\beta$ and $\beta = 3m$. Furthermore, we include a nugget effect $\gamma > 0$ so the covariance matrix is $\Sigma = \sigma^2(R + \gamma I)$ where the (i, j)th element of R is the correlation between the ith and jth outputs. Note that the GP with the inputs and outputs specified above is not an internally consistent model for the observations, in the sense that the observations can not follow a joint normal distribution since they appear both as the outputs and inputs of the GP. A similar situation appears in [32] which considers emulation of dynamic computer codes using GP. We nonetheless use GP as a tool to fit a predictor by estimating parameters through maximizing the ostensible Gaussian likelihood and making predictions through interpolation. Denote the outputs as $\mathbf{s} = (s_{m+1}(\nu), \ldots, s_N(\nu))$ and the range parameters as $\theta = (\theta_1, \ldots, \theta_{3m})$, then the log-likelihood function by profiling over the scale parameter σ^2 satisfies

$$\begin{aligned} 2l(\theta, \gamma) = & -\log|R(\theta) + \gamma I| - n \log \mathbf{s}^T (R(\theta) + \gamma I)^{-1} \mathbf{s} \\ & - n \log(2\pi) + n \log n - n, \end{aligned} \tag{5.3.2}$$

and

$$\hat{\sigma}^2 = \frac{\mathbf{s}^T (R(\theta) + \gamma I)^{-1} \mathbf{s}}{n},$$

where $n = N - m$. Then (5.3.2) is maximized to obtain maximum likelihood estimates (MLEs) $\hat{\theta}$ and $\hat{\gamma}$. At time step t with input $(s_{t-m}, \ldots, s_{t-1})$, the prediction of $s_t(\nu)$ is made by the empirical best linear predictor (EBLP):

$$\hat{s}_t(\nu) = \Sigma^T_{\mathbf{s}s_t(\nu)}(\hat{\theta}, \hat{\gamma}) \Sigma^{-1}_{\mathbf{ss}}(\hat{\theta}, \hat{\gamma}) \mathbf{s}.$$

A second model we experiment with is a feed-forward neural network, a detailed account of which can be found in [49]. We fit a NN model consisting of 1 hidden layer with 20 neurons using the same set of inputs and outputs as those for fitting the GP model. The estimation and prediction are carried out using MATLAB Neural Network Toolbox 9.1 [15]. Note that for both GP and NN, three separate models are fitted for predicting the three components

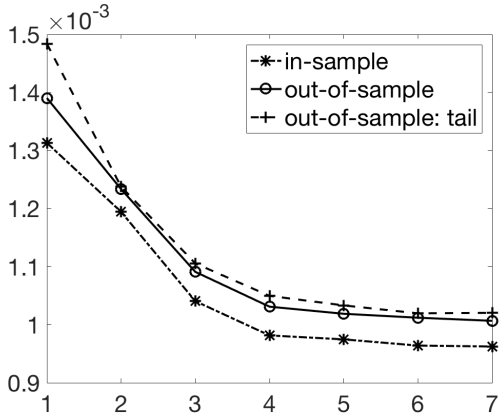

Figure 5.7: RMSEs at each embedding length m for predicting the initial N observations (in-sample), the next N observations (out-of-sample) and the tail N observations (out-of-sample: tail) using the GP model. $N = 2000$, $m = 1, \dots, 7$.

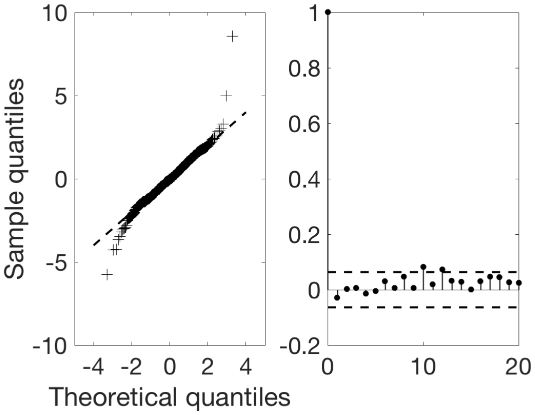

Figure 5.8: QQ-plot and autocorrelations of the out-of-sample successive standardized errors for predicting the 3rd component when $m = 7$.

of the observation.

For each component ν, we fit GP and NN models as discussed above based on the initial N observations, and evaluate the root mean squared errors (RMSEs) for predicting in-sample and out-of-sample. We consider two sets of observations when evaluating out-of-sample: one is the next N observations after the initial N observations used to fit the model, and the other is the N observations at the tail end of the data set. Since the data are collected in a chronicled order and the properties of the electrical components may change with environmental conditions such as temperature as the circuit operates, we expect the tail part of the data set to be the most disparate to the in-sample data, and hence may provide information about possibly time-varying parameters.

Figure 5.7 shows the in-sample and the two types of out-of-sample RMSEs for the GP model fitted with the initial $N = 2000$ observations. Prediction errors decrease and gradually stabilize as the embedding length increases. The out-of-sample error evaluated with the tail part of the data set is only slightly larger than that evaluated with the second N observations, suggesting little variation of the parameters across time. The results for the NN model are only slightly worse; the differences of RMSEs between the two models are on the order of 10^{-5}, which is two orders of magnitude smaller than the prediction error itself. Figures D.1 and D.2 in Supplement D.1 show the RMSEs for the NN model and the comparison between the two models.

The use of GP provides a statistical basis for a model diagnostic. Under the Gaussian process model assumption for $\{s_k(\nu)|k = m+1, \ldots, N\}$, for $\nu \in \{1, 2, 3\}$, we have that the successive standardized errors are independent and normally distributed. In other words, we expect that

$$\frac{s_k(\nu) - E(s_k(\nu)|s_{1:k-1}(\nu))}{\sqrt{\mathrm{Var}(s_k(\nu)|s_{1:k-1}(\nu))}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

under the GP model. Figure 5.8 shows the QQ-plot and the autocorrelations of the out-of-sample successive standardized errors for predicting the third component ($\nu = 3$) when $m = 7$. The successive standardized errors are well approximated by a normal distribution except for a few outliers, and the independence condition is not obviously violated.

Note that both the GP and NN models are constructed and fitted for one-step prediction based on m previous observations, while our aim is to understand the underlying long-term dynamics based on the observations. As a result, we investigate the capacities of the previously developed models in long-term prediction. Specifically, we propagate the fitted models out-of-sample from multiple starting points, and compare the trajectories of the model propagation and observations. One quantitative measure of the long-term predictability is hence the area enclosed by the two trajectories. For $\nu \in \{1, 2, 3\}$, denote the model propagation and observations from the ith starting point as $\{p_k^{(i)}(\nu)\}_{k=1:T}$ and $\{s_k^{(i)}(\nu)\}_{k=1:T}$, respectively, for a tracking length of T. We define the area $A^{(i)}(\nu)$ between the two trajectories as the sum of the areas of the trapezoids whose vertices are consecutive observation and model propagation points, i.e.,

$$A^{(i)}(\nu) = \sum_{k=1}^{T-1} \Bigg[\frac{1}{2} \Big(\max(s_{k+1}^{(i)}(\nu), p_{k+1}^{(i)}(\nu)) + \max(s_{k}^{(i)}(\nu), p_{k}^{(i)}(\nu)) \Big) \\ - \frac{1}{2} \Big(\min(s_{k+1}^{(i)}(\nu), p_{k+1}^{(i)}(\nu)) + \min(s_{k}^{(i)}(\nu), p_{k}^{(i)}(\nu)) \Big) \Bigg]. \tag{5.3.3}$$

An illustration of this definition of area is shown in Figure D.3 in Supplement D.1. We calculate the average area over all components and all starting points defined by

$$\overline{A} = \frac{1}{3L} \sum_{i=1}^{L} \sum_{\nu=1}^{3} A^{(i)}(\nu), \tag{5.3.4}$$

and use that as a measure of the long-term predictability of a model.

Figure 5.9 shows the average area between the model propagation and observations over $L = 500$ out-of-sample starting points each tracking out a length of $T = 2000$. We choose the starting points every 200 steps to make the starting points fairly evenly distributed over the attractor. For both models, when increasing the embedding length, the long-term predictability improves for small m, then deteriorates. This pattern suggests an embedding length being too large ($m > 2$ for GP and $m > 3$ for NN) may produce a model well suited for one-step prediction while performing worse for long-term tracking with our parameter estimates. So in the following, we focus on the GP and NN models with $m = 2$ and $m = 3$, respectively.

Another interesting feature of the long-term prediction is the region on the state space where model propagations tend to lose track of the observations. To measure such a divergence of the model propagation from the observations, we use a moving window of length 50 and calculate the average area over three components between the two trajectories inside the window. An average window area exceeding 65, by empirical experiments, seems to be a good indicator of a divergence between the model propagations and observations. Figure 5.10 shows one example of such a divergence pattern. Note that following the divergence,

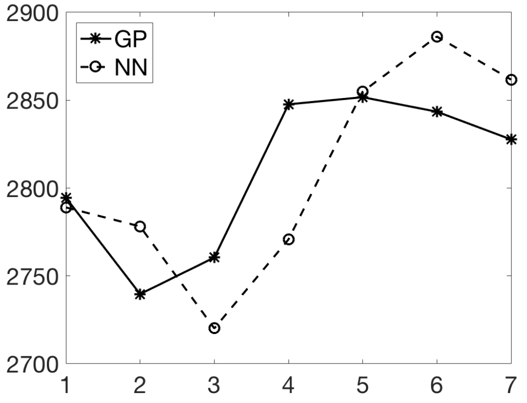

Figure 5.9: Average area $\overline{A}$, defined in (5.3.4), between observations and model propagations in 2000 steps for 500 out-of-sample starting points when $m = 1, \ldots, 7$.

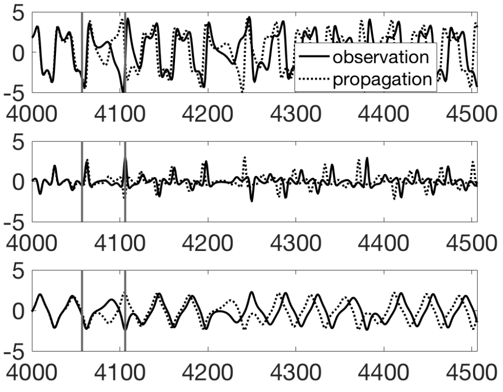

Figure 5.10: Trajectories for observations and NN model propagations when $m = 3$. Starting time point is 3801. The interval $[4057, 4106]$ surrounded by vertical lines is the first window whose average area over the three components exceeds 65.

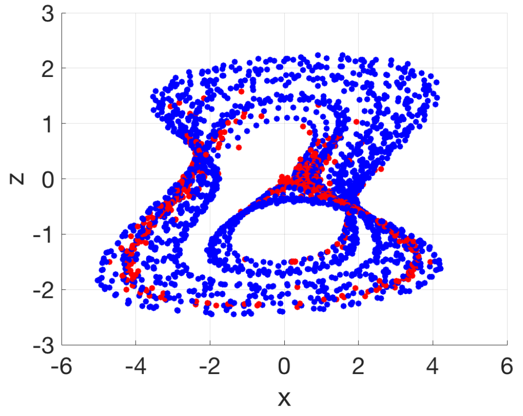

Figure 5.11: Blue is a subset of the observations displayed to represent the attractor; red are points where the average area in a window of length 50 between the observations and GP model propagations with $m = 2$ first exceeds 65 for the 500 pairs of trajectories.

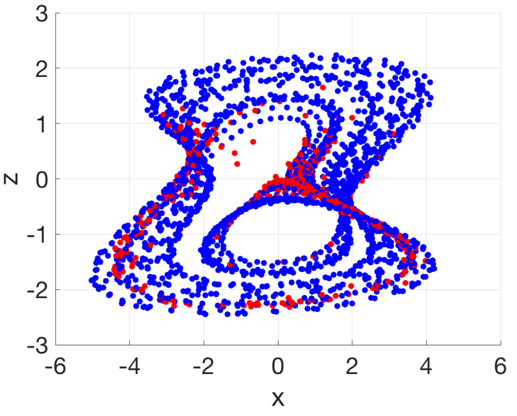

Figure 5.12: Same as Figure 5.11 but using NN model propagations with $m = 3$.

although the model propagations sometimes get back on track with the observations briefly, the tracking becomes noticeably worse. Figures 5.11 and 5.12 show, for GP and NN models, the regions on the state space where a divergence first occurs for the 500 pairs of trajectories. Note that both models suggest the divergence largely concentrates in the region with the most significant nonlinearity, which is approximately $\{(x, y, z) || z| < 0.5, 0 < x < 1\}$.

5.4 Simulation results

[18] shows that the likelihood function for the initial state of a chaotic logistic map exhibits complex and irregular behaviors. In this section, we explore, through simulations, the extent to which similar behaviors of the likelihood functions arise for systems whose realizations are qualitatively similar to our observations. In addition, we compare the long-term predictability of the fitted GP and NN models in Section 5.3 with simulations generated with an electronic noise term based on measurements from a reconstructed physical circuit.

We consider the following state space formulation:

$$u_{t+1} = M(u_t) + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, Q) \tag{5.4.1a}$$

$$w_t = H u_t + \eta_t, \tag{5.4.1b}$$

where (5.4.1a) models the evolution of the underlying states u_t with a stochastic model error ε_t, and (5.4.1b) models the noisy observations w_t with an observational error η_t. In this section, we investigate the effects of both i.i.d. and temporally correlated observational errors. We simulate observations according to (5.4.1) with two choices of the model mapping M. One is obtained from augmenting the nominal model (5.2.1) by including in all three equations of the right-hand-side all monomials up to the third order and two fourth order terms z^4, xz^3. The coefficient for each monomial is estimated via fitting the resulting ODE model to the initial 10,000 observations by minimizing one-step ahead prediction errors. This augmented ODE model leads to smaller and less structured prediction errors than the nominal model. For example, the RMSE of the augmented ODE model for predicting the tail 2000 observations is 2.4×10^{-3} while that of the nominal model is 9×10^{-2}. In this case, the model mapping M_P is obtained by integrating the fitted ODE model using the Runge-Kutta 4th order method, the covariance for the model error is $Q_P = \sigma_\varepsilon^2 I_3$, and the observation mapping is $H_P = I_3$.

For the second choice, we train a feed-forward NN model (2 hidden layers and 10 neurons each layer) with embedding length $m = 3$ using the initial 400,000 observations. The two-layer NN configuration was chosen from multiple configurations since it gave the smallest prediction errors in this case. We use a large training set here because we are interested in obtaining the best approximation we can to the circuit's behavior as a basis for further simulations. In contrast, in Section 5.3, where we used a training set of only 2000 observations, our goal was to show that it is possible to fit an accurate model with a modest sample size. Furthermore, we wanted to make a fair comparison to the GP approach for which maximizing (5.3.2) for larger sample sizes is computationally difficult. The results in Section 5.3 indicate that $m = 3$ appears to perform well in long-term predictions for NN models fitted by minimizing the one-step prediction errors, and by using a large training set, we expect this model to be able to capture most of the dynamics in the data. Note that with this larger training set, we indeed obtain a better fit to the data. For example, the RMSE evaluated using the tail 2000 observations decreases by 8% compared to that of the NN model fitted with only the initial 2000 observations considered in Section 5.3.

Since the NN model predicts based on three previous observations, in this case, we have an augmented state vector $u_t^T = (\widetilde{u}_{t-2}^T, \widetilde{u}_{t-1}^T, \widetilde{u}_t^T)$, a model mapping M_N, and an observational mapping H_N defined as

$$\begin{bmatrix} \widetilde{u}_{t-1} \\ \widetilde{u}_t \\ \widetilde{u}_{t+1} \end{bmatrix} = M_N\left(\begin{bmatrix} \widetilde{u}_{t-2} \\ \widetilde{u}_{t-1} \\ \widetilde{u}_t \end{bmatrix}\right) + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \widetilde{\varepsilon}_t \end{bmatrix}, \qquad \widetilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I_3)$$
$$\widetilde{u}_{t+1} = f_{NN}(\widetilde{u}_{t-2}, \widetilde{u}_{t-1}, \widetilde{u}_t), \qquad H_N = \begin{bmatrix} \mathbf{0} & \mathbf{0} & I_3 \end{bmatrix}, \tag{5.4.2}$$

where $f_{NN}(\cdot)$ is the fitted NN predictor.

5.4.1 Independent and identically distributed observational error

In this subsection, we consider the observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, and investigate the effects of model error on the likelihood functions. When there is no model error, the log-likelihood function of the initial state u_0 satisfies

$$2l_n(u_0) = -\sum_{k=0}^{n-1} \left\| w_k - HM^k(u_0) \right\|^2 / \sigma_\eta^2 - 3n\log(\sigma_\eta^2) - 3n\log(2\pi), \tag{5.4.3}$$

and the profile log-likelihood function obtained by profiling over σ_η^2 satisfies

$$2\widetilde{l}_n(u_0) = -3n\log(\hat{\sigma}_\eta^2) - 3n\log(2\pi) - 3n,$$
$$\hat{\sigma}_\eta^2 = \frac{1}{3n}\sum_{k=0}^{n-1} \left\| w_k - HM^k(u_0) \right\|^2. \tag{5.4.4}$$

In our experiments, the standard deviation for the observational error is $\sigma_\eta = 10^{-3}$, and that for the model error is chosen as small as $\sigma_\varepsilon = 10^{-12}$. We are interested in comparing the likelihoods (5.4.3) and (5.4.4) for observations generated with and without model errors. Specifically, we investigate the likelihood functions of the last element of the initial state while fixing the other elements at their true values. Note that although the dimension of the state vector is three for the ODE model and nine for the NN model, the last element of

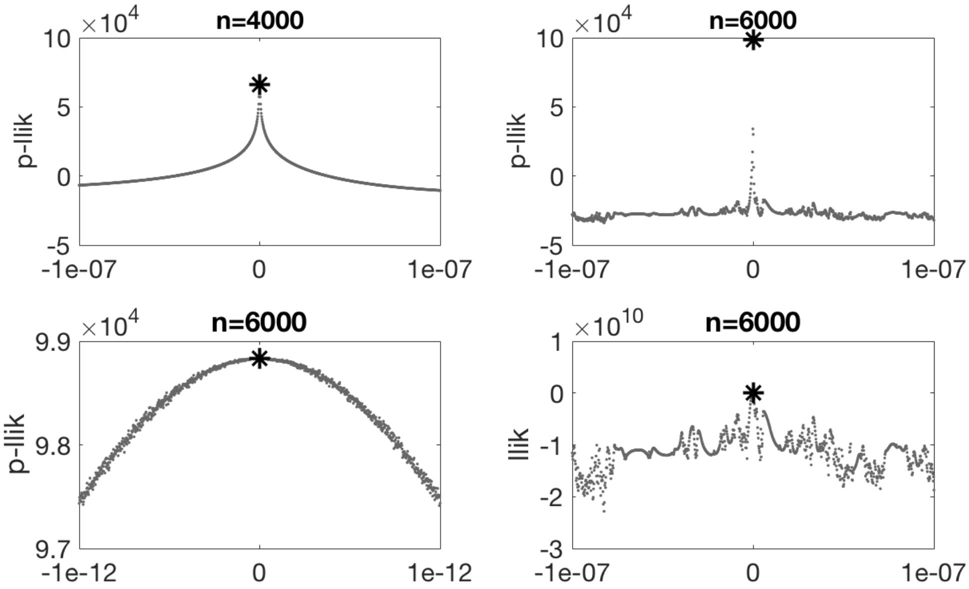

Figure 5.13: Profile log-likelihoods (top row and bottom left panel) and log-likelihood (bottom right panel) of the 3rd element of the initial state for the simulated ODE model M_P with no model error and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$. Labels on the horizontal axes are differences of the simulation points minus the true initial state. Asterisk indicates true initial state.

a state for both models corresponds to the z coordinate of the most recent observation. For both the ODE and NN models, the simulated observations manifest a qualitatively similar attractor as the real data. Figures D.4–D.7 in Supplement D.1 show realizations of the simulation models.

For the ODE model, Figure 5.13 shows, with no model error, the likelihoods of the third element of the initial state with the other two elements fixed at their true values. We observe the same general pattern as in [18]: the likelihood functions on a fixed interval of initial states become more jagged and irregular as we increase the number of observations, and only show some smoothness when focusing on a narrower interval (see, for example, the top right and bottom left panels of Figure 5.13). In addition, much more information about the true initial state can be learned with more observations, since the relative difference of the profile log-likelihoods between the true initial state and the neighboring simulation points becomes more pronounced. Note that for a fixed number of observations, the profile log-likelihood is

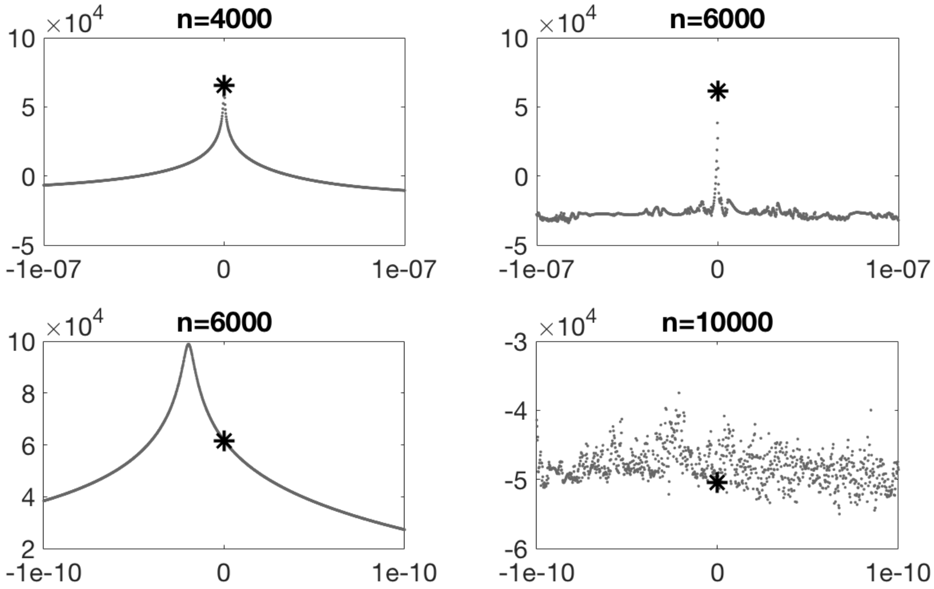

Figure 5.14: Profile log-likelihoods of the 3rd element of the initial state for the simulated ODE model M_P with model error $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I_3)$, $\sigma_\varepsilon = 10^{-12}$ and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$. Labels on the horizontal axes are differences of the simulation points minus the true initial state. Asterisk indicates true initial state.

dramatically sharper than the log-likelihood at the true initial state with σ_η^2 fixed at its true value, and hence contains more information about the truth.

The top right panel of Figure 5.13 seems to indicate the true initial state maximizes the profile log-likelihood. Admittedly, the simulation is based on finitely many discrete points in some neighborhood of the true initial state, and hence does not fully represent the entire likelihood function. However, to add evidence to the remark that the true initial state can be identified as the maximizer of the likelihood function under no model error, in the bottom left panel of Figure 5.13, we evaluate the profile log-likelihood of the third element of the initial state for 1000 points in the interval $(u_0(3) - 10^{-12}, u_0(3) + 10^{-12})$ when $n = 6000$. In this case, the distance between the true initial state and the closest simulation point is on the order of 10^{-15}, and the true initial state still has the largest likelihood in the simulation interval. Note that due to the finite precision of computers, the model being simulated essentially has a discrete state space, and a difference of 10^{-15} is fairly close to the machine precision. As a result, the true initial state may maximize the likelihood for a large enough number of observations given the effectively discrete state space. In Section 5.4.2, we will look more closely into the comparison of the likelihoods of the true initial state and its neighbors for an increasing number of observations.

Figure 5.14 shows the log-likelihoods of the third element of the initial state under a small model error with standard deviation $\sigma_\varepsilon = 10^{-12}$. Similar to the case without model error, we observe wilder behavior of the likelihood functions with more observations. However, increasing the number of observations from 4000 to 6000 no longer helps much in identifying the true initial state. In the top row of Figure 5.14, although more observations make the profile log-likelihood more concentrated around the true initial state, the difference in log-likelihoods between the true initial state and the neighboring points only increases slightly, compared with the sharp increase under no model error (top row of Figure 5.13). The pattern is more obvious on a smaller scale: see the bottom row of Figure 5.14 where the true initial state clearly no longer maximizes the likelihood. When the number of observations increases, we do not gain more information about the true initial state.

Figures 5.15 and 5.16 show, with no model error and a small model error as defined in (5.4.2) respectively, the likelihoods of the 9th element of the initial state with the other elements fixed at their true values for the simulated NN model. Similarly to the ODE model, for a fixed range of the initial states, the likelihoods become more jagged for a larger number of observations, which indicates the NN model captures the chaotic character in the data. Moreover, more observations provide more information for the true initial state under no model error, and when there is even a small model error, the true initial state is no longer the MLE and more observations do not give more information. An interesting feature to note in this case is that on the bottom row of Figure 5.16, the profile log-likelihood attains local maxima at different points when $n = 4000$ and $n = 6000$. It shows that, even with the correct dynamics for the deterministic part of the model, there does not exist a starting point that can track this simulated NN system with even a tiny stochastic model error for

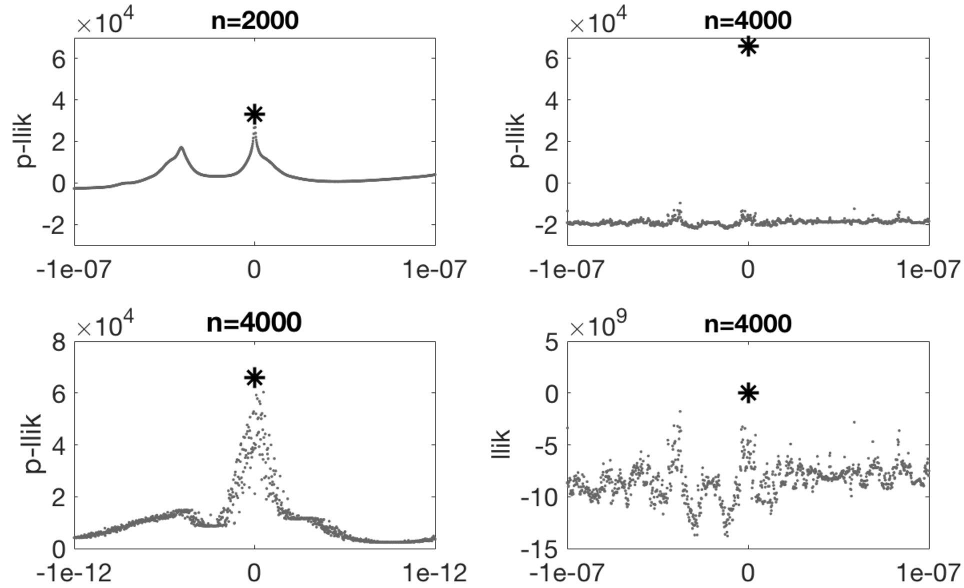

Figure 5.15: Profile log-likelihoods (top row and bottom left panel) and log-likelihood (bottom right panel) of the 9th element of the initial state for the simulated NN model M_N with no model error and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$. Labels on the horizontal axes are differences of the simulation points minus the true initial state. Asterisk indicates true initial state.

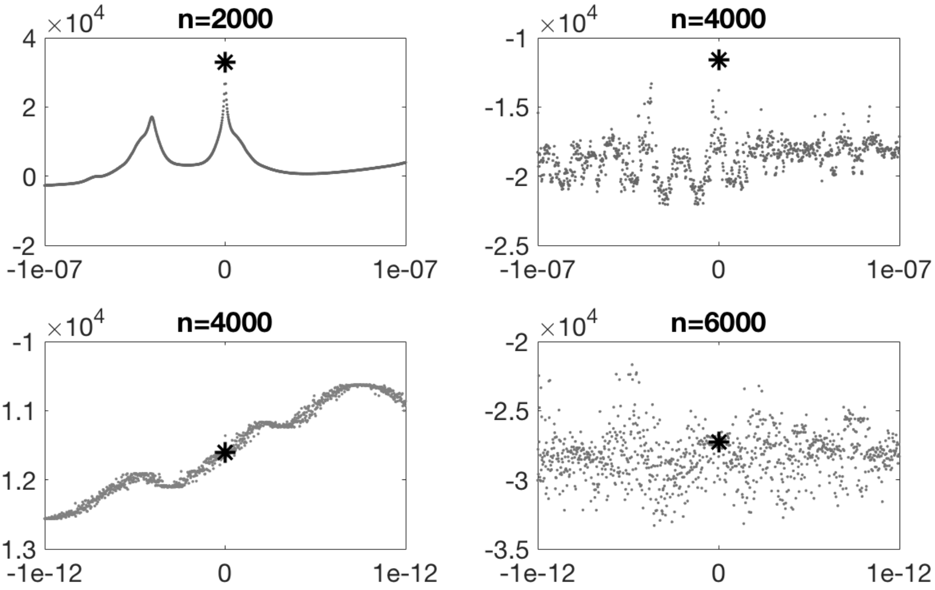

Figure 5.16: Profile log-likelihoods of the 9th element of the initial state for the simulated NN model M_N with model error $\widetilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I_3)$ as defined in (5.4.2), $\sigma_\varepsilon = 10^{-12}$ and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$. Labels on the horizontal axes are differences of the simulation points minus the true initial state. Asterisk indicates true initial state.

an indefinitely long time.

5.4.2 Temporally correlated observational error

In this subsection, we consider observational errors that are temporally correlated with an AR(1) structure. That is, we have $\eta_t = \phi\eta_{t-1} + v_t$ for some $\phi \in \mathbb{R}$ and $v_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_v^2 I)$. A temporally correlated observational error can be used to model dependences in the measurement process resulted from, for example, a limited instantaneous rate of response to changes in signal for the measurement device. When there is no model error, the profile log-likelihood function of the initial state obtained by profiling over σ_η^2 satisfies

$$2\widetilde{l}_n(u_0) = -3n\log(\hat{\sigma}_\eta^2) - \log(|R|) - 3n\log(2\pi) - 3n, \tag{5.4.5}$$

where

$$\hat{\sigma}_\eta^2 = \frac{\mathbf{e}^T R^{-1}\mathbf{e}}{3n}, \quad R = \text{toeplitz}(1, \phi, \ldots, \phi^{n-1}) \otimes I_3, \quad \mathbf{e} = \Big[w_k - HM^k(u_0)\Big]_{k=0}^{n-1}.$$

We take $\phi = 0.9$ and $\sigma_v = 10^{-3}$ as the standard deviation for the innovation. Realizations of the simulated models under this temporally correlated observational error again form an attractor similar to that of the real data and are shown in Figures D.8 and D.9 in Supplement D.1.

Figures 5.13 and 5.15 showed that more information about the true initial state can be learned with more observations under i.i.d. observational error. We now investigate this finding in more detail and compare the effects of different observational error structures. We use the difference in the profile log-likelihood function between the true initial state and the neighboring simulation points as a measure of identifiability of the true initial state. The neighboring simulation points are taken on a regular grid centered at the true initial state for the ODE model, and at the last three elements of the true initial state for the NN model. With this specification, we have a total of 26 simulation points excluding the true initial state. Denote the shortest distance from the simulation points to the true initial state as d. An illustration of this setup of the simulation points is shown in Figure D.10 in Supplement D.1.

Figure 5.17 shows, for the ODE model, the differences in the profile log-likelihood functions (5.4.4) and (5.4.5) under i.i.d. and AR(1) observational errors, respectively. For each observational error type, we experiment with simulation points closer to the true initial state with $d = 2 \times 10^{-15}$ and those further away with $d = 10^{-5}$. In accordance with the feature shown in the top row of Figure 5.13 that, as the number of observations increases, the difference of the profile log-likelihood between the true initial state and the neighboring points increases under both types of observational errors. Note that the linear rate of the increase in Figure 5.17 implies an exponential rate of increase in the likelihood ratio. In addition, as indicated by the larger slope of the dashed lines, the information about the true initial state grows faster in the number of observations for the i.i.d. observations than for the AR(1) observational errors measured with the same set of simulation points.

For both types of observational errors, the difference in the profile log-likelihood is larger between the true initial state and the simulation points further away than those closer. However, this discrepancy between points at different distances to the true initial state diminishes as number of observations increases. This pattern shows that the neighboring points, regardless of their distance to the true initial state, become similarly worse in tracking the observations relative to the true initial state. This is due to the chaotic feature of the system so that even a tiny departure from the true initial state can lead to significant divergence in the propagation with a long enough horizon. In other words, an initial state with a tiny but non-zero departure from the truth, in the long run, does not have much advantage in tracking the observations over an initial state with a large departure. Figure 5.18 shows the differences of the profile log-likelihood functions for the NN model. We observe similar patterns as for the ODE model that, as the number of observations increases, i.i.d. observations provide more information to the true initial state with a higher rate. In addition,

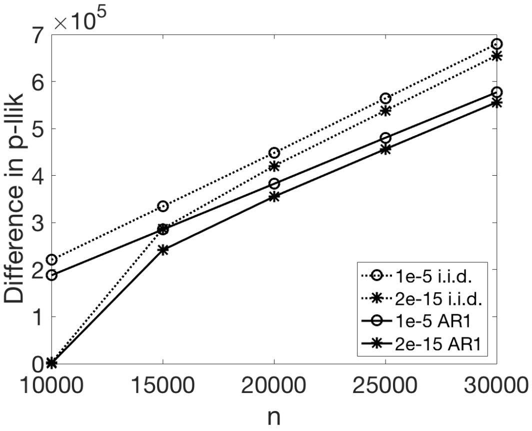

Figure 5.17: Difference in the profile log-likelihood function between the true initial state and the neighboring simulation points for the ODE model with no model error for various n. Numbers in the legend are the distances d between the simulation points and the true initial state.

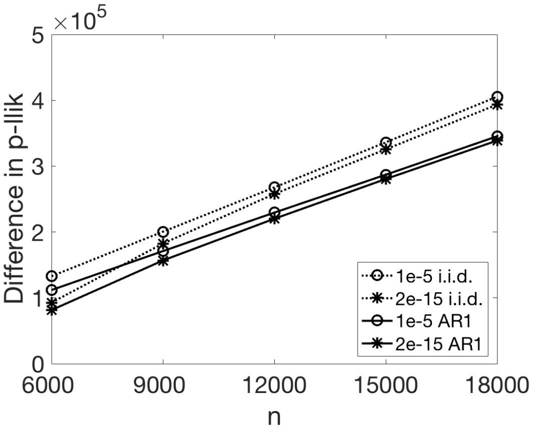

Figure 5.18: Same as Figure 5.17 but for the NN model.

the simulation points with different distances to the true initial state become more similar relative to the truth in terms of their abilities to track the observations.

We note a theoretical result [74, Theorem 3] regarding signal extraction that is relevant to what Figures 5.17 and 5.18 show. That result proves, for an invertible chaotic system Ψ under no model error and Gaussian observational noise, it is impossible to consistently infer any single state from the infinite two-sided observation time series. The result is established based on the existence of homoclinic points. Two points $x_0 \neq x_0'$ constitute a homoclinic pair [e.g., 59, 74, 75] if $\lim_{|n|\to\infty}(1+\alpha)^{|n|}|\Psi^n(x_0) - \Psi^n(x_0')| = 0$ for some $\alpha > 0$. In words, the trajectories of the two distinct points approach each other exponentially fast both forward and backward in time. Consequently, if the observational noise is unbounded, for example Gaussian, there is a positive probability that no matter how many observations one has, the true state x_0 that generates the observations has a lower likelihood than its homoclinic point x_0', and hence can not be inferred from the data. However, for all realizations of the ODE and NN models we have experimented with, we find that the true initial state has the highest likelihood among the simulation points and we appear to gain exponentially growing information about the truth with more observations. One possible resolution for this apparent discrepancy between the theory and our empirical results lies in the relatively small observational noise used in the simulations, which leads to a perhaps tiny, though positive, probability that the homoclinic point gives a higher likelihood than the true initial state. As a result, we do not observe the effects of homoclinic points in our limited number of simulations.

5.4.3 Simulations with electronic noise

In this section, we consider the simulated NN model M_N with no observational error and a model error estimated from a reconstructed physical circuit. The circuit was built using the same components as those in [84], and was fully covered to be thermally controlled at 20 °C. We expect this reconstructed circuit to mimic the circuit system in [84], and thus provide an estimate of the model error encountered in reality. The model error in this case is taken to be the electronic noise in the circuit, which consists mostly of the Johnson (thermal) noise [65]. The Johnson noise is approximately white and has very nearly a Normal distribution [12]. The standard deviation for the Johnson noise was measured for one component in the circuit expected to contribute the most to the entire noise level, and was estimated to be 100µV with a 3dB bandwidth of 1kHz set by the integrator of the circuit. Thus, we maintain that the noise level for the entire circuit in [84] should fall in the range 100µV to 400µV by taking into account the noise contributed from other components and that the experiment in [84] was conducted with a higher room temperature. The simulated NN model with a model error implied by the physical circuit can be regarded as at least a fair approximation to the underlying dynamics of the circuit system. As a result, the performance of two realizations generated from this simulation model in tracking each other sets a benchmark on how well any fitted model can possibly track the observations. In the following, we revisit the fitted GP and NN models in Section 5.3, and evaluate their tracking abilities by comparing with the simulation model under the estimated range of the model error.

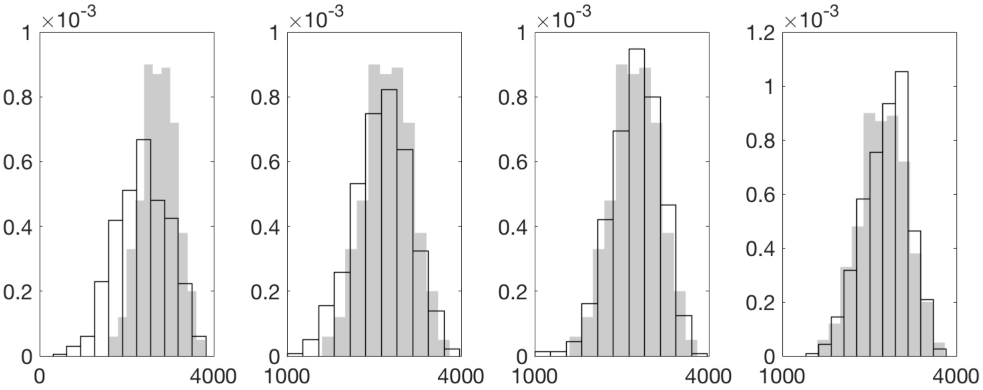

Figure 5.19: Histograms (gray) of average areas between real observations and GP model propagations (embedding length $m = 2$) in 2000 steps for the 500 starting points considered in Section 5.3; histograms (white with solid bar outlines) of average areas between two realizations in 2000 steps from (5.4.2) under model error with standard deviation σ_ε and no observational error for the same 500 starting points. From left to right: $\sigma_\varepsilon = 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}$. Histograms are normalized so that the areas of bars in each histogram sum to one.

Figure 5.19 shows the histograms of the average areas between the fitted GP model propagations and the real observations, and those between two realizations of the simulated model (5.4.2) for the same 500 starting points considered in Section 5.3. The results for the fitted NN model in Section 5.3 are very similar, and a comparison of both models with the simulations is shown in Figure D.11 in Supplement D.1. For a smaller estimate of the model error $\sigma_\varepsilon = 10^{-4}$, the tracking performance of the GP model is clearly inferior to the simulation model. It indicates that the fitted models do not seem to fully capture the dynamics if the model error is indeed this small. However, as the model error increases, we see much better agreement between the fitted models and the simulations. As discussed previously, the smaller end of the model error estimate $\sigma_\varepsilon = 10^{-4}$ is measured with one component of the reconstructed circuit, while the noise level of the entire circuit in [84] is expected to be larger due to contributions from other components and a higher temperature. Consequently, the fitted GP model may approach the limit of how well one can track the system in the presence of a stochastic model error in the circuit. Table 5.1 additionally supports this point by showing the proportions of starting points from which the simulation model produces a smaller tracking area than the fitted models tracking the observations. When the model error $\sigma_\varepsilon = 10^{-4}$ is small, the tracking performance of the fitted models is worse than that of the simulations for around 65% of the starting points. However, for larger model errors, for example when $\sigma_\varepsilon = 3 \times 10^{-4}$, the fitted models track better compared with the simulations for about half of the starting points. Thus, if the true model error has this plausible standard deviation of 3×10^{-4}, then the GP and NN models track the system as well as possible.

σ_ε	10^{-4}	2×10^{-4}	3×10^{-4}	4×10^{-4}
GP	68.8%	56.2%	49.4%	43.2%
NN	65.4%	54.8%	46.8%	46.2%

Table 5.1: Proportions of starting points for which the average area between two trajectories is smaller for the simulations than for the GP and NN model propagations tracking the real observations. The simulation trajectories for each starting point are two realizations from (5.4.2) under model error with standard deviation σ_ε and no observational error.

5.5 SPICE simulation

SPICE (Simulation Program with Integrated Circuit Emphasis) is a general-purpose electrical circuit simulator used in circuit design to verify the circuit operation at transistor level and predict the behavior of the designed circuit [92]. It enables designers to simulate the circuit even before prototyping. We use LTspice [41], a version of the SPICE simulator developed by Linear Technology, to simulate the circuit in [84] in an attempt to obtain a good representation of the underlying dynamics. However, we were not able to obtain simulation results close to the observations. In particular, for all SPICE simulations with various initial values and component values the same as or slightly perturbed from their nominal values in [84], the resulting attractors in some cases do not qualitatively resemble that in Figure 5.2 and the simulations always have systematically much larger, typically twice larger, output ranges for the three voltages than the observations.

There are three major difficulties in carrying out the SPICE simulation to mimic the behavior of the observations. First, the circuit contains potentiometers used to tune the circuit into a parameter space with chaotic behavior, which introduces extra free variables to be adjusted. However, neither the given nominal values of the potentiometers nor the other values tested generate simulations matching the observations. Secondly, the particular multiplier AD534J[1] used in the circuit does not have an associated SPICE model that can be used in the simulation. As a substitute, we used a drop-in replacement multiplier AD734[2] for which an associated SPICE model exists. The SPICE model is configured to have the same transfer function as that of AD534J used in the experiment. However, the model AD734 does differ from the AD534J and has a higher slew rate and a lower noise figure. Finally, the SPICE models for the operational amplifier and multiplier set unrealistically low values for the lead inductance and capacitance, which can give rise to numerical instability in the simulation. Although we were unable to construct a matching model of the circuit using SPICE, the fact that both GP and NN models provide an excellent fit to the data shows that the circuit's behavior is well-described by a dynamical system.

5.6 Discussion

The analysis of observations generated from unknown nonlinear and chaotic dynamical systems poses significant challenges to modeling the underlying dynamics and characterizing the stochastic model and observational errors. We looked into some aspects of these problems with voltage measurement data generated by a laboratory-built electrical circuit. The nominal model of the measurements shows notable deficiencies in capturing the dynamics in the observations, so we turned to GP and NN models that are trained for one-step prediction. With the squared exponential covariance for GP and the feed-forward NN configurations

[1] Product description and details can be found in `http://www.analog.com/en/products/linear-products/analog-multipliers-dividers/ad534`.

[2] Product description and details can be found in `http://www.analog.com/en/products/linear-products/analog-multipliers-dividers/ad734`.

we considered, both models perform similarly in one-step prediction. The prediction error decreases as the embedding length increases. However, good short-term prediction does not necessarily lead to a good representation of the underlying dynamics. As we show for both GP and NN models, the best performance in tracking the observations long-term occurs at a moderately valued embedding length, while a larger embedding length results in quick deterioration in the tracking abilities. The GP and NN models we fitted are both deterministic predictors, while an alternative is to fit a state space model. We experimented with fitting one assuming no model error and a Normal observational error using likelihood method. However, we were not able to obtain results better than the fitted GP and NN predictors. The inclusion of a model error in the state space formulation introduces more parameters to be estimated, and we did not experience success with this approach under a nonlinear dynamics.

The fact that the embedding length leading to the best performance in both one-step prediction and long-term tracking is larger than one for both GP and NN models is an indication that the state of the circuit system may not solely consist of the three voltage measurements at one time point. In fact, the circuit components have different bandwidths, and hence they respond to signal with different rates. For example, the operational amplifiers respond faster than the multipliers. As a result, a single measurement frequency is not able to capture all possible behaviors of every component, especially those occurring at higher frequency. In this case, using an embedding length larger than one, i.e., incorporating previous observations, in the prediction can be seen roughly as a way to at least partially recover the frequency information, which presumably is also part of the state.

We investigated the effects of model and observational errors on the likelihood function of the initial state through simulations. The simulated data were generated by our fitted ODE and NN models, which manifest a low-dimensional attractor and chaotic features similar to the observations. We found that the effects of model and observational errors on inferring the initial state are quite different. In the absence of model error, the true initial state appears to maximize the likelihood function, and the likelihood ratio between the true initial state and the neighboring points grows exponentially in the number of observations. In other words, we have exponentially increasing information about the true initial state with more observations. Temporally correlated observational errors result in a slower rate of increase for the likelihood ratio but preserve the exponential growth. However, in the presence of a model error, even one with a standard deviation as small as 10^{-12}, an increasing number of observations no longer provides unboundedly increasing information about the true initial state through the likelihood function. Moreover, in this case, there does not seem to exist an initial state capable of tracking the observations indefinitely long even with the correct dynamics.

Note that the standard deviation of the model error considered in Section 5.4.1 is significantly smaller than the measured electronic noise in the circuit, yet it is still impossible, through the likelihood function, to identify the initial state and hence track the observations in the long term even with the correct deterministic part of the dynamics. This points out that chaotic systems in reality, which almost always contain stochastic model errors, do not behave as ideally as the deterministic chaos in terms of the identifiability of the initial state under increasing observations. However, our comparisons in Section 5.4.3 show that it may be possible, at least in some cases like ours, to construct deterministic predictors that generate long-term predictions approaching the tracking limit of the system given the existence of a stochastic model error in the dynamics.

Appendices

APPENDIX A

SUPPLEMENT TO CHAPTER 2

A.1 Multiple shooting with zero observability

In this section, we prove that for a class of linear systems, under zero observability, the condition number of the Hessian matrix of augmented Lagrangian has an exponential lower bound. Hence the multiple shooting method is not stable if there are no observations. We consider the model propagation mapping to be time independent, that is, $M(x_j) = Ax_j$, and $B_j = \mathbf{0}$ for $0 \leq j \leq N$. We assume A has at least one real eigenvalue with modulus strictly larger than 1. With this model specification, J_i are identical for all $1 \leq i \leq d$ and so are Λ_i. For simplicity, we denote them respectively as J_1 and Λ_1 for $1 \leq i \leq d$. The expanded forms of J_1 and Λ_1 are

$$J_1 = \begin{bmatrix} A^TQ^{-1}A & -A^TQ^{-1} & & \mathbf{0} \\ -Q^{-1}A & A^TQ^{-1}A+Q^{-1} & \ddots & \\ & \ddots & \ddots & \\ & \ddots & A^TQ^{-1}A+Q^{-1} & -A^TQ^{-1} \\ \mathbf{0} & & -Q^{-1}A & Q^{-1} \end{bmatrix},$$

$$\Lambda_1 = \begin{bmatrix} \mathbf{0} & I \\ -QA^{-T}Q^{-1}A & A+QA^{-T}Q^{-1} \\ \vdots & \vdots \\ L^{(P_i-1)}_{P_{i+1}}(x_{P_i-1}, x_{P_i}) & L^{(P_i)}_{P_{i+1}}(x_{P_i-1}, x_{P_i}) \end{bmatrix}.$$

For $p = P_i, P_i - 1$, adapting the optimality recursions (2.2.7) to the linear system under consideration and applying the chain rule, we have that the recursion of $L^{(p)}_{P_i+j}$ for $0 \leq j \leq k-1$ is

$$L^{(p)}_{P_i+j+1} = (A + QA^{-T}Q^{-1})L^{(p)}_{P_i+j} - QA^{-T}Q^{-1}AL^{(p)}_{P_i+j-1}. \tag{A.1.1}$$

Denote L_1 to be the last two block rows of Λ_1.

Lemma A.1.1. *Denote* $\hat{\Lambda} = \Lambda_1 \begin{bmatrix} I \\ A \end{bmatrix}$. *Then,* $J_1\hat{\Lambda} = \mathbf{0}$.

Proof. We first prove that for $1 \leq j \leq k$, the jth block of $\hat{\Lambda}$ is $(\hat{\Lambda})_j = A^j$ by induction. It is evident for $j = 1, 2$. Suppose it is true for all $j \leq j_0$, $2 \leq j_0 \leq k-1$. Then by recursion (A.1.1),

$$\begin{aligned}
(\hat{\Lambda})_{j_0+1} &= L^{(P_i-1)}_{P_i+j_0} + L^{(P_i)}_{P_i+j_0}A \\
&= (A + QA^{-T}Q^{-1})(L^{(P_i-1)}_{P_i+j_0-1} + L^{(P_i)}_{P_i+j_0-1}A) \\
&- (QA^{-T}Q^{-1}A)(L^{(P_i-1)}_{P_i+j_0-2} + L^{(P_i)}_{P_i+j_0-2}A) \\
&= (A + QA^{-T}Q^{-1})(\hat{\Lambda})_{j_0} - (QA^{-T}Q^{-1}A)(\hat{\Lambda})_{j_0-1} \\
&= A^{j_0+1}.
\end{aligned}$$

A direct multiplication completes the proof. □

Proposition A.1.2. *Let* $|\lambda| > 1$, $\lambda \in \mathbb{R}$ *be an eigenvalue of* A. *Denote* $\lambda_k = \lambda^{k-1}$. *Then, for the linear system under consideration, we have that*

$$\kappa\Big(\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)\Big) \geq \frac{\lambda_{min}(Q_B^{-1})}{\mu}|\lambda_k|^{2(d-1)}.$$

Proof. For any $s = (s_1, \ldots, s_{2d+1}) \in \mathbb{R}^{(2d+1)J}$, denote $\hat{s}_0 = s_1$, $\hat{s}_i = (s_{2i}, s_{2i+1})$ for $1 \leq i \leq d$. Then from Theorem 2.2.2 (b) we have that

$$\begin{aligned}
s^T\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)s &= s_1^T\Lambda_0^T J_0\Lambda_0 s_1 + \sum_{i=1}^{d}\hat{s}_i^T\Lambda_1^T J_1\Lambda_1\hat{s}_i \qquad \text{(A.1.2)} \\
&+ \mu\|\hat{s}_1 - L_0 s_1\|^2 + \mu\sum_{i=2}^{d}\|\hat{s}_i - L_1\hat{s}_{i-1}\|^2.
\end{aligned}$$

Consider $s = (s_1, \ldots, s_{2d+1}) \in \mathbb{R}^{(2d+1)J}$ such that $s_1 = 0$, $s_{2i} = \lambda_k^{i-1}s_2$, $s_{2i+1} = \lambda s_{2i}$ for

$1 \leq i \leq d$, and let $\|s_2\| = 1$ be the eigenvector of A corresponding to λ, that is, $As_2 = \lambda s_2$. Then $\hat{s}_i = \begin{bmatrix} I \\ A \end{bmatrix} s_{2i}$ for $1 \leq i \leq d$, which gives that

$$\hat{s}_i^T \Lambda_1^T J_1 \Lambda_1 \hat{s}_i \;=\; s_{2i}^T \hat{\Lambda}^T J_1 \hat{\Lambda} s_{2i} = 0, \tag{A.1.3}$$

where the last equality follows from Lemma A.1.1.

Since L_1 consists of the last two block rows of Λ_1, we have that $L_1 \begin{bmatrix} I \\ A \end{bmatrix} = \begin{bmatrix} A^{k-1} \\ A^k \end{bmatrix}$. Hence by the definition of s for $2 \leq i \leq d$, we obtain that

$$\hat{s}_i - L_1 \hat{s}_{i-1} \;=\; \begin{bmatrix} I \\ A \end{bmatrix} s_{2i} - \begin{bmatrix} A^{k-1} \\ A^k \end{bmatrix} s_{2(i-1)} = \mathbf{0}. \tag{A.1.4}$$

Using (A.1.3) and (A.1.4) in (A.1.2), we obtain that

$$s^T \nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu) s = \mu \|\hat{s}_1\|^2 \leq \mu(\|s_2\|^2 + |\lambda|^2 \|s_2\|^2) = \mu(1 + |\lambda|^2).$$

From the definition of s, we have that

$$\|s\|^2 = \sum_{i=1}^{d} \|s_{2i}\|^2 + \|s_{2i+1}\|^2 = (1 + |\lambda|^2) \sum_{i=1}^{d} |\lambda_k|^{2(i-1)} \geq (1 + |\lambda|^2)|\lambda_k|^{2(d-1)}.$$

Hence we have that

$$\begin{aligned} \lambda_{min}(\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)) &\leq \frac{s^T \nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu) s}{\|s\|^2} \\ &\leq \mu |\lambda_k|^{-2(d-1)}. \end{aligned} \tag{A.1.5}$$

On the other hand, let $t = (t_1, \ldots, t_{2d+1}) \in \mathbb{R}^{(2d+1)J}$ be such that $\|t_1\| = 1$ and $t_i = 0$ for all $2 \leq i \leq 2d + 1$. J_0 differs from J_1 by only the (1,1)th block element so that

$(J_0)_{(1,1)} = (J_1)_{(1,1)} + Q_B^{-1}$. Then

$$\begin{aligned} t^T \nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu) t &= t_1^T \Lambda_0^T J_0 \Lambda_0 t_1 + \mu t_1^T L_0^T L_0 t_1 \\ &\geq t_1^T \Lambda_0^T J_0 \Lambda_0 t_1 \geq \lambda_{min}(Q_B^{-1}). \end{aligned}$$

Hence we have that

$$\begin{aligned} \lambda_{max}(\nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu)) &\geq \frac{t^T \nabla_x^2 L_A(x^*, \lambda^*, \psi^*, \mu) t}{\|t\|^2} \\ &\geq \lambda_{min}(Q_B^{-1}). \end{aligned} \tag{A.1.6}$$

Combining equation (A.1.5) and (A.1.6) completes the proof. □

A.2 Single shooting condition number

In this section, we prove that for a certain class of linear systems that satisfy the observability condition, the condition number of the Hessian matrix for the single shooting function (2.1.7) has an exponential lower bound in N. Hence the single shooting method is not stable for this class of systems.

We consider linear time-independent systems such that $M(x_i) = Ax_i$ and $H(x_i) = Bx_i$. Denote $C_1 = QA^{-T}Q^{-1} + A + QA^{-T}B^TR^{-1}B$ and $C_2 = QA^{-T}Q_B^{-1} + A + QA^{-T}B^TR^{-1}B$. We have the following.

Proposition A.2.1. *For linear systems satisfying*

(a) $C_1C_2 - I = C_2^2$,

(b) there exist eigenvalues λ_1 *and* λ_2 *of* C_2 *such that* $|\lambda_1| > 1$ *and* $|\lambda_1| > |\lambda_2| \neq 0$,

(c) $QA^{-T}Q^{-1}A = I_J$.

We have

$$\kappa\Big(\nabla^2_{x_0}\hat{\Gamma}(x_0^*)\Big) \geq \begin{cases} \frac{C}{N}\left|\frac{\lambda_1}{\lambda_2}\right|^{2(N-1)}, & |\lambda_2| \geq 1 \\ \frac{C}{N}|\lambda_1|^{2(N-1)}, & |\lambda_2| < 1 \end{cases}$$

for some constant $C > 0$*, where* x_0^* *is the first component of a local minimizer of* $\Gamma(x_{0:N})$ (2.1.3).

Note: At the end of this section, we give an example of a linear system satisfying conditions (a)–(c) with observation matrix B being full rank, namely, with full observability.

Proof. It is shown in [7, Theorem 3] that x_0^* is a local minimizer of $\hat{\Gamma}(x_0)$ and that

$$\nabla_{x_0}\hat{\Gamma}(x_0^*) = \theta_0(x_0^*, \lambda_1) + \sum_{j=1}^{N-1} {L_j^{(0)}}^T \theta_j(\lambda_{j-1}, \lambda_j, \lambda_{j+1}) + {L_N^{(0)}}^T \theta_N(\lambda_{N-1}, \lambda_N), \quad \text{(A.2.1)}$$

where $L_j^{(0)}$, $0 \leq j \leq N$ are as defined in Definition 2.2.1(b).

Applying the chain rule and the optimality conditions (2.1.4), (2.1.5), and (2.1.6) to (A.2.1), we obtain that the Hessian matrix for the single shooting function (2.1.7) is

$$\nabla^2_{x_0}\hat{\Gamma}(x_0^*) = \Lambda_s^T J_s \Lambda_s, \quad \text{(A.2.2)}$$

where Λ_s is $(N+1)J \times J$ dimensional and J_s is $(N+1)J \times (N+1)J$ dimensional. They are defined as

$$\begin{aligned} \Lambda_s^T &= \begin{bmatrix} {L_0^{(0)}}^T & {L_1^{(0)}}^T & \dots & {L_N^{(0)}}^T \end{bmatrix} \\ J_s &= \begin{bmatrix} Q_B^{-1} + B^T R^{-1} B + A^T Q^{-1} A & -A^T Q^{-1} & & & \mathbf{0} \\ -Q^{-1}A & C_3 + A^T Q^{-1} A & \ddots & & \\ & \ddots & \ddots & & \\ & \ddots & C_3 + A^T Q^{-1} A & -A^T Q^{-1} & \\ \mathbf{0} & & -Q^{-1}A & C_3 & \end{bmatrix}, \end{aligned}$$

where $C_3 := Q^{-1} + B^T R^{-1} B$. Denote $d_j(x) = \begin{bmatrix} L_j^{(0)} x \\ L_{j-1}^{(0)} x \end{bmatrix}$, for $1 \leq j \leq N$. Then, for $1 \leq j \leq N-1$, from the recursion for the derivatives (2.3.8) and (2.3.9), we have that

$$\begin{aligned} d_{j+1}(x) &= \begin{bmatrix} QA^{-T}Q^{-1} + A + QA^{-T}B^TR^{-1}B & -QA^{-T}Q^{-1}A \\ I_J & \mathbf{0} \end{bmatrix} d_j(x) \\ &= \begin{bmatrix} QA^{-T}Q^{-1} + A + QA^{-T}B^TR^{-1}B & -I_J \\ I_J & \mathbf{0} \end{bmatrix} d_j(x) \\ &= \begin{bmatrix} C_1 & -I_J \\ I_J & \mathbf{0} \end{bmatrix} d_j(x) := Dd_j(x), \end{aligned}$$

and

$$\begin{aligned} d_1(x) &= \begin{bmatrix} QA^{-T}B^TR^{-1}B + A + QA^{-T}Q_B^{-1} \\ I_J \end{bmatrix} x \\ &= \begin{bmatrix} C_2 \\ I_J \end{bmatrix} x := \hat{C}_2 x. \end{aligned}$$

For any eigenvector v of C_2 with corresponding eigenvalue λ, we have from condition (a) that $Dd_1(v) = \lambda d_1(v)$. Hence for $1 \leq j \leq N$, we have that $d_j(v) = \lambda^{j-1} d_1(v)$. Denoting $\widetilde{Q} = (I, -A)^T Q^{-1} (I, -A)$ and using (A.2.2), we have that

$$\begin{aligned} v^* \nabla_{x_0}^2 \hat{\Gamma}(x_0^*) v &= v^* \Lambda_s^T J_s \Lambda_s v \\ &= v^* Q_B^{-1} v + \sum_{j=0}^{N} (L_j^{(0)} v)^* B^T R^{-1} B (L_j^{(0)} v) \\ &\quad + \sum_{j=0}^{N-1} (L_{j+1}^{(0)} v - AL_j^{(0)} v)^* Q^{-1} (L_{j+1}^{(0)} v - AL_j^{(0)} v) \end{aligned}$$

$$
\begin{aligned}
&= v^* Q_B^{-1} v + \sum_{j=0}^{N} (L_j^{(0)} v)^* B^T R^{-1} B (L_j^{(0)} v) + \sum_{j=1}^{N} d_j(v)^* \widetilde{Q} d_j(v) \\
&= v^* Q_B^{-1} v + \sum_{j=0}^{N} (L_j^{(0)} v)^* B^T R^{-1} B (L_j^{(0)} v) + v^* \hat{C}_2^T \widetilde{Q} \hat{C}_2 v \sum_{j=1}^{N} |\lambda|^{2(j-1)},
\end{aligned}
$$

where $\hat{C}_2^T \widetilde{Q} \hat{C}_2 = (Q_B^{-1} + B^T R^{-1} B)^T A^{-1} Q A^{-T} (Q_B^{-1} + B^T R^{-1} B)$ is positive definite.

Let v_1 and v_2 be eigenvectors of C_2 corresponding respectively to λ_1 and λ_2 as defined in condition (b), and $\|v_1\| = 1$, $\|v_2\| = 1$. Then we have

$$
\begin{aligned}
\lambda_{max}(\nabla_{x_0}^2 \hat{\Gamma}(x_0^*)) &\geq \frac{v_1^* \nabla_{x_0}^2 \hat{\Gamma}(x_0^*) v_1}{\|v_1\|^2} \\
&\geq v_1^* \hat{C}_2^T \widetilde{Q} \hat{C}_2 v_1 \sum_{j=1}^{N} |\lambda_1|^{2(j-1)} \\
&\geq \lambda_{min}(\hat{C}_2^T \widetilde{Q} \hat{C}_2) |\lambda_1|^{2(N-1)}
\end{aligned}
\tag{A.2.3}
$$

and

$$
\begin{aligned}
\lambda_{min}(\nabla_{x_0}^2 \hat{\Gamma}(x_0^*)) &\leq \frac{v_2^* \nabla_{x_0}^2 \hat{\Gamma}(x_0^*) v_2}{\|v_2\|^2} \\
&\leq \lambda_{max}(Q_B^{-1}) + \lambda_{max}(B^T R^{-1} B) \|d_1(v_2)\|^2 \sum_{j=1}^{N} |\lambda_2|^{2(j-1)} \\
&+ \lambda_{max}(B^T R^{-1} B) + v_2^* \hat{C}_2^T \widetilde{Q} \hat{C}_2 v_2 \sum_{j=1}^{N} |\lambda_2|^{2(j-1)} \\
&\leq 2U + 2U \sum_{j=1}^{N} |\lambda_2|^{2(j-1)} \\
&\leq \begin{cases} 4UN|\lambda_2|^{2(N-1)}, & |\lambda_2| \geq 1 \\ 4UN, & |\lambda_2| < 1 \end{cases},
\end{aligned}
\tag{A.2.4}
$$

where

$$U = \max\left(\lambda_{max}(Q_B^{-1}), \lambda_{max}(B^TR^{-1}B)\lambda_{max}(\hat{C}_2^T\hat{C}_2), \lambda_{max}(B^TR^{-1}B), \lambda_{max}(\hat{C}_2^T\widetilde{Q}\hat{C}_2)\right).$$

Equations (A.2.3) and (A.2.4) give that

$$\kappa\left(\nabla_{x_0}^2\hat{\Gamma}(x_0^*)\right) \geq \begin{cases} \frac{\lambda_{min}(\hat{C}_2^T\widetilde{Q}\hat{C}_2)}{4UN}\left|\frac{\lambda_1}{\lambda_2}\right|^{2(N-1)}, & |\lambda_2| \geq 1 \\ \frac{\lambda_{min}(\hat{C}_2^T\widetilde{Q}\hat{C}_2)}{4UN}|\lambda_1|^{2(N-1)}, & |\lambda_2| < 1 \end{cases}.$$

Letting $C = \frac{\lambda_{min}(\hat{C}_2^T\widetilde{Q}\hat{C}_2)}{4U}$ completes the proof. □

Consider an example for which $Q = A = diag(2, 1, \ldots, 1)$ for all $0 \leq j \leq N-1$, $Q_B = diag(4, \frac{3}{2}, \ldots, \frac{3}{2})$, and $B^TR^{-1}B = diag(\frac{7}{4}, \frac{4}{3}, \ldots, \frac{4}{3})$ such that B is full rank. Then, $C_1 = diag(\frac{17}{4}, \frac{10}{3}, \ldots, \frac{10}{3})$ and $C_2 = diag(4, 3, \ldots, 3)$ so that all three conditions in Proposition A.2.1 are satisfied. For this example, even with full observability, single shooting is not stable.

APPENDIX B

SUPPLEMENT TO CHAPTER 3

B.1 Proofs of results in Sections 3.2 and 3.3

B.1.1 Proof of Proposition 3.2.4

With dynamic programming, the "cost-to-go" value functions for a problem started at k with state x_k, $J_k(x_k)$, satisfy

$$\begin{aligned} J_{n_2}(x_{n_2}) &= (x_{n_2} - d_{n_2})^T Q_{n_2}(x_{n_2} - d_{n_2}) \\ J_k(x_k) &= \min_{v_k} (x_k - d_k)^T Q_k (x_k - d_k) + v_k^T \hat{R}_k v_k + J_{k+1}(A_k x_k + \hat{B}_k v_k + f_k) \end{aligned}$$

for $n_1 \leq k \leq n_2 - 1$. We claim that

$$J_k(x_k) = x_k^T K_k x_k - 2\sum_{i=k}^{n_2} d_i^T M_i^k x_k - 2\sum_{i=k}^{n_2-1} f_i^T S_i^k x_k + T_k, \qquad n_1 \leq k \leq n_2, \quad \text{(B.1.1)}$$

where T_k is some constant matrix.

Now we prove (B.1.1) by reverse induction and show that (3.2.6) holds whenever (B.1.1) holds at $k+1$. When $k = n_2$, (B.1.1) holds with $K_{n_2} = Q_{n_2}$ by definition. Suppose (B.1.1) holds for $J_{k+1}(x_{k+1})$ for some $n_1 \leq k \leq n_2 - 1$. Then, replacing the induction hypothesis formula in the cost-to-go function, we obtain

$$\begin{aligned} J_k(x_k) &= \min_{v_k} \left\{ v_k^T W_k v_k + 2 r_k^T v_k \right\} + x_k^T \left(A_k^T K_{k+1} A_k + Q_k \right) x_k \\ &\quad -2 \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} A_k x_k - 2 d_k^T Q_k x_k \\ &\quad -2 \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} A_k x_k + 2 f_k^T K_{k+1} A_k x_k + T_k, \end{aligned}$$

where

$$\begin{aligned}
W_k &= \hat{R}_k + \hat{B}_k^T K_{k+1} \hat{B}_k, && \text{(B.1.2)} \\
r_k^T &= x_k^T A_k^T K_{k+1} \hat{B}_k - \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} \hat{B}_k - \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} \hat{B}_k + f_k^T K_{k+1} \hat{B}_k && \text{(B.1.3)}
\end{aligned}$$

The optimal control law, which is the solution of the preceding optimization problem, is hence

$$\begin{aligned}
v_k^*(x_k) &= -W_k^{-1} r_k \\
&= L_k x_k + W_k^{-1} \sum_{i=k+1}^{n_2} \hat{B}_k^T \left(M_i^{k+1}\right)^T d_i \\
&\quad + W_k^{-1} \sum_{i=k+1}^{n_2-1} \hat{B}_k^T \left(S_i^{k+1}\right)^T f_i - W_k^{-1} \hat{B}_k K_{k+1} f_k.
\end{aligned}$$

Therefore (3.2.6) holds true at k.

Substituting $v_k^* = -W_k^{-1} r_k$, we obtain that $v_k^{*T} W_k v_k^* + 2 r_k^T v_k^* = -r_k^T W_k^{-1} r_k$. As a result,

$$\begin{aligned}
J_k(x_k) = & -r_k^T W_k^{-1} r_k + x_k^T \left(A_k^T K_{k+1} A_k + Q_k\right) x_k \\
& - 2 \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} A_k x_k - 2 d_k^T Q_k x_k \\
& - 2 \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} A_k x_k + 2 f_k^T K_{k+1} A_k x_k + T_k.
\end{aligned} \tag{B.1.4}$$

Substituting the expression of r_k defined in (B.1.3), we have that, up to a constant term,

$$\begin{aligned}
&- r_k^T W_k^{-1} r_k + x_k^T \left(A_k^T K_{k+1} A_k + Q_k \right) x_k \\
&= x_k^T \left(A_k^T (K_{k+1} - K_{k+1} \hat{B}_k W_k^{-1} \hat{B}_k^T K_{k+1}) A_k + Q_k \right) x_k \\
&\quad + 2 \left(\sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} \hat{B}_k + \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} \hat{B}_k - f_k^T K_{k+1} \hat{B}_k \right) W_k^{-1} \hat{B}_k K_{k+1} A_k x_k \\
&= x_k^T K_k x_k - 2 \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} \hat{B}_k L_k x_k - 2 \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} \hat{B}_k L_k x_k + 2 f_k^T K_{k+1} \hat{B}_k L_k x_k,
\end{aligned} \tag{B.1.5}$$

by (3.2.7b) and (3.2.7d). Now we substitute (B.1.5) back into (B.1.4) and have that

$$\begin{aligned}
J_k(x_k) &= x_k^T K_k x_k - 2 \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} \hat{B}_k L_k x_k - 2 \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} \hat{B}_k L_k x_k + 2 f_k^T K_{k+1} \hat{B}_k L_k x_k \\
&\quad - 2 \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} A_k x_k - 2 d_k^T Q_k x_k - 2 \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} A_k x_k + 2 f_k^T K_{k+1} A_k x_k + T_k \\
&\overset{(3.2.7e)}{=} x_k^T K_k x_k - 2 \sum_{i=k+1}^{n_2} d_i^T M_i^{k+1} D_k x_k - 2 d_k^T Q_k x_k \\
&\quad - 2 \sum_{i=k+1}^{n_2-1} f_i^T S_i^{k+1} D_k x_k + 2 f_k^T K_{k+1} D_k x_k + T_k \\
&= x_k^T K_k x_k - 2 \sum_{i=k}^{n_2} d_i^T M_i^k x_k - 2 \sum_{i=k}^{n_2-1} f_i^T S_i^k x_k + T_k,
\end{aligned}$$

by (3.2.7f) and (3.2.7g). This proves the induction hypothesis at (B.1.1) at step k. Since we proved above that (3.2.6) holds true at k, this completes the induction step and proves that both (B.1.1) and (3.2.6) hold for all k.

B.1.2 Proof of Proposition 3.2.6

The recursion (3.2.5c) and optimal control law (3.2.6) imply that x_k^* has the form of (3.2.8) for some C_i^k and F_i^k. When $k = n_1 + 1$, we have that

$$\begin{aligned}
x_{n_1+1}^* &= A_{n_1} x_{n_1} + \hat{B}_{n_1} v_{n_1}^* + f_{n_1} \\
&= D_{n_1} x_{n_1} + E_{n_1} \sum_{i=n_1+1}^{n_2} \left(M_i^{n_1+1}\right)^T d_i + E_{n_1} \sum_{i=n_1}^{n_2-1} \left(S_i^{n_1+1}\right)^T f_i - E_{n_1} K_{n_1+1} f_{n_1} + f_{n_1}.
\end{aligned} \tag{B.1.6}$$

Applying recursion (3.2.5c) and the optimal control law (3.2.6) gives

$$\begin{aligned}
x_{k+1}^* \;=\; & \left(\prod_{i=n_1}^{k} D_i\right) x_{n_1} + \sum_{i=n_1+1}^{n_2} D_k C_i^k d_i + \sum_{i=n_1}^{n_2-1} D_k F_i^k f_i \\
& + E_k \sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + E_k \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i - E_k K_{k+1} f_k + f_k.
\end{aligned}$$

Combining with (B.1.6), we obtain the recursions

$$\begin{aligned}
C_i^{k+1} &= \begin{cases} D_k C_i^k, & n_1 + 1 \le i \le k \\ D_k C_i^k + E_k \left(M_i^{k+1}\right)^T, & k+1 \le i \le n_2 \end{cases} \\
C_i^{n_1+1} &= E_{n_1} \left(M_i^{n_1+1}\right)^T, \qquad n_1 + 1 \le i \le n_2,
\end{aligned} \tag{B.1.7}$$

and

$$F_i^{k+1} = \begin{cases} D_k F_i^k, & n_1 \le i \le k-1 \\ D_k F_i^k - E_k K_{k+1} + I, & i = k \\ D_k F_i^k + E_k \left(S_i^{k+1}\right)^T, & k+1 \le i \le n_2 - 1 \end{cases} \tag{B.1.8}$$

$$F_i^{n_1+1} = \begin{cases} E_{n_1} \left(S_i^{n_1+1}\right)^T, & n_1+1 \le i \le n_2 - 1 \\ -E_{n_1} K_{n_1+1} + I, & i = n_1. \end{cases}$$

Now we prove that C_i^k and F_i^k defined in (3.2.9) satisfy the recursions (B.1.7) and (B.1.8), respectively.

(a) Proof that C_i^k defined in (3.2.9) satisfies (B.1.7).

When $k = n_1 + 1$, since $i \ge n_1 + 1$, we have that

$$C_i^{n_1+1} = \sum_{s=n_1}^{n_1} \left(\prod_{l=s+1}^{n_1} D_l \right) E_s \left(M_i^{s+1}\right)^T = E_{n_1} \left(M_i^{n_1+1}\right)^T,$$

which satisfies (B.1.7).

When $k > n_1 + 1$, if $i \le k$, then $i \le k+1$, then we have that

$$\begin{aligned} C_i^{k+1} &= \sum_{s=n_1}^{i-1} \left(\prod_{l=s+1}^{k} D_l \right) E_s \left(M_i^{s+1}\right)^T \\ &= D_k \sum_{s=n_1}^{i-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(M_i^{s+1}\right)^T \\ &= D_k C_i^k, \end{aligned}$$

and if $i \geq k+1$, then $i \geq k$, so there follows that

$$\begin{aligned} C_i^{k+1} &= \sum_{s=n_1}^{k} \left(\prod_{l=s+1}^{k} D_l \right) E_s \left(M_i^{s+1} \right)^T \\ &= D_k \sum_{s=n_1}^{k-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(M_i^{s+1} \right)^T + E_k \left(M_i^{k+1} \right)^T \\ &= D_k C_i^k + E_k \left(M_i^{k+1} \right)^T, \end{aligned}$$

which both satisfy (B.1.7).

(b) Proof that F_i^k defined in (3.2.9) satisfies (B.1.8).

When $k = n_1 + 1$, if $i = n_1$, we have from (3.2.9) that

$$F_{n_1}^{n_1+1} = I - E_{n_1} K_{n_1+1},$$

and if $i \geq n_1 + 1 = k$, there follows that

$$F_i^{n_1+1} = \sum_{s=n_1}^{n_1} \left(\prod_{l=s+1}^{n_1} D_l \right) E_s \left(S_i^{s+1} \right)^T = E_{n_1} \left(S_i^{n_1+1} \right)^T,$$

which satisfies (B.1.8).

When $k > n_1 + 1$, if $i \leq k-1$, then $i \leq k \leq k+1$, and then we have that

$$\begin{aligned} F_i^{k+1} &= \sum_{s=n_1}^{i-1} \left(\prod_{l=s+1}^{k} D_l \right) E_s \left(S_i^{s+1} \right)^T + \left(\prod_{l=i+1}^{k} D_l \right) (I - E_i K_{i+1}) \\ &= D_k \left(\sum_{s=n_1}^{i-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(S_i^{s+1} \right)^T + \left(\prod_{l=i+1}^{k-1} D_l \right) (I - E_i K_{i+1}) \right) \\ &= D_k F_i^k. \end{aligned}$$

If $i = k$, then $k+1 \geq i+1$, and hence

$$\begin{aligned} F_i^{k+1} &= \sum_{s=n_1}^{i-1} \left(\prod_{l=s+1}^{k} D_l \right) E_s \left(S_i^{s+1} \right)^T + \left(I - E_i K_{i+1} \right) \\ &= D_k \sum_{s=n_1}^{i-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(S_i^{s+1} \right)^T + \left(I - E_k K_{k+1} \right) \\ &= D_k F_i^k + \left(I - E_k K_{k+1} \right), \end{aligned}$$

and if $i \geq k+1$, then $i \geq k$, and we have that

$$\begin{aligned} F_i^{k+1} &= \sum_{s=n_1}^{k} \left(\prod_{l=s+1}^{k} D_l \right) E_s \left(S_i^{s+1} \right)^T \\ &= D_k \sum_{s=n_1}^{k-1} \left(\prod_{l=s+1}^{k-1} D_l \right) E_s \left(S_i^{s+1} \right)^T + E_k \left(S_i^{k+1} \right)^T \\ &= D_k F_i^k + E_k \left(S_i^{k+1} \right)^T, \end{aligned}$$

which all satisfy (B.1.8).

B.1.3 Proof of Lemma 3.2.9

For $n_1 \leq k \leq n_2 - 1$, Definition 3.2.5, Assumption 3.2.1, and the definition of W_k (3.2.7c) imply that $\|W_k^{-1}\|_2 \leq \frac{1}{\lambda_{min}(R_k)} \leq \frac{1}{\lambda_R}$ and thus $\|E_k\|_2 \leq \frac{C_B^2}{\lambda_R} \triangleq C_E$. Proposition 3.2.7 and (3.2.7d) imply that $\|L_k\|_2 \leq \beta C_A C_B / \lambda_R \triangleq C_L$. Lastly, Assumption 3.2.1 and (3.2.4) give that

$$\|f_k\|_2 \leq \left(C_B + \frac{C_B C_R}{\lambda_R} \right) \|\widetilde{b}_i\|_2 \leq 2 \left(C_B + \frac{C_B C_R}{\lambda_R} \right) U \triangleq l_0.$$

B.1.4 Proof of Lemma 3.2.10

Proposition 3.2.6 gives the following:

$$\begin{aligned}
C_i^k &= \sum_{s=n_1}^{\min(i,k)-1} \left(\prod_{l=s+1}^{k-1} D_l\right) E_s \left(\prod_{l=s+1}^{i-1} D_l\right)^T Q_i, \\
F_i^k &= -\sum_{s=n_1}^{\min(i,k)-1} \left(\prod_{l=s+1}^{k-1} D_l\right) E_s \left(\prod_{l=s+1}^{i} D_l\right)^T K_{i+1} \\
&\quad + \left(\prod_{l=i+1}^{k-1} D_l\right) (I - E_i K_{i+1})\, \mathbf{1}_{(k \geq i+1)},
\end{aligned}$$

where $E_s = \hat{B}_s^T W_s^{-1} \hat{B}_s$ (Definition 3.2.5). Lemma 3.2.9 gives that $\|E_s\|_2 \leq C_E$. Using Proposition 3.2.8, the triangle inequality and properties of norms, we have that

$$\begin{aligned}
\|C_i^k\|_2 &\leq \sum_{s=n_1}^{\min(i,k)-1} C_E C_Q C_1^2 \rho^{k-s-1} \rho^{i-s-1} \\
&\leq C_E C_Q C_1^2 \begin{cases} \rho^{k-i} \sum_{s=n_1}^{i-1} \rho^{2i-2s-2}, & i \leq k \\ \rho^{i-k} \sum_{s=n_1}^{k-1} \rho^{2k-2s-2}, & k < i \end{cases} \\
&= C_E C_Q C_1^2 \begin{cases} \rho^{k-i} \sum_{t=0}^{i-n_1-1} \rho^{2t}, & i \leq k \\ \rho^{i-k} \sum_{t=0}^{k-n_1-1} \rho^{2t}, & k < i \end{cases} \\
&\leq \frac{C_E C_Q C_1^2}{1-\rho^2} \rho^{|k-i|}.
\end{aligned}$$

Similarly for F_i^k, we have that

$$\begin{aligned}
\|F_i^k\|_2 &\leq \sum_{s=n_1}^{\min(i,k)-1} C_E C_1 \beta C_1^2 \rho^{k-s-1} \rho^{i-s-1} + (1 + C_E \beta) C_1 \rho^{k-i-1} \mathbf{1}_{(k \geq i+1)} \\
&\leq \frac{C_E C_1 \beta C_1^2}{1-\rho^2} \rho^{|k-i|} + \frac{C_1(1 + C_E \beta)}{\rho} \rho^{|k-i|}.
\end{aligned}$$

Letting

$$C_2 = \max\left(\frac{C_E C_Q C_1^2}{1-\rho^2}, \frac{C_E C_1 \beta C_1^2}{1-\rho^2} + \frac{C_1(1+C_E\beta)}{\rho}\right)$$

completes the proof.

B.1.5 Proof of Proposition 3.2.16

The Karush-Kuhn-Tucker (KKT) conditions for problem (3.2.1) are

$$2R_k u_k^* - C_k^T \lambda_k^* + B_k^T \phi_k^* = 0, \qquad n_1 \le k \le n_2 - 1 \tag{B.1.9a}$$

$$2Q_k(x_k^* - d_k) + A_k^T \phi_k^* - \phi_{k-1}^* = 0, \qquad n_1 + 1 \le k \le n_2 - 1 \tag{B.1.9b}$$

$$2Q_{n_2}(x_{n_2}^* - d_{n_2}) - \phi_{n_2-1}^* = 0, \tag{B.1.9c}$$

$$x_{k+1}^* = A_k x_k^* + B_k u_k^*, \qquad n_1 \le k \le n_2 - 1 \tag{B.1.9d}$$

$$l_k \le u_k^* \le b_k, \qquad n_1 \le k \le n_2 - 1 \tag{B.1.9e}$$

$$\lambda_k^* \ge 0, \qquad n_1 \le k \le n_2 - 1, \tag{B.1.9f}$$

where λ_k^* is the optimal Lagrange multipliers associated with the active constraints $C_k u_k^* = \bar{b}_k$.

We prove the result by induction starting from the rightmost endpoint. When $k = n_2 - 1$, KKT condition (B.1.9c) gives

$$\phi_{n_2-1}^* \;=\; 2Q_{n_2} x_{n_2}^* - 2Q_{n_2} d_{n_2},$$

which satisfies (3.2.18) because $M_{n_2}^{n_2} = Q_{n_2}$ as defined in (3.2.7f). Suppose (3.2.18) is true for k. Then for $k-1$, (B.1.9b) gives

$$\phi_{k-1}^* \;=\; A_k^T \phi_k^* + 2Q_k(x_k^* - d_k).$$

Then by substituting the induction hypothesis and (B.1.9d), we have that

$$\begin{aligned}\phi_{k-1}^* &= 2A_k^T K_{k+1}(A_k x_k^* + B_k u_k^*) + 2Q_k(x_k^* - d_k)\\ &\quad - 2A_k^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i \right)\\ &= 2A_k^T K_{k+1}(A_k x_k^* + \hat{B}_k v_k^* + f_k) + 2Q_k(x_k^* - d_k)\\ &\quad - 2A_k^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i \right),\end{aligned}$$

since (3.2.4) implies that $\hat{B}_k v_k^* + f_k = \hat{B}_k \hat{u}_k^* + \widetilde{B}_k \widetilde{b}_k = B_k u_k^*$. Substituting v_k^* from the optimal control law (3.2.6) then gives the following:

$$\begin{aligned}\phi_{k-1}^* &= 2\left(A_k^T K_{k+1} A_k + Q_k + A_k^T K_{k+1} \hat{B}_k L_k\right) x_k^*\\ &\quad - 2A_k^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i \right) - 2Q_k d_k + 2A_k^T K_{k+1} f_k\\ &\quad + 2A_k^T K_{k+1} \hat{B}_k W_k^{-1} \hat{B}_k^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i - K_{k+1} f_k \right)\end{aligned}$$

$$\begin{aligned}&\overset{(3.2.7b),(3.2.7d)}{=} 2K_k x_k^* - 2A_k^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i \right) - 2Q_k d_k\\ &\quad + 2A_k^T K_{k+1} f_k - 2\left(\hat{B}_k L_k\right)^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i - K_{k+1} f_k \right)\\ &\overset{(3.2.7e)}{=} 2K_k x_k^* - 2Q_k d_k - 2D_k^T \left(\sum_{i=k+1}^{n_2} \left(M_i^{k+1}\right)^T d_i + \sum_{i=k+1}^{n_2-1} \left(S_i^{k+1}\right)^T f_i - K_{k+1} f_k \right)\\ &= 2K_k x_k^* - 2\sum_{i=k}^{n_2} \left(M_i^k\right)^T d_i - 2\sum_{i=k}^{n_2-1} \left(S_i^k\right)^T f_i,\end{aligned}$$

where the last equality follows from (3.2.7f) and (3.2.7g). This completes the proof.

B.1.6 Proof of Lemma 3.2.17

Propositions 3.2.8, 3.2.16, and Corollary 3.2.15 give the following:

$$\begin{aligned}
\|\phi_k^*\|_2 &\leq 2\beta C_g + 2m_0 \sum_{i=k+1}^{n_2} \|M_i^{k+1}\|_2 + 2l_0 \sum_{i=k+1}^{n_2-1} \|S_i^{k+1}\|_2 \\
&\leq 2\beta C_g + 2m_0 C_Q \sum_{i=k+1}^{n_2} C_1 \rho^{i-k-1} + 2l_0\beta \sum_{i=k+1}^{n_2-1} C_1 \rho^{i-k} \\
&\leq 2\beta C_g + \frac{2C_1(m_0 C_Q + \beta l_0)}{1-\rho} \triangleq C_\phi,
\end{aligned}$$

where m_0 is the bound on the reference trajectory in Assumption 3.2.12 and l_0 is the bound on $\|f_i\|$ derived in Lemma 3.2.9.

B.1.7 Proof of Lemma 3.3.5

Let

$$\begin{aligned}
L(y,\theta) = y^T G y/2 + y^T c(\theta) + \lambda^T (Ay - r) + \phi^T (By - d(\theta)) \\
+ \theta^T F \theta + y^T c_1 + \theta^T c_2 + C
\end{aligned} \tag{B.1.10}$$

be the Lagrangian of problem (3.3.8). Then we have that

$$\nabla^2_{(y,\theta)} L = \begin{bmatrix} G & \nabla_\theta c \\ \nabla_\theta^T c & * \end{bmatrix}.$$

Since G and F are positive definite and LICQ holds at y_0, then from [22, Theorem 5.53] and [22, Remark 5.55] we have that

$$\begin{aligned} D_p y(\theta_0) &= \mathrm{argmin}_{h \in S} \begin{bmatrix} h^T & p^T \end{bmatrix} \left(\nabla^2_{(y,\theta)} L(y_0, \theta_0) \right) \begin{bmatrix} h \\ p \end{bmatrix} \\ &= \mathrm{argmin}_{h \in S}\ h^T G h/2 + p^T \left(\nabla^T_\theta c(\theta_0) \right) h, \end{aligned} \tag{B.1.11}$$

where S is the solution of the following linearized problem,

$$\begin{aligned} \min_h \quad & (G y_0 + c(\theta_0) + c_1)^T h + \left(\nabla^T_\theta c(\theta_0) y_0 + 2F\theta_0 + c_2 \right)^T p \\ \text{s.t.} \quad & Bh - (\nabla_\theta d(\theta_0))\, p = 0 \\ & A_{I(y_0,\theta_0)} h \leq 0, \end{aligned} \tag{B.1.12}$$

and S is given by

$$S = \left\{ h : \begin{bmatrix} B & -\nabla_\theta d(\theta_0) \end{bmatrix} \begin{bmatrix} h \\ p \end{bmatrix} = 0,\ \begin{bmatrix} A_{I_+(y_0,\theta_0,\bar{\lambda})} & 0 \end{bmatrix} \begin{bmatrix} h \\ p \end{bmatrix} = 0,\ \begin{bmatrix} A_{I_0(y_0,\theta_0,\bar{\lambda})} & 0 \end{bmatrix} \begin{bmatrix} h \\ p \end{bmatrix} \leq 0 \right\}.$$

Thus the directional derivative $D_p y(\theta_0)$ of $y(\theta)$ along direction p at θ_0 is the solution of the problem

$$\begin{aligned} \min_h \quad & h^T G h/2 + p^T \left(\nabla^T_\theta c(\theta_0) \right) h \\ \text{s.t.} \quad & Bh - (\nabla_\theta d(\theta_0))\, p = 0 \\ & A_{I_+(y_0,\theta_0,\bar{\lambda})} h = 0 \\ & A_{I_0(y_0,\theta_0,\bar{\lambda})} h \leq 0. \end{aligned} \tag{B.1.13}$$

Let I_1 be the set of active inequality constraints of problem (B.1.13). Then $I_1 \subset I_0(y_0, \theta_0, \bar{\lambda})$ and let $I'(\theta_0) = I_1 \cup I_+(y_0, \theta_0, \bar{\lambda})$. The KKT condition of problem (B.1.13) is hence

$$\widetilde{G} \triangleq \begin{bmatrix} G & A_{I'(\theta_0)}^T & B^T \\ A_{I'(\theta_0)} & 0 & 0 \\ B & 0 & 0 \end{bmatrix}, \qquad \widetilde{G} \begin{bmatrix} h^* \\ \phi_1^* \\ \phi_2^* \end{bmatrix} = \begin{bmatrix} \nabla_\theta c(\theta_0) p \\ 0 \\ \nabla_\theta d(\theta_0) p \end{bmatrix}$$

for some Lagrange multipliers ϕ_1^* and ϕ_2^*. Since LICQ holds at y_0, rows of $A_{I'(\theta_0)}$ and B are linearly independent. Together with the fact that G is positive definite, we have that $\widetilde{G}$ is invertible. Denote the first row of $\widetilde{G}^{-1}$ to be $\begin{bmatrix} p_{11} & p_{12} & p_{13} \end{bmatrix}$. Then we have that

$$D_p y(\theta_0) = h^* = \left(-p_{11} \nabla_\theta c(\theta_0) + p_{13} \nabla_\theta d(\theta_0)\right) p.$$

On the other hand, for problem (3.3.9) with $I'(\theta_0)$ constructed above, the KKT condition is

$$\widetilde{G} \begin{bmatrix} y_{I'(\theta_0)}^*(\theta) \\ \psi_1^* \\ \psi_2^* \end{bmatrix} = \begin{bmatrix} -c(\theta) \\ r' \\ d(\theta) \end{bmatrix},$$

for some Lagrange multipliers ψ_1^* and ψ_2^*. Since $\widetilde{G}$ is invertible, we have that $y_{I'(\theta_0)}^*(\theta) = -p_{11} c(\theta) + p_{12} r' + p_{13} d(\theta)$. It follows that

$$\left. \frac{d y_{I'(\theta_0)}^*(\theta)}{d\theta} \right|_{\theta=\theta_0} = -p_{11} \nabla_\theta c(\theta_0) + p_{13} \nabla_\theta d(\theta_0).$$

As a result, we have that

$$D_p y(\theta_0) = \left(\left. \frac{d y_{I'(\theta_0)}^*(\theta)}{d\theta} \right|_{\theta=\theta_0} \right) p,$$

which proves the claim.

APPENDIX C

SUPPLEMENT TO CHAPTER 4

C.1 Proofs of statements in Sections 2 − 4

In order to prove Proposition 4.2.1, Lemma 4.2.2, Theorems 4.2.3 and 4.2.4, we need the following lemmas.

Lemma C.1.1. *For* $0 \leq r \leq m$ *and* $i \geq 1$,

(a) $\left[{m \atop r}\right]_q = \left[{m \atop m-r}\right]_q$

(b) $\left[{m \atop r}\right]_q = q^r \left[{m-1 \atop r}\right]_q + \left[{m-1 \atop r-1}\right]_q = \left[{m-1 \atop r}\right]_q + q^{(m-r)} \left[{m-1 \atop r-1}\right]_q$

(c) $\lim_{q \to 1} \left[{m \atop r}\right]_q = \binom{m}{r}$

(d) $\sum_{j=1}^{i} (-w)^{i-j} \left[{i-1 \atop j-1}\right]_{w^2} = \prod_{k=1}^{i-1} \left(1 + (-w)^k\right)$

We refer the readers to [5] for the proof of Lemma C.1.1.

Lemma C.1.2. *If* $a_n \sim b_n$, $a_n > 0$ *and* $\sum_{n=1}^{\infty} a_n = \infty$, *then* $\sum_{n=1}^{N} a_n \sim \sum_{n=1}^{N} b_n$ *as* $N \to \infty$.

Proof. We carry out a standard $\varepsilon - N$ proof.

Since $a_n \sim b_n$, for any $\varepsilon > 0$, there exists $N_0(\varepsilon) > 0$ such that for any $n \geq N_0(\varepsilon)$,

$$|a_n - b_n| < \varepsilon a_n / 2,$$

then for any $N \geq N_0(\varepsilon)$,

$$\sum_{n=N_0(\varepsilon)}^{N} |a_n - b_n| < \frac{\varepsilon}{2} \sum_{n=N_0(\varepsilon)}^{N} a_n. \tag{C.1.1}$$

Since $\sum_{n=1}^{\infty} a_n = \infty$, there exists $N_1(\varepsilon) \geq N_0(\varepsilon)$ such that for any $N \geq N_1(\varepsilon)$,

$$\sum_{n=1}^{N_0(\varepsilon)} |a_n - b_n| < \frac{\varepsilon}{2} \sum_{n=N_0(\varepsilon)}^{N} a_n. \tag{C.1.2}$$

As a result, for any $N \geq N_1(\varepsilon) \geq N_0(\varepsilon)$,

$$\frac{\left|\sum_{n=1}^{N}(a_n - b_n)\right|}{\sum_{n=1}^{N} a_n} \leq \frac{\sum_{n=1}^{N_0(\varepsilon)-1} |a_n - b_n| + \sum_{n=N_0(\varepsilon)}^{N} |a_n - b_n|}{\sum_{n=N_0(\varepsilon)}^{N} a_n} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

where the second inequality follows from (C.1.2) and (C.1.1). □

Lemma C.1.3. *$\sum_{j=1}^{i}(-w)^{i-j}\left[\begin{smallmatrix} i-1 \\ j-1 \end{smallmatrix}\right]_{w^2} j^p - \sum_{j=1}^{i-1}(-w)^{i-j-1}\left[\begin{smallmatrix} i-2 \\ j-1 \end{smallmatrix}\right]_{w^2} j^p = A_{ip}(w) + B_{ip}(w)$, where*

$$\begin{aligned} A_{ip}(w) &= (-w)^{i-1} \sum_{j=1}^{i-1} (-w)^{i-1-j} \begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2} (i-1-j)^p \\ B_{ip}(w) &= \sum_{j=1}^{i} (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} (j^p - (j-1)^p) \end{aligned}$$

for $p \geq 0$ and $i - 1 > p$.

Proof.

$$\begin{aligned} & \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p - \sum_{j=1}^{i-1}(-w)^{i-j-1}\begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2} j^p \\ = \;& (-w)^{i-1} + (i^p - (i-1)^p) + \sum_{k=1}^{i-2}(-w)^{i-k-1}\left(\begin{bmatrix} i-1 \\ k \end{bmatrix}_{w^2}(k+1)^p - \begin{bmatrix} i-2 \\ k-1 \end{bmatrix} k^p\right) \\ = \;& (-w)^{i-1} + (i^p - (i-1)^p) \\ & + \sum_{k=1}^{i-2}(-w)^{i-k-1}\left(w^{2k}\begin{bmatrix} i-2 \\ k \end{bmatrix}_{w^2} k^p + \begin{bmatrix} i-1 \\ k \end{bmatrix}_{w^2}((k+1)^p - k^p)\right) \\ = \;& (-w)^{i-1}\sum_{k=1}^{i-2}(-w)^k\begin{bmatrix} i-2 \\ k \end{bmatrix}_{w^2} k^p + \sum_{k=0}^{i-1}(-w)^{i-k-1}\begin{bmatrix} i-1 \\ k \end{bmatrix}_{w^2}((k+1)^p - k^p) \end{aligned}$$

$$
\begin{aligned}
&= (-w)^{i-1}\sum_{j=1}^{i-1}(-w)^{i-1-j}\begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2}(i-1-j)^p + \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2}(j^p-(j-1)^p) \\
&= A_{ip}(w)+B_{ip}(w)
\end{aligned}
$$

where the second equality is obtained using Lemma C.1.1 (b), and the fourth equality is obtained by a change of variable $j = i-1-k$ for $A_{ip}(w)$ and $j = k+1$ for $B_{ip}(w)$. □

The following lemma gives a factorization of $\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p$ that enables simplification of $\hat{\theta}_0$.

Lemma C.1.4.

$$
\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p = \begin{cases} (1-w)^{\frac{i-p-1}{2}} f_{ip}(w), & i-p \quad \text{odd} \\ (1-w)^{\frac{i-p}{2}} g_{ip}(w), & i-p \quad \text{even} \end{cases}
$$

for any $w \in (0,1)$, where, for $p \geq 0$ and $i > p$, $f_{ip}(w)$ is a polynomial, $f_{ip}(1) = \frac{(i-1)!}{(\frac{i-p-1}{2})!}$, $g_{ip}(w)$ is a polynomial, and $g_{ip}(1) = \frac{i!(p+1)}{2(\frac{i-p}{2})!}$.

Proof. We denote for simplicity $m_{ip}(w) := \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p$. Here we make induction on (i,p) where $p \geq 0$ and $i > p$. Specifically, we prove the following three steps of which the first two serve as induction basis:

(a) The statement holds for any $p = 0$ and $i > 0$.

With Lemma C.1.1 (d), we have

$$
\begin{aligned}
m_{i0}(w) &= \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} = \prod_{k=1}^{i-1}\left(1+(-w)^k\right) \\
&= \begin{cases} (1-w)^{\frac{i-1}{2}} f_{i0}(w), & i \quad \text{odd} \\ (1-w)^{\frac{i}{2}} g_{i0}(w), & i \quad \text{even} \end{cases}
\end{aligned}
$$

where $f_{i0}(w)$ and $g_{i0}(w)$ are polynomials and

$$\begin{aligned} f_{i0}(1) &= 2^{\frac{i-1}{2}}(i-2)!! = \frac{(i-1)!}{(\frac{i-1}{2})!}, \\ g_{i0}(1) &= 2^{\frac{i-2}{2}}(i-1)!! = \frac{i!}{2(\frac{i}{2})!}. \end{aligned}$$

(b) The statement holds for any $i = p+1$ and $p \geq 1$.

When $i = p+1$, $i-p$ is odd and $m_{p+1,p}(w)$ is a polynomial by definition. Then by Lemma C.1.1 (c),

$$\begin{aligned} m_{p+1,p}(1) &= \sum_{j=1}^{p+1}(-1)^{p+1-j}\binom{p}{j-1}j^p \\ &= (-1)^p\sum_{k=0}^{p}(-1)^k\binom{p}{k}(k+1)^p \\ &= (-1)^p(-1)^p p! \\ &= p!. \end{aligned}$$

(c) Suppose the statement holds for any (i', p') such that ($0 \leq p' < p$ and $i' > p'$) or ($p' = p$ and $p' < i' < i$), then the statement also holds for (i, p).

Define the following polynomials in w:

$$\begin{aligned} h_{ip}(w) &= (-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}\Bigg((1-w)^{\frac{k-3}{2}}g_{i-1,p-k}(w)\mathbf{1}_{(k \text{ odd})} \\ &\quad +(1-w)^{\frac{k-4}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(k \text{ even})}\Bigg), \\ r_{ip}(w) &= \sum_{k=0}^{p-2}(-1)^{p-k+1}\binom{p}{k}\Bigg((1-w)^{\frac{p-k-3}{2}}f_{ik}(w)\mathbf{1}_{(i-k \text{ odd})} \\ &\quad +(1-w)^{\frac{p-k-2}{2}}g_{ik}(w)\mathbf{1}_{(i-k \text{ even})}\Bigg), \end{aligned}$$

$$\begin{aligned}
u_{ip}(w) &= (-w)^{i-1}\sum_{k=2}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}\Bigg((1-w)^{\frac{k-3}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(k\text{ odd})}\\
&\quad +(1-w)^{\frac{k-2}{2}}g_{i-1,p-k}\mathbf{1}_{(k\text{ even})}\Bigg),\\
v_{ip}(w) &= \sum_{k=0}^{p-1}(-1)^{p-k+1}\binom{p}{k}\Bigg((1-w)^{p-k-2}f_{ik}(w)\mathbf{1}_{(i-k\text{ odd})}\\
&\quad +(1-w)^{p-k-1}g_{ik}(w)\mathbf{1}_{(i-k\text{ even})}\Bigg).
\end{aligned}$$

When $i-p$ is even, letting $A_{ip}(w)$ and $B_{ip}(w)$ be defined in Lemma C.1.3 and from binomial expansion, we have

$$\begin{aligned}
A_{ip}(w) &= (-w)^{i-1}\sum_{j=1}^{i-1}(-w)^{i-1-j}\begin{bmatrix}i-2\\ j-1\end{bmatrix}_{w^2}(i-1-j)^p\\
&= (-w)^{i-1}\sum_{j=1}^{i-1}(-w)^{i-1-j}\begin{bmatrix}i-2\\ j-1\end{bmatrix}_{w^2}\\
&\quad\times\Bigg((-1)^p j^p+(-1)^{p-1}p(i-1)j^{p-1}+(-1)^{p-2}(i-1)^2\binom{p}{2}j^{p-2}\\
&\quad\quad+\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}j^{p-k}\Bigg)\\
&= (-w)^{i-1}\Bigg((-1)^p m_{i-1,p}(w)+(-1)^{p-1}p(i-1)m_{i-1,p-1}(w)\\
&\quad\quad+(-1)^{p-2}(i-1)^2\binom{p}{2}m_{i-1,p-2}(w)\Bigg)\\
&\quad+(-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}m_{i-1,p-k}(w)\\
&= (-w)^{i-1}\Bigg((-1)^p(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w)\\
&\quad\quad+(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\Bigg)\\
&\quad+(-w)^{i-1}\left((-1)^{p-2}(i-1)^2\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)\right)
\end{aligned}$$

$$
\begin{aligned}
& +(-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^{k}\binom{p}{k} \\
& \quad \times\Bigg((1-w)^{\frac{i-1-p+k}{2}}g_{i-1,p-k}(w)\mathbf{1}_{(i-p+k-1 \text{ even})} \\
& \qquad +(1-w)^{\frac{i-p+k-2}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(i-p+k-1 \text{ odd})}\Bigg) \\
= \; & (-w)^{i-1}\Bigg((-1)^{p}(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w) \\
& \quad +(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\Bigg) \\
& +(-w)^{i-1}\left((-1)^{p-2}(i-1)^{2}\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)\right) \\
& +(-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^{k}\binom{p}{k}(1-w)^{\frac{i-p}{2}+1} \\
& \quad \times\left((1-w)^{\frac{k-3}{2}}g_{i-1,p-k}(w)\mathbf{1}_{(k \text{ odd})}+(1-w)^{\frac{k-4}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(k \text{ even})}\right) \\
= \; & (-w)^{i-1}\Bigg((-1)^{p}(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w) \\
& \quad +(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\Bigg) \\
& +(-w)^{i-1}\left((-1)^{p-2}(i-1)^{2}\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)\right) \\
& +(1-w)^{\frac{i-p}{2}+1}h_{ip}(w),
\end{aligned}
$$

and

$$
\begin{aligned}
B_{ip}(w) &= \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^{2}}\left(j^{p}-(j-1)^{p}\right) \\
&= \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^{2}}\left(\sum_{k=0}^{p-1}\binom{p}{k}j^{k}(-1)^{(p-k+1)}\right) \\
&= \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^{2}}\left(pj^{p-1}+\sum_{k=0}^{p-2}\binom{p}{k}j^{k}(-1)^{(p-k+1)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= pm_{i,p-1}(w) + \sum_{k=0}^{p-2}(-1)^{(p-k+1)}\binom{p}{k}m_{ik}(w) \\
&= p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w) \\
&\quad + \sum_{k=0}^{p-2}(-1)^{p-k+1}\binom{p}{k}\Bigg((1-w)^{\frac{i-k-1}{2}}f_{ik}(w)\mathbf{1}_{(i-k \text{ odd})} \\
&\qquad +(1-w)^{\frac{i-k}{2}}g_{ik}(w)\mathbf{1}_{(i-k \text{ even})}\Bigg) \\
&= p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w) \\
&\quad + \sum_{k=0}^{p-2}(-1)^{p-k+1}\binom{p}{k}(1-w)^{\frac{i-p}{2}+1}\Bigg((1-w)^{\frac{p-k-3}{2}}f_{ik}(w)\mathbf{1}_{(i-k \text{ odd})} \\
&\qquad +(1-w)^{\frac{p-k-2}{2}}g_{ik}(w)\mathbf{1}_{(i-k \text{ even})}\Bigg) \\
&= p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w) + (1-w)^{\frac{i-p}{2}+1}r_{ip}(w).
\end{aligned}
$$

Then by Lemma C.1.3, we have

$$
\begin{aligned}
m_{ip}(w) &= m_{i-1,p}(w) + A_{ip}(w) + B_{ip}(w) \\
&= (1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w) - w^{i-1}(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w) \\
&\quad +w^{i-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w) \\
&\quad -w^{i-1}(i-1)^2\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w) + p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w) \\
&\quad +(1-w)^{\frac{i-p}{2}+1}h_{ip}(w) + (1-w)^{\frac{i-p}{2}+1}r_{ip}(w) \\
&= (1-w)^{\frac{i-p}{2}}\Bigg(f_{i-1,p}(w)\left(\frac{1-w^{i-1}}{1-w}\right) + p(i-1)w^{i-1}g_{i-1,p-1}(w) \\
&\quad -\binom{p}{2}(i-1)^2w^{i-1}f_{i-1,p-2}(w) + pf_{i,p-1}(w) \\
&\quad +(1-w)h_{ip}(w) + (1-w)r_{ip}(w)\Bigg).
\end{aligned}
$$

Let

$$\begin{aligned} g_{ip}(w) &= f_{i-1,p}(w)\left(\frac{1-w^{i-1}}{1-w}\right) + p(i-1)w^i g_{i-1,p-1}(w) \\ &\quad -\binom{p}{2}(i-1)^2 w^{i-1} f_{i-1,p-2}(w) + p f_{i,p-1}(w) \\ &\quad +(1-w)h_{ip}(w) + (1-w)r_{ip}(w) \end{aligned}$$

which is a polynomial by induction hypothesis, and

$$\begin{aligned} g_{ip}(1) &= (i-1)\frac{(i-2)!}{(\frac{i-p-2}{2})!} + p(i-1)\frac{(i-1)!p}{2(\frac{i-p}{2})!} - \binom{p}{2}(i-1)^2\frac{(i-2)!}{(\frac{i-p}{2})!} + p\frac{(i-1)!}{(\frac{i-p}{2})!} \\ &= \frac{(i-1)!}{(\frac{i-p}{2})!}\left(\frac{i-p}{2} + p + \frac{p^2(i-1)}{2} - \frac{p(p-1)(i-1)}{2}\right) \\ &= \frac{i!(p+1)}{2(\frac{i-p}{2})!}. \end{aligned}$$

When $i-p$ is odd, similarly by binomial expansion and induction hypothesis, we have

$$\begin{aligned} A_{ip}(w) &= (-w)^{i-1}\Bigg((-1)^p(1-w)^{\frac{i-p-1}{2}} g_{i-1,p}(w) \\ &\qquad +(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p-1}{2}} f_{i-1,p-1}(w)\Bigg) \\ &\quad +(-w)^{i-1}\sum_{k=2}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k} m_{i-1,p-k}(w) \\ &= (-w)^{i-1}\Bigg((-1)^p(1-w)^{\frac{i-p-1}{2}} g_{i-1,p}(w) \\ &\qquad +(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p-1}{2}} f_{i-1,p-1}(w)\Bigg) \\ &\quad +(-w)^{i-1}\sum_{k=2}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}(1-w)^{\frac{i-p+1}{2}} \\ &\qquad \times\left((1-w)^{\frac{k-3}{2}} f_{i-1,p-k}(w)\mathbf{1}_{(k \text{ odd})} + (1-w)^{\frac{k-2}{2}} g_{i-1,p-k}\mathbf{1}_{(k \text{ even})}\right) \end{aligned}$$

$$\begin{aligned} &= (-w)^{i-1}\Bigg((-1)^p(1-w)^{\frac{i-p-1}{2}}g_{i-1,p}(w) \\ &\quad +(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p-1}{2}}f_{i-1,p-1}(w)\Bigg) + (1-w)^{\frac{i-p+1}{2}}u_{ip}(w), \end{aligned}$$

and

$$\begin{aligned} B_{ip}(w) &= \sum_{k=0}^{p-1}(-1)^{p-k+1}\binom{p}{k}m_{ik}(w) \\ &= \sum_{k=0}^{p-1}(-1)^{p-k+1}\binom{p}{k}(1-w)^{\frac{i-p+1}{2}} \\ &\quad \times \left((1-w)^{p-k-2}f_{ik}(w)\mathbf{1}_{(i-k \text{ odd})} + (1-w)^{p-k-1}g_{ik}(w)\mathbf{1}_{(i-k \text{ even})}\right) \\ &= (1-w)^{\frac{i-p+1}{2}}v_{ip}(w). \end{aligned}$$

Then similarly by Lemma C.1.3, we have

$$\begin{aligned} m_{ip}(w) &= m_{i-1,p}(w) + A_{ip}(w) + B_{ip}(w) \\ &= (1-w)^{\frac{i-p-1}{2}}\left(g_{i-1,p}(w) + w^{i-1}g_{i-1,p}(w) - p(i-1)w^{i-1}f_{i-1,p-1}(w)\right) \\ &\quad +(1-w)^{\frac{i-p-1}{2}}\left((1-w)u_{ip}(w) + (1-w)v_{ip}(w)\right). \end{aligned}$$

Let

$$\begin{aligned} f_{ip}(w) &= g_{i-1,p}(w) + w^{i-1}g_{i-1,p}(w) - p(i-1)w^{i-1}f_{i-1,p-1}(w) \\ &\quad +(1-w)u_{ip}(w) + (1-w)v_{ip}(w) \end{aligned}$$

which is a polynomial, and

$$\begin{aligned} f_{ip}(1) &= 2\frac{(i-1)!(p+1)}{2(\frac{i-p-1}{2})!} - p(i-1)\frac{(i-2)!}{(\frac{i-p-1}{2})!} \\ &= \frac{(i-1)!}{(\frac{i-p-1}{2})!}. \end{aligned}$$

□

Lemma C.1.5. *$a_{i0}(w)$ monotonically decreases for $w \in (0,1)$ and any $i \geq 1$.*

Proof. With Lemma C.1.1 (d), we have

$$\begin{aligned} a_{i0}(w) &= \frac{\prod_{k=1}^{i-1}\left(1+(-w)^k\right)^2}{\prod_{k=1}^{i-1}(1-w^{2k})} \\ &= \prod_{k=1}^{i-1}\frac{1+(-w)^k}{1-(-w)^k}. \end{aligned}$$

For $k \geq 1$, let

$$f_k(w) = \frac{(1-w^k)(1+w^{k+1})}{(1+w^k)(1-w^{k+1})}, \tag{C.1.8}$$

then $f_k(w) = \left(\frac{2}{1+w^k} - 1\right)\left(\frac{2}{1-w^{k+1}} - 1\right)$ and

$$\begin{aligned} f_k'(w) &= \frac{-2kw^{k-1}}{(1+w^k)^2}\left(\frac{2}{1-w^{k+1}} - 1\right) + \frac{2(k+1)w^k}{(1-w^{k+1})^2}\left(\frac{2}{1+w^k} - 1\right) \\ &= \frac{1}{(1+w^k)^2(1-w^{k+1})^2}2w^{k-1}\left(kw^{2k+2} - (k+1)w^{2k+1} + (k+1)w - k\right) \end{aligned}$$

For $g_k(w) = kw^{2k+2} - (k+1)w^{2k+1} + (k+1)w - k$, we know $g_k(1) = 0$ and

$$g_k'(w) = (k+1)(1-w)(1 + w + \cdots + w^{2k-1} - 2kw^{2k}) > 0$$

for $w \in (0,1)$. As a result, $g_k(w) < 0$ for $w \in (0,1)$ and hence $f_k(w)$ monotonically decreases on $(0,1)$. Because

$$a_{i0}(w) = \begin{cases} \prod_{k=1}^{\frac{i-1}{2}} f_{2k-1}(w), & i \qquad \text{odd} \\ \prod_{k=1}^{\frac{i-2}{2}} f_{2k-1}(w)\frac{1-w^{i-1}}{1+w^{i-1}}, & i \qquad \text{even} \end{cases}$$

$a_{i0}(w)$ is monotonically decreasing over $(0,1)$. □

Lemma C.1.6. *Denote $w_i^* := e^{-\frac{1}{2(i-2)}}$, then $a_{i1}(w)$ monotonically decreases for $w \in (w_i^*, 1)$ and any $i \geq 2$ and i even.*

Proof. We prove by induction. Note that by the definition of w, we have $n^2 = \left(\theta_1 \log \frac{1}{w}\right)^{-1}$.

When $i = 2$,

$$\begin{aligned} a_{21}(w) &= \frac{(2-w)^2 \theta_1 \log \frac{1}{w}}{(1-w^2)}, \\ a_{21}'(w) &= \frac{(2-w)\theta_1}{(1-w^2)^2} \left(-2(1-w^2) \log \frac{1}{w} + (2-w) \left(2w \log \frac{1}{w} + w - \frac{1}{w} \right) \right). \end{aligned}$$

Consider $z(w) = 2w \log \frac{1}{w} + w - \frac{1}{w}$, then $z(1) = 0$ and $z'(w) = \frac{1}{w^2} + 2 \log \frac{1}{w} - 1 > 0$ for $w \in (0, 1)$. So $z(w) < 0$ for $w \in (0, 1)$. As a result, $a_{21}'(w) < 0$ for $w \in (0, 1)$. Note that $w_2^* = 0$. We denote for simplicity $m_{ip}(w) := \sum_{j=1}^{i} (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p$. Suppose the statement holds for some even i, then for $i+2$, by repeatedly applying Lemma C.1.3, we obtain

$$m_{i+2,1}(w) = (1 + w^{i+1})(1 - w^i) m_{i1}(w) + \left(i w^i (1-w) + (1 + w^i)(2 - w^{i+1}) \right) m_{i0}(w).$$

We consider, by using Lemma C.1.1 (d),

$$\begin{aligned} \sqrt{a_{i+2,1}(w)} &= \frac{m_{i+2,1}(w)}{n \sqrt{\prod_{k=1}^{i+1} (1 - w^{2k})}} \\ &= \frac{(1 + w^{i+1})(1 - w^i) m_{i1}(w)}{n \sqrt{(1 - w^{2i})(1 - w^{2(i+1)}) \prod_{k=1}^{i-1} (1 - w^{2k})}} \\ &\quad + \frac{m_{i0}(w) \left(i w^i (1-w) + (1 + w^i)(2 - w^{i+1}) \right)}{n \sqrt{\prod_{k=1}^{i+1} (1 - w^{2k})}} \\ &= \sqrt{a_{i1}(w)} \sqrt{\frac{(1 - w^i)(1 + w^{i+1})}{(1 + w^i)(1 - w^{i+1})}} \\ &\quad + \sqrt{a_{i0}(w)} \frac{i w^i (1-w) + (1 + w^i)(2 - w^{i+1})}{n \sqrt{(1 - w^{2i})(1 - w^{2(i+1)})}} \end{aligned}$$

$$
\begin{aligned}
&= \sqrt{a_{i1}(w)}\sqrt{\frac{(1-w^i)(1+w^{i+1})}{(1+w^i)(1-w^{i+1})}} \\
&\quad + \sqrt{a_{i+2,0}(w)}\frac{iw^i(1-w)+(1+w^i)(1-w^{i+1})+(1+w^i)}{n(1+w^i)(1-w^{i+1})} \\
&= \sqrt{a_{i1}(w)}\sqrt{\frac{(1-w^i)(1+w^{i+1})}{(1+w^i)(1-w^{i+1})}} \\
&\quad + \frac{w^i}{n}\frac{i\sqrt{a_{i+2,0}}}{(1+w^i)(1+w+\cdots+w^i)} + \frac{\sqrt{a_{i+2,0}}}{n} + \frac{\sqrt{a_{i+2,0}}}{n(1-w^{i+1})} \\
&:= I_1(w)+I_2(w)+I_3(w)+I_4(w).
\end{aligned}
$$

Now we show the monotonicity for each $I_i(w)$, $i = 1, \ldots, 4$. Firstly,

$$I_1(w) = \sqrt{a_{i1}(w)f_i(w)}$$

where $f_i(w)$ is defined in (C.1.8). $I_1(w)$ monotonically decreases for $w > w_i^*$ by the induction hypothesis and the fact that $f_i(w)$ decreases (proved in Lemma C.1.5). Secondly, $I_2(w)$ monotonically decreases for $w > w_{i+2}^*$ by Lemma C.1.5 and the fact that $\frac{w^{2i}}{n^2}$ monotonically decreases for $w > w_{i+2}^* = e^{-\frac{1}{2i}}$, which follows from

$$
\begin{aligned}
\frac{w^{2i}}{n^2} &= \theta_1 w^{2i}\log\frac{1}{w}, \\
\frac{d}{dw}\left(w^{2i}\log\frac{1}{w}\right) &= w^{2i-1}(2i\log\frac{1}{w}-1) < 0 \\
&\Leftrightarrow\ w > e^{-\frac{1}{2i}}.
\end{aligned}
$$

Thirdly, $I_3(w)$ monotonically decreases for $w \in (0,1)$ by Lemma C.1.5. Finally,

$$I_4(w) = \frac{\sqrt{a_{i+1,0}(1-w^{i+1})}}{n(1-w^{i+1})\sqrt{1+w^{i+1}}} = \frac{\sqrt{a_{i+1,0}}}{n\sqrt{1-w^{2(i+1)}}},$$

where $a_{i+1,0}$ monotonically decreases for $w \in (0,1)$ by Lemma C.1.5. Then we show that

$n^2(1-w^{2(i+1)})$ monotonically increases over $(0,1)$ as follows:

$$\begin{aligned} n^2(1-w^{2(i+1)}) &= \frac{1-w^{2(i+1)}}{\theta_1 \log\frac{1}{w}}, \\ \frac{d}{dw}\left(\frac{1-w^{2(i+1)}}{\log\frac{1}{w}}\right) &= \frac{1}{w\log^2\frac{1}{w}}\left(1-w^{2(i+1)}-2(i+1)w^{2(i+1)}\log\frac{1}{w}\right), \end{aligned}$$

letting $t(w) = 1-w^{2(i+1)}-2(i+1)w^{2(i+1)}\log\frac{1}{w}$, then $t(1)=0$ and $t'(w) = -(2i+2)^2 w^{2i+1}\log\frac{1}{w} < 0$ for $w \in (0,1)$, so we have $t(w) > 0$ for $w \in (0,1)$. So $I_4(w)$ monotonically decreases for $w \in (0,1)$. As a result, we conclude that $a_{i+2,1}(w)$ monotonically decreases for $w > w^*_{i+2}$. □

C.1.1 Proof of Proposition 4.2.1

With Lemma C.1.4, we can cancel some factors to obtain

$$\frac{\left(\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})} = \begin{cases} \frac{f_{ip}^2(w)}{(1-w)^p n^{2p}\prod_{k=1}^{i-1} z_k(w)}, & i-p \quad \text{odd} \\ \frac{g_{ip}^2(w)(1-w)}{(1-w)^p n^{2p}\prod_{k=1}^{i-1} z_k(w)}, & i-p \quad \text{even} \end{cases}$$

where $z_k(w) = (1+w^k)(1+w+\cdots+w^{k-1})$.

Note that

$$(1-w)n^2 \to 1/\theta_1$$

as $n \to \infty$. Then we have

$$l_{ip} = \begin{cases} \frac{f_{ip}^2(1)\theta_1^p}{\prod_{k=1}^{i-1} z_k(1)}, & i-p \quad \text{odd} \\ 0, & i-p \quad \text{even} \end{cases}$$

which equals the right hand side of (4.2.5) by using Lemma C.1.4.

C.1.2 Proof of Lemma 4.2.2

When $i-p$ is odd, by Stirling's Approximation, as $i \to \infty$, we have

$$\begin{aligned} l_{ip} &\sim \frac{\sqrt{2\pi(i-1)}(\frac{i-1}{e})^{i-1}\theta_1^p}{2^{i-1}\pi(i-p-1)(\frac{i-p-1}{2e})^{i-p-1}} \\ &\sim \sqrt{\frac{2}{\pi}}\frac{i^{p-\frac{1}{2}}\theta_1^p}{2^p}. \end{aligned}$$

Since $\sum_{i=1}^{\infty} i^{p-1/2} = \infty$ for all $p \geq 0$, by Lemma C.1.2, as $n \to \infty$, we have

$$\begin{aligned} \sum_{i=p+1}^{n} l_{ip} &= \sum_{k=1}^{\lceil\frac{n-p}{2}\rceil} l_{p+2k-1,p} \\ &\sim \sum_{k=1}^{\lceil\frac{n-p}{2}\rceil} \sqrt{\frac{2}{\pi}}\frac{(2k)^{p-\frac{1}{2}}\theta_1^p}{2^p} \\ &\sim \sqrt{\frac{2}{\pi}}\frac{2^{p-\frac{1}{2}}\theta_1^p}{2^p}\sum_{k=1}^{\lceil\frac{n-p}{2}\rceil} k^{p-\frac{1}{2}} \\ &\sim \sqrt{\frac{2}{\pi}}\frac{2^{p-\frac{1}{2}}\theta_1^p}{2^p}\frac{1}{(p+\frac{1}{2})}(\frac{n-p}{2})^{p+\frac{1}{2}} \\ &\sim \frac{n^{p+\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})} \end{aligned}$$

and Lemma 4.2.2 follows.

C.1.3 Proof of Theorem 4.2.3

Let $s(x) = \frac{1-w^x}{x}$ for some $w \in (0,1)$. Note that $s(x)$ monotonically decreases for $x > 0$. The series expansion is $s(x) = \sum_{l=1}^{\infty} \frac{(-x)^{l-1}}{l!}(\frac{1}{\theta_1 n^2})^l$ for $w = e^{-1/\theta_1 n^2}$. Then

$$s(2k-1) - s(2k) = \sum_{l=1}^{\infty} \frac{(-1)^l}{\theta_1^l l!}\frac{(2k)^{l-1} - (2k-1)^{l-1}}{n^{2l}}$$

$$\begin{aligned} &\leq \sum_{l=1}^{\infty} \frac{1}{\theta_1^{2l}(2l)!} \frac{(2k)^{2l-1}-(2k-1)^{2l-1}}{n^{4l}} \\ &\leq \sum_{l=1}^{\infty} \frac{1}{\theta_1^{2l}(2l)!} \frac{2l\binom{2l-1}{l}(2k)^{2l-2}}{n^{4l}} \\ &\leq \sum_{l=1}^{\infty} \frac{1}{\theta_1^{2l}l!} \frac{(2k)^{2l-2}}{n^{4l}} \\ &\leq e^{1/\theta_1^2} \sum_{l=1}^{\infty} \frac{(2k)^{2l-2}}{n^{4l}}, \end{aligned}$$

where the second inequality comes from the binomial expansion.

Referring back to Definition 4.3.1, since, when n is odd, $(n-1)!! = 2^{\frac{n-1}{2}}\left(\frac{n-1}{2}\right)!$ and $(n-1)! = (n-1)!!(n-2)!!$, we have

$$l_{n0} = \frac{(n-1)!}{2^{n-1}\left(\frac{n-1}{2}!\right)^2} = \frac{(n-2)!!}{(n-1)!!},$$

so

$$\frac{a_{n0}(w)}{l_{n0}} = \prod_{k=1}^{\frac{n-1}{2}} \frac{1+w^{2k}}{1+w^{2k-1}} \frac{s(2k-1)}{s(2k)} \leq \prod_{k=1}^{\frac{n-1}{2}} \frac{s(2k-1)}{s(2k)},$$

and

$$\begin{aligned}
\log \frac{a_{n0}(w)}{l_{n0}} &= \sum_{k=1}^{\frac{n-1}{2}} \log\left(1 + \frac{s(2k-1)-s(2k)}{s(2k)}\right) \\
&\leq \sum_{k=1}^{\frac{n-1}{2}} \frac{s(2k-1)-s(2k)}{s(n-1)} \\
&\leq \frac{e^{1/\theta_1^2}}{s(n-1)} \sum_{k=1}^{\frac{n-1}{2}} \sum_{l=1}^{\infty} \frac{(2k)^{2l-2}}{n^{4l}} \\
&= \frac{e^{1/\theta_1^2}}{s(n-1)} \sum_{l=1}^{\infty} \sum_{k=1}^{\frac{n-1}{2}} \frac{(2k)^{2l-2}}{n^{4l}} \\
&\leq \frac{e^{1/\theta_1^2}}{s(n-1)} \sum_{l=1}^{\infty} \frac{n^{2l-1}}{n^{4l}} \\
&= \frac{e^{1/\theta_1^2}}{s(n-1)} \frac{n^2}{n^3(n^2-1)} \\
&= \frac{1}{(n+1)} \frac{e^{1/\theta_1^2}}{(1-w^{n-1})n} \to 0
\end{aligned} \tag{C.1.9}$$

as $n \to \infty$, since $(1-w^{n-1})n \to 1/\theta_1$ as $n \to \infty$.

By Lemma C.1.5, $\frac{a_{n0}(w)}{l_{n0}} \geq 1$ for $n \geq 1$, and combined with (C.1.9), we obtain that for n odd,

$$\frac{a_{n0}(w)}{l_{n0}} \to 1$$

as $n \to \infty$. For n even,

$$a_{n0}(w) \quad = \quad a_{n-1,0}(w)\frac{1-w^{n-1}}{1+w^{n-1}} \sim \frac{1}{2\theta_1 n} l_{n-1,0}$$

as $n \to \infty$.

As a result, denoting $w_i = e^{-1/\theta_1 i^2}$, we have, as $n \to \infty$,

$$\sum_{i=1}^{n} a_{i0}(w_i) \quad \sim \quad \sum_{i=1}^{n} l_{i0} + \sum_{i=1}^{n} \frac{1}{2\theta_1 i} l_{i0} \sim \sum_{i=1}^{n} l_{i0}. \tag{C.1.10}$$

Lemma C.1.5 implies

$$\sum_{i=1}^{n} l_{i0} \leq \sum_{i=1}^{n} a_{i0}(w) \leq \sum_{i=1}^{n} a_{i0}(w_i). \tag{C.1.11}$$

Combining (C.1.10) and (C.1.11), we have

$$\hat{\theta}_0 \;=\; \frac{1}{n}\sum_{i=1}^{n} a_{i0}(w) \sim \frac{1}{n}\sum_{i=1}^{n} l_{i0} \sim \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}},$$

where the last asymptotic equivalence is obtained by taking $p = 0$ in Lemma 4.2.2.

C.1.4 Proof of Theorem 4.2.4

Since there exists $N(\theta_1)$ such that for any $n > N(\theta_1)$,

$$e^{-\frac{1}{2(n-2)}} < e^{-\frac{1}{\theta_1 n^2}} = w,$$

with Lemma C.1.6, for any $n > N(\theta_1)$,

$$w > w_n^* \geq w_i^*$$

for any $2 \leq i \leq n$ and i even.

As a result,

$$\sum_{i=2}^{n} a_{i1}(w) \geq \sum_{i=2}^{n} l_{i1},$$

then

$$\frac{\hat{\theta}_0}{\sqrt{n}} = \frac{\sum_{i=1}^{n} a_{i1}}{n^{3/2}} \geq \frac{\sum_{i=2}^{n} l_{i1}}{n^{3/2}}.$$

Theorem 4.2.4 follows by taking limit infimum on both sides of the above inequality and setting $p = 1$ in Lemma 4.2.2.

C.1.5 Proof of Proposition 4.2.5

For convenience, consider n to be even in the subsequent proof. The arguments only need to be slightly modified for n odd. Denote k as the index of the first non-zero element in the observations, then $k = n/2 + 1$ and $f(k/n) = g(1/n)$. The observations are $\mathbf{z} = (0, \ldots, 0, g(1/n), \ldots, g(1/2))^T$. Denote the (k, k)th element of the inverse Cholesky factor by $C^{-1}_{(k,k)}$. Referring back to (4.2.3) gives

$$\begin{aligned}\hat{\theta}_0 &= \frac{1}{n}\sum_{i=1}^{n} \|C^{-1}\mathbf{z}\|^2 \\ &\geq \frac{1}{n}\left(C^{-1}_{(k,k)} g(1/n)\right)^2 \\ &= \frac{g^2(1/n)}{n\prod_{l=1}^{k-1}(1-w^{2l})}.\end{aligned} \tag{C.1.12}$$

Let $L = p + 1$, then for all n sufficiently large,

$$\prod_{l=1}^{k-1}(1-w^{2l}) < \prod_{l=1}^{L}(1-w^{2l}). \tag{C.1.13}$$

Since $w = e^{1/(\theta_1 n^2)}$, $\prod_{l=1}^{L}(1-w^{2l}) \sim \frac{2^L L!}{\theta_1^L n^{2L}}$ as $n \to \infty$. Combining (C.1.12) and (C.1.13) gives

$$\begin{aligned}\liminf_{n\to\infty} \frac{\hat{\theta}_0}{n} &\geq \liminf_{n\to\infty} \frac{g^2(1/n) n^{2p}}{n^{2p+2}\prod_{l=1}^{L}(1-w^{2l})} \\ &= \frac{\theta_1^L c}{2^L L!} > 0.\end{aligned}$$

C.1.6 Proof of a statement in Section 2

We state and prove the following proposition.

Proposition C.1.7. *With the exponential covariance function* $\mathrm{Cov}(f(x), f(y)) = \theta_0 e^{-|x-y|/\theta_1}$*, if the observations are* $\boldsymbol{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \ldots, f(1))^T$ *for some* f *having a* *bounded second derivative on* $[0,1]$*, then as* $n \to \infty$*,*

$$\hat{\theta}_0 \sim \frac{1}{n} f(0)^2 + \frac{1}{2\theta_1 n} \int_0^1 \big(f(x) + \theta_1 f'(x)\big)^2 \, dx.$$

Proof. Denote the correlation matrix as R and its Cholesky decomposition as $R = CC^T$ for some C lower triangular, then $R_{ij} = \rho^{|i-j|}$ where $\rho = e^{-1/n\theta_1}$. The Cholesky and inverse Cholesky factors are

$$C = \begin{bmatrix} 1 & & & & \\ \rho & \sqrt{1-\rho^2} & & \mathbf{0} & \\ \rho^2 & \rho\sqrt{1-\rho^2} & \sqrt{1-\rho^2} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho^{n-1} & \rho^{n-2}\sqrt{1-\rho^2} & \rho^{n-3}\sqrt{1-\rho^2} & \dots & \sqrt{1-\rho^2} \end{bmatrix},$$

$$C^{-1} = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} \sqrt{1-\rho^2} & & & \mathbf{0} \\ -\rho & 1 & & \\ & \ddots & \ddots & \\ \mathbf{0} & & -\rho & 1 \end{bmatrix},$$

so

$$\begin{aligned} \hat{\theta}_0 &= \frac{1}{n} \boldsymbol{z}^T R^{-1} \boldsymbol{z} = \frac{1}{n} \|C^{-1}\boldsymbol{z}\|^2 \\ &= \frac{1}{n} f(0)^2 + \frac{1}{n} \sum_{j=2}^{n} \frac{\Big(f(\frac{j}{n}) - \rho f(\frac{j-1}{n})\Big)^2}{1-\rho^2}. \end{aligned}$$

A Taylor expansion gives that for some $x_j \in (\frac{j-1}{n}, \frac{j}{n})$, $2 \leq j \leq n$,

$$\begin{aligned} f(\frac{j}{n}) - \rho f(\frac{j-1}{n}) &= f(\frac{j-1}{n}) + \frac{1}{n} f'(\frac{j-1}{n}) + \frac{1}{n^2} f''(x_j) - f(\frac{j-1}{n}) \left(1 - \frac{1}{\theta_1 n} + \frac{\alpha_n}{n^2}\right) \\ &= \frac{1}{\theta_1 n} f(\frac{j-1}{n}) + \frac{1}{n} f'(\frac{j-1}{n}) + \frac{r_{jn}}{n^2}, \end{aligned} \tag{C.1.14}$$

where $|\alpha_n| \leq \alpha = 1/(2\theta_1^2)$ and

$$r_{jn} = f''(x_j) - \alpha_n f(\frac{j-1}{n}).$$

Since $f''(x)$ is bounded on $[0,1]$, $f(x)$ and $f'(x)$ are continuous and bounded on $[0,1]$. Denote $A := \sup_{x\in[0,1]} \left\{|f(x)|, |f'(x)|, |f''(x)|\right\}$, then

$$|r_{jn}| \;\; \leq \;\; A + A\alpha.$$

(C.1.14) gives that

$$\begin{aligned} \sum_{j=2}^{n} \left(f(\frac{j}{n}) - \rho f(\frac{j-1}{n}) \right)^2 \;\; &= \;\; \sum_{j=2}^{n} \left(\frac{1}{\theta_1 n} f(\frac{j-1}{n}) + \frac{1}{n} f'(\frac{j-1}{n}) \right)^2 + \frac{\sum_{j=2}^{n} r_{jn}^2}{n^4} \\ &\quad + \sum_{j=2}^{n} \frac{2 r_{jn}}{n^2} \left(\frac{1}{\theta_1 n} f(\frac{j-1}{n}) + \frac{1}{n} f'(\frac{j-1}{n}) \right), \end{aligned}$$

where as $n \to \infty$,

$$\begin{aligned} \sum_{j=2}^{n} \left(\frac{1}{\theta_1 n} f(\frac{j-1}{n}) + \frac{1}{n} f'(\frac{j-1}{n}) \right)^2 \;\; &= \;\; \frac{1}{n} \sum_{j=2}^{n} \frac{1}{n} \left(\frac{1}{\theta_1} f(\frac{j-1}{n}) + f'(\frac{j-1}{n}) \right)^2 \\ &\sim \;\; \frac{1}{n} \int_0^1 \left(\frac{1}{\theta_1} f(\frac{j-1}{n}) + f'(\frac{j-1}{n}) \right)^2 dx, \end{aligned}$$

since the integrand is continuous so the sum converges to the corresponding Riemann integral,

and

$$\frac{\sum_{j=2}^{n} r_{jn}^2}{n^4} \quad \leq \quad \frac{A^2(1+\alpha)^2}{n^3},$$

$$\begin{aligned}\left|\sum_{j=2}^{n} \frac{2r_{jn}}{n^2}\left(\frac{1}{\theta_1 n} f(\frac{j-1}{n}) + \frac{1}{n} f'(\frac{j-1}{n})\right)\right| &\leq \sum_{j=2}^{n} \frac{2|r_{jn}|}{n^2}\left(\frac{A}{\theta_1 n} + \frac{A}{n}\right) \\ &\leq \frac{2A^2(1+\alpha)}{n^2}\left(1 + \frac{1}{\theta_1}\right).\end{aligned}$$

As a result, as $n \to \infty$,

$$\sum_{j=2}^{n}\left(f(\frac{j}{n}) - \rho f(\frac{j-1}{n})\right)^2 \sim \frac{1}{n}\int_0^1 \left(\frac{1}{\theta_1} f(\frac{j-1}{n}) + f'(\frac{j-1}{n})\right)^2 dx,$$

together with $\rho = e^{-1/\theta_1 n}$, $1 - \rho^2 \sim \frac{2}{\theta_1 n}$ complete the proof. □

C.1.7 Proof of Proposition 4.3.2

First of all, we prove the following lemma that is used frequently in the subsequent proofs.

Lemma C.1.8. *For $0 \leq p \leq m-1$ and $m \geq 1$,*

(a) $\sum_{l=0}^{m}(-1)^l\binom{m}{l} l^p = 0$

(b) $\sum_{l=0}^{m}(-1)^l\binom{m}{l} l^m = (-1)^m m!$

Proof. The Stirling numbers of the second kind can be expressed as the sum [16]:

$$S(m,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^m.$$

Lemma C.1.8 follows from $S(m,m) = 1$ and $S(p,m) = 0$ for $0 \leq p < m$. □

For convenience we write $R_1(\theta_1, n) = R_n$ and $D(\theta_1, n) = D_n$. Since both R_n and D_n are nested, we use induction to prove. First of all, when $n = 1$, $R_1 = D_1 = 1$. Suppose Proposition 4.3.2 is true for n, i.e. $D_n^T D_n R_n = I_n$, then for $n+1$, partition R_{n+1} and D_{n+1} as

$$R_{n+1} := \begin{bmatrix} R_n & r_{n+1} \\ r_{n+1}^T & R_{n+1,n+1} \end{bmatrix}$$

$$D_{n+1} := \begin{bmatrix} D_n & \mathbf{0} \\ d_{n+1}^T & D_{n+1,n+1} \end{bmatrix},$$

then

$$D_{n+1}^T D_{n+1} R_{n+1} = \begin{bmatrix} D_n^T D_n R_n + d_{n+1} A_n & C_n \\ D_{n+1,n+1} A_n & B_n \end{bmatrix},$$

where

$$\begin{aligned} A_n &= d_{n+1}^T R_n + D_{n+1,n+1} r_{n+1}^T, \\ B_n &= D_{n+1,n+1}(d_{n+1}^T r_{n+1} + D_{n+1,n+1} R_{n+1,n+1}), \\ C_n &= D_n^T D_n r_{n+1} + d_{n+1} d_{n+1}^T r_{n+1} + d_{n+1} D_{n+1,n+1} R_{n+1,n+1}. \end{aligned}$$

Here we claim that $A_n = \mathbf{0}^T$, $B_n = 1$ and $C_n = \mathbf{0}$ so that along with the induction hypothesis, we have

$$D_{n+1}^T D_{n+1} R_{n+1} = I_{n+1}.$$

(a) Proof of $A_n = \mathbf{0}^T$: Note that if n and j are of the same parity, then $(d_{n+1}^T R_n)_j = (r_{n+1})_j = 0$, so we have that the jth element of A_n is 0. If n and j are of different parity,

$$(d_{n+1}^T R_n)_j = \sum_{i=1,(i+j)\text{even}}^{n} \frac{\sqrt{n!}2^{\frac{j-1}{2}}(i+j-3)!!(-1)^{i+\frac{i+j}{2}}}{\theta_1^{\frac{j-1}{2}}(i-1)!(n+1-i)!!}$$

$$-(D_{n+1,n+1}r_{n+1})_j \quad = \quad -\frac{2^{\frac{j-1}{2}}(-1)^{n+1+\frac{n+1+j}{2}}(n+j-2)!!}{\theta_1^{\frac{j-1}{2}}\sqrt{n!}}$$

(i) If n odd, j even and $j < n$, applying Lemma C.1.8 and making change of variable $l = \frac{i}{2} - 1$ gives,

$$\begin{aligned}
(d_{n+1}^T R_n)_j &= (2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!} \\
&\quad \times \sum_{l=0}^{\frac{n-1}{2}-1}(-1)^{l+1}\frac{(2l+1)!!}{(2l+1)!(n-1-2l)!!}\prod_{m=1}^{\frac{j-2}{2}}(2l+2m+1) \\
&= -\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\sum_{l=0}^{\frac{n-1}{2}-1}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{j-2}{2}}(2l+2m+1) \\
&= -\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!} \\
&\quad \times \left(\sum_{l=0}^{\frac{n-1}{2}}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{j-2}{2}}(2l+2m+1) - (-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{j-2}{2}}(n+2m)\right) \\
&= -\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\left(0 + (-1)^{\frac{n+1}{2}}\prod_{m=1}^{\frac{j-2}{2}}(n+2m)\right) \\
&= (2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{n+j-1}{2}}\frac{\sqrt{n!}}{(n-1)!!}\prod_{m=1}^{\frac{j-2}{2}}(n+2m) \\
&= (2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{n+j-1}{2}}\frac{(n+j-2)!!}{\sqrt{n!}} \\
&= -(D_{n+1,n+1}r_{n+1})_j.
\end{aligned}$$

(ii) If n even, j odd and $j < n$, similarly, applying change of variable $l = \frac{i-1}{2}$ gives

$$(d_{n+1}^T R_n)_j \quad = \quad -\sqrt{-1}\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n}{2}}(\frac{n}{2})!}\sum_{l=0}^{\frac{n}{2}-1}(-1)^l\binom{\frac{n}{2}}{l}\prod_{m=1}^{\frac{j-1}{2}}(2l+2m-1)$$

$$
\begin{aligned}
&= \sqrt{-1}\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n}{2}}(\frac{n}{2})!}(-1)^{\frac{n}{2}}\prod_{m=1}^{\frac{j-1}{2}}(n+2m-1)\\
&= \frac{(-1)^{\frac{n+j+1}{2}}(2/\theta_1)^{\frac{j-1}{2}}(n+j-2)!!}{\sqrt{n!}}\\
&= -(D_{n+1,n+1}r_{n+1})_j.
\end{aligned}
$$

(b) Proof of $B_n = 1$: Equation (4.3.1) gives

$$
r_{n+1}^T d_{n+1} = \sum_{i=1,(n+i)\text{odd}}^{n} (2/\theta_1)^{\frac{n}{2}}\sqrt{n!}\frac{(-1)^{n+1+\frac{n+1+i}{2}}(n+i-2)!!}{(i-1)!(n+1-i)!!}.
$$

(i) If n odd, applying change of variable $l = \frac{i}{2} - 1$ and using Lemma C.1.8,

$$
\begin{aligned}
& r_{n+1}^T d_{n+1}\\
&= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!}\sum_{i=1,(i)\text{even}}^{n}\frac{(-1)^{\frac{n+1+i}{2}}(i-1)!!\prod_{m=1}^{\frac{n-1}{2}}(i+2m-1)}{(i-1)!(n+1-i)!!}\\
&= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}\sum_{l=0}^{\frac{n-1}{2}-1}\frac{(-1)^l(2l+1)!!}{(2l+1)!(n-1-2l)!!}\prod_{m=1}^{\frac{n-1}{2}}(2l+2m+1)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\sum_{l=0}^{\frac{n-1}{2}-1}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{n-1}{2}}(2l+2m+1)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\\
&\quad\times\left(\sum_{l=0}^{\frac{n-1}{2}}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{n-1}{2}}(2l+2m+1)-(-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{n-1}{2}}(n+2m)\right)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\left(\sum_{l=0}^{\frac{n-1}{2}}(-1)^l\binom{\frac{n-1}{2}}{l}(2l)^{\frac{n-1}{2}}-(-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{n-1}{2}}(n+2m)\right)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\left((-2)^{\frac{n-1}{2}}\left(\frac{n-1}{2}\right)!-(-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{n-1}{2}}(n+2m)\right)
\end{aligned}
$$

$$\begin{aligned} &= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!} - \frac{(2\theta_2)^{\frac{n}{2}}}{\sqrt{n!}}(2n-1)!! \\ &= \frac{1}{D_{n+1,n+1}} - D_{n+1,n+1}R_{n+1,n+1}. \end{aligned}$$

(ii) If n even, similarly, applying change of variable $l = \frac{i-1}{2}$ gives

$$\begin{aligned} r_{n+1}^T d_{n+1} &= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n}{2}}}{2^{\frac{n}{2}}(\frac{n}{2})!}\sum_{l=0}^{\frac{n}{2}-1}(-1)^l\binom{\frac{n}{2}}{l}\prod_{m=1}^{\frac{n}{2}}(2l+2m-1) \\ &= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n}{2}}}{2^{\frac{n}{2}}(\frac{n}{2})!}\left((-2)^{\frac{n}{2}}(\frac{n}{2})! - (-1)^{\frac{n}{2}}\prod_{m=1}^{\frac{n}{2}}(n+2m-1)\right) \\ &= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!} - \frac{(2/\theta_1)^{\frac{n}{2}}}{\sqrt{n!}}(2n-1)!! \\ &= \frac{1}{D_{n+1,n+1}} - D_{n+1,n+1}R_{n+1,n+1}. \end{aligned}$$

(c) Proof of $C_n = \mathbf{0}$:

$$\begin{aligned} R_n^T C_n &= r_{n+1} + R_n^T d_{n+1}(d_{n+1}^T r_{n+1} + D_{n+1,n+1}R_{n+1,n+1}) \\ &= r_{n+1} + R_n^T d_{n+1}\frac{B_n}{D_{n+1,n+1}} \\ &= \frac{1}{D_{n+1,n+1}}(D_{n+1,n+1}r_{n+1} + R_n^T d_{n+1}) \\ &= \frac{1}{D_{n+1,n+1}}A_n^T \\ &= \mathbf{0}. \end{aligned}$$

Since R_n is nonsingular, $C_n = \mathbf{0}$.

C.1.8 Proof of a statement in Section 3

We state and prove the following proposition.

Proposition C.1.9. *If $\{R_n\}$ is a sequence of nested positive definite matrices and let $R_n^{-1} = D_n^T D_n$ be the reverse Cholesky decomposition of R_n^{-1}, then $\{D_n\}$ is nested.*

Proof. Letting $R_n = C_n C_n^T$ be the Cholesky decomposition of R_n, then $D_n = C_n^{-1}$ since $R_n^{-1} = D_n^T D_n$ and D_n is lower triangular as required. Since $\{R_n\}$ is nested, by construction of the Cholesky decompostion, C_n is a nested sequence of lower triangular matrices. By inspection of the relationship $D_n C_n = I_n$ when both D_n and C_n are lower triangular, it is apparent that $\{D_n\}$ is a nested sequence of matrices. □

C.1.9 Proof of Theorem 4.3.3

Consider $k \geq i$ and $k + i$ is even (so that $k - i$ is also even and $d_{ki} \neq 0$). By Stirling's Approximation, as $k \to \infty$,

$$\begin{aligned}
d_{ki}^2 &= \frac{(k-1)!}{(2/\theta_1)^{i-1}\left((i-1)!\right)^2\left((k-i)!!\right)^2} \\
&= \frac{(k-1)!}{(2/\theta_1)^{i-1}\left((i-1)!\right)^2 2^{k-i}\left(\frac{k-i}{2}!\right)^2} \\
&\sim \frac{\theta_1^{i-1}}{\left((i-1)!\right)^2 2^{k-1}} \frac{\sqrt{2\pi(k-1)}(\frac{k-1}{e})^{k-1}}{\pi(k-i)(\frac{k-i}{2e})^{k-i}} \\
&\sim \frac{\theta_1^{i-1}}{\left((i-1)!\right)^2 2^{k-1}} \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{k}} 2^{k-i} e^{1-i} (k-1)^{i-1} (1 + \frac{i-1}{k-i})^{k-i} \\
&\sim \sqrt{\frac{2}{\pi}} \frac{k^{i-\frac{3}{2}} \theta_1^{i-1}}{2^{i-1}\left((i-1)!\right)^2}.
\end{aligned}$$

Since $f(x) = x^p$, the nth order derivatives when $x = 0$ are all 0 except for $n = p$, when it is $p!$. As $n \to \infty$,

$$\begin{aligned}
\hat{\theta}_0 &= \frac{1}{n} \|D(\theta_1, n) z\|^2 \\
&= \frac{(p!)^2}{n} \sum_{k=1}^{n} d_{k,p+1}^2 \\
&= \frac{(p!)^2}{n} \sum_{k=0}^{\left\lfloor \frac{n-p}{2} \right\rfloor} d_{p+1+2k,p+1}^2
\end{aligned}$$

$$
\begin{aligned}
&\sim \quad \frac{(p!)^2}{n} \sum_{k=0}^{\left\lfloor \frac{n-p}{2} \right\rfloor} \sqrt{\frac{2}{\pi}} \frac{(p+1+2k)^{p-\frac{1}{2}} \theta_1^p}{2^p (p!)^2} \\
&\sim \quad \sqrt{\frac{2}{\pi}} \frac{2^{p-\frac{1}{2}} \theta_1^p}{2^p} \frac{(\lfloor \frac{n-p}{2} \rfloor)^{p+\frac{1}{2}}}{n(p+\frac{1}{2})} \\
&\sim \quad \frac{n^{p-\frac{1}{2}} \theta_1^p}{\sqrt{2\pi} 2^p (p+\frac{1}{2})}
\end{aligned}
$$

where the first asymptotic equivalence follows from Lemma C.1.2.

C.1.10 Proof of a statement in Section 4

We state and prove the following proposition.

Proposition C.1.10. *For some $m > 1$, consider a $(2m-1) \times (2m-1)$ regular grid on $[0,1] \times [0,1]$. Observations $\boldsymbol{z}$ are taken on the $m \times m$ regular sub-grid. When $m = 12$, the setup is shown in Figure 4.4. Denote $p_{i,j}$ as the predictand at location (i, j) for some j odd, and $\hat{p}_{i,j}$ as the EBLP defined in (4.4.7), then*

$$
\hat{p}_{i,j} = v_j^T(\hat{\theta}_2) \boldsymbol{z}_{(j-1)m+1:jm},
$$

where $v_j^T(\hat{\theta}_2) = r_j(\hat{\theta}_2) R^{-1}(\hat{\theta}_2, m)$ for some $r_j(\hat{\theta}_2) \in \mathbb{R}^{1 \times m}$ depending only on $\hat{\theta}_2$. That is, $\hat{p}_{i,j}$ only depends on observations on the jth column of the grid, and the range parameter estimate along columns.

Proof. Note that the covariance of $p_{i,j}$ and the observations $\boldsymbol{z}$ is

$$
\mathrm{Cov}\left(p_{i,j}, \boldsymbol{z}^T\right) = \hat{\theta}_0 R(\hat{\theta}_1, m)_{j,.} \otimes r_j(\hat{\theta}_2),
$$

where $R(\hat{\theta}_1, m)_{j,.}$ is the jth row of $R(\hat{\theta}_1, m)$ and $r_j(\hat{\theta}_2)$ is the correlation of $p_{i,j}$ and observations on the jth column. Then we have

$$\begin{aligned}
\hat{p}_{i,j} &= \operatorname{Cov}\left(p_{i,j}, \boldsymbol{z}^T\right) \operatorname{Cov}\left(\boldsymbol{z}, \boldsymbol{z}^T\right)^{-1} \boldsymbol{z} \\
&= \left(R(\hat{\theta}_1, m)_{j,.} \otimes r_j(\hat{\theta}_2)\right) \left(R^{-1}(\hat{\theta}_1, m) \otimes R^{-1}(\hat{\theta}_2, m)\right) \boldsymbol{z} \\
&= \left(R(\hat{\theta}_1, m)_{j,.} R^{-1}(\hat{\theta}_1, m)\right) \otimes \left(r_j(\hat{\theta}_2) R^{-1}(\hat{\theta}_2, m)\right) \boldsymbol{z} \\
&= \left(\boldsymbol{e}_j^T \otimes r_j(\hat{\theta}_2) R^{-1}(\hat{\theta}_2, m)\right) \boldsymbol{z} \\
&= v_j^T(\hat{\theta}_2) \boldsymbol{z}_{(j-1)m+1:jm},
\end{aligned}$$

where $\boldsymbol{e}_j$ is the jth standard base and $v_j^T(\hat{\theta}_2) = r_j(\hat{\theta}_2) R^{-1}(\hat{\theta}_2, m)$. □

APPENDIX D

SUPPLEMENT TO CHAPTER 5

D.1 Supplementary figures

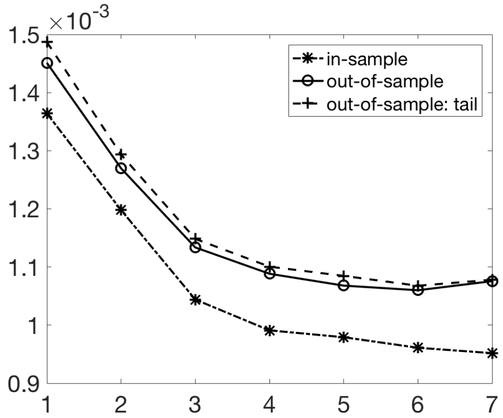

Figure D.1: RMSEs at each embedding length m for predicting the initial N observations (in-sample), the next N observations (out-of-sample) and the tail N observations (out-of-sample: tail) using the NN model. $N = 2000$, $m = 1, \ldots, 7$.

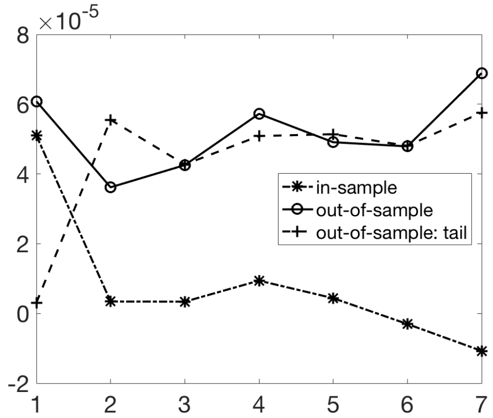

Figure D.2: Differences of RMSEs of the NN model minus those of the GP model at each embedding length m for predicting the initial N observations (in-sample), the next N observations (out-of-sample) and the tail N observations (out-of-sample: tail). $N = 2000$, $m = 1, \ldots, 7$.

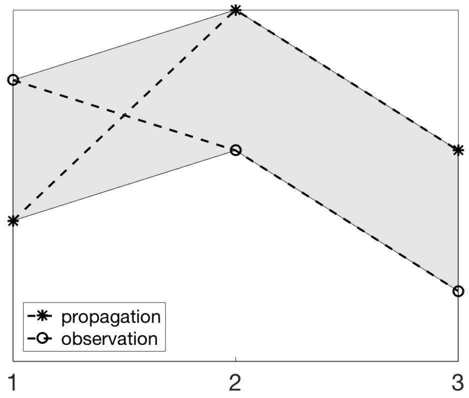

Figure D.3: Illustration of the area (shaded region) between model propagations in three steps and observations as defined by (5.3.3).

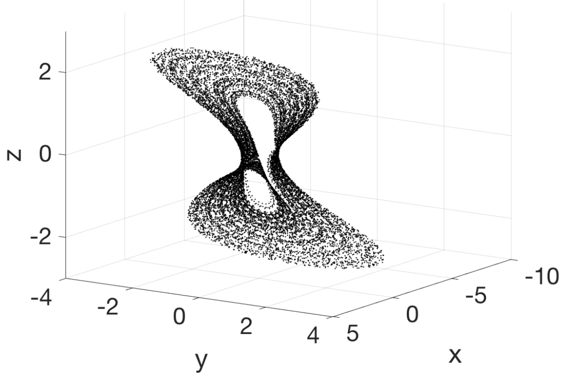

Figure D.4: Simulated observations from the ODE model M_P with no model error and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$.

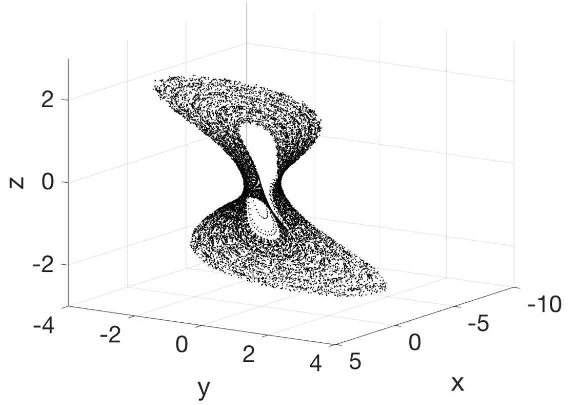

Figure D.5: Simulated observations from the ODE model M_P with model error $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I_3)$, $\sigma_\varepsilon = 10^{-12}$ and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$.

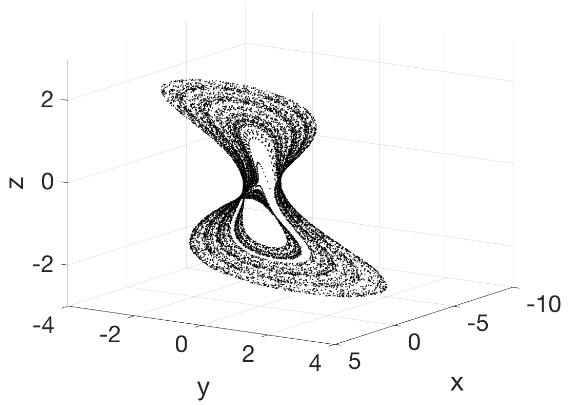

Figure D.6: Simulated observations from the NN model M_N with no model error and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$.

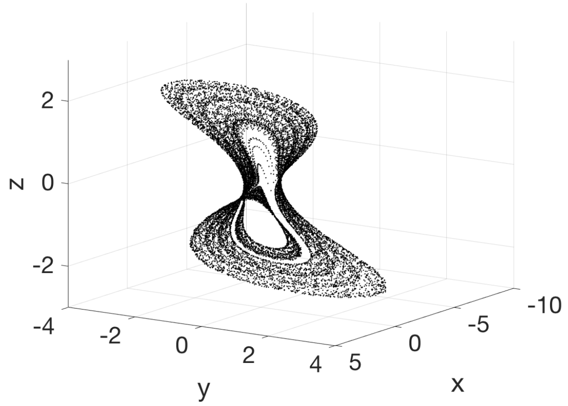

Figure D.7: Simulated observations from the NN model M_N with model error $\widetilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I_3)$ as defined in (5.4.2), $\sigma_\varepsilon = 10^{-12}$ and with observational error $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3)$, $\sigma_\eta = 10^{-3}$.

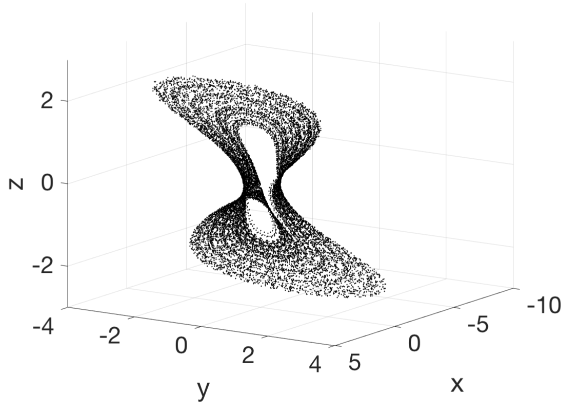

Figure D.8: Simulated observations from the ODE model M_P with no model error and with observational error $\eta_t = \phi\eta_{t-1} + v_t$, $\phi = 0.9$, $v_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_v^2 I_3)$, $\sigma_v = 10^{-3}$.

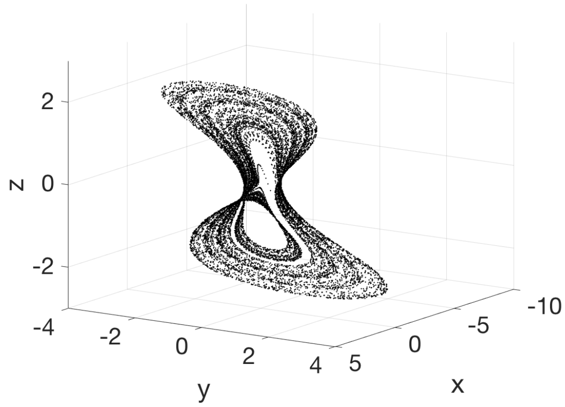

Figure D.9: Simulated observations from the NN model M_N with no model error and with observational error $\eta_t = \phi\eta_{t-1} + v_t$, $\phi = 0.9$, $v_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma_v^2 I_3)$, $\sigma_v = 10^{-3}$.

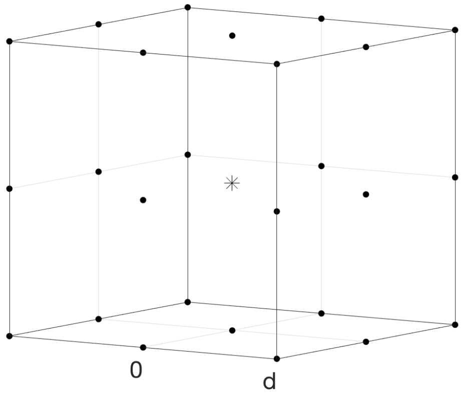

Figure D.10: The simulation points (dot) are taken on a regular grid centered at the true initial state for the ODE model, and at the last three elements of the true initial state for the NN model. The shortest distance from the simulation points to the true initial state (asterisk) is d.

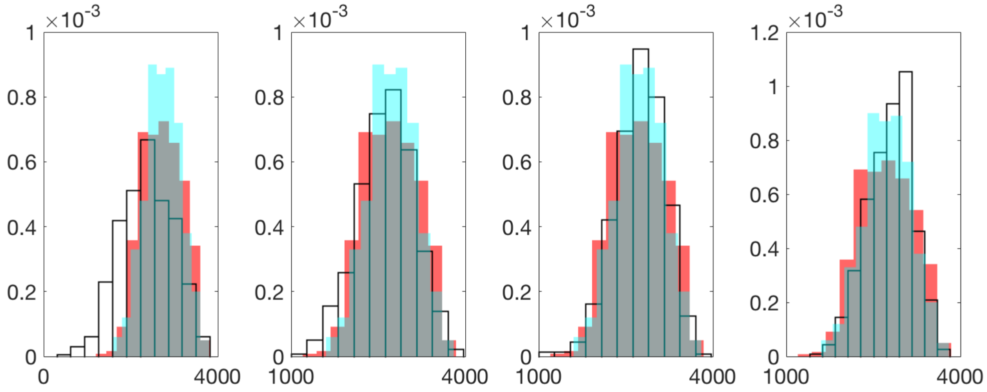

Figure D.11: Histograms of average areas between real observations and model propagations (red: NN with embedding length $m = 3$, blue: GP with $m = 2$) in 2000 steps for the 500 starting points considered in Section 5.3; histograms (white with solid bar outlines) of average areas between two realizations in 2000 steps from (5.4.2) under model error with standard deviation σ_ε and no observational error for the same 500 starting points. From left to right: $\sigma_\varepsilon = 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}$. Histograms are normalized so that the areas of bars in each histogram sum to one.

REFERENCES

[1] Henry Abarbanel. *Analysis of observed chaotic data.* Springer Science & Business Media, 2012.

[2] Henry DI Abarbanel, Reggie Brown, John J Sidorowich, and Lev Sh Tsimring. The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, 65(4):1331, 1993.

[3] Markus Abt and William J Welch. Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes. *Canadian Journal of Statistics*, 26(1):127–137, 1998.

[4] Anant Agarwal and Jeffrey Lang. *Foundations of analog and digital electronic circuits.* Morgan Kaufmann, 2005.

[5] George E Andrews. *The Theory of Partitions.* Number 2. Cambridge University Press, Cambridge, UK, 1998.

[6] Ioannis Andrianakis and Peter G Challenor. The effect of the nugget on Gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12):4215–4228, 2012.

[7] Mihai Anitescu, Xiaoyan Zeng, and Emil M Constantinescu. A low-memory approach for best-state estimation of hidden Markov models with model error. *SIAM Journal on Numerical Analysis*, 52(1):468–495, 2014.

[8] A. Apte, D. Auroux, and M. Ramaswamy. Variational data assimilation for discrete Burgers equation. In *Electronic Journal of Differential Equations*, volume 19, pages 15–30, 2010.

[9] François Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013.

[10] NJ Balmforth and RV Craster. Synchronizing moore and spiegel. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 7(4):738–752, 1997.

[11] Clayton Barrows, Marissa Hummon, Wesley B Jones, and Elaine T Hale. *Time Domain Partitioning of Electricity Production Cost Simulations*. National Renewable Energy Laboratory, 2014.

[12] John R Barry, Edward A Lee, and David G Messerschmitt. *Digital communication*. Springer Science & Business Media, 2012.

[13] Matthew H Bassett, Joseph F Pekny, and Gintaras V Reklaitis. Decomposition techniques for the solution of large-scale scheduling problems. *AIChE Journal*, 42(12):3373–3387, 1996.

[14] BJ Bayly, I Goldhirsch, and Steven A Orszag. Independent degrees of freedom of dynamical systems. *Journal of Scientific Computing*, 2(2):111–121, 1987.

[15] Mark Hudson Beale, Martin T Hagan, and Howard B Demuth. Neural network toolbox user's guide. *The Mathworks Inc*, 1992.

[16] Robert A Beeler. *How to Count: An Introduction to Combinatorics and Its Applications*. Springer, Cham, Switzerland, 2015.

[17] Thomas Bengtsson, Peter Bickel, Bo Li, et al. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and statistics: Essays in honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics, 2008.

[18] L Mark Berliner. Likelihood and bayesian prediction of chaotic systems. *Journal of the American Statistical Association*, 86(416):938–952, 1991.

[19] PE Berry and RM Dunnett. Contingency constrained economic dispatch algorithm for transmission planning. In *Generation, Transmission and Distribution, IEE Proceedings C*, volume 136, pages 238–244. IET, 1989.

[20] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific, Belmont, MA, 1995.

[21] Lorenz T Biegler and Victor M Zavala. Large-scale nonlinear programming using Ipopt: An integrating framework for enterprise-wide dynamic optimization. *Computers & Chemical Engineering*, 33(3):575–582, 2009.

[22] J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[23] FH Branin and SK Hoo. A method for finding multiple extrema of a function of n variables. *Numerical Methods*, pages 231–237, 1972.

[24] Matt Bromberg, Tsu-Shuan Chang, and PB Luh. Decomposition and coordination for non-convex optimal control problems with parallel algorithm. In *Decision and Control, 1987. 26th IEEE Conference on*, volume 26, pages 1468–1475. IEEE, 1987.

[25] Paul Bryant, Reggie Brown, and Henry DI Abarbanel. Lyapunov exponents from observed time series. *Physical Review Letters*, 65(13):1523, 1990.

[26] Arthur Earl Bryson. *Dynamic optimization*, volume 1. Prentice Hall, 1999.

[27] Martin Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35(3):335–356, 1989.

[28] Tsu-Shuan Chang, Shi-Chung Chang, and Peter B Luh. A hierarchical decomposition for large scale optimal control problems with parallel processing capability. In *American Control Conference, 1986*, pages 1995–2000. IEEE, 1986.

[29] Isabelle Charpentier, Claude Dal Cappello, and Jean Utke. Efficient higher-order derivatives of the hypergeometric function. In *Advances in Automatic Differentiation*, pages 127–137. Springer, Berlin, Germany, 2008.

[30] Haiyan Cheng and Adrian Sandu. Collocation least-squares polynomial chaos method. In *Proceedings of the 2010 Spring Simulation Multiconference*. Society for Computer Simulation International, 2010.

[31] Stephen E Cohn and David F Parrish. The behavior of forecast error covariances for a kalman filter in two dimensions. *Monthly Weather Review*, 119(8):1757–1785, 1991.

[32] Stefano Conti, John Paul Gosling, Jeremy E Oakley, and Anthony O'hagan. Gaussian process emulation of dynamic computer codes. *Biometrika*, 96(3):663–676, 2009.

[33] John L Crassidis and John L Junkins. *Optimal estimation of dynamic systems*. CRC press, 2011.

[34] James P Crutchfield and Bruce S McNamara. Equations of motion from a data series. *Complex Systems*, 1:417–452, 1987.

[35] Predrag Cvitanović. Invariant measurement of strange sets in terms of cycles. *Physical Review Letters*, 61(24):2729, 1988.

[36] Roger Daley. *Atmospheric data analysis*. Number 2. Cambridge University Press, 1993.

[37] DP Dee. Testing the perfect-model assumption in variational data assimilation. In *Proc. Second Int. Symp. on Assimilation of Observations in Meteorology and Oceanography*, pages 225–228. Citeseer, 1995.

[38] François-xavier Le Dimer and Olivier Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus A*, 38(2):97–110, 1986.

[39] A. Doucet and A.M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *The Oxford Handbook of Nonlinear Filtering, Oxford University Press.*, 2009.

[40] A Elkamel, M Zentner, F Pekny, and GV Reklaitis. A decomposition heuristic for scheduling the general batch chemical plant. *Engineering Optimization+ A35*, 28(4):299–330, 1997.

[41] M Engelhardt. LTSpice/SwitcherCAD IV. *Linear Technology Corporation*, 2011.

[42] Jacob Engwerda. *LQ dynamic optimization and differential games.* John Wiley & Sons, 2005.

[43] J Doyne Farmer and John J Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59(8):845, 1987.

[44] Alexander Forrester, Andras Sobester, and Andy Keane. *Engineering design via surrogate modeling: a practical guide.* John Wiley & Sons, 2008.

[45] Richard Franke. A critical comparison of some methods for interpolation of scattered data. Technical report, DTIC Document, 1979.

[46] Carlos E Garcia, David M Prett, and Manfred Morari. Model predictive control: theory and practice – a survey. *Automatica*, 25(3):335–348, 1989.

[47] James Glimm, Shuling Hou, Yoon-Ha Lee, David H Sharp, and Kenny Ye. Sources of uncertainty and error in the simulation of flow in porous media. *Computational & Applied Mathematics*, 23(2-3):109–120, 2004.

[48] Facundo A Gómez, Christopher E Coleman-Smith, Brian W O'Shea, Jason Tumlinson, and Robert L Wolpert. Characterizing the formation history of milky way like stellar halos with model emulators. *The Astrophysical Journal*, 760(2):112, 2012.

[49] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[50] Henry Gould and Jocelyn Quaintance. Double fun with double factorials. *Mathematics Magazine*, 85(3):177–192, 2012.

[51] Andreas Griewank. Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation. *Optimization Methods and Software*, 1(1):35–54, 1992.

[52] Andreas Griewank, Jean Utke, and Andrea Walther. Evaluating higher derivative tensors by forward propagation of univariate Taylor series. *Mathematics of Computation*, 69(231):1117–1130, 2000.

[53] Andreas Griewank and Andrea Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, 2008.

[54] Ignacio Grossmann. Enterprise-wide optimization: A new frontier in process systems engineering. *AIChE Journal*, 51(7):1846–1857, 2005.

[55] Guoxiang Gu. *Discrete-Time Linear Systems: Theory and Design with Applications*. Springer Science & Business Media, 2012.

[56] John Guckenheimer and Philip Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42. Springer Science & Business Media, 2013.

[57] Ben Haaland, Peter ZG Qian, et al. Accurate emulators for large-scale computer experiments. *The Annals of Statistics*, 39(6):2974–3002, 2011.

[58] William W. Hager. Lipschitz continuity for constrained processes. *SIAM Journal on Control and Optimization*, 17(3):321–338, 1979.

[59] Morris W Hirsch, Stephen Smale, and Robert L Devaney. *Differential equations, dynamical systems, and an introduction to chaos.* Academic press, 2012.

[60] Paul Horowitz and Winfield Hill. *The art of electronics.* Cambridge Univ. Press, 1989.

[61] Wanzhen Huang and William J Welch. Properties of parameters in a stochastic process model for computer experiments. Unpublished M. Math thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, 2000.

[62] PJM Interconnection. Estimated hourly load data. `http://www.pjm.com/markets-and-operations/energy/real-time/loadhryr.aspx`. Accessed: 2016-05-28.

[63] Jennifer R Jackson and Ignacio E Grossmann. Temporal decomposition scheme for nonlinear multisite production planning and distribution models. *Industrial & engineering chemistry research*, 42(13):3045–3055, 2003.

[64] Andrew H Jazwinski. *Stochastic processes and filtering theory.* Courier Corporation, 2007.

[65] John Bertrand Johnson. Thermal agitation of electricity in conductors. *Physical Review*, 32(1):97, 1928.

[66] Kevin Judd and Alistair Mees. Modeling chaotic motions of a string from experimental data. *Physica D*, 92(3):221–236, 1996.

[67] Rudolf Kalman. On the general theory of control systems. *IRE Transactions on Automatic Control*, 4(3):110–110, 1959.

[68] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[69] Eugenia Kalnay. *Atmospheric modeling, data assimilation, and predictability.* Cambridge University Press, 2003.

[70] SS a Keerthi and Elmer G Gilbert. Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving-horizon approximations. *Journal of optimization theory and applications*, 57(2):265–293, 1988.

[71] Javad Khazaei, Golbon Zakeri, and Shmuel S Oren. Market clearing mechanisms under uncertainty. Technical report, Working Paper, Princeton University, University of Auckland, University of California at Berkeley, 2014.

[72] Wook Hyun Kwon and Soo Hee Han. *Receding horizon control: model predictive control for state models.* Springer Science & Business Media, 2006.

[73] William Lahoz, Boris Khattatov, and Richard Menard. *Data assimilation: making sense of observations.* Springer Science & Business Media, 2010.

[74] Steven P Lalley. Beneath the noise, chaos. *Annals of Statistics*, 27(2):461–479, 1999.

[75] Steven P Lalley. Removing the noise from chaos plus noise. In *Nonlinear Dynamics and Statistics*, pages 233–244. Springer, 2001.

[76] A Lapedes and R Farber. Nonlinear signal processing using neural network: Prediction and system modeling. los alamos nat. Technical report, Lab, Technical Report. LA-UR-872662, 1987.

[77] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 15. SIAM, 1995.

[78] John M Lewis, Sivaramakrishnan Lakshmivarahan, and Sudarshan Dhall. *Dynamic data assimilation: A least squares approach*, volume 13. Cambridge University Press, 2006.

[79] Yong B Lim, Jerome Sacks, WJ Studden, and William J Welch. Design and analysis of computer experiments when the output is highly correlated over the input space. *Canadian Journal of Statistics*, 30(1):109–126, 2002.

[80] Magnus Lindskog, Dick Dee, Yannick Tremolet, Erik Andersson, Gabor Radnoti, and Mike Fisher. A weak-constraint four-dimensional variational analysis system in the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, 135(640):695–706, 2009.

[81] Wei-Liem Loh and Tao-Kai Lam. Estimating structured correlation matrices in smooth Gaussian random field models. *The Annals of Statistics*, 28(3):880–904, 2000.

[82] Edward N Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646, 1969.

[83] Miles Lubin and Iain Dunning. Computing in operations research using Julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.

[84] Reason L Machete. *Modeling a Moore-Spiegel Electronic Circuit: the imperfect model scenario.* PhD thesis, University of Oxford, 2007.

[85] Andrew J Majda, John Harlim, and Boris Gershgorin. Mathematical strategies for filtering turbulent dynamical systems. *Discrete Contin. Dyn. Syst*, 27(2):441–486, 2010.

[86] Jay D Martin and Timothy W Simpson. On the use of kriging models to approximate deterministic computer models. In *ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 481–492. American Society of Mechanical Engineers, 2004.

[87] MJ Martin, MJ Bell, and Nancy K Nichols. Estimation of systematic error in an

equatorial ocean model using data assimilation. *International Journal for Numerical Methods in Fluids*, 40(3-4):435–444, 2002.

[88] Jacob Mattingley, Yang Wang, and Stephen Boyd. Receding horizon control. *IEEE Control Systems*, 31(3):52–65, 2011.

[89] Peter S Maybeck. *Stochastic models, estimation, and control*, volume 3. Academic press, 1982.

[90] Derek W Moore and Edward A Spiegel. A thermally excited non-linear oscillator. *The Astrophysical Journal*, 143:871, 1966.

[91] Francisco D Munoz, Enzo E Sauma, and Benjamin F Hobbs. Approximations in power transmission planning: implications for the cost and performance of renewable portfolio standards. *Journal of Regulatory Economics*, 43(3):305–338, 2013.

[92] Laurence William Nagel and Donald O Pederson. *SPICE: Simulation program with integrated circuit emphasis.* Electronics Research Laboratory, College of Engineering, University of California, 1973.

[93] IM Navon. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. *Dynamics of Atmospheres and Oceans*, 27(1):55–79, 1998.

[94] Jorge Nocedal and Stephen Wright. *Numerical optimization.* Springer Science & Business Media, 2006.

[95] Jeremy Oakley. Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society: Series C*, 53(1):83–93, 2004.

[96] Olga Obrezanova, Gábor Csányi, Joelle MR Gola, and Matthew D Segall. Gaussian processes: A method for automatic QSAR modeling of ADME properties. *Journal of Chemical Information and Modeling*, 47(5):1847–1857, 2007.

[97] David Orrell, L Smith, J Barkmeijer, and TN Palmer. Model error in weather forecasting. *Nonlinear processes in geophysics*, 8(6):357–371, 2001.

[98] TN Palmer, GJ Shutts, R Hagedorn, FJ Doblas-Reyes, Thomas Jung, and M Leutbecher. Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences*, 33:163–193, 2005.

[99] Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013.

[100] A Pikovsky. Discrete-time dynamic noise filtering. *Sov J Commun Technol Electron*, 31:81, 1986.

[101] Planning Division Transmission Planning Department. PJM region transmission planning process. Manual PJM Manual 14B, PJM, 2016.

[102] Michael JD Powell. Radial basis functions for multivariable interpolation: a review. In *Algorithms for approximation*, pages 143–167. Clarendon Press, 1987.

[103] Pritam Ranjan, Ronald Haynes, and Richard Karsten. A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4):366–378, 2011.

[104] Jerome Sacks, Susannah B Schiller, and William J Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989.

[105] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical Science*, pages 409–423, 1989.

[106] Masaki Sano and Yasuji Sawada. Measurement of the lyapunov spectrum from a chaotic time series. *Physical Review Letters*, 55(10):1082, 1985.

[107] Thomas J Santner, Brian J Williams, and William I Notz. *The design and analysis of computer experiments*. Springer Science & Business Media, 2013.

[108] Dan Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[109] Timothy W Simpson, Timothy M Mauery, John J Korte, and Farrokh Mistree. Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA Journal*, 39(12):2233–2241, 2001.

[110] Steve Smale. Mathematical problems for the next century. *The Mathematical Intelligencer*, 20(2):7–15, 1998.

[111] Leonard A Smith, C Ziehmann, and K Fraedrich. Uncertainty dynamics and predictability in chaotic systems. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2855–2886, 1999.

[112] Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

[113] Qinghu Tang, Ying Bin Lau, Shuangquan Hu, Wenjin Yan, Yanhui Yang, and Tao Chen. Response surface methodology using Gaussian processes: Towards optimizing the trans-stilbene epoxidation over Co 2+–NaX catalysts. *Chemical Engineering Journal*, 156(2):423–431, 2010.

[114] Yannick Trémolet. Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 132(621):2483–2504, 2006.

[115] Yannick Trémolet. Model-error estimation in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1267–1280, 2007.

[116] Francesco Uboldi and Masafumi Kamachi. Time-space weak-constraint data assimilation for nonlinear models. *Tellus A*, 52(4):412–421, 2000.

[117] Mathias Wagner, Andrea Walther, and Bernd-Jochen Schaefer. On the efficient computation of high-order derivatives for implicitly defined functions. *Computer Physics Communications*, 181(4):756–764, 2010.

[118] Alan Wolf, Jack B Swift, Harry L Swinney, and John A Vastano. Determining lyapunov exponents from a time series. *Physica D*, 16(3):285–317, 1985.

[119] Stephen Wolfram. *The MATHEMATICA® Book, version 4*. Cambridge University Press, Cambridge, UK, 1999.

[120] Wanting Xu and Mihai Anitescu. A limited-memory multiple shooting method for weakly constrained variational data assimilation. *SIAM Journal on Numerical Analysis*, 54(6):3300–3331, 2016.

[121] Wanting Xu and Michael L Stein. Maximum likelihood estimation for a smooth gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):138–175, 2017.

[122] Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

[123] Wei Zhang, Alessandro Abate, and Jianghai Hu. Efficient suboptimal solutions of switched LQR problems. In *American Control Conference, 2009. ACC'09.*, pages 1084–1091. IEEE, 2009.

[124] Wei Zhang, Jianghai Hu, and Alessandro Abate. On the value functions of the discrete-time switched LQR problem. *IEEE Transactions on Automatic Control*, 54(11):2669–2674, 2009.

[125] Wei Zhang, Jianghai Hu, and Alessandro Abate. A study of the discrete-time switched LQR problem. Technical report, Paper 384, Purdue University, 2009.

[126] Dusanka Zupanski. A general weak constraint applicable to operational 4dvar data assimilation systems. *Monthly Weather Review*, 125(9):2274–2292, 1997.

[127] Milija Zupanski, Dusanka Zupanski, Tomislava Vukicevic, Kenneth Eis, and Thomas Vonder Haar. CIRA/CSU four-dimensional variational data assimilation system. *Monthly Weather Review*, 133(4):829–843, 2005.