THE UNIVERSITY OF CHICAGO

IDENTIFICATION OF DYNAMICAL SYSTEMS: IDENTIFIABILITY TO STOCHASTIC OPTIMIZATION

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

DEPARTMENT OF STATISTICS

BY

VIVAK PATEL

CHICAGO, ILLINOIS

AUGUST 2018

To my grandparents, Somabhai and Madhukanta Patel

To my parents, Rajendra and Praghna Patel

To my sister, Pooja Patel

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

First and foremost, I am deeply indebted to Mihai Anitescu. His intellectual prowess, work ethic, affability, and keen emotional intelligence taught me how to be a better scholar, a better colleague, and a better leader. This work would have never been possible without his measured honesty about my progress, patience and encouragement for my own exploration, and his unwavering challenge for me to be better. I thank Dan Nicolae, who has been supportive and generous to me with his time and friendship. His cheerfulness at both six in the morning and six at night have always been some of the best moments of my days at the University. I also thank Rina Foygel Barber, whose kindness and leniency towards me have been unquestionably undeserved, and has, by example, given me an unsurpassed role model for being an academic.

While I am grateful for all of my classmates, I especially thank Mathiyar Bonakdarpour and Chris McKennan with whom my daily interactions always brought to light new ideas and questions, and for encouraging me to not eat lunch over my notebooks at least once a week. I thank Mohammad Jahangoshahi who suffered with me through courses, who always sincerely considered my crazy ideas, and who has taught me much. It has been my deepest honor to count him as one of my dearest friends.

Moreover, as the last few weeks before writing this have demonstrated, I have made countless friends that have gotten me to the end of this journey. In particular, I thank Daniel Adrian Maldonado, who started out as a collaborator and grew into a close friend. I also thank Ram Sabaratnam, whose quirkiness, cheer and patience has helped me mend from my knee surgery. I thank Ilana Ventura for her unwavering support and fierce affection. She has seen me through my toughest times over the past five years, and for this I am eternally grateful.

Finally, I thank my family. My grandparents, my parents and my sister have been my

greatest teachers and they have all sacrificed immeasurably for my successes. For this reason, I attribute my best qualities to them, and I attribute my worst qualities to not listening to them.

Abstract

The central theme of this thesis is to understand two related question. When can a differential system model be identifiable from observations? If the model is identifiable, how can we identify it practically? While these questions are by no means new, we study them in a modern context where systems and models are more complex, observations are more frequent, and the stochastic nature of the underlying phenomenon must be considered. Chapter 1 discusses the nuances of these two questions in this modern context. Chapters 2 and 3 delve into first question by refining notions of identifiability and by contributing necessary conditions for identifiability of certain differential equation models. Chapters 4 to 7 delve into the second question from the perspective of designing computable estimators to handle the higher frequency of observations. Chapter 8 also addresses the second question by designing a novel optimization framework to address phenomenon with a stochastic nature.

1 | Introduction

From biology to engineering, differential equation models are a common tool for understanding time-varying phenomena. Moreover, just as many other models in science or engineering, differential equation models will have unknown parameters that must be inferred from observation of the phenomenon. Thus, as a special case of model inference, differential equation model inference is subject to standard mathematical and statistical questions, such as:

1. (Identifiability) Is it possible to determine the unknown parameter from the given observations or an idealized version of the observations?

2. (Estimation) What is an appropriate estimator for the unknown parameter? What are the properties of the estimator?

3. (Computation) Given actual measurements, how can we practically compute the estimator for the unknown parameter?

However, owing to the additional structure and nuances of differential equation models, unique opportunities and challenges arise in the study of identifiability, estimation and computation for differential equation model inference. The following three examples illustrate these different and unique aspects of differential equation models.

Plasma Cholesterol Kinetics. In a study by Goodman and Noble (1968), a two-dimensional dynamical system model,

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -k_{xx} & k_{yx} \\ k_{xy} & -k_{yy} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} s_x \\ s_y \end{bmatrix}, \tag{1.1}$$

was used to understand blood plasma cholesterol kinetics in patients, where x and y represent the concentrations of cholesterol in two groups of bodily tissue; k_{xx} and k_{yy} represent the clearance rate of cholesterol by the tissue groups; k_{xy} and k_{yx} represent the rate of cholesterol

exchange between the two tissue groups; and s_x and s_y represent the uptake rate of cholesterol by the tissue groups from the blood. The plasma cholesterol kinetics model is illustrated in Fig. 1.1.

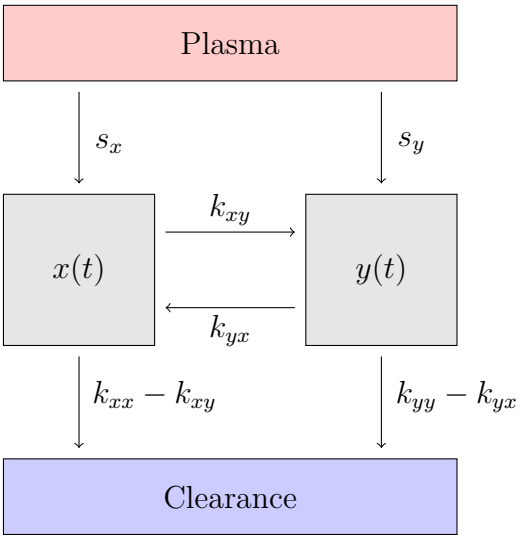


Figure 1.1: Diagram of cholesterol kinetics between blood plasma and two tissue groups.

This dynamical system model of plasma cholesterol kinetics is particularly useful: a completely specified model for a given patient can be used as a non-invasive diagnostic tool to determine potential pathologies in the two tissue groups or in the plasma. However, while the model is useful, the model does not reflect an actual physical system, but rather an abstraction of the physical system based on key biological functions. As a result, the model's parameters, $\{k_{xx}, k_{yy}, k_{xy}, k_{yx}, s_x, s_y\}$, cannot be directly measured from a patient.

Although the parameters cannot be measured directly, the sum of the concentrations of the two groups, $x(t) + y(t)$, can be measured by taking the difference in radioactively-tagged cholesterol concentrations in an initial intravenous injection and periodic blood samples. The ability to measure some information about the system raises an important question: given the measurements of the concentrations, is it possible to infer the model parameters? In fact, this question can be idealized: given ideal measurements (i.e., noise-free measurements of $x(t) + y(t)$ for all $t \in [0, T]$, where $T > 0$), is there only one set of parameters that produces

the given measurements?

In statistics, the study of this question is known as identifiability, and the statistics literature provides general, sufficient conditions that can answer this question affirmatively (see Van der Vaart, 1998, Ch. 5). Unfortunately, the sufficient conditions in the statistical literature are mostly of theoretical value, as they are usually not computable. On the other hand, as we will show, identifiability conditions for differential equation models can be made practical owing to the additional structure in the model, and these conditions can be used for planning sensor systems (i.e., design of experiments), determining regularization schemes, and for justifying estimation and computation schemes for determining the parameter.

As the preceding discussion suggests, the identifiability of differential equation models can be viewed as an extension of existing results in the statistics literature. However, this is not a complete picture: as the next example shows, the identifiability of differential equation models often has features that are not typically considered in standard statistical models.

Power System Frequency Stability. Power system operators are required to maintain the frequency and voltage of the power system within exceptionally tight margins. Indeed, even minor persistent deviations can result in unexpected outcomes. For example, the persistent frequency deviation in the Continental European Power System, caused by political conflicts in the Balkans, has, almost comically, resulted in many Continental European clocks being behind by almost six minutes (ENTSOE, 2018). However, frequency instability is often more sinister, and can cause more catastrophic problems (Newell et al., 2015).

Importantly, as distributed renewable energy generators (e.g., solar panels, wind farms) become more prevalent, maintaining frequency stability is becoming a pressing issue (Newell et al., 2015). To understand why, note that the frequency of the system is established by the rotation of generators at a certain angular frequency, and, during events such as faults, is maintained by keeping a certain amount of inertia, contributed by these rotating generators, in the system. However, as renewable energy generators either do not contribute inertia to

the system or contribute an unknown (virtual) inertia to the system (Tamrakar et al., 2017), the total inertia is now an unknown parameter of the power system. Moreover, because renewable generators are usually distributed and privately owned, the inertia contribution of the individual generators might not be directly measurable.

Thus, just as in the plasma cholesterol kinetics example, the inertia of the system must be determined from measurements of the state dynamics of the system, which is characterized by a differential equation model,

$$\dot{x} = f(t, x, \theta, \eta), \tag{1.2}$$

where x is the state of the system (e.g., rotor angles, rotor speeds, potentials and currents); θ is a parameter (e.g., system inertias, initial state); and $\{\eta(t) : t \in [0, T]\}$ represents an input into the system, such as energy usage by end users, which is also called load (Wang and Sun, 2015).

Again, just as in the plasma cholesterol kinetics example, a natural question to ask is: given an idealized version of the measurements, it is possible to infer the model parameters? However, unlike the plasma cholesterol kinetics example, the power system model has a (potentially) time-varying input $\{\eta(t) : t \in [0, T]\}$, and the nature of this input will drastically change how we formalize the question of identifiability. Owing to these inputs, the identifiability of differential equation models and the notions of identifiability commonly considered in the statistics literature no longer coincide, and the disciplines diverge. However, even in the differential equation identifiability literature, the question of how the inputs impact identifiability has been scantly explored. Accordingly, the first part of this thesis is on the identifiability of differential equation models and the role that inputs play in identifiability.

Once the question of identifiability is settled, the next natural question is to design an estimator for the unknown parameters and understand the properties of these estimators. For the power system model and differential equation models in general, a standard estimator

is the maximum (quasi) likelihood estimator, $\hat{\theta}$, which solves

$$\begin{aligned} \min_{\theta} \sum_{i=1}^{N} \|y_i - Ox(t_i)\|_2^2 \\ \dot{x} = f(t, x, \theta, \eta) \\ x(0) = \tilde{x}(\theta), \end{aligned} \tag{1.3}$$

where O is a matrix, called the measurement or observation operator; $\{y_i : i = 1, \ldots, N\}$, are measurements taken at equally spaced time points $\{t_i : i = 1, \ldots, N\}$ on the interval $[0, T]$ for $T > 0$ such that

$$y_i = Ox(t_i) + \epsilon_i, \tag{1.4}$$

where $\{\epsilon_i : i = 1, \ldots, N\}$ are independent and identically distributed with $\mathbb{E}[\epsilon_1] = 0$ and $\mathbb{V}[\epsilon_1] = \sigma^2 I$ for some $\sigma^2 > 0$. Additionally, for this standard estimator, consistency and asymptotic normality can be readily established using the usual approaches for M-estimators with some modifications (see Ljung, 1999, Ch. 8 & 9).[1]

Moreover, to compute $\hat{\theta}$, a standard gradient-based, iterative optimization routines can be used for which the gradient is computed using the adjoint differential system (see Cao et al., 2003). The basic iteration for such an optimization procedure requires the following steps. First, the differential system is numerically integrated with step size $h_f \leq T/N$ to compute the objective function, and to compute the intermediate states $\{x(t_i) : i = 1, \ldots, N\}$, which must be stored.[2] Then, using the intermediate states, the adjoint system can be defined and numerically integrated with a step size of $h_b \leq T/N$ to compute the gradient of the objective function. Using the gradient information, a new iterate can be generated.

The expensive parts of the optimization are integrating the forward system (i.e., the

[1] Note, Ljung (1999) considers the case where t_i are equally spaced over the entire non-negative real line. However, with some care, the results of Ljung (1999) can be extended to the case here where the time interval is fixed, and the density of measurements in the interval increases.

[2] It is also possible to store fewer states and then recalculate the states that were not stored as they are needed.

differential equations model) to evaluate the objective function, and integrating the backward system (i.e., the adjoint differential equations model) to evaluate the gradient. Moreover, as the forward and backward step size is bounded by T/N, the computational costs grow as the density of measurements increases (i.e., N is, by some metric, large), even for basic integration schemes such as Forward Euler. Indeed, for power systems, which are poised to be saturated with high-frequency measurement devices such as synchrophasors and automated metering devices, how to effectively and quickly estimate the unknown parameters with all of the available data is a realistic challenge.[3] Accordingly, the second part of this thesis is on estimators uniquely suited to this challenge — called incremental estimators or stochastic gradient methods — and their statistical and optimization properties.

As the preceding discussion suggests, with our identifiability work and our incremental estimator work, we can state and compute a valid estimator for the parameters of a dynamical system with an observed input, $\{\eta(t) : t \in [0, T]\}$. However, as the next example shows, the input, $\{\eta(t) : t \in [0, T]\}$, is not always observed and may only be characterized through a stochastic process model, in which case, alternative estimators must be stated and computed.

Seismically Retrofitted Systems. In the design of large buildings or critical structures such as bridges or roadways, engineers must design systems that can withstand seismic events or hazardous weather events. To ensure the safety and stability of these structures, engineers employ seismic retrofitting techniques. Typically, a seismic retrofitting technique's efficacy is determined in a laboratory setting by testing the technique to failure (ElGawady et al., 2004); however, the efficacy of a seismic retrofitting technique that is currently deployed in a bridge or building cannot be tested in the same way without damaging the structure.

As a realistic alternative, the efficacy of a seismic retrofitting technique can be evaluated

[3] This problem becomes more apparent for parameter tracking. Traditionally, when we do not pose a model for the parameter dynamics, parameter tracking requires the assumption that the parameter is fixed over some interval with a sufficiently high density of measurements. However, as the frequency of measurements increases, the assumption on the parameter will no longer be driven by the measurement frequency, but rather the needs of the applications, which may only require updated parameter estimates, say, every 15 seconds. In this case, we are now in the setting just described.

by inference through a dynamical systems model for the structure, such as the (simplified) vibrator-oscillator system depicted in Fig. 1.2, whose dynamics are described by

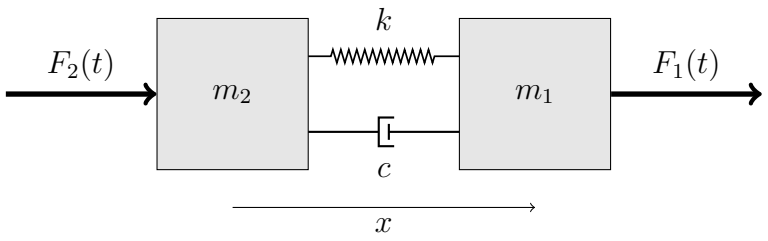


Figure 1.2: A diagram of the vibration-oscillator system.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & m_1 & 0 \\ 0 & 0 & 0 & m_2 \end{bmatrix} \begin{bmatrix} \dot{p}_1 \\ \dot{p}_2 \\ \dot{v}_1 \\ \dot{v}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k & k & -c & c \\ k & -k & c & -c \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ v_1 \\ v_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} F_1(t) \\ F_2(t) \end{bmatrix}, \tag{1.5}$$

where m_1 is the mass of the retrofitting technique; m_2 is the mass of the structure; the state variables p_1 and p_2 represent the positions (in the x direction) of masses m_1 and m_2, respectively; the state variables v_1 and v_2 represent the velocities of masses m_1 and m_2, respectively; the parameters k and c are the spring and damping constants of the technique; and $F_1(t)$ and $F_2(t)$ are the typical, day-to-day forces acting on the system such minor vibrations caused by passing vehicles or wind on the structure (see Xie, 2010, Ch. 10).

Thus, just as for the preceding two examples, a natural question is whether or not the parameters, k and c, can be determined from observations of the system state. However, unlike the preceding two examples, this question is complicated by the fact that the inputs are best thought of as stochastic processes. For example, the incident wind velocity on the structure can be observed at certain points, but the entire incident wind velocity field on the structure is typically not observable. Moreover, owing to the turbulent and stochastic nature of wind, the incident wind velocity field is best modeled as a stochastic process. To reiterate, in the seismic retrofitting example, the question of identifiability is complicated by

the stochastically-modeled input to the system. Accordingly, the identifiability of dynamical systems with stochastic inputs will be addressed in the first part of this thesis.

Along with identifiability, we must also consider estimation and computation of the unknown parameters of the differential equation model with stochastic inputs. Owing to the stochastic inputs, we no longer have a clear notion of a likelihood nor, in turn, a clear notion of the maximum likelihood estimator. Thus, we will choose an estimator depending on the needs of the application. For example, we may choose a robust estimator, which solves

$$\begin{aligned} \min_{\theta} \ \max_{\eta \in \mathfrak{N}} & \sum_{i=1}^{N} \|y_i - Ox(t_i)\|_2^2 \\ & \dot{x} = f(t, x, \theta, \eta) \\ & x(0) = \tilde{x}(\theta), \end{aligned} \tag{1.6}$$

where $\mathfrak{N}$ is a subset of the space of input functions. However, one of the challenges with (1.6) is that we do not take into account the stochastic process model for η.

Thus, we may choose an alternative estimator that does take this into account, such as the estimator that solves

$$\begin{aligned} \min_{\theta} \ & \mathbb{E}_{\eta} \left[\sum_{i=1}^{N} \|y_i - Ox(t_i)\|_2^2 \right] \\ & \dot{x} = f(t, x, \theta, \eta) \\ & x(0) = \tilde{x}(\theta), \end{aligned} \tag{1.7}$$

where $\mathbb{E}_{\eta}$ represents the expectation with respect to the stochastic process model for the input, η.[4] Moreover, we can achieve more "robustness" by adding additional risk terms (e.g., variance, value-at-risk) to the objective function.

An important feature about (1.7) is that it is a special case of a more generic optimization

[4] In a sense, we can reconcile (1.6) and (1.7). Consider the extreme value distribution induced on (1.3) for the given stochastic model of η restricted to $\mathfrak{N}$. Now, we can compute the estimator that minimizes the marginalized version of (1.3) over this extreme value distribution. Then, we are again in the case of (1.7), but with a different distribution.

problem,

$$\min_{\theta} \ \mathbb{E}\left[f(\theta, \xi)\right], \tag{1.8}$$

where ξ is a random variable, which is found in any number of probabilistic and deterministic problems such as mixed effect model inference, stochastic optimal control, empirical risk minimization in machine learning, image reconstruction, and robust PDE-constrained optimization. Owing to its generality, the third part of this thesis is on methodologies for solving (1.8), and, as a special case, (1.7).

To summarize, the first part of this thesis is on the identifiability of dynamical systems. In this part, our main contribution is a careful study of necessary conditions for identifiability for the different types of inputs discussed above. The second part of this thesis is on incremental estimators. In this part, our main contributions are the local analysis of stochastic gradient descent and a statistical analysis of the Kalman Filter as an incremental estimator. The third part of this thesis is on solving (1.8). In this part, our main contribution is a novel framework for integrating statistical estimation and deterministic optimizers to solve (1.8). Overall, this thesis furthers the study of inference for a time-invariant linear differential equation model with uncontrollable or stochastic inputs, and, in doing so, develops mathematics and statistics that can be applied to a host of other disciplines.

Identifiability of Dynamical Systems

2 | System Identifiability and Inputs

System identifiability is the study of whether a given set of observations can uniquely determine a model from a family of dynamical system models. Owing to the multitude of ways of making observations and defining dynamical system models, system identifiability can often be subdivided across a number of dimensions such as the model formulation, the stability of the dynamics, the frequency of observations, the nonlinearity of the dynamics, and the nature of the inputs.

In this work, system identifiability will be divided across the latter dimension — the nature of the inputs. One classification of possible inputs is motivated by example in Section 2.1. This classification of the inputs is based on their physical properties in the context of the identification problem (Zadeh, 1956; Lory et al., 1959). However, we will revise this classification in Section 2.2, based on the mathematical properties of the inputs in the context of system identifiability theory. Guided by this mathematical characterization, in Section 2.3, we define different notions of system identifiability, review and refine the main results corresponding to these different notions of system identifiability, and construct some insightful counterexamples. Finally, in Section 2.4, we compare the results presented in the preceding section and contextualize our contributions, which are presented in subsequent chapters.

2.1 Rotor Dynamical System

A rotor dynamical system is a shaft being rotated at some angular frequency, as shown in Fig. 2.1. Rotor dynamical systems are found in a number of applications from Uranium-enriching centrifuges to household blenders and washing machines. Thus, the understanding

Disclaimer: Portions of text in this and the following chapter are pulled directly from a co-authored manuscript currently in preparation.

of rotor dynamics can be motivated from any number of perspectives.

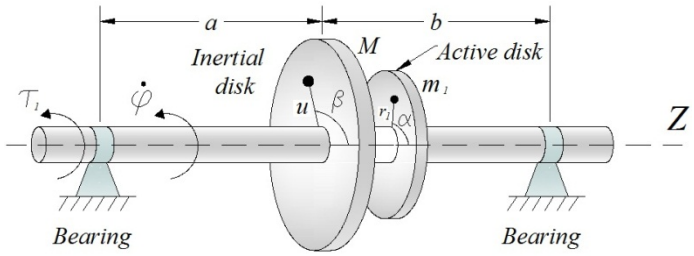


Figure 2.1: A diagram of the rotor dynamical system from Blanco-Ortega et al. (2012).

One aspect of rotor dynamics that is of particular interest is the system's vibrations in the plane perpendicular to the shaft, which are (under simplifications) described by the Föppl-Jeffcott model (see Yoon et al., 2012, Ch. 2),

$$\begin{aligned} \ddot{p}_1 + 2\zeta_1\omega_n\dot{p}_1 + \omega_n^2 p_1 &= u\omega(t)^2\cos(\omega(t)t) \\ \ddot{p}_2 + 2\zeta_2\omega_n\dot{p}_2 + \omega_n^2 p_2 &= u\omega(t)^2\cos(\omega(t)t), \end{aligned} \tag{2.1}$$

where p_1 and p_2 are the positions (along an orthonormal basis) of the shaft in the plane perpendicular to the system's rotation; ζ_1 and ζ_2 are the damping coefficients of the system; ω_n is the critical frequency of the shaft; u is the distance between the geometric center and the center of mass of the shaft; and $\omega(t)$ is the driving frequency (i.e., input) of the shaft (note, in Fig. 2.1, $\omega(t) = \dot{\varphi}(t)$.)

These vibrations are typically unwanted as they can cause damage to the rotor dynamical system or to the encasing system. Thus, understanding the vibrations is a critical step in mitigating the damage to, and increasing the efficiency of, the rotor dynamical system. In the simple Föppl-Jeffcott model, the vibrational behavior of a rotor dynamical system is characterized by the parameters $\zeta_1, \zeta_2, \omega_n$ and u, which motivates the need to identify these parameters. Importantly, the identifiability of these parameters is governed by the nature of the inputs.

Zero Input. For example, consider the situation in which the system is initially at rest and there is no input to the system (i.e., $\omega(t) = 0$). Then, in this case, it would be impossible to distinguish between distinct sets of parameters because we will only observe the system at rest. Now consider the situation in which the system is already operating at some angular frequency, and the driving frequency is turned off (i.e., $\omega(t) = 0$). Then, by observing the decay of the vibrations and by assuming idealized behavior, the rotor's damping coefficients, ζ_1, ζ_2, can be computed. For both contexts, the system identification is carried out for $\omega(t) = 0$, which is referred to as a system identification with *zero input*.[1]

Nonzero Input. Now consider a system for which $\omega(t)$ is not identically zero for all time points. The task of determining the parameters in this context is called system identification with *nonzero input*. Moreover, we can further divide the case of a nonzero input into the controllable input or the uncontrollable input cases.

Controllable Input. The critical frequency, ω_n, can be determined by varying the driving frequency of the shaft and observing at which angular frequency the amplitude of the vibrations is maximized. Then, from this experimentation and assumptions of idealized behavior, the critical frequency parameter can be computed. Because the parameter is being determined by manipulating the driving frequency, this is an example of a system identification with a *controllable input*.

Uncontrollable Input. Now consider the task of determining the parameters in an operational setting; that is, the driving frequency is determined by the needs of the application and not by the needs of the identification problem. Here, the parameters (e.g., critical frequency, damping coefficients) can only be determined from observations of the system without any intervention on the driving frequency from the observer. Because the parameters are being determined from observations of the system without intervention from the observer, this is

[1] A dynamical system with zero inputs is sometimes referred to as an uncontrollable system. Note, an uncontrollable system and system with uncontrollable inputs are not the same.

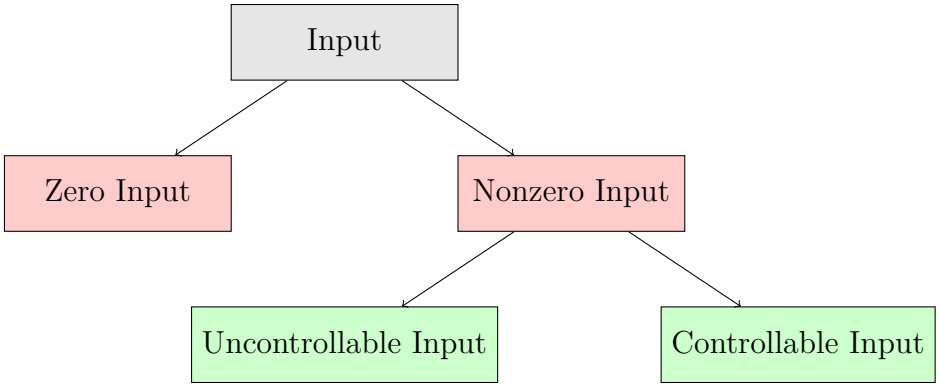


Figure 2.2: A classification of inputs based on their physical properties.

an example of a system identification with an *uncontrollable input*. There are several nuances regarding uncontrollable inputs that must be pointed out. First, it is logically possible for uncontrollable inputs to be zero. However, the case in which the input is identically zero, even if it is uncontrollable, will be considered a zero input system. Therefore, we are enforcing a strict distinction between the zero input and nonzero input cases. Second, whether the system identifiability problem has controllable or uncontrollable inputs depends entirely on the context. For example, there may be a system operator who has control over the inputs to the system, yet the observer performing the identification cannot influence the system operator, and, consequently, the parameter determination is subject to an uncontrollable input. Indeed, this is exactly the case for power systems: there is a system operator who is required by law to maintain the systems operations within very tight margins, and there are a whole host of other agents who are interested in inferring some information about the system, but who have no control over the system inputs.

The preceding discussion of the different types of inputs leads to the classification shown in Fig. 2.2. Indeed, such a classification is rather natural based on the physical interpretation of the inputs. However, this classification is misleading when it comes to system identifiability theory, as we discuss next.

2.2 System Identifiability and the Input Dimension

Following from Fig. 2.2, the system identifiability literature has primarily focused on the cases of zero inputs (Kalman, 1959; Reid, 1977; Denis-Vidal et al., 2001; Evans et al., 2002; Denis-Vidal and Joly-Blanchard, 2004; Stanhope et al., 2014) and controllable inputs (Kalman, 1959; Bellman and Åström, 1970; Hájek, 1972; Grewal et al., 1974; Thowsen, 1978; Ljung, 1999). For the case of uncontrollable inputs, the system identifiability literature has been rather scant. Glover and Willems (1974) provide conditions for system identifiability with white noise inputs; however, their results were based on informal and somewhat contradictory conditions such as the system being in steady state while also being driven by white noise. Moreover, in Chapter 8, Ljung (1999) studies very particular uncontrollable inputs that are perturbed by white-noise processes, but does not provide a complete characterization.

The system identifiability literature's bias towards zero and controllable inputs raises the question: is there a need to consider systems with uncontrollable inputs? Following from the power system and seismic retrofitting examples in Chapter 1, and the rotor dynamical system example of Section 2.1, there is a clear need to consider systems with uncontrollable inputs. Then, why has the system identifiability literature not explored the topic of uncontrollable inputs more completely? At least partially, the lack of system identifiability for the uncontrollable input case can be explained by the fact that, for a one-dimensional input, the system identifiability theory for a controllable input or an uncontrollable input coincide. Therefore, believing that the one-dimensional case extends trivially to multi-dimensional input case (e.g., Bellman and Åström, 1970; Reid, 1977; Gargash and Mital, 1980), the issue of multi-dimensional inputs was not further investigated. However, as we now explain, whereas a one-dimensional input does not, a multi-dimensional input requires considering the controllable and uncontrollable input cases separately.

One-dimensional Input. To see how the theory coincides for the one-dimensional input case, consider a dynamical system with inputs $\eta : [0, T] \to \mathbb{R}$ and observations $y : [0, T] \to \mathbb{R}^n$, and

let $H(s)$ and $Y(s)$ denote the Laplace transformations of η and y, respectively. Now, consider a family of transfer function models, $\{G(s,\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, for the dynamical system that relate the inputs and outputs by $Y(s) = G(s,\theta)H(s)$. According to Section 4.6 of Ljung (1999), a particular model $\theta \in \Theta$ is identifiable from all other models, $\psi \in \Theta \setminus \{\theta\}$, whenever $G(s,\theta) \neq G(s,\psi)$ (i.e., two different models generate different outputs $Y(s)$). Equivalently, since $H(s)$ is one-dimensional, we can reformulate the requirement that $G(s,\theta) \neq G(s,\psi)$ to $G(s,\theta)H(s) \neq G(s,\psi)H(s)$, regardless of whether or not $\eta(t)$ is controllable or uncontrollable (so long as it is nonzero). Thus, in the one-dimensional case, the controllability of $H(s)$ does not play a role in the definition of identifiability. Hence, the system identifiability theory for the one-dimensional controllable and uncontrollable input cases coincide.

Multi-dimensional Input. Now, to see how the system identifiability theory for the controllable and uncontrollable inputs cases diverge in multiple dimensions, let the input, $\eta : [0,T] \to \mathbb{R}^q$, be multi-dimensional (i.e., $q > 1$), and denote its Laplace transform by $H(s)$. Moreover, let y and Y be the same as above. Now, consider a family of transfer function models, $\{G(s,\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, for the dynamical system that relate the inputs and outputs by $Y(s) = G(s,\theta)H(s)$. Again, according to Section 4.6 of Ljung (1999), a particular model $\theta \in \Theta$ is identifiable from all other models, $\psi \in \Theta \setminus \{\theta\}$, whenever $G(s,\theta) \neq G(s,\psi)$. That is, θ is identifiable whenever there exists an $H(s)$ such that $G(s,\theta)H(s) \neq G(s,\psi)H(s)$. In the case where η is controllable, we can try all possible variations of $H(s)$ to check if $G(s,\theta)H(s) \neq G(s,\psi)H(s)$ holds. However, in the case where η is uncontrollable and multi-dimensional, for any given input, $H(s)$, we cannot guarantee that $G(s,\theta) \neq G(s,\psi)$ implies $G(s,\theta)H(s) \neq G(s,\psi)H(s)$. Thus, we see that the definition of identifiability in Ljung (1999) only applies to the identifiability for the controllable input case.

As the preceding examples and discussion demonstrate, the mathematical interaction between inputs and dynamical systems warrant an alternative classification for inputs. In particular, the case of a nonzero input must be subdivided into a one-dimensional input and

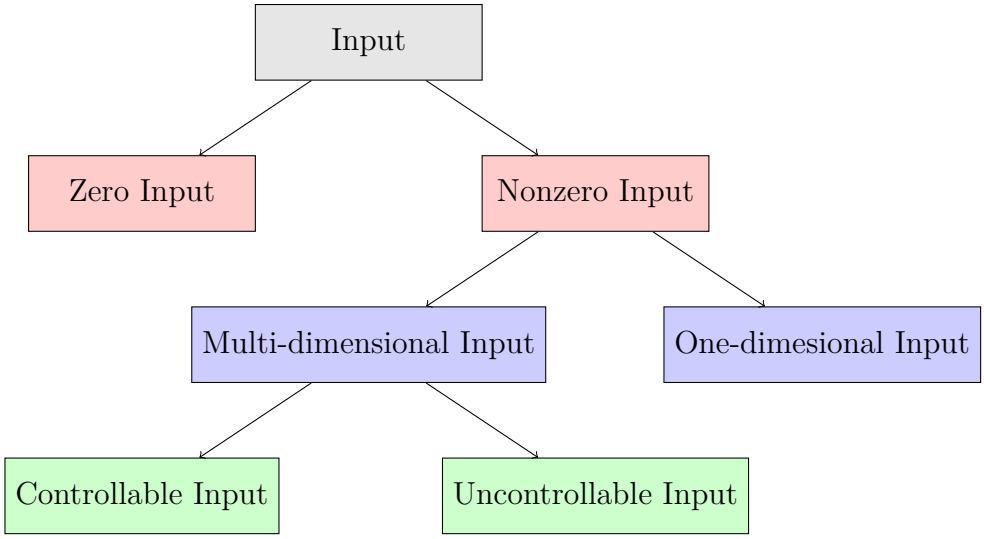


Figure 2.3: A classification of inputs based on their mathematical properties.

a multi-dimensional input. Moreover, the one-dimensional input does not need any further subdivision, since, as we have shown, the definitions for the one-dimensional, controllable input and the one-dimensional, uncontrollable input coincide. On the other hand, the multi-dimensional input does need to be subdivided into controllable and uncontrollable input cases. This mathematics-based classification is summarized in Fig. 2.3. Based on the interpretation in Fig. 2.3, in the next section, we will present, refine and reinterpret key results from system identifiability theory.

2.3 One-dimensional Identifiability Theory

The theory discussed here will focus on state-space models such as (1.1) and (2.1), rather than the transfer function models championed by Ljung (1999). The primary reason for this is that state-space models are a more common formulation for real phenomena such as general weather circulation (Roeckner et al., 2003) and rotors (Yoon et al., 2012).

In particular, we will consider the class of linear state space models over $t \in [0, T]$, defined as

$$\dot{x} = A(\theta)x(t) + B(\theta)\eta(t), \quad x(0) = \tilde{x}(\theta), \tag{2.2}$$

Table 2.1: Notation for Identifiability.

$[0,T]$	Time interval of interest	n	Dimension of observation
$x(t)$	State on $[0,T]$	$\eta(t)$	Input on $[0,T]$
$X(s)$	Laplace transformation of state	$H(s)$	Laplace transformation of input
d	Dimension of state	q	Dimension of input
$y(t)$	Observation on $[0,T]$	θ	Parameter
$Y(s)$	Laplace transformation of observation	p	Dimension of parameter

where $x(t) \in \mathbb{R}^d$ is the state variable; $\eta(t) \in \mathbb{R}^q$ is the input with $q \geq 1$; $\theta \in \mathbb{R}^p$ is the unknown parameter; and $A : \mathbb{R}^p \to \mathbb{R}^{d\times d}$, $B : \mathbb{R}^p \to \mathbb{R}^{d\times q}$ and $\tilde{x} : \mathbb{R}^p \to \mathbb{R}^d$ are smooth functions of the parameter θ (at least in some open set of interest). Second, we consider observations over $t \in [0,T]$, defined as

$$y(t) = Ox(t), \tag{2.3}$$

where $O \in \mathbb{R}^{n\times d}$ is called the observation operator. A summary of the notation can be found in Table 2.1.

For models of the type (2.2), system identifiability can be traced to the study of a simpler family of models given by

$$\dot{x} = Ax(t), \quad x(0) = \theta, \tag{2.4}$$

where $\theta \in \mathbb{R}^d$ (i.e., $p = d$). In particular, system identifiability for the family of models given by (2.4) is called observability, and is defined presently.

Definition 2.1 (Observability)**.** *A model, θ, in the class (2.4) with observations (2.3) is called observable if for any other model, $\psi \neq \theta$, the sets of observations generated by θ and ψ are distinct.*

The main observability (i.e., identifiability) result for a family given by (2.4) with observations given by (2.3) was formulated by Kalman (1959), and is reproduced below.

Theorem 2.1 (Kalman (1959)). *A model in the class specified by (2.4) is observable (i.e., identifiable) from the observations defined by (2.3) if and only if*

$$\texttt{null}\left(\mathcal{M}(O,A)\right) = \{0\}, \tag{2.5}$$

where $\texttt{null}(\cdot)$ *denotes the null space of a matrix; and* $\mathcal{M}(\cdot,\cdot)$ *is defined below in (2.11).*

Proof. Solving (2.4) and using (2.3), we have that $y(t;\theta) = O\exp(At)\theta$ and $y(t;\psi) = O\exp(At)\psi$ for two parameters θ and ψ. Then, from our definition, the identifiability of θ is equivalent the statement that if $y(t;\theta) = y(t;\psi)$ for all $t \in [0,T]$ then $\theta = \psi$. Now,

$$\begin{aligned}
0 = y(t;\theta) - y(t;\psi) &\Leftrightarrow 0 = O\exp(At)(\theta-\psi) && (2.6)\\
&\Leftrightarrow 0 = \sum_{j=0}^{\infty} \frac{t^j}{j!} OA^j(\theta-\psi) && (2.7)\\
&\Leftrightarrow 0 = OA^j(\theta-\psi),\ \forall j = 0,1,\ldots && (2.8)\\
&\Leftrightarrow 0 = OA^j(\theta-\psi),\ \forall j = 0,1,\ldots,d-1 && (2.9)\\
&\Leftrightarrow 0 = \mathcal{M}(O,A)(\theta-\psi), && (2.10)
\end{aligned}$$

where (2.7) follows from a Taylor expansion of $\exp(At)$; (2.8) follows from the linear independence of the monomials $\{t^j : j = 0,1,\ldots\}$ on $[0,T]$; (2.9) follows from an application of the Cayley-Hamilton theorem; and (2.10) follows from using the definition in (2.11). Hence, the identifiability of θ is equivalent to the statement that $0 = \mathcal{M}(O,A)(\theta-\psi)$ implies $\theta = \psi$. The result follows. ■

In the system identifiability literature, the function $\mathcal{M}(\cdot,\cdot)$ plays a special role. The

function is defined by

$$\mathcal{M}(M_1, M_2) = \begin{bmatrix} M_1 \\ M_1 M_2 \\ \vdots \\ M_1 M_2^{d-1} \end{bmatrix}, \tag{2.11}$$

where M_1 and M_2 are appropriately dimensioned and d is the dimension of the state variable. As seen in the proof of Theorem 2.1, the function comes from the first d terms in the Taylor expansion of the function $M_1 \exp(M_2 t)$, which, in turn, specifies the relationship between the observations and the parameter for the model class (2.4); that is,

$$y(t) = O \exp(At)\theta. \tag{2.12}$$

Therefore, from the definition of $\mathcal{M}$, Theorem 2.1 connects identifiability of the model to whether or not the mapping induced by $M_1 \exp(M_2 t)$ is an injection. Moreover, Theorem 2.1 has two other noteworthy features. First, Theorem 2.1 presents a computable condition for determining if $M_1 \exp(M_2 t)$ is, in fact, an injection. Second, with a small extension, Theorem 2.1 can be shown to not depend on inputs, and, consequently, it is the foundational result for both system identifiability for zero and all nonzero inputs.

The system identifiability result of Kalman was later extended, in a sense, by the seminal work of Bellman and Åström (1970), who studied the model class (2.2) with several simplifications: O is the identity matrix; $\tilde{x}(\theta) = 0$; $B(\theta) \in \mathbb{R}^d$; and

$$A(\theta) = e_{11}\theta_1 + e_{12}\theta_2 + \cdots + e_{dd}\theta_{d^2} \quad \text{and} \quad B(\theta) = e_1\theta_{d^2+1} + \cdots e_d\theta_{d^2+d}, \tag{2.13}$$

where e_{ij} is the standard basis for $\mathbb{R}^{d\times d}$ and e_i is the standard basis for $\mathbb{R}^d$. That is, the system identifiability results of Bellman and Åström (1970) apply to the model family and

observations

$$\begin{aligned} \dot{x} &= A(\theta)x(t) + B(\theta)\eta(t), \quad x(0) = 0, \\ y(t) &= x(t), \end{aligned} \tag{2.14}$$

where $\eta : [0, T] \to \mathbb{R}$. For this class of models (and for (2.2) with a one-dimensional input), system identifiability is now defined.

Definition 2.2 (One-dimensional Identifiability)**.** *A model, θ, in class (2.2) with a one-dimensional input (i.e., $q = 1$) and with observations (2.3) is called one-dimensional identifiable if for any other distinct model, $\psi \neq \theta$, the sets of observations generated by θ and ψ are distinct.*

Remark 2.1. *If we restrict ψ to being within some neighborhood of θ, then we will refer to this case as local, one-dimensional identifiability.*

The system identifiability result of Bellman and Åström (1970) is stated below and proved with additional care for the input.

Theorem 2.2 (Bellman and Åström (1970))**.** *Suppose any input to the system, $\eta : [0, T] \to \mathbb{R}$, is piecewise continuous and not identically zero. A model in the class specified by (2.14) is one-dimensional identifiable if*

$$\mathtt{null}\left(\begin{bmatrix} B(\theta) & A(\theta)B(\theta) & \cdots & A(\theta)^{d-1}B(\theta) \end{bmatrix}\right) = \{0\}. \tag{2.15}$$

Remark 2.2. *Whenever we consider a nonzero, piecewise continuous input, we will always assume that it is nontrivial. That is, there is an open interval on which the function is nonzero.*

Proof. By solving (2.14), we have that for any $t \in [0, T]$,

$$y(t; \theta) = \int_0^t \exp(A(\theta)(t - \tau))B(\theta)\eta(\tau)d\tau = (h(\cdot; \theta) * \eta)(t), \tag{2.16}$$

where $h(t;\theta) = \exp(A(\theta)t)B(\theta)$ is the impulse response function. First, we will show that, under the conditions on η, distinct impulse response functions generate distinct observations. Then, we will show that condition (2.15) implies that the impulse response function is uniquely determined by a parameter under the parameterization in (2.13).

Now, let $t_1 = \inf\{t \in (0,T) : \lim_{s\to t} \eta(s) \neq 0\}$ (see Remark 2.2). Then, by, for example, Theorem 7 of Titchmarsh (1926), $h(t;\theta) = h(t;\psi)$ for all $t \in [0, T-t_1)$. Therefore, if $h(t;\theta) \neq h(t;\psi)$ for some $t \in [0, T-t_1)$ then there must be a $t \in [0,T]$ such that $y(t;\theta) \neq y(t;\psi)$.

Now, we must show that (2.15) implies that $h(t;\theta)$ is uniquely determined by θ. That is, if (2.15) holds then $h(t;\theta) = h(t;\psi)$ for $t \in [0,T]$ implies $\theta = \psi$. Note, $B(\theta) = h(0;\theta) = h(0;\psi) = B(\psi)$. Therefore, by the choice of parametrization, $h(0;\theta) = h(0;\psi)$ implies $B(\theta) = B(\psi)$. Now, since $h(t;\theta)$ is analytic, then its derivatives about 0 uniquely determine $h(t;\theta)$ for $[0,T]$. Moreover, for any $j \in \mathbb{N}$,

$$\frac{\partial^j}{\partial t^j} h(t;\theta) = A(\theta) \frac{\partial^{j-1}}{\partial t^{j-1}} h(t;\theta). \tag{2.17}$$

Therefore,

$$\left[\frac{\partial}{\partial t} h(t;\theta)\Big|_{t=0} \quad \cdots \quad \frac{\partial^d}{\partial t^d} h(t;\theta)\Big|_{t=0} \right] = A(\theta) \left[h(0;\theta) \quad \cdots \quad \frac{\partial^{d-1}}{\partial t^{d-1}} h(t;\theta)\Big|_{t=0} \right], \tag{2.18}$$

where the matrix on the right hand side is invertible by (2.15). Therefore, $h(t;\theta) = h(t;\psi)$ on $t \in [0,t]$ implies that $B(\theta) = B(\psi)$ and $A(\theta) = A(\psi)$. Hence, by the choice of parametrization, $\theta = \psi$. ■

Just as for Theorem 2.1, Theorem 2.2 is a computable condition for determining whether or not the mapping from the parameter space to the space of impulse responses, $\theta \mapsto \{\exp(A(\theta)t)B(\theta) : t \in [0,T]\}$, is an injection. However, unlike Theorem 2.1, Theorem 2.2 only applies for a one-dimensional, nonzero input. Indeed, when the input, η, is identically zero, then $y(t;\theta) = 0$ for any θ by (2.16), and, consequently, there is no information available

to discern between two distinct parameters.

To complete the one-dimensional input case, we develop the necessary and sufficient conditions for local (see Remark 2.1), one-dimensional identifiability presented by Thowsen (1978). As we have done for Theorems 2.1 and 2.2, we prove the results of Thowsen (1978) using the theory of simultaneous equations, but we will need slightly more analysis, which we develop presently.

Lemma 2.1. *Let $z : \mathbb{R}^p \to \mathbb{R}^N$ have a continuous Jacobian, $J : \mathbb{R}^p \to \mathbb{R}^{N \times p}$. If for some $\theta \in \mathbb{R}^p$, $\texttt{rank}(J(\theta)) = p$ then there exists a neighborhood of θ, $\mathcal{N}$, such that z restricted to $\mathcal{N}$ is an injection.*

Proof. Since $\texttt{rank}(J(\theta)) = p$, without loss of generality, the first p rows of $J(\theta)$ form a $p \times p$ submatrix that is invertible. Let $J_j : \mathbb{R}^p \to \mathbb{R}^p$ be the vector equal to the j^{th} row of J, and let $z_j : \mathbb{R}^p \to \mathbb{R}$ denote the j^{th} row of z. Now, by the mean value theorem, for any $\psi \in \mathcal{N}$ and for any z_j, there is a ψ_j (on the segment connecting θ and ψ) such that $z_j(\psi) - z_j(\theta) = J_j(\psi_j)'(\psi - \theta)$. Let $\bar{J}(\psi, \theta)$ be the matrix whose rows are $J_1(\psi_1), \ldots, J_p(\psi_p)$. Since $\texttt{det}(\cdot)$ is a continuous function and the first p rows of $J(\theta)$ have a nonzero determinant, there is a neighborhood of θ, $\mathcal{N}$, such that for any $\psi \in \mathcal{N}$, $\texttt{det}(\bar{J}(\psi, \theta)) \neq 0$. Therefore, if the first p components of $z(\psi)$ and $z(\theta)$ are identical, then $\theta = \psi$. ■

Lemma 2.2. *Let $z : \mathbb{R}^p \to \mathbb{R}^N$ have a continuous Jacobian, $J : \mathbb{R}^p \to \mathbb{R}^{N \times p}$. If for some $\theta \in \mathbb{R}^p$, there is a neighborhood of θ, $\mathcal{N}$, such that $\texttt{rank}(J(\psi)) = \texttt{rank}(J(\theta))$ for all $\psi \in \mathcal{N}$ and z restricted to $\mathcal{N}$ is an injection, then $\texttt{rank}(J(\theta)) = p$.*

Proof. See Fisher (1966) Theorem 5.A.1. ■

Clearly, Lemmas 2.1 and 2.2 are nearly converses, but Lemma 2.2 is restricted to points for which there exists a neighborhood in which the Jacobian of z has constant rank. While the restriction to such points, called rank regular points, may seem superfluous in light of

the continuity of the Jacobian of z, the following counterexample shows that the restriction to rank regular points is necessary for the conclusion to hold.

Consider $z : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $z(\theta) = (\theta_1^3, \theta_2^3, \theta_3^3)$. Then, z is clearly an injection since $x \mapsto x^3$ is an injection. However, the Jacobian, which is not only continuous but also differentiable, has zero rank at $\theta = 0$. Thus, in order to conclude that the Jacobian has full rank p, we need to be at a regular point.[2] Fortunately, the restriction to rank regular points is not overly burdensome. Indeed, the set of rank regular points is open and dense if the Jacobian of z is continuous (Lewis, 2009). Moreover, the complement of the set of rank regular points is not only measure zero, but also the intersection of the roots of finitely many analytic functions whenever $z : \mathbb{R}^p \to \mathbb{R}^N$ is an analytic function of θ (Lewis, 2009). Hence, the set of points *that are not* rank regular points is, in a sense, negligible. With these two lemmas, we are now ready to state and prove the result of Thowsen (1978), but with additional care for the one-dimensional input and some minor refinements.

Theorem 2.3 (Thowsen (1978))**.** *Consider the family of models defined by (2.2) with $\tilde{x}(\theta) = 0$ and with observations (2.3). Suppose any input into the system, $\eta : [0, T] \to \mathbb{R}$, is piecewise continuous and not identically zero. Let $z_l(\psi) = OA(\psi)^l B(\psi)$ for $l = 0, 1, \ldots$ and let*

$$z(\psi) = \begin{bmatrix} z_0(\psi) \\ z_1(\psi) \\ \vdots \\ z_{2d-1}(\psi) \end{bmatrix}. \tag{2.19}$$

Finally, suppose that $A : \mathbb{R}^p \to \mathbb{R}^{d \times d}$ are $B : \mathbb{R}^p \to \mathbb{R}^d$ are continuously differentiable in a neighborhood of θ.

1. *The model θ is locally, one-dimensional identifiable if the nullity of the Jacobian of z*

[2] From our counterexample, it may be possible that z is a local injection about a point θ if and only if higher-order derivatives of z evaluated at θ have "full rank." The sufficient part of the preceding statement would supply an interesting extension to the implicit function theorem.

evaluated at θ is zero.

2. If θ is a rank regular point of z and the model θ is locally, one-dimensional identifiable then the nullity of the Jacobian of z evaluated at θ is zero.

Proof. Let $h(t;\theta) = O\exp(A(\theta)t)B(\theta)$. Then, just as in the proof of Theorem 2.2, θ is locally identifiable if and only if there is a neighborhood of θ, $\mathcal{N}$, such that for any $\psi \in \mathcal{N} \setminus \{\theta\}$, $\exists t \in [0,T]$ such that $h(t;\theta) \neq h(t;\psi)$. Now, (1) by the linear independence of the monomials and (2) the Taylor expansion of $h(t;\cdot)$, $h(t;\theta) \neq h(t;\psi)$ if and only if $z_l(\theta) \neq z_l(\psi)$ for $l = 0, 1, \ldots$. By Theorem 2 of Thowsen (1978), $z_l(\theta) \neq z_l(\psi)$ for $l = 0, 1, \ldots$ if and only if $z(\theta) \neq z(\psi)$ (where z is defined in (2.19)). To summarize, θ is locally identifiable if and only if there is a neighborhood of θ, $\mathcal{N}$, such that z restricted to $\mathcal{N}$ is an injection.

By Lemma 2.1, if the rank of the Jacobian of z evaluated at θ has rank p, then there is a neighborhood of θ, $\mathcal{N}$, such that z restricted to $\mathcal{N}$ is an injection, and, in turn, θ is locally identifiable.

Now, if θ is locally identifiable then there is a neighborhood of θ, $\mathcal{N}$, such that z restricted to $\mathcal{N}$ is an injection. Moreover, if θ is a rank regular point, then the Jacobian of z evaluated at θ has rank p by Lemma 2.2. ■

Remark 2.3. *Note, Thowsen (1978) considers an observation model in which the observation operator, O, also depends on the parameter θ. While this case is equally challenging to what we have considered so far, we will not add this complexity as it is uncommon in the physical and engineering systems that motivate this work.*

We finish this section by constructing an example to demonstrate the importance of the rank regular point requirement for concluding that the Jacobian of z at a model of interest has nullity zero if the model is one-dimensional identifiable. In order to construct this example, we will first need the following result.

Lemma 2.3. *Let $N \in \mathbb{R}^{d\times d}$ be a nilpotent matrix with an eigenvector $v \in \mathbb{R}^d$, and let $h(t;N,v) = \exp(Nt)v$. For any $E \in \mathbb{R}^{d\times d}$ that commutes with N and $w \in \mathbb{R}^d$, $h(t;N+$*

$E, v+w) = h(t; N, v)$ *for all* $t \in [0, T]$ *if and only if* $w = 0$ *and* E *has a zero eigenvalue with corresponding eigenvector* v.

Proof. Note, since N is nilpotent, $\exp(Nt) = \sum_{j=0}^{d} \frac{t^j}{j!} N^j$. Moreover, since v is an eigenvector of N, then $Nv = 0$. Therefore, $h(t; N, v) = \sum_{j=0}^{d} \frac{t^j}{j!} N^j v = v$. Now, suppose $w = 0$ and E has a zero eigenvalue with corresponding eigenvector v, then

$$\begin{aligned} h(t; N+E, v+w) = h(t; N+E, v) &= \exp(Et)\exp(Nt)v \\ &= \exp(Et)v = v = h(t; N, v). \end{aligned} \tag{2.20}$$

Now, suppose that there is an E that commutes with N and w such that $h(t; N+E, v+w) = h(t; N, v)$ for all $t \in [0, T]$. Then, $h(0; N+E, v+w) = h(0; N, v)$ implies that $w = 0$. So, we must show that $h(t; N+E, v) = h(t; N, v)$ implies that E has a zero eigenvalue with a corresponding eigenvector v. Note, since $h(t; N, v) = v$, we have

$$v = h(t; N+E, v) = \exp(Et)\exp(Nt)v = \exp(Et)v. \tag{2.21}$$

If $Ev \neq 0$, then we immediately have a contradiction. Therefore, E must have a zero eigenvalue with a corresponding eigenvector v. ■

Theorem 2.4. *Consider the family of models defined by (2.14) with* $d = 2$. *Suppose any input into the system,* $\eta : [0, T] \to \mathbb{R}$, *is piecewise continuous and not identically zero. Let*

$$A(\theta) = N + \theta^3 I, \text{ where } N = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \text{ and } I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{2.22}$$

and let $B(\theta) = v := (1, 2)$ *(a vector of dimension two). Then, (1) the model* $\theta = 0$ *is locally, one-dimensional identifiable, (2)* $\theta = 0$ *is not a rank regular point of* z *(defined in (2.19)), and (3) the Jacobian of* z *evaluated at* $\theta = 0$ *has rank* 0.

Proof. From the proof of Theorem 2.2, we have shown that θ is one-dimensional identifiable

if and only if $h(t;\theta) = h(t;\psi)$ implies $\theta = \psi$. By Lemma 2.3, since N is nilpotent with eigenvector v, $h(t; N + \theta^3 I, v) = h(t; N, v)$ if and only if I has a zero eigenvalue and a corresponding eigenvector v. However, the identity, I, does not have a zero eigenvalue. Therefore, for any $\theta \neq 0$, $h(t; N + \theta^3 I, v) \neq h(t; N, v)$. Thus, we conclude that (1) the model $\theta = 0$ is locally, one-dimensional identifiable. Now, the Jacobian of z is

$$J(\psi) = \begin{bmatrix} 0 \\ 3\psi^2 v \\ 6\psi^5 v \\ 9\psi^8 v \end{bmatrix}. \tag{2.23}$$

Note, $J(\psi)$ is just a vector. Thus, for any $\psi \neq 0$, $J(\psi)$ has a nonzero rank, but at $\psi = 0$, $J(0) = 0$ and has rank zero. Therefore, we have shown (2) and (3). ■

2.4 Comparing One-Dimensional Identifiability Results

With Theorems 2.1 to 2.3 in hand, it is particularly insightful to compare these three results, and use these comparisons to preview and contextualize our main contributions below. First, all three of these results provide sufficient conditions for the identifiability of a model from a family of models. However, the family of models is distinct for Theorem 2.1 and the latter two results. In particular, Theorems 2.2 and 2.3 ignore the possibility of a nonzero initial state. Fortunately, as Reid (1977) points out, the identifiability of a model in the general model class, (2.2), with observations, (2.3), can be viewed as the the "sum" of the identifiability results for two separate model classes: one without inputs and one with a zero initial condition. While this is true, requiring each model class to be separately identifiable is overly burdensome, especially when the input is multi-dimensional and uncontrollable. Thus, it is better to work directly with (2.2). However, doing so is nontrivial, and even recent results (e.g. Stanhope et al., 2014) have only considered the zero input case. Indeed,

even here, we will only be partially successful.

Second, Theorems 2.1 and 2.3 provide necessary and sufficient, *computable* conditions for the weak identifiability of a model from a family of models. This is particularly extraordinary since there is no precedence in the statistical literature for reducing the infinite dimensional question of identifiability (i.e., showing $y(t;\theta) = y(t;\psi)$ for all $t \in [0,T]$) to an equivalent, verifiable finite dimensional condition. In our more general results, we will not preserve this computability unless there is some additional structure for the input such as the input is controllable or is a nondegenerate stochastic process.

Now, comparing Theorems 2.2 and 2.3, we see that Theorem 2.2 requires significantly fewer terms in the expansion than Theorem 2.3. The reason for this is straightforward: Theorem 2.2 is predicated on observing the entire state $x(t)$, whereas Theorem 2.3 allows for very limited information about $x(t)$ to be observed. Thus, while it is still an incredible achievement, the need for so few terms to determine the identifiability in the case of full observability (i.e., O has full column rank) is not surprising. On the other hand, the fact that there is an upper bound on the number of terms needed to determine identifiability regardless of the observability operator, O, is singularly incredible. Again, in our results below, we will see the same upper limit on the number of terms needed in the expansion, regardless of the observability operator.

Continuing with Theorems 2.2 and 2.3, we might ask if Theorem 2.3 simplifies to Theorem 2.2 for the parametrization given in (2.13) and when O has full column rank. The following result shows that the conditions of Theorems 2.2 and 2.3 are algebraically equivalent.

Lemma 2.4. *Under parametrization (2.13) with model class and observations defined by (2.14), the conditions that the Jacobian of z, defined in (2.19), has rank $p = d^2 + d$ and (2.15) are equivalent.*

Proof. We start by computing the Jacobian of z. Let $\mathbb{D}_j(Q) = \sum_{k=0}^{j} A^{j-k}(\theta) Q A^k(\theta)$. Then,

the derivative of $A(\theta)^{j+1}B(\theta)$ with respect to θ_1 is $\mathbb{D}_j(e_{11})B(\theta)$ (recall from the parametrization, (2.13), $B(\theta)$ does not depend on $\theta_1, \ldots, \theta_{d^2}$). Therefore,

$$\nabla z(\psi)|_{\psi=\theta} = \begin{bmatrix} 0 & \cdots & 0 & I \\ \mathbb{D}_0(e_{11})B(\theta) & \cdots & \mathbb{D}_0(e_{dd})B(\theta) & A(\theta) \\ \mathbb{D}_1(e_{11})B(\theta) & \cdots & \mathbb{D}_1(e_{dd})B(\theta) & A(\theta)^2 \\ \vdots & & \vdots & \vdots \\ \mathbb{D}_{2d-2}(e_{11})B(\theta) & \cdots & \mathbb{D}_{2d-2}(e_{dd})B(\theta) & A(\theta)^{2d-2} \end{bmatrix}, \tag{2.24}$$

where 0 is the zero element in $\mathbb{R}^d$. Moreover, using the Cayley-Hamilton theorem and the fact that $\mathbb{D}_j(Q) = A\mathbb{D}_{j-1}(Q) + QA^j(\theta)$, a block Gaussian elimination (i.e., multiplying the block row above by A and subtracting it from the one below) will reduce the Jacobian of z evaluated at θ to

$$\nabla z(\psi)|_{\psi=\theta} = \begin{bmatrix} 0 & \cdots & 0 & I \\ e_{11}B(\theta) & \cdots & e_{dd}B(\theta) & \mathbf{0} \\ e_{11}A(\theta)B(\theta) & \cdots & e_{dd}B(\theta) & \mathbf{0} \\ \vdots & & \vdots & \vdots \\ e_{11}A(\theta)^{d-1}B(\theta) & \cdots & e_{dd}A(\theta)^{d-1}B(\theta) & \mathbf{0} \\ 0 & \cdots & 0 & \mathbf{0} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \mathbf{0} \end{bmatrix}, \tag{2.25}$$

where $\mathbf{0}$ is the zero element in $\mathbb{R}^{d\times d}$. From (2.15), let

$$C(\theta)' = \begin{bmatrix} B(\theta) & A(\theta)B(\theta) & \cdots & A(\theta)^{d-1}B(\theta) \end{bmatrix}, \tag{2.26}$$

where the prime indicates a transpose. Then, up to a permutation, the nonzero part of the

Jacobian of z evaluated at θ is

$$\begin{bmatrix} I & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & C(\theta) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & C(\theta) & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & C(\theta) \end{bmatrix}, \tag{2.27}$$

which has rank $d^2 + d$ if and only if $\texttt{rank}(C(\theta)) = d$. ■

Therefore, using Lemma 2.4, we can conclude that (2.15) is not only sufficient for identifiability, but it is also a necessary condition at any rank regular point. In fact, using the strategy of Lemma 2.4, we can actually make specific statements for arbitrary affine parametrization choices for A and B. While such statements will coincide with existing results for particular affine parametrization choices when the observability operator has full column rank (see Gargash and Mital, 1980), the value in Lemma 2.4 is that it provides a direct algebraic link to such results. However, this direct algebraic link is nontrivial when the observability operator does not have full column rank.

Finally, although it is often claimed that these one-dimensional identifiability results trivially extend to the multi-dimensional input case (Bellman and Åström, 1970; Glover and Willems, 1974; Reid, 1977; Gargash and Mital, 1980), we have already shown that such a claim is suspect since the multi-dimensional case must be considered separately under the controllable input and uncontrollable input cases. In fact, even extending these results to the multi-dimensional controllable input case is nontrivial, as we now demonstrate.

Consider the same model class studied by Bellman and Åström (1970), but now let $B(\theta) \in \mathbb{R}^{d \times q}$ for $q > 1$ (i.e., a multi-dimensional input) with the analogous parametrization

to (2.13), defined by

$$A(\theta) = e_{11}\theta_1 + e_{12}\theta_2 + \cdots + e_{dd}\theta_{d^2} \quad \text{and} \quad B(\theta) = e_{11}\theta_{d^2+1} + \cdots e_{dq}\theta_{d^2+dq}, \tag{2.28}$$

where, with an abuse of notation, e_{ij} is either understood to be the standard basis of $\mathbb{R}^{d\times d}$ or $\mathbb{R}^{d\times q}$ (note, it will always be clear within the context). Moreover, let $B_1(\theta), \ldots, B_q(\theta)$ denote the columns of $B(\theta)$. Now suppose that the multi-dimensional input is controllable, that is, $\eta : [0,T] \to \mathbb{R}^q$ can be arbitrarily selected. Then, of course, we can choose $\eta(t)$ to have only one nonzero component, in which case, we have simplified from the multi-dimensional input case to the one-dimensional input case. Thus, we can apply Theorem 2.2 with the appropriate column of $B(\theta)$ to find a sufficient condition for system identifiability with a controllable, multi-dimensional input. However, owing to the multitude of ways of controlling the input, such a condition could hardly be necessary.

Moreover, to reiterate the point that the one-dimensional case does not extend trivially to the multi-dimensional case, suppose now that condition (2.15) holds for $\{B_j(\theta) : j = 1, \ldots, q-1\}$, but does not hold for $j = q$. Then, for any piecewise continuous input that is nonzero in just one component indexed by $j = 1, \ldots, q-1$, Theorem 2.2 implies that the system is identifiable (i.e., two distinct parameters cannot produce identical observations). Therefore, if we can control $\eta(t)$, then we can guarantee identifiability. However, now suppose that $\eta(t)$ is uncontrollable and just so happens to be nonzero in just the last component, $j = q$. Then, Theorem 2.2 is no longer applicable. Thus, if we cannot control $\eta(t)$, then we can no longer even have a sufficient condition for identifiability under Theorem 2.2.

Following this discussion, we emphasize that system identifiability theory is distinct for one-dimensional and multi-dimensional inputs. Moreover, we see that system identifiability for multi-dimensional inputs has a clear dependence on whether the input, $\eta(t)$, is controllable or uncontrollable. Therefore, to make their dependence on the input explicit, we will denote the observations and states by $y(t;\theta,\eta)$ or $x(t;\theta,\eta)$ for a model θ. With this notation, we

can now define system identifiability for a controllable input, which we will refer to as weak identifiability, and for an uncontrollable input, which we will refer to as strong identifiability.

Definition 2.3 (Local, Weak Identifiability)**.** *Let $\theta \in \mathbb{R}^p$. If there exists a neighbohood of θ, $\mathcal{N}$, such that for any $\psi \in \mathcal{N}$, there is an input $\eta : [0,T] \to \mathbb{R}^q$ such that $y(t;\theta,\eta) \neq y(t;\psi,\eta)$ for some $t \in [0,T]$, then we say that the model, θ, is locally, weakly identifiable in the class defined by (2.2) with observations defined by (2.3).*

Remark 2.4. *We will continue to restrict ourselves to piecewise continuous inputs, which almost always occur for real phenomena. However, we will still refer to this case as weak identifiability without any further qualification.*

It is valuable to explicitly state how Definition 2.3 is specific to systems with controllable inputs. First, note that, in the definition, we begin by choosing θ and a ψ in a neighborhood of θ. Second, in the definition, we then *choose* (i.e., there exists) an input $\eta : [0,T] \to \mathbb{R}^q$ that results in distinct sets of observations for the two distinct models. Hence, because the ability to choose the input is the defining feature of a system with controllable inputs, we see that Definition 2.3 is specific to systems with controllable inputs.

Before formally defining strong identifiability, there are several other features of Definition 2.3 worth highlighting. First, Definition 2.3 considers identifiability locally. Therefore, Definition 2.3 also applies to the case where the parameter is restricted to some open $\Theta \subset \mathbb{R}^p$ without having to explicitly consider this open set. Moreover, the focus of Definition 2.3 on local identifiability allows for a greater range of parametrization choices, which will extend the utility of the theory. Finally, when $q = 1$, Definition 2.3 coincides with local analogues of the two preceding notions of identifiability.

Definition 2.4 (Local, Strong Identifiability)**.** *Let $\theta \in \mathbb{R}^p$. Let $\eta : [0,T] \to \mathbb{R}^q$. If there exists a neighborhood of θ, $\mathcal{N}$, such that $\forall \psi \in \mathcal{N} \setminus \{\theta\}$, $y(t;\theta,\eta) \neq y(t;\psi,\eta)$ for some $t \in [0,T]$, then we say that the model, θ, is locally, strongly identifiable in the class defined by (2.2) with observations defined by (2.3).*

Remark 2.5. *Just as we do for weak identifiability, we will restrict ourselves to piecewise continuous inputs, yet we will still refer to this case as strong identifiability without any further qualification.*

Just as we did for weak identifiability, it is worth underscoring how Definition 2.4 applies to uncontrollable inputs. In Definition 2.4, we start by fixing θ and $\eta : [0, T] \to \mathbb{R}^q$ before considering any alternative local models, $\psi \in \mathcal{N}$. Therefore, because the input is fixed before we consider identifiability, Definition 2.4 applies to uncontrollable inputs. Moreover, just as for weak identifiability, Definition 2.4 is specified locally, which allows for choosing an arbitrary, open parameter set $\Theta \subset \mathbb{R}^p$ instead of considering all of $\mathbb{R}^p$. Additionally, when $q = 1$, Definition 2.4 agrees with the definitions of observability and one-dimensional identifiability.

In the next chapter, we present necessary conditions for strong identifiability. We close this chapter with a simple lemma that should help clarify why we have labeled one notion as strong identifiability and the other as weak identifiability.

Lemma 2.5. *If a model is locally, strongly identifiable in the class (2.2) with observations (2.3), then the model is locally, weakly identifiable in the same class with the same observations.*

Proof. Let the model be denoted by θ. Then, for the given input, $\eta : [0, T] \to \mathbb{R}^q$, there is a neighborhood of θ, $\mathcal{N}$, in which for any $\psi \in \mathcal{N} \setminus \{\theta\}$, $y(t; \theta, \eta) \neq y(t; \psi, \eta)$. Then, we just use this same input, η, in the definition of weak identifiability to conclude that strong identifiability implies weak identifiability. ■

3 | Strong System Identifiability

In this chapter, we develop necessary conditions for when a model is locally, strongly identifiable — a concept that was formulated in Definition 2.4. Importantly, our development of these conditions will differ from the simultaneous equation approach that we used to prove Theorems 2.2 and 2.3; specifically, we develop the equivalent conditions using an optimization formulation. While this approach is not novel and it has parallels with the simultaneous equation approach, the optimization formulation approach is useful because, as we will see, it helps streamline the analysis for uncontrollable inputs.

The optimization formulation is presented in Section 3.1. While this optimization approach streamlines the results, it still recasts identifiability as an infinite dimensional problem. Therefore, in Section 3.2, we state fundamental lemmas for reducing the infinite dimensional problem to a lower dimensional problem. Using these lemmas, in Section 3.3, we state several necessary conditions for strong identifiability, and connect them to the classical results for weak identifiability.

3.1 An Optimization Formulation

We begin by recasting strong identifiability as a minimization problem. In order to do so, we must formulate an appropriate function to minimize. Fortunately, as we will see, choosing this function is rather straightforward. Then, our goal will be to use optimality criteria to find a condition that is equivalent to a model being locally, strongly identifiable. We begin by stating the appropriate objective function.

Lemma 3.1 (Strong Identifiability Function)**.** *For two distinct parameters ψ and θ, define the strong identifiability function by*

$$f(\psi;\theta,\eta) = \int_0^T \|y(t;\psi,\eta) - y(t;\theta,\eta)\|_2^2 \, dt. \tag{3.1}$$

Suppose $\eta : [0,T] \to \mathbb{R}^q$ is piecewise continuous. The model θ is locally, strongly identifiable in the class (2.2) with observations (2.3) if and only if there exists a neighborhood of θ, $\mathcal{N}$, such that $f(\psi;\theta,\eta) > 0$ for all $\psi \in \mathcal{N} \setminus \{\theta\}$ — that is, θ is a strict local minimizer of $f(\psi;\theta,\eta)$.

Proof. Suppose θ is locally, strongly identifiable. Then, there is a neighborhood of θ, $\mathcal{N}$, such that for any $\psi \in \mathcal{N} \setminus \{\theta\}$, $\exists \tau \in [0,T]$ such that $y(\tau;\psi,\eta) \neq y(\tau;\theta,\eta)$. Since η is piecewise continuous, $\exists \epsilon > 0$ such that $\|y(t;\psi,\eta) - y(t;\theta,\eta)\|_2^2 > 0$ for all $t \in [\max\{\tau-\epsilon,0\}, \min\{\tau+\epsilon,T\}]$. Therefore, $\forall \psi \in \mathcal{N} \setminus \{\theta\}$, $f(\psi;\theta,\eta) > 0$.

Now suppose there is a neighborhood of θ, $\mathcal{N}$, such that $\forall \psi \in \mathcal{N} \setminus \{\theta\}$, $f(\psi;\theta,\eta) > 0$. Therefore, for all $\psi \in \mathcal{N} \setminus \{\theta\}$, there is a set of nonzero measure on which $y(t;\psi,\eta) \neq y(t;\theta,\eta)$. ■

Before proceeding with the optimization formulation, we briefly digress by comparing the optimization problem for strong identifiability with a similar formulation for weak identifiability, which we state presently.

Lemma 3.2 (Weak Identifiability Function)**.** *For two distinct parameters ψ and θ, define the weak identifiability function by*

$$g(\psi,\eta;\theta) = \int_0^T \|y(t;\psi,\eta) - y(t;\theta,\eta)\|_2^2 \, dt \tag{3.2}$$

Suppose any input, $\eta : [0,T] \to \mathbb{R}^q$, is piecewise continuous. The model θ is locally, weakly identifiable in the class (2.2) with observations (2.3) if and only if there exists a neighborhood of θ, $\mathcal{N}$, such that for any $\psi \in \mathcal{N} \setminus \{\theta\}$, $\exists \eta : [0,T] \to \mathbb{R}^q$ such that $g(\psi,\eta;\theta) > 0$.

Proof. The proof is similar to the proof of Lemma 3.1. ■

Comparing the strong identifiability function, $f(\psi;\theta,\eta)$, and the weak identifiability function, $g(\psi,\eta;\theta)$, we see that the strong identifiability function has the input as a given, whereas

the weak identifiability function has the input as an argument. Again, by this comparison, we reinforce the concept that strong identifiability applies to an uncontrollable or given input, whereas weak identifiability applies to a controllable or selectable input.

Resuming with our main discussion, the reformulation of strong identifiability as an optimality point is useful if there is a simple way of verifying optimality. For example, if the strong identifiability function is twice continuously differentiable, we can use optimality criteria for strict local minimizers to verify if a particular model is a strict local minimizer, or, equivalently by Lemma 3.1, if it is strongly identifiable. However, the use of such optimality criteria requires verifying that the strong identifiability function is differentiable. The next result provides generic criteria for guaranteeing that functions similar to the strong identifiability function are differentiable.

Lemma 3.3 (Existence of a Stationary Point)**.** *Let $\theta \in \mathbb{R}^p$, and suppose $\psi \mapsto y(t,\psi)$ is twice continuously differentiable for all $t \in [0,T]$ in a neighborhood of θ, $\mathcal{N}$. Moreover, suppose that*

$$\int_0^T \sup_{\phi\in\mathcal{N}} \left\| \nabla_\psi^j y(t,\psi)\big|_{\psi=\phi} \right\|_2^2 dt < \infty,\ j=0,1,2, \tag{3.3}$$

where $\nabla_\psi^j y(t,\psi)\big|_{\psi=\phi}$ is the gradient of order j of $y(t,\psi)$ with respect to ψ and evaluated at ϕ. Let $f(\psi;\theta) = \int_0^T \|y(t,\psi) - y(t,\theta)\|_2^2\, dt$. Then, $\psi \mapsto f(\psi,\theta)$ is twice continuously differentiable in $\mathcal{N}$, $\nabla_\psi f(\psi;\theta)|_{\psi=\theta} = 0$, and

$$\nabla_\psi^2 f(\psi;\theta)\big|_{\psi=\theta} = \int_0^T \left[\nabla_\psi y(t,\psi)|_{\psi=\theta}\right]' \left[\nabla_\psi y(t,\psi)|_{\psi=\theta}\right] dt. \tag{3.4}$$

Proof. Showing that $f(\psi;\theta,\eta)$ is twice continuously differentiable requires a standard application of the mean-value theorem and the dominated convergence theorem using the condition in (3.3). Then, we have that the derivative and the integral commute, and, consequently, we have a closed form for the derivatives of f. We now use these closed forms to compute the gradient and Hessians of f.

For $j = 0, 1, 2$, let $y_j(t, \phi) = \nabla_\psi^j y(t, \psi)\big|_{\psi=\phi}$ and $f_j(\phi; \theta) = \nabla_\psi^j f(\psi; \theta)\big|_{\psi=\phi}$. Then,

$$f_1(\phi; \theta) = \int_0^T y_1(t, \phi)'(y(t, \phi) - y(t, \theta))dt, \tag{3.5}$$

and

$$f_2(\phi; \theta) = \int_0^T y_1(t, \phi)' y_1(t, \phi) + y_2(t, \phi)'(y(t, \phi) - y(t, \theta))dt. \tag{3.6}$$

Then, evaluating f_1 and f_2 at θ, we conclude that $f_1(\theta; \theta) = 0$ and $f_2(\theta; \theta)$ is (3.4). ■

Remark 3.1. *While we have not explicitly pointed out the dependence of y on η in this result, we know that when η is fixed, showing the dependence of y on η is just for greater clarity.*

We now bring these different results together to state an optimality criteria condition that is equivalent to strong identifiability. Specifically, we use Lemma 3.1 to restate strong identifiability as an optimality problem. Then, we use Lemma 3.3 to verify that the strong identifiability function is twice continuously differentiable. Finally, we use second-order optimality criteria for strict minimizers to state an equivalent condition for strong identifiability. This logic is summarized in the following result.

Theorem 3.1. *Suppose $\eta : [0, T] \to \mathbb{R}^q$ is piecewise continuous. Let $\theta \in \mathbb{R}^p$. Let $x(t; \phi, \eta)$ denote the solution to a model, ψ, in the family (2.2). Moreover, suppose $x_j(t; \phi, \eta) := \nabla_\psi^j x(t; \psi, \eta)\big|_{\psi=\phi}$ for $j = 0, 1, 2$ exist and are continuous in some neighborhood of θ, $\mathcal{N}$, and suppose $\int_0^T \sup_{\phi\in\mathcal{N}} \|x_j(t; \phi, \eta)\|_2^2 \, dt < \infty$ for $j = 0, 1, 2$. The model, θ, is locally, strongly identifiable if and only if*

$$\int_0^T x_1(t; \theta, \eta)' O' O x_1(t; \theta, \eta)dt \tag{3.7}$$

is positive definite.

Proof. By Lemma 3.1, θ is locally, strongly identifiable if and only if θ is a strict local minimizer of (3.1). Using Lemma 3.3 and standard optimality criteria (e.g., see Bertsekas,

1999, Section 1.1.2), θ is a strict local minimizer if and only if (3.7) is positive definite. ■

While Theorem 3.1 provides an equivalent, optimality-based condition for local, strong identifiability, we are left with verifying when a family of models satisfy the regularity conditions of the theorem. Fortunately, these regularity conditions hold under a rather general setting as we prove in the next result.

Lemma 3.4. *Suppose $\eta : [0, T] \to \mathbb{R}^q$ is piecewise continuous and fix $\theta \in \mathbb{R}^p$. Then, there exists a neighborhood of θ, $\mathcal{N}$, such that the observations generated by (2.3) for models defined in (2.2) satisfy (3.3) if $\psi \mapsto A(\psi)$, $\psi \mapsto B(\psi)$ and $\psi \mapsto \tilde{x}(\psi)$ are twice continuously differentiable in an open neighborhood of $\overline{\mathcal{N}}$.*

Proof. Note that

$$y(t; \psi, \eta) = O \exp(A(\psi)t) \left[\tilde{x}(\psi) + \int_0^t \exp(-A(\psi)\tau) B(\psi)\eta(\tau) d\tau \right]. \tag{3.8}$$

It follows readily that $(\psi, t) \mapsto O \exp(A(\psi)t)\tilde{x}(\psi)$ and $(\psi, t, \tau) \mapsto \exp(-A(\psi)\tau)B(\psi)\eta(\tau)$ are twice continuously differentiable in ψ in a neighborhood of $\overline{\mathcal{N}}$. Moreover, by piecewise continuity and compactness, the norms of these functions and their two derivatives with respect to ψ are bounded above on $\overline{\mathcal{N}} \times [0, T]$ or $\overline{\mathcal{N}} \times [0, T] \times [0, T]$, respectively.

Hence, working component-wise and using a standard mean-value theorem and dominated convergence argument, we can show that $y(t; \psi, \eta)$ is twice continuously differentiable with respect to ψ in $\mathcal{N}$. Then, again using the bounds on the norms of the mappings above and their derivatives, we can show that (3.3) holds. ■

Using Lemma 3.4, we can greatly simplify Theorem 3.1 when $A(\psi), B(\psi)$ and $\tilde{x}(\psi)$ are twice continuously differentiable. The next result states this simplification.

Corollary 3.1. *Suppose $\eta : [0, T] \to \mathbb{R}^q$ is piecewise continuous. Let $\theta \in \mathbb{R}^p$ and suppose $A(\psi), B(\psi), \tilde{x}(\psi)$ are twice continuously differentiable in a neighborhood of θ. The model θ in*

the class specified by (2.2) with observations specified by (2.3) is locally, strongly identifiable if and only if the following condition, which is well-specified, holds:

$$\forall m \in \mathbb{R}^p \setminus \{0\}, \exists t \in [0,T], \ \ s.t. \ 0 \neq Ox_1(t;\theta,\eta)m. \tag{3.9}$$

Proof. By Lemma 3.4, we have that x_1 exists and that the hypotheses of Theorem 3.1 hold. Therefore, θ is locally, strongly identifiable if and only if $\forall m \in \mathbb{R}^p \setminus \{0\}$,

$$\int_0^T \|Ox_1(t;\theta,\eta)m\|_2^2 \, dt > 0. \tag{3.10}$$

Using the continuity of x_1 as a function of t gives the desired result. ■

Equation 3.9 will be the keystone for our main results below. However, (3.9) will be needed in several different forms, which we will state after introducing some notation. Suppose $A(\psi), B(\psi), \tilde{x}(\psi)$ are twice continuously differentiable. Then, for $j = 1, \ldots, p$, denote

$$\frac{\partial}{\partial \psi_j} A(\psi)\bigg|_{\psi=\theta} = A_j(\psi), \quad \frac{\partial}{\partial \psi_j} B(\psi)\bigg|_{\psi=\theta} = B_j(\psi), \text{ and } \ \nabla_\psi \tilde{x}(\psi)|_{\psi=\theta} = \tilde{x}_1(\theta), \tag{3.11}$$

and let

$$\gamma(t;\theta,\eta) = \begin{bmatrix} A_1(\theta)x(t;\theta,\eta) + B_1(\theta)\eta(t) & \cdots & A_p(\theta)x(t;\theta,\eta) + B_p(\theta)\eta(t) \end{bmatrix}. \tag{3.12}$$

Also, for any $m \in \mathbb{R}^p$, let $\bar{A}(\theta, m) = \sum_{j=1}^p m_j A_j(\theta)$ and $\bar{B}(\theta, m) = \sum_{j=1}^p m_j B_j(\theta)$.

Corollary 3.2. *Suppose $\eta : \mathbb{R} \to \mathbb{R}^q$ is continuous on $[0,T]$ and zero outside of $[0,T]$. Let $\theta \in \mathbb{R}^p$ and suppose $A(\psi), B(\psi), \tilde{x}(\psi)$ are twice continuously differentiable in a neighborhood of θ. The following conditions are equivalent to (3.9):*

1. *For all nonzero $m \in \mathbb{R}^p$, $\exists s \in \mathbb{C}$ where s is not an eigenvalue of $A(\theta)$ such that*

$$0 \neq O\left(sI - A(\theta)\right)^{-1}\left[\tilde{x}_1(\theta) + \Gamma(s;\theta,\eta)\right]m, \tag{3.13}$$

where Γ denotes the Laplace transformation of γ and H is the Laplace transformation of η.

2. *For all nonzero $m \in \mathbb{R}^p$, $\exists s \in \mathbb{C}$ where s is not an eigenvalue of $A(\theta)$ such that*

$$\begin{aligned} 0 \neq & O\left(sI - A(\theta)\right)^{-1}\left[\tilde{x}_1(\theta)m + \bar{B}(\theta,m)H(s)\right] \\ & + O\left(sI - A(\theta)\right)^{-1}\bar{A}(\theta,m)\left(sI - A(\theta)\right)^{-1}\left[\tilde{x}(\theta) + B(\theta)H(s)\right]. \end{aligned} \tag{3.14}$$

Proof. First, recall that $x(t;\theta,\eta) = \tilde{x}(\theta) + \int_0^t A(\theta)x(\tau;\theta,\eta) + B(\theta)\eta(\tau)d\tau$. Therefore, by the regularity of $x(t;\theta,\eta)$,

$$\begin{aligned} \dot{x}_1(t;\theta,\eta)m &= \gamma(t;\theta,\eta)m + A(\theta)x_1(t;\theta,\eta)m \\ x_1(t;\theta,\eta)m &= \tilde{x}_1(\theta)m. \end{aligned} \tag{3.15}$$

Then, computing the Laplace transformation of x_1, denoted by X_1, which is well defined for any $s \in \mathbb{C}$ such that $\Re(s) > \|A\|$, we have

$$X_1(s;\theta,H)m = \left(sI - A(\theta)\right)^{-1}\left[\tilde{x}_1(\theta)m + \Gamma(s;\theta,\eta)m\right], \tag{3.16}$$

which we can extend by analytic continuation to all $s \in \mathbb{C}$ such that s is not an eigenvalue of A. Noting that a function's Laplace transformation is zero if and only if the function is zero, (3.9) is equivalent to the first condition.

Now, note that $\gamma(t;\theta,\eta)m = \bar{A}(\theta,m)x(t;\theta,\eta) + \bar{B}(\theta,m)\eta(t)$, and

$$X(s;\theta,H) = \left(sI - A(\theta)\right)^{-1}\left[\tilde{x}(\theta) + B(\theta)H(s)\right], \tag{3.17}$$

we can replace $\Gamma(s;\theta,\eta)$ in the first condition to conclude the equivalence with the second

condition. ■

With the equivalence of strongly identifiability established in Corollary 3.2, we now state several lemmas that will be useful in reducing the dimensionality of the necessary conditions for strong identifiability.

3.2 Fundamental Lemmas for Finiteness

In the one-dimensional identifiability results in Chapter 2, there were two key ingredients to restate one-dimensional identifiability as a finite-dimensional condition. First, one-dimensional identifiability is shown to be equivalent to the impulse response function being an injection. Second, using a Taylor expansion of the impulse response function, the independence of the monomials, and the Cayley-Hamilton theorem, the requirement that the impulse response function is an injection is reduced to verifying that a few terms of the impulse response function have a trivial null space.

Similar principles will be applied for the strong identifiability results presented below. For example, the next result is again the Cayley-Hamilton theorem.

Lemma 3.5. *Let $O \in \mathbb{R}^{n\times d}$, $A \in \mathbb{R}^{d\times d}$, and $F : \mathbb{C} \to \mathbb{C}^{d\times q}$ (or possibly defined on a subset of $\mathbb{C}$). $0 = OA^jF(z)$ for $j = 0, 1, \ldots$ if $0 = OA^jF(z)$ for $j = 0, \ldots, d-1$.*

Proof. We need to show that $0 = OA^jF(z)$ for $j = 0, \ldots, d-1$ implies that $0 = OA^jF(z)$ for $j \geq d$. By the Cayley-Hamilton Theorem, there exist constants $c_0, \ldots, c_{d-1}$ that depend on A, such that $A^d = c_0I + c_1A + \cdots c_{d-1}A^{d-1}$. Then, $OA^dF(z) = \sum_{k=0}^{d-1} c_kOA^kF(z) = 0$. Suppose $0 = OA^lF(z)$ for $l = d, \ldots, j-1$. Then, by the Cayley-Hamilton theorem, $OA^jF(z) = \sum_{k=0}^{d-1} c_kOA^{j-d+k}F(z) = 0$. ■

However, it will not always be the case that the Cayley-Hamilton theorem will appear in such a straightforward manner. In fact, as the next result shows, the Cayley-Hamilton theorem may appear in less obvious ways.

Lemma 3.6. *Let $O \in \mathbb{R}^{n\times d}$, $A, B \in \mathbb{R}^{d\times d}$, and $E_1, E_2 : \mathbb{C} \to \mathbb{C}^{d\times q}$. Now, define*

$$h_j(z) = OA^{j+1}E_1(z) + \sum_{k=0}^{j} OA^{j-k}BA^kE_2(z). \tag{3.18}$$

If $0 = OE_1(z)$ and $0 = h_j(z)$ for $j = 0, \ldots, 2(d-1)$ then $0 = h_j(z)$ for $j = 0, 1, \ldots$.

Proof. Define $\mathbb{D}_{-1} = 0$, $\mathbb{D}_0 = B$ and $\mathbb{D}_j = \mathbb{D}_{j-1}A + A^jB$ for $j \in \mathbb{N}$. Then,

$$h_j(z) = OA^{j+1}E_1(z) + O\mathbb{D}_jE_2(z). \tag{3.19}$$

We need only show that if $0 = OE(z)$ and $0 = OA^{j+1}E_1(z)+O\mathbb{D}_jE_2(z)$ for $j = 0, \ldots, 2(d-1)$ then $0 = OA^{j+1}E_1(z)+O\mathbb{D}_jE_2(z)$ for $j \geq 2d-1$. We will start with two general observations. First, by the Cayley-Hamilton theorem, there exist scalars $c_0, \ldots, c_{d-1}$ such that $A^d = c_0I + \cdots c_{d-1}A^{d-1}$. Therefore, for any $j \geq d-1$,

$$\begin{aligned} OA^{j+1}E_1(z) + O\mathbb{D}_jE_2(z) = &\sum_{k=0}^{d-1} c_k \left(OA^{j+1-d+k}E_1(z) + O\mathbb{D}_{j-d+k}E_2(z)\right) \\ &+ O\mathbb{D}_jE_2(z) - \sum_{k=0}^{d-1} c_kO\mathbb{D}_{j-d+k}E_2(z). \end{aligned} \tag{3.20}$$

Second, by the recursive definition of $\mathbb{D}_j$ and the Cayley-Hamitlon theorem, for $j \geq d$,

$$\begin{aligned} \mathbb{D}_j - \sum_{k=0}^{d-1} c_k\mathbb{D}_{j-d+k} &= \mathbb{D}_{j-1}A - \sum_{k=0}^{d-1} c_k\mathbb{D}_{j-1-d+k}A \\ &= \cdots = \mathbb{D}_{d-1}A^{j-d+1} - \sum_{k=0}^{d-1} c_k\mathbb{D}_{k-1}A^{j-d+1}. \end{aligned} \tag{3.21}$$

Now, applying our hypotheses and (3.20) and (3.21), we have that

$$0 = O\mathbb{D}_{d-1}A^lE_2(z) - \sum_{k=0}^{d-1} c_kO\mathbb{D}_{k-1}A^lE_2(z),\ l = 0, 1, \ldots, d-1. \tag{3.22}$$

We now proceed by induction. The base case is $j = 2d-1$. Applying the hypotheses and (3.20) and (3.21), we have

$$\begin{aligned} OA^{2d}E_1(z) + O\mathbb{D}_{2d-1}E_2(z) &= O\mathbb{D}_{2d-1}E_2(z) - \sum_{k=0}^{d-1} c_k O\mathbb{D}_{d+k-1}E_2(z) \\ &= O\left(\mathbb{D}_{d-1}A^d - \sum_{k=0}^{d-1} c_k \mathbb{D}_{k-1}A^d\right) E_2(z). \end{aligned} \tag{3.23}$$

Now, applying the Cayley-Hamilton theorem to A^d and using (3.22), we have that $h_{2d-1}(z) = 0$. If we assume that this holds up to some $j = l-1$ (for an appropriate l), then extending it to the case of $j = l$ follows the same argument. Thus, by induction, we have the result. ■

With these two results in hand, we are ready to prove our necessary conditions for strong identifiability.

3.3 Necessary Conditions for Strong Identifiability

Here, we will state necessary conditions for local, strong identifiability. These necessary conditions will require quite a bit of notation, and we will introduce the notation as we introduce the conditions. Using (2.11), let

$$W = \mathcal{M}(O, A(\theta)). \tag{3.24}$$

Theorem 3.2. *Suppose $\eta : \mathbb{R} \to \mathbb{R}^q$ is continuous and nontrivial on $[0,T]$ and zero elsewhere. Let $\theta \in \mathbb{R}^p$ and suppose $A(\psi)$, $B(\psi)$, and $\tilde{x}(\psi)$ are twice continuously differentiable in a neighborhood of θ. If the model θ in the class (2.2) with observations (2.3) is locally, strongly identifiable then*

$$\{0\} = \mathtt{null}\left(W\tilde{x}_1(\theta)\right) \cap \left\{ \bigcap_{t\in[0,T]} \mathtt{null}\left(W\gamma(t;\theta,\eta)\right) \right\}. \tag{3.25}$$

Proof. We will prove the inverse statement. Therefore, we suppose that there exists an $m \in \mathbb{R}^p \setminus \{0\}$ such that $W\tilde{x}_1(\theta)m = 0$ and $W\gamma(t;\theta,\eta) = 0$ for all $t \in [0,T]$. Computing the Laplace transform, we have that $W\tilde{x}_1(\theta)m = 0$ and $W\Gamma(s;\theta,\eta) = 0$ for $s \in \mathbb{C}$ that is not an eigenvalue of $A(\theta)$. Therefore, by Lemma 3.5, $OA(\theta)^j\tilde{x}_1(\theta)m = 0$ and $OA(\theta)^j\Gamma(s;\theta,\eta) = 0$ for $j = 0, 1, \ldots$ for $s \in \mathbb{C}$ that is not an eigenvalue of $A(\theta)$. Therefore, for any $|s| > \|A\|_2$,

$$0 = \sum_{j=0}^{\infty} \frac{1}{s^{j+1}} OA(\theta)^j \left[\tilde{x}_1(\theta) + \Gamma(s;\theta,\eta)\right] m. \tag{3.26}$$

Because Γ is analytic, by analytic continuation, we see that we have satisfied the first negation of the first condition of Corollary 3.2, which implies that θ is not locally, strongly identifiable. ■

Theorem 3.2 is particularly insightful owing to the following interpretation: if the parameter is strongly identifiable, then this is encoded in *local* information about the parameter. For example, the first term, $W\tilde{x}_1(\theta)$ is local information derived from the initial condition (see Theorem 2.1). Similarly, the remaining term, $W\gamma(t;\theta,\eta)$ reflects local information about the remainder of the trajectory system. However, we can improve on this result by recognizing that the local information about $x(t;\theta,\eta)$ encoded in $\gamma(t;\theta,\eta)$ depends only on the initial state condition and the input $\eta(t)$. Therefore, taking advantage of this information we have the following result (presented after some notation).

Let

$$\mathcal{I} = \begin{bmatrix} & W\tilde{x}_1(\theta) & \\ \mathcal{M}(WA_1(\theta), A(\theta))\tilde{x}(\theta) & \cdots & \mathcal{M}(WA_p(\theta), A(\theta))\tilde{x}(\theta) \end{bmatrix}, \tag{3.27}$$

and let the nullity of $\mathcal{I}$ be I. Let

$$\mathcal{T}_j = \begin{bmatrix} WB_j(\theta) \\ \mathcal{M}(WA_j(\theta), A(\theta))B(\theta) \end{bmatrix} \tag{3.28}$$

for $j = 1, \ldots, p$, and denote $\texttt{null}(\mathcal{T}) = \{m \in \mathbb{R}^p : \sum_{j=1}^p m_j \mathcal{T}_j = 0\}$. Finally, if $I > 0$, let Z be the matrix whose columns form an orthonormal basis for $\texttt{null}(\mathcal{I})$ and let $\tau_i = \sum_{j=1}^p Z_{ji} \mathcal{T}_j$ for $i = 1, \ldots, I$.

Theorem 3.3. *Under the setting of Theorem 3.2, if the model θ in the class (2.2) with observations (2.3) is locally, strongly identifiable then either $I = 0$ or both $\texttt{null}(\mathcal{I}) \cap \texttt{null}(\mathcal{T})$ is trivial and $\exists u \in \texttt{row}(\tau_1) + \cdots + \texttt{row}(\tau_I)$ such that $\int_0^T (\eta' u)^2 dt \neq 0$.*

Proof. Again, we will consider the inverse statement. That is, if $I > 0$ and either (case 1) $\texttt{null}(\mathcal{I}) \cap \texttt{null}(\mathcal{T})$ is nontrivial or (case 2) $\texttt{null}(\mathcal{I}) \cap \texttt{null}(\mathcal{T})$ is trivial and $\forall u \in \texttt{row}(\tau_1) + \cdots + \texttt{row}(\tau_I)$, $\int_0^T (\eta' u)^2 dt = 0$, then θ is not locally, strongly identifiable. By Theorem 3.2, it is sufficient for us to show that there is a nonzero $m \in \mathbb{R}^p$ such that either $0 = W \tilde{x}_1(\theta) m$ and $W \Gamma(s; \theta, H) m = 0$ for all s not an eigenvalue of $A(\theta)$. By substituting in for Γ, it is enough to show that $0 = W \tilde{x}_1(\theta) m$ and

$$0 = W \bar{A}(\theta, m) \left(sI - A(\theta)\right)^{-1} \left[\tilde{x}(\theta) + B(\theta) H(s)\right] + W \bar{B}(\theta, m) H(s) \tag{3.29}$$

for all s not an eigenvalue of $A(\theta)$.

Now, for Case 1, suppose there is a nonzero m in $\texttt{null}(\mathcal{I}) \cap \texttt{null}(\mathcal{T})$. Then $0 = W \tilde{x}_1(\theta) m$, $0 = W \bar{A}(\theta, m) A(\theta)^j \tilde{x}(\theta)$ for $j = 0, 1, \ldots$ (by Cayley-Hamilton), $0 = W B(\theta, m)$ and $0 = W \bar{A}(\theta, m) A(\theta)^j B(\theta)$. Therefore, using the same expansion and analytic continuation argument as in Theorem 3.2, the conclusion follows.

For Case 2, we have that for any nonzero $m \in \texttt{null}(\mathcal{I})$, $0 = W \tilde{x}_1(\theta) m$ and $0 = W \bar{A}(\theta, m) A(\theta)^j \tilde{x}(\theta)$ for $j = 0, 1, \ldots$, but $0 \neq \sum_{j=1}^p m_j \mathcal{T}_j$. Note, by the definition of $\tau_1, \ldots, \tau_I$, this is equivalent to saying that for all nonzero $\alpha \in \mathbb{R}^I$, $0 \neq \sum_{i=1}^I \alpha_i \tau_i$. However, by the condition on η, $0 = \sum_{i=1}^I \alpha_i \tau_i H(s)$. Therefore, using the same expansion and analytic continuation argument as in Theorem 3.2, the conclusion follows. ■

The preceding two results are more similar in form to Theorems 2.1 and 2.2, but what

about Theorem 2.3? The next result presents a necessary condition that is similar in structure to Theorem 2.3.

For $F : \mathbb{R}^p \to \mathbb{R}^{d\times d}$, let $\mathbb{D}_j(\theta, F) = \sum_{k=0}^{j} A(\theta)^{j-k} F(\theta) A(\theta)^k$ for $j = 0, 1, \ldots$. Let

$$\mathcal{J} = \begin{bmatrix} O\tilde{x}_1(\theta) \\ OA(\theta)\tilde{x}_1(\theta) \\ OA(\theta)^2\tilde{x}_1(\theta) \\ \vdots \\ OA(\theta)^{2d-1}\tilde{x}_1(\theta) \end{bmatrix} + \begin{bmatrix} 0 & \cdots & 0 \\ O\mathbb{D}_0(\theta, A_1)\tilde{x}(\theta) & \cdots & O\mathbb{D}_0(\theta, A_p)\tilde{x}(\theta) \\ O\mathbb{D}_1(\theta, A_1)\tilde{x}(\theta) & \cdots & O\mathbb{D}_1(\theta, A_p)\tilde{x}(\theta) \\ \vdots & & \vdots \\ O\mathbb{D}_{2d-2}(\theta, A_1)\tilde{x}(\theta) & \cdots & O\mathbb{D}_{2d-2}(\theta, A_p)\tilde{x}(\theta) \end{bmatrix}, \tag{3.30}$$

and let the nullity of $\mathcal{J}$ be J. Let

$$\mathcal{U}_j = \begin{bmatrix} OB_j(\theta) \\ OA(\theta)B_j(\theta) + O\mathbb{D}_0(\theta, A_j)B(\theta) \\ \vdots \\ OA(\theta)^{2d-1}B_j(\theta) + O\mathbb{D}_{2d-2}(\theta, A_j)B(\theta) \end{bmatrix} \tag{3.31}$$

for $j = 1, \ldots, p$, and denote $\texttt{null}(\mathcal{U}) = \{m \in \mathbb{R}^p : \sum_{j=1}^{p} m_j \mathcal{U}_j = 0\}$. Finally, if $J > 0$, let Ω be the matrix whose columns form an orthonormal basis for $\texttt{null}(\mathcal{J})$ and let $\nu_i = \sum_{j=1}^{p} \Omega_{ji}\mathcal{U}_j$ for $i = 1, \ldots, J$.

Theorem 3.4. *Under the setting of Theorem 3.2, if the model θ in the class (2.2) with observations (2.3) is locally, strongly identifiable then either $J = 0$ or both $\texttt{null}(\mathcal{J}) \cap \texttt{null}(\mathcal{U})$ is trivial and $\exists u \in \texttt{row}(\nu_1) + \cdots + \texttt{row}(\nu_J)$ such that $\int_0^T (\eta' u)^2 dt \neq 0$.*

Proof. Again, we will prove the inverse statement. By the second condition of Corollary 3.2 and analytic continuation, it is sufficient to prove that there exists a nonzero $m \in \mathbb{R}^p$ such

that

$$0 = \frac{1}{s} O(\tilde{x}_1(\theta)m + \bar{B}(\theta, m)H(s)) + \sum_{j=0}^{\infty} \frac{1}{s^{j+2}} \left\{ OA(\theta)^{j+1}(\tilde{x}_1(\theta)m + \bar{B}(\theta, m)H(s)) + O\mathbb{D}_j(\theta, \bar{A}(\cdot, m))(\tilde{x}(\theta) + B(\theta)H(s) \right\} \tag{3.32}$$

for $|s| > \|A\|_2$. We will prove that each term is zero under our conditions, which by Lemma 3.6, reduces to showing that the term for $1/s$ is zero and the first $2d-2$ terms in the sum are zero. The proof now proceeds in the same way as in Theorem 3.3. ■

The necessary conditions for strong identifiability presented above demonstrate the complexities of the case for multi-dimensional input and how the sufficient conditions for the one dimensional input fall short of completely addressing the identifiability for this problem. Thus, an important task is to establish sufficient conditions for strong identifiability, but we will not pursue this further here.

Incremental Estimation

4 | Computability

As is typical in the theory and practice of statistics, useful concepts such as hypotheses, experiments, observational units and estimators can only be defined and appraised within the context of an application. For example, consider the following definition of an estimator.

Definition 4.1 (Estimator)**.** *Let $\mathcal{Y}$ be a set called the observation space. Let Θ be a set called the parameter space. An estimator (of order N) is a mapping from $\mathcal{Y}^N$ to Θ.*

Because Definition 4.1 depends on the application-dependent notions of an observation space and a parameter space, an estimator is only well-specified in the context of an application. Moreover, because Definition 4.1 is simply stating that an estimator is a mapping, it does not provide any additional structure to say something interesting about estimators. For example, consider the observation space of $\mathcal{Y} = \mathbb{R}$ and the parameter space of $\Theta = \mathbb{R}_{\geq 0}$ that relates an observation, y, to a parameter, θ, by $y \sim \mathcal{N}(\theta, 1)$. Given a set of observations, $y_1, \ldots, y_N$, an example of an estimator is the mapping $(y_1, \ldots, y_N) \mapsto 0$. Another example of an estimator is the mapping $(y_1, \ldots, y_N) \mapsto |y_1| + |y_N|$. Importantly, both of these estimators might be incredibly useful or completely useless, and making this determination depends on the application in which the estimator is employed.

Because the utility of an estimator is application specific, we seem to preclude the development of a generic framework and theory for analyzing and evaluating estimators. Indeed, no such generic estimation theory framework exists; however, there are themes that run across applications that have allowed for the development of several useful estimation theories. One popular theme, formalized by Fisher (1925), is either when (1) the observations are assumed to be generated by a specific model in a family of models and the application

Disclaimer: Parts of this chapter are taken from a manuscript that I wrote, which is currently posted to the arXiv (Patel, 2017b).

requires determining this specific model, or (2) the distribution of the observations specifies a specific model in a family of models by some optimality criteria and the application requires determining this specific model. In both cases, we refer to the specific model in the family as the *true model*. In such a context, the utility of an estimator can be evaluated across general criteria such as bias, variance, admissibility, consistency, efficiency and sufficiency (Fisher, 1925; Shao, 2003). Importantly, evaluating estimators using these general criteria has led to rather rich classes of estimators and to equally rich theories of estimators, which, together, have been indispensable to the development of medicine, science and engineering over the past half century.

However, the past decade has witnessed the rise of applications that demand another criteria: computability. Computability is by no means a new or even recent criteria for evaluating estimators, as evidenced by the works of Cotes (1722); Legendre (1805); Gauss (1809), who, under their severe limits in computational power, invented what we call stochastic gradient descent[1] and recursive least squares.[2] In fact, computability was also a criteria considered at the foundations of modern statistics. For instance, under the family of probability distribution function models

$$y \sim \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2}, \tag{4.1}$$

and given independent observations $y_1, \ldots, y_N$ from a true model, θ^*, Fisher (1925) defined an *alternative estimator* to the maximum likelihood estimator defined by

$$\overline{\theta} = \underline{\theta} + \frac{2}{N} \sum_{i=1}^{N} \frac{2(y_i - \underline{\theta})}{1 + (y_i - \underline{\theta})^2}, \tag{4.2}$$

[1] See Gowing (2002) p. 107. Cotes describes a procedure of correcting observations of a celestial object to find the "most probable" position of the object. Importantly, this correction procedure is based on *local* rather than *global* errors in the observation, which is the procedure of stochastic approximation.

[2] See Appendix A of Simon (2006).

where $\underline{\theta}$ is selected as a median of the sample distribution.[3] Importantly, the estimator defined in (4.2) can be computed using a single pass through the observations after determining $\underline{\theta}$, whereas computing the maximum likelihood estimator would require multiple passes through the observations. Thus, the estimator in (4.2) is computationally more favorable in comparison to the maximum likelihood estimator.

Interestingly, a careful consideration of the computability notion of Fisher and the computability notions of Cotes, Gauss or Legendre reveals that these notions of computability are defined relative to the computing environment. Thus, because the notion of computability depends on the application's computing environment, we immediately preclude the possibility of a general estimation theory with respect to arbitrary notions of computability. However, just as before, there are themes across different applications that can be used to develop a theory. For example, the theme of applications with a large number of observations or streaming observations can be used to motivate the following definition for computability.

Definition 4.2 (Computability)**.** *Let $\mathfrak{F} = \{\theta_N : N \in \mathbb{N}\}$ be a sequence of estimators of order N. $\mathfrak{F}$ is computable if the complexity of computing θ_{N+1} given θ_N can be bounded independently of $N \in \mathbb{N}$.*

Remark 4.1. *In the definition of computability, we do not specify the type of complexity that we are considering. This vagueness is intentional, as it allows for us to choose the type of complexity such as storage, computational or communication. Moreover, we also do not specify how we measure complexity. Again, this vagueness is intentional, as it allows for us to choose how we measure complexity such as worst-case, average-case or smoothed.*

To give an example of applying Definition 4.2, consider two sequences of estimators, $\mathfrak{F}_1$ and $\mathfrak{F}_2$, in which an estimator of order N is defined according to (4.2). However, for the

[3]We stress that Fisher is defining an alternative estimator and *not* a computational algorithm for computing the maximum likelihood estimator. This distinction is important as it reminds us that we can only compare estimators to other estimators and computational algorithms to other computational algorithms. However, this distinction is not always well-posed as we will show.

estimators in $\mathfrak{F}_2$, use a fixed $\underline{\theta}$ for all of the estimators. To be explicit, denoting $\theta_N(\mathfrak{F}_j)$ as the order N estimator in the sequence $\mathfrak{F}_j$ for $j = 1, 2$, the two sequences of estimators are defined by

$$\theta_N(\mathfrak{F}_1) = \tilde{\theta}_N + \frac{2}{N}\sum_{i=1}^{N} \frac{2(y_i - \tilde{\theta}_N)}{1 + (y_i - \tilde{\theta}_N)^2}, \text{ where } \tilde{\theta}_N \in \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} |y_i - \theta|, \tag{4.3}$$

and, for some fixed $\underline{\theta} \in \Theta$,

$$\theta_N(\mathfrak{F}_2) = \underline{\theta} + \frac{2}{N}\sum_{i=1}^{N} \frac{2(y_i - \underline{\theta})}{1 + (y_i - \underline{\theta})^2}. \tag{4.4}$$

If we use worst-case complexity in our notion of computability, we easily see that $\mathfrak{F}_1$ is *not* computable: regardless of $\theta_{N-1}(\mathfrak{F}_1)$, the determination of $\tilde{\theta}_N$ will require a $\mathcal{O}(N)$ method to add an element to a list of sorted observations, and $\mathcal{O}(N)$ arithmetic operations to determine $\theta_N(\mathfrak{F}_1)$. On the other hand, we see that $\mathfrak{F}_2$ is computable because

$$\theta_N(\mathfrak{F}_2) = \frac{1}{N}\left(\underline{\theta} + \frac{2(y_N - \underline{\theta})}{1 + (y_N - \underline{\theta})^2}\right) + \frac{N-1}{N}\theta_{N-1}(\mathfrak{F}_2), \tag{4.5}$$

where the expression on the right hand side can be evaluated with a fixed computational and storage complexity regardless of N.[4]

To give another illustrative of example of Definitions 4.1 and 4.2, consider the space of observations $\mathcal{Y} = \mathbb{R}^N$. Moreover, let the first observation be $(y_1, \ldots, y_N)$ as generated from (4.1) from a true model, θ^*, and all subsequent observations be $0 \in \mathbb{R}^N$. Now, let the sequence $\{\theta_N : N \in \mathbb{N}\}$ be the iterates generated by gradient descent with Armijo backtracking for computing the maximum likelihood estimate of θ^* from $(y_1, \ldots, y_N)$. Then, using Definition 4.1, $\{\theta_N\}$ define a valid sequence of estimators. Moreover, using Definition 4.2,

[4]Recall that computability alone is not necessarily a good metric for an estimator. Thus, while $\mathfrak{F}_2$ might be computable, it might not be consistent, which, depending on the application, may be an even more important criteria for an estimator.

$\{\theta_N : N \in \mathbb{N}\}$ is computable. By a straightforward extension, any sequence of iterates generated by a practical optimization method will generate computable estimators according to our definitions.

At this point, we might ask what the difference is between the computable estimators generated by (4.4) and the computable estimators generated by optimizers. One obvious difference is the distinct goals of the estimators: the estimator generated by (4.4) is, presumably, trying to determine θ^*, whereas the estimator generated by an optimizers is, presumably, trying to determine the maximum likelihood estimator of order N for θ^*. While this difference may be sufficient to distinguish between these two types of estimators, consider the situation of estimating the maximum likelihood estimator of order N using stochastic gradient descent with i.i.d. samples with replacement over $\{y_1, \ldots, y_N\}$. We will then have two computable estimators (one from stochastic gradient descent, and one from an optimizer) and their goals are now identical. For many researchers, these two estimators intuitively belong to the same class, and, consequently should be comparable. However, this is a fundamentally flawed comparison as we now argue.

Consider the situation of computing the average of $\mathcal{Y} \subset \mathbb{R}$, where $\mathcal{Y}$ is finite with M unique elements. Suppose we now take three distinct approaches to computing the mean of $\mathcal{Y}$.

1. We sample the elements of $\mathcal{Y}$ uniformly without replacement. Then, knowing that we have M elements in $\mathcal{Y}$, we compute the mean of $\mathcal{Y}$ once we have sampled all M elements.

2. We sample the elements of $\mathcal{Y}$ independently and uniformly with replacement, and we keep track of the labels of each element sampled. Then, knowing that we have M elements in $\mathcal{Y}$, we compute the mean of $\mathcal{Y}$ once we have sampled all M.

3. We sample the elements of $\mathcal{Y}$ independently and uniformly with replacement. Then, *not knowing* that we have M elements in $\mathcal{Y}$, we compute the sample average of the current samples.

In the first approach, we will exactly compute the mean of $\mathcal{Y}$ within M samples. In the second approach, the probability of failing to compute the mean of $\mathcal{Y}$ by sample j, for j sufficiently large, decays exponentially with respect to j.[5] In the third approach, the probability that the absolute difference between sample mean at by sample j and the true mean exceeds any $\epsilon > 0$ decays as j^{-1}.

Clearly, all three approaches are computing the same value, but the first and second approaches are using the finiteness of $\mathcal{Y}$ that is unavailable in the third approach. Thus, the first and second approaches have faster rates of convergence owing to an information advantage in comparison to the third approach. This parallels the computable estimators discussed above: while the estimators are superficially comparable because they have the same stated purpose, the optimizer-based estimator is making use of the finiteness of the observation space, while the stochastic gradient descent estimator does not have this information. Thus, comparing these two estimators is based on a flawed premise.

In this work, motivated by the dynamical system identification and streaming data problems, we will consider estimators that operate without the assumption of a finite cardinality on the set of all possible observations. Moreover, owing to the large data and streaming data context, we will also focus on estimators that are computable. With this in mind, we consider the following estimators.

Definition 4.3 (Incremental Estimator)**.** *Let $\mathcal{Y}$ be an observation space (i.e., the set of all possible observations). Let $\Theta \subset \mathbb{R}^p$ be open and denote the parameter space. Let $\mathfrak{F} = \{\theta_N : N \in \mathbb{N}\}$ be a sequence of estimators. The estimators in $\mathfrak{F}$ are incremental estimators if $\mathfrak{F}$ is computable and the estimators do not make use of $|\mathcal{Y}|$.*

Just as with other criteria for evaluating estimators, computability alone will not be sufficient for a sequence of estimators to be useful; specifically, in large or streaming data applications, convergence to the true model at a practical rate will also be a valuable criteria.

[5] We simply bound the probability of failing to sample a subset of $\mathcal{Y}$, and show that the resulting failure probability decays exponentially.

Owing to our interest in a true model, we will assume that such a true model exists and can be specified as the minimizer (or root) of some estimating equation, m. Thus, we have two approaches to designing incremental estimators: (1) we can start with classical, consistent estimators and try to find incremental analogues, or (2) we can develop incremental estimators and then study the desired consistency and convergence rate properties. Here, we will take the latter approach. Accordingly, in Section 4.1, we develop a surprisingly powerful approach for generating incremental estimators. Then, in Section 4.2, we discuss one mechanism by which these incremental estimators will fail in practice. In Section 4.3, we discuss a generic approach for mitigating this mode of failure.

4.1 An Ansatz

Our ansatz consists of four elements: two principles from statistics and two principles from numerical optimization. The two statistical principles are based on the asymptotic normality of M-estimators and the plug-in principle, and the two optimization principles are based on Hessian approximation techniques and local models in iterative optimization. The statistical principles will be used to define an incremental likelihood function, and then the optimization principles will be used to derive computationally favorable approximations to the likelihood function. In concert, these principles will generate a collection of incremental estimators, such as stochastic gradient descent, AdaGrad and Natural Gradient Descent, which have typically been motivated from a purely optimization perspective. It is essential to keep in mind that while we will make assumptions below to establish the ansatz, the ansatz is a heuristic and our assumptions do not need to be satisfied in order to apply it to generate incremental estimators.

To begin our ansatz, we start with the standard M-estimation problem. Let $y_1, y_2, \ldots \in \mathcal{Y}$ be observations, let $\Theta \subset \mathbb{R}^p$ be an open set of parameters, let $m : \mathcal{Y} \times \Theta \to \mathbb{R}$, and define an estimator of order N by

$$\theta_N \in \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} m(y_i, \theta). \tag{4.6}$$

The estimator defined by (4.6) should be familiar as such estimators arise from empirical risk minimization in machine learning problems and generalized estimating equations in statistics.

Now, suppose we have observations $y_1, \ldots, y_N$ and we generate an estimator θ_N by (4.6). Moreover, suppose we have the estimator's distribution function (with respect to Lebesgue measure), $p_N(\theta)$, and, given this distribution function, we (recklessly) decide to throw away the observations. Now suppose we collect a new observation y_{N+1}, and we want to update our estimate θ_N to account for the new observation. How can we go about performing this update without $y_1, \ldots, y_N$?

One natural approach comes by considering the case in which (1) there exists a probability density function model for y given θ, (2) m is the negative log-likelihood of this probability density function, and (3) $y_1, \ldots, y_{N+1}$ are i.i.d. from the model θ^*. Then, the maximum aposteriori estimator for θ can be defined by

$$\theta_{N+1} \in \arg\min_{\theta \in \Theta} m(y_{N+1}, \theta) - \log p_N(\theta). \tag{4.7}$$

If $p_N(\theta)$ can be represented independently of $y_1, \ldots, y_N$ for all $N \in \mathbb{N}$ and if the minimization problem can be solved with a bounded complexity, then (4.7) defines a sequence of incremental estimators. Unfortunately, with the exception of very special cases, p_N cannot be represented independently of $y_1, \ldots, y_N$, nor can the minimization be solved with bounded complexity for arbitrary $m : \mathcal{Y} \times \Theta \to \mathbb{R}$. Consequently, (4.7) will not result in incremental estimators. However, we can partially salvage (4.7) if we are willing to make approximations. Our first approximation will be to replace $p_N(\theta)$ with the asymptotic distribution for estimators generated by (4.6).

4.1.1 Asymptotic Normality of M-Estimators

For the estimation functions, m, and the resulting estimators, $\mathfrak{F}$, we can show that $p_N(\theta)$ can be well-approximated (in the limit) by a normal distribution with a specific mean and variance, under the following mild assumptions.

Assumption 4.1 (M-Estimation Assumptions)**.** *Let $\Theta \subset \mathbb{R}^p$ be open, $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{Y}$ be an observation space, and $Y : \Omega \to \mathcal{Y}$ be a random variable. Moreover, let $m : \mathcal{Y} \times \Theta \to \mathbb{R}$ that satisfies*

1. *(Existence) The function $\mathbb{E}[m(Y, \theta)]$ has a local minimizer at $\theta^* \in \Theta$.*
2. *(Differentiability in Expectation) The function $\mathbb{E}[m(Y, \theta)]$ is twice continuously differentiable in a neighborhood of θ^*.*
3. *(Differentiability Almost Surely) Let $m_\psi(y, \theta) = \nabla_\psi m(y, \psi)|_{\psi=\theta}$. $m_\psi(Y, \theta^*)$ exists almost surely.*
4. *(Lipschitz) For every θ_1 and θ_2 in a neighborhood of θ^*,*

$$|m(Y, \theta_1) - m(Y, \theta_2)| \leq L(Y) \|\theta_1 - \theta_2\| \tag{4.8}$$

almost surely, where $\mathbb{E}[L(Y)^2] < \infty$.

Under these assumptions, if $\{\theta_N\}$ converges in probability to θ^*, then $p_N(\theta)$ can be approximated by the normal distribution as specified in the following result.

Theorem 4.1 (5.23 in Van der Vaart (1998))**.** *Suppose $m : \mathcal{Y} \times \Theta \to \mathbb{R}$, θ^* and Y satisfy Assumption 4.1. Let $Y_1, Y_2, \ldots$ be independent copies of Y. Moreover, let $\{\theta_N : N \in \mathbb{N}\}$ be estimators generated by (4.6) (with y_i replaced by Y_i) such that $\theta_N \to \theta^*$ in probability. Then, the distribution of θ_N converges to a normal distribution with mean θ^* and variance*

$$\frac{V_*^{-1} \Sigma_* V_*^{-1}}{N} \tag{4.9}$$

where V_* is the second derivative of $\mathbb{E}[m(Y,\theta)]$ evaluated at θ^*, and

$$\Sigma_* = \mathbb{E}\left[m_\psi(Y,\theta^*)m_\psi(Y,\theta^*)'\right]. \tag{4.10}$$

Under the usual regularity assumptions, $V^* = \mathbb{E}[m_{\psi\psi}(y,\theta^*)]$ where $m_{\psi\psi}(y,\theta)$ is equal to $\nabla^2_\psi m(y,\psi)\big|_{\psi=\theta}$. Moreover, if the second Bartlett Identity is satisfied (see Bartlett, 1953b,a), then $V_* = \Sigma_*$, and, consequently, the variance is

$$\frac{\Sigma_*^{-1}}{N}. \tag{4.11}$$

Therefore, using Theorem Theorem 4.1, we will replace $p_N(\theta)$ with the normal approximation to generate an estimator defined by

$$\theta_{N+1} \in \arg\min_{\theta\in\Theta} m(y_{N+1},\theta) + \frac{N}{2}\left\|\theta-\theta^*\right\|^2_{V_*\Sigma_*^{-1}V_*}. \tag{4.12}$$

Notice that (4.12) does not explicitly use $y_1,\ldots,y_N$ to define the subsequent estimator θ_{N+1}. Consequently, (4.12) is closer to defining an incremental estimator. However and quite obviously, we do not have θ^*, V_* nor Σ_*. Therefore, while (4.12) is an improvement over (4.7) with regard to computability, it is still not practical.

4.1.2 Plug-in Principle

In order to make (4.12) more practical, we will apply the plug-in principle to replace θ^* with θ_N. Thus, (4.12) becomes

$$\theta_{N+1} \in \arg\min_{\theta\in\Theta} m(y_{N+1},\theta) + \frac{N}{2}\left\|\theta-\theta_N\right\|^2_{V_*\Sigma_*^{-1}V_*}. \tag{4.13}$$

Note, the usual application and use of the plug-in principle requires Slutsky's theorem, which is a consequence of the continuous mapping theorem. For example, suppose we have i.i.d.

random variables, $X_1, X_2, \ldots, X_N$, and we use the sample mean to estimate the population mean. We know, from a standard calculation, that the expected squared error between the sample mean and population mean is $\mathbb{V}[X_1]/N$. Thus, if we wanted to provide a measure of error on our sample mean estimator, we cannot do so unless the true population variance, $\mathbb{V}[X_1]$, is known.

Fortunately, we can provide an approximate measure of error by using the plug-in principle; that is, we replace $\mathbb{V}[X_1]$ with the sample variance. Importantly, by the weak law of large numbers and Slutsky's theorem, the asymptotic distribution of the mean estimate with the sample variance in place of the actual variance will converge to a normal distribution. However, Slutsky's theorem cannot justify the plug-in principle in (4.13), since we do not know if $\{\theta_N\}$ converges to θ^* in probability. Regardless, as this is a heuristic approach, we are not concerned if the intermediate steps can be justified, so long as the outcomes have value. Yet, knowing how Slutsky's theorem will break down is useful when we are interested in proving properties of the generated incremental estimators. We now turn our attention to V_* and Σ_*.

4.1.3 Hessian Approximations

There are several approaches to handling the unknown values of V_* and Σ_*, which we term zero-order, first-order and second-order approximations. Zero-order approximations replace $V_*\Sigma_*^{-1}V_*/N$ with a scalar for each $N \in \mathbb{N}$. Owing to their simplicity, zero-order approximations are a rather popular approach (Murata, 1998; Bertsekas, 2011; Schaul et al., 2013; Toulis et al., 2014). In fact, zero-order approximations are better known as learning rates.

First order approximations replace V_* and Σ_* or the product $V_*\Sigma_*^{-1}V_*/N$ with a diagonal or nearly-diagonal approximation. First-order approximations include the popular techniques of Duchi et al. (2011); Tieleman and Hinton (2012); Zeiler (2012).

Second order approximations attempt to generate convergent estimators of V_* and Σ_*. For example, if mild regularity conditions are satisfied by $m_{\psi\psi}$, a second order approximation

might replace V_* and Σ_* by

$$V_N = \frac{1}{\eta} m_{\psi\psi}(y_N, \theta_{N-1}) + V_{N-1} \text{ and } \Sigma_N = \frac{1}{\gamma} m_\psi(y_N, \theta_{N-1}) m_\psi(y_N, \theta_{N-1})' + \Sigma_{N-1}, \quad (4.14)$$

where η and γ are some turning parameters. However, to our knowledge, there are no such second order approximation methods in the literature. In the special case when the second Bartlett identity is satisfied (such as for quasi-likelihood models), we need only estimate Σ_N. In this case, second order approximations are considered by the natural gradient descent method (Amari et al., 2000) and our approach (Patel, 2016).

Therefore, letting C_N denote either the zero, first or second order approximation to the unknown variance terms, (4.13) can be restated as

$$\theta_{N+1} \in \arg\min_{\theta \in \Theta} m(y_{N+1}, \theta) + \frac{1}{2} \|\theta - \theta_N\|^2_{C_N^{-1}}. \quad (4.15)$$

While we have derived (4.15) from predominantly statistical principles, we recognize that it has connections to the numerical optimization literature. For example, (4.15) is the proximal operator (Bertsekas, 2011), but allows for a different metric for each N. More generally, (4.15) can be viewed as a penalized form of the celebrated variable metric techniques (e.g., quasi-Newton) in numerical optimization (Davidon, 1991; Fletcher, 1970; Goldfarb, 1970), or as a preconditioned variant of Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963).

4.1.4 Local Model

By exploiting this connection to the optimization literature, we will consider another optimization principle to make (4.15) even more practical. In general, iterative numerical optimization methods generate a sequence of subproblems in order to solve the original optimization problem. These subproblems often take the form of local linear or quadratic model for the original objective function, which are less expensive to solve and can be shown to still converge to an optimal point.

In order to reduce the effort needed to determine θ_{N+1}, we can also apply this principle in (4.15) by replacing $m(y_{N+1}, \theta)$ with a local model, such as linear, quadratic or regression-like local models. To demonstrate these local models, we derive some examples.

Local Linear Model. Suppose we use a local linear model surrogate for $m(y_{N+1}, \theta)$. A natural choice for the point of expansion is θ_N. In this case, (4.15) becomes

$$\theta_{N+1} \in \arg\min_{\theta \in \Theta} m(y_{N+1}, \theta_N) + m_\psi(y_{N+1}, \theta_N)'(\theta - \theta_N) + \frac{1}{2} \|\theta - \theta_N\|^2_{C_N^{-1}}, \tag{4.16}$$

which, when $\Theta = \mathbb{R}^p$, can be explicitly solved to recover

$$\theta_{N+1} = \theta_N - C_N m_\psi(y_{N+1}, \theta_N). \tag{4.17}$$

Importantly, when C_N is a scalar, we recognize (4.17) as stochastic gradient descent (Chung, 1954). When C_N is a scaling of a fixed matrix for all N, we recognize (4.17) as preconditioned stochastic gradient descent (Murata, 1998).

Local Quadratic Model. Suppose we use a local quadratic model surrogate for $m(y_{N+1}, \theta)$. A natural choice for the point of expansion is θ_N. In this case, (4.15) becomes

$$\begin{aligned} \theta_{N+1} \in \arg\min_{\theta \in \Theta} & m(y_{N+1}, \theta_N) + m_\psi(y_{N+1}, \theta_N)'(\theta - \theta_N) \\ & + \frac{1}{2} \|\theta - \theta_N\|^2_{m_{\psi\psi}(y_{N+1}, \theta_N) + C_N^{-1}}, \end{aligned} \tag{4.18}$$

which, when $\Theta = \mathbb{R}^p$, can be explicitly solved to recover

$$\theta_{N+1} = \theta_N - \left(m_{\psi\psi}(y_{N+1}, \theta_N) + C_N^{-1}\right)^{-1} m_\psi(y_{N+1}, \theta_N). \tag{4.19}$$

To our knowledge, there are no methods in the literature that use this approach.

Local Lipschitz Model. Suppose we know that the gradient of $m(y_{N+1}, \theta)$ is Lipschitz continuous with a constant L (or even the Hessian is Lipschitz continuous). Then, we can use a higher-order surrogate model for $m(y_{N+1}, \theta)$ (see Birgin et al., 2017). In this case, (4.15) becomes

$$\theta_{N+1} \in \arg\min_{\theta \in \Theta} m(y_{N+1}, \theta_N) + m_\psi(y_{N+1}, \theta_N)'(\theta - \theta_N) + \frac{1}{2} \|\theta - \theta_N\|^2_{LI + C_N^{-1}}, \tag{4.20}$$

which, when $\Theta = \mathbb{R}^p$, can be explicitly solved to recover

$$\theta_{N+1} = \theta_N - \left(LI + C_N^{-1}\right)^{-1} m_\psi(y_{N+1}, \theta_N). \tag{4.21}$$

When C_N is a scalar proportional to N^{-1}, we recover the approach considered in Bottou et al. (2016), but with different interpretations for the constants.

Local Regression Model. Suppose, for a given choice of m and distribution over the observations, the expected Hessian of m is equal to the expected outer-product of the gradients of m (e.g., quasi-likelihood structure). In this case, (4.15) can be approximated by

$$\begin{aligned} \theta_{N+1} \in \arg\min_{\theta \in \Theta} & \, m(y_{N+1}, \theta_N) + m_\psi(y_{N+1}, \theta_N)'(\theta - \theta_N) \\ & + \frac{1}{2} \left[m_\psi(y_{N+1}, \theta_N)'(\theta - \theta_N)\right]^2 + \frac{1}{2} \|\theta - \theta_N\|^2_{C_N^{-1}}, \end{aligned} \tag{4.22}$$

which, when $\Theta = \mathbb{R}^p$, can be explicitly solved to recover (using the Sherman-Morrison Formula)

$$\theta_{N+1} = \theta_N - \frac{1}{1 + m_\psi(y_{N+1}, \theta_N)' C_N m_\psi(y_{N+1}, \theta_N)} C_N m_\psi(y_{N+1}, \theta_N). \tag{4.23}$$

Under some additional regularity conditions, this estimator is considered by natural gradient descent (Amari et al., 2000) and Kalman-based Stochastic Gradient Descent (Patel, 2016).

Remark 4.2. *It is worth noting that, depending on the problem, $m(y, \theta)$ may have a sufficiently simple structure to warrant solving (4.15) directly. In the case where C_N is a scalar, this approach is discussed by Bertsekas (2011) and Toulis et al. (2014).*

4.1.5 Computability of the Estimators

To summarize, we have proposed an ansatz to generate incremental estimators. Our approach generates these estimators by leveraging the asymptotic normality for M-estimators, the plug-in principle, Hessian approximations, and local modeling. However, we have not yet verified that the resulting estimators are indeed incremental estimators; that is, we have not yet verified if the estimators generated by our ansatz are computable in the sense of Definition 4.2.

In general, we cannot verify if the estimators are computable without discussing individual design choices. However, we can generally state that if C_N is computable, and (4.15) (with our without a surrogate local model) can be solved with a bounded complexity independent of N, then the sequence of estimators generated by our ansatz is computable. For example, if C_N is simply N^{-1} times a fixed scalar or matrix for all $N \in \mathbb{N}$, then the estimators in our examples, (4.17), (4.19), (4.21), and (4.23), are computable.

4.1.6 A Note on Consistency

When $Y, Y_1, Y_2, \ldots$ are identically distributed and $\mathbb{E}\left[m(Y, \theta)\right]$ is strongly convex, there is a standard approach to proving that incremental estimators of the form

$$\theta_{N+1} = \theta_N - M_{N+1} m_\psi(Y_{N+1}, \theta_N), \tag{4.24}$$

where $\{M_N : N \in \mathbb{N}\}$ are scalars or matrices, converge to some region of the minimizer of $\mathbb{E}\left[m(Y, \theta)\right]$.

Let θ^* denote the unique minimizer of $\mathbb{E}\left[m(Y, \theta)\right]$. The first step is to expand the Eu-

clidean distance between an estimator of order $N+1$ and θ^* in terms of θ_N:

$$\|\theta_{N+1} - \theta^*\|_2^2 = \|\theta_N - \theta^*\|_2^2 - 2(\theta_N - \theta^*)' M_{N+1} m_\psi(Y_{N+1}, \theta_N) + \|M_N m_\psi(Y_{N+1}, \theta_N)\|_2^2 . \quad (4.25)$$

Second, letting $\mathcal{F}_N = \sigma(\theta_0, \theta_1, \ldots, \theta_N)$, we must establish the following inequalities. There is a sequence of scalar-valued random variables, $\{\alpha_N\}$, adapted to $\mathcal{F}_N$, such that

$$\mathbb{E}\left[2(\theta_N - \theta^*)' M_{N+1} m_\psi(Y_{N+1}, \theta_N) \middle| \mathcal{F}_N\right] \geq \alpha_N \|\theta_N - \theta^*\|_2^2 . \quad (4.26)$$

There are nonnegative sequences of scalar-valued random variables, $\{\beta_N : N \in \mathbb{N}\}$ and $\{\gamma_N : N \in \mathbb{N}\}$, such that

$$\mathbb{E}\left[\|M_{N+1} m_\psi(Y_{N+1}, \theta_N)\|_2^2 \middle| \mathcal{F}_N\right] \leq \beta_N \|\theta_N - \theta^*\|_2^2 + \gamma_N . \quad (4.27)$$

Third, we apply these two inequalities to (4.25) to conclude

$$\mathbb{E}\left[\|\theta_{N+1} - \theta^*\|_2^2 \middle| \mathcal{F}_N\right] \leq (1 - \alpha_N + \beta_N) \|\theta_N - \theta^*\|_2^2 + \gamma_N . \quad (4.28)$$

Finally, as a consequence of the supermartingale convergence theorem, if $\sum_{N=0}^{\infty} \max\{\beta_N - \alpha_N, 0\} < \infty$ and $\sum_{N=0}^{\infty} \gamma_N < \infty$ almost surely then $\lim_{N\to\infty} \|\theta_N - \theta^*\|_2^2$ exists almost surely and is finite almost surely. Further, if there exists a nonnegative deterministic sequence $\{\delta_N\}$ such that $\delta_N \to 0$, $\sum_{N=0}^{\infty} \delta_N < \infty$ and $\mathbb{E}[\gamma_N]/\delta_N \to 0$ as $N \to \infty$, and

$$\mathbb{E}\left[\|\theta_{N+1} - \theta^*\|_2^2\right] \leq (1 - \delta_N)\mathbb{E}\left[\|\theta_N - \theta^*\|_2^2\right] + \mathbb{E}[\gamma_N] , \quad (4.29)$$

then $\mathbb{E}\left[\|\theta_N - \theta^*\|_2^2\right] \to 0$ as $N \to \infty$ by Lemma 1.5.1 of Bertsekas (1999).

Thus, when all of these steps can be verified, we can prove that the estimators not only converge but are consistent. However, following this standard approach is usually only

possible for rather simple incremental estimators. For more complex situations, such as when M_N is a random variable, this standard approach is typically futile, and we will require more sophisticated approaches.

4.2 Stagnation

While the incremental estimators generated by our ansatz can be found in the literature and may even be widely used, they are often subject to two forms of failure: stagnation and divergence. Stagnation refers to the situation in which a sequence of estimators effectively stops making progress towards the true model, θ^*. Divergence refers to the situation in which a the norms of a sequence of estimators diverges to infinity. Whether these phenomena occur will depend on the choice of $\{C_N\}$ and whether Θ is a bounded, convex subset of $\mathbb{R}^p$. In the latter case, divergence can be avoided by projecting the estimator back onto Θ. Moreover, when Θ is a bounded and convex, stagnation can be avoided with additional knowledge of Θ and the methodology introduced by Nemirovski et al. (2009). However, when we do not make such assumptions on Θ, the modes of failure are completely determined by the choice of $\{C_N\}$. In this chapter, we will focus on stagnation. When we discuss stochastic gradient descent in later chapters, we will discuss divergence.

As mentioned, stagnation is a phenomenon in which the a sequence of estimators effectively stops making progress towards the true model despite guarantees of convergence in theory. One mechanism behind stagnation is the incorrect choice in scaling when determining C_N. To illustrate, we reproduce the deterministic example of Nemirovski et al. (2009). Consider minimizing $m(\theta) = \theta^2/10$ with gradient descent and a scheduled step size of c/k where $c > 0$ is a tuning parameter and k is the iteration. Suppose that $\theta_0 = 1$, then $\{\theta_N : N \in \mathbb{N}\}$ are given by

$$\theta_N = \theta_{N-1} - \frac{c}{N} m_\psi(\theta_{N-1}) = \theta_{N-1}\left[1 - \frac{c}{5N}\right] = \cdots = \prod_{k=1}^{N}\left[1 - \frac{c}{5k}\right]. \tag{4.30}$$

A straightforward bound of the right hand side shows that

$$|\theta_N| \leq \mathcal{O}\left[N^{-c/5}\right]. \tag{4.31}$$

Therefore, when $c \geq 5$, we can guarantee relatively fast rates of convergence when using this scheduled step sizes of the given form. When $c \in (0, 5)$, the upper bound becomes impractically slow. Importantly, as shown in the example of Nemirovski et al. (2009) for $c = 1$, the order of the upper bound is also the order of the lower bound: $\theta_N > 0.8N^{-1/5}$. Thus, even after $N = 10^{10}$ iterations, the absolute error will be on the order of 10^{-2}.

This mechanism of stagnation naturally extends to the infinite observation space as well. For example, let $Q \in \mathbb{R}^{p\times p}$ be an orthonormal matrix. Let $X, X_1, X_2, \ldots \in \mathbb{R}^d$ be independent random vectors that are drawn from the columns of Q such that each column has an equal probability of being drawn. Now, let $\epsilon, \epsilon_1, \epsilon_2, \ldots \in \mathbb{R}$ be independent random variables that are uniformly distributed over $(-1, 1)$. Let $\theta^* \in \mathbb{R}^p$ be an arbitrary vector, and define

$$Z = X'\theta^* + \epsilon \quad \text{and} \quad Z_N = X'_N\theta^* + \epsilon_N \quad \forall N \in \mathbb{N}. \tag{4.32}$$

Given the observations $Y = (Z, X)$ and $\{Y_N = (Z_N, X_N) : N \in \mathbb{N}\} \subset [-1, 1] \times \{Q_1, \ldots, Q_p\}$, where Q_j denotes the j^{th} column of Q, we consider the ordinary least squares estimating equation, $m(Y, \theta) = \frac{1}{2}(Y - X'\theta)^2$.

Given that the Hessian of the expected value of $m(Y, \theta)$ is the identity matrix, we will attempt to estimate θ^* using a zero-order, linear model incremental estimator (i.e., stochastic gradient descent) with $C_N = N^{-\nu}$ for $\nu \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Implementing a random instantiation of this problem with $p = 100$ and running each case of the incremental estimator from the same initialization $\theta_0 = 0 \in \mathbb{R}^{100}$, Fig. 4.1 shows that the stagnation phenomenon persists.

To further substantiate this point, for each of the learning rates, we run the incremental

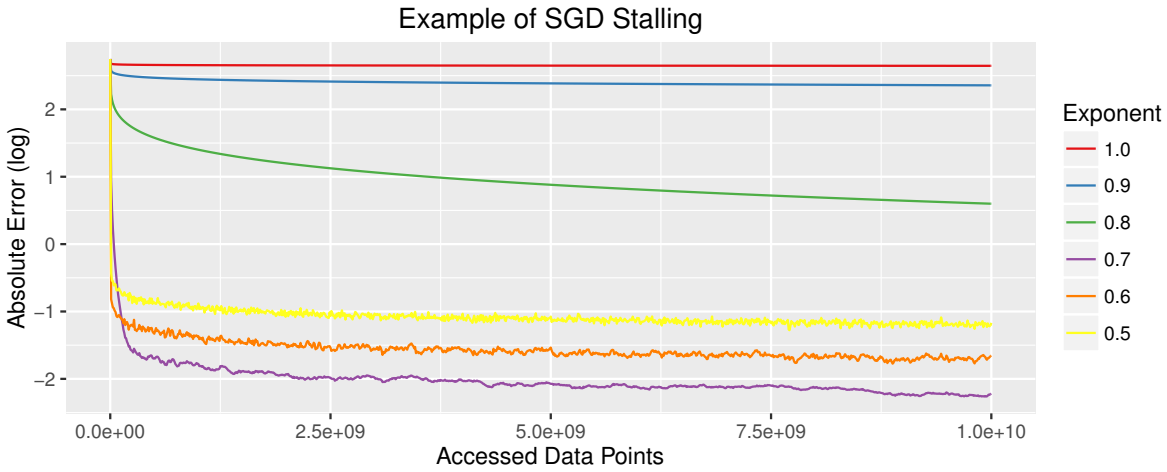


Figure 4.1: The absolute error (logarithmic scale) between the zero-order, linear model incremental estimator (i.e., SGD) and the true parameter for different choices of C_N (i.e., learning rates). We observe that all of the incremental estimators stagnate.

estimator 100 times independently on 10^8 observations. In Table 4.1, we report the summary statistics of the absolute errors for each of the learning rates, and we report the fraction of estimators that achieve an absolute error less than 10^{-1}. For comparison, we also report the same values when we solve the ordinary least squares (OLS) problem exactly. Table 4.1 reiterates that the learning rate plays a dominant role in determining the convergence rate of the incremental estimator.

Table 4.1: A Summary of SGD Convergence Statistics for the Linear Regression Problem

Method	Statistics					
	Mean	Median	Variance	Max	Min	Fraction
OLS	0.0287	0.0287	4.7116e-6	0.0361	0.0233	1.0
SGD, $\nu = 1.0$	466.9528	467.8384	29.7957	477.0266	446.7251	0.0
SGD, $\nu = 0.9$	330.2472	330.0679	13.7920	338.6541	320.6750	0.0
SGD, $\nu = 0.8$	80.8682	80.9932	1.4741	83.0688	77.4359	0.0
SGD, $\nu = 0.7$	0.1394	0.1392	2.1745e-5	0.1515	0.1282	0.0
SGD, $\nu = 0.6$	0.0814	0.0807	3.6933e-5	0.1008	0.0640	0.99
SGD, $\nu = 0.5$	0.2022	0.2017	2.1454e-4	0.2559	0.1625	0.0

While the examples that we have given thus far have been specific to stochastic gradient descent, stagnation is a general issue that impacts nearly all incremental estimators. For example, we will consider solving a simple linear system problem with a first-order, linear model approximation to (4.15). The linear system problem has a true parameter, $\theta^* = 1$, and two-dimensional observations in the space $\mathcal{Y} = \{(x, x) : x \in [1, 2]\}$, where the first component represents the coefficient and the second component represents the constant. That is, for any $y \in \mathcal{Y}$, the first component $y_{\cdot,1}$ and the second component $y_{\cdot,2}$ are related by $y_{\cdot,2} = \theta^* y_{\cdot,1}$.

Now, for this problem, we can define $m : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ to be $m(y, \theta) = \frac{1}{2}(y_{\cdot,2} - \theta y_{\cdot,1})^2$. Consequently, $m_\psi(y, \theta) = -y_{\cdot,1}(y_{\cdot,2} - \theta y_{\cdot,1})$. Now, let $Y_1, Y_2, \ldots \in \mathcal{Y}$ be independent, uniformly distributed random variables over $\mathcal{Y}$. For this problem, motivated by the estimation of Σ_*, we will consider the first-order linear approximation to (4.15) that generates a sequence of estimators given by

$$\theta_{N+1} = \theta_N - C_N m_\psi(Y_{N+1}, \theta_N), \tag{4.33}$$

where θ_0 is arbitrary and

$$C_N = \frac{\eta}{\sum_{j=1}^N m_\psi(Y_{j+1}, \theta_j)^2} \tag{4.34}$$

where $\eta > 0$ is a tuning parameter (note, this estimator is computable as the denominator can be stored and updated at each iteration). In the context of the linear systems problem, the update is

$$\theta_{N+1} = \theta_N + \eta \frac{Y_{N+1,1}(Y_{N+1,2} - \theta_N Y_{N+1,1})}{\sum_{j=1}^N Y_{j+1,1}^2 (Y_{j+1,2} - \theta_j Y_{j+1,1})^2}, \tag{4.35}$$

which has an absolute error of

$$|\theta_{N+1} - \theta^*| = |\theta_N - \theta^*| \left| 1 - \frac{\eta Y_{N+1,1}^2}{\sum_{j=0}^N Y_{j+1,1}^2 (\theta_j - \theta^*)^2} \right|. \tag{4.36}$$

From (4.36), we see that if θ_0 is far from θ^* and η is not correctly selected, the rate of

convergence of $\{\theta_N\}$ will be impractically slow. In fact, if $\theta_0 - \theta^*$ is sufficiently large, then we will not make any progress owing to finite arithmetic precision issues.

At this point, we might ask how we can choose C_N for incremental estimators in order to avoid stagnation. Well, if we have oracle knowledge of the problem, there are very simple ways of selecting C_N to ensure that an incremental estimator converges (see Bottou et al., 2016, Section 4). However, in any real problem, having oracle information is unrealistic. Therefore, while they will not relieve incremental estimators of the problem completely, we now discuss a generic tool for mitigating stagnation.

4.3 Restarts

From the example of Nemirovski et al. (2009), we notice that the first few iterations of the incremental estimator generate the most progress towards the true model. Similarly, from our numerical examples on linear regression, we also notice that the first few iterations of the incremental estimator generate the most progress towards the true model. Thus, we might naturally conclude that stagnation is a large iteration issue: that is, we only observe this problem when the number of iterations grows.

Moreover, we might conclude that if we repeatedly use the learning rate in the first few iterations, then we can avoid stagnation. That is, we might choose to *restart* C_N back to C_1 at a given iteration. To be specific, suppose we have an incremental estimator that generates θ_{N+1} from θ_N according to the rule $\mathcal{A} : \mathcal{Y} \times \Theta \times \mathcal{S} \to \Theta \times \mathcal{S}$, where $\mathcal{S}$ represents the space of hyperparameters (e.g., the choices of C_N). Under this notation, Algorithm 1 presents a generic algorithm for resetting such incremental estimators at arbitrary iteration intervals (called triggers) $\tau_1, \tau_2, \ldots \in \mathbb{N} \cup \{\infty\}$.

The idea of restarts is not particularly novel. For example, restarts are used in nonlinear conjugated gradient methods to avoid excess computational costs and take advantage of negative curvature (see Nocedal and Wright, 2006, Chapter 5.2). Thus, using restarts for incremental estimators, owing to their connection to numerical optimization techniques, is

Algorithm 1: Generic Reset Algorithm

Data: Parameter θ_0, Hyper Parameter S_0, Increasing restart triggers $\tau_1, \tau_2, \cdots \in \mathbb{N} \cup \{\infty\}$
Result: Parameter θ

```
θ, S ← θ_0, S_0
k, j ← 0, 1
while true do
    Read new observation, Y
    θ, S ← A(Y, θ, S)
    k ← k + 1
    if k == τ_j then
        k, j ← 0, j + 1
        S ← S_0
    end
end
```

not groundbreaking by any means. In fact, we used restarts in a numerical example in our work on Kalman Filtering (Patel, 2016). This topic was later explored experimentally by Loshchilov and Hutter (2016).

However, unlike the deterministic case, restarts for incremental estimation require carefully chosen restart triggers. In particular, in order for an incremental estimator to converge (in probability), we need the the variance of the estimator to converge to 0. Thus, if we do not allow the values of C_N to decay, then we will prevent the estimator from converging. In Patel (2017b), we proved that, when C_N are allowed to decay, SGD with restarts converges. Here, we generalize this result.

Theorem 4.2. *Let $\{Y_N : N \in \mathbb{N}\} \subset \mathcal{Y}$ be i.i.d random variables. Let $(\theta_0, S_0) \in \Theta \times \mathcal{S}$ be an arbitrary initialization and let $(\theta_{N+1}, S_{N+1}) = \mathcal{A}(Y_{N+1}, \theta_N, \mathcal{S}_N)$ for $N = 0, 1, \ldots$. Let $\alpha \geq 1$ and suppose the following three statements hold.*

1. *(Convergence) Suppose for every $(\theta_0, S_0) \in \Theta \times \mathcal{S}$, there is a point $\theta^* \in \Theta$ such that*

$$\lim_{N \to \infty} \mathbb{E}\left[\|\theta_N - \theta^*\|_\alpha^\alpha\right] = 0. \tag{4.37}$$

2. (Stability) For any point of convergence $\theta^* \in \Theta$ *and* $S_0 \in \mathcal{S}$*, let* $L(\theta^*, S_0)$ *denote all* $\theta_0 \in \Theta$ *such that the resulting sequence of estimators converges in the sense of (4.37). Suppose for any* $\theta_0 \in L(\theta^*, S_0)$*,* $\{\theta_N : N \in \mathbb{N}\} \subset L(\theta^*, S_0)$ *almost surely.*

3. (Uniform Rate Control) Fix $L(\theta^*, S_0)$*. For any* $\epsilon > 0$*,* $\exists K \in \mathbb{N}$ *such that if* $k \geq K$ *and* $\theta_0 \in L(\theta^*, S_0)$ *then*

$$\mathbb{E}\left[\|\theta_N - \theta^*\|_\alpha^\alpha\right] < \epsilon. \tag{4.38}$$

Let $\tau_1 < \tau_2 < \cdots \in \mathbb{N} \cup \{\infty\}$ *be stopping times relative to the filtration generated by the estimators and hyperparameters, and let* $T_j = \sum_{l=1}^{j} \tau_l$*. Then, for any* $(\theta_0, S_0) \in \Theta \times \mathcal{S}$ *with* $\{\theta_N, S_N\}$ *generated according to Algorithm 1, there is a* $\theta^* \in \Theta$ *such that*

$$\lim_{j \to \infty} \mathbb{E}\left[\left\|\theta_{T_j} - \theta^*\right\|_\alpha^\alpha\right] = 0. \tag{4.39}$$

Proof. We begin by establishing several facts.

1. First, fix $(\theta_0, S_0) \in \Theta \times \mathcal{S}$. Then, by the convergence and stability properties, there is a point $\theta^* \in \Theta$ such that $\theta_0 \in L(\theta^*, S_0)$. By the the stability property and because the observations are exchangeable, $\{\theta_{T_j} : j \in \mathbb{N}\} \subset L(\theta^*, S_0)$ almost surely.

2. Second, by the uniform rate control property, for any $\epsilon > 0$ there exists $K \in \mathbb{N}$ such that for any $\theta_0 \in L(\theta^*, S_0)$ and $k \geq K$, $\mathbb{E}\left[\|\theta_k - \theta^*\|_\alpha^\alpha\right] < \epsilon$. Therefore, because $\theta_{T_j} \in L(\theta^*, S_0)$ for the sequence of estimators, $\{\psi_{N,j} : N \in \mathbb{N}\}$, that are initialized with $\psi_{0,j} = \theta_{T_j}$ and S_0, $\mathbb{E}\left[\|\psi_{k,j} - \theta^*\|_\alpha^\alpha \middle| \psi_{0,j}\right] < \epsilon$ for any $k \geq K$.

3. Third, because $\tau_1, \tau_2 \in \mathbb{N} \cup \{\infty\}$ are integer-valued and increasing, $j \leq \tau_{j+1}$ for a $j \in \mathbb{N} \cup \{0\}$.

Applying our three facts, for any $\epsilon > 0$, there is a $K \in \mathbb{N}$ such that for any $j \geq K$,

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_{T_{j+1}} - \theta^*\right\|_\alpha^\alpha \middle| \theta_{T_j}\right] &= \sum_{l=j}^{\infty} \mathbb{E}\left[\left\|\psi_{l,j} - \theta^*\right\|_\alpha^\alpha \middle| \theta_{T_j}, \tau_{j+1} = l\right] \mathbb{P}\left[\tau_{j+1} = l \middle| \theta_{T_j}\right] \\ &< \sum_{l=j}^{\infty} \epsilon \mathbb{P}\left[\tau_{j+1} = l \middle| \theta_{T_j}\right] < \epsilon.\end{aligned} \tag{4.40}$$

Because this bound holds with probability one, we conclude that $\mathbb{E}\left[\left\|\theta_{T_{j+1}} - \theta^*\right\|_\alpha^\alpha\right] < \epsilon$ for all $j \geq K$. ■

Theorem 4.2 simply guarantees that if we started with a convergent incremental estimator, then we end with a convergent incremental estimator after employing restarts. Hence, while Theorem 4.2 tells us that a restarted incremental estimator is still convergent (under some stringent requirements), it does not convince us that the rate of convergence is any better than the original estimator. That is, Theorem 4.2 does not definitively say that restarting mitigates stagnation. Showing this is an ongoing effort. However, for now, we use two numerical examples to demonstrate that restarting mitigates stagnation in practice.

4.3.1 Linear Regression

First, recall the linear regression problem above that we solved using stochastic gradient descent with $C_N = N^{-\nu}$ where $\nu \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Here, we solve same problem solved with stochastic gradient descent with the same learning rates but with deterministic restart triggers. Fig. 4.2 shows the analogous output of Fig. 4.1. Similarly, Table 4.2 shows the analogous output of Table 4.1. Comparing these results, we observe that the restart strategy generally mitigates stagnation, but does not entirely eliminate it.

4.3.2 Neutrino Classification

Data Set, Task and Model. In the following experiments, we use a data set available on the UCI Repository which was generated by a Fermi Lab experiment used to test techniques for differentiating between electron neutrinos, considered the signal, and muon neutrinos,

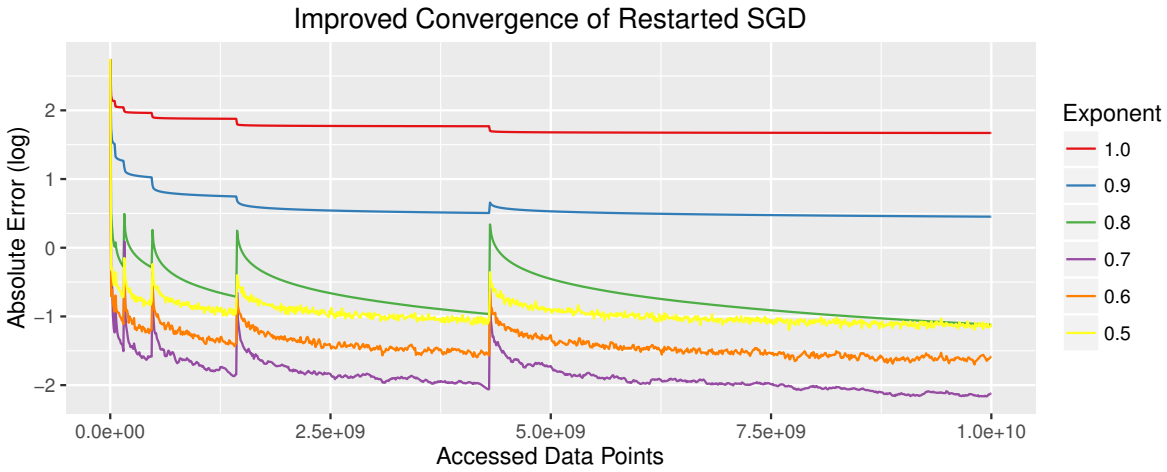


Figure 4.2: The absolute error (logarithmic scale) between the zero-order, linear model incremental estimator (i.e., SGD) and the true parameter for different choices of C_N (i.e., learning rates) with restarts. We observe that all the restarted incremental estimators have mitigated stagnation in comparison to the standard incremental estimators in Fig. 4.1.

considered the background (Roe et al., 2005). The data set contains 130,064 examples where the first 36,499 examples correspond to electron neutrinos, and the remaining 93,565 correspond to muon neutrinos, and each example has a fifty dimensional feature vector. The task is to discern between the electron and muon neutrinos using the feature vector.

The data set was preprocessed by prepending a 1 to the example if it corresponded to a signal, and a 0 otherwise. The data set was randomly shuffled and 91,044 examples

Table 4.2: A Summary of Restarted SGD Convergence Statistics for the Linear Regression Problem

Method	Statistics					
	Mean	Median	Variance	Max	Min	Fraction
SGD, $\nu = 1.0$	117.9137	118.5475	31.4797	130.7890	105.7674	0.0
SGD, $\nu = 0.9$	20.9212	21.0031	1.1371	23.5030	18.1436	0.0
SGD, $\nu = 0.8$	0.8612	0.8177	0.0265	1.3330	0.5676	0.0
SGD, $\nu = 0.7$	0.0476	0.0474	1.1066e-5	0.0563	0.0406	1.0
SGD, $\nu = 0.6$	0.1029	0.1018	5.0506e-5	0.1216	0.0851	0.34
SGD, $\nu = 0.5$	0.2460	0.2480	3.4880e-4	0.2810	0.1913	0.0

(approximately seventy percent) of the data was used as the training set, and the remaining 39,020 examples were left as the testing set.

The training data set was used to train a two-layer feed forward neural network with the following architecture:

1. The observation layer contained one neuron with a logistic activation function with five inputs and one output.

2. The hidden layer contained five neurons, each with a logistic activation function. The output of these five neurons fed into the observation layer neuron. Each of the five neurons was arbitrarily assigned ten of the feature vectors without overlap.

The resulting model had $p = 61$ dimensional parameter vector to be learned.

Experimental Set Up. The experiments listed below were run on a machine with an Intel i5 Processor (3.33 GHz) with nearly 4 GB of memory.

Seven incremental estimators were used to train the model using the training data with exactly 30 epochs and were all initialized at exactly the same random value: SGD (zero-order, linear model, see Chung (1954)), restarted SGD, AdaGrad (first-order, linear model with variance control,[6] see Duchi et al. (2011)), restarted AdaGrad, kSGD (second-order, regression model, see Patel (2016)), restarted kSGD, and BFGS (see Nocedal and Wright, 2006, Chapter 8).

As the goal of the experiment is to compare standard methods against their restarted analogues, there was little effort to optimize the hyperparameters for the methods. The learning rate for SGD and restarted SGD were arbitrarily set to $C_N = N^{-0.7}$. For AdaGrad and restarted AdaGrad, the multiplicative factor was set to $\eta = 0.001$. For restarted AdaGrad, the adaptive learning rate was set to the vector of ones at each trigger. For kSGD and restarted kSGD, the hyperparameter was set to $\gamma^2 = 0.01$. For restarted kSGD, the

[6] Although it is not viewed as such, the square root in AdaGrad is used to mitigate stagnation as well, but through a different mechanism. We refer to such mechanisms as variance control. We are working on a more complete discussion of these techniques, but we will not include them in this work.

covariance estimate is reset to the identity at each trigger. For BFGS, the maximal line search length was $\alpha_0 = 1$, the line search reduction factor was $\rho = 0.5$, and the Armijo condition parameter was set to $c = 0.0001$.

The triggering events were set to deterministic values. In particular the first triggering event occurred at iteration 100. All future triggering iteration occurred a factor of 1.56 times the previous triggering iteration. The factor of 1.56 was selected to ensure a large number of restarts occurred (each method restarted 22 times) and that the method was just shy of another restart by the end of training. This factor was selected before any experimentation was done.

For each of the stochastic methods, the parameter was recorded every 5,000 iterations and at the last iteration. For BFGS, the parameter was recorded at the end of each iteration. For each recorded parameter, the training and testing error were computed. For the last recorded parameter, the total gradient was computed.

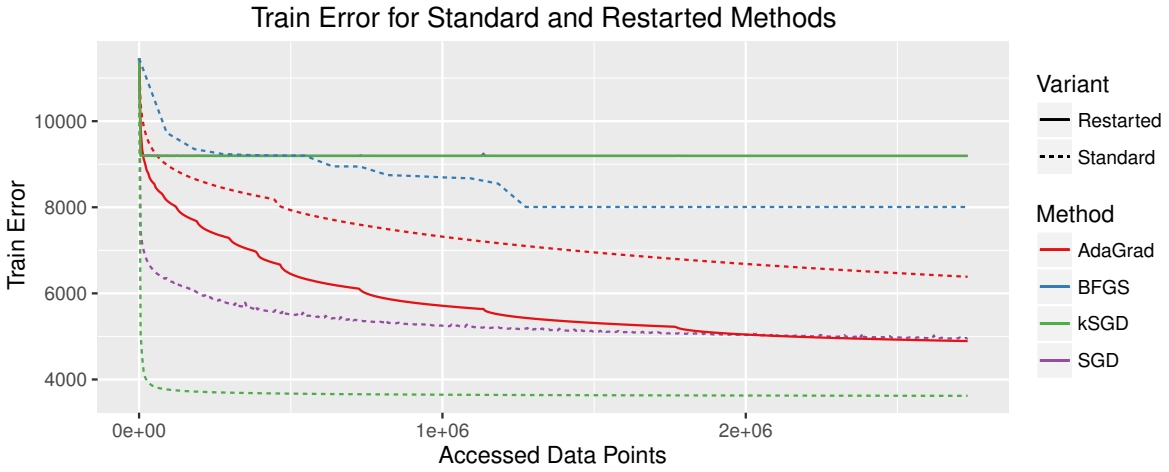


Figure 4.3: Training error for the neural network model on the neutrino data which was learned using the standard SGD, AdaGrad, kSGD and BFGS methods, and restarted SGD, AdaGrad and kSGD methods.

Results and Discussion. In Figs. 4.3 and 4.4, the training and testing error for the different methods are plotted. From these figures, it seems that the restart methods for SGD and

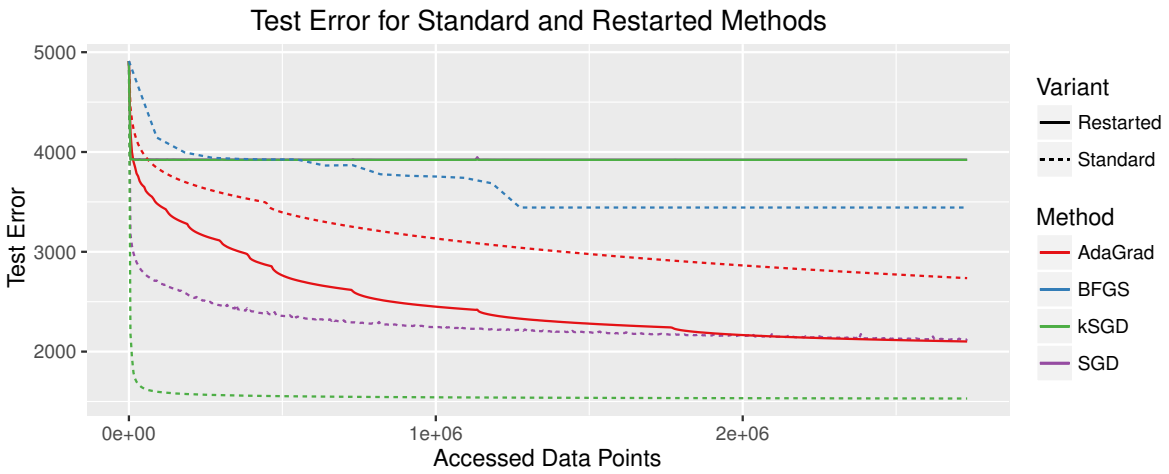


Figure 4.4: Testing error for the neural network model on the neutrino data which was learned using the standard SGD, AdaGrad, kSGD and BFGS methods, and restarted SGD, AdaGrad and kSGD methods.

Table 4.3: Gradient Norms of the Training Error for the Last Iteration.

Comparison of Total Gradients for Neutrino Problem				
Variants	SGD	AdaGrad	kSGD	BFGS
Standard	0.08176	0.01909	0.05436	0.02065
Restart	0.00015	0.01018	8.176e-6	—

kSGD do much worse than their standard counterparts, while the restart method for AdaGrad outperforms its standard counterpart. However, the nonlinear nature of the neural network and the local search behavior of these learning methods is confounding the results. As evidenced by the training error and testing error of BFGS, many of these methods, owing to their random sampling nature, may end up in rather different local minima which have different training and testing errors. Therefore, the correct quantity to consider is the norm of the total gradient of the model on the training error for the final measured parameter. This quantity describes if a stationary point has actually been found by the different learning methods. These values are tabulated in Table 4.3. Thus, in light of Table 4.3, the restarted SGD and kSGD variants perform quite well: indeed, both seem to converge very quickly to their local stationary points in comparison to their standard variants. Similarly,

the restarted AdaGrad also has a better total gradient norm in comparison to the standard AdaGrad method.

To summarize this chapter, we discussed the notions of estimators, computability and a specific type of computable estimators called incremental estimators. We then derived a unified approach based on statistical and numerical optimization principles to generate incremental estimators, and connected the incremental estimators generated by our approach to those currently in the literature. We then briefly discussed a pernicious phenomenon, called stagnation, that is endemic to incremental estimators, and we discussed a strategy for mitigating the impact of this mode of failure, which retains the convergence properties of a consistent incremental estimator and works well experimentally. In the next two chapters, we turn to the detailed analysis of two specific incremental estimators: stochastic gradient descent and the Kalman Filter.

5 | Stochastic Gradient Descent

Stochastic gradient descent (SGD) is a zero-order, linear model incremental estimator. While the methodology can be traced back to Cotes (1722), SGD's rigorous development for one-dimensional parameters came much later with the efforts of Robbins and Monro (1951), who established consistency, and Chung (1954), who established convergence rates of the estimator's moments and asymptotic normality of the estimator. Moreover, SGD has continued to grow in popularity owing to its trivial implementability and its success at nonconvex optimization for machine learning applications (see Keskar et al., 2016).

SGD's success on nonconvex optimization has generated a number of recent works in understanding the properties of SGD for such problems, specifically in how SGD appears to avoid saddle points and how SGD appears to choose flat minimizers. Unfortunately, many of these works modify SGD into an alternative methodology in order to theoretically justify the experimentally observed properties of SGD (Ge et al., 2015; Jin et al., 2017; Mou et al., 2017; Zhang et al., 2017; Kleinberg et al., 2018). Thus, these works have not explained SGD's properties, but rather an alternative method's properties for nonconvex problems.

Another set of works have attempted to study SGD directly. Hardt et al. (2015) considered SGD's properties for nonconvex problems in the context of a notion of algorithmic stability, but this work restricted the set of problems to the point where the results could not be applied to the canonical linear regression problem. Bottou et al. (2016) (Section 4) establishes the consistency of SGD for convex, strongly convex and nonconvex problems under more realistic conditions. While these statements are useful, they do not provide insight into SGD's convergence behavior on nonconvex problems; for example, such results cannot explain SGD's geometric preferences for choosing flatter minimizers—those minimiz-

Disclaimer: Portions of this work are reproduced or adapted from Patel (2017a).

ers at which the Hessian of the objective function has small eigenvalues—in comparison to deterministic gradient descent methods (Keskar et al., 2016).

Here, our goal will be to provide insight into this phenomenon. The widely accepted explanation for SGD's preference for flat minimizers is the *stochastic mechanism*: the stochasticity of SGD will "push" SGD out of sharp basins, whereas the stochasticity of SGD is insufficient at "pushing" SGD out of flat basins (e.g. Keskar et al., 2016). To further illustrate this point, consider SGD on a nonconvex optimization problem whose expected objective function is given by Fig. 5.1. Moreover, suppose the noise process is generated by i.i.d. mean zero, bounded random variables added to the derivative of the expected objective function. Then, it stands to reason that the noise has a lower probability of "pushing" SGD estimates out of the basin on the left (the flat minimizer's basin) in comparison to the minimizer on the right (the sharp minimizer).

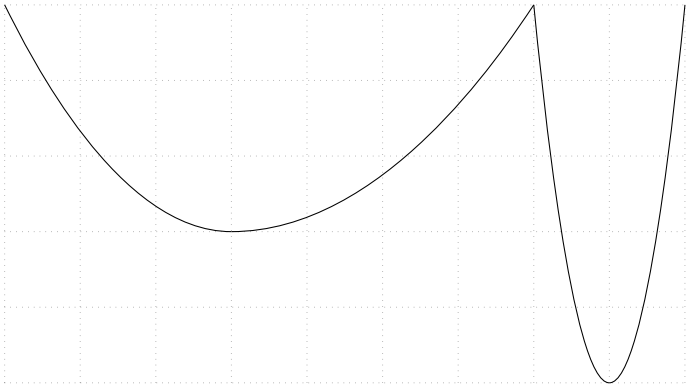


Figure 5.1: A nonconvex objective function with two minimizers. The objective function's Hessian at the minimizer on the left has smaller eigenvalues relative to the objective function's Hessian at the minimizer on the right. The minimizer on the left is flatter than the minimizer on the right. The minimizer on the right is sharper than the minimizer on the left.

While the stochastic mechanism is rather intuitive, it confounds the *size* of the basin about a minimizer and the *sharpness/flatness* of the basin. For example, if the expected objective function of the nonconvex problem is the one shown in Fig. 5.2, then the stochastic mechanism predicts that the SGD estimates will have a lower probability of converging to the flatter minimizer on the left in comparison to the sharper minimizer on the right. However,

in practice, as will show on more complex nonconvex problems, this is not observed: SGD is more likely to converge to the flatter minimizer for generic learning rate choices. Thus, the stochastic mechanism is insufficient in explaining why SGD prefers flatter minimizers over sharper minimizers in general.

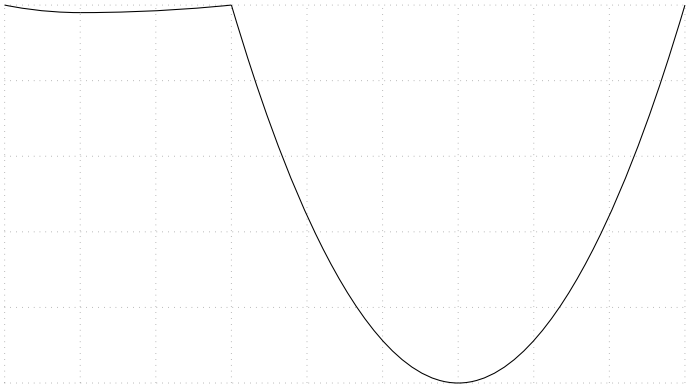


Figure 5.2: A nonconvex objective function with two minimizers. The minimizer on the left is the flatter minimizer, whereas the minimizer on the right is the sharper minimizer.

Here, we propose an alternative mechanism that generalizes the mechanisms of divergence for classical numerical optimizers. In particular, suppose we apply classical gradient descent with a constant step size for a quadratic optimization problem. Then, if the step size is larger than twice the smallest eigenvalue, the gradient descent iterates will diverge. Fig. 5.3 shows the first few iterations for minimizing $f(x) = x^2$ with gradient descent with a constant step size of 1.1. Here, divergence from the minimizer occurs by repeatedly overshooting the minimizer. The resulting behavior can be observed as an *exponential divergence* from the minimizer.

The mechanism for divergence that we develop here is similar to the classical mechanism in that there will be a threshold above which our mechanism will predict that SGD will not converge to this minimizer. To develop this mechanism, in Section 5.1, we begin by analyzing a rather generic linear regression problem, for which we establish the usual sufficient conditions for convergence and we state a necessary condition for convergence.[1]

[1] I have not seen these necessary conditions expressed explicitly before.

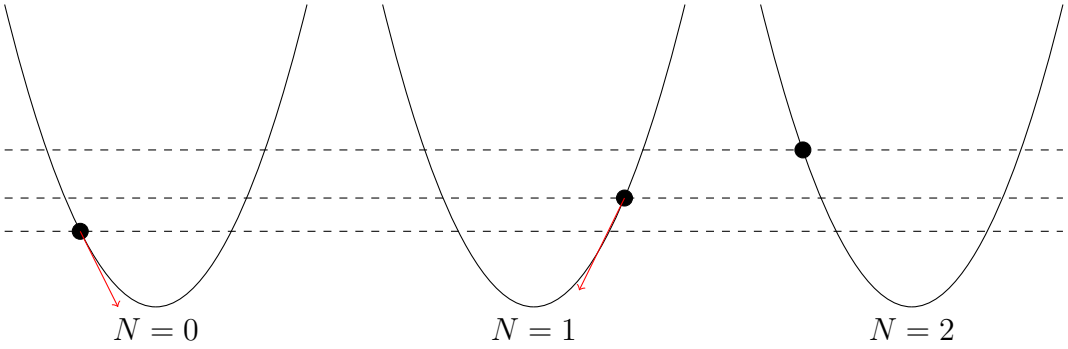


Figure 5.3: The divergence of gradient descent for objective function $f(x) = x^2$ with a step size of 1.1. The iterations are initialized at $x = -1$. N indicates the iteration and the red arrows indicate the direction of the negative gradient.

Moreover, we prove that there is a threshold on the step sizes above which the SGD iterates will exhibit divergence. More importantly, the mechanism predicts a pattern of divergence, namely, exponential divergence, for SGD that we observe for classical gradient descent.

To increase in complexity, in Section 5.2, we study batch SGD on a general quadratic problem. Again, just as for SGD on linear regression, we will derive thresholds for divergence based on the expected local geometry and batch size. Importantly, our bounds include SGD on linear regression as a special case *and* include classical gradient descent as a special case. Moreover, by establishing the result on the general quadratic problem, we will be able to locally model nonconvex problems and apply our analysis to these nonconvex problems.

Accordingly, in Section 5.3, we numerically study nonconvex problems using our theory for the general quadratic problem. In particular, we locally approximate the regions around minimizers by quadratic problems and observe the convergence and divergence behavior of SGD. Based on our theory, we compute the thresholds for the local quadratic approximations, and numerically observe that for learning rates above these thresholds will result in exponential divergence.

5.1 Linear Regression Problem

Owing to its simple structure and widespread use, the linear regression problem is the natural starting point for analyzing an incremental estimator. Here, we will consider a specific choice of the the linear regression problem, which is defined in Problem 5.1

Problem 5.1 (Linear Regression)**.** *Let $Q \in \mathbb{R}^{p\times p}$ such that $Q = Q'$ and $Q \succ 0$. Let $X, X_1, X_2, \ldots \in \mathbb{R}^p$ be independent, $\mathcal{N}(0, Q)$-distributed random variables. Let $\epsilon, \epsilon_1, \epsilon_2, \ldots \in \mathbb{R}$ be independent, identically distributed, mean-zero random variables with a finite, non-zero second-moment. Let $\theta^* \in \mathbb{R}^p$, and define*

$$Z = X'\theta^* + \epsilon \text{ and } Z_j = X_j'\theta^* + \epsilon_j, j \in \mathbb{N}. \tag{5.1}$$

Let $Y = (Z, X)$ and $Y_j = (Z_j, X_J)$ for $j \in \mathbb{N}$. The linear regression problem is to determine θ^* from $\{Y_j : j \in \mathbb{N}\}$.

Remark 5.1. *The specialization of X to a normally distributed random variable is exploited in two ways. First, it ensures that the components of X are independent under some rotation. Second, we have a closed form for the fourth order moment. We can remove the requirement of normality and simply require that these two conditions hold.*

Now, we will define SGD for our linear regression problem. SGD is defined relative to an estimating function m. While there are many choices for such functions (see Bean et al., 2013), we will consider the standard ordinary least squares estimating equation given by

$$m(y, \theta) = \frac{1}{2}(z - x'\theta)^2, \text{ where } y = (z, x) \in \mathbb{R} \times \mathbb{R}^p. \tag{5.2}$$

Accordingly, the expected estimating equation, $R(\theta)$, is

$$R(\theta) = \mathbb{E}\left[m(Y, \theta)\right] = \frac{1}{2}\mathbb{V}\left[\epsilon\right] + \frac{1}{2}(\theta - \theta^*)'Q(\theta - \theta^*). \tag{5.3}$$

Moreover, for the estimating equation and an arbitrary $\theta_0 \in \mathbb{R}^p$, SGD is given by

$$\theta_N = \theta_{N-1} + C_N X_N (Z_N - X_N' \theta_{N-1}), \tag{5.4}$$

where $N \in \mathbb{N}$; and C_N is a nonnegative scalar.

We now show that SGD results in a consistent estimator in the sense that $\theta_N \to \theta^*$ in some sense. Note, here we can either show that $\theta_N \to \theta^*$ or, equivalently, $R(\theta_N) - R(\theta^*) \to 0$. In general, we will consider the latter case as it is also a good metric when $R(\theta)$ is not assumed to be strongly convex. Moreover, note that the sufficient conditions for consistency are rather well-known, but the necessary condition has not been stated to our knowledge. The proofs of the main results can be found at the end of this section.

Theorem 5.1. *Suppose SGD, (5.4), is applied to solve Problem 5.1.*

1. $\mathbb{E}\left[R(\theta_N) - R(\theta^*)\right] \to 0$ *if and only if*

$$\sum_{k=1}^{\infty} C_k = \infty \text{ and } \lim_{k \to \infty} C_k = 0. \tag{5.5}$$

2. $R(\theta_N) - R(\theta^*) \to 0$ *in probability if*

$$\sum_{k=1}^{\infty} C_k = \infty \text{ and } \lim_{k \to \infty} C_k = 0. \tag{5.6}$$

3. $R(\theta_N) - R(\theta^*) \to 0$ *with probability one if*

$$\sum_{k=1}^{\infty} C_k = \infty \text{ and } \sum_{k=1}^{\infty} C_k^2 < \infty. \tag{5.7}$$

While the conditions on the learning rates stated in Theorem 5.1 will theoretically guarantee convergence, they do not guarantee that the gap between the estimator and the true parameter will decay monotonically. In fact, we can state the following result.

Theorem 5.2. *Suppose SGD, (5.4), is applied to solve Problem 5.1 with $p > 2$. If C_N is larger than*

$$\frac{2\lambda_{\max}(Q)}{2\lambda_{\max}(Q)^2 + \mathbf{tr}\left[Q^2\right]}, \tag{5.8}$$

where $\lambda_{\max}(Q)$ indicates the largest eigenvalue of Q, then

$$\mathbb{E}\left[R(\theta_N) - R(\theta^*)\right] > \rho_N \mathbb{E}\left[R(\theta_{N-1}) - R(\theta^*)\right] + \frac{1}{2} C_N^2 \mathbf{tr}\left[Q^2\right] \mathbb{V}\left[\epsilon\right], \tag{5.9}$$

where $\rho_N > 1$ grows quadratically with C_N larger than the threshold.

Thus, while Theorem 5.1 guarantees that a given sequence of step sizes ensures consistency of the estimator, Theorem 5.2 shows that the estimators may not always be progressing towards the the true parameter. More importantly, Theorem 5.2 is similar in nature to classical results for the divergence of gradient in that (1), based on the geometry of the problem, it supplies a threshold for the step sizes above which we predict divergence of the iterates, and (2) it concludes that the rate of divergence is exponential. In the next section, we will show that, indeed, this result for SGD and the classical result for gradient descent are two extremes of the spectrum.

However, before moving to this result, we would like to briefly (numerically) study the practical consequences of exponential divergence. Specifically, Theorem 5.2 suggests that the exponential divergence can be catastrophic. That is, even if we choose a sequence of learning rates that satisfy the sufficient conditions for convergence in Theorem 5.1, it is possible for the intermediate step sizes to diverge the SGD estimates to numerical infinity. Unfortunately, as the next experiment show, this intermediate step is not just a theoretical possibility, it will occur even for generic choices of learning rates.

Experimental Setup. We generate a sequence of random manifestations of Problem 5.1 with all combinations of the following variables.

1. (Dimension) The dimension of the parameter θ is either 10, 100 or $1,000$.

2. (Conditioning) The condition number of Q is set to 10^3 or 10^6.

3. (Max Eigenvalue) The maximum eigenvalue of Q is set to 10^k for $k = -3, \ldots, 4$.

Note that by fixing the maximum eigenvalue and condition number of Q, the smallest eigenvalue is determined. The remaining eigenvalues of Q were selected from a uniform random variable between these two extremes.

For each combination, SGD was run with one of five possible learning rates $C_N = N^{-\gamma}$ where $\gamma = 0.2, 0.4, 0.6, 0.8, 1.0$, all of which satisfy the necessary and sufficient condition for convergence in expectation according to Theorem 5.1. This results in two-hundred and forty independent runs of SGD.

For each run, SGD was allowed to run for 10^6 iterations. The following variables were collected:

1. Initial Parameter Optimality Gap. The euclidean distance between the randomly generated initial estimators, θ_0, and the true parameter, θ^*.

2. Maximum Parameter Optimality Gap. The maximum over all euclidean distances between the estimates generated by SGD and the true parameter.

3. Divergence Period. The number of iterations needed to achieve the Maximum Parameter Optimality Gap.

4. Recovery Period. The minimum between the maximum allowed iterations and the number of iterations after the divergence period for the estimates generated by SGD to drop below the initial optimality gap, and remain below this value for 10^4 iterations. (This 10^4 confirmation period is not included in the recovery period.)

Once the data was collected, each run was categorized into one of three categories:

1. Diverged. Within the first 10^6 iterations, the SGD parameter optimality gap was greater than 10^{300}.

2. Converged. Within the first 10^6 iterations, the SGD parameter optimality gap *never* exceeded the initial parameter optimality gap.

3. Recovered. Within the first 10^6 iterations, the SGD parameter optimality gap exceeded the initial parameter optimality gap, and eventually began to decrease.

Results and Discussion. Table 5.1 summarizes the number of runs which diverged, converged and recovered relative to the maximum eigenvalue. From the summary in Table 5.1, the maximum eigenvalue captures the number of runs which diverge, converge and recover. The primary message of Table 5.1 is the following: although we may choose learning rates that ensure convergence in theory, we cannot choose them blindly as they may lead to estimators that exhibit numerical divergence.

Table 5.1: SGD Runs that Diverge, Converge and Recover by the Largest Eigenvalue.

Max Eigenvalue	Diverged	Recovered	Converged
0.001	0	10	20
0.01	0	15	15
0.1	2	18	10
1	6	24	0
10	16	14	0
100	24	6	0
1000	30	0	0
10000	30	0	0

As it is also insightful, we now consider the ratio between the recovery period and the divergence period for the runs which recovered. Figs. 5.4 to 5.8 compare the ratio of the recovery period to the divergence period (log-scale) against the problem dimension, problem conditioning, maximum eigenvalue, learning rate exponent, and the maximum parameter optimality gap for the runs which had a recovery period. Therefore, as seen from Table 5.1 and Fig. 5.4, there is no data for when the maximum eigenvalues are 10^3 and 10^4 because all such runs diverged.

We underscore several observations. First, the ratio between the recovery period and the divergence period does not seem to depend on the problem dimension (Fig. 5.5), the

problem conditioning (Fig. 5.6) nor the learning rate (Fig. 5.7). Second, the ratio between the recovery period and divergence period is favorable (i.e., about or less than one) for problems with smaller maximal eigenvalues (Fig. 5.4). Third, if ratio grows quite rapidly with a small increase in the maximum optimality gap (Fig. 5.8).

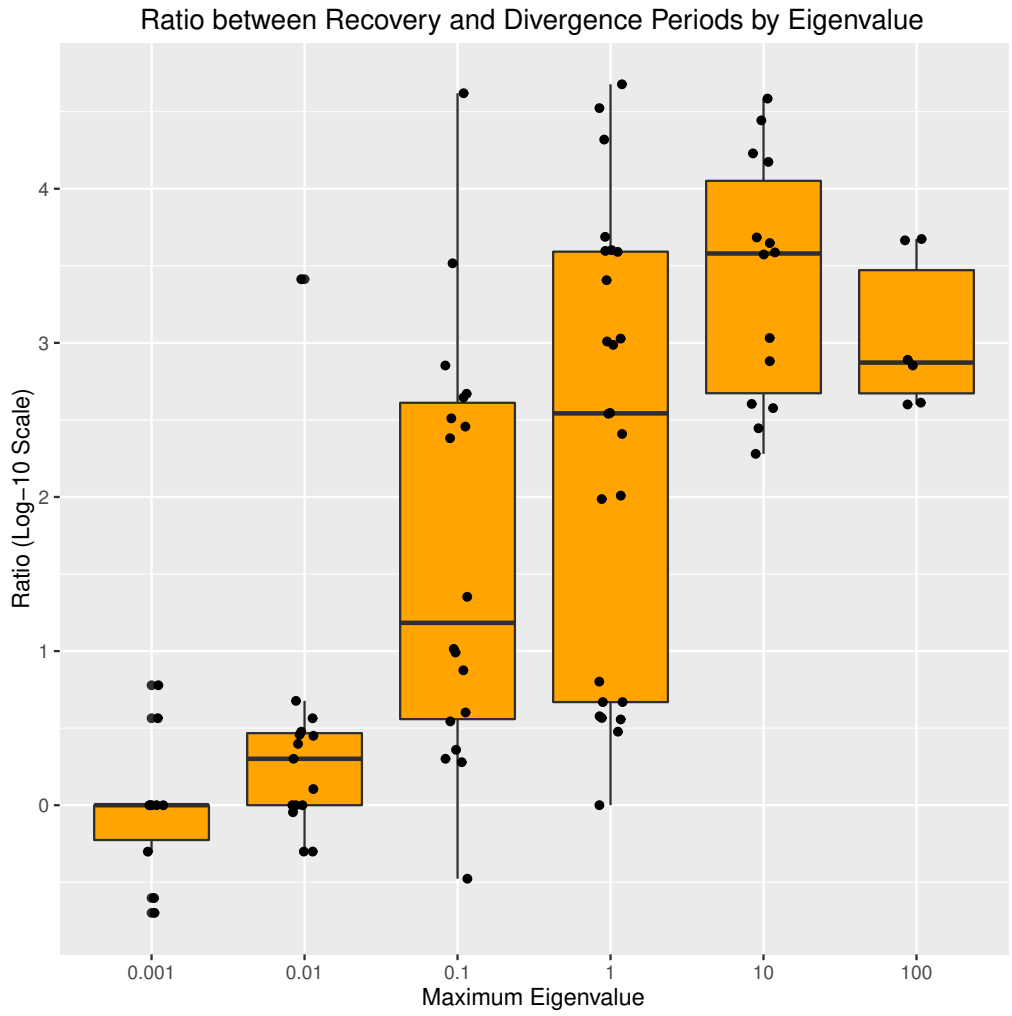


Figure 5.4: The behavior of the recovery-divergence ratio by the largest eigenvalue of the problem.

In summary, Table 5.1 and Fig. 5.4 show that the divergence and recovery of SGD will depend on the maximum eigenvalue of the Hessian, which supports Theorem 5.2. Moreover, by Fig. 5.8, even small amounts of divergence become very costly as the recovery period grows rapidly with respect to small increases in maximum optimality gap. Therefore, if a

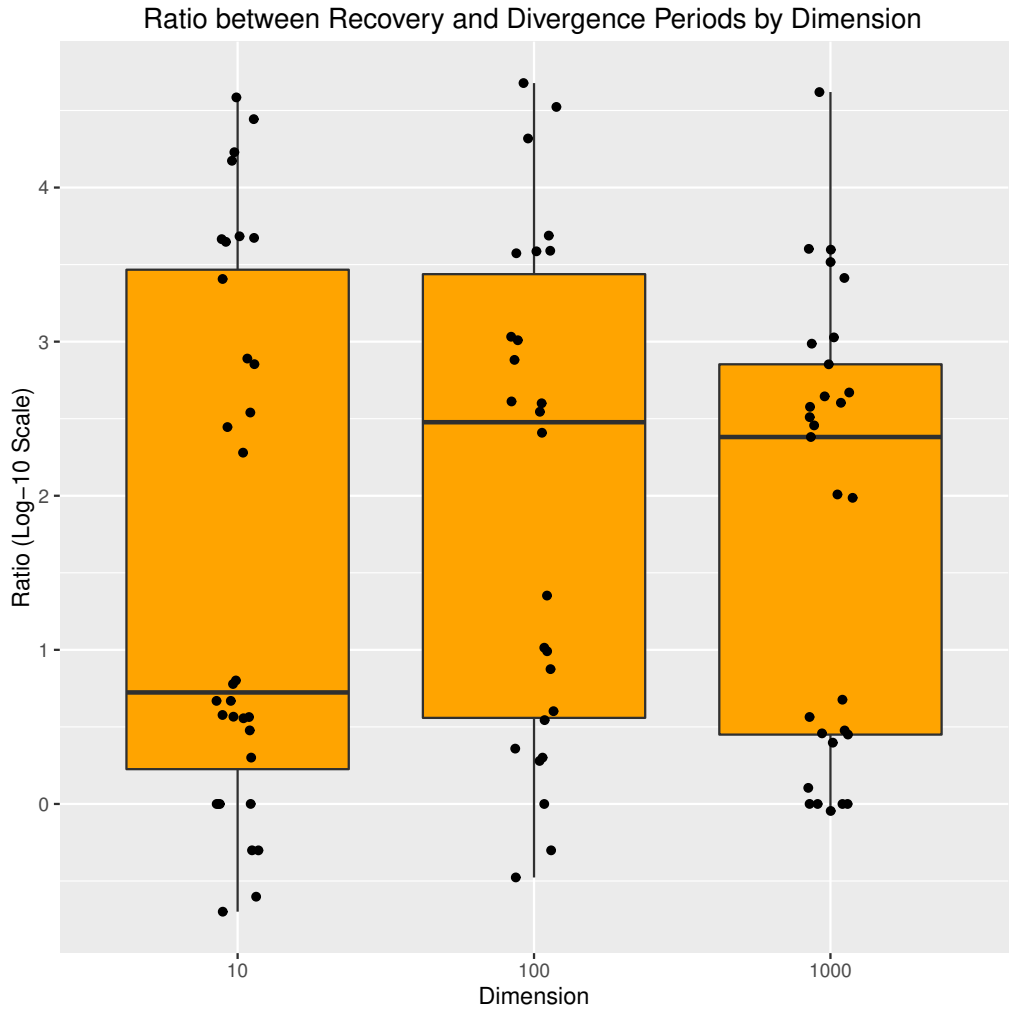


Figure 5.5: The behavior of the recovery-divergence ratio by the dimension of the problem.

minimizer has a Hessian with a large eigenvalue, then we must carefully choose the learning rate of SGD to remain below (5.8) in order to avoid divergence and reduce wasted recovery iterations. Unfortunately, as we already saw, scaling down the learning rate to remain below (5.8) can also lead to stagnation.

The preceding theory and experiments demonstrate that the convergence of the SGD estimates for the linear regression problem is highly sensitive to the maximum eigenvalue of the problem and the choice of learning rates. In fact, if the learning rates are chosen poorly (see Theorem 5.2), we can have long recovery periods or even catastrophic divergence of the

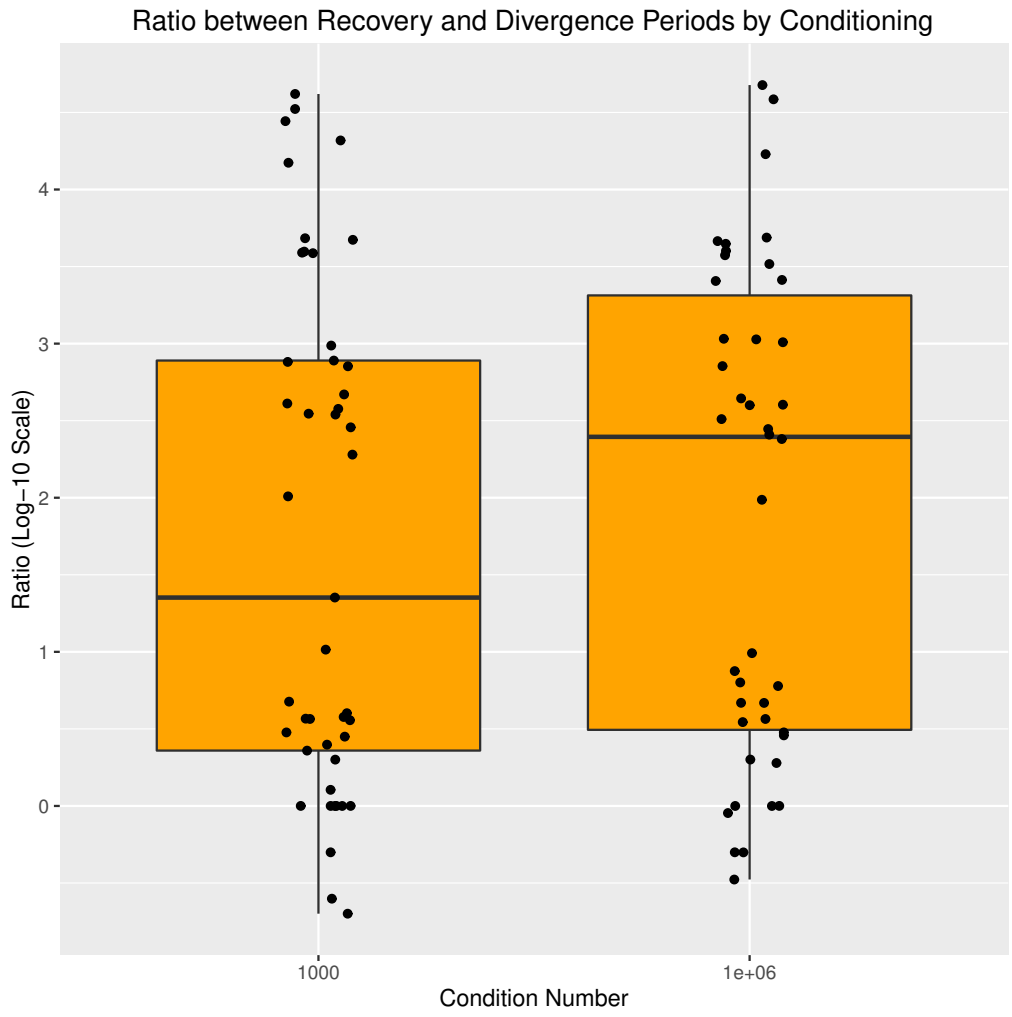


Figure 5.6: The behavior of the recovery-divergence ratio by the condition number of the problem.

estimators. This behavior will also hold when we study more generic quadratic problems in Section 5.2. However, when the problem is nonconvex, divergence from a particular minimizer will *not* automatically be terrible: it might allow the estimators to explore the landscape to find a minimizer to which it can converge. We finish this section by proving Theorems 5.1 and 5.2.

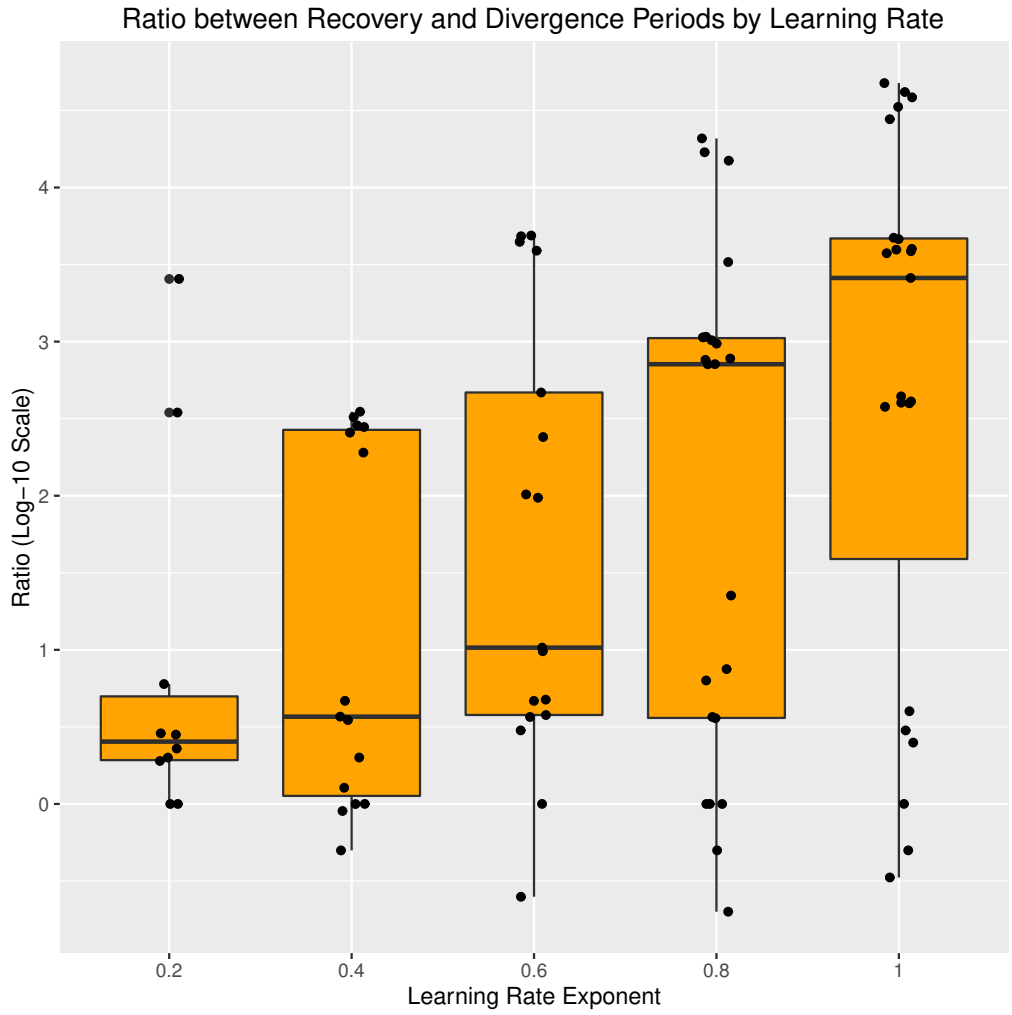


Figure 5.7: The behavior of the recovery-divergence ratio by the learning rate of SGD.

5.1.1 A Recursion Lemma

Our first goal is to relate the error between $R(\theta_{N+1}) - R(\theta^*)$ and $R(\theta_N) - R(\theta^*)$. It follows by a direct calculation, and using the subsequent lemma.

Lemma 5.1. *Suppose SGD is used to solve Problem 5.1. Let $U\Lambda U'$ be the spectral decomposition of Q, and let M be a matrix such that $U'MU$ is a diagonal matrix (note, M and Q*

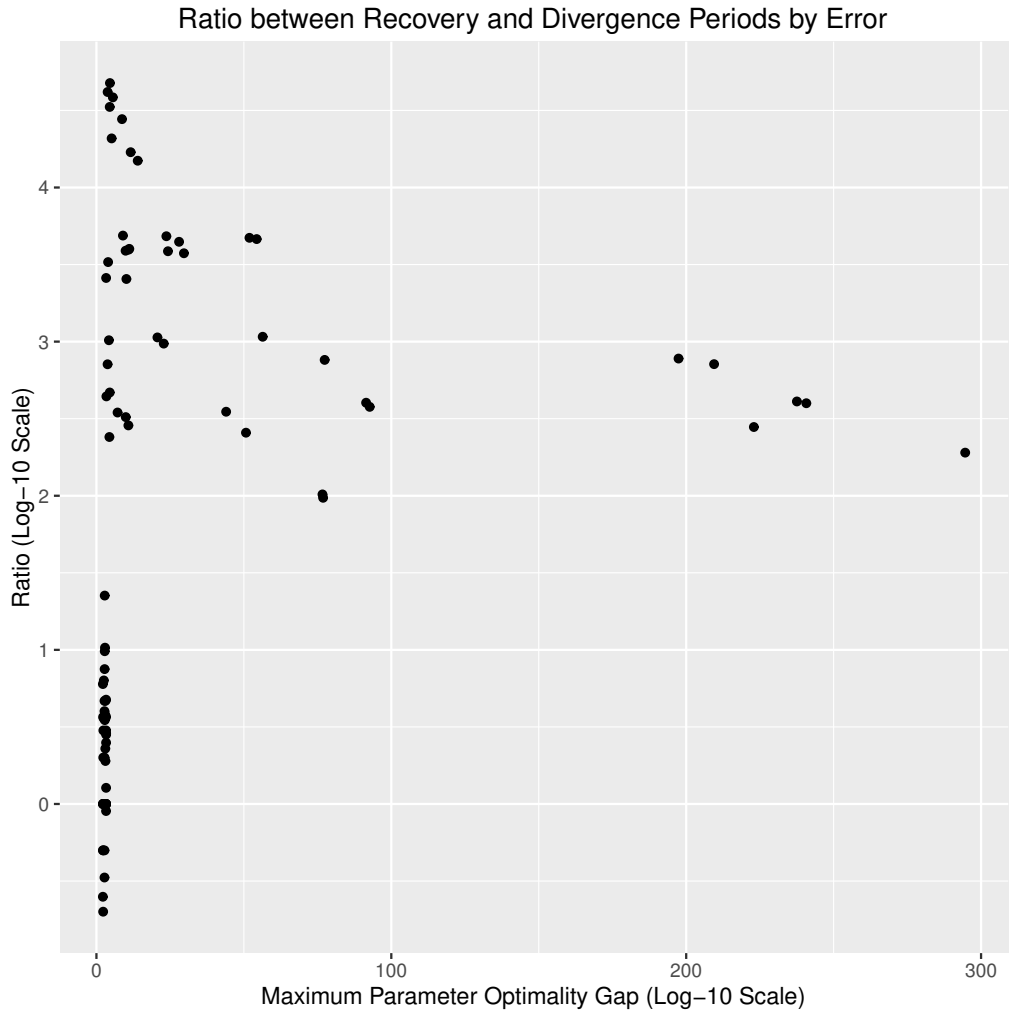


Figure 5.8: The behavior of the recovery-divergence ratio by the maximum parameter optimality gap achieved.

commute). Then

$$\begin{aligned}\mathbb{E}\left[(\theta_{N+1}-\theta^*)'M(\theta_{N+1}-\theta^*)|\,\theta_N\right] &= (\theta_N-\theta^*)'M(\theta_N-\theta^*)\\ &- 2C_{N+1}(\theta_N-\theta^*)'MQ(\theta_N-\theta^*) + 2C_{N+1}^2(\theta_N-\theta^*)'MQ^2(\theta_N-\theta^*)\\ &+ C_{N+1}^2\mathbf{tr}\left[MQ\right](\theta_N-\theta^*)'Q(\theta_N-\theta^*) + C_{N+1}^2\mathbf{tr}\left[MQ\right]\mathbb{V}\left[\epsilon\right]\end{aligned} \tag{5.10}$$

Lemma 5.2. *Suppose $X \sim \mathcal{N}(0,Q)$ with Q positive definite and symmetric, and let $U\Lambda U'$ be the spectral decomposition of Q. Let $M = ULU'$ where L is an arbitrary diagonal matrix.*

Then,

$$\mathbb{E}\left[XX'MXX'\right] = 2MQ^2 + \mathbf{tr}\left[MQ\right]Q. \tag{5.11}$$

Proof. Let the components of Λ be denoted $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, and the components of L be denoted $l_1, l_2, \ldots, l_p$. Let the components of $H := U'X$ be denoted $h_1, h_2, \ldots, h_p$. Therefore:

1. $X = \sum_{i=1}^{p} h_i u_i$, where u_i are the columns of U.
2. $h_i \sim \mathcal{N}(0, \lambda_i)$, for $i = 1, \ldots, p$.
3. $\{h_1, h_2, \ldots, h_p\}$ are independent random variables.

Therefore,

$$\begin{aligned} XX'MXX' &= \sum_{e=1}^{p}\sum_{f=1}^{p}\sum_{g=1}^{p}\sum_{i=1}^{p} h_e h_f h_g h_i u_e u_f' U L U' u_g u_i' \\ &= \sum_{e=1}^{p}\sum_{f=1}^{p}\sum_{g=1}^{p}\sum_{i=1}^{p} h_e h_f h_g h_i l_g u_e u_f' u_g u_i' \\ &= \sum_{e=1}^{p}\sum_{f=1}^{p}\sum_{i=1}^{p} h_e h_f^2 h_i l_f u_e u_i'. \end{aligned} \tag{5.12}$$

Taking expectations,

$$\mathbb{E}\left[XX'MXX'\right] = \sum_{e=1}^{p}\sum_{f=1}^{p}\sum_{i=1}^{p} \mathbb{E}\left[h_e h_f^2 h_i\right] l_f u_e u_i'. \tag{5.13}$$

We now specifically look at $\mathbb{E}\left[h_e h_i h_f^2\right]$. If $e \neq i$, then, by independence, the expectation is 0. Hence, we have

$$\mathbb{E}\left[XX'MXX'\right] = \sum_{i=1}^{p}\sum_{f=1}^{p} \mathbb{E}\left[h_i^2 h_f^2\right] l_f u_i u_i'. \tag{5.14}$$

In this case, using the properties of the normal distribution, $\mathbb{E}\left[h_i^2 h_f^2\right] = 3\lambda_i^2 \mathbf{1}\left[i = f\right] +$

$\lambda_i\lambda_f \mathbf{1}\left[i \neq f\right] = 2\lambda_i^2 \mathbf{1}\left[i = f\right] + \lambda_i\lambda_f$. Therefore,

$$\begin{aligned}\mathbb{E}\left[XX'MXX'\right] &= \sum_{i=1}^{p} 2l_i\lambda_i^2 u_i u_i' + \sum_{i=1}^{p}\sum_{f=1}^{p} l_f\lambda_f\lambda_i u_i u_i \\ &= 2UL\Lambda^2U' + U\Lambda U'\mathbf{tr}\left[L\Lambda\right] \\ &= 2MQ^2 + \mathbf{tr}\left[MQ\right]Q.\end{aligned} \tag{5.15}$$

■

Using Lemma 5.1, we can derive the following upper and lower bounds on the recursion.

Lemma 5.3. *Suppose SGD is used to solve Problem 5.1. Let $U\Lambda U'$ be the spectral decomposition of Q and let M be a matrix such that $U'MU = L$ is a positive definite diagonal matrix. Let $s^+ = \max_{i=1,\ldots,d}\{\lambda_i l_i^{-1}\}$, $s^- = \min_{i=1,\ldots,d}\{\lambda_i l_i^{-1}\}$, and*

$$e_k = (\theta_k - \theta^*)'M(\theta_k - \theta^*). \tag{5.16}$$

For an upper bound,

1. *When $C_{k+1}(\lambda_1 + \lambda_p) > 1$,*

$$\mathbb{E}\left[e_{k+1}|\,\theta_k\right] \leq e_k\left(1 - 2C_{k+1}\lambda_1 + 2C_{k+1}^2\lambda_1^2 + C_{k+1}^2\mathbf{tr}\left[MQ\right]s^+\right) + C_{k+1}^2\mathbf{tr}\left[MQ\right]\mathbb{V}\left[\epsilon\right]. \tag{5.17}$$

2. *When $C_{k+1}(\lambda_1 + \lambda_p) \leq 1$,*

$$\mathbb{E}\left[e_{k+1}|\,\theta_k\right] \leq e_k\left(1 - 2C_{k+1}\lambda_p + 2C_{k+1}^2\lambda_p^2 + C_{k+1}^2\mathbf{tr}\left[MQ\right]s^+\right) + C_{k+1}^2\mathbf{tr}\left[MQ\right]\mathbb{V}\left[\epsilon\right]. \tag{5.18}$$

For a lower bound,

1. *When $C_{k+1}(\lambda_p + \lambda_{p-1}) > 1$,*

$$e_k\left(1 - 2C_{k+1}\lambda_p + 2C_{k+1}^2\lambda_p^2 + C_{k+1}^2\mathbf{tr}\left[MQ\right]s^-\right) + C_{k+1}^2\mathbf{tr}\left[MQ\right]\mathbb{V}\left[\epsilon\right] \leq \mathbb{E}\left[e_{k+1}|\,\theta_k\right]. \tag{5.19}$$

2. *For* $j = p, p-1, \ldots, 3$, *when* $C_{k+1}(\lambda_{j-1} + \lambda_{j-2}) \geq 1 > C_{k+1}(\lambda_j + \lambda_{j-1})$,

$$e_k \left(1 - 2C_{k+1}\lambda_{j-1} + 2C_{k+1}^2\lambda_{j-1}^2 + C_{k+1}^2 \mathbf{tr}\left[MQ\right] s^-\right) + C_{k+1}^2 \mathbf{tr}\left[MQ\right] \mathbb{V}\left[\epsilon\right] \leq \mathbb{E}\left[e_{k+1} \middle| \theta_k\right]. \tag{5.20}$$

3. *When* $1 \geq C_{k+1}(\lambda_2 + \lambda_1)$,

$$e_k \left(1 - 2C_{k+1}\lambda_1 + 2C_{k+1}^2\lambda_1^2 + C_{k+1}^2 \mathbf{tr}\left[MQ\right] s^-\right) + C_{k+1}^2 \mathbf{tr}\left[MQ\right] \mathbb{V}\left[\epsilon\right] \leq \mathbb{E}\left[e_{k+1} \middle| \theta_k\right]. \tag{5.21}$$

Proof. From Lemma 5.1,

$$\begin{aligned}\mathbb{E}\left[(\theta_{k+1} - \theta^*)' M (\theta_{k+1} - \theta^*) \middle| \theta_k\right] &= C_{k+1}^2 \mathbf{tr}\left[MQ\right] \mathbb{V}\left[\epsilon\right] \\ &+ (\theta_k - \theta^*)' M (\theta_k - \theta^*) \left(1 - 2C_{k+1} \sum_{i=1}^p w_i \lambda_i + 2C_{k+1}^2 \sum_{i=1}^p w_i \lambda_i^2 + C_{k+1}^2 \mathbf{tr}\left[MQ\right] \sum_{i=1}^p w_i \frac{\lambda_i}{l_i}\right),\end{aligned} \tag{5.22}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of Q, $l_i > 0$ are the diagonal elements of L, and

$$w_i = \frac{l_i \left[u_i'(\theta_k - \theta^*)\right]^2}{(\theta_k - \theta^*)' M (\theta_k - \theta^*)}, \tag{5.23}$$

where u_i is the i^{th} column of U. Note, $\sum_{i=1}^d w_i = 1$. We can finish proving the lemma by finding upper and lower bounds on the quantity

$$\sum_{i=1}^p w_i \left(C_{k+1}\lambda_i^2 - \lambda_i\right). \tag{5.24}$$

Since this is a convex combination, it will be lower bounded by the minimal element in the combination, and upper bounded by the maximal element. By differentiating, we see that the expression $C_{k+1}\lambda^2 - \lambda$ is minimized when $\lambda = (2C_{k+1})^{-1}$. Therefore, for the upper bound, we have four cases to consider in the range $\lambda \in [\lambda_p, \lambda_1]$:

1. When $1 \leq 2C_{k+1}\lambda_p$, then (5.24) is maximized by λ_1 in the range.
2. When $2C_{k+1}\lambda_p < 1 < 2C_{k+1}\lambda_1$, and $C_{k+1}(\lambda_1 + \lambda_p) > 1$, then (5.24) is maximized by

λ_1 in the range.

3. When $2C_{k+1}\lambda_p < 1 < 2C_{k+1}\lambda_1$, and $C_{k+1}(\lambda_1 + \lambda_p) \leq 1$, then (5.24) is maximized by λ_p in the range.

4. When $2C_{k+1}\lambda_1 \leq 1$, then (5.24) is maximized by λ_p in the range.

For the lower bound, we have many more cases to consider in the range $\lambda = \lambda_p, \ldots, \lambda_1$:

1. When $1 \leq 2C_{k+1}\lambda_p$, then (5.24) is minimized by λ_p in the range.

2. When $2C_{k+1}\lambda_j \leq 1 \leq 2C_{k+1}\lambda_{j-1}$ and $C_{k+1}(\lambda_j + \lambda_{j-1}) > 1$, then (5.24) is minimized by λ_j in the range, for $j = p, p-1, \ldots, 2$.

3. When $2C_{k+1}\lambda_j \leq 1 \leq 2C_{k+1}\lambda_{j-1}$ and $C_{k+1}(\lambda_j + \lambda_{j-1}) \leq 1$, then (5.24) is minimized by λ_j in the range, for $j = p, p-1, \ldots, 2$.

4. When $2C_{k+1}\lambda_1 \leq 1$, then (5.24) is minimized by λ_1 in the range.

■

5.1.2 Proofs of the Main Results

Proof of Theorem 5.1. We deal with the sufficient conditions first. Using Lemma 5.3 with $M = Q$, and letting $K \in \mathbb{N}$ such that for all $k \geq K$

$$C_k \leq \min\left\{\frac{1}{\lambda_1 + \lambda_p}, \frac{\lambda_p}{2\lambda_p^2 + \mathbf{tr}\left[Q^2\right]}\right\}, \tag{5.25}$$

then, for $k \geq K$,

$$\mathbb{E}\left[R(\theta_{k+1}) - R(\theta^*)\right] \leq (1 - \lambda_p C_{k+1})\mathbb{E}\left[R(\theta_k) - R(\theta^*)\right] + \frac{1}{2}C_{k+1}^2 \mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right]. \tag{5.26}$$

Letting $\gamma_k = \lambda_p C_k$ and using Lemma 1.5.1 in Bertsekas (1999), for a learning rate such that $\sum C_k = \infty$ and $C_k \to 0$,

$$\mathbb{E}\left[R(\theta_k) - R(\theta^*)\right] \to 0 \tag{5.27}$$

as $k \to \infty$. Accordingly, convergence in probability follows by an application of Markov's Inequality. For learning rates that also satisfy $\sum C_k^2 < \infty$, we apply Theorem 1 in Robbins and Siegmund (1985) to show that the sequence converges with probability one.

We now consider the necessary condition. Suppose that

$$\mathbb{E}\left[R(\theta_{k+1}) - R(\theta^*)\right] \to 0, \tag{5.28}$$

as $k \to \infty$. From Lemma 5.3,

$$\begin{aligned}
\mathbb{E}\left[R(\theta_{k+1}) - R(\theta^*)\right] &\geq \mathbb{E}\left[R(\theta_k) - R(\theta^*)\right]\left(1 - 2C_{k+1}\lambda_1 + 2C_{k+1}^2\lambda_1^2 + C_{k+1}^2 \mathbf{tr}\left[Q^2\right]\right) \\
&\quad + \frac{1}{2}C_{k+1}^2 \mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right] \\
&\geq \mathbb{E}\left[R(\theta_k) - R(\theta^*)\right]\left(1 - 2C_{k+1}\lambda_1 + C_{k+1}^2\lambda_1^2\right) \\
&\quad + \frac{1}{2}C_{k+1}^2 \mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right] \\
&\geq \mathbb{E}\left[R(\theta_k) - R(\theta^*)\right]\left(1 - \gamma_k + \frac{1}{2}\gamma_k^2\right) + \frac{1}{8\lambda_1^2}\gamma_k^2 \mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right]
\end{aligned} \tag{5.29}$$

where $\gamma_k = 2C_{k+1}\lambda_1$. Iterating, we have

$$\begin{aligned}
\mathbb{E}\left[R(\theta_{k+1}) - R(\theta^*)\right] \geq \left[R(\theta_0) - R(\theta^*)\right]\prod_{j=0}^{k}\left(1 - \gamma_j + \frac{1}{2}\gamma_j^2\right) \\
+ \frac{\mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right]}{8\lambda_1^2}\left[\gamma_k^2 + \sum_{j=0}^{k-1}\gamma_j^2\prod_{l=j+1}^{k}\left(1 - \gamma_l + \frac{1}{2}\gamma_l^2\right)\right]
\end{aligned} \tag{5.30}$$

Since the expected optimality gap must converge to zero, this implies that $\gamma_k \to 0$ as $k \to \infty$. Now, we must show that $\sum_{j=0}^{\infty}\gamma_j = \infty$. Since $\gamma_k \to 0$, there is a $K \in \mathbb{N}$ such that for $k \geq K$, $\gamma_k < 1$. Again, since the expected optimality gap converges to zero and γ_j are positive for all j, this implies for every $\epsilon > 0$ there is an $M > K$ such that for $m \geq M$

$$\epsilon \geq \prod_{l=K}^{m}\left(1 - \gamma_l + \frac{1}{2}\gamma_l^2\right) \geq \prod_{l=K}^{m}\exp(-\gamma_l) \geq \exp(-\sum_{l=K}^{m}\gamma_l). \tag{5.31}$$

Hence, since $\epsilon > 0$ is arbitrary and γ_j are nonnegative, we conclude that $\sum_{j=0}^{\infty} \gamma_j = \infty$. The conclusion follows by recalling that $\gamma_k = 2C_{k+1}\lambda_1$. ■

Proof of Theorem 5.2. Under the conditions of Lemma 5.3, for any C, $0 \leq 1 - C(2\lambda_j) + C^2(2\lambda_j^2 + \mathbf{tr}\left[MQ\right] s^-)$. This follows because the discriminant of the quadratic is nonpositive:

$$0 \geq 4\lambda_j^2 - 4(2\lambda_j^2 + \mathbf{tr}\left[MQ\right] s^-) \geq -\lambda_j^2 - \mathbf{tr}\left[MQ\right] s^-, \tag{5.32}$$

which is always true since M is positive definite.

Now, Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ denote the eigenvalues of Q. We will need the following calculations in the cases examined below. Since $p > 2$, for $j = p, \ldots, 2$, $2\lambda_j\lambda_{j-1} \leq \lambda_j^2 + \lambda_{j-1}^2$. In turn, $2\lambda_j\lambda_{j-1} + 2\lambda_j^2 < \mathbf{tr}\left[Q^2\right] + 2\lambda_j^2$. Finally, it follows that

$$\frac{2\lambda_j}{2\lambda_j^2 + \mathbf{tr}\left[Q^2\right]} < \frac{1}{\lambda_j + \lambda_{j-1}}. \tag{5.33}$$

From Lemma 5.3, we have several cases which we must consider to show the result.

1. (Case 1) Suppose $(\lambda_p + \lambda_{p-1})^{-1} < C$. Then, to show the result, it is sufficient to have

$$1 < 1 - C(2\lambda_p) + C^2(2\lambda_p^2 + \mathbf{tr}\left[Q^2\right]). \tag{5.34}$$

Rearranging, it is enough to have $\frac{2\lambda_p}{2\lambda_p^2 + \mathbf{tr}[Q^2]} < C$. By (5.33), this is always true for the interval under consideration. Therefore, applying Lemma 5.3,

$$\mathbb{E}\left[e_{k+1} | \theta_k\right] \geq e_k(1 - C_{k+1}(2\lambda_p) + C_{k+1}^2(2\lambda_p^2 + \mathbf{tr}\left[Q^2\right])) + C_{k+1}^2 \mathbf{tr}\left[Q^2\right] \mathbb{V}\left[\epsilon\right], \tag{5.35}$$

from which the result follows for this case.

2. (Case 2) For $j = p, p-1, \ldots, 3$, suppose $(\lambda_{j-1} + \lambda_{j-2})^{-1} \leq C < (\lambda_j + \lambda_{j-1})^{-1}$. Then,

to show the result, we must have

$$1 < 1 - C(2\lambda_{j-1}) + C^2(2\lambda_{j-1}^2 + \mathbf{tr}\left[Q^2\right]). \tag{5.36}$$

Rearranging, we must have $\frac{2\lambda_{j-1}}{2\lambda_{j-1}^2+\mathbf{tr}[Q^2]} < C$. By (5.33), this is always true for the interval under consideration. Therefore, applying Lemma 5.3,

$$\mathbb{E}\left[e_{k+1}|\,\theta_k\right] \geq e_k(1 - C_{k+1}(2\lambda_{j-1}) + C_{k+1}^2(2\lambda_{j-1}^2 + \mathbf{tr}\left[Q^2\right])) + C_{k+1}^2\mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right], \tag{5.37}$$

from which the result follows for this case.

3. (Case 3) Finally, suppose $C < (\lambda_1 + \lambda_2)^{-1}$. Then, we must have $\frac{2\lambda_1}{2\lambda_1^2+\mathbf{tr}[Q^2]} < C$. In fact, when this inequality holds,

$$\mathbb{E}\left[e_{k+1}|\,\theta_k\right] \geq e_k(1 - C_{k+1}(2\lambda_1) + C_{k+1}^2(2\lambda_1^2 + \mathbf{tr}\left[Q^2\right])) + C_{k+1}^2\mathbf{tr}\left[Q^2\right]\mathbb{V}\left[\epsilon\right], \tag{5.38}$$

from which the result follows for this case.

■

5.2 General Quadratic Problem

Armed with the intuition from the linear regression problem, we begin a local analysis of nonconvex problems. The primary principle behind such an analysis is to approximate a region around a minimizer by a quadratic function, and analyze the properties of the estimator on this quadratic function. In doing this analysis, we will see that the results for SGD on the linear regression problem and classical results for gradient descent are two extremes on a spectrum. Moreover, we will again predict an exponential divergence pattern for the SGD estimates.

Although our analysis will not completely carry over for arbitrary stochastic nonconvex problems, the consequences for the nonconvex problem will still hold as our numerical

experiments will demonstrate.

5.2.1 The Quadratic Problem, its Properties and Two Types of Minimizers

We will begin by defining the quadratic problem, derive some of its relevant properties, and discuss the two types of minimizers that can occur for the quadratic problem.

Problem 5.2 (Quadratic Sums)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $Q \in \mathbb{R}^{p\times p}$ be a nonzero, symmetric, positive semi-definite random matrix and $r \in \mathbb{R}^p$ be a random vector in the image space of Q (with probability one) such that (1) $\mathbb{E}[Q] \prec \infty$, (2) $\mathbb{E}[Q\mathbb{E}[Q]\, Q] \prec \infty$, (3) $\mathbb{E}[Q\mathbb{E}[Q]\, r]$ is finite, and (4) $\mathbb{E}[r'Qr]$ is finite. Let $Y = (Q, r)$, and let $\{Y_N : N \in \mathbb{N}\}$ be independent copies of Y. The quadratic sums problem is to use $\{Y_N : N \in \mathbb{N}\}$ to determine a θ^* such that*

$$\theta^* \in \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\theta'\mathbb{E}[Q]\,\theta + \mathbb{E}[r]'\,\theta. \tag{5.39}$$

Note, in defining the quadratic sums problem, we have not required that the solution, θ^*, be unique. That is, we are not requiring a strongly convex problem. This generality is important as it will allow us to better approximate more exotic problems, such as those arising from neural networks or differential equations, which may have minimizers whose basins are not locally, strongly convex. We now establish some basic geometric properties about the quadratic sums problem.

Lemma 5.4. *The objective function in (5.39) is equivalent to (up to a constant)*

$$\frac{1}{2}(\theta - \theta^*)'\mathbb{E}[Q]\,(\theta - \theta^*), \tag{5.40}$$

for any θ^* satisfying (5.39).

Proof. If θ^* is a minimizer of the objective in (5.39), then θ^* must be a stationary point of the objective function. This implies that $-\mathbb{E}[Q]\,\theta^* = r$. Therefore, the objective in (5.39)

is, up to an additive constant,

$$\frac{1}{2}\theta'\mathbb{E}\left[Q\right]\theta - \theta'\mathbb{E}\left[Q\right]\theta^* + \frac{1}{2}(\theta^*)'\mathbb{E}\left[Q\right]\theta^* = \frac{1}{2}(\theta-\theta^*)'\mathbb{E}\left[Q\right](\theta-\theta^*), \tag{5.41}$$

■

Note, because Q is symmetric, positive semi-definite with probability one, $\mathbb{E}\left[Q\right]$ will also be symmetric, positive semi-definite. That is, $m := \texttt{rank}(\mathbb{E}\left[Q\right]) \leq p$. For future reference, we will denote the nonzero eigenvalues of $\mathbb{E}\left[Q\right]$ by $\lambda_1 \geq \cdots \lambda_m > 0$. Beyond the eigenvalues, we will also need some higher order curvature information, namely, s_Q and t_Q, which are given by

$$t_Q = \sup\left\{\frac{v'\mathbb{E}\left[Q\mathbb{E}\left[Q\right]Q\right]v - v'\mathbb{E}\left[Q\right]^3 v}{v'\mathbb{E}\left[Q\right]v} : v \in \mathbb{R}^p,\ v'\mathbb{E}\left[Q\right]v \neq 0\right\}, \tag{5.42}$$

and

$$s_Q = \inf\left\{\frac{v'\mathbb{E}\left[Q\mathbb{E}\left[Q\right]Q\right]v - v'\mathbb{E}\left[Q\right]^3 v}{v'\mathbb{E}\left[Q\right]v} : v \in \mathbb{R}^p,\ v'\mathbb{E}\left[Q\right]v \neq 0\right\}. \tag{5.43}$$

By the following result, it is straightforward to show that $s_Q > 0$ if $\mathbb{P}\left[Q = \mathbb{E}\left[Q\right]\right] < 1$.

Lemma 5.5. *Let Q be as in Problem 5.2. Then, $\mathbb{E}\left[Q\mathbb{E}\left[Q\right]Q\right] \succeq \mathbb{E}\left[Q\right]^3$. Moreover, if there is an $x \in \mathbb{R}^p$ such that $x'\mathbb{E}\left[Q\right]x = 0$ then $x'\mathbb{E}\left[Q\mathbb{E}\left[Q\right]Q\right]x = 0$.*

Proof. For the first part, note that for any $x \in \mathbb{R}^q$, the function $f(M) = x'M\mathbb{E}\left[Q\right]Mx$ is convex over the space of all symmetric, positive semi-definite matrices. Therefore, by Jensen's inequality, $\mathbb{E}\left[f(Q)\right] \geq f(\mathbb{E}\left[Q\right])$. Moreover, since x is arbitrary, the conclusion follows.

For the second part, when $x = 0$ the result follows trivially. Suppose then that there is an $x \neq 0$ such that $x'\mathbb{E}\left[Q\right]x = 0$. Then, since $Q \succeq 0$, $x'Qx = 0$ almost surely. Therefore, x is in the null space of all Q on a probability one set. Therefore, $x'Q\mathbb{E}\left[Q\right]Qx = 0$ with probability one. The conclusion follows. ■

Note that if $\mathbb{P}[Q = \mathbb{E}[Q]] = 1$ then there exists a solution to (5.39) such that $Q\theta^* + r = 0$ with probability one. This phenomenon of the existence of a single solution for all pairs (Q, r) with probability one will play a special role. This motivates the following definition.

Definition 5.1 (Homogeneous and Inhomogeneous Solutions)**.** *A solution to θ^* satisfying (5.39) is called homogeneous if*

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\theta' Q\theta + r'\theta \text{ with probability one.} \tag{5.44}$$

Otherwise, the solution is called inhomogeneous.

In the case of a homogeneous minimizer, the objective function in (5.44) can be rewritten as (up to an additive constant)

$$\frac{1}{2}(\theta - \theta^*)' Q(\theta - \theta^*), \tag{5.45}$$

which follows from the same reasoning as the expected value case. Importantly, SGD will behave differently for problems with homogeneous minimizers and inhomogeneous minimizer. As we will show, for homogeneous minimizers, SGD behaves rather similarly to classical gradient descent in terms of convergence and divergence, whereas for inhomogeneous minimizers, SGD will behave markedly differently. To show these results, we will first define SGD-k (i.e., mini-batch SGD) for the quadratic sums problem. Note, SGD-k is still a zero-order, linear model incremental estimator.

Definition 5.2 (SGD-k)**.** *Let $\theta_0 \in \mathbb{R}^p$ be arbitrary. SGD-k generates a sequence of estimators defined by*

$$\theta_{N+1} = \theta_N - \frac{C_{N+1}}{k} \sum_{j=Nk+1}^{(N+1)k} Q_j\theta_N + r_j, \tag{5.46}$$

where $\{C_N : N \in \mathbb{N}\}$ is a sequence of scalars.

Note, when $k = 1$, we have the usual notion of stochastic gradient descent and, as $k \to \infty$, we recover gradient descent (under additional moment assumptions). Therefore, the generalized notion of SGD considered in Definition 5.2 will allow us to generate results that explore the complete range of stochastic to deterministic methods, and bridge the relationship between stochastic gradient descent and gradient descent.

5.2.2 Characterizing SGD-k for the Quadratic Sums Problem

Because the connections to classical gradient descent will be the most obvious, we will begin our characterization of SGD-k for the quadratic sums problem when the quadratic sums problem has a homogeneous minimizer. Note, we relegate the technical lemmas for proving these statements to the next subsection.

Theorem 5.3 (Homogeneous Minimizer)**.** *Let $\theta_0 \in \mathbb{R}^p$ be arbitrary and let $\{\theta_N : N \in \mathbb{N}\}$ be the estimators generated by SGD-k for Problem 5.2. Let θ^* be a homogeneous solution to the quadratic sums problem. Denote by $e_N = (\theta_N - \theta^*)'\mathbb{E}[Q](\theta_N - \theta^*)$.*

If $0 < C_{N+1}$ and

$$C_{N+1} < \begin{cases} \frac{2\lambda_1}{\lambda_1^2 + t_Q/k} & k > t_Q/(\lambda_1\lambda_m) \\ \frac{2\lambda_m}{\lambda_m^2 + t_Q/k} & k \leq t_Q/(\lambda_1\lambda_m), \end{cases} \tag{5.47}$$

then $\mathbb{E}[e_{N+1}] < \mathbb{E}[e_N]$. Moreover, if $\{C_{N+1}\}$ satisfy the upper bound and are uniformly bounded away from zero, then $\mathbb{E}[e_N]$ decays exponentially to zero.

Moreover, there exists $\rho_N > 1$ such that $\mathbb{E}[e_{N+1}] > \rho_N \mathbb{E}[e_N]$, if either $C_{N+1} < 0$ or

$$C_{N+1} > \frac{2\lambda_j}{\lambda_j^2 + s_Q/k}, \tag{5.48}$$

where

$$j = \begin{cases} 1 & k \leq \frac{s_Q}{\lambda_1 \lambda_2} \\ l \ (l \in \{2, \ldots, m-1\}) & \frac{s_Q}{\lambda_l \lambda_{l-1}} < k \leq \frac{s_Q}{\lambda_{l+1} \lambda_l} \\ m & k > \frac{s_Q}{\lambda_m \lambda_{m-1}}. \end{cases} \tag{5.49}$$

Proof. The upper and lower bounds follow from Lemma 5.6 and Lemma 5.10. For the upper bound, note that $k > t_Q/(\lambda_1 \lambda_m)$ implies

$$\frac{2\lambda_j}{\lambda_j^2 + t_Q/k} > \frac{2}{\lambda_1 + \lambda_m}, \tag{5.50}$$

for $j = 1, m$. Moreover, when $C_{N+1} > 0$ and strictly less than expression on the left hand side of (5.50), the upper bound in Lemma 5.10 is strictly less than one. Thus, if $\{C_{N+1}\}$ are uniformly bounded away from zero, the exponential convergence rate of $\mathbb{E}[e_N]$ follows trivially.

For the lower bound, we make note of several facts. First, if $k \geq \frac{s_Q}{\lambda_l \lambda_{l+1}}$ then $k \geq \frac{s_Q}{\lambda_j \lambda_{j+1}}$ for $j = 1, 2, \ldots, l$. Moreover, when $k \geq \frac{s_Q}{\lambda_l \lambda_{l+1}}$, then

$$\frac{2\lambda_l}{\lambda_l^2 + s_Q/k} \geq \frac{2}{\lambda_l + \lambda_{l+1}}. \tag{5.51}$$

Since the left hand side of this inequality is the lower bound on C_{N+1} that guarantees that $\mathbb{E}[e_{N+1}] > \mathbb{E}[e_N]$ (by Lemma 5.10), then whenever $k \geq \frac{s_Q}{\lambda_l \lambda_{l+1}}$, $\mathbb{E}[e_N]$ is not guaranteed to diverge. On the other hand, if $k < \frac{s_Q}{\lambda_l \lambda_{l-1}}$ then $k < \frac{s_Q}{\lambda_j \lambda_{j-1}}$ for $j = l, l+1, \ldots, m$. Moreover, if $k < \frac{s_Q}{\lambda_l \lambda_{l-1}}$ then

$$\frac{2\lambda_l}{\lambda_l^2 + s_Q/k} < \frac{2}{\lambda_l + \lambda_{l-1}}. \tag{5.52}$$

Therefore, since the left hand side of this inequality is the lower bound on C_{N+1} that guarantees that $\exists \rho_N > 1$ such that $\mathbb{E}[e_{N+1}] > \rho_N \mathbb{E}[e_N]$, the lower bound is guaranteed to diverge for C_{N+1} larger than the right hand side.

Thus, by the monotonicty of the eigenvalues, for the $l \in \{2, \dots, m-1\}$ such that $\frac{s_Q}{\lambda_l \lambda_{l-1}} < k \leq \frac{s_Q}{\lambda_l \lambda_{l+1}}$, $\mathbb{E}[e_{N+1}] > \mathbb{E}[e_N]$ if

$$C_{N+1} > \frac{2\lambda_l}{\lambda_l^2 + s_Q/k}. \tag{5.53}$$

Note, we can handle the edge cases similarly. ■

Theorem 5.3 generalizes the known results for classical gradient descent. In particular, as $k \to \infty$, the upper bound on the step size to induce convergence is $\frac{2}{\lambda_1}$ and the lower bound on the step size to induce divergence is $\frac{2}{\lambda_m}$, which are the known bounds for gradient descent to convergence or diverge, respectively (see Bertsekas, 1999, Chapter 1). Furthermore, Theorem 5.3 also captures the exponential divergence property that we observed for linear regression (see Theorem 5.2) and that we observed for gradient descent (see Fig. 5.3).

However, Theorem 5.3 shows that SGD-k is distinct from classical gradient descent in several ways. First, SGD-k has distinct phases for convergence and divergence depending on the batch-size and expected geometry of the quadratic sums problem as represented by the eigenvalues, t_Q and s_Q. Thus, as opposed to classical gradient descent, SGD-k requires a more complex convergence and divergence statement to capture these distinct phases. We now state the analogous statement for the inhomogeneous problem.

Theorem 5.4 (Inhomogeneous Minimizer)**.** *Let $\theta_0 \in \mathbb{R}^p$ be arbitrary and let $\{\theta_N : N \in \mathbb{N}\}$ be the estimators generated by SGD-k for Problem 5.2. Suppose θ^* is an inhomogeneous solution to the quadratic sums problem. Denote by $e_N = (\theta_N - \theta^*)' \mathbb{E}[Q](\theta_N - \theta^*)$.*

If $0 < C_{N+1}$ and

$$C_{N+1} < \begin{cases} \frac{2\lambda_1^2}{(1+1/k)\lambda_1^2 + t_Q/k} & k+1 > t_Q/(\lambda_1 \lambda_m) \\ \frac{2\lambda_m^2}{(1+1/k)\lambda_m^2 + t_Q/k} & k+1 \leq t_Q/(\lambda_1 \lambda_m), \end{cases} \tag{5.54}$$

then $\mathbb{E}[e_{N+1}] < \rho_N \mathbb{E}[e_N] + C_{N+1}^2 \frac{\psi}{k}$, where $\rho_N < 1$ and is uniformly bounded away from zero for all N, and $\psi > 0$. Moreover, if $\{C_N : N \in \mathbb{N}\}$ are nonnegative, converge to zero and

sum to infinity then $\mathbb{E}[e_N] \to 0$ *as* $N \to \infty$.

Furthermore, suppose $C_{N+1} < 0$ *or*

$$C_{N+1} > \frac{2(\lambda_j + \gamma)}{\lambda_j^2 + s_Q/k}, \tag{5.55}$$

where $4\gamma^2 \in \left(0, \frac{s_Q}{k}\right]$ *and*

$$j = \begin{cases} 1 & k \leq \frac{s_Q}{\lambda_1\lambda_2 + \gamma(\lambda_1+\lambda_2)} \\ l \ (l \in \{2, \ldots, m-1\}) & \frac{s_Q}{\lambda_l\lambda_{l-1} + \gamma(\lambda_l+\lambda_{l-1})} < k \leq \frac{s_Q}{\lambda_l\lambda_{l+1} + \gamma(\lambda_l+\lambda_{l+1})} \\ m & k > \frac{s_Q}{\lambda_m\lambda_{m-1} + \gamma(\lambda_m+\lambda_{m-1})}. \end{cases} \tag{5.56}$$

There $\exists \delta > 0$ *such that if* $\|\mathbb{E}[Q](\theta_N - \theta^*)\| < \delta$ *then* $\mathbb{E}[e_{N+1}|\,\theta_N] \geq \rho_N e_N + C_{N+1}^2\psi$ *with probability one, where* $\rho_N > 1$ *and* $\psi > 0$ *independently of* N.

Proof. For the upper bound, we apply Lemmas 5.9 and 5.10 to conclude that

$$\mathbb{E}[e_{N+1}|\,\theta_N] \leq e_N\left(1 - 2C_{N+1}\lambda_j + \left(1 + \frac{1}{k}\right)C_{N+1}^2\lambda_j^2 + C_{N+1}^2\frac{t_Q}{k}\right) + C_{N+1}^2\frac{\psi}{k}, \tag{5.57}$$

where $\psi > 0$ is independent of N and

$$j = \begin{cases} 1 & C_{N+1} \leq 0 \text{ or } \frac{2}{(1+1/k)(\lambda_1+\lambda_m)} \\ m & \text{otherwise.} \end{cases} \tag{5.58}$$

Moreover, the component multiplying e_N is less than one if $C_{N+1} > 0$ and

$$C_{N+1} < \frac{2\lambda_j}{\left(1 + \frac{1}{k}\right)\lambda_j^2 + \frac{t_Q}{k}}. \tag{5.59}$$

When $k + 1 > t_Q/(\lambda_1\lambda_m)$, the right hand side of (5.59) is larger than $\frac{2}{(1+1/k)(\lambda_1+\lambda_m)}$. The upper bound follows.

For the lower bound, note that since θ^* is inhomogeneous, $\mathbb{P}[Q = \mathbb{E}[Q]] < 1$. Therefore, $s_Q > 0$. Now, we apply Lemmas 5.9 and 5.10 to conclude that for any $\gamma > 0$ there exists ψ_N such that

$$\mathbb{E}[e_{N+1}|\,\theta_N] \geq e_N \left(\left(1 - \frac{4\gamma^2}{\lambda_j^2 + \frac{1}{k}s_Q} \right) - 2C_{N+1}\lambda_j + C_{N+1}^2\lambda_j^2 + C_{N+1}^2 \frac{s_Q}{k} \right) + C_{N+1}^2 \frac{\psi_N}{k}, \quad (5.60)$$

where ψ_N are nonnegative and uniformly bounded from zero for C_{N+1} sufficiently small, and

$$j = \begin{cases} 1 & C_{N+1} \in \left(0, \frac{2}{\lambda_1 + \lambda_2}\right] \\ l \ (l \in \{2, \ldots, m-1\}) & C_{N+1} \in \left(\frac{2}{\lambda_{l-1} + \lambda_l}, \frac{2}{\lambda_l + \lambda_{l+1}}\right] \\ m & C_{N+1} \leq 0 \text{ or } \frac{2}{\lambda_m + \lambda_{m-1}} < C_{N+1}. \end{cases} \quad (5.61)$$

The term multiplying e_N can be rewritten as

$$1 - \frac{4\gamma^2}{\lambda_j^2 + \frac{1}{k}s_Q} - \frac{\lambda_j^2}{\lambda_j^2 + \frac{1}{k}s_Q} + \left(\lambda_j^2 + \frac{1}{k}s_Q\right) \left(C_{N+1} - \frac{\lambda_j}{\lambda_j^2 + \frac{1}{k}s_Q} \right)^2, \quad (5.62)$$

which nonnegative when $4\gamma^2$ is in the interval given by in the statement of the result. Moreover, if

$$C_{N+1} > \frac{2(\lambda_j + \gamma)}{\lambda_j^2 + \frac{s_Q}{k}}, \quad (5.63)$$

then

$$C_{N+1} > \frac{\lambda_j}{\lambda_j^2 + \frac{s_Q}{k}} + \frac{\sqrt{\lambda_j^2 + 4\gamma^2}}{\lambda_j^2 + \frac{s_Q}{k}}, \quad (5.64)$$

which implies

$$\left(C_{N+1} - \frac{\lambda_j}{\lambda_j^2 + \frac{s_Q}{k}} \right)^2 > \frac{\lambda_j^2 + 4\gamma^2}{\left(\lambda_j^2 + \frac{s_Q}{k}\right)^2}. \quad (5.65)$$

From this inequality, we conclude that if (5.63) holds then (5.62) is strictly larger than 1.

Lastly, for any $\tilde{j} \neq j$,

$$\frac{s_Q}{\lambda_j \lambda_{\tilde{j}} + \gamma(\lambda_j + \lambda_{\tilde{j}})} > k \tag{5.66}$$

is equivalent to

$$\frac{2}{\lambda_j + \lambda_{\tilde{j}}} > \frac{2(\lambda_j + \gamma)}{\lambda_j^2 + \frac{1}{k} s_Q}. \tag{5.67}$$

With these facts, the conclusion for the lower bound just as in the proof of Theorem 5.3. ■

Comparing Theorems 5.3 and 5.4, we see that inhomogeneity exacts an additional price. To be specific, call the term multiplying e_N in the bounds as the systematic component and the additive term as the variance component. First, the inhomogeneous case contains a variance component that was not present for the homogeneous case. Of course, this variance component is induced by the inhomogeneity: at each update, SGD-k is minimizing an approximation to the quadratic problem that possibly has a distinct solution, which requires decaying the step-sizes to remove the variability of solutions induced by these distinct approximations. Second, the inhomogeneous case exacts an additional price: the inhomogeneity shrinks the threshold for step sizes that ensure a reduction in the systematic component, and inflates the threshold for step sizes that ensure an increase in the systematic component.

Continuing to compare the results for the homogeneous and inhomogeneous case, we see that the systematic component grows exponentially. Of course, Theorem 5.4 is not as satisfying as Theorem 5.3 in that we require the gradient of the expected objective function to be small. However, this is also the precise region in which such a result is important: it says that for step sizes that are too large, any estimate near the solution set is highly unstable. In other words, if we initialize the SGD estimates near the solution set, our estimates will diverge exponentially fast from this region.

Now, using the quadratic problem to locally approximate nonconvex problems near minimizers, the preceding results predict some interesting phenomenon. First, owing to the different phases, as k increases, the threshold for divergence increases. Hence, identical

learning rates applies to minibatch SGD with different batch sizes can result in distinct minimizers. In particular, the SGD-k with the smaller batch size (in comparison to SGD-k with the larger batch size) will be less likely to converge to minimizers whose quadratic approximations have large eigenvalues. Thus, these results explain the phenomenon observed by Keskar et al. (2016) of why small batch methods are more likely to converge to flatter minimizers. In the next section, while we can formalize these results for the nonconvex problem, here we numerically study whether or not our mechanism is the appropriate one for nonconvex problems. Before moving to this topic, we prove the technical lemmas needed to prove the main two results above.

5.2.3 Technical Lemmas for SGD-k for the Quadratic Sums Problem

Lemma 5.6. *Let $\theta_0 \in \mathbb{R}^p$ be arbitrary and let $\{\theta_N : N \in \mathbb{N}\}$ be the estimators generated by SGD-k for Problem 5.2. Let θ^* be a solution to the quadratic sums problem, and let $e_N = (\theta_N - \theta^*)'\mathbb{E}[Q](\theta_N - \theta^*)$. Then, with probability one,*

$$\begin{aligned}\mathbb{E}[e_{N+1}|\,\theta_N] = e_N &- 2C_{N+1}(\theta_N - \theta^*)'\mathbb{E}[Q]^2(\theta_N - \theta^*) + C_{N+1}^2(\theta_N - \theta^*)'\mathbb{E}[Q]^3(\theta_N - \theta^*) \\ &+ \frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\bar{M}(\theta_N - \theta^*) + 2\frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}[Q\mathbb{E}[Q](Q\theta^* + r)] \\ &+ \frac{C_{N+1}^2}{k}\mathbb{E}[(Q\theta^* + r)'\mathbb{E}[Q](Q\theta^* + r)],\end{aligned} \tag{5.68}$$

where $M = \mathbb{E}[Q\mathbb{E}[Q]Q] - \mathbb{E}[Q]^3 \succeq 0$. Moreover, if θ^* is a homogeneous minimizer, then, with probability one,

$$\begin{aligned}\mathbb{E}[e_{N+1}|\,\theta_N] = e_N &- 2C_{N+1}(\theta_N - \theta^*)'\mathbb{E}[Q]^2(\theta_N - \theta^*) + C_{N+1}^2(\theta_N - \theta^*)\mathbb{E}[Q]^3(\theta_N - \theta^*) \\ &+ \frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\bar{M}(\theta_N - \theta^*).\end{aligned} \tag{5.69}$$

Proof. The result is a straightforward calculation from the properties of SGD-k and the quadratic sums problem. In the case of the homogeneous minimizer, recall that $Q\theta^* + r = 0$

with probability one. ■

Lemma 5.7. *Let Q and r be as in Problem 5.2. Then, letting $(\cdot)^\dagger$ denote the Moore-Penrose pseudo-inverse,*

$$\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right] = \mathbb{E}\left[Q\right]\mathbb{E}\left[Q\right]^\dagger \mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]. \tag{5.70}$$

Proof. Note that for any $x \in \texttt{row}(\mathbb{E}\left[Q\right])$, $x = \mathbb{E}\left[Q\right]\mathbb{E}\left[Q\right]^\dagger x$. Thus, we must show that $\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]$ is in $\texttt{row}(\mathbb{E}\left[Q\right])$. Recall, that if there exists a $v \in \mathbb{R}^p$ such that $\mathbb{E}\left[Q\right]v = 0$, then $v'\mathbb{E}\left[Q\right]v = 0$. In turn, $v'Qv = 0$ with probability one, which implies that $Qv = 0$ with probability one. Hence, $\texttt{null}(\mathbb{E}\left[Q\right]) \subset \texttt{null}(Q)$ with probability one. Let the set of probability one be denoted by Ω'. Then,

$$\texttt{row}(\mathbb{E}\left[Q\right]) = \texttt{null}(\mathbb{E}\left[Q\right])^\perp \supset \bigcup_{\omega\in\Omega'} \texttt{null}(Q)^\perp = \bigcup_{\omega\in\Omega'} \texttt{col}(Q) \supset \bigcup_{\omega\in\Omega'} \texttt{col}(Q\mathbb{E}\left[Q\right]), \tag{5.71}$$

where $\texttt{row}(Q) = \texttt{col}(Q)$ by symmetry. Therefore, $Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right) \in \texttt{row}(\mathbb{E}\left[Q\right])$ with probability one. Hence, its expectation is in $\texttt{row}(\mathbb{E}\left[Q\right])$. ■

Lemma 5.8. *Under the setting of Lemma 5.6, for any $\phi > 0$ and $j \in \mathbb{N}$,*

$$\begin{aligned}
&2\left|\frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]\right| \\
&\quad\leq 2\frac{C_{N+1}^2}{k}\left\|\mathbb{E}\left[Q\right]^{j/2}\left(\theta_N - \theta^*\right)\right\|_2\left\|(\mathbb{E}\left[Q\right]^{j/2})^\dagger\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]\right\|_2 \\
&\quad\leq \frac{\phi C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^j\left(\theta_N - \theta^*\right) \\
&\qquad + \frac{C_{N+1}^2}{\phi k}\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]Q\right](\mathbb{E}\left[Q\right]^j)^\dagger\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right],
\end{aligned} \tag{5.72}$$

for any $N = 0, 1, 2, \ldots$.

Proof. We will make use of three facts: Lemma 5.7, Hölder's inequality, and the fact that

for any $\phi > 0$, $\varphi \geq 0$ and $x \in \mathbb{R}$, $2|\varphi x| \leq \phi x^2 + \phi^{-1}\varphi^2$. Using these facts, we have

$$\begin{aligned}
&2\left|\frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]\right| \\
&\quad = 2\left|\frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^{j/2}\left(\mathbb{E}\left[Q\right]^{j/2}\right)^{\dagger}\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]\right| \\
&\quad \leq 2\frac{C_{N+1}^2}{k}\left\|\mathbb{E}\left[Q\right]^{j/2}\left(\theta_N - \theta^*\right)\right\|_2\left\|\left(\mathbb{E}\left[Q\right]^{j/2}\right)^{\dagger}\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]\right\|_2 \\
&\quad \leq \frac{\phi C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^{j}\left(\theta_N - \theta^*\right) \\
&\quad + \frac{C_{N+1}^2}{\phi k}\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]Q\right]\left(\mathbb{E}\left[Q\right]^{j}\right)^{\dagger}\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right].
\end{aligned} \tag{5.73}$$

■

Lemma 5.9. *Under the setting of Lemma 5.6,*

$$\begin{aligned}
\mathbb{E}\left[e_{N+1}|\,\theta_N\right] \leq\ & e_N - 2C_{N+1}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^2\left(\theta_N - \theta^*\right) + C_{N+1}^2(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^3\left(\theta_N - \theta^*\right) \\
&+ \frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\bar{M}(\theta_N - \theta^*) + \frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^3\left(\theta_N - \theta^*\right) \\
&+ \frac{C_{N+1}^2}{k}\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right] \\
&+ \frac{C_{N+1}^2}{k}\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]Q\right]\left(\mathbb{E}\left[Q\right]^3\right)^{\dagger}\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]
\end{aligned} \tag{5.74}$$

Now, for any $\varphi > 0$, $\exists \psi \in \mathbb{R}$ such that

$$\begin{aligned}
\mathbb{E}\left[e_{N+1}|\,\theta_N\right] \geq\ & \left(1 - \frac{\varphi}{k}\right)e_N - 2C_{N+1}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^2\left(\theta_N - \theta^*\right) \\
&+ C_{N+1}^2(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^3\left(\theta_N - \theta^*\right) + \frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\bar{M}(\theta_N - \theta^*) \\
&+ \frac{C_{N+1}^2}{k}\psi,
\end{aligned} \tag{5.75}$$

where $\psi \geq 0$ if

$$\varphi\frac{\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]}{\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]Q\right]\mathbb{E}\left[Q\right]^{\dagger}\mathbb{E}\left[Q\mathbb{E}\left[Q\right]\left(Q\theta^* + r\right)\right]} \geq C_{N+1}^2. \tag{5.76}$$

Proof. When $C_{N+1} = 0$, the statements hold by Lemma 5.6. Suppose $C_{N+1} \neq 0$. Using Lemma 5.6 and applying Lemma 5.8 with $\phi = 1$ and $j = 3$, the upper bound follows. Now, using Lemma 5.6 and applying Lemma 5.8 with $\phi = C_{N+1}^{-2}\varphi$ for $\varphi > 0$ and $j = 1$, we have

$$\begin{aligned}\mathbb{E}\left[e_{N+1}|\,\theta_N\right] \geq & \left(1 - \frac{\varphi}{k}\right) e_N - 2C_{N+1}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^2(\theta_N - \theta^*) \\ & + C_{N+1}^2(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^3(\theta_N - \theta^*) + \frac{C_{N+1}^2}{k}(\theta_N - \theta^*)'\bar{M}(\theta_N - \theta^*) \\ & + \frac{C_{N+1}^2}{k}\psi,\end{aligned} \tag{5.77}$$

where

$$\begin{aligned}\psi = & -\frac{C_{N+1}^2}{\varphi}\mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right]Q\right]\mathbb{E}\left[Q\right]^\dagger\mathbb{E}\left[Q\mathbb{E}\left[Q\right](Q\theta^* + r)\right] \\ & + \mathbb{E}\left[(Q\theta^* + r)'\mathbb{E}\left[Q\right](Q\theta^* + r)\right].\end{aligned} \tag{5.78}$$

Therefore, if φ is given as selected in the statement of the result, then $\psi \geq 0$. ■

Lemma 5.10. *Let t_Q and s_Q be defined as in (5.42) and (5.43). In the setting of Lemma 5.6, denote $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ as the nonnegative eigenvalues of $\mathbb{E}\left[Q\right]$, where m is the rank of $\mathbb{E}\left[Q\right]$. Then, for $\alpha_j \geq 0$ for $j = 0, 1, 2, 3$,*

$$\begin{aligned}&\alpha_0(\theta_N - \theta^*)'\mathbb{E}\left[Q\right](\theta_N - \theta^*) - \alpha_1 C_{N+1}(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^2(\theta_N - \theta^*) \\ &+ \alpha_2 C_{N+1}^2(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^3(\theta_N - \theta^*) + \alpha_3 C_{N+1}^2(\theta_N - \theta^*)'\bar{M}(\theta_N - \theta^*)\end{aligned} \tag{5.79}$$

is bounded above by

$$(\theta_N - \theta^*)'\mathbb{E}\left[Q\right](\theta_N - \theta^*)\left[\alpha_0 - \alpha_1 C_{N+1}\lambda_j + \alpha_2 C_{N+1}^2\lambda_j^2 + \alpha_3 C_{N+1}^2 t_Q\right], \tag{5.80}$$

where

$$j = \begin{cases} 1 & C_{N+1} > \frac{\alpha_1}{\alpha_2}\frac{1}{\lambda_1 + \lambda_m} \text{ or } C_{N+1} \leq 0 \\ m & \text{otherwise.} \end{cases} \tag{5.81}$$

Moreover, (5.79) is bounded below by

$$(\theta_N - \theta^*)'\mathbb{E}\left[Q\right](\theta_N - \theta^*)\left[\alpha_0 - \alpha_1 C_{N+1}\lambda_j + \alpha_2 C_{N+1}^2\lambda_j^2 + \alpha_3 C_{N+1}^2 s_Q\right], \tag{5.82}$$

where

$$j = \begin{cases} 1 & C_{N+1} \in \left(0, \frac{\alpha_1}{\alpha_2}\frac{1}{\lambda_1+\lambda_2}\right] \\ l \ (l \in \{2,\ldots,m-1\}) & C_{N+1} \in \left(\frac{\alpha_1}{\alpha_2}\frac{1}{\lambda_{l-1}+\lambda_l}, \frac{\alpha_1}{\alpha_2}\frac{1}{\lambda_l+\lambda_{l+1}}\right] \\ m & C_{N+1} > \frac{\alpha_1}{\alpha_2}\frac{1}{\lambda_{m-1}+\lambda_m} \text{ or } C_{N+1} \leq 0. \end{cases} \tag{5.83}$$

Proof. When $(\theta_N - \theta^*)\mathbb{E}\left[Q\right]^j(\theta_N - \theta^*) = 0$ for $j = 1$, then it holds for $j = 1, 2$ and $(\theta_N - \theta^*)'M(\theta_N - \theta^*) = 0$ by Lemma 5.5. Therefore, in this case, the bounds hold trivially. So, we will assume that $(\theta_N - \theta^*)\mathbb{E}\left[Q\right](\theta_N - \theta^*) > 0$.

Let $u_1, \ldots, u_m$ be the orthonormal eigenvectors corresponding to $\lambda_1, \ldots, \lambda_m$. Then,

$$\frac{(\theta_N - \theta^*)'\mathbb{E}\left[Q\right]^j(\theta_N - \theta^*)}{(\theta_N - \theta^*)'\mathbb{E}\left[Q\right](\theta_N - \theta^*)} = \sum_{l=1}^{m} \lambda_l^{j-1}\frac{\lambda_l(u_l'(\theta_N - \theta^*))^2}{(\theta_N - \theta^*)'\mathbb{E}\left[Q\right](\theta_N - \theta^*)}. \tag{5.84}$$

Denote the ratio on the right hand side by w_l, and note that $\{w_l : l = 1, \ldots, m\}$ sum to one and are nonnegative. With this notation, we bound (5.79) from above by

$$(\theta_N - \theta^*)'\mathbb{E}\left[Q\right](\theta_N - \theta^*)\left[\alpha_0 - \alpha_1 C_{N+1}\sum_{l=1}^{m}\lambda_l w_l + \alpha_2 C_{N+1}^2\sum_{l=1}^{m}\lambda_l^2 w_l + \alpha_3 C_{N+1}^2 t_Q\right], \tag{5.85}$$

and from below by the same equation but with t_Q replaced by s_Q. By the properties of $\{w_l : l = 1, \ldots, m\}$, we see that the bounds are composed of convex combinations of quadratics of the eigenvalues. Thus, for the upper bound, if we assign all of the weight to the eigenvalue that maximizes the polynomial $-\alpha_1 C_{N+1}\lambda + \alpha_2 C_{N+1}^2\lambda^2$, then we will have the upper bound presented in the result. To compute the lower bounds, we do the analogous calculation. ■

5.3 Nonconvex Problem

We now numerically explore whether our analysis of the quadratic problem can be used as a local analysis for nonconvex problems by studying two distinct, contrived nonconvex problems. In general, using a quadratic local approximation and applying Theorems 5.3 and 5.4, we will observe that (1) the estimated thresholds based on the quadratic local approximation are sufficiently accurate, and (2) when SGD estimates diverge from minimizers, the divergence follows an exponential pattern. With these two observations, we will conclude that the generalization of the deterministic mechanism for divergence is sufficient to explain the observed geometric preferences of SGD for nonconvex problems.

In Subsection 5.3.1, we will formulate the first nonconvex problem call the quadratic-circle sums problem. In Subsection 5.3.2, we experiment with SGD-k on the quadratic-circle sums problem. In Subsection 5.3.3, we study SGD-k on a modification of the Styblinski-Tang problem. In Subsection 5.3.4, we briefly summarize our results.

5.3.1 Quadratic-Circle Problem

We begin by studying the SGD-k on a nonconvex problem with homogeneous minimizers. The nonconvex problem, called the quadratic-circle sums problem, is the sum of a structurally similar contrived functions with rich, modifiable surfaces that also have a tractable analytic form. The tractable analytic form allows us to better analyze the SGD estimates.

Each quadratic-circle function is the minimum of two functions: a quadratic basin function and a circular basin function. The quadratic basin function is defined as

$$g(x) = q_1(x_2 - q_2 x_1^2 - q_3)^2. \tag{5.86}$$

The quadratic basin function, g, traces out a parabola in the plane of its arguments where $g(x)$ is zero and increases quadratically as x_2 deviates from this parabola. The parameters q_1 and q_2 are non-negative. The parameter q_1 determines how quickly $g(x)$ increases off of

the parabola (i.e., its sharpness), and the remaining parameters determine the shape of the parabola. Examples of the quadratic basin function are shown in Fig. 5.9.

The circular basin function is a piecewise function defined as

$$h(x) = \begin{cases} c_4 & \|x\| \leq c_2 \\ c_4 + c_1 & \|x\| \geq c_3 \\ c_4 + c_1 \left(\frac{\|x\|-c_2}{c_3-c_2}\right)^3 \left[6\left(\frac{\|x\|-c_2}{c_3-c_2}\right)^2 - 15\left(\frac{\|x\|-c_2}{c_3-c_2}\right) + 10\right] & \text{otherwise} \end{cases}, \tag{5.87}$$

where all of the parameters are nonnegative. The circular basin function, h, defines a flat region encompassed by a circle of radius c_2 and defines a flat region outside of a circle of radius c_3. Between these two circles and for $c_2 > 0$, we have a twice differentiable function which ensures that $h(x)$ is twice continuously differentiable. When the ratio between c_3 and c_2 decreases, the steepness of the basin's walls increases (i.e., the walls are sharper). Increasing the distance between c_4 and c_1 has a similar outcome. Examples of the circular basin function are shown in Fig. 5.10.

The quadratic-circle function is then defined to be the minimum of the quadratic basin function and the circular basin function:

$$f(x) = \min\{h(x), g(x)\}. \tag{5.88}$$

Examples of the quadratic-circle function are shown in Fig. 5.11.

Remark 5.2. *The quadratic-circle function has curves of non-differentiability. If an iterate is at a point of non-differentiability, the gradient will always be selected to be the quadratic component. However, any point randomly selected from the plane using a measure which admits a probability density function with respect to the Lebesgue measure has a probability of zero of ever hitting a non-differentiable point.*

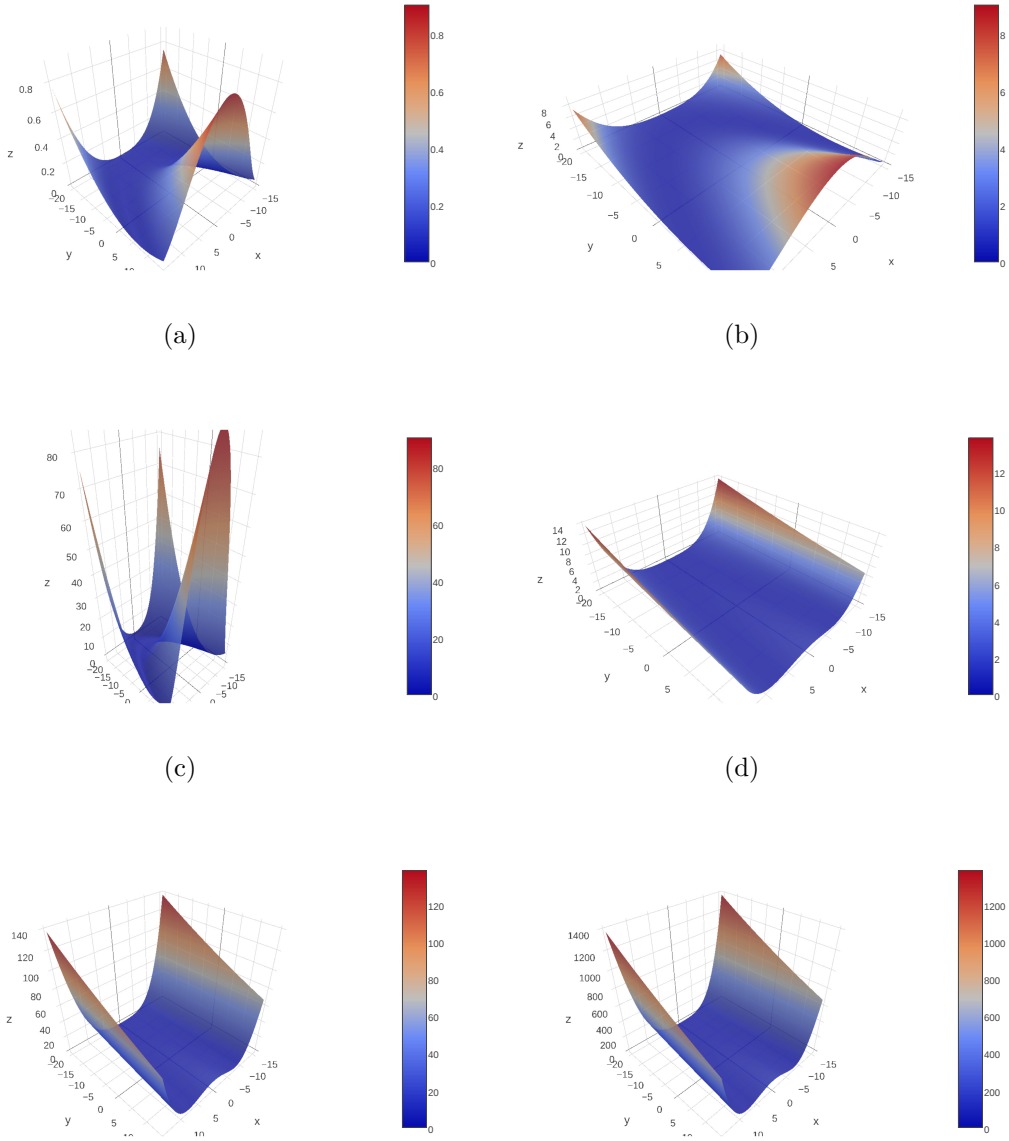
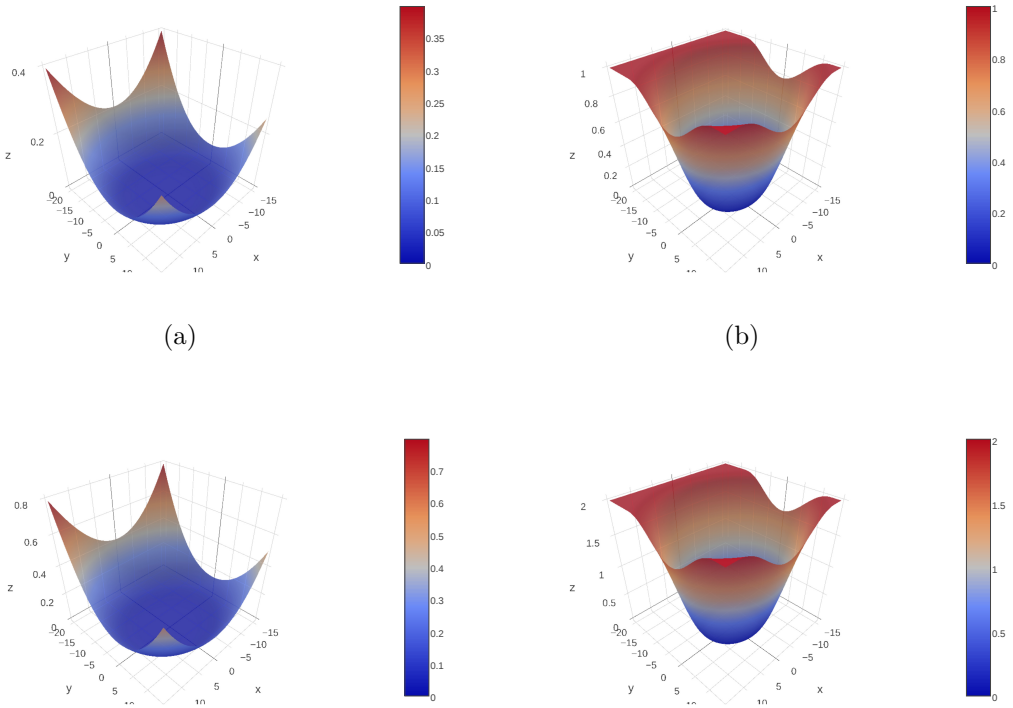


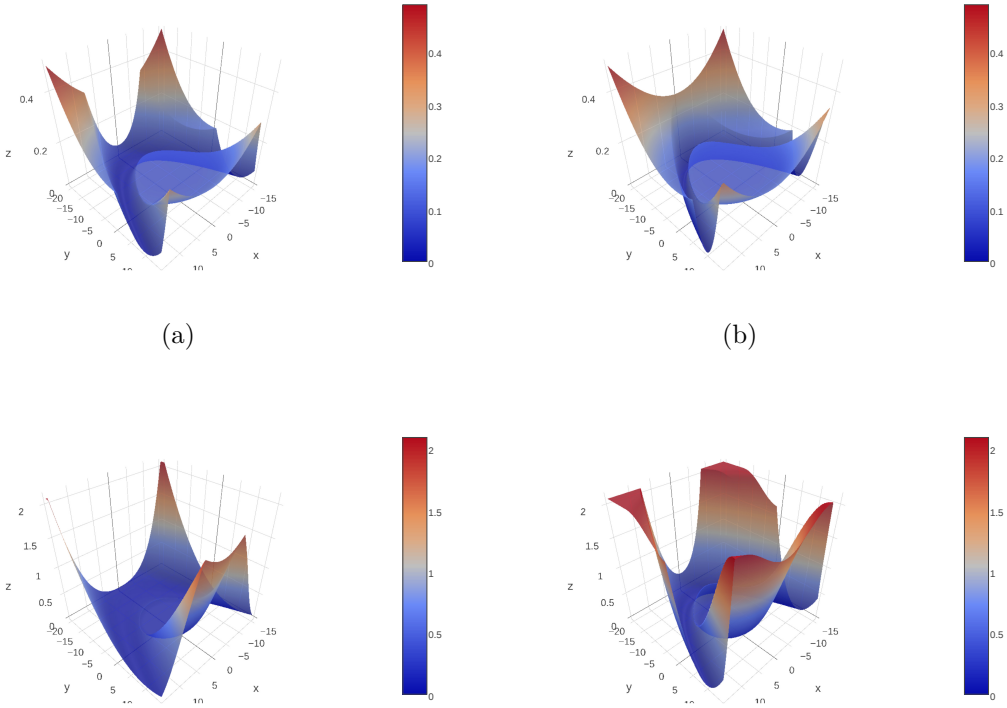
Figure 5.9: Examples of the quadratic basin functions used in constructing the Quadratic-Circle problem.



(a) (b)

(c) (d)

Figure 5.10: Examples of the circular basin functions used in constructing the Quadratic-Circle Problem.



(a) (b)

(c) (d)

Figure 5.11: Examples of the quadratic-circle function.

The nonconvex objective function that we will use is a summation of N functions of the form (5.88) with different parameters for the quadratic and circular components. For later reference, we define this objective function presently.

Problem 5.3 (Quadratic-Circle Sums)**.** *Let $g_1, \ldots, g_N$ be functions as specified (5.86) and $h_1, \ldots, h_N$ be functions as specified by (5.87). The quadratic-circle sums objective function is*

$$\sum_{i=1}^{N} p_i \min\{g_i(x), h_i(x)\}, \tag{5.89}$$

where p_i are positive valued and sum to one. Let Y be a random variable taking values $\{(g_i, h_i) : i = 1, \ldots, N\}$ with the probability of sampling $(g_i, h_i) = p_i$. Let $Y_1, Y_2, \ldots$ be independent copies of Y. The quadratic-circle sums problem is to use $\{Y_1, Y_2, \ldots\}$ to minimize

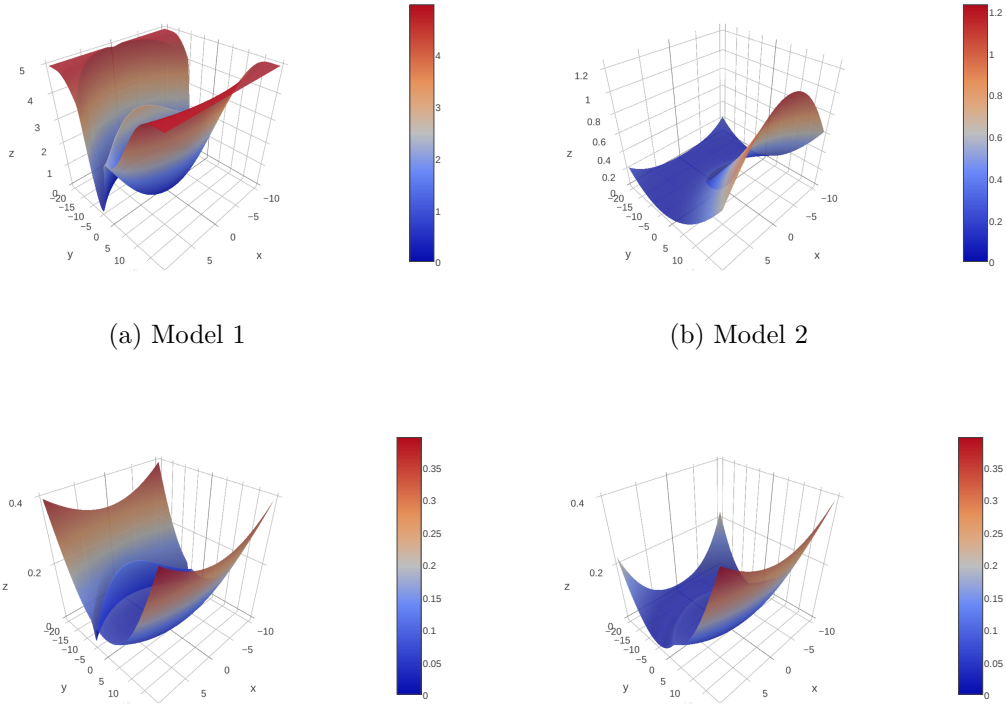
(5.89) *restricted to* $(-10, 10) \times (-20, 15) \subset \mathbb{R}^2$.

5.3.2 SGD-k on the Quadratic-Circle Sums Problem

In general, we will design four model quadratic-circular sums problems to analyze. All of the model functions have two minimizers in the region of interest: one for the circle-basin component at $(0, 0)$ and one for quadratic basin component at $(0, -15)$. However, the sharpness of the basins about each minimizers distinguishes the four model quadratic-circle sums problems. The first model, Model 1, is characterized by having at least one large eigenvalue about each minimizer. The second model, Model 2, is characterized by having at least one large eigenvalue for the circular basin minimizer and only small eigenvalues for the quadratic basin minimizer. Model 3 is characterized by having at least one large eigenvalue for the quadratic basin minimizer and only small eigenvalues for the circular basin minimizer. Model 4 is characterized by having only small eigenvalues for both minimizers. Note, the size of these eigenvalues are relative to the other models and not some general baseline. For clarity, each of the four model expected quadratic-circle sums objective functions are visualized in Fig. 5.12.

Now, in order to use our local analysis to study SGD-k on these four model problems, we will need to estimate s_Q, t_Q, the lower bound on the learning rate for divergence and the upper bound on the learning rate for convergence as defined in Theorem 5.3. Because the minimizers are not exactly quadratics, we will estimate the geometric parameters by averaging over quadratic approximations to points within a small neighborhood of the minimizer and then compute the lower bounds for divergence and upper bounds for convergence from these geometric parameter estimates. The lower bound estimates are reported in Table 5.2. Note, for the lower bounds, $k_{\max}$ is the largest integer smaller than $\frac{s_Q}{\lambda_{m-1}\lambda_m}$. The upper bounds are reported in Table 5.3. Note, for the upper bounds, $k_{\max}$ is the the integer closest to $\frac{t_Q}{\lambda_m\lambda_1}$.

There are several notable features in Tables 5.3 and 5.3, but since the main points are



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 5.12: The expected objective function for the four model quadratic-circle sums objective functions.

transferable between these two cases, we focus on the lower bound results. First, the flatter minima tend to have a larger lower bound for divergence. For example, since the circular-basin minima in Models 3 and 4 are flatter than the circular-basin minima in Models 1 and 2, the lower bound for divergence is nearly ten thousand fold larger. Similarly, since the quadratic minima in Models 2 and 4 are flatter than the quadratic basin minima in Models 1 and 3, the lower bound for divergence is over a hundred times larger. Second, for the circular

Table 5.2: Estimates of (5.48) for the Quadratic-Circle Sums Problem Models

	Circ. Minimum				Quad. Minimum	
	$k = 1$	$k = 0.99k_{\max}$	$k = 2.0k_{\max}$	$k = \infty$	$k = 1$	$k = \infty$
Model 1	$3.560e+11$	$3.583e+11$	$3.583e+11$	$3.583e+11$	$4.977e+00$	$4.981e+00$
Model 2	$3.339e+11$	$3.360e+11$	$3.360e+11$	$3.360e+11$	$9.954e+02$	$9.961e+02$
Model 3	$1.266e+15$	$1.273e+15$	$1.273e+15$	$1.273e+15$	$4.977e+00$	$4.981e+00$
Model 4	$1.269e+15$	$1.277e+15$	$1.277e+15$	$1.277e+15$	$9.954e+02$	$9.961e+02$

Table 5.3: Estimates of (5.47) for the Quadratic-Circle Sums Problem Models

	Circ. Minimum				Quad. Minimum	
	$k = 1$	$k = 0.99k_{\max}$	$k = 2.0k_{\max}$	$k = \infty$	$k = 1$	$k = \infty$
Model 1	$8.318e-05$	$1.541e+11$	$2.013e+11$	$2.874e+11$	$4.977e+00$	$4.981e+00$
Model 2	$8.456e-05$	$1.624e+11$	$2.157e+11$	$3.180e+11$	$9.954e+02$	$9.961e+02$
Model 3	$5.655e-06$	$6.222e+14$	$8.282e+14$	$1.226e+15$	$4.977e+00$	$4.981e+00$
Model 4	$5.511e-06$	$6.267e+14$	$8.264e+14$	$1.201e+15$	$9.954e+02$	$9.961e+02$

basin minimum, there is a very minor difference between the lower bounds for divergence between the $k = 1$ and $k = \infty$ cases. This is to be expected since the circular symmetry of the basin implies that its condition number is one.

Guided by the quantities in Tables 5.3 and 5.3, we now run SGD-1 and SGD-∞ (i.e. Gradient Descent) on the four models with different learning rates and starting points, and analyze the properties of these iterates in order to determine how well Theorem 5.3 generalizes to this nonconvex problem.

Experimental Procedure. We study four experimental factors: the model, the learning method, the initialization point, and the learning rate. The model factor has four levels as given by the four model objective functions shown in Fig. 5.12. The learning method has two levels which are SGD-1 or GD. The initialization point also has two levels: either the initialization point is selected randomly with a uniform probability from a disc of radius 10^{-8} centered about the circular basin minimizer, or the initialization point is selected randomly with a uniform probability from a disc of radius 10^{-8} centered about the quadratic basin minimizer. The learning rate has levels that are conditional on the initialization point. If the initialization point is near the minimizer of the circular basin, then the learning rate takes on values 10^{10}, $5(10^{10})$, 10^{11}, $5(10^{11})$, 10^{12}, and $5(10^{12})$. If the initialization point is near the minimizer of the circular basin, then the learning rate takes on values $1, 4, 16, 64, 256, 1024$.

For each unique collection of the levels of the four factors, one hundred independent runs are executed with at most twenty iterations. For each run, the euclidean distance between the iterates and the circular basin minimizer and the euclidean distance between the iterates and the quadratic basin minimizer are recorded.

Results and Discussion. Note, we will primarily focus on the divergence perspective because the convergence perspective is an almost identical discussion. Also note that the results of SGD-k on Models 1 and 2 and Models 3 and 4 are similar when initialized around the circular basin minimizer, and the results of SGD-k on Models 1 and 3 and Models 2 and 4 are similar when initialized around the quadratic basin minimizer. Hence, when we discuss the circular basin minimizer, we will compare Models 1 and 3, but we could have just as easily replaced Model 1 with Model 2 or Model 3 with Model 4 and the discussion would be identical. Similarly, when we discuss the quadratic basin minimizer, we will compare Models 1 and 2, but we could have replaced the results of Model 1 with Model 3 or Model 2 with Model 4 and the discussion would be identical. We now describe several observations and discuss how well Theorem 5.3 anticipates these experimental observations.

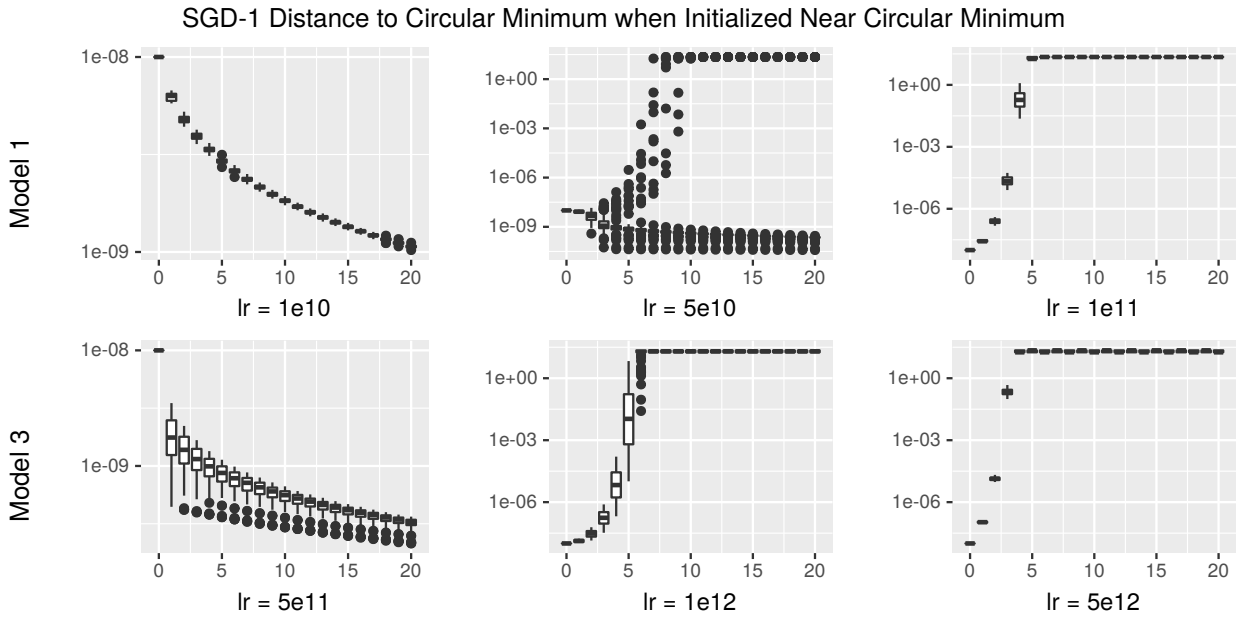


Figure 5.13: The behavior of SGD-1 on Models 1 and 3 when initialized near the circular minimum. The y-axis shows the distance (in logarithmic scale) between the estimates and the circular minimum for all runs of the specified model and the specified learning rate.

Fig. 5.13 shows the distance between the iterates and the circular basin minimizer for the one hundred independent runs of SGD-1 for the specified model and the specified learning rate when initialized near the circular basin minimizer. The learning rates that are displayed

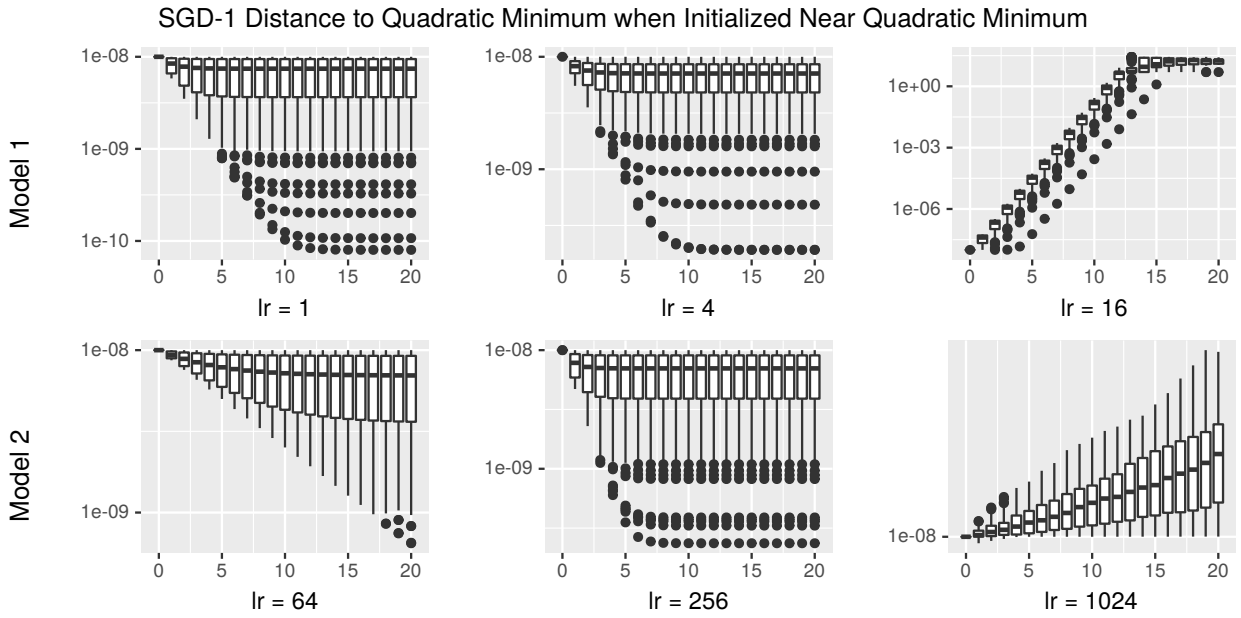


Figure 5.14: The behavior of SGD-1 on Models 1 and 2 when initialized near the quadratic minimum. The y-axis shows the distance (in logarithmic scale) between the iterates and the quadratic minimum for all runs of the specified model and the specified learning rate.

are the ones where a transition in the convergence-divergence behavior of the method occur for the specific model. Specifically, SGD-1 begins to diverge for learning rates between 5×10^{10} and 10^{11} for Model 1 and between 5×10^{11} and 10^{12} for Model 3. Similarly, Fig. 5.14 shows the distance between the iterates and the quadratic basin minimizer for the one hundred independent runs of SGD-1 for the specified model and the specified learning rate when initialized near the quadratic basin minimizer. SGD-1 begins to diverge for learning rates between 4 and 16 for Model 1 and between 256 and 1024 for Model 2.

From these observations we see that relatively flatter minimizers enjoy larger thresholds for divergence of SGD-1 in comparison to sharp minimizers. Moreover, while the bounds computed in Table 5.2, which are based on Theorem 5.3 are conservative, they are still rather informative, especially in the case of the quadratic basin. Finally, we also observe that the exponential divergence predicted by Theorem 5.3 holds when divergence occurs for both minimizers.

Fig. 5.15 shows the distance between the iterates and the circular basin minimizer for the

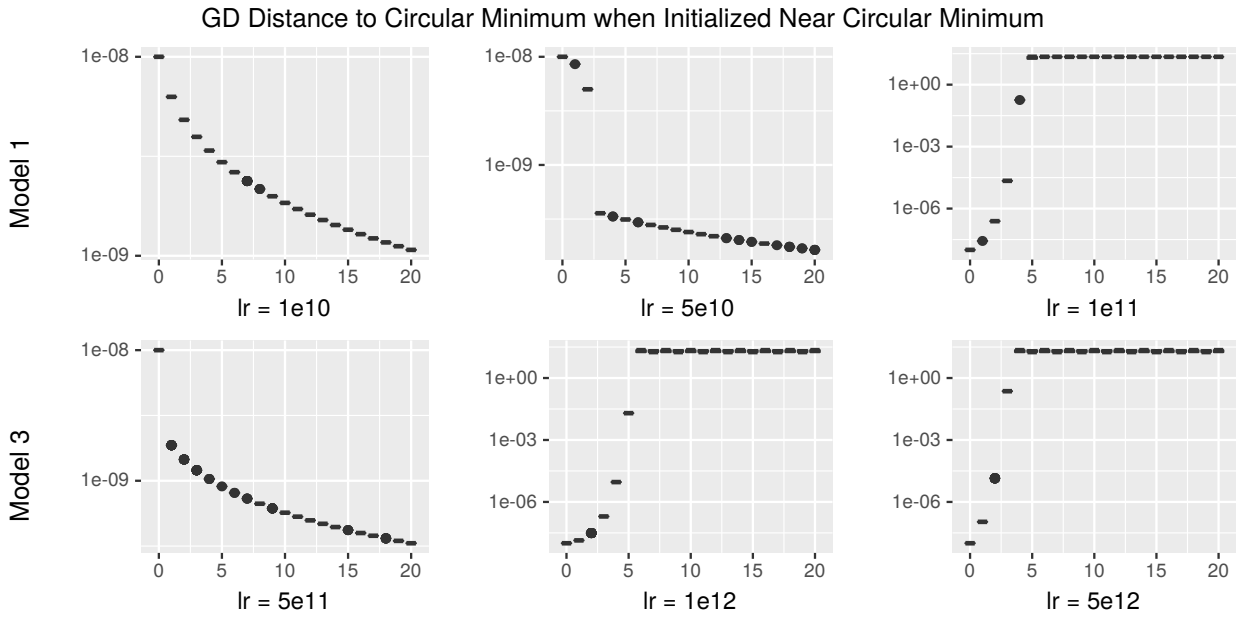


Figure 5.15: The behavior of GD on Models 1 and 3 when initialized near the circular basin minimizer. The y-axis shows the distance (in logarithmic scale) between the iterates and the circular minimizer for all runs of the specified model and the specified learning rate.

one hundred independent runs of GD for the specified model and the specified learning rate when initialized near the circular basin minimizer. If we compare Figs. 5.13 and 5.15, we notice that, for Model 1 and learning rate 5×10^{10}, the runs of GD converge whereas some of the runs for SGD-1 diverge. Although we do not report the results for all of the models or all of the learning rates, we note the boundary for divergence-convergence for GD are smaller than those of SGD-1. In light of Theorem 5.3, this behavior is expected: according to Theorem 5.3, as the batch-size, k, increases, the lower bound on the learning rates for divergence increases. Therefore, we should expect that at those boundary learning rates where some SGD-1 runs are stable and others diverge, for GD, we should only see stable runs, and, indeed, this is what the comparison of Figs. 5.13 and 5.15 shows.

Fig. 5.16 shows the distance between the iterates and the circular basin minimizer for the one hundred independent runs of SGD-1 and the one hundred independent runs of GD for the specified method and the specified learning rate when initialized near the quadratic basin minimizer. In Fig. 5.16, we see that for the learning rates that lead to divergence

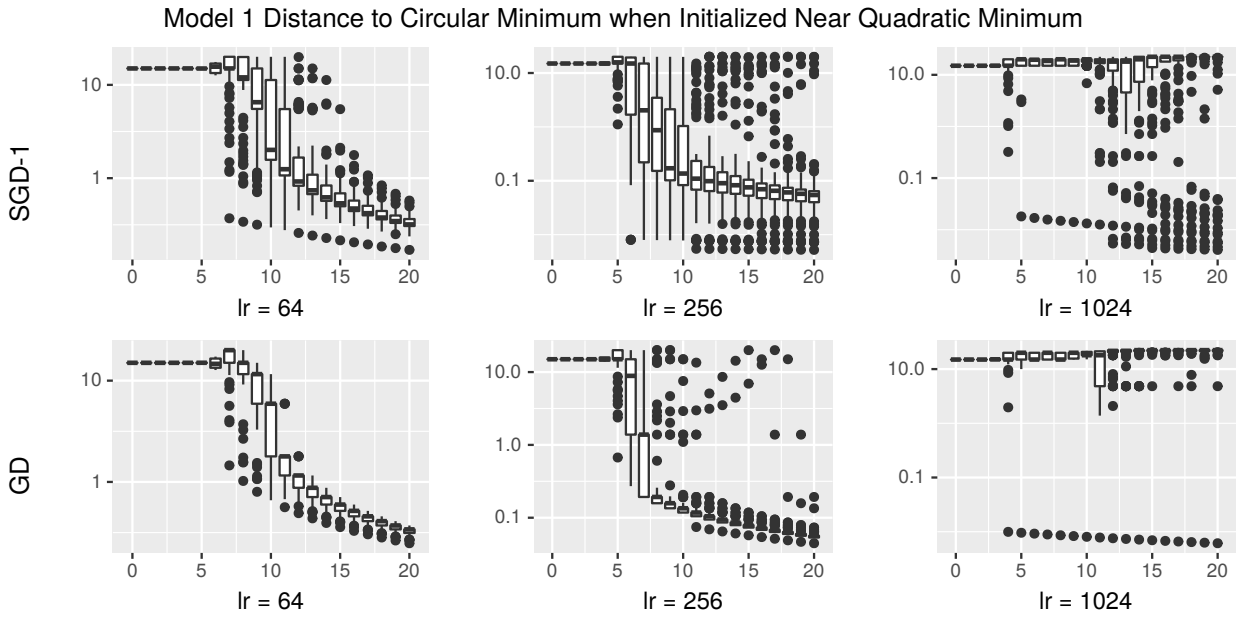


Figure 5.16: The behavior of SGD-1 and GD on Model 1 when initialized near the quadratic minimum for select learning rates. The y-axis shows the distance (in logarithmic scale) between the iterates and the circular minimizer for all runs of the specified method and the specified learning rate.

from the quadratic basin minimizer (compare to the top three subplots of Fig. 5.13) are in some cases able to converge to the circular minimizer. Again, in light of Theorem 5.3, this is expected: according to Theorem 5.3, the learning rates shown in Fig. 5.16 guarantee divergence from the quadratic minimizer and are sufficiently small that they can lead to convergence to the circular minimizer. However, we notice in Fig. 5.16 that as the learning rate increases, even though the learning rate is below the divergence bound for the circular minimizer, the iterates for SGD-1 and GD are diverging from both the circular and quadratic minima and are converging to the corners of the feasible region.

To summarize, we see that convergence-divergence behavior of SGD-k for the quadratic-circle sums nonconvex problem is captured by Theorem 5.3. In particular, the bounds estimated by Theorem 5.3 are able to predict several observed phenomenon. Specifically, Theorem 5.3 (1) predicts the bounds for when divergence from a minimizer will occur, and (2) captures that this divergence is exponential. In order to further elaborate our results,

we now consider a less trivial nonconvex optimization problem which has multiple minima, a more complex surface, and is much higher dimensional.

5.3.3 Styblinksi-Tang Problem

We now explore the modification of Styblinksi-Tang (ST) function (Styblinski and Tang, 1990), which was originally introduced in the neural network literature and served as a test function owing to its complex surface yet simple mathematical form. The p-dimensional ST function is given by

$$f(x) = \frac{1}{2}\sum_{i=1}^{p} c_{i,1}x_i^4 + c_{i,2}x_i^2 + c_{i,3}x_i, \tag{5.90}$$

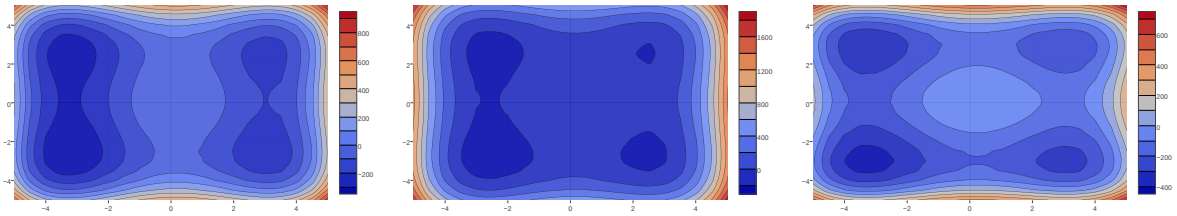
where $c_{i,1}, c_{i,3}$ are non-negative scalars, and $c_{i,2}$ is a non-positive scalar. Using this equation we can define the following problem, for which two dimensional examples are plotted in Fig. 5.17.

Problem 5.4 (Styblinski-Tang Sums)**.** *Let $f_1, \ldots, f_N$ be functions as specified by (5.90) of some dimension p. The Styblinksi-Tang sums objective function is*

$$\sum_{i=1}^{N} p_i f_i(x), \tag{5.91}$$

where p_i are positive valued and sum to one. Let Y be a random variable taking values $\{f_i : i = 1, \ldots, N\}$ with the probability of sampling $f_i = p_i$. Let $Y_1, Y_2, \ldots$ be independent copies of Y. The Styblinski-Tang Sums problem is to use $\{Y_1, Y_2, \ldots\}$ to minimize (5.91) restricted to $(-5, 5)^p \subset \mathbb{R}^p$.

While the analysis of Problem 5.4 may seem superfluous given the analysis of Problem 5.3, it is a fundamentally different problem because the component functions of Problem 5.4 do not have a common minimizer; that is, Problem 5.4 is a nonconvex problem with exclusively inhomogeneous minimizers. Therefore, in this case, the relevant theory is supplied by Theorem 5.4. However, our approach to studying SGD-k on Problem 5.4 will be similar to our



(a) (b) (c)

Figure 5.17: Contour plots of examples of (5.91) for dimension $p = 2$.

study of SGD-k on Problem 5.3 in the sense that we will use constant step sizes. Therefore, the use of constant step sizes will prevent SGD-k from converging to the minimizer, but we should expect stability around the minimizer. Importantly, by using constant step sizes, we remove the added variable of how to schedule the learning rate appropriately.

To study the stability and divergence properties of SGD-k on Problem 5.4, we consider three randomly generated realizations of the problem, which we label Model 1, Model 2, and Model 3. Model 1 is of dimension ten, has 2^{10} minima, and has $N = 200$ components; Model 2 is of dimension fifty, has 2^{50} minima, and has $N = 1000$ components; Model 3 is of dimension one hundred, has 2^{100} minima, and has $N = 2000$ components.

For each model, the flattest and sharpest minimizers are identified. For each minimizer, the phase boundary of $\frac{t_Q}{\lambda_1 \lambda_m}$ is estimated and reported in Table 5.4. Using these phase boundaries, estimates for the upper bounds on convergence and lower bounds on divergence for the learning rates, according to Theorem 5.4, are reported in Tables 5.5 and 5.6. In general, the flat minimizers have a higher tolerance with respect to divergence and convergence for larger learning rates in comparison to the sharp minimizers. However, this relationship is not guaranteed as we see in Table 5.6 for Model 3. Here, we see that the the sharpest minimizer has a slightly higher tolerance for larger learning rates. This is due to the t_Q term in the bounds, indicating that it cannot be neglected. We now look at how well our estimates based on our theory compare to numerical experiments of SGD-k for different values of k on the

three models.

Table 5.4: Estimates of the phase boundary for the Styblinski-Tang Sums problem for the flattest and sharpest minimizers for each of the three models.

Minimizer	Model 1	Model 2	Model 3
Flat	$5.096e+00$	$1.343e+02$	$3.920e+02$
Sharp	$3.260e+00$	$1.186e+02$	$2.668e+02$

Table 5.5: Estimates of (5.55) (with $\gamma = 0$) for Styblinski-Tang Sums Problem models.

	Model 1		Model 2		Model 3	
k	Flat	Sharp	Flat	Sharp	Flat	Sharp
1.0	$9.001e-03$	$7.682e-03$	$2.134e-02$	$1.951e-02$	$2.598e-02$	$1.871e-02$
100.0	$9.814e-02$	$8.133e-02$	$6.831e-01$	$6.117e-01$	$1.221e+00$	$8.561e-01$
200.0	$1.033e-01$	$8.546e-02$	$8.100e-01$	$7.225e-01$	$1.590e+00$	$1.106e+00$
350.0	$1.057e-01$	$8.737e-02$	$8.800e-01$	$7.833e-01$	$1.827e+00$	$1.264e+00$
500.0	$1.067e-01$	$8.815e-02$	$9.116e-01$	$8.106e-01$	$1.942e+00$	$1.341e+00$
∞	$1.091e-01$	$9.004e-02$	$9.947e-01$	$8.823e-01$	$2.279e+00$	$1.563e+00$

Table 5.6: Estimates of (5.54) for Styblinski-Tang Sums Problem models.

	Model 1		Model 2		Model 3	
k	Flat	Sharp	Flat	Sharp	Flat	Sharp
1.0	$2.319e-03$	$2.250e-03$	$8.603e-04$	$8.334e-04$	$7.587e-04$	$9.089e-04$
100.0	$4.268e-02$	$3.781e-02$	$7.925e-02$	$7.621e-02$	$7.345e-02$	$8.594e-02$
200.0	$4.632e-02$	$4.087e-02$	$1.174e-01$	$1.049e-01$	$1.423e-01$	$1.629e-01$
350.0	$4.808e-02$	$4.234e-02$	$1.329e-01$	$1.180e-01$	$2.379e-01$	$2.303e-01$
500.0	$4.882e-02$	$4.296e-02$	$1.403e-01$	$1.243e-01$	$2.812e-01$	$2.539e-01$
∞	$5.064e-02$	$4.447e-02$	$1.613e-01$	$1.418e-01$	$3.742e-01$	$3.337e-01$

Experimental Procedure. We study four experimental factors: the model, the batch size, the initialization point, and the learning rate. The model factor has three levels that correspond to the three models described above. The batch size will take the values corresponding to SGD-1, SGD-200, SGD-500, and Gradient Descent. The initialization point will be randomly selected from a uniform distribution on a ball whose radius will take values $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and is centered at either the sharpest minimizer or the flattest minimizer. The learning rates will be linear combinations of the upper and lower bounds. Specifically, if u is (5.54) and l is (5.55), then the learning rates will be $1.5l$, $0.5(u+l)$, or $0.5u$. For each unique collection of the levels of the four factors, one hundred independent runs are executed with at most twenty iterations. For each run, the euclidean distances between the iterates and the minimizer near the initialization point are recorded.

Results and Discussion We now summarize our experimental results and discuss how Theorem 5.4 explains the important aspects of our experimental results. Note, the results Models 1, 2 and 3 are nearly identical for the purposes of our discussion, and so we will feature Model 1 only in our discussion below.

Fig. 5.18 shows the distance between SGD-k iterates and the flat minimizer on Model 1 for different batch sizes and different learning rates when SGD-k is initialized near the flat minimizer. We note that, regardless of batch size, the iterates are diverging from the flat minimizer and are converging to the corners of the feasible region for learning rates $1.5l$ and $0.5(u+l)$. On the other hand, for the learning rate $0.5u$, we see stability for $k = 1, 200, 500$ about the minimizer and we see convergence for GD (i.e., $k = \infty$). Similarly, Fig. 5.19 shows the distance between SGD-k iterates and the sharp minimizer on Model 1 for different batch sizes and different learning rates when SGD-k is initialized near the sharp minimizer. We note that, regardless of batch size, the iterates are diverging from the sharp minimizer and converging to the corners of the feasible region for learning rates $1.5l$ and $0.5(u+l)$. On the other hand, for the learning rate $0.5u$, we see stability for $k = 1, 200, 500$ about the sharp minimizer and we see convergence for GD (i.e., $k = \infty$).

Taking the results in Figs. 5.18 and 5.19 together, we see that we are able to use Theorem 5.4 to find learning rates that either ensure divergence from or stability about a minimizer if we know its local geometric properties. Consequently, we again have evidence that our deterministic mechanism can correctly predict the behavior of SGD-k for nonconvex problems. Moreover, when divergence does occur, Figs. 5.18 and 5.19 also display exponential divergence as predicted by Theorem 5.4.

For a different perspective, Fig. 5.20 shows the distance between SGD-k iterates and the flat minimizer on Model 1 for different starting radii and different learning rates when SGD-k is initialized near the flat minimizer. We note that, regardless of the starting radius, the iterate are diverging from the flat minimizer and are converging to the corners of the feasible

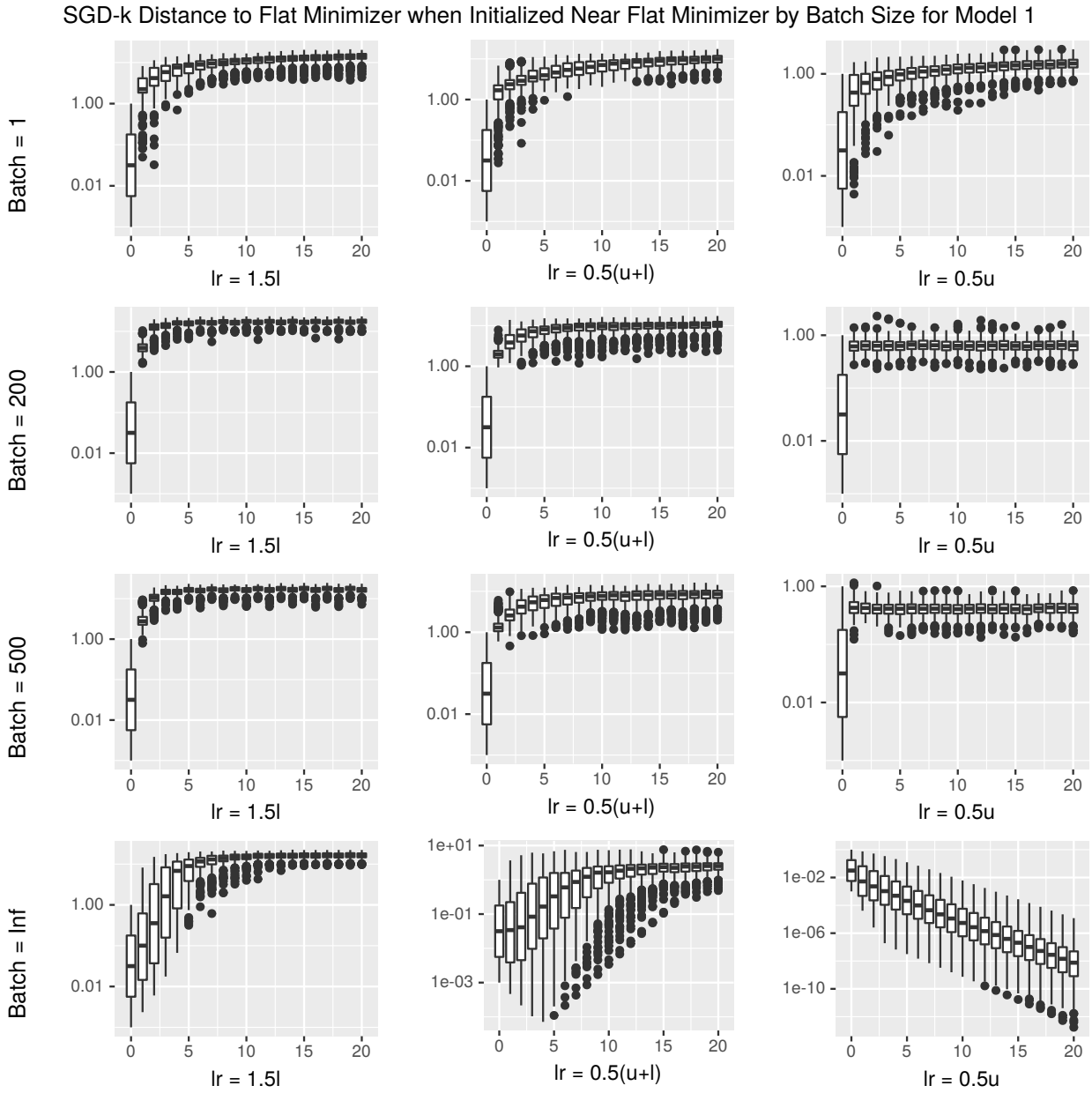


Figure 5.18: The behavior of SGD-k on Model 1 for different batch sizes when initialized near the flat minimizer. The y-axis shows the distance (in logarithmic scale) between the iterates and the flat minimizer for all runs of the specified batch size and specified learning rate.

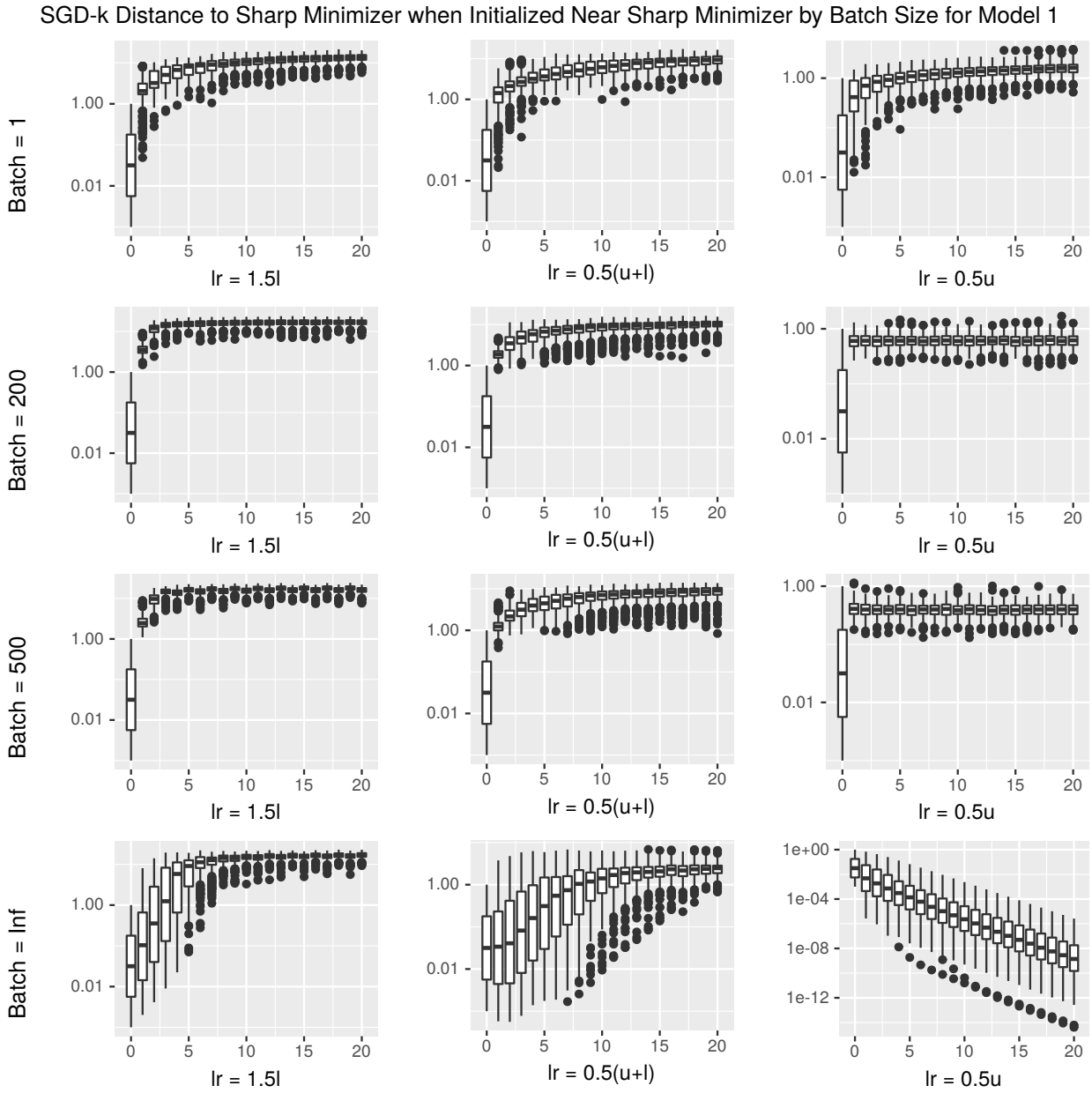


Figure 5.19: The behavior of SGD-k on Model 1 for different batch sizes when initialized near the sharp minimizer. The y-axis shows the distance (in logarithmic scale) between the iterates and the sharp minimizer for all runs of the specified batch size and specified learning rate.

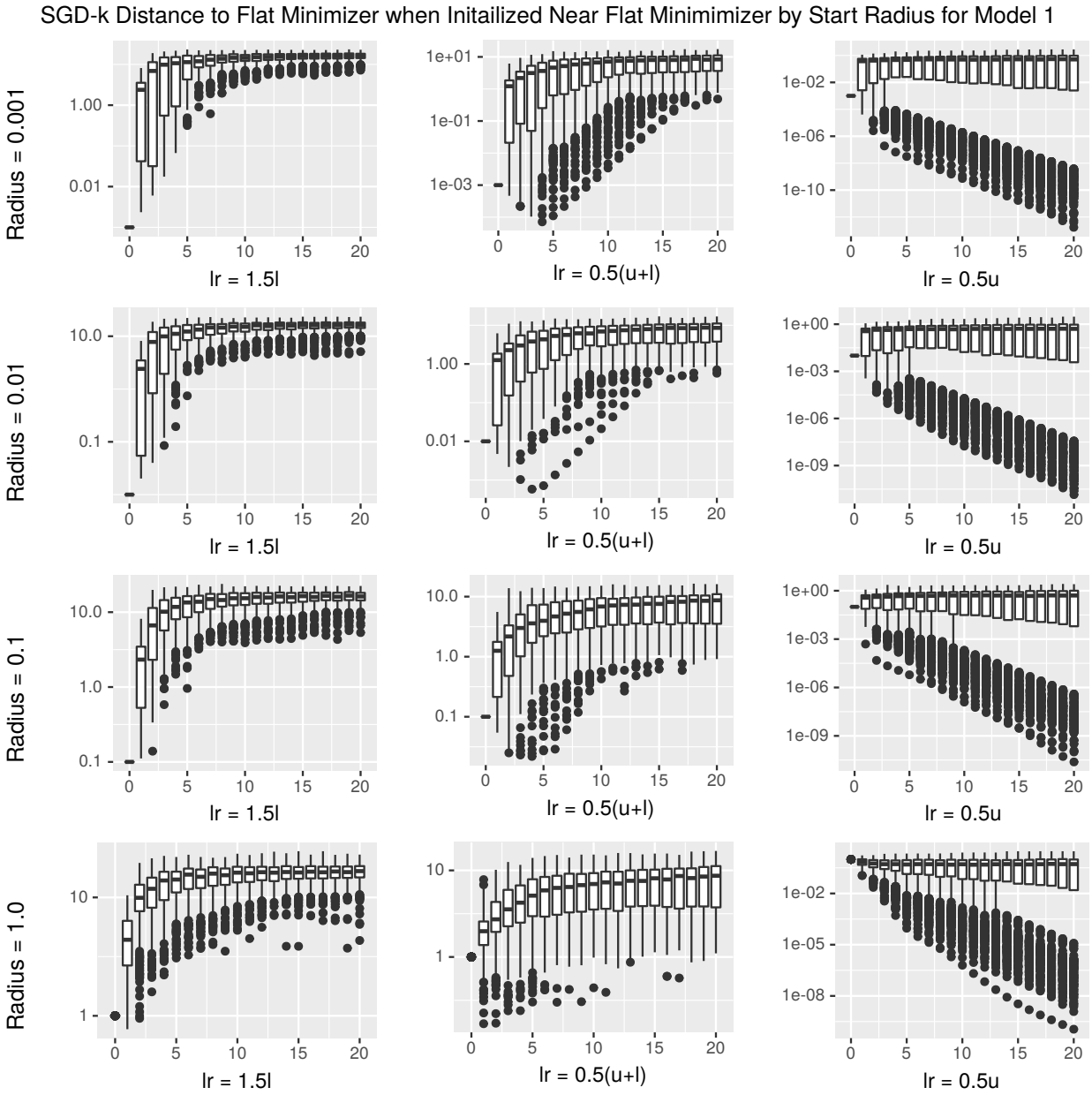


Figure 5.20: The behavior of SGD-k on Model 1 for different starting radii when initialized near the flat minimizer. The y-axis shows the distance (in logarithmic scale) between the iterates and the flat minimizer for all runs of the specified starting radius and specified learning rate.

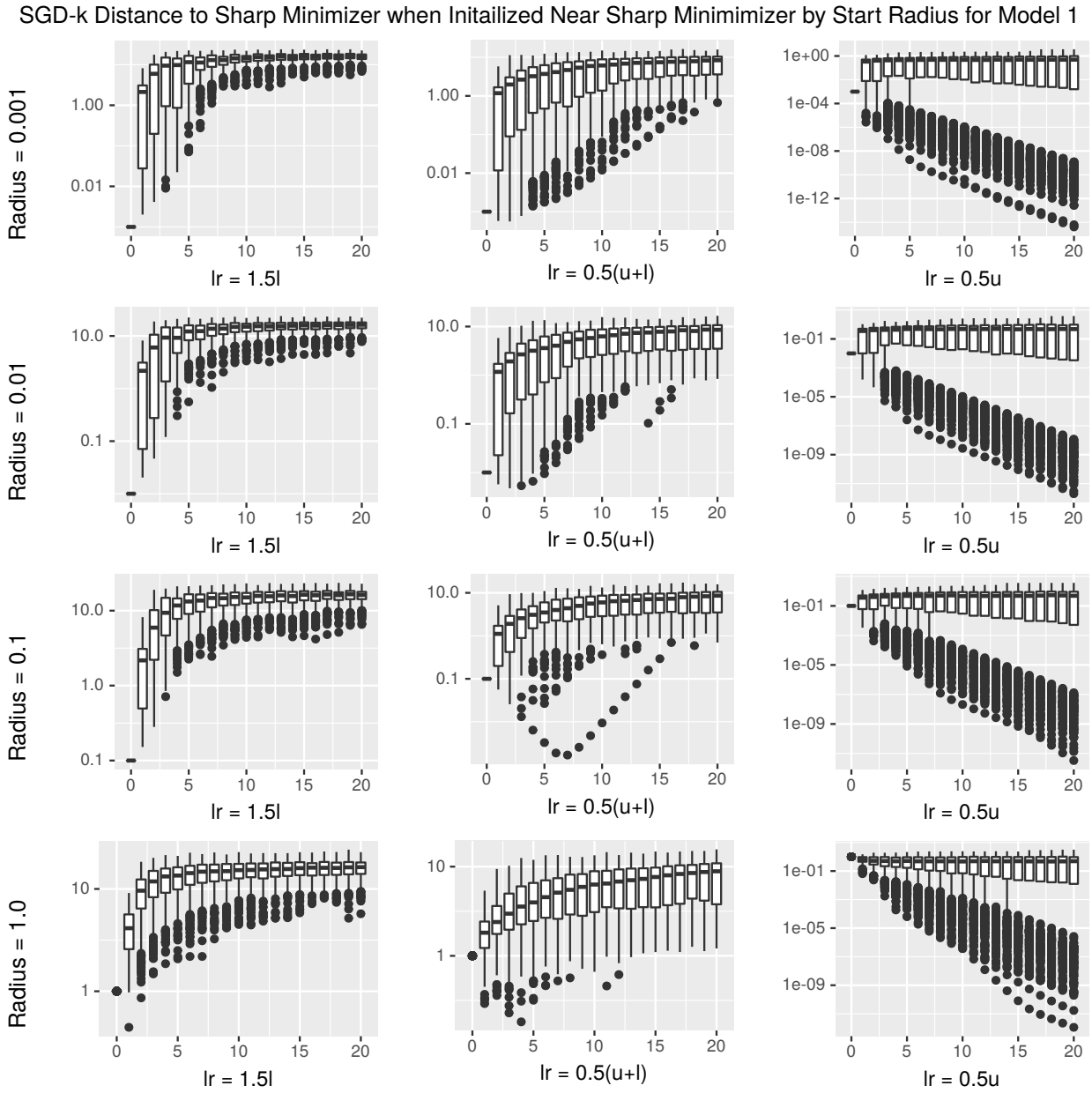


Figure 5.21: The behavior of SGD-k on Model 1 for different starting radii when initialized near the sharp minimizer. The y-axis shows the distance (in logarithmic scale) between the iterates and the sharp minimizer for all runs of the specified starting radius and specified learning rate.

region for learning rates $1.5l$ and $0.5(u+l)$. On the other hand, for the learning rate $0.5u$, we see stability and even convergence for all of the runs regardless of the starting radius. Similarly, Fig. 5.21 shows the distance between SGD-k iterates and the sharp minimizer on Model 1 for different starting radii and different learning rates when SGD-k is initialized near the sharp minimizer. We note that, regardless of the starting radius, the iterates are diverging from the sharp minimizer and are converging to the corners of the feasible region for learning rates $1.5l$ and $0.5(u+l)$. On the other hand, for learning rate $0.5u$, we see stability about and even convergence to the sharp minimizer.

Taking the results in Figs. 5.20 and 5.21 together, we see that we are able to use Theorem 5.4 to find learning rates that either ensure divergence from or stability about a minimizer if we know its local geometric properties. Consequently, we again have evidence that our deterministic mechanism can correctly predict the behavior of SGD-k for nonconvex problems. Moreover, when divergence does occur, Figs. 5.18 and 5.19 also display exponential divergence as predicted by Theorem 5.4.

5.3.4 Summary of Results

To summarize, we wanted to explore the mechanism by which SGD is captured by minimizers with distinct geometric properties. We discussed the widely-accepted stochastic mechanism for this observed phenomenon, and we pointed out some intuitive challenges with this mechanism. We then proposed a deterministic mechanism for how SGD escapes from different minimizers, which was based on the properties of classical gradient descent. In the subsequent two sections, we showed how this mechanism operated on the linear regression problem and a general quadratic problem. In particular, we showed that the the behavior of gradient descent is a special case of our deterministic mechanism, and we highlighted how our deterministic mechanism predicts exponential divergence when appropriate. We then used our quadratic analysis as a local approximation for two nonconvex problems, and showed that our mechanism and its prediction of exponential divergence were valid for these noncon-

vex problems, which further supports our deterministic mechanism over the widely-accepted stochastic mechanism.

6 | Kalman-based Stochastic Gradient Descent

In the previous chapter, we saw that stochastic gradient descent's convergence and divergence properties are rather sensitive to the local geometry of the estimation problem. Consequently, if SGD's learning rate is selected incorrectly, we can observe phenomenon such as stagnation or catastrophic divergence. One way to avoid these issues is to reduce SGD's sensitivity to the local geometry of the estimation problem, which naturally leads to incremental estimators with higher-order approximations of the variance and more complex local models.

Unfortunately, making use of such local geometric information is rather challenging, especially without oracle knowledge of the problem. However, for estimation problems in which the local geometric information is highly structured, we can design incremental estimators that perform better than incremental estimators that do not use such information. In particular, here we consider regression problems that have a quasi-likelihood structure (Wedderburn, 1974), which include linear regression, nonparametric regression, generalized linear regression, and many deep neural networks learning problems as special cases.

A regression problem with quasi-likelihood structure is characterized by the gradient of the estimation function having the form

$$m_\psi(y, \theta) = -\frac{z - \mu(x, \theta)}{V(x, \theta)} \mu_\psi(x, \theta), \tag{6.1}$$

where $y = (z, x) \in \mathbb{R} \times \mathbb{R}^d$; $\mu : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is called the mean function; $\mu_\psi(x, \theta) = \nabla_\psi \mu(x, \psi)|_{\psi=\theta}$; and $V : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ is called the variance function. Importantly, if $\exists \theta^* \in \Theta$ such that $Y = (Z, X)$ is a random variable, $\mathbb{E}[Z|X] = \mu(x, \theta^*)$ and $\mathbb{V}[Z|X] = V(X, \theta^*)$,

Disclaimer: Parts of this work are adapted from or reproduced from published work (see Patel, 2016).

then under some mild regularity conditions on μ and V,

$$\mathbb{E}\left[m_\psi(Y,\theta^*)\middle|\, X\right] = 0, \tag{6.2}$$

and

$$\mathbb{E}\left[m_\psi(Y,\theta^*)m_\psi(Y,\theta)'\middle|\, X\right] = \mathbb{E}\left[m_{\psi\psi}(Y,\theta)\middle|\, X\right] = \frac{1}{V(X,\theta^*)}\mu_\psi(X,\theta^*)\mu_\psi(X,\theta^*)'. \tag{6.3}$$

For regression problems with this quasi-likelihood structure, we can take advantage of the second-order information by using the second-order, regression-like model incremental estimator specified in Algorithm 2. We refer to this incremental estimator as Kalman-based Stochastic Gradient Descent (Patel, 2016). Here, our main effort is to prove the consistency of this estimator in the linear regression setting.

Algorithm 2: Kalman-based Stochastic Gradient Descent

Data: Parameter θ_0, Positive Hyper-parameters γ_1, γ_2, Mean Function μ, Variance Function V
Result: Parameter θ

$\theta \leftarrow \theta_0$
$M \leftarrow p \times p$ identity matrix
while *true* **do**
  Read new observation, $Y = (Z, X)$
  $g \leftarrow \mu_\psi(X, \theta)$
  $m \leftarrow \mu(X, \theta)$
  $v \leftarrow Mg$
  $s \leftarrow \min\{\gamma_1, \max\{\gamma_2, V(X,\theta)\}\} + v'g$
  $\theta \leftarrow \theta + \frac{Z-m}{s}v$
  $M \leftarrow M - \frac{1}{s}vv'.$
end

In Section 6.1, we will begin by defining the linear regression problem that we will use throughout the remainder of this chapter. In Section 6.2, we begin by placing our incremental estimator within the literature on Kalman Filtering (Kalman, 1960), Natural Gradient

Descent (Amari and Cardoso, 1997; Amari et al., 1997) and Recursive Least Squares (Gauss, 1809). In Section 6.3, we derive the convergence behavior of kSGD. In Section 6.4, we numerically compare kSGD against other methodologies. Finally, in Section 6.5, we make some comments on using kSGD in the low-memory environment.

6.1 A Linear Regression Problem

In order to formulate the linear regression problem, the linear model must be specified:

Assumption 6.1. *Suppose that* $Y = (Z, X), Y_1 = (Z_1, X_1), Y_2 = (Z_2, X_2), \ldots \in \mathbb{R} \times \mathbb{R}^p$ *are independent, identically distributed, and* $\exists \theta^* \in \mathbb{R}^p$ *such that:*

$$Z_i = X_i' \theta^* + \epsilon_i, \tag{6.4}$$

where ϵ_i *are independent, identically distributed mean zero random variables with variance* $\sigma^2 \in (0, \infty)$*, and are independent of all* $\{X_j : j \in \mathbb{N}\}$*.*

Remark 6.1. *The model does not assume a distribution for the errors,* ϵ_i*; hence, the results presented will hold even if the model is misspecified with a reinterpretation of* σ^2 *as the limiting mean residuals squared. In addition, if the model has heteroscedasticity, the convergence of the kSGD parameter estimate to* θ^* *will still hold in the results below as long as the supremum over all variances is bounded.*

Informally, the linear regression problem is the task of determining θ^* from the data $Y_1, Y_2, \ldots$. To formalize this, the linear regression problem can be restated as minimizing an estimating equation over the data, which we will define as

$$m(Y, \theta) = \frac{1}{2} \left(Z - \theta' X \right)^2 \tag{6.5}$$

up to a positive multiplicative constant and conditioned on X. The linear regression objective

function is taken to be the expected value of $m(Y, \theta)$:

$$R(\theta) = \mathbb{E}\left[m(Y, \theta)\right] = R(\theta^*) + \frac{1}{2}(\theta - \theta^*)' Q_* (\theta - \theta^*), \tag{6.6}$$

where $Q_* = \mathbb{E}\left[XX'\right]$.

Thus, the linear regression problem is the task of minimizing $R(\theta)$. However, on its own, the linear regression problem is ill-posed for several reasons. First, the linear regression objective's Hessian, Q_*, may not be well-specified, that is, $Q_* \not\prec \infty$. One way to ensure that $Q_* \prec \infty$ is to require that

$$\lambda_{\max}(Q_*) \leq \mathbf{tr}\left[Q_*\right] = \mathbb{E}\left[\mathbf{tr}\left[XX'\right]\right] = \mathbb{E}\left[\|X\|_2^2\right] < \infty. \tag{6.7}$$

This requirement is collected in the next assumption:

Assumption 6.2. *$X \in L^2$. That is, $\mathbb{E}\left[\|X\|_2^2\right] < \infty$.*

For some results, Assumption 6.2 is strengthened by:

Assumption 6.3. *$X \in L^\infty$. That is, $\|X\|_\infty < \infty$ almost surely.*

Second, θ^* may not be identifiable. To ensure that θ^* is identifiable, we will require that $Q_* \succ 0$ by using the next assumption.

Assumption 6.4. *The linear span of the image of X is $\mathbb{R}^n$. Specifically, for all unit vectors $v \in \mathbb{R}^n$, $\mathbb{P}\left[|X'v| = 0\right] < 1$.*

To see how this assumption implies that $Q_* \succ 0$, suppose there is a unit vector $v \in \mathbb{R}^n$ such that $0 \geq v' Q_* v$. Then $0 \geq v' Q_* v = \mathbb{E}\left[(X'v)^2\right]$. Hence, $X'v = 0$ almost surely, which contradicts Assumption 6.4.

With either Assumptions 6.1, 6.2 and 6.4 or Assumptions 6.1, 6.3 and 6.4, we have a well-specified linear regression problem. We now turn our attention to using the linear regression

problem to elucidate the relationship between kSGD and techniques in statistical filtering and machine learning.

6.2 Related Methodologies

Kalman-based Stochastic Gradient Descent is related to a number of methodologies in statistical filtering and machine learning. For example, kSGD is closely related to the Kalman Filter first proposed by Kalman (1960). Specifically, kSGD can be derived by computing the optimal Kalman Gain, just as is done to derive the Kalman Filter. We demonstrate this relationship in detail.

Let $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the σ-algebra of the random variables $X_1, \ldots, X_k$, and consider the following general update scheme

$$\theta_{k+1} = \theta_k + G_{k+1}(Z_{k+1} - \theta_k' X_{k+1}), \tag{6.8}$$

where G_{k+1} is a random variable in $\mathbb{R}^p$ and is measurable with respect to $\mathcal{F}_{k+1}$.

Remark 6.2. *We note that since Z_{k+1} is not measurable with respect to $\mathcal{F}_{k+1}$, then θ_{k+1} is also not measurable with respect to $\mathcal{F}_{k+1}$. Therefore, in order to simplify the calculations below, we construct G_{k+1} without information from $Z_1, \ldots, Z_k$ and $\theta_1, \ldots, \theta_k$.*

Using Assumption 6.1, (6.8) can be rewritten as

$$\theta_{k+1} = \theta_k - G_{k+1} X_{k+1}'(\theta_k - \theta^*) + G_{k+1}\epsilon_{k+1}. \tag{6.9}$$

We will choose an optimal G_{k+1} in the sense that it minimizes the l^2 error between θ_{k+1} and θ^* given $\mathcal{F}_{k+1}$. Noting that G_{k+1} is measurable with respect to $\mathcal{F}_{k+1}$, and using the

independence, first moment and second moment properties of ϵ_k,

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_{k+1}-\theta^*\right\|^2\middle|\mathcal{F}_{k+1}\right] &= \mathbf{tr}\left[\mathbb{E}\left[(\theta_k-\theta^*)(\theta_k-\theta^*)'\middle|\mathcal{F}_k\right]\right] \\ &- \mathbf{tr}\left[G_{k+1}X_{k+1}'\mathbb{E}\left[(\theta_k-\theta^*)(\theta_k-\theta^*)'\middle|\mathcal{F}_k\right]\right] \\ &- \mathbf{tr}\left[\mathbb{E}\left[(\theta_k-\theta^*)(\theta_k-\theta^*)'\middle|\mathcal{F}_k\right]X_{k+1}G_{k+1}'\right] \\ &+ \mathbf{tr}\left[G_{k+1}X_{k+1}'\mathbb{E}\left[(\theta_k-\theta^*)(\theta_k-\theta^*)'\middle|\mathcal{F}_k\right]X_{k+1}G_{k+1}'\right] + \sigma^2\mathbf{tr}\left[G_{k+1}G_{k+1}'\right].\end{aligned} \tag{6.10}$$

We now write $\mathcal{M}_k = \mathbb{E}\left[(\theta_k-\theta^*)(\theta_k-\theta^*)'\middle|\mathcal{F}_k\right]$, which gives us that

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_{k+1}-\theta^*\right\|^2\middle|\mathcal{F}_{k+1}\right] &= \mathbf{tr}\left[\mathcal{M}_k\right] - \mathbf{tr}\left[G_{k+1}X_{k+1}'\mathcal{M}_k\right] - \mathbf{tr}\left[\mathcal{M}_kX_{k+1}G_{k+1}'\right] \\ &+ \mathbf{tr}\left[G_{k+1}X_{k+1}'\mathcal{M}_kX_{k+1}G_{k+1}'\right] + \sigma^2\mathbf{tr}\left[G_{k+1}G_{k+1}'\right].\end{aligned} \tag{6.11}$$

Differentiating with respect to G_{k+1} and solving for G_{k+1} when this expression is set to zero, we have that

$$G_{k+1} = \frac{\mathcal{M}_kX_{k+1}}{\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1}}. \tag{6.12}$$

Moreover, the derivation gives us an update scheme for $\mathcal{M}_k$ as well, which is

$$\begin{aligned}\mathcal{M}_{k+1} &= \mathcal{M}_k - G_{k+1}X_{k+1}'\mathcal{M}_k - \mathcal{M}_kX_{k+1}G_{k+1}' + (\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1})G_{k+1}G_{k+1}' \\ &= \mathcal{M}_k - 2\frac{\mathcal{M}_kX_{k+1}X_{k+1}'\mathcal{M}_k}{\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1}} \\ &\quad + \frac{(\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1})\mathcal{M}_kX_{k+1}X_{k+1}'\mathcal{M}_k}{(\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1})^2}.\end{aligned} \tag{6.13}$$

To summarize, we have the optimal stochastic gradient update scheme is given by

$$\theta_{k+1} = \theta_k + \frac{\mathcal{M}_kX_{k+1}}{\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1}}(Z_{k+1} - \theta_k'X_{k+1}), \tag{6.14}$$

where

$$\mathcal{M}_{k+1} = \left(I - \frac{\mathcal{M}_kX_{k+1}X_{k+1}'}{\sigma^2 + X_{k+1}'\mathcal{M}_kX_{k+1}}\right)\mathcal{M}_k. \tag{6.15}$$

Because σ^2 and $\mathcal{M}_k$ are not known, we take the kSGD method to be given by

$$\theta_k = \theta_{k-1} + \frac{M_{k-1}X_k}{\gamma_k^2 + X_k' M_{k-1} X_k}\left(Z_k - \theta_{k-1}' X_k\right), \tag{6.16}$$

where

$$M_k = \left(I - \frac{M_{k-1}X_k X_k'}{\gamma_k^2 + X_k' M_{k-1} X_k}\right) M_{k-1}, \tag{6.17}$$

where $\theta_0 \in \mathbb{R}^n$ is arbitrary; and M_0 can be any positive definite matrix, but for simplicity, we will take it to be the identity. Moreover, we have that the method described by (6.16) and (6.17) is a special case of Algorithm 2. For future reference, the sequence $\{\gamma_k^2\}$ replaces the unknown σ^2 value and will be referred to as tuning parameters. We refer to θ_k as a parameter estimate, $\mathcal{M}_k$ as the true covariance of the parameter estimate, and M_k as the estimated covariance of the parameter estimate.

While kSGD is derived in a manner similar to the Kalman Filter for the linear regression problem, it diverges in an important way: kSGD is an estimator for a parameter, whereas the Kalman Filter is an estimate for a complete distribution. For this reason, kSGD can be applied more broadly to regression problems with quasi-likelihood structure, whereas the Kalman Filter requires assumptions about linearity and normality in order to remain valid.

Another closely related method is natural gradient descent (Amari and Cardoso, 1997; Amari et al., 1997), which has an algebraic structure that is equivalent to the Kalman Filter.[1] The important distinction between kSGD and Natural Gradient Descent is in how M_k is updated. In particular, kSGD allows M_k to decay to zero naturally, whereas Natural Gradient Descent introduces hyperparameters to prevent the covariance estimate from converging. Another important distinction is in rigor: below, we show that kSGD, under the linear regression problem above, is actually insensitive to the conditioning of the problem, which was only heuristically derived for natural gradient descent.

[1] This is based on recent work by Yann Ollivier.

Finally, at least for linear regression, kSGD and the recursive least squares method are nearly identical, except that the recursive least squares method employs a so-called forgetting factor. A consequence of this forgetting factor is that the covariance of the parameter estimate never converges to zero, which, consequently, implies that the parameter estimate never converges. If the noise is ignored (reducing the problem to solving a linear system), then several works have shown that recursive least squares converges R-linearly (i.e., exponentially) to the true parameter (Johnstone et al., 1982; Bittanti et al., 1990; Parkum et al., 1992; Cao and M Schwartz, 2003). However, we can actually show that in the noise-free case, kSGD will converge in an infinite number of iterations with exponentially decaying probability.[2] Thus, the interesting case is the one with noise, and was not carefully treated until our work, which we reproduce presently.[3]

6.3 Convergence Analysis

In the analysis below, we use the following conventions and notation. Recalling that $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, we consider two types of error in our analysis:

$$e_k = \mathbb{E}\left[\theta_k | \mathcal{F}_k\right] - \theta^* \tag{6.18}$$

and

$$E_k = \theta_k - \theta^* \tag{6.19}$$

which are related by $e_k = \mathbb{E}\left[E_k | \mathcal{F}_k\right]$.

6.3.1 Convergence of the Estimated Covariance and Estimated Parameter

Informally, the main result of this section, Theorem 6.1, states that M_k bounds $\mathcal{M}_k$ from above and below arbitrarily well in the limit. To establish Theorem 6.1, we will need two

[2] This is not a hard proof, and it is shown in an earlier version of Patel (2016) on the arXiv.

[3] It is worth mentioning the work of Bertsekas (1996), in which he applies the Kalman Filter for nonlinear regression. However, in this work, he is in a completely deterministic setting as the data is assumed to be fixed and finite.

basic calculations collected in Lemmas 6.1 and 6.2. In the calculations below, we recall that M_0 is taken to be the identity for simplicity.

Lemma 6.1. *If for $j = 1, \ldots, k+1$, $0 < \gamma_j^2 < \infty$ then M_{k+1} is a symmetric, positive definite matrix and*

$$M_{k+1}^{-1} = M_k^{-1} + \frac{1}{\gamma_{k+1}^2} X_{k+1} X_{k+1}'. \tag{6.20}$$

Proof. By the Sherman-Morrison-Woodbury matrix identity:

$$M_1 = I - \frac{X_1 X_1'}{\gamma_1^2 + X_1' X_1} = \left(I + \frac{1}{\gamma_1^2} X_1 X_1' \right)^{-1}. \tag{6.21}$$

Hence, $M_1^{-1} = I + \frac{1}{\gamma_1^2} X_1 X_1'$. So M_1 is symmetric and positive definite. Suppose this is true up to some k. By induction and using the Sherman-Morrison-Woodbury matrix identity, we conclude that

$$M_{k+1} = M_k - \frac{M_k X_{k+1} X_{k+1}' M_k}{\gamma_{k+1}^2 + X_{k+1}' M_k X_{k+1}} = \left(M_k^{-1} + \frac{1}{\gamma_{k+1}^2} X_{k+1} X_{k+1}' \right)^{-1}. \tag{6.22}$$

■

Lemma 6.2. *If for $j = 1, \ldots, k+1$, $0 < \gamma_j^2 < \infty$ then*

$$M_{k+1}^{-1} E_{k+1} = M_k^{-1} E_k + X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2} \quad \text{and} \quad M_{k+1}^{-1} E_{k+1} = E_0 + \sum_{j=1}^{k+1} X_j \frac{\epsilon_j}{\gamma_j^2}. \tag{6.23}$$

and, recalling $\mathcal{M}_k = \mathbb{E}\left[E_k E_k' \middle| \mathcal{F}_k \right]$ where $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$,

$$\mathcal{M}_{k+1} = M_{k+1} E_0 E_0' M_{k+1} + M_{k+1} \left(\sum_{j=1}^{k+1} \frac{\sigma^2}{\gamma_j^2} \frac{1}{\gamma_j^2} X_j X_j' \right) M_{k+1}. \tag{6.24}$$

Proof. Using (6.16) and Assumption 6.1:

$$E_{k+1} = \left(I - \frac{M_k X_{k+1} X'_{k+1}}{\gamma_{k+1}^2 + X'_{k+1} M_k X_{k+1}} \right) E_k + M_k X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2 + X'_{k+1} M_k X_{k+1}}. \tag{6.25}$$

Now, premultiplying E_k by $M_k M_k^{-1}$ and using (6.17),

$$\begin{aligned} E_{k+1} &= M_{k+1} M_k^{-1} E_k + M_k X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2 + X'_{k+1} M_k X_{k+1}} \\ &= M_{k+1} \left(M_k^{-1} E_k + M_{k+1}^{-1} M_k X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2 + X'_{k+1} M_k X_{k+1}} \right). \end{aligned} \tag{6.26}$$

By applying the recurrence relation in Lemma 6.1,

$$\begin{aligned} E_{k+1} &= M_{k+1} \left(M_k^{-1} E_k + X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2 + X'_{k+1} M_k X_{k+1}} + X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2} \frac{X'_{k+1} M_k X_{k+1}}{\gamma_{k+1}^2 + X'_{k+1} M_k X_{k+1}} \right) \\ &= M_{k+1} \left(M_k^{-1} E_k + X_{k+1} \frac{\epsilon_{k+1}}{\gamma_{k+1}^2} \right). \end{aligned} \tag{6.27}$$

Using the recursion, we have that $E_{k+1} = M_{k+1} \left(E_0 + \sum_{j=1}^{k+1} X_j \frac{\epsilon_j}{\gamma_j^2} \right)$. Thus, calculating $E_{k+1} E'_{k+1}$, taking its conditional expectation with respect to $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, and recalling that $\{\epsilon_j\}$ are independent of $\{X_j\}$ by Assumption 6.1, we establish that

$$\mathcal{M}_{k+1} = M_{k+1} E_0 E'_0 M_{k+1} + M_{k+1} \left(\sum_{j=1}^{k+1} \frac{\sigma^2}{\gamma_j^2} \frac{1}{\gamma_j^2} X_j X'_j \right) M_{k+1}. \tag{6.28}$$

■

Lemmas 6.1 and 6.2 suggest a natural condition on the tuning parameters in order to ensure that these calculations hold for all k; namely, we require for some δ^2 and Δ^2,

$$0 < \delta^2 \leq \inf_k \gamma_k^2 \leq \sup_k \gamma_k^2 \leq \Delta^2 < \infty. \tag{6.29}$$

Using this condition and Lemmas 6.1 and 6.2, we can also bound $\mathcal{M}_k$ by M_k for all $k \in \mathbb{N}$

as

$$\begin{aligned} M_{k+1}E_0E_0'M_{k+1} + \frac{\sigma^2}{\Delta^2}M_{k+1}\left(\sum_{j=1}^{k+1}\frac{1}{\gamma_j^2}X_jX_j'\right)M_{k+1} &\preceq \mathcal{M}_{k+1} \\ \preceq M_{k+1}E_0E_0'M_{k+1} + \frac{\sigma^2}{\delta^2}M_{k+1}\left(\sum_{j=1}^{k+1}\frac{1}{\gamma_j^2}X_jX_j'\right)M_{k+1}&. \end{aligned} \tag{6.30}$$

Using Lemma 6.1,

$$\begin{aligned} M_{k+1}E_0E_0'M_{k+1} + \frac{\sigma^2}{\Delta^2}M_{k+1}\left(M_{k+1}^{-1} - I\right)M_{k+1} &\preceq \mathcal{M}_{k+1} \\ \preceq M_{k+1}E_0E_0'M_{k+1} + \frac{\sigma^2}{\delta^2}M_{k+1}\left(M_{k+1}^{-1} - I\right)M_{k+1}&. \end{aligned} \tag{6.31}$$

Using $0 \preceq M_{k+1}E_0E_0'M_{k+1} \preceq M_{k+1}^2\left\|E_0\right\|_2^2$, we conclude that

$$\frac{\sigma^2}{\Delta^2}M_{k+1} - \frac{\sigma^2}{\Delta^2}M_{k+1}^2 \preceq \mathcal{M}_{k+1} \preceq \left\|E_0\right\|_2^2 M_{k+1}^2 + \frac{\sigma^2}{\delta^2}M_{k+1}. \tag{6.32}$$

Because the true covariance, $\mathcal{M}_k$, is a measure of the error between the estimated, θ_k, and true parameter, θ^*, then we want $\mathcal{M}_k \to 0$, which inequality (6.32) suggests occurs if $M_k \to 0$. The next theorem formalizes this claim.

Theorem 6.1. *If Assumptions 6.1, 6.2 and 6.4 hold, and the tuning parameters satisfy (6.29), then $M_k \to 0$ almost surely. Moreover, for all $\epsilon > 0$, almost surely there exists a $K \in \mathbb{N}$ such that for $k \geq K$,*

$$\frac{1-\epsilon}{\Delta^2}M_k \preceq \frac{1}{\sigma^2}\mathcal{M}_k \preceq \frac{1+\epsilon}{\delta^2}M_k. \tag{6.33}$$

Remark 6.3. *There is a conventional difference between "almost surely there exists a K" and "there exists a K almost surely." The first statement implies that for each outcome, ω, on a probability one set there is a $K(\omega)$. The latter statement implies $\exists K$ for all ω. Thus, "almost surely there exists a K" is a weaker result than "there exists a K almost surely," but it is stronger than convergence in probability.*

Proof. To show that $M_k \to 0$, it is equivalent to prove that $\lambda_{\max}(M_k)$, the maximum eigenvalue of M_k, goes to 0. This is equivalent to showing that the minimum eigenvalue of M_k^{-1}, $\lambda_{\min}(M_k^{-1}) = \lambda_{\max}(M_k)^{-1}$, diverges to infinity. Moreover, by Lemma 6.1 and the Courant-Fischer Principle (see Courant and Hilbert, 1962, Chapter 4), $\lambda_{\min}(M_k^{-1})$ is a nondecreasing sequence. Hence, it is sufficient to show that a subsequence diverges to infinity, which we will define using the following stochastic process. Define the stochastic process $\{S_k : k+1 \in \mathbb{N}\}$ by $S_0 = 0$ and

$$S_k = \min\left\{m > S_{k-1} : span[X_{S_{k-1}+1}, \ldots, X_m] = \mathbb{R}^n\right\}. \tag{6.34}$$

We will now show that the sequence $\{\lambda_{\min}(M_{S_k}^{-1})\}$ diverges. By Lemma 6.1,

$$M_{S_{k+1}}^{-1} = M_{S_k}^{-1} + \sum_{s=S_k+1}^{S_{k+1}} \frac{1}{\gamma_s^2} X_s X_s' \succeq M_{S_k}^{-1} + \frac{1}{\Delta^2} \sum_{s=S_k+1}^{S_{k+1}} X_s X_s' \succeq I + \frac{1}{\Delta^2} \sum_{j=0}^{k} \mathcal{X}_j, \tag{6.35}$$

where

$$\mathcal{X}_k = \sum_{j=S_k+1}^{S_{k+1}} X_j X_j' \quad \forall k+1 \in \mathbb{N}. \tag{6.36}$$

By the Courant-Fischer Principle,

$$\lambda_{\min}(M_{S_{k+1}}^{-1}) \geq 1 + \frac{1}{\Delta^2} \sum_{j=0}^{k} \lambda_{\min}(\mathcal{X}_j). \tag{6.37}$$

Thus, we are left with showing that $\sum_{j=0}^{\infty} \lambda_{\min}(\mathcal{X}_j)$ diverges almost surely. To this end, we will show that $\lambda_{\min}(\mathcal{X}_j)$ will be greater than some $\alpha > 0$ infinitely often. As a result, the sum must diverge to infinity. To show this, we will use a standard Borel-Cantelli argument (see Durrett, 2010, Section 2.3). Consider the events $A_j = \{\lambda_{\min}(\mathcal{X}_j) \geq \alpha\}$ where the choice of α comes from Lemma 6.6, in which $\inf_j \mathbb{E}[\lambda_{\min}(\mathcal{X}_j)] \geq \alpha > 0$. For such an α, $\mathbb{P}[A_j] > 0$ else we would have a contradiction. By Lemma 6.5, we have that A_j are independent and that $\mathbb{P}[A_j] = \mathbb{P}[A_0] > 0$ for all $j \in \mathbb{N}$. Thus by the (Second) Borel-Cantelli Lemma (see

Durrett, 2010, Theorem 2.3.5),

$$\sum_{j=0}^{\infty} \mathbb{P}\left[A_j\right] = \infty. \tag{6.38}$$

Hence, $\lambda_{\min}(\mathcal{X}_j) \geq \alpha$ infinitely often, and thus $\lambda_{\min}(M_{S_{k+1}}^{-1}) \to \infty$ as $k \to \infty$ almost surely. Now, let $\epsilon > 0$, then almost surely there is a $K \in \mathbb{N}$ such that if $k \geq K$,

$$\lambda_{\max}(M_k) \leq \min\left\{\epsilon, \frac{\sigma^2 \epsilon}{\|E_0\|^2 \delta^2}\right\}. \tag{6.39}$$

Applying this to inequality (6.32), we can conclude the result. ■

As a simple corollary to Theorem 6.1, we prove that E_k and e_k converge to 0 in some sense.

Corollary 6.1. *If Assumptions 6.1, 6.2 and 6.4 hold, and the tuning parameters satisfy (6.29), then*

1. $\mathbb{E}\left[\|E_k\|_2 \middle| \mathcal{F}_k\right] \leq \sqrt{\mathbb{E}\left[\|E_k\|_2^2 \middle| \mathcal{F}_k\right]} \to 0$ *as* $k \to \infty$ *almost surely.*
2. $\|e_k\|_2 \to 0$ *as* $k \to \infty$ *almost surely.*

Proof. Note, by Jensen's Inequality,

$$\mathbb{E}\left[\|E_k\|_2 \middle| \mathcal{F}_k\right]^2 \leq \mathbb{E}\left[\|E_k\|_2^2 \middle| \mathcal{F}_k\right] = \mathbf{tr}\left[\mathcal{M}_k\right]. \tag{6.40}$$

By (6.32) and Theorem 6.1, $\mathbf{tr}\left[\mathcal{M}_k\right] \leq \|E_0\|_2^2 \mathbf{tr}\left[M_k^2\right] + \frac{\sigma^2}{\delta^2}\mathbf{tr}\left[M_k\right] \to 0$ as $k \to \infty$ almost surely. By Jensen's Inequality, $\|e_k\|_2 \leq \mathbb{E}\left[\|E_k\|_2 \middle| \mathcal{F}_k\right]$ almost surely, thus, $\|e_k\|_2 \to 0$ almost surely. ■

Remark 6.4. *One of several interpretations of this result is as follows. By Lemma 6.1 and (6.32), it follows that* $\mathbb{E}\left[\|E_k\|_2^2 \middle| \mathcal{F}_k\right]$ *are bounded random variables. Hence, by Corollary 6.1,* $E_k \to 0$ *in* L^2*. The usual results from measure theory then give that* $\theta_k \to \theta^*$ *in the joint* (Z, X) *probability.*

6.3.2 Insensitivity to Conditioning

The main result of this section, Theorem 6.2, states that M_k approximates a scaling of the inverse Hessian. As a consequence, Corollary 6.2 states that kSGD's convergence rate does not depend on the conditioning of the Hessian, thereby addressing the second algorithmic challenge faced by SGD. Moreover, Corollary 6.2 states that kSGD becomes arbitrarily close to the optimal statistical estimator's convergence rate in the limit (see Murata (1998), Ch.5 in Van der Vaart (1998)).

Theorem 6.2. *If Assumptions 6.1, 6.3 and 6.4 hold, and $\gamma_k^2 = \gamma^2 \in (0, \infty)$ for all $k \in \mathbb{N}$, then for $\epsilon, \upsilon \in (0,1)$ the event,*

$$\mathcal{E}_{k,\epsilon} = \left\{ \frac{\gamma^2(1-\epsilon)}{k} Q_*^{-1} \preceq M_k \preceq \frac{\gamma^2(1+\epsilon)}{k} Q_*^{-1} \right\}, \tag{6.41}$$

has probability at least $1 - \mathcal{O}\left[k^{-3+\upsilon}\right]$.

Proof. Let $\{\epsilon_k\}$ be a non-negative sequence. Define $\mathcal{Y}_k = X_k X_k' - Q_*$ and $\bar{\mathcal{Y}}_k = \frac{1}{k}\sum_{j=1}^k \mathcal{Y}_j$, and, using Lemma 6.1, define the event

$$\begin{aligned}
\mathcal{B}_k &= \left\{ M_k^{-1} \preceq \frac{k}{\gamma^2} Q_* + I(1-\epsilon_k k) \right\} \cup \left\{ \frac{k}{\gamma^2} Q_* + I(1+\epsilon_k k) \preceq M_k^{-1} \right\} \\
&= \left\{ \bar{\mathcal{Y}}_k \preceq -\epsilon_k \gamma^2 I \right\} \cup \left\{ \epsilon_k \gamma^2 I \preceq \bar{\mathcal{Y}}_k \right\} \\
&= \bigcap_{\|u\|_2 = 1} \left\{ \left| u' \bar{\mathcal{Y}}_k u \right| \geq \epsilon_k \gamma^2 \right\} \\
&\subseteq \left\{ \left| v' \bar{\mathcal{Y}} v \right| > \epsilon_k \gamma^2 \right\},
\end{aligned} \tag{6.42}$$

for an arbitrary unit vector $v \in \mathbb{R}^p$. Note that, by construction, $v'\mathcal{Y}_j v$ are independent, mean zero random variables. By Lemma 6.7, $-pC^2 \leq v'\mathcal{Y}_j v \leq pC^2$ almost surely, where $C = \|X\|_\infty$. Therefore, by Markov's Inequality and Lemma 6.8,

$$\mathbb{P}\left[\mathcal{B}_k\right] \leq \mathbb{P}\left[\left| v' \bar{\mathcal{Y}}_k v \right| > \epsilon_k \gamma^2\right] = \mathcal{O}\left[\frac{p^6 C^{12}}{\epsilon_k^6 \gamma^{12} k^3} \right]. \tag{6.43}$$

We now turn to relating $\mathcal{E}_{k,\epsilon}$ to $\mathcal{B}_k$. Let $\tilde{M}_k = Q_*^{1/2} M_k Q_*^{1/2}$. Then, $\mathcal{B}_k$

$$= \left\{ M_k^{-1} \preceq \frac{k}{\gamma^2} Q_* + I(1 - \epsilon_k k) \right\} \cup \left\{ \frac{k}{\gamma^2} Q_* + I(1 + \epsilon_k k) \preceq M_k^{-1} \right\} \tag{6.44}$$

$$= \left\{ M_k \preceq \frac{\gamma^2}{k} \left(Q_* + I \left[\frac{\gamma^2}{k} + \gamma^2 \epsilon_k \right] \right)^{-1} \right\} \cup \left\{ \frac{\gamma^2}{k} \left(Q_* + I \left[\frac{\gamma^2}{k} - \gamma^2 \epsilon_k \right] \right)^{-1} \preceq M_k \right\} \tag{6.45}$$

$$= \left\{ \tilde{M}_k \preceq \frac{\gamma^2}{k} \left(I + \gamma^2 Q_*^{-1} \left[\frac{1}{k} + \epsilon_k \right] \right)^{-1} \right\} \cup \left\{ \frac{\gamma^2}{k} \left(I + \gamma^2 Q_*^{-1} \left[\frac{1}{k} - \epsilon_k \right] \right)^{-1} \preceq \tilde{M}_k \right\} \tag{6.46}$$

$$\supseteq \left\{ \tilde{M}_k \preceq \frac{\gamma^2}{k} \frac{\lambda_{\min}(Q_*)}{\lambda_{\min}(Q_*) + \gamma^2/k + \gamma^2 \epsilon_k} I \right\} \cup \left\{ \frac{\gamma^2}{k} \frac{\lambda_{\max}(Q_*)}{\lambda_{\max}(Q_*) + \gamma^2/k - \gamma^2 \epsilon_k} I \preceq \tilde{M}_k \right\}. \tag{6.47}$$

Now take $\epsilon_k = \frac{pC^2}{\gamma^2 k^\upsilon}$ where $\upsilon \in (0,1)$. Then, for every $\epsilon > 0$ there is a $K \in \mathbb{N}$ such that for $k \geq K$,

$$1 - \epsilon \leq \frac{\lambda_{\min}(Q_*)}{\lambda_{\min}(Q_*) + \gamma^2/k + \gamma^2 \epsilon_k} \frac{\lambda_{\max}(Q_*)}{\lambda_{\max}(Q_*) + \gamma^2/k - \gamma^2 \epsilon_k} \leq 1 + \epsilon. \tag{6.48}$$

Therefore,

$$\mathcal{B}_k \supseteq \left\{ \tilde{M}_k \preceq \frac{\gamma^2(1-\epsilon)}{k} I \right\} \cup \left\{ \frac{\gamma^2(1+\epsilon)}{k} I \preceq \tilde{M}_k \right\} = \mathcal{E}_{k,\epsilon}^C. \tag{6.49}$$

Using (6.43), $\mathbb{P}\left[\mathcal{E}_{k,\epsilon}\right] \geq 1 - \mathcal{O}\left[k^{-3+\upsilon}\right]$. ■

Remark 6.5. *Given Assumption 6.3, it is likely that with a bit more work the probability of $1 - \mathcal{O}\left[k^{-3+\upsilon}\right]$ can be extended to some arbitrary $-A + \upsilon$ where $A \in \mathbb{N}_{>3}$. Also, we can extend this result to a convergence in expectation by using the fact that $0 \preceq M_k \preceq I$.*

Corollary 6.2. *If Assumptions 6.1, 6.3 and 6.4 hold, and $\gamma_k^2 = \gamma^2 \in (0, \infty)$ for all $k \in \mathbb{N}$, then for any $\epsilon, \upsilon \in (0,1)$ the event,*

$$\frac{p\sigma^2(1-\epsilon)}{k} \leq \mathbb{E}\left[R(\theta_k) \middle| \mathcal{F}_k \right] - R(\theta^*) \leq \frac{p\sigma^2(1+\epsilon)}{k}, \tag{6.50}$$

occurs with probability at least $1 - \mathcal{O}\left[k^{-3+\upsilon}\right]$.

Proof. Let $\epsilon' = \epsilon/4$. On the event $\mathcal{E}_{k,\epsilon'}$, we have a uniform bound on the rate at which

$M_k \to 0$. Thus, on the event $\mathcal{E}_{k,\epsilon'}$, we can strengthen Theorem 6.1 to $\exists K \in \mathbb{N}$ almost surely such that for any $k \geq K$

$$\frac{1-\epsilon'}{\gamma^2} M_k \preceq \frac{1}{\sigma^2}\mathcal{M}_k \preceq \frac{1+\epsilon'}{\gamma^2} M_k. \tag{6.51}$$

Combining this with Theorem 6.2, on the event $\mathcal{E}_{k,\epsilon'}$,

$$\frac{\gamma^2(1-\epsilon')^2}{k} Q_*^{-1} \preceq (1-\epsilon')M_k \preceq \frac{\gamma^2}{\sigma^2}\mathcal{M}_k \preceq (1+\epsilon')M_k \preceq \frac{\gamma^2(1+\epsilon')^2}{k} Q_*^{-1}. \tag{6.52}$$

Noting, $1-\epsilon \leq (1-\epsilon')^2 \leq (1+\epsilon')^2 \leq 1+\epsilon$, the event,

$$\frac{\sigma^2(1-\epsilon)}{k} I \preceq Q_*^{1/2}\mathcal{M}_k Q_*^{1/2} \preceq \frac{\sigma^2(1+\epsilon)}{k} I, \tag{6.53}$$

contains $\mathcal{E}_{k,\epsilon'}$. Now, using (6.6),

$$\begin{aligned} \mathbb{E}\left[R(\theta_k)\middle|\mathcal{F}_k\right] - r(\theta^*) &= \mathbb{E}\left[(\theta_k-\theta^*)'Q_*(\theta_k-\theta^*)\middle|\mathcal{F}_k\right] \\ &= \mathbf{tr}\left[Q_*^{1/2}\mathbb{E}\left[(\theta_k-\theta^*)(\theta_k-\theta^*)'\middle|\mathcal{F}_k\right]Q_*^{1/2}\right] \\ &= \mathbf{tr}\left[Q_*^{1/2}\mathcal{M}_k Q_*^{1/2}\right]. \end{aligned} \tag{6.54}$$

Therefore, on a set containing $\mathcal{E}_{k,\epsilon'}$,

$$\frac{p\sigma^2(1-\epsilon)}{k} \leq \mathbb{E}\left[R(\theta_k)\middle|\mathcal{F}_k\right] - R(\theta^*) \leq \frac{p\sigma^2(1+\epsilon)}{k}. \tag{6.55}$$

■

Now that we have shown that kSGD is insensitive to the local geometry of the problem, we turn our attention to carefully analyzing the effect that the tuning parameters have on convergence.

6.3.3 Conditions on Tuning Parameters

The tuning parameter condition, (6.29)), raises two natural questions:

1. Is the tuning parameter condition, (6.29), the necessary condition on tuning parameters in order to guarantee convergence?

2. Given the wide range of possible tuning parameters, is there an optimal strategy for choosing the tuning parameters?

As will be shown in Theorem 6.3, the first question can be answered negatively. For example, Theorem Theorem 6.3 suggests that if the method converges the tuning parameters could have been $\gamma_k^2 = k^p$ for $p \in (0, 1]$, which is an example not covered by condition (6.29).

Theorem 6.3. *Suppose* $\{\gamma_k^2\}$ *are deterministic. If Assumptions* 6.1 *and* 6.4 *hold,* $\mathbb{E}\left[\|X\|_2^4\right] < \infty$*, and for any* e_0 *we have that* $e_k \to 0$ *then* $\sum_{k=1}^{\infty} \gamma_k^{-2}$ *diverges almost surely.*

Proof. Using (6.25), premultiplying E_k by $M_k M_k^{-1}$, and taking conditional expectation with respect to $\mathcal{F}_{k+1}$, it follows that

$$e_{k+1} = M_{k+1} M_k^{-1} e_k. \tag{6.56}$$

Repeatedly applying the recursion, $e_{k+1} = M_{k+1} e_0$. Thus, $e_{k+1} \to 0$ for any e_0 is equivalent to $\lambda_{\max}(M_k) \to 0$. Note,

$$\lambda_{\max}(M_k) = \frac{1}{\lambda_{\min}(M_k^{-1})} \geq \frac{1}{\mathbf{tr}\left[M_k^{-1}\right]}. \tag{6.57}$$

We will prove that if $\sum_{j=1}^{\infty} \gamma_j^{-2} < \infty$ then the supremum of the trace of M_k^{-1} is finite, and therefore $\lambda_{\max}(M_k) > 0$, which gives a contradiction. That is, we will show the supremum over all k of

$$\mathbf{tr}\left[M_k^{-1}\right] = \mathbf{tr}\left[I + \sum_{j=1}^{k} \frac{1}{\gamma_j^2} X_j X_j'\right] = n + \sum_{j=1}^{k} \frac{\|X_j\|_2^2}{\gamma_j^2} \tag{6.58}$$

is almost surely finite, where we used Lemma 6.1 recursively to compute

$$M_k^{-1} = M_{k-1}^{-1} + \frac{1}{\gamma_k^2} X_k X_k' = I + \sum_{j=1}^{k} \frac{1}{\gamma_j^2} X_j X_j'. \tag{6.59}$$

The main tool used is Kolmogorov's Three Series Theorem (see Durrett, 2010, Theorem

2.5.4). Suppose that $\sum_{k=1}^{\infty} \gamma_j^{-2} < \infty$, and let $A > 0$. By Markov's Inequality,

$$\sum_{j=1}^{\infty} \mathbb{P}\left[\|X_j\|_2^2 > A\gamma_j^2\right] \leq \sum_{j=1}^{\infty} \frac{\mathbb{E}\left[\|X_j\|_2^2\right]}{\gamma_j^2 A} < \infty, \tag{6.60}$$

$$\sum_{j=1}^{\infty} \frac{\mathbb{E}\left[\|X_j\|_2^2 \mathbf{1}\left[\|X_j\|_2^2 \leq A\gamma_j^2\right]\right]}{\gamma_j^2} \leq \sum_{j=1}^{\infty} \frac{\mathbb{E}\left[\|X_j\|_2^2\right]}{\gamma_j^2} < \infty, \tag{6.61}$$

$$\sum_{j=1}^{\infty} \frac{\mathbb{V}\left[\|X_j\|_2^2 \mathbf{1}\left[\|X_j\|_2^2 \leq A\gamma_j^2\right]\right]}{\gamma_j^4} \leq \sum_{j=1}^{\infty} \frac{\mathbb{E}\left[\|X_j\|_2^4\right]}{\gamma_j^4} < \infty. \tag{6.62}$$

Thus, by Kolmogorov's Three Series Theorem, $\sup_k \mathbf{tr}\left[M_k^{-1}\right] < \infty$. Hence, $e_k \not\to 0$ and we have a contradiction. ■

Remark 6.6. *Using a condition such as $\sum_{k=1}^{\infty} \gamma_k^{-2}$ diverges or the sum over any subsequence diverges in place of (6.29) seems appealing. Although $\mathcal{M}_k$ can be upper bounded by M_k using such a condition and Lemma 6.2, the main theoretical difficulty occurs in proving that the $\sum_{k=1}^{\infty} \mathbb{P}[A_k] = \infty$ in the proof ofTheorem 6.1.*

The second question can be understood by examining the two roles tuning parameters have in Theorem 6.1. First, the tuning parameters determine how quickly M_k^{-1} diverges: if the tuning parameters take on very small values, then, in light of (6.37), $\lambda_{\min}(M_k^{-1})$ will diverge quickly. Therefore, the tuning parameters should be selected to be a fixed small value when the goal is to converge quickly. Second, the tuning parameter bounds, δ^2 and Δ^2, determine how tightly M_k bounds $\mathcal{M}_k$: if δ^2 and Δ^2 are close to σ^2 then M_k will have a tight lower and upper bound on $\mathcal{M}_k$. Therefore, from an algorithmic perspective, if the tuning parameter bounds are close to σ^2, then M_k will be a better stop condition.

In the case when σ^2 is of moderate size, the tuning parameters can be selected to be small thereby ensuring that fast convergence is achieved and that M_k is a strong stop condition. On the other hand, if σ^2 is large, the tuning parameters cannot satisfy both fast convergence and ensure that M_k is a strong stop condition. These opposing tuning parameter goals can be reconciled with the following result.

Theorem 6.4. *Suppose Assumptions Assumptions 6.1, 6.2 and 6.4 hold, and* $\sigma^2 \in (\delta^2, \Delta^2)$ *for some* $\Delta^2 \geq \delta^2 > 0$*. If there exists a sequence of tuning parameters satisfying (6.29) and satisfying for any* $\epsilon \in (0,1)$*, almost surely* $\exists K \in \mathbb{N}$ *such that for* $k \geq K$ *implies* $|\gamma_k^2 - \sigma^2| < \epsilon\sigma^2$ *then almost surely there exists a* $K' \in \mathbb{N}$ *such that for* $k \geq K'$

$$\mathcal{M}_k \preceq \frac{1+\epsilon}{1-\epsilon} M_k. \tag{6.63}$$

Proof. Let $\epsilon \in (0,1)$. Then by assumption, almost surely there exists a $K \in \mathbb{N}$ for which $\gamma_k^2 \geq \sigma^2(1-\epsilon)$. From Lemma 6.2 with $k \geq K$ and because $\sigma^2 \geq \delta^2$,

$$\begin{aligned}
\mathcal{M}_k &= M_k E_0 E_0' M_k + M_k \left(\sum_{j=1}^{k} \frac{\sigma^2}{\gamma_j^2} \frac{1}{\gamma_j^2} X_j X_j' \right) M_k \\
&= M_k E_0 E_0' M_k + M_k \left(\sum_{j=1}^{K-1} \left(\frac{\sigma^2}{\gamma_j^2} - 1 \right) \frac{1}{\gamma_j^2} X_j X_j' \right) M_k \\
&\quad + M_k \left(\sum_{j=1}^{K-1} \frac{1}{\gamma_j^2} X_j X_j' + \sum_{j=K}^{k} \frac{\sigma^2}{\gamma_j^2} \frac{1}{\gamma_j^2} X_j X_j' \right) M_k \\
&\preceq M_k^2 \left\| E_0 \right\|_2^2 + \left(\frac{\sigma^2}{\delta^2} - 1 \right) M_k M_{K-1}^{-1} M_k \\
&\quad + M_k \left(\sum_{j=1}^{K-1} \frac{1}{\gamma_j^2} X_j X_j' + \frac{1}{1-\epsilon} \sum_{j=K}^{k} \frac{1}{\gamma_j^2} X_j X_j' \right) M_k \\
&\preceq M_k^2 \left\| E_0 \right\|_2^2 + \left(\frac{\sigma^2}{\delta^2} - 1 \right) M_k M_{K-1}^{-1} M_k + \frac{1}{1-\epsilon} M_k.
\end{aligned} \tag{6.64}$$

By Theorem 6.1, $M_k \to 0$ almost surely. Thus, almost surely there is a $K' \in \mathbb{N}$, which we can take larger than K, such that if $k \geq K'$ then

$$M_k^2 \left\| E_0 \right\|_2^2 + \left(\frac{\sigma^2}{\delta^2} - 1 \right) M_k M_{K-1}^{-1} M_k \preceq \frac{\epsilon}{1-\epsilon} M_k. \tag{6.65}$$

Combining the inequalities gives the result. ■

The construction of such a tuning parameter strategy is quite difficult because a sequence

which almost surely converges to σ^2 will never be known apriori. In practice, the construction of γ_k^2 will then have to depend on the data $(X_1, Y_1), (X_2, Y_2), \ldots$; therefore, a data dependent tuning parameter strategy will introduce measurability issues in Theorem 6.4. Thus, we only use Theorem 6.4 as motivation for constructing a tuning parameter strategy which estimates σ^2. The construction of such tuning parameters is the content of the next subsection.

6.3.4 Adaptive Choice of Tuning Parameters

We will define a sequence of tuning parameters $\{\gamma_k^2\}$ which satisfy condition (6.29) and incrementally estimate σ^2 when $\sigma^2 \in (\delta^2, \Delta^2)$. This choice of tuning parameters will be determined by a two-step process:

1. By defining a sequence of unbounded estimators $\{\xi_k^2\}$ which converge to σ^2 in some sense.

2. Then, by defining $\gamma_k^2 = \phi_\tau(\xi_k^2)$ for a $\tau \in \mathbb{N}$ where

$$\phi_\tau(x) = \tau \mathbf{1}\left[x > \tau\right] + \tau^{-1}\mathbf{1}\left[x < \tau^{-1}\right] + x\mathbf{1}\left[\tau^{-1} \leq x \leq \tau\right] \tag{6.66}$$

to ensure that condition (6.29) is satisfied.

Remark 6.7. *The choice of ϕ_τ imposes $\delta^2 = \tau^{-1}$ and $\Delta^2 = \tau$ in (6.29). The use of a single parameter, τ, for the bounds is strictly a matter of convenience. Using different lower and upper bounds is completely satisfactory from a theoretical perspective and can be a strategy to negotiate the trade-offs in choosing the tuning parameters.*

We will use the following general form for estimators $\{\xi_k^2\}$:

$$\xi_1^2 = r_1^2 \quad \text{and} \quad \xi_{k+1}^2 = \frac{1}{k+1} f_{k+1} r_{k+1}^2 + \left(1 - \frac{1}{k+1}\right)\xi_k^2, \tag{6.67}$$

where, the residuals, r_k, satisfy $r_k = Z_k - X_k'\theta_{k-1}$, and the hyperparameters, f_k, are non-negative. One natural choice for the hyperparameters is $f_k = 1$ for all k. Indeed, such a

choice has the following nice theoretical guarantee, which is a consequence of Proposition 6.1 below.

Corollary 6.3. *If Assumptions 6.1, 6.2 and 6.4 hold and $f_k = 1$ for all k then, almost surely,*

$$\lim_{k\to\infty} \mathbb{E}\left[\xi_k^2 \middle| \mathcal{F}_k\right] = \sigma^2. \tag{6.68}$$

Moreover, almost surely,

$$\phi_\tau(\sigma^2) \leq \liminf_{k\to\infty} \mathbb{E}\left[\gamma_k^2 \middle| \mathcal{F}_k\right] \leq \limsup_{k\to\infty} \mathbb{E}\left[\gamma_k^2 \middle| \mathcal{F}_k\right] \leq \phi_\tau(\sigma^2 + \tau^{-1}). \tag{6.69}$$

Although this result calls into question why f_k are even considered in (6.67), it is misleading: if the initial residuals, $r_1^2, r_2^2, \ldots$, deviate from σ^2 significantly, then a large number of cases must be assimilated in order for the estimator of ξ_k^2 to converge to σ^2. A common strategy to overcome this is to shrink the initial residuals — similarly, introduce forgetting factors — which converge to unity. That is, we consider a positive sequence $f_k \leq 1$ such that $f_k \to 1$. However, because f_k can be designed, this procedure raises the question of how and when f_k should approach 1. Indeed, the only guidance which can be given on this choice is to make f_k depend on M_k: once M_k is sufficiently small, we have an assurance that r_k^2 should be faithful estimators of σ^2, and so f_k should be near 1 in this regime. Since M_k are measurable with respect to $\mathcal{F}_k$, a more flexible condition is to allow f_k to be measurable with respect to $\mathcal{F}_k$ as well.

Remark 6.8. *Despite restricting f_k to be $\mathcal{F}_k$ measurable, there are still many strategies for choosing f_k. For example, f_k can be set using a hard or soft thresholding strategy depending on the $\mathbf{tr}\left[M_k\right]$. Regardless, the strategy for choosing f_k should depend on the problem, and an uninformed choice is ill-advised unless a sufficient amount of data is being processed.*

An additional strategy is to delay the index at which ξ_1^2 is calculated. To be specific, we can define a stopping time V with respect to $\mathcal{F}_k$ such that V is finite almost surely, and

let $r_k = Z_{V+k} - X'_{V+k}\theta_{V+k-1}$. Indeed, then the process will only be started once a specific condition has been met, such as if $\mathbf{tr}\left[M_k\right]$ is below a threshold. This offers more flexibility than manipulating f_k alone. These considerations on f_k and V are collected in the following assumption.

Assumption 6.5. *V is a stopping time with respect to $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ such that V is almost surely finite with stopped σ-algebra $\mathcal{F}_V$ (Durrett, 2010, p. 156). Also, take f_k to satisfy:*

1. *$f_k \in [0, 1]$ and $f_k \to 1$ as $k \to \infty$ almost surely.*
2. *f_k are $\mathcal{F}_{V+k}$ measurable.*

and let $r_k = Z_{V+k} - X'_{V+k}\theta_{V+k-1}$, ξ_k^2 be defined as in (6.67), and $\gamma_{V+k}^2 = \phi_\tau(\xi_k^2)$.

We are now ready to show that a tuning parameter strategy using Assumption 6.5 does indeed approximate σ^2 in the limit.

Proposition 6.1. *If Assumptions 6.1, 6.2, 6.4 and 6.5 hold, then, almost surely,*

$$\lim_{k\to\infty} \mathbb{E}\left[\xi_k^2 \middle| \mathcal{F}_{V+k}\right] = \sigma^2. \tag{6.70}$$

Moreover, almost surely,

$$\phi_\tau(\sigma^2) \leq \liminf_{k\to\infty} \mathbb{E}\left[\gamma_{V+k}^2 \middle| \mathcal{F}_{V+k}\right] \leq \limsup_{k\to\infty} \mathbb{E}\left[\gamma_{V+k}^2 \middle| \mathcal{F}_{V+k}\right] \leq \phi_\tau(\sigma^2 + \tau^{-1}). \tag{6.71}$$

Proof. Applying (6.67) repeatedly and using Assumption 6.1,

$$\begin{aligned}
\xi_{k+1}^2 &= \sum_{l=1}^{\infty} \xi_{k+1}^2 \mathbf{1}\left[V = l\right] \\
&= \frac{1}{k+1} \sum_{l=1}^{\infty} \sum_{j=1}^{k+1} f_j \left(Y_{l+j} - X'_{l+j}\theta_{l+j-1}\right)^2 \mathbf{1}\left[V = l\right] \\
&= \frac{1}{k+1} \sum_{j=1}^{k+1} f_j \sum_{l=1}^{\infty} \left(\epsilon_{l+j} + X'_{l+j}(\theta^* - \theta_{l+j-1})\right)^2 \mathbf{1}\left[V = l\right].
\end{aligned} \tag{6.72}$$

Taking the conditional expectations, using the Conditional Monotone Convergence Theorem (see Durrett, 2010, Theorem 5.1.2), and using the facts that f_j and $\mathbf{1}\left[V = l\right]$ are measurable with respect to $\mathcal{F}_{V+k+1}$ by construction:

$$\mathbb{E}\left[\xi_{k+1}^2 \middle| \mathcal{F}_{V+k+1}\right] \tag{6.73}$$

$$= \frac{1}{k+1}\sum_{j=1}^{k+1} f_j \sum_{l=1}^{\infty} \mathbb{E}\left[\left(\epsilon_{l+j} + X_{l+j}'(\theta^* - \theta_{l+j-1})\right)^2 \mathbf{1}\left[V = l\right] \middle| \mathcal{F}_{V+k+1}\right] \tag{6.74}$$

$$= \frac{1}{k+1}\sum_{j=1}^{k+1} f_j \sum_{l=1}^{\infty} \mathbf{1}\left[V = l\right] \mathbb{E}\left[\epsilon_{l+j}^2 \middle| \mathcal{F}_{V+k+1}\right] + \mathbf{1}\left[V = l\right] X_{l+j}' \mathcal{M}_{l+j-1} X_{l+j} \tag{6.75}$$

$$+ 2\mathbf{1}\left[V = l\right] \mathbb{E}\left[\epsilon_{l+j} X_{l+j}'(\theta^* - \theta_{l+j-1}) \middle| \mathcal{F}_{V+k+1}\right] \tag{6.76}$$

$$= \frac{1}{k+1}\sum_{j=1}^{k+1} f_j \sum_{l=1}^{\infty} \mathbf{1}\left[V = l\right] \sigma^2 + \mathbf{1}\left[V = l\right] X_{l+j}' \mathcal{M}_{l+j-1} X_{l+j} \tag{6.77}$$

$$= \frac{1}{k+1}\sum_{j=1}^{k+1} f_j \sigma^2 + f_j X_{V+j}' \mathcal{M}_{V+j-1} X_{V+j}, \tag{6.78}$$

where we use the facts that (1) $\mathbb{E}\left[(\theta^* - \theta_{l+j-1})(\theta^* - \theta_{l+j-1})' \middle| \mathcal{F}_{V+k+1}\right] = \mathcal{M}_{l+j-1}$ because $\mathcal{F}_{V+j-1} \subset \mathcal{F}_{V+k-1}$ and we are restricting to the event $\{V = l\}$, and (2) ϵ_j are independent of $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, thus, ϵ_j are independent of $\mathcal{F}_{V+k}$. Using this equality, we will establish a lower and upper bound on $\mathbb{E}\left[\xi_j^2 \middle| \mathcal{F}_{V+j}\right]$.

Lower Bound. Since $\mathcal{M}_{V+j} \succeq 0$, we have $\mathbb{E}\left[\xi_{k+1}^2 \middle| \mathcal{F}_{k+1}\right] \geq \frac{\sigma^2}{k+1}\sum_{j=1}^{k+1} f_j$. Let $\epsilon > 0$. Because V is almost surely finite and $f_k \to 1$ almost surely, almost surely there is a $K \in \mathbb{N}$ such that if $k \geq K$ then $f_k \geq (1 - \epsilon)$. Therefore,

$$\liminf_{k\to\infty} \mathbb{E}\left[\xi_{k+1}^2 \middle| \mathcal{F}_{V+k+1}\right] \geq \sigma^2(1-\epsilon) \tag{6.79}$$

almost surely.

Upper Bound. For an upper bound, using the same $\epsilon > 0$ as for the lower bound, because V is almost surely finite and by Theorem 6.1, almost surely there is a $K \in \mathbb{N}$ such

that for $k \geq K$

$$\mathcal{M}_{V+k} \preceq \tau\sigma^2(1+\epsilon)M_{V+k} \quad \text{and} \quad \lambda_{\max}(M_{V+k}) < \frac{\epsilon}{\tau\sigma^2(1+\epsilon)\mathbb{E}\left[\|X_1\|^2\right]}. \tag{6.80}$$

Since $f_k \leq 1$,

$$\begin{aligned}
&\mathbb{E}\left[\xi_k^2 \middle| \mathcal{F}_{V+k}\right] \\
&\leq \sigma^2 + \frac{1}{k}\sum_{j=K+1}^{k} X'_{V+j}\mathcal{M}_{V+j-1}X_{V+j} + \frac{1}{k}\sum_{j=1}^{K} X'_{V+j}\mathcal{M}_{V+j-1}X_{V+j} &(6.81)\\
&\leq \sigma^2 + \frac{\tau\sigma^2(1+\epsilon)}{k}\sum_{j=K+1}^{k} X'_{V+j}M_{V+j-1}X_{V+j} + \frac{1}{k}\sum_{j=1}^{K} X'_{V+j}\mathcal{M}_{V+j-1}X_{V+j} &(6.82)\\
&\leq \sigma^2 + \frac{\epsilon}{\mathbb{E}\left[\|X_1\|^2\right]}\frac{1}{k}\sum_{j=K+1}^{k}\|X_k\|^2 + \frac{1}{k}\sum_{j=1}^{K} X'_{V+j}\mathcal{M}_{V+j-1}X_{V+j}. &(6.83)
\end{aligned}$$

As $k \to \infty$, the third term will vanish and the second term will converge to ϵ almost surely by the Strong Law of Large Numbers. Therefore, almost surely

$$\sigma^2(1-\epsilon) \leq \liminf_{k\to\infty} \mathbb{E}\left[\xi_{k+1}^2 \middle| \mathcal{F}_{k+1}\right] \leq \limsup_{k\to\infty} \mathbb{E}\left[\xi_{k+1}^2 \middle| \mathcal{F}_{k+1}\right] \leq \sigma^2 + \epsilon$$

Since $\epsilon > 0$ is arbitrary, it follows that the limit of the sequence exists and is equal to σ^2 almost surely.

Bounds on γ_{V+k}^2. Define

$$\nu_\tau(x) = \begin{cases} \tau & x > \tau \\ x & x \leq \tau, \end{cases} \quad \text{and} \quad \psi_\tau(x) = \begin{cases} x & x > \tau^{-1} \\ \tau^{-1} & x \leq \tau^{-1}. \end{cases} \tag{6.84}$$

From this definition, we have that for all $x \in \mathbb{R}$, $\nu_\tau \leq \phi_\tau \leq \psi_\tau$. In the next statement, for the maximum of a set of values, we use $\vee$, and for the minimum we use $\wedge$. Applying this

relationship to γ_{V+k}^2,

$$\mathbb{E}\left[\nu_\tau(\xi_k^2)\middle|\mathcal{F}_{V+k}\right] \vee \tau^{-1} \leq \mathbb{E}\left[\gamma_k^2\middle|\mathcal{F}_{V+k}\right] \leq \mathbb{E}\left[\psi_\tau(\xi_k^2)\middle|\mathcal{F}_{V+k}\right] \wedge \tau. \tag{6.85}$$

First, manipulating the right hand side,

$$\begin{aligned}\mathbb{E}\left[\psi_\tau(\xi_k^2)\middle|\mathcal{F}_{V+k}\right] &= \mathbb{E}\left[\xi_k^2\mathbf{1}\left[\xi_k^2 > \tau^{-1}\right]\middle|\mathcal{F}_{V+k}\right] + \tau^{-1}\mathbb{P}\left[\xi_k^2 < \tau^{-1}\middle|\mathcal{F}_{V+k}\right] \\ &\leq \mathbb{E}\left[\xi_k^2\middle|\mathcal{F}_{V+k}\right] + \tau^{-1}.\end{aligned} \tag{6.86}$$

Applying the first part of Proposition 6.1, $\limsup_{k\to\infty}\mathbb{E}\left[\psi_\tau(\xi_k^2)\middle|\mathcal{F}_{V+k}\right] \leq \sigma^2 + \tau^{-1}$ almost surely. For the reverse inequality:

$$\mathbb{E}\left[\nu_\tau(\xi_k^2)\middle|\mathcal{F}_{V+k}\right] = \tau\mathbb{P}\left[\xi_k^2 > \tau\middle|\mathcal{F}_{V+k}\right] + \mathbb{E}\left[\xi_k^2\mathbf{1}\left[\xi_k^2 < \tau\right]\middle|\mathcal{F}_{V+k}\right] \tag{6.87}$$

$$\geq \mathbb{E}\left[\xi_k^2\middle|\mathcal{F}_{V+k}\right] - \mathbb{E}\left[\left(\xi_k^2 - \tau\right)_+\middle|\mathcal{F}_{V+k}\right] \tag{6.88}$$

$$\geq \mathbb{E}\left[\xi_k^2\middle|\mathcal{F}_{V+k}\right] - \mathbb{E}\left[\left(\xi_k^2 - \sigma^2\right)_+\middle|\mathcal{F}_{V+k}\right] - \left(\sigma^2 - \tau\right)_+ \tag{6.89}$$

$$\geq \mathbb{E}\left[\xi_k^2\middle|\mathcal{F}_{V+k}\right] - \mathbb{E}\left[\left|\xi_k^2 - \sigma^2\right|\middle|\mathcal{F}_{V+k}\right] - \left(\sigma^2 - \tau\right)_+. \tag{6.90}$$

By the first part of the result, the first term converges to σ^2, and we are left with showing that the second term converges to 0. Note,

$$\begin{aligned}&\mathbb{E}\left[\left|\xi_k^2 - \sigma^2\right|\middle|\mathcal{F}_{V+k}\right] \\ &\leq \mathbb{E}\left[\left|\frac{1}{k}\sum_{j=1}^k f_j\epsilon_{V+j}^2 - \sigma^2\right|\middle|\mathcal{F}_{V+k}\right] + \frac{2}{k}\mathbb{E}\left[\sum_{j=1}^k f_j\left|\epsilon_{V+j}X_{V+j}'(\theta^* - \theta_{V+j-1})\right|\middle|\mathcal{F}_{V+k}\right] \\ &\quad + \frac{1}{k}\sum_{j=1}^k f_j X_{V+j}'\mathcal{M}_{V+j-1}X_{V+j}.\end{aligned} \tag{6.91}$$

For the first term, we note that (1) the argument is bounded by $\frac{1}{k}\sum_{j=1}^k \epsilon_{V+j}^2 + \sigma^2$, which are uniformly integrable with expectation $2\sigma^2$, and (2) because V is finite almost surely

and by the strong law of large numbers, almost surely $\lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} f_j \epsilon_j^2 = \sigma^2$. Therefore, by the conditional Dominated Convergence Theorem, the first term converges to 0 almost surely.

For the cross term, using Cauchy-Schwarz,

$$\begin{aligned}
&\frac{2}{k}\mathbb{E}\left[\sum_{j=1}^{k} f_j \left|\epsilon_{V+j} X'_{V+j}(\theta^* - \theta_{V+j-1})\right| \middle| \mathcal{F}_{V+k}\right] \\
&\quad \leq \frac{2}{k}\sum_{j=1}^{k} \mathbb{E}\left[\left|\epsilon_{V+j}\right| \left\|X_{V+j}\right\|_2 \left\|\theta^* - \theta_{V+j-1}\right\|_2 \middle| \mathcal{F}_{V+k}\right] \qquad (6.92)\\
&\quad \leq \frac{2}{k}\sum_{j=1}^{k}\sum_{l=1}^{\infty} \mathbb{E}\left[\left|\epsilon_{l+j}\right| \left\|X_{l+j}\right\|_2 \left\|\theta_{l+j-1} - \theta^*\right\|_2 \mathbf{1}\left[V = l\right] \middle| \mathcal{F}_{V+k}\right],
\end{aligned}$$

where, in the last line, the expectation and summation are alternated by the conditional Monotone Convergence Theorem. Using an identical conditioning argument as in the first part of the proof, and noting $\mathbb{E}\left[|\epsilon_j|\right]^2 \leq \mathbb{E}\left[|\epsilon_j|^2\right] = \sigma^2$ by Jensen's inequality, the cross term is bounded by

$$\frac{2}{k}\mathbb{E}\left[\sum_{j=1}^{k} f_j \left|\epsilon_{V+j} X'_{V+j}(\theta^* - \theta_{V+j-1})\right| \middle| \mathcal{F}_{V+k}\right] \leq \frac{2\sigma}{k}\sum_{j=1}^{k} \left\|X_{V+j}\right\|_2 \mathbf{tr}\left[\mathcal{M}_{V+j-1}\right]^{1/2}. \qquad (6.93)$$

Using the same argument as for the upper bound of $\mathbb{E}\left[\xi_k^2 \middle| \mathcal{F}_{V+k}\right]$, that is, using (6.80) in order to bound $\mathbf{tr}\left[\mathcal{M}_{V+j-1}\right]$, we have that for any $\epsilon > 0$,

$$\limsup_{k\to\infty} \frac{2}{k}\mathbb{E}\left[\sum_{j=1}^{k} \left|\epsilon_j X'_j(\theta^* - \theta_{j-1})\right| \middle| \mathcal{F}_k\right] \leq \epsilon. \qquad (6.94)$$

Therefore, the limit exists and is 0. To summarize,

$$\liminf_{k\to\infty} \mathbb{E}\left[\nu_\tau(\xi_k^2) \middle| \mathcal{F}_k\right] \geq \sigma^2 - \left(\sigma^2 - \tau\right)_+ = \sigma^2 \wedge \tau. \qquad (6.95)$$

■

6.3.5 Renewal Process

We show that the stochastic process $\{S_k : k+1 \in \mathbb{N}\}$ defined in (6.34) is a renewal process (see Durrett, 2010, Section 4.4). We define the inter-arrival times, $T_k = S_k - S_{k-1}$ for all $k \in \mathbb{N}$.

Lemma 6.3. *If $X_1, X_2, \ldots$ are independent and identically distributed, then $T_1, T_2, \ldots$ are independent and identically distributed.*

Proof. Let $k, j \in \mathbb{N}$

$$\mathbb{P}\left[T_k \geq j\right] = \sum_{s=1}^{\infty} \mathbb{P}\left[T_k \geq j | S_{k-1} = s\right] \mathbb{P}\left[S_{k-1} = s\right] \tag{6.96}$$

$$= \sum_{s=1}^{\infty} \mathbb{P}\left[span[X_{s+1}, \ldots, X_{s+j}] = \mathbb{R}^n | S_{k-1} = s\right] \mathbb{P}\left[S_{k-1} = s\right]. \tag{6.97}$$

Note that $\sigma(X_{s+1}, \ldots, X_{s+j})$ is independent of $\sigma(X_1, \ldots, X_s)$, and so:

$$\mathbb{P}\left[T_k \geq j\right] = \sum_{s=1}^{\infty} \mathbb{P}\left[span[X_{s+1}, \ldots, X_{s+j}] = \mathbb{R}^n\right] \mathbb{P}\left[S_{k-1} = s\right]. \tag{6.98}$$

Finally, $X_1, \ldots, X_j$ has the same distribution as $X_{s+1}, \ldots, X_{s+j}$. Therefore:

$$\mathbb{P}\left[T_k \geq j\right] = \sum_{s=1}^{\infty} \mathbb{P}\left[span[X_1, \ldots, X_j] = \mathbb{R}^n\right] \mathbb{P}\left[S_{k-1} = s\right] \tag{6.99}$$

$$= \mathbb{P}\left[T_1 \geq j\right] \sum_{s=1}^{\infty} \mathbb{P}\left[S_{k-1} = s\right] \tag{6.100}$$

$$= \mathbb{P}\left[T_1 \geq j\right]. \tag{6.101}$$

We have established that $T_1, T_2, \ldots$ are identically distributed. Now let $k_1 < k_2 < \cdots < k_r$ be positive integers with $r \in \mathbb{N}$. Let $j_1, \ldots, j_r \in \mathbb{N}$.

$$\mathbb{P}\left[T_{k_r} = j_r, \ldots, T_{k_1} = j_1\right] \tag{6.102}$$

$$= \sum_{s=1}^{\infty} \mathbb{P}\left[T_{k_r} = j_r | S_{k_r-1} = s, T_{k_{r-1}} = j_{r-1}, \ldots, T_{k_1} = j_1\right] \tag{6.103}$$

$$\times \mathbb{P}\left[S_{k_r-1} = s, T_{k_{r-1}} = j_{r-1}, \ldots, T_{k_1} = j_1\right]. \tag{6.104}$$

Just as above, $\sigma(X_{s+1}, \ldots, X_{s+j_r})$ is independent of $\sigma(X_1, \ldots, X_s)$ and so:

$$\mathbb{P}\left[T_{k_r} = j_r, \ldots, T_{k_1} = j_1\right]$$

$$= \sum_{s=1}^{\infty} \mathbb{P}\left[T_{k_r} = j_r\right] \mathbb{P}\left[S_{k_r-1} = s, T_{k_{r-1}} = j_{r-1}, \ldots, T_{k_1} = j_1\right] \tag{6.105}$$

$$= \mathbb{P}\left[T_{k_r} = j_r\right] \sum_{s=1}^{\infty} \mathbb{P}\left[S_{k_r-1} = s, T_{k_{r-1}} = j_{r-1}, \ldots, T_{k_1} = j_1\right] \tag{6.106}$$

$$= \mathbb{P}\left[T_{k_r} = j_r\right] \mathbb{P}\left[T_{k_{r-1}} = j_{r-1}, \ldots, T_{k_1} = j_1\right]. \tag{6.107}$$

Applying the argument recursively, we have established independence. ■

Lemma 6.4. *If $X_1, X_2, \ldots$ are independent, identically distributed, and satisfy Assumption 6.4, then $\mathbb{E}[T_1] < \infty$ and $\mathbb{E}[S_k] = k\mathbb{E}[T_1]$ for all k.*

Proof. We prove that T_1 is bounded by a geometric random variable and so its expectation must exist. Let $\mathcal{S}_m = span[X_1, \ldots, X_m]$. In this notation, we have that $T_1 = \inf\{m > 0 : \mathcal{S}_m = n\}$. We will now decompose T_1 into $P_1, \ldots, P_n$, where $P_k = \inf\{m > P_{k-1} : \dim(\mathcal{S}_m) = k\}$ with $P_0 = 0$. By construction,

$$0 = P_0 < P_1 < \cdots < P_n = T_1. \tag{6.108}$$

Let $U_1, \ldots, U_n$ denote the inter-arrival times defined by $U_k = P_k - P_{k-1}$. On the event that $U_k = j$, we have that $\exists v \in \mathbb{R}^n$ with $\|v\| = 1$ such that $\mathcal{S}_{P_{k-1}+j-1}$ is orthogonal to v, and by Assumption 6.4, there is a $p = 1 - \mathbb{P}X_1'v = 0 > 0$. Then, $\mathbb{P}[U_k = j] \leq (1-p)^{j-1}$. Thus, $\mathbb{E}[U_k] < \infty$. Therefore, $\mathbb{E}[T_1] = \mathbb{E}[U_1] + \mathbb{E}[U_2] + \cdots \mathbb{E}[U_n] < \infty$. Now, by Lemma 6.3, $T_1, \ldots, T_k$ are independent and identically distributed. Therefore, $\mathbb{E}[S_k] = \mathbb{E}[T_1] + \cdots +$

$\mathbb{E}[T_k] = k\mathbb{E}[T_1]$. ■

Lemma 6.5. *If $X_1, X_2, \ldots$ are independent and identically distributed, then $\mathcal{X}_0, \mathcal{X}_1, \ldots$ defined in (6.36) are independent and identically distributed. In particular, the eigenvalues of $\mathcal{X}_0, \mathcal{X}_1, \ldots$ are independent and identically distributed.*

Proof. Let A and be a measurable set.

$$\begin{aligned}
&\mathbb{P}[\mathcal{X}_k \in A] \\
&= \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \mathbb{P}\left[X_{s+1}X_{s+1}' + \cdots + X_{s+t}X_{s+t}' \in A | T_{k+1} = t, S_k = s\right] \\
&\quad \times \mathbb{P}[T_{k+1} = t | S_k = s]\,\mathbb{P}[S_k = s] && (6.109)\\
&= \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \mathbb{P}[X_1X_1' + \cdots + X_tX_t' \in A | T_1 = t]\,\mathbb{P}[T_1 = t]\,\mathbb{P}[S_k = s] && (6.110)\\
&= \mathbb{P}[\mathcal{X}_0 \in A] && (6.111)
\end{aligned}$$

Thus, $\mathcal{X}_0, \mathcal{X}_1, \ldots$ are identically distributed. By the independence of $X_1, X_2, \ldots$ and the independence of $T_1, T_2, \ldots$, which is established in Lemma 6.3, $\mathcal{X}_0, \mathcal{X}_1, \ldots$ are independent because they are functions of independent random variables. Finally, since the eigenvalues of $\mathcal{X}_k$ can be calculated using the Courant-Fisher Principle, we have that they too are independent and identically distributed. ■

Lemma 6.6. *If $X_1, X_2, \ldots$ are independent and identically distributed, and satisfy Assumptions 6.2 and 6.4, then $\exists \alpha > 0$ such that for all k, $\mathbb{E}[\lambda_{\min}(\mathcal{X}_k)] \geq \alpha > 0$.*

Proof. By Lemma 6.5, we need only consider $\mathcal{X}_0$. Suppose there exists a $v \in \mathbb{R}^n$ with $\|v\| = 1$ such that $\mathbb{E}[v'\mathcal{X}_0 v] = 0$. Note, by Assumption 6.2 and Cauchy-Schwarz, the expectation is well defined. Since $\mathcal{X}_0 \succeq 0$ by construction, almost surely

$$0 = v'\mathcal{X}_0 v = \sum_{s=1}^{S_1} (X_s' v)^2. \tag{6.112}$$

Thus, $X_s'v = 0$ for all $s = 1, \ldots, S_1$. However, by construction, Assumption 6.4 and Lemma 6.4, $X_1, \ldots, X_{S_1}$ span $\mathbb{R}^n$ and $S_1 < \infty$ almost surely, hence there is an $s \leq S_1$ such that $X_s'v \neq 0$ almost surely. Therefore, we have a contradiction. ■

6.3.6 Some Properties of L^∞ Random Variables

In this section, we establish some useful properties of L^∞ random variables.

Lemma 6.7. *Suppose $X \in L^\infty$ random variable taking values in $\mathbb{R}^p$. Then, for $C = \|X\|_\infty$,*

$$0 \preceq XX' \preceq pC^2 I \quad \text{and} \quad -pC^2 I \preceq XX' - \mathbb{E}[XX'] \preceq pC^2 I \tag{6.113}$$

almost surely.

Proof. The lower bound is straightforward. For the upper bound, let v be any unit vector. By Cauchy-Schwartz,

$$v'XX'v = (X'v)^2 \leq \|X\|_2^2 \|v\|_2^2 \leq pC^2 \|v\|_2^2 \leq pC^2, \tag{6.114}$$

where $C = \|X\|_\infty$. Thus the first set of inequalities holds almost surely. Moreover, the first set of inequalities imply $0 \preceq \mathbb{E}[XX'] \preceq pC^2 I$. Hence,

$$\begin{aligned} -\mathbb{E}[XX'] \preceq XX' - \mathbb{E}[XX'] \preceq pC^2 I - \mathbb{E}[XX'] \\ -pC^2 I \preceq XX' - \mathbb{E}[XX'] \preceq pC^2 I. \end{aligned} \tag{6.115}$$

■

Lemma 6.8. *Let $Z_1, Z_2, \ldots, Z_k$ be mean zero, independent random variables with $\|Z_1\|_\infty = \cdots = \|Z_k\|_\infty = D > 0$. Then*

$$\mathbb{E}\left[\left(\sum_{j=1}^{k} Z_j\right)^6\right] \leq \mathcal{O}\left[D^6 k^3\right]. \tag{6.116}$$

Proof. Note that $\mathbb{E}[Z_j] = 0$ and Z_j are independent. Hence, in the polynomial expansion, any monomial with a term that has unity exponent is going to have a zero expectation. So, we need to count all monomials whose terms have exponents at least two in the expansion:

1. There are k terms of the form $(Z_j)^6$.
2. There are $\binom{k}{2}$ terms of the form $(Z_j)^2(Z_i)^4$ with $\frac{6!}{4!2!} = 15$ ways of choosing the exponents.
3. There are $\binom{k}{2}$ terms of the form $(Z_j)^3(Z_i)^3$ with $\frac{6!}{3!3!} = 20$ ways of choosing the exponents.
4. There are $\binom{k}{3}$ terms of the form $(Z_j)^2(Z_i)^2(Z_l)^2$ with $\frac{6!}{2!2!2!} = 90$ ways of choosing the exponents.

Since $-D \leq Z_j \leq D$ almost surely by assumption, we have that $\mathbb{E}\left[\left(\sum_{j=1}^k Z_j\right)^6\right] \leq (k + 35k^2 + 90k^3) D^6$. ■

6.4 Numerical Experiments

Three problems were experimented on: a linear regression problem on medical claims payment by CMS (Centers for Medicare & Medicaid, 2010; Belgium Network of Open Source Analytical Consultants, 2012), an additive nonparametric Haar wavelet regression problem on gas sensor readings (UCI Machine Learning Repository, 2015; Fonollosa et al., 2015), and a logistic regression problem on adult income classification (UCI Machine Learning Respository, 1996; Kohavi, 1996). For each problem, the dimension of the unknown parameter (p), number of observations (N), condition number (κ) of the Hessian at the minimizer, and the optimization methods implemented on the problem are collected in Table 6.1.

Remark 6.9. *For the linear and Haar wavelet regression problems, the Hessian does not depend on the parameter, and so it can be calculated directly. For the logistic regression problem, the minimizer was first calculated using generalized Gauss Newton (GN) (Wedderburn, 1974), and confirmed by checking that the composite gradient at the minimizer had a*

euclidean norm no larger than 10^{-10}*; then, the Hessian was calculated at this approximate minimizer.*

Table 6.1: A tabulation of the number of parameters (n), number of observations (N), condition number (κ) of the Hessian at the minimizer, and optimization methods implemented for each of the three problem types. Note, the Haar wavelet regression problem's maximum eigenvalue was 28.7, but its smallest eigenvalue was within numerical precision of zero.

Problem	N	n	κ	Methods
CMS-Linear	$2,801,660$	34	2.44×10^6	kSGD,SGD,SQN
Gas-Haar	$4,177,004$	$1,263$	–	kSGD, SGD, SQN
Income-Logistic	$30,162$	29	1.96×10^{24}	kSGD, SGD, SQN, GN

For each method, intermediate parameter values, elapsed compute time, and data points assimilated (ADP) were periodically stored. Once the method terminated, the objective function was calculated at each stored parameter value using the entire data set. The methods are compared along two dimensions:

1. Efficiency: the number of data points assimilated (ADP) to achieve the objective function value. The higher the efficiency of a method, the less information it needs to minimize the objective function. Thus, higher efficiency methods require fewer data points or fewer epochs in order to achieve the same objective function value in comparison to a lower efficiency method.

2. Effort: the elapsed time (in seconds) to achieve the objective function value. This metric is a proxy for the cost of gradient evaluations, Hessian evaluations, floating point operations, and I/O latencies. Thus, higher effort methods require more resources or more time in order to achieve the same objective function value in comparison to a lower effort method.

The methods were implemented in the Julia Programming Language (v0.4.5). For the linear and logistic regression problem, the methods were run on an Intel i5 (3.33GHz) CPU with 3.7 Gb of memory; for the Haar Wavelet regression problem, the methods were run on an Intel X5650 (2.67GHz) CPU with 10 Gb of memory.

Remark 6.10. *The objective function for the linear and Haar wavelet regression is the mean of the residuals squared (MRS). Therefore, for these problems, the results are discussed in terms of MRS.*

6.4.1 Linear Regression for CMS Payment Data

We modeled the medical claims payment as a linear combination of the patient's sex, age and place of service. Because the explanatory variables were categorical, there were $p = 34$ parameters. The optimal MRS was determined to be $38,142.6$ using an incremental QR algorithm (Miller, 1992).

The three methods, SGD, kSGD and SQN, were initialized at zero. For SGD, the learning rate was taken to be of the general form

$$\eta(k, p, c_1, c_2, c_3) = c_1 \mathbf{1}\left[k \leq c_2\right] + \frac{c_3}{(k - c_2)^p} \mathbf{1}\left[k > c_2\right], \tag{6.117}$$

where k is the ADP, $p \in (0.5, 1]$, $c_1 \in [0, \infty)$, $c_2 \in [0, \infty]$, and $c_3 \in (0, \infty)$. SGD was implemented for learning rates over a grid of values for p, c_1, c_2 and c_3. The best learning rate, $(p = 0.75, c_1 = 0.01, c_2 = 10^5, c_3 = 1)$, achieved the smallest MRS. Note, this learning rate took only 0.01 seconds longer per epoch than the fastest learning rate. For SQN, the parameters b, b_h and L were allowed to vary between the recommended values $b = 100, 1000$, $b_h = 300, 600$, $L = 10, 20$ (Byrd et al., 2016), M was allowed to take values $10, 20, 34$, and the learning rate constant, c, was allowed to vary over a grid of positive numbers. The best set of parameters, $(b = 1000, b_h = 300, L = 20, M = 34, c = 2)$, came within 2% of the optimal MRS with the smallest ADP and least amount of time. The tuning parameters for kSGD, summarized in Table 6.2, were selected to reflect the concepts discussed above and were not determined based on any results from running the method.

Figure Fig. 6.1 visualizes the differences between SGD, kSGD and SQN in terms of efficiency and effort. In terms of efficiency, kSGD-1, kSGD-3 and SQN are comparable,

whereas kSGD-2 and SGD perform quite poorly in comparison. For kSGD-2, this behavior is to be expected for large choices of γ_k^2 as discussed below. For SGD, despite an optimal choice in the learning rate, it still does not come close to the optimal MRS after a single epoch. Even when SGD is allowed to complete multiple epochs so that its total elapsed time is greater than kSGD's single epoch elapsed time (Fig. 6.1, right), SGD does not make meaningful improvements towards the optimal MRS. Indeed, this is to be expected as the rate of convergence of SGD is highly sensitive to the local geometry. We also see in Fig. 6.1 (right) that kSGD-1 and kSGD-3 require much less effort to calculate the minimizer in comparison to SQN.

Subsection 6.4.1 visualizes the behavior of the covariance estimates for each of the three kSGD tuning parameter choices. We highlight the sluggish behavior of M_k for kSGD-2, which underscores one of the ideas discussed in above: if σ^2 is large, choosing γ_k^2 to approximate σ^2 for all k will slow down the convergence of the algorithm. Another important property is that, despite some variability, $\mathbf{tr}\left[M_k\right]$ is reflective of the decay in MRS; this property empirically reinforces the result in Theorem 6.1, and the claim that M_k can be used as a stop condition in practice.

6.4.2 Nonparametric Wavelet Regression on Gas Sensor Readings

We modeled the ethylene concentration in a mixture of Methane, Ethylene and Air as an additive nonparametric function of time and sixteen gas sensor voltage readings. Because

Table 6.2: Tuning parameter selection for kSGD method. kSGD-1 uses a tuning parameter based on Theorem 6.3. kSGD-2 uses a tuning parameter approximating the MRS based on (6.15). kSGD-3 uses a tuning parameter to increase the speed of convergence based on the discussion in Subsection 6.3.3.

Label	γ_k^2
kSGD-1	k^{-1}
kSGD-2	$38,000$
kSGD-3	0.0001

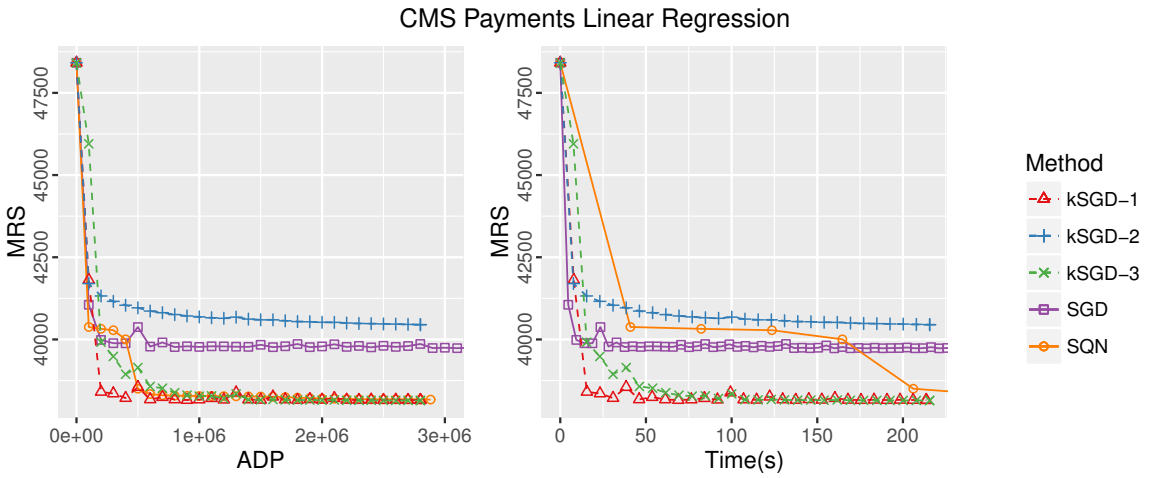


Figure 6.1: A comparison of the performance of SGD, kSGD and SQN for the linear regression problem. *Left:* In terms of efficiency, kSGD-1, kSGD-3 and SQN are comparable, whereas kSGD-2 and SGD perform quite poorly in comparison. *Right:* kSGD-1 and kSGD-3 produce nearly optimal estimates within the first 50 seconds, which is approximately the amount of time SGD needs to complete one epoch. Also, kSGD-1 and kSGD-3 require less effort than SQN.

the response and explanatory variables were in bounded intervals, the function space was approximated using Haar wavelets without shifts (Haar, 1910). The resolution for the time component was 9, and the resolution for each gas sensor was 3, which resulted in features and a parameter of dimension $p = 1,263$.

Remark 6.11. *Because of their high-cost of calculation, the features were calculated in advance of running the methods.*

Again, SGD, kSGD and SQN were implemented and initialized at zero. For SGD, using the same criteria as described in Subsection 6.4.1, the best learning rate was found to be $(p = 0.8, c_1 = 0.0, c_2 = 0.0, c_3 = 1)$. For SQN, regardless of the choice of parameters (over a grid larger than the one used in Subsection 6.4.1), the BFGS estimates quickly became unstable and caused the parameter estimate to diverge. For kSGD, the method was implemented with $\gamma_k^2 = 0.0001$.

Figure Subsection 6.4.2 compares SGD and kSGD. Although kSGD is much more efficient than SGD, it is remarkably slower than SGD because it is performing $\mathcal{O}\left[1.6e6\right]$ operations

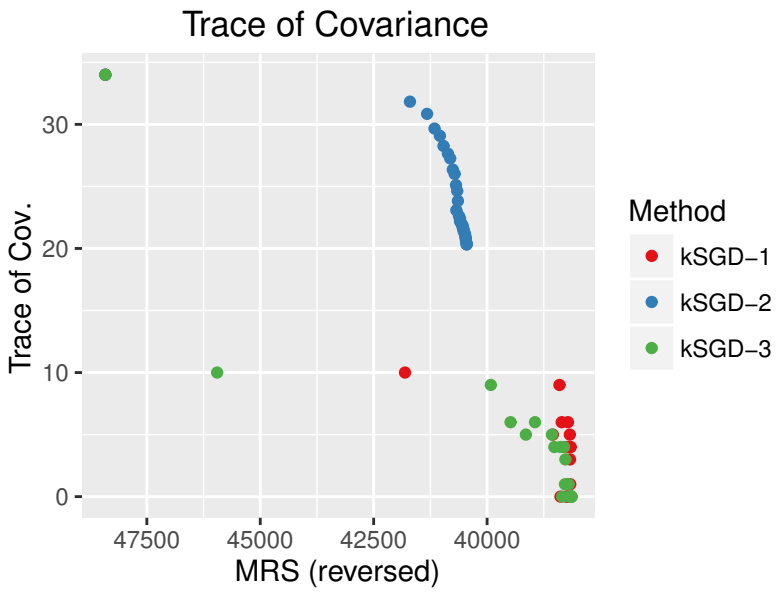


Figure 6.2: A comparison of the MRS and trace of the covariance. The rapid decay of kSGD-1's and kSGD-3's covariance is reflected in the rapid decay of their MRS. On the other hand, kSGD-2's covariance is decaying slowly and this too is reflected in the slow decay of the MRS.

per iteration in comparison to SGD's $\mathcal{O}[1.2e3]$ operations per iteration. For this problem, this difference in effort can be reduced by using sparse matrix techniques since at most 74 of the $1,263$ components in each feature vector are non-zero; however, for dense problems this issue can only be resolved by parallelizing the floating point operations at each iteration or using low memory methods discussed in the next section.

6.4.3 Logistic Regression on Adult Incomes

We modeled two income classes as a logistic model with eight demographic explanatory variables. Four of the demographic variables were continuous and the remaining four were categorical, which resulted in $p = 29$ parameters.

SGD, kSGD, SQN and GN were implemented and initialized at zero. For SGD, the best learning rate was found to be $(p = 0.5, c_1 = 0.0, c_2 = 0.0, c_3 = 0.01)$. For SQN, the best parameter set was found to be $(b = 1000, b_h = 300, L = 10, M = 29, c = 10)$. For GN, there were no tuning parameters. kSGD was implemented as in Algorithm 2 within Algorithm 1,

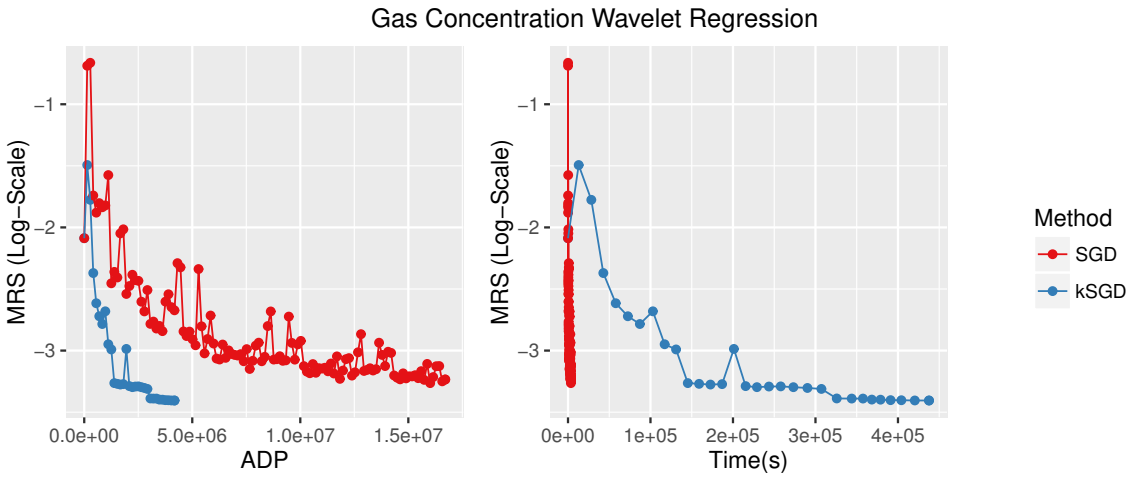


Figure 6.3: A comparison of SGD and kSGD on the Haar wavelet regression problem. *Left:* kSGD is more efficient than SGD. *Right:* kSGD requires significantly more effort.

where the triggers were determined by the covariance estimate dropping below decaying (by a factor of 5) thresholds. The threshold was arbitrarily intialized at 15, and γ_k^2 was started at 0.0001 and increased by a factor of 10 until the method did not fail; the method succeeded when $\gamma_k^2 = 0.1$.

Figure Subsection 6.4.3 visualizes the efficiency and effort of the four methods. In general, the behavior of kSGD, SGD and SQN follow the trends in the two preceding examples. An interesting feature is that kSGD had greater efficiency and required less effort than GN. This is due to the fact that kSGD incompletely solves the GN subproblem away from the minimizer, while GN solves the subproblem exactly at each iteration. However, it is important to note that the choice of kSGD tuning parameters is not as straightforward for the logistic regression problem as it is for the linear regression problem, and appropriate choices will not be discussed further in this paper.

6.5 Reduced Complexity Modifications

As the nonparametric regression example demonstrated, while kSGD is more efficient than SGD, kSGD requires drastically more floating point operations and storage in comparison to SGD for large dimensional problems. In modern applications such as training deep neural

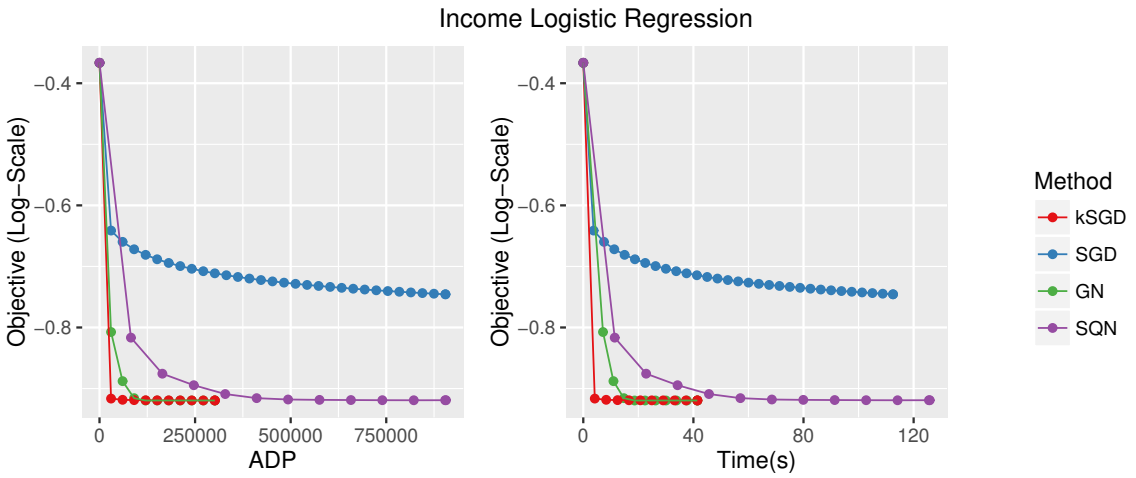


Figure 6.4: A comparison of SGD, kSGD, SQN and GN on the logistic regression problem. *Left:* kSGD is more efficient than all three methods. *Right:* kSGD requires less effort than all three methods.

networks or running as a subroutine in 4-dimensional variational data assimilation problems, kSGD's limited scalability renders it impractical.

This drawback raises a natural question: is it possible to vary kSGD in such a way as to reduce its floating point and storage complexity while retaining its efficiency? By examining Algorithm 2, we see that the main challenge is in storing and performing operation, or with, the matrix M. Therefore, any attempt to improve kSGD's scalability will depend on reducing the cost of storing and manipulating M.

There are several standard approaches that we can borrow from numerical optimization such as diagonal block approximations to M or low-rank approximations. While these approaches are successful at reducing the storage complexity associated with M, naive approaches still require extensive floating point operations to update M in a systematic manner. However, if we take advantage of the inherent structure of M, namely that

$$M_{k+1}^{-1} = M_k^{-1} + \frac{1}{V(X_{k+1}, \theta_k)} \mu_\psi(X_{k+1}, \theta_k) \mu_\psi(X_{k+1}, \theta_k)', \tag{6.118}$$

we recognize that M_k^{-1} can be composed using a sequence of vectors. Moreover, we would

recognize this update as having an analogous structure to the limited-memory Quasi-Newton update (see Nocedal and Wright, 2006, Chapter 7.2), or, more accurately, an approximate Gram-Schmidt orthogonalization. Leveraging this structure and adapting it to kSGD, we can derive a collection of reduced kSGD techniques based on the relative dimensions of p and d.

The complexities of several reduced kSGD approaches are summarized in Table 6.3, where p is the dimension of the parameter, θ, d is the dimension of the features, X, and $m \geq 0$ is an integer valued quantity controlling the number of vectors to be stored. A detailed summary of each of these approaches is given in the next subsection, which is then followed by some simple numerical experiments comparing the reduced kSGD variants to kSGD and other incremental estimators.

Table 6.3: Summary of Reduced kSGD Approaches

Label	Condition	Memory Storage	Floating Point	Grad Evaluation
FMM	$d < p$	$\mathcal{O}\left[m(d+p)+p\right]$	$\mathcal{O}\left[mp\right]$	$\mathcal{O}\left[m\right]$
FLM	$d < p$	$\mathcal{O}\left[md+p\right]$	$\mathcal{O}\left[m^3p\right]$	$\mathcal{O}\left[m^3\right]$
PMM	$d \geq p$	$\mathcal{O}\left[2(m+1)p\right]$	$\mathcal{O}\left[mp\right]$	0
PLM	$d \geq p$	$\mathcal{O}\left[(m+2)p\right]$	$\mathcal{O}\left[m^3p\right]$	0

6.5.1 Some Reduced kSGD Approaches

Here, we discuss and produce algorithms for the four reduced kSGD approaches summarized in Table 6.3. For each approach, we store an implicit representation of M based on a "seed" vector c_0, which is a seed vector for the matrix M, and a collection of feature and parameter vectors to update M based on recent information. A generic reduced kSGD subroutine is given in Algorithm 3, in which two of the calculations depend on which of the four approaches that we take. Note, Algorithm 3 is referred to as a subroutine because it will will be most useful if embedded within Algorithm 1. Also note, we have ignored the variance term in place of a hyper-parameter γ for notational simplicity.

Algorithm 3: A Generic Reduced kSGD Subroutine

Data: Example (Z, X), Iterate θ, Hyper-parameter $\gamma > 0$
Data: Mean function μ
Data: Implicit representation of M
Result: Updated Iterate, θ, Updated Implicit Representation of M

$g_{k+1} \leftarrow \mu_\psi(X, \theta)$
$m \leftarrow \mu(X, \theta)$
Compute v_{k+1} using implicit representation //Approach Dependent
$s \leftarrow \gamma + v'_{k+1} g_{k+1}$
$\theta \leftarrow \theta + \frac{Z-m}{s} v_{k+1}$
Update Implicit Representation of M //Approach Dependent
return *Estimate θ, Representation of M*

Feature-based, Moderate Memory (FMM). For the FMM-kSGD approach, suppose that $d < p$ and we can store m vectors of size d and $m + 2$ vectors of size p. In computing θ_{k+1}, the m vectors of size d are allocated to storing $\{X_{k+1-m}, \ldots, X_{k+2-m}, \ldots, X_k\}$. One of the vectors of size p is used to store the seed vector for the matrix M, denoted c_0; m of the vectors of size p are used to store $\{v_{k+1-m}, \ldots, v_k\}$, which represent the product between the covariance and derivative of μ. In this approach, we do *not* stored the vectors $\{g_{k+1-m}, \ldots, g_k\}$, which we will recompute by evaluating the gradient, μ_ψ, m times *at the current parameter iterate* and the stored observations $\{X_{k+1-m}, \ldots, X_{k+2-m}\}$. Algorithm 4 details how to compute v_{k+1} from this implicit representation of M, and includes an update for the implicit representation. Note, some operations are element-wise as indicated by the comments.

Feature-based, Low Memory (FLM). For the FLM-kSGD approach, suppose that $d < p$ and we can store m vectors of size d and a of size p. For computing θ_{k+1}, the m vectors of size d are allocated to storing $\{X_{k+1-m}, \ldots, X_k\}$. The vector of size p is allocated to storing the seed vector c_0. Here, we do store neither $\{g_{k+1-m}, \ldots, g_k\}$ nor $\{v_{k+1-m}, \ldots, v_k\}$, which means that we will have to compute these quantities recursively. This will require performing $\mathcal{O}[m^3]$ gradient evaluations and many floating point operations. Thus, we are

Algorithm 4: FMM-kSGD Matrix-Vector Product and Update Subroutine

Data: Matrix seed c_0, Observations $\{X_{k+1-m}, \ldots, X_k, X_{k+1}\}$, Directions $\{v_{k+1-m}, \ldots, v_k\}$
Data: Estimate θ, Gradient g_{k+1}, Mean function μ, Hyper-parameter γ
Result: Approximate Matrix-Vector Product q
Result: Updated Representation of M

$q \leftarrow c_0 \times g_{k+1}$ // Element-wise multiplication
for $i = 1 : m$ **do**
  $g \leftarrow \mu_\psi(X_{k+i-m}, \theta)$
  if $i == 1$ **then**
    $c_0 \leftarrow \left(c_0^{-1} + \frac{1}{\gamma} g^2\right)^{-1}$ //Element-wise operations
  end
  $\alpha \leftarrow \frac{g'q}{\gamma + v'_{k+i-m} g}$
  $q \leftarrow q - \alpha v_i$
end
return *Matrix-Vector Product* q
return *Implicit Representation* c_0, $\{X_{k+2-m}, \ldots, X_k, X_{k+1}\}$, $\{v_{k+2-m}, \ldots, v_k, q\}$

trading off on memory storage by increasing the computational complexity of the estimation scheme. Algorithm 5 details the main recursive algorithm. Moreover, we update c_0 after v_{k+1} is computed and we use the same update approach as in Algorithm 4.

Algorithm 5: FLM-kSGD Matrix-Vector Product Subroutine

Data: Matrix seed c_0, Observations $\{X_{k+1-m}, \ldots, X_{k+1-j}\}$ $(j \in \{1, \ldots, m\})$
Data: Estimate θ, Mean function μ, Hyper-parameter γ
Result: Approximate Matrix-Vector Product q

$q \leftarrow c_0 \times \mu_\psi(X_{k+1-j}, \theta)$ // Element-wise multiplication
for $i = 1 : (m - j + 1)$ **do**
  $g \leftarrow \mu_\psi(X_{k+i-m}, \theta)$
  Compute v recursively with c_0, $\{X_{k+1-m}, \ldots, X_{k+(i-1)-m}\}$ and g.
  $\alpha \leftarrow \frac{g'q}{\gamma + v'g}$
  $q \leftarrow q - \alpha v$
end
return *Matrix-Vector Product* q

Parameter-based, Moderate Memory (PMM). For the PMM-kSGD approach, suppose that $d \geq p$ and $2m+1$ vectors of size p can be stored. Then, one vector is allocated to storing the

seed vector c_0; $2m$ vectors are allocated to storing $\{(g_{k+1-m}, v_{k+1-m}), \ldots, (g_k, v_k)\}$. Because the gradient vectors are being stored, there is no need to recompute them for each update. Algorithm 6 details this approach.

Algorithm 6: PMM-kSGD Matrix-Vector Product Subroutine

Data: Matrix seed c_0, $\{g_{k+1-m}, \ldots, g_k, g_{k+1}\}$, $\{v_{k+1-m}, \ldots, v_k\}$
Data: Hyper-parameter $\gamma > 0$
Result: Approximate Matrix-Vector Product q
Result: Updated Representation of M

$q \leftarrow c_0. \times g_{k+1}$ // Element-wise multiplication
for $i = 1 : m$ **do**
  $\alpha \leftarrow \frac{g'_{k+i-m} q}{\gamma + v'_{k+i-m} g_{k+i-m}}$
  $q \leftarrow q - \alpha v_i$
end
$c_0 \leftarrow \left(c_0^{-1} + \frac{1}{\gamma} g_{k+1-m}^2\right)^{-1}$ //Element-wise operations
return *Matrix-Vector Product* q
return *Implicit Representation* c_0, $\{g_{k+2-m}, \ldots, g_k, g_{k+1}\}$, $\{v_{k+2-m}, \ldots, v_k, q\}$

Parameter-based, Low Memory (PLM). For the PLM-kSGD approach, suppose that $d \geq p$ and $m + 1$ vectors of size p can e stored. Then, one vector is allocated to storing the seed vector c_0, and the remaining m vectors are allocated to storing $\{g_{k+1-m}, \ldots, g_k\}$. Again, just as for PMM-kSGD, we store the gradients so that we do not have to recompute them. However, just as for FLM-kSGD, we are not storing the directions $\{v_{k+1-m}, \}, v_k\}$, which requires that we recompute them at each iteration. Algorithm 7 details this approach. We also update c_0 as we do in Algorithm 6 once θ has been updated. Note, it is unlikely that this variant is particularly useful, unless the floating point calculations can be done extremely inexpensively.

6.5.2 Numerical Experiments

The goal of the experiments is to evaluate the performance of the reduced kSGD variants against other popular methods. Specifically, we compare PMM-kSGD (with $m = 0$ and $m = 5$) to SGD, AdaGrad, full kSGD, and zero-order, full model incremental estimator

Algorithm 7: PLM-kSGD Matrix-Vector Product Subroutine

Data: Matrix seed c_0, Observations $\{g_{k+1-m}, \ldots, g_{k+1-j}\}$ $(j \in \{1, \ldots, m\})$
Data: Hyper-parameter γ
Result: Approximate Matrix-Vector Product q

$q \leftarrow c_0 \times g_{k+1-j}$ // Element-wise multiplication
for $i = 1 : (m - j + 1)$ **do**
  Compute v_{k+i-m} recursively with c_0, $\{g_{k+1-m}, \ldots, g_{k+(i-1)-m}\}$ and g_{k+i-m}.
  $\alpha \leftarrow \frac{g'_{k+i-m} q}{\gamma + v'_{k+i-m} g_{k+i-m}}$
  $q \leftarrow q - \alpha v$
end
return *Matrix-Vector Product* q

(proximal method). Each method is embedded in the Algorithm 1 methodology where $\tau_1 = 1000$ and $\tau_{k+1} = 2\tau_k$ for $k \in \mathbb{N}$ (note, without the restart methodology, SGD and AdaGrad perform significantly worse than with restart methodology). Each method was initialized at the zero vector.

The two problems described below are of small dimension because we wanted to ensure that the dimension is manageable for kSGD and the proximal method. For each method, the parameter, the number of accessed examples, and the elapsed time were periodically recorded. We then use these variables to compare the performance of the different algorithms.

The details of each method are as follows.

- SGD. We implement stochastic gradient descent with the following learning rate:

$$\frac{1000}{k + 2000},$$

  where k is the iteration. We do not attempt to find an optimal learning rate as this requires oracle knowledge of the problem, and, instead, consider the realistic case where a learning rate is simply selected which performs sufficiently well.

- AdaGrad. We implement the diagonal variant of AdaGrad with the tuning parameter $\eta = 1e - 3$. This was the largest such parameter (over a grid) which behaved well for

the algorithm for both problems.

- KF. We implemented the usual kSGD with the hyper-parameter $\gamma = 1$. For both problems, such a choice of γ is too conservative, but we did not modify it.
- KF0. We implemented PMM-kSGD with memory parameter $m = 0$ and hyper-parameter $\gamma = 1$.
- KF5. We implemented PMM-kSGD with memory parameter $m = 5$ and hyper-parameter $\gamma = 1$.
- Prox. We implemented a zero-order, full model incremental estimator with an SGD warm start and the same learning rate as the SGD implementation. The proximal operator was solved using Newton's method.

For both the logistic regression and neural network training problems, the methods saw the examples in exactly the same order. Moreover, for both problems, each method saw exactly the same number of examples (approximately two epochs).

Logistic Regression on Adult Incomes. The data set used comes from the UCI repository (UCI Machine Learning Respository, 1996). The data set was preprocessed by removing all examples with missing information, which results in 30,162 examples remaining. The model states two income classes as a logistic model with eight demographic variables as features. The model contains twenty-nine dimensional parameter to be learned. For each recorded parameter, the empirical mean of the negative log-likelihood, which we refer to as the objective function, of the data set was computed.

Figs. 6.5 and 6.6 show the objective function value against the accessed examples (efficiency) and against the elapsed time (effort), respectively. There are several takeaways from these results. First, the runtime of SGD, AdaGrad and the low-memory KF variants are approximately the same (note, the SGD run time is shifted to the right by about three

seconds because of the programming languages precompilation and optimization features). Moreover, SGD is clearly the best method in terms of elapsed time and accessed examples. We attribute this to two features: first, we were lucky in our choice of learning rate, as other learning rates did not perform as well, and, second, the reset routine greatly improved the convergence behavior of SGD – without it, SGD and AdaGrad perform very poorly. Another important takeaway is that the reduced kSGD perform reasonably well in comparison to kSGD and the proximal method, but do so with a much lower runtime.

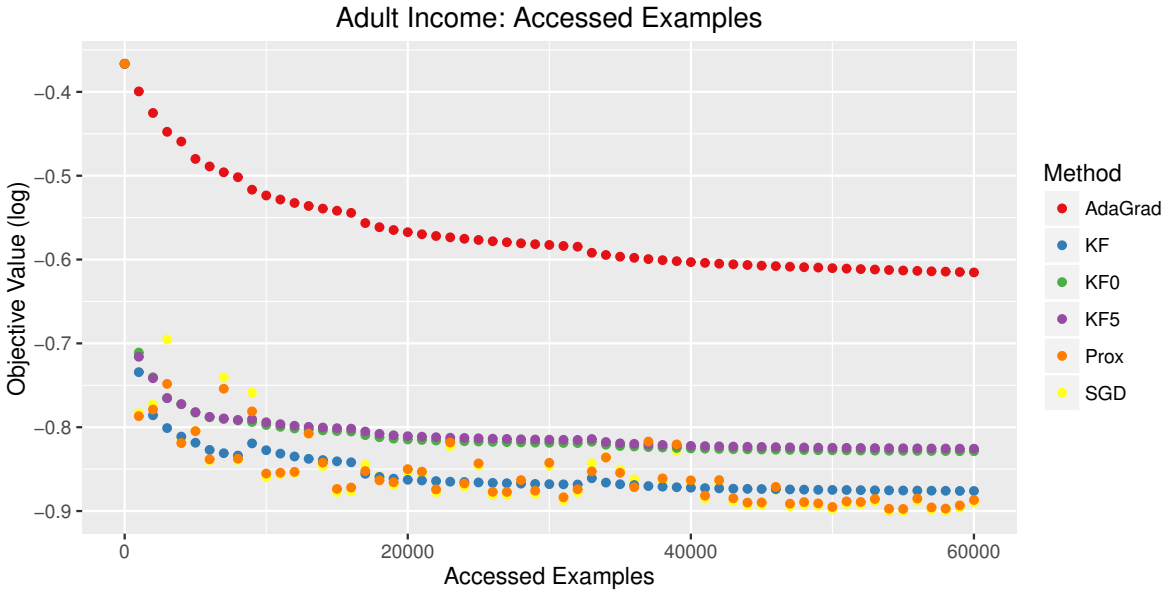


Figure 6.5: A comparison of the objective function value against the number of accessed examples for the logistic regression on the adult income data.

Neural Network for Neutrino Classifier. The data set used comes from the UCI repository (Roe et al., 2005). The data was preprocessed by prepending each feature vector with an indicator variable for being a signal or background, and the data was split into a training data set (seventy percent) and a testing data set (thirty percent). The model was a two layer neural network with logistic activation functions at each node, resulting in a sixty-one dimensional parameter. For each recorded parameter, the testing error was calculated using the testing data set.

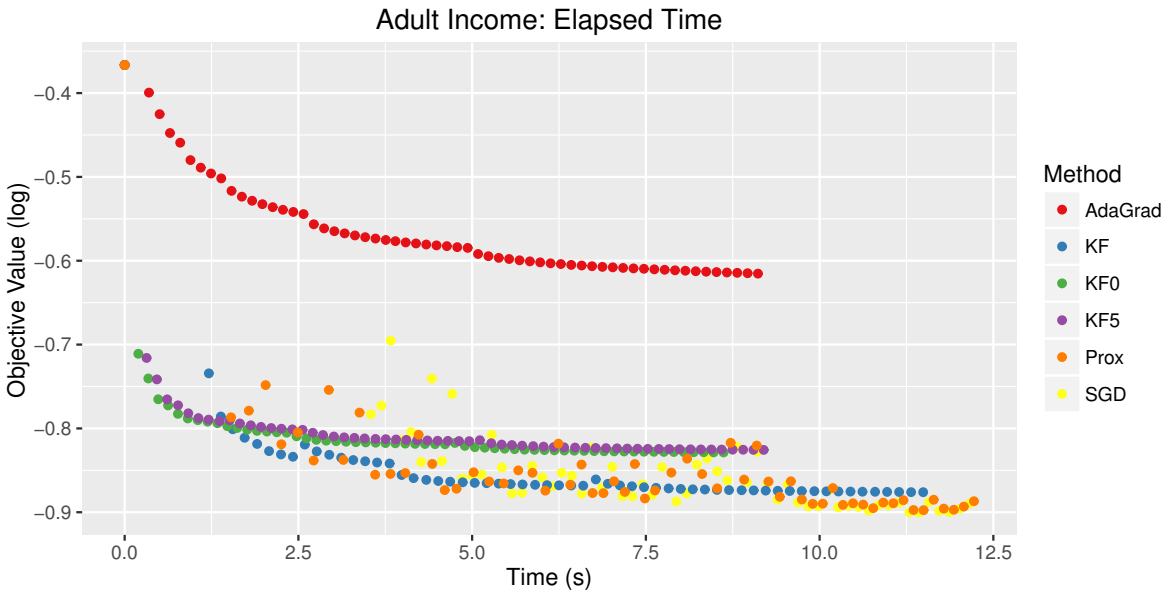


Figure 6.6: A comparison of the objective function value against the elapsed time for the logistic regression on the adult income data.

Figs. 6.7 and 6.8 show the objective function value against the accessed examples and against the elapsed time, respectively. Again, we see time shift for SGD owing to the precompilation and optimizations of the programming language. However, here, AdaGrad performs much better, but it is out performed by KF0, KF5 and KF. Moreover, we also see that KF5 outperforms KF0 in this case, which is what we expect for the added costs associated with using the KF5 method over the KF0 method. While neither of these methods perform as well as the full kSGD, they do offer a significant improvement over the competing methods and do so with a comparable compute time. (Note, the proximal method is left out as it requires a Hessian to be computed, which is rather difficult for neural networks.)

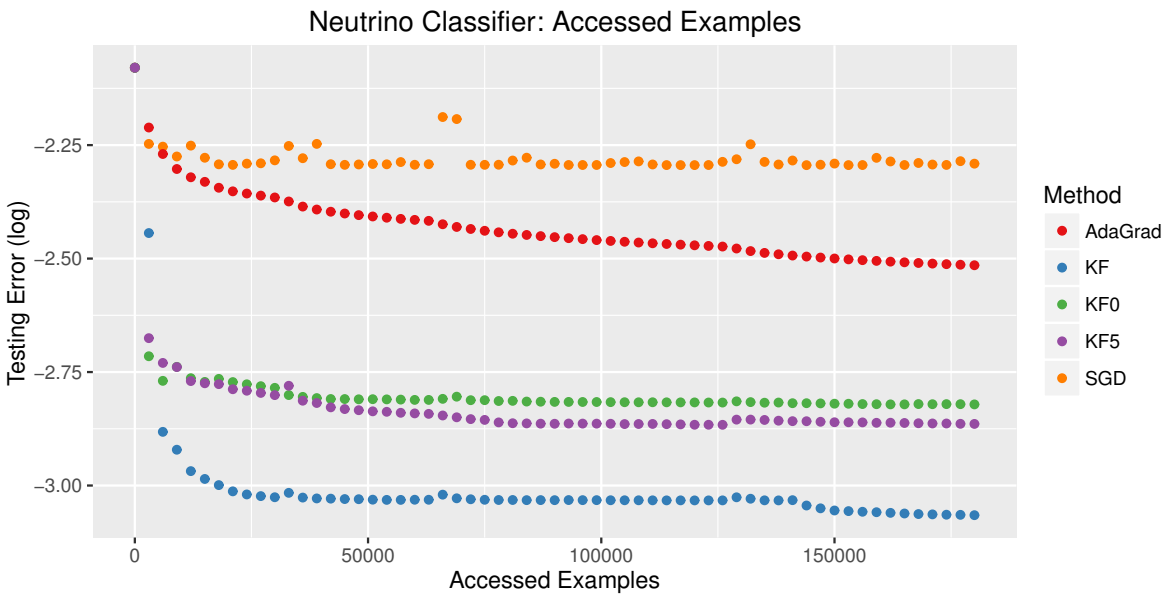


Figure 6.7: A comparison of the testing error against the number of accessed examples for the two-layer neural network on the neutrino data set.

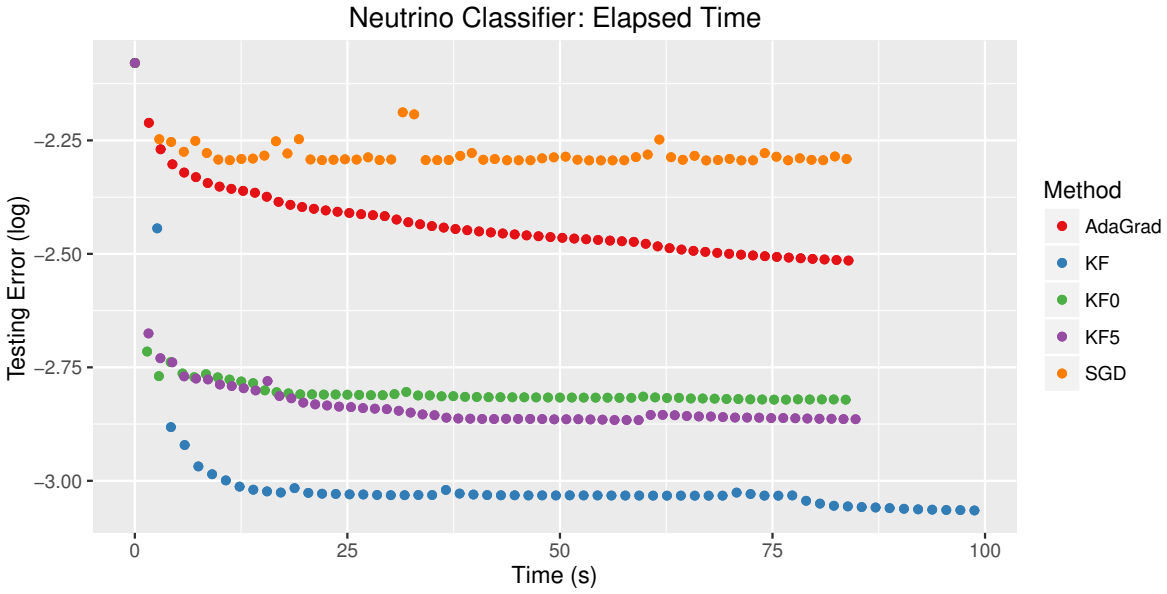


Figure 6.8: A comparison of the testing error against the elapsed time for the two-layer neural network on the neutrino data set.

7 | Incremental Estimation for Dynamical Systems

Suppose that we have a family of dynamical system models indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^p$, which is given by

$$\begin{aligned} \dot{x} &= f(t, x, \theta, \eta) \\ x(0) &= \tilde{x}(\theta), \end{aligned} \tag{7.1}$$

where $f : [0, T] \times \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^d$ is referred to as the dynamics; $x : [0, T] \to \mathbb{R}^d$ is called the state; $\eta : [0, T] \to \mathbb{R}^q$ is called the input; and $\tilde{x} : \mathbb{R}^p \to \mathbb{R}^d$. Moreover, suppose that an unknown, identifiable model in this family, θ^*, generates states $\{x(t; \theta^*, \eta) : t \in [0, T]\}$, and observations

$$y_i = Ox(t_i; \theta^*, \eta) + \epsilon_i, \;\; i = 1, \ldots, N, \tag{7.2}$$

where $t_1 \leq t_2 \leq \cdots \leq t_N \in [0, T]$ are equally spaced observation times; $O \in \mathbb{R}^{n \times d}$ is called the measurement or observation operator; and $\{\epsilon_i : i = 1, \ldots, N\}$ are independent and identically distributed with $\mathbb{E}[\epsilon_1] = 0$ and $\mathbb{V}[\epsilon_1] = \sigma^2 I$ for some $\sigma^2 > 0$. Given observations $\{y_i : i = 1, \ldots, N\}$, our goal is to estimate θ^*.

Although we mentioned this only briefly in Chapter 1, (7.1) and (7.2) are often approximations to the more common problem of dynamical system tracking; that is, the situation in which the parameter is changing over time. However, tracking the parameter is often challenging because (1) there is no model for the dynamics of the parameters (which, even if it existed, would have its own parameters that must be estimated), and (2) the observation at time t is (usually) insufficient to invert the parameter at time t. Therefore, the parameter tracking approximation, specified by (7.1) and (7.2) over interval $[0, T]$, trades off between exactly tracking the parameter and having a sufficient number of observations to ensure that the parameter is identifiable, without having to model the dynamics of the parameter.

The parameter tracking approximation imposes several practical constraints, which are

determined by the size of the approximation interval $[0, T]$. In an ideal situation, $[0, T]$ would be sufficiently small such that some measure of the error between the estimated parameter and the temporally-changing true parameter is within some allowable tolerance. As the observation rate in physical (e.g., Cooper, 1998) and engineering (e.g., Von Meier et al., 2014) applications increases, small intervals will have a sufficient number of observations to ensure that the parameter is identifiable. However, reducing the interval size requires reducing the time needed to compute the estimates; otherwise, the estimates may have too large of a lag between the interval over which they are computed and the current interval.

While there have been important strides in leveraging the structure of this estimation problem to increase computational efficiency of the optimization (e.g., Zavala and Anitescu, 2010), the primary bottleneck is not caused by the optimization, but rather is caused by the numerical integration needed to compute the gradient as discussed in Chapter 1. To summarize this discussion, if we have N equally spaced observations in the interval $[0, T]$, then the computation of the gradient requires a time step no larger than T/N, and, when N is large (i.e., a high observation rate), this bound on the time step imposes a significant computational burden. This suggests that if we used a subsampled gradient with an incremental estimator, then we could increase this upper bound on the time step and save computational time during the gradient calculation.

However, to ensure the computational benefit of using subsampled gradients, we must be careful about how these gradients are computed. Moreover, to use subsampled gradients in kSGD, we also need to carefully compute the μ_ψ term in Algorithm 2. These considerations are discussed in Section 7.1. Moreover, to demonstrate that the resulting incremental estimators actually supply the stated computational benefit, we report on several numerical experiments in Section 7.2.[1]

[1]The use of incremental estimators for such problems is not novel (Ho and Lee, 1964; Ljung, 1979), although I have not seen them used in the form that we are suggesting. These earlier incremental estimators, however, tend to perform somewhat poorly in practice.

7.1 Gradient Computation

Let

$$z(t) = \begin{bmatrix} x(t) \\ \theta \end{bmatrix} \tag{7.3}$$

denote the augmented state variable, whose dynamics are given by $\dot{z} = g(t, z)$, where

$$g(t, z) = \begin{bmatrix} f(t, x, \theta, \eta) \\ 0 \end{bmatrix}; \tag{7.4}$$

and $0 \in \mathbb{R}^p$. Moreover, let

$$H = \begin{bmatrix} O & 0 \end{bmatrix}, \tag{7.5}$$

where $0 \in \mathbb{R}^{n \times p}$. Then, using standard approaches from calculus of variations (see Dreyfus, 1962, 1965),

$$\nabla_\psi \sum_{i=1}^{N} \frac{1}{2} \left\| y_i - H z(t_i) \right\|_2^2 \Big|_{\psi=\theta} = \lambda(0)' \begin{bmatrix} \nabla_\psi \tilde{x}(\psi)|_{\psi=\theta} \\ I \end{bmatrix}, \tag{7.6}$$

where $\lambda(0)$ is the solution to

$$\begin{aligned} \dot{\lambda} &= \nabla_\zeta g(t, \zeta)'|_{\zeta=z(t)} \, \lambda \\ \lambda(t_i^-) &= \lambda(t_i^+) - H'(y_i - H z(t_i)) \; i = 1, \ldots, N-1 \\ \lambda(t_N) &= -H'(y_N - H z(t_N)), \end{aligned} \tag{7.7}$$

over $t_N > t > 0$. The dynamical system specified by (7.7) is known as the adjoint system and, because $\lambda(0)$ is computed by integrating the adjoint system from t_N to 0, the calculation of $\lambda(0)$ is known as the backwards integration. Note, (7.7) requires storing the forward system, $\{z(t) : t \in [0, T]\}$, in order to formulate the adjoint system. Thus, a naive solution of (7.7) would require a storage complexity of $\mathcal{O}[N]$. Fortunately, as demonstrated by Griewank (1992), an efficient solution of (7.7), requiring a storage and additional computa-

tional compleixty of $\mathcal{O}\left[\log(N)\right]$, can be achieved by checkpointing some of the forward states and reintegrating the forward system from these checkpoints as needed in order to formulate (7.7).

Equation 7.7 can be straightforwardly modified for a subsampled gradient: let $\mathcal{S} \subset \{1, \ldots, N\}$ be an ordered set, then the subsampled gradient is given by the right hand side of (7.6) where $\lambda(0)$ is computed by solving

$$\begin{aligned} \dot{\lambda} &= \nabla_{\zeta} g(t, \zeta)'|_{\zeta = z(t)} \lambda \\ \lambda(t_J) &= -\frac{1}{p_J} H'(y_J - Hz(t_J)), \ J = \max\{\mathcal{S}\} \\ \lambda(t_j^-) &= \lambda(t_j^+) - \frac{1}{p_j} H'(y_j - Hz(t_j)) \ j \in \mathcal{S} \setminus \{J\} \end{aligned} \tag{7.8}$$

from $t_J > t > 0$, where $p_i > 0$ is the probability that $i \in \mathcal{S}$ for $i = 1, \ldots, N$.

To see why these inclusion probabilities are needed, note that by solving (7.8), we have that

$$\lambda(0) = -\sum_{i=1}^{N} \exp\left(\int_0^{t_i} \nabla_{\zeta} g(t, \zeta)'|_{\zeta = z(t)} \, dt \right) H'(y_i - Hz(t_i)). \tag{7.9}$$

Now, let $p_i > 0$ be the probability of including the observation indexed by i in a random subsample, $\mathcal{S}$. Then,

$$\begin{aligned} \lambda(0) &= -\sum_{i=1}^{N} \exp\left(\int_0^{t_i} \nabla_{\zeta} g(t, \zeta)'|_{\zeta = z(t)} \, dt \right) H'(y_i - Hz(t_i)) \\ &= -\sum_{i=1}^{N} \frac{p_i}{p_i} \exp\left(\int_0^{t_i} \nabla_{\zeta} g(t, \zeta)'|_{\zeta = z(t)} \, dt \right) H'(y_i - Hz(t_i)) \\ &= -\mathbb{E}\left[\sum_{i=1}^{N} \frac{\mathbf{1}\left[i \in \mathcal{S}\right]}{p_i} \exp\left(\int_0^{t_i} \nabla_{\zeta} g(t, \zeta)'|_{\zeta = z(t)} \, dt \right) H'(y_i - Hz(t_i)) \right] \\ &= -\mathbb{E}\left[\sum_{j \in \mathcal{S}} \frac{1}{p_j} \exp\left(\int_0^{t_j} \nabla_{\zeta} g(t, \zeta)'|_{\zeta = z(t)} \, dt \right) H'(y_j - Hz(t_j)) \right], \end{aligned} \tag{7.10}$$

from which we see that scaling by the reciprocal of the inclusion probabilities produces an

unbiased estimate of the gradient.

While the subsampled gradient can be directly used within SGD, it is not sufficient for using kSGD. For kSGD, we need the structural component of the derivative, μ_ψ (see Algorithm 2), which, from (7.9), is (up to a scaling)

$$\mu_\psi(t_i) := \exp\left(\int_0^{t_i} \nabla_\zeta g(t,\zeta)'|_{\zeta=z(t)}\, dt\right) H', \tag{7.11}$$

which is the solution to

$$\begin{aligned} \dot{\mu}_\psi(t) &= \left(\nabla_\zeta g(t,\zeta)'|_{\zeta=z(t)}\right) \mu_\psi(t) \\ \mu_\psi(0) &= H'. \end{aligned} \tag{7.12}$$

Note that we can integrate (7.12) in parallel with the forward system (with a small time lag). Thus, we can then compute the gradient in parallel to solving the forward integration without the need for checkpoints. However, integrating (7.12) requires forwarding integrating n differential systems, which is more efficient then using just the subsampled gradient when $\log(|\mathcal{S}|) \gg n$. Therefore, if we have a sufficiently high density of observations in a time interval $[0, T]$, then it is better to compute the derivative this way.

7.2 Numerical Experiments

We now report on several experiments on simple systems of ordinary differential equations, in which we compare an approximate computation of the MLE against SGD and kSGD with varying subsample sizes.

Experimental Setup. We will repeat an identical experiment on three different differential equation models, which are described in detail below. For each experiment, we assume that the initial state is a fixed, known value. We then simulate observations from a "true" model, where the observations are recorded every 0.01 time units over an interval of $[0, 10]$, and each observation is a perturbation of the state by independent, normally distributed random variables with a variance matrix that is the identity.

Then, we apply several methods to estimate the true model. The first method is gradient descent (GD); the second method is stochastic gradient descent (SGD) with either ten, thirty or fifty percent of the data subsampled per iteration; and the third method is Kalman-based SGD (kSGD) with either ten, thirty or fifty percent of the data subsampled per iteration. Each estimator is initialized at the same point, and each estimator is run up to twenty-five iterations. The elapsed time, the absolute error between the estimates and the "true" model, and the mean squared error are recorded.

7.2.1 FitzHugh-Nagumo Model

The FitzHugh-Nagumo ODE system is a model of the excitation patterns of nerve cells (FitzHugh, 1961; Nagumo et al., 1962). The FitzHugh-Nagumo model is specified by a two-dimensional state vector $x = (x_1, x_2)$ and a four-dimensional parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ with temporal dynamics

$$\begin{aligned} \dot{x}_1 &= x_1 - \frac{x_1^2}{3} - x_2 - \theta_1 \\ \dot{x}_2 &= \frac{1}{\theta_4}\left(x_1 + \theta_2 - \theta_3 x_2\right), \end{aligned} \tag{7.13}$$

with an initial state of $x(0) = (1.5, 1.0)$. For the true parameter $\theta = (-0.5, 0.7, 0.8, 12.5)$, the phase portrait and corresponding observations are shown in Fig. 7.1.

The elapsed times for the different estimators are reported in Table 7.1. We see that there are tremendous savings when SGD and kSGD are used in comparison to GD, and we see that this savings begins to reduce as the percentage of the data subsampled increases. This is expected because increasing the size of the subsample reduces the size of the time step used in the forward and backward integration.

The absolute error and mean squared error (MSE) are reported in Fig. 7.2. We notice that the kSGD estimates are finding alternative minimizers that have a much lower MSE in comparison to GD and SGD. However, owing to the nonconvexity of the problem, it is hard

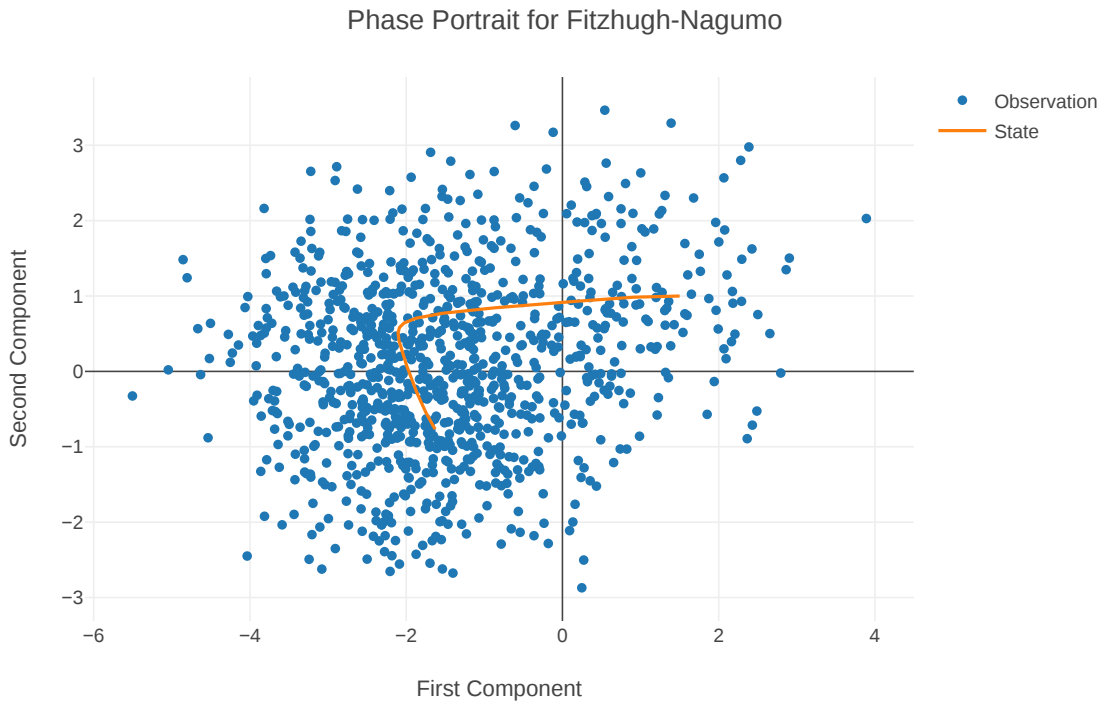


Figure 7.1: The phase-portrait of the simulated FitzHugh-Nagumo model state and the corresponding observations over a ten second interval.

to reason about why this is observed.

7.2.2 Lotka-Volterra Model

The Lotka-Volterra ODE system is a model describing the ecological variations of two populations, where one of the two populations is a predator of the other (Lotka, 1910; Volterra, 1928). The Lotka-Volterra model is specified by a two-dimensional state vector $x = (x_1, x_2)$ and a four-dimensional parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ with temporal dynamics

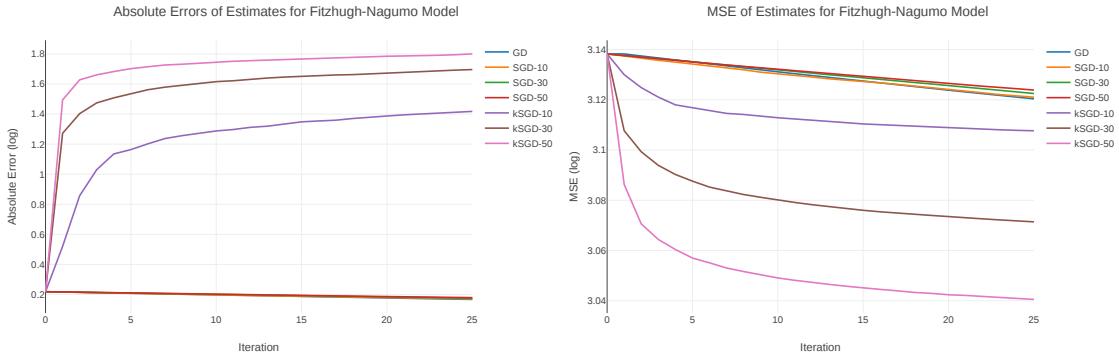
$$\begin{aligned} \dot{x}_1 &= \theta_1 x_1 - \theta_2 x_1 x_2 \\ \dot{x}_2 &= \theta_3 x_1 x_2 - \theta_4 x_2, \end{aligned} \tag{7.14}$$

with an initial state of $x(0) = (0.9, 0.9)$. For the true parameter $\theta = (2/3, 0.8, 1, 1)$, the phase portrait and corresponding observations are shown in Fig. 7.3.

The elapsed times for the different estimators are reported in Table 7.2. Again, we see

Table 7.1: Elapsed Time Comparison for FitzHugh-Nagumo Experiment

Method	% Data	Time (s)	% Time of GD
GD	100%	66.42	100.00
SGD	50%	35.26	53.09
SGD	30%	22.99	34.62
SGD	10%	9.32	14.04
kSGD	50%	42.70	64.29
kSGD	30%	26.52	39.93
kSGD	10%	10.52	15.84



(a) (b)

Figure 7.2: The absolute error (left) and mean-squared-error (right) for the estimates generated by the different estimation methodologies for the FitzHugh-Nagumo experiment.

that there are tremendous savings when SGD and kSGD are used in comparison to GD, and we see that this savings begins to reduce as the percentage of the data subsampled increases. Again, this is expected because increasing the size of the subsample reduces the size of the time step used in the forward and backward integration.

The absolute error and mean squared error (MSE) are visualized in Fig. 7.4. Again, we notice that the kSGD estimates are finding alternative minimizers that have a much lower MSE in comparison to GD and SGD.

7.2.3 Van der Pol Model

The Van der Pol ODE system is a model of a nonlinearly damped spring (Van der Pol, 1934). The Van der Pol model is specified by a two-dimensional state vector $x = (x_1, x_2)$ and a

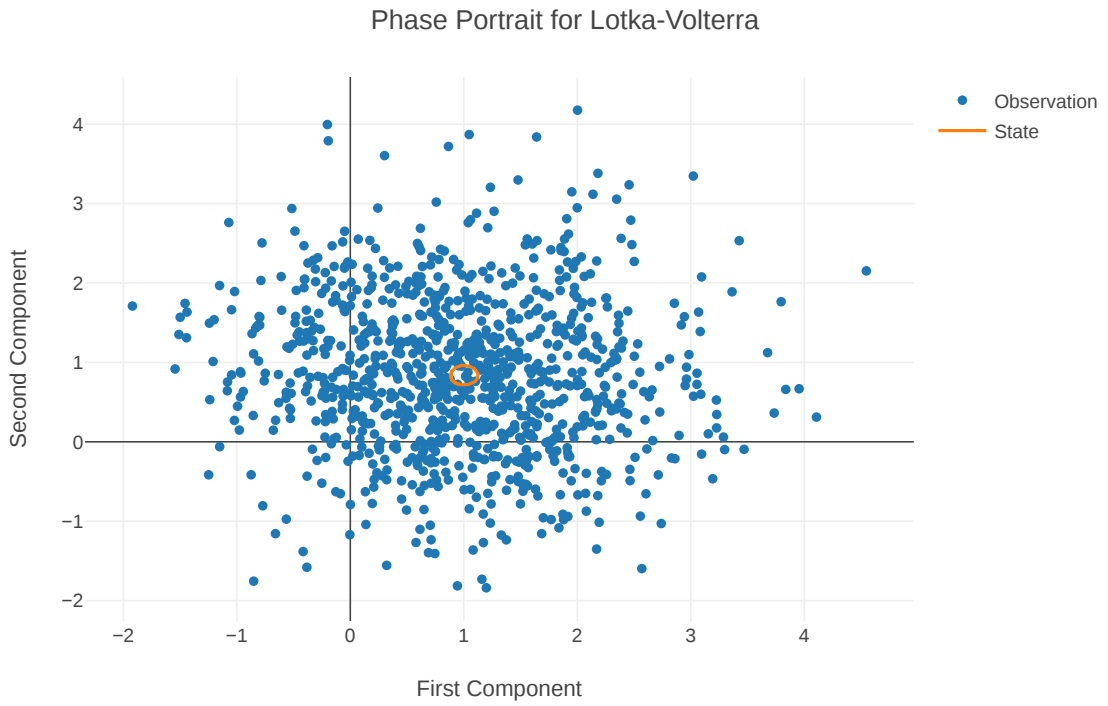


Figure 7.3: The phase-portrait of the simulated Lokta-Volterra model state and the corresponding observations over a ten second interval.

scalar parameter θ with temporal dynamics

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \theta(1 - x_1^2)x_2 - x_1 \end{aligned} \tag{7.15}$$

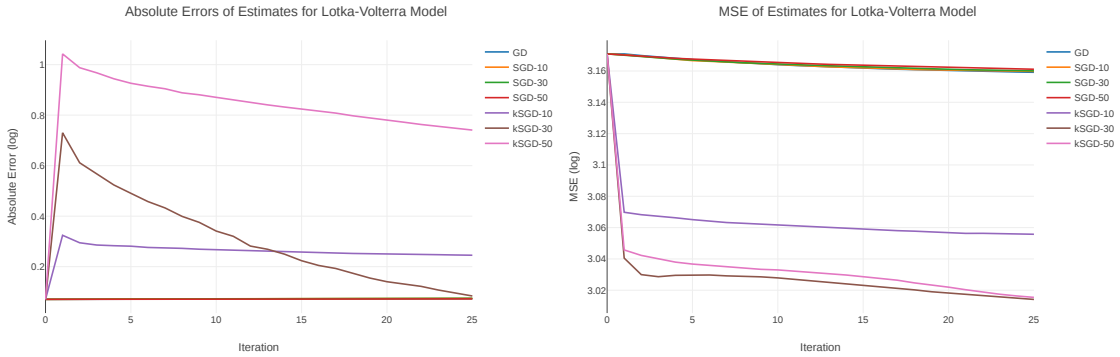
with an initial state of $x(0) = (0.5, -0.5)$. For the true parameter $\theta = 3.5$, the phase portrait and corresponding observations are shown in Fig. 7.5.

The elapsed times for the different estimators are reported in Table 7.3. Again, we see that there are tremendous savings when SGD and kSGD are used in comparison to GD, and we see that this savings begins to reduce as the percentage of the data subsampled increases.

The absolute error and mean squared error (MSE) are visualized in Fig. 7.6. Unlike the previous two examples, kSGD seems to be making very little progress for this example. It is not entirely clear why this is happening.

Table 7.2: Elapsed Time Comparison for Lokta-Volterra Experiment

Method	% Data	Time (s)	% Time of GD
GD	100%	87.12	100.00
SGD	50%	37.83	43.42
SGD	30%	26.68	30.62
SGD	10%	13.02	14.94
kSGD	50%	48.10	55.21
kSGD	30%	33.20	38.11
kSGD	10%	19.23	22.08



(a) (b)

Figure 7.4: The absolute error (left) and mean-squared-error (right) for the estimates generated by the different estimation methodologies for the Lokta-Volterra experiment.

7.2.4 A Brief Summary

From the preceding experiments, we underscore several patterns. First, there is a tremendous savings in computational effort when using subsampled gradients in comparison to full gradients. Of course, as long as the gradient calculation is implemented sensibly, this is an expected outcome. Second, kSGD tends to find estimates that have a lower MSE than either SGD and GD, but as we saw with the Van der Pol example, this is not always guaranteed. These two observations warrant a more detailed theoretical exploration of the capabilities of incremental estimators for dynamical system tracking.

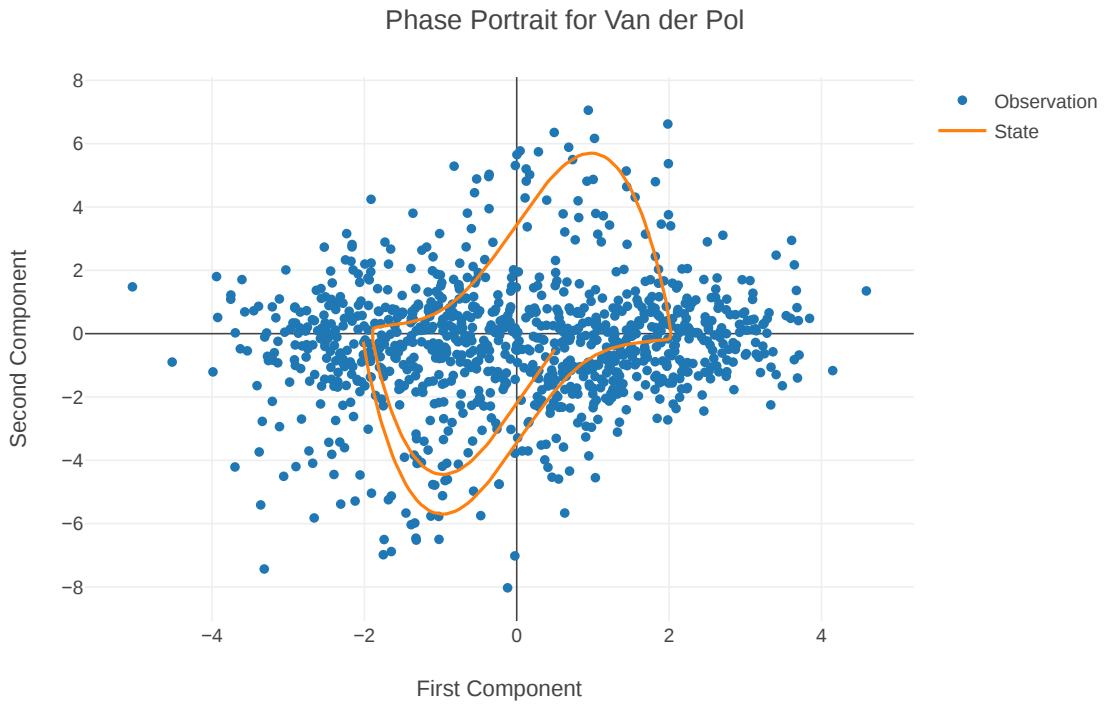
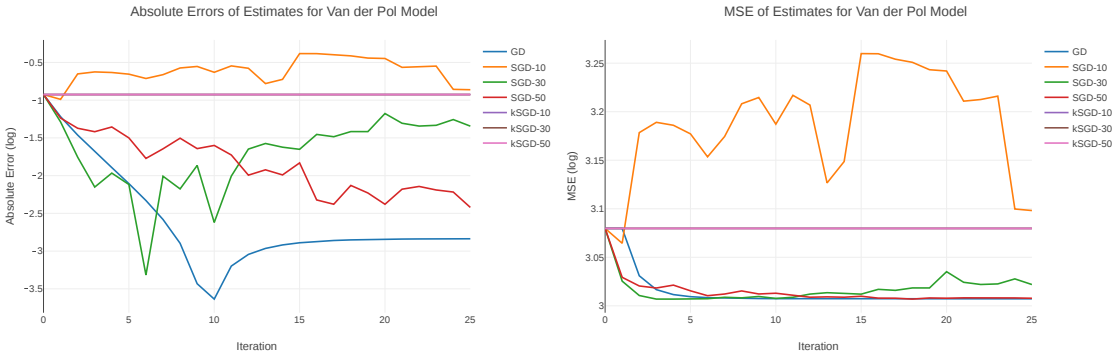


Figure 7.5: The phase-portrait of the simulated Van der Pol model state and the corresponding observations over a ten second interval.



(a) (b)

Figure 7.6: The absolute error (left) and mean-squared-error (right) for the estimates generated by the different estimation methodologies for the Van der Pol experiment.

Table 7.3: Elapsed Time Comparison for Van der Pol Experiment

Method	% Data	Time (s)	% Time of GD
GD	100%	24.58	100.00
SGD	50%	10.30	41.92
SGD	30%	6.75	27.46
SGD	10%	2.94	11.95
kSGD	50%	9.60	39.06
kSGD	30%	5.70	23.19
kSGD	10%	3.18	12.94

Marginalized Optimization

8 | Statistical Filtering & Optimization

As discussed in Chapter 1, from statistics to control, a ubiquitous optimization problem is

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}\left[f(\theta, \xi)\right], \tag{8.1}$$

where ξ is an $\mathbb{R}^d$-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$; $\mathbb{E}[\cdot]$ denotes the expectation operator with respect to the probability space; and $f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function and is sufficiently regular such that $\nabla_\theta \mathbb{E}\left[f(\theta, \xi)\right] = \mathbb{E}\left[\nabla_\theta f(\theta, \xi)\right]$. Currently, this optimization problem is addressed using three broad approaches, each with its unique advantages and challenges.

1. In the sample average approximation (SAA) approach, the objective function is replaced with an empirical mean generated from a sample of the random variable (Shapiro et al., 2009). The resulting objective function is deterministic and can be solved with mature deterministic solvers. However, for more complex variants of (8.1), the SAA objective function with only, say, a thousand samples becomes prohibitively expensive to manipulate per iteration (e.g., Kouri and Surowiec, 2016).

2. In the Bayesian Optimization approach, the objective function is nonparametrically estimated using a Gaussian process prior that is updated at each iteration using an evaluation of the function f at some point in $\mathbb{R}^p$ (Brochu et al., 2010). Under mild conditions, the estimated objective function converges to the actual objective function in the limit. However, the estimated objective function update requires inverting a dense matrix whose dimension is equal to the iteration number, which becomes prohibitively expensive as the number of iterations grows. Moreover, this issue is exacerbated if the gradient function is also using a Gaussian process.

3. In the Incremental Estimation approach (e.g., Stochastic Gradient Descent), the ob-

jective function is ignored and sampled gradient information is used exclusively to perform the optimization (Bottou et al., 2016). The incremental estimation approach is distinguished by its inexpensive per iteration costs and simplicity of use. However, as we have seen, while the iterates can be shown to converge to a stationary point, such methods are extremely difficult to tune and they do not provide stopping criteria (with the exception of kSGD).

Despite the challenges of these three approaches, each has highly desirable advantages such as the ability to use deterministic solvers, the convergence of surrogates to the original problem, and inexpensive per iteration costs. In an ideal situation, there would be an optimization methodology that retains all of these benefits. Here, we introduce a novel framework for developing such solvers. Our key insight is to treat the objective function and the gradient function as a Hidden Markov Model *over function spaces*, and to integrate statistical filters and deterministic iterative solvers to practically estimate and minimize (8.1).

The remainder of this chapter describes this framework. In Section 8.1, we review the basic principles of HMMs. In Section 8.2, we develop our framework. In Section 8.3, we demonstrate three simple optimizers that can be derived from our framework against state-of-the-art approaches for a problem from statistics, a problem from machine learning, and a problem from stochastic optimal control.

8.1 Principles of Statistical Filtering

A statistical filter is an estimation methodology for a Hidden Markov Model (HMM) (Simon, 2006; Shumway and Stoffer, 2006). A basic HMM is defined by three components: an initial distribution, μ, a Markov transition kernel, p, and an observation distribution, q. The first two components define a Markov Chain, $\{X_0, X_1, X_2, \ldots\} \subset \mathbb{R}^h$, where $X_0 \sim \mu$ and $\mathbb{P}\left[X_{i+1} \in A \middle| X_i\right] = p(X_i, A)$ for a measurable set A. The third component defines a sequence of observations $\{Y_1, Y_2, \ldots\} \subset \mathbb{R}^b$ such that $\mathbb{P}\left[Y_i \in B \middle| X_i\right] = q(B|X_i)$ for a measurable set B.

The sequential estimation of a HMM follows from Bayes' Rule. In particular, given a

state X_i and an observation Y_{i+1}, the distribution of X_{i+1} is

$$\mathbb{P}\left[X_{i+1}\middle|\, X_i, Y_{i+1}\right] \propto \mathbb{P}\left[Y_{i+1}\middle|\, X_i\right]\mathbb{P}\left[X_{i+1}\middle|\, X_i\right], \tag{8.2}$$

for $i \geq 0$. In general, (8.2) does not have a closed form and must be propagated by a numerical approximation method. Moreover, (8.2) relies on knowing X_i, which is estimated from X_{i-1} by (8.2), which, in turn, relies on an estimate of X_{i-2} by (8.2). Hence, estimating X_{i+1} requires a numerical method that not only approximates (8.2), but also propagates the uncertainty in the estimates from $X_i, X_{i-1}, \ldots, X_0$. When such an estimation is performed, the estimation procedure is called a statistical filter. Popular statistical filters for this case include the extended Kalman Filter, the unscented Kalman Filter, the ensemble Kalman Filter and the Particle Filter, which are discussed in detail by Simon (2006).

Under certain circumstances, (8.2) can be propagated in closed form, which gives rise to the Kalman Filter. To describe the Kalman Filter, consider the situation where the transitions between hidden states are defined by $X_{i+1} = f(X_i, i) + \epsilon_i$, where $f : \mathbb{R}^h \times \mathbb{N} \cup \{0\} \to \mathbb{R}^h$; and $\epsilon_i \sim \mathcal{N}(0, \Sigma_i)$ are non-degenerate, independent random variables for all $i \geq 0$. Moreover, consider the observations defined by $Y_i = LX_i + \eta_i$, where $L \in \mathbb{R}^{h\times b}$; and $\eta_i \sim \mathcal{N}(0, \Gamma_i)$ are non-degenerate, independent random variables for all $i > 0$. Then, using Bayes' rule, we have that $\mathbb{P}\left[X_{i+1}\middle|\, X_i, Y_{i+1}\right]$ is

$$\mathcal{N}\left(f(X_i, i) + \Sigma_i L'\left(\Gamma_{i+1} + L\Sigma_i L'\right)^{-1}\left[Y_{i+1} - Lf(X_i, i)\right], \left[L'\Gamma_{i+1}^{-1}L + \Sigma_i^{-1}\right]^{-1}\right). \tag{8.3}$$

Thus, we have a description for X_{i+1} that involves another quantity, X_i.

However, as mentioned above, we must also propagate the uncertainty of the estimate of X_i into X_{i+1}. Unfortunately, this cannot be done in closed form unless $f(X_i, i)$ is a linear function of X_i; in which case, the uncertainty can be propagated in closed form, and $\mathbb{P}\left[X_{i+1}\middle|\, Y_{i+1}, Y_i, \ldots, Y_1, X_0\right]$ has the form of (8.3) but with Σ_i replaced by $\tilde{\Sigma}_i$ that contains

the propagated uncertainty of the estimate of X_i. This estimation procedure is called the Kalman Filter.

The Kalman Filter has several glaring challenges. First, the Kalman Filter requires that X_0 is known precisely, or, more generally, that the mean of X_0 and its covariance are known when X_0 is a random variable. Fortunately, the Kalman Filter is robust to this lack of precision (Patel, 2016). Second, the Kalman Filter requires that the underlying relationship between X_i and X_{i+1} (i.e., $f(X_i, i)$) is linear, which is a highly specialized case. However, the Kalman Filter can be "extended" to the case of nonlinearity by using the Jacobian of $f(x, i)$ evaluated at X_i in $\tilde{\Sigma}_i$ and ignoring the additional variance added by the Jacobian. In general, this (Extended) Kalman Filter, has been extremely successful in practice (see Simon, 2006).

Importantly, the Kalman Filter's success is not limited to replacing nonlinear relationships with an approximate linear relationship, but also to a majority of cases with systematic errors in the relationships between the hidden states, and to a majority of cases with the inclusion of deterministic chaos in the hidden state dynamics (Simon, 2006). Intuitively, the Kalman Filter's ability to navigate these errors and approximations in the systematic error comes down to reinterpreting the Kalman Filter's estimate of X_{i+1} as a compromise between the approximate dynamics, which serve as a regularizing force in the estimation, and the observations, which serve as a corrective force in the estimation. Specifically, we can restate the Kalman Filter estimate, $\hat{X}_{i+1}$, as a proximal operator,

$$\hat{X}_{i+1} = \text{argmin}_z \left\{ \|Y_{i+1} - LZ\|^2_{\Gamma^{-1}_{i+1}} + \|Z - f(X_i, i)\|^2_{\tilde{\Sigma}_i^{-1}} \right\}. \tag{8.4}$$

From (8.4), the Kalman Filter is clearly balancing the information from the observation with information from the dynamics by their relative variances, Γ_i and $\tilde{\Sigma}_i$. Therefore, the Kalman Filter can reduce emphasis on inaccuracies in f by inflating its variance and giving more responsibility to Y_{i+1} in determining X_{i+1}. This interpretation of the Kalman Filter will be

essential to the development of our framework.

8.2 Our Framework

Recall, in the previous section, we reviewed the principles behind HMMs and Statistical filters. In this section, we extend the HMMs and Statistical filters for the purposes of optimization. As a stepping stone, we first develop this extension of HMMs and Statistical filters for functions along an iterate sequence.

8.2.1 Statistical Filtering of Functions along an Iterate Sequence

We begin by carefully developing the HMM that we will use to solve (8.1), and then we will develop the statistical filtering procedure for estimating this particular HMM. We conclude with several stability and convergence results.

The Hidden Markov Model. To specify the HMM, we need to define (a) the states, (b) the relationship between the states, and (c) the observations. While the definition of the observations is rather straightforward, the key insights are in how we define the states of the HMM and how we define the relationship between the states.

To define the states, let $\{\theta_0, \theta_1, \ldots\} \subset \mathbb{R}^p$. We define the states by

$$\{\mathbb{E}\left[f(\theta_i, \xi)\right], \mathbb{E}\left[\nabla f(\theta_i, \xi)\right] : i = 0, 1, \ldots\}, \tag{8.5}$$

where the gradients are evaluated with respect to the argument θ and evaluated at θ_i. Moreover, we define the relationship between states by the approximation by

$$\begin{bmatrix} \mathbb{E}\left[f(\theta_{i+1}, \xi)\right] \\ \mathbb{E}\left[\nabla f(\theta_{i+1}, \xi)\right] \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[f(\theta_i, \xi)\right] \\ \mathbb{E}\left[\nabla f(\theta_i, \xi)\right] \end{bmatrix} + \begin{bmatrix} \mathbb{E}\left[\nabla f(\theta_i, \xi)\right]' \\ 0 \end{bmatrix} (\theta_{i+1} - \theta_i) + \begin{bmatrix} \epsilon_i \\ \lambda_i \end{bmatrix}, \tag{8.6}$$

where the dynamics for the objective function are given by Taylor's theorem; and ϵ_i and λ_i are terms representing the systematic errors incurred by the approximation. We make two

important remarks regarding (8.6). First, we can readily integrate higher-order derivatives into (8.6). Second, we can easily bound the approximation errors, ϵ_i and λ_i, as we now state.

Lemma 8.1. *Suppose that* $\nabla \mathbb{E}\left[f(\theta, \xi)\right]$ *is* L*-Lipschitz continuous. Then,*

$$|\epsilon_i| \leq \frac{1}{2} L \left\|\theta_{i+1} - \theta_i\right\|_2^2 \text{ and } \left\|\lambda_i\right\|_2 \leq L \left\|\theta_{i+1} - \theta_i\right\|_2 . \tag{8.7}$$

Proof. By assumption,

$$\left\|\lambda_i\right\|_2 = \left\|\nabla \mathbb{E}\left[f(\theta_{k+1}, \xi)\right] - \nabla \mathbb{E}\left[f(\theta_k), \xi\right]\right\|_2 \leq L \left\|\theta_{i+1} - \theta_i\right\|_2 . \tag{8.8}$$

Moreover, by Taylor's Theorem,

$$\begin{aligned} |\epsilon_i| &= \left|\mathbb{E}\left[f(\theta_{k+1}, \xi)\right] - \mathbb{E}\left[f(\theta_k, \xi)\right] - (\theta_{k+1} - \theta_k)' \nabla \mathbb{E}\left[f(\theta_k, \xi)\right]\right| \\ &= \left| \int_0^1 (x_{k+1} - x_k)' \left(\nabla \mathbb{E}\left[f(\theta_k + t(\theta_{k+1} - \theta_k), \xi)\right] - \nabla \mathbb{E}\left[f(\theta_k, \xi)\right]\right) dt \right| \\ &\leq L \left\|\theta_{k+1} - \theta_k\right\|_2^2 \int_0^1 t dt. \end{aligned} \tag{8.9}$$

■

For the observations and their relationship to the states, we observe the concatenation of $f(\theta_i, \xi)$ and $\nabla f(\theta_i, \xi)$ (for independent copies of ξ, when we make sequential observations), which we assume are unbiased estimates of the objective function and its gradient. Moreover, we assume that the joint variance of these observations is given by $\Gamma(\theta_i) \succ 0$. Note, we will sometimes write $\Gamma_i = \Gamma(\theta_i)$. To reiterate, (8.5), (8.6) and the observations define a HMM for the objective function and its gradient *along a specific sequence of iterates* with approximate dynamics.

The Statistical Filter. We can now apply the (extended) Kalman Filter to estimate our HMM. The filter has three components that were implicitly described in (8.4) that we now

define separately: the initial state estimate, $\{\hat{F}_0(\theta_0), \hat{G}_0(\theta_0)\}$, and its covariance, Σ_0; the analysis states, $\{\hat{F}_i^a(\theta_i), \hat{G}_i^a(\theta_i) : i \in \mathbb{N}\}$, and their covariances, $\{\bar{\Sigma}_i : i \in \mathbb{N}\}$; and the filtered states, $\{\hat{F}_i(\theta_i), \hat{G}_i(\theta_i) : i \in \mathbb{N}\}$, and their covariances, $\{\Sigma_i : i \in \mathbb{N}\}$.

Let $\{\xi_i : i = 0, 1, \ldots\}$ be independent random variables with the distribution of ξ. We now define the initial state as $\hat{F}_0(\theta_0) = f(\theta_0, \xi_0)$ and $\hat{G}_0(\theta_0) = \nabla f(\theta_0, \xi_0)$, from which it follows that the initial state has variance $\Sigma_0 = \Gamma_0$. From (8.6), we define the analysis states by

$$\begin{bmatrix} \hat{F}_i^a(\theta_i) \\ \hat{G}_i^a(\theta_i) \end{bmatrix} = \begin{bmatrix} \hat{F}_{i-1}(\theta_{i-1}) \\ \hat{G}_{i-1}(\theta_{i-1}) \end{bmatrix} + \begin{bmatrix} \hat{G}_{i-1}(\theta_{i-1})' \\ 0 \end{bmatrix} (\theta_i - \theta_{i-1}). \tag{8.10}$$

The covariance for the analysis states are computed in the following, straightforward result.

Proposition 8.1. *Given the covariance of $(\hat{F}_i(x_i), \hat{G}_i(x_i))$, Σ_i, the covariance for the analysis state, $(\hat{F}_{i+1}^a(x_{i+1}), \hat{G}_{i+1}^a(x_{i+1}))$, is*

$$\bar{\Sigma}_{i+1}(h) = \begin{bmatrix} 1 & h' \\ 0 & I \end{bmatrix} \Sigma_i \begin{bmatrix} 1 & 0 \\ h & I \end{bmatrix}, \tag{8.11}$$

where $h = \theta_{i+1} - \theta_i$. Moreover, if the filtered state is normally distributed, then the analysis state is normally distributed.

Proof. We can rewrite (8.10) as

$$\begin{bmatrix} \hat{F}_i^a(\theta_i) \\ \hat{G}_i^a(\theta_i) \end{bmatrix} = \begin{bmatrix} \hat{F}_{i-1}(\theta_{i-1}) \\ \hat{G}_{i-1}(\theta_{i-1}) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ h & I \end{bmatrix}, \tag{8.12}$$

from which the covariance formula and normal distribution properties (if the filtered state is normally distributed) follow. ■

While Proposition 8.1 gives an exact form for the covariance of the analysis states, this is not how we will use it. In view of (8.6), (8.10) is approximate and, following from the

discussion in Section 8.1, we will need to modify $\tilde{\Sigma}_i$ to account for the approximate state relationship. Specifically, we add to the covariance in Proposition 8.1 a symmetric, positive definite matrix $T_i(h)$. We will refer to the inflated version as $\tilde{\Sigma}_{i+1}(h)$.

Finally, the filtered state follows from (8.3) with $L = I$ and Σ_{i-1} replaced by $\tilde{\Sigma}_i = \tilde{\Sigma}_i(\theta_{i+1} - \theta_i)$, which is given by

$$\begin{bmatrix} \hat{F}_i(\theta_i) \\ \hat{G}_i(\theta_i) \end{bmatrix} = \underset{z \in \mathbb{R}^{d+1}}{\arg\min} \left\{ \left\| z - \begin{bmatrix} f(\theta_i, \xi_i) \\ \nabla f(\theta_i, \xi_i) \end{bmatrix} \right\|^2_{\Gamma_i^{-1}} + \left\| z - \begin{bmatrix} \hat{F}_i^a(\theta_i) \\ \hat{G}_i^a(\theta_i) \end{bmatrix} \right\|^2_{\tilde{\Sigma}_i^{-1}} \right\}. \tag{8.13}$$

Proposition 8.2. *Let $h = \theta_i - \theta_{i-1}$. Given the analysis state's covariance, $\bar{\Sigma}_i = \bar{\Sigma}_i(h)$, and denoting $T_i = T_i(h) = \tilde{\Sigma}_i - \bar{\Sigma}_i$, the covariance of the filtered state, defined by (8.13), is*

$$\left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right)^{-1} - \left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right)^{-1} \tilde{\Sigma}_i^{-1} T_i \tilde{\Sigma}_i^{-1} \left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right)^{-1}. \tag{8.14}$$

Moreover, if the analysis is normally distributed, then the filtered state is normally distributed.

Proof. By solving (8.13), we have that

$$\left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right) \begin{bmatrix} \hat{F}_i(\theta_i) \\ \hat{G}_i(\theta_i) \end{bmatrix} = \Gamma_i^{-1} \begin{bmatrix} f(\theta_i, \xi_i) \\ \nabla f(\theta_i, \xi_i) \end{bmatrix} + \tilde{\Sigma}_i^{-1} \begin{bmatrix} \hat{F}_i^a(\theta_i) \\ \hat{G}_i^a(\theta_i) \end{bmatrix}. \tag{8.15}$$

Therefore,

$$\begin{aligned} \left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right) \Sigma_i \left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right) &= \Gamma_i^{-1} + \tilde{\Sigma}_i^{-1} \bar{\Sigma}_i \tilde{\Sigma}_i^{-1} \\ &= \Gamma_i^{-1} + \tilde{\Sigma}_i^{-1} + \tilde{\Sigma}_i^{-1} \left(\bar{\Sigma}_i - \tilde{\Sigma}_i\right) \tilde{\Sigma}_i^{-1}, \end{aligned} \tag{8.16}$$

from which the calculation follows. ∎

Stability and Convergence of Estimates. The approximations used in defining the HMM and the statistical filter raise the question of whether our estimates actually track the objective and gradient function. The following two results describe the behavior of the estimates within the relevant optimization context, that is, when $\{\theta_0, \theta_1, \ldots\}$ is a converging sequence. The first result states that when our approximation-motivated covariance correction term, $T_i(h)$ is $c\|h\|_2 I$ for some $c > 0$, we can guarantee that the variance of our estimates decays to zero. The second result states that when our covariance correction term is cI for $c > 0$, then we can guarantee that the bias of our estimates decays to zero. Together, these two results suggest that there is some limiting bias-variance trade-off that prevents the convergence of our estimators. We aim to address this concern in future efforts.

Theorem 8.1. *Let $\{\theta_0, \theta_1, \ldots\} \subset \mathbb{R}^d$ be a sequence converging to θ^*. Moreover, suppose that $\exists \gamma \in (0, \infty)$ such that $\sup\{\|\Gamma(\theta_i)\|_2\} < \gamma$. Now, let the analysis states be given by (8.10) and filtered states be given by (8.13), where $\tilde{\Sigma}_i(h) - \bar{\Sigma}_i(h) = c\|h\|_2 I$ for some $c > 0$. Then, $\lim_{i\to\infty} \|\Sigma_i\|_2 = 0$.*

Proof. Let $h_i = \theta_{i+1} - \theta_i$. We will begin by establishing several facts, and then we will prove the main result. By Proposition 8.1 and Lemma 8.2,

$$\left\|\tilde{\Sigma}_{i+1}\right\|_2 \leq \|\Sigma_i\|_2 (1 + \|h_i\|_2 + \|h_i\|_2^2) + c\|h_i\|_2 . \tag{8.17}$$

By Proposition 8.2 and Lemma 8.3,

$$\|\Sigma_{i+1}\|_2 \leq \frac{\left[\|\Sigma_i\|_2 \left(1 + \|h_i\|_2 + \|h_i\|_2^2\right) + c\|h_i\|_2\right]\gamma}{\left[\|\Sigma_i\|_2 \left(1 + \|h_i\|_2 + \|h_i\|_2^2\right) + c\|h_i\|_2\right] + \gamma} \tag{8.18}$$

Therefore, bounding the right hand side of (8.18) by $\|\Sigma_i\|_2 \frac{\gamma}{\delta+\gamma}$, we conclude that for $\delta_i \in$

$[0, \|\Sigma_i\|_2)$, if $\|h_i\|_2$ satisfies

$$\begin{aligned} 0 \geq \|\Sigma_i\|_2 \left[\gamma + \delta_i - \|\Sigma_i\|_2\right] \|h_i\|_2^2 + \left[\gamma + \delta_i - \|\Sigma_i\|_2\right] \left[c + \|\Sigma_i\|_2\right] \|h_i\|_2 \\ + \|\Sigma_i\| \left[\delta_i - \|\Sigma_i\|_2\right], \end{aligned} \tag{8.19}$$

then $\|\Sigma_{i+1}\|_2 \leq \|\Sigma_i\|_2 \frac{\gamma}{\delta+\gamma} \leq \|\Sigma_i\|_2$.

We are now ready to prove the result. For a contradiction, suppose that $\{\|\Sigma_i\|_2\}$ does not converge to 0. Then, there is an $\epsilon > 0$ such that for a subsequence $\{i_k\}$, $\|\Sigma_{i_k}\|_2 > 2\epsilon$. Then, with $\delta_{i_k} = \epsilon$, if $\|h_{i_k}\|_2$ satisfies (8.19), then $\|\Sigma_{i_k}\|_2 \leq \|\Sigma_{i_k-1}\|_2 \frac{\gamma}{\epsilon+\gamma} < \|\Sigma_{i_k-1}\|_2$.

Because $\{\theta_k\}$ is a convergent sequence, $\|h_i\| \to 0$. Therefore, for sufficiently large j, if $k \geq j$ then $\|h_{i_k}\|_2$ will satisfy (8.19) and so $\|\Sigma_{i_k}\|_2 \leq \|\Sigma_{i_k-1}\|_2$. Therefore, we conclude that for sufficiently large j, if $i \geq j$ then $\|h_i\|_2$ will satisfy (8.19) and $\|\Sigma_i\| > 2\epsilon$.

However, for $i \geq j$, because (8.19) is satisfied,

$$\|\Sigma_i\|_2 \leq \|\Sigma_j\|_2 \left(\frac{\gamma}{\epsilon + \gamma}\right)^{i-j} \tag{8.20}$$

which will be smaller than 2ϵ for $i - j$ large enough. This is our contradiction, and so $\|\Sigma_i\|_2 \to 0$. ■

Theorem 8.2. *Let $\{\theta_0, \theta_1, \ldots\} \subset \mathbb{R}^d$ be a sequence converging to θ^*. Moreover, suppose that $\exists \gamma \in (0, \infty)$ such that $\sup\{\|\Gamma(x_i)\|_2\} < \gamma$. Now, let the analysis states be given by (8.10) and filtered states be given by (8.13), where $\tilde{\Sigma}_i(h) - \bar{\Sigma}_i(h) = cI$ for some $c > 0$. If $\nabla \mathbb{E}[f(\cdot, \xi)]$ is Lipschitz continuous then the bias of $\hat{F}_j(\theta_j)$ and $\hat{G}_j(\theta_j)$ converge to zero.*

Proof. First, let B_i denote the Euclidean norm of the bias of $(\hat{F}_i(x_i), \hat{G}_i(x_i))$. Then, using Lemmas 8.1 and 8.2,

$$B_i \leq \left\|\Gamma_i(\Gamma_i + \tilde{\Sigma}_i)^{-1}\right\|_2 \left[(1 + \|h_i\|_2)B_{i-1} + \frac{1}{2}L\|h_i\|_2^2 + L\|h_i\|_2\right]. \tag{8.21}$$

Now, since $\tilde{\Sigma}_i \succeq cI$, using Lemma 8.4, we have that

$$B_i \leq \frac{\gamma(1+\|h_i\|_2)}{\gamma+c}B_{i-1} + \frac{\gamma}{\gamma+c}\left[\frac{1}{2}L\|h_i\|_2^2 + L\|h_i\|_2\right]. \tag{8.22}$$

Finally, since $\|h_i\|_2 \to 0$, we have that $B_i \to 0$. ■

8.2.2 Statistical Filtering for Optimization

With the developments of the previous subsection, we are now ready to state our novel methodology for solving (8.1) that avoids the disadvantages of current techniques and retains their advantage. We will state this methodology in two steps.

First, we state how to extend the statistical filter to function spaces for optimization. Suppose we have $(\hat{F}_{i-1}(\theta_{i-1}), \hat{G}_{i-1}(\theta_{i-1}))$ with variance Σ_{i-1}. Then, we define function $\hat{F}_i(\theta)$ and $\hat{G}_i(\theta)$ for each θ by

$$\begin{bmatrix}\hat{F}_i(\theta)\\ \hat{G}_i(\theta)\end{bmatrix} = \underset{z\in\mathbb{R}^{d+1}}{\arg\min}\left\{\left\|z - \begin{bmatrix}f(\theta,\xi_i)\\ \nabla f(\theta,\xi_i)\end{bmatrix}\right\|^2_{\Gamma(\theta)^{-1}} + \left\|z - \begin{bmatrix}\hat{F}_i^a(\theta)\\ \hat{G}_i^a(\theta)\end{bmatrix}\right\|^2_{\tilde{\Sigma}(\theta-\theta_{i-1})^{-1}}\right\}. \tag{8.23}$$

By Lemma 8.1, (8.23) is only useful locally about θ_{i-1}. However, this is exactly what we need because we will use an iterative optimization procedure, which is inherently local, to define θ_i.

That is, for the second step, we define θ_i to be an exact or approximate solution of

$$\begin{aligned}&\min_{\theta}\ \hat{F}_i(\theta)\\ &\text{subject to: } e_1'\Sigma_i(\theta)e_1 \leq \gamma e_1'\Sigma_i(\theta_{i-1})e_1,\end{aligned} \tag{8.24}$$

where $e_1 \in \mathbb{R}^{p+1}$ is the basis vector that has a one in its first component; and $\gamma > 1$. Effectively, the constraint requires that the variance of the objective does not grow too severely. Indeed, by the definition of $\tilde{\Sigma}_i(\theta)$, we can guarantee that there is a neighborhood

of θ_{i-1} in which this condition is guaranteed to hold. Therefore, we are guaranteed that a standard optimization solver can be used to solve (8.24) either completely or inexactly.

Once we have generated θ_i, we can now repeat the procedure with $\hat{F}_i(\theta_i)$ and $\hat{G}_i(\theta_i)$ to, first, determine $\hat{F}_{i+1}(\theta)$ and $\hat{G}_{i+1}(\theta)$ using the statistical filter, and, second, to compute θ_{i+1} using a standard deterministic optimizer. Thus, we have described a complete optimization methodology for addressing (8.1).

8.2.3 Technical Lemmas

Lemma 8.2. *Let $h \in \mathbb{R}^p$. Then,*

$$\left\| \begin{bmatrix} 1 & h' \\ 0 & I \end{bmatrix} \right\|_2 \leq \sqrt{1 + \|h\|_2 + \|h\|_2^2}. \tag{8.25}$$

Proof. Let H denote the matrix of interest. Recall that $\|H\|_2^2 = \|H * H'\|_2$, where

$$HH' = \begin{bmatrix} 1 + \|h\|_2^2 & h' \\ h & I \end{bmatrix}. \tag{8.26}$$

Let $\lambda, x \in \mathbb{R}$, $y \in \mathbb{R}^p$ such that

$$\begin{cases} (1 - \lambda + \|h\|_2^2)x + h'y = 0 \\ xh + (1 - \lambda)y = 0, \end{cases} \tag{8.27}$$

then λ is an eigenvalue of HH' and the concatenation of x and y is an eigenvector of HH'. From the second equation of (8.27), if $x \neq 0$ then $y = h$ and $1 - \lambda = -x$. Plugging this into the first equation, we have that $x^2 - \|h\|_2^2 x - \|h\|_2^2 = 0$, which we can solve to conclude that both of the eigenvalues corresponding to the case when $x \neq 0$ are smaller than $1 + \|h\|_2 + \|h\|_2^2$ when $x \neq 0$. Now, when $x = 0$, then $\lambda = 1$ and y must be orthogonal to h. Thus, we have accounted for $p + 1$ eigenvalues of HH'. Therefore, $\|H\|_2 \leq \sqrt{1 + \|h\|_2 + \|h\|_2^2}$. ■

Lemma 8.3. *Suppose A, B are symmetric, invertible matrices of the same dimension. Let $\|A\|_2 \leq a < \infty$ and $\|B\| \leq b < \infty$. Then,*

$$\left(A^{-1} + B^{-1}\right)^{-1} \preceq \frac{ab}{a+b} I. \tag{8.28}$$

Proof. By definition, $A^{-1} \succeq a^{-1} I$ and $B^{-1} \succeq b^{-1} I$. Hence, $A^{-1} + B^{-1} \succeq (a^{-1} + b^{-1}) I$. Therefore, $(A^{-1} + B^{-1})^{-1} \preceq (a^{-1} + b^{-1})^{-1} I$. ■

Lemma 8.4. *Suppose A, B are symmetric, invertible matrices of the same dimension such that $\|A\| \leq a < \infty$ and $0 \prec bI < B$. Then,*

$$A(A+B)^{-2}A \preceq \left(\frac{a}{a+b}\right)^2 I. \tag{8.29}$$

Therefore, $\|A(A+B)^{-1}\|_2 \leq \frac{a}{a+b}$.

Proof. Note, $I \succeq a^{-1}A$. Therefore, $B \succeq bI \succeq (b/a)A$. Hence, $A + Bsucceq(1 + b/a)A$. Therefore,

$$A(A+B)^{-2}A \succeq A\left(\frac{a}{a+b}\right)^2 A^{-2}A. \tag{8.30}$$

Moreover, $\|A(A+B)^{-1}\|_2 = \sqrt{\lambda_{\max}(A(A+B)^{-1}(A+B)^{-1}A)}$, from which the rest of the result follows. ■

8.3 Numerical Experiments

Here, we compare a naive optimizer generated by our framework against standard optimizers used to solve (8.1) as it appears in three different application areas: statistics, machine learning and stochastic optimal control.

8.3.1 Estimation of Generalized Linear Mixed Effect Models

Consider a study of 250 patients with the same disease who are each treated by one of 91 different physicians. In this study, we record the patient's state after treatment (not cured

or cured); gender, g, (male or female); disposition prior to treatment, d, (not optimistic or optimistic); level of exercise, e (no exercise or does exercise); and the treating physician. Based on the outcomes of our study, we want to state a model that describes the probability of being cured based on the patient's disposition and level of exercise, while accounting for the possible variations between physicians (not just our 91 physicians). In this case, we might choose to model the probability of being cured as

$$p_{cured} = \frac{1}{1 + \exp\left[\beta_0 + \beta_d d + \beta_e e + \beta_g g + \rho\right]},$$

where $\beta_0, \beta_d,\ \beta_e$ and β_g are unknown coefficients; d is one if the patient's disposition is optimistic and zero otherwise; e is one if the patient exercises at all and is zero otherwise; g is one if the patient is a female and zero otherwise; and ρ is a mean-zero, normally distributed random variable with an unknown variance σ_ρ^2 that represents the random deviation from the systematic behavior (described by the intercept, patient's behavior and patient's exercise level) based on the physician (see Bates et al. (2015) for short overview of such models).

Let $p_{cured}^{(i)}$ denote the probability of patient i being cured as described by the model and let $y^{(i)}$ denote the state of the patient at the end of the treatment; specifically $y^{(i)} = 1$ if the patient is cured and is zero otherwise. Then, the likelihood function is

$$\mathcal{L}(\beta_0, \beta_d, \beta_e, \sigma_\rho) = \mathbb{E}\left[\prod_{i=1}^{250} (p_{cured}^{(i)})^{y^{(i)}} (1 - p_{cured}^{(i)})^{1-y^{(i)}}\right]. \tag{8.31}$$

We compute our estimates by minimizing $-\mathcal{L}(\beta_0, \beta_d, \beta_d, \sigma_\rho)$, which, in our notation is an example of (8.1).

Techniques for maximizing (8.31) are of two varieties: a finite approximation to the integral, or an approximation of the integrand that results in an analytic form of the objective function (Tuerlinckx et al., 2006). Unfortunately, these techniques are known to perform poorly when the dimension of the random components increase, when a normal approx-

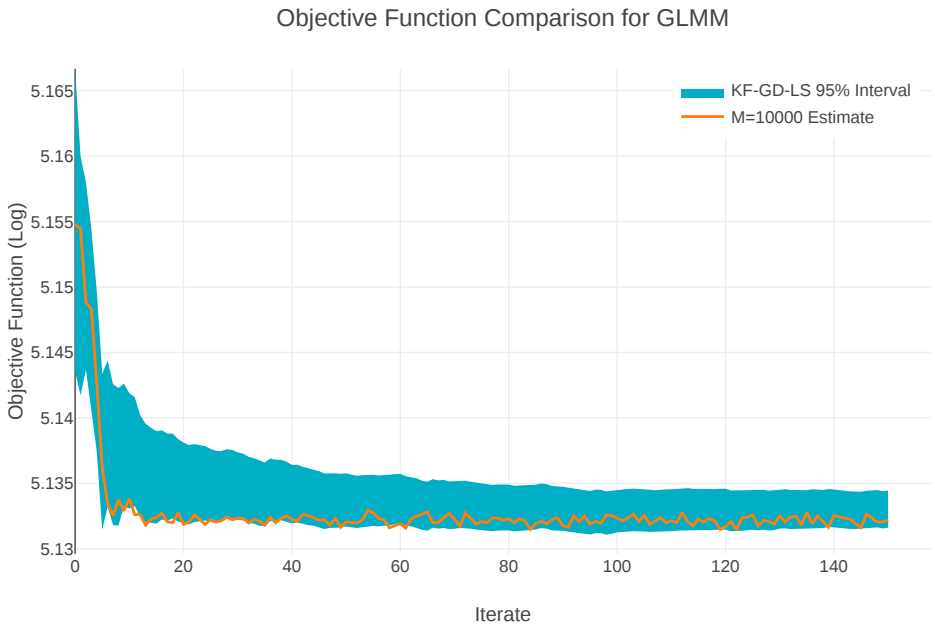
imation to the integrand is invalid, and when using a large number of groups with few measurements per group (Tuerlinckx et al., 2006).

Experimental Setup. For our experiment, we simulate this study and compare two techniques for solving (8.31). We use the de facto standard solver implemented in the R Programming Language in the package `lme4` (Bates et al., 2015). Using our framework, we generate an optimizer that leverages the Kalman Filter with $T(h) = c\|h\|_2 I$ with $c > 0$ to generate the subproblems and Gradient Descent with Line Search to solve the subproblem, which we will denote by KF-GD-LS. The observations for our solver are generated by a Monte Carlo approximation to the integral in (8.31) with 100 random samples. In order to compare the techniques to some "truth" we compute (8.31) and its derivative using a Monte Carlo approximation with 10,000 random samples, which has a variance on the order of 10^{-4}.

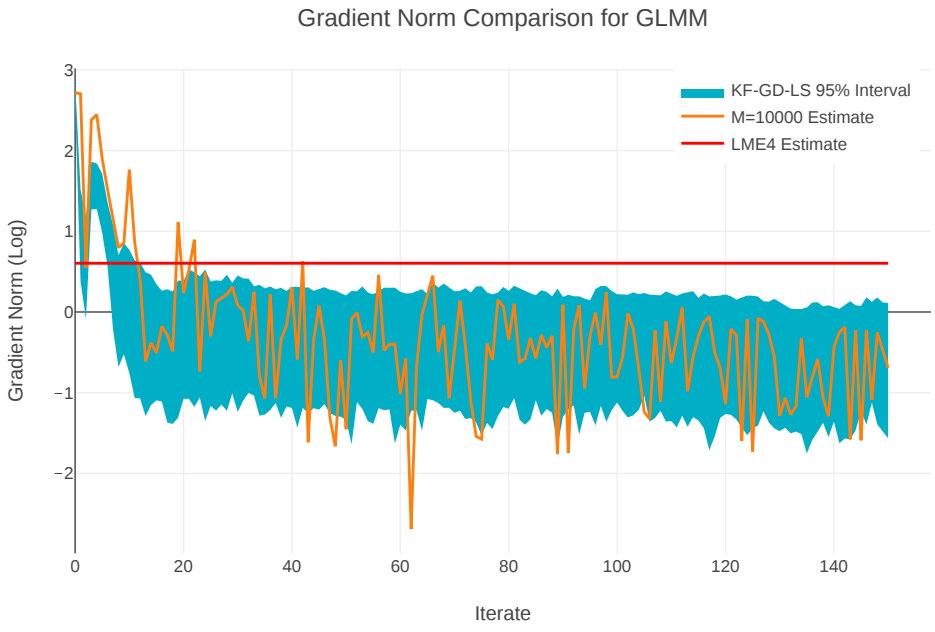
Results & Discussion. The comparison of our solvers to the "truth" and to the `lme4` solution is presented in Subsection 8.3.1. There are several notable features. First, despite the bias of KF-GD-LS, as predicted by Theorem 8.1, the optimizer's confidence interval seems to well approximate the "true" behavior of the objective function and its gradient. Second, based on the gradient norm, our naive approach drastically outperforms the `lme4` solver in finding a stationary point.

8.3.2 Empirical Risk Minimization for Training Machine Learning Models

Experimental Setup. We revisit the neutrino classification problem that was considered several times in the previous chapters. Again, we will be fitting a two-layer feed forward neural network just as before. Here, we train the neural network with stochastic gradient descent (SGD) with a decaying learning rate. Using our framework, we also train the network using a Kalman Filter with $T(h) = c\|h\|_2$ for $c > 0$ to generate the subproblems, and we solve the subproblem using a single gradient step with the same learning rate as for (SGD). For the iterates generated by the two methods, we report the training error and the testing error.



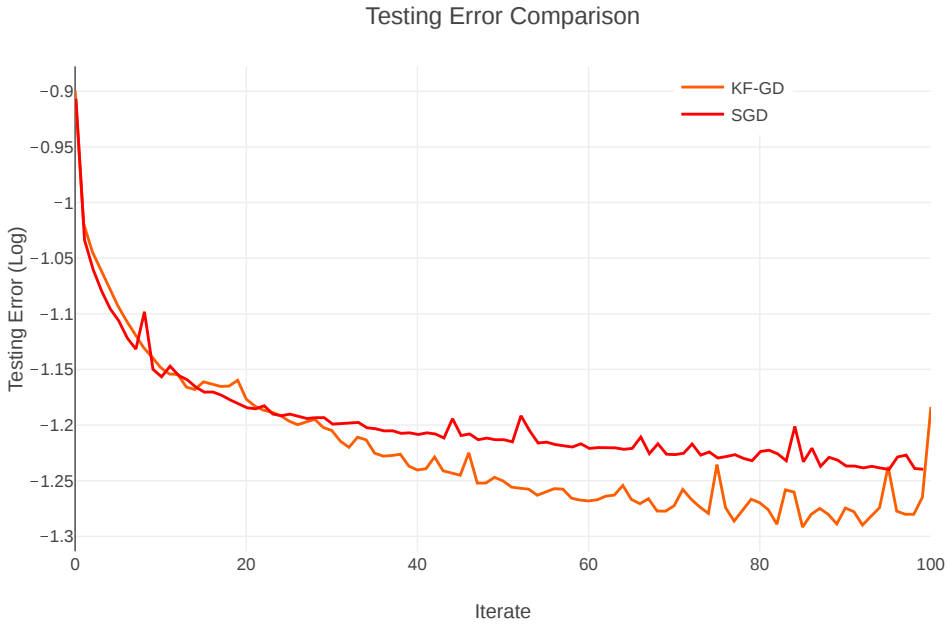
(a)



(b)

Figure 8.1: A comparison of the de facto solver and our KF-GD-LS approach. A 95% confidence interval is computed for the KF-GD-LS approach based on its covariance estimates.

(a)



(b)

Figure 8.2: A comparison of SGD and KF-GD for training a two-layer neural network in detecting neutrino signals.

Results & Discussion. The training error and testing error are reported in Fig. 8.2. We underscore two observations. First, the estimated training error of our optimizer matches quite well with the true training error generated by our iterates. Second, up until the end, our optimizer outperforms SGD despite having the same learning rate and observing random samples of the examples. Importantly, this improved performance exists for both the training and testing error, which indicates that our optimizers might confer some additional benefits in comparison to SGD.

8.3.3 Stochastic Optimal Control of Inventory

Consider the problem of planning inventory at a distribution center over a discretized time period $t = 0, 1, \ldots, T$. Let I_t denote the inventory at time t. If the value I_t is positive, then there is a surplus of inventory being stored at the distribution center, and this storage has a cost, $c_1 I_t^2$, associated with it. If the value I_t is negative, then there is a shortage of inventory being stored at the distribution center, and this shortage has an identical cost of $c_1 I_t^2$ associated with it. Moreover, suppose that the distribution center is subject to a random Poisson distributed demand, S_t, that is independent at each time point $t = 0, \ldots, T-1$. Additionally, suppose that the mean of S_t is a known value λ. Finally, suppose that facility controls its ability to receive inventory, P_t, at each time period from $t = 0, \ldots, T-1$. However, the receipt of inventory has a cost $c_2 P_t$ associated with it. To summarize, the inventory evolves with dynamics

$$I_{t+1} = I_t + P_t - S_t \tag{8.32}$$

with a total cost over the time interval

$$\sum_{t=0}^{T} c_1 I_t^2 + \sum_{t=0}^{T-1} c_2 P_t. \tag{8.33}$$

The goal of the planning then is to minimize the cost, which can be stated as

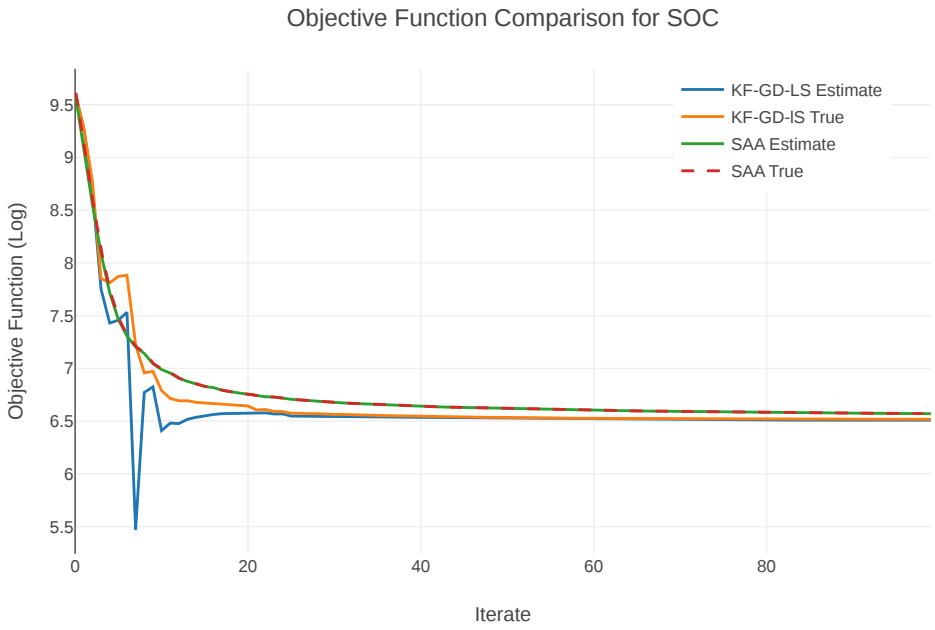
$$\begin{aligned} &\min_{P_0,\ldots,P_t,I_0} \mathbb{E}\left[\sum_{t=0}^{T} c_1 I_t^2 + \sum_{t=0}^{T-1} c_2 P_t\right] \\ &\text{subject to } \; I_{t+1} = I_t + P_t - S_t, \; t = 0,\ldots,T-1 \\ &\qquad\qquad P_t \geq 0, t = 0,\ldots,T-1. \end{aligned} \tag{8.34}$$

Clearly, (8.34) is an example of (8.1).

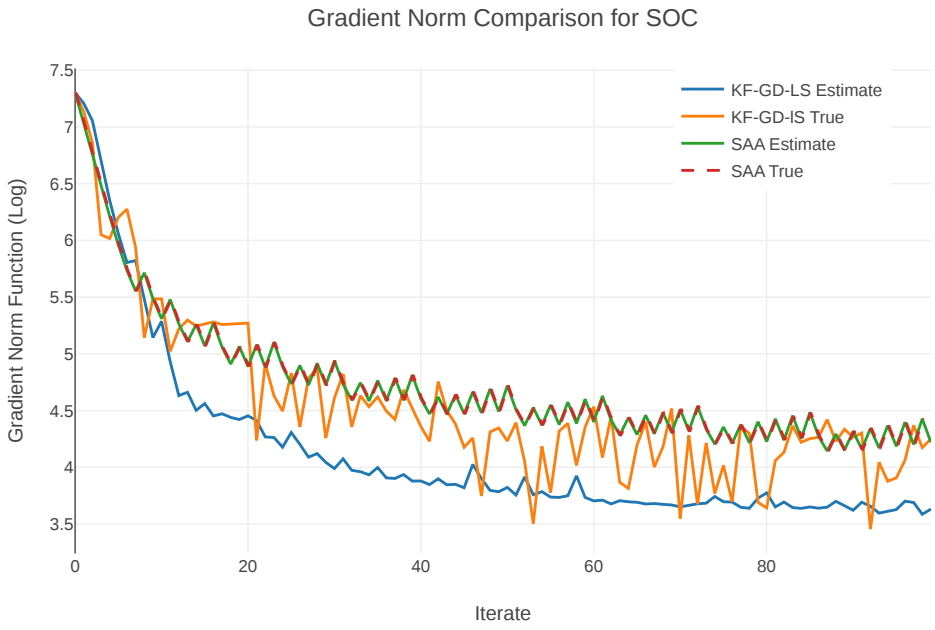
The most common technique for solving (8.34) makes use of sample average approximations to the objective function (Shapiro et al., 2009). The main drawbacks of this approach are that the original problem has been replaced by a sample approximation, which converges as the size of the sample tends to infinity, and that a sample average requires increasing the number of constraints that must be satisfied. Thus, the sample average approximation can drastically increase the dimensionality of the problem.

Experimental Setup. For our experiment, we simulate the stochastic inventory problem for arbitrary with $\lambda = 100$, $T = 100$, $c_1 = \$12$, and $c_2 = \$6$. We apply the SAA approach with $50,000$ samples of the stochastic demand process, and we minimize the resulting objective function using gradient descent with a line search subproblem that is solved using backtracking. Using our framework, we apply a Kalman Filter with $T(h) = c\|h\|_2$ with $c > 0$ to generate the subproblems, and we applying gradient descent with line search solved by backtracking up to one iteration to solve the subproblems. For our optimizer, we also only use $50,000$ samples. Importantly, we can compute the true objective function and gradient function in closed form owing to the specific form of the stochastic demand process, which we can use as the baseline "truth."

Results & Discussion. The results of this experiment are reported in Fig. 8.3. Both optimizers perform well in estimating the objective function. However, our optimizer does much worse at correctly tracking the gradient norm correctly. Despite this, our optimizer results



(a)



(b)

Figure 8.3: A comparison of SAA and KF-GD-LS for the stochastic inventory control problem.

in a savings of approximately $\$430,000$ for the distribution center in comparison to the SAA approach. Moreover, our optimizer is over fifty times faster than the SAA approach.

8.3.4 A Brief Summary

In the preceding three examples, we compared simple optimizers generated by our framework against standard optimizers. In all of these examples, we saw that our simple optimizers tended to outperform their counterparts and usually provided some additional benefit, such as the ability to track the objective function and the gradient with uncertainty bounds.

9 | Conclusions

In this thesis, we were motivated by the problem of identifying a dynamical system model with multi-dimensional uncontrollable inputs. In studying this problem, we achieved several important outcomes. First, we argued that the multi-dimensional input case is nontrivial, and we derived necessary conditions for when a dynamical system model with a multi-dimensional could be identified. Second, we discussed the possibility of using incremental estimators to determine the unknown model. Here, we introduced an ansatz, motivated predominantly by statistical ideas, to generate such incremental estimators. We then explored the topic of restarts and two particular incremental estimators in greater detail, and we then experimented on using these estimators for the task of estimating dynamical system models. Third, we introduced a novel framework for generating optimizers to solve stochastic optimization problems, and we demonstrated the efficacy of the optimizers generated by this framework on several nontrivial problems. Below, we discuss the future direction and outlook for this work.

Strong Identifiability. One of the main shortcomings of the above work is the lack of sufficient conditions for strong identifiability, which would be particularly useful for the case of stating when a dynamical system with a multidimensional stochastic input is guaranteed to be strongly identifiability with probability one. The main obstacle in establishing sufficient conditions is in understanding the zero patterns of the Laplace transformations of the inputs. Once this is resolved, the sufficient conditions can be easily established.

Moreover, once necessary and sufficient conditions are established for the linear system, the next natural step is to consider nonlinear systems. There are many approaches to extend from the linear to the nonlinear system such as using local linear approximations to using algebraic mappings, as have been done in related works. However, these techniques must be extended in order to apply them for the cases of multidimensional inputs.

Finally, we can also use these results to study specific application areas such as integrated sensor networks (i.e., design of experiments), and appropriate regularization techniques for the identification of dynamical systems.

Incremental Estimators. While we defined the notions of computability and incremental estimators, which are not particularly novel, and we developed a unique statistically motivated procedure for generating such incremental estimators, we did not provide a unifying theory either in terms of asymptotic properties nor in terms of computational complexity for these estimators. The development of this theory will require melding several areas from mathematical statistics to complexity theory.

In addition to the ansatz, we also discussed two modes of failures for incremental estimators — stagnation and divergence. We also suggested the technique of restarts to mitigate the impact of stagnation on incremental estimators, and provided a simple result stating that restarts will not impact the convergence of the incremental estimator. However, we have not taken full advantage of this restart structure. In particular, restarts induce a (time-invariant) Markov structure on top of the incremental estimators. This Markov structure can be used during estimation to find optimal hyper parameters that avoid stagnation and divergence, and can be used to establish convergence diagnostics.

With regards to the work on stochastic gradient descent, a rigorous extension to nonconvex stochastic programs is clearly a logical next step for the work introduced. However, this is particularly challenging because we can have nonconvex problems for which the Hessians of the random mappings about a minimizer are not positive semi-definite, even though the expected Hessian is positive semi-definite. Therefore, a natural restriction is to consider nonconvex stochastic programs where the Hessians of the random mappings about a minimizer are positive semi-definite and establishing the results for the nonconvex case rigorously.

With regards to the Kalman-based Stochastic Gradient Descent work, the next logical steps for this work include rigorous consistency and efficiency proofs for generalized linear

models and, more generally, quasi-likelihood models. While the consistency can be achieved in a somewhat straightforward manner, demonstrating the efficiency of the estimator is much more challenging for such problems. Another area to extend to is when the parameter is restricted to a convex set. Here, we can take advantage of the fact that generalized linear models can be used as self-concordant barrier functions for such problems, and integrate kSGD with the principles of interior-point methods to generate incremental estimators that work for high-dimensional problems.

Stochastic Filtering & Optimization. We believe this work has the most promise. The flexibility of the framework allows for us to integrate constraints, integer programming problems, model-based techniques, and low-memory approaches with very little effort. However, we have only understood some of the statistical aspects of this approach. We must still explore the optimization aspects and see if we can improve this framework even for gradient-based methods. However, if we can resolve these issues, then we believe that this could be a very substantial contribution.

Overall, we have introduced many tools and ideas in this work to address the problem of identifying linear dynamical systems. Although we have not completely solved this problem (i.e., the sufficient condition for identifiability is still missing), we have made progress on this topic. Moreover, by pursuing this problem, we have introduced a number of mathematical, statistical and algorithmic tools to solve a much broader range of problems. However, these tools and ideas are far from complete, and are far from being considered substantial contributions. The hope is that this thesis is a foundation from which many research programs can grow into substantial contributions.

The woods are lovely, dark and deep,
But I have promises to keep,
And miles to go before I sleep,
And miles to go before I sleep.[1]

[1] From "Stopping by the Woods on a Snowy Evening" by Robert Frost.

Bibliography

Amari, S.-I. and J.-F. Cardoso (1997). Blind source separation-semiparametric statistical approach. *IEEE Transactions on Signal Processing 45*(11), 2692–2700.

Amari, S.-I., T.-P. Chen, and A. Cichocki (1997). Stability analysis of learning algorithms for blind source separation. *Neural Networks 10*(8), 1345–1351.

Amari, S.-I., H. Park, and K. Fukumizu (2000). Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Computation 12*(6), 1399–1409.

Bartlett, M. (1953a). Approximate confidence intervals. *Biometrika*, 12–19.

Bartlett, M. S. (1953b). Approximate confidence intervals. ii. more than one unknown parameter. *Biometrika*, 306–317.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences 110*(36), 14563–14568.

Belgium Network of Open Source Analytical Consultants (2012). Biglm on your big data in open source r, it just works – similar as in sas.

Bellman, R. and K. J. Åström (1970). On structural identifiability. *Mathematical Biosciences 7*(3), 329–339.

Bertsekas, D. P. (1996). Incremental least squares methods and the extended kalman filter. *SIAM Journal on Optimization 6*(3), 807–822.

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena scientific Belmont.

Bertsekas, D. P. (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning 2010*, 1–38.

Birgin, E. G., J. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint (2017). Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming 163*(1-2), 359–368.

Bittanti, S., P. Bolzern, and M. Campi (1990). Convergence and exponential convergence of identification algorithms with directional forgetting factor. *Automatica 26*(5), 929–932.

Blanco-Ortega, A., G. Silva-Navarro, J. Colín-Ocampo, M. Oliver-Salazar, and G. Vela-Valdés (2012). Automatic balancing of rotor-bearing systems. In *Advances on Analysis and Control of Vibrations-Theory and Applications*. InTech.

Bottou, L., F. E. Curtis, and J. Nocedal (2016). Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*.

Brochu, E., V. M. Cora, and N. De Freitas (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Byrd, R., S. Hansen, J. Nocedal, and Y. Singer (2016). A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization 26*(2), 1008–1031.

Cao, L. and H. M Schwartz (2003). Exponential convergence of the kalman filter based parameter estimation algorithm. *International Journal of Adaptive Control and Signal Processing 17*(10), 763–783.

Cao, Y., S. Li, L. Petzold, and R. Serban (2003). Adjoint sensitivity analysis for differential-algebraic equations: The adjoint dae system and its numerical solution. *SIAM Journal on Scientific Computing 24*(3), 1076–1089.

Centers for Medicare & Medicaid (2010). Basic stand alone carrier line items public use files.

Chung, K. L. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics*, 463–483.

Cooper, P. S. (1998). High speed data acquisition. In *AIP Conference Proceedings CONF-9707140*, Volume 422, pp. 3–13. AIP.

Cotes, R. (1722). *Harmonia mensurarum*.

Courant, R. and D. Hilbert (1962). Methods of mathematical physics (interscience, new york, 1953), vol. 1. *Google Scholar*, 351–388.

Davidon, W. C. (1991). Variable metric method for minimization. *SIAM Journal on Optimization 1*(1), 1–17.

Denis-Vidal, L. and G. Joly-Blanchard (2004). Equivalence and identifiability analysis of uncontrolled nonlinear dynamical systems. *Automatica 40*(2), 287–292.

Denis-Vidal, L., G. Joly-Blanchard, and C. Noiret (2001). Some effective approaches to check the identifiability of uncontrolled nonlinear systems. *Mathematics and computers in simulation 57*(1), 35–44.

Dreyfus, S. (1962). The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications 5*(1), 30–45.

Dreyfus, S. E. (1965). Dynamic programming and the calculus of variations. Technical report, RAND CORP SANTA MONICA CA.

Duchi, J., E. Hazan, and Y. Singer (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research 12*(Jul), 2121–2159.

Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

ElGawady, M., P. Lestuzzi, and M. Badoux (2004). A review of conventional seismic retrofitting techniques for urm. In *13th international brick and block masonry conference*, pp. 1–10.

ENTSOE (2018, March). continuing frequency deviation in the continental european power system originating in serbia-kosovo.

Evans, N. D., M. J. Chapman, M. J. Chappell, and K. R. Godfrey (2002). Identifiability of uncontrolled nonlinear rational systems. *Automatica 38*(10), 1799–1805.

Fisher, F. M. (1966). *The identification problem in econometrics*. McGraw-Hill.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 22, pp. 700–725. Cambridge University Press.

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal 1*(6), 445–466.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal 13*(3), 317–322.

Fonollosa, J., S. Sheik, R. Huerta, and S. Marco (2015). Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors and Actuators B: Chemical 215*, 618–629.

Gargash, B. and D. Mital (1980). A necessary and sufficient condition of global structural identifiability of compartmental models. *Computers in biology and medicine 10*(4), 237–242.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss*. sumtibus Frid. Perthes et IH Besser.

Ge, R., F. Huang, C. Jin, and Y. Yuan (2015). Escaping from saddle pointsâĂŤonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842.

Glover, K. and J. Willems (1974). Parametrizations of linear dynamical systems: Canonical forms and identifiability. *IEEE Transactions on Automatic Control 19*(6), 640–646.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation 24*(109), 23–26.

Goodman, D. S. and R. P. Noble (1968). Turnover of plasma cholesterol in man. *Journal of Clinical Investigation 47*(2), 231.

Gowing, R. (2002). *Roger Cotes-natural philosopher*, Volume 50. Cambridge University Press.

Grewal, M., G. Bekey, and H. Payne (1974). Parameter identification of dynamical systems. In *Proc. of the 1974 IEEE Conference on Decision and Control, Phoenix, AZ*.

Griewank, A. (1992). Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation. *Optimization Methods and software 1*(1), 35–54.

Haar, A. (1910). On the theory of orthogonal function systems. *Math. Ann 69*(3), 331–371.

Hájek, M. (1972). A contribution to the parameter estimation of a certain class of dynamical systems. *Kybernetika 8*(2), 165–173.

Hardt, M., B. Recht, and Y. Singer (2015). Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*.

Ho, Y.-C. and R. C. Lee (1964). Identification of linear dynamic systems. In *Adaptive Processes, 1964. Third Symposium on*, Volume 3, pp. 86–101. IEEE.

Jin, C., R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan (2017). How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*.

Johnstone, R. M., C. R. Johnson, R. R. Bitmead, and B. D. Anderson (1982). Exponential convergence of recursive least squares with exponential forgetting factor. In *Decision and Control, 1982 21st IEEE Conference on*, pp. 994–997. IEEE.

Kalman, R. (1959). On the general theory of control systems. *IRE Transactions on Automatic Control 4*(3), 110–110.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering 82*(1), 35–45.

Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kleinberg, R., Y. Li, and Y. Yuan (2018). An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, Volume 96, pp. 202–207. Citeseer.

Kouri, D. P. and T. M. Surowiec (2016). Risk-averse pde-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization 26*(1), 365–396.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics 2*(2), 164–168.

Lewis, A. D. (2009). Semicontinuity of rank and nullity and some consequences. *AuthorâĂŹs notes on semicontinuity of rank and nullity*.

Ljung, L. (1979). Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems. *IEEE Transactions on Automatic Control 24*(1), 36–50.

Ljung, L. (1999). *System Identification: Theory for the User* (2nd ed.). Prentice Hall.

Lory, H., D. Lai, and W. Huggins (1959). On the use of growing harmonic exponentials to identify static nonlinear operators. *IRE Transactions on Automatic Control 4*(2), 91–99.

Loshchilov, I. and F. Hutter (2016). Sgdr: stochastic gradient descent with restarts. *Learning 10*, 3.

Lotka, A. J. (1910). Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry 14*(3), 271–274.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics 11*(2), 431–441.

Miller, A. J. (1992). Algorithm as 274: Least squares routines to supplement those of gentleman. *Applied Statistics*, 458–478.

Mou, W., L. Wang, X. Zhai, and K. Zheng (2017). Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*.

Murata, N. (1998). A statistical study of on-line learning. *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK*, 63–92.

Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE 50*(10), 2061–2070.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization 19*(4), 1574–1609.

Newell, S. A., R. Carroll, P. Ruiz, and W. Gorman (2015). *Cost-Benefit Analysis of ERCOTâĂŹs Future Ancillary Services (FAS) Proposal*. The Brattle Group.

Nocedal, J. and S. J. Wright (2006). *Numerical optimization 2nd ed.* Springer.

Parkum, J., N. K. Poulsen, and J. Holst (1992). Recursive forgetting algorithms. *International Journal of Control 55*(1), 109–128.

Patel, V. (2016). Kalman-based stochastic gradient method with stop condition and insensitivity to conditioning. *SIAM Journal on Optimization 26*(4), 2620–2648.

Patel, V. (2017a). The impact of local geometry and batch size on the convergence and divergence of stochastic gradient descent. *arXiv preprint arXiv:1709.04718*.

Patel, V. (2017b). On sgd's failure in practice: Characterizing and overcoming stalling. *arXiv preprint arXiv:1702.00317*.

Reid, J. (1977). Structural identifiability in linear time-invariant systems. *IEEE Transactions on Automatic Control 22*(2), 242–246.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Robbins, H. and D. Siegmund (1985). A convergence theorem for nonnegative almost supermartingales and some applications. js rustagi, optimizing methods in statistics.

Roe, B. P., H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 543*(2-3), 577–584.

Roeckner, E., G. Bäuml, L. Bonaventura, R. Brokopf, M. Esch, M. Giorgetta, S. Hagemann, I. Kirchner, L. Kornblueh, E. Manzini, et al. (2003). The atmospheric general circulation model echam 5. part i: Model description.

Schaul, T., S. Zhang, and Y. LeCun (2013). No more pesky learning rates. *ICML (3) 28*, 343–351.

Shao, J. (2003). *Mathematical statistics*. Springer New York.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

Shumway, R. H. and D. S. Stoffer (2006). *Time series analysis and its applications: with R examples*. Springer Science & Business Media.

Simon, D. (2006). *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons.

Stanhope, S., J. Rubin, and D. Swigon (2014). Identifiability of linear and linear-in-parameters dynamical systems from a single trajectory. *SIAM Journal on Applied Dynamical Systems 13*(4), 1792–1815.

Styblinski, M. and T.-S. Tang (1990). Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing. *Neural Networks 3*(4), 467–483.

Tamrakar, U., D. Shrestha, M. Maharjan, B. P. Bhattarai, T. M. Hansen, and R. Tonkoski (2017). Virtual inertia: Current trends and future directions. *Applied Sciences 7*(7), 654.

Thowsen, A. (1978). Identifiability of dynamic systems. *International Journal of Systems Science 9*(7), 813–825.

Tieleman, T. and G. Hinton (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning 4*(2).

Titchmarsh, E. C. (1926). The zeros of certain integral functions. *Proceedings of the London Mathematical Society 2*(1), 283–302.

Toulis, P., E. Airoldi, and J. Rennie (2014). Statistical analysis of stochastic gradient methods for generalized linear models. In *ICML*, pp. 667–675.

Tuerlinckx, F., F. Rijmen, G. Verbeke, and P. Boeck (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology 59*(2), 225–255.

UCI Machine Learning Repository (2015). Gas sensor array under dynamic gas mixtures data set.

UCI Machine Learning Respository (1996). Adult data set.

Van der Pol, B. (1934). The nonlinear theory of electric oscillations. *Proceedings of the Institute of Radio Engineers 22*(9), 1051–1086.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge university press.

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science 3*(1), 3–51.

Von Meier, A., D. Culler, A. McEachern, and R. Arghandeh (2014). Micro-synchrophasors for distribution systems. In *Innovative Smart Grid Technologies Conference (ISGT), 2014 IEEE PES*, pp. 1–5. IEEE.

Wang, B. and K. Sun (2015). Power system differential-algebraic equations. *arXiv preprint arXiv:1512.05185*.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussâĂŤnewton method. *Biometrika 61*(3), 439–447.

Xie, W.-C. (2010). *Differential Equations for Engineers*. Cambridge University Press.

Yoon, S. Y., Z. Lin, and P. E. Allaire (2012). *Control of surge in centrifugal compressors by active magnetic bearings: Theory and implementation*. Springer Science & Business Media.

Zadeh, L. (1956). On the identification problem. *IRE Transactions on Circuit Theory 3*(4), 277–281.

Zavala, V. M. and M. Anitescu (2010). Real-time nonlinear optimization as a generalized equation. *SIAM Journal on Control and Optimization 48*(8), 5444–5467.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, Y., P. Liang, and M. Charikar (2017). A hitting time analysis of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1702.05575*.