ADAPTIVE SEQUENTIAL OPTIMIZATION WITH APPLICATIONS TO MACHINE LEARNING

BY

CRAIG WILSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Venugopal V. Veeravalli, Chair
Associate Professor Angelia Nedić
Professor Rayadurgam Srikant
Professor Tamer Başar

Abstract

The focus of this thesis is on solving a sequence of optimization problems that change over time in a structured manner. This type of problem naturally arises in contexts as diverse as channel estimation, target tracking, sequential machine learning, and repeated games. Due to the time-varying nature of these problems, it is necessary to determine new solutions as the problems change in order to ensure good solution quality. However, since the problems change over time in a structured manner, it is beneficial to exploit solutions to the previous optimization problems in order to efficiently solve the current optimization problem.

The first problem considered is sequentially solving minimization problems that change slowly, in the sense that the gap between successive minimizers is bounded in norm. The minimization problems are solved by sequentially applying a selected optimization algorithm, such as stochastic gradient descent (SGD), based on drawing a number of samples in order to carry out a desired number of iterations. Two tracking criteria are introduced to evaluate approximate minimizer quality: one based on being accurate with respect to the mean trajectory, and the other based on being accurate in high probability (IHP). Knowledge of the bound on how the minimizers change, combined with properties of the chosen optimization algorithm, is used to select the number of samples needed to meet the desired tracking criterion.

Next, it is not assumed that the bound on how the minimizers change is known. A technique to estimate the change in minimizers is provided along with analysis to show that eventually the estimate upper bounds the change in minimizers. This estimate of the change in minimizers is combined with the previous analysis to provide sample size selection rules to ensure that the mean or IHP tracking criterion is met. Simulations are used to confirm that the estimation approach provides the desired tracking accuracy in practice.

An application of this framework to machine learning problems is considered next. A cost-based approach is introduced to select the number of samples with

a cost function for taking a number of samples and a cost budget over a fixed horizon. An extension of this framework is developed to apply cross validation for model selection. Finally, experiments with synthetic and real data are used to confirm that this approach performs well for machine learning problems.

The next model considered is solving a sequence of minimization problems with the possibility that there can be abrupt jumps in the minimizers mixed in with the normal slow changes. Alternative approaches are introduced to estimate the changes in the minimizers and select the number of samples. A simulation experiment demonstrates the effectiveness of this approach.

Finally, a variant of this framework is applied to learning in games. A sequence of repeated games is considered in which the underlying stage games themselves vary slowly over time in the sense that the pure strategy Nash equilibria change slowly. Approximate pure-strategy Nash equilibria are learned for this sequence of zero sum games. A technique is introduced to estimate the change in the Nash equilibiria as for the sequence of minimization problems. Applications to a synthetic game and a game based on a surveillance network problem are introduced to demonstrate the game framework.

To my mother, father, and brother.

Acknowledgments

I am grateful to my advisor Prof. Venu Veeravalli. He has spent a considerable amount of time helping me find an interesting topic for my thesis and developing the work within. He has taught me a number of valuable lessons about approaching research that will be useful for the rest of my career. Prof. Angelia Nedić has also played an enormous role in developing this thesis. She has in particular played a crucial role in developing the topic of this thesis and determining some interesting extensions of this work.

I also give thanks for my family including my mother, father, and brother who have supported me throughout this whole long process. They have supported me throughout and encouraged me every step of the way.

I am also grateful to the rest of my committee, Prof. Başar and Prof. Srikant, for taking the time to read my thesis and provide valuable feedback. I have had the pleasure of working with a number of great fellow graduate students during my time at UIUC including Jonathan, Yun, Taposh, and Aly. These individuals have provided substantial amounts of useful feedback for refining my work and presentation. I also want to thank Chris Dunaway who has helped carry out some simulation work for my thesis. The various CSL office support specialists that I have worked with including Terry Hovde, Peggy Wells, Barbara Horner, and Brenda Roy have always been eager to help and have made dealing with administrative issues painless.

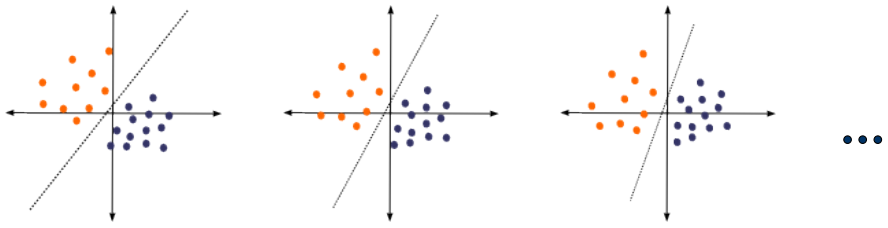
Table of Contents

Chapter 1

Introduction

1.1 A Motivating Example

Consider a sequence of classification problems that change slowly over time as in Figure 1.1. Given the location of a point, we want to decide whether the point in



(a) p_1 (b) p_2 (c) p_3

Figure 1.1: Slowly changing classification problems

question is orange or blue. In this simple example, we assume that the points are linearly separable, i.e., it is possible to draw a line such that the orange points fall on one side and the blue points fall on the other side. Thus, our goal is to find a line that can separate the orange and blue points. The problem of finding a linear classifier can be reduced to solving an optimization problem.

The key feature of this problem is that over time the relationship between whether a point is orange or blue changes as in Figure 1.1. Therefore, the classifier developed for the first time instant will perform poorly as the underlying distribution changes. To maintain good classification performance, it is essential that we update our classifier as the underlying problem changes. However, since the problem has only changed slightly over time, we want to take advantage of the previously developed high-performance classifier. Since, as mentioned above, finding a classifier can be carried out by solving an optimization problem, we want to solve a sequence of optimization problems that change in a structured manner. More precisely, the optimization problems change slowly in the sense that the

linear boundary that separates the two classes changes slowly.

The problems we consider in this thesis all share the same features of this problem. We want to solve a sequence of optimization problems, to find a classifier in the above example, that change over time in a structured manner. By exploiting the structure of the change in the optimization problems, the previous solution, and properties of the method used to solve the optimization problems, we produce a good approximate solution to each problem.

1.2 Thesis Outline

The work in this thesis can be divided into two different areas. First, we consider solving stochastic optimization problems that change over time. Second, we consider solving a repeated zero sum game in which the stage games themselves change slowly.

1.2.1 Adaptive Sequential Optimization

Problems involving optimizing a sequence of functions that slowly vary over time naturally arise in many different contexts including channel estimation, parameter tracking, and sequential learning. To describe and analyze such problems, we consider solving a sequence of optimization problems

$$\min_{x \in \mathscr{X}} f_n(\boldsymbol{x}) \tag{1.1}$$

with $\boldsymbol{x} \in \mathscr{X} \subset \mathbb{R}^d$. We call our process for efficiently solving a sequence of optimization problems *adaptive sequential optimization*.

In this thesis, we consider several different models for the change in the minimizers. Section 2.2.3 contains a discussion of why we pose the idea of slow changes in terms of the minimizers. It turns out that several different reasonable models for changes in f_n can be reduced to a bound on the change in the minimizers. First, to capture the idea that the sequence of functions in (1.1) is changing slowly we assume a bound on the minimizers of the form

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_{L_2} \leq \rho \tag{1.2}$$

or

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_2 \leq \rho \tag{1.3}$$

where

$$\boldsymbol{x}_n^* = \arg\min_{\boldsymbol{x}\in\mathscr{X}} f_n(\boldsymbol{x})$$

and $\|X\|_{L_q} = (\mathbb{E}[X^q])^{1/q}$. We assume that $\boldsymbol{x}_n^*$ is unique for each n. Rather than using a Markov chain model or other Bayesian model for the changes in $\{\boldsymbol{x}_n^*\}_{n=1}^{\infty}$, we only use the bound (1.2) in our analysis. The formulation in (1.2) involves an expectation to allow for the possibility that the functions $\{f_n(\boldsymbol{x})\}$ themselves evolve stochastically. If the functions $\{f_n(\boldsymbol{x})\}$ evolve stochastically, then the minimizers $\boldsymbol{x}_n^*$ are random variables. We will consider the case when ρ is known or unknown separately. When ρ is unknown to us, we will develop an estimate of ρ.

Next, we consider abrupt changes in the sense that

$$\rho_n \triangleq \|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 \in \{\rho^{(1)}, \rho^{(2)}\} \quad \forall n \geq 2 \tag{1.4}$$

where $\rho^{(2)} \gg \rho^{(1)}$. The change $\rho^{(1)}$ corresponds to small, slow changes and the change $\rho^{(2)}$ corresponds to large, abrupt change.

Given a sequence of functions $\{f_n(\boldsymbol{x})\}_{n=1}^{\infty}$, we want to efficiently, sequentially minimize each of the functions to within a desired accuracy. We look at solving this problem by applying an optimization algorithm such as SGD that uses K_n stochastic gradient steps. To be precise, suppose that we have a function $\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z})$ such that, given $\boldsymbol{z}_n \sim p_n$ from an auxiliary distribution p_n, it holds that

$$\mathbb{E}_{\boldsymbol{z}_n\sim p_n}[\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n) \mid \boldsymbol{x}] = \nabla f_n(\boldsymbol{x}) \tag{1.5}$$

For example, if the functions $\{f_n(\boldsymbol{x})\}_{n=1}^{\infty}$ are of the form

$$f_n(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_n\sim p_n}[\ell(\boldsymbol{x}, \boldsymbol{z}_n)] \tag{1.6}$$

where $\ell(\boldsymbol{x}, \boldsymbol{z})$ can correspond to a loss in machine learning, then we can set

$$\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n) = \nabla_{\boldsymbol{x}} \ell(\boldsymbol{x}, \boldsymbol{z}_n)$$

under suitable technical conditions. We want to understand the trade-off between the solution accuracy and the complexity, represented by the number of stochastic gradients K_n. In effect, we want to understand how many samples $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}$ are

necessary to achieve a desired level of accuracy.

We introduce two different types of criteria to characterize approximate minimizers of (1.1), denoted $\boldsymbol{x}_n$ for each n. First, we define a *mean criterion*

$$\mathbb{E}\left[f_n(\boldsymbol{x}_n)\right] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon \tag{1.7}$$

and second, we define an *in high probability (IHP) criterion*

$$\mathbb{P}\left\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\right\} \leq r \tag{1.8}$$

with the expectation and probability taken over the samples $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}$. In the context of machine learning, the mean criterion is referred to as an *excess risk* criterion.

The goal of the analysis for adaptive sequential optimization is to connect the number of samples K_n to either the mean criterion in (1.7) or the IHP criterion in (1.8) being met. This connection makes it possible to select the number of samples to achieve a desired mean criterion or IHP criterion.

1.2.2 Time-Varying Stage Game in a Repeated Game

Given a particular solution concept for a game, such as a Nash equilibrium, a natural issue that arises is how the players learn the solution of a game. This problem is of particular consequence when the players have incomplete knowledge of the game and can only learn through repeatedly playing the game. In a repeated game, the players repeatedly play a fixed stage game observing outcomes such as the utility achieved by the player and the strategies of the other players. For repeated games, there are a variety of techniques that can discover solutions or approximate solutions generally for games with finite strategy spaces [1].

We consider playing a sequence of finite horizon repeated games with a slowly changing stage game G_n. The sequence of games played is as follows:

$$G_1, \ldots, G_1, \ldots, G_n, \ldots, G_n, \ldots$$

Note that each group of games with the same stage game, i.e., $G_n, \ldots, G_n$, is itself a finite horizon repeated game. The slowly changing nature is captured by the idea that stage game G_n is played a number of times before changing where the

solutions of the stage games change slowly. The notion of slow change will be made more precise later. To learn the solution to each stage game, we exploit the solution to the previous stage game. Using knowledge of the previous stage games is crucial, since collecting the feedback needed to carry out the learning dynamics for each stage game on its own may be expensive. When the players play the n^{th} stage game K_n times to determine a solution, the sequence of games is as follows:

$$\underbrace{G_1,\ldots,G_1}_{\substack{\text{learning phase}\\ K_1 \text{ times}}},\ \underbrace{G_1,\ldots,G_1}_{\text{additional plays}},\ \ldots,\ \underbrace{G_n,\ldots,G_n}_{\substack{\text{learning phase}\\ K_n \text{ times}}},\ \underbrace{G_n,\ldots,G_n}_{\text{additional plays}},\ \ldots$$

There is a learning phase consisting of K_n plays of G_n in which the players find a solution to the current stage game and an exploitation phase in which the players use their knowledge until the stage game changes. Since the players only play each game a finite number of times, they can generally only find approximate solutions. Therefore, we seek a connection between the number of times K_n the stage game must be played to learn an approximate solution and the desired quality of the approximate solution.

1.3 Organization of Thesis

This thesis is organized as follows. Chapters 2, 3, 4, and 5 consider solving a sequence of minimization problems with various models on how the minimizers of each problem change. Chapter 2 considers the problems of solving this sequence of minimization problems when the change in the minimizers is known. Chapter 3 considers the same problem except without knowledge of the change in the minimizers and instead introduces an approach to estimate this quantity. Chapter 4 applies the techniques in Chapters 2 and 3 to machine learning problems. In addition, a cost based approach to selecting the number of samples is introduced. Chapter 5 considers the case where the change in the minimizers need not be small but can have abrupt jumps mixed in. New methods to estimate the change in the minimizer and select the number of samples are developed based on the methods of Chapter 3. In Chapter 6, we consider playing a sequence of repeated games in which the stage games vary slowly. We introduce a technique for each player to learn the Nash equilibrium of each stage game while exploiting knowledge of the previous stage games. Finally, in Chapter 7, we consider some broad future

research directions.

Chapter 2

Change in Minimizers Known

We consider solving the sequence of optimization problems described in Section 1.2.1. In this chapter, we work under the assumption that

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_{L_2} \leq \rho$$

and the change in the minimizers, ρ, is known to us. We develop rules to select the number of stochastic gradients K_n to achieve a desired mean criterion or IHP criterion in (1.7) and (1.8). In Chapter 3, we will drop the assumption that ρ is known and extend the results in this chapter. The content of this chapter corresponds to the work in [2] and [3].

2.1 Related Work

There has been some work on similar problems, but general optimization theory tools to deal with time-varying optimization problems under (1.2) have yet to be developed. One relevant approach is online optimization in which a sequence of functions arrive. In the online optimization approach, generally no knowledge is available about the incoming functions other than that all the functions come from a specified class of functions, i.e., linear or convex functions with uniformly bounded gradients. Online optimization models do not include the notion of a desired tracking accuracy at each time instant such as (1.7) and (1.8). Instead, only bounds on the worst case performance of the best estimators are investigated through regret formulations [4–13].

For the problem of online optimization, the idea of controlling the variation of the sequence of functions has been studied in [14] and [15]. In [15], regret is minimized subject to a bound, say G_b, on the total variation of the gradients over

a time interval T of interest, i.e.,

$$\sum_{n=2}^{T} \max_{\boldsymbol{x} \in \mathscr{X}} \|\nabla f_n(\boldsymbol{x}) - \nabla f_{n-1}(\boldsymbol{x})\|_2^2 \leq G_b \tag{2.1}$$

If all the functions $\{f_n(x)\}$ are strongly convex with the same parameter m, then by the optimality conditions (see Theorem 2F.10 in [16]) relation (2.1) implies that

$$\sum_{n=2}^{T} \|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2^2 \leq \tilde{G}_b$$

with $\tilde{G}_b$ a function of G_b. Therefore, the work in [15] can be seen as studying the regret while controlling the total variation in the optimal solutions over T time instants. In contrast, we control the variation of the optimal solutions at each time instant with (1.2) and then seek to maintain a tracking criterion such as (1.7) and (1.8) at each time instant.

Additionally, there is other work that has some of the ingredients of our proposed problem formulation. In [17], a sequence of quadratic functions is considered and treated within the domain of estimation theory; however, the authors only examine the least mean squares (LMS) algorithm (corresponding to $K_n = 1$ for all n). The work in [18, 19] considers a sequence $\{f_n\}$ of convex objective functions converging to some limit function f, where all the functions f_n have the same set of minima. However, aside from considering time-varying objective functions, these works have nothing else in common with the work described here. There has also been work in [20] considering the limit as the rate of change of the functions goes to zero and for the defined just above LMS algorithm in [21]. The results in [20] and [21] both require a Bayesian model for the changes in the function sequence, which we do not require.

If we have a quadratic loss centered at $\boldsymbol{x}_n^*$ and a linear state space evolution for the optimal solution $\boldsymbol{x}_n^*$, then we could apply the Kalman filter [22]. If the function we seek to optimize is non-linear, another approach we can consider under a Bayesian framework is particle filtering [23]. For particle filtering, it is harder to provide exact guarantees on performance similar to those given in (1.7) and (1.8).

To conclude, there are no existing approaches within optimization theory or estimation theory that allow us to solve a sequence of time-varying problems, subject to adhering to a pre-specified tracking error criterion such as (1.7) or (1.8)

under only assumption (1.2). In this work, we fill in this gap and provide methods to solve such problems.

2.2 Problem Formulation

2.2.1 A Simple Motivating Example

We motivate our problem formulation by analyzing a simple scalar quadratic tracking problem based on the analysis of the least means squares (LMS) algorithm in [21]. This example demonstrates the importance of knowing the properties of the functions $f_n(\boldsymbol{x})$ in order to choose the number K_n to achieve a good mean-tracking performance.

At time n, we observe K_n realizations of a signal consisting of the pair $y_n(k)$ and $w_n(k)$, with $y_n(k)$ given by

$$y_n(k) = \eta_n w_n(k) + e_n(k) \qquad k = 1, \ldots, K_n$$

Our goal is to estimate η_n. In this example, $y_n(k)$, η_n, $w_n(k)$, and $e_n(k)$ are all scalars. We take $w_n(k) \sim \mathcal{N}(0,1)$ and $e_n(k) \sim \mathcal{N}(0, \sigma_e^2)$. We assume that the collection of all η_n, $w_n(k)$, and $e_n(k)$ over n and k are independent. The random variables η_n are generated by a random walk model

$$\eta_{n+1} - \eta_n \sim \mathcal{N}(0, \sigma_\delta^2) \tag{2.2}$$

with η_1 a fixed constant. In order to estimate η_n, we minimize

$$f_n(x) = \mathbb{E}_{(w_n, e_n) \sim p_n} \left[\frac{1}{2} (x w_n - y_n)^2 \right] \tag{2.3}$$

as a function of x. Thus, we have $x_n^* = \eta_n$, and so $\|x_{n+1}^* - x_n^*\|_{L_2} = \sigma_\delta$ where $\|X\|_{L_2} = \sqrt{\mathbb{E}\|X\|^2}$ due to the assumption in (2.2). Therefore, this problem satisfies (1.2) with $\rho = \sigma_\delta$.

For this signal model, we set $z_n(k) = [w_n(k)\ e_n(k)]^\top$ and define the stochastic gradients as

$$\boldsymbol{g}_n(\boldsymbol{x}, z_n(k)) = -(y_n(k) - x^\top w_n(k)) w_n(k)$$

As required in (1.5), it holds that

$$\mathbb{E}_{z_n(k)\sim p_n}\left[\boldsymbol{g}_n\left(\boldsymbol{x}, z_n(k)\right) \mid \boldsymbol{x}\right] = \nabla f_n(\boldsymbol{x})$$

We generate an approximate minimizer x_n of (2.3) to approximate η_n using K_n steps of stochastic gradient descent

$$\begin{aligned} \boldsymbol{x}_n(k) &= \Pi_{\mathscr{X}}[\boldsymbol{x}_n(k-1) - \mu_n(k)\boldsymbol{g}_n\left(\boldsymbol{x}, \boldsymbol{z}_n(k)\right)] \\ \boldsymbol{x}_n(0) &\triangleq \boldsymbol{x}_{n-1} \end{aligned} \tag{2.4}$$

for $k = 1, \ldots, K_n$. For a generic application of SGD, we choose $\boldsymbol{x}_n$ as a function of $\{\boldsymbol{x}_n(0), \ldots, \boldsymbol{x}_n(K_n)\}$ such as selecting the last iterate $\boldsymbol{x}_n(K_n)$ or averaging the iterates [24]. In this section, we simply choose constant step sizes $\mu_n(k) = \mu$ and the last iterate $\boldsymbol{x}_n \triangleq \boldsymbol{x}_n(K_n)$. Finally, $\Pi_{\mathscr{X}}$ is the projection onto $\mathscr{X}$.

We follow the analysis in [21] for non-stationary LMS adapted to this multiple iteration version. By expanding the definition of $f_n(x)$ in (2.3), it holds that

$$f_n(x) = \frac{1}{2}\sigma_e^2 + \frac{1}{2}\mathbb{E}(\boldsymbol{x} - \boldsymbol{x}^*)^2$$

Using this observation, define

$$a_n(k) \triangleq \mathbb{E}\left[\frac{1}{2}\left(x_n(k) - x_n^*\right)^2\right]$$

The quantity $a_n(k)$ satisfies the following recursion from equation (5.8) in [21]:

$$a_n(k+1) = (1 - 2\mu + 3\mu^2)a_n(k) + \frac{1}{2}\mu^2\sigma_e^2 \tag{2.5}$$

This recurrence relation can be solved to yield

$$a_n(K_n) = C_1 a_n(0) + C_2$$

with $C_1 \triangleq (1 - 2\mu + 3\mu^2)^{K_n}$ and $C_2 \triangleq \frac{\mu(1-C_1)}{4-6\mu}\sigma_e^2$. This analysis captures the behavior of the multiple iteration LMS algorithm in (2.4). We then have

$$a_{n+1}(0) \;=\; a_n(K_n) + \frac{1}{2}\sigma_\delta^2$$

and so it follows that

$$\mathbb{E}\left[\frac{1}{2}(x_{n+1}-x_{n+1}^*)^2\right] \leq C_1\left(\mathbb{E}\left[\frac{1}{2}(x_n-x_n^*)^2\right]+\frac{1}{2}\sigma_\delta^2\right)+C_2$$

This first order, inhomogeneous recurrence relation can be solved. The asymptotic mean performance achieved is then given by

$$\begin{aligned}
&\mathscr{E}_\infty(\mu)\\
&\triangleq \lim_{n\to\infty}\left(\mathbb{E}_{\boldsymbol{z}_n\sim p_n}\left[\frac{1}{2}\left(y_n-x_n w_n\right)^2\right]-\frac{1}{2}\sigma_e^2\right)\\
&= \lim_{n\to\infty}\mathbb{E}\left[\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{x}_n^*)^2\right]\\
&= \lim_{n\to\infty} a_n(K_n)\\
&= \frac{\frac{1}{2}C_1\sigma_\delta^2+C_2}{1-C_1}
\end{aligned}$$

Note that if $K_n = 1$, then $\mathscr{E}_\infty(\mu)$ is the same as equation (5.22) in [21] for LMS.

Unknown Parameters

If $\mu > \frac{2}{3}$, then by the preceding analysis in (2.5) $\mathbb{E}[(y_n - x_n w_n)^2] \to \infty$. Therefore, we need to know the function structure and choose the algorithm parameters carefully just to guarantee a finite mean tracking criterion (1.7) at all times.

Known Parameters

A choice of μ in an appropriate range is important to guarantee finite mean tracking criterion (1.7); however, we also want good mean tracking performance guarantees, which requires optimizing $\mathscr{E}_\infty(\mu)$ over μ for each K. To demonstrate this issue, we plot $\mathscr{E}_\infty(\mu)$ versus μ for $\sigma_e^2 = \sigma_\delta^2 = 1$ and various values of $K_n = K$ in Figure 2.1. Minimizing the asymptotic mean tracking quality depends crucially on the choice of step size μ. A bad choice of step size can result in mean tracking error orders of magnitude larger than the optimal choice. Figure 2.2 shows the asymptotic mean tracking quality optimized over μ versus K. We need to be careful in selecting μ for each K to obtain good mean tracking performance.

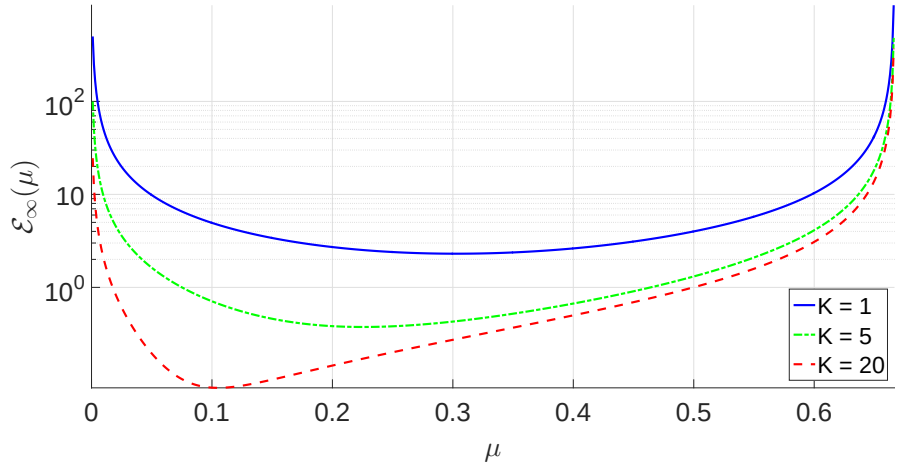


Figure 2.1: $\mathscr{E}_\infty(\mu)$ vs. μ

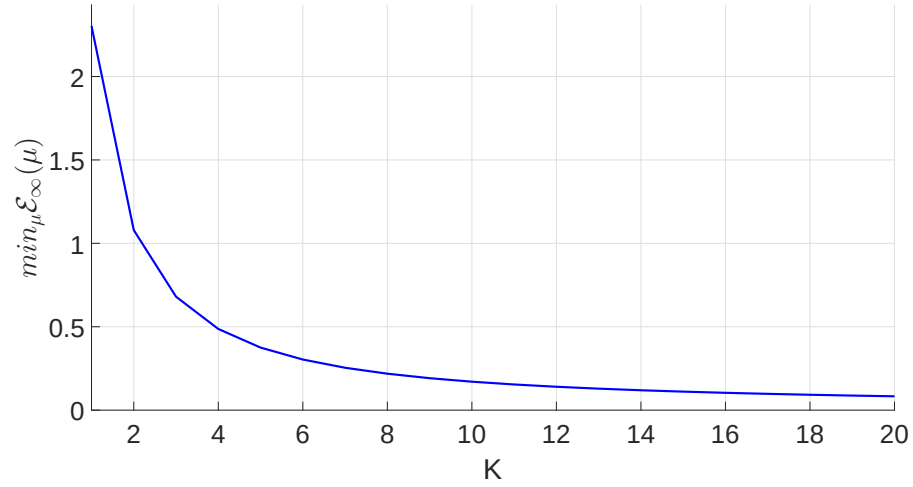


Figure 2.2: $\min_\mu \mathscr{E}_\infty(\mu)$ vs. K

2.2.2 Assumptions

We make several assumptions to proceed. First, let $\mathscr{X}$ be closed and convex with $\text{diam}(\mathscr{X}) < +\infty$. Define the σ-algebra

$$\mathscr{F}_i \triangleq \sigma\left(\bigcup_{j=1}^{i}\bigcup_{k=1}^{K_j} \boldsymbol{z}_j(k)\right) \tag{2.6}$$

where by convention $\mathscr{F}_0$ is the trivial σ-algebra. We suppose that each function $f_n(\boldsymbol{x})$ satisfies the following conditions:

A.1 For each n, $f_n(\boldsymbol{x})$ is twice continuously differentiable with respect to $\boldsymbol{x}$.

A.2 For each n, $f_n(\boldsymbol{x})$ is strongly convex with a parameter $m > 0$, i.e.,

$$\nabla^2 f_n(\boldsymbol{x}) \succeq m\boldsymbol{I} \quad \forall \boldsymbol{x} \in \mathscr{X}, \forall n. \tag{2.7}$$

A.3 For each n, we can draw stochastic gradients $\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n)$ such that (1.5) holds.

A.4 Given an optimization algorithm that generates an approximate minimizer

$$\boldsymbol{x}_n \triangleq \mathscr{A}(\boldsymbol{x}_{n-1}, \{\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}(k))\}_{k=1}^{K_n})$$

using K_n stochastic gradients $\{\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n(k))\}_{k=1}^{K_n}$, there exists a function $b(d_0, K_n)$ such that the following conditions hold:

1. With K_n and d_0 both $\mathscr{F}_{n-1}$-measurable, it holds that

$$\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_n^*\|^2 \leq d_0 \;\Rightarrow\; \mathbb{E}[f_n(\boldsymbol{x}_n) \mid \mathscr{F}_{n-1}] - f_n(\boldsymbol{x}_n^*) \leq b(d_0, K_n)$$

2. With K_n a constant, it holds that $\mathbb{E}[b(d_0, K_n)] = b(\mathbb{E}[d_0], K_n)$.
3. The bound $b(d_0, K_n)$ is non-decreasing in d_0 and non-increasing in K_n.

A.5 There exist constants $A, B \geq 0$ such that

$$\mathbb{E}\left[\|\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n)\|_2^2 \mid \mathscr{F}_{n-1}\right] \leq A + B\|\boldsymbol{x} - \boldsymbol{x}_n^*\|_2^2 \tag{2.8}$$

A.6 Initial approximate minimizers $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfy

$$f_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i^*) \leq \varepsilon_i \quad i = 1, 2$$

with ε_1 and ε_2 known.

For assumption A.4, we generally look at SGD defined in (2.4). Given the iterates, we choose $\boldsymbol{x}_n$ as a convex combination of the iterates $\{\boldsymbol{x}_n(k)\}_{k=0}^{K_n}$ generated by SGD

$$\boldsymbol{x}_n = \sum_{k=0}^{K_n} \lambda_n(k)\boldsymbol{x}_n(k)$$

One simple choice is setting $\boldsymbol{x}_n = \boldsymbol{x}_n(K_n)$, which corresponds to setting $\lambda_n(K_n) = 1$ and

$$\lambda_n(0) = \cdots = \lambda_n(K_n - 1) = 0$$

Appendix A discusses several applicable bounds $b(d_0, K)$ for SGD and choices of convex combinations $\{\lambda_n(k)\}$. In practice, we may not know the parameters such as the strong convexity parameter m from Assumption A.2 and the gradient parameters A and B from Assumption A.5. Appendix D introduces several techniques to estimate these parameters using the stochastic gradients in A.3.

In our assumptions, we condition on the σ-algebra $\mathscr{F}_{n-1}$, since this captures all of the information available at the beginning of time n. Later, we will select K_n as a function of the stochastic gradients $\{\boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k))\}_{k=1}^{K_i}$ for $i = 1, \ldots, n-1$. This implies that K_n is $\mathscr{F}_{n-1}$ measurable. In this case, where K_n is itself a random variable Assumption A.4 is crucial to our analysis.

Finally, for Assumption A.6, we generally must select K_1 and K_2 blindly in the sense that we have no information about ρ defined in (1.2). We can only make a choice such as

$$K_i = \min \left\{ K \geq 1 \;\middle|\; b\left(\operatorname{diam}^2(\mathscr{X}), K\right) \leq \varepsilon \right\} \quad i = 1, 2$$

or fixed initial choices for K_1 and K_2. Regardless of our choice of K_1 and K_2, we can set

$$\varepsilon_i \triangleq b\left(\operatorname{diam}^2(\mathscr{X}), K\right) \quad i = 1, 2$$

In order to have $\varepsilon_i \leq \varepsilon$ for $i = 1, 2$, we may need to draw significantly more samples up front to find points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ due to using $\operatorname{diam}(\mathscr{X})$.

2.2.3 Constructing a Bound on the Change in Minimizers

We look at the justification behind our choice of controlling the change in functions through the minimizers $\boldsymbol{x}_n^*$ by showing that several other reasonable ways to control how the functions change can be reduced to a bound on the change in minimizers. In Section 2.3, we show that bounds on the change in the minimizer can be used to select the number of stochastic gradients K_n.

Change in f

Suppose that we instead bound the change in the optimal function values, in the following manner:

$$f_n(\boldsymbol{x}_{n-1}^*) - f_n(\boldsymbol{x}_n^*) \leq \tilde{\rho}$$

This bounds the loss incurred as a result of using the minimizer of a previous function f_{n-1} and the next function f_n. By the strong convexity Assumption A.2, it holds that

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 \leq \sqrt{\frac{2}{m}\left(f_n(\boldsymbol{x}_{n-1}^*) - f_n(\boldsymbol{x}_n^*)\right)} \leq \sqrt{\frac{2}{m}\tilde{\rho}}$$

Therefore, a bound on the optimal function values can be translated into a bound on the change in the minimizers.

Change in Distribution

As mentioned before, for learning problems in (1.6), we can generally write our functions as an expectation of a loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$, i.e., $f_n(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_n \sim p_n}[\ell(\boldsymbol{x}, \boldsymbol{z}_n)]$. Therefore, the source of change in this problem is the changes in the distributions p_n. We can control change by making an assumption on how p_n changes through an appropriate probability metric or pseudo-metric. Given a class of functions $\mathscr{F}$ mapping from $\mathscr{Z} \to \mathbb{R}$, an integral probability metric [25] between two distributions p and q on $\mathscr{Z}$ is defined as

$$\gamma_{\mathscr{F}}(p, q) \triangleq \sup_{h \in \mathscr{F}} |\mathbb{E}_{\boldsymbol{z} \sim p}[h(\boldsymbol{z})] - \mathbb{E}_{\tilde{\boldsymbol{z}} \sim p}[h(\tilde{\boldsymbol{z}})]|$$

The following lemma shows that under an inclusion condition on the loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$, the integral probability metric bounds can lead to bounds on the change in minimizers.

Lemma 1. *If the class $\{\ell(\boldsymbol{x}, \cdot) \mid \boldsymbol{x} \in \mathscr{X}\} \subset \mathscr{F}$ of loss functions is such that $\gamma_{\mathscr{F}}(p_n, p_{n-1}) \leq \tilde{\rho}$ for all $n \geq 1$, then it holds that*

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 \leq \sqrt{\frac{2}{m}\tilde{\rho}} \qquad \textit{for all } n \geq 1$$

Proof. Applying the strong convexity Assumption (A.2) to $f_n(\boldsymbol{x})$ and $f_{n-1}(\boldsymbol{x})$, for the solutions $\boldsymbol{x}_n^*$ and $\boldsymbol{x}_{n-1}^*$, we obtain

$$\begin{aligned} f_n(\boldsymbol{x}_{n-1}^*) &\geq f_n(\boldsymbol{x}_n^*) + \frac{1}{2} m \|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2^2 \\ f_{n-1}(\boldsymbol{x}_n^*) &\geq f_{n-1}(\boldsymbol{x}_{n-1}^*) + \frac{1}{2} m \|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2^2 \end{aligned}$$

By adding these two inequalities and rearranging, it holds that

$$\begin{aligned} m\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2^2 & \\ &\leq \big(f_n(\boldsymbol{x}_{n-1}^*) - f_n(\boldsymbol{x}_n^*)\big) + \big(f_{n-1}(\boldsymbol{x}_n^*) - f_{n-1}(\boldsymbol{x}_{n-1}^*)\big) \\ &= \big(f_n(\boldsymbol{x}_{n-1}^*) - f_{n-1}(\boldsymbol{x}_{n-1}^*)\big) + (f_{n-1}(\boldsymbol{x}_n^*) - f_n(\boldsymbol{x}_n^*)) \end{aligned}$$

Now, examine the term $f_n(\boldsymbol{x}_{n-1}^*) - f_{n-1}(\boldsymbol{x}_{n-1}^*)$. By relation (1.6), we have

$$\begin{aligned} & f_n(\boldsymbol{x}_{n-1}^*) - f_{n-1}(\boldsymbol{x}_{n-1}^*) \\ &\quad \leq |\mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\ell(\boldsymbol{x}_{n-1}^*, \boldsymbol{z}_n)\right] - \mathbb{E}_{\boldsymbol{z}_{n-1} \sim p_{n-1}}\left[\ell(\boldsymbol{x}_{n-1}^*, \boldsymbol{z}_{n-1})\right]| \\ &\quad \leq \sup_{f \in \mathscr{F}} |\mathbb{E}_{\boldsymbol{z}_n \sim p_n}[f(\boldsymbol{z}_n)] - \mathbb{E}_{\boldsymbol{z}_{n-1} \sim p_{n-1}}[f(\boldsymbol{z}_{n-1})]| \\ &\quad = \gamma_{\mathscr{F}}(p_n, p_{n-1}) \end{aligned}$$

Similarly, we can see that the same estimate holds for the term $f_{n-1}(\boldsymbol{x}_n^*) - f_n(\boldsymbol{x}_n^*)$. Therefore,

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 \leq \sqrt{\frac{2}{m}\gamma_{\mathscr{F}}(p_n, p_{n-1})} \leq \sqrt{\frac{2}{m}\tilde{\rho}}$$

□

Thus, we see that we can translate a bound on the change in distributions through an integral probability metric to a bound on the change in minimizers.

Parameterized Functions

Finally, we examine the case in which the functions $\{f_n(\boldsymbol{x})\}$ come from a parameterized class of functions $f(\boldsymbol{x}, \boldsymbol{\theta})$, i.e.,

$$f_n(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}_n)$$

Furthermore, we assume that the parameters themselves change slowly

$$\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|_2 \leq \delta$$

We look at cases in which we can translate the bound δ on the parameters to changes in the minimizers.

One simple case is when the function $f(\boldsymbol{x},\boldsymbol{\theta})$ satisfies

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x},\tilde{\boldsymbol{\theta}})\|_2 \leq \tilde{M}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2$$

and the stochastic gradients for our function $f(\boldsymbol{x},\theta)$ are the exact gradients plus noise

$$\boldsymbol{g}_n(\boldsymbol{x},\boldsymbol{z}_n,\boldsymbol{\theta}_n) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{\theta}_n) + \eta_n$$

where η_n is a mean zero random variable with $\mathbb{E}\|\eta_n\|_2^2 \leq \sigma^2$. The following bound follows from optimality principle (see Theorem 2F.10 in [16]) and the uniform strong convexity:

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_{L_2} \leq \frac{1}{m}\|\nabla_x f(\boldsymbol{x}_n^*,\boldsymbol{\theta}_n) - \nabla_x f(\boldsymbol{x}_n^*,\boldsymbol{\theta}_{n-1})\|_{L_2}$$

Taking the expectation over $\boldsymbol{\eta}_n$ and $\boldsymbol{\eta}_{n-1}$ yields

$$\begin{aligned}
&\|\nabla_x f(\boldsymbol{x}_n^*,\boldsymbol{\theta}_n) - \nabla_x f(\boldsymbol{x}_n^*,\boldsymbol{\theta}_{n-1})\|_{L_2} \\
&\quad = \|\nabla_x \tilde{f}(\boldsymbol{x}_n^*,\boldsymbol{\theta}_n) + \boldsymbol{\eta}_n - (\nabla_x \tilde{f}(\boldsymbol{x}_n^*,\boldsymbol{\theta}_{n-1}) + \boldsymbol{\eta}_{n-1})\|_{L_2} \\
&\quad \leq \tilde{M}\|\theta_{n+1} - \theta_n\|_2 + \|\boldsymbol{\eta}_n - \boldsymbol{\eta}_{n-1}\|_{L_2} \\
&\quad = \tilde{M}\delta + 2\sigma
\end{aligned}$$

This implies that

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_{L_2} \leq \frac{1}{m}\left(\tilde{M}\delta + 2\sigma\right)$$

A second case is one in which we control the change through the implicit function theorem. The optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$ for a fixed $\boldsymbol{\theta}$ must satisfy

$$\nabla_x f(\boldsymbol{x}^*(\boldsymbol{\theta}),\boldsymbol{\theta}) = 0.$$

Applying the implicit function theorem [16] to this problem, we have

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{x}^*(\boldsymbol{\theta}) = -\nabla_{xx} f(\boldsymbol{x}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\nabla_{x\theta} f(\boldsymbol{x}^*(\boldsymbol{\theta}),\boldsymbol{\theta}).$$

If we can uniformly bound the matrix on the right-hand side in the spectral norm by G, i.e.,

$$\| - \nabla_{xx} f(\boldsymbol{x}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\nabla_{x\theta} f(\boldsymbol{x}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\|_2 \leq G$$

then $\boldsymbol{x}^*(\boldsymbol{\theta})$ is Lipschitz in $\boldsymbol{\theta}$ with modulus G. Therefore,

$$\|\boldsymbol{x}^*_{n+1} - \boldsymbol{x}^*_n\|_{L_2} \leq G\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|_2 \leq G\delta$$

In both cases, we recover a bound on the change in the minimizers.

2.3 Tracking Analysis with Change in Minimizers Known

In this section we combine the bound $b(d_0, K)$ in Assumption A.4 with our model for the changes in functions in (1.2) to choose the number of stochastic gradients K needed to achieve desired mean criterion ε and IHP criterion (t, r) in (1.7) and (1.8). In this section, we assume that ρ is known. In Chapter 3, we will consider the case when ρ is unknown.

2.3.1 Mean Criterion Analysis

We show how to choose K to achieve a target mean criterion ε for all n. The idea behind the analysis is to proceed by induction using Assumption A.6 as a base case. Suppose that

$$\mathbb{E}[f_{n-1}(\boldsymbol{x}_{n-1})] - f_{n-1}(\boldsymbol{x}^*_{n-1}) \leq \varepsilon$$

Denote the distance from the initial point $\boldsymbol{x}_{n-1}$ to the minimizer $\boldsymbol{x}^*_n$ by $d_n(0)$, i.e.,

$$d_n(0) = \|\boldsymbol{x}_{n-1} - \boldsymbol{x}^*_n\|_2^2 \tag{2.9}$$

To bound $\mathbb{E}[d_n(0)]$ we first use the triangle inequality and (1.2) to get

$$\begin{aligned}\sqrt{\mathbb{E}[d_n(0)]} &\leq \|\boldsymbol{x}_{n-1} - \boldsymbol{x}^*_{n-1}\|_{L_2} + \|\boldsymbol{x}^*_{n-1} - \boldsymbol{x}^*_n\|_{L_2} \\ &\leq \|\boldsymbol{x}_{n-1} - \boldsymbol{x}^*_{n-1}\|_{L_2} + \rho\end{aligned}$$

By the strong convexity Assumption (A.2), we have

$$\frac{m}{2}\|\boldsymbol{x}_{n-1} - \boldsymbol{x}^*_{n-1}\|_2^2 \leq f_n(\boldsymbol{x}_{n-1}) - f_n(\boldsymbol{x}^*_{n-1}) \tag{2.10}$$

yielding

$$\mathbb{E}\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_{n-1}^*\|_2^2 \leq \frac{2}{m}\left(\mathbb{E}[f_n(\boldsymbol{x}_{n-1})] - f_n(\boldsymbol{x}_{n-1}^*)\right) \leq \frac{2\varepsilon}{m}$$

Putting everything together we have

$$\mathbb{E}[d_n(0)] \leq \left(\sqrt{\frac{2\varepsilon}{m}} + \rho\right)^2 \tag{2.11}$$

Therefore, we have a bound

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + \rho\right)^2, K\right)$$

and we can set

$$K^* = \min\left\{K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + \rho\right)^2, K\right) \leq \varepsilon\right\} \tag{2.12}$$

to ensure that

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon \quad \forall n \geq 1$$

2.3.2 IHP Tracking Error Analysis

For the IHP criterion, we assume that Assumptions A.1-A.6 hold. We seek an upper bound $r(t,K)$ such that

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\} \leq r(t,K) \quad \forall n \geq 1 \tag{2.13}$$

Using the mean criterion bounds of the previous section, we know that for all n

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon$$

Then by Markov's inequality, it holds that

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\} \leq \frac{\varepsilon}{t} \tag{2.14}$$

Although this bound always holds, we look at a way to tighten this bound. As before, we proceed by induction. As a base case, we can set

$$\mathbb{P}\{f_1(\boldsymbol{x}_1) - f_1(\boldsymbol{x}_1^*) > t\} \leq \frac{\varepsilon}{t}$$

Now, suppose that

$$\mathbb{P}\left\{f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*) > t\right\} \leq r_{n-1}(t)$$

and we want to construct a bound $r_n(t)$ on $\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\}$. We proceed by conditioning on $\{d_n(0) \leq \delta\}$ and $\{d_n(0) > \delta\}$ with $d_n(0)$ defined in (2.9) using the law of total probability:

$$\begin{aligned}
&\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\} \\
&\quad = \mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t \mid d_n(0) \leq \delta\}\,\mathbb{P}\{d_n(0) \leq \delta\} \\
&\qquad + \mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t \mid d_n(0) > \delta\}\,\mathbb{P}\{d_n(0) > \delta\}
\end{aligned}$$

For the first term, it holds that

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t \mid d_n(0) \leq \delta\} \leq \frac{1}{t} b(\delta, K) \triangleq \psi(t, \delta) \tag{2.15}$$

and

$$\mathbb{P}\{d_n(0) \leq \delta\} \leq 1$$

For the second term, it holds that

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t \mid d_n(0) > \delta\} \leq \psi(t, \operatorname{diam}^2(\mathscr{X}))$$

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\}$$
$$\leq \inf_{0<\delta\leq \mathrm{diam}^2(\mathscr{X})} \left\{ \psi(t,\delta) + \psi(t,\mathrm{diam}^2(\mathscr{X})) r_{n-1}\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)\right\} \tag{2.16}$$

$$r_n(t) = \inf_{0<\delta\leq \mathrm{diam}^2(\mathscr{X})} \left\{ \psi(t,\delta) + \psi(t,\mathrm{diam}^2(\mathscr{X})) r\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)\right\} \tag{2.17}$$

---

and

$$\begin{aligned}
&\mathbb{P}\{d_n(0) > \delta\} \\
&= \mathbb{P}\left\{\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_n^*\|_2^2 > \delta\right\} \\
&= \mathbb{P}\left\{\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_n^*\|_2 > \sqrt{\delta}\right\} \\
&\leq \mathbb{P}\left\{\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_{n-1}^*\|_2 + \rho > \sqrt{\delta}\right\} \\
&\leq \mathbb{P}\left\{\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_{n-1}^*\|_2 > \left(\sqrt{\delta}-\rho\right)_+\right\} \\
&\leq \mathbb{P}\left\{\sqrt{f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*)} > \sqrt{\frac{2}{m}}\left(\sqrt{\delta}-\rho\right)_+\right\} \\
&\leq r_{n-1}\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)
\end{aligned}$$

where $(x)_+ = \max\{x, 0\}$. Combining these bounds yields an overall bound

$$\begin{aligned}
&\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\} \\
&\quad \leq \psi(t,\delta) + \psi(t,\mathrm{diam}^2(\mathscr{X})) r_{n-1}\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)
\end{aligned}$$

We can optimize this bound over δ to yield the bound in (2.16). The quantity $\psi(t,\delta)$ defined in (2.15) can be replaced by any bound that also satisfies the inequality in (2.15). Therefore, we can set

$$r_n(t) = \inf_{0<\delta\leq \mathrm{diam}^2(\mathscr{X})} \left\{ \psi(t,\delta) + \psi(t,\mathrm{diam}^2(\mathscr{X})) r\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)\right\} \tag{2.18}$$

resulting in

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\} \leq r_n(t)$$

Bound at a Finite Number of Points

The bound of the preceding section is exact but difficult to compute. In this section, we introduce a computationally simpler bound. Computing the entire sequence of functions $r_n(t)$ is generally difficult, so we look at bounding

$$\mathbb{P}\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t\}$$

at a finite number of points $t^{(1)}, \ldots, t^{(N)}$ ordered in increasing order. We want to compute bounds $r_n(1), \ldots, r_n(N)$ such that

$$\mathbb{P}\left\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t^{(i)}\right\} \leq r_n(i) \quad i = 1, \ldots, N$$

We define an initial bound

$$r_1(i) = \frac{\varepsilon}{t^{(i)}}$$

Suppose that

$$\mathbb{P}\left\{f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*) > t^{(i)}\right\} \leq r_{n-1}(i)$$

Then as in (2.18), it follows that

$$\begin{aligned}
&\mathbb{P}\left\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t^{(i)}\right\} \\
&\leq \inf_{0<\delta\leq \mathrm{diam}^2(\mathscr{X})} \left\{\psi(t^{(i)}, \delta) + \psi(t^{(i)}, \mathrm{diam}^2(\mathscr{X}))\mathbb{P}\{d_n(0) > \delta\}\right\} \qquad (2.19)
\end{aligned}$$

The key then is to bound $\mathbb{P}\{d_n(0) > \delta\}$ in terms of $\{r_{n-1}(i)\}_{i=1}^N$. Define the function

$$t(\delta) = \max\{t^{(i)} \mid t^{(i)} \leq \delta\}$$

to be point $t^{(i)}$ closest to δ but not greater. Provided that $\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2 \geq t^{(1)}$ and $t^{(i)} = t\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)$ it holds that

$$\begin{aligned}
&\mathbb{P}\{d_n(0) > \delta\} \\
&\quad \leq \mathbb{P}\left\{f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}^*_{n-1}) > \frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right\} \\
&\quad \leq \mathbb{P}\left\{f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}^*_{n-1}) > t^{(i)}\right\} \\
&\quad \leq r_{n-1}(i) \qquad (2.20)
\end{aligned}$$

Otherwise, if $\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2 < t^{(1)}$, then

$$\mathbb{P}\{d_n(0) > \delta\} \leq \frac{\varepsilon}{\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+}$$

Define the overall bound for the term $\mathbb{P}\{d_n(0) > \delta\}$ as follows:

$$\phi_n(\delta) \triangleq \begin{cases} r_{n-1}\left(t\left(\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2\right)\right), & t^{(1)} \leq \frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2 \\ \frac{\varepsilon}{\frac{2}{m}\left(\sqrt{\delta}-\rho\right)_+^2}, & \text{else} \end{cases} \qquad (2.21)$$

Then we can set

$$r_n^{(i)} = \inf_{\delta>0}\left\{\psi(t^{(i)},\delta) + \psi(t^{(i)},\mathrm{diam}^2(\mathscr{X}))\phi_n(\delta)\right\}$$

This algorithm is summarized in Algorithm 1. In practice, once the bound $\psi(t^{(i)},\mathrm{diam}^2(\mathscr{X}))$ is less than one, then the gains are significant. Figure 2.3 plots a comparison of the bound produced by Algorithm 1 against the Markov inequality bound from (2.14) applied to the motivating example problem in Section 2.2.1 with $K_n = 300$.

Either the Markov bound of (2.14) or Algorithm 1 will produce valid upper bounds of the form

$$\mathbb{P}\left\{f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}^*_n) > t^{(i)}\right\} \leq r_n(i) \quad i = 1,\ldots,N$$

Suppose that the set $\{t^{(1)},\ldots,t^{(N)}\}$ contains t at index i^*. These bounds can in

Algorithm 1 Calculate IHP bounds

Require: Points $t^{(1)},\ldots,t^{(N)}$

1. Set
$$r_1^{(i)} = \frac{\varepsilon}{t^{(i)}} \quad i = 1,\ldots,N$$

2. Compute
$$r_n^{(i)} = \min_{0<\delta\leq \mathrm{diam}^2(\mathscr{X})} \left\{ \psi(t^{(i)},\delta) + \psi(t^{(i)},\mathrm{diam}^2(\mathscr{X}))\phi_n(\delta) \right\}$$

for $i = 1,\ldots,N$

3. $n \leftarrow n+1$ and go back to step 2

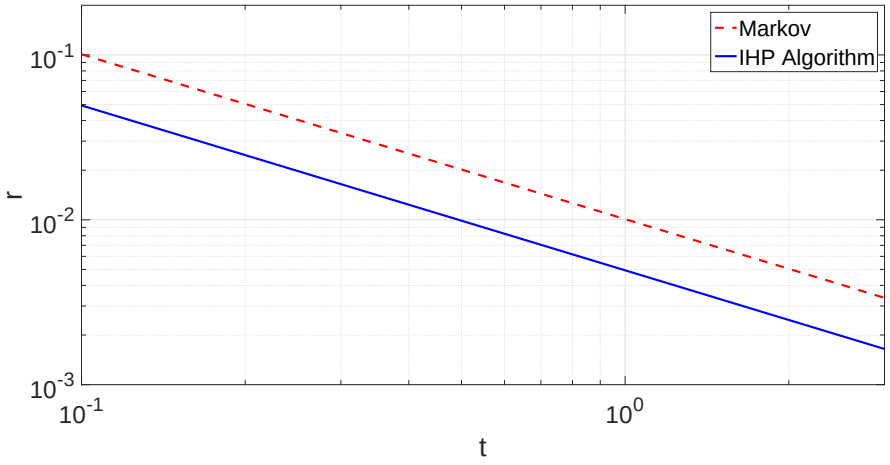


Figure 2.3: IHP algorithm plot

turn be used to select K_n to achieve a target (t,r) pair by selecting the smallest K_n such that

$$r_n(i^*) \leq r$$

Chapter 3

Change in Minimizers Unknown

We consider solving the sequence of optimization problems described in Section 1.2.1. In this chapter, we work under the assumption that

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_{L_2} \leq \rho$$

and the change in the minimizers, ρ, is unknown to us. First, we develop a method to estimate ρ and provide some theoretical guarantees for this estimate. We also consider an alternate condition on the minimizers

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_2 \leq \rho \tag{3.1}$$

using the Euclidean norm. This condition implies the L_2 condition, so any estimate of ρ for the L_2 condition also works for this conditions. However, in practice, we can provide an alternate, tighter estimate of ρ when the Euclidean norm condition holds. Using our estimates of ρ, we extend the analysis of Chapter 2 and provide similar guarantees for the mean criterion and IHP criterion. Finally, we discuss a few alternate methods to estimate ρ that do not seem to work as well in practice as the direct estimate. We include them in case useful applications are found at a later date. We make the same assumptions as Chapter 2 given in Assumptions A.1-A.6. The content of this chapter covers parts of the work in [26], [27], and [28].

3.1 Estimating the Change in Minimizers

In practice, we do not know ρ, so we must construct an estimate $\hat{\rho}_n$ using the stochastic gradients $\{\boldsymbol{g}_n(\boldsymbol{x}, z_n(k))\}_{k=1}^{K_n}$. First, we construct estimates $\tilde{\rho}_i$ for the one-step changes $\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|_2$ for each i. Next, we combine the one-step estimates to construct an overall estimate $\hat{\rho}_n$ for ρ. As an intermediary step, we also

look at a special case in which either

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_2 = \rho \tag{3.2}$$

or

$$\|\boldsymbol{x}_{n+1}^* - \boldsymbol{x}_n^*\|_{L_2} = \rho \tag{3.3}$$

We show that for our estimate $\hat{\rho}_n$ and appropriately chosen sequences $\{t_n\}$ for all n large enough $\hat{\rho}_n + t_n \geq \rho$ almost surely. With this property, analysis similar to that in Section 2.3 of Chapter 2 holds.

3.1.1 One-Step Changes

We construct an estimate $\tilde{\rho}_i$ for the one-step changes $\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|$. As a consequence of the strong convexity of $f_i(\boldsymbol{x})$, we have the following lemma:

Lemma 2. *It holds that*

$$\|\boldsymbol{x} - \boldsymbol{x}_i^*\|_2 \leq \frac{1}{m}\|\nabla f_i(\boldsymbol{x})\|_2 \quad \forall i \geq 1 \quad \forall \boldsymbol{x} \in \mathscr{X}$$

Proof. Since our functions $f_i(\boldsymbol{x})$ are convex, it holds that

$$\langle \nabla f_i(\boldsymbol{x}_i^*), \boldsymbol{x} - \boldsymbol{x}_i^* \rangle \geq 0 \quad \forall i \geq 1 \quad \forall \boldsymbol{x} \in \mathscr{X}$$

By the strong coercivity of the gradient, a consequence of strong convexity [29], it holds that

$$\langle \nabla f_i(\boldsymbol{x}) - \nabla f_i(\tilde{\boldsymbol{x}}), \boldsymbol{x} - \tilde{\boldsymbol{x}} \rangle \geq m\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2$$

Plugging in $\tilde{\boldsymbol{x}} = \boldsymbol{x}_i^*$ and $\boldsymbol{x}$ yields

$$\langle \nabla f_i(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_i^* \rangle \geq m\|\boldsymbol{x} - \boldsymbol{x}_i^*\|_2^2$$

Applying the Cauchy-Schwarz inequality yields the result. □

This in turn by way of the triangle inequality proves that

$$\begin{aligned}
&\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|_2 \\
&\quad \leq \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2 + \|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|_2 + \|\boldsymbol{x}_{i-1} - \boldsymbol{x}_{i-1}^*\|_2 \\
&\quad \leq \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2 + \frac{1}{m}\|\nabla f_i(\boldsymbol{x}_i)\|_2 + \frac{1}{m}\|\nabla f_{i-1}(\boldsymbol{x}_{i-1})\|_2
\end{aligned} \tag{3.4}$$

Motivated by this bound, we define the following estimate denoted the *direct estimate* by approximating the gradients

$$\tilde{\rho}_i \triangleq \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2 + \frac{1}{m}\|\hat{G}_i\|_2 + \frac{1}{m}\|\hat{G}_{i-1}\|_2 \tag{3.5}$$

where

$$\hat{G}_i = \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}_i, \boldsymbol{z}_i(k))$$

3.1.2 Combining with Constant Change of Minimizers

As a special case, we look at combining the one-step estimates when either (3.2) or (3.3) holds.

Euclidean Norm Condition

Under (3.2), we construct an estimate by averaging the one-step estimates

$$\hat{\rho}_n \triangleq \frac{1}{n-1}\sum_{i=2}^{n} \tilde{\rho}_i \tag{3.6}$$

We want to show that for an appropriate sequence $\{t_n\}$ and for all n large enough

$$\hat{\rho}_n + t_n \geq \rho$$

almost surely under (3.2) or (3.3). The difficulty in actually proving this condition for (3.5) is that when we compute

$$\hat{G}_i = \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}_i, \boldsymbol{z}_i(k))$$

$\boldsymbol{x}_i$ and $\{\boldsymbol{z}_i(k)\}_{k=1}^{K_i}$ are dependent. To get around this issue, we consider performing a second independent draw of samples $\{\tilde{\boldsymbol{z}}_i(k)\}_{k=1}^{K_i}$. Note that we do not need to actually draw new independent samples; this is purely for the sake of analysis. We start from $\boldsymbol{x}_{i-1}$ and produce $\tilde{\boldsymbol{x}}_i$ using these new samples. For example, with SGD, we have

$$\begin{aligned} \boldsymbol{x}_i(k) &= \Pi_{\mathscr{X}}\left[\boldsymbol{x}_i(k-1) - \mu(k)\boldsymbol{g}_i\left(\boldsymbol{x}_i(k-1), \boldsymbol{z}_i(k)\right)\right] \\ \tilde{\boldsymbol{x}}_i(k) &= \Pi_{\mathscr{X}}\left[\tilde{\boldsymbol{x}}_i(k-1) - \mu(k)\boldsymbol{g}_i\left(\boldsymbol{x}_i(k-1), \tilde{\boldsymbol{z}}_i(k)\right)\right] \end{aligned} \tag{3.7}$$

for $k = 1, \ldots, K_i$ with $\boldsymbol{x}_i(0) = \tilde{\boldsymbol{x}}_i(0) = \boldsymbol{x}_{i-1}$. Then we copy the form of the direct estimate using $\tilde{\boldsymbol{x}}_i$ in place of $\boldsymbol{x}_i$ by defining

$$\tilde{\rho}_i^{(2)} = \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_{i-1}\|_2 + \frac{1}{m}\|\tilde{G}_i\|_2 + \frac{1}{m}\|\tilde{G}_{i-1}\|_2 \tag{3.8}$$

with

$$\tilde{G}_i = \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i\left(\tilde{\boldsymbol{x}}_i, \boldsymbol{z}_i(k)\right)$$

In this case, $\tilde{\boldsymbol{x}}_i$ and $\{\boldsymbol{z}_i(k)\}_{k=1}^{K_i}$ are independent, so $\mathbb{E}[\tilde{\rho}_i^{(2)}] \geq \rho$ by Lemma 2. Under (3.2), using a dependent sub-Gaussian concentration inequality from [30] similar to Hoeffding's inequality, we then argue that $\hat{\rho}_n$ from (3.6) is close to

$$\hat{\rho}_n^{(2)} = \frac{1}{n-1}\sum_{i=2}^{n} \tilde{\rho}_i^{(2)}$$

which in turn upper bounds ρ for all n large enough almost surely. Similarly, under (3.3), we show that $\hat{\rho}_n^2$ from (3.10) is close to $(\hat{\rho}_n^{(2)})^2$, which in turn upper bounds ρ^2 for all n large enough almost surely.

To proceed with our analysis, suppose that the following conditions hold:

B.1 Suppose there exist functions $C_i(K_i)$ such that

$$\mathbb{E}\left[\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2^2 \mid \mathscr{F}_{i-1}\right] \leq C_i^2(K_i)$$

B.2 Suppose that it holds that

$$\mathbb{E}\left[\|\boldsymbol{g}_i\left(\boldsymbol{x}, \boldsymbol{z}_i\right) - \boldsymbol{g}_i\left(\tilde{\boldsymbol{x}}, \boldsymbol{z}_i\right)\|_2 \mid \mathscr{F}_{i-1}\right] \leq M\mathbb{E}\left[\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 \mid \mathscr{F}_{i-1}\right]$$

$\forall \boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathscr{X}$ and

$$\mathbb{E}\left[\|\boldsymbol{g}_i\left(\boldsymbol{x}, \boldsymbol{z}_i\right) - \nabla f_i(\boldsymbol{x})\|_2 \mid \mathscr{F}_{i-1}\right] \leq \boldsymbol{\sigma} \;\; \forall \boldsymbol{x} \in \mathscr{X}$$

$$D_n = \frac{1}{n-1}\left[\left(1+\frac{M}{m}\right)C_1 + \sqrt{\frac{\sigma}{K_1}} + 2\sum_{i=2}^{n-1}\left(\left(1+\frac{M}{m}\right)C_i + \sqrt{\frac{\sigma}{K_i}}\right) + \left(1+\frac{M}{m}\right)C_n + \sqrt{\frac{\sigma}{K_n}}\right] \tag{3.9}$$

B.3 Suppose that the gradients are bounded in the sense that

$$\|\boldsymbol{g}_n(\boldsymbol{x},\boldsymbol{z})\|_2 \leq G \quad \forall \boldsymbol{x} \in \mathscr{X}, \boldsymbol{z} \in \mathscr{Z}$$

Assumption B.1 is a bound on the difference between two independent outputs of the optimization algorithm $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_i$ starting from $\boldsymbol{x}_{i-1}$. Assumption B.2 controls how the gradient grows for two points $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$. Finally, Assumption B.3 is reasonable if the space $\mathscr{Z}$ that contains the $\boldsymbol{z}_n$ has finite diameter. In this case, it holds that

$$\sup_{\boldsymbol{x}\in\mathscr{X},\boldsymbol{z}\in\mathscr{Z}} \|\boldsymbol{g}_n(\boldsymbol{x},\boldsymbol{z})\|_2 < +\infty$$

We now show that the direct estimate from (3.6) upper bounds ρ from (3.2) eventually.

Theorem 1. *Provided that B.1-B.3 hold and our sequence* $\{t_n\}$[1] *satisfies*

$$\sum_{n=2}^{\infty}\left(\exp\left\{-\frac{(n-1)t_n^2}{18 diam^2(\mathscr{X})}\right\} + 2\exp\left\{-\frac{m^2(n-1)t_n^2}{72G^2}\right\}\right) < +\infty$$

it holds that for all n large enough

$$\hat{\rho}_n + D_n + t_n \geq \rho$$

almost surely with D_n *defined in* (3.9)

Proof. See Appendix B □

From now on, for notational convenience, we absorb D_n into the t_n term and refer only to $\hat{\rho}_n + t_n$.

[1]Note that a choice of t_n that is no greater than $1/\sqrt{n-1}$ works here.

L_2 Norm Condition

Under (3.3), we construct an estimate by averaging the squares of the one-step estimates and taking a square root

$$\hat{\rho}_n \triangleq \sqrt{\frac{1}{n-1}\sum_{i=2}^{n}\tilde{\rho}_i^2} \tag{3.10}$$

We now show that the direct estimate from (3.10) upper bounds ρ from (3.3) eventually.

Theorem 2. *Provided that B.1-B.3 hold and our sequence $\{t_n\}$ satisfies*

$$\sum_{n=2}^{\infty}\left(\exp\left\{-\frac{(n-1)t_n^2}{18diam^2(\mathscr{X})}\right\}+2\exp\left\{-\frac{m^2(n-1)t_n^2}{72G^2}\right\}\right)<+\infty$$

it holds that for all n large enough

$$\sqrt{(\hat{\rho}_n)^2+\tilde{D}_n+t_n}\geq\rho$$

almost surely with $\tilde{D}_n = 2diam(\mathscr{X})D_n$.

Proof. See Appendix B □

3.1.3 Combining with Bounded Changes of Minimizers

We examine estimating ρ in the case that either (1.2) or (3.1) holds. We denote the exact one-step time changes by $\rho_i \triangleq \|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|$. The simplest way to combine these estimates would be to set

$$\hat{\rho}_n = \max\{\tilde{\rho}_2,\ldots,\tilde{\rho}_n\}$$

For the sake of argument, suppose that $\tilde{\rho}_i = \rho_i + e_i$ with independent $e_i \sim \mathscr{N}(0,\sigma^2)$. Then it follows that [31]

$$\mathbb{E}[\hat{\rho}_n]\geq\mathbb{E}[\max\{e_2,\ldots,e_n\}]$$

For independent Gaussian random variables, it holds that $\mathbb{E}[\max\{e_2,\ldots,e_n\}]\to\infty$, so this estimate is guaranteed to blow up. We do produce an upper bound, but it

increases to the trivial bound $\text{diam}^2(\mathscr{X})$. Next, we examine how to avoid this issue.

Euclidean Norm Condition

Suppose that the following conditions hold.

B.4 We have estimates $\hat{h}_W : \mathbb{R}^W \to \mathbb{R}$ that are non-decreasing in their arguments such that

$$\mathbb{E}[\hat{h}_W(\rho_j, \ldots, \rho_{j-W+1})] \geq \rho$$

B.5 There exists absolute constants $\{b_i\}_{i=1}^W$ for any fixed $W \geq 1$ such that $\forall \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^W_{\geq 0}$

$$|\hat{h}_W(p_1, \ldots, p_W) - \hat{h}_W(q_1, \ldots, q_W)| \leq \sum_{i=1}^W b_i |p_i - q_i|$$

For example, if $\rho_i \overset{\text{iid}}{\sim} \text{Unif}[0, \rho]$, then

$$\hat{h}_W(\rho_i, \rho_{i+1}, \ldots, \rho_{i+W-1}) = \frac{W+1}{W} \max\{\rho_i, \rho_{i+1}, \ldots, \rho_{i+W-1}\}$$

is an estimate of ρ with the required properties with $b_i = 1 + \frac{1}{W}$. In this case, we compute the max over a sliding window and then average the maximums. This estimate will not blow up but will eventually upper bound ρ as we will see later.

Given an estimate satisfying Assumptions B.4-B.5, we compute

$$\bar{\rho}^{(i)} = \hat{h}_W(\tilde{\rho}_i, \tilde{\rho}_{i-1}, \ldots, \tilde{\rho}_{i-W+1})$$

and produce an estimate $\hat{\rho}_n$ that is an average of maximums of sliding windows of one-step estimates

$$\begin{aligned} \hat{\rho}_n &= \frac{1}{n-W} \sum_{i=W+1}^n \bar{\rho}^{(i)} \\ &= \frac{1}{n-W} \sum_{i=W+1}^n \hat{h}_{\min\{W, i-1\}}(\tilde{\rho}_i, \tilde{\rho}_{i-1}, \ldots, \tilde{\rho}_{\max\{i-W+1, 2\}}) \end{aligned} \tag{3.11}$$

Under Assumptions B.1-B.5, we can then show that

$$\hat{\rho}_n = \frac{1}{n-W} \sum_{i=W+1}^n \bar{\rho}^{(i)} \tag{3.12}$$

eventually upper bounds ρ.

Theorem 3. *Provided that B.1-B.5 hold and our sequence $\{t_n\}$ satisfies*

$$\sum_{n=2}^{\infty}\left(\exp\left\{-\frac{(n-W)^2t_n^2}{18(n-1)diam^2(\mathscr{X})\left(\sum_{j=1}^{W}b_j\right)^2}\right\}+2\exp\left\{-\frac{m^2(n-W)^2t_n^2}{72(n-1)G^2\left(\sum_{j=1}^{W}b_j\right)^2}\right\}\right)<+\infty \tag{3.13}$$

it holds that for all n large enough

$$\hat{\rho}_n+\left(\frac{n-1}{n-W}\sum_{j=1}^{W}b_j\right)D_n+t_n\geq\rho$$

with D_n from Theorem 1.

Proof. The proof in this case is similar to the proof for the equality assumption on ρ in (3.2) and is provided in Appendix B. □

As before, we will absorb $\left(\frac{n-1}{n-W}\sum_{j=1}^{W}b_j\right)D_n$ into t_n.

L_2 Norm Condition

Suppose that the following conditions hold, which are analogs of B.4-B.5:

B.6 We have estimates $\hat{h}_W:\mathbb{R}^W\to\mathbb{R}$ that are non-decreasing in their arguments such that

$$\mathbb{E}[\hat{h}_W(\rho_j^2,\ldots,\rho_{j-W+1}^2)]\geq\rho^2$$

B.7 There exist absolute constants $\{b_i\}_{i=1}^W$ for any fixed $W\geq 1$ such that $\forall \boldsymbol{p},\boldsymbol{q}\in\mathbb{R}_{\geq 0}^W$

$$|\hat{h}_W(p_1^2,\ldots,p_W^2)-\hat{h}_W(q_1^2,\ldots,q_W^2)|\leq\sum_{i=1}^{W}b_i|p_i^2-q_i^2|$$

For example, if $\rho_i\overset{\text{iid}}{\sim}\text{Unif}[0,\rho]$, then

$$\hat{h}_W\left(\rho_i^2,\rho_{i+1}^2,\ldots,\rho_{i+W-1}^2\right)=\frac{W+2}{W}\max\{\rho_i^2,\rho_{i+1}^2,\ldots,\rho_{i+W-1}^2\}$$

is an estimate of ρ with the required properties with $b_i = \frac{W+2}{W}$. In this case, we compute the max over a sliding window and then average the maximums. This estimate will not blow up but will eventually upper bound ρ as we will see later.

Given an estimate satisfying Assumptions B.4-B.5, we compute

$$\bar{\rho}^{(i)} = \sqrt{\hat{h}_W(\tilde{\rho}_i^2, \tilde{\rho}_{i-1}^2, \ldots, \tilde{\rho}_{i-W+1}^2)}$$

Under Assumptions B.1-B.3 and B.6-B.7, we can then show that

$$\hat{\rho}_n = \sqrt{\frac{1}{n-W} \sum_{i=W+1}^{n} \left(\bar{\rho}^{(i)}\right)^2} \tag{3.14}$$

eventually upper bounds ρ.

Theorem 4. *Provided that B.1-B.3 and B.6-B.7 hold and our sequence $\{t_n\}$ satisfies*

$$\sum_{n=2}^{\infty} \left(\exp\left\{ -\frac{(n-W)^2 t_n^2}{18(n-1)diam^2(\mathscr{X}) \left(\sum_{j=1}^{W} b_j\right)^2} \right\} + 2\exp\left\{ -\frac{m^2(n-W)^2 t_n^2}{72(n-1)G^2 \left(\sum_{j=1}^{W} b_j\right)^2} \right\} \right) < +\infty \tag{3.15}$$

it holds that for all n large enough

$$\sqrt{(\hat{\rho}_n)^2 + \left(\frac{n-1}{n-W} \sum_{j=1}^{W} b_j\right) \tilde{D}_n + t_n} \geq \rho$$

with $\tilde{D}_n$ from Theorem 2.

Proof. The proof in this case is similar to the proof for the equality assumption on ρ in (3.2) and is provided in Appendix B. □

3.2 Tracking Analysis with Change in Minimizers Unknown

We now examine the case with ρ unknown. We extend the work of Section 2.3 of Chapter 2 using the estimate of ρ in Section 3.1. Our analysis depends on the following crucial assumptions:

C.1 For appropriate sequences $\{t_n\}$, for all n sufficiently large it holds that $\hat{\rho}_n + t_n \geq \rho$ almost surely.

C.2 The bound $b(d_0, K_n)$ defined in Assumption A.4 factors as $b(d_0, K_n) = \alpha(K_n)d_0 + \beta(K_n)$

We have demonstrated that Assumption C.1 holds for the direct estimate of ρ.

In this section, we assume that either of the L_2 conditions (3.1) or (3.3) holds. Our analysis is not affected by which one is true. We note that the Euclidean norm conditions (1.2) and (3.2) imply the L_2 conditions (3.1) and (3.3) respectively. We use the following result, proved in Appendix C, to derive rules to pick K_n when ρ is unknown:

Theorem 5. *Under Assumptions C.1- C.2, with $K_n \geq K^*$ for all n large enough where*

$$K^* = \min \left\{ K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + \rho \right)^2, K \right) \leq \varepsilon \right\}, \tag{3.16}$$

we have

$$\limsup_{n \to \infty} \left(\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \right) \leq \varepsilon$$

almost surely.

Proof. See Appendix C. □

3.2.1 Update Past Mean Criterion Bounds

We first consider updating all past mean criterion bounds as we go. At time n, we plug-in $\hat{\rho}_{n-1} + t_{n-1}$ in place of ρ and follow the analysis of Section 2.3 of Chapter 2. Define

$$\hat{\varepsilon}_i^{(n)} \;=\; b\left(\left(\sqrt{\frac{2}{m}\hat{\varepsilon}_{i-1}^{(n)}} + (\hat{\rho}_{n-1} + t_{n-1}) \right)^2, K_i \right) \quad i = 1, \ldots, n$$

If it holds that $\hat{\rho}_{n-1} + t_{n-1} \geq \rho$, then $\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \hat{\varepsilon}_n^{(i)}$ for $i = 1, \ldots, n$. Assumption C.1 guarantees that this holds for all n large enough almost surely. We can thus set K_n equal to

$$K_n = \min \left\{ K \;\middle|\; b\left(\left(\sqrt{\frac{2}{m} \max\{\hat{\varepsilon}_{n-1}^{(n-1)}, \varepsilon\}} + (\hat{\rho}_{n-1} + t_{n-1}) \right)^2, K \right) \leq \varepsilon \right\} \tag{3.17}$$

for all $n \geq 3$ to achieve mean criterion ε. The maximum in this definition ensures that when $\hat{\rho}_{n-1} + t_{n-1} \geq \rho$, $K_n \geq K^*$ with K^* from (3.16). We can therefore apply Theorem 5.

3.2.2 Do Not Update Past Mean Criterion Bounds

Updating all past estimates of the mean criterion bounds from time 1 up to n imposes a computational and memory burden. Suppose that instead for all $n \geq 3$ we set

$$K_n = \min \left\{ K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + (\hat{\rho}_{n-1} + t_{n-1}) \right)^2, K \right) \leq \varepsilon \right\} \tag{3.18}$$

This is the same form as the choice in (3.16) with $\hat{\rho}_{n-1} + t_{n-1}$ in place of ρ. Due to Assumption C.1, for all n large enough it holds that $\hat{\rho}_n + t_n \geq \rho$ almost surely. Then by the monotonicity Assumption in A.4, for all n large enough we pick $K_n \geq K^*$ almost surely. We can therefore apply Theorem 5.

3.2.3 In High Probability Bounds

We can adopt the same approach as with ρ known by substituting $\hat{\rho}_{n-1} + t_{n-1}$ in place of ρ. As soon as $\rho_{n-1} + t_{n-1} \geq \rho$, Algorithm 1 will produce valid upper bounds of the form

$$\mathbb{P}\left\{ f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t^{(i)} \right\} \leq r_n(i) \quad i = 1, \ldots, N$$

Suppose that the set $\{t^{(1)}, \ldots, t^{(N)}\}$ contains t at index i^*. These bounds can in turn be used to select K_n to achieve a target (t, r) pair by selecting the smallest K_n

such that

$$r_n(i^*) \leq r$$

3.3 Experiment

We apply our framework to a mean-squared vector estimation problem similar to the scalar problem in Section 2.2.1 of Chapter 2. In Chapter 4, we apply the framework developed in this chapter to a variety of machine learning problems similar to (1.6) using real data. We fix the following signal model:

$$y_n = \boldsymbol{\eta}_n^\top \boldsymbol{w}_n + e_n.$$

Our goal is to estimate $\boldsymbol{\eta}_n$. We consider minimizing the following functions to estimate $\boldsymbol{\eta}_n$:

$$f_n(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\frac{1}{2}(y_n - \boldsymbol{x}^\top \boldsymbol{w}_n)^2\right] \tag{3.19}$$

By simple algebraic manipulation, it holds that

$$f_n(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\eta}_n)^\top \mathbb{E}[\boldsymbol{w}_n \boldsymbol{w}_n^\top](\boldsymbol{x} - \boldsymbol{\eta}_n) + \frac{1}{2}\mathbb{E}[e_n^2]. \tag{3.20}$$

It is easy to see then that $\boldsymbol{x}_n^* = \boldsymbol{\eta}_n$. Set

$$\boldsymbol{z}_n \triangleq [\boldsymbol{w}_n^\top \; e_n]^\top \tag{3.21}$$

and define the stochastic gradients

$$\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n) \triangleq -(y_n - \boldsymbol{x}^\top \boldsymbol{w}_n)\boldsymbol{w}_n$$

which satisfy the required condition in (1.5). To find approximate minimizers $\boldsymbol{x}_n$, we apply SGD using the inverse step size averaging technique discussed in Appendix A.

Let $\boldsymbol{w}_n \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma_w^2}{d}\boldsymbol{I}\right)$ where $\boldsymbol{w}_n \in \mathbb{R}^d$ and $e_n \sim \mathcal{N}(0, \sigma_e^2)$. We assume $\boldsymbol{\eta}_n$ is a deterministic sequence satisfying

$$\|\boldsymbol{\eta}_{n+1} - \boldsymbol{\eta}_n\|_2 \leq \rho \tag{3.22}$$

Since $\boldsymbol{x}_n^* = \boldsymbol{\eta}_n$, the minimizer change condition in (1.2) is satisfied and we use the

ρ estimate in (3.6). Note that $\{\boldsymbol{\eta}_n\}$ is deterministic, so we cannot apply a Kalman filter. Furthermore, we suppose that all $\boldsymbol{w}_n$, e_n, and $\boldsymbol{\eta}_n$ over all time instants are independent.

With this choice of model combined with the form of the functions in (3.20), it is clear that the functions $f_n(\boldsymbol{x})$ are strongly convex with $m = \sigma_{\boldsymbol{w}}^2/d$ satisfying Assumption A.2. By applying the inequality $(a+b) \leq 2a^2 + 2b^2$, it follows that

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n)\|_2^2 \\
&\quad = \mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}^*, \boldsymbol{z}_n) + (\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n) - \boldsymbol{g}_n(\boldsymbol{x}^*, \boldsymbol{z}_n))\|_2^2 \\
&\quad \leq 2\mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}^*, \boldsymbol{z}_n)\|_2^2 + 2\mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n) - \boldsymbol{g}_n(\boldsymbol{x}^*, \boldsymbol{z}_n)\|_2^2
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}_n^*, \boldsymbol{z}_n)\|_2^2 &= \mathbb{E}\|e_n \boldsymbol{w}_n\|_2^2 \\
&= \mathbb{E}[e_n^2]\mathbb{E}\|\boldsymbol{w}_n\|_2^2 \\
&= \sigma_e^2 \sigma_{\boldsymbol{w}}^2
\end{aligned}$$

and for the second term, we have

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n) - \boldsymbol{g}_n(\boldsymbol{x}^*, \boldsymbol{z}_n)\|_2^2 \\
&\quad = \mathbb{E}\left[\|\boldsymbol{w}_n\|_2^2 (\boldsymbol{x} - \boldsymbol{x}^*) \boldsymbol{w}_n \boldsymbol{w}_n^\top (\boldsymbol{x} - \boldsymbol{x}^*)\right] \\
&\quad \leq \mathbb{E}\left[\|\boldsymbol{w}_n\|_2^4\right] \|\boldsymbol{x} - \boldsymbol{x}^*)\|_2^2 \\
&\quad \leq 3\sigma_w^4 \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2
\end{aligned}$$

The last inequality follows due to the fact the $\mathbb{E}|x|^4 \leq 3\mathbb{E}|x|^2$ for x a centered Gaussian. This implies that

$$\mathbb{E}\|\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z}_n)\|_2^2 \leq 2\sigma_e^2 \sigma_{\boldsymbol{w}}^2 + 6\sigma_w^4 \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2$$

Therefore, for Assumption A.5, we can set

$$\begin{aligned}
A &= 2\sigma_e^2 \sigma_{\boldsymbol{w}}^2 \\
B &= 6\sigma_w^4
\end{aligned}$$

Putting it together, we have the parameters summarized in Table 3.1. For this simulation, we choose $d = 2$, $\sigma_w^2 = 0.5$, $\sigma_e^2 = 0.5$, and $\rho = 1$.

Table 3.1: Parameter Table

Parameter	Value
m	σ_w^2/d
A	$2\sigma_e^2\sigma_w^2$
B	$6\sigma_w^4$

3.3.1 Mean Tracking Criterion

First, we assume that ρ and all the parameters in Table 3.1 are known. We focus on the mean tracking criterion in (1.7). Figure 3.1 shows the trade-off for the optimal ε versus K^* defined in (3.16). Any pair (ε, K) located above this curve can be achieved in the sense that by setting $K_n = K^*$, we achieve

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon$$

For comparison, we plot the mean criterion curve achievable using the choice

$$K_n = \min\left\{K \geq 1 \;\middle|\; b\left(\text{diam}^2(\mathscr{X}), K\right) \leq \varepsilon\right\}$$

For a fixed value of K_n, the mean criterion achievable using K^* is substantially smaller than the value achieved using the $\text{diam}^2(\mathscr{X})$ bound.

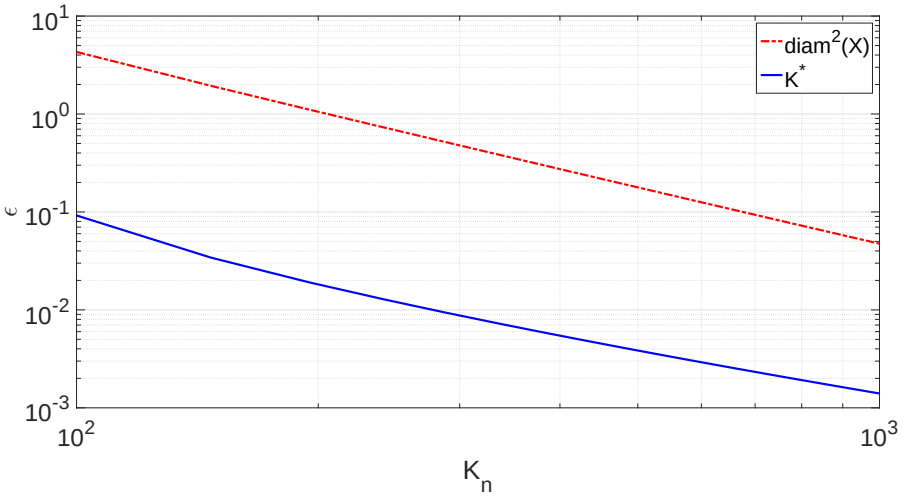


Figure 3.1: ε vs. K

Next, we examine the case where ρ and the parameters in Table 3.1 are unknown. We estimate ρ using the techniques introduced in Section 3.1, specifically

(3.12), select K_n using the rule in (3.18), and estimate the parameters using the techniques in Appendix D. We target several different values of the mean tracking accuracy ε from (1.7) including 0.001, 0.01, and 0.03. For the problem in this section, we can compute an estimate of the mean criterion to evaluate our methods. First, we have $f_n(\boldsymbol{x}_n^*) = \frac{1}{2}\sigma_e^2$. Second, for the sake of evaluation, we draw additional samples $\{\tilde{\boldsymbol{z}}_n(k)\}_{k=1}^{T_n} \overset{\text{iid}}{\sim} p_n$ and compute

$$\frac{1}{T_n}\frac{1}{2}\sum_{k=1}^{T_n}\left(\tilde{y}_n(k) - \boldsymbol{x}_n^\top \tilde{\boldsymbol{w}}_n(k)\right)^2 \tag{3.23}$$

to estimate $f_n(\boldsymbol{x}_n)$. With these two pieces, we can estimate the mean criterion by computing

$$\frac{1}{T_n}\frac{1}{2}\sum_{k=1}^{T_n}\left(\tilde{y}_n(k) - \boldsymbol{x}_n^\top \tilde{\boldsymbol{w}}_n(k)\right)^2 - \frac{1}{2}\sigma_e^2 \tag{3.24}$$

Table 3.2 shows an estimate of the actual achieved mean criterion for three different ε mean criterion targets averaged over $n = 1$ to 100. In all cases, we meet our mean criterion target on average.

Table 3.2: Estimate of Mean Criterion

ε	Mean Criterion Estimate
0.001	0.0008 ± 0.0002
0.01	0.0073 ± 0.0012
0.03	0.022 ± 0.0022

Figure 3.2 shows the estimate of ρ. Our estimates of ρ upper bound the true value of $\rho = 1$ as desired. With a smaller mean criterion target, we produce a tighter estimate of ρ. Figure 3.3 shows the selected number of samples K_n for each mean tracking target ε. As expected, for smaller choices of ε, K_n is larger. In addition, for any given choice of ε, K_n settles down and is roughly constant for large n. Figure 3.4 shows the estimate $\hat{\varepsilon}_{i,n}$ of the mean criterion achieved computed by updating the past. In comparison to Table 3.2, we see that the mean criterion estimates in Figure 3.4 are reasonable upper bounds.

Finally, we examine estimation of the parameters m, A, and B used in the above simulations. We carry out the methods describe in Appendix D and compare them to the true values in Table 3.1. Due to space constraints, we only include the estimate of m in Figures 3.5. As desired, we have a lower bound on the strong convexity parameter m.

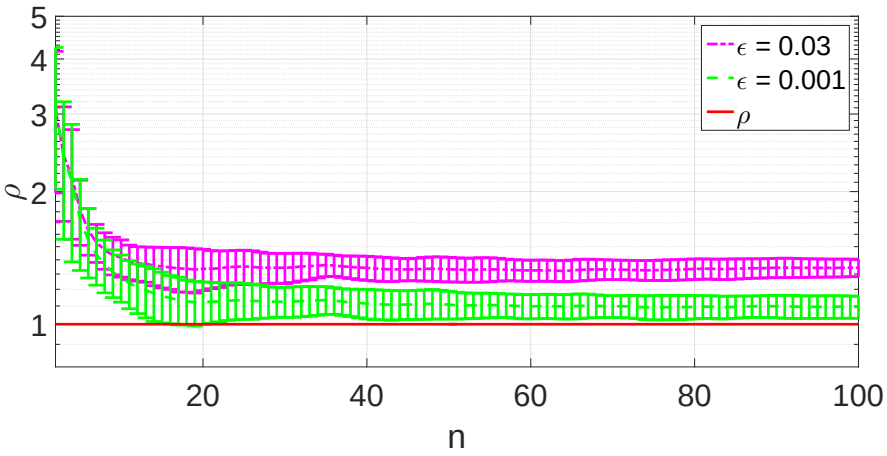


Figure 3.2: Estimate of ρ from (1.2)

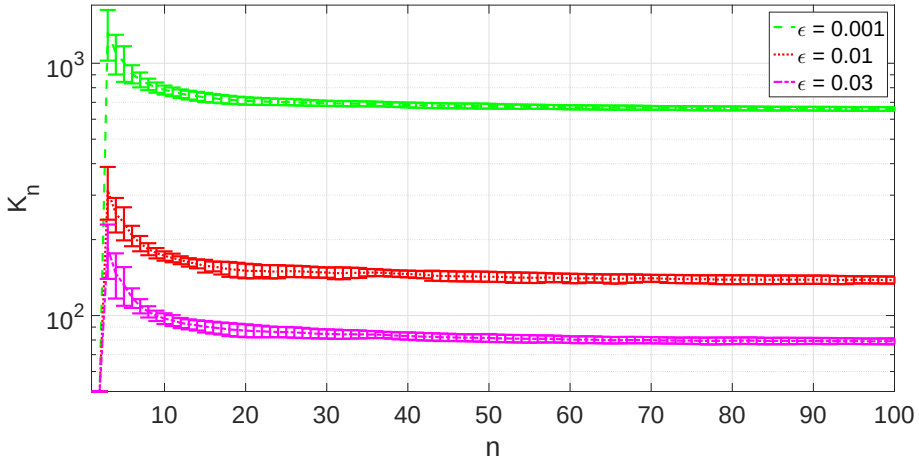


Figure 3.3: Selected K_n

3.3.2 IHP Tracking Criterion

Figure 3.6 plots r vs. ε for several values of K by applying the IHP algorithm. The IHP bounds appear to be loose in general as we need fairly large values of K to get non-trivial bounds for reasonable ε and small r. The looseness of these bounds is not surprising, since we are only using the first moment of the tracking error to bound.

We choose K_n by targeting $t = 0.1$ and $r = 0.25$. Figure 3.7 shows the resulting empirical probability. As mentioned above, we can compute $f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*)$ exactly, so we can calculate the fraction of the time that the loss violates the $t = 0.1$

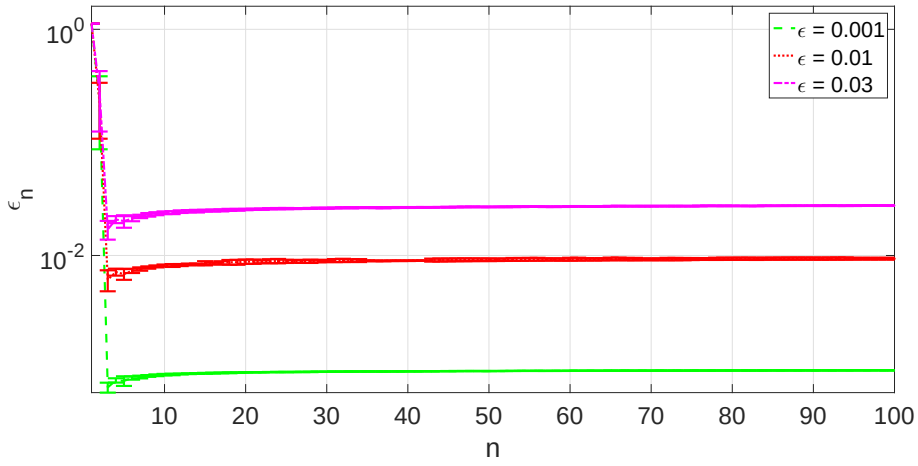


Figure 3.4: Estimate of mean tracking accuracy

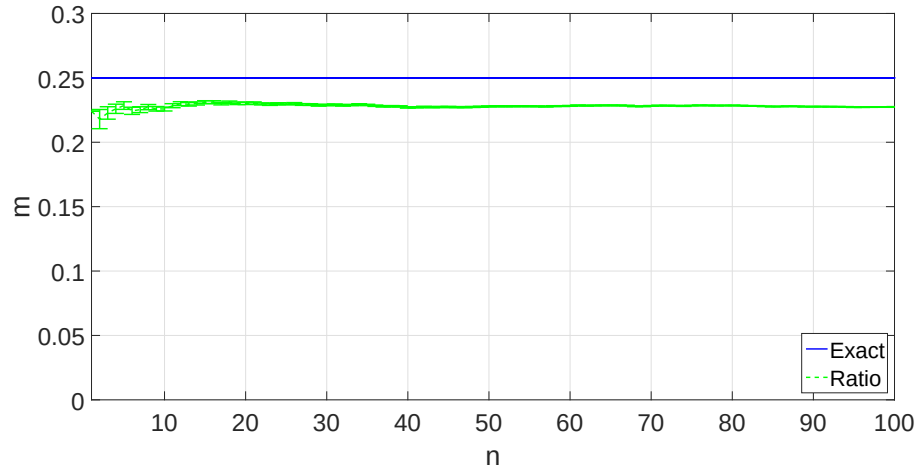


Figure 3.5: Estimate of m

constraint. The empirical probability that

$$f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) > t$$

satisfies our target value of $r = 0.25$.

3.3.3 Kalman Filter Comparison

We now consider a slight modification of our model, so that we can apply the Kalman filter. As mentioned above, since we assume that $\boldsymbol{\eta}_n$ is generated as a

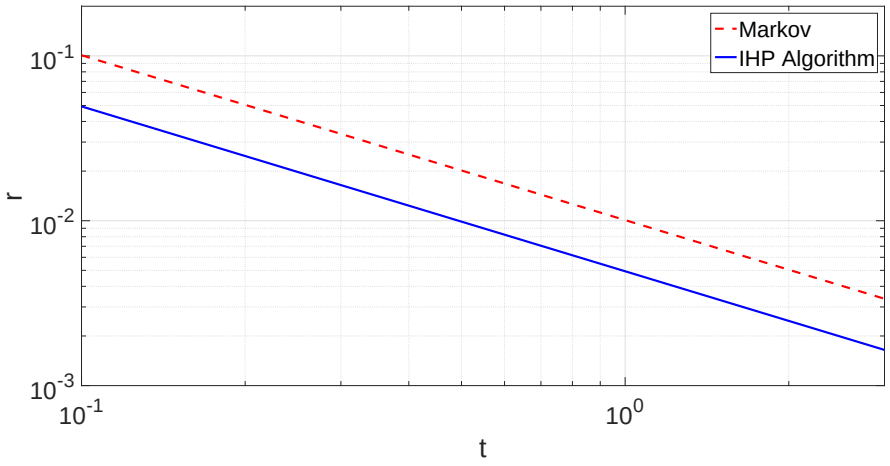


Figure 3.6: r vs. t

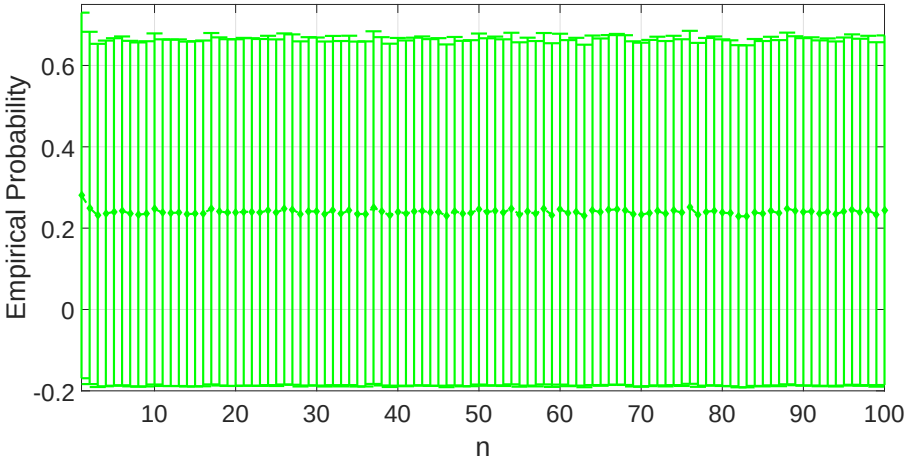


Figure 3.7: Empirical probability

deterministic sequence, we cannot apply the Kalman filter. In this section, we instead assume that

$$\boldsymbol{\eta}_n - \boldsymbol{\eta}_{n-1} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

and $\boldsymbol{\eta}_1$ is fixed as in (2.2). Then it holds that

$$\|\boldsymbol{\eta}_n - \boldsymbol{\eta}_{n-1}\|_{L_2} \leq \sigma \triangleq \rho$$

as in (2.2). We satisfy (3.1) and use the estimate of ρ in (3.10).

To apply the Kalman filter, we take $\boldsymbol{\eta}_n$ to be the state of the system. The state

evolution equation is given by

$$\boldsymbol{\eta}_n(k) = \begin{cases} \boldsymbol{\eta}_1(1), & \text{fixed} \\ \boldsymbol{\eta}_{n-1}(K_{n-1}) + \zeta_n, & k = 1 \\ \boldsymbol{\eta}_n(k-1), & 1 < k \leq K_n \end{cases}$$

with $\zeta_n \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$. The observation equation is given by the pair $(\boldsymbol{w}_n(k), y_n(k))$ with

$$y_n(k) = \boldsymbol{\eta}_n^\top(k)\boldsymbol{w}_n(k) + e_n(k)$$

Let $\hat{\boldsymbol{\eta}}_n(k|\tilde{k})$ be the estimate of $\boldsymbol{\eta}_n$ at time k of epoch n given all the information up to $\tilde{k}$ with $k \geq \tilde{k}$. Let $P_n(k|\tilde{k})$ be the estimate of the covariance. The prediction equations for the state estimate and covariance estimate are given by [32]

$$\begin{aligned} \hat{\boldsymbol{\eta}}_n(k|k-1) &= \hat{\boldsymbol{\eta}}_n(k-1|k-1) \\ P_n(k|k-1) &= P_n(k-1|k-1) + \sigma^2 \boldsymbol{I} \mathbb{1}_{\{k=1\}} \end{aligned} \tag{3.25}$$

The update equations are given by

$$\begin{aligned} \hat{\boldsymbol{\eta}}_n(k|k) &= \hat{\boldsymbol{\eta}}_n(k|k-1) + \boldsymbol{G}_n(k)(y_n(k) - \hat{\boldsymbol{\eta}}_n^\top(k|k-1)\boldsymbol{w}_n(k)) \\ P_n(k|k) &= (\boldsymbol{I} - G_n(k)\boldsymbol{w}_n^\top(k))P_n(k|k-1) \\ G_n(k) &= P_n(k|k-1)\boldsymbol{w}_n(k)\left(\sigma_e^2 + \boldsymbol{w}_n^\top(k)P_n(k|k-1)\boldsymbol{w}_n(k)\right)^{-1} \end{aligned}$$

where $G_n(k)$ is the Kalman gain. We have the initial conditions

$$\begin{aligned} \hat{\boldsymbol{\eta}}_n(1|0) &= \hat{\boldsymbol{\eta}}_{n-1}(K_{n-1}|K_{n-1}) \\ P_n(1|0) &= P_{n-1}(K_{n-1}|K_{n-1}) \end{aligned}$$

Figure 3.8 shows a comparison of the Kalman filter against our SGD based approach both with exact and mismatched parameters for the Kalman filter. Table 3.3 uses the technique from (3.24) to estimate the mean criterion for all three methods. The Kalman filter receives the number of samples K_n chosen by the SGD approach. With correct parameters for the Kalman filter, both methods achieve similar performance, but the SGD method is able to control its desired accuracy. With incorrect parameters, the Kalman filter's performance is worse.

Table 3.3: Kalman Filter Comparison

Method	Mean Criterion Estimate
Direct Estimate	$1.9 \times 10^{-2} \pm 1.1 \times 10^{-3}$
Kalman Filter	$1.8 \times 10^{-2} \pm 5.7 \times 10^{-3}$
Kalman Filter - Mismatch	$9.4 \times 10^{-2} \pm 5.3 \times 10^{-2}$

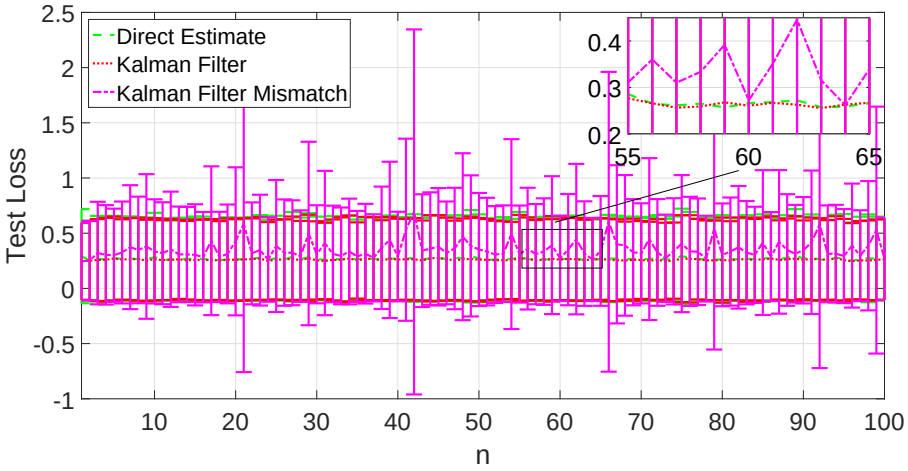


Figure 3.8: Comparison of Kalman filter and our approach

3.4 Alternate Methods to Estimate the Change in the Minimizers

We present three additional methods to estimate the one-step change in the minimizers, ρ, in this section. These methods generally are looser than the direct estimate and are more complicated, so they are not as useful. We include them for completeness and to provide an idea of possible alternatives. Some theorems analogous to those in Theorems 1 and 3 are provided in Appendix B.4.

3.4.1 One-Step Changes

Direct Estimate with Resampling

We introduce a *resampled direct estimate* that is computationally more complex but easier to analyze. To motivate this alternative approach, note that since $\boldsymbol{x}_i$ is

computed using the same samples used to compute

$$\hat{G}_i \triangleq \frac{1}{K_i} \sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}_i, \boldsymbol{z}_i(k))$$

it generally holds that

$$\mathbb{E}\left[\frac{1}{K_i} \sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}_i, \boldsymbol{z}_i(k)) \;\middle|\; \boldsymbol{x}_i \right] \neq \nabla f_i(\boldsymbol{x}_i)$$

If $\boldsymbol{x}_i$ was independent of the samples used to compute $\hat{G}_i$, then it would hold that

$$\mathbb{E}\left[\hat{G}_i \mid \boldsymbol{x}_i\right] = \nabla f_i(\boldsymbol{x}_i) \tag{3.26}$$

We consider a resampling based method that does satisfy (3.26).

Fix a positive integer $\Delta K > 0$ such that $\Delta K < K_i$ for all i. We choose R subsets $T_i^1, \ldots, T_i^R$ of $\{1, \ldots, K_i\}$ of size ΔK with no repeats in a set. We compute approximate minimizers $\boldsymbol{x}^{(T_i^b)}$ generated from the samples not in T_i^b, i.e.,

$$\left\{ \boldsymbol{z}(k) \mid k \notin T_i^b \right\}$$

As before, it holds that

$$\begin{aligned} \|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\| \leq \|\boldsymbol{x}_i^{(T_i^b)} - \boldsymbol{x}_{i-1}^{(T_{i-1}^b)}\| &+ \frac{1}{m} \left\| \nabla_{\boldsymbol{x}} f_i\left(\boldsymbol{x}_i^{(T_i^b)}\right) \right\| \\ &+ \frac{1}{m} \left\| \nabla_{\boldsymbol{x}} f_{i-1}\left(\boldsymbol{x}_i^{(T_i^b)}\right) \right\| \end{aligned}$$

This suggests that we define the estimate

$$\begin{aligned} \tilde{\rho}_i \triangleq \frac{1}{R} \sum_{b=1}^{R} \Bigg(\|\boldsymbol{x}_i^{(T_i^b)} - \boldsymbol{x}_{i-1}^{(T_{i-1}^b)}\| &+ \frac{1}{m} \left\| \frac{1}{|T_i^b|} \sum_{\boldsymbol{z} \in T_i^b} \boldsymbol{g}_i\left(\boldsymbol{x}_i^{(T_i^b)}, \boldsymbol{z}_i\right) \right\| \\ &+ \frac{1}{m} \left\| \frac{1}{|T_{i-1}^b|} \sum_{\boldsymbol{z} \in T_{i-1}^b} \boldsymbol{g}_{i-1}\left(\boldsymbol{x}_{i-1}^{(T_{i-1}^b)}, \boldsymbol{z}_{i-1}\right) \right\| \Bigg) \end{aligned} \tag{3.27}$$

which satisfies $\mathbb{E}[\tilde{\rho}_i] \geq \rho$.

Integral Probability Metric Estimate

In this section, we consider functions $f_n(\boldsymbol{x})$ of the form

$$f_n(\boldsymbol{x}) \triangleq \mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\ell(\boldsymbol{x}, \boldsymbol{z}_n)\right]$$

This naturally arises in the machine learning context when $\ell(\boldsymbol{x}, \boldsymbol{z})$ corresponds to a loss function where $\boldsymbol{x}$ parameterizes the model and $\boldsymbol{z}_n$ is the piece of data at time n.

Scalar Integral Probability Metric Estimate Given a class of functions $\mathscr{F}$ where each $h \in \mathscr{F}$ maps $\mathscr{Z} \to \mathbb{R}$, an integral probability metric (IPM) [33] between two distributions p and q on $\mathscr{Z}$ is defined to be

$$\gamma_{\mathscr{F}}(p, q) \triangleq \sup_{h \in \mathscr{F}} \left| \mathbb{E}_{\boldsymbol{z} \sim p}[h(\boldsymbol{z})] - \mathbb{E}_{\tilde{\boldsymbol{z}} \sim q}[h(\tilde{\boldsymbol{z}})] \right| \tag{3.28}$$

Lemma 3 shows how an IPM can be used to bound the change in minimizer at time i as long as the class of functions $\mathscr{F}$ is rich enough with respect to the loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$.

Lemma 3. *Assume that* $\{\ell(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathscr{X}\} \subset \mathscr{F}$. *Then*

$$\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\| \leq \frac{2}{m} \gamma_{\mathscr{F}}(p_i, p_{i-1})$$

Proof. By our strong convexity Assumption A.2 and [34],

$$\begin{aligned} f_i(\boldsymbol{x}_{i-1}^*) &\geq f_i(\boldsymbol{x}_i^*) + \frac{1}{2} m \|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|^2 \\ f_{i-1}(\boldsymbol{x}_i^*) &\geq f_{i-1}(\boldsymbol{x}_{i-1}^*) + \frac{1}{2} m \|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|^2 \end{aligned}$$

Adding and rearranging these inequalities yields

$$m\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|^2 \leq (f_i(\boldsymbol{x}_i^*) - f_{i-1}(\boldsymbol{x}_i^*)) + \left(f_{i-1}(\boldsymbol{x}_{i-1}^*) - f_i(\boldsymbol{x}_{i-1}^*)\right) \tag{3.29}$$

Since $\{\ell(\boldsymbol{x}_i^*, \cdot) : \boldsymbol{x} \in \mathscr{X}\} \subset \mathscr{F}$, for the first term on the right-hand side of (3.29)

we have

$$\begin{aligned} f_i(\boldsymbol{x}_i^*) - f_{i-1}(\boldsymbol{x}_i^*) \\ &\leq \left|\mathbb{E}_{\boldsymbol{z}_i \sim p_i}[\ell(\boldsymbol{x}_i^*, \boldsymbol{z}_i)] - \mathbb{E}_{\boldsymbol{z}_{i-1} \sim p_{i-1}}[\ell(\boldsymbol{x}_i^*, \boldsymbol{z}_{i-1})]\right| \\ &\leq \gamma_{\mathscr{F}}(p_i, p_{i-1}) \end{aligned}$$

Repeating this argument for the other term in (3.29) and rearranging completes the proof. □

Since we do not know p_i and p_{i-1}, we cannot compute $\gamma_{\mathscr{F}}(p_i, p_{i-1})$. Instead, we estimate the IPM. With K_i and K_{i-1} samples from p_i and p_{i-1} respectively, we can plug in the empirical distributions $\hat{p}_i$ and $\hat{p}_{i-1}$ to yield the estimate

$$\frac{2}{m}\gamma_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1}) \tag{3.30}$$

It is easy to see that empirical IPM is biased upward, i.e.,
$\mathbb{E}\left[\frac{2}{m}\gamma_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1})\right] \geq \frac{2}{m}\gamma_{\mathscr{F}}(p_i, p_{i-1})$.

This estimate is not in a closed form, since the IPM in (3.30) still involves taking the supremum over $\mathscr{F}$. However, for certain classes of functions, evaluating this supremum can be reduced to solving an optimization problem in a finite-dimensional space. Suppose that class of functions $\mathscr{F}$ is of the form

$$\mathscr{F} = \left\{ f \;\middle|\; |f(\boldsymbol{z}) - f(\tilde{\boldsymbol{z}})| \leq r(\boldsymbol{z}, \tilde{\boldsymbol{z}}) \quad \forall \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathscr{Z} \right\}$$

for a function $r(\boldsymbol{z}, \tilde{\boldsymbol{z}})$. Plugging empiricals into (3.4.1) shows that evaluating the IPM in (3.30) is equivalent to evaluating

$$\sup_{h \in \mathscr{F}} \left| \frac{1}{K_i} \sum_{k=1}^{K_i} h(\tilde{\boldsymbol{z}}_i(k)) - \frac{1}{K_{i-1}} \sum_{k=1}^{K_{i-1}} h(\tilde{\boldsymbol{z}}_i(K_i + k)) \right|$$

We relax by optimizing over the value of the function h at $\tilde{\boldsymbol{z}}(k)$ denoted α_k. Consider the following linear program (LP) where the constraints constitute a necessary condition for the values $\{\alpha_i\}$ to correspond to a 1-Lipschitz function:

$$\begin{aligned} &\text{maximize } \frac{1}{K_i} \sum_{k=1}^{K_i} \alpha_k - \frac{1}{K_{i-1}} \sum_{k=1}^{K_{i-1}} \alpha_{K_i + k} \\ &\text{subject to } \alpha_k - \alpha_j \leq r(\tilde{\boldsymbol{z}}_i(k), \tilde{\boldsymbol{z}}_i(j)) \quad \forall k \neq j \end{aligned}$$

For any function $h \in \mathscr{F}$, the point $\alpha_k = h(\tilde{z}_i(k))$ is feasible; therefore, this LP gives an upper bound on the IPM. In fact, authors in [33] showed that the value of this LP is actually the IPM for $\mathscr{F}$ the class of 1-Lipschitz functions.

Vector Integral Probability Metric Estimate Given a class of functions $\mathscr{F}$ where each $h \in \mathscr{F}$ maps $\mathscr{Z} \to \mathbb{R}$, an integral probability metric (IPM) [33] between two distributions p and q on $\mathscr{Z}$ is defined to be

$$\gamma_{\mathscr{F}}(p,q) \triangleq \sup_{h \in \mathscr{F}} \left| \mathbb{E}_{z \sim p}[h(z)] - \mathbb{E}_{\tilde{z} \sim q}[h(\tilde{z})] \right|$$

We consider an extension of this idea, which we call a *vector IPM*, in which the class of functions $\mathscr{F}$ maps $\mathscr{Z} \to \mathscr{X}$:

$$\gamma^{\mathrm{V}}_{\mathscr{F}}(p,q) \triangleq \sup_{f \in \mathscr{F}} \| \mathbb{E}_{z \sim p}[f(z)] - \mathbb{E}_{\tilde{z} \sim q}[f(\tilde{z})] \| \tag{3.31}$$

Lemma 4 shows that a vector IPM can be used to bound the change in minimizer at time i and follows from variational inequalities in [16] and the assumption that $\{\nabla_x \ell(x,\cdot) : x \in \mathscr{X}\} \subset \mathscr{F}$.

Lemma 4. *Assume that* $\{\nabla_x \ell(x,\cdot) : x \in \mathscr{X}\} \subset \mathscr{F}$. *Then we have*

$$\|x_i^* - x_{i-1}^*\| \leq \frac{1}{m} \gamma^{V}_{\mathscr{F}}(p_i, p_{i-1})$$

Proof. By exploiting variational inequalities from [16], we can show that

$$\begin{aligned}
&\|x_i^* - x_{i-1}^*\| \\
&\quad \leq \frac{1}{m} \|\nabla_x f_i(x_{i-1}^*) - \nabla_x f_{i-1}(x_{i-1}^*)\| \\
&\quad = \frac{1}{m} \|\mathbb{E}_{z_i \sim p_i}\left[\nabla_x \ell(x_{i-1}^*, z_i)\right] - \mathbb{E}_{z_{i-1} \sim p_{i-1}}\left[\nabla_x \ell(x_{i-1}^*, z_{i-1})\right] \|
\end{aligned}$$

By assumption $\{\nabla_x \ell(x_{i-1}^*,\cdot) : x \in \mathscr{X}\} \subset \mathscr{F}$, so

$$\begin{aligned}
&\|\nabla_x f_i(x_{i-1}^*) - \nabla_x f_{i-1}(x_{i-1}^*)\| \\
&\qquad = \|\mathbb{E}_{z_i \sim p_i}\left[\ell(x_{i-1}^*, z_i)\right] - \mathbb{E}_{z_{i-1} \sim p_{i-1}}\left[\ell(x_{i-1}^*, z_{i-1})\right] \| \\
&\qquad \leq \sup_{f \in \mathscr{F}} \|\mathbb{E}_{z_i \sim p_i}[f(z_i)] - \mathbb{E}_{z_{i-1} \sim p_{i-1}}[f(z_{i-1})] \| \\
&\qquad = \gamma^{\mathrm{V}}_{\mathscr{F}}(p_i, p_{i-1})
\end{aligned}$$

□

We cannot compute this vector IPM, since we do not know the distributions p_i and p_{i-1}. Instead, we plug in the empiricals $\hat{p}_i$ and $\hat{p}_{i-1}$ to yield the estimate $\frac{1}{m}\gamma^{\mathrm{V}}_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1})$. This estimate is biased upward, which ensures that $\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\| \leq \mathbb{E}\left[\frac{1}{m}\gamma^{\mathrm{V}}_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1})\right]$.

Our estimate is still not in a closed form since there is a supremum over $\mathscr{F}$ in the computation of $\gamma^{\mathrm{V}}_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1})$. For the class of functions

$$\mathscr{F} = \left\{h \;\middle|\; \|h(\boldsymbol{z}) - h(\tilde{\boldsymbol{z}})\| \leq r(\boldsymbol{z}, \tilde{\boldsymbol{z}})\right\} \tag{3.32}$$

we can compute an upper bound Γ_i on $\gamma^{\mathrm{V}}_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1})$ yielding a computable estimate $\tilde{\rho}_i = \frac{1}{m}\Gamma_i$. Set $\tilde{\boldsymbol{z}}_i(k) = \boldsymbol{z}_i(k)$ if $1 \leq k \leq K_i$ and $\tilde{\boldsymbol{z}}_i(k) = \boldsymbol{z}_{i-1}(k)$
if $K_i + 1 \leq k \leq K_i + K_{i-1}$. From (3.31), we have

$$\gamma^{\mathrm{V}}_{\mathscr{F}}(\hat{p}_i, \hat{p}_{i-1}) = \sup_{h \in \mathscr{F}} \left\| \frac{1}{K_i}\sum_{k=1}^{K_i} h(\tilde{\boldsymbol{z}}_i(k)) - \frac{1}{K_{i-1}}\sum_{k=1}^{K_{i-1}} h(\tilde{\boldsymbol{z}}_i(K_i + k)) \right\|$$

We can relax this supremum by maximizing over the function value $f(\tilde{\boldsymbol{z}}_i(k))$ denoted by α_k in the following non-convex quadratically constrained quadratic program (QCQP):

$$\begin{aligned} &\text{maximize} \left\| \frac{1}{K_i}\sum_{k=1}^{K_i} \alpha_k - \frac{1}{K_{i-1}}\sum_{k=1}^{K_{i-1}} \alpha_{K_i + k} \right\| \\ &\text{subject to } \|\alpha_k - \alpha_j\| \leq r(\tilde{\boldsymbol{z}}_i(k), \tilde{\boldsymbol{z}}_i(j)) \quad \forall k < j \end{aligned}$$

The constraints are imposed to ensure that the function values α_k can correspond to a function in $\mathscr{F}$ from (3.32). The value of this QCQP may not exactly equal the vector IPM but at least provides an upper bound.

Comparison of Estimates

The direct estimate is easier to compute but may be loose if $\|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|$ is large. If $\|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|$ is large, then the IPM approaches are generally tighter. However, the IPM estimates are more difficult to compute due to need to solve an LP or QCQP and check the inclusion conditions in Lemma 4. Also, the number of constraints in the LP or QCQP grows quadratically in the number of samples.

3.4.2 Combining the One-Step Estimates

Direct Estimate with Resampling

For the resampling direct estimate from (3.27), the analysis is straightforward. The following lemma guarantees that averages eventually upper bound ρ.

Theorem 6. *For any sequence* $\{t_n\}$ *such that*

$$\sum_{n=2}^{\infty} \exp\left\{-\frac{(n-1)t_n^2}{2diam^2(\mathscr{X})}\right\} < +\infty$$

it holds that for all n large enough

$$\hat{\rho}_n + t_n \geq \rho$$

almost surely

Proof. The proof in this case is similar to the proof for the equality assumption on ρ in (3.2) and is provided in Appendix B. □

This estimate is computationally more complex, but is substantially easier to analyze. The case with the inequality constraint definition of ρ is straightforward.

IPM Estimates

We have the following lemma characterizing the performance of the IPM estimates with a constant change in minimizers.

Lemma 5. *For both IPM estimates and any sequence* $\{t_n\}$ *such that*

$$\sum_{n=2}^{\infty} \exp\left\{-\frac{nt_n^2}{4diam(\mathscr{X})^2}\right\} < \infty$$

for all n large enough it holds that $\hat{\rho}_n + t_n \geq \rho$ *almost surely.*

Proof. Define the random variables

$$V_i = \tilde{\rho}_i - \mathbb{E}\left[\tilde{\rho}_i \mid \mathscr{F}_{i-2}\right]$$

with $\{\mathscr{F}_i\}_{i=1}^n$ defined in (2.6). We have

$$-\text{diam}(\mathscr{X}) \leq V_i \leq \text{diam}(\mathscr{X})$$

Clearly, V_i is $\mathscr{F}_i$-measurable and $\mathbb{E}[V_i \mid \mathscr{F}_{i-2}] = 0$. Now, we can apply Lemma 19 of Appendix B with $W = 2$ to yield

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^n V_i < -nt\right\} &\leq \exp\left\{-\frac{2(nt)^2}{(2)\left(4n\text{diam}^2(\mathscr{X})\right)}\right\} \\
&= \exp\left\{-\frac{nt^2}{4\text{diam}^2(\mathscr{X})}\right\}
\end{aligned}$$

None of the random variables $\{z_i(k)\}_{k=1}^{K_i}$ and $\{z_{i-1}(k)\}_{k=1}^{K_{i-1}}$ are $\mathscr{F}_{i-2}$ measurable. Also, regardless of how many samples K_i and K_{i-1} are taken, the IPM estimate is biased upward. Thus, it holds that

$$\mathbb{E}[\tilde{\rho}_i \mid \mathscr{F}_{i-2}] \geq \rho$$

Therefore, it follows that

$$\begin{aligned}
\mathbb{P}\{\hat{\rho}_n < \rho - t\} &\leq \mathbb{P}\left\{\sum_{i=1}^n \tilde{\rho}_i < \sum_{i=1}^n \mathbb{E}[\tilde{\rho}_i \mid \mathscr{F}_{i-2}] - nt\right\} \\
&= \mathbb{P}\left\{\sum_{i=1}^n V_i < -nt\right\} \\
&\leq \exp\left\{-\frac{nt^2}{4\text{diam}^2(\mathscr{X})}\right\}
\end{aligned}$$

Note that we pay a price of two in the exponent due to $\tilde{\rho}_i$ and $\tilde{\rho}_{i-1}$ both depending on the samples from p_{i-1}. Since

$$\sum_{n=2}^{\infty} \exp\left\{-\frac{nt_n^2}{4\text{diam}(\mathscr{X})^2}\right\} < \infty$$

it follows that

$$\sum_{n=2}^{\infty} \mathbb{P}\{\hat{\rho}_n + t < \rho\} < +\infty$$

This in turn guarantees by way of the Borel-Cantelli lemma that for n large enough

$$\hat{\rho}_n + t_n \geq \rho$$

almost surely. □

We have the following lemma characterizing the performance of the IPM estimates with a bounded change in minimizers.

Lemma 6 (IPM One-Step Estimates)**.** *For the estimate in* (3.11) *computed using either IPM estimate for $\tilde{\rho}_i$ and any sequence $\{t_n\}$ such that*

$$\sum_{n=2}^{\infty} \exp\left\{-\frac{2(n-1)t_n^2}{(W+1)diam(\mathscr{X})^2}\right\} < \infty$$

it holds that for all n large enough $\hat{\rho}_n + t_n \geq \rho$ almost surely.

Proof. We copy the proof of Lemma 5 of Appendix B with $W+1$ in place of 2 and note that $\tilde{\rho}^{(i)}$ and $\tilde{\rho}^{(j)}$ with $|i-j| > W+1$ do not depend on the same samples. Lemma 19 and some simple algebra yield

$$\mathbb{P}\{\hat{\rho}_n < \rho - t\} \leq \exp\left\{-\frac{2(n-1)t^2}{(W+1)\text{diam}(\mathscr{X})^2}\right\}$$

We pay a price of $W+1$ in the denominator of the exponent due to the dependence of the $\tilde{\rho}^{(i)}$. By the Borel-Cantelli Lemma, for all n large enough it holds that $\hat{\rho}_n + t_n \geq \rho$ almost surely as long as

$$\sum_{n=2}^{\infty} \exp\left\{-\frac{2(n-1)t_n^2}{(W+1)\text{diam}(\mathscr{X})^2}\right\} < \infty$$

□

3.4.3 Experiment

To compare the various one-step estimates of ρ, we run the simulation example of Section 3.3 again with $d = 1$ under the assumption that

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 \leq \rho$$

from (3.6). We compare the direct estimate, direct estimate with resampling, and both IPM methods in Figure 3.9. The IPM estimates are the loosest, the direct estimate with resampling is the second closest, and the direct estimate is the tightest.

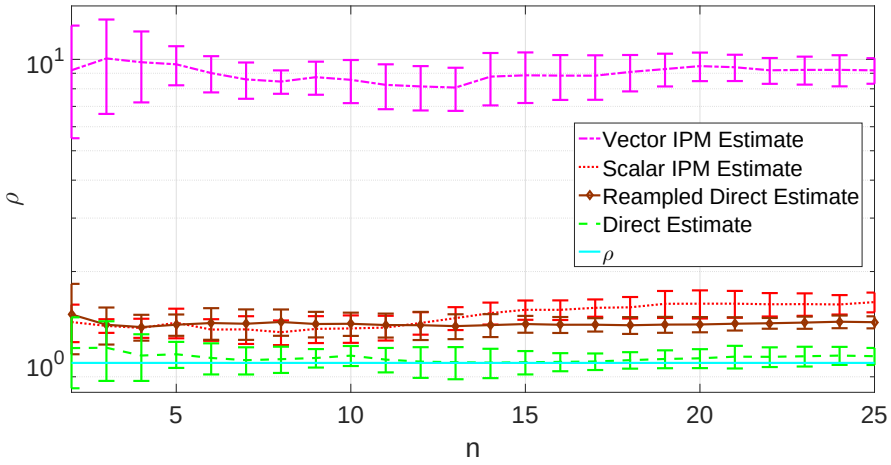


Figure 3.9: Various one step estimates

Chapter 4

Applications to Machine Learning

We consider a specialization of the model introduced in Section 1.2.1 to machine learning problems. Consider solving a sequence of machine learning problems such as regression or classification by minimizing the expected value of a fixed loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$ at each time n:

$$\min_{\boldsymbol{x} \in \mathscr{X}} \left\{ f_n(\boldsymbol{x}) \triangleq \mathbb{E}_{\boldsymbol{z}_n \sim p_n} \left[\ell(\boldsymbol{x}, \boldsymbol{z}_n) \right] \right\} \quad \forall n \geq 1 \tag{4.1}$$

For regression, $\boldsymbol{z}_n$ corresponds to the predictors and response pair at time n and $\boldsymbol{x}$ parameterizes the regression model. For classification $\boldsymbol{z}_n$ corresponds to the feature and label pair at time n and $\boldsymbol{x}$ parameterizes the classifier. Although motivated by regression and classification, our framework works for any loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$ that satisfies certain properties discussed in Chapter 2. In the learning context, a *task* consists of the loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$ and the distribution p_n, and so our problem can be viewed as learning a sequence of tasks.

This chapter is a specialization of the work in Chapters 2 and 3 for general functions $f_n(\boldsymbol{x})$ to the specific form in (4.1) applied to real world data sets. We also consider a few extensions that are useful for machine learning problems. In the machine learning context, the mean criterion is referred to as *excess risk*. We make the same assumptions as Chapter 2 given in Assumptions A.1-A.6. The content of this chapter covers parts of the work in [26], [27], and [35].

4.1 Related Work

We introduce some prior work that is relevant to the machine learning context. The prior work in this section complements the prior work in Section 2.1.

Our problem has connections with *multi-task learning* (MTL) and *transfer learning*. In multi-task learning, one tries to learn several tasks simultaneously as

in [36], [37], and [38] by exploiting the relationships between the tasks. In transfer learning, knowledge from one source task is transferred to another target task either with or without additional training data for the target task [39]. Multi-task learning could be applied to our problem by running a MTL algorithm each time a new task arrives, while remembering all prior tasks. However, this approach incurs a memory and computational burden. Transfer learning lacks the sequential nature of our problem. For multi-task and transfer learning, there are theoretical guarantees on regret for some algorithms [40].

We can also consider the *concept drift* problem in which we observe a stream of incoming data that potentially changes over time, and the goal is to predict some property of each piece of data as it arrives. After prediction, we incur a loss that is revealed to us. For example, we could observe a feature $\boldsymbol{w}_n$ and predict the label y_n as in [41]. Some approaches for concept drift use iterative algorithms such as SGD, but without specific models on how the data changes. As a result, only simulation results showing good performance are available. There are also some bandit approaches in which one of a finite number of predictors must be applied to the data as in [42]. For this approach, there are regret guarantees using techniques for analyzing bandit problems.

Another relevant model is *sequential supervised learning* (see [43]) in which we observe a stream of data consisting of feature/label pairs $(\boldsymbol{w}_n, y_n)$ at time n, with $\boldsymbol{w}_n$ being the feature vector and y_n being the label. At time n, we want to predict y_n given $\boldsymbol{x}_n$. One approach to this problem, studied in [44] and [45], is to look at L consecutive pairs $\{(\boldsymbol{w}_{n-i}, y_{n-i})\}_{i=1}^{L}$ and develop a predictor at time n by applying a supervised learning algorithm to this training data. Another approach is to assume that there is an underlying hidden Markov model (HMM) [46]. The label y_n represents the hidden state and the pair $(\boldsymbol{w}_n, \bar{y}_n)$ represents the observation with $\bar{y}_n$ being a noisy version of y_n. HMM inference techniques are used to estimate y_n.

None of the prior work discussed in this section involves choosing the number of samples K_n at each time n to control the excess risk. Most approaches instead focus on bounding the regret or provide no guarantees.

4.2 Extensions for Real-World Applications

4.2.1 Cost Approach

In our experiments in Section 4.3, we run our algorithm to choose a number of samples $\{K_n\}_{n=1}^T$ over a horizon of length T using the choice in (3.18). We then compare against taking

$$\sum_{n=1}^{T} K_n$$

samples at time $n = 1$ and no samples at the other $T-1$ time instants. In this section, we consider a different type of comparison based on assuming that there is a cost $c(K_n)$ of taking K_n samples. For example, we could have

$$c(K) = C_0 \mathbb{1}_{\{K>0\}} + C_1 K \tag{4.2}$$

This implies we pay a fixed cost of C_0 any time we take at least one sample and a marginal cost of C_1 per sample. We want to control the excess risk by deciding when to take samples, and how many samples to take with a total budget C over a horizon of length T, i.e.,

$$\sum_{n=1}^{T} c(K_n) \leq C \tag{4.3}$$

We can compare our approach against any other approach that respects the cost budget in (4.3). One option is to again take all samples up front

$$K_n = \begin{cases} \max\left\{K \geq 1 \;\middle|\; c(K) \leq C\right\}, & n = 1 \\ 0, & 2 \leq n \leq T \end{cases} \tag{4.4}$$

Another option is to sample every ΔT time instants and divide the cost budget evenly over the times that we take samples using

$$K_n = \begin{cases} \max\left\{K \geq 1 \;\middle|\; c(K) \leq \left\lfloor \frac{C}{T/\Delta T} \right\rfloor \right\}, & \Delta T | (n-1) \\ 0, & \text{else} \end{cases} \tag{4.5}$$

where the notation $a|b$ means a divides b.

For analysis, we need Assumption C.1 and the following additional assumptions:

E.1 There exists a function $e(\|\boldsymbol{x}-\boldsymbol{x}_n^*\|_2^2)$ such that

$$f_n(\boldsymbol{x}) - f_n(\boldsymbol{x}_n^*) \leq e(\|\boldsymbol{x}-\boldsymbol{x}_n^*\|_2^2)$$

For example, suppose that the functions $f_n(\boldsymbol{x})$ have Lipschitz continuous gradients with modulus M and $\boldsymbol{x}_n^* \in \text{int}(\mathscr{X})$ for all $n \geq 1$ where $\text{int}(\mathscr{X})$ is the interior of $\mathscr{X}$. By the descent lemma [34], we have

$$\begin{aligned} f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) &\leq \langle \nabla f_n(\boldsymbol{x}_n^*), \boldsymbol{x}_n - \boldsymbol{x}_n^* \rangle + \frac{1}{2} M \|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|_2^2 \\ &= \frac{1}{2} M \|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|_2^2 \end{aligned}$$

Thus, we can set

$$e(\|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|_2^2) = \frac{1}{2} M \|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|_2^2$$

Since we need to consider the possibility that $K_n = 0$ for some n in $\{1,\ldots,T\}$ but still provide estimates of the excess risk, we need an alternate version of the bound in (2.11). Define

$$t_s(n) = \max\{m \mid 1 \leq m \leq n \text{ and } K_m > 0\}$$

where $t_s(n)$ is the last time no later than n at which samples were taken. If no samples have been taken so far, then by convention $t_s(n) = +\infty$. We construct the recursively defined function $\tilde{b}_n(\rho, K_n)$ by considering the following four cases:

1. No samples have been taken by time n:

$$\tilde{b}_n(\rho, K_n) \triangleq e(\text{diam}(\mathscr{X}))$$

2. Samples taken at time n for the first time

$$\tilde{b}_n(\rho, K_n) \triangleq b(\text{diam}^2(\mathscr{X}), K_n)$$

3. No samples taken at time n but samples have been taken previously

$$\tilde{b}_n(\rho, K_n) \triangleq e\left(\left(\sqrt{\frac{2}{m}\tilde{b}_{t_s(n-1)}} + ((n - t_s(n-1))\rho)\right)^2\right)$$

4. Samples taken at time n and samples have been taken previously

$$\tilde{b}_n(\rho, K_n) \triangleq b\left(\frac{4}{m}\tilde{b}_{t_s(n-1)} + 2\left((n - t_s(n-1))\rho\right)^2, K_n\right)$$

Suppose that over a time horizon of length T we have a total cost budget C with respect to the number of samples $\{K_n\}_{n=1}^T$ as in (4.3). Define the *excess risk gaps*

$$\xi_n = \left(\tilde{b}_n(\rho, K_n) - \varepsilon\right)_+$$

with $(x)_+ = \max\{x, 0\}$. The variable ξ_n is the extent to which the target excess risk of ε is violated upwards. If our excess risk is below our target level ε, then we set $\xi_n = 0$. Our goal is to minimize the size of the ξ_n, while taking into account the cost constraint in (4.3). To control the size of ξ_n, suppose that we have a function $\phi : \mathbb{R}^T \to \mathbb{R}$ that describes the cumulative loss of the excess risk gaps $\xi_1, \ldots, \xi_T$. We provide a few possible choices for $\phi(\xi_1, \ldots, \xi_T)$

1. $\phi(\xi_1, \ldots, \xi_T) = \frac{1}{T}\sum_{n=1}^T \xi_t$
2. $\phi(\xi_1, \ldots, \xi_T) = \max\{\xi_1, \ldots, \xi_T\}$
3. Set
$$\phi(\xi_1, \ldots, \xi_T) = \max_{(a,b)\in\tau} \sum_{n=a}^{b} \xi_n$$
with
$$\tau = \{(a,b) \mid a < b\ ,\ \xi_a \leq \xi_{a+1} \leq \cdots \leq \xi_b\}$$

The first two conditions penalize the average and maximum excess risk gaps respectively. In practice, with the first two choices, we will stop taking samples before the horizon T resulting in relatively poor performance towards the end of the horizon. To combat this problem we introduce the third criterion that penalizes large increasing runs of excess risk gaps. This penalty does not have the drawbacks of the first two penalties and tends to favor a more uniform choice of the number of samples K_n.

We consider the case when ρ is known to us and plan over the horizon of length

T by solving the following optimization problem:

$$\begin{aligned}
&\underset{K_1,\ldots,K_T}{\text{minimize}}\ \phi(\xi_1,\ldots,\xi_T) \\
&\text{subject to}\ \sum_{n=1}^{T} c(\rho,K_n) \le C \\
&\qquad \mathbb{1}_{\{K_1>0\}} \le \mathbb{1}_{\{K_2>0\}} \\
&\qquad \mathbb{1}_{\{K_n>0\}} \le \mathbb{1}_{\{K_{n-1}>0\}} + \mathbb{1}_{\{K_{n+1}>0\}} \quad n=2,\ldots,T-1 \\
&\qquad \mathbb{1}_{\{K_{T-1}>0\}} \le \mathbb{1}_{\{K_T>0\}} \\
&\qquad K_n \in \mathbb{Z}_{\ge 0} \qquad\qquad\qquad\qquad n=1,\ldots,T
\end{aligned} \tag{4.6}$$

The idea of this problem is to satisfy the excess risk bound ε with minimal violation $\phi(\xi_1,\ldots,\xi_T)$. To estimate ρ, we need samples from consecutive time instants. Therefore, we impose the constraint that if we take samples at time n, then we must take samples either at either time $n-1$ or time $n+1$ through the constraint

$$\mathbb{1}_{\{K_n>0\}} \le \mathbb{1}_{\{K_{n-1}>0\}} + \mathbb{1}_{\{K_{n+1}>0\}}$$

This is a mixed integer non-linear programming problem (MINLP). There are no general methods to efficiently solve MINLP, so we consider a relaxation of this problem later. In the case that we know ρ, we can plan the number of samples ahead of time before any samples have been taken.

When ρ is unknown, we cannot plan over the entire horizon. Instead, at each time instant m we have to plan over the remaining time horizon of length $T-m+1$, while using the estimate $\hat{\rho}_{m-1}+t_{m-1}$ in place of ρ and the remaining cost budget

$$C - \sum_{n=1}^{m-1} c(K_n)$$

Under Assumption C.1, we know that we eventually use an upper bound on ρ and produce conservative estimates of the excess risk. We consider the cost-to-go

problem

$$
\begin{aligned}
&\underset{K_m,\ldots,K_T}{\text{minimize}}\ \phi(\xi_m,\ldots,\xi_T) \\
&\text{subject to}\ \sum_{n=m}^{T} c(K_n) \le C - \sum_{n=1}^{m-1} c(K_n) \\
&\qquad \mathbb{1}_{\{K_m>0\}} \le \mathbb{1}_{\{K_{m+1}>0\}} \\
&\qquad \mathbb{1}_{\{K_m>0\}} \le \mathbb{1}_{\{K_{m-1}>0\}} + \mathbb{1}_{\{K_{m+1}>0\}} \quad n = m+1,\ldots,T-1 \\
&\qquad \mathbb{1}_{\{K_{T-1}>0\}} \le \mathbb{1}_{\{K_T>0\}} \\
&\qquad K_n \in \mathbb{Z}_{\ge 0} \qquad\qquad\qquad\qquad n = m,\ldots,T
\end{aligned}
\tag{4.7}
$$

This is the same form as (4.6) except over the time horizon from $n = m,\ldots,T$ taking into account the portion of the cost budget that has been expended. In this problem, we only optimize over $K_m,\ldots,K_T$. This problem is again a MINLP.

Next, we look at approximate solutions to (4.6) and (4.7). The major difficulties in solving these programs are that the decision variables $\{K_n\}_{n=1}^{T}$ are integer-valued and the cost function $c(K)$ may be discontinuous at zero due to fixed costs. We consider relaxing K_n to be real-valued and introduce a piecewise approximation $\hat{c}(K)$ of the cost functions $c(K)$:

$$
\hat{c}(K) = \left(\frac{c(K_0)K}{K_0}\right)\mathbb{1}_{\{K\le K_0\}} + c(K)\mathbb{1}_{\{K>K_0\}}
$$

Generally, we pick $0 < K_0 < 1$. We consider the relaxed program

$$
\begin{aligned}
&\underset{K_1,\ldots,K_T}{\text{minimize}}\ \phi(\xi_1,\ldots,\xi_T) \\
&\text{subject to}\ \sum_{n=1}^{T} \hat{c}(\rho,K_n) \le C \\
&\qquad K_1 \le K_2 \\
&\qquad K_n \le K_{n-1} + K_{n+1} \quad n = 2,\ldots,T-1 \\
&\qquad K_{T-1} \le K_T \\
&\qquad K_n \in \mathbb{R}_{\ge 0} \qquad\qquad n = 1,\ldots,T
\end{aligned}
\tag{4.8}
$$

We also relax the indicator constraints to inequality to encourage taking samples

at consecutive times. In practice, this forces more gradual changes in samples K_n and makes it easier to solve these problems. This problem can be readily solved by gradient based solvers such as IPOPT [47].

4.2.2 Cross Validation

We can also apply cross validation for model selection. Suppose we have loss functions $\ell_\lambda(\boldsymbol{x}, \boldsymbol{z})$ parameterized by λ, which controls the model complexity. For example, we could have a quadratic penalty term

$$\ell_\lambda(\boldsymbol{x}, \boldsymbol{z}) = \tilde{\ell}(\boldsymbol{x}, \boldsymbol{z}) + \frac{1}{2}\lambda\|\boldsymbol{x}\|_2^2$$

The value of $\lambda = 0$ corresponds to the true loss function that we want to minimize. Suppose we have C different values $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(C)}$ of λ under consideration. For each $\lambda^{(i)}$, we generate an approximate minimizer $\boldsymbol{x}_n^{(i)}$ of

$$\mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\ell_{\lambda^{(i)}}(\boldsymbol{x}, \boldsymbol{z}_n)\right] \tag{4.9}$$

We want to select the value $\lambda^{(i)}$ and corresponding $\boldsymbol{x}_n^{(i)}$ that achieves the smallest loss

$$\mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\ell_0(\boldsymbol{x}_n^{(i)}, \boldsymbol{z}_n)\right] \tag{4.10}$$

We generate an approximate minimizer $\boldsymbol{x}_n^{(i)}$ for each problem in (4.9) starting from $\boldsymbol{x}_{n-1}^{(i)}$. To select the best choice of $\lambda^{(i^*)}$ in terms of minimizing (4.10), we apply cross validation and set $\boldsymbol{x}_n = \boldsymbol{x}_n^{(i^*)}$ [48].

The idea behind cross validation is to divide the training samples $\{\boldsymbol{z}_n(k)]\}_{k=1}^{K_n}$ into P equal sized pieces. For every $P-1$ out of P pieces, we use the $P-1$ pieces of the training set to generate an approximate solution $\tilde{\boldsymbol{x}}_n^{(i)}$ to (4.9). We use the remaining piece of the training set to evaluate the empirical test loss achieved by $\tilde{\boldsymbol{x}}_n^{(i)}$ as in (3.23). We do this for every possible choice of $P-1$ out of P pieces and average the empirical test loss estimates. We then select the value $\lambda^{(i^*)}$ that achieves the smallest empirical test loss.

To apply cross validation to our framework, we run C parallel versions of our approach and at time n we generate C different choices for the number of samples $K_n^{(i)}$. We then choose

$$K_n = \max\{K_n^{(1)}, \ldots, K_n^{(C)}\}$$

After choosing K_n, we apply the usual cross validation approach to select $\lambda^{(i)}$ for time n. Figure 4.1 shows this approach for two values of λ.

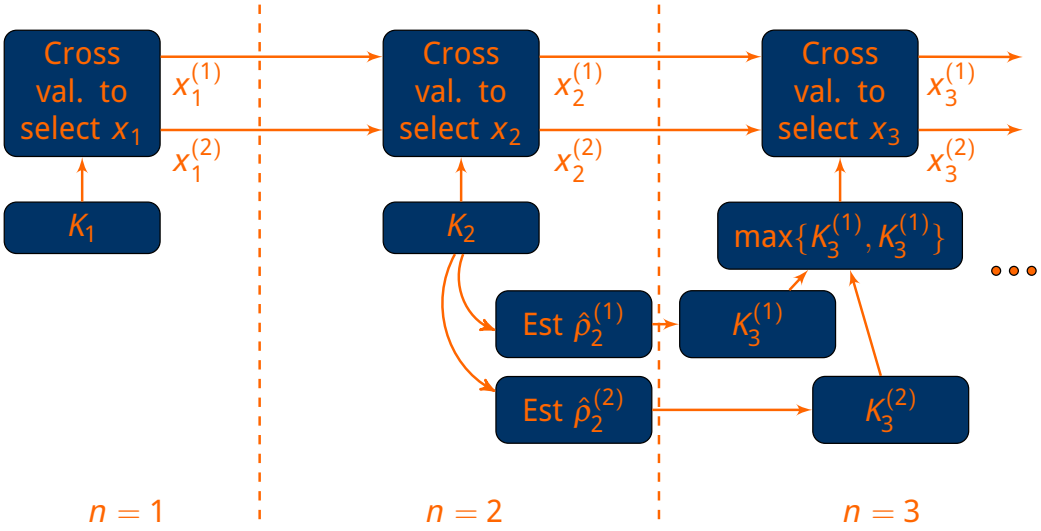


Figure 4.1: Cross validation approach

4.3 Experiments

We focus on two regression applications for synthetic and real data as well as two classification applications for synthetic and real data. For the synthetic regression problem, we can explicitly compute ρ and $\boldsymbol{x}_n^*$ and exactly evaluate the performance of our method. It is straightforward to check that all requirements in A.1-A.6 are satisfied for the problems considered in this section. We apply the do not update past excess risk choice of K_n here.

4.3.1 Synthetic Regression

Consider a regression problem with synthetic data using the penalized quadratic loss

$$\ell(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{2}\left(y - \boldsymbol{w}^\top \boldsymbol{x}\right)^2 + \frac{1}{2}\lambda\|\boldsymbol{x}\|_2^2$$

with $\boldsymbol{z} = (\boldsymbol{w}, y) \in \mathbb{R}^{d+1}$. The distribution of $\boldsymbol{z}_n$ is zero mean Gaussian with covariance matrix

$$\begin{bmatrix} \sigma_{\boldsymbol{w}}^2 \boldsymbol{I} & r_{\boldsymbol{w}_n, y_n} \\ r_{\boldsymbol{w}_n, y_n}^\top & \sigma_{y_n}^2 \end{bmatrix}$$

Under these assumptions, we can analytically compute minimizers $\boldsymbol{x}_n^*$ of $f_n(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_n \sim p_n}[\ell(\boldsymbol{x}, \boldsymbol{z}_n)]$. We change only $r_{\boldsymbol{w}_n, y_n}$ and $\sigma_{y_n}^2$ appropriately to ensure that $\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 = \rho$ holds for all n. We find approximate minimizers using SGD with $\lambda = 0.1$. We estimate ρ using the direct estimate.

We let n range from 1 to 25 with $\rho = 1$, a target excess risk $\varepsilon = 0.1$, and K_n from (3.18). We average over twenty runs of our algorithm. Figure 4.2 shows $\hat{\rho}_n$, our estimate of ρ, which is above ρ in general. Figure 4.3 shows the number of samples K_n, which settles down. We can exactly compute $f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*)$, and so by averaging over the twenty runs of our algorithm, we can estimate the excess risk (denoted "sample average estimate"), shown in Figure 4.4. We average over the time horizon from $n = 1$ to 25 to yield the sample average estimate excess risk given by $2.797 \times 10^{-2} \pm 1.071 \times 10^{-2}$. Therefore, we achieve our desired excess risk.

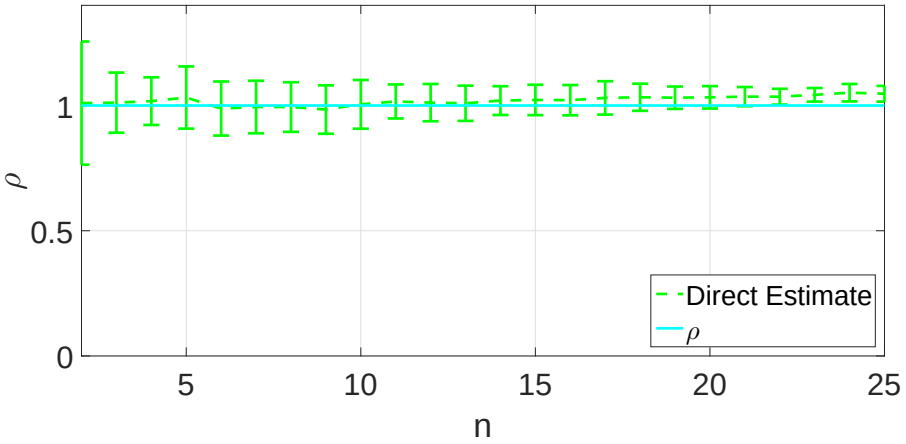


Figure 4.2: ρ Estimate

Cost Approach

We consider applying the cost approach in Section 4.2.1 to the synthetic regression problem with the cost in (4.2). We compare the optimal cost approach introduced in (4.8) of Section 4.2.1 to the approach in (3.18), taking all samples at time $n = 1$ as in (4.4), and taking samples every five time instants as in (4.5). Note that the method from (3.18) does not satisfy the cost budget. Figure 4.5 shows the test loss of these approaches. We achieve similar test loss to the method in (3.18) and better than the other two methods. Figure 4.6 shows the number of samples

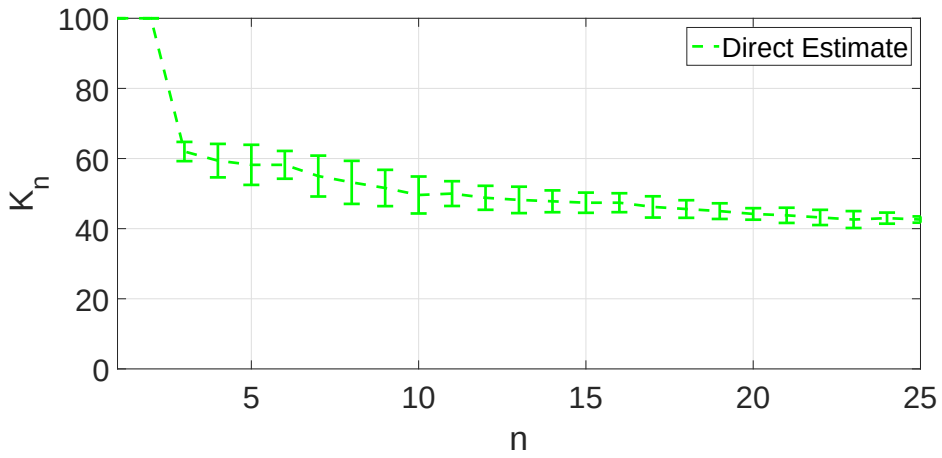


Figure 4.3: K_n

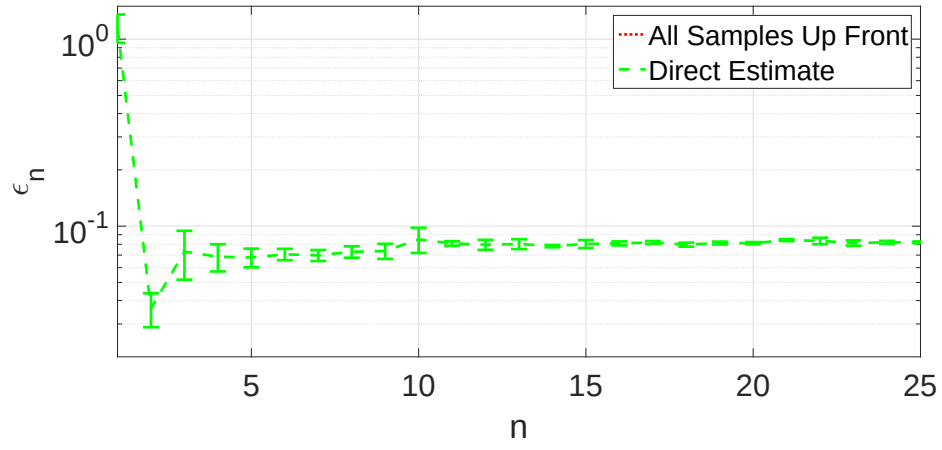


Figure 4.4: Excess risk

selected for both methods. At some time instants, our optimal cost approach does not take samples.

4.3.2 Synthetic Classification

Consider a binary classification problem using

$$\ell(\boldsymbol{x},\boldsymbol{z}) = \frac{1}{2}(1-y(\boldsymbol{w}^\top \boldsymbol{x}))_+^2 + \frac{1}{2}\lambda\|\boldsymbol{x}\|_2^2$$

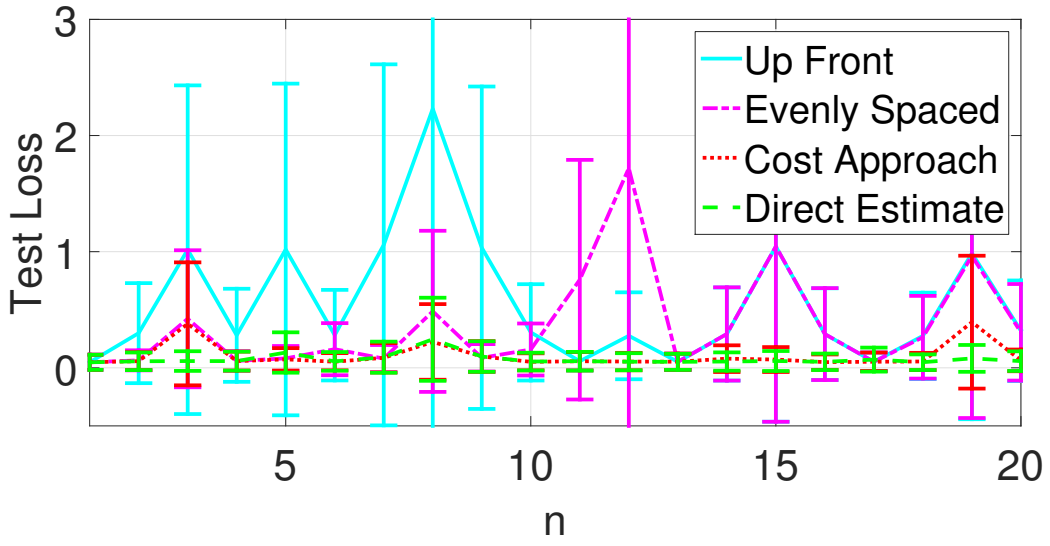


Figure 4.5: Test loss

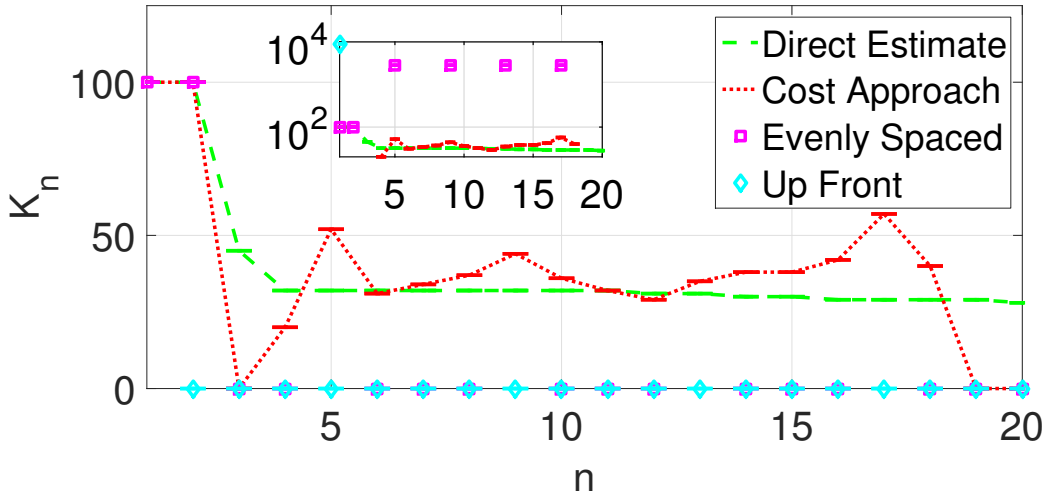


Figure 4.6: Cost choice of K_n

with $\boldsymbol{z} = (\boldsymbol{w}, y) \in \mathbb{R}^d \times \mathbb{R}$ and $(y)_+ = \max\{y, 0\}$. This is a smoothed version of the hinge loss used in support vector machines (SVM) [48]. We suppose that at time n, the two classes have features drawn from a Gaussian distribution with covariance matrix $\sigma^2 \boldsymbol{I}$ but different means $\mu_n^{(1)}$ and $\mu_n^{(2)}$, i.e.,

$$\boldsymbol{w}_n \mid \{y_n = i\} \sim \mathcal{N}(\mu_n^{(i)}, \sigma^2 \boldsymbol{I})$$

The class means move slowly over uniformly spaced points on a unit sphere in $\mathbb{R}^d$ as in Figure 4.7 to ensure that the constant Euclidean norm condition defined in (3.2) holds. We find approximate minimizers using SGD with $\lambda = 0.1$. We estimate ρ using the direct estimate with $t_n \propto 1/n^{3/8}$.

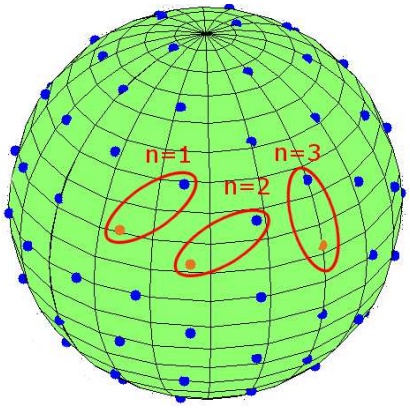


Figure 4.7: Evolution of class means

We let n range from 1 to 25 and target a excess risk $\varepsilon = 0.1$. We average over twenty runs of our algorithm. As a comparison, if our algorithm takes $\{K_n\}_{n=1}^{25}$ samples, then we consider taking $\sum_{n=1}^{25} K_n$ samples up front at $n = 1$. This is what we would do if we assumed that our problem is not time varying. Figure 4.8 shows $\hat{\rho}_n$, our estimate of ρ. Figure 4.9 shows the average test loss for both sampling strategies. To compute test loss we draw T_n additional samples $\{\boldsymbol{z}_n^{\text{test}}(k)\}_{k=1}^{T_n}$ from p_n and compute $\frac{1}{T_n}\sum_{k=1}^{T_n} \ell(\boldsymbol{x}_n, \boldsymbol{z}_n^{\text{test}}(k))$. We see that our approach achieves substantially smaller test loss than taking all samples up front. We do not draw the error bars on this plot as it makes it difficult to see the actual losses achieved.

To further evaluate our approach we look at the receiver operating characteristic (ROC) of our classifiers. The ROC is a plot of the probability of a true positive against the probability of a false positive. The area under the curve (AUC) of the ROC equals the probability that a randomly chosen positive instance ($y = 1$) will be rated higher than a negative instance ($y = -1$) [49]. Thus, a large AUC is desirable. Figure 4.10 plots the AUC of our approach against taking all samples up front. Our sampling approach achieves a substantially larger AUC.

4.3.3 Panel Study on Income Dynamics Income - Regression

The Panel Study of Income Dynamics (PSID) surveyed individuals every year to gather demographic and income data annually from 1974 to 2012 [50]. We want

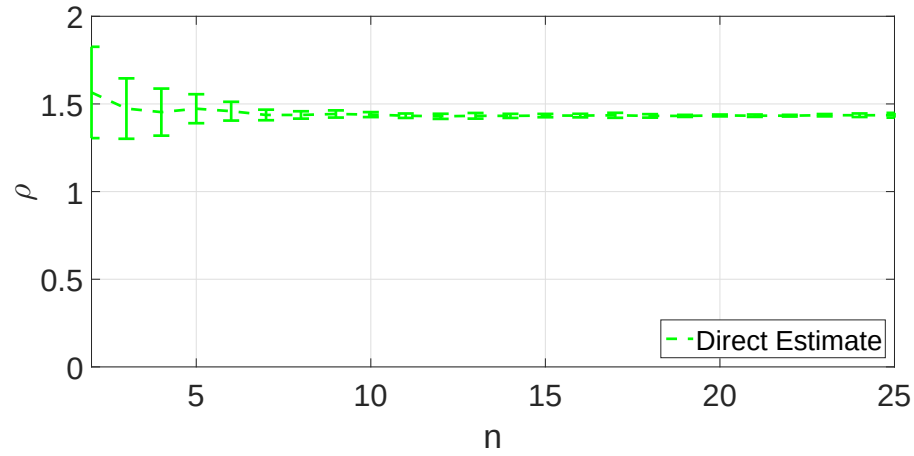


Figure 4.8: ρ Estimate

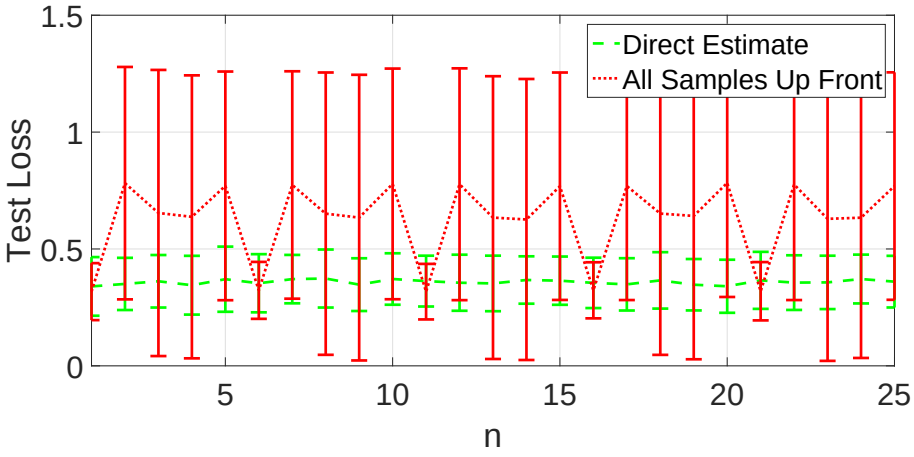


Figure 4.9: Test Loss

to predict an individual's annual income (y) from several demographic features ($\boldsymbol{w}$) including age, education, work experience, etc. chosen based on previous economic studies in [51]. The idea of this problem conceptually is to rerun the survey process and determine how many samples we would need if we wanted to solve this regression problem to within a desired excess risk criterion ε.

We use the same loss function, direct estimate for ρ, and minimization algorithm as the synthetic regression problem. The primary difference is that we set the parameter λ in the loss function by using cross validation as in Section 4.2.2. The values of λ that we consider are 0.1, 0.15, and 0.2. The income is adjusted for inflation to 2012 dollars with mean \$35,176. We only consider individuals with income below \$300,000. Including individuals with much higher income

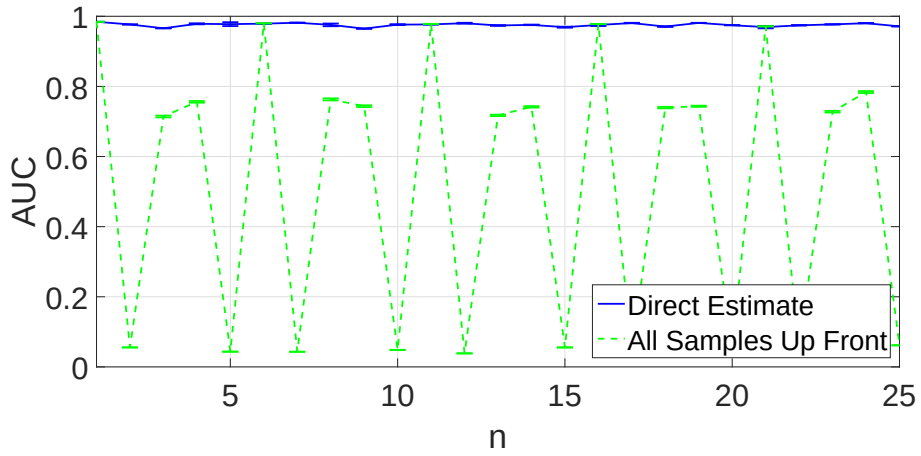


Figure 4.10: Area under the curve

degrades the regression model's performance. We average over twenty runs of our algorithm by resampling [48]. We compare to taking an equivalent number of samples up front.

Figure 4.11 shows the number of samples K_n, which settles down quickly. Figure 4.12 shows $\hat{\rho}_n$, which appears to converge. Figure 4.13 shows the test losses over time evaluated over twenty percent of the available samples. The test loss for our approach is substantially less than taking the same number of samples up front. The square roots of the average test losses over this time period for our approach and all samples up front are $\$2254 \pm 798$ and $\$4194 \pm 425$ respectively in 2012 dollars.

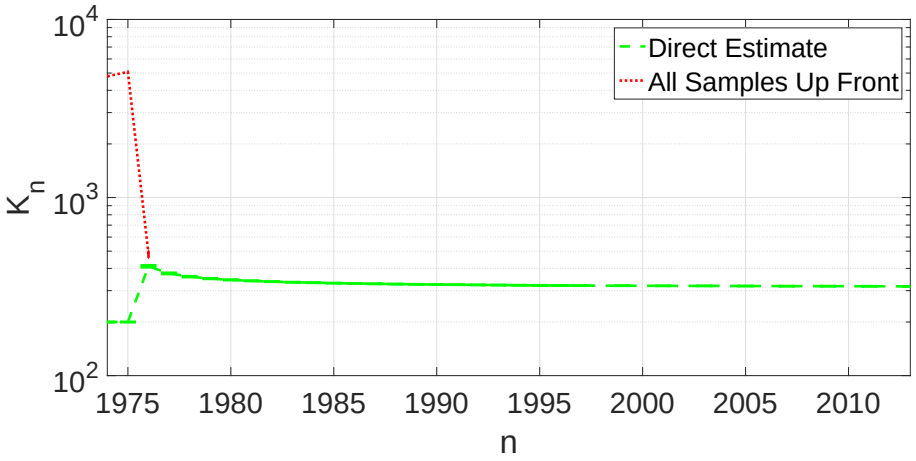


Figure 4.11: K_n

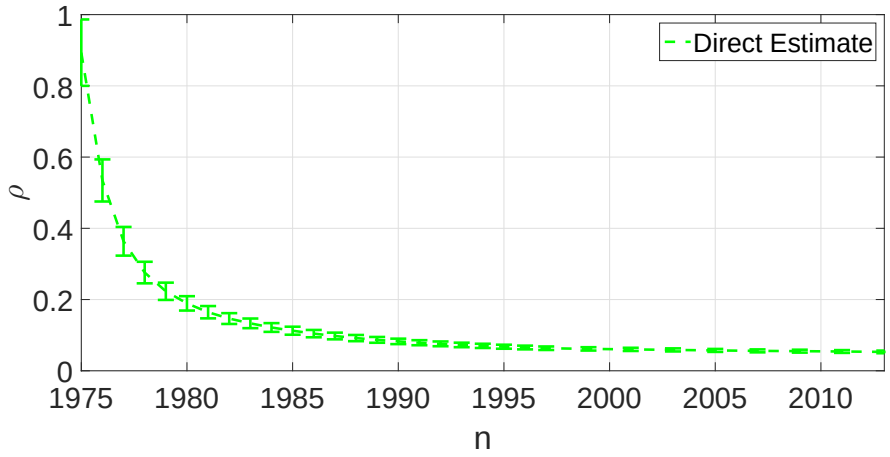


Figure 4.12: ρ Estimate

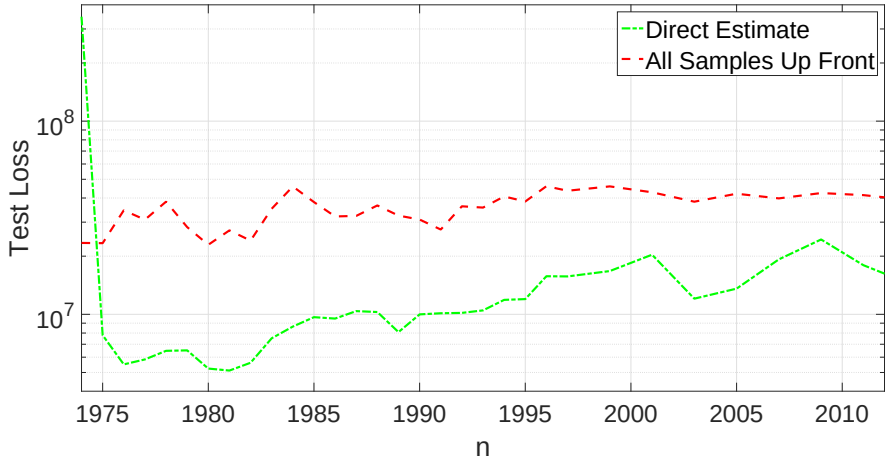


Figure 4.13: Test loss

4.3.4 General Social Survey - Classification

The General Social Survey (GSS) surveyed individuals every year to gather socio-economic data annually from 1981-2013 [52]. We want to predict an individual's marital status (y) from several demographic features ($\boldsymbol{w}$) including age, education, etc. We model this as a binary classification problem using the loss

$$\ell(\boldsymbol{x},\boldsymbol{z}) = -y\boldsymbol{w}^\top\boldsymbol{x} + \log\left(1+e^{\boldsymbol{w}^\top\boldsymbol{x}}\right) + \frac{1}{2}\lambda\|\boldsymbol{x}\|_2^2$$

with $\boldsymbol{z} = (\boldsymbol{w}, y) \in \mathbb{R}^d \times \mathbb{R}$. This loss corresponds to logistic regression with a quadratic penalty [48]. We find approximate minimizers using SGD with $\lambda = 0.1$.

Figure 4.14 shows the estimate of ρ for the GSS data set. Figure 4.15 shows the test loss. We see that our approach achieves smaller test loss than taking all samples up front. We plot the AUC against time in Figure 4.16. Our approach has a larger AUC than the samples up front method especially for the later years.

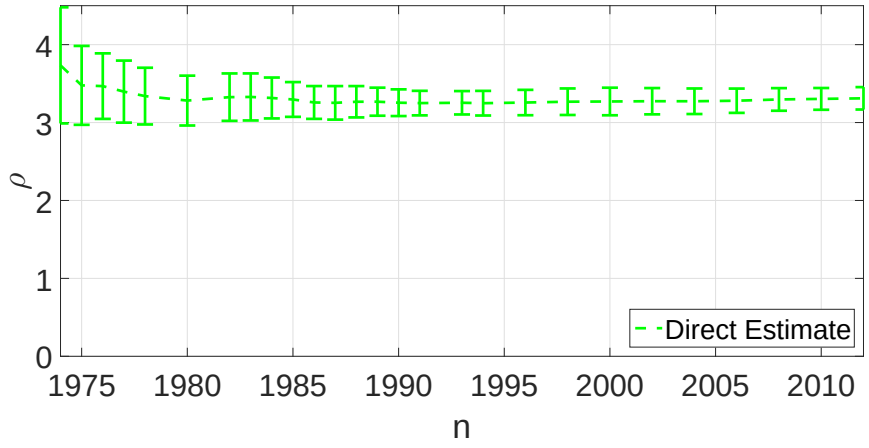


Figure 4.14: ρ Estimate

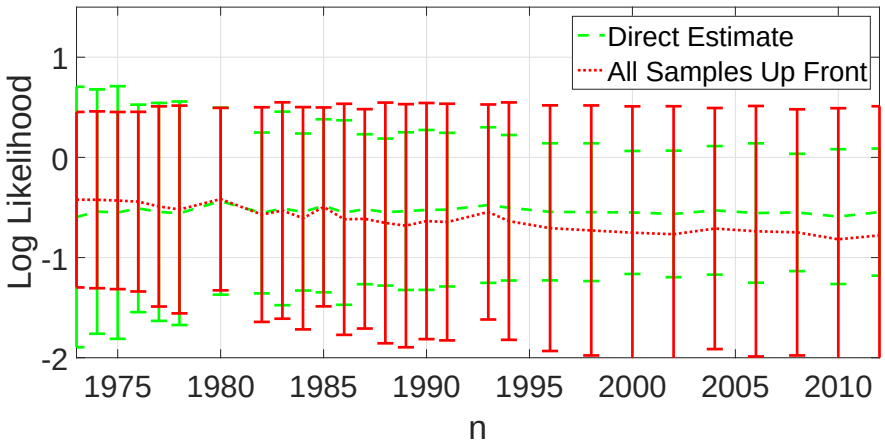


Figure 4.15: Test Loss

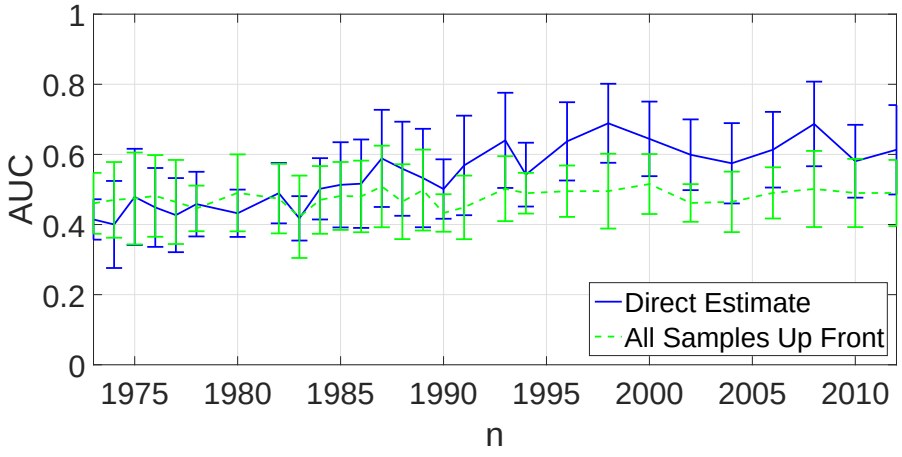


Figure 4.16: Area under the curve

Chapter 5

Abrupt Changes in the Minimizers

In this section, we again solve the sequence of minimization problems of Section 1.2.1, but under a different assumption on how the minimizers change. The minimizers change abruptly in the sense that

$$\rho_n \triangleq \|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2 \in \{\rho^{(1)}, \rho^{(2)}\} \quad \forall n \geq 2 \tag{5.1}$$

where $\rho^{(2)} \gg \rho^{(1)}$. The case with $\rho_n = \rho^{(1)}$ corresponds to small, slow changes and the case with $\rho_n = \rho^{(2)}$ corresponds to large, abrupt changes.

First, we study the case where $\rho^{(1)}$ and $\rho^{(2)}$ are known in order to characterize the impact of abrupt changes. We then introduce alternative K_n selection rules to respond to abrupt changes. Next, we consider the case where $\rho^{(1)}$ and $\rho^{(2)}$ are unknown and develop estimates for $\rho^{(1)}$ and $\rho^{(2)}$ extending the work in Chapter 3. Building on our new estimates for $\rho^{(1)}$ and $\rho^{(2)}$, we develop selection rules for K_n when $\rho^{(1)}$ and $\rho^{(2)}$ are unknown. We do not have theoretical guarantees for the estimates of $\rho^{(1)}$ and $\rho^{(2)}$ and the K_n selection rules that rely on these estimates. We only have rigorous analysis for the case when $\rho^{(1)}$ and $\rho^{(2)}$ are known. Finally, an experiment on synthetic data demonstrates that our methods work. We make the same assumptions A.1-A.6 from Chapter 2.

5.1 Optimization with Changes in Minimizers Known

Consider first the case when the changes $\rho^{(1)}$ and $\rho^{(2)}$ are known. We further assume that the change at time n, ρ_n, is revealed to us at the beginning of time $n+1$. In later sections, we will replace the assumption that ρ_n is revealed to us a

time $n+1$ with an estimate of ρ_n at time $n+1$. Recall the function

$$\phi_{K,\rho}(v) = \alpha(K)\left(\sqrt{\frac{2v}{m}}+\rho\right)^2 + \beta(K) = b\left(\left(\sqrt{\frac{2v}{m}}+\rho\right)^2, K\right)$$

defined in (C.1) of Appendix C. If either K or ρ is implicit, then we drop the subscripts in ϕ. With this notation, it holds that

$$K^* = \min\{K \geq 1 \mid \phi_K(\varepsilon) \leq \varepsilon\}$$

From previous analysis in Lemma 21 of Appendix C, we know that $\phi_{K,\rho}(v)$ has a unique positive fixed point $\bar{v}$ with $0 < \phi'_{K,\rho}(\bar{v}) < 1$. Furthermore, the fixed point iteration

$$v_n = \phi_{K^*,\rho}(v_{n-1})$$

converges to $\bar{v}$ with $\bar{v} \leq \varepsilon$. We consider the following recursion:

$$\varepsilon_n = \phi_{K_n,\rho_n}(\varepsilon_{n-1}) \tag{5.2}$$

where ε_n bounds the mean criterion at time n. It is difficult to quantify the effect of the abrupt jumps $\rho^{(2)}$ from (5.1), so we consider a linear upper bound on (5.2). Since $\phi_K(v)$ is a concave function in v, it holds that

$$\begin{aligned}\phi_K(v) &\leq \phi_K(\bar{v}) + \phi'_K(\bar{v})(v-\bar{v}) \\ &= \bar{v} + \phi'_K(\bar{v})(v-\bar{v}) \\ &= \phi'_K(\bar{v})v + (1-\phi'_K(\bar{v}))\bar{v}\end{aligned}$$

This in turn implies that

$$\varepsilon_n \leq \phi'_{K_n}(\bar{v}_n)\varepsilon_{n-1} + (1-\phi'_{K_n}(\bar{v}_n))\bar{v}_n \tag{5.3}$$

Define

$$g_n = \phi'_{K_n,\rho_n}(\bar{v}_n) \tag{5.4}$$

which is a function of K_n and ρ_n. By Lemma 21, it holds that $0 < g_n < 1$. Then it holds that

$$\varepsilon_n \leq g_n\varepsilon_{n-1} + (1-g_n)\bar{v}_n$$

If K_n is selected to drive $\bar{v}_n$ below ε and $\varepsilon_{n-1} \leq \varepsilon$, then it holds that $\varepsilon_n \leq \varepsilon$. For any n_0, by induction applied to (5.3), it holds that

$$\varepsilon_n \leq \left(\prod_{i=n_0}^{n} g_i\right) \varepsilon_{n_0-1} + \sum_{i=n_0}^{n} \left(\prod_{j=i+1}^{n} g_j\right) (1-g_i)\bar{v}_i \tag{5.5}$$

For the recursion in (5.2), it is difficult to quantify the impact of abrupt changes. For the upper bound in (5.5), due to the linearity in ε_{n_0-1}, it is substantially easier to characterize and quantify the effect of abrupt changes.

To understand some properties of the bound in (5.5) and possible ways to respond to abrupt changes, we provide Lemma 7 that develops useful properties of g and $\bar{v}$ viewed as functions of ρ and K.

Lemma 7. *It holds that*[1]

1. $g = \phi'_{K,\rho}(\bar{v})$ *as a function of* ρ *is non-decreasing in* ρ
2. *If* $\alpha(K) = \beta(K) = \Theta(K^{-\delta})$ *for* $\delta \in (0,1]$, *then* $g \to 0$ *as* $K \to \infty$
3. $(1-g)\bar{v}$ *as a function of* ρ *is non-decreasing in* ρ
4. *If* $\alpha(K) \to 0$ *and* $\beta(K) \to 0$ *as* $K \to \infty$, *then* $(1-g)\bar{v} \to 0$ *as* $K \to \infty$

Proof. First, by solving the quadratic

$$\phi_{K,\rho}(\bar{v}) = \bar{v}$$

we have

$$\sqrt{\bar{v}} = \frac{2\alpha\rho\sqrt{2/m}}{1-2\alpha/m} + \frac{1}{2(1-2\alpha/m)}\sqrt{\frac{8}{m}\rho^2\alpha^2 + 4(1-2\alpha/m)(\alpha\rho^2+\beta)} \tag{5.6}$$

We have dropped the K argument here due to space. It also holds that

$$\phi'(v) = \frac{2\alpha(K)}{m}\left(1 + \frac{\rho}{\sqrt{\frac{2}{m}v}}\right) \tag{5.7}$$

By examining the form of this derivative, we see that g depends on ρ through $\frac{\rho}{\sqrt{\bar{v}}}$,

[1] $f(K) = \Theta(g(K))$ iff $f(K) = \mathcal{O}(g(K))$ and $g(K) = \mathcal{O}(f(K))$

which equals

$$\frac{\rho}{\sqrt{\bar{v}}} = \frac{1}{\frac{2\alpha\sqrt{2/m}}{1-2\alpha/m} + \frac{1}{2(1-2\alpha/m)}\sqrt{\frac{8}{m}\alpha^2 + 4(1-2\alpha/m)(\alpha+\frac{\beta}{\rho^2})}}$$

It is clear that $\phi'(\bar{v})$ is non-decreasing in ρ.

Second, suppose that $\alpha = \beta = \Theta(K^{-\delta})$ for $\delta \in (0,1]$. Then from (5.7) it holds that

$$\phi'(\bar{v}) = \frac{2\alpha(K)}{m} + \frac{2\rho\alpha(K)}{m\sqrt{\frac{2}{m}\bar{v}}}$$

Since $\alpha(K) = \Theta(K^{-\delta})$, the first term $\frac{2\alpha(K)}{m} \to 0$ as $K \to \infty$. By applying the assumption that $\alpha(K) = \beta(K) = \Theta(K^{-\delta})$ with (5.6), we have

$$\frac{2\alpha(K)\rho}{m\sqrt{\bar{v}}} = \Theta\left(K^{-\delta}\right)$$

This implies that $\phi'(\bar{v}) \to 0$ as $K \to \infty$ by (5.7).

For the third claim, we have by (5.7)

$$(1-g)\bar{v} = (1-\phi'(\bar{v}))\bar{v} \tag{5.8}$$

Plugging in (5.6) and simplifying shows that

$$\begin{aligned}
&(1-g)\bar{v} \\
&= \alpha\rho\left(\left(\frac{4\sqrt{2}\sqrt{\alpha^2m^2\rho^2\left(-2\alpha\beta+\alpha m\rho^2+\beta m\right)}}{m(m-2\alpha)^2}\right.\right. \\
&\qquad \left.+\frac{2m\left(\beta(m-2\alpha)+\alpha\rho^2(2\alpha+m)\right)}{m(m-2\alpha)^2}\right)^{1/2} \\
&\qquad \left.+\frac{4\alpha\rho}{m-2\alpha}+\rho\right)+\frac{2\sqrt{2}\sqrt{\alpha^2m^2\rho^2\left(-2\alpha\beta+\alpha m\rho^2+\beta m\right)}}{m^2-2\alpha m}+\beta
\end{aligned}$$

It then follows that $(1-g)\bar{v}$ is non-decreasing in ρ.

For the fourth claim, first note that

$$|(1-g)\bar{v}| \leq \bar{v}$$

Then by examining (5.6) and using the assumption that $\alpha, \beta \to 0$ as $K \to \infty$, it follows that

$$(1-g)\bar{v} \to 0$$

as $K \to \infty$

□

We now apply the tools we have developed to a few different models for abrupt changes.

5.1.1 No Abrupt Changes

Consider the case in which there are no abrupt jumps, which is equivalent to $\rho^{(2)} = \rho^{(1)}$. In this special case, we set $K_n = K^*$, with K^* defined in (3.16), yielding

$$g = g_n = \phi'_{K^*,\rho^{(1)}}(\bar{v})$$

from (5.4) and so it holds that

$$\varepsilon_n \leq g^{n-n_0+1}\varepsilon_{n_0-1} + (1-g^{n-n_0+1})\bar{v}$$

By our choice of K^*, it follows that $\bar{v} \leq \varepsilon$. Provided that $\varepsilon_{n_0-1} \leq \varepsilon$, it holds that $\varepsilon_n \leq \varepsilon$. Therefore, in this simple case, we continue to meet the mean criterion for all n.

5.1.2 Periodic Groups of Jumps

Consider a periodic model in which blocks of jumps occur. A block of ΔJ_1 $\rho^{(1)}$ slow changes occur followed by a block of ΔJ_2 $\rho^{(2)}$ abrupt jumps as follows:

$$\underbrace{\rho^{(1)},\ldots,\rho^{(1)}}_{\Delta J_1 \text{ times}},\underbrace{\rho^{(2)},\ldots,\rho^{(2)}}_{\Delta J_2 \text{ times}},\underbrace{\rho^{(1)},\ldots,\rho^{(1)}}_{\Delta J_1 \text{ times}},\underbrace{\rho^{(2)},\ldots,\rho^{(2)}}_{\Delta J_2 \text{ times}},\ldots \tag{5.9}$$

We start with the bound in (5.5). Suppose that n_0 is the start of a run of ΔJ_2 abrupt jumps associated with $\rho^{(2)}$. To ensure that the mean criterion does not blow up, we need a stability condition of the following form called the *mean criterion cycle condition*:

$$\varepsilon_{n_0+\Delta J_2+\Delta J_1-1} \leq \varepsilon_{n_0-1} \tag{5.10}$$

This ensures that over the course of a cycle of $\rho^{(2)}$ and $\rho^{(1)}$ jumps, the mean criterion returns to at least the level it was at before the cycle began. Using (5.5), this condition can be satisfied if it holds that

$$\begin{aligned}
&\varepsilon_{n_0+\Delta J_2+\Delta J_1-1} \\
&\quad \leq \left(\prod_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1} g_i\right)\varepsilon_{n_0-1} + \sum_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1}\left(\prod_{j=i+1}^{n_0+\Delta J_2+\Delta J_1-1} g_j\right)(1-g_i)\bar{v}_i \\
&\quad \leq \varepsilon_{n_0-1}
\end{aligned}$$

By rearranging, this condition becomes

$$\varepsilon_{n_0-1} \leq \frac{\sum_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1}\left(\prod_{j=i+1}^{n_0+\Delta J_2+\Delta J_1-1} g_j\right)(1-g_i)\bar{v}_i}{1-\prod_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1} g_i}$$

If we want the mean criterion gap to be no greater than ε at the end of a $\rho^{(2)}$ and $\rho^{(1)}$ cycle, then we need

$$\frac{\sum_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1}\left(\prod_{j=i+1}^{n_0+\Delta J_2+\Delta J_1-1} g_j\right)(1-g_i)\bar{v}_i}{1-\prod_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1} g_i} \leq \varepsilon \tag{5.11}$$

Now, suppose that we use the choice of K_n in (3.16) with no modifications. It holds that $g_i \in \{g^{(1)}, g^{(2)}\}$ and $\bar{v}_i \in \{\bar{v}^{(1)}, \bar{v}^{(2)}\}$ where $g^{(i)}$ and $\bar{v}^{(i)}$ are computed under $\rho^{(i)}$ with $K = K^*$. Using this observation, it holds that

$$\prod_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1} g_i = \left(g^{(1)}\right)^{\Delta J_1}\left(g^{(2)}\right)^{\Delta J_2}$$

In addition, it follows that

$$\begin{aligned}
&\sum_{i=n_0}^{n_0+\Delta J_2+\Delta J_1-1} \left(\prod_{j=i+1}^{n_0+\Delta J_2+\Delta J_1-1} g_j\right)(1-g_i)\bar{v}_i \\
&= \sum_{i=n_0}^{n_0+\Delta J_2-1} \left(\prod_{j=i+1}^{n_0+\Delta J_2+\Delta J_1-1} g_j\right)(1-g_i)\bar{v}_i \\
&\qquad + \sum_{i=n_0+\Delta J_2}^{n_0+\Delta J_2+\Delta J_1-1} \left(\prod_{j=i+1}^{n_0+\Delta J_2+\Delta J_1-1} g_j\right)(1-g_i)\bar{v}_i \\
&= \sum_{i=1}^{\Delta J_2} \left(g^{(1)}\right)^{\Delta J_1} \left(g^{(2)}\right)^{\Delta J_2-i}(1-g^{(2)})\bar{v}^{(2)} + \sum_{i=1}^{\Delta J_1} \left(g^{(1)}\right)^{\Delta J_1-i}(1-g^{(1)})\bar{v}^{(1)} \\
&\leq \left(g^{(1)}\right)^{\Delta J_1}\left(1-\left(g^{(2)}\right)^{\Delta J_2}\right)\bar{v}^{(2)} + \left(1-\left(g^{(1)}\right)^{\Delta J_1}\right)\bar{v}^{(1)} \\
&\leq \bar{v}^{(1)} + \left(g^{(1)}\right)^{\Delta J_1}\left(\bar{v}^{(2)}-\bar{v}^{(1)}\right)
\end{aligned}$$

This in turn implies that

$$\varepsilon_n \leq \left(g^{(1)}\right)^{\Delta J_1}\left(g^{(2)}\right)^{\Delta J_2}\varepsilon_{n_0-1} + \left(g^{(1)}\right)^{\Delta J_1}\left(\bar{v}^{(2)}-\bar{v}^{(1)}\right) + \bar{v}^{(1)}$$

By our choice of $K = K^*$, it follows that $\bar{v}^{(1)} \leq \varepsilon$. Therefore, we want

$$\left(g^{(1)}\right)^{\Delta J_1}\left(g^{(2)}\right)^{\Delta J_2}\varepsilon_{n_0-1} + \left(g^{(1)}\right)^{\Delta J_1}\left(\bar{v}^{(2)}-\bar{v}^{(1)}\right) + \varepsilon \leq \varepsilon_{n_0-1}$$

This in turn implies that across an entire cycle we can satisfy the cycle condition (5.10) with mean criterion at the end of a cycle equal to

$$\frac{\varepsilon}{1-\left(g^{(1)}\right)^{\Delta J_1}\left(g^{(2)}\right)^{\Delta J_2}} + \frac{\left(g^{(1)}\right)^{\Delta J_1}\left(\bar{v}^{(2)}-\bar{v}^{(1)}\right)}{1-\left(g^{(1)}\right)^{\Delta J_1}\left(g^{(2)}\right)^{\Delta J_2}} \tag{5.12}$$

From the analysis of Lemma 7, it follows that $\bar{v}^{(2)} \geq \bar{v}^{(1)}$, so the second term in this expression is non-negative. In regards to the cycle condition, we roughly have mean criterion given by

$$\varepsilon + \left(g^{(1)}\right)^{\Delta J_1}\left(\bar{v}^{(2)}-\bar{v}^{(1)}\right)$$

From this expression, it is clear that the primary effect of the abrupt changes is in the term $\bar{v}^{(2)} - \bar{v}^{(1)}$, which is suppressed by the $\left(g^{(1)}\right)^{\Delta J_1}$ term. Finally, as the

time between abrupt jumps ΔJ_1 becomes large, we have

$$\limsup_{\Delta J_1 \to \infty} \left(\frac{\varepsilon}{1 - \left(g^{(1)}\right)^{\Delta J_1} \left(g^{(2)}\right)^{\Delta J_2}} + \frac{\left(g^{(1)}\right)^{\Delta J_1} \left(\bar{v}^{(2)} - \bar{v}^{(1)}\right)}{1 - \left(g^{(1)}\right)^{\Delta J_1} \left(g^{(2)}\right)^{\Delta J_2}} \right) \leq \varepsilon$$

Therefore, for large n, we will roughly meet the mean criterion target ε in the sense of (5.10).

Finally, we look at bounding the mean criterion for n not corresponding to abrupt jumps. Suppose that n satisfies

$$\Delta J_2 < n - n_0 \leq \Delta J_2 + \Delta J_1$$

This condition on n ensures that this time instant occurs between rounds of $\rho^{(2)}$ jumps. Then by similar analysis, it follows that

$$\varepsilon_n \leq \left(g^{(1)}\right)^{n-n_0+1-\Delta J_2} \left(g^{(2)}\right)^{\Delta J_2} \varepsilon_{n_0-1} + \left(g^{(1)}\right)^{n-n_0+1-\Delta J_2} \left(\bar{v}^{(2)} - \bar{v}^{(1)}\right) + \varepsilon \tag{5.13}$$

This provides us with an idea of the effect of a $\rho^{(2)}$ jump. There is a spike in the mean criterion due to the $\rho^{(2)}$ jump captured in the $\bar{v}^{(2)} - \bar{v}^{(1)}$ term. This spike is exponentially suppressed during the $\rho^{(1)}$ phase. This shows that even just using $K_n = K*$ provides some protection against abrupt jumps provided they are sufficiently rare.

5.1.3 New Sample Selection Rules

We consider alternative rules for choosing K_n now to counteract $\rho^{(2)}$ jumps. If we knew whether ρ_n equals $\rho^{(1)}$ or $\rho^{(2)}$ at the beginning of the n^{th} time instant, then we would choose

$$K_n = \begin{cases} K^{(1)}, & \rho_n = \rho^{(1)} \\ K^{(2)}, & \rho_n = \rho^{(2)} \end{cases}$$

with

$$\begin{aligned} K^{(1)} &= \min\left\{K \geq 1 \mid \phi_{K,\rho^{(1)}}(\varepsilon) \leq \varepsilon\right\} \\ K^{(2)} &= \min\left\{K \geq 1 \mid \phi_{K,\rho^{(2)}}(\varepsilon) \leq \varepsilon\right\} \end{aligned}$$

With this choice, it is obvious that the mean criterion is always less than ε for all n.

Update the Past

A more reasonable assumption is that we can determine whether ρ_n equals $\rho^{(1)}$ or $\rho^{(2)}$ at the beginning of the $(n+1)^{\text{th}}$ time instant but not during the n^{th} time instant. The bound ε_{n-1} from (5.2) can be computed at time n. Suppose that we choose K_n by hypothesizing that $\rho_n = \rho_{n-1}$

$$K_n \triangleq \min\left\{ K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon_{n-1}}{m}} + \rho_{n-1}\right)^2, K\right) \leq \varepsilon \right\} \tag{5.14}$$

If it is true that $\rho_n \leq \rho_{n-1}$, then it holds that

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon$$

When $\rho_n > \rho_{n-1}$, we have no guarantees. Thus, in general, we will not satisfy our mean criterion target when $\rho_n > \rho_{n-1}$. However, since our bound ε_{n-1} is always correct at time n, as soon as $\rho_n \leq \rho_{n-1}$, we will again satisfy our mean criterion target ε. For the periodic jump model in Section 5.1.2, the only time we fail to meet the mean criterion is when $\rho_{n-1} = \rho^{(1)}$ and $\rho_n = \rho^{(2)}$.

Do Not Update the Past

Suppose that we compute

$$K^{(i)} = \min\left\{ K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + \rho^{(i)}\right)^2, K\right) \leq \varepsilon \right\} \quad i = 1,2 \tag{5.15}$$

and set

$$K_n \triangleq \begin{cases} K^{(1)}, & \rho_{n-1} = \rho^{(1)} \\ K^{(2)}, & \rho_{n-1} = \rho^{(2)} \end{cases} \tag{5.16}$$

Note that $K^{(1)} = K^*$ with K^* defined in (3.16).

The following analysis characterizes the performance of this rule for the peri-

odic model. By way of comparison, define the rule

$$\tilde{K}_n = \begin{cases} K^{(1)}, & \rho_{n-1} = \rho^{(1)}, \rho_{n-2} = \rho^{(1)} \\ K^{(2)}, & \rho_{n-2} = \rho^{(2)} \\ K^{(2)}, & \rho_{n-1} = \rho^{(2)}, \rho_{n-2} = \rho^{(1)} \end{cases} \tag{5.17}$$

This rule is identical to the rule in (5.16) except $K^{(2)}$ is used when the jump size transitions from $\rho^{(1)}$ to $\rho^{(2)}$. With this rule, the mean criterion is always less than ε provided that $\varepsilon_1 \leq \varepsilon$. Define g_i and $\bar{v}_i$ with respect to the choice of K_n in (5.16) and $\tilde{g}_i$ and $\tilde{v}_i$ with respect to the choice of K_n in (5.17).

With the choice of K_n in (5.17), we have

$$\varepsilon_n \leq \left(\prod_{i=n_0}^{n} \tilde{g}_i \right) \varepsilon_{n_0-1} + \sum_{i=n_0}^{n} \left(\prod_{j=i+1}^{n} \tilde{g}_j \right) (1-\tilde{g}_i)\tilde{v}_i$$

With the choice of K_n in (5.16), we have

$$\begin{aligned} \varepsilon_n &\leq \left(\prod_{i=n_0}^{n} g_i \right) \varepsilon_{n_0-1} + \sum_{i=n_0}^{n} \left(\prod_{j=i+1}^{n} g_j \right) (1-g_i)\bar{v}_i \\ &= \left(\prod_{i=n_0}^{n} \tilde{g}_i \right) \varepsilon_{n_0-1} + \sum_{i=n_0}^{n} \left(\prod_{j=i+1}^{n} \tilde{g}_j \right) (1-\tilde{g}_i)\tilde{v}_i \\ &\quad + \left(\prod_{i=n_0+1}^{n} g_i \right) (g_{n_0} - \tilde{g}_{n_0})\, \varepsilon_{n_0} + \left(\prod_{i=n_0+1}^{n} g_i \right) ((1-g_{n_0})\bar{v}_{n_0} + (1-\tilde{g}_{n_0})\tilde{v}_{n_0}) \end{aligned}$$

This shows that the gap between the choice of K_n in (5.17) and the choice of K_n in (5.16) is bounded by

$$\left(\prod_{i=n_0+1}^{n} g_i \right) (g_{n_0} - \tilde{g}_{n_0})\, \varepsilon_{n_0} + \left(\prod_{i=n_0+1}^{n} g_i \right) ((1-g_{n_0})\bar{v}_{n_0} + (1-\tilde{g}_{n_0})\tilde{v}_{n_0}) \tag{5.18}$$

This bound shows that as ΔJ_1 becomes large, the mean criterion of the rule in (5.16) approaches ε. From Lemma 7, we have

$$g_{n_0} \geq \tilde{g}_{n_0}$$

and

$$(1-g_{n_0})\bar{v}_{n_0} \geq (1-\tilde{g}_{n_0})\tilde{v}_{n_0}$$

This shows that the gap (5.18) is non-negative as expected.

The mean criterion performance of the K_n rule from (5.16) is superior to $K_n = K^*$ rule, since the rule in (5.16) always selects at least $K^{(1)} = K^*$ samples. In comparison to the rule in (5.14), we cannot guarantee that the mean criterion is met at all time instants except for the case when $\rho_n > \rho_{n-1}$.

5.2 Estimating Changes in Minimizers

We now examine the case when we do not know $\rho^{(1)}$ and $\rho^{(2)}$ and must instead estimate them. We develop methods to combine the one-step estimates from Chapter 3 to yield estimates for $\rho^{(1)}$ and $\rho^{(2)}$.

Suppose that we have one-step estimates $\tilde{\rho}_i$ of $\rho_i = \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2$. First, we consider combining the one-step estimates $\tilde{\rho}_i$ to estimate $\rho^{(1)}$ and $\rho^{(2)}$ as follows:

$$\begin{aligned} \hat{\rho}_n^{(1)} &= \frac{1}{n-1} \sum_{i=2}^{n} \tilde{\rho}_i \\ \hat{\rho}_n^{(2)} &= \frac{1}{n-W} \sum_{i=W+1}^{n} \hat{h}_{\min\{W,i-1\}}(\tilde{\rho}_i, \tilde{\rho}_{i-1}, \ldots, \tilde{\rho}_{\max\{i-W+1,2\}}) \end{aligned} \tag{5.19}$$

These two methods are precisely the ones developed in Chapter 3 defined in (3.5) and (3.11). Under the conditions of Theorems 1 and 3 of Chapter 3, it follows that for appropriate sequences t_n and n large enough

1. $\hat{\rho}_n^{(1)} + t_n \geq \rho^{(1)}$
2. $\hat{\rho}_n^{(2)} + t_n \geq \rho^{(2)}$

almost surely. Therefore, with these two estimates, we have upper bounds on $\rho^{(1)}$ and $\rho^{(2)}$.

Although both of these estimates work, there is a slight complication with the estimation of $\rho^{(1)}$. Suppose that the fraction of time that $\rho^{(2)}$ occurs is $\lambda \in (0,1)$. Then by examining the analysis of $\hat{\rho}_n^{(1)}$ from Chapter 3, it holds that

$$\hat{\rho}_n^{(1)} + t_n \geq \rho^{(1)} + \lambda(\rho^{(2)} - \rho^{(1)}) \tag{5.20}$$

almost surely. If the product $\lambda(\rho^{(2)} - \rho^{(1)})$ is large, then our estimate of $\rho^{(1)}$ can be far from the true value. When our estimate of $\rho^{(1)}$ is substantially above $\rho^{(1)}$, we take substantially more samples than are necessary to meet our mean criterion

target when $\rho_n = \rho^{(1)}$. In the following sections, we examine alternative ways to estimate $\rho^{(1)}$ and $\rho^{(2)}$ to avoid this issue.

5.2.1 Trimmed Mean Estimate

To combat this issue, we compute an estimate of $\rho^{(1)}$ while trying to remove those one-step estimates $\tilde{\rho}_n$ that correspond to $\rho_n = \rho^{(2)}$ through an α-trimmed mean estimate with $\alpha \in (0, \frac{1}{2})$. We remove $100\alpha\%$ of the samples from each side and compute the mean. This estimate is defined as

$$\hat{\rho}_n^{(1)} = \frac{1}{(n-1)(1-2\alpha)} \sum_{i=2+(n-1)\alpha}^{1+(n-1)(1-\alpha)} \tilde{\rho}_{(i)} \tag{5.21}$$

with order statistics $\tilde{\rho}_{(2)} \leq \cdots \leq \tilde{\rho}_{(n)}$. The i^{th}-order statistic $\tilde{\rho}_{(i)}$ is the $(i-1)^{\text{th}}$ smallest value in the set $\{\tilde{\rho}_2, \ldots, \tilde{\rho}_n\}$. The following lemma from [53] describes how far an order statistic can be from the sample mean.

Lemma 8. *For any collection of random variables $X_1, \ldots, X_n$, it holds that*

$$X_{(i)} \geq A_n - B_n \max\left\{ \frac{(n-1)(i-1)}{n(n-i+1)}, \frac{(n-1)(n-i)}{ni} \right\}$$

with

$$A_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$B_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - A_n)^2}$$

Applying Lemma 8 and (5.20), for all n large enough, it follows that

$$\hat{\rho}_n^{(1)} + t_n \geq \rho^{(1)} + \lambda(\rho^{(2)} - \rho^{(1)}) - \mathbb{E}[B_n] \frac{1}{(n-1)(1-2\alpha)} \sum_{i=1}^{(n-1)(1-2\alpha)} h(i, n-1)$$

almost surely where

$$h(i,n) \triangleq \max\left\{ \frac{(n-1)(i-1)}{n(n-i+1)}, \frac{(n-1)(n-i)}{ni} \right\}$$

This provides a little control on the possible overshoot when estimating $\rho^{(1)}$. However, we have no proof that this estimate upper bounds $\rho^{(1)}$ for n large enough.

All we can say is that if we select α such that

$$\lambda(\rho^{(2)} - \rho^{(1)}) - \mathbb{E}[B_n]\frac{1}{(n-1)(1-2\alpha)}\sum_{i=1}^{(n-1)(1-2\alpha)} h(i, n-1) \geq 0 \qquad (5.22)$$

then it does indeed hold that for all n large enough

$$\hat{\rho}_n^{(1)} + t_n \geq \rho^{(1)}$$

almost surely. Unfortunately, we have no way to select α such that (5.22) holds.

Selecting α for the Trimmed Mean Approach

To apply a trimmed mean filter, we must select a fraction α to trim. In this section, we look at data dependent methods to select α known as α-adaptive trimmed mean filters. There are no strong theoretical guarantees on the correctness of α-adaptive trimmed mean filters, but they do work well in practice.

We examine two methods proposed in [54]. We first introduce some basic properties of trimmed means. Suppose that we have $n-1$ one-step estimates $\tilde{\rho}_2, \ldots, \tilde{\rho}_n \overset{\text{iid}}{\sim} F$. Our estimates are not in fact iid, but far apart $\tilde{\rho}_i$ are only weakly dependent, so this assumption is not too far from the truth. Define the exact trimmed mean

$$a(\alpha) \triangleq \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} \rho dF(\rho)$$

and the approximate trimmed means

$$a_n(\alpha) \triangleq \frac{1}{n-2[\alpha n]}\sum_{j=[\alpha n]+1}^{n-[\alpha n]} \tilde{\rho}_{(j)}$$

The following central limit theorem (CLT) type result from [55] relating $a_n(\alpha)$ to $a(\alpha)$ holds:

$$\sqrt{n}(a_n(\alpha) - a(\alpha)) \xrightarrow{D} \mathcal{N}(0, V(\alpha))$$

with variance

$$V(\alpha) = \frac{1}{(1-2\alpha)^2}\left(\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (\rho - a(\alpha))^2 d\rho \right. \\ \left. + \alpha(F^{-1}(\alpha) - M(\alpha))^2 + \alpha(F^{-1}(1-\alpha) - M(\alpha))^2\right) \tag{5.23}$$

The first method, known as Jaeckel's estimate, is based on this CLT result and chooses α to minimize the variance $V(\alpha)$. Since we cannot compute $V(\alpha)$ in a closed form, we instead compute a trimmed estimate of the variance as a proxy

$$V_n(\alpha) = \frac{1}{(1-2\alpha)^2}\left(\frac{1}{n}\sum_{j=[\alpha n]+1}^{n-[\alpha n]} (\tilde{\rho}_{(j)} - a_n(\alpha))^2 \right. \\ \left. + \alpha(\tilde{\rho}_{[\alpha n]-1} - a_n(\alpha))^2 + \alpha(\tilde{\rho}_{n-[\alpha n]} - a_n(\alpha))^2\right)$$

and choose

$$\alpha^* = \arg\min_{\alpha \in [0, \frac{1}{2}]} V_n(\alpha)$$

The second method, referred to as Otën's estimate, is motivated by noting that the reason for using a trimmed mean is to remove atypical outliers that bias the estimate. We want to find the part of the distribution F that is roughly symmetric. This removes outliers since they introduce a large degree of skew in the distribution. If the distribution F^{-1} is symmetric between $[\alpha, 1-\alpha]$, then $a(\alpha)$ is the median. Thus, it holds that

$$\frac{(F^{-1}(\alpha) - a(\alpha)) + (F^{-1}(1-\alpha) - a(\alpha))}{1-2\alpha} = 0$$

By a simple computation, it is easy to see that

$$\frac{da}{d\alpha} = \frac{(F^{-1}(\alpha) - a(\alpha)) + (F^{-1}(1-\alpha) - a(\alpha))}{1-2\alpha}$$

The condition for symmetry is thus

$$\frac{da}{d\alpha} = 0$$

We want to pick a value for α such that $\frac{da}{d\alpha}$ is small. Since $\frac{da}{d\alpha}(\alpha)$ cannot be computed exactly, we need to estimate this quantity. Given a parameter $T > 0$ that

controls the desired degree of symmetry, we select

$$\alpha^* = \min\left\{ \alpha \;\middle|\; \frac{1}{1/n}|a_n(\alpha) - a_n(\alpha - 1/n)| \le T, \alpha n \in \mathbb{Z}, \frac{1}{n} \le \alpha < \frac{1}{2} \right\}$$

As a result of this choice of α, it holds that

$$\frac{1}{1/n}|a_n(\alpha^*) - a_n(\alpha^* - 1/n)| \approx \frac{da}{d\alpha}(\alpha^*)$$

5.2.2 Heuristic Methods for Estimating the Change in Minimizers

We apply the expectation-maximization (EM) algorithm and k-means clustering to estimate $\rho^{(1)}$ and $\rho^{(2)}$. We have no theoretical guarantees for these techniques, but they work well in practice.

K-Means Clustering Approach

To combine the one-step estimates, we apply the k-means++ algorithm [56], a variant of the k-means algorithm with good choices for the initial clusters, to determine $\rho^{(1)}$ and $\rho^{(2)}$. The k-means algorithm clusters the one-step estimates $\tilde{\rho}_2, \ldots, \tilde{\rho}_n$ into two clusters with centroids $\hat{\rho}^{(1)}$ and $\hat{\rho}^{(2)}$. The k-means algorithm works by alternatively assigning the one-step estimates to the closest current cluster centroid $\hat{\rho}^{(i)}$ and averaging the one-step estimates currently assigned to a cluster to produce a new centroid $\hat{\rho}^{(i)}$. Finally, we decide that $\rho_i = \rho^{(j)}$ if $\tilde{\rho}_i$ is assigned to the cluster with centroid $\hat{\rho}^{(j)}$.

Expectation-Maximization Approach

We apply the expectation-maximization algorithm to estimate $\rho^{(1)}$ and $\rho^{(2)}$ and the value of ρ_n defined in (5.1) under a Gaussian mixture model (GMM). We model the abrupt changes using a Gaussian mixture model where ρ_n is a $\{\rho^{(1)}, \rho^{(2)}\}$-valued random variable with distribution $\boldsymbol{\alpha} = (\alpha^{(1)}, 1 - \alpha^{(1)})$. For notational convenience, we also define $\alpha^{(2)} = 1 - \alpha^{(1)}$. For the GMM model, we assume that our estimates of ρ_i satisfy

$$\tilde{\rho}_i = \rho_i + e_i$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$. This will not be true in practice and is only an approximate model. For a fixed value of ρ, the pdf of the observation is given by

$$p(x, \rho) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\rho)^2}{2\sigma^2} \right\}$$

For a GMM, the pdf of $\tilde{\rho}_i$ is given by

$$\alpha^{(1)} p(x, \rho^{(1)}) + (1 - \alpha^{(1)}) p(x, \rho^{(2)})$$

We apply the EM algorithm [57] to estimate $\rho^{(1)}$ and α. For this particular GMM model, the expectation step is equivalent to computing the following quantities:

$$w_{ik} = \frac{\alpha^{(k)} p(\tilde{\rho}_i, \hat{\rho}^{(k)})}{\alpha^{(1)} p(\tilde{\rho}_i, \hat{\rho}^{(1)}) + (1 - \alpha^{(1)}) p(\tilde{\rho}_i, \hat{\rho}^{(2)})}$$

The quantity w_{ik} is interpreted as the *responsibility* of the k^{th} mixture component for the i^{th} observation. The maximization step is comprised of the following updates:

$$\begin{aligned}
\hat{N}_n^j &= \sum_{i=2}^{n} w_{ij} \\
\hat{\alpha}_n^{(j)} &= \frac{\hat{N}_n^j}{n-1} \\
\hat{\rho}_n^{(j)} &= \frac{1}{\hat{N}_n^j} \sum_{i=2}^{n} w_{ij} \tilde{\rho}_i \\
\hat{\sigma}_n^2 &= \frac{1}{2} \sum_{j=1}^{2} \frac{1}{\hat{N}_n^j} \sum_{i=2}^{n} w_{ij} (\tilde{\rho}_i - \hat{\rho}_n^{(j)})^2
\end{aligned}$$

To decide whether a $\rho^{(2)}$ change occurred at time n, we check whether it holds that

$$(1 - \hat{\alpha}_n^{(1)}) p(\tilde{\rho}_n, \hat{\rho}_n^{(2)}) > \hat{\alpha}_n^{(1)} p(\tilde{\rho}_n, \hat{\rho}_n^{(1)})$$

This is a maximum likelihood estimate of which change occurred at time n. The EM algorithm provides an effective approach to estimate $\rho^{(i)}$ for $i = 1, 2$ and to determine which change occurred. In practice, this method seems to work well and provides accurate estimate of $\rho^{(i)}$ for $i = 1, 2$.

5.3 Optimization with Changes Unknown

We now consider the problem of selecting K_n when $\rho^{(1)}$ and $\rho^{(2)}$ are unknown and must be estimated. We introduce alternative rules to select K_n under these conditions by modifying the selection rules in (5.14) and (5.16). We consider the estimates in (5.19), the trimmed mean, k-means, and EM approaches to estimate $\rho^{(1)}$ and $\rho^{(2)}$.

5.3.1 Update the Past

First, we first consider an analog of the rule in (5.14). We estimate the mean criterion achieved at times $1,\ldots,n$ using

$$\hat{\varepsilon}_{i,n} = b\left(\left(\sqrt{\frac{2\hat{\varepsilon}_{i-1,n}}{m}} + \rho^{(j_{i,n})}\right)^2, K_i\right) \quad i = 2,\ldots,n$$

with

$$j_{i,n} = \begin{cases} 1, & \tilde{\rho}_i \geq \hat{\rho}_n^{(2)} - t_{TH} \\ 2, & \text{else} \end{cases}$$

In this recursion, we use the estimates of $\rho^{(1)}$ and $\rho^{(2)}$ at time n and decide the value of ρ_i by checking if $\tilde{\rho}_i$ is sufficiently close to $\hat{\rho}_n^{(2)}$. Then we set

$$K_n = \min\left\{K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\hat{\varepsilon}_{n-1,n-1}}{m}} + \hat{\rho}_{n-1}^{(j_{n-1,n-1})}\right)^2, K\right) \leq \varepsilon\right\} \tag{5.24}$$

We cannot provide any theoretical guarantees for this method, but it behaves similarly to the rule in (5.14).

5.3.2 Do Not Update the Past

We now consider an analog of the rule in (5.16). We compute

$$K_n^{(i)} \triangleq \min\left\{K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + \hat{\rho}_{n-1}^{(i)} + t_{n-1}\right)^2, K\right) \leq \varepsilon\right\} \tag{5.25}$$

and set

$$K_n = \begin{cases} K_n^{(1)}, & \tilde{\rho}_{n-1} \leq \hat{\rho}_{n-1}^{(1)} + t_{TH} \\ K_n^{(2)}, & \tilde{\rho}_{n-1} > \hat{\rho}_{n-1}^{(1)} + t_{TH} \end{cases} \tag{5.26}$$

We cannot provide any theoretical guarantees for this method, but it behaves similarly to the rule in (5.16).

5.4 Experiment

We present a synthetic example to demonstrate that our sample selection and ρ estimation rules work as expected. We use the example from Section 4.3.1 of Chapter 4 with the change model in (5.1). We use the periodic jump model with $\Delta J_1 = 10$, $\Delta J_2 = 4$, $\rho^{(1)} = 1$, and $\rho^{(2)} = 8$. We average over twenty runs. We test all ρ estimation methods. To select K_n, we compare the selection rule in (5.16) with $\rho^{(1)}$ and $\rho^{(2)}$ known against the rules in (5.24) and (5.26) with $\rho^{(1)}$ and $\rho^{(2)}$ unknown. We do not plot the K_n selection rule in (5.14), since its performance is nearly identical to that in (5.16).

Figure 5.1 and 5.2 show the various estimates of $\rho^{(1)}$ and $\rho^{(2)}$ respectively. All of the estimates upper bound their respective ρ values of $\rho^{(1)} = 1$ and $\rho^{(2)} = 8$ respectively. The k-means and EM methods produce estimates closest to the true values of ρ. The trimmed mean methods produce looser estimates with Jaeckel's method for selecting α the tightest of the trimmed mean methods. Simply averaging the one-step estimates to estimate $\rho^{(1)}$ produces the loosest estimate.

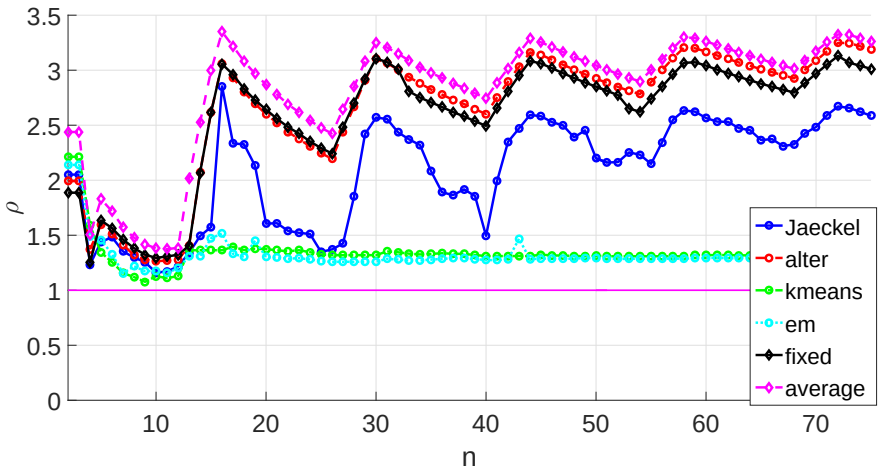


Figure 5.1: Abrupt $\rho^{(1)}$ estimates

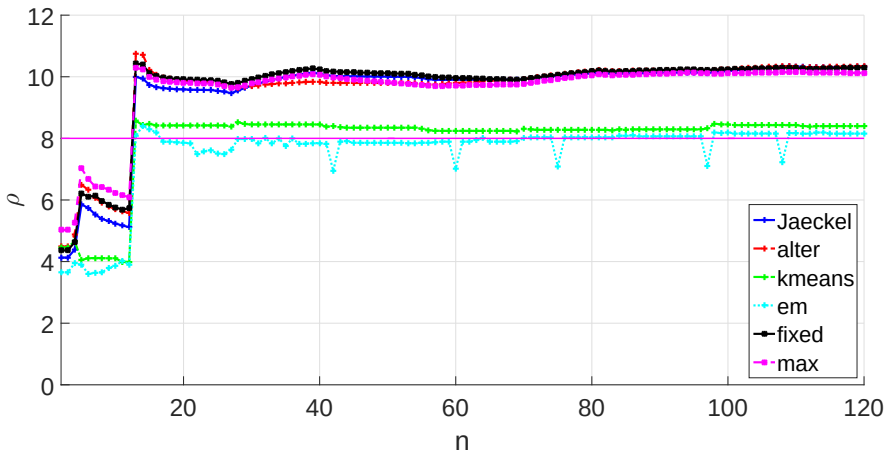


Figure 5.2: Abrupt $\rho^{(2)}$ estimates

Figure 5.3 shows the resulting choice of K_n using the ρ for a variety of choices of K_n averaged over a cycle of $\rho^{(1)}$ and $\rho^{(2)}$ changes. The ρ known method implements the method in (5.14). The ρ unknown methods implement the update past rule from (5.24) and the current rule from (5.26). The only $\rho^{(1)}$ method uses the selection rule from (3.18) with

$$\hat{\rho}_n = \frac{1}{n-1}\sum_{i=2}^{n}\tilde{\rho}_i \tag{5.27}$$

Finally, the exact $\rho^{(1)}$ method uses the rule in (3.16) using $\rho = \rho^{(1)}$ with $\rho^{(1)}$ known. These two choice of K_n corresponds to applying the methods of Chapters 2 and 3 ignoring the change model in (5.1).

The exact $\rho^{(1)}$ method uses the fewest samples. The only $\rho^{(1)}$ method uses substantially more samples due to the mixing of estimates of $\rho^{(1)}$ and $\rho^{(2)}$ in $\hat{\rho}_n$. All of the abrupt change methods select a similar number of samples. These methods select a small number of samples when $\rho_n = \rho^{(1)}$ and takes more when $\rho_n = \rho^{(2)}$. Table 5.1 shows the average number of samples selected for each method. We see that the new abrupt selection rules take slightly fewer samples than the only $\rho^{(1)}$ approach using the methods of Chapter 3 naively.

Figure 5.4 compares the test losses of the various K_n selection methods using the EM ρ estimation method. During the period of $\rho^{(1)}$ small changes all of the K_n choices offer similar performance. Since the only $\rho^{(1)}$ and abrupt change methods with ρ unknown take more samples during this period they offer slightly better performance, but there is not a large gap in test loss. During a period of

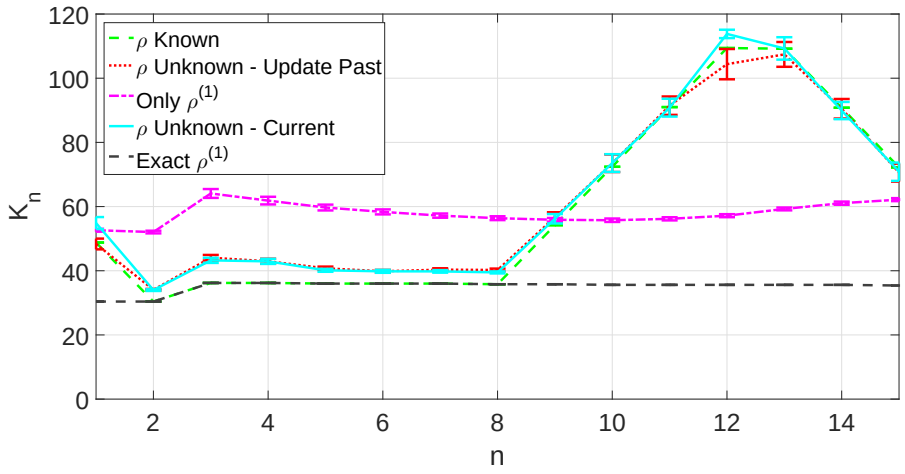


Figure 5.3: Abrupt K_n

Table 5.1: Average K_n

Method	K_n
Exact $\rho^{(1)}$	35
Only $\rho^{(1)}$	64 ± 3
ρ Known	60 ± 3
ρ Unknown - Update	61 ± 2
ρ Unknown - Current	60 ± 3

$\rho^{(2)}$ jumps, the exact $\rho^{(1)}$ method offers the worst performance with a substantial increase in test loss. The only $\rho^{(1)}$ method is less sensitive to the first $\rho^{(2)}$ jump, since it take more samples in the $\rho^{(1)}$ regime. However, for subsequent $\rho^{(2)}$ jumps the test loss remains higher than desired. All of the abrupt changes show the ability to respond more aggressively to $\rho^{(2)}$ changes and decrease the test loss after the first jump.

In summary, the abrupt K_n selection rules offer better test loss performance without taking more samples in light of Table 5.1. In fact, the abrupt change rules use slightly fewer samples than the only $\rho^{(1)}$ approach, which uses the tools of Chapter 3 naively. Both of the abrupt K_n selection rules with ρ unknown offer similar performance to the case when ρ is known.

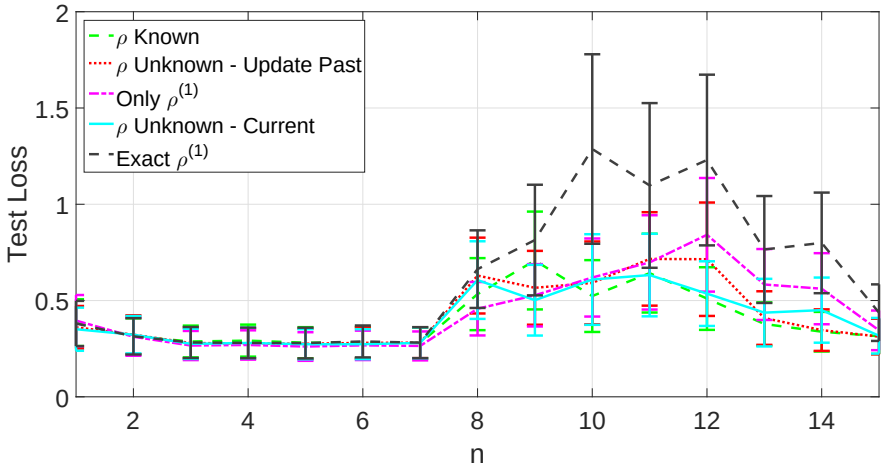


Figure 5.4: Abrupt test losses

Chapter 6

Slowly Changing Dynamic Games

In this chapter, we consider the game problem introduced in Section 1.2.2. We examine the problem of finding solutions to a sequence of repeated games in which the stage game slowly varies over time. We find solutions to each stage game by finding an approximate saddle point for each stage game. We develop gradient based learning dynamics to find an approximate Nash equilibrium of each stage game. We apply our framework to a simple quadratic zero-sum game and an estimation problem in a wireless sensor network.

6.1 Related Work

The problem of learning in games with continuous state spaces has been studied in several settings. The work in [58] studied the problem of learning in a game in which each player has a utility concave in its action. Gradient based strategies were developed to asymptotically learn a Nash equilibrium for the deterministic case. Learning under different concavity assumptions on the utilities, which allows for more general, not gradient based, learning techniques, has been studied [59]. For this class of games, there are some guarantees on recovering an approximate Nash equilibrium. Prior work on finding saddle points such as the Arrow-Hurwicz-Uzawa algorithm [60] can be viewed as finding the Nash equilibrium of a zero sum game.

There is also a rich body of work on learning in games with finite strategy spaces. We give a brief overview of prior work summarized in [1]. The simplest learning dynamic is *fictitious play* in which a player assumes that its opponents play stationary strategies. The player can treat the empirical distribution of the opponent's plays as the opponent's true (mixed) strategy. For an oblivious opponent, an opponent that does not react to its opponent's play, fictitious play will converge to a Nash equilibrium. However, for non-oblivious opponents, there are simple

counterexamples such as Shapley's game [1] in which the play does not converge to a Nash equilibrium. There has been some work on repeated matrix games in which the players employ gradient based learning approaches that converge to Nash equilibria in [61].

In the *replicator dynamic* approach to learning, the share of players using a given strategy increases if the strategy produces good results relative to other strategies in play. A thorough overview of learning techniques is provided in [1]. The authors of [62, 63] consider the problem of a finite zero sum game in which the two players may employ different learning strategies due to differences in rationality and information. For this model, when both players use gradient based approaches to update their strategies, it is possible to show that the strategies of the players converge to a Nash equilibrium. This model has applications to security problems in which the security system and the attacker have differing levels of knowledge. Finally, there has been some work on the case in which the stage games of a repeated game change focused on 2×2 matrix games using an extension of fictitious play [64].

Our work is novel and focuses on slowly changing games with continuous strategy spaces in which we try to transfer knowledge between the changing games. Most prior work has focused only on finding solutions for repeated games with a fixed stage game.

6.2 Zero Sum Games

Consider a sequence of two player zero-sum games modeled as a sequence of saddle point problems. For a saddle point problem we have a function $\mathscr{L}^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n)$ with $\boldsymbol{x}_1^n \in \mathscr{X}_1$ the action of player one at time n and $\boldsymbol{x}_2^n \in \mathscr{X}_2$ the action of player two at time n. We assume that this function is of the form

$$\mathscr{L}^n(\boldsymbol{x}) \triangleq \mathbb{E}\left[\ell^n(\boldsymbol{x})\right] \tag{6.1}$$

where $\ell^n(\boldsymbol{x})$ is a random function. We seek a saddle point $(\bar{\boldsymbol{x}}_1^n, \bar{\boldsymbol{x}}_2^n)$ such that

$$\mathscr{L}^n(\bar{\boldsymbol{x}}_1^n, \boldsymbol{x}_2) \leq \mathscr{L}^n(\bar{\boldsymbol{x}}_1^n, \bar{\boldsymbol{x}}_2^n) \leq \mathscr{L}^n(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2^n) \quad \forall \boldsymbol{x}_1 \in \mathscr{X}_1, \forall \boldsymbol{x}_2 \in \mathscr{X}_2$$

A saddle point corresponds to a pure strategy Nash equilibrium (PNE) of this game, since player one cannot unilaterally decrease $\mathscr{L}^n(\bar{\boldsymbol{x}}_1^n, \bar{\boldsymbol{x}}_2^n)$, and player two

cannot unilaterally increase $\mathscr{L}^n(\bar{\boldsymbol{x}}_1^n, \bar{\boldsymbol{x}}_2^n)$.

We look to find a PNE by iterative, gradient based methods. Since we generally cannot exactly recover a PNE, we need to consider an approximate PNE. Therefore, our goal is to find a pure strategy ε-Nash equilibrium (ε-PNE) $(\tilde{\boldsymbol{x}}_1^n, \tilde{\boldsymbol{x}}_2^n)$. An ε-PNE $(\tilde{\boldsymbol{x}}_1^n, \tilde{\boldsymbol{x}}_2^n)$ is defined by the following two conditions:

$$\begin{aligned} \mathscr{L}^n(\tilde{\boldsymbol{x}}_1^n, \tilde{\boldsymbol{x}}_2^n) - \mathscr{L}^n(\boldsymbol{x}_1, \tilde{\boldsymbol{x}}_2^n) \leq \varepsilon \qquad & \forall \boldsymbol{x}_1 \in \mathscr{X}_1 \\ \mathscr{L}^n(\tilde{\boldsymbol{x}}_1^n, \boldsymbol{x}_2) - \mathscr{L}^n(\tilde{\boldsymbol{x}}_1^n, \tilde{\boldsymbol{x}}_2^n) \leq \varepsilon \qquad & \forall \boldsymbol{x}_2 \in \mathscr{X}_2 \end{aligned} \tag{6.2}$$

This ensures that no player can improve its utility by more than ε by deviating from the ε-PNE.

We capture the idea of slow change in the stage games by assuming that the PNE of subsequent stage games are close in the sense that

$$\|\bar{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^{n-1}\|_2 \leq \rho \tag{6.3}$$

Under assumptions that we will later impose on $\mathscr{L}^n(\boldsymbol{x})$, it holds that the Nash equilibrium $\bar{\boldsymbol{x}}^n$ is unique. Therefore, the slow change condition in (6.3) is well-defined. Under this slow change assumption, we choose the number of times, K_n, the game must be played to produce a sequence of ε-PNE $\tilde{\boldsymbol{x}}^n$.

We assume that the players have access to random functions $\nabla_i \ell(\boldsymbol{x})$, denoted *stochastic gradients*, such that

$$\mathbb{E}\left[\nabla_i \ell(\boldsymbol{x}) \mid \boldsymbol{x}\right] = \nabla_i \mathscr{L}(\boldsymbol{x}) \qquad i = 1,2 \tag{6.4}$$

The players' learning dynamics are based on using K_n stochastic gradients. We assume that both players know the saddle point functions and that both can compute $\nabla_1 \ell(\boldsymbol{x})$ and $\nabla_2 \ell(\boldsymbol{x})$. As a result, each player can independently compute the Nash equilibrium without cooperation of the player. We can view this as a case where there is a cost to receiving feedback useful for learning a NE, which each player wants to control while still producing a useful ε-PNE.

Since each player's ε-PNE $\tilde{\boldsymbol{x}}$ is thus a random variable, the criteria we consider are as follows:

$$\begin{aligned} \mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\mathscr{L}^n(\tilde{\boldsymbol{x}}_1^n, \tilde{\boldsymbol{x}}_2^n) - \mathscr{L}^n(\boldsymbol{x}_1, \tilde{\boldsymbol{x}}_2^n)\right] \leq \varepsilon \qquad & \forall \boldsymbol{x}_1 \in \mathscr{X}_1 \\ \mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\mathscr{L}^n(\tilde{\boldsymbol{x}}_1^n, \boldsymbol{x}_2) - \mathscr{L}^n(\tilde{\boldsymbol{x}}_1^n, \tilde{\boldsymbol{x}}_2^n)\right] \leq \varepsilon \qquad & \forall \boldsymbol{x}_2 \in \mathscr{X}_2 \end{aligned} \tag{6.5}$$

This condition ensures that (6.2) holds on average. We are in effect trying to control the average quality of the ε-PNE that we produce. This criterion is a direct analog of the mean criterion for optimization problems from (1.7). For convenience, when we refer to generating an ε-PNE $\tilde{\boldsymbol{x}}$, we mean in the sense of (6.5).

Assumptions: Suppose that $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is m_1 strongly convex in $\boldsymbol{x}_1$, M_1 Lipschitz gradients in $\boldsymbol{x}_1$, m_2-strongly concave in $\boldsymbol{x}_2$, M_2 Lipschitz gradients in $\boldsymbol{x}_2$,

$$\mathbb{E}\|\nabla_i \mathscr{L}(\boldsymbol{x})\|_2^2 \leq A_i + B_i\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2 \quad i = 1,2 \tag{6.6}$$

and

$$\max_{\boldsymbol{x}_1 \in \mathscr{X}_1, \boldsymbol{x}_2 \in \mathscr{X}_2} \|\nabla_{12}^2 \mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)\|_2 \leq H \tag{6.7}$$

In addition, we assume that $\nabla_i \mathscr{L}(\bar{\boldsymbol{x}}) = 0$ for $i = 1,2$, and we set $m = \min\{m_1, m_2\}$ and $M = \max\{M_1, M_2\}$. We generally need $m > H$ for our work.

6.3 Learning for Fixed Zero Sum Games

We examine how players can learn ε-PNE for a fixed stage game using a gradient-based learning dynamic. In this section, we drop the index n, since we only consider a single stage game.

First, we give conditions for a unique saddle point to exist. By the continuity and the convexity/concavity of $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and the compactness of $\mathscr{X}_1$ and $\mathscr{X}_2$, we can apply Sion's minimax theorem [65] to show the existence of a saddle point. Furthermore, we want each of our saddle point functions to possess a unique saddle point in order for (6.3) to be well-defined. Lemma 9 guarantees the uniqueness of the saddle point under the strong convexity and strong concavity conditions.

Lemma 9. *Suppose that the saddle point function $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is m_1-strongly convex in $\boldsymbol{x}_1$ and m_2-strongly concave in $\boldsymbol{x}_2$. Then $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ possesses a unique saddle point.*

Proof. Suppose that $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ possesses two distinct saddle points $\bar{\boldsymbol{x}}$ and $\check{\boldsymbol{x}}$. Without loss of generality, suppose that $\bar{\boldsymbol{x}}_1 \neq \check{\boldsymbol{x}}_1$. Since $\bar{\boldsymbol{x}}$ is a saddle point, it follows that

$$\langle \nabla_1 \mathscr{L}(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2), \check{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_1 \rangle \geq 0 \quad \forall \check{\boldsymbol{x}}_1 \in \mathscr{X}_1$$

By the strong convexity of $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in $\boldsymbol{x}_1$, it holds that

$$\begin{aligned}
\mathscr{L}(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2) &< \mathscr{L}(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2) + \frac{1}{2} m_1 \|\breve{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_1\|_2^2 \\
&\leq \mathscr{L}(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2) + \langle \nabla_1 \mathscr{L}(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2), \breve{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_1 \rangle + \frac{1}{2} m_1 \|\breve{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_1\|_2^2 \\
&\leq \mathscr{L}(\breve{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2) \\
&\leq \mathscr{L}(\breve{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2)
\end{aligned}$$

Similarly, starting from $\breve{\boldsymbol{x}}$, it holds that

$$\mathscr{L}(\breve{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2) < \mathscr{L}(\breve{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2) \leq \mathscr{L}(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2)$$

This is a contradiction, so $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ possesses a unique saddle point. □

Second, we look at the simple case of finding one saddle point of $\mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Consider the following gradient learning dynamics with step sizes $\mu(k)$:

$$\begin{aligned}
\boldsymbol{x}_1(k+1) &= \Pi_{\mathscr{X}_1}\left[\boldsymbol{x}_1(k) - \mu(k+1)\nabla_1 \ell(\boldsymbol{x}(k))\right] \\
\boldsymbol{x}_2(k+1) &= \Pi_{\mathscr{X}_2}\left[\boldsymbol{x}_2(k) + \mu(k+1)\nabla_2 \ell(\boldsymbol{x}(k))\right]
\end{aligned} \quad k = 0, \ldots, K-1 \tag{6.8}$$

As noted in the related work section, this is the stochastic version of the Arrow-Hurwicz-Uzawa algorithm for finding a saddle point [60].

The analysis in Lemma 10 is a first step to connecting the number of iterations K in (6.8) to the ε-PNE quality. Lemma 10 connects the distance of the iterates from the PNE to the number of iterations K.

Lemma 10. *Provided that* (6.10) *and* (6.7) *hold with $m > H$, it follows that*

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_1(k+1) - \bar{\boldsymbol{x}}_1\|_2^2 \leq (1 - 2m_1\mu(k+1) + B\mu^2(k+1))\mathbb{E}\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2^2 \\
+ 2H\mu(k+1)\mathbb{E}\left[\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2 \|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2\right] \\
+ B\mu^2(k+1)\mathbb{E}\|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2^2 + A\mu^2(k+1) \\
\mathbb{E}\|\boldsymbol{x}_2(k+1) - \bar{\boldsymbol{x}}_2\|_2^2 \leq (1 - 2m\mu(k+1) + B\mu^2(k+1))\mathbb{E}\|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|^2 \\
+ 2H\mu(k+1)\mathbb{E}\left[\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2 \|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2\right] \\
+ B\mu^2(k+1)\mathbb{E}\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2^2 + A\mu^2(k+1)
\end{aligned}$$

Proof. We have

$$\begin{aligned}\|\boldsymbol{x}_1(k+1)-\bar{\boldsymbol{x}}_1\|_2^2 &\leq \|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1-\mu(k+1)\nabla_1\ell(\boldsymbol{x}(k))\|_2^2 \\ &\leq \|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\|_2^2-2\mu(k+1)\left\langle\nabla_1\ell(\boldsymbol{x}(k)),\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\right\rangle \\ &\quad+\mu^2(k+1)\|\nabla_1\ell(\boldsymbol{x}(k))\|_2^2\end{aligned}$$

Define the σ-algebra

$$\mathscr{F}(k)=\sigma\left(\nabla_1\ell(\boldsymbol{x}(1)),\ldots,\nabla_1\ell(\boldsymbol{x}(k))\right)$$

Then it holds that

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}_1(k+1)-\bar{\boldsymbol{x}}_1\|_2^2\mid\mathscr{F}(k)\right] \\ &\quad\leq\|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\|_2^2+2\mu(k+1)\left\langle\nabla_1\mathscr{L}(\boldsymbol{x}(k)),\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\right\rangle \\ &\qquad+\mu^2(k+1)\mathbb{E}\left[\|\nabla_1\ell(\boldsymbol{x}(k))\|_2^2\mid\mathscr{F}(k)\right]\end{aligned}$$

Taking the unconditional expectation over $\mathscr{F}(k)$ yields

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}_1(k+1)-\bar{\boldsymbol{x}}_1\|_2^2\leq\mathbb{E}\|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\|_2^2-2\mu(k+1)\mathbb{E}\left\langle\nabla_1\mathscr{L}(\boldsymbol{x}(k)),\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\right\rangle \\ +\mu^2(k)\mathbb{E}\|\nabla_1\ell(\boldsymbol{x}(k))\|_2^2\end{aligned}\tag{6.9}$$

By assumption, we have

$$\mathbb{E}\left[\|\nabla_1\ell(\boldsymbol{x}(k))\|_2^2\right]\leq A_1+B_1\|\boldsymbol{x}(k)-\bar{\boldsymbol{x}}\|_2^2\tag{6.10}$$

By Taylor's theorem, for some point $\tilde{\boldsymbol{x}}$, we have

$$\nabla_1\mathscr{L}(\boldsymbol{x}(k)) \;=\; \nabla_1\mathscr{L}(\bar{\boldsymbol{x}})+\nabla_{11}^2\mathscr{L}(\tilde{\boldsymbol{x}})(\boldsymbol{x}_1(k)-\tilde{\boldsymbol{x}}_1)+\nabla_{12}^2\mathscr{L}(\tilde{\boldsymbol{x}})(\boldsymbol{x}_2(k)-\tilde{\boldsymbol{x}}_2)$$

This in turn implies that

$$\begin{aligned}
&\langle \nabla_1 \mathscr{L}(\boldsymbol{x}(k)), \boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1 \rangle \\
&\quad = \langle \nabla_1 \mathscr{L}(\bar{\boldsymbol{x}}), \boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1 \rangle + \left\langle \nabla_{11}^2 \mathscr{L}(\tilde{\boldsymbol{x}})(\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1), \boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1 \right\rangle \\
&\qquad\qquad + \left\langle \nabla_{12}^2 \mathscr{L}(\tilde{\boldsymbol{x}})(\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2), \boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1 \right\rangle \\
&\quad \geq \langle \nabla_1 \mathscr{L}(\bar{\boldsymbol{x}}), \boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1 \rangle + m_1 \|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2^2 \\
&\qquad\qquad - H \|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2 \|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2
\end{aligned} \tag{6.11}$$

Substituting the bounds in (6.10) and (6.11) into (6.9), it holds that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_1(k+1) - \bar{\boldsymbol{x}}_1\|_2^2 \leq (1 - 2m_1\mu(k+1) + B_1\mu^2(k+1))\mathbb{E}\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2^2 \\
+ 2H\mu(k+1)\mathbb{E}\left[\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2 \|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2\right] \\
+ B_1\mu^2(k+1)\mathbb{E}\|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2^2 + A_1\mu^2(k)
\end{aligned}$$

Similarly, for the second player, it holds that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_2(k+1) - \bar{\boldsymbol{x}}_2\|_2^2 \leq (1 - 2m_2\mu(k+1) + B_2\mu^2(k+1))\mathbb{E}\|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2^2 \\
+ 2H\mu(k+1)\mathbb{E}\left[\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2 \|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2\right] \\
+ B_2\mu^2(k+1)\mathbb{E}\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2^2 + A_2\mu^2(k)
\end{aligned}$$

□

Define the pair of recursions

$$\begin{aligned}
a(k+1) &= (1 - 2m_1\mu(k+1) + B_1\mu^2(k+1))a(k) + 2H\mu(k+1)\sqrt{a(k)b(k)} \\
&\qquad + B_1\mu^2(k+1)b(k) + A_1\mu^2(k+1) \\
b(k+1) &= (1 - 2m_2\mu(k+1) + B_2\mu^2(k+1))b(k) + 2H\mu(k+1)\sqrt{a(k)b(k)} \\
&\qquad + B_2\mu^2(k+1)a(k) + A_2\mu^2(k+1)
\end{aligned} \tag{6.12}$$

with

$$\mathbb{E}\|\boldsymbol{x}_1(k) - \bar{\boldsymbol{x}}_1\|_2^2 \leq a(0)$$

and

$$\mathbb{E}\|\boldsymbol{x}_2(k) - \bar{\boldsymbol{x}}_2\|_2^2 \leq b(0)$$

From the Cauchy-Schwarz inequality, it follows that

$$\mathbb{E}\left[\|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\|_2\|\boldsymbol{x}_2(k)-\bar{\boldsymbol{x}}_2\|_2\right] \quad \leq \quad \left(\mathbb{E}\|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\|_2^2\right)^{1/2}\left(\mathbb{E}\|\boldsymbol{x}_2(k)-\bar{\boldsymbol{x}}_2\|_2^2\right)^{1/2}$$

Therefore, using this observation combined with the bounds in Lemma 10, it follows that

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}_1(k)-\bar{\boldsymbol{x}}_1\|_2^2 &\leq a(k)\\ \mathbb{E}\|\boldsymbol{x}_2(k)-\bar{\boldsymbol{x}}_2\|_2^2 &\leq b(k)\end{aligned} \tag{6.13}$$

Remark: Suppose that $m_1 = m_2 = m$, $A_1 = A_2$, and $B_1 = B_2$. Consider minimizing a $m-H$-strongly convex function $f(\boldsymbol{x})$ by applying stochastic gradient descent using the gradient bound

$$\mathbb{E}\|\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x})\|_2^2 \leq A + 2B\|\boldsymbol{x}-\boldsymbol{x}^*\|_2^2$$

where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathscr{X}}\phi(\boldsymbol{x})$. Then it holds that

$$\mathbb{E}\|\boldsymbol{x}(K)-\boldsymbol{x}^*\|_2^2 \leq a(K)$$

This shows that the analysis of the saddle point problem in this chapter is similar to the analysis of the optimization problem in previous chapters.

Define

$$b(a(0),b(0),K) = a(K)+b(K) \tag{6.14}$$

This bound in turn satisfies

$$\mathbb{E}\|\boldsymbol{x}(K)-\bar{\boldsymbol{x}}\|_2^2 \leq b(a(0),b(0),K)$$

This gives us control on the strategies produced by gradient ascent and their closeness to the PNE, but does not guarantee that $\boldsymbol{x}(K)$ is an ε-PNE. We examine a way to translate from a strategy's proximity to an ε-PNE to guaranteeing that the strategy is in fact an ε-PNE. Applying strong convexity, it holds that

$$\mathscr{L}(\boldsymbol{x}_1,\boldsymbol{x}_2(K))-\mathscr{L}(\boldsymbol{x}(K)) \quad \geq \quad \langle\nabla_1\mathscr{L}(\boldsymbol{x}(K)),\boldsymbol{x}_1-\boldsymbol{x}_1(K)\rangle+\frac{1}{2}m_1\|\boldsymbol{x}_1-\boldsymbol{x}_1(K)\|_2^2$$

This lower bound is a positive definite quadratic in $\boldsymbol{x}_1$. Minimizing over $\boldsymbol{x}_1$ and

rearranging yields

$$\mathscr{L}(\boldsymbol{x}(K)) - \mathscr{L}(\boldsymbol{x}_1, \boldsymbol{x}_2(K)) \;\leq\; \frac{1}{2m_1}\|\nabla_1 \mathscr{L}(\boldsymbol{x}(K))\|_2^2 \;\; \forall \boldsymbol{x}_1 \in \mathscr{X}_1$$

This guarantees that if $\|\nabla_1 \mathscr{L}(\boldsymbol{x}(K))\|_2$ is small, then player one cannot decrease $\mathscr{L}(\boldsymbol{x}(K))$ by a large amount. Similarly, for player two it holds that

$$\mathscr{L}(\boldsymbol{x}_1(K), \boldsymbol{x}_2) - \mathscr{L}(\boldsymbol{x}(K)) \;\leq\; \frac{1}{2m_2}\|\nabla_2 \mathscr{L}(\boldsymbol{x}(K))\|_2^2 \;\; \forall \boldsymbol{x}_2 \in \mathscr{X}_2$$

If it holds that

$$\max\left\{\frac{1}{2m_1}\|\nabla_1 \mathscr{L}(\boldsymbol{x}(K))\|_2^2, \frac{1}{2m_2}\|\nabla_2 \mathscr{L}(\boldsymbol{x}(K))\|_2^2\right\} \leq \varepsilon \tag{6.15}$$

then $\boldsymbol{x}(K)$ is an ε-PNE as desired. By the Lipschitz gradient property applied at the PNE $\bar{\boldsymbol{x}}$, it holds that

$$\begin{aligned}\|\nabla_1 \mathscr{L}(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)\|_2^2 &= \|\nabla_1 \mathscr{L}(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2) - \nabla_1 \mathscr{L}(\bar{\boldsymbol{x}})\|_2^2 \\ &\leq M_1^2 \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1\|_2^2\end{aligned}$$

and

$$\begin{aligned}\|\nabla_2 \mathscr{L}(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)\|_2^2 &= \|\nabla_2 \mathscr{L}(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2) - \nabla_2 \mathscr{L}(\bar{\boldsymbol{x}})\|_2^2 \\ &\leq M_2^2 \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2\|_2^2\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\max\left\{\frac{1}{2m_1}\|\nabla_1 \mathscr{L}(\boldsymbol{x}(K))\|_2^2, \frac{1}{2m_2}\|\nabla_2 \mathscr{L}(\boldsymbol{x}(K))\|_2^2\right\} \\ &\qquad \leq \frac{1}{2\min\{m_1, m_2\}}\left(\|\nabla_1 \mathscr{L}(\boldsymbol{x}(K))\|_2^2 + \|\nabla_2 \mathscr{L}(\boldsymbol{x}(K))\|_2^2\right) \\ &\qquad \leq \frac{\max\{M_1, M_2\}^2}{2\min\{m_1, m_2\}}\|\boldsymbol{x}(K) - \bar{\boldsymbol{x}}\|_2^2 \end{aligned} \tag{6.16}$$

and

$$\mathbb{E}\left[\max\left\{\frac{1}{2m_1}\|\nabla_1\mathscr{L}(\boldsymbol{x}(K))\|_2^2, \frac{1}{2m_2}\|\nabla_2\mathscr{L}(\boldsymbol{x}(K))\|_2^2\right\}\right] \\ \leq \frac{\max\{M_1,M_2\}^2}{2\min\{m_1,m_2\}} b\left(\mathbb{E}\|\boldsymbol{x}_1(0)-\bar{\boldsymbol{x}}_1\|_2^2, \mathbb{E}\|\boldsymbol{x}_2(0)-\bar{\boldsymbol{x}}_2\|_2^2, K\right) \tag{6.17}$$

Provided that

$$\frac{\max\{M_1,M_2\}^2}{2\min\{m_1,m_2\}} b\left(\mathbb{E}\|\boldsymbol{x}_1(0)-\bar{\boldsymbol{x}}_1\|_2^2, \mathbb{E}\|\boldsymbol{x}_2(0)-\bar{\boldsymbol{x}}_2\|_2^2, K\right) \leq \varepsilon$$

it follows that (6.15) holds yielding an ε-PNE.

We have control on the distance between the strategies produced by the gradient learning dynamics and the distance between the PNE and the resulting ε-PNE. By setting

$$K^* = \min\left\{K \geq 1 \;\middle|\; \frac{\max\{M_1,M_2\}^2}{2\min\{m_1,m_2\}} b\left(\mathbb{E}\|\boldsymbol{x}_1(0)-\bar{\boldsymbol{x}}_1\|_2^2, \mathbb{E}\|\boldsymbol{x}_2(0)-\bar{\boldsymbol{x}}_2\|_2^2, K\right) \leq \varepsilon\right\}$$

we produce an ε-PNE. With this control we will be able to learn an ε-PNE for each stage game in the next section.

6.4 Learning for Time-Varying Zero Sum Games

We apply the same gradient ascent learning dynamics as in (6.8) except with a time index n for the current stage game

$$\begin{aligned} \boldsymbol{x}_1^n(k+1) &= \Pi_{\mathscr{X}}\left[\boldsymbol{x}_1^n(k) - \mu_n(k+1)\nabla_1\ell^n(\boldsymbol{x}^n(k))\right] \\ \boldsymbol{x}_2^n(k+1) &= \Pi_{\mathscr{X}}\left[\boldsymbol{x}_2^n(k) + \mu_n(k+1)\nabla_2\ell^n(\boldsymbol{x}^n(k))\right] \\ \boldsymbol{x}^n(0) &\triangleq \boldsymbol{x}^{n-1}(K_{n-1}) \\ \tilde{\boldsymbol{x}}^n &\triangleq \boldsymbol{x}^n(K_n) \end{aligned} \tag{6.18}$$

with $k = 0, \ldots, K_n - 1$. At time n, we start with the strategy generated by the learning dynamics for the previous stage game. Since the stage games change slowly according to (6.3), this is a reasonable starting point. The last iterate $\boldsymbol{x}^n(K_n)$ is our ε-PNE $\tilde{\boldsymbol{x}}^n$. Note that each player can unilaterally apply this method due to its

knowledge of the saddle point function.

In light of (6.17), to generate a sequence of ε-PNE, it is sufficient to generate a sequence of iterates $\boldsymbol{x}^n(K_n)$ such that

$$\mathbb{E}\|\boldsymbol{x}^n(K_n) - \bar{\boldsymbol{x}}^n\|_2^2 \leq \tilde{\varepsilon}$$

with

$$\tilde{\varepsilon} = \frac{2\min\{m_1, m_2\}\varepsilon}{\max\{M_1, M_2\}^2}$$

We start with the trivial bounds

$$\mathbb{E}\|\boldsymbol{x}_1^1(0) - \bar{\boldsymbol{x}}_1^1\|_2^2 \leq \operatorname{diam}^2(\mathscr{X})$$

and

$$\mathbb{E}\|\boldsymbol{x}_2^1(0) - \bar{\boldsymbol{x}}_2^1\|_2^2 \leq \operatorname{diam}^2(\mathscr{X})$$

We set

$$K_1 = \min\left\{K \geq 1 \;\middle|\; b\left(\operatorname{diam}^2(\mathscr{X}_1), \operatorname{diam}^2(\mathscr{X}_2), K\right) \leq \tilde{\varepsilon}\right\}$$

Proceeding inductively, suppose that at the beginning of time instant n, we have

$$\mathbb{E}\|\boldsymbol{x}^n(0) - \bar{\boldsymbol{x}}^{n-1}\|_2^2 \leq \tilde{\varepsilon}$$

By the triangle inequality for $i = 1, 2$, we have

$$\begin{aligned}
\left(\mathbb{E}\|\boldsymbol{x}_i^n(0) - \bar{\boldsymbol{x}}_i^n\|_2^2\right)^{1/2} &\leq \left(\mathbb{E}\|\boldsymbol{x}^n(0) - \bar{\boldsymbol{x}}^n\|_2^2\right)^{1/2} \\
&= \left(\mathbb{E}\|\boldsymbol{x}^{n-1} - \bar{\boldsymbol{x}}^n\|_2^2\right)^{1/2} \\
&\leq \left(\mathbb{E}\|\boldsymbol{x}^{n-1} - \bar{\boldsymbol{x}}^{n-1}\|_2^2\right)^{1/2} + \|\bar{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^{n-1}\|_2 \\
&\leq \sqrt{\tilde{\varepsilon}} + \rho
\end{aligned}$$

Using this bound and setting

$$K_n = \min\left\{K \geq 1 \;\middle|\; b\left(\left(\sqrt{\tilde{\varepsilon}} + \rho\right)^2, \left(\sqrt{\tilde{\varepsilon}} + \rho\right)^2, K\right) \leq \tilde{\varepsilon}\right\}$$

yields an ε-PNE $\boldsymbol{x}^n(K_n)$.

Disagreement Among Players

Assuming that all players know the saddle point functions $\mathscr{L}^n(\boldsymbol{x})$, they can compute an ε-PNE to their desired accuracy. However, the players may choose different ε. As a result player one discovers an ε_1-PNE $\tilde{\boldsymbol{x}}$ and plays $\tilde{\boldsymbol{x}}_1$. Player two discovers an ε_2-PNE $\breve{\boldsymbol{x}}$ and plays $\breve{\boldsymbol{x}}_2$. The question then is what can we say about the combination of strategies $(\tilde{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2)$? We know that for player one

$$\mathscr{L}(\breve{\boldsymbol{x}}) - \mathscr{L}(\tilde{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2) \leq \varepsilon_2$$

and for player two

$$\mathscr{L}(\tilde{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2) - \mathscr{L}(\tilde{\boldsymbol{x}}) \leq \varepsilon_1$$

This ensures that player one's deviation from the ε_2-PNE computed by player two cannot reduce $\mathscr{L}(\breve{\boldsymbol{x}})$ by more than ε_2. Similarly, player two's deviation from player one's ε_1-PNE cannot increase $\mathscr{L}(\breve{\boldsymbol{x}})$ by more than ε_1. Therefore, a player's choice of ε constrains the benefit that its opponent can accrue by deviating from the player's choice of ε-PNE. A player cannot, however, guarantee that it is impossible to substantially improve its own reward.

6.5 Estimating the Change in the Nash Equilibrium

It is also possible that we do not know ρ and must estimate it. We look at the special case with

$$\|\bar{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^{n-1}\|_2 = \rho$$

By using the ε-PNE generated through (6.18) and the triangle inequality, we have

$$\|\boldsymbol{x}^n - \boldsymbol{x}^{n-1}\|_2 \leq \|\tilde{\boldsymbol{x}}^n - \tilde{\boldsymbol{x}}^{n-1}\|_2 + \|\tilde{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^n\|_2 + \|\tilde{\boldsymbol{x}}^{n-1} - \bar{\boldsymbol{x}}^{n-1}\|_2 \tag{6.19}$$

Lemma 11 provides a tool to bound $\|\boldsymbol{x} - \bar{\boldsymbol{x}}^n\|_2$.

Lemma 11. *Provided that* $\min\{m_1, m_2\} > H$*, for any strategies* $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$*, it holds that*

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}^n\|_2 \leq \frac{1}{\min\{m_1, m_2\} - H} \left(\|\nabla_1 \mathscr{L}^n(\boldsymbol{x})\|_2 + \|\nabla_2 \mathscr{L}^n(\boldsymbol{x})\|_2\right)$$

Proof. By Taylor's theorem, for some point $\breve{\boldsymbol{x}}^n$ it holds that

$$\begin{aligned}
&\langle \boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n, \nabla_1 \mathscr{L}^n(\boldsymbol{x}) \rangle \\
&\quad = \langle \boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n, \nabla_1 \mathscr{L}^n(\bar{\boldsymbol{x}}^n) \rangle + \left\langle \boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n, \nabla_{11}^2 \mathscr{L}^n(\breve{\boldsymbol{x}}^n)(\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n) \right\rangle \\
&\qquad\qquad + \left\langle \boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n, \nabla_{12}^2 \mathscr{L}^n(\breve{\boldsymbol{x}}^n)(\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n) \right\rangle \\
&\quad \geq m_1 \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2^2 - H \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 \cdot \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2
\end{aligned}$$

This in turn implies that

$$\|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 \|\nabla_1 \mathscr{L}^n(\boldsymbol{x})\|_2 \geq m_1 \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2^2 - H \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 \cdot \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2$$

and so

$$\|\nabla_1 \mathscr{L}^n(\boldsymbol{x})\|_2 \geq m_1 \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 - H \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2$$

Similarly, we have

$$\|\nabla_2 \mathscr{L}^n(\boldsymbol{x})\|_2 \geq m_2 \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2 - H \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2$$

This in turn implies that

$$(m_1 - H)\|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 + (m_2 - H)\|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2 \leq \|\nabla_1 \mathscr{L}^n(\boldsymbol{x})\|_2 + \|\nabla_2 \mathscr{L}^n(\boldsymbol{x})\|_2$$

By the sub-additivity of the square root function, it holds that

$$\begin{aligned}
\|\boldsymbol{x} - \bar{\boldsymbol{x}}^n\|_2 &= \sqrt{\|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2^2 + \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2^2} \\
&\leq \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 + \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2
\end{aligned}$$

This in turn implies that

$$\begin{aligned}
(\min\{m_1, m_2\} - H)\, \|\boldsymbol{x} - \bar{\boldsymbol{x}}^n\|_2 &\leq (\min\{m_1, m_2\} - H)\, (\|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 + \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2) \\
&\leq (m_1 - H)\|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1^n\|_2 + (m_2 - H)\|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2^n\|_2 \\
&\leq \|\nabla_1 \mathscr{L}^n(\boldsymbol{x})\|_2 + \|\nabla_2 \mathscr{L}^n(\boldsymbol{x})\|_2
\end{aligned}$$

□

Using Lemma 11 and (6.19), we have the bound

$$\begin{aligned}
&\|\bar{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^{n-1}\|_2 \\
&\quad \leq \|\tilde{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^n\|_2 + \|\tilde{\boldsymbol{x}}^n - \tilde{\boldsymbol{x}}^{n-1}\|_2 + \|\tilde{\boldsymbol{x}}^{n-1} - \bar{\boldsymbol{x}}^{n-1}\|_2 \\
&\quad \leq \|\tilde{\boldsymbol{x}}^n - \tilde{\boldsymbol{x}}^{n-1}\|_2 \\
&\qquad + \frac{1}{\min\{m_1, m_2\} - H} \left(\|\nabla_1 \mathscr{L}^n(\tilde{\boldsymbol{x}}^n)\|_2 + \|\nabla_2 \mathscr{L}^n(\tilde{\boldsymbol{x}}^n)\|_2\right) \\
&\qquad + \frac{1}{\min\{m_1, m_2\} - H} \left(\|\nabla_1 \mathscr{L}^{n-1}(\tilde{\boldsymbol{x}}^{n-1})\|_2 + \|\nabla_2 \mathscr{L}^{n-1}(\tilde{\boldsymbol{x}}^{n-1})\|_2\right)
\end{aligned}$$

Therefore, we can define the one-step estimate of ρ

$$\begin{aligned}
\tilde{\rho}_n \triangleq\ & \|\tilde{\boldsymbol{x}}^n - \tilde{\boldsymbol{x}}^{n-1}\|_2 \\
&+ \frac{1}{\min\{m_1, m_2\} - H} \left(\left\| \frac{1}{K_n} \sum_{k=1}^{K_n} \nabla_1 \ell_k^n(\tilde{\boldsymbol{x}}^n) \right\|_2 + \left\| \frac{1}{K_n} \sum_{k=1}^{K_n} \nabla_2 \ell_k^n(\tilde{\boldsymbol{x}}^n) \right\|_2 \right) \\
&+ \frac{1}{\min\{m_1, m_2\} - H} \left(\left\| \frac{1}{K_{n-1}} \sum_{k=1}^{K_{n-1}} \nabla_1 \ell_k^{n-1}(\tilde{\boldsymbol{x}}^{n-1}) \right\|_2 + \right. \\
&\qquad \left. \left\| \frac{1}{K_{n-1}} \sum_{k=1}^{K_{n-1}} \nabla_2 \ell_k^{n-1}(\tilde{\boldsymbol{x}}^{n-1}) \right\|_2 \right)
\end{aligned}$$

We can then combine the one-step estimates $\tilde{\rho}_i$ by averaging to yield an estimate

$$\hat{\rho}_n = \frac{1}{n-1} \sum_{i=2}^{n} \tilde{\rho}_i$$

By using Lemma 16 as in Theorems 1 and 3, it is possible to prove the for all n large enough

$$\hat{\rho}_n + t_n \geq \rho$$

almost surely for appropriate choices of sequences $\{t_n\}$. This inures that we satisfy Assumption C.1. Since the proof is nearly identical, we omit it here.

By examining our bound $b(d_0, d_0, K)$ defined in (6.14), it can be seen that bound factors as in Assumption C.2 of Chapter 3. Thus, we can apply Theorem 5 to argue that we eventually produce a sequence of ε_n-PNE $\tilde{\boldsymbol{x}}^n$ with

$$\limsup_{n \to \infty} \varepsilon_n \leq \varepsilon$$

6.6 Experiments

We examine a simple synthetic example of a zero-sum game and an example of an adversary attacking a surveillance network.

Synthetic Zero-Sum Game

Consider the following stochastic convex-concave, zero-sum game corresponding to finding the saddle point of

$$\mathscr{L}^n(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \mathbb{E}_{(a^n, b^n) \sim p_n} \left[\ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2) \right]$$

with

$$\ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \frac{1}{2} m_1 \|\boldsymbol{x}_1\|_2^2 + \langle a^n, \boldsymbol{x}_1 \rangle + H \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle - \langle b^n, \boldsymbol{x}_2 \rangle - \frac{1}{2} m_2 \|\boldsymbol{x}_2\|_2^2$$

The random variables a^n and b^n are unknown to the players. Note that this includes the case where a^n and b^n are deterministic as well. This function satisfies the assumptions of Section 6.2. The stochastic gradients in this problem are of the form

$$\begin{aligned} \nabla_1 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2)(k) &= m_1 \boldsymbol{x}_1 + a^n(k) + H \boldsymbol{x}_2 \\ \nabla_2 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2)(k) &= -m_2 \boldsymbol{x}_2 + b^n(k) + H \boldsymbol{x}_1 \end{aligned} \qquad k = 1, \ldots, K_n$$

with $(a^n(k), b^n(k)) \overset{\text{iid}}{\sim} p_n$. In effect, each time we play the game, we can elect to receive a realization (a^n, b^n), which can be used to carry out the learning strategy from (6.18).

We can find the PNE in a closed form

$$\bar{\boldsymbol{x}}^n = \begin{bmatrix} \frac{m_2 \mathbb{E}[a^n] - H \mathbb{E}[b^n]}{m_1 m_2 + H^2} \\ -\frac{m_1 \mathbb{E}[b^n] + H \mathbb{E}[a^n]}{m_1 m_2 + H^2} \end{bmatrix}$$

This in turn implies that

$$\begin{aligned}
&\|\bar{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^{n-1}\|_2^2 \\
&\quad = \left\| \frac{m_2 \mathbb{E}[a^n] - H\mathbb{E}[b^n]}{m_1 m_2 + H^2} - \frac{m_2 \mathbb{E}[a^{n-1}] - H\mathbb{E}[b^{n-1}]}{m_1 m_2 + H^2} \right\|_2^2 \\
&\qquad\qquad + \left\| - \frac{m_1 \mathbb{E}[b^n] + H\mathbb{E}[a^n]}{m_1 m_2 + H^2} + \frac{m_1 \mathbb{E}[b^{n-1}] + H\mathbb{E}[a^{n-1}]}{m_1 m_2 + H^2} \right\|_2^2 \\
&\quad \leq \frac{2(m_2^2 + H^2)}{(m_1 m_2 + H^2)^2} \mathbb{E}\|a^n - a^{n-1}\|_2^2 + \frac{2(m_1^2 + H^2)}{(m_1 m_2 + H^2)^2} \mathbb{E}\|b^n - b^{n-1}\|_2^2
\end{aligned}$$

and so it follows that

$$\|\bar{\boldsymbol{x}}^n - \bar{\boldsymbol{x}}^{n-1}\|_2 \leq \frac{\sqrt{2(m_2^2 + H^2)}}{m_1 m_2 + H^2} \|a^n - a^{n-1}\|_{L_2} + \frac{\sqrt{2(m_1^2 + H^2)}}{m_1 m_2 + H^2} \|b^n - b^{n-1}\|_{L_2}$$

with $\|\boldsymbol{x}\|_{L_2} = \left(\mathbb{E}\|\boldsymbol{x}\|_2^2\right)^{1/2}$. Therefore, by controlling the size of the changes in a^n and b^n using the L_2-norm, it follows that the PNE changes slowly according to (6.3).

For this problem, clearly we have $m_1 = M_1$, $m_2 = M_2$, and H. For the gradient growth condition in (6.10), we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_1 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2)\|_2^2 \\
&\quad \leq 2\mathbb{E}\|\nabla_1 \ell^n(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2)\|_2^2 + 2\mathbb{E}\|\nabla_1 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2) - \nabla_1 \ell^n(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2)\|_2^2 \\
&\quad = 2\mathrm{tr}\left\{\mathrm{Cov}\left(a^n\right)\right\} + 4m_1^2 \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1\|_2^2 + 4H^2 \|\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2\|_2^2
\end{aligned}$$

A similar bound holds for $\nabla_2 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2)$, so we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_1 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2)\|_2^2 + \mathbb{E}\|\nabla_2 \ell^n(\boldsymbol{x}_1, \boldsymbol{x}_2)\|_2^2 \\
&\qquad\qquad \leq 2\mathrm{tr}\left\{\mathrm{Cov}\left(a^n\right)\right\} + 2\mathrm{tr}\left\{\mathrm{Cov}\left(b^n\right)\right\} + 4\left(m_1^2 + H^2\right)\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2
\end{aligned}$$

Figure 6.1 shows the estimate of ρ. As desired, this estimate upper bounds the true $\rho = 1$. Figure 6.2 shows the chosen number of rounds of play (K_n). This appears to settle down. Figure 6.3 shows the quality of the achieved PNE. We achieve our target value of $\varepsilon = 0.2$.

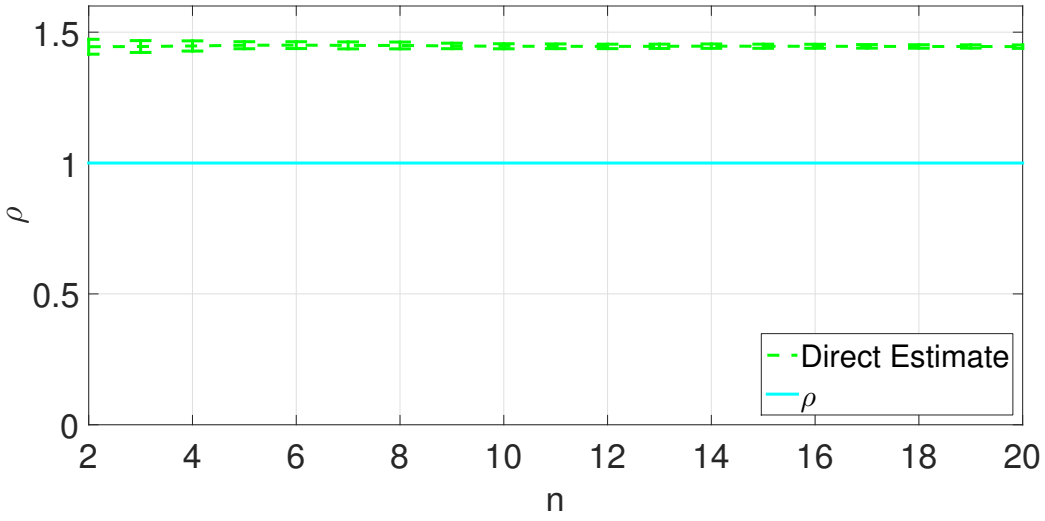


Figure 6.1: ρ Estimate

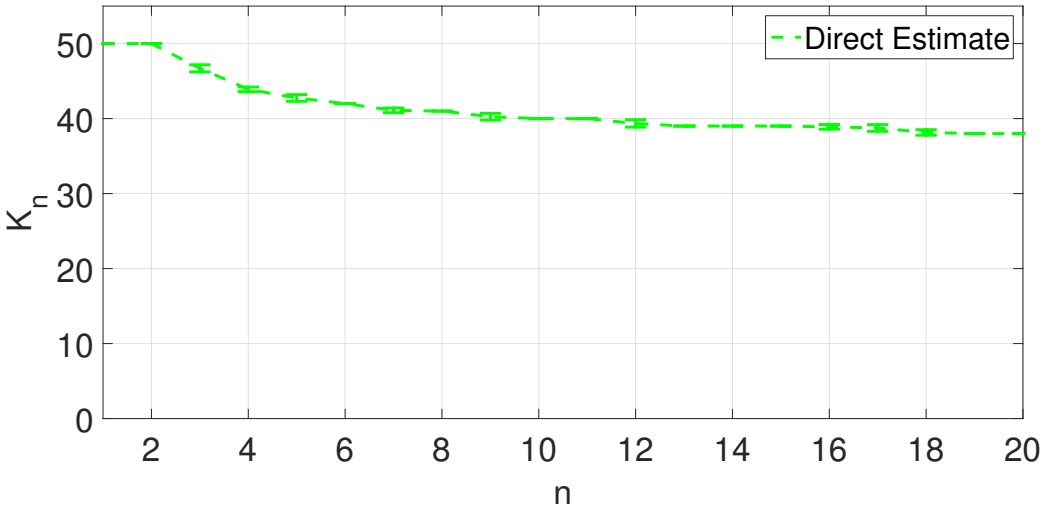


Figure 6.2: Choice of K_n

Surveillance Network Game

Consider an estimation problem in which a fusion center estimates a scalar quantity w^n during epoch n using the observations of two sensor nodes. An adversary has compromised one of these sensor nodes and sends a false signal to the fusion center. Figure 6.4 shows the surveillance network and adversary structure. We draw the compromised sensor closer to the signal w^n to visually indicate that it is compromised by the adversary. The goal of the fusion center is to estimate w^n and the goal of the adversary is to impede the estimation process, while avoiding detection of its actions. Avoiding detection is crucial, since otherwise the fusion center can ignore the compromised sensor node or reset it to regain control.

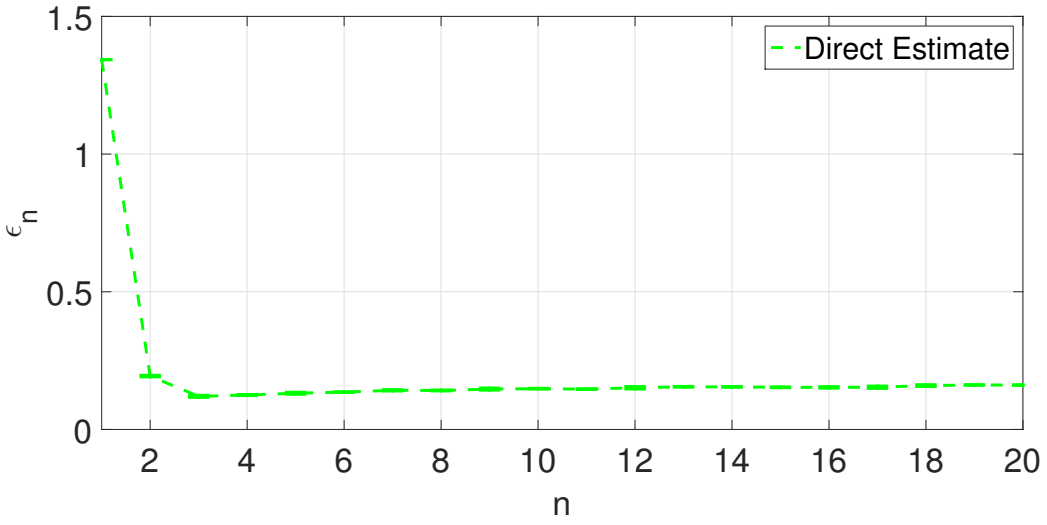


Figure 6.3: Quality of PNE

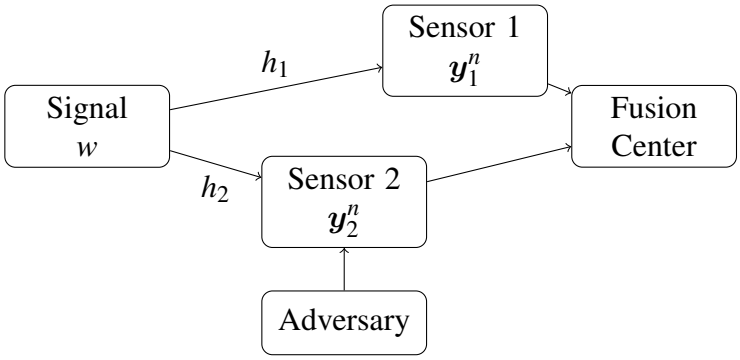


Figure 6.4: Surveillance network

We model this problem as a zero sum game that corresponds to finding the saddle point of the function

$$\mathscr{L}^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n) \triangleq \mathbb{E}_{y_1^n, y_2^n}\left[\phi(\boldsymbol{x}_1^n, y_1^n, y_2^n(\boldsymbol{x}_2^n)) + \psi(\boldsymbol{x}_2^n)\right]$$

where $\boldsymbol{x}_1^n$ is the action of the fusion center, $\boldsymbol{x}_2^n$ is the action of the adversary, y_1^n is the observation of the not compromised sensor node, and $y_2^n(\boldsymbol{x}_2^n)$ is the observation of the compromised sensor node. The function $\phi(\boldsymbol{x}_1^n, y_1^n, y_2^n(\boldsymbol{x}_2^n))$ is the estimation loss of the sensor network. The function $\psi(\boldsymbol{x}_2^n)$ penalizes the actions of the adversary based on how likely it is that the fusion center can detect the adversary's action.

We consider the specific problem with

$$\phi(\boldsymbol{x}_1^n, y_1^n, y_2^n(\boldsymbol{x}_2^n)) = \frac{1}{2}\left(w^n - (\boldsymbol{x}_1^n)^\top \begin{bmatrix} y_1^n \\ y_2^n(\boldsymbol{x}_2^n) \end{bmatrix}\right)^2 \tag{6.20}$$

and

$$\psi(\boldsymbol{x}_2^n) = -\frac{\lambda}{2}(\boldsymbol{x}_2^n)^2 \tag{6.21}$$

We model these observations as

$$\begin{aligned} y_1^n(k) &= w^n &&+ e_1^n(k) \\ y_2^n(k) &= w^n + \boldsymbol{x}_2^n &&+ e_2^n(k) \end{aligned} \qquad k = 1, \ldots, K_n$$

where $w^n \sim \mathcal{N}(\mu^n, \sigma_w^2)$, $e_1^n(k), e_2^n(k) \sim \mathcal{N}(0, \sigma_e^2)$, and $e_1^n(k), e_2^n(k), w^n$ are independent. Note that at each time instant only one w^n is drawn. We capture the idea of slow change by assuming that μ^n changes slowly. We do not have a closed form for the saddle point, but our simulation indicates that this slow change assumption on μ^n actually produces slow changes in the game's PNE.

The estimation loss $\phi(\boldsymbol{x}_1^n, y_1^n, y_2^n(\boldsymbol{x}_2^n))$ in this choice is mean squared error. To understand the motivation for the detection penalty, we consider a particular scheme for the fusion center to detect the adversary. Suppose that given K_n observations $y_1^n(1), \ldots, y_1^n(K_n)$ and $y_2^n(1), \ldots, y_2^n(K_n)$ the fusion center computes the test statistic

$$T_n \triangleq \frac{1}{2\sigma_e^2} \sum_{k=1}^{K_n} (y_2^n(k) - y_1^n(k))^2$$

Under our signal model, this test statistic becomes

$$T_n = \frac{1}{2\sigma_e^2} \sum_{k=1}^{K_n} (\boldsymbol{x}_2^n + e_2^n(k) - e_1^n(k))^2$$

If the adversary no action, then the statistic T_n is χ^2-distributed with K_n degrees of freedom. If the adversary takes an action, then T_n is noncentral χ^2-distributed with K_n degrees of freedom and noncentrality parameter

$$\lambda_n = \frac{K_n \left(\boldsymbol{x}_2^n\right)^2}{2\sigma_e^2} \tag{6.22}$$

Therefore, the fusion center can carry out the hypothesis test

$$\begin{aligned} H_0 &: T_n \sim \chi^2_{K_n} \\ H_1 &: T_n \sim \chi^2_{K_n}(\lambda_n) \end{aligned} \tag{6.23}$$

using a threshold test

$$\delta(y_2^n(1),\ldots,y_2^n(K_n)) = \begin{cases} 0, & T_n \leq \tau \\ 1, & T_n > \tau \end{cases} \tag{6.24}$$

The complementary CDF of the test statistic T_n depends on the Marcum Q-function [66], i.e.,

$$\mathbb{P}_{\boldsymbol{x}_2^n}\{T_n > \tau\} = Q_{\frac{K}{2}}\left(\sqrt{\lambda_n}, \sqrt{\tau}\right)$$

The false alarm probability of the test in (6.24) can be set at α by finding τ satisfying

$$\mathbb{P}_0\{T_n > \tau\} = Q_{\frac{K}{2}}\left(0, \sqrt{\tau}\right) = \alpha$$

The Marcum Q-function is monotone increasing in its first argument [66], so we have

$$\lim_{\lambda_n \to \infty} \mathbb{P}_{\boldsymbol{x}_2^n}\{T_n > \tau\} = 1$$

monotonically. Thus, the size of the noncentrality parameter directly controls the power of the test. From the form of the noncentrality parameter in (6.22), it follows that $\left(\boldsymbol{x}_2^n\right)^2$ directly controls the power of the test in (6.24). This justifies the detection penalty in (6.21), since $\left(\boldsymbol{x}_2^n\right)^2$ directly affects the power of the test in (6.24).

Expanding this model and evaluating the expectation, it holds that

$$\mathscr{L}^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n) = \frac{1}{2}\left((1 - \boldsymbol{x}_{11}^n - \boldsymbol{x}_{12}^n)w + \boldsymbol{x}_{12}^n \boldsymbol{x}_2^n\right)^2 + \frac{1}{2}\sigma_e^2 \|\boldsymbol{x}_1^n\|_2^2 - \frac{1}{2}\lambda\left(\boldsymbol{x}_2^n\right)^2 \tag{6.25}$$

This model does not fit into the framework exactly, so we consider an approximation of our saddle point function of the following form:

$$\begin{aligned} \mathscr{L}^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n) = \frac{1}{2}(1 - \boldsymbol{x}_{11}^n - \boldsymbol{x}_{12}^n)^2\left((\mu^n)^2 + \sigma_w^2\right) + (1 - \boldsymbol{x}_{11}^n - \boldsymbol{x}_{12}^n)\boldsymbol{x}_2^n \mu^n \\ + \gamma\sqrt{\left(\boldsymbol{x}_{12}^n \boldsymbol{x}_2^n\right)^2 + \delta^2} + \frac{1}{2}\sigma_e^2 \|\boldsymbol{x}_1^n\|_2^2 - \frac{1}{2}\lambda\left(\boldsymbol{x}_2^n\right)^2 \end{aligned} \tag{6.26}$$

Figure 6.5 shows the estimation loss, MSE, for the fusion center

$$\frac{1}{2}\left((1-\boldsymbol{x}_{11}-\boldsymbol{x}_{12})\left((\mu^n)^2+\sigma_w^2\right)-\boldsymbol{x}_{12}\boldsymbol{x}_2\right)^2+\frac{1}{2}\sigma_e^2\|\boldsymbol{x}_1^n\|_2^2$$

plotted against λ for both the exact saddle point function in (6.25) and the approximate saddle point function in (6.26). We see that the approximate saddle point problem roughly captures the trade-off between detection probability, through λ, and MSE of the exact saddle point problem. Finally, we assume that we have

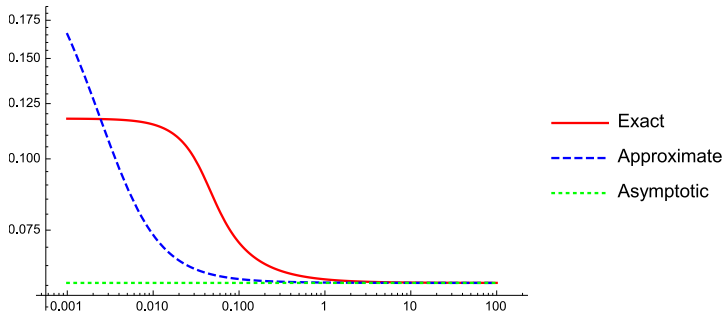


Figure 6.5: Mean squared error vs. λ

access to stochastic gradients of $\nabla_i \mathscr{L}^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n)$ of the form

$$\nabla_i \ell^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n) = \nabla_i \mathscr{L}^n(\boldsymbol{x}_1^n, \boldsymbol{x}_2^n) + g_i$$

with $g_i \sim \mathscr{N}(0, \sigma_e^2)$.

Figure 6.6 plots the estimate of ρ. We do not know what the true value of ρ is, but our estimate settles down. Figure 6.7 plots the choice of numbers of rounds of the game, K_n, selected by our method. Our choice of K_n settles down too. Figure 6.8 plots the Nash equilibrium quality. We achieve close to our target of $\varepsilon = 0.2$. Figure 6.9 plots the estimation loss achieved by the approach in this chapter, an approach where only the fusion center modifies its strategy, and an approach where only the adversary modifies its strategy. Note that we plot the exact MSE estimation loss and not the approximate loss. If the fusion center stops playing, then the adversary can substantially increase the MSE. If the adversary stops playing, then the fusion center can substantially lower the MSE. Our game approach performs in between these two approaches.

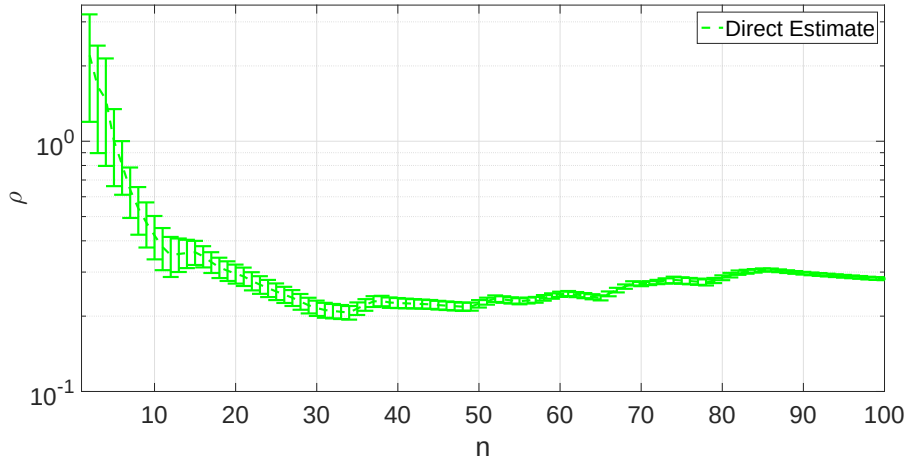


Figure 6.6: Surveillance game ρ

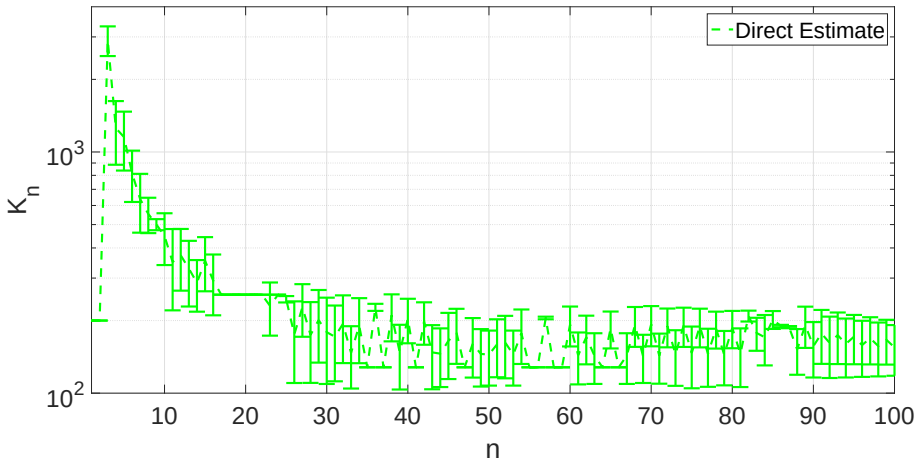


Figure 6.7: Surveillance game ρ

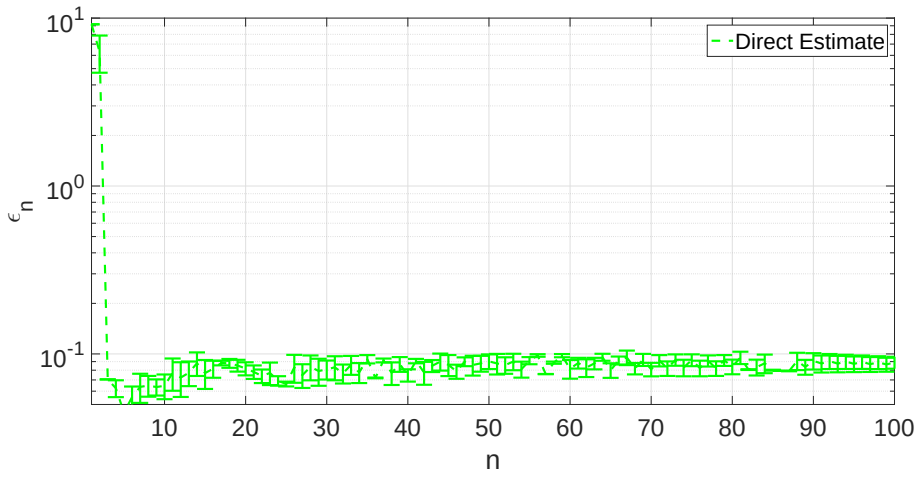


Figure 6.8: Surveillance game Nash equilibrium quality

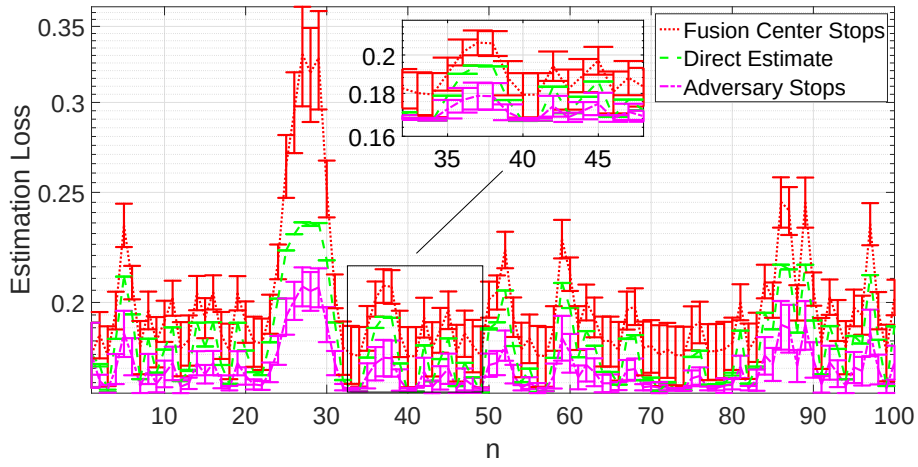


Figure 6.9: Surveillance game loss

Chapter 7

Conclusion

We have introduced general techniques to solve a sequence of time varying optimization problems and to find solutions of repeated games with time-varying stage games. We have introduced data dependent ways to select either the number of samples for the optimization problem or the number of rounds for the game problem to find approximate solutions under a variety of different models for how the problems change. We have provided theoretical guarantees on the accuracy of our approximate solutions. Tests involving both synthetic and real data have demonstrated the efficacy of our approaches in this thesis.

7.1 Future Work

In this section, we outline some broad future research directions.

7.1.1 Model Free Approach

We look at an approach to controlling the mean criterion when we do not know $b(d_0, K_n)$. Since we do not know $b(d_0, K)$, we will not estimate the change in the minimizers, ρ, in contrast to the approach of Chapters 3, 4, and 5. Suppose that Assumptions A.1-A.6 and the following assumptions hold:

1. The functions $f_n(\boldsymbol{x})$ have uniformly Lipschitz continuous gradients in $\boldsymbol{x}$ with modulus M, i.e.,

$$\mathbb{E}_{\boldsymbol{z}_n \sim p_n} \| \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z}_n) - \boldsymbol{g}(\tilde{\boldsymbol{x}}, \boldsymbol{z}_n) \| \leq M \| \boldsymbol{x} - \tilde{\boldsymbol{x}} \|$$

2. The minimizers satisfy $\boldsymbol{x}_n^* \in \text{int}(\mathscr{X})$ for all $n \geq 1$.

As an aside, we first consider the problem of variable structure control (VSC) from control theory, which has connections to our problem [67].

Variable Structure Control Consider controlling a discrete-time linear system

$$\boldsymbol{x}_{n+1} = A\boldsymbol{x}_n + B\boldsymbol{u}_n + \boldsymbol{d}_n \tag{7.1}$$

where $\boldsymbol{x}_n$ is the state, $\boldsymbol{u}_n$ is the control input, and $\boldsymbol{d}_n$ is a disturbance signal. We now examine a special type of control law relevant to our work. From [67], the design of a sliding mode controller consists of the following two steps:

1. Find a switching function $s(\boldsymbol{x})$ such that the sliding mode $s(\boldsymbol{x}_n) = 0$ is stable
2. Determine a control law

$$\boldsymbol{u}(\boldsymbol{x}) = \begin{cases} \boldsymbol{u}^+(\boldsymbol{x}), & s(\boldsymbol{x}) > 0 \\ \boldsymbol{u}^-(\boldsymbol{x}), & s(\boldsymbol{x}) < 0 \end{cases}$$

   with $\boldsymbol{u}_n = \boldsymbol{u}(\boldsymbol{x}_n)$

A sliding mode controller tries to drive the system to $s(\boldsymbol{x}) = 0$ by switching between two control laws, $\boldsymbol{u}^+(\boldsymbol{x})$ and $\boldsymbol{u}^-(\boldsymbol{x})$, depending on whether the system is above the sliding plane, $s(\boldsymbol{x}) > 0$, or below, $s(\boldsymbol{x}) < 0$. In general for a discrete-time system, it is impossible to achieve $s(\boldsymbol{x}_n) = 0$, since crossings of the sliding mode need not occur at an exact discrete time instant. Instead, the best we can hope for is that the state $\boldsymbol{x}_n$ remains close to the sliding mode in the sense that $|s(\boldsymbol{x}_n)|$ is small (known as a quasi-sliding mode band (QSMB)). For linear and non-linear sliding mode controllers, there are some results and theoretical guarantees for sliding mode controllers [67, 68].

For our problem, we can view the state of discrete time system as the approximate minimizer $\boldsymbol{x}_n$

$$\boldsymbol{x}_n = \mathscr{A}\left(\boldsymbol{x}_{n-1}, \{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}\right)$$

where $\mathscr{A}$ captures the behavior of our optimization algorithm. In this section, we treat $\mathscr{A}\left(\boldsymbol{x}_{n-1}, \{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}\right)$ as a black box that is completely unknown to us. As a consequence of the descent lemma [34], we have

$$f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) \leq \frac{1}{2}M\|\boldsymbol{x}_n - \boldsymbol{x}_n^*\|^2$$

As a consequence of strong convexity, we have

$$\|\boldsymbol{x}_n - \boldsymbol{x}_n^*\| \leq \frac{1}{m}\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x}_n)\|$$

This in turn shows that

$$\begin{aligned} f_n(\boldsymbol{x}_n) - f_n(\boldsymbol{x}_n^*) &\leq \frac{1}{2}M\left(\frac{1}{m}\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x}_n)\|\right)^2 \\ &\leq \frac{M}{2m^2}\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x}_n)\|^2 \end{aligned}$$

and so

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \frac{M}{2m^2}\mathbb{E}\left[\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x}_n)\|^2\right]$$

Therefore, if we have

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x}_n)\|^2\right] \leq \frac{2m^2\varepsilon}{M} \triangleq \delta$$

then it holds that

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n) \leq \varepsilon$$

This suggests that we choose the switching function

$$s_n(\boldsymbol{x}) = \|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})\|^2$$

Our goal is to find a QSMB such that

$$s_n(\boldsymbol{x}_n) \leq \delta$$

almost surely. This will in turn imply that

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon$$

as desired.

With K_1 fixed, we consider using the controller

$$K_n = \max\left\{\begin{cases} 1, & \|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})\|^2 > \delta \\ -1, & \|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})\|^2 < \delta \end{cases}, 1\right\}$$

If we are above our target δ for $\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})\|^2$, then we increase the number of samples by one. If we are below our target δ for $\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})\|^2$, then we decrease the number of samples by one.

In practice, we do not know $\|\nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})\|^2$, so we must approximate this quantity.

Define

$$\hat{G}_n \triangleq \frac{1}{K_n}\sum_{k=1}^{K_n}\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}_n,\boldsymbol{z}_n(k))$$

Suppose we start out with a fixed K_1. We choose K_n from K_{n-1} by setting

$$\Delta K_n = \begin{cases} +1, & \hat{G}_{n-1}^2 > \delta \\ -1, & \hat{G}_{n-1}^2 \leq \delta \end{cases}$$

and for $n \geq 2$

$$\begin{aligned} K_n &= K_{n-1} + \Delta K_n \\ &= K_1 + \sum_{i=2}^{n} \Delta K_i \end{aligned}$$

We have no theoretical guarantees for this approach, but this is of interest as it is computationally simple. Future work would consist of studying properties of this approach and trying to provide theoretical guarantees.

We consider a simple test of this method on the synthetic regression problem from Section 4.3.1 of Chapter 4. Figure 7.1 shows the number of samples selected using our method against the number of samples selected by the model free method. This method eventually settles down and select K_n close to what the methods of this thesis pick. Figure 7.2 shows the loss achieved by the direct estimate method against the model free approach of this section.

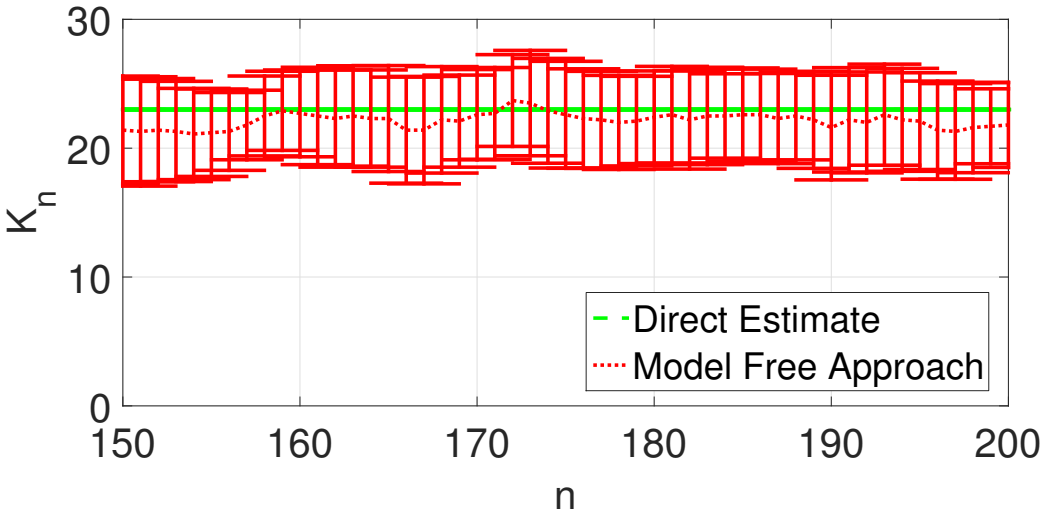


Figure 7.1: Model free K_n

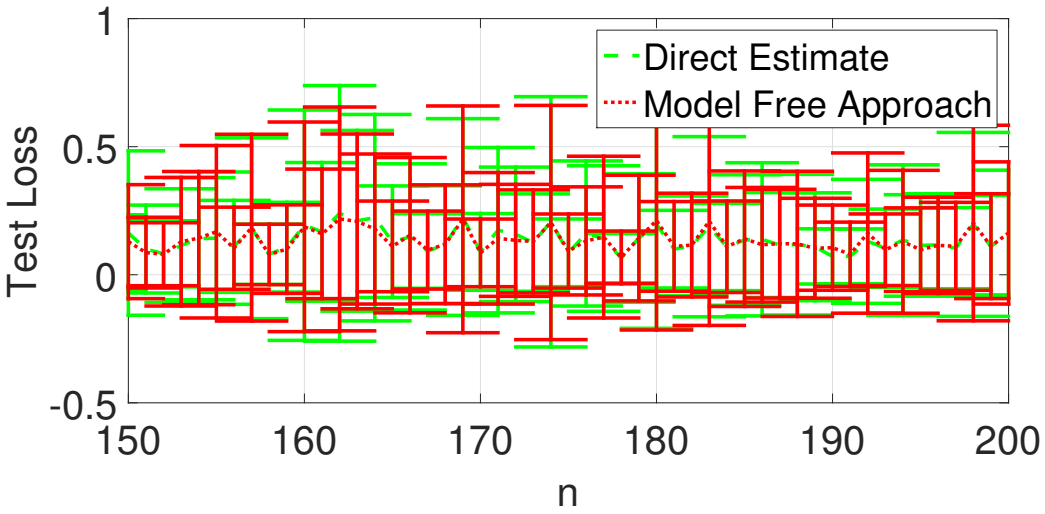


Figure 7.2: Model free test loss

7.1.2 Extension of the Game Problem

We consider several extensions of the game problem in Chapter 6 to non-zero-sum games. Consider a P-person continuous game where player p has utility $u_p^n(\boldsymbol{x})$ for the n^{th} stage game. The strategy space for player p is $\mathscr{X}_p \subset \mathbb{R}^{d_p}$. We impose various convexity/concavity conditions on the utilities. As before, we want to find a pure strategy ε-Nash equilibrium for each stage game n. A pure strategy ε-Nash equilibrium is a pure strategy $\bar{\boldsymbol{x}}$ such that

$$u_p^n(\boldsymbol{x}_p, \bar{\boldsymbol{x}}_{-p}) - u_p^n(\bar{\boldsymbol{x}}) \leq \varepsilon \quad \forall \boldsymbol{x}_p, \forall p$$

This means that any player can only unilaterally increase its utility by ε. With $\varepsilon = 0$, an ε-PNE is just a PNE. We focus on ε-Nash equilibria with $\varepsilon > 0$, since there are polynomial time algorithms to find such equilibria for games with continuous strategy spaces. The problem of finding a Nash equilibrium, $\varepsilon = 0$, is more challenging and no general, time efficient algorithm is known [4]. We focus on pure strategies in this section due to the difficulty of finding mixed equilibrium for continuous games. In the most general case, we need to deal with distributions over the entire continuous strategy space, which is challenging. For some continuous games with utilities that can be expressed as the sum of polynomials, the mixed strategy Nash equilibria are supported on a finite subset of the strategy space [69]. For such games, it may be possible to efficiently find a mixed strategy ε-Nash equilibrium; however, we do not consider this problem in this section.

Diagonally Strict Concave Utilities First, suppose that the utilities $u_p(\boldsymbol{x}_p, \boldsymbol{x}_{-p})$ are concave in the action $\boldsymbol{x}_p$ of player p. Note that certain forms of Cournot competitions are concave games [58]. Then it holds that there exists a pure strategy Nash equilibrium [58]. However, the pure strategy Nash equilibrium need not be unique. As an example, consider the two player game with utilities

$$\begin{aligned} u_1(\boldsymbol{x}_1, \boldsymbol{x}_2) &= x_1 x_2 - \frac{1}{2} x_1^2 \\ u_2(\boldsymbol{x}_1, \boldsymbol{x}_2) &= x_1 x_2 - \frac{1}{2} x_2^2 \end{aligned}$$

and $\mathscr{X}_1 = \mathscr{X}_2 = \mathscr{X}$. The best responses are given by

$$\begin{aligned} \mathrm{BR}_1(x_2) &= x_2 \\ \mathrm{BR}_2(x_1) &= x_1 \end{aligned}$$

where $\mathrm{BR}_i(x_{-i})$ is the best response of player i to the actions of all the other players, i.e.,

$$\mathrm{BR}_i(x_{-i}) = \arg\max_{\boldsymbol{x}_i \in \mathscr{X}_i} u_i\left(\boldsymbol{x}_i, \boldsymbol{x}_{-i}\right)$$

Thus, it follows that any pair $(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}, \boldsymbol{x})$ is a Nash equilibrium.

We want to focus on the case when our game admits a unique PNE, so that it is straightforward to discuss how the PNE is changing. If the PNE are not unique, then defining the change in stage games is more difficult. From the previous example, we see that concavity in the players' strategy is not enough to guarantee a unique PNE. To guarantee a unique pure strategy Nash equilibrium, we need stronger conditions on the utilities. The utilities are said to be *diagonally strictly concave* if there exist constants $r_p > 0$ such that the pseudo-gradient map

$$g(\boldsymbol{x}, \boldsymbol{r}) = \begin{bmatrix} r_1 \nabla_1 u_1(\boldsymbol{x}) \\ \vdots \\ r_P \nabla_P u_P(\boldsymbol{x}) \end{bmatrix}$$

satisfies

$$(\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top g(\boldsymbol{x}, \boldsymbol{r}) + (\boldsymbol{x} - \tilde{\boldsymbol{x}})^\top g(\tilde{\boldsymbol{x}}, \boldsymbol{r}) > 0 \quad \forall \boldsymbol{x}, \tilde{\boldsymbol{x}}$$

If the utilities are diagonally strictly concave, then the pure strategy Nash equilibrium is unique [58]. In order to check for diagonally strict concave utilities, it is sufficient that $G(\boldsymbol{x}, \boldsymbol{r}) + G(\boldsymbol{x}, \boldsymbol{r})^\top$ is negative definite with $G(\boldsymbol{x}, \boldsymbol{r})$ the Jacobian

associated with the mapping $g(\boldsymbol{x}, \boldsymbol{r})$. Again, certain forms of Cournot competitions have diagonally strictly concave utilities.

For a fixed stage game, the analysis in [58] shows that if all players use gradient ascent

$$\boldsymbol{x}_p(k) = \Pi_{\mathscr{X}} \left[\boldsymbol{x}_p(k-1) + \mu(k) \nabla_p u_p(\boldsymbol{x}(k)) \right] \qquad k = 1, \ldots, K$$

then the player's strategies converge to the unique pure strategy Nash equilibrium $\bar{\boldsymbol{x}}$. To compute the gradient ascent iteration, all players need to know their own utilities plus the pure strategy played by the other players at the previous time instant. The analysis in [58] only proved that gradient ascent converges to the pure strategy Nash equilibrium $\bar{\boldsymbol{x}}$ but does not provide rates. We can look at extending the analysis in [58] to our time varying stage game problem.

Socially Concave Utilities Next, we consider a variant of concave games with fewer restrictions on the learning dynamics. A socially concave game is a game for which there exists positive constants $\{r_p\}_{p=1}^P$ such that

$$\sum_{p=1}^{P} r_p u_p(\boldsymbol{x})$$

is concave and $u_p(\boldsymbol{x}_p, \boldsymbol{x}_{-p})$ is convex in $\boldsymbol{x}_{-p}$ for each p [59]. If the utilities are twice differentiable, then the game is also a concave game. Certain Cournot competitions and resource allocation games are socially concave games.

The following theorem captures several nice properties of learning in socially concave games from [59]:

Theorem 7. *If all players play strategies that achieve bounded external regret, i.e.,*

$$\max_{\boldsymbol{x}_p} \sum_{k=1}^{K} \left(u_p(\boldsymbol{x}_p, \boldsymbol{x}_{-p}(k)) - u_p(\boldsymbol{x}(k)) \right) \leq R_p(K),$$

then it holds that

1. *The average strategy* $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}(k)$ *is a* $\left(\frac{1}{r_{min}} \sum_{p=1}^{P} \frac{r_p R_p(K)}{K} \right)$*-Nash equilibrium with* $r_{min} = \min_i r_i$
2. $\left| u_p(\boldsymbol{x}(K)) - u_p\left(\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}(k) \right) \right| \leq \frac{1}{r_p} \sum_{p=1}^{P} \frac{r_p R_p(K)}{K}$

This shows that the average strategy is an ε-Nash equilibrium, and the utility at time K is close to the utility of the ε-Nash equilibrium. Any method that achieves low regret in the sense that $R_p(K) = o(K)$ is sufficient. Applying gradient ascent achieves $O(\sqrt{K})$ regret if the utilities are concave [13] or $O(\log K)$ regret if the utilities are strongly concave [7]. Gradient ascent thus meets the requirements of a socially concave game but any low regret method can be used.

Future work consists of extending Theorem 7 to the time varying problem to track the PNE.

Folk Theorem Extension Thus far, we have looked at tracking the equilibria of the stage games that constitute the n^{th} repeated game. However, the repeated game formed from the stage game may have Nash equilibria not present in the stage game. A classic example is the prisoner's dilemma game. In the stage game, the unique Nash equilibrium is for both players to defect. Consider the following strategy know as tit for tat: a player initially cooperates and then repeats the action of the other player for every time instant afterwards. Both players using tit for tat constitutes a Nash equilibrium of the repeated game and results in both players cooperating [1].

In the field of repeated games, there are results known as folk theorems that characterize the type of new Nash equilibria that can appear in repeated games that are not present in the stage game [70]. This suggests that as our stage games change slowly over time, we need to examine Nash equilibria of the overall repeated game and not just the stage game. For the simple case in which the stage game does not change there has been a little work on finding equilibria of the repeated game [71].

Appendix A

Examples of $b(d_0, K)$ Bound for SGD

We examine bounds $b(d_0, K)$ satisfying assumption A.4 for SGD (2.4). We form a convex combination of the iterates to yield a final approximate minimizer

$$\bar{\boldsymbol{x}}(K) \quad = \quad \sum_{k=0}^{K} \lambda(k)\boldsymbol{x}(k)$$

Note that this includes the case where $\bar{\boldsymbol{x}}(K) = \boldsymbol{x}(K)$ by selecting $\lambda(K) = 1$ and $\lambda(0) = \cdots = \lambda(K-1) = 0$. We consider the following bounds:

1. Lipschitz gradient bound with $\bar{\boldsymbol{x}} = \boldsymbol{x}(K)$ (Lemma 13)

$$b(d_0, K) = \frac{1}{2}M \left(\prod_{k=1}^{K} (1 - 2m\mu(k) + B\mu^2(k)) d(0) \right.$$
$$\left. + A \sum_{k=1}^{K} \prod_{i=k+1}^{K} (1 - 2m\mu(i) + B\mu^2(i))\mu^2(k) \right)$$

2. Inverse step size averaging (Lemma 14)

$$b(d_0, K) = \frac{(1+B)d(0) + B\sum_{k=1}^{K} \gamma(k) + (K+1)A}{m(K+1)(K+4)}$$

3. Special quadratic bound with averaging (Lemma 15)

$$b(d_0, K) = \frac{M}{2} \left(\frac{1}{m^{1/2}} \sum_{k=1}^{K-1} \left| \frac{1}{\mu(k+1)} - \frac{1}{\mu(k)} \right| (\gamma(k))^{1/2} \right.$$
$$+ \frac{1}{m^{1/2}\mu(1)} (d(0))^{1/2} + \frac{1}{m^{1/2}\mu(K)} (\gamma(K))^{1/2}$$
$$\left. + \sqrt{\frac{A}{mK}} + \sqrt{\frac{2B}{mK^2} \sum_{k=1}^{K} \mathbb{E}[d(k-1)]} \right)^2$$

We now examine the proofs of these three bounds, which are mostly small extensions of previous results. Define

$$d(k) \triangleq \|\boldsymbol{x}(k) - \boldsymbol{x}^*\|^2 \tag{A.1}$$

First, we bound $\mathbb{E}[d(k)]$ in Lemma 12, which follows the classic Lyapunov function analysis of SGD [72].

Lemma 12. *It holds that*

$$\mathbb{E}[d(k)] \leq \prod_{k=1}^{K}(1 - 2m\mu(k) + B\mu^2(k))d(0) + A\sum_{k=1}^{K}\prod_{i=k+1}^{K}(1 - 2m\mu(i) + B\mu^2(i))\mu^2(k)$$

Proof. See [72]. □

It is possible to further upper bound the bound in Lemma 12 to yield a closed form given in [73]; however, the bound in Lemma 12 is generally tighter. Next, we apply Lemma 12 along with a Lipschitz gradient assumption on $f(\boldsymbol{x})$ to produce a simple $b(d_0, K)$ bound.

Lemma 13. *With arbitrary step sizes, assuming that $f(\boldsymbol{x})$ has Lipschitz continuous gradients with modulus M, and $\lambda(K) = 1$, it holds that*

$$\mathbb{E}[f(\bar{\boldsymbol{x}}(K))] - f(\boldsymbol{x}^*) \leq \frac{1}{2}M\mathbb{E}[d(K)]$$

and therefore, it holds that

$$b(d_0, K) = \frac{M}{2}\left(\prod_{k=1}^{K}(1 - 2m\mu(k) + B\mu^2(k))d(0) + A\sum_{k=1}^{K}\prod_{i=k+1}^{K}(1 - 2m\mu(i) + B\mu^2(i))\mu^2(k)\right)$$

satisfies the requirements of A.4.

Proof. Using the descent lemma from [34], it holds that $\mathbb{E}[f(\boldsymbol{x})] - f(\boldsymbol{x}^*) \leq \frac{1}{2}M\mathbb{E}[d(K)]$. Plugging in the bound from Lemma 12 yields the bound $b(d_0, K)$. □

Next, we consider an extension of the averaging scheme derived with $B = 0$ in [74] to the case with $B > 0$ using the bounds in Lemma 12. This averaging scheme puts weight

$$\lambda(k) = \frac{\frac{1}{\mu(k)}}{\sum_{j=1}^{K} \frac{1}{\mu(j)}}$$

on the iterate $\boldsymbol{x}(k)$ with step size $\mu(k) = \mathscr{O}(k^{-1})$. Therefore, this averaging puts increasing weight on later iterates.

Lemma 14. *With the choice of step sizes given by*

$$\mu(k) = \frac{1}{m(k+1)} \quad \forall k \geq 1$$

it holds that

$$b(d_0, K) = \frac{(1+B)d(0) + B\sum_{k=1}^{K} \gamma(k) + (K+1)A}{m(K+1)(K+4)}$$

satisfies assumption A.4 where $\mathbb{E}[d(k)] \leq \gamma(k)$.

Proof. This proof is a straightforward extension of the proof in [74]. We have using standard analysis of SGD (see [72] for example).

$$\begin{aligned}\mathbb{E}[d(k)] \leq (1 - 2m\mu(k) + B\mu^2(k))\mathbb{E}[d(k-1)] \\ - 2\mu(k)(\mathbb{E}[f(\boldsymbol{x}(k-1))] - f(\boldsymbol{x}^*)) + A\mu^2(k)\end{aligned}$$

Then dividing by $\mu^2(k)$, we have

$$\begin{aligned}\frac{1}{\mu^2(k)}\mathbb{E}[d(k)] \leq \left(\frac{1-2m\mu(k)}{\mu^2(k)} + B\right)\mathbb{E}[d(k-1)] \\ - \frac{2}{\mu(k)}(\mathbb{E}[f(\boldsymbol{x}(k-1))] - f(\boldsymbol{x}^*)) + A\end{aligned}$$

It holds that

$$\frac{1-2m\mu(k)}{\mu^2(k)} \leq \frac{1}{\mu^2(k-1)}$$

This implies that

$$\begin{aligned}&\frac{1}{\mu^2(k)}\mathbb{E}[d(k)]-\frac{1}{\mu^2(k-1)}\mathbb{E}[d(k-1)]\\&\qquad\leq B\mathbb{E}[d(k-1)]-\frac{2}{\mu(k)}(\mathbb{E}[f(\boldsymbol{x}(k-1))]-f(\boldsymbol{x}^*))+A\end{aligned}$$

Summing from $k=1$ to $K+1$ and rearranging yields

$$\begin{aligned}&\sum_{k=0}^{K}\frac{1}{\mu(k+1)}\left(\mathbb{E}[f(\boldsymbol{x}(k))]-f(\boldsymbol{x}^*)\right)\\&\qquad\leq\frac{1}{2}(1+B)d(0)+\frac{1}{2}B\sum_{k=1}^{K}\mathbb{E}[d(k)]+\frac{1}{2}(K+1)A\end{aligned}$$

With the weights

$$\lambda(k)=\frac{\frac{1}{\mu(k+1)}}{\sum_{\tau=0}^{K}\frac{1}{\mu(j+1)}}$$

we have

$$\mathbb{E}[f(\bar{\boldsymbol{x}}(K))]-f(\boldsymbol{x}^*)\leq\frac{\frac{1}{2}(1+B)d(0)+\frac{1}{2}B\sum_{k=1}^{K}\mathbb{E}[d(k)]+\frac{1}{2}(K+1)A}{\sum_{\tau=0}^{K}\frac{1}{\mu(\tau)}}$$

Then it holds that

$$\sum_{\tau=0}^{K}\frac{1}{\mu(\tau+1)}=\sum_{\tau=0}^{K}m(\tau+2)=\frac{1}{2}m(K+1)(K+4)$$

so

$$\begin{aligned}&\mathbb{E}[f(\bar{\boldsymbol{x}}(K))]-f(\boldsymbol{x}^*)\\&\qquad\leq\frac{(1+B)d(0)+B\sum_{k=1}^{K}\mathbb{E}[d(k)]+(K+1)A}{m(K+1)(K+4)}\\&\qquad\leq\frac{(1+B)d(0)+B\sum_{k=1}^{K}\gamma(k)+(K+1)A}{m(K+1)(K+4)}\end{aligned}$$

□

To get the required $\gamma(k)$ bounds, we use Lemma 12. For the choice of step sizes

in Lemma 14 from Lemma 12, it holds that $\mathbb{E}[d(k)] = \mathscr{O}\left(\frac{1}{k}\right)$. Since

$$\sum_{k=1}^{K} \frac{1}{k} = \mathscr{O}(\log K)$$

it holds that

$$\mathbb{E}[f(\bar{\boldsymbol{x}}(K))] - f(\boldsymbol{x}^*) = \mathscr{O}\left(\frac{d(0)}{K^2} + \frac{\log(K)}{K^2} + \frac{1}{K}\right)$$

The $\mathscr{O}(\frac{1}{K})$ rate is minimax optimal for stochastic minimization of a strongly convex function [75].

Next, we look at a special case of averaging from [73] for stochastic gradients such that

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z}) - \boldsymbol{g}(\tilde{\boldsymbol{x}},\boldsymbol{z}) - \boldsymbol{g}^{(2)}(\boldsymbol{x},\boldsymbol{z})(\boldsymbol{x}-\tilde{\boldsymbol{x}})\|^2 = 0 \quad \forall \boldsymbol{x},\tilde{\boldsymbol{x}} \in \mathscr{X}$$

where $\boldsymbol{g}^{(2)}(\boldsymbol{x},\boldsymbol{z})$ is an unbiased stochastic second derivative with respect to $\boldsymbol{x}$. Quadratic objectives satisfy this condition.

Lemma 15. *Assuming that*

1. $\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z}) - \boldsymbol{g}(\tilde{\boldsymbol{x}},\boldsymbol{z}) - \boldsymbol{g}^{(2)}(\boldsymbol{x},\boldsymbol{z})(\boldsymbol{x}-\tilde{\boldsymbol{x}})\|^2 = 0 \quad \forall \boldsymbol{x},\tilde{\boldsymbol{x}} \in \mathscr{X}$
2. $\mu(k) = Ck^{-\alpha}$ *with* $\alpha \geq 1/2$
3. $\lambda(0) = 0$ *and* $\lambda(k) = 1/K$ *for* $1 \leq k \leq K$
4. $\mathbb{E}[d(k)] \leq \gamma(k)$

it holds that

$$\begin{aligned}\left(\mathbb{E}[\bar{d}(K)]\right)^{1/2} \leq{}& \frac{1}{m^{1/2}} \sum_{k=1}^{K-1} \left| \frac{1}{\mu(k+1)} - \frac{1}{\mu(k)} \right| (\gamma(k))^{1/2} \\ &+ \frac{1}{m^{1/2}\mu(1)} (d(0))^{1/2} + \frac{1}{m^{1/2}\mu(K)} (\gamma(K))^{1/2} \\ &+ \sqrt{\frac{A}{mK}} + \sqrt{\frac{2B}{mK^2} \sum_{k=1}^{K} \mathbb{E}[d(k-1)]}\end{aligned}$$

with $\bar{d}(K) = \|\bar{\boldsymbol{x}}(K) - \boldsymbol{x}^*\|^2$. *If* f *has Lipschitz continuous gradients with modulus*

M, then it holds that

$$b(d_0,K) = \frac{M}{2}\left(\frac{1}{m^{1/2}}\sum_{k=1}^{K-1}\left|\frac{1}{\mu(k+1)}-\frac{1}{\mu(k)}\right|(\gamma(k))^{1/2} + \frac{1}{m^{1/2}\mu(1)}(d(0))^{1/2} + \frac{1}{m^{1/2}\mu(K)}(\gamma(K))^{1/2} + \sqrt{\frac{A}{mK}} + \sqrt{\frac{2B}{mK^2}\sum_{k=1}^{K}\mathbb{E}[d(k-1)]}\right)^2$$

5 satisfies Assumption A.4.

Proof. See [73] for the proof. □

This decays at rate $\mathscr{O}\left(\frac{1}{K}\right)$ as long as $\mu(k) = Ck^{-\alpha}$ with $\frac{1}{2} \leq \alpha \leq 1$. To get the bounds $\gamma(k)$, we can again apply Lemma 12.

Appendix B

ρ Estimation Proofs

For our analysis of minimizer change estimation in this appendix and parameter estimation in Appendix D, we need to introduce a few results for sub-Gaussian random variables including the following key technical lemma from [30]. This lemma controls the concentration of sums of random variables that are sub-Gaussian conditioned on a particular filtration $\{\mathscr{F}_i\}_{i=0}^n$. Such a collection of random variables is referred to as a *sub-Gaussian Martingale sequence*.

Lemma 16 (Theorem 7.5 of [30])**.** *Suppose we have a collection of random variables $\{V_i\}_{i=1}^n$ and a filtration $\{\mathscr{F}_i\}_{i=0}^n$ such that for each random variable V_i it holds that*

1. $\mathbb{E}\left[e^{s\left(V_i-\mathbb{E}\left[V_i \,\middle|\, \mathscr{F}_{i-1}\right]\right)} \,\middle|\, \mathscr{F}_{i-1}\right] \leq e^{\frac{1}{2}\sigma_i^2 s^2}$ *with σ_i^2 a constant*
2. *V_i is $\mathscr{F}_i$-measurable*

Then for every $\boldsymbol{a} \in \mathbb{R}^n$ it holds that

$$\mathbb{P}\left\{\sum_{i=1}^n a_i V_i > \sum_{i=1}^n a_i \mathbb{E}\left[V_i \,\middle|\, \mathscr{F}_{i-1}\right] + t\right\} \leq \exp\left\{-\frac{t^2}{2\nu}\right\}$$

with $\nu = \sum_{i=1}^n \sigma_i^2 a_i^2$. The other tail is similar.

Remark: If we can upper bound the conditional expectations $\mathbb{E}\left[V_i \,\middle|\, \mathscr{F}_{i-1}\right] \leq C_i$ by $\mathscr{F}_{i-1}$-measurable random variables C_i, then we have

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^n a_i V_i > \sum_{i=1}^n a_i C_i + t\right\} &\leq \mathbb{P}\left\{\sum_{i=1}^n a_i V_i > \sum_{i=1}^n a_i \mathbb{E}\left[V_i \,\middle|\, \mathscr{F}_{i-1}\right] + t\right\} \\ &\leq \exp\left\{-\frac{t^2}{2\nu}\right\}\end{aligned}$$

For our analysis, we generally cannot compute $\mathbb{E}\left[V_i \,\middle|\, \mathscr{F}_{i-1}\right]$, but we can find "nice" C_i.

To find σ_i^2 for use in Lemma 16, we employ the following conditional version of Hoeffding's lemma.

Lemma 17 (Conditional Hoeffding's Lemma)**.** *If a random variable V and a σ-algebra $\mathscr{F}$ satisfy $a \leq V \leq b$ and $\mathbb{E}[V|\mathscr{F}] = 0$, then*

$$\mathbb{E}\left[e^{sV} \mid \mathscr{F}\right] \leq \exp\left\{\frac{1}{8}(b-a)^2 s^2\right\}$$

Proof. Follows from the standard proof of Hoeffding's Lemma from [76]. □

Using these concentration tools, we can analyze averages of the direct estimate.

B.1 Euclidean Norm Condition

As a reminder, we consider running our optimization algorithm used to generate $\boldsymbol{x}_i$ again using independent samples $\{\tilde{\boldsymbol{z}}_i(k)\}_{k=1}^{K_i}$ to yield a second approximate minimizer $\tilde{\boldsymbol{x}}_i$. For SGD, the process to generate $\tilde{x}_i$ is summarized in (3.7). We connect $\tilde{\rho}_i$ to $\tilde{\rho}_i^{(2)}$ with $\tilde{\rho}_i^{(2)}$ defined in (3.8).

Proof of Theorem 1: To proceed, we compare the three single step estimates:

1. $\tilde{\rho}_i = \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2 + \frac{1}{m}\|G_i\|_2 + \frac{1}{m}\|G_{i-1}\|_2$
2. $\tilde{\rho}_i^{(2)} = \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_{i-1}\|_2 + \frac{1}{m}\|\tilde{G}_i\|_2 + \frac{1}{m}\|\tilde{G}_{i-1}\|_2$
3. $\tilde{\rho}_i^{(3)} = \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_{i-1}\|_2 + \frac{1}{m}\|\nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 + \frac{1}{m}\|\nabla f_{i-1}(\tilde{\boldsymbol{x}}_{i-1})\|_2$

where

$$\hat{G}_i = \frac{1}{K_i}\sum_{k=1}^{K_i} \nabla_{\boldsymbol{x}} \ell(\boldsymbol{x}_i, \boldsymbol{z}_i(k))$$

and

$$\tilde{G}_i = \frac{1}{K_i}\sum_{k=1}^{K_i} \nabla_{\boldsymbol{x}} \ell(\tilde{\boldsymbol{x}}_i, \boldsymbol{z}_i(k))$$

Define $\hat{\rho}_n^{(2)}$ and $\hat{\rho}_n^{(3)}$ analogously to $\hat{\rho}_n$ as an average of the corresponding single step estimates. Using the triangle inequality and the reverse triangle inequality, it

holds that

$$
\begin{aligned}
&|\hat{\rho}_n - \hat{\rho}_n^{(3)}| \\
&\quad = |\hat{\rho}_n - \hat{\rho}_n^{(2)} + \hat{\rho}_n^{(2)} - \hat{\rho}_n^{(3)}| \\
&\quad \leq |\hat{\rho}_n - \hat{\rho}_n^{(2)}| + |\hat{\rho}_n^{(2)} - \hat{\rho}_n^{(3)}| \\
&\quad \leq \frac{1}{n-1} \sum_{i=2}^{n} \left(\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2 + \|\boldsymbol{x}_{i-1} - \tilde{\boldsymbol{x}}_{i-1}\|_2 + \frac{1}{m} \|\hat{G}_i - \tilde{G}_i\|_2 \right. \\
&\qquad\qquad \left. + \frac{1}{m} \|\hat{G}_{i-1} - \tilde{G}_{i-1}\|_2 \right) \\
&\qquad + \frac{1}{n-1} \sum_{i=2}^{n} \left(\frac{1}{m} \|\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 + \frac{1}{m} \|\tilde{G}_{i-1} - \nabla f_{i-1}(\tilde{\boldsymbol{x}}_{i-1})\|_2 \right) \\
&\quad \leq \frac{1}{n-1} \left(\|\boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1\|_2 + 2 \sum_{i=2}^{n-1} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2 + \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|_2 \right) \\
&\qquad + \frac{1}{m(n-1)} \left(\|\hat{G}_1 - \tilde{G}_1\|_2 + 2 \sum_{i=2}^{n-1} \|\hat{G}_i - \tilde{G}_i\|_2 + \|\hat{G}_n - \tilde{G}_n\|_2 \right) \\
&\qquad + \frac{1}{m(n-1)} \left(\|\tilde{G}_1 - \nabla f_1(\tilde{\boldsymbol{x}}_1)\|_2 + 2 \sum_{i=2}^{n-1} \|\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 \right. \\
&\qquad\qquad \left. + \|\tilde{G}_n - \nabla f_n(\tilde{\boldsymbol{x}}_n)\|_2 \right)
\end{aligned}
$$

Define

$$
U_n = \frac{1}{n-1} \left(\|\boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1\|_2 + 2 \sum_{i=2}^{n-1} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2 + \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|_2 \right)
$$

and

$$
V_n = \frac{1}{m(n-1)} \left(\|\hat{G}_1 - \tilde{G}_1\|_2 + 2 \sum_{i=2}^{n-1} \|\hat{G}_i - \tilde{G}_i\|_2 + \|\hat{G}_n - \tilde{G}_n\|_2 \right)
$$

and

$$
\begin{aligned}
W_n = \frac{1}{m(n-1)} \left(\|\tilde{G}_1 - \nabla f_1(\tilde{\boldsymbol{x}}_1)\|_2 + 2 \sum_{i=2}^{n-1} \|\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 \right. \\
\left. + \|\tilde{G}_n - \nabla f_n(\tilde{\boldsymbol{x}}_n)\|_2 \right)
\end{aligned}
$$

Then it holds that

$$
|\hat{\rho}_n - \hat{\rho}_n^{(3)}| \leq U_n + V_n + W_n
$$

We will apply Lemma 16 to each of U_n, V_n, and W_n. Define the following sums of conditional expectations closely related to U_n, V_n, and W_n:

$$\begin{aligned}
\breve{U}_n &= \frac{1}{n-1}\Bigg(\mathbb{E}\left[\|\boldsymbol{x}_1-\tilde{\boldsymbol{x}}_1\|_2 \mid \mathscr{F}_0\right]+2\sum_{i=2}^{n-1}\mathbb{E}\left[\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\qquad +\mathbb{E}\left[\|\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n\|_2 \mid \mathscr{F}_{n-1}\right]\Bigg) \\
\breve{V}_n &= \frac{1}{m(n-1)}\Bigg(\mathbb{E}\left[\|\hat{G}_1-\tilde{G}_1\|_2 \mid \mathscr{F}_0\right]+2\sum_{i=2}^{n-1}\mathbb{E}\left[\|\hat{G}_i-\tilde{G}_i\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\qquad +\mathbb{E}\left[\|\hat{G}_n-\tilde{G}_n\|_2 \mid \mathscr{F}_{n-1}\right]\Bigg) \\
\breve{W}_n &= \frac{1}{m(n-1)}\Bigg(\mathbb{E}\left[\|\tilde{G}_1-\nabla f_1(\tilde{\boldsymbol{x}}_1)\|_2 \mid \mathscr{F}_0\right]+2\sum_{i=2}^{n-1}\mathbb{E}\left[\|\tilde{G}_i-\nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\qquad +\mathbb{E}\left[\|\tilde{G}_n-\nabla f_n(\tilde{\boldsymbol{x}}_n)\|_2 \mid \mathscr{F}_{n-1}\right]\Bigg)
\end{aligned}$$

Bounding these sums of conditional expectations will be useful in terms of the remark after Lemma 16.

Next, we look at bounding $\mathbb{E}[\|\boldsymbol{x}_i-\boldsymbol{x}_{i-1}\|_2 \mid \mathscr{F}_{i-1}]$, $\mathbb{E}[\|\hat{G}_i-\tilde{G}_i\|_2 \mid \mathscr{F}_{i-1}]$, and $\mathbb{E}[\|\tilde{G}_i-\nabla f_i(\tilde{\boldsymbol{x}})_i\|_2 \mid \mathscr{F}_{i-1}]$. First, by assumption B.1, it holds that

$$\mathbb{E}[\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|_2 \mid \mathscr{F}_{i-1}] \leq C_i$$

Second, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{G}_i-\tilde{G}_i\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\quad \leq \frac{1}{K_i}\sum_{k=1}^{K_i}\mathbb{E}\left[\|\boldsymbol{g}_i(\boldsymbol{x}_i,\boldsymbol{z}_i(k))-\boldsymbol{g}_i(\tilde{\boldsymbol{x}}_i,\boldsymbol{z}_i(k))\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\quad = \mathbb{E}\left[\|\boldsymbol{g}_i(\boldsymbol{x}_i,\boldsymbol{z}_i(1))-\boldsymbol{g}_i(\tilde{\boldsymbol{x}}_i,\boldsymbol{z}_i(1))\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\quad \leq M\mathbb{E}\left[\|\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\quad \leq MC_i
\end{aligned}$$

Third, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 \mid \mathscr{F}_{i-1}\right] \\
&\quad \leq \left(\mathbb{E}\left[\left\|\frac{1}{K_i}\sum_{k=1}^{K_i}\left(\boldsymbol{g}_i\left(\tilde{\boldsymbol{x}}_i, \boldsymbol{z}_i(k)\right) - \nabla f_i(\tilde{\boldsymbol{x}}_i)\right)\right\|_2^2 \,\middle|\, \mathscr{F}_{i-1}\right]\right)^{1/2} \\
&\quad \leq \left(\mathbb{E}\left[\frac{1}{K_i^2}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i\left(\tilde{\boldsymbol{x}}_i, \boldsymbol{z}_i(k)\right) - \nabla f_i(\tilde{\boldsymbol{x}}_i)\right\|_2^2 \,\middle|\, \mathscr{F}_{i-1}\right]\right)^{1/2} \\
&\quad \leq \left(\frac{\sigma}{K_i}\right)^{1/2}
\end{aligned}$$

We can now produce bounds on $\breve{U}_n$, $\breve{V}_n$, and $\breve{W}_n$ denoted $\bar{U}_n$, $\bar{V}_n$, and $\bar{W}_n$ as follows:

1. $\bar{U}_n = \frac{1}{n-1}\left(C_1 + 2\sum_{i=2}^{n-1} C_i + C_n\right)$
2. $\bar{V}_n = \frac{M}{m(n-1)}\left(C_1 + 2\sum_{i=2}^{n-1} C_i + C_n\right)$
3. $\bar{W}_n = \frac{1}{m(n-1)}\left(\left(\frac{\sigma}{K_1}\right)^{1/2} + 2\sum_{i=2}^{n-1}\left(\frac{\sigma}{K_1}\right)^{1/2} + \left(\frac{\sigma}{K_n}\right)^{1/2}\right)$

Then it holds that

$$\begin{aligned}
&\mathbb{P}\left\{|\hat{\rho}_n - \hat{\rho}_n^{(3)}| > (\breve{U}_n + \breve{V}_n + \breve{W}_n) + t_n\right\} \\
&\quad \leq \mathbb{P}\left\{U_n + V_n + W_n > (\bar{U}_n + \bar{V}_n + \bar{W}_n) + t_n\right\} \\
&\quad \leq \mathbb{P}\left\{U_n > \bar{U}_n + \frac{1}{3}t_n\right\} + \mathbb{P}\left\{V_n > \bar{V}_n + \frac{1}{3}t_n\right\} \\
&\qquad\qquad + \mathbb{P}\left\{W_n > \bar{W}_n + \frac{1}{3}t_n\right\}
\end{aligned}$$

Now, we bound each of these three probabilities using Lemma 16. First, we have

$$0 \leq \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2 \leq \mathrm{diam}(\mathscr{X})$$

so applying Lemmas 17 and 16 with $\sigma_i^2 = \frac{1}{4}\mathrm{diam}^2(\mathscr{X})$ and

$$\begin{aligned}
a_1 &= a_n = \frac{1}{n-1} \\
a_2 &= \cdots = a_{n-2} = \frac{2}{n-1}
\end{aligned}$$

yields

$$\begin{aligned}
\nu_U &= \frac{1}{4}\text{diam}^2(\mathscr{X})\sum_{i=1}^{n} a_i^2 \\
&= \frac{1}{4}\text{diam}^2(\mathscr{X})\left(\frac{1}{n-1}\right)^2 + \sum_{i=2}^{n-1}\left(\frac{2}{n-1}\right)^2 + \left(\frac{1}{n-1}\right)^2 \\
&\leq \frac{1}{n-1}\text{diam}^2(\mathscr{X}) \qquad \text{(B.1)}
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
\mathbb{P}\left\{U_n > \bar{U}_n + \frac{1}{3}t_n\right\} &\leq \exp\left\{-\frac{(t_n/3)^2}{2\nu_U}\right\} \\
&= \exp\left\{-\frac{(n-1)t_n^2}{18\text{diam}^2(\mathscr{X})}\right\}
\end{aligned}$$

Since

$$0 \leq \|\hat{G}_i - \tilde{G}_i\|_2 \leq 2G$$

and

$$0 \leq \|\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{x}}_i)\|_2 \leq 2G$$

we can apply Lemmas 17 and 16 to V_n and W_n to yield

$$\mathbb{P}\left\{V_n > \bar{V}_n + \frac{1}{3}t_n\right\} \leq \exp\left\{-\frac{(t_n/3)^2}{2\nu_V}\right\} = \exp\left\{-\frac{m^2(n-1)t_n^2}{72G^2}\right\}$$

and

$$\mathbb{P}\left\{W_n > \bar{W}_n + \frac{1}{3}t_n\right\} \leq \exp\left\{-\frac{(t_n/3)^2}{2\nu_W}\right\} = \exp\left\{-\frac{m^2(n-1)t_n^2}{72G^2}\right\}$$

Define

$$D_n = \bar{U}_n + \bar{V}_n + \bar{W}_n$$

which is the definition in (3.9). It follows that

$$\begin{aligned}
&\mathbb{P}\left\{\hat{\rho}_n < \hat{\rho}_n^{(3)} - D_n - t_n\right\} \\
&\qquad \leq \exp\left\{-\frac{(n-1)t_n^2}{18\text{diam}^2(\mathscr{X})}\right\} + 2\exp\left\{-\frac{m^2(n-1)t_n^2}{72G^2}\right\} \qquad \text{(B.2)}
\end{aligned}$$

Then it follows that

$$\begin{aligned}
&\sum_{n=2}^{\infty} \mathbb{P}\left\{\hat{\rho}_n < \hat{\rho}_n^{(3)} - D_n - t_n\right\} \\
&\qquad \leq \sum_{n=2}^{\infty} \left(\exp\left\{ -\frac{(n-1)t_n^2}{18\mathrm{diam}^2(\mathscr{X})} \right\} \right. \\
&\qquad\qquad \left. +2\exp\left\{ -\frac{m^2(n-1)t_n^2}{72G^2} \right\} \right) \\
&\qquad < +\infty
\end{aligned}$$

Therefore, by the Borel-Cantelli lemma, for all n large enough it holds that

$$\hat{\rho}_n + D_n + t_n \geq \hat{\rho}_n^{(3)}$$

almost surely. Finally, by (3.4), it holds that $\hat{\rho}_n^{(3)} \geq \rho$, which proves the result. □

Looking at the form of D_i, it follows that in this case

$$D_n = \mathscr{O}\left(\frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{\sqrt{K_i}} \right)$$

In the case where $K_i = K^*$, this implies that

$$D_n = \mathscr{O}\left(\frac{1}{\sqrt{K^*}} \right)$$

We can also prove the result for the inequality constraint of (1.2) using similar techniques in Theorem 3.

Proof of Theorem 3: Define $\bar{\rho}_i^{(2)}$ and $\bar{\rho}_i^{(3)}$ analogous to the equality case proof and

the pair $\hat{\rho}_i^{(2)}$ and $\hat{\rho}_i^{(3)}$ of the form in (3.12). First, we have by assumptions B.4-B.5

$$\begin{aligned}
&|\hat{\rho}_n - \hat{\rho}_n^{(3)}| \\
&\quad \leq \frac{1}{n-W} \sum_{i=W+1}^{n} |\bar{\rho}^{(i)} - \bar{\rho}_3^{(i)}| \\
&\quad \leq \frac{1}{n-W} \sum_{i=W+1}^{n} \sum_{j=i-W+1}^{i} b_j |\tilde{\rho}_j - \tilde{\rho}_j^{(3)}| \\
&\quad \leq \frac{1}{n-W} \sum_{i=W+1}^{n} \sum_{j=i-W+1}^{i} b_j \left(|\tilde{\rho}_j - \tilde{\rho}_j^{(2)}| + |\tilde{\rho}_j^{(2)} - \tilde{\rho}_j^{(3)}| \right) \\
&\quad \leq \frac{\sum_{j=1}^{W} b_j}{n-W} \sum_{i=2}^{n} \left(|\tilde{\rho}_i - \tilde{\rho}_i^{(2)}| + |\tilde{\rho}_i^{(2)} - \tilde{\rho}_i^{(3)}| \right) \\
&\quad \leq \left(\frac{n-1}{n-W} \sum_{j=1}^{W} b_j \right) \frac{1}{n-1} \sum_{i=2}^{n} \left(|\tilde{\rho}_i - \tilde{\rho}_i^{(2)}| + |\tilde{\rho}_i^{(2)} - \tilde{\rho}_i^{(3)}| \right)
\end{aligned}$$

In comparison, we looked at controlling

$$\frac{1}{n-1} \sum_{i=2}^{n} \left(|\tilde{\rho}_i - \tilde{\rho}_i^{(2)}| + |\tilde{\rho}_i^{(2)} - \tilde{\rho}_i^{(3)}| \right)$$

in the proof of Theorem 1. The quantity of interest here is the same scaled by

$$\frac{n-1}{n-W} \sum_{j=1}^{W} b_j$$

We know that this term is finite, so we can argue that $\hat{\rho}_n$ upper bounds $\hat{\rho}_n^{(3)}$ for large enough n as in Theorem 1. The only remaining piece of the proof is to argue that $\hat{\rho}_n^{(3)}$ eventually upper bounds ρ.

By construction, we always have $\tilde{\rho}_i^{(3)} \geq \rho_i$. Therefore, by assumptions B.4-B.5, it follows that

$$\hat{h}_W(\tilde{\rho}_i^{(3)}, \tilde{\rho}_{i-1}^{(3)}, \ldots, \tilde{\rho}_{i-W+1}^{(3)}) \geq \hat{h}_W(\rho_i, \rho_{i-1}, \ldots, \rho_{i-W+1})$$

and

$$\begin{aligned}
\mathbb{E}\left[\hat{h}_W(\tilde{\rho}_i^{(3)}, \tilde{\rho}_{i-1}^{(3)}, \ldots, \tilde{\rho}_{i-W+1}^{(3)}) \mid \mathscr{F}_{i-W} \right] &\geq \mathbb{E}\left[\hat{h}_W(\rho_i, \rho_{i-1}, \ldots, \rho_{i-W+1}) \mid \mathscr{F}_{i-W} \right] \\
&\geq \rho \qquad \text{(B.3)}
\end{aligned}$$

We know that the random variable $\hat{h}_W(\tilde{\rho}_i^{(3)}, \tilde{\rho}_{i-1}^{(3)}, \ldots, \tilde{\rho}_{i-W+1}^{(3)})$ is $\mathscr{F}_i$ measurable and we can bound its expectation conditioned on $\mathscr{F}_{i-W}$. Under these conditions, we can apply a slight modification of Lemma 16 to show that for all n large enough

$$\hat{\rho}_n^{(3)} + t_n \geq \rho$$

This observation combined with a nearly identical proof to equality case shows that for all n large enough and appropriate $\{t_n\}$

$$\hat{\rho}_n + \left(\frac{n-1}{n-W} \sum_{j=1}^{W} b_j \right) D_n + t_n \geq \rho$$

almost surely. □

B.2 L_2 Norm Condition

Now, we look at analyzing $\hat{\rho}_n$ from (3.10) under the L_2 condition. First, we consider the condition in (3.3). Define the averaged estimate

$$\hat{\rho}_n^{(3)} \triangleq \sqrt{\frac{1}{n-1} \sum_{i=2}^{n} \left(\tilde{\rho}_i^{(3)} \right)^2}$$

and analogously $\left(\hat{\rho}_n^{(2)} \right)^2$. The following lemma shows that $\hat{\rho}_n^{(3)}$ upper bounds ρ eventually.

Lemma 18. *For all sequences $\{t_n\}$ such that*

$$\sum_{n=2}^{\infty} \exp\left\{ -\frac{2(n-1)t_n^2}{diam^2(\mathscr{X})} \right\} < +\infty$$

it holds that for all n large enough

$$\sqrt{\left(\hat{\rho}_n^{(3)} \right)^2 + t_n} \geq \rho$$

almost surely.

Proof. First, for all i it holds that

$$\tilde{\rho}_i^{(3)} \geq \|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|$$

This in turn implies that

$$\mathbb{E}\left[\left(\tilde{\rho}_i^{(3)}\right)^2 \;\middle|\; \mathscr{F}_{i-1}\right] \geq \mathbb{E}\left[\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|^2 \mid \mathscr{F}_{i-1}\right] = \rho^2$$

Second, it holds that $0 \leq \left(\tilde{\rho}_i^{(3)}\right)^2 \leq \text{diam}^2(\mathscr{X})$. Applying Lemmas 17 and 16 yields

$$\begin{aligned}
&\mathbb{P}\left\{\left(\hat{\rho}_n^{(3)}\right)^2 < \rho^2 - t_n\right\} \\
&\quad \leq \mathbb{P}\left\{\left(\hat{\rho}_n^{(3)}\right)^2 < \frac{1}{n-1}\sum_{i=2}^{n}\mathbb{E}\left[\left(\tilde{\rho}_i^{(3)}\right)^2 \;\middle|\; \mathscr{F}_i\right] - t_n\right\} \\
&\quad \leq \exp\left\{-\frac{2(n-1)t_n^2}{\text{diam}^2(\mathscr{X})}\right\}
\end{aligned}$$

By the Borel-Cantelli lemma, this in turn implies that for n sufficiently large

$$\sqrt{\left(\hat{\rho}_n^{(3)}\right)^2 + t_n} \geq \rho$$

□

We can now follow the proof technique of Theorem 1 and Lemma 18 to prove Theorem 2.

Proof of Theorem 2: This is a straightforward extension of the proof of Theorem 1 using the observation that

$$\begin{aligned}
&|(\hat{\rho}_n)^2 - (\hat{\rho}_n^{(3)})^2| \\
&\quad \leq |(\hat{\rho}_n)^2 - (\hat{\rho}_n^{(2)})^2| + |(\hat{\rho}_n^{(2)})^2 - (\hat{\rho}_n^{(3)})^2| \\
&\quad \leq |\hat{\rho}_n + \hat{\rho}_n^{(2)}||\hat{\rho}_n - \hat{\rho}_n^{(2)}| + |\hat{\rho}_n^{(2)} + \hat{\rho}_n^{(3)}||\hat{\rho}_n^{(2)} - \hat{\rho}_n^{(3)}| \\
&\quad \leq 2\text{diam}(\mathscr{X})\left(|\hat{\rho}_n - \hat{\rho}_n^{(2)}| + |\hat{\rho}_n^{(2)} - \hat{\rho}_n^{(3)}|\right)
\end{aligned}$$

We can now follow the proof technique of Theorem 1. □

Proof of Theorem 4: This is a straightforward extension of the proof of Theorem 3 along the lines of Theorem 2 □

B.3 Effect of Parameter Estimation

Our analysis of estimating ρ assumes that we know the parameters of the functions and in particular the strong convexity parameter m. We now argue that the effect of using estimated parameters ψ from Appendix D of [3] instead is minimal. This happens because we know that for all n large enough it holds that

$$\hat{\psi}_n + t_n \mathbf{1} + o_{\mathbb{P}}(1) \geq \psi^*$$

almost surely. Applying Lemma 25 with $\phi_i = \tilde{\rho}_i$ and $\pi_i = \hat{\psi}_i + t_i$ (the parameters such as strong convexity parameter) guarantees that estimating ρ with the parameters unknown works as with the parameters known. Therefore, the analysis in this section is not restrictive.

B.4 Proofs for Alternate One-Step Estimates

For the IPM estimates, we need a version of Hoeffding's inequality that allows for some dependence. Given an integer W, we construct a cover of $\{1,2,\ldots,n\}$ by dividing the set into W groups of integers spaced by W, i.e.,

$$\mathscr{A}_j = \left\{ j, j+W, j+2W \ldots, j + \left\lfloor \frac{n-j}{W} \right\rfloor W \right\} \qquad j = 1,\ldots,W \tag{B.4}$$

Note that

$$\{1,2,\ldots,n\} = \bigcup_{j=1}^{W} \mathscr{A}_j$$

and $\mathscr{A}_i \cap \mathscr{A}_j = \emptyset$ for $i \neq j$. The proof of Lemma 19 is nearly identical to the proof of the extension of Hoeffding's inequality from [77] with Lemma 16 used instead. We assume that if we refer to a filtration $\mathscr{F}_i$ with $i < 0$, then we implicitly refer to $\mathscr{F}_0$.

Lemma 19 (Dependent Hoeffding's Inequality)**.** *Suppose we are given a collection of random variable $\{V_i\}_{i=1}^n$ and a filtration $\{\mathscr{F}\}_{i=0}^n$ such that*

1. $a_i \leq V_i \leq b_i$ *for constants* a_i *and* $b_i \quad i = 1, \ldots, n$

2. V_i *is* $\mathscr{F}_i$*-measurable* $\quad i = 1, \ldots, n$

3. *Given an integer* W *and a cover* $\{\mathscr{A}_j\}_{j=1}^{W}$ *as in* (B.4) *for each* j *it holds that*

$$\mathbb{E}\left[V_{j+iW} \;\middle|\; \mathscr{F}_{j+(i-1)W}\right] = 0 \quad i = 1, \ldots, \left\lfloor \frac{n-j}{W} \right\rfloor$$

and

$$\mathbb{E}\left[V_j \;\middle|\; \mathscr{F}_0\right] = 0$$

Then it holds that

$$\mathbb{P}\left\{\sum_{i=1}^{n} V_i > t\right\} \leq \exp\left\{-\frac{2t^2}{W\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n} V_i < -t\right\} \leq \exp\left\{-\frac{2t^2}{W\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

Proof. Define

$$U_j \triangleq \sum_{i=0}^{\left\lfloor \frac{n-j}{W} \right\rfloor} V_{j+iW}$$

for $j = 1, \ldots, W$. Let $\{p_j\}_{j=1}^{W}$ be a probability distribution on $\{1, \ldots, W\}$ to be specified later. By Jensen's inequality, we have

$$\begin{aligned} \exp\left\{s\sum_{i=1}^{n} V_i\right\} &= \exp\left\{\sum_{j=1}^{W} p_j \frac{s}{p_j} U_j\right\} \\ &\leq \sum_{j=1}^{W} p_j \exp\left\{\frac{s}{p_j} U_j\right\} \end{aligned}$$

Then it holds that

$$\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n} V_i\right\}\right] \leq \sum_{j=1}^{W} p_j \mathbb{E}\left[\exp\left\{\frac{s}{p_j} U_j\right\}\right]$$

Now consider one term

$$\mathbb{E}\left[\exp\left\{\frac{s}{p_j}U_j\right\}\right] = \mathbb{E}\left[\exp\left\{\frac{s}{p_j}\sum_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor} V_{j+iW}\right\}\right]$$

Since $a_{j+iW} \leq V_{j+iW} \leq b_{j+iW}$ and

$$\mathbb{E}\left[V_{j+iW} \;\middle|\; \mathscr{F}_{j+(i-1)W}\right] = 0,$$

we can apply the conditional version Hoeffding's lemma given in Lemma 17 to yield

$$\mathbb{E}\left[e^{sV_{j+iW}} \;\middle|\; \mathscr{F}_{j+(i-1)W}\right] \leq \exp\left\{\frac{1}{8}\left(b_{j+iW}-a_{j+iW}\right)^2 s^2\right\}$$

Then we can apply Lemma 16 to $\{V_{j+iW}\}_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}$ and $\{\mathscr{F}_{j+iW}\}_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}$ to yield

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\frac{s}{p_j}U_j\right\}\right] &\leq \exp\left\{\frac{s^2}{8p_j^2}\sum_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}(b_{j+iW}-a_{j+iW})^2\right\} \\
&= \prod_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}\exp\left\{\frac{s^2}{8p_j^2}(b_\alpha-a_\alpha)^2\right\}
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n}V_i\right\}\right] &\leq \sum_{j=1}^{W}p_j\prod_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}\exp\left\{\frac{s^2}{8p_j^2}(b_\alpha-a_\alpha)^2\right\} \\
&= \sum_{j=1}^{W}p_j\exp\left\{\frac{s^2c_j}{8p_j^2}\right\}
\end{aligned}$$

with

$$c_j = \sum_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}(b_{j+iW}-a_{j+iW})^2$$

Let $p_j = \sqrt{c_j}/T$ and

$$T = \sum_{j=1}^{W}\sqrt{c_j}$$

Therefore, we have

$$\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n}V_i\right\}\right] \leq \exp\left\{\frac{1}{8}T^2s^2\right\}$$

Applying the Chernoff bound [76] and optimizing yields

$$\mathbb{P}\left\{\sum_{i=1}^{n}V_i > t\right\} \leq \exp\left\{-2t^2/T^2\right\}$$

Bounding T with Cauchy-Schwarz yields

$$T^2 \leq \left(\sum_{j=1}^{W}1\right)\left(\sum_{j=1}^{W}c_j\right) = W\sum_{i=1}^{n}(b_i-a_i)^2$$

and the results follows. The proof for the other tail is nearly identical. □

Remark: If we do not have the condition 3 of Lemma 19, then it holds that

$$\mathbb{P}\left\{\sum_{i=1}^{n}V_i > \sum_{j=1}^{W}\sum_{i=0}^{\left\lfloor\frac{n-j}{W}\right\rfloor}\mathbb{E}\left[V_{j+iW}\;\middle|\;\mathscr{F}_{j+(i-1)W}\right]+t\right\}$$
$$\leq \exp\left\{-\frac{2t^2}{W\sum_{i=1}^{n}(b_i-a_i)^2}\right\} \tag{B.5}$$

If we can bound the conditional expectation

$$\mathbb{E}\left[V_{j+iW}\;\middle|\;\mathscr{F}_{j+(i-1)W}\right] \leq C_{j+iW},$$

by a $\mathscr{F}_{j+(i-1)W}$-measurable random variable, then we have

$$\mathbb{P}\left\{\sum_{i=1}^{n}V_i > \sum_{i=1}^{n}C_i+t\right\} \leq \exp\left\{-\frac{2t^2}{W\sum_{i=1}^{n}(b_i-a_i)^2}\right\}$$

This remark is similar to the remark after Lemma 16.

Appendix C

Proofs for Analysis with Change in Minimizers Unknown

We prove a general result showing that for any choice of K_n such that $K_n \geq K^*$ for all n large enough, with K^* from (3.16), the mean criterion is controlled in the sense that

$$\limsup_{n\to\infty} (\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*)) \leq \varepsilon$$

Consider the function

$$\phi_K(v) = \alpha(K)\left(\sqrt{\frac{2}{m}}v + \rho\right)^2 + \beta(K) = b\left(\left(\sqrt{\frac{2}{m}}v + \rho\right)^2, K\right) \tag{C.1}$$

from Assumption C.2. Note that as a function of v, $\phi_K(v)$ is clearly increasing and strictly concave. If we select $K_n = K^*$, then by definition it holds that

$$\phi_{K^*}(\varepsilon) \leq \varepsilon \tag{C.2}$$

First, we study fixed points of the function $\phi_{K^*}(v)$. We need Theorem 3.3 of [78] to proceed

Lemma 20 (Theorem 3.3 of [78])**.** *Suppose that f is an increasing and strictly concave function mapping from $\mathbb{R}$ to $\mathbb{R}$ such that $f(0) \geq 0$ and there exist points $0 < a < b$ such that $f(a) > a$ and $f(b) < b$. Then f has unique positive fixed point.*

Proof. See [78] for the proof. □

We consider the fixed points of the function $\phi_{K^*,\rho}(v) + \delta$ with $\delta \geq 0$. We add the term δ for reasons that will become clear later in the proof of Theorem 5.

Lemma 21. *Provided that $\alpha(K) > 0$ for all $K > 0$, $\rho > 0$, and $\delta \geq 0$, the function $\phi_{K^*,\rho}(v) + \delta$ has a unique positive fixed point $\bar{v}_\delta$ with the following properties:*

1. $\bar{v}_0 = \phi_{K^*,\rho}(\bar{v}_0) \leq \varepsilon$

2. $\phi'_{K^*,\rho}(\bar{v}_\delta) < 1$

3. $\bar{v}_\delta$ *is non-decreasing in* δ *and*

$$\lim_{\delta \searrow 0} \bar{v}_\delta = \bar{v}_0$$

Proof. Since

$$\lim_{v \to 0} (\phi_{K^*}(v) + \delta) = \phi_{K^*}(0) + \delta$$

and

$$\phi_{K^*}(0) + \delta = \alpha(K^*)\rho^2 + \beta(K^*) + \delta > 0$$

for all $\delta \geq 0$, there exists a positive a sufficiently small that

$$\phi_{K^*}(a) + \delta > a$$

Next, expanding $\phi_K(v)$ yields

$$\phi_K(v) = \frac{2}{m}\alpha(K)v + 2\alpha(K)\rho\sqrt{\frac{2}{m}}\sqrt{v} + \alpha(K)\rho^2 + \beta(K)$$

Since $\phi_{K^*}(\varepsilon) \leq \varepsilon$, we obviously must have $\frac{2}{m}\alpha(K^*) \leq 1$. Suppose that

$$\frac{2}{m}\alpha(K^*) = 1$$

Then it holds that

$$\phi_{K^*}(\varepsilon) = \varepsilon + \sqrt{2m}\rho\sqrt{\varepsilon} + \frac{m}{2}\rho^2 + \beta(K) > \varepsilon$$

This contradicts (C.2), so it holds that

$$\frac{2}{m}\alpha(K^*) < 1$$

It is thus readily apparent that

$$v - (\phi_{K^*}(v) + \delta) \to \infty$$

as $v \to \infty$. Therefore, there exists a point $b > a$ such that

$$\phi_{K^*}(b) + \delta < b$$

In addition, it is easy to check that $\phi_{K^*}(v) + \delta$ is increasing and strictly concave. Therefore, we can apply Lemma 20 from [78] to conclude that there exists a unique, positive fixed point $\bar{v}_\delta$ of $\phi_{K^*}(v) + \delta$.

Next, suppose that $\phi'_{K^*}(\bar{v}_\delta) > 1$. Then by continuity for $v > \bar{v}_\delta$ sufficiently close to $\bar{v}_\delta$, we have

$$\phi_{K^*}(v) + \delta > v$$

However, we know that as $v \to \infty$, it holds that $v - (\phi_{K^*}(v) + \delta) \to \infty$. By the Intermediate Value Theorem, this implies that there is another fixed point on $[v, b]$. This is a contradiction, since $\bar{v}_\delta$ is the unique, positive fixed point. Therefore, it holds that $\phi'_{K^*}(\bar{v}_\delta) \leq 1$. Now, suppose that $\phi'_{K^*}(\bar{v}_\delta) = 1$. Since $\phi_{K^*}(v)$ is strictly concave, its derivative is decreasing [29]. Therefore, on $[0, \bar{v}_\delta)$, it holds that

$$\phi'_{K^*}(v) \geq 1$$

This implies that

$$\begin{aligned} \bar{v}_\delta &= \phi_{K^*}(\bar{v}_\delta) + \delta \\ &= \phi_{K^*}(0) + \int_0^{\bar{v}_\delta} \phi'_{K^*}(v) dv + \delta \\ &\geq \phi_{K^*}(0) + \delta + \bar{v}_\delta \\ &> \bar{v}_\delta \end{aligned}$$

This is a contradiction, so it must be that $\phi'_{K^*}(\bar{v}_\delta) < 1$.

Since there is a unique positive fixed point $\bar{v}_\delta$ and $v - (\phi_{K^*}(v) + \delta) \to \infty$, it must hold that $\phi_{K^*}(x) + \delta \leq x$ iff $x \geq \bar{v}_\delta$. Since $\phi_{K^*}(\varepsilon) \leq \varepsilon$, it holds that $\bar{v}_0 \leq \varepsilon$.

Finally, for $\delta' \geq \delta$, it holds that

$$\begin{aligned} \bar{v}_\delta &= \phi_{K^*}(\bar{v}_\delta) + \delta \\ &= \phi_{K^*}(\bar{v}_\delta) + \delta' + \underbrace{(\delta - \delta')}_{<0} \\ &< \phi_{K^*}(\bar{v}_\delta) + \delta' \end{aligned} \tag{C.3}$$

By the observation above, we then have $\bar{v}_\delta \leq \bar{v}_{\delta'}$. This monotonicity and the implicit function theorem [79] in turn imply that

$$\lim_{\delta \searrow 0} \bar{v}_\delta = \bar{v}_0$$

□

As a simple consequence of the concavity of $\phi_{K^*}(v)$, we can study a fixed point iteration involving $\phi_K(v)$. Define the n-fold composition mapping

$$(\phi_K + \delta)^{(n)}(v) \triangleq ((\phi_K + \delta) \circ \cdots \circ (\phi_K + \delta))(v)$$

Lemma 22. *For any $v > 0$, it holds that*

$$\lim_{n\to\infty} (\phi_{K^*} + \delta)^{(n)}(v) = \bar{v}_\delta$$

Proof. Following [80], for any fixed point $\bar{v}$, it holds that

$$|\phi_{K^*}(v) + \delta - \bar{v}_\delta| \leq \phi'_{K^*}(\bar{v})|v - \bar{v}_\delta|$$

Therefore, applying the fixed point property repeatedly yields

$$|(\phi_{K^*} + \delta)^{(n)}(v) - \bar{v}_\delta| \leq (\phi'_{K^*}(\bar{v}))^n |v - \bar{v}_\delta|$$

By Lemma 21, it holds that

$$\phi'_{K^*}(\bar{v}) < 1$$

and so the result follows. □

This implies that if we select K^* stochastic gradients at every time instant, and we start from any v, then it holds that

$$\phi^{(n)}_{K^*,\rho}(v) \to \bar{v}_0$$

with $\bar{v}_0 \leq \varepsilon$.

Now, we show that we control the mean criterion defined in (1.7) when we estimate ρ. In Section 2.3.1 of [3], we pick a deterministic choice of $K_n = K^*$ and

proceed with the analysis. Then it holds that

$$\begin{aligned}
&\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \\
&\quad \leq \mathbb{E}\left[b\left(\left(\sqrt{\frac{2}{m}(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*))} + \rho \right)^2, K_n \right) \right] \\
&\quad = \mathbb{E}\left[\alpha(K_n) \left(\sqrt{\frac{2}{m}(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*))} + \rho \right)^2 \right] \\
&\qquad\qquad + \mathbb{E}[\beta(K_n)] \qquad \text{(C.4)} \\
&\quad = \alpha(K_n)\mathbb{E}\left[\left(\sqrt{\frac{2}{m}(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*))} + \rho \right)^2 \right] \\
&\qquad\qquad + \beta(K_n) \qquad \text{(C.5)}
\end{aligned}$$

We can bound

$$\mathbb{E}\left[\left(\sqrt{\frac{2}{m}(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*))} + \rho \right)^2 \right]$$

using (2.11) and recover (C.1). However, in Chapter 3, K_n and $\boldsymbol{x}_{n-1}$ are dependent random variables, so (C.5) does not hold in general. Instead, only (C.4) holds. To get around this issue, we need a more sophisticated analysis using the observation that $K_n \geq K^*$ for all n large enough. This property implies that K_n behaves like a constant for n large enough and a close analog of the analysis in Section 2.3.1 applies.

Proof of Theorem 5: We know that for all n large enough, we pick $K_n \geq K^*$ almost surely. This in turn implies that there exists a finite almost surely random variable $\tilde{N}$ such that

$$n \geq \tilde{N} \;\; \Rightarrow \;\; K_n \geq K^*$$

Since $\tilde{N}$ is finite almost surely, we know that

$$\lim_{n\to\infty} \mathbb{P}\left\{\tilde{N} > n\right\} = 0$$

By the compactness of $\mathscr{X}$, it follows that there is a constant $C > 0$ such that

$$\max_{\boldsymbol{x} \in \mathscr{X}} \phi_{K^*,\rho} \left(f_n(\boldsymbol{x}) - f_n(\boldsymbol{x}_n^*) \right) \leq C \ \ \forall n \geq 1$$

Then it follows that

$$\begin{aligned}
&\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \\
&\quad = \mathbb{E}\left[\phi_{K_n,\rho} \left(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*) \right) \right] \\
&\quad = \mathbb{E}\left[\phi_{K_n,\rho} \left(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*) \right) \mathbb{1}_{\{n \geq \tilde{N}\}} \right] \\
&\qquad + \mathbb{E}\left[\phi_{K_n,\rho} \left(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*) \right) \mathbb{1}_{\{n < \tilde{N}\}} \right] \\
&\quad \leq \mathbb{E}\left[\phi_{K^*,\rho} \left(f_{n-1}(\boldsymbol{x}_{n-1}) - f_{n-1}(\boldsymbol{x}_{n-1}^*) \right) \mathbb{1}_{\{n \geq \tilde{N}\}} \right] \\
&\qquad + C\mathbb{P}\left\{ \tilde{N} > n \right\} \\
&\quad \leq \phi_{K^*,\rho} \left(\mathbb{E}[f_{n-1}(\boldsymbol{x}_{n-1})] - f_{n-1}(\boldsymbol{x}_{n-1}^*) \right) + C\mathbb{P}\left\{ \tilde{N} > n \right\}
\end{aligned}$$

To bound the mean criterion, we consider the recursion

$$\varepsilon_n = \phi_{K^*,\rho} \left(\varepsilon_{n-1} \right) + C\mathbb{P}\left\{ \tilde{N} > n \right\} \qquad \forall n \geq \tilde{N} \tag{C.6}$$

which satisfies

$$\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon_n \qquad \forall n \geq \tilde{N}$$

By assumption, we know that as $n \to \infty$

$$C\mathbb{P}\left\{ \tilde{N} > n \right\} \to 0$$

Fix $\delta > 0$. Then there exists a random variable $\tilde{N}_\delta \geq \tilde{N}$ such that

$$n \geq \tilde{N}_\delta \ \ \Rightarrow \ \ C\mathbb{P}\left\{ \tilde{N} > n \right\} \leq \delta$$

Then we consider the recursion

$$\begin{aligned}
\tilde{\varepsilon}_n &= \phi_{K^*,\rho} \left(\tilde{\varepsilon}_{n-1} \right) + \delta \qquad \forall n \geq \tilde{N}_\delta \\
\tilde{\varepsilon}_{\tilde{N}_\delta} &= \varepsilon_{\tilde{N}_\delta}
\end{aligned} \tag{C.7}$$

By construction, we have $\varepsilon_n \leq \tilde{\varepsilon}_n$ for all $n \geq \tilde{N}_\delta$. As a consequence of Lemmas 21

and 22, we have

$$\begin{aligned}\limsup_{n\to\infty}\left(\mathbb{E}[f_n(\boldsymbol{x}_n)]-f_n(\boldsymbol{x}_n^*)\right) &\leq \limsup_{n\to\infty}\varepsilon_n\\ &\leq \limsup_{n\to\infty}\tilde{\varepsilon}_n\\ &\leq \bar{v}_\delta\end{aligned}$$

Since $\delta > 0$ was arbitrary and $\bar{v}_\delta \searrow \bar{v}_0$ as $\delta \searrow 0$ from Lemma 21, it follows that

$$\limsup_{n\to\infty}\left(\mathbb{E}[f_n(\boldsymbol{x}_n)]-f_n(\boldsymbol{x}_n^*)\right) \leq \bar{v}_0 \leq \varepsilon$$

□

Appendix D

Parameter Estimation

We may need to estimate parameters of the functions $\{f_n(\boldsymbol{x})\}$ such as the strong convexity parameter m to compute the bound $b(d_0, K)$ from assumption A.4. In this section, we assume that the bound $b(d_0, K, \psi)$ is parameterized by $\psi \in \mathscr{P}$, which depends on properties of the functions $f_n(\boldsymbol{x})$. In most cases, we have the parameters

$$\psi = \begin{bmatrix} 1/m & M & A & B \end{bmatrix}^\top$$

where m is the parameter of strong convexity, M is the Lipschitz gradient modulus, and the pair (A, B) controls gradient growth as in assumption A.5, i.e.,

$$\mathbb{E}\|\nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z})\|^2 \leq A + B\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \tag{D.1}$$

We parameterize using $1/m$, since a smaller m generally increases the bound $b(d_0, K, \psi)$. Therefore, if the true parameters are ψ^* for all functions $\{f_n(\boldsymbol{x})\}$, then we want to find an estimate $\hat{\psi}$ such that $\hat{\psi} \geq \psi^*$. It is generally true that $b(d_0, K, \psi)$ is increasing in ψ, so the estimate $\hat{\psi}$ produces a more conservative bound on the mean criterion. With a more conservative bound on the mean criterion, the methods of Chapters 2-6 work.

There is a slight complication in estimating the gradient parameters A and B in (D.1), since the choice of A and B is generally not unique. As an example, since the space $\mathscr{X}$ is compact, for any fixed $B > 0$ it holds that

$$\max_{\boldsymbol{x} \in \mathscr{X}} \left\{ \mathbb{E}\|\nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z})\|^2 - B\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \right\} < +\infty$$

Therefore, we fix any $B > 0$ and set $A = \max_{\boldsymbol{x} \in \mathscr{X}} \left\{ \mathbb{E}\|\nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z})\|^2 - B\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \right\}$. In practice, a choice of (A, B) with A or B too large leads to a much larger choice of K_n. We need to find a trade-off between a choice of A and B such that neither is too large. Finally, when comparing estimates of A and B, since there is no unique choice of A and B, we need to be careful in evaluating a particular estimate.

As in estimating ρ, we produce one time instant estimates $\tilde{m}_i$, $\tilde{M}_i$, $\tilde{A}_i$, and $\tilde{B}_i$ at time i and combine them by averaging to yield

1. $\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{m}_i$
2. $\hat{M}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{M}_i$
3. $\hat{A}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{A}_i$
4. $\hat{B}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{B}_i$

We can analogously define maximum combinations of the single time instant estimates as with ρ estimation in section 3.1.3 of Chapter 3. The details of this approach are not included but the results are analogous to the direct estimate in (3.5). In this case, we can assume that $f_n(\boldsymbol{x})$ has true parameters ψ_n such that $\psi_n \leq \psi^*$.

We make the following assumptions for our analysis:

D.1 The parameters $\psi \in \mathscr{P}$ with $\mathscr{P}$ compact and there exists a true set of parameters ψ^*

D.2 The bound $b(d_0, K, \tilde{\psi})$ is non-decreasing in ψ, i.e.,

$$\psi \leq \tilde{\psi} \;\Rightarrow\; b(d_0, K, \psi) \leq b(d_0, K, \tilde{\psi})$$

D.3 $\nabla f_n(\boldsymbol{x}_n)$ has Lipschitz continuous gradients with modulus M and

$$\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{g}(\tilde{\boldsymbol{x}}, \boldsymbol{z})\|^2 \mid \boldsymbol{x}, \tilde{\boldsymbol{x}}\right] \leq L^2 \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2$$

D.4 $f_n(\boldsymbol{x})$ is twice differentiable and there exist stochastic second derivatives with respect to $\boldsymbol{x}$, $\boldsymbol{g}^{(2)}(\boldsymbol{x}, \boldsymbol{z})$, such that

$$\mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\boldsymbol{g}_n^{(2)}(\boldsymbol{x}, \boldsymbol{z}_n) \mid \boldsymbol{x}\right] = \nabla^2_{\boldsymbol{x}\boldsymbol{x}} f_n(\boldsymbol{x})$$

D.5 The space $\mathscr{Z}$ is compact and there exists a constant G such that

$$\|\boldsymbol{g}_n(\boldsymbol{x}, \boldsymbol{z})\| \leq G \quad \forall \boldsymbol{x}\, \forall \boldsymbol{z}\, \forall n$$

D.6 We have access to functions $\hat{f}_n(\boldsymbol{x}, \boldsymbol{z})$ such that

$$\mathbb{E}[\hat{f}_n(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{x}] = f_n(\boldsymbol{x})$$

D.1 Estimating Strong Convexity Parameter and Lipschitz Gradient Modulus

We seek one-step estimates $\tilde{m}_i$ and $\tilde{M}_i$, of the parameter of strong convexity and Lipschitz gradient modulus respectively, such that $\mathbb{E}[\tilde{m}_i \mid \mathscr{F}_{i-1}] \leq m$ and $\mathbb{E}[\tilde{M}_i \mid \mathscr{F}_{i-1}] \geq M$. In this section, we focus on estimates for the strong convexity parameter. The methods can be extended to estimating the Lipschitz gradient modulus trivially.

D.1.1 Penalty Method

Suppose that our functions are of the form

$$f_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_i \sim p_i}\left[\ell(\boldsymbol{x}, \boldsymbol{z}_i)\right]$$

with

$$\ell(\boldsymbol{x}, \boldsymbol{z}) = \tilde{\ell}(\boldsymbol{x}, \boldsymbol{z}) + \frac{1}{2}\lambda \|\boldsymbol{x}\|^2$$

Then it holds, trivially, that $m \geq \lambda$ yielding a simple estimate of m.

D.1.2 Hessian Method

Due to the strong convexity condition in assumption A.1, we have

$$\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f_i(\boldsymbol{x}) \succeq m\boldsymbol{I} \quad \forall \boldsymbol{x} \in \mathscr{X}$$

This in turn implies that

$$\lambda_{\min}\left(\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f_i(\boldsymbol{x})\right) \geq m \quad \forall \boldsymbol{x} \in \mathscr{X}$$

where $\lambda_{\min}(\boldsymbol{X})$ is the smallest eigenvalue of the matrix $\boldsymbol{X}$. For convenience, we assume that

$$\min_{\boldsymbol{x} \in \mathscr{X}} \lambda_{\min}\left(\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f_n(\boldsymbol{x})\right) = m$$

This suggests that given $\{\boldsymbol{z}_i(k)\}_{k=1}^{K_i}$ we set

$$\tilde{m}_i \triangleq \min_{\boldsymbol{x}\in\mathscr{X}} \lambda_{\min}\left(\frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x},\boldsymbol{z}_i(k)\right)\right)$$

Since

$$\lambda_{\min}(A) = \min_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \langle A\boldsymbol{v},\boldsymbol{v}\rangle$$

$\lambda_{\min}(A)$ is a concave function of A. By Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}[\tilde{m}_i \mid \mathscr{F}_{i-1}] &= \mathbb{E}\left[\min_{\boldsymbol{x}\in\mathscr{X}} \lambda_{\min}\left(\frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x},\boldsymbol{z}_i(k)\right)\right) \;\middle|\; \mathscr{F}_{i-1}\right] \\
&\leq \min_{\boldsymbol{x}\in\mathscr{X}} \mathbb{E}\left[\lambda_{\min}\left(\frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x},\boldsymbol{z}_i(k)\right)\right) \;\middle|\; \mathscr{F}_{i-1}\right] \\
&\leq \min_{\boldsymbol{x}\in\mathscr{X}} \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x},\boldsymbol{z}_i(k)\right) \;\middle|\; \mathscr{F}_{i-1}\right]\right) \\
&= \min_{\boldsymbol{x}\in\mathscr{X}} \lambda_{\min}\left(\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f_i(\boldsymbol{x})\right) \\
&= m
\end{aligned}$$

Similarly, we can set

$$\tilde{M}_i \triangleq \max_{\boldsymbol{x}\in\mathscr{X}} \lambda_{\max}\left(\frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x},\boldsymbol{z}_i(k)\right)\right)$$

where $\lambda_{\max}(\boldsymbol{X})$ is the largest eigenvalue of $\boldsymbol{X}$. Since

$$\lambda_{\max}(A) = \max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \langle A\boldsymbol{v},\boldsymbol{v}\rangle$$

$\lambda_{\max}(A)$ is a convex function of A. By Jensen's inequality, it holds that $\mathbb{E}[\tilde{M}_i \mid \mathscr{F}_{i-1}] \geq M$.

Gradient Descent Approach: To minimize over $\boldsymbol{x}$ and compute $\tilde{m}_i$ and $\tilde{M}_i$, we can use gradient descent by exploiting eigenvalue perturbation results [81]. First, suppose that we are given a matrix valued function $\boldsymbol{T}(\boldsymbol{x})$, and we want to compare the eigenspectrum of $\boldsymbol{T}(\boldsymbol{x})$ and $\boldsymbol{T}(\boldsymbol{x}_0)$ for a fixed point $\boldsymbol{x}_0$. Given eigenvectors $\boldsymbol{v}_{0i}$ and eigenvalues λ_{0i} of the matrix $\boldsymbol{T}(\boldsymbol{x}_0)$, we want to efficiently find eigenvectors

$\boldsymbol{v}_i$ and eigenvalues λ_i of $\boldsymbol{T}(\boldsymbol{x})$. From [81], it holds that

$$\frac{\partial \lambda_i}{\partial \boldsymbol{T}_{kj}}(\boldsymbol{x}) = \boldsymbol{v}_i^{(k)}(\boldsymbol{x})\boldsymbol{v}_i^{(j)}(\boldsymbol{x})(2-\delta_{ij})$$

where $\boldsymbol{v}_i^{(k)}(\boldsymbol{x})$ is the k^{th} entry of the eigenvector corresponding to $\lambda_i(\boldsymbol{x})$. Then it holds that

$$\begin{aligned}\nabla_{\boldsymbol{x}}\lambda_{\min}(\boldsymbol{T}(\boldsymbol{x})) &= \sum_{i,j}\frac{\partial \lambda_{\min}}{\partial \boldsymbol{T}_{ij}}\nabla_{\boldsymbol{x}}\boldsymbol{T}_{ij}(\boldsymbol{x}) \\ &= \sum_{i,j}\boldsymbol{v}_{\min}^{(i)}(\boldsymbol{x})\boldsymbol{v}_{\min}^{(j)}(\boldsymbol{x})(2-\delta_{ij})\nabla_{\boldsymbol{x}}\boldsymbol{T}_{ij}(\boldsymbol{x})\end{aligned}$$

With this observation, we can use gradient descent to solve the optimization problem

$$\min_{\boldsymbol{x}\in\mathscr{X}}\lambda_{\min}\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\boldsymbol{g}_i^{(2)}(\boldsymbol{x},\boldsymbol{z}_i(k))\right)$$

by setting

$$\boldsymbol{T}(\boldsymbol{x}) = \frac{1}{K_i}\sum_{k=1}^{K_i}\boldsymbol{g}_i^{(2)}(\boldsymbol{x},\boldsymbol{z}_i(k))$$

Starting from any $\boldsymbol{m}_i(0)\in\mathscr{X}$, we compute

$$\boldsymbol{m}_i(p) = \Pi_{\mathscr{X}}\left[\boldsymbol{m}_i(p-1) - \mu(p)\nabla_{\boldsymbol{x}}\lambda_{\min}\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\boldsymbol{g}_i^{(2)}(\boldsymbol{m}_i(p-1),\boldsymbol{z}_i(k))\right)\right]$$

for $p=1,\ldots,P$ and set

$$\hat{m}_i \triangleq \lambda_{\min}\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\boldsymbol{g}_i^{(2)}(\boldsymbol{m}_i(P),\boldsymbol{z}_i(k))\right) \tag{D.2}$$

D.1.3 Ratio Method

For any two points $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, by strong convexity we have

$$f_i(\tilde{\boldsymbol{x}}) \geq f_i(\boldsymbol{x}) + \langle \nabla f_i(\boldsymbol{x}), \tilde{\boldsymbol{x}}-\boldsymbol{x}\rangle + \frac{1}{2}m\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|^2 \quad \forall \boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}$$

which implies that

$$m \leq \frac{f_i(\tilde{\boldsymbol{x}}) - f_i(\boldsymbol{x}) - \langle \nabla f_i(\boldsymbol{x}), \tilde{\boldsymbol{x}} - \boldsymbol{x} \rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2}$$

We suppose that for all n

$$m = \min_{\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathscr{X}} \frac{f_i(\tilde{\boldsymbol{x}}) - f_i(\boldsymbol{x}) - \langle \nabla f_i(\boldsymbol{x}), \tilde{\boldsymbol{x}} - \boldsymbol{x} \rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2}$$

This is not restrictive, since any $m > 0$ that satisfies

$$m \leq \min_{\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathscr{X}} \frac{f_i(\tilde{\boldsymbol{x}}) - f_i(\boldsymbol{x}) - \langle \nabla f_i(\boldsymbol{x}), \tilde{\boldsymbol{x}} - \boldsymbol{x} \rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2}$$

can be taken as a parameter of strong convexity for the function $f_i(\boldsymbol{x})$. Consider the following estimate of m:

$$\tilde{m}_i \triangleq \min_{\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathscr{X}} \left\{ \frac{\frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\tilde{\boldsymbol{x}}, \boldsymbol{z}_i(k)) - \frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\boldsymbol{x}, \boldsymbol{z}_i(k))}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} - \frac{\left\langle \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k)), \tilde{\boldsymbol{x}} - \boldsymbol{x} \right\rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} \right\}$$

This estimate satisfies [1]

$$
\begin{aligned}
&\mathbb{E}[\tilde{m}_i \mid \mathscr{F}_{i-1}] \\
&\quad = \mathbb{E}\left[\min_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \left\{ \frac{\frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\tilde{\boldsymbol{x}}, \boldsymbol{z}_i(k)) - \frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\boldsymbol{x}, \boldsymbol{z}_i(k))}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} \right.\right. \\
&\qquad\qquad \left.\left. - \frac{\left\langle \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k)), \tilde{\boldsymbol{x}} - \boldsymbol{x} \right\rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} \right\}\right] \\
&\quad \leq \min_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \mathbb{E}\left[\frac{\frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\tilde{\boldsymbol{x}}, \boldsymbol{z}_i(k)) - \frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\boldsymbol{x}, \boldsymbol{z}_i(k))}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} \right. \\
&\qquad\qquad \left. - \frac{\left\langle \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k)), \tilde{\boldsymbol{x}} - \boldsymbol{x} \right\rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} \right] \\
&\quad = \min_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \frac{f_i(\tilde{\boldsymbol{x}}) - f_i(\boldsymbol{x}) - \langle \nabla f_i(\boldsymbol{x}), \tilde{\boldsymbol{x}} - \boldsymbol{x}\rangle}{\frac{1}{2}\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2} \\
&\quad = m
\end{aligned}
$$

We can approximately solve this minimization problem by using a numerical solver or gradient descent. Additionally, since computing the minimum here is difficult and is generally a non-convex problem, we can instead look at an approximate method. Suppose that we have N points $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(N)$. Then we know that for any two distinct points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$

$$
m \leq \frac{f_i(\boldsymbol{x}(i)) - f_i(\boldsymbol{x}(j)) - \langle \nabla f_i(\boldsymbol{x}(j)), \boldsymbol{x}(i) - \boldsymbol{x}(j)\rangle}{\frac{1}{2}\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2}
$$

This suggests the estimate

$$
\begin{aligned}
\hat{m}_i \triangleq \min_{i\neq j} \left\{ \frac{\frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\boldsymbol{x}(i), \boldsymbol{z}_i(k)) - \frac{1}{K_i}\sum_{k=1}^{K_i} \hat{f}_i(\boldsymbol{x}(j), \boldsymbol{z}_i(k))}{\frac{1}{2}\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2} \right. \\
\left. - \frac{\left\langle \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}(j), \boldsymbol{z}_i(k)), \boldsymbol{x}(i) - \boldsymbol{x}(j) \right\rangle}{\frac{1}{2}\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2} \right\}
\end{aligned} \tag{D.3}
$$

[1] We ignore measurability issues here.

for the strong convexity parameter. Then, by similar reasoning, we have

$$\mathbb{E}[\hat{m}_i \mid \mathscr{F}_{i-1}] \leq \min_{i \neq j} \frac{f_i(\boldsymbol{x}(i)) - f_i(\boldsymbol{x}(j)) - \langle \nabla f_i(\boldsymbol{x}(j)), \boldsymbol{x}(i) - \boldsymbol{x}(j) \rangle}{\frac{1}{2}\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2}$$

It is difficult to compare this estimator to m exactly. All we can say is that

$$m \leq \min_{i \neq j} \frac{f_i(\boldsymbol{x}(i)) - f_i(\boldsymbol{x}(j)) - \langle \nabla f_i(\boldsymbol{x}(j)), \boldsymbol{x}(i) - \boldsymbol{x}(j) \rangle}{\frac{1}{2}\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2}$$

In practice, this method produces estimates close to m.

D.1.4 Problem Specific Estimate

The methods of this appendix are general in the sense that they can be applied to any function $\{f_n(\boldsymbol{x})\}$ satisfying assumptions **D.1**-**D.6**. For any specific problem, there may be alternate estimates of the function parameters based on the specific form of the functions. As an example, suppose that

$$f_n(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_n \sim p_n}\left[\ell\left(\boldsymbol{x}, \boldsymbol{z}_n\right)\right]$$

with $\boldsymbol{z} = \left[\ \boldsymbol{w}^\top y\ \right]^\top$ and

$$\ell\left(\boldsymbol{x}, \boldsymbol{z}\right) = \frac{1}{2}\left(y - \boldsymbol{w}^\top \boldsymbol{x}\right)^2 + \frac{1}{2}\lambda\|\boldsymbol{x}\|^2$$

Then the strong convexity parameter is given by

$$m = \lambda_{\min}\left(\lambda \boldsymbol{I} + \mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^\top\right]\right)$$

where $\lambda_{\min}(\boldsymbol{X})$ is the largest eigenvalue of the matrix $\boldsymbol{X}$. This suggests that we set

$$\hat{m}_i = \lambda_{\min}\left(\lambda \boldsymbol{I} + \frac{1}{K_i}\sum_{k=1}^{K_i} \boldsymbol{w}_i(k)\boldsymbol{w}_i^\top(k)\right)$$

Since the smallest eigenvalue $\lambda_{\min}(\boldsymbol{X})$ is a concave function of the matrix $\boldsymbol{X}$ as argued above, it follows that $\mathbb{E}\left[\hat{m}_i \mid \mathscr{F}_{i-1}\right] \leq m$. For other classes of functions $\{f_n(\boldsymbol{x})\}$, there may be other useful estimates of the strong convexity, Lipschitz gradient, and gradient parameters.

D.2 Estimating Gradient Parameters

We seek (A, B) such that

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z})\|^2 \leq A + B\|\boldsymbol{x} - \boldsymbol{x}^*\|^2$$

We consider three different approaches to estimate (A, B). The first method estimates B through an intermediate quantity and then estimates A. The second method estimates (A, B) jointly by searching for a pair (A, B) that satisfy assumption A.5. The final method is a heuristic method that estimates (A, B) through testing a finite number of points akin to the ratio method of section D.1.3 .

D.2.1 Estimate (A, B) through an Intermediate Step

From assumption **D.6**, it holds that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z})\|^2 &= \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{z}) + (\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{z}))\|^2 \\
&\leq 2\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{z})\|^2 + 2\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{z})\|^2 \\
&\leq 2\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{z})\|^2 + 2L^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2
\end{aligned}$$

Thus, we can set $B = 2L^2$ and $A = 2\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}^*, \boldsymbol{z})\|^2$. This suggests that given an estimate $\tilde{L}_i$ for L such that $\mathbb{E}[\tilde{L}_i \mid \mathscr{F}_{i-1}] \geq L$, we set

$$\tilde{B}_i = 2\tilde{L}_i^2$$

Then by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}[\tilde{B}_i \mid \mathscr{F}_{i-1}] &= 2\mathbb{E}[\tilde{L}_i^2 \mid \mathscr{F}_{i-1}] \\
&\geq 2\left(\mathbb{E}[\tilde{L}_i \mid \mathscr{F}_{i-1}]\right)^2 \\
&\geq 2L^2 \\
&= B
\end{aligned}$$

Given this observation, we now look at methods to estimate L and thus B.

L Estimation - Ratio Method

First, we consider a ratio estimator like the strong convexity ratio method. It clearly holds that

$$L^2 \leq \max_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \frac{\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)-\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i)\|^2}{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2}$$

To that end define

$$\tilde{L}_i^2 = \max_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \frac{\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))-\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i(k))\|^2}{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2}$$

This estimate satisfies[2]

$$\begin{aligned}
\mathbb{E}\left[\tilde{L}_i^2 \mid \mathscr{F}_{i-1}\right] &= \mathbb{E}\left[\max_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \frac{\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))-\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i(k))\|^2}{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2} \,\middle|\, \mathscr{F}_{i-1}\right] \\
&\geq \max_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \mathbb{E}\left[\frac{\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))-\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i(k))\|^2}{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2} \,\middle|\, \mathscr{F}_{i-1}\right] \\
&= \max_{\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathscr{X}} \frac{\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)-\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i)\|^2}{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2} \\
&\geq L^2
\end{aligned}$$

We can approximately evaluate this estimate using a numerical solver or gradient descent. As with the case of estimating strong convexity, we can construct an alternate estimate based on only a finite number of points $\boldsymbol{x}(1),\ldots,\boldsymbol{x}(N)$, i.e.,

$$\tilde{L}_i^2 = \max_{i,j\in\{1,\ldots,N\}} \frac{\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x}(i),\boldsymbol{z}_i(k))-\boldsymbol{g}_i(\boldsymbol{x}(j),\boldsymbol{z}_i(k))\|^2}{\|\boldsymbol{x}(i)-\boldsymbol{x}(j)\|^2}$$

This is not exact but close in practice.

L Estimation - M Method

For our second estimate, we start from Taylor's theorem, which guarantees that for some $\bar{\boldsymbol{x}} \in \mathscr{X}$

$$\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i) = \boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i) + \boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)(\tilde{\boldsymbol{x}}-\boldsymbol{x})$$

[2]We ignore measurability issues for this maximum.

This in turn implies that

$$\|\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i)-\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)\|^2 \leq \|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)\|_F^2\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|^2$$

with $\|\boldsymbol{X}\|_F$ the Frobenius norm. By simple computation, we have

$$\begin{aligned}\|\nabla^2_{\boldsymbol{xx}}f(\bar{\boldsymbol{x}})\|_F^2 &\leq \lambda^2_{\max}\left(\nabla^2_{\boldsymbol{xx}}f(\bar{\boldsymbol{x}})\right)\\ &\leq M^2\end{aligned}$$

This in turn implies that

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)-\nabla^2_{\boldsymbol{xx}}f_i(\bar{\boldsymbol{x}})\|_F^2 \mid \mathscr{F}_{i-1}\right]\\ &\quad=\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)\|_F^2 \mid \mathscr{F}_{i-1}\right]-\|\nabla^2_{\boldsymbol{xx}}f_i(\bar{\boldsymbol{x}})\|_F^2\\ &\quad\geq\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)\|_F^2 \mid \mathscr{F}_{i-1}\right]-\lambda^2_{\max}\left(\nabla^2_{\boldsymbol{xx}}f_i(\bar{\boldsymbol{x}})\right)\\ &\quad\geq\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)\|_F^2 \mid \mathscr{F}_{i-1}\right]-M^2\end{aligned}$$

Finally, we have

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{g}_i(\tilde{\boldsymbol{x}},\boldsymbol{z}_i)-\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)\|^2 \mid \mathscr{F}_{i-1}\right]\\ &\quad\leq\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)\|_F^2\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|^2 \mid \mathscr{F}_{i-1}\right]\\ &\quad\leq\left(M^2+\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\bar{\boldsymbol{x}},\boldsymbol{z}_i)-\nabla^2_{\boldsymbol{xx}}f_i(\bar{\boldsymbol{x}})\|_F^2 \mid \mathscr{F}_{i-1}\right]\right)\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|^2\end{aligned}$$

This in turn implies that

$$L^2\leq M^2+\max_{\boldsymbol{x}\in\mathscr{X}}\mathbb{E}\left[\|\boldsymbol{g}_i^{(2)}(\boldsymbol{x},\boldsymbol{z}_i)-\nabla^2_{\boldsymbol{xx}}f_i(\boldsymbol{x})\|_F^2 \mid \mathscr{F}_{i-1}\right]$$

Finally, we arrive at the estimate

$$\tilde{L}_i^2\triangleq(\hat{M}_{i-1}+t_{i-1})^2+\max_{\boldsymbol{x}\in\mathscr{X}}\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i^{(2)}(\boldsymbol{x},\boldsymbol{z}_i(k))-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i^{(2)}(\boldsymbol{x},\boldsymbol{z}_i(j))\right\|^2$$

This estimate satisfies

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{L}_i^2 \mid \mathscr{F}_{i-1}\right] \\
&\quad = \mathbb{E}\left[(\hat{M}_{i-1} + t_{i-1})^2 \mid \mathscr{F}_{i-1}\right] \\
&\qquad + \mathbb{E}\left[\max_{\boldsymbol{x}\in\mathscr{X}} \frac{1}{K_i - 1} \sum_{k=1}^{K_i} \left\| \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x}, \boldsymbol{z}_i(k)\right) - \frac{1}{K_i} \sum_{j=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x}, \boldsymbol{z}_i(j)\right) \right\|^2 \,\middle|\, \mathscr{F}_{i-1}\right] \\
&\quad \geq M^2 + \max_{\boldsymbol{x}\in\mathscr{X}} \mathbb{E}\left[\frac{1}{K_i - 1} \sum_{k=1}^{K_i} \left\| \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x}, \boldsymbol{z}_i(k)\right) - \frac{1}{K_i} \sum_{j=1}^{K_i} \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x}, \boldsymbol{z}_i(j)\right) \right\|^2 \,\middle|\, \mathscr{F}_{i-1}\right] \\
&\quad \geq M^2 + \max_{\boldsymbol{x}\in\mathscr{X}} \mathbb{E}\left[\| \boldsymbol{g}_i^{(2)}\left(\boldsymbol{x}, \boldsymbol{z}_i\right) - \nabla^2_{\boldsymbol{x}\boldsymbol{x}} f_i(\boldsymbol{x}) \|_F^2 \mid \mathscr{F}_{i-1}\right] \\
&\quad \geq L^2
\end{aligned}
$$

A Estimate - Maximization

We drop the i index temporarily for convenience. We have

$$\mathbb{E}\| \boldsymbol{g}\left(\boldsymbol{x}, \boldsymbol{z}\right) \|^2 \leq A + 2L^2 \|\boldsymbol{x} - \boldsymbol{x}^*\|^2$$

First, we seek A such that

$$\mathbb{E}\| \boldsymbol{g}\left(\boldsymbol{x}, \boldsymbol{z}\right) \|^2 \leq A + 2\left(\frac{L}{M}\right)^2 \|\nabla f(\boldsymbol{x})\|^2 \tag{D.4}$$

Since

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \geq \frac{1}{M} \|\nabla f(\boldsymbol{x})\|$$

for any A such that (D.4) holds, we have

$$\mathbb{E}\| \boldsymbol{g}\left(\boldsymbol{x}, \boldsymbol{z}\right) \|^2 \leq A + 2\left(\frac{L}{M}\right)^2 \|\boldsymbol{x} - \boldsymbol{x}^*\|^2$$

Therefore, we have

$$\begin{aligned} A &= \max_{\boldsymbol{x}\in\mathscr{X}} \left(\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 - 2\left(\frac{L}{M}\right)^2 \|\nabla f(\boldsymbol{x})\|^2 \right) \\ &= \max_{\boldsymbol{x}\in\mathscr{X}} \left(\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 - 2\left(\frac{L}{M}\right)^2 \left(\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 - \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z}) - \nabla f(\boldsymbol{x})\|^2\right) \right) \\ &= \max_{\boldsymbol{x}\in\mathscr{X}} \left(\left(1 - 2\left(\frac{L}{M}\right)^2\right) \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 + 2\left(\frac{L}{M}\right)^2 \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z}) - \nabla f(\boldsymbol{x})\|^2 \right) \end{aligned}$$

It holds that

1. $\frac{L}{M} \leq \frac{L}{m}$
2. $-\frac{L}{M} \leq -\frac{m}{M}$

Our estimates of L and M produce upper bounds and our estimate of m produces a lower bound. Therefore, we can upper bound both of the above quantities: $\frac{L}{m}$ and $-\frac{m}{M}$.

Returning back to the i index, this suggests that we set

$$\begin{aligned} \tilde{A}_i = \max_{\boldsymbol{x}\in\mathscr{X}} \Bigg(& \left(1 - 2\left(\frac{\hat{m}_{i-1} + t_{i-1}}{\hat{M}_{i-1} - t_{i-1}}\right)^2\right) \frac{1}{K_i} \sum_{k=1}^{K_i} \|\boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k))\|^2 \\ & + 2\left(\frac{\hat{L}_{i-1} + t_{i-1}}{\hat{m}_{i-1} - t_{i-1}}\right)^2 \frac{1}{K_i - 1} \sum_{k=1}^{K_i} \left\| \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k)) - \frac{1}{K_i} \sum_{j=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(j)) \right\|^2 \Bigg) \end{aligned}$$

We will show later that for i large enough $\hat{L}_{i-1} + t_{i-1} \geq L$, $\hat{M}_{i-1} - t_{i-1} \geq M$, and $\hat{m}_{i-1} - t_{i-1} \leq m \leq M$. Therefore, for i large enough, it holds that

$$\begin{aligned} \tilde{A}_i \geq \max_{\boldsymbol{x}\in\mathscr{X}} \Bigg(& \left(1 - 2\left(\frac{L}{M}\right)^2\right) \frac{1}{K_i} \sum_{k=1}^{K_i} \|\boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k))\|^2 \\ & + 2\left(\frac{L}{M}\right)^2 \frac{1}{K_i - 1} \sum_{k=1}^{K_i} \left\| \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(k)) - \frac{1}{K_i} \sum_{j=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}, \boldsymbol{z}_i(j)) \right\|^2 \Bigg) \end{aligned}$$

Then for i large enough, it holds that

$$\begin{aligned}
&\mathbb{E}[\tilde{A} \mid \mathscr{F}_{i-1}] \\
&\quad \geq \mathbb{E}\left[\max_{\boldsymbol{x}\in\mathscr{X}}\left(\left(1-2\left(\frac{L}{M}\right)^2\right)\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))\|^2\right.\right. \\
&\qquad\qquad \left.\left.+2\left(\frac{L}{M}\right)^2\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(j))\right\|^2\right)\right] \\
&\quad = \max_{\boldsymbol{x}\in\mathscr{X}}\mathbb{E}\left[\left(1-2\left(\frac{L}{M}\right)^2\right)\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))\|^2\right. \\
&\qquad\qquad \left.+2\left(\frac{L}{M}\right)^2\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(j))\right\|^2\right] \\
&\quad = \max_{\boldsymbol{x}\in\mathscr{X}}\left(\left(1-2\left(\frac{L}{M}\right)^2\right)\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)\|^2+2\left(\frac{L}{M}\right)^2\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)-\nabla f_i(\boldsymbol{x})\|^2\right) \\
&\quad = \max_{\boldsymbol{x}\in\mathscr{X}}\left(\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)\|^2-2\left(\frac{L}{M}\right)^2\|\nabla f_i(\boldsymbol{x})\|^2\right) \\
&\quad = A
\end{aligned}$$

Alternatively, we can consider the estimate

$$\begin{aligned}
\tilde{A}_i = \max_{\boldsymbol{x}\in\mathscr{X}}&\left(\left(1-2\left(\frac{\hat{L}_{i-1}+t_{i-1}}{\hat{M}_{i-1}-t_{i-1}}\right)^2\right)\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))\|^2\right. \\
&\left.+2\left(\frac{\hat{L}_{i-1}+t_{i-1}}{\hat{M}_{i-1}-t_{i-1}}\right)^2\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(k))-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i(j))\right\|^2\right)
\end{aligned}$$

which consists of plugging in our estimates of L and M into (D.5). We have no guarantees for this method, but it seems to work well in practice.

D.2.2 Search Method

For this estimate, we assume that there is a convex region $\mathscr{G} \subset \mathbb{R}^2_+$ such that for any $(A,B) \in \mathscr{G}$, it holds that

$$\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}_i)\|^2 \leq A + B\|\boldsymbol{x}-\boldsymbol{x}_i^*\|^2 \quad \forall \boldsymbol{x} \in \mathscr{X}, \forall n \geq 1$$

and for all $t_A, t_B > 0$

$$(A,B) \in \mathscr{G} \;\; \Rightarrow \;\; (A+t_A, B+t_B) \in \mathscr{G}$$

In light of the discussion at the beginning of this appendix, this is a reasonable assumption. We again drop the i index again for convenience. Due to the following bound [29]

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \geq \frac{1}{M}\|\nabla f(\boldsymbol{x})\|$$

it holds that

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 \leq A + \frac{B}{M^2}\|\nabla f(\boldsymbol{x})\|^2 \;\; \Rightarrow \;\; \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 \leq A + B\|\boldsymbol{x} - \boldsymbol{x}^*\|^2$$

Therefore if we can find a pair (A,B) such that

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 \leq A + \frac{B}{M^2}\|\nabla f(\boldsymbol{x})\|^2$$

holds, then the pair (A,B) also works for the gradient condition. In practice, we must use an estimate $\hat{M}_i$ of M that satisfies

$$\hat{M}_i + t_i \geq M$$

for all i large enough almost surely. This in turn implies that

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 \leq A + \frac{B}{(\hat{M}_i + t_i)^2}\|\nabla f(\boldsymbol{x})\|^2 \;\; &\Rightarrow \;\; \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 \leq A + \frac{B}{M^2}\|\nabla f(\boldsymbol{x})\|^2 \\ &\Rightarrow \;\; \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\|^2 \leq A + B\|\boldsymbol{x} - \boldsymbol{x}^*\|^2\end{aligned}$$

Therefore, using the estimate $\hat{M}_i$ in place of M produces more conservative estimates. Thus, we develop the search estimate of this section using the true M.

We have

$$\mathbb{E}\left[\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(k))\|^2\right] = \mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z})\|^2$$

and

$$\mathbb{E}\left[\frac{1}{K_i-1}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(k))\|^2-\frac{1}{K_i}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(k))-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(j))\right\|^2\right]$$
$$=\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z})\|^2-\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z})-\nabla f_i(\boldsymbol{x})\|^2$$
$$=\|\nabla f_i(\boldsymbol{x})\|^2$$

Define the function

$$\psi_{\boldsymbol{x}}(A,B)=A+\frac{B}{M^2}\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(k))\|^2\right.$$
$$\left.-\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(k))-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(j))\right\|^2\right)$$
$$-\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z}(k))\|^2$$

We have the following lemma that characterizes the performance of a simple estimate of $(\tilde{A},\tilde{B})$.

Lemma 23. *Suppose that we always choose $\tilde{A}_i$ and $\tilde{B}_i$ such that $\min_{\boldsymbol{x}\in\mathscr{X}}\psi_{\boldsymbol{x}}(\tilde{A},\tilde{B})\geq 0$. Then it holds that*

$$\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{x},\boldsymbol{z})\|^2\leq\mathbb{E}[\tilde{A}_i]+\mathbb{E}[\tilde{B}_i]\|\boldsymbol{x}-\boldsymbol{x}^*\|^2\quad\forall\boldsymbol{x}\in\mathscr{X}$$

Proof. First, by the monotone convergence theorem, it holds that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\tilde{A}+\frac{\tilde{B}}{M^2}\|\nabla f(\boldsymbol{x})\|^2-\mathbb{E}\|\boldsymbol{g}\left(\boldsymbol{x},\boldsymbol{z}\right)\|^2\right)-\psi_{\boldsymbol{x}}(\tilde{A},\tilde{B})\right] \\
&=\mathbb{E}\left[\frac{\tilde{B}}{M^2}\left(\|\nabla f(\boldsymbol{x})\|^2\right.\right. \\
&\quad\left.\left.-\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)\|^2-\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(j)\right)\right\|^2\right)\right)\right] \\
&=\mathbb{E}\left[\lim_{q\to\infty}\frac{\tilde{B}\vee\frac{1}{q}}{M^2}\left(\|\nabla f(\boldsymbol{x})\|^2\right.\right. \\
&\quad\left.\left.-\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)\|^2-\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(j)\right)\right\|^2\right)\right)\right] \\
&=\lim_{q\to\infty}\mathbb{E}\left[\frac{\tilde{B}\vee\frac{1}{q}}{M^2}\left(\|\nabla f(\boldsymbol{x})\|^2\right.\right. \\
&\quad\left.\left.-\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)\|^2-\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(j)\right)\right\|^2\right)\right)\right] \\
&\geq\limsup_{q\to\infty}\frac{1}{qM^2}\mathbb{E}\left[\|\nabla f_i(\boldsymbol{x})\|^2\right. \\
&\quad\left.-\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)\|^2-\frac{1}{K_i-1}\sum_{k=1}^{K_i}\left\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(k)\right)-\frac{1}{K_i}\sum_{j=1}^{K_i}\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}(j)\right)\right\|^2\right)\right] \\
&=0
\end{aligned}
$$

This in turn implies the following chain of inequalities

$$
\begin{aligned}
&\mathbb{E}[\tilde{A}_i]+\mathbb{E}[\tilde{B}_i]\|\boldsymbol{x}-\boldsymbol{x}^*\|^2-\mathbb{E}\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}\right)\|^2 \\
&\quad=\mathbb{E}\left[\tilde{A}_i+\tilde{B}_i\|\boldsymbol{x}-\boldsymbol{x}^*\|^2-\mathbb{E}\|\boldsymbol{g}\left(\boldsymbol{x},\boldsymbol{z}\right)\|^2\right] \\
&\quad\geq\mathbb{E}\left[\tilde{A}_i+\frac{\tilde{B}_i}{M^2}\|\nabla f_i(\boldsymbol{x})\|^2-\mathbb{E}\|\boldsymbol{g}_i\left(\boldsymbol{x},\boldsymbol{z}\right)\|^2\right] \\
&\quad\geq\mathbb{E}\left[\psi_{\boldsymbol{x}}(\tilde{A}_i,\tilde{B}_i)\right] \\
&\quad\geq\mathbb{E}\left[\min_{\boldsymbol{x}\in\mathscr{X}}\psi_{\boldsymbol{x}}(\tilde{A}_i,\tilde{B}_i)\right] \\
&\quad\geq 0
\end{aligned}
$$

□

This lemma shows that the pair $(\mathbb{E}[\tilde{A}_i],\mathbb{E}[\tilde{B}_i])\in\mathscr{G}$. From the convexity assump-

tion, it holds that

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\tilde{A}_i],\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\tilde{B}_i]\right)\in\mathscr{G}$$

To find $(\tilde{A}_i,\tilde{B}_i)$ that satisfy $\min_{\boldsymbol{x}\in\mathscr{X}}\psi_{\boldsymbol{x}}(\tilde{A},\tilde{B})\geq 0$, we start out with $(\tilde{A}_i(0),\tilde{B}_i(0))$ and growth rates $\alpha,\beta>1$. We then set

$$\begin{aligned}\tilde{A}_i(k) &= \alpha\tilde{A}_i(k-1)\\ \tilde{B}_i(k) &= \beta\tilde{B}_i(k-1)\end{aligned}$$

and stop when

$$\min_{\boldsymbol{x}\in\mathscr{X}}\psi_{\boldsymbol{x}}(\tilde{A}_i(k),\tilde{B}_i(k))\geq 0$$

This minimum can be evaluated using numerical optimization methods. We are then guaranteed that $(\mathbb{E}\tilde{A}_i(k),\mathbb{E}\tilde{B}_i(k))\in\mathscr{G}$

D.2.3 Finite Point Approximation

Suppose that we select N points $\boldsymbol{x}(1),\ldots,\boldsymbol{x}(N)\in\mathscr{X}$. We want to find A and B such that

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}(j),\boldsymbol{z})\|^2\leq A+B\|\boldsymbol{x}(j)-\boldsymbol{x}^*\|^2\quad j=1,\ldots,N$$

The following implication holds

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}(j),\boldsymbol{z})\|^2\leq A+\frac{B}{M^2}\|\nabla f(\boldsymbol{x}(j))\|^2\\ \Rightarrow\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}(j),\boldsymbol{z})\|^2\leq A+B\|\boldsymbol{x}(j)-\boldsymbol{x}^*\|^2\end{aligned}$$

As discussed in section D.2.2, it follows that plugging in the estimate $\hat{M}_i$ in place of M produces larger estimates of (A,B), so plugging in $\hat{M}_i$ in place of M does not negatively affect the methods of this section.

We look for (A,B) such that

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}(j),\boldsymbol{z})\|^2\leq A+\frac{B}{M^2}\|\nabla f(\boldsymbol{x}(j))\|^2\quad j=1,\ldots,N$$

Define

$$s_i(j)\triangleq\frac{1}{K_i}\sum_{k=1}^{K_i}\|\boldsymbol{g}_i(\boldsymbol{x}(j),\boldsymbol{z}_i(k))\|^2$$

and

$$d_i(j) \triangleq g_i(j) - \frac{1}{K_i - 1} \sum_{k=1}^{K_i} \left\| \boldsymbol{g}_i(\boldsymbol{x}(j), \boldsymbol{z}_i(k)) - \frac{1}{K_i} \sum_{p=1}^{K_i} \boldsymbol{g}_i(\boldsymbol{x}(p), \boldsymbol{z}_i(p)) \right\|^2$$

We want to find (A, B) such that

$$s_i(j) \leq A + \frac{B}{(\hat{M}_{i-1} + t_{i-1})^2} d_i(j) \quad j = 1, \ldots, N$$

Suppose that we are given a function $\phi(A, B)$ that controls the size of (A, B). For example, we may have $\phi(A, B) = \frac{1}{2}A^2 + \frac{1}{2}B^2$ or $\phi(A, B) = \lambda A^2 + (1 - \lambda)B^2$ with $0 < \lambda < 1$. We solve

$$\begin{aligned} \underset{\tilde{A}_i, \tilde{B}_i}{\text{minimize}} \quad & \phi(\tilde{A}_i, \tilde{B}_i) \\ \text{subject to} \quad & s_i(j) \leq \tilde{A}_i + \frac{\tilde{B}_i}{(\hat{M}_{i-1} + t_{i-1})^2} d_i(j), \quad j = 1, \ldots, N \\ & \tilde{A}_i \geq 0 \, , \, \tilde{B}_i \geq 0 \end{aligned}$$

to generate approximate $(\tilde{A}_i, \tilde{B}_i)$. This allows us to choose to emphasize large or smaller A or B.

D.3 Combining One-Step Estimates and ρ Estimation

One issue in parameter estimation is that there may be some dependencies among the various estimates. For example, the estimate of A relies on an estimate of L and m. However, the actual estimates we compute, $\tilde{A}_i$, depend on $\hat{L}_{i-1}$ and $\hat{m}_{i-1}$, which may not be above L and below m respectively. Fortunately, due to controlling the conditional expectations of our estimates we can argue that the averaged estimates eventually upper bound the desired quantity. First, we present a result showing that if we plug in the true parameters that our estimates work.

Lemma 24. *Suppose that we want to estimate ϕ^* by combining one-step estimates $\phi_i(\pi^*)$ where π^* are the true parameters on which the estimate ϕ_i depends and the following conditions hold:*

1. $|\phi_i(\pi^*)| \leq C$
2. $\mathbb{E}[\phi_i(\pi^*) \mid \mathscr{F}_{i-1}] \geq \phi^*$

3. $\sum_{n=1}^{\infty} \exp\left\{-\frac{2nt_n^2}{C^2}\right\} < +\infty$

Then for all n large enough, it holds that

$$\frac{1}{n}\sum_{i=1}^{n} \phi_i(\pi^*) + t_n \geq \phi^*$$

almost surely.

Proof. Since $|\phi_i(\pi^*)| \leq C$, by applying Lemma 17 it holds that

$$\mathbb{E}\left[e^{s(\phi_i(\pi^*) - \mathbb{E}[\phi_i(\pi^*) \mid \mathscr{F}_{i-1}])} \mid \mathscr{F}_{i-1}\right] \leq \exp\left\{\frac{1}{2}\frac{C^2}{4}s^2\right\}$$

Then by Lemma 16, it holds that

$$\begin{aligned}
\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_i(\pi^*) < \phi^* - t_n\right\} &= \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_i(\pi^*) < \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\phi_i(\pi^*) \mid \mathscr{F}_{i-1}\right] - t_n\right\} \\
&\leq \exp\left\{-\frac{2nt_n^2}{C^2}\right\}
\end{aligned}$$

Since it holds that

$$\sum_{n=1}^{\infty}\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_i(\pi^*) < \phi^* - t_n\right\} \leq \sum_{n=1}^{\infty}\exp\left\{-\frac{2nt_n^2}{C^2}\right\} < +\infty$$

by the Borel-Cantelli lemma, it follows that for all n large enough

$$\frac{1}{n}\sum_{i=1}^{n}\phi_i(\pi^*) + t_n \geq \phi^*$$

□

Since our estimates of m and M do not depend on any parameters π, this lemma shows that both of these estimates averaged do lower and upper bound m and M respectively. We bootstrap from this result to show that the estimates of L, B and A work using Lemma 25. The random variables X_n is $o_{\mathbb{P}}(1)$ if

$$\lim_{n\to\infty}\mathbb{P}\{|X_n| \geq t\} = 0 \quad \forall t > 0$$

Lemma 25. *Suppose that we want to estimate ϕ^* by combining one-step estimates $\phi_i(\pi_i)$ where π_i are the estimates of the parameters on which the estimate*

ϕ_i *depends and the following hold:*

1. $|\phi_i(\pi)| \leq C$
2. *For all n large enough* $\pi_n \geq \pi^*$ *almost surely*
3. $\pi \leq \tilde{\pi} \;\Rightarrow\; \phi_i(\pi) \leq \phi_i(\tilde{\pi})$
4. *For appropriate sequences* t_n, $\frac{1}{n}\sum_{i=1}^n \phi_i(\pi^*) + t_n \geq \phi^*$

Then for all n large enough, it holds that

$$\frac{1}{n}\sum_{i=1}^n \phi_i(\pi_i) + t_n \geq \phi^* + o_{\mathbb{P}}(1)$$

almost surely.

Proof. There exists a finite almost surely random variable $\tilde{N}$ such that

$$n \geq \tilde{N} \;\Rightarrow\; \pi_i \geq \pi^*$$

It holds that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \phi_i(\pi_i) &= \frac{1}{n}\sum_{i=1}^{\tilde{N}-1} \phi_i(\pi_i) + \frac{1}{n}\sum_{i=\tilde{N}}^n \phi_i(\pi_i) \\
&= \frac{1}{n}\sum_{i=1}^{\tilde{N}-1} \phi_i(\pi_i) + \frac{1}{n}\sum_{i=\tilde{N}}^n \phi_i(\pi^*) \\
&= \frac{1}{n}\sum_{i=1}^{\tilde{N}-1} (\phi_i(\pi_i) - \phi_i(\pi^*)) + \frac{1}{n}\sum_{i=1}^n \phi_i(\pi^*)
\end{aligned}$$

By the boundedness of $\phi(\pi)$, this implies that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \phi_i(\pi_i) + t_n &= \left(\frac{1}{n}\sum_{i=1}^n \phi_i(\pi^*) + t_n\right) + o_{\mathbb{P}}(1) \\
&\geq \phi^* + o_{\mathbb{P}}(1)
\end{aligned}$$

□

Using Lemma 25, we have constructed estimates $\hat{\psi}_n$ such that for all n large enough it holds that

$$\hat{\psi}_n + t_n \mathbf{1} + o_{\mathbb{P}}(1) \geq \psi^*$$

almost surely with $\mathbf{1}$ a vector all ones. Therefore, by assumption for all n large enough it holds that

$$b(d_0, K, \psi^*) \leq b(d_0, K, \hat{\psi}_n + t_n \mathbf{1} + o_{\mathbb{P}}(1))$$

D.4 Experiment

We carry out a simple experiment based on the simulation in section 3.3 and section 4.3.1 to test our various estimates of the strong convexity parameter m and the gradient parameters (A, B). In estimating the strong convexity parameter, the Hessian method is from section D.1.2 and the ratio and finite ratio methods are from section D.1.3. In estimating the gradient parameter method, the search method is from section D.2.2, the finite method is from section D.2.3, the L estimation methods max and ratio are from sections D.2.1 and D.2.1 respectively.

Figure D.1 shows the estimate of the strong convexity parameter m. We see that all methods eventually produce estimates that lower bound m. Note that even when the estimates are above the true value m, the gap is small. Figure D.2 and Figure D.3 show estimates of the gradient parameters A and B. The L estimation methods for A and B produce estimates quite close to the values computed from section 3.3. As noted in the beginning of this appendix, the choice of A and B is not unique. The fixed value corresponds to the values of A and B in section 3.3. Finally, we note that the fixed values of A and B from section 3.3 use the inequality

$$(a+b)^2 \leq 2a^2 + 2b^2$$

so there may be some slack in the fixed values computed using section 3.3. Due to this slackness, the smaller values of A and B chosen by the search method and the fixed method may be better choices of the gradient parameters.

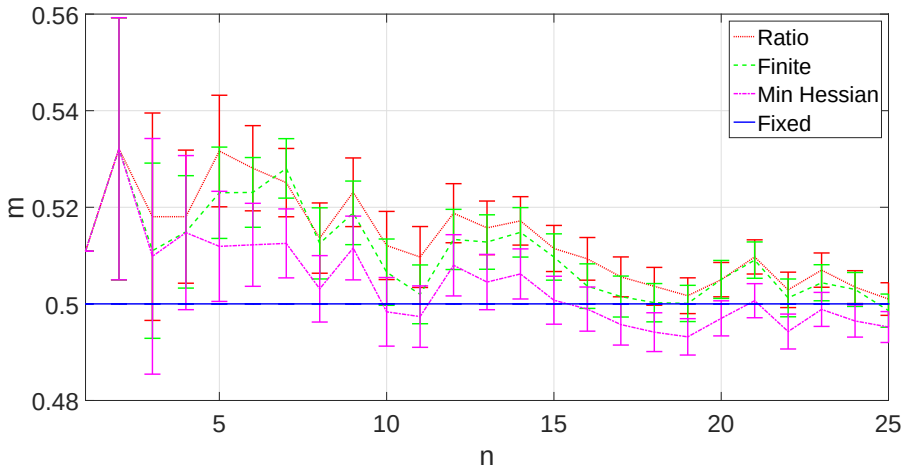


Figure D.1: Strong convexity parameter - *m*

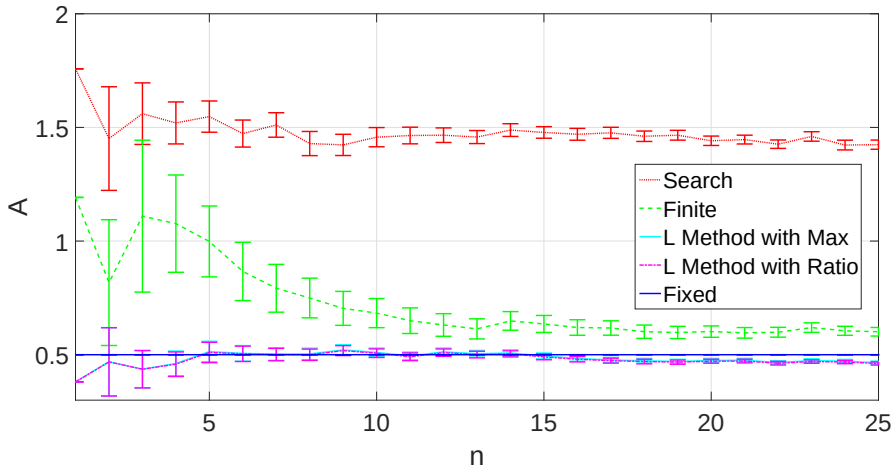


Figure D.2: Gradient parameters - *A*

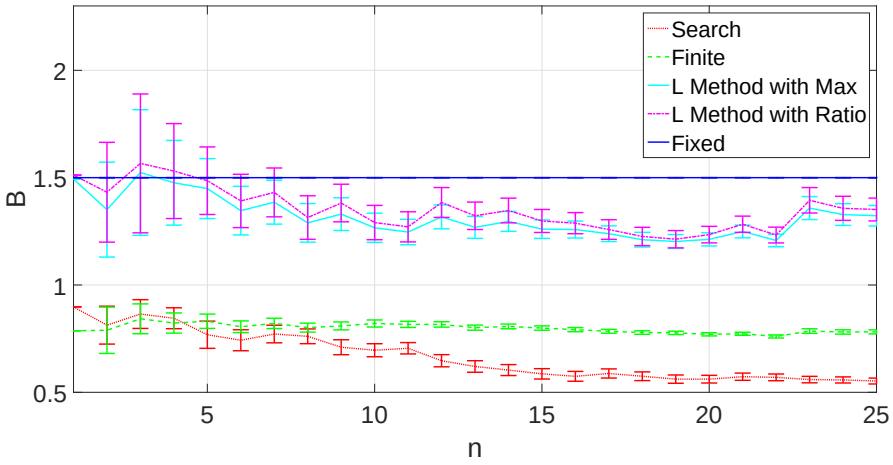


Figure D.3: Gradient parameters - *B*

References

[1] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, ser. MIT press series on economic learning and social evolution. Cambridge (Mass.), London: The MIT Press, 1998.

[2] C. Wilson, V. V. Veeravalli, and A. Nedić, "Dynamic stochastic optimization," in *CDC*, 2014.

[3] C. Wilson, V. Veeravalli, and A. Nedić, "Adaptive sequential optimization - part I: Known change in minimizers," submitted to *IEEE Transactions on Automatic Control*, 2016.

[4] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, N.Y., USA: Cambridge University Press, 2006.

[5] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[6] J. Duchi and Y. Singer, "Efficient online and batch learning using forward backward splitting," *Journal of Machine Learning Research*, vol. 10, pp. 2899–2934, 2009.

[7] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, pp. 169–192, 2007.

[8] E. H. P. Bartlett and A. Rakhlin, "Adaptive online gradient descent," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA, USA: MIT Press, 2008, vol. 20, pp. 65–72.

[9] S. Shalev-Shwartz and S. Kakade, "Mind the duality gap: Logarithmic regret algorithms for online optimization," in *Advances in Neural Information Processing Systems*, D. Koller, D.Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21, MIT Press, 2009, pp. 1457–1464.

[10] S. Shalev-Shwartz and Y. Singer, "Convex repeated games and Fenchel duality," in *Advances in Neural Information Processing Systems*, vol. 19, MIT Press, Cambridge, MA, USA, 2006, pp. 1265–1271.

[11] S. Shalev-Shwartz and Y. Singer, "Logarithmic regret algorithms for strongly convex repeated games," 2007, in The Hebrew University Leibniz Center Technical Reports.

[12] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," Microsoft, Tech. Rep. no. MSR-TR-2010-23, March 2010.

[13] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 2003, pp. 928–936.

[14] A. Rakhlin and K. Sridharan, "Online learning with predictable sequences," in *arXiv:1208.3728*, Aug. 2012.

[15] C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu, "Online optimization with gradual variations," in *COLT*, 2012.

[16] A. Dontchev and R. Rockafellar, *Implicit Functions and Solution Mappings: A View from Variational Analysis*. New York, New York: Springer, 2009.

[17] N. Takahashi, I. Yamada, and A. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4795–4810, 2010.

[18] I. Yamada and N. Ogura, "Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions," *Numerical Functional Analysis and Optimization*, vol. 25, no. (7-8), pp. 593–617, 2005.

[19] K. Slavakis, I. Yamada, and N. Ogura, "The adaptive projected subgradient method over the fixed point set of strongly attracting nonexpansive mappings," *Numerical Functional Analysis and Optimization*, vol. 27, no. (7-8), pp. 905–930, 2006.

[20] H. Kushner and J. Yang, "Analysis of adaptive step size SA algorithms for parameter tracking," in *Proceedings of the Conference on Decision and Control*, vol. 1, 1994, pp. 730–737.

[21] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[22] A. Sayed, *Adaptive Filters*. Hoboken, New Jersey, USA: Wiley & Sons, Inc., 2008.

[23] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, vol. 12, no. 656-704, p. 3, 2009.

[24] B. Polyak and A. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, pp. 838–855, 1992.

[25] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429+, June 1997.

[26] C. Wilson and V. Veeravalli, "Adaptive sequential optimization with applications to machine learning," in *ICASSP*, Mar. 2016.

[27] C. Wilson and V. Veeravalli, "Adaptive sequential optimization with applications to machine learning," *arXiv:1509.07422*, Sep. 2015.

[28] C. Wilson, V. Veeravalli, and A. Nedić, "Adaptive sequential optimization - part II: Unknown change in minimizers," submitted to *IEEE Transactions on Automatic Control*, 2016.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[30] R. Antonini and Y. Kozachenko, "A note on the asymptotic behavior of sequences of generalized subgaussian random vectors," *Random Op. and Stoch. Equ.*, vol. 13, pp. 39–52, 2005.

[31] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2011.

[32] S. Haykin, *Adaptive Filter Theory*. Springer, 2002.

[33] B. Sriperumbudur, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, pp. 1550–1599, 2012.

[34] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[35] C. Wilson and V. Veeravalli, "Adaptive sequential optimization with applications to machine learning," 2016.

[36] A. Agarwal, H. Daumé, and S. Gerber, "Learning multiple tasks using manifold regularization," in *NIPS*, 2011, pp. 46–54.

[37] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 109–117.

[38] Y. Zhang and D. Yeung, "A convex formulation for learning task relationships in multi-task learning," *CoRR*, vol. abs/1203.3536, 2012.

[39] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[40] A. Agarwal, A. Rakhlin, and P. Bartlett, "Matrix regularization techniques for online multitask learning," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2008-138, Oct 2008.

[41] Z. Towfic, J. Chu, and A. Sayed, "Online distributed online classification in the midst of concept drifts," *Neurocomputing*, vol. 112, pp. 138–152, 2013.

[42] C. Tekin, L. Canzian, and M. van der Schaar, "Context adaptive big data stream mining," in *Allerton Conference*, 2014, pp. 46–54.

[43] T. Dietterich, "Machine learning for sequential data: A review," in *Structural, Syntactic, and Statistical Pattern Recognition*, 2002, pp. 15–30.

[44] T. Fawcett and F. Provost, "Adaptive fraud detection." *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 291–316, 1997.

[45] N. Qian and T. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, pp. 865–884, Aug 1988.

[46] Y. Bengio and P. Frasconi, "Input-output HMM's for sequence processing," *IEEE Transactions on Neural Networks*, vol. 7(5), pp. 1231–1249, 1996.

[47] A. Wächter and L. T. Biegler, "On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, 2006.

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001.

[49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, 2006.

[50] "Panel study of income dynamics: public use dataset," Univ. of Michigan Survey Research Center, 2015. [Online]. Available: https://psidonline.isr.umich.edu/

[51] K. Murphy and F. Welch, "Empirical age-earnings profiles," *Journal of Labor Economics*, vol. 8, no. 2, pp. 202–29, 1990.

[52] "General social survey," Univ. of Chicago National Opinion Research Center, 2015. [Online]. Available: http://gss.norc.org/

[53] N. Papadatos, "Distribution and expectation bounds on order statistics from possibly dependent variates," *Statistics & Probability Letters*, vol. 54, no. 1, pp. 21–31, 2001.

[54] R. Öten and R. J. P. de Figueiredo, "Adaptive alpha-trimmed mean filters under deviations from assumed noise model," *IEEE Trans. on Image Processing*, pp. 627–639, 2004.

[55] S. M. Stigler, "The asymptotic distribution of the trimmed mean," *Annals of Statistics*, vol. 1, pp. 472–477, 1973.

[56] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," in *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[57] B. C. Levy, *Principles of signal detection and parameter estimation*. New York: Springer, 2008.

[58] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

[59] E. Even-Dar, Y. Mansour, and U. Nadav, "On the convergence of regret minimization dynamics in concave games," in *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, ser. STOC '09. New York, NY, USA: ACM, 2009. [Online]. Available: http://doi.acm.org/10.1145/1536414.1536486 pp. 523–532.

[60] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Nonlinear Programming*. Stanford: Stanford University Press, 1958.

[61] S. Shamma and G. Arslan, "Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria," *IEEE Trans Automatic Control*, vol. 50, no. 3, pp. 312–327, Mar 2005.

[62] H. T. Q. Zhu and T. Başar, "Heterogeneous learning in zero-sum stochastic games with incomplete information," in *IEEE CDC*, 2010, pp. 219–224.

[63] H. T. Q. Zhu and T. Başar, "Hybrid learning in stochastic games and its applications in network security," in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, F. L. Lewis and D. Liu, Eds. IEEE Press/Wiley.

[64] M. LiCalzi, "Fictitious play by cases," *Games and Economic Behavior*, vol. 11, no. 1, pp. 64–89, 1995.

[65] M. Sion, "On general minimax theorems," *Pac. J. Math*, pp. 171–176, 1958.

[66] V. M. Kapinas, S. K. Mihos, and G. K. Karagiannidis, "On the monotonicity of the generalized Marcum and Nuttall Q-functions," *IEEE Transactions on Communications*, vol. 55, no. 8, pp. 3701–3710, Aug 2009.

[67] W. Gao, Y. Wang, and A. Homaifa, "Discrete-time variable structure control systems," *IEEE Transactions on Industrial Electronics*, vol. 42, no. 2, pp. 117–122, Apr 1995.

[68] J.-H. Kim and D. il Cho, "Discrete-time variable structure control using recursive switching function," in *Proceedings of the American Control Conference*, vol. 2, 2000, pp. 1113–1117 vol.2.

[69] N. Stein, A. Ozdaglar, and P. Parrilo, "Separable and low-rank continuous games," *International Journal of Game Theory*, pp. 475–504, 2008.

[70] J. Ratliff, "A folk theorem sampler," 1996. [Online]. Available: http://www.virtualperfection.com/gametheory/

[71] D. de Farias and N. Megiddo, "How to combine expert (or novice) advice when actions impact the environment," in *In Advances in Neural Information Processing Systems 16*, 2004.

[72] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, pp. 1574–1609, 2009.

[73] F. Bach and E. Moulines, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *Advances in Neural Information Processing Systems (NIPS)*, Spain, 2011.

[74] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[75] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Norwell, Massachusetts, USA: Kluwer Academic Publishers, 2004.

[76] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[77] S. Janson, "Large deviations for sums of partly dependent random variables," *Random Structures Algorithms*, vol. 24, pp. 234–248, 2004.

[78] J. Kennan, "Uniqueness of positive fixed points for increasing concave functions on rn: An elementary result," *Review of Economic Dynamics*, vol. 4, pp. 893–899, 2001.

[79] C. H. Edward, *Advanced Calculus of Several Variables*. New York: Dover Publications, 1994.

[80] A. Granas and J. Dugundji, *Fixed Point Theory*. Springer-Verlag, 2003.

[81] L. Trefethen, *Numerical Linear Algebra*. SIAM, 1997.