THE UNIVERSITY OF CHICAGO

CONSTRAINED AND LOCALIZED FORMS OF STATISTICAL MINIMAX THEORY

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

DEPARTMENT OF STATISTICS

BY
YUANCHENG ZHU

CHICAGO, ILLINOIS
AUGUST 2016

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I must offer my most sincere gratitude to my advisor, John Lafferty. He is so far the best mentor in my life, academically and personally. I thank him for his patience and encouragement, which has provided a phenomenal environment for me and has helped me overcome obstacles and frustrations in my research over the past four years. His erudition and deep insight never stops to surprise and inspire me, and his guidance about how to approach research, how to write, how to give a talk and how to teach has been invaluable to me. This thesis wouldn't have been possible without his support and guidance. I am extremely lucky and proud to have become his student.

I am grateful to my committee members, Rina Foygel Barber, Steven Lalley, and Tracy Ke. My conversation and collaboration with them has alway been delightful and inspiring. Their feedback and suggestion on research topics as well as advice on research approaches and career choice has been really helpful. Thanks as well to John Duchi and Sabyasachi Chatterjee for their invaluable contribution and collaboration on the second part of this thesis. I would also like to thank Hoyt Long and Richard Jean So, with whom I got to practice the statistical methodologies that I learned to answer meaningful questions from other disciplines.

My appreciation also goes to my fellow students and friends, including Zhe, Wei, Siqi, Qinqing, Liwen, Matt, Dinah and Irene, for many conversations, both academic and cathartic. It is because of them that my life in Hyde Park has become so colorful and memorable.

Finally, I take this opportunity to express the profound gratitude from my deep heart to my beloved parents, Haiyang and Hong, for their love and unconditional support. Thank you!

ABSTRACT

Statistical minimax theory is a fundamental quantity used to assess the difficulty of various statistical tasks. We consider two variants on traditional minimax theory to alleviate some of its deficiencies. The first variant, a constrained form of minimax theory, puts computational constraints on the procedures and leads to minimax complexities that are achievable by computationally efficient methods. We illustrate this by an example of nonparametric estimation with storage constraint. We show how the minimax risk varies with the number of bits that is allowed to be used to represent the estimate. This establishes the Pareto optimal minimax tradeoff between storage and risk under quantization constraints for Sobolev spaces. As for the second variant, we extend the traditional minimax analysis by introducing a localized form of minimax complexity for individual instances. The formulation is based on the "hardest local alternative." As an example, we derive the local minimax complexity for stochastic optimization of convex functions. The local minimax complexity is expressed in terms of a localized and computational analogue of the modulus of continuity. We show how the computational modulus of continuity can be explicitly calculated in concrete cases, and relates to the curvature of the function at the optimum. We also prove a superefficiency result that demonstrates it is a meaningful benchmark, acting as a computational analogue of the Fisher information in statistical estimation. The nature and practical implications of the results are demonstrated in simulations.

CHAPTER 1

INTRODUCTION

I am prepared for the worst, but hope for the best. (Benjamin Disraeli)

Minimax theory acts as the cornerstone of many statistical analyses. Typically, it is used to quantify the hardness of a set of statistical problems, such as estimation, testing, and optimization. For any procedure that is designed to complete the particular statistical task, we look at its worst-case performance when applied to the set of problems. Then the procedure that has the best worst-case performance gives the minimax risk or complexity of the problem. Although such worst-case analyses have gained in popularity and been of great importance, their usage has also been criticized due to some deficiencies of the formulation.

1.1 Minimax Analyses

Before we proceed to present the modifications of minimax theory, we set the stage by giving a brief review of minimax analysis in its full generality. Suppose that there exists some data generating mechanism P drawn from $\mathcal{P}$. Such $\mathcal{P}$ could be a family of probability distributions, or a class of stochastic oracles. The data generating mechanism P, upon being queried, returns a random variable or vector. For example, when P is a probability distribution, then a total number of n queries give us n samples from the distribution; when P is some stochastic oracle, it returns samples depending on its underlying parameter as well as the query input. We call a mapping $\theta : \mathcal{P} \to \Theta$ a parameter of P and our goal is to estimate the parameter $\theta(P)$ for the underlying P. We will simply write θ as the parameter when its dependence on P is clear from the context. Suppose that $\mathcal{A}_n$ is the collection of procedures that make up to n queries to the data generating mechanism P. In the most common case where P is a probability distribution, $\mathcal{A}_n$ is simply the set of estimators of θ measurable with respect to n i.i.d. data points. In order to assess the quality of any procedure, we

define $\text{err} : \mathcal{A}_n \times \Theta \to [0, \infty)$ to be some error measure of using $A \in \mathcal{A}_n$ to estimate θ. The associated risk is simply the expected error $\mathbb{E}_P \text{ err}(A, \theta(P))$, where the expectation is taken with respect to P and possibly the randomness in the procedure A. We can then rank the procedures by their worst-case risk with P ranging in $\mathcal{P}$, i.e., $\sup_{P \in \mathcal{P}} \mathbb{E}_P \text{ err}(A, \theta(P))$. The *minimax risk* is defined as this worst-case risk of the optimal procedure

$$R_n(\mathcal{P}) = \inf_{A \in \mathcal{A}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \text{ err}(A, \theta(P)). \tag{1.1}$$

When each $P \in \mathcal{P}$ can be indexed by the parameter θ, we will write

$$R_n(\Theta) = \inf_{A \in \mathcal{A}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \text{ err}(A, \theta). \tag{1.2}$$

Let us first consider two illustrative examples of this definition and related results from two different areas.

Example 1.1.1 (Normal means estimation)**.** Let n be a positive integer. Suppose that

$$X_i \overset{\text{ind.}}{\sim} \mathcal{N}(\theta_i, \sigma_n^2) \text{ for } i = 1, 2, \ldots, n.$$

We assume the variance $\sigma_n^2 = \sigma^2/n$ is known and would like to estimate the means $\theta = (\theta_1, \ldots, \theta_n)$. Suppose that the mean vector is known to be contained in $\Theta_n(c) = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \theta_i^2 \leq c^2\}$, the ℓ_2 ball in $\mathbb{R}^n$ centered at the origin with radius c. This estimation problem corresponds to the afore-defined setup with a data generating mechanism P, which at the ith query returns $X_i \sim \mathcal{N}(\theta_i, \sigma_n^2)$, and $\mathcal{P}$ contains all such P with $\theta \in \Theta_n(c)$. Consider the ℓ_2 loss as our error measure, i.e., $\text{err}(\widehat{\theta}, \theta) = \|\theta - \widehat{\theta}\|^2$. Thus, the minimax risk for this estimation problem can be written as

$$R_n(\Theta_n(c)) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_n(c)} \mathbb{E}_\theta \|\theta - \widehat{\theta}\|_2^2. \tag{1.3}$$

The infimum is taken over all estimators $\widehat{\theta}$ that are measurable with respect to the data $(X_1, \ldots, X_n)$. This normal means model is a centerpiece of nonparametric estimation. It arises naturally when representing an estimator in terms of an orthogonal basis; see Brown and Low (1996a) and Johnstone (2015). Pinsker (1980) show that

$$\liminf_{n\to\infty} R_n(\Theta_n(c)) = \frac{\sigma^2 c^2}{\sigma^2 + c^2}. \tag{1.4}$$

Note that the maximum likelihood estimator, which, in this case, is the sample $(X_1, \ldots, X_n)$ itself has a risk of σ^2, regardless of θ. Therefore, the MLE is not minimax for the parameter space $\Theta_n(c)$. In fact, appropriate amount of shrinkage on the MLE, which trades bias for less variance, can be proved to be minimax optimal in this case.

A more interesting case is to estimate the mean vector from a Sobolev ellipsoid. A Sobolev ellipsoid of order m and radius c is defined as

$$\Theta(m, c) = \left\{ \theta \in \ell_2 : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{1.5}$$

with $a_j = (\pi j)^m$. Suppose that our model is still given by

$$X_j \sim \mathcal{N}(\theta_j, \sigma_n^2) \text{ for } j = 1, 2, \ldots \tag{1.6}$$

where $\sigma_n^2 = \sigma^2/n$. Note that a slight difference is that here we have an infinite sequence of data as well as mean components. Still, n is our effective sample size, as this formulation arises from settings where we have noisy evaluations of a function at n evenly spaced locations. The Sobolev ellipsoid comes up when we convert the observations to an orthonormal basis, and assume that the function satisfies certain smoothness conditions. The minimax risk for estimating the normal mean vector in a Sobolev ellipsoid can be then formulated as

$$R_n(\Theta(m, c)) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta(m,c)} \mathbb{E}_\theta \|\theta - \widehat{\theta}\|_2^2. \tag{1.7}$$

It is shown in Pinsker (1980) that

$$\liminf_{n\to\infty} n^{\frac{2m}{2m+1}} R_n(\Theta(m,c)) = \left(\frac{\sigma}{\pi}\right)^{\frac{2m}{2m+1}} c^{\frac{2}{2m+1}} \left(\frac{m}{m+1}\right)^{\frac{2m}{2m+1}} (2m+1)^{\frac{1}{2m+1}}. \tag{1.8}$$

This is referred to as Pinsker's Theorem, which characterizes the convergence rate and leading constant of the nonparametric estimation problem to a constant level. However, we must notice that here as the mean vector has infinite length, so the minimax optimal estimator could possibly take up very large or theoretically infinite storage.

Example 1.1.2 (First-order stochastic convex optimization)**.** Let $\mathcal{F}$ be a collection of Lipschitz convex functions defined on a compact convex set $\mathcal{C} \subset \mathbb{R}^d$. Given a function $f \in \mathcal{F}$, our goal is to find a minimum point, $x_f^* \in \arg\min_{x\in\mathcal{C}} f(x)$. However, our knowledge about f can only be gained through a first-order oracle P. The oracle, upon being queried with $x \in \mathcal{C}$, returns $f'(x) + \xi$, where $f'(x)$ is a subgradient of f at x and $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$. In our previous setup of minimax analyses, the corresponding family $\mathcal{P}$ in this case contains such stochastic oracles for functions f in $\mathcal{F}$. Consider optimization algorithms that make a total of n queries to this first-order oracle, and let $\mathcal{A}_n$ be the collection of all such algorithms. For $A \in \mathcal{A}_n$, the ith query point x_i is a random vector measurable with respect to the previous query points $x_1, \ldots, x_{i-1}$ and the previous query responses $Y_1, \ldots, Y_{i-1}$. Denote by $\widehat{x}_A$ the output of the algorithm A, which is the estimated minimum point of the underlying function. We can consider either a function value error, $\mathrm{err}(A, f) = f(\widehat{x}_A) - \inf_{x\in\mathcal{C}} f(x)$, or a point value error $\mathrm{err}(A, f) = \inf_{y\in\arg\min f(x)} \|y - \widehat{x}_A\|$. Either way, the minimax complexity can be defined as

$$R_n(\mathcal{F}) = \inf_{A\in\mathcal{A}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f \,\mathrm{err}(A, f). \tag{1.9}$$

The expectation $\mathbb{E}_f$ denotes the average with respect to the randomness introduced by the oracle and any additional randomness injected by the algorithm itself. The minimax risk $R_n(\mathcal{F})$ characterizes the hardness of the entire class $\mathcal{F}$. Nemirovsky and Yudin (1983) show that for the function value error, the minimax complexity $R_n(\mathcal{F})$ scales as $O(1/\sqrt{n})$ when $\mathcal{F}$

contains all Lipschitz convex functions. They also show that for the set of strongly convex functions the minimax complexity decreases at a faster rate $O(1/n)$. The classical results indicate that strongly convex functions are relatively easy to optimize.

1.2 Deficiencies and Variants

Although minimax analyses have gained in popularity and are of great importance in many fields, there are always some critiques against the usage of this notion as a characterization of the hardness of the task. Among all the critiques, the following two are probably the most common, one against its optimism, and the other, on the contrary, against its pessimism.

First, traditional minimax analyses are optimistic, in the sense that, any procedure is included for consideration as long as it is measurable with respect to the data, regardless of the computation, storage, or communication cost. Consequently, the optimal procedure that achieves the minimax risk can be practically infeasible, hence leaving the benchmark meaningless. For example, it can allow algorithms that scale arbitrarily fast with the problem dimension and sample size. It is of interest to consider procedures that have polynomial-time complexity and ask what is the optimal procedure amongst them. Moreover, it assumes that the algorithm, or the final estimator produced by the algorithm, can use an infinite amount of space for storage or representation. Oftentimes, such resources are limited by the storage space or the precision of the computation system. It is then again interesting to understand how much is lost if we limit our choice to those procedures with storage constraints. In the first part of the thesis, we will introduce a constrained form of minimax analyses, in which the infimum is taken over those procedures satisfying certain conditions. We will give a brief review in Chapter 2 before delving into a concrete examples. We build on the normal means estimation problem described in Example 1.1.1 and consider estimation problems with storage constraints in Chapter 3.

On the other hand, another criticism of the theory is that the minimax benchmark is too pessimistic. In fact, minimax analyses quantify the hardness of a family of problems. For

a particular problem that is of interest, it can belong to multiple families whose minimax complexities can be quite different. It is not immediately clear how well we should expect or hope to solve individual problems by applying such worst-case analyses. We consider a definition of "local" minimax risk by looking at the hardest local alternative. Such a two-point formulation becomes meaningful benchmark only if some criteria can be shown, such as achievability by adaptive algorithms, superefficiency, etc. We give a brief introduction of the formulation and the criteria in Chapter 4. In Chapter 5, we illustrate the application of the framework in the context of stochastic convex optimization.

Part I

Constrained Forms of Minimax Theory

CHAPTER 2

CONSTRAINED FORMS OF MINIMAX THEORY: INTRODUCTION AND RELATED WORK

Statistical minimax theory quantifies the hardness of many statistical tasks. It is based on a saddle point formulation where the statistician chooses a procedure to minimize the worst-case performance of the procedure when applied to the all the problem instances. The minimax risk or complexity thus defined can sometimes underestimate the real difficulty of the problem, as the procedure that achieves the minimax risk can be unrealistic. The only requirement on the procedure in the traditional minimax formulation is that it is measurable with respect to the data available at the time of making final decision. Consequently this allows the use of procedures that can be practically infeasible due to computational reasons. For example, the maximum likelihood estimator for some model with combinatorial assumptions such as sparsity has exponential computational complexity as it usually requires an enumeration of all combinations of features which scales exponentially with the dimension. Nonparametric estimation problems, on the other hand, usually take up very large amount of space to store and represent the estimates, due to the assumption of an infinite dimensional truth. It is then natural to ask if we limit ourselves to procedures that are computationally efficient, how much harder will the problem become. To put it more explicitly, we write the constrained form of minimax risk as

$$R_n(\mathcal{P}; B) = \inf_{A \in \mathcal{A}_n : \mathcal{C}(A) \le B} \sup_{P \in \mathcal{P}} \mathbb{E}_P \, \mathrm{err}(A, \theta(P))$$

where we use the notation from Chapter 1, and $\mathcal{C}(A)$ denotes the computational cost of procedure A, such as running time, number of floating point operations, number of bits used to store or construct the procedure, etc. Here we only consider the procedures whose computational cost fall within a budget B. It is of interest and importance to understand the role of the constraint and how the budget B influences the minimax risk.

The quest of the constrained forms of minimax analyses can be viewed as an approach to understand the tradeoff between statistical accuracy and computational efficiency. With the recent development of large-scale and high-dimensional statistical analyses, computation becomes too significant a component to be ignored, making it more urgent and important to understand this tradeoff. In the recent decade, there have been an increasing number of work devoted to characterize such tradeoff.

Many studies have been focusing on the tradeoff between statistical risk and computational runtime. One approach to examine the relationship is to study different procedures with different statistical accuracies and computational runtimes. Chandrasekaran and Jordan (2013) describe a computational framework based on convex relaxation, which does this in a principled way. Thus, to achieve a desired risk, various methods with different level of relaxation leads to different computational complexity and require different number of samples. Figure 2.1 is a replicate of a plot in Chandrasekaran and Jordan (2013), which characterizes the phenomenon. On the computation-sample plane, each point corresponds to a procedure that require a certain runtime and a certain number of samples to achieve the desired risk. We would expect that for a specific statistical task there is a feasible region on this plane. Designing a sequence of algorithms with various computation and sample complexities is like probing the feasible region. In addition to using convex relaxation to study the tradeoff, Bruer et al. (2014) analyze the tradeoff between time and data by aggressively smoothing the optimization problem, and Lucic et al. (2015) using k-means to describe the data. Other than such analyses using a particular set of techniques to trade off accuracy for computation, many other studies of algorithms for statistical inference can be thought of as revealing feasible points on the plane. For example, in the sparse PCA problem where one tries to recover the k-sparse principle component of a p-dimensional multivariate distribution based on i.i.d. samples, various algorithms are studied. The statistically optimal procedure searches through all possible combinations; it requires $\binom{p}{k}$ many operations, and $k \log(p-k)$ many samples to be able to recover the truth principle component. A computationally more

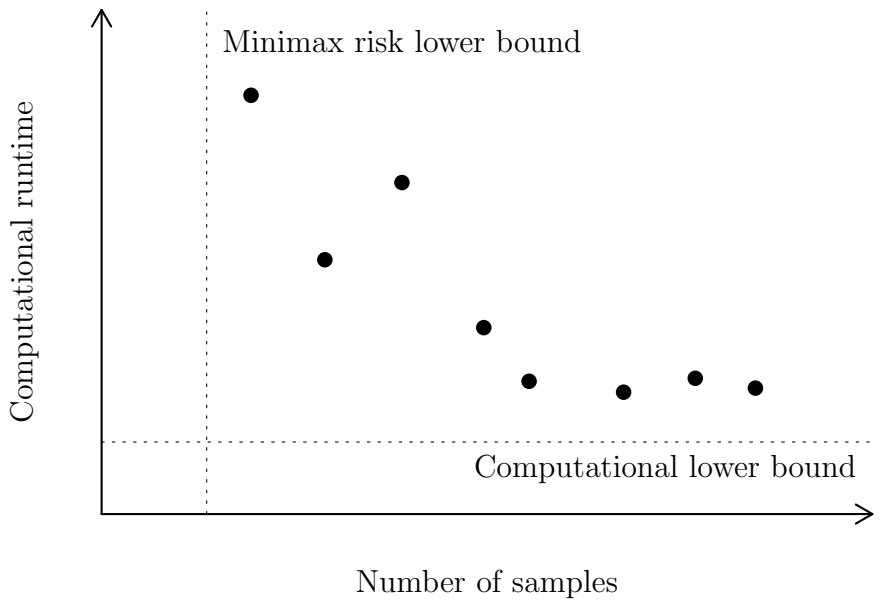

Figure 2.1: Replicate of Figure 1 in Chandrasekaran and Jordan (2013). Tradeoff between the runtime and sample complexity in a parameter estimation problem. The risk is assumed to be fixed to some desired level, and the points in the plot refer to different procedures that require a certain runtime and a certain number of samples to achieve the desired risk. The vertical and horizontal lines refer to lower bounds in sample complexity and in runtime, respectively.

efficient method based on thresholding requires only $np \log p$ many operations, but needs $k^2 \log(p-k)$ samples (Johnstone and Lu, 2012). We thus locate two points in the feasible region for this problem.

Despite all the fruitful results on the tradeoff between time and data, few negative results have been proved. That is, it is hard to give a sharp characterization of the Pareto frontier of the feasible region. Hence, we seldom understand the optimality of a procedure – whether it is the optimal amongst the ones with the same or less amount of computation resources. Getting such negative results is equivalent to drawing the boundary of the feasible region. Any procedures that are exactly on the boundary are in some sense optimal; those in the interior are sub-optimal – we could be better off by either spending less computational resource, or collecting less samples, or both. See Figure 2.2. One of the few examples of such

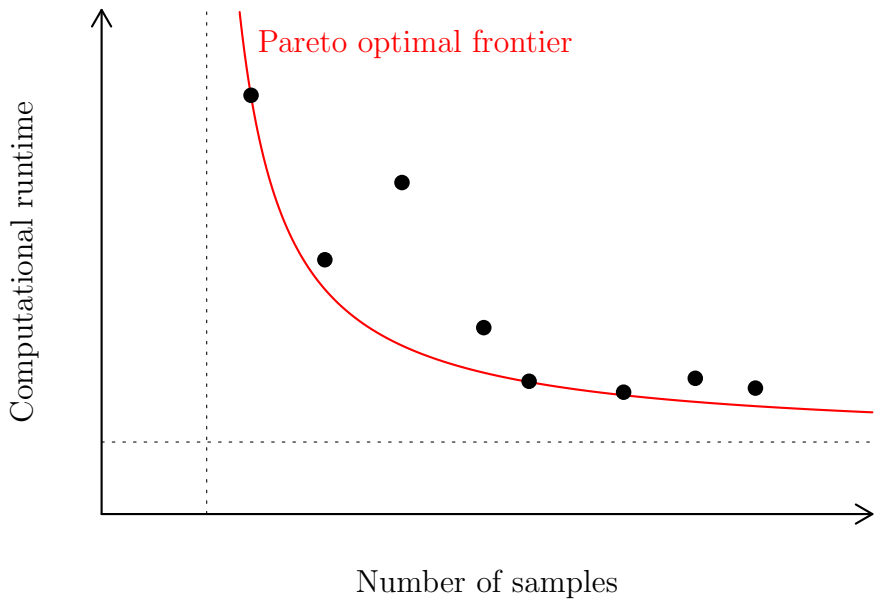

Figure 2.2: Pareto optimal frontier of the computation and sample complexity tradeoff. An algorithm is optimal if its corresponding point falls on the frontier.

negative results is given in Berthet and Rigollet (2013) and Wang et al. (2014), in which it is shown that for the sparse PCA problem $k^2 \log p$ is actually the best sample complexity if we restrict our choices to the procedures that have a polynomial runtime. Similar work that shows such a "computational barrier" include Ma et al. (2015) for sub-matrix detection, and Gao et al. (2014) for sparse CCA, amongst others. However, we must note that such results are not enough to fully characterize the Pareto curve, and neither does it gives an explicit form of the time-constrained minimax risk. In fact, it seems a hard task and requires some unconventional techniques.

In some other problems, it is in fact possible to get a sharp characterization of the computational-statistical tradeoff in terms of constrained minimax risk. In a expository article, Wainwright (2014) gives three examples of such constrained minimax risks, including communication-constrained distributed estimation (Zhang et al., 2013), privacy-constrained estimation (Duchi et al., 2013), and results on polynomial-time sparse regression (Zhang

et al., 2014). Zhu and Lafferty (2014) and Zhu and Lafferty (2015) give explicit formula for the storage constrained minimax risks for Gaussian sequence estimation in different parameter families.

CHAPTER 3

STATISTICAL ESTIMATION WITH STORAGE CONSTRAINTS

3.1 Introduction

In this chapter we introduce a minimax framework for statistical estimation under storage constraints. In the classical statistical setting, the minimax risk for estimating a function f from a function class $\mathcal{F}$ using a sample of size n places no constraints on the estimator $\widehat{f}_n$, other than requiring it to be a measurable function of the data. However, if the estimator is to be constructed with restrictions on the computational resources used, it is of interest to understand how the error can degrade. Letting $C(\widehat{f}_n) \leq B_n$ indicate that the computational resources $C(\widehat{f}_n)$ used to construct $\widehat{f}_n$ are required to fall within a budget B_n, the constrained minimax risk is

$$R_n(\mathcal{F}, B_n) = \inf_{\widehat{f}_n : C(\widehat{f}_n) \leq B_n} \sup_{f \in \mathcal{F}} R(\widehat{f}_n, f).$$

Minimax lower bounds on the risk as a function of the computational budget thus determine a feasible region for computation constrained estimation, and a Pareto optimal tradeoff for risk versus computation as B_n varies.

In this chapter we treat the case where the complexity $C(\widehat{f}_n)$ is measured by the storage or space used by the procedure. Specifically, we limit the number of bits used to represent the estimator $\widehat{f}_n$. We focus on the setting of nonparametric regression under standard smoothness assumptions, and study how the excess risk depends on the storage budget B_n.

We view the study of quantized estimation as a theoretical problem of fundamental interest. But quantization may arise naturally in future applications of large scale statistical estimation. For instance, when data are collected and analyzed on board a remote satellite, the estimated values may need to be sent back to Earth for further analysis. To limit communication costs, the estimates can be quantized, and it becomes important to understand

what, in principle, is lost in terms of statistical risk through quantization. A related scenario is a cloud computing environment where data are processed for many different statistical estimation problems, with the estimates then stored for future analysis. To limit the storage costs, which could dominate the compute costs in many scenarios, it is of interest to quantize the estimates, and the quantization-risk tradeoff again becomes an important concern. A related problem is to distribute the estimation over many parallel processors, and to then limit the communication costs of the submodels to the central host. Estimates are always quantized to some degree in practice. But to impose energy constraints on computation, future processors may limit precision in arithmetic computations more significantly (Galal and Horowitz, 2011); the cost of limited precision in terms of statistical risk must then be quantified.

We study risk-storage tradeoffs in the normal means model of nonparametric estimation assuming the target function lies in a Sobolev space. The problem is intimately related to classical rate distortion theory (Gallager, 1968), and our results rely on a marriage of minimax theory and rate distortion ideas. We thus build on and refine the connection between function estimation and lossy source coding that was elucidated in David Donoho's 1998 Wald Lectures (Donoho, 2000).

We work in the Gaussian white noise model

$$dX(t) = f(t)dt + \varepsilon dW(t), \quad 0 \le t \le 1, \tag{3.1}$$

where W is a standard Wiener process on $[0,1]$, ε is the standard deviation of the noise, and f lies in the periodic Sobolev space $\tilde{W}(m,c)$ of order m and radius c. (We discuss the nonperiodic Sobolev space $W(m,c)$ in Section 3.4.) In this classical setting, the minimax risk of estimation

$$R_\varepsilon(m,c) = \inf_{\widehat{f}_\varepsilon} \sup_{f \in \tilde{W}(m,c)} \mathbb{E}\|f - \widehat{f}_\varepsilon\|_2^2$$

is well known to satisfy

$$\lim_{\varepsilon\to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m,c) = \left(\frac{c^2(2m+1)}{\pi^{2m}}\right)^{\frac{1}{2m+1}} \left(\frac{m}{m+1}\right)^{\frac{2m}{2m+1}} \triangleq \mathsf{P}_{m,c} \tag{3.2}$$

where $\mathsf{P}_{m,c}$ is Pinsker's constant (Nussbaum, 1999). The constrained minimax risk for quantized estimation becomes

$$R_\varepsilon(m,c,B_\varepsilon) = \inf_{\widehat{f}_\varepsilon, C(\widehat{f}_\varepsilon)\le B_\varepsilon} \sup_{f\in\tilde{W}(m,c)} \mathbb{E}\|f-\widehat{f}_\varepsilon\|_2^2$$

where $\widehat{f}_\varepsilon$ is a *quantized estimator* that is required to use storage $C(\widehat{f}_\varepsilon)$ no greater than B_ε bits in total. Our main result identifies three separate quantization regimes.

- In the *over-sufficient regime*, the number of bits is very large, satisfying $B_\varepsilon \gg \varepsilon^{-\frac{2}{2m+1}}$ and the classical minimax rate of convergence $R_\varepsilon \asymp \varepsilon^{\frac{4m}{2m+1}}$ is obtained. Moreover, the optimal constant is the Pinsker constant $\mathsf{P}_{m,c}$.

- In the *sufficient regime*, the number of bits scales as $B_\varepsilon \asymp \varepsilon^{-\frac{2}{2m+1}}$. This level of quantization is just sufficient to preserve the classical minimax rate of convergence, and thus in this regime $R_\varepsilon(m,c,B_\varepsilon) \asymp \varepsilon^{\frac{4m}{2m+1}}$. However, the optimal constant degrades to a new constant $\mathsf{P}_{m,c}+\mathsf{Q}_{m,c,d}$, where $\mathsf{Q}_{m,c,d}$ is characterized in terms of the solution of a certain variational problem, depending on $d = \lim_{\varepsilon\to 0} B_\varepsilon \varepsilon^{\frac{2}{2m+1}}$.

- In the *insufficient regime*, the number of bits scales as $B_\varepsilon \ll \varepsilon^{-\frac{2}{2m+1}}$, with however $B_\varepsilon \to \infty$. Under this scaling the number of bits is insufficient to preserve the unquantized minimax rate of convergence, and the quantization error dominates the estimation error. We show that the quantized minimax risk in this case satisfies

$$\lim_{\varepsilon\to 0} B_\varepsilon^{2m} R_\varepsilon(m,c,B_\varepsilon) = \frac{c^2 m^{2m}}{\pi^{2m}}.$$

Thus, in the insufficient regime the quantized minimax rate of convergence is B_ε^{-2m}, with optimal constant as shown above.

By using an upper bound for the family of constants $\mathsf{Q}_{m,c,d}$, the three regimes can be combined together to view the risk in terms of a decomposition into estimation error and quantization error. Specifically, we can write

$$R_\varepsilon(m,c,B_\varepsilon) \;\approx\; \underbrace{\mathsf{P}_{m,c}\,\varepsilon^{\frac{4m}{2m+1}}}_{\text{estimation error}} \;+\; \underbrace{\frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}}_{\text{quantization error}} .$$

When $B_\varepsilon \gg \varepsilon^{-\frac{2}{2m+1}}$, the estimation error dominates the quantization error, and the usual minimax rate and constant are obtained. In the insufficient case $B_\varepsilon \ll \varepsilon^{-\frac{2}{2m+1}}$, only a slower rate of convergence is achievable. When B_ε and $\varepsilon^{-\frac{2}{2m+1}}$ are comparable, the estimation error and quantization error are on the same order. The threshold $\varepsilon^{-\frac{2}{2m+1}}$ should not be surprising, given that in classical unquantized estimation the minimax rate of convergence is achieved by estimating the first $\varepsilon^{-\frac{2}{2m+1}}$ Fourier coefficients and simply setting the remaining coefficients to zero. This corresponds to selecting a smoothing bandwidth that scales as $h \asymp n^{-\frac{1}{2m+1}}$ with the sample size n.

At a high level, our proof strategy integrates elements of minimax theory and source coding theory. In minimax analysis one computes lower bounds by thinking in Bayesian terms to look for least-favorable priors. In source coding analysis one constructs worst case distributions by setting up an optimization problem based on mutual information. Our quantized minimax analysis requires that these approaches be carefully combined to balance the estimation and quantization errors. To show achievability of the lower bounds we establish, we likewise need to construct an estimator and coding scheme together. Our approach is to quantize the blockwise James-Stein estimator, which achieves the classical Pinsker bound. However, our quantization scheme differs from the approach taken in classical rate distortion theory, where the generation of the codebook is determined once the source distribution is

known. In our setting, we require the allocation of bits to be adaptive to the data, using more bits for blocks that have larger signal size. We therefore design a quantized estimation procedure that adaptively distributes the communication budget across the blocks. Assuming only a lower bound m_0 on the smoothness m and an upper bound c_0 on the radius c of the Sobolev space, our quantization-estimation procedure is adaptive to m and c in the usual statistical sense, and is also adaptive to the coding regime. In other words, given a storage budget B_ε, the coding procedure achieves the optimal rate and constant for the unknown m and c, operating in the corresponding regime for those parameters.

In the following section we establish some notation, outline our proof strategy, and present some simple examples. In Section 3.3 we state and prove our main result on quantized minimax lower bounds, relegating some of the technical details to an appendix. In Section 3.4 we show asymptotic achievability of these lower bounds, using a quantized estimation procedure based on adaptive James-Stein estimation and quantization in blocks, again deferring proofs of technical lemmas to the supplementary material. This is followed by a presentation of some results from experiments in Section 3.5, illustrating the performance and properties of the proposed quantized estimation procedure.

3.2 Quantized estimation and minimax risk

Suppose that $(X_1, \ldots, X_n) \in \mathcal{X}^n$ is a random vector drawn from a distribution P_n. Consider the problem of estimating a functional $\theta_n = \theta(P_n)$ of the distribution, assuming θ_n is restricted to lie in a parameter space Θ_n. To unclutter some of the notation, we will suppress the subscript n and write θ and Θ in the following, keeping in mind that nonparametric settings are allowed. The subscript n will be maintained for random variables. The minimax ℓ_2 risk of estimating θ is then defined as

$$R_n(\Theta) = \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\theta - \widehat{\theta}_n\|^2$$

where the infimum is taken over all possible estimators $\widehat{\theta}_n : \mathcal{X}^n \to \Theta$ that are measurable with respect to the data $X_1, \ldots, X_n$. We will abuse notation by using $\widehat{\theta}_n$ to denote both the estimator and the estimate calculated based on an observed set of data. Among numerous approaches to obtaining the minimax risk, the Bayesian method is best aligned with quantized estimation. Consider a prior distribution $\pi(\theta)$ whose support is a subset of Θ. Let $\delta(X_{1:n})$ be the posterior mean of θ given the data $X_1, \ldots, X_n$, which minimizes the integrated risk. Then for any estimator $\widehat{\theta}_n$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\theta - \widehat{\theta}_n\|^2 \geq \int_\Theta \mathbb{E}_\theta \|\theta - \widehat{\theta}_n\|^2 d\pi(\theta) \geq \int_\Theta \mathbb{E}_\theta \|\theta - \delta(X_{1:n})\|^2 d\pi(\theta).$$

Taking the infimum over $\widehat{\theta}_n$ yields

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\theta - \widehat{\theta}_n\|^2 \geq \int_\Theta \mathbb{E}_\theta \|\theta - \delta(X_{1:n})\|^2 d\pi(\theta) \triangleq R_n(\Theta; \pi).$$

Thus, any prior distribution supported on Θ gives a lower bound on the minimax risk, and selecting the least-favorable prior leads to the largest lower bound provable by this approach.

Now consider constraints on the storage or communication cost of our estimate. We restrict to the set of estimators that use no more than a total of B_n bits; that is, the estimator takes at most 2^{B_n} different values. Such *quantized estimators* can be formulated by the following two-step procedure. First, an *encoder* maps the data $X_{1:n}$ to an index $\phi_n(X_{1:n})$, where

$$\phi_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{B_n}\}$$

is the *encoding function*. The *decoder*, after receiving or retrieving the index, represents the estimates based on a *decoding function*

$$\psi_n : \{1, 2, \ldots, 2^{B_n}\} \to \Theta,$$

mapping the index to a codebook of estimates. All that needs to be transmitted or stored

is the B_n-bit-long index, and the quantized estimator $\widehat{\theta}_n$ is simply $\psi_n \circ \phi_n$, the composition of the encoder and the decoder functions. Denoting by $C(\widehat{\theta}_n)$ the storage, in terms of the number of bits, required by an estimator $\widehat{\theta}_n$, the minimax risk of quantized estimation is then defined as

$$R_n(\Theta, B_n) = \inf_{\widehat{\theta}_n, C(\widehat{\theta}_n) \leq B_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\theta - \widehat{\theta}_n\|^2,$$

and we are interested in the effect of the constraint on the minimax risk. Once again, we consider a prior distribution $\pi(\theta)$ supported on Θ and let $\delta(X_{1:n})$ be the posterior mean of θ given the data. The integrated risk can then be decomposed as

$$\begin{aligned} \int_\Theta \mathbb{E}_\theta \|\theta - \widehat{\theta}_n\|^2 d\pi(\theta) &= \mathbb{E}\|\theta - \delta(X_{1:n}) + \delta(X_{1:n}) - \widehat{\theta}_n\|^2 \\ &= \mathbb{E}\|\theta - \delta(X_{1:n})\|^2 + \mathbb{E}\|\delta(X_{1:n}) - \widehat{\theta}_n\|^2 \end{aligned} \tag{3.3}$$

where the expectation is with respect to the joint distribution of $\theta \sim \pi(\theta)$ and $X_{1:n} \,|\, \theta \sim P_\theta$, and the second equality is due to

$$\begin{aligned} &\mathbb{E}\langle \theta - \delta(X_{1:n}), \delta(X_{1:n}) - \widehat{\theta}_n \rangle \\ &= \mathbb{E}\left(\mathbb{E}\left(\langle \theta - \delta(X_{1:n}), \delta(X_{1:n}) - \widehat{\theta}_n \rangle \,|\, X_{1:n} \right) \right) \\ &= \mathbb{E}\left(\langle \mathbb{E}(\theta - \delta(X_{1:n}) \,|\, X_{1:n}), \mathbb{E}(\delta(X_{1:n}) - \widehat{\theta}_n \,|\, X_{1:n}) \rangle \right) \\ &= \mathbb{E}\left(\langle 0, \mathbb{E}(\delta(X_{1:n}) - \widehat{\theta}_n \,|\, X_{1:n}) \rangle \right) = 0, \end{aligned}$$

using the fact that $\theta \to X_{1:n} \to \widehat{\theta}_n$ forms a Markov chain. The first term in the decomposition (3.3) is the Bayes risk $R_n(\Theta; \pi)$. The second term can be viewed as the excess risk due to quantization.

Let $T_n = T(X_1, \ldots, X_n)$ be a sufficient statistic for θ. The posterior mean can be expressed in terms of T_n and we will abuse notation and write it as $\delta(T_n)$. Since the

quantized estimator $\widehat{\theta}_n$ uses at most B_n bits, we have

$$B_n \geq H(\widehat{\theta}_n) \geq H(\widehat{\theta}_n) - H(\widehat{\theta}_n \,|\, \delta(T_n)) = I(\widehat{\theta}_n; \delta(T_n)),$$

where H and I denote the Shannon entropy and mutual information, respectively. Now consider the optimization

$$\begin{aligned} \inf_{P(\cdot \,|\, \delta(T_n))} \quad & \mathbb{E}\|\delta(T_n) - \tilde{\theta}_n\|^2 \\ \text{such that} \quad & I(\tilde{\theta}_n; \delta(T_n)) \leq B_n \end{aligned}$$

where the infimum is over all conditional distributions $P(\tilde{\theta}_n \,|\, \delta(T_n))$. This parallels the definition of the distortion rate function, minimizing the distortion under a constraint on mutual information (Gallager, 1968). Denoting the value of this optimization by $Q_n(\Theta, B_n; \pi)$, we can lower bound the quantized minimax risk by

$$R_n(\Theta, B_n) \geq R_n(\Theta; \pi) + Q_n(\Theta, B_n; \pi).$$

Since each prior distribution $\pi(\theta)$ supported on Θ gives a lower bound, we have

$$R_n(\Theta, B_n) \geq \sup_{\pi} \Big\{ R_n(\Theta; \pi) + Q_n(\Theta, B_n; \pi) \Big\}$$

and the goal becomes to obtain a least favorable prior for the quantized risk.

Before turning to the case of quantized estimation over Sobolev spaces, we illustrate this technique on some simpler, more concrete examples.

Example 3.2.1 (Normal means in a hypercube)**.** Let $X_i \sim \mathcal{N}(\theta, \sigma^2 I_d)$ for $i = 1, 2, \ldots, n$. Suppose that σ^2 is known and $\theta \in [-\tau, \tau]^d$ is to be estimated. We choose the prior $\pi(\theta)$ on

θ to be a product distribution with density

$$\pi(\theta) = \prod_{j=1}^{d} \frac{3}{2\tau^3}\big(\tau - |\theta_j|\big)_+^2.$$

It is shown in Johnstone (2015) that

$$R_n(\Theta;\pi) \geq \frac{\sigma^2 d}{n}\frac{\tau^2}{\tau^2 + 12\sigma^2/n} \geq c_1\frac{\sigma^2 d}{n}$$

where $c_1 = \frac{\tau^2}{\tau^2+12\sigma^2}$. Turning to $Q_n(\Theta, B_n;\pi)$, let $T^{(n)} = (T_1^{(n)},\ldots,T_d^{(n)}) = \mathbb{E}(\theta|X_{1:n})$ be the posterior mean of θ. In fact, by the independence and symmetry among the dimensions, we know $T_1,\ldots,T_d$ are independently and identically distributed. Denoting by $T_0^{(n)}$ this common distribution, we have

$$Q_n(\Theta, B_n;\pi) \geq d \cdot q(B_n/d)$$

where $q(B)$ is the distortion rate function for $T_0^{(n)}$, i.e., the value of the following problem

$$\begin{aligned} \inf_{P(\widehat{T}\,|\,T_0^{(n)})} \quad & \mathbb{E}(T_0^{(n)} - \widehat{T})^2 \\ \text{such that} \quad & I(\widehat{T}; T_0^{(n)}) \leq B. \end{aligned}$$

Now using the Shannon lower bound (Cover and Thomas, 2006), we get

$$Q_n(\Theta, B_n;\pi) \geq \frac{d}{2\pi e}\cdot 2^{h(T_0^{(n)})}\cdot 2^{-\frac{2B_n}{d}}.$$

Note that as $n\to\infty$, $T_0^{(n)}$ converges to θ in distribution, so there exists a constant c_2 independent of n and d such that

$$R_n(\Theta, B_n) \geq c_1\frac{\sigma^2 d}{n} + c_2 d\, 2^{-\frac{2B_n}{d}}.$$

This lower bound intuitively shows the risk is regulated by two factors, the estimation error and the quantization error; whichever is larger dominates the risk. The scaling behavior of this lower bound (ignoring constants) can be achieved by first quantizing each of the d intervals $[-\tau, \tau]$ using B_n/d bits each, and then mapping the MLE to its closest codeword.

Example 3.2.2 (Binomial)**.** Let $X_i \sim \text{Bern}(\theta)$ be independent samples from a Bernoulli distribution, for $i = 1, 2, \ldots, n$, and take $\pi(\theta) = 1$ to be the uniform prior on $[0, 1]$. Then $T_n = \overline{X}_n$ and

$$\delta(T_n) = \frac{n}{n+2}\overline{X}_n + \frac{2}{n+1} \cdot \frac{1}{2}$$

with $I(\tilde{\theta}_n, \delta(T_n)) = I(\tilde{\theta}_n, \overline{X}_n)$. In this case it can be shown that

$$R_n(\Theta, B_n) \geq \frac{c_1}{n} + c_2 H^{-1}\left(1 - \frac{B_n}{n}\right)$$

for constants c_1 and c_2, where H^{-1} is the inverse of the binary entropy function on $[0, \frac{1}{2}]$.

Example 3.2.3 (Gaussian sequences in Euclidean balls)**.** In the example shown above, the lower bound is tight only in terms of the scaling of the key parameters. In some instances, we are able to find an asymptotically tight lower bound for which we can show achievability of both the rate and the constants. Estimating the mean vector of a Gaussian sequence with an ℓ_2 norm constraint on the mean is one of such case, as we showed in previous work (Zhu and Lafferty, 2014).

Specifically, let $X_i \sim \mathcal{N}(\theta_i, \sigma_n^2)$ for $i = 1, 2, \ldots, n$, where $\sigma_n^2 = \sigma^2/n$. Suppose that the parameter $\theta = (\theta_1, \ldots, \theta_n)$ lies in the Euclidean ball $\Theta_n(c) = \left\{\theta : \sum_{i=1}^n \theta_i^2 \leq c^2\right\}$. Furthermore, suppose that $B_n = nB$. Then using the prior $\theta_i \sim \mathcal{N}(0, c^2)$ it can be shown that

$$\liminf_{n\to\infty} R_n(\Theta_n(c), B_n) \geq \frac{\sigma^2 c^2}{\sigma^2 + c^2} + \frac{c^4 2^{-2B}}{\sigma^2 + c^2}.$$

The asymptotic estimation error $\sigma^2 c^2/(\sigma^2 + c^2)$ is the well-known Pinsker bound for the Euclidean ball case. As shown in Zhu and Lafferty (2014), an explicit quantization scheme

can be constructed that asymptotically achieves this lower bound, realizing the smallest possible quantization error $c^4 2^{-2B}/(\sigma^2 + c^2)$ for a budget of $B_n = nB$ bits.

The Euclidean ball case is clearly relevant to the Sobolev ellipsoid case, but new coding strategies and proof techniques are required. In particular, as will be made clear in the sequel, we will use an adaptive allocation of bits across blocks of coefficients, using more bits for blocks that have larger estimated signal size. Moreover, determination of the optimal constants requires a detailed analysis of the worst case prior distributions and the solution of a series of variational problems.

3.3 Quantized estimation over Sobolev spaces

Recall that the *Sobolev space of order m and radius c* is defined by

$$W(m,c) = \Big\{ f \in [0,1] \to \mathbb{R} : f^{(m-1)} \text{ is absolutely continuous and } \int_0^1 (f^{(m)}(x))^2 dx \le c^2 \Big\}.$$

The *periodic Sobolev space* is defined by

$$\tilde{W}(m,c) = \left\{ f \in W(m,c) : f^{(j)}(0) = f^{(j)}(1),\ j = 0,1,\ldots,m-1 \right\}. \tag{3.4}$$

The white noise model (3.1) is asymptotically equivalent to making n equally spaced observations along the sample path, $Y_i = f(i/n) + \sigma\epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,1)$ (Brown and Low, 1996a). In this formulation, the noise level in the formulation (3.1) scales as $\epsilon^2 = \sigma^2/n$, and the rate of convergence takes the familiar form $n^{-\frac{2m}{2m+1}}$ where n is the number of observations.

To carry out quantized estimation we now require an encoder

$$\phi_\varepsilon : \mathbb{R}^{[0,1]} \to \{1, 2, \ldots, 2^{B_\varepsilon}\}$$

which is a function applied to the sample path $X(t)$. The decoding function then takes the form

$$\psi_\varepsilon : \{1, 2, \ldots, 2^{B_\varepsilon}\} \to \mathbb{R}^{[0,1]}$$

and maps the index to a function estimate. As in the previous section, we write the composition of the encoder and the decoder as $\widehat{f}_\varepsilon = \psi_\varepsilon \circ \phi_\varepsilon$, which we call the quantized estimator. The communication or storage $C(\widehat{f}_\varepsilon)$ required by this quantized estimator is no more than B_ε bits.

To recast quantized estimation in terms of an infinite sequence model, let $(\varphi_j)_{j=1}^\infty$ be the trigonometric basis, and let

$$\theta_j = \int_0^1 \varphi_j(t) f(t) dt, \quad j = 1, 2, \ldots,$$

be the Fourier coefficients. It is well known (Tsybakov, 2008) that $f = \sum_{j=1}^\infty \theta_j \varphi_j$ belongs to $\tilde{W}(m, c)$ if and only if the Fourier coefficients θ belong to the *Sobolev ellipsoid* defined as

$$\Theta(m, c) = \left\{ \theta \in \ell_2 : \sum_{j=1}^\infty a_j^2 \theta_j^2 \leq \frac{c^2}{\pi^{2m}} \right\} \tag{3.5}$$

where

$$a_j = \begin{cases} j^m, & \text{for even } j, \\ (j-1)^m, & \text{for odd } j. \end{cases}$$

Although this is the standard definition of a Sobolev ellipsoid, for the rest of the paper we will set $a_j = j^m$, $j = 1, 2, \ldots$ for convenience of analysis. All of the results hold for both definitions of a_j. Also note that (3.5) actually gives a more general definition, since m is no longer assumed to be an integer, as it is in (3.4). Expanding with respect to the same orthonormal basis, the observed path $X(t)$ is converted into an infinite Gaussian sequence

$$Y_j = \int_0^1 \varphi_j(t)\, dX(t), \quad j = 1, 2, \ldots,$$

with $Y_j \sim \mathcal{N}(\theta_j, \varepsilon^2)$. For an estimator $(\widehat{\theta}_j)_{j=1}^{\infty}$ of $(Y_j)_{j=1}^{\infty}$, an estimate of f is obtained by

$$\widehat{f}(x) = \sum_{j=1}^{\infty} \widehat{\theta}_j \varphi_j(x)$$

with squared error $\|\widehat{f} - f\|_2^2 = \|\widehat{\theta} - \theta\|_2^2$. In terms of this standard reduction, the quantized minimax risk is thus reformulated as

$$R_\varepsilon(m, c, B_\varepsilon) = \inf_{\widehat{\theta}_\varepsilon, C(\widehat{\theta}_\varepsilon) \le B_\varepsilon} \sup_{\theta \in \Theta(m,c)} \mathbb{E}_\theta \|\theta - \widehat{\theta}_\varepsilon\|_2^2. \tag{3.6}$$

To state our result, we need to define the value of the following variational problem:

$$\mathsf{V}_{m,c,d} \triangleq \tag{3.7}$$
$$\max_{(\sigma^2, x_0) \in \mathcal{F}(m,c,d)} \int_0^{x_0} \frac{\sigma^2(x)}{\sigma^2(x)+1} dx + x_0 \exp\left(\frac{1}{x_0} \int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0} \right)$$

where the feasible set $\mathcal{F}(m, c, d)$ is the collection of increasing functions $\sigma^2(x)$ and values x_0 satisfying

$$\int_0^{x_0} x^{2m} \sigma^2(x) dx \le c^2$$
$$\frac{\sigma^4(x)}{\sigma^2(x)+1} \ge \exp\left(\frac{1}{x_0} \int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0} \right) \text{ for all } x \le x_0.$$

The significance and interpretation of the variational problem will become apparent as we outline the proof of this result.

Theorem 3.3.1. *Let $R_\varepsilon(m, c, B_\varepsilon)$ be defined as in (3.6), for $m > 0$ and $c > 0$.*

(i) If $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to \infty$ as $\varepsilon \to 0$, then

$$\liminf_{\varepsilon \to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m, c, B_\varepsilon) \ge \mathsf{P}_{m,c}$$

where $\mathsf{P}_{m,c}$ is Pinker's constant defined in (3.2).

(ii) If $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to d$ *for some constant* d *as* $\varepsilon \to 0$*, then*

$$\liminf_{\varepsilon \to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m, c, B_\varepsilon) \geq \mathsf{P}_{m,c} + \mathsf{Q}_{m,c,d} = \mathsf{V}_{m,c,d}$$

where $\mathsf{V}_{m,c,d}$ *is the value of the variational problem* (3.7)*.*

(iii) If $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to 0$ *and* $B_\varepsilon \to \infty$ *as* $\varepsilon \to 0$*, then*

$$\liminf_{\varepsilon \to 0} B_\varepsilon^{2m} R_\varepsilon(m, c, B_\varepsilon) \geq \frac{c^2 m^{2m}}{\pi^{2m}}.$$

In the first regime where the number of bits B_ε is much greater than $\varepsilon^{-\frac{2}{2m+1}}$, we recover the same convergence result as in Pinsker's theorem, in terms of both convergence rate and leading constant. The proof of the lower bound for this regime can directly follow the proof of Pinsker's theorem, since the set of estimators considered in our minimax framework is a subset of all possible estimators.

In the second regime where we have "just enough" bits to preserve the rate, we suffer a loss in terms of the leading constant. In this "Goldilocks regime," the optimal rate $\varepsilon^{-\frac{4m}{2m+1}}$ is achieved but the constant in front of the rate is Pinsker's constant $\mathsf{P}_{m,c}$ plus a positive quantity $\mathsf{Q}_{m,c,d}$ determined by the variational problem.

While the solution to this variational problem does not appear to have an explicit form, it can be computed numerically. We discuss this term at length in the sequel, where we explain the origin of the variational problem, compute the constant numerically and approximate it from above and below. The constants $\mathsf{P}_{m,c}$ and $\mathsf{Q}_{m,c,d}$ are shown graphically in Figure 3.1. Note that the parameter d can be thought of as the average number of bits per coefficient used by an optimal quantized estimator, since $\varepsilon^{-\frac{2}{2m+1}}$ is asymptotically the number of coefficients needed to estimate at the classical minimax rate. As shown in Figure 3.1, the constant for quantized estimation quickly approaches the Pinsker constant as d increases—when $d = 3$ the two are already very close.

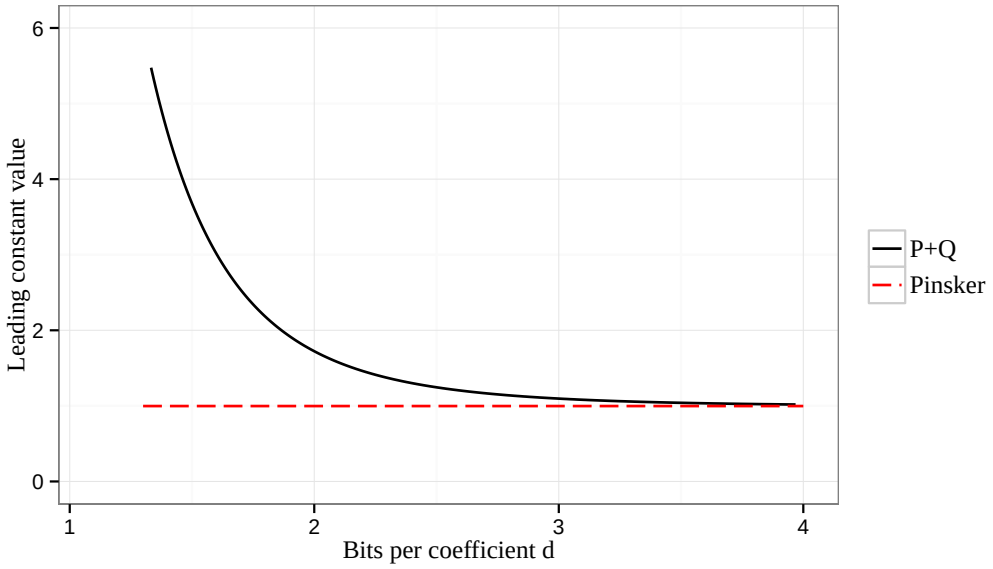

Figure 3.1: The constants $\mathsf{P}_{m,c} + \mathsf{Q}_{m,c,d}$ as a function of quantization level d in the sufficient regime, where $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to d$. The parameter d can be thought of as the average number of bits per coefficient used by an optimal quantized estimator, because $\varepsilon^{-\frac{2}{2m+1}}$ is asymptotically the number of coefficients needed to estimate at the classical minimax rate. Here we take $m = 2$ and $c^2/\pi^{2m} = 1$. The curve indicates that with only 2 bits per coefficient, optimal quantized minimax estimation degrades by less than a factor of 2 in the constant. With 3 bits per coefficient, the constant is very close to the classical Pinsker constant.

In the third regime where the communication budget is insufficient for the estimator to achieve the optimal rate, we obtain a sub-optimal rate which no longer depends explicitly on the noise level ε of the model. In this regime, quantization error dominates, and the risk decays at a rate of $B^{-\frac{1}{2m}}$ no matter how fast ε approaches zero, as long as $B \ll \varepsilon^{-\frac{2}{2m+1}}$. Here the analogue of Pinsker's constant takes a very simple form.

Proof of Theorem 3.3.1. Consider a Gaussian prior distribution on $\theta = (\theta_j)_{j=1}^\infty$ with $\theta_j \sim \mathcal{N}(0, \sigma_j^2)$ for $j = 1, 2, \ldots,$ in terms of parameters $\sigma^2 = (\sigma_j^2)_{j=1}^\infty$ to be specified later. One requirement for the variances is

$$\sum_{j=1}^{\infty} a_j^2 \sigma_j^2 \le \frac{c^2}{\pi^{2m}}.$$

We denote this prior distribution by $\pi(\theta; \sigma^2)$, and show in Section 3.7 that it is asymptotically

concentrated on the ellipsoid $\Theta(m, c)$. Under this prior the model is

$$\theta_j \sim \mathcal{N}(0, \sigma_j^2)$$
$$Y_j \,|\, \theta_j \sim \mathcal{N}(\theta_j, \varepsilon^2), \quad j = 1, 2, \ldots$$

and the marginal distribution of Y_j is thus $\mathcal{N}(0, \sigma_j^2 + \varepsilon^2)$. Following the strategy outlined in Section 3.2, let δ denote the posterior mean of θ given Y under this prior, and consider the optimization

$$\inf \;\; \mathbb{E}\|\delta - \tilde{\theta}\|^2$$
$$\text{such that} \;\; I(\delta; \tilde{\theta}) \leq B_\epsilon$$

where the infimum is over all distributions on $\tilde{\theta}$ such that $\theta \to Y \to \tilde{\theta}$ forms a Markov chain. Now, the posterior mean satisfies $\delta_j = \gamma_j Y_j$ where $\gamma_j = \sigma_j^2/(\sigma_j^2 + \epsilon^2)$. Note that the Bayes risk under this prior is

$$\mathbb{E}\|\theta - \delta\|_2^2 = \sum_{j=1}^{\infty} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2}.$$

Define

$$\mu_j^2 \triangleq \mathbb{E}(\delta_j - \tilde{\theta}_j)^2.$$

Then the classical rate distortion argument (Cover and Thomas, 2006) gives that

$$\begin{aligned} I(\delta; \tilde{\theta}) &\geq \sum_{j=1}^{\infty} I(\gamma_j Y_j; \tilde{\theta}_j) \\ &\geq \sum_{j=1}^{\infty} \frac{1}{2} \log_+ \left(\frac{\gamma_j^2 (\sigma_j^2 + \varepsilon^2)}{\mu_j^2} \right) \\ &= \sum_{j=1}^{\infty} \frac{1}{2} \log_+ \left(\frac{\sigma_j^4}{\mu_j^2 (\sigma_j^2 + \varepsilon^2)} \right) \end{aligned}$$

where $\log_+(x) = \max(\log x, 0)$. Therefore, the quantized minimax risk is lower bounded by

$$R_\varepsilon(m, c, B_\varepsilon) = \inf_{\widehat{\theta}_\varepsilon, C(\widehat{\theta}_\varepsilon) \leq B_\varepsilon} \sup_{\theta \in \Theta(m,c)} \mathbb{E}\|\theta - \widehat{\theta}_\varepsilon\|^2 \geq V_\varepsilon(B_\varepsilon, m, c)(1 + o(1))$$

where $V_\varepsilon(B_\varepsilon, m, c)$ is the value of the optimization

$$\begin{aligned} \max_{\sigma^2} \min_{\mu^2} \quad & \sum_{j=1}^{\infty} \mu_j^2 + \sum_{j=1}^{\infty} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2} \\ \text{such that} \quad & \sum_{j=1}^{\infty} \frac{1}{2} \log_+ \left(\frac{\sigma_j^4}{\mu_j^2(\sigma_j^2 + \varepsilon^2)} \right) \leq B_\varepsilon \\ & \sum_{j=1}^{\infty} a_j^2 \sigma_j^2 \leq \frac{c^2}{\pi^{2m}} \end{aligned} \tag{$\mathcal{P}_1$}$$

and the $(1 + o(1))$ deviation term is analyzed in the supplementary material.

Observe that the quantity $V_\varepsilon(B_\varepsilon, m, c)$ can be upper and lower bounded by

$$\max\Big\{R_\varepsilon(m, c), Q_\varepsilon(m, c, B_\varepsilon)\Big\} \leq V_\varepsilon(m, c, B_\varepsilon) \leq R_\varepsilon(m, c) + Q_\varepsilon(m, c, B_\varepsilon) \tag{3.8}$$

where the estimation error term $R_\varepsilon(m, c)$ is the value of the optimization

$$\begin{aligned} \max_{\sigma^2} \quad & \sum_{j=1}^{\infty} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2} \\ \text{such that} \quad & \sum_{j=1}^{\infty} a_j^2 \sigma_j^2 \leq \frac{c^2}{\pi^{2m}} \end{aligned} \tag{$\mathcal{R}_1$}$$

and the quantization error term $Q_\varepsilon(m, c, B_\varepsilon)$ is the value of the optimization

$$\begin{aligned} \max_{\sigma^2} \min_{\mu^2} \quad & \sum_{j=1}^{\infty} \mu_j^2 \\ \text{such that} \quad & \sum_{j=1}^{\infty} \frac{1}{2} \log_+ \left(\frac{\sigma_j^4}{\mu_j^2(\sigma_j^2 + \varepsilon^2)} \right) \leq B_\varepsilon \\ & \sum_{j=1}^{\infty} a_j^2 \sigma_j^2 \leq \frac{c^2}{\pi^{2m}}. \end{aligned} \tag{$\mathcal{Q}_1$}$$

The following results specify the leading order asymptotics of these quantities.

Lemma 3.3.2. *As* $\varepsilon \to 0$,

$$R_\varepsilon(m,c) = \mathsf{P}_{m,c}\,\varepsilon^{\frac{4m}{2m+1}}(1+o(1)).$$

Lemma 3.3.3. *As* $\varepsilon \to 0$,

$$Q_\varepsilon(m,c,B_\varepsilon) \le \frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}(1+o(1)). \tag{3.9}$$

Moreover, if $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to 0$ *and* $B_\varepsilon \to \infty$,

$$Q_\varepsilon(m,c,B_\varepsilon) = \frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}(1+o(1)).$$

This yields the following closed form upper bound.

Corollary 3.3.4. *Suppose that* $B_\varepsilon \to \infty$ *and* $\varepsilon \to 0$. *Then*

$$V_\varepsilon(m,c,B_\varepsilon) \le \left(\mathsf{P}_{m,c}\,\varepsilon^{\frac{4m}{2m+1}} + \frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}\right)(1+o(1)). \tag{3.10}$$

In the insufficient regime $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to 0$ and $B_\varepsilon \to \infty$ as $\varepsilon \to 0$, equation (3.8) and Lemma 3.3.3 show that

$$V_\varepsilon(m,c,B_\varepsilon) = \frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}(1+o(1)).$$

Similarly, in the over-sufficient regime $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to \infty$ as $\varepsilon \to 0$, we conclude that

$$V_\varepsilon(m,c,B_\varepsilon) = \mathsf{P}_{m,c}\,\varepsilon^{\frac{4m}{2m+1}}(1+o(1)).$$

We now turn to the sufficient regime $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to d$. We begin by making three observations about the solution to the optimization $(\mathcal{P}_1)$. First, we note that the series $(\sigma_j^2)_{j=1}^\infty$

that solves $(\mathcal{P}_1)$ can be assumed to be decreasing. If (σ_j^2) were not in decreasing order, we could rearrange it to be decreasing, and correspondingly rearrange (μ_j^2), without violating the constraints or changing the value of the optimization. Second, we note that given (σ_j^2), the optimal (μ_j^2) is obtained by the "reverse water-filling" scheme (Cover and Thomas, 2006). Specifically, there exists $\eta > 0$ such that

$$\mu_j^2 = \begin{cases} \eta & \text{if } \dfrac{\sigma_j^4}{\sigma_j^2 + \varepsilon^2} \geq \eta \\ \dfrac{\sigma_j^4}{\sigma_j^2 + \varepsilon^2} & \text{otherwise,} \end{cases}$$

where η is chosen so that

$$\frac{1}{2} \sum_{j=1}^{\infty} \log_+ \left(\frac{\sigma_j^4}{\mu_j^2(\sigma_j^2 + \varepsilon^2)} \right) \leq B_\varepsilon.$$

Third, there exists an integer $J > 0$ such that the optimal series (σ_j^2) satisfies

$$\frac{\sigma_j^4}{\sigma_j^2 + \varepsilon^2} \geq \eta, \text{ for } j = 1, \ldots, J \quad \text{and} \quad \sigma_j^2 = 0, \text{ for } j > J,$$

where η is the "water-filling level" for (μ_j^2). Using these three observations, the optimization $(\mathcal{P}_1)$ can be reformulated as

$$\begin{aligned} \max_{\sigma^2, J} \quad & J\eta + \sum_{j=1}^{J} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2} \\ \text{such that} \quad & \frac{1}{2} \sum_{j=1}^{J} \log_+ \left(\frac{\sigma_j^4}{\eta(\sigma_j^2 + \varepsilon^2)} \right) = B_\varepsilon \\ & \sum_{j=1}^{J} a_j^2 \sigma_j^2 \leq \frac{c^2}{\pi^{2m}} \\ & (\sigma_j^2) \text{ is decreasing and } \frac{\sigma_J^4}{\sigma_J^2 + \varepsilon^2} \geq \eta. \end{aligned} \tag{$\mathcal{P}_2$}$$

To derive the solution to $(\mathcal{P}_2)$, we use a continuous approximation of σ^2, writing

$$\sigma_j^2 = \sigma^2(jh)h^{2m+1}$$

where h is the bandwidth to be specified and $\sigma^2(\cdot)$ is a function defined on $(0, \infty)$. The constraint that $\sum_{j=1}^{\infty} a_j^2 \sigma_j^2 \leq \frac{c^2}{\pi^{2m}}$ becomes the integral constraint

$$\int_0^{\infty} x^{2m}\sigma^2(x)dx \leq \frac{c^2}{\pi^{2m}}.$$

We now set the bandwidth so that $h^{2m+1} = \varepsilon^2$. This choice of bandwidth will balance the two terms in the objective function, and thus gives the hardest prior distribution. Applying the above three observations under this continuous approximation, we transform problem $(\mathcal{P}_2)$ to the following optimization:

$$\begin{aligned} \max_{\sigma^2, x_0} \quad & x_0\eta + \int_0^{x_0} \frac{\sigma^2(x)}{\sigma^2(x)+1}dx \\ \text{such that} \quad & \int_0^{x_0} \frac{1}{2}\log_+\left(\frac{\sigma^4(x)}{\eta(\sigma^2(x)+1)}\right) = d \\ & \int_0^{x_0} x^{2m}\sigma^2(x)dx \leq \frac{c^2}{\pi^{2m}} \\ & \sigma^2(x) \text{ is decreasing and } \frac{\sigma^4(x)}{\sigma^2(x)+1} \geq \eta \text{ for all } x \leq x_0. \end{aligned} \tag{$\mathcal{P}_3$}$$

Note that here we omit the convergence rate $h^{2m} = \varepsilon^{\frac{4m}{2m+1}}$ in the objective function. The asymptotic equivalence between $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$ can be established by a similar argument to

Theorem 3.1 in Donoho (2000). Solving the first constraint for η yields

$$\begin{aligned} \max_{\sigma^2, x_0} \quad & \int_0^{x_0} \frac{\sigma^2(x)}{\sigma^2(x)+1} dx + x_0 \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right) \\ \text{such that} \quad & \int_0^{x_0} x^{2m}\sigma^2(x) dx \le \frac{c^2}{\pi^{2m}} \\ & \sigma^2(x) \text{ is decreasing} \\ & \frac{\sigma^4(x)}{\sigma^2(x)+1} \ge \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right) \\ & \qquad \text{for all } x \le x_0. \end{aligned} \tag{$\mathcal{P}_4$}$$

The following is proved using a variational argument in the supplementary material.

Lemma 3.3.5. *The solution to ($\mathcal{P}_4$) satisfies*

$$\frac{1}{(\sigma^2(x)+1)^2} + \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right) \frac{\sigma^2(x)+2}{\sigma^2(x)(\sigma^2(x)+1)} = \lambda x^{2m}$$

for some $\lambda > 0$.

Fixing x_0, the lemma shows that by setting

$$\alpha = \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right)$$

we can express $\sigma^2(x)$ implicitly as the unique positive root of a third-order polynomial in y,

$$\lambda x^{2m} y^3 + (2\lambda x^{2m} - \alpha) y^2 + (\lambda x^{2m} - 3\alpha - 1) y - 2\alpha.$$

This leads us to an explicit form of $\sigma^2(x)$ for a given value α. However, note that α still depends on $\sigma^2(x)$ and x_0, so the solution $\sigma^2(x)$ might not be compatible with α and x_0. We can either search through a grid of values of α and x_0, or, more efficiently, use an iterative method to find the pair of values that gives us the solution. We omit the details on how to calculate the values of the optimization as it is not main purpose of the paper.

To summarize, in the regime $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to d$ as $\varepsilon \to 0$, we obtain

$$V_\varepsilon(m, c, B_\varepsilon) = (\mathsf{P}_{m,c} + \mathsf{Q}_{m,c,d})\, \varepsilon^{\frac{4m}{2m+1}}(1 + o(1)),$$

where we denote by $\mathsf{P}_{m,c} + \mathsf{Q}_{m,c,d}$ the values of the optimization $(\mathcal{P}_4)$. □

3.4 Achievability

We now show that the lower bounds in Theorem 3.3.1 are achievable by a quantized estimator using a random coding scheme. The basic idea of our quantized estimation procedure is to conduct blockwise estimation and quantization together, using a quantized form of the Stein estimator.

Before we set the stage for the quantized form of James-Stein estimator, let us first look at a class of simple procedures. Suppose that $\widehat{\theta} = \widehat{\theta}(X)$ is an estimator of $\theta \in \Theta(m, c)$ without quantization. We assume that $\widehat{\theta} \in \Theta(m, c)$, as projection always reduces mean squared error. To design a B-bit quantized estimator, let $\check{\Theta}$ be the optimal δ-covering of the parameter space $\Theta(m, c)$ such that $|\check{\Theta}| \le 2^B$, that is,

$$\delta = \delta(B) = \inf_{\check{\Theta} \subset \Theta : |\check{\Theta}| \le 2^B} \sup_{\theta \in \Theta} \inf_{\theta' \in \check{\Theta}} \|\theta - \theta'\|.$$

The quantized estimator is then defined to be

$$\check{\theta} = \check{\theta}(X) = \arg\min_{\theta' \in \check{\Theta}} \|\widehat{\theta}(X) - \theta'\|.$$

Now the mean squared error

$$\mathbb{E}_\theta \|\check{\theta} - \theta\|^2 = \mathbb{E}_\theta \|\check{\theta} - \widehat{\theta} + \widehat{\theta} - \theta\|^2 \le 2\mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2 + 2\mathbb{E}_\theta \|\check{\theta} - \widehat{\theta}\|^2 \le 2 \sup_{\theta'} \mathbb{E}_{\theta'} \|\widehat{\theta} - \theta'\|^2 + 2\delta(B)^2.$$

If we pick $\widehat{\theta}$ to be the minimax estimator for Θ, the first term here gives the minimax

risk for estimating θ in the parameter space Θ. The second term is closely related to the metric entropy of the parameter space $\Theta(m, c)$. In fact, for the Sobolev ellipsoid $\Theta(m, c)$, it is shown in Donoho (2000) that $\delta(B)^2 = \frac{c^2 m^{2m}}{\pi^{2m}} B^{-2m}(1+o(1))$ as $B \to \infty$. Thus, with an extra constant factor of 2, the mean squared error of this quantized estimator is decomposed into the minimax risk for Θ and an error term due to quantization. In addition to the fact that the aforementioned procedure does not achieve the exact lower bound of the minimax risk for the constrained estimation problem, it is not clear how such an ε-net can be generated. In what follows we will describe a quantized estimation procedure which we will show to achieve the lower bound up to exact constants, and also adapt to the unknown parameters.

We begin by defining the block system to be used, which is usually referred to as the *weakly geometric system of blocks* (Tsybakov, 2008). Let $N_\varepsilon = \lfloor 1/\varepsilon^2 \rfloor$ and $\rho_\varepsilon = (\log(1/\varepsilon))^{-1}$. Let $J_1, \ldots, J_K$ be a partition of the set $\{1, \ldots, N_\varepsilon\}$ such that

$$\bigcup_{k=1}^{K} J_k = \{1, \ldots, N_\varepsilon\}, \quad J_{k_1} \cap J_{k_2} = \emptyset \text{ for } k_1 \neq k_2,$$
$$\text{and } \min\{j : j \in J_k\} > \max\{j : j \in J_{k-1}\}.$$

Let T_k be the cardinality of the kth block and suppose that $T_1, \ldots, T_k$ satisfy

$$\begin{aligned} T_1 &= \lceil \rho_\varepsilon^{-1} \rceil = \lceil \log(1/\varepsilon) \rceil, \\ T_2 &= \lfloor T_1(1+\rho_\varepsilon) \rfloor, \\ &\vdots \\ T_{K-1} &= \lfloor T_1(1+\rho_\varepsilon)^{K-2} \rfloor, \\ T_K &= N_\varepsilon - \sum_{k=1}^{K-1} T_k. \end{aligned} \tag{3.11}$$

For an infinite sequence $x \in \ell_2$, denote by $x_{(k)}$ the vector $(x_j)_{j \in J_k} \in \mathbb{R}^{T_k}$. We also write $j_k = \sum_{l=1}^{k-1} T_l + 1$, which is the smallest index in block J_k. The weakly geometric system of blocks is defined such that the size of the blocks does not grow too quickly (the ratio between

the sizes of the neighboring two blocks goes to 1 asymptotically), and that the number of the blocks is on the logarithmic scale with respect to $1/\varepsilon$ ($K \lesssim \log^2(1/\varepsilon)$). See Lemma 3.7.3.

We are now ready to describe the quantized estimation scheme. We first give a high-level description of the scheme, and then the precise specification. In contrast to rate distortion theory, where the codebook and allocation of the bits are determined once the source distribution is known, here the codebook and allocation of bits are adaptive to the data—more bits are used for blocks having larger signal size. The first step in our quantization scheme is to construct a "base code" of 2^{B_ε} randomly generated vectors of maximum block length T_K, with $\mathcal{N}(0,1)$ entries. The base code is thought of as a $2^{B_\varepsilon} \times T_K$ random matrix $\mathcal{Z}$; it is generated before observing any data, and is shared between the sender and receiver. After observing data (Y_j), the rows of $\mathcal{Z}$ are apportioned to different blocks $k = 1, \ldots, K$, with more rows being used for blocks having larger estimated signal size. To do so, the norm $\|Y_{(k)}\|$ of each block k is first quantized as a discrete value $\check{S}_k$. A subcodebook $\mathcal{Z}_k$ is then constructed by normalizing the appropriate rows and the first T_k columns of the base code, yielding a collection of random points on the unit sphere $\mathbb{S}^{T_k-1}$. To form a quantized estimate of the coefficients in the block, the codeword $\check{Z}_{(k)} \in \mathcal{Z}_k$ having the smallest angle to $Y_{(k)}$ is then found. The appropriate indices are then transmitted to the receiver. To decode and reconstruct the quantized estimate, the receiver first recovers the quantized norms $(\check{S}_k)$, which enables reconstruction of the subdivision of the base code that was used by the encoder. After extracting for each block k the appropriate row of the base code, the codeword $\check{Z}_{(k)}$ is reconstructed, and a James-Stein type estimator is then calculated.

The quantized estimation scheme is detailed below.

STEP 1. *Base code generation.*

1.1. Generate codebook $\mathcal{S}_k = \left\{\sqrt{T_k\varepsilon^2} + i\varepsilon^2 \ : \ i = 0, 1, \ldots, s_k\right\}$ where $s_k = \left\lceil \varepsilon^{-2} c (j_k \pi)^{-m} \right\rceil$, for $k = 1, \ldots, K$.

1.2. Generate base code $\mathcal{Z}$, a $2^B \times T_K$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries.

$(\mathcal{S}_k)$ and $\mathcal{Z}$ are shared between the encoder and the decoder, before seeing any data.

STEP 2. *Encoding.*

2.1. *Encoding block radius.* For $k = 1, \ldots, K$, encode

$\check{S}_k = \arg\min \{|s - S_k| : s \in \mathcal{S}_k\}$ where

$$S_k = \begin{cases} \sqrt{T_k \varepsilon^2} & \text{if } \|Y_{(k)}\| < \sqrt{T_k \varepsilon^2} \\ \sqrt{T_k \varepsilon^2} + c(j_k \pi)^{-m} & \text{if } \|Y_{(k)}\| > \sqrt{T_k \varepsilon^2} + c(j_k \pi)^{-m} \\ \|Y_{(k)}\| & \text{otherwise.} \end{cases}$$

2.2. *Allocation of bits.* Let $(\tilde{b}_k)_{k=1}^K$ be the solution to the optimization

$$\begin{aligned} \min_{\bar{b}} \quad & \sum_{k=1}^{K} \frac{(\check{S}_k^2 - T_k \varepsilon^2)^2}{\check{S}_k^2} \cdot 2^{-2\bar{b}_k} \\ \text{such that} \quad & \sum_{k=1}^{K} T_k \bar{b}_k \leq B, \ \bar{b}_k \geq 0. \end{aligned} \tag{3.12}$$

2.3. *Encoding block direction.* Form the data-dependent codebook as follows.

Divide the rows of $\mathcal{Z}$ into blocks of sizes $2^{\lceil T_1 \tilde{b}_1 \rceil}, \ldots, 2^{\lceil T_K \tilde{b}_K \rceil}$. Based on the kth block of rows, construct the data-dependent codebook $\tilde{\mathcal{Z}}_k$ by keeping only the first T_k entries and normalizing each truncated row; specifically, the jth row of $\tilde{\mathcal{Z}}_k$ is given by

$$\tilde{\mathcal{Z}}_{k,j} = \frac{\mathcal{Z}_{i,1:T_k}}{\|\mathcal{Z}_{i,1:T_k}\|} \in \mathbb{S}_{T_k - 1}$$

where i is the appropriate row of the base code $\mathcal{Z}$ and $\mathcal{Z}_{i,1:t}$ denotes the first t entries of the row vector. A graphical illustration is shown below in Figure 3.2.

With this data-dependent codebook, encode

$$\check{Z}_{(k)} = \arg\max\{\langle z, Y_{(k)}\rangle : z \in \tilde{\mathcal{Z}}_k\}$$

for $k = 1, \ldots, K$.

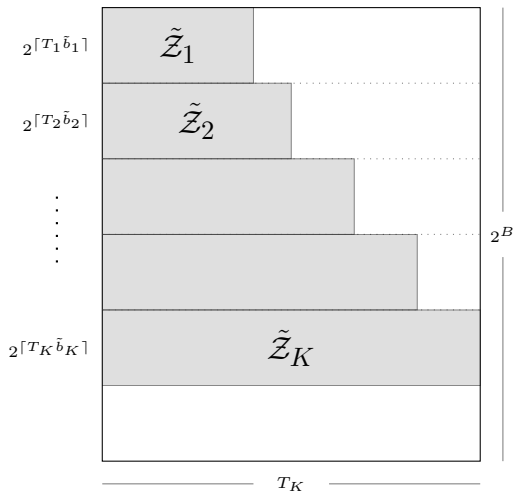

Figure 3.2: An illustration of the data-dependent codebook. The big matrix represents the base code $\mathcal{Z}$, and the shaded areas are $(\tilde{\mathcal{Z}}_k)$, sub-matrices of size $T_k \times 2^{\lceil T_k\tilde{b}_k\rceil}$ with rows normalized.

Step 3. *Transmission.* Transmit or store $(\check{S}_k)_{k=1}^K$ and $(\check{Z}_{(k)})_{k=1}^K$ by their corresponding indices.

Step 4. *Decoding & Estimation.*

4.1. Recover $(\check{S}_k)$ based on the transmitted or stored indices and the common codebook $(\mathcal{S}_k)$.

4.2. Solve (3.12) and get $(\tilde{b}_k)$. Reconstruct $(\tilde{\mathcal{Z}}_k)$ using $\mathcal{Z}$ and $(\tilde{b}_k)$.

4.3. Recover $(\check{Z}_{(k)})$ based on the transmitted or stored indices and the reconstructed codebook $(\tilde{\mathcal{Z}}_k)$.

4.4. Estimate $\theta_{(k)}$ by

$$\check{\theta}_{(k)} = \frac{\check{S}_k^2 - T_k\varepsilon^2}{\check{S}_k}\sqrt{1 - 2^{-2\tilde{b}_k}} \cdot \check{Z}_{(k)}.$$

4.5. Estimate the entire vector θ by concatenating the $\check{\theta}_{(k)}$ vectors and padding with zeros; thus,

$$\check{\theta} = \left(\check{\theta}_{(1)}, \ldots, \check{\theta}_{(K)}, 0, 0, \ldots\right).$$

The following theorem establishes the asymptotic optimality of this quantized estimator.

Theorem 3.4.1. *Let $\check{\theta}$ be the quantized estimator defined above.*

(i) If $B\varepsilon^{\frac{2}{2m+1}} \to \infty$, then

$$\lim_{\varepsilon\to 0} \varepsilon^{-\frac{4m}{2m+1}} \sup_{\theta\in\Theta(m,c)} \mathbb{E}\|\theta - \check{\theta}\|^2 = \mathsf{P}_{m,c}.$$

(ii) If $B\varepsilon^{\frac{2}{2m+1}} \to d$ for some constant d as $\varepsilon \to 0$, then

$$\lim_{\varepsilon\to 0} \varepsilon^{-\frac{4m}{2m+1}} \sup_{\theta\in\Theta(m,c)} \mathbb{E}\|\theta - \check{\theta}\|^2 = \mathsf{P}_{m,c} + \mathsf{Q}_{d,m,c}.$$

(iii) If $B\varepsilon^{\frac{2}{2m+1}} \to 0$ and $B(\log(1/\varepsilon))^{-3} \to \infty$, then

$$\lim_{\varepsilon\to 0} B^{2m} \sup_{\theta\in\Theta(m,c)} \mathbb{E}\|\theta - \check{\theta}\|^2 = \frac{c^2 m^{2m}}{\pi^{2m}}.$$

The expectations are with respect to the random quantized estimation scheme Q and the distribution of the data.

We pause to make several remarks on this result before outlining the proof.

Remark 3.4.1.1. The total number of bits used by this quantized estimation scheme is

$$\begin{aligned}\sum_{k=1}^{K} \lceil T_k \tilde{b}_k \rceil + \sum_{k=1}^{K} \log\lceil \varepsilon^{-2} c (j_k \pi)^{-m} \rceil &\le \sum_{k=1}^{K} \lceil T_k \tilde{b}_k \rceil + \sum_{k=1}^{K} \log\lceil \varepsilon^{-2} c \rceil \\ &\le B + K + 2K\rho_\varepsilon^{-1} + K \log\lceil c \rceil \\ &= B + O((\log(1/\varepsilon))^3),\end{aligned}$$

where we use the fact that $K \lesssim \log^2(1/\varepsilon^2)$ (See Lemma 3.7.3). Therefore, as long as $B(\log(1/\varepsilon))^{-3} \to \infty$, the total number of bits used is asymptotically no more than B, the given communication budget.

Remark 3.4.1.2. The quantized estimation scheme does not make essential use of the parameters of the Sobolev space, namely the smoothness m and the radius c. The only exception is that in Step 1.1 the size of the codebook $\mathcal{S}_k$ depends on m and c. However, suppose that we know a lower bound on the smoothness m, say $m \ge m_0$, and an upper bound on the radius c, say $c \le c_0$. By replacing m and c by m_0 and c_0 respectively, we make the codebook independent of the parameters. We shall assume $m_0 > 1/2$, which leads to continuous functions. This modification does not, however, significantly increase the number of bits; in fact, the total number of bits is still $B + O(\rho_\varepsilon^{-3})$. Thus, we can easily make this quantized estimator minimax adaptive to the class of Sobolev ellipsoids $\{\Theta(m, c) : m \ge m_0,\ c \le c_0\}$, as long as B grows faster than $(\log(1/\varepsilon))^3$. More formally, we have

Corollary 3.4.2. *Suppose that B_ε satisfies $B_\varepsilon(\log(1/\varepsilon))^{-3} \to \infty$. Let $\check{\theta}'$ be the quantized estimator with the modification described above, which does not assume knowledge of m and c. Then for $m \ge m_0$ and $c \le c_0$,*

$$\lim_{\varepsilon \to 0} \frac{\sup_{\theta \in \Theta(m,c)} \mathbb{E}\|\theta - \check{\theta}'\|^2}{\inf_{\widehat{\theta}, C(\widehat{\theta}) \le B} \sup_{\theta \in \Theta(m,c)} \mathbb{E}\|\theta - \widehat{\theta}\|^2} = 1,$$

where the expectation in the numerator is with respect to the data and the randomized coding scheme, while the expectation in the denominator is only with respect to the data.

Remark 3.4.2.1. When B grows at a rate comparable to or slower than $(\log(1/\varepsilon))^3$, the lower bound is still achievable, just no longer by the quantized estimator we described above. The main reason is that when B does not grow faster than $\log(1/\varepsilon)^3$, the block size $T_1 = \lceil \log(1/\varepsilon) \rceil$ is too large. The blocking needs to be modified to get achievability in this case.

Remark 3.4.2.2. In classical rate distortion (Cover and Thomas, 2006; Gallager, 1968), the probabilistic method applied to a randomized coding scheme shows the existence of a code achieving the rate distortion bounds. According to Theorem 3.3.1, the expected risk, averaged over the randomness in the codebook, similarly achieves the quantized minimax lower bound. However, note that the average over the codebook is inside the supremum over the Sobolev space, implying that the code achieving the bound may vary over the ellipsoid. In other words, while the coding scheme generates a codebook that is used for different θ, it is not known whether there is one code generated by this randomized scheme that is "universal," and achieves the risk lower bound with high probability over the ellipsoid. The existence or non-existence of such "universal codes" is an interesting direction for further study.

Remark 3.4.2.3. We have so far dealt with the periodic case, i.e., functions in the periodic Sobolev space $\tilde{W}(m, c)$ defined in (3.4). For the Sobolev space $W(m, c)$, where the functions are not necessarily periodic, the lower bound given in Theorem 3.3.1 still holds, since $\tilde{W}(m, c)$ is a subset of the larger class $W(m, c)$. To extend the achievability result to $W(m, c)$, we again need to relate $W(m, c)$ to an ellipsoid. Nussbaum et al. (1985) shows using spline theory that the non-periodic space can actually be expressed as an ellipsoid, where the length of the jth principal axis scales as $(\pi^2 j)^m$ asymptotically. Based on this link between $W(m, c)$ and the ellipsoid, the techniques used here to show achievability apply, and since the principal axes scale as in the periodic case, the convergence rates remain the same.

Proof of Theorem 3.4.1 We now sketch the proof of Theorem 3.4.1, deferring the full details to Section 3.7. To provide only an informal outline of the proof, we shall write

$A_1 \approx A_2$ as a shorthand for $A_1 = A_2(1+o(1))$, and $A_1 \lesssim A_2$ for $A_1 \leq A_2(1+o(1))$, without specifying here what these $o(1)$ terms are.

To upper bound the risk $\mathbb{E}\|\check{\theta} - \theta\|^2$, we adopt the following sequence of approximations and inequalities. First, we discard the components whose index is greater than N and show that

$$\mathbb{E}\|\check{\theta} - \theta\|^2 \approx \mathbb{E}\sum_{k=1}^{K} \|\check{\theta}_{(k)} - \theta_{(k)}\|^2.$$

Since $\check{S}_k$ is close enough to S_k, we can then safely replace $\check{\theta}_{(k)}$ by $\widehat{\theta}_{(k)} = \frac{S_k^2 - T_k\varepsilon^2}{S_k}\sqrt{1 - 2^{-2\tilde{b}_{(k)}}}\cdot \check{Z}_{(k)}$ and obtain

$$\approx \mathbb{E}\sum_{k=1}^{K} \|\widehat{\theta}_{(k)} - \theta_{(k)}\|^2.$$

Writing $\lambda_k = \frac{S_k^2 - T_k\varepsilon^2}{S_k^2}$, we further decompose the risk into

$$= \mathbb{E}\sum_{k=1}^{K} \Big(\|\widehat{\theta}_{(k)} - \lambda_k Y_{(k)}\|^2 + \|\lambda_k Y_{(k)} - \theta_{(k)}\|^2 + 2\langle \widehat{\theta}_{(k)} - \lambda_k Y_{(k)}, \lambda_k Y_{(k)} - \theta_{(k)} \rangle \Big).$$

Conditioning on the data Y and taking the expectation with respect to the random codebook yields

$$\lesssim \mathbb{E}\sum_{k=1}^{K} \left(\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k} + \|\lambda_k Y_{(k)} - \theta_{(k)}\|^2 \right).$$

By two oracle inequalities upper bounding the expectations with respect to the data, and the fact that $\tilde{b}$ is the solution to (3.12),

$$\lesssim \min_{b \in \Pi_{\mathrm{blk}}(B)} \sum_{k=1}^{K} \left(\frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} 2^{-2\bar{b}_k} + \frac{\|\theta_{(k)}\|^2 T_k \varepsilon^2}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} \right).$$

Showing that the blockwise constant oracles are almost as good as the monotone oracle, we get for some $B' \approx B$

$$\lesssim \min_{b \in \Pi_{\mathrm{mon}}(B'),\ \omega \in \Omega_{\mathrm{mon}}} \sum_{j=1}^{N} \left(\frac{\theta_j^4}{\theta_j^2 + \varepsilon^2} 2^{-2b_j} + (1-\omega_j)^2 \theta_j^2 + \omega_j^2 \varepsilon^2 \right),$$

where $\Pi_{\mathrm{blk}}(B)$, $\Pi_{\mathrm{mon}}(B)$ are the classes of blockwise constant and monotone allocations of the bits defined in (3.18), (3.19), and Ω_{mon} is the class of monotone weights defined in (3.21). The proof is then completed by Lemma 3.7.8 showing that the last quantity is equal to $V_\varepsilon(m, c, B)$.

3.5 Experiments

Here we illustrate the performance of the proposed quantized estimation scheme. We use the function

$$f(x) = \sqrt{x(1-x)} \sin\left(\frac{2.1\pi}{x + 0.3} \right), \quad 0 \le x \le 1,$$

which we shall refer to as the "damped Doppler function," shown in Figure 3.3 (the gray lines). Note that the value 0.3 differs from the value 0.05 in the usual Doppler function used to illustrate spatial adaptation of methods such as wavelets. Since we do not address spatial adaptivity in this paper, we "slow" the oscillations of the Doppler function near zero in our illustrations.

We use this f as the underlying true mean function and generate our data according to

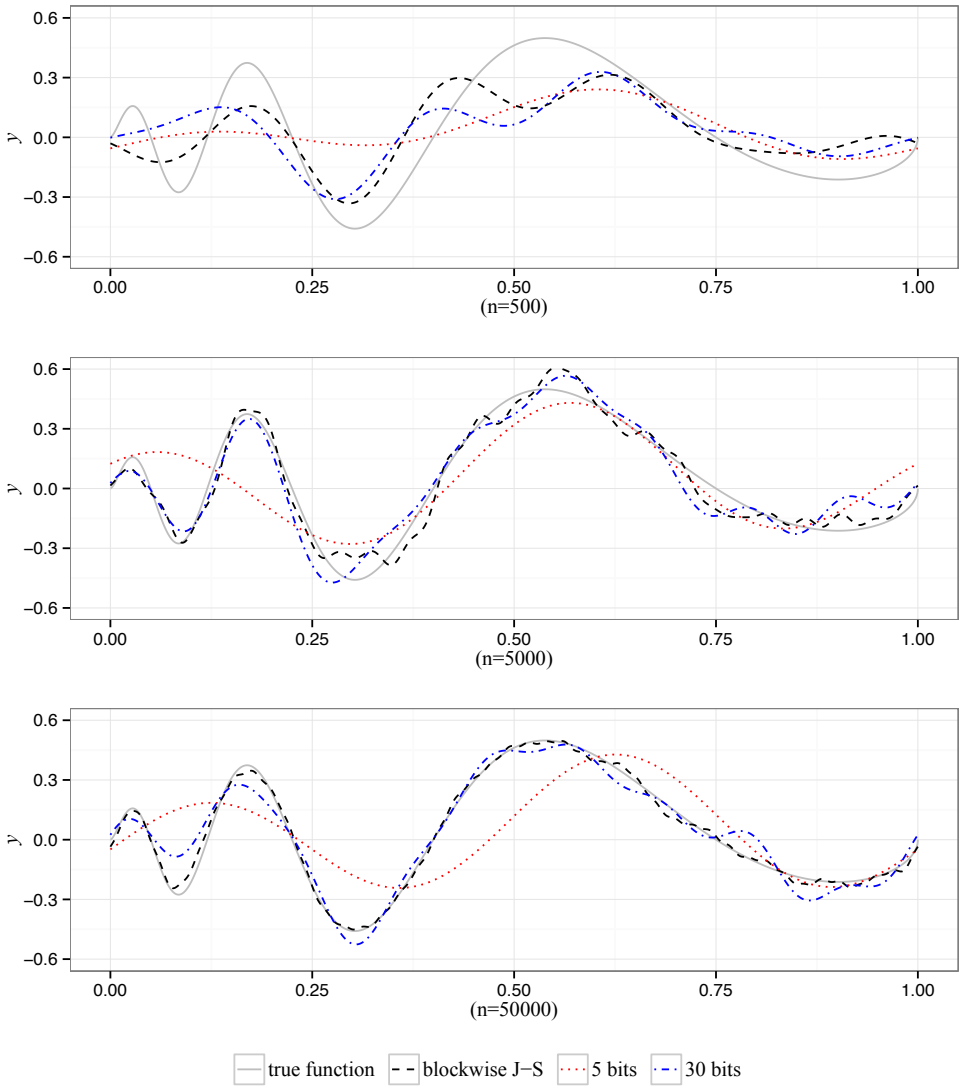

Figure 3.3: The damped Doppler function (solid gray) and typical realizations of the estimators under different noise levels ($n = 500$, 5000, and 50000). Three estimators are used: the blockwise James-Stein estimator (dashed black), and two quantized estimator with budgets of 5 bits (dashed red) and 30 bits (dashed blue). The 5-bit budget appears to be "sufficient" in the first setting but "insufficient" in the latter two, while the 30-bit one changes from "over-sufficient" to "sufficient" and finally "insufficient."

the corresponding white noise model (3.1). Recall that the white noise model is defined as

$$dX(t) = f(t)dt + \varepsilon dW(t), \quad 0 \le t \le 1.$$

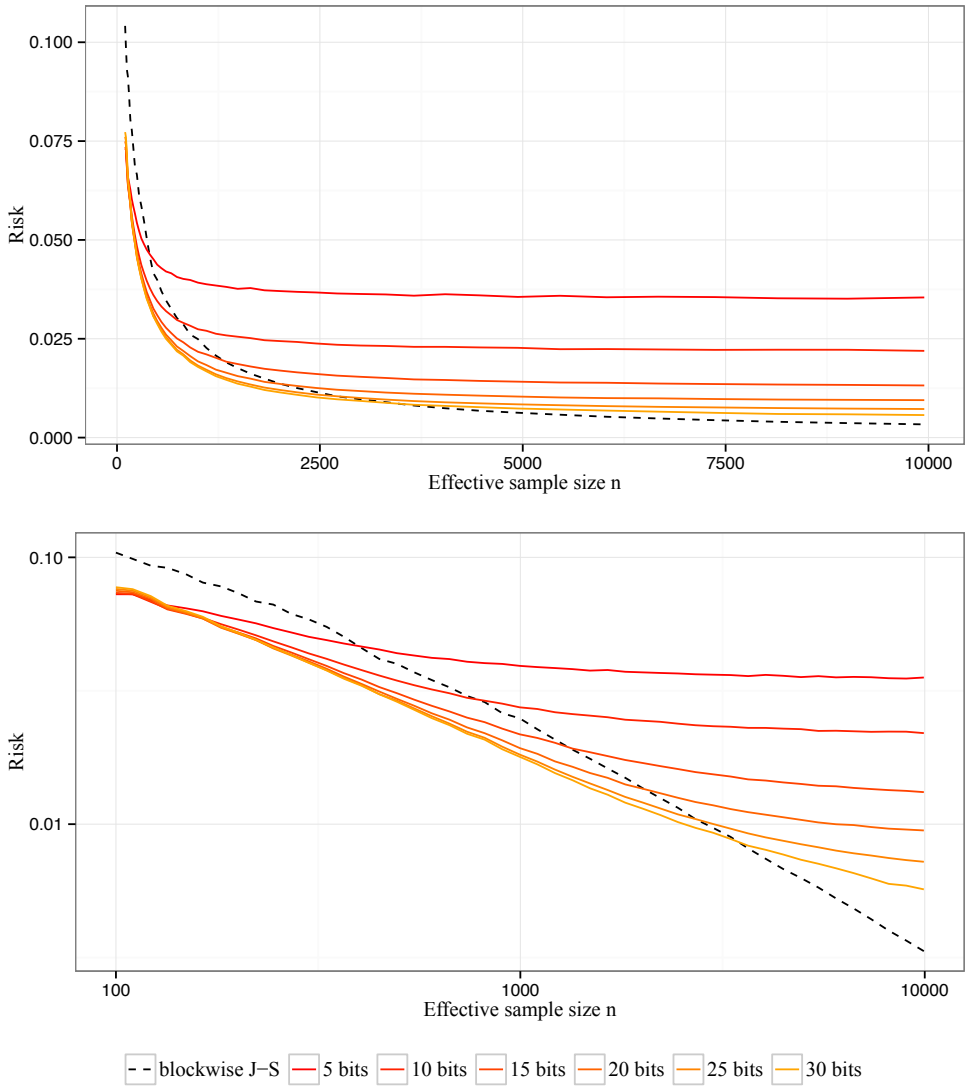

Figure 3.4: Risk versus effective sample size $n = 1/\varepsilon^2$ for estimating the damped Doppler function with different estimators. The dashed line represents the risk of the blockwise James-Stein estimator, and the solid ones are for the quantized estimators with different budgets. The budgets are 5, 10, 15, 20, 25, and 30 bits, corresponding to the lines from top to bottom. The two plots are the same curves on the original scale and the log-log scale.

We apply the blockwise James-Stein estimator, as well as the proposed quantized estimator with different communication budgets. We also vary the noise level ε and, equivalently, the effective sample size $n = 1/\varepsilon^2$.

We first show in Figure 3.3 some typical realizations of these estimators on data generated under different noise levels ($n = 500$, 5000, and 50000 respectively). To keep the plots

succinct, we show only the true function, the blockwise James-Stein estimates and quantized estimates using total bit budgets of 5 and 30 bits. We observe, in the first plot, that both quantized estimates deviate from the true function, and so does the blockwise James-Stein estimates. This is when the noise is relatively large and any quantized estimates work similarly poorly no matter how large a budget is given. Both 5 bits and 30 bits seems to be "sufficient/over-sufficient" here. In the second plot, the blockwise James-Stein estimate is close to the quantized estimate with a budget of 30 bits while with a budget of 5 bits it fails to capture the fluctuations of the true function. Thus, a budget of 30 bits is still "sufficient," but 5 bits apparently becomes "insufficient." In the third plot, the blockwise James-Stein estimate gives a better fit than the two quantized estimates, as both budgets become "insufficient" to achieve the optimal risk.

Next, in Figure 3.4 we plot the risk as a function of sample size n, averaging over 2000 simulations. Note that the bottom plot is the just the first plot on a log-log scale. In this set of plots, we are able to observe the phase transition for the quantized estimators. For relatively small values of n, all quantized estimators yield a similar error rate, with risks that are close to (or even smaller than) that of the blockwise James-Stein estimator. This is the over-sufficient regime—even the smallest budget suffices to achieve the optimal risk. As n increases, the curves start to separate, with estimators having smaller bit budgets leading to worse risks compared to the blockwise James-Stein estimator, and compared to estimators with larger budgets. This can be seen as the sufficient regime for the small-budget estimators—the risks are still going down, but at a slower rate than optimal. The six quantized estimators all end up in the insufficient regime—as n increases, their risks stay constant, while the risk of the blockwise James-Stein estimator continues to decrease.

3.6 Related work and future directions

Concepts related to quantized nonparametric estimation appear in multiple communities. As mentioned in the introduction, Donoho's 1997 Wald Lectures (on the eve of the 50th

anniversary of Shannon's 1948 paper), drew sharp parallels between rate distortion, metric entropy and minimax rates, focusing on the same Sobolev function spaces we treat here. One view of the present work is that we take this correspondence further by studying how the risk continuously degrades with the level of quantization. We have analyzed the precise leading order asymptotics for quantized regression over the Sobolev spaces, showing that these rates and constants are realized with coding schemes that are adaptive to the smoothness m and radius c of the ellipsoid, achieving automatically the optimal rate for the regime corresponding to those parameters given the specified communication budget. Our detailed analysis is possible due to what Nussbaum (Nussbaum, 1999) calls the "Pinsker phenomenon," referring to the fact that linear filters attain the minimax rate in the over-sufficient regime. It will be interesting to study quantized nonparametric estimation in cases where the Pinsker phenomenon does not hold, for example over Besov bodies and different L_p spaces.

Many problems of rate distortion type are similar to quantized regression. The standard "reverse water filling" construction to quantize a Gaussian source with varying noise levels plays a key role in our analysis, as shown in Section 3.3. In our case the Sobolev ellipsoid is an infinite Gaussian sequence model, requiring truncation of the sequence at the appropriate level depending on the targeted quantization and estimation error. In the case of Euclidean balls, Draper and Wornell (2004) study rate distortion problems motivated by communication in sensor networks; this is closely related to the problem of quantized minimax estimation over Euclidean balls that we analyzed in Zhu and Lafferty (2014). The essential difference between rate distortion and our quantized minimax framework is that in rate distortion the quantization is carried out for a random source, while in quantized estimation we quantize our estimate of the deterministic and unknown basis coefficients. Since linear estimators are asymptotically minimax for Sobolev spaces under squared error (the "Pinsker phenomenon"), this naturally leads to an alternative view of quantizing the observations, or said differently, of compressing the data before estimation.

Statistical estimation from compressed data has appeared previously in different commu-

nities. In Zhou et al. (2009) a procedure is analyzed that compresses data by random linear transformations in the setting of sparse linear regression. Zhang and Berger (1988) study estimation problems when the data are communicated from multiple sources; Ahlswede and Csiszár (1986) consider testing problems under communication constraints; the use of side information is studied by Ahlswede and Burnashev (1990); other formulations in terms of multiterminal information theory are given by Han and Amari (1998); nonparametric problems are considered by Raginsky (2007). In a distributed setting the data may be divided across different compute nodes, with distributed estimates then aggregated or pooled by communicating with a central node. The general "CEO problem" of distributed estimation was introduced by Berger et al. (1996), and has been recently studied in parametric settings in Zhang et al. (2013) and Garg et al. (2014). These papers take the view that the data are communicated to the statistician at a certain rate, which may introduce distortion, and the goal is to study the degradation of the estimation error. In contrast, in our setting we can view the unquantized data as being fully available to the statistician at the time of estimation, with communication constraints being imposed when communicating the estimated model to a remote location.

Finally, our quantized minimax analysis shows achievability using random coding schemes, which are not computationally efficient. A natural problem is to develop practical coding schemes that come close to the quantized minimax lower bounds. In our view, the most promising approach currently is to exploit source coding schemes based on greedy sparse regression Venkataramanan et al. (2013), applying such techniques blockwise according to the procedure we developed in Section 3.4.

3.7 Proofs of technical results

In this section, we provide proofs for Theorems 3.3.1 and 3.4.1.

3.7.1 Proof of Theorem 3.3.1

We first show

Lemma 3.7.1. *The quantized minimax risk is lower bounded by $V_\varepsilon(m, c, B_\varepsilon)$, the value of the optimization ($\mathcal{P}_1$).*

Proof. As will be clear to the reader, $V_\varepsilon(m, c, B_\varepsilon)$ is achieved by some σ^2 that is non-increasing and finitely supported. Let σ^2 be such that

$$\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0 = \sigma_{n+1} = \ldots, \ \sum_{j=1}^{n} a_j^2 \sigma_j^2 = \frac{c^2}{\pi^{2m}},$$

and let

$$\Theta_n(m, c) = \{\theta \in \ell_2 : \sum_{j=1}^{n} a_j^2 \theta_j^2 \leq \frac{c^2}{\pi^{2m}}, \ \theta_j = 0 \text{ for } j \geq n+1\} \subset \Theta(m, c).$$

For $\tau \in (0, 1)$, write $s_j^2 = (1-\tau)\sigma_j^2$ and let $\pi_\tau(\theta; \sigma^2)$ be a the prior distribution on θ such that

$$\begin{aligned} &\theta_j \sim \mathcal{N}(0, s_j^2), \quad j = 1, \ldots, n, \\ &\mathbb{P}(\theta_j = 0) = 1, \quad j \geq n+1. \end{aligned}$$

We observe that

$$\begin{aligned} R_\varepsilon(m, c, B_\varepsilon) &\geq \inf_{\widehat{\theta}, C(\widehat{\theta}) \leq B_\varepsilon} \sup_{\theta \in \Theta_n(m,c)} \mathbb{E}\|\theta - \widehat{\theta}\|^2 \\ &\geq \inf_{\widehat{\theta}, C(\widehat{\theta}) \leq B_\varepsilon} \int_{\Theta_n(m,c)} \mathbb{E}\|\theta - \widehat{\theta}\|^2 d\pi_\tau(\theta; \sigma^2) \\ &\geq I - r \end{aligned}$$

where I is the integrated risk of the optimal quantized estimator

$$I = \inf_{\widehat{\theta}, C(\widehat{\theta}) \leq B_\varepsilon} \int_{\mathbb{R}^n \otimes \{0\}^\infty} \mathbb{E}\|\theta - \widehat{\theta}\|^2 d\pi_\tau(\theta; \sigma^2)$$

and r is the residual

$$r = \sup_{\widehat{\theta} \in \Theta(m,c)} \int_{\overline{\Theta(m,c)}} \mathbb{E}\|\theta - \widehat{\theta}\|^2 d\pi_\tau(\theta; \sigma^2)$$

where $\overline{\Theta(m,c)} = (\mathbb{R}^n \otimes \{0\}^\infty) \backslash \Theta_n(m,c)$. As shown in Section 3.3, $\lim_{\tau \to 0} I$ is lower bounded by the value of the optimization

$$\begin{aligned} \min_{\mu^2} \quad & \sum_{j=1}^{\infty} \mu_j^2 + \sum_{j=1}^{\infty} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2} \\ \text{such that} \quad & \sum_{j=1}^{\infty} \frac{1}{2} \log_+ \left(\frac{\sigma_j^4}{\mu_j^2 (\sigma_j^2 + \varepsilon^2)} \right) \leq B_\varepsilon. \end{aligned}$$

It then suffices to show that $r = o(I)$ as $\varepsilon \to 0$. Let $d_n = \sup_{\theta \in \Theta_n(m,c)} \|\theta\|$. We have

$$\begin{aligned} r &= \sup_{\widehat{\theta} \in \Theta(m,c)} \int_{\overline{\Theta_n(m,c)}} \mathbb{E}\|\theta - \widehat{\theta}\|^2 d\pi_\tau(\theta; \sigma^2) \\ &\leq 2 \int_{\overline{\Theta_n(m,c)}} (d_n^2 + \mathbb{E}\|\theta\|^2) d\pi_\tau(\theta; \sigma^2) \\ &= 2 \left(d_n^2 \, \mathbb{P}\left(\theta \notin \Theta_n(m,c)\right) + \left(\mathbb{P}(\theta \notin \Theta_n(m,c)) \mathbb{E}\|\theta\|^4 \right)^{1/2} \right) \end{aligned}$$

where we use the Cauchy-Schwarz inequality. Noticing that

$$\begin{aligned}\mathbb{E}\|\theta\|^4 &= \mathbb{E}\Big(\Big(\sum_{j=1}^n \theta_j^2\Big)^2\Big) \\ &= \sum_{j_1\neq j_2} \mathbb{E}(\theta_{j_1}^2)\mathbb{E}(\theta_{j_2}^2) + \sum_{j=1}^n \mathbb{E}(\theta_j^4) \\ &\leq \sum_{j_1\neq j_2} s_{j_1}^2 s_{j_2}^2 + 3\sum_{j=1}^n s_j^4 \\ &\leq 3\Big(\sum_{j=1}^n s_j^2\Big)^2 \leq 3d_n^4,\end{aligned}$$

we obtain

$$\begin{aligned}r &\leq 2d_n^2\left(\mathbb{P}\left(\theta\notin\Theta_n(m,c)\right) + \sqrt{3\mathbb{P}\left(\theta\notin\Theta_n(m,c)\right)}\right) \\ &\leq 6d_n^2\sqrt{\mathbb{P}\left(\theta\notin\Theta_n(m,c)\right)}.\end{aligned}$$

Thus, we only need to show that $\sqrt{\mathbb{P}\left(\theta\notin\Theta_n(m,c)\right)} = o(I)$. In fact,

$$\begin{aligned}&\mathbb{P}(\theta\notin\Theta_n(m,c)) \\ &\quad= \mathbb{P}\left(\sum_{j=1}^n a_j^2\theta_j^2 > \frac{c^2}{\pi^{2m}}\right) \\ &\quad= \mathbb{P}\left(\sum_{j=1}^n a_j^2(\theta_j^2 - \mathbb{E}(\theta_j^2)) > \frac{c^2}{\pi^{2m}} - (1-\tau)\sum_{j=1}^n a_j^2\sigma_j^2\right) \\ &\quad= \mathbb{P}\left(\sum_{j=1}^n a_j^2(\theta_j^2 - \mathbb{E}(\theta_j^2)) > \frac{\tau c^2}{\pi^{2m}}\right) \\ &\quad= \mathbb{P}\left(\sum_{j=1}^n a_j^2 s_j^2(Z_j^2 - 1) > \frac{\tau}{1-\tau}\sum_{j=1}^n a_j^2 s_j^2\right)\end{aligned}$$

where $Z_j \sim \mathcal{N}(0,1)$. By Lemma 3.7.2, we get

$$\mathbb{P}(\theta\notin\Theta_n(m,c)) \leq \exp\left(-\frac{\tau^2}{8(1-\tau)^2}\frac{\sum_{j=1}^n a_j^2 s_j^2}{\max_{1\leq j\leq n} a_j^2 s_j^2}\right)$$

For the σ^2 that achieves $V_\varepsilon(m, c, B_\varepsilon)$, we have that $\sqrt{\mathbb{P}(\theta \notin \Theta_n(m, c))} = o(I)$. □

Lemma 3.7.2 (Lemma 3.5 in Tsybakov (2008))**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(0, 1)$. For $t \in (0, 1)$ and $\omega_j > 0$, $j = 1, \ldots, n$, we have*

$$\mathbb{P}\left(\sum_{j=1}^{n} \omega_j (X_j^2 - 1) > t \sum_{j=1}^{n} X_j\right) \leq \exp\left(-\frac{t^2 \sum_{j=1}^{n} \omega_j}{8 \max_{1 \leq j \leq n} \omega_j}\right).$$

Proof of Lemma 3.3.2. This is in fact Pinsker's theorem, which gives the exact asymptotic minimax risk of estimation of normal means in the Sobolev ellipsoid. The proof can be found in Nussbaum (1999) and Tsybakov (2008). □

Proof of Lemma 3.3.3. As argued in Section 3.3 for the lower bound in the sufficient regime, optimization problem $(\mathcal{Q}_1)$ can be reformulated as

$$\begin{aligned} \max_{\sigma^2, J} \quad & J\eta \\ \text{such that} \quad & \frac{1}{2} \sum_{j=1}^{J} \log_+ \left(\frac{\sigma_j^4}{\eta(\sigma_j^2 + \varepsilon^2)}\right) \leq B_\varepsilon \\ & \sum_{j=1}^{J} a_j^2 \sigma_j^2 \leq \frac{c^2}{\pi^{2m}} \\ & (\sigma_j^2) \text{ is decreasing and } \frac{\sigma_J^4}{\sigma_J^2 + \varepsilon^2} \geq \eta. \end{aligned} \tag{$\mathcal{Q}_2$}$$

Now suppose that we have a series (σ_j^2) which satisfies the last constraint and is supported

on $\{1, \ldots, J\}$. By the first constraint, we have that

$$\begin{aligned}
J\eta &= J \exp\left(-\frac{2B_\varepsilon}{J}\right) \left(\prod_{j=1}^{J} \frac{\sigma_j^4}{\sigma_j^2 + \varepsilon^2}\right)^{\frac{1}{J}} \\
&\leq J \exp\left(-\frac{2B_\varepsilon}{J}\right) \left(\prod_{j=1}^{J} \sigma_j^2\right)^{\frac{1}{J}} \\
&= J \exp\left(-\frac{2B_\varepsilon}{J}\right) \left(\prod_{j=1}^{J} a_j^2 \sigma_j^2\right)^{\frac{1}{J}} \left(\prod_{j=1}^{J} a_j^{-2}\right)^{\frac{1}{J}} \\
&\leq \exp\left(-\frac{2B_\varepsilon}{J}\right) \left(\sum_{j=1}^{J} a_j^2 \sigma_j^2\right) \left(\prod_{j=1}^{J} a_j^{-2}\right)^{\frac{1}{J}} \\
&\leq \frac{c^2}{\pi^{2m}} \exp\left(-\frac{2B_\varepsilon}{J}\right) \left(\prod_{j=1}^{J} a_j^{-2}\right)^{\frac{1}{J}} \\
&= \frac{c^2}{\pi^{2m}} \left(\exp\left(\frac{B_\varepsilon}{m}\right) J!\right)^{-\frac{2m}{J}}. \qquad (3.13)
\end{aligned}$$

This provides a series of upper bounds for $Q_\varepsilon(m, c, B_\varepsilon)$ parameterized by J. Minimizing (3.13) over J, we obtain that the optimal J satisfies

$$\frac{J^J}{J!} < \exp\left(\frac{B_\varepsilon}{m}\right) \leq \frac{(J+1)^{J+1}}{(J+1)!}. \qquad (3.14)$$

Denote this optimal J by J_ε. By Stirling's approximation, we have

$$\lim_{\varepsilon \to 0} \frac{B_\varepsilon / m}{J_\varepsilon} = 1,$$

and plugging this asymptote into (3.13), we get as $\varepsilon \to 0$

$$\frac{c^2}{\pi^{2m}} \left(\exp\left(\frac{B_\varepsilon}{m}\right) J_\varepsilon!\right)^{-\frac{2m}{J_\varepsilon}} \sim \frac{c^2}{\pi^{2m}} J_\varepsilon^{-2m} \sim \frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}.$$

This gives the desired upper bound (3.9).

Next we show that the upper bound (3.9) is asymptotically achievable when $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to 0$

and $B_\varepsilon \to \infty$. It suffices to find a feasible solution that attains (3.9). Let

$$\tilde{\sigma}_j^2 = \frac{c^2/\pi^{2m}}{J_\varepsilon a_j^2},\ j = 1, \ldots, J_\varepsilon.$$

Note that the entire sequence of $(\tilde{\sigma}_j^2)_{j=1}^{J_\varepsilon}$ does not qualify for a feasible solution, since the first constraint in $(\mathcal{Q}_2)$ won't be satisfied for any $\eta \leq \frac{\tilde{\sigma}_{J_\varepsilon}^4}{\tilde{\sigma}_{J_\varepsilon}^2+\varepsilon^2}$. We keep only the first J'_ε terms of $(\tilde{\sigma}_j^2)$, where J'_ε is the largest j such that

$$\frac{\tilde{\sigma}_j^4}{\tilde{\sigma}_j^2 + \varepsilon^2} \geq \tilde{\sigma}_{J_\varepsilon}^2. \tag{3.15}$$

Thus,

$$\sum_{j=1}^{J'_\varepsilon} \frac{1}{2} \log_+ \left(\frac{\frac{\tilde{\sigma}_j^4}{\tilde{\sigma}_j^2+\varepsilon^2}}{\tilde{\sigma}_{J_\varepsilon}^2} \right) \leq \sum_{j=1}^{J'_\varepsilon} \frac{1}{2} \log_+ \left(\frac{\tilde{\sigma}_j^2}{\tilde{\sigma}_{J_\varepsilon}^2} \right) \leq \sum_{j=1}^{J_\varepsilon} \frac{1}{2} \log_+ \left(\frac{\tilde{\sigma}_j^2}{\tilde{\sigma}_{J_\varepsilon}^2} \right) \leq B_\varepsilon,$$

where the last inequality is due to (3.14). This tells us that setting $\eta = \tilde{\sigma}_{J_\varepsilon}^2$ leads to a feasible solution to $(\mathcal{Q}_2)$. As a result,

$$Q_\varepsilon(m, c, B_\varepsilon) \geq J'_\varepsilon \tilde{\sigma}_{J_\varepsilon}^2. \tag{3.16}$$

If we can show that $J'_\varepsilon \sim J_\varepsilon$, then

$$J'_\varepsilon \tilde{\sigma}_{J_\varepsilon}^2 \sim J_\varepsilon \tilde{\sigma}_{J_\varepsilon}^2 \sim \frac{c^2 m^{2m}}{\pi^{2m}} B_\varepsilon^{-2m}. \tag{3.17}$$

To show that $J'_\varepsilon \sim J_\varepsilon$, it suffices to show that $a_{J'_\varepsilon} \sim a_{J_\varepsilon}$. Plugging the formula of $\tilde{\sigma}_j^2$ into (3.15) and solving for $a_{J'_\varepsilon}^2$, we get

$$a_{J'_\varepsilon}^2 \sim \frac{-\frac{c^2}{\pi^{2m} J_\varepsilon} + \sqrt{(\frac{c^2}{\pi^{2m} J_\varepsilon})^2 + 4\frac{c^2}{\pi^{2m} J_\varepsilon} \varepsilon^2 a_{J_\varepsilon}^2}}{2\varepsilon^2} \sim a_{J_\varepsilon}^2$$

where the last equivalence is due to the assumption $B_\varepsilon \varepsilon^{\frac{2}{2m+1}} \to 0$ and L'Hôpital's rule. $\square$

Proof of Lemma 3.3.5. Suppose that $\sigma^2(x)$ with x_0 solves $(\mathcal{P}_4)$. Consider function $\sigma^2(x) + \xi v(x)$ such that it is still feasible for $(\mathcal{P}_4)$, and thus we have

$$\int_0^{x_0} x^{2m} v(x) dx \leq 0.$$

Now plugging $\sigma^2(x) + \xi v(x)$ for $\sigma^2(x)$ in the objective function of $(\mathcal{P}_4)$, taking derivative with respect to ξ, and letting $\xi \to 0$, we must have

$$\int_0^{x_0} \frac{v(x)}{(\sigma^2(x)+1)^2} dx + x_0 \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right) \frac{1}{x_0} \int_0^{x_0} \frac{2v(x)}{\sigma^2(x)} - \frac{v(x)}{\sigma^2(x)+1} dx \leq 0,$$

which, after some calculation and rearrangement of terms, yields

$$\int_0^{x_0} v(x) \left(\frac{1}{(\sigma^2(x)+1)^2} + \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right) \frac{\sigma^2(x)+2}{\sigma^2(x)(\sigma^2(x)+1)} \right) dx \leq 0.$$

Thus, we obtain that, for some λ

$$\frac{1}{(\sigma^2(x)+1)^2} + \exp\left(\frac{1}{x_0}\int_0^{x_0} \log \frac{\sigma^4(x)}{\sigma^2(x)+1} dx - \frac{2d}{x_0}\right) \frac{\sigma^2(x)+2}{\sigma^2(x)(\sigma^2(x)+1)} = \lambda x^{2m}.$$

$\square$

3.7.2 Proof of Theorem 3.4.1

Now we give the details of the proof of Theorem 3.4.1. For the purpose of our analysis, we define two allocations of bits, the monotone allocation and the blockwise constant allocation,

$$\Pi_{\text{blk}}(B) = \left\{ (b_j)_{j=1}^\infty : \sum_{j=1}^\infty b_j \leq B,\ b_j = \bar{b}_k \text{ for } j \in J_k,\ 0 \leq b_j \leq b_{\max} \right\}, \tag{3.18}$$

$$\Pi_{\text{mon}}(B) = \left\{ (b_j)_{j=1}^\infty : \sum_{j=1}^\infty b_j \leq B,\ b_{j-1} \geq b_j,\ 0 \leq b_j \leq b_{\max} \right\}, \tag{3.19}$$

where $b_{\max} = 2\log(1/\varepsilon)$. We also define two classes of weights, the monotonic weights and the blockwise constant weights,

$$\Omega_{\text{blk}} = \left\{(\omega_j)_{j=1}^{\infty} : \ \omega_j = \bar{\omega}_k \text{ for } j \in J_k, \ 0 \le \omega_j \le 1\right\}, \tag{3.20}$$

$$\Omega_{\text{mon}} = \left\{(\omega_j)_{j=1}^{\infty} : \ \omega_{j-1} \ge \omega_j, \ 0 \le \omega_j \le 1\right\}. \tag{3.21}$$

We will also need the following results from Tsybakov (2008) regarding the weakly geometric system of blocks.

Lemma 3.7.3. *Let $\{J_k\}$ be a weakly geometric block system defined by (3.11). Then there exists $0 < \varepsilon_0 < 1$ and $C > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$,*

$$K \le C\log^2(1/\varepsilon),$$
$$\max_{1\le k\le K-1} \frac{T_{k+1}}{T_k} \le 1 + 3\rho_\varepsilon.$$

We divide the proof into four steps.

Step 1. Truncation and replacement

The loss of the quantized estimator $\check{\theta}$ can be decomposed into

$$\|\check{\theta} - \theta\|^2 = \sum_{k=1}^{K} \|\check{\theta}_{(k)} - \theta_{(k)}\|^2 + \sum_{j=N+1}^{\infty} \theta_j^2,$$

where the remainder term satisfies

$$\sum_{j=N+1}^{\infty} \theta_j^2 \le N^{-2m} \sum_{j=N+1}^{\infty} a_j^2\theta_j^2 = O(N^{-2m}).$$

If we assume that $m > 1/2$, which corresponds to classes of continuous functions, the remainder term is then $o(\varepsilon^2)$. If $m \le 1/2$, the remainder term is on the order of $O(\varepsilon^{4m})$, which is still negligible compared to the order of the lower bound $\varepsilon^{\frac{4m}{2m+1}}$. To ease the notation, we

will assume that $m > 1/2$, and write the remainder term as $o(\varepsilon^2)$, but need to bear in mind that the proof works for all $m > 0$. We can thus discard the remainder term in our analysis. Recall that the quantized estimate for each block is given by

$$\check{\theta}_{(k)} = \frac{\check{S}_k^2 - T_k\varepsilon^2}{\check{S}_k}\sqrt{1 - 2^{-2\tilde{b}_k}}\check{Z}_{(k)},$$

and consider the following estimate with $\check{S}_k$ replaced by S_k

$$\widehat{\theta}_{(k)} = \frac{S_k^2 - T_k\varepsilon^2}{S_k}\sqrt{1 - 2^{-2\tilde{b}_k}}\check{Z}_{(k)}.$$

Notice that

$$\begin{aligned}
\|\widehat{\theta}_{(k)} - \check{\theta}_{(k)}\| &= \left|\frac{\check{S}_k^2 - T_k\varepsilon^2}{\check{S}_k} - \frac{S_k^2 - T_k\varepsilon^2}{S_k}\right|\sqrt{1 - 2^{-2\tilde{b}_k}}\|\check{Z}_{(k)}\| \\
&\leq \left|\frac{\check{S}_k S_k + T_k\varepsilon^2}{\check{S}_k S_k}\right|\left|\check{S}_k - S_k\right| \\
&\leq 2\varepsilon^2
\end{aligned}$$

where the last inequality is because $\check{S}_k S_k \geq T_k\varepsilon^2$ and $\left|\check{S}_k - S_k\right| \leq \varepsilon^2$. Thus we can safely replace $\check{\theta}_{(k)}$ by $\widehat{\theta}_{(k)}$ because

$$\begin{aligned}
&\|\check{\theta}_{(k)} - \theta_{(k)}\|^2 \\
&\quad = \|\check{\theta}_{(k)} - \widehat{\theta}_{(k)} + \widehat{\theta}_{(k)} - \theta_{(k)}\|^2 \\
&\quad \leq \|\check{\theta}_{(k)} - \widehat{\theta}_{(k)}\|^2 + \|\widehat{\theta}_{(k)} - \theta_{(k)}\|^2 + 2\|\check{\theta}_{(k)} - \widehat{\theta}_{(k)}\|\|\widehat{\theta}_{(k)} - \theta_{(k)}\| \\
&\quad = \|\widehat{\theta}_{(k)} - \theta_{(k)}\|^2 + O(\varepsilon^2).
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\|\check{\theta} - \theta\|^2 = \mathbb{E}\sum_{k=1}^{K}\|\widehat{\theta}_{(k)} - \theta_{(k)}\|^2 + O(K\varepsilon^2).$$

Step 2. Expectation over codebooks

Now conditioning on the data Y, we work under the probability measure introduced by the random codebook. Write

$$\lambda_k = \frac{S_k^2 - T_k\varepsilon^2}{S_k^2} \text{ and } Z_{(k)} = \frac{Y_{(k)}}{\|Y_{(k)}\|}.$$

We decompose and examine the following term

$$\begin{aligned} A_k &= \|\widehat{\theta}_{(k)} - \theta_{(k)}\|^2 \\ &= \|\widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)} + \lambda_k S_k Z_{(k)} - \theta_{(k)}\|^2 \\ &= \underbrace{\|\widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}\|^2}_{A_{k,1}} + \underbrace{\|\lambda_k S_k Z_{(k)} - \theta_{(k)}\|^2}_{A_{k,2}} \\ &\qquad + \underbrace{2\langle \widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}, \lambda_k S_k Z_{(k)} - \theta_{(k)} \rangle}_{A_{k,3}}. \end{aligned}$$

To bound the expectation of the first term $A_{k,1}$, we need the following lemma, which bounds the probability of the distortion of a codeword exceeding the desired value.

Lemma 3.7.4. *Suppose that $Z_1, \ldots, Z_n$ are independent and each follows the uniform distribution on the t-dimensional unit sphere $\mathbb{S}^{t-1}$. Let $y \in \mathbb{S}^{t-1}$ be a fixed vector, and*

$$Z^* = \arg\min_{z \in Z_{1:n}} \left\| \sqrt{1 - 2^{-2b}} z - y \right\|^2.$$

If $n = 2^{qt}$, then

$$\mathbb{E} \left\| \sqrt{1 - 2^{-2q}} Z^* - y \right\|^2 \leq 2^{-2q} (1 + \nu) + 2e^{-2t}$$

where

$$\nu = \frac{3 \log t + 5}{t - 3 \log t - 6}.$$

Observe that

$$\begin{aligned} A_{k,1} &= \left\| \widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)} \right\|^2 \\ &= \left\| \lambda_k S_k \sqrt{1 - 2^{-2\tilde{b}_k}} \check{Z}_{(k)} - \lambda_k S_k Z_{(k)} \right\|^2 \\ &= \lambda_k^2 S_k^2 \left\| \sqrt{1 - 2^{-2\tilde{b}_k}} \check{Z}_{(k)} - Z_{(k)} \right\|^2 . \end{aligned}$$

Then, it follows as a result of Lemma 3.7.4 that

$$\begin{aligned} \mathbb{E}\left(A_{k,1} \,|\, Y_{(k)}\right) &\leq \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} \left(2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + 2e^{-2T_k}\right) \\ &\leq \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} \left(2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + 2e^{-2T_1}\right) \\ &\leq \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + \frac{2c^2}{(j_k\pi)^{2m}}\varepsilon^2, \end{aligned}$$

where $\nu_\varepsilon = \frac{3\log T_1 - 5}{T_1 - 3\log T_1 - 6}$. Since $A_{k,2}$ only depends on $Y_{(k)}$, $\mathbb{E}\left(A_{k,2} \,|\, Y_{(k)}\right) = A_{k,2}$. Next we consider the cross term $A_{k,3}$. Write $\gamma_k = \frac{\langle \theta_{(k)}, Y_{(k)} \rangle}{\|Y_{(k)}\|^2}$ and

$$\begin{aligned} A_{k,3} &= 2\left\langle \widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}, \lambda_k S_k Z_{(k)} - \theta_{(k)} \right\rangle \\ &= \underbrace{2\left\langle \widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}, \gamma_k Y_{(k)} - \theta_{(k)} \right\rangle}_{A_{k,3a}} \\ &\qquad + \underbrace{2\left\langle \widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}, \lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)} \right\rangle}_{A_{k,3b}}. \end{aligned}$$

The quantity γ_k is chosen such that $\langle Y_{(k)}, \gamma_k Y_{(k)} - \theta_{(k)} \rangle = 0$ and therefore

$$\begin{aligned} A_{k,3a} &= 2\left\langle \widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}, \gamma_k Y_{(k)} - \theta_{(k)} \right\rangle \\ &= 2\left\langle \Pi_{Y_{(k)}^\perp}(\widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}), \gamma_k Y_{(k)} - \theta_{(k)} \right\rangle \end{aligned}$$

where $\Pi_{Y_{(k)}^{\perp}}$ denotes the projection onto the orthogonal complement of $Y_{(k)}$. Due to the choice of $\check{Z}_{(k)}$, the projection $\Pi_{Y_{(k)}^{\perp}}(\widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)})$ is rotation symmetric and hence $\mathbb{E}\left(A_{k,3a} \,|\, Y_{(k)}\right) = 0$. Finally, for $A_{k,3b}$ we have

$$\begin{aligned}
&\mathbb{E}\left(A_{k,3b} \,|\, Y_{(k)}\right) \\
&\quad \le 2\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \mathbb{E}\left(\|\widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}\| \,|\, Y_{(k)}\right) \\
&\quad \le 2\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \sqrt{\mathbb{E}\left(\|\widehat{\theta}_{(k)} - \lambda_k S_k Z_{(k)}\|^2 \,|\, Y_{(k)}\right)} \\
&\quad \le 2\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \sqrt{\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + \frac{2c^2}{(j_k\pi)^{2m}}\varepsilon^2}.
\end{aligned}$$

Combining all the analyses above, we have

$$\begin{aligned}
&\mathbb{E}\left(A_k \,|\, Y_{(k)}\right) \\
&\le \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + \frac{2c^2}{(j_k\pi)^{2m}}\varepsilon^2 + \|\lambda_k S_k Z_{(k)} - \theta_{(k)}\|^2 \\
&\quad + 2\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \sqrt{\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + \frac{2c^2}{(j_k\pi)^{2m}}\varepsilon^2},
\end{aligned}$$

and summing over k we get

$$\begin{aligned}
&\mathbb{E}\left(\|\check{\theta} - \theta\|^2 \,|\, Y\right) \\
&\le \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + \sum_{k=1}^{K} \|\lambda_k S_k Z_{(k)} - \theta_{(k)}\|^2 \\
&\quad + 2\sum_{k=1}^{K} \|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \sqrt{\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + O(\varepsilon^2)} + O(K\varepsilon^2).
\end{aligned} \tag{3.22}$$

Step 3. Expectation over data

First we will state three lemmas, which bound the deviation of the expectation of some particular functions of the norm of a Gaussian vector to the desired quantities. The proofs are given in Section 3.7.2.

Lemma 3.7.5. *Suppose that $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$ independently for $i = 1, \ldots, n$, where $\|\theta\|^2 \leq c^2$. Let S be given by*

$$S = \begin{cases} \sqrt{n\sigma^2} & \text{if } \|X\| < \sqrt{n\sigma^2} \\ \sqrt{n\sigma^2} + c & \text{if } \|X\| > \sqrt{n\sigma^2} + c \\ \|X\| & \text{otherwise.} \end{cases}$$

Then there exists some absolute constant C_0 such that

$$\mathbb{E}\left(\frac{S^2 - n\sigma^2}{S} - \frac{\langle \theta, X \rangle}{\|X\|}\right)^2 \leq C_0 \sigma^2.$$

Lemma 3.7.6. *Let X and S be the same as defined in Lemma 3.7.5. Then for $n > 4$*

$$\mathbb{E}\frac{(S^2 - n\sigma^2)^2}{S^2} \leq \frac{\|\theta\|^4}{\|\theta\|^2 + n\sigma^2} + \frac{4n}{n-4}\sigma^2.$$

Lemma 3.7.7. *Let X and S be the same as defined in Lemma 3.7.5. Define*

$$\widehat{\theta}_+ = \left(\frac{\|X\|^2 - n\sigma^2}{\|X\|^2}\right)_+ X, \quad \widehat{\theta}_\dagger = \frac{S^2 - n\sigma^2}{S\|X\|} X.$$

Then

$$\mathbb{E}\|\widehat{\theta}_\dagger - \theta\|^2 \leq \mathbb{E}\|\widehat{\theta}_+ - \theta\|^2 \leq \frac{n\sigma^2\|\theta\|^2}{\|\theta\|^2 + n\sigma^2} + 4\sigma^2.$$

We now take the expectation with respect to the data on both sides of (3.22). First, by the Cauchy-Schwarz inequality

$$\begin{aligned} &\mathbb{E}\left(\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \sqrt{\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1 + \nu_\varepsilon) + O(\varepsilon^2)}\right) \\ &\leq \sqrt{\mathbb{E}\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\|^2} \sqrt{\mathbb{E}\left(\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1 + \nu_\varepsilon) + O(\varepsilon^2)\right)}. \end{aligned} \tag{3.23}$$

We then calculate

$$\begin{aligned}
&\sqrt{\mathbb{E}\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\|^2} \\
&= \sqrt{\mathbb{E}\left\| \frac{S_k^2 - T_k\varepsilon^2}{S_k} \frac{Y_{(k)}}{\|Y_{(k)}\|} - \frac{\langle \theta_{(k)}, Y_{(k)} \rangle}{\|Y_{(k)}\|} \frac{Y_{(k)}}{\|Y_{(k)}\|} \right\|^2} \\
&= \sqrt{\mathbb{E}\left(\frac{S_k^2 - T_k\varepsilon^2}{S_k} - \frac{\langle \theta_{(k)}, Y_{(k)} \rangle}{\|Y_{(k)}\|} \right)^2} \\
&\leq C_0 \varepsilon,
\end{aligned}$$

where the last inequality is due to Lemma 3.7.5, and C_0 is the constant therein. Plugging this in (3.23) and summing over k, we get

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\left(\|\lambda_k S_k Z_{(k)} - \gamma_k Y_{(k)}\| \sqrt{\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + O(\varepsilon^2)} \right) \\
&\qquad \leq C_0 \varepsilon \sum_{k=1}^{K} \sqrt{\mathbb{E}\left(\frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + O(\varepsilon^2) \right)} \\
&\qquad \leq C_0 \sqrt{K} \varepsilon \sqrt{\mathbb{E} \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + O(K\varepsilon^2)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\|\check{\theta} - \theta\|^2 \\
&\leq \underbrace{\mathbb{E} \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}}_{B_1}(1+\nu_\varepsilon) + \underbrace{\mathbb{E} \sum_{k=1}^{K} \|\lambda_k S_k Z_{(k)} - \theta_{(k)}\|^2}_{B_2} \\
&\quad + C_0 \sqrt{K} \varepsilon \sqrt{\mathbb{E} \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k}(1+\nu_\varepsilon) + O(K\varepsilon^2)} \\
&\quad + O(K\varepsilon^2).
\end{aligned}$$

Now we deal with the term B_1. Recall that the sequence $\tilde{b}$ solves problem (3.12), so for any

sequence $b \in \Pi_{\text{blk}}$

$$\sum_{k=1}^{K} \frac{(\check{S}_k^2 - T_k\varepsilon^2)^2}{\check{S}_k^2} 2^{-2\tilde{b}_k} \leq \sum_{k=1}^{K} \frac{(\check{S}_k^2 - T_k\varepsilon^2)^2}{\check{S}_k^2} 2^{-2\bar{b}_k}.$$

Notice that

$$\left| \frac{(\check{S}_k^2 - T_k\varepsilon^2)^2}{\check{S}_k^2} - \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} \right| = \left| \check{S}_k^2 - S_k^2 \right| \left| \frac{\check{S}_k^2 S_k^2 - T_k\varepsilon^2}{\check{S}_k^2 S_k^2} \right| = O(\varepsilon^2)$$

and thus,

$$\sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k} \leq \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\bar{b}_k} + O(K\varepsilon^2).$$

Taking the expectation, we get

$$\mathbb{E} \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k} \leq \sum_{k=1}^{K} \mathbb{E} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\bar{b}_k} + O(K\varepsilon^2).$$

Applying Lemma 3.7.6, we get for $T_k > 4$

$$\mathbb{E} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} \leq \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} + \frac{4T_k}{T_k - 4}\varepsilon^2$$

and it follows that

$$\mathbb{E} \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k} \leq \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} 2^{-2\bar{b}_k} + O(K\varepsilon^2).$$

Since $b \in \Pi_{\text{blk}}$ is arbitrary,

$$\mathbb{E} \sum_{k=1}^{K} \frac{(S_k^2 - T_k\varepsilon^2)^2}{S_k^2} 2^{-2\tilde{b}_k} \leq \min_{b \in \Pi_{\text{blk}}} \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} 2^{-2\bar{b}_k} + O(K\varepsilon^2).$$

Turning to the term B_2, as a result of Lemma 3.7.7 we have

$$\|\lambda_k S_k Z_{(k)} - \theta_{(k)}\|^2 \leq \frac{\|\theta_{(k)}\|^2 T_k \varepsilon^2}{\|\theta_{(k)}\|^2 + T_k \varepsilon^2} + 4\varepsilon^2.$$

Combining the above results, we have shown that

$$\mathbb{E}\|\check{\theta} - \theta\|^2 \leq M + O(K\varepsilon^2) + C_0\sqrt{K}\varepsilon\sqrt{M + O(K\varepsilon^2)} \tag{3.24}$$

where

$$\begin{aligned} M &= (1+\nu_\varepsilon) \min_{b \in \Pi_{\text{blk}}(B)} \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} 2^{-2\bar{b}_k} + \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^2 T_k \varepsilon^2}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} \\ &= (1+\nu_\varepsilon) \min_{b \in \Pi_{\text{blk}}(B)} \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} 2^{-2\bar{b}_k} \\ &\quad + \min_{\omega \in \Omega_{\text{blk}}} \sum_{k=1}^{K} \left((1-\bar{\omega}_k)^2 \|\theta_{(k)}\|^2 + \bar{\omega}_k^2 T_k \varepsilon^2 \right). \end{aligned}$$

Step 4. Blockwise constant is almost optimal

We now show that in terms of both bit allocation and weight assignment, blockwise constant is almost optimal. Let's first consider bit allocation. Let $B' = \frac{1}{1+3\rho_\varepsilon}(B - T_1 b_{\max})$. We are going to show that

$$\min_{b \in \Pi_{\text{blk}}(B)} \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k\varepsilon^2} 2^{-2\bar{b}_k} \leq \min_{b \in \Pi_{\text{mon}}(B')} \sum_{j=1}^{N} \frac{\theta_j^4}{\theta_j^2 + \varepsilon^2} 2^{-2b_j}.$$

In fact, suppose that $b^* \in \Pi_{\text{mon}}(B')$ achieves the minimum on the right hand side, and define $b^\star$ by

$$b_j^\star = \begin{cases} \max_{i \in B_k} b_i^* & j \in B_k \\ 0 & j \geq N \end{cases}.$$

The sum of the elements in $b^\star$ then satisfies

$$\begin{aligned}
\sum_{j=1}^{\infty} b_j^\star &= \sum_{k=0}^{K-1} T_{k+1} \max_{j \in B_{k+1}} b_j^* \\
&= T_1 b_1^\star + \sum_{k=1}^{K-1} T_{k+1} \max_{j \in B_{k+1}} b_j^* \\
&\leq T_1 b_{\max} + \sum_{k=1}^{K-1} \frac{T_{k+1}}{T_k} \sum_{j \in B_k} b_j^* \\
&\leq T_1 b_{\max} + (1 + 3\rho_\varepsilon) \sum_{k=1}^{K-1} \sum_{j \in B_k} b_j^* \\
&\leq T_1 b_{\max} + (1 + 3\rho_\varepsilon) B' \\
&= B,
\end{aligned}$$

which means that $b^\star \in \Pi_{\text{blk}}(B)$. It then follows that

$$\begin{aligned}
&\min_{b \in \Pi_{\text{blk}}(B)} \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k \varepsilon^2} 2^{-2\bar{b}_k} \\
&\leq \sum_{k=1}^{K} \frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2 + T_k \varepsilon^2} 2^{-2\bar{b}_k^\star} \\
&\leq \sum_{k=1}^{K} \sum_{j \in B_k} \frac{\theta_j^4}{\theta_j^2 + \varepsilon^2} 2^{-2b_j^\star} \qquad (3.25) \\
&\leq \sum_{j=1}^{N} \frac{\theta_j^4}{\theta_j^2 + \varepsilon^2} 2^{-2b_j^*} \\
&= \min_{b \in \Pi_{\text{mon}}(B')} \sum_{j=1}^{N} \frac{\theta_j^4}{\theta_j^2 + \varepsilon^2} 2^{-2b_j},
\end{aligned}$$

where (3.25) is due to Jensen's inequality on the convex function $\frac{x^2}{x+\varepsilon^2}$

$$\frac{\left(\frac{1}{T_k}\|\theta_{(k)}\|^2\right)^2}{\frac{1}{T_k}\|\theta_{(k)}\|^2 + \varepsilon^2} \leq \frac{1}{T_k} \sum_{j \in B_k} \frac{\theta_j^4}{\theta_j^2 + \varepsilon^2}.$$

Next, for the weights assignment, by Lemma 3.11 in Tsybakov (2008), we have

$$\begin{aligned}\min_{\omega\in\Omega_{\text{blk}}} \sum_{k=1}^{K} \left((1-\bar{\omega}_k)^2\|\theta_{(k)}\|^2 + \bar{\omega}_k^2 T_k\varepsilon^2\right) \\ \leq (1+3\rho_\varepsilon)\left(\min_{\omega\in\Omega_{\text{mon}}} \sum_{k=1}^{K}\left((1-\omega_j)^2\theta_j^2 + \omega_j^2\varepsilon^2\right)\right) + T_1\varepsilon^2.\end{aligned} \tag{3.26}$$

Combining (3.7.2) and (3.26), we get

$$\begin{aligned}M &= (1+\nu_\varepsilon)\min_{b\in\Pi_{\text{blk}}(B)}\sum_{k=1}^{K}\frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2+T_k\varepsilon^2}2^{-2\bar{b}_k} \\ &\quad + \min_{\omega\in\Omega_{\text{blk}}}\sum_{k=1}^{K}\left((1-\bar{\omega}_k)^2\|\theta_{(k)}\|^2+\bar{\omega}_k^2T_k\varepsilon^2\right) \\ &\leq (1+\nu_\varepsilon)\min_{b\in\Pi_{\text{blk}}(B)}\sum_{k=1}^{K}\frac{\|\theta_{(k)}\|^4}{\|\theta_{(k)}\|^2+T_k\varepsilon^2}2^{-2\bar{b}_k} \\ &\quad + (1+3\rho_\varepsilon)\min_{\omega\in\Omega_{\text{mon}}}\sum_{k=1}^{K}\left((1-\bar{\omega}_k)^2\|\theta_{(k)}\|^2+\bar{\omega}_k^2T_k\varepsilon^2\right) + T_1\varepsilon^2 \\ &\leq (1+\nu_\varepsilon)\Bigg(\min_{b\in\Pi_{\text{mon}}(B')}\sum_{j=1}^{N}\frac{\theta_j^4}{\theta_j^2+\varepsilon^2}2^{-2b_j} \\ &\quad + \min_{\omega\in\Omega_{\text{mon}}}\sum_{j=1}^{N}\left((1-\omega_j)^2\theta_j^2+\omega_j^2\varepsilon^2\right)\Bigg) + T_1\varepsilon^2.\end{aligned}$$

Then by Lemma 3.7.8,

$$M \leq (1+\nu_\varepsilon)V_\varepsilon(m,c,B') + T_1\varepsilon^2.$$

which, plugged into (3.24), gives us

$$\begin{aligned}\mathbb{E}\|\check{\theta}-\theta\|^2 &\leq (1+\nu_\varepsilon)V_\varepsilon(m,c,B') + O(K\varepsilon^2) \\ &\quad + C_0\sqrt{K}\varepsilon\sqrt{(1+\nu_\varepsilon)V_\varepsilon(m,c,B') + O(K\varepsilon^2)}.\end{aligned}$$

Recall that

$$\nu_\varepsilon = O\left(\frac{\log\log(1/\varepsilon)}{\log(1/\varepsilon)}\right), \quad K = O(\log^2(1/\varepsilon)),$$

and that

$$\lim_{\varepsilon\to 0}\frac{B'}{B}=\lim_{\varepsilon\to 0}\frac{1}{1+3\rho_\varepsilon}\left(1-\frac{T_1 b_{\max}}{B}\right)=1.$$

Thus,

$$\lim_{\varepsilon\to 0}\frac{V_\varepsilon(m,c,B')}{V_\varepsilon(m,c,B)}=1.$$

Also notice that no matter how B grows as $\varepsilon\to 0$, $V_\varepsilon(m,c,B)=O(\varepsilon^{\frac{4m}{2m+1}})$. Therefore,

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{\mathbb{E}\|\check{\theta}-\theta\|^2}{V_\varepsilon(B,m,c)}\\
&\quad\le \lim_{\varepsilon\to 0}\Bigg((1+\nu_\varepsilon)\frac{V_\varepsilon(B',m,c)}{V_\varepsilon(B,m,c)}+\frac{O(K\varepsilon^2)}{V(B,m,c)}\\
&\qquad + C_0\sqrt{(1+\nu_\varepsilon)\frac{K\varepsilon^2}{V_\varepsilon(B,m,c)}\frac{V_\varepsilon(B',m,c)}{V_\varepsilon(B,m,c)}+\left(\frac{O(K\varepsilon^2)}{V_\varepsilon(B,m,c)}\right)^2}\Bigg)\\
&\quad = 1
\end{aligned}$$

which concludes the proof.

Lemma 3.7.8. *Let V_1 be the value of the optimization*

$$\begin{aligned}
&\max_{\theta}\min_{b}\quad \sum_{j=1}^{N}\left(\frac{\theta_j^4}{\theta_j^2+\varepsilon^2}2^{-2b_j}+\frac{\theta_j^2\varepsilon^2}{\theta_j^2+\varepsilon^2}\right)\\
&\textit{such that}\quad \sum_{j=1}^{N}b_j\le B,\ b_j\ge 0,\ \sum_{j=1}^{J}a_j^2\theta_j^2\le\frac{c^2}{\pi^{2m}},
\end{aligned}\tag{A_1}$$

and let V_2 be the value of the optimization

$$\begin{aligned}
&\max_{\theta}\min_{b,\omega}\quad \sum_{j=1}^{N}\left(\frac{\theta_j^4}{\theta_j^2+\varepsilon^2}2^{-2b_j}+(1-\omega_j)^2\theta_j^2+\omega_j^2\varepsilon^2\right)\\
&\textit{such that}\quad \sum_{j=1}^{N}b_j\le B,\ b_{j-1}\ge b_j,\ 0\le b_j\le b_{\max},\ \omega_{j-1}\ge\omega_j,\\
&\qquad\qquad \sum_{j=1}^{J}a_j^2\theta_j^2\le\frac{c^2}{\pi^{2m}}.
\end{aligned}\tag{A_2}$$

Then $V_1 = V_2$.

Proof of Lemmas

Proof of Lemma 3.7.4. Let $\zeta(t)$ be a positive function of t to be specified later. Let

$$p_0 = \mathbb{P}\left(\left\|\sqrt{1-2^{-2q}}Z_1 - y\right\| \leq 2^{-q}(1+\zeta(t)^{-1})\right).$$

Using a result from Sakrison (1968), p_0 can be bounded by

$$p_0 \geq \frac{\Gamma(\frac{t}{2}+1)}{\pi t \Gamma(\frac{t+1}{2})} 2^{-q(t-1)}(1+\zeta(t)^{-1})^{t-1}.$$

We obtain that

$$\begin{aligned}
&\mathbb{E}\left\|\sqrt{1-2^{-2b}}Z^* - y\right\|^2 \\
&\quad \leq 2^{-2q}(1+\zeta(t)^{-1})^2 + 2\mathbb{P}\left(\left\|\sqrt{1-2^{-2b}}Z_* - y\right\| > 2^{-q}(1+\zeta(t)^{-1})\right) \\
&\quad \leq 2^{-2q}(1+2\zeta(t)^{-1}) + 2(1-p_0)^n.
\end{aligned}$$

To upper bound $(1-p_0)^n$, we consider

$$\begin{aligned}
\log\left((1-p_0)^n\right) = n\log(1-p_0) &\leq -np_0 \\
&\leq -2^{qt}\frac{\Gamma(\frac{t}{2}+1)}{\pi t\Gamma(\frac{t+1}{2})}2^{-q(t-1)}(1+\zeta(t)^{-1})^{t-1} \\
&\leq -2^{q}\frac{\Gamma(\frac{t}{2}+1)}{\pi t\Gamma(\frac{t+1}{2})}(1+\zeta(t)^{-1})^{(\zeta(t)+1)\frac{t-1}{\zeta(t)+1}} \\
&\leq -\frac{\sqrt{2\pi}(\frac{t}{2})^{\frac{t}{2}+\frac{1}{2}}e^{-\frac{t}{2}}}{\pi t e(\frac{t}{2}-\frac{1}{2})^{\frac{t}{2}}e^{-(\frac{t}{2}-\frac{1}{2})}}e^{\frac{t-1}{\zeta(t)+1}} \\
&= -\frac{1}{\sqrt{\pi e^3}}t^{-\frac{1}{2}}\left(\frac{t}{t-1}\right)^{\frac{t}{2}}e^{\frac{t-1}{\zeta(t)+1}} \\
&\leq -\frac{1}{\sqrt{\pi}e}t^{-\frac{1}{2}}e^{\frac{t-1}{\zeta(t)+1}}
\end{aligned}$$

where we have used the Stirling's approximation

$$\sqrt{2\pi}z^{z+1/2}e^{-z} \leq \Gamma(z+1) \leq ez^{z+1/2}e^{-z}.$$

In order for $(1-p_0)^n \leq e^{-2t}$ to hold, we need

$$-2t = -\frac{1}{\sqrt{\pi}e}t^{-\frac{1}{2}}e^{\frac{t-1}{\zeta(t)+1}},$$

which leads to the choice of $\zeta(t)$

$$\zeta(t) = \frac{t-1}{\log(2\sqrt{\pi}et^{\frac{3}{2}})} - 1.$$

Observing that

$$2\zeta(t)^{-1} \leq \frac{3\log t + 5}{t - 3\log t - 6}$$

completes the proof. □

Proof of Lemma 3.7.5. We first claim that

$$\mathbb{E}\left(\frac{S^2-n\sigma^2}{S} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2 \leq \mathbb{E}\left(\frac{\|X\|^2-n\sigma^2}{\|X\|} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2.$$

In fact, writing $\mathbb{E}_r(\cdot)$ for the conditional expectation $\mathbb{E}(\cdot \mid \|X\| = r)$, it suffices to show that for $r < \sqrt{n\sigma^2}$ and $r > \sqrt{n\sigma^2} + c$

$$\mathbb{E}_r\left(\frac{S^2-n\sigma^2}{S} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2 \leq \mathbb{E}_r\left(\frac{\|X\|^2-n\sigma^2}{\|X\|} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2.$$

When $r < \sqrt{n\sigma^2}$, it is equivalent to

$$\mathbb{E}_r\left(\frac{\langle\theta, X\rangle}{\|X\|}\right)^2 \leq \mathbb{E}_r\left(\frac{\langle\theta, X\rangle}{\|X\|} - \frac{\|X\|^2-n\sigma^2}{\|X\|}\right)^2$$

It is then sufficient to show that $\mathbb{E}_r\langle\theta, X\rangle \geq 0$. This can be obtained by following a similar argument as in Lemma A.6 in Tsybakov (2008). When $r > \sqrt{n\sigma^2} + c$, we need to show that

$$\mathbb{E}_r\left(\frac{(\sqrt{n\sigma^2}+c)^2 - n\sigma^2}{\sqrt{n\sigma^2}+c} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2 \leq \mathbb{E}_r\left(\frac{\|X\|^2 - n\sigma^2}{\|X\|} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2,$$

which, after some algebra, boils down to

$$\frac{(\sqrt{n\sigma^2}+c)^2 - n\sigma^2}{\sqrt{n\sigma^2}+c} + \frac{r^2 - n\sigma^2}{r} \geq \frac{2}{r}\mathbb{E}_r\langle\theta, X\rangle.$$

This holds because

$$\begin{aligned}
&r\left(\frac{(\sqrt{n\sigma^2}+c)^2 - n\sigma^2}{\sqrt{n\sigma^2}+c} + \frac{r^2 - n\sigma^2}{r} - \frac{2}{r}\mathbb{E}_r\langle\theta, X\rangle\right) \\
&\geq \|\theta\|^2 + r^2 - n\sigma^2 - 2\mathbb{E}_r\langle\theta, X\rangle \\
&\geq \mathbb{E}_r\|X - \theta\|^2 - n\sigma^2 \\
&\geq 0
\end{aligned}$$

where we have used the assumption that $r > \sqrt{n\sigma^2} + c$, $\|\theta\| \leq c$ and that

$$\mathbb{E}_r\|X - \theta\| \geq \mathbb{E}_r\|X\| - \|\theta\| \geq \sqrt{n\sigma^2}.$$

Now that we have shown (3.7.2) and noting that

$$\mathbb{E}\left(\frac{\|X\|^2 - n\sigma^2}{\|X\|} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2 = \sigma^2\mathbb{E}\left(\frac{\|X/\sigma\|^2 - n}{\|X/\sigma\|} - \frac{\langle\theta/\sigma, X/\sigma\rangle}{\|X/\sigma\|}\right)^2,$$

we can assume that $X \sim N(\theta, I_n)$ and equivalently show that there exists a universal constant C_0 such that

$$\mathbb{E}\left(\frac{\|X\|^2 - n}{\|X\|} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2 \leq C_0$$

holds for any n and θ. Letting $Z = X - \theta$ and writing $\|\theta\|^2 = \xi$, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{\|X\|^2 - n}{\|X\|} - \frac{\langle\theta, X\rangle}{\|X\|}\right)^2 \\
&= \mathbb{E}\left(\frac{\|Z+\theta\|^2 - n - \xi}{\|Z+\theta\|} - \frac{\langle\theta, Z\rangle}{\|Z+\theta\|}\right)^2 \\
&\le 2\mathbb{E}\left(\frac{\|Z+\theta\|^2 - n - \xi}{\|Z+\theta\|}\right)^2 + 2\mathbb{E}\left(\frac{\langle\theta, Z\rangle}{\|Z+\theta\|}\right)^2 \\
&\le 2\mathbb{E}\|Z+\theta\|^2 - 4(n+\xi) + 2\mathbb{E}\frac{(n+\xi)^2}{\|Z+\theta\|^2} + 2\mathbb{E}\left(\frac{\langle\theta, Z\rangle}{\|Z+\theta\|}\right)^2 \\
&\le 2(n+\xi) - 4(n+\xi) + 2\frac{(n+\xi)^2}{n+\xi-4} + 2\mathbb{E}\left(\frac{\langle\theta, Z\rangle}{\|Z+\theta\|}\right)^2 \\
&= \frac{8(n+\xi)}{n+\xi-4} + 2\mathbb{E}\left(\frac{\langle\theta, Z\rangle}{\|Z+\theta\|}\right)^2.
\end{aligned}$$

where the last inequality is due to Lemma 3.7.9. To bound the last term, we apply the Cauchy-Schwarz inequality and get

$$\begin{aligned}
\mathbb{E}\left(\frac{\langle\theta, Z\rangle}{\|Z+\theta\|}\right)^2 &\le \sqrt{\mathbb{E}\frac{1}{\|Z+\theta\|^4}\mathbb{E}\langle\theta, Z\rangle^4} \\
&\le \sqrt{\frac{3(n-4)\xi^2}{(n-6)(n+\xi-4)(n+\xi-6)}}
\end{aligned}$$

where the last inequality is again due to Lemma 3.7.9. Thus we just need to take C_0 to be

$$\sup_{n\ge 7, \xi \ge 0} \frac{8(n+\xi)}{n+\xi-4} + 2\sqrt{\frac{3(n-4)\xi^2}{(n-6)(n+\xi-4)(n+\xi-6)}},$$

which is apparently a finite quantity. □

Proof of Lemma 3.7.6. Since the function $(x^2 - n\sigma^2)^2/x^2$ is decreasing on $(0, \sqrt{n\sigma^2})$ and increasing on $(\sqrt{n\sigma^2}, \infty)$, we have

$$\frac{(S^2 - n\sigma^2)^2}{S^2} \le \frac{(\|X\|^2 - n\sigma^2)^2}{\|X\|^2},$$

and it follows that if $n > 4$

$$\mathbb{E}\frac{(S^2 - n\sigma^2)^2}{S^2} \leq \mathbb{E}\frac{(\|X\|^2 - n\sigma^2)^2}{\|X\|^2} \tag{3.27}$$

$$= \mathbb{E}\|X\|^2 - 2n\sigma^2 + n^2\sigma^4\mathbb{E}\left(\frac{1}{\|X\|^2}\right) \tag{3.28}$$

$$\leq \|\theta\|^2 - n\sigma^2 + \frac{n^2\sigma^4}{\|\theta\|^2 + n\sigma^2 - 4\sigma^2} \tag{3.29}$$

$$\leq \frac{\|\theta\|^4}{\|\theta\|^2 + n\sigma^2} + \frac{4n}{n-4}\sigma^2 \tag{3.30}$$

where (3.29) is due to Lemma 3.7.9, and (3.30) is obtained by

$$\begin{aligned}
&\|\theta\|^2 - n\sigma^2 + \frac{n^2\sigma^4}{\|\theta\|^2 + n\sigma^2 - 4\sigma^2} - \frac{\|\theta\|^4}{\|\theta\|^2 + n\sigma^2} \\
&= \frac{\|\theta\|^4 + 4\sigma^2(n\sigma^2 - \|\theta\|^2)}{\|\theta\|^2 + n\sigma^2 - 4\sigma^2} - \frac{\|\theta\|^4}{\|\theta\|^2 + n\sigma^2} \\
&= \frac{4n^2\sigma^6}{(\|\theta\|^2 + n\sigma^2 - 4\sigma^2)(\|\theta\|^2 + n\sigma^2)} \\
&\leq \frac{4n}{n-4}\sigma^2.
\end{aligned}$$

□

Proof of Lemma 3.7.7. First, the second inequality

$$\mathbb{E}\|\widehat{\theta}_+ - \theta\|^2 \leq \frac{n\sigma^2\|\theta\|^2}{\|\theta\|^2 + n\sigma^2} + 4\sigma^2$$

is given by Lemma 3.10 from Tsybakov (2008). We thus focus on the first inequality. For convenience we write

$$g_+(x) = \left(\frac{\|x\|^2 - n\sigma^2}{\|x\|^2}\right)_+, \quad g_\dagger(x) = \frac{s(x)^2 - n\sigma^2}{s(x)\|x\|}$$

with

$$s(x) = \begin{cases} \sqrt{n\sigma^2} & \text{if } \|x\| < \sqrt{n\sigma^2} \\ \sqrt{n\sigma^2} + c & \text{if } \|x\| > \sqrt{n\sigma^2} + c \\ \|x\| & \text{otherwise} \end{cases}.$$

Notice that $g_+(x) = g_\dagger(x)$ when $\|x\| \le \sqrt{n\sigma^2} + c$ and $g_+(x) > g_\dagger(x)$ when $\|x\| > \sqrt{n\sigma^2} + c$. Since $g_\dagger$ and g_+ both only depend on $\|x\|$, we sometimes will also write $g_\dagger(\|x\|)$ for $g_\dagger(x)$ and $g_+(\|x\|)$ for $g_+(x)$. Setting $\mathbb{E}_r(\cdot)$ to denote the conditional expectation $\mathbb{E}(\cdot \mid \|X\| = r)$ for brevity, it suffices to show that for $r \ge \sqrt{n\sigma^2} + c$

$$\begin{aligned} & \mathbb{E}_r\left(\|g_\dagger(X)X - \theta\|^2\right) \le \mathbb{E}_r\left(\|g_+(X)X - \theta\|^2\right) \\ \iff \quad & g_\dagger(r)^2 r^2 - 2g_\dagger(r)\mathbb{E}_r\langle X, \theta\rangle \le g_+(r)^2 r^2 - 2g_+(r)\mathbb{E}_r\langle X, \theta\rangle \\ \iff \quad & (g_\dagger(r)^2 - g_+(r)^2) r^2 \ge 2(g_\dagger(r) - g_+(r))\mathbb{E}_r\langle X, \theta\rangle \\ \iff \quad & (g_\dagger(r) + g_+(r)) r^2 \ge 2\mathbb{E}_r\langle X, \theta\rangle. \end{aligned} \tag{3.31}$$

On the other hand, we have

$$\begin{aligned} (g_\dagger(r) + g_+(r)) r^2 &\ge \left(\frac{\|\theta\|^2}{r^2} + \frac{r^2 - n\sigma^2}{r^2}\right) r^2 \\ &= \|\theta\|^2 + r^2 - n\sigma^2 \\ &= \|\theta\|^2 + r^2 - 2\mathbb{E}_r\langle X, \theta\rangle - n\sigma^2 + 2\mathbb{E}_r\langle X, \theta\rangle \\ &= \mathbb{E}_r\|X - \theta\|^2 - n\sigma^2 + 2\mathbb{E}_r\langle X, \theta\rangle \\ &\ge 2\mathbb{E}_r\langle X, \theta\rangle \end{aligned}$$

where the last inequality is because

$$\|X - \theta\|^2 \ge (\|X\| - \|\theta\|)^2 \ge n\sigma^2.$$

Thus, (3.31) holds and hence $\mathbb{E}\|\widehat{\theta}_\dagger - \theta\|^2 \leq \mathbb{E}\|\widehat{\theta}_+ - \theta\|^2$. □

Proof of Lemma 3.7.8. It is easy to see that $V_1 \leq V_2$, because for any θ the inside minimum is smaller for (A_1) than for (A_2). Next, we will show $V_1 \geq V_2$.

Suppose that θ^* achieves the value V_2, with corresponding b^* and ω^*. We claim that θ^* is non-increasing. In fact, if θ^* is not non-increasing, then there must exist an index j such that $\theta_j^* < \theta_{j+1}^*$ and for simplicity let's assume that $\theta_1^* < \theta_2^*$. We are going to show that this leads to $b_1^* = b_2^*$ and $\omega_1^* = \omega_2^*$. Write

$$s_1 = \frac{\theta_1^{*4}}{\theta_1^{*2} + \varepsilon^2}, \quad s_2 = \frac{\theta_2^{*4}}{\theta_2^{*2} + \varepsilon^2}.$$

We have $s_1 < s_2$. Let $\bar{b}^* = \frac{b_1^* + b_2^*}{2}$ and observe that $b_1^* \geq \bar{b}^* \geq b_2^*$. Notice that

$$\begin{aligned}
&\left(s_1 2^{-2b_1^*} + s_2 2^{-2b_2^*}\right) - \left(s_1 2^{-2\bar{b}^*} + s_2 2^{-2\bar{b}^*}\right) \\
&= s_1\left(2^{-2b_1^*} - 2^{-2\bar{b}^*}\right) + s_2\left(2^{-2b_2^*} - 2^{-2\bar{b}^*}\right) \\
&\geq s_2\left(2^{-2b_1^*} - 2^{-2\bar{b}^*}\right) + s_2\left(2^{-2b_2^*} - 2^{-2\bar{b}^*}\right) \\
&\geq s_2\left(2^{-2b_1^*} + 2^{-2b_2^*} - 2 \cdot 2^{-2\bar{b}^*}\right) \\
&\geq 0,
\end{aligned}$$

where equality holds if and only if $b_1^* = b_2^*$, since $s_2 > s_1 \geq 0$. Hence, b_1^* and b_2^* have to be equal, or otherwise it would contradict with the assumption that b^* achieves the inside minimum of (A_2). Now turn to ω^*. Write $\bar{\omega}^* = \frac{\omega_1^* + \omega_2^*}{2}$ and note that $\omega_1^* \geq \bar{\omega}^* \geq \omega_2^*$.

Consider

$$
\begin{aligned}
&\left((1-\omega_1^*)^2\theta_1^{*2}+\omega_1^{*2}\varepsilon^2\right)+\left((1-\omega_2^*)^2\theta_2^{*2}+\omega_2^{*2}\varepsilon^2\right)-\left((1-\bar{\omega}^*)^2(\theta_1^{*2}+\theta_2^{*2})+2\bar{\omega}^{*2}\varepsilon^2\right)\\
&=\left((1-\omega_1^*)^2-(1-\bar{\omega}^*)^2\right)\theta_1^{*2}+\left((1-\omega_2^*)^2-(1-\bar{\omega}^*)^2\right)\theta_2^{*2}+\left(\omega_1^{*2}+\omega_2^{*2}-2\bar{\omega}^{*2}\right)\varepsilon^2\\
&\geq\left((1-\omega_1^*)^2-(1-\bar{\omega}^*)^2\right)\theta_2^{*2}+\left((1-\omega_2^*)^2-(1-\bar{\omega}^*)^2\right)\theta_2^{*2}+\left(\omega_1^{*2}+\omega_2^{*2}-2\bar{\omega}^{*2}\right)\varepsilon^2\\
&=\left((1-\omega_1^*)^2+(1-\omega_2^*)^2-2(1-\bar{\omega}^*)^2\right)\theta_2^{*2}+\left(\omega_1^{*2}+\omega_2^{*2}-2\bar{\omega}^{*2}\right)\varepsilon^2\\
&\geq 0,
\end{aligned}
$$

where the equality holds if and only if $\omega_1^*=\omega_2^*$. Therefore, ω_1^* and ω_2^* must be equal. Now, with $b_1^*=b_2^*$ and $\omega_1^*=\omega_2^*$, we can switch θ_1^* and θ_2^* without increasing the objective function and violating the constraints. Thus, our claim that θ^* is non-increasing is justified.

Next, we will show that $b_1^*<b_{\max}$. If $b_1^*=b_{\max}$, then for $j=1,\ldots,N$

$$
\frac{\theta_j^{*4}}{\theta_j^{*2}+\varepsilon^2}2^{-2b_j^*}\leq\frac{\theta_1^{*4}}{\theta_1^{*2}+\varepsilon^2}2^{-2b_1^*}\leq\theta_1^{*2}2^{-2b_{\max}}\leq c^2 2^{-4\log(1/\varepsilon)}=c^2\varepsilon^4,
$$

and therefore

$$
\sum_{j=1}^{N}\frac{\theta_j^{*4}}{\theta_j^{*2}+\varepsilon^2}2^{-2b_j^*}\leq Nc^2\varepsilon^4=o(\varepsilon^{\frac{4m}{2m+1}}).
$$

Now we have shown that the solution triplet (θ^*,b^*,ω^*) to (A_2) satisfy that θ^* is non-decreasing and $b_1^*<b_{\max}$. In order to prove $V_1\geq V_2$, it then suffices to show that if we take $\theta=\theta^*$ in (A_1), the minimizer $b^\star$ is non-increasing and $b_1^\star\leq b_{\max}$. In fact, if so, we will have $b^\star=b^*$ as well as $\omega^*=\frac{\theta_j^{*2}}{\theta_j^{*2}+\varepsilon^2}$ and then

$$
V_1\geq\min_{b:\sum_{j=1}^N b_j\leq B}\sum_{j=1}^{N}\left(\frac{\theta_j^{*4}}{\theta_j^{*2}+\varepsilon^2}2^{-2b_j}+\frac{\theta_j^{*2}\varepsilon^2}{\theta_j^{*2}+\varepsilon^2}\right)\geq V_2,
$$

completing the proof. □

Lemma 3.7.9. *Suppose that $W_{n,\xi}$ follows a non-central chi-square distribution with n degrees*

of freedom and non-centrality parameter ξ. We have for $n \geq 5$

$$\mathbb{E}\left(W_{n,\xi}^{-1}\right) \leq \frac{1}{n+\xi-4},$$

and for $n \geq 7$

$$\mathbb{E}\left(W_{n,\xi}^{-2}\right) \leq \frac{n-4}{(n-6)(n+\xi-4)(n+\xi-6)}.$$

Proof. It is well known that the non-central chi-square random variable $W_{n,\xi}$ can be written as a Poisson-weighted mixture of central chi-square distributions, i.e., $W_{n,\xi} \sim \chi^2_{n+2K}$ with $K \sim \text{Poisson}(\xi/2)$. Then

$$\begin{aligned}\mathbb{E}\left(W_{n,\xi}^{-1}\right) &= \mathbb{E}\left(\mathbb{E}(W_{n,\xi}^{-1} \mid K)\right) = \mathbb{E}\left(\frac{1}{n+2K-2}\right) \\ &\geq \frac{1}{n+2\mathbb{E}K-2} = \frac{1}{n+\xi-2}\end{aligned}$$

where we have used the fact that $\mathbb{E}(1/\chi^2_n) = n-2$ and Jensen's inequality. Similarly, we have

$$\begin{aligned}\mathbb{E}\left(W_{n,\xi}^{-2}\right) &= \mathbb{E}\left(\mathbb{E}(W_{n,\xi}^{-2} \mid K)\right) = \mathbb{E}\left(\frac{1}{(n+2K-2)(n+2K-4)}\right) \\ &\geq \frac{1}{(n+2\mathbb{E}K-2)(n+2\mathbb{E}K-4)} \\ &= \frac{1}{(n+\xi-2)(n+\xi-4)}\end{aligned}$$

Using the Poisson-weighted mixture representation, the following recurrence relation can be derived (Chattamvelli and Jones, 1995)

$$1 = \xi\mathbb{E}\left(W_{n+4,\xi}^{-1}\right) + n\mathbb{E}\left(W_{n+2,\xi}^{-1}\right), \tag{3.32}$$

$$\mathbb{E}\left(W_{n,\xi}^{-1}\right) = \xi\mathbb{E}\left(W_{n+4,\xi}^{-2}\right) + n\mathbb{E}\left(W_{n+2,\xi}^{-2}\right), \tag{3.33}$$

for $n \geq 3$. Thus,

$$\begin{aligned}\mathbb{E}\left(W^{-1}_{n+4,\xi}\right) &= \frac{1}{\xi} - \frac{n}{\xi}\mathbb{E}\left(W^{-1}_{n+2,\xi}\right) \\ &\leq \frac{1}{\xi} - \frac{n}{\xi}\frac{1}{n+\xi} \\ &= \frac{1}{n+\xi}.\end{aligned}$$

Replacing n by $n-4$ proves (3.7.9). On the other hand, rearranging (3.32), we get

$$\begin{aligned}\mathbb{E}\left(W^{-1}_{n+2,\xi}\right) &= \frac{1}{n} - \frac{\xi}{n}\mathbb{E}\left(W^{-1}_{n+4,\xi}\right) \\ &\leq \frac{1}{n} - \frac{\xi}{n}\frac{1}{n+\xi+2} \\ &= \frac{n+2}{n(n+\xi+2)}.\end{aligned}$$

Now using (3.33), we have

$$\begin{aligned}\mathbb{E}\left(W^{-2}_{n+4,\xi}\right) &= \frac{1}{\xi}\mathbb{E}\left(W^{-1}_{n,\xi}\right) - \frac{n}{\xi}\mathbb{E}\left(W^{-2}_{n+2,\xi}\right) \\ &\leq \frac{n}{\xi(n-2)(n+\xi)} - \frac{n}{\xi(n+\xi)(n+\xi-2)} \\ &= \frac{n}{(n-2)(n+\xi)(n+\xi-2)}.\end{aligned}$$

Replacing n by $n-4$ proves (3.7.9). □

Part II

Localized Forms of Minimax Theory

CHAPTER 4

LOCALIZED FORMS OF MINIMAX THEORY: INTRODUCTION AND RELATED WORK

Statistical minimax theory, as a measure of hardness of statistical task, has been criticized by many for being too conservative. It is indeed so, as it looks at the worst-case risk in a parameter family, regardless of the problem in hand. Given a particular problem instance, to utilize minimax analyses for quantifying its difficulty, we can put it into some natural classes of problems and then calculate the minimax risk for the class. However, since a problem instance can belong to a range of classes, for which the minimax risk can be quite different, it is not immediately clear how well we should expect or hope to solve each individual problem. As a response to such concerns there has been a great effort to develop adaptive procedures that are simultaneously minimax over a collection of parameter spaces. This point of view and history is particularly well explained in Donoho et al. (1995) in the context of global estimation of functions under integrated mean squared error. It should however be stressed such an approach is still provided by considering the worst-case risk over large parameter spaces, and it is not clear how the "collection of parameter spaces" could or should be chosen. It is then of interest to design a benchmark that is focused on the level of individual problem instance.

Cai and Low (2015) propose a framework for assessing the difficulty of solving individual problem, which is still based on minimax formulation, in the context of nonparametric estimation of convex function at a point. In order to assess the difficulty for each individual instance one must at least consider an additional function since otherwise the problem is degenerate. For a given instance $P \in \mathcal{P}$ it is natural to choose the other instance, say P', to be the one which is most difficult to distinguish from P. The benchmark $R_n(P; \mathcal{P})$ can

then be expressed as

$$R_n(P;\mathcal{P}) = \sup_{P'\in\mathcal{P}} \inf_{A\in\mathcal{A}_n} \max_{Q\in\{P,P'\}} \mathbb{E}_Q \, \mathrm{err}(A,Q)$$

using the same notation as established in Chapter 1. Apparently, this formulation gives a lower bound on the traditional minimax risk for the class $\mathcal{P}$–it is in fact the largest lower bound using a two-point subclass. It is though not guaranteed that this provides anything meaningful, not to mention as a measure of the hardness of the problem. However, as is shown in Cai and Low (2015) and in at least one other application, this actually provides a meaningful characterization of difficulty for solving individual problem, in the sense that

- $R_n(P;\mathcal{P})$ varies considerably over the collection of convex functions;
- There is a procedure that has a risk uniformly within a constant factor of $R_n(P;\mathcal{P})$ for every instance P and every n;
- Outperforming the benchmark $R_n(P;\mathcal{P})$ at some instance P leads to worse performance at other instance.

It is the combination of these three factors that make $R_n(P;\mathcal{P})$ a useful benchmark.

As we see in Cai and Low (2015) and in the chapter that follows, such a local formulation of minimax risk can usually be expressed by a modulus of continuity of the individual instance with $1/\sqrt{T}$ plugged in. A modulus of continuity is of the form

$$\omega_P(\epsilon) = \sup\{d(P,Q) : Q \in \mathcal{P},\, \kappa(P,Q) \le \epsilon\}$$

where d and κ are two (semi)metric defined on $\mathcal{P}$. d quantifies the distance between P and Q in a way related to the error measure, while κ measures the dissimilarity in the information input space between P and Q. Thus, with $1/\sqrt{T}$ plugged in, this modulus of continuity acts as a bridge from the input space to the output metric. We note that such modulus of continuity has also been studied before, in a global sense, by Donoho and Liu (1987) in the

context of log concave density estimation. In both the global case and our local case, the geometric quantity, modulus of continuity, characterizes the difficulty of statistical tasks.

CHAPTER 5

LOCAL MINIMAX COMPLEXITY OF CONVEX OPTIMIZATION

5.1 Introduction

The traditional analysis of algorithms is based on a worst-case, minimax formulation. One studies the running time, measured in terms of the smallest number of arithmetic operations required by any algorithm to solve any instance in the family of problems under consideration. Classical worst-case complexity theory focuses on discrete problems. In the setting of convex optimization, where the problem instances require numerical rather than combinatorial optimization, Nemirovsky and Yudin (1983) developed an approach to minimax analysis based on a first order oracle model of computation. In this model, an algorithm to minimize a convex function can make queries to a first-order "oracle," and the complexity is defined as the smallest error achievable using some specified minimum number of queries needed. Specifically, the oracle is queried with an input point $x \in \mathcal{C}$ from a convex domain $\mathcal{C}$, and returns an unbiased estimate of a subgradient vector to the function f at x. After T calls to the oracle, an algorithm A returns a value $\widehat{x}_A \in \mathcal{C}$, which is a random variable due to the stochastic nature of the oracle, and possibly also due to randomness in the algorithm. The Nemirovski-Yudin analysis reveals that, in the worst case, the number of calls to the oracle required to drive the expected error $\mathbb{E}(f(\widehat{x}_A) - \inf_{x \in \mathcal{C}} f(x))$ below ϵ scales as $T = O(1/\epsilon)$ for the class of strongly convex functions, and as $T = O(1/\epsilon^2)$ for the class of Lipschitz convex functions.

In practice, one naturally finds that some functions are easier to optimize than others. Intuitively, if the function is "steep" near the optimum, then the subgradient may carry a

great deal of information, and a stochastic gradient descent algorithm may converge relatively quickly. A minimax approach to analyzing the running time cannot take this into account for a particular function, as it treats the worst-case behavior of the algorithm over all functions. It would be of considerable interest to be able to assess the complexity of solving an individual convex optimization problem. Doing so requires a break from traditional worst-case thinking.

In this paper we revisit the traditional view of the complexity of convex optimization from the point of view of a type of localized minimax complexity. In local minimax, our objective is to quantify the intrinsic difficulty of optimizing a specific convex function f. With the target f fixed, we take an alternative function g within the same function class $\mathcal{F}$, and evaluate how the maximum expected error decays with the number of calls to the oracle, for an optimal algorithm designed to optimize either f or g. The local minimax complexity $R_T(f;\mathcal{F})$ is defined as the least favorable alternative g:

$$R_T(f;\mathcal{F}) = \sup_{g\in\mathcal{F}} \inf_{A\in\mathcal{A}_T} \max_{h\in\{f,g\}} \text{error}(A,h) \tag{5.1}$$

where $\text{error}(A,h)$ is some measure of error for the algorithm applied to function h. In contrast, the traditional global worst-case performance of the best algorithm, as defined by the minimax complexity $R_T(\mathcal{F})$ of Nemirovsky and Yudin, is

$$R_T(\mathcal{F}) = \inf_{A\in\mathcal{A}_T} \sup_{g\in\mathcal{F}} \text{error}(A,g). \tag{5.2}$$

The local minimax complexity can be thought of as the difficulty of optimizing the hardest alternative to the target function. Intuitively, a difficult alternative is a function g for which querying the oracle with g gives results similar to querying with f, but for which the value of $x \in \mathcal{C}$ that minimizes g is far from the value that minimizes f.

Our analysis ties this function-specific notion of complexity to a localized and computational analogue of the modulus of continuity that is central to statistical minimax analysis (Donoho and Liu, 1987, 1991). We show that the local minimax complexity gives a mean-

ingful benchmark for quantifying the difficulty of optimizing a specific function by proving a superefficiency result; in particular, outperforming this benchmark at some function must lead to a larger error at some other function. Furthermore, we propose an adaptive algorithm in the one-dimensional case that is based on binary search, and show that this algorithm automatically achieves the local minimax complexity, up to a logarithmic factor. Our study of the algorithmic complexity of convex optimization is motivated by the work of Cai and Low (2015), who propose an analogous definition in the setting of statistical estimation of a one-dimensional convex function. The present work can thus be seen as exposing a close connection between statistical estimation and numerical optimization of convex functions. In particular, our results imply that the local minimax complexity can be viewed as a computational analogue of Fisher information in classical statistical estimation.

In the following section we establish our notation, and give a technical overview of our main results, which characterize the local minimax complexity in terms of the computational modulus of continuity. In Section 5.2.1, we demonstrate the phenomenon of superefficiency of the local minimax complexity. In Section 5.3 we present the algorithm that adapts to the benchmark, together with an analysis of its theoretical properties. We also present simulations of the algorithm and comparisons to traditional stochastic gradient descent. Finally, we conclude with a brief review of related work and a discussion of future research directions suggested by our results.

5.2 Local minimax complexity

In this section, we first establish notation and define a modulus of continuity for a convex function f. We then state our main result, which links the local minimax complexity to this modulus of continuity.

Let $\mathcal{F}$ be the collection of Lipschitz convex functions defined on a compact convex set $\mathcal{C} \subset \mathbb{R}^d$. Given a function $f \in \mathcal{F}$, our goal is to find a minimum point, $x_f^* \in \arg\min_{x \in \mathcal{C}} f(x)$. However, our knowledge about f can only be gained through a first-order oracle. The oracle,

upon being queried with $x \in \mathcal{C}$, returns $f'(x) + \xi$, where $f'(x)$ is a subgradient of f at x and $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$. When the oracle is queried with a non-differentiable point x of f, instead of allowing the oracle to return an arbitrary subgradient at x, we assume that it has a deterministic mechanism for producing $f'(x)$. That is, when we query the oracle with x twice, it should return two random vectors with the same mean $f'(x)$. Such an oracle can be realized, for example, by taking $f'(x) = \arg\min_{z \in \partial f(x)} \|z\|$.

Consider optimization algorithms that make a total of T queries to this first-order oracle, and let $\mathcal{A}_T$ be the collection of all such algorithms. For $A \in \mathcal{A}_T$, denote by $\widehat{x}_A$ the output of the algorithm. We write $\mathrm{err}(x, f)$ for a measure of error for using x as the estimate of the minimum point of $f \in \mathcal{F}$. In this notation, the usual minimax complexity is defined as

$$R_T(\mathcal{F}) = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}_f \, \mathrm{err}(\widehat{x}_A, f). \tag{5.3}$$

Note that the algorithm A queries the oracle at up to T points $x_t \in \mathcal{C}$ selected sequentially, and the output $\widehat{x}_A$ is thus a function of the entire sequence of random vectors $v_t \sim N(f'(x_t), \sigma^2 I_d)$ returned by the oracle. The expectation $\mathbb{E}_f$ denotes the average with respect to this randomness (and any additional randomness injected by the algorithm itself). The minimax risk $R_T(\mathcal{F})$ characterizes the hardness of the entire class $\mathcal{F}$. To quantify the difficulty of optimizing an individual function f, we consider the following local minimax complexity, comparing f to its hardest local alternative

$$R_T(f; \mathcal{F}) = \sup_{g \in \mathcal{F}} \inf_{A \in \mathcal{A}_T} \max_{h \in \{f, g\}} \mathbb{E}_h \, \mathrm{err}(\widehat{x}_A, h). \tag{5.4}$$

We now proceed to define a computational modulus of continuity that characterizes the local minimax complexity. Let $\mathcal{X}_f^* = \arg\min_{x \in \mathcal{C}} f(x)$ be the set of minimum points of function f. We consider $\mathrm{err}(x, f) = \inf_{y \in \mathcal{X}_f^*} \|x - y\|$ as our measure of error. Define $d(f, g) = \inf_{x \in \mathcal{X}_f^*, y \in \mathcal{X}_g^*} \|x - y\|$ for $f, g \in \mathcal{F}$. It is easy to see that $\mathrm{err}(x, f)$ and $d(f, g)$ satisfy

the *exclusion inequality*

$$\mathrm{err}(x, f) < \frac{1}{2} d(f, g) \quad \text{implies} \quad \mathrm{err}(x, g) \geq \frac{1}{2} d(f, g). \tag{5.5}$$

Next we define

$$\kappa(f, g) = \sup_{x \in \mathcal{C}} \|f'(x) - g'(x)\| \tag{5.6}$$

where $f'(x)$ is the unique subgradient of f that is returned as the mean by the oracle when queried with x. For example, if we take $f'(x) = \arg\min_{z \in \partial f(x)} \|z\|$, we have

$$\kappa(f, g) = \sup_{x \in \mathcal{C}} \| \operatorname*{Proj}_{\partial f(x)}(0) - \operatorname*{Proj}_{\partial g(x)}(0) \| \tag{5.7}$$

where $\mathrm{Proj}_B(z)$ is the projection of z to the set B. Thus, $d(f, g)$ measures the dissimilarity between two functions in terms of the distance between their minimizers, whereas $\kappa(f, g)$ measures the dissimilarity by the largest separation between their subgradients at any given point.

Given d and κ, we define the *modulus of continuity* of d with respect to κ at the function f by

$$\omega_f(\epsilon) = \sup \{ d(f, g) : g \in \mathcal{F}, \kappa(f, g) \leq \epsilon \}. \tag{5.8}$$

We now show how to calculate the modulus for some specific functions.

Example 5.2.1. Suppose that f is a convex function on a one-dimensional interval $\mathcal{C} \subset \mathbb{R}$. Then we have

$$\omega_f(\epsilon) = \sup \left\{ \inf_{x \in \mathcal{X}_f^*} |x - y| : y \in \mathcal{C}, |f'(y)| < \epsilon \right\}. \tag{5.9}$$

This essentially says that the modulus of continuity measures the size (in fact, the larger half-width) of the the "flat set" where the magnitude of the subderivative is smaller than ϵ. See Figure 5.1 for an illustration Thus, for the class of symmetric functions $f(x) = \frac{1}{k}|x|^k$

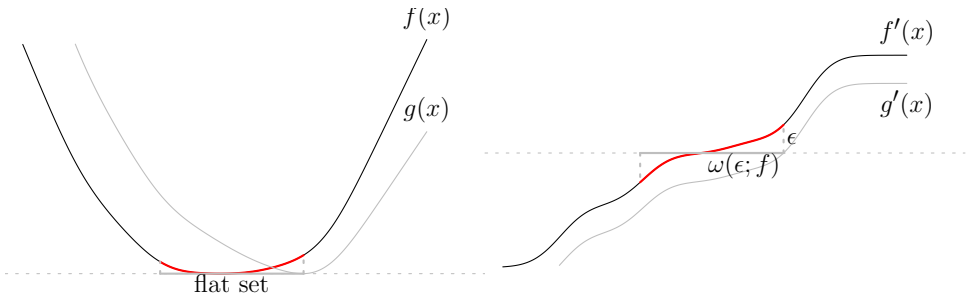

Figure 5.1: Illustration of the flat set and the modulus of continuity. Both the function f (left) and its derivative f' (right) are shown (black curves), along with one of the many possible alternatives, g and its derivative g' (solid gray curves), that achieve the sup in the definition of $\omega_f(\epsilon)$. The flat set contains all the points for which $|f'(x)| < \epsilon$, and $\omega_f(\epsilon)$ is the larger half width of the flat set.

over $\mathcal{C} = [-1, 1]$, with $k > 1$,

$$\omega_f(\epsilon) = \epsilon^{\frac{1}{k-1}}. \tag{5.10}$$

For the asymmetric case $f(x) = \frac{1}{k_l}|x|^{k_l} I(-1 \le x \le 0) + \frac{1}{k_r}|x|^{k_r} I(0 < x \le 1)$ with $k_l, k_r > 1$,

$$\omega_f(\epsilon) = \epsilon^{\frac{1}{k_l \vee k_r - 1}}. \tag{5.11}$$

That is, the size of the flat set depends on the flatter side of the function.

Local minimax is characterized by the modulus

We now state our main result linking the local minimax complexity to the modulus of continuity. We say that the modulus of the continuity has *polynomial growth* if there exists $\alpha > 0$ and ϵ_0, such that for any $c \ge 1$ and $\epsilon \le \epsilon_0/c$

$$\omega_f(c\epsilon) \le c^\alpha \omega_f(\epsilon). \tag{5.12}$$

Our main result below shows that the modulus of continuity characterizes the local minimax complexity of optimization of a particular convex function, in a manner similar to how the modulus of continuity quantifies the (local) minimax risk in a statistical estimation setting

Cai and Low (2015); Donoho and Liu (1987, 1991), relating the objective to a geometric property of the function.

Theorem 5.2.1. *Suppose that $f \in \mathcal{F}$ and that $\omega_f(\epsilon)$ has polynomial growth. Then there exist constants C_1 and C_2 independent of T and $T_0 > 0$ such that for all $T > T_0$*

$$C_1\, \omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \leq R_T(f;\mathcal{F}) \leq C_2\, \omega_f\left(\frac{\sigma}{\sqrt{T}}\right). \tag{5.13}$$

Remark 5.2.1.1. We use the error metric $\mathrm{err}(x,f) = \inf_{y\in\mathcal{X}_f^*} \|x-y\|$ here. For a given a pair (err, d) that satisfies the exclusion inequality (5.5), our proof technique applies to yield the corresponding lower bound. For example, we could use $\mathrm{err}(x,f) = \inf_{y\in\mathcal{X}_f^*} |v^T(x-y)|$ for some vector v. This error metric would be suitable when we wish to estimate $v^T x_f^*$, for example, the first coordinate of x_f^*. Another natural choice of error metric is $\mathrm{err}(x,f) = f(x) - \inf_{x\in\mathcal{C}} f(x)$, with a corresponding distance $d(f,g) = \inf_{x\in\mathcal{C}} |f(x) - \inf_x f(x) + g(x) - \inf_x g(x)|$. For this case, while the proof of the lower bound stays exactly the same, further work is required for the upper bound, which is beyond the scope of this paper.

Remark 5.2.1.2. Although the theorem gives an upper bound for the local minimax complexity, this does not guarantee the existence of an algorithm that achieves the local complexity for any function. Therefore, it is important to design an algorithm that adapts to this benchmark for each individual function. We solve this problem in the one-dimensional case in Section 5.3.

The proof of this theorem is given in the appendix. We now illustrate the result with examples that verify the intuition that different functions should have different degrees of difficulty for stochastic convex optimization.

Example 5.2.2. For the function $f(x) = \frac{1}{k}|x|^k$ with $x \in [-1,1]$ for $k > 1$, we have $R_T(f;\mathcal{F}) = O\big(T^{-\frac{1}{2(k-1)}}\big)$. When $k = 2$, we recover the strongly convex case, where the (global) minimax complexity is $O\big(1/\sqrt{T}\big)$ with respect to the error $\mathrm{err}(x,f) = \inf_{y\in\mathcal{X}_f^*} \|x - y\|$. We see a faster rate of convergence for $k < 2$. As $k \to \infty$, we also see that the error

fails to decrease as T gets large. This corresponds to the worst case for any Lipschitz convex function. In the asymmetric setting with $f(x) = \frac{1}{k_l}|x|^{k_l} I(-1 \le x \le 0) + \frac{1}{k_r}|x|^{k_r} I(0 < x \le 1)$ with $k_l, k_r > 1$, we have $R_T(f; \mathcal{F}) = O(T^{-\frac{1}{2(k_l \vee k_r - 1)}})$.

The following example illustrates that the local minimax complexity and modulus of continuity are consistent with known behavior of stochastic gradient descent for strongly convex functions.

Example 5.2.3. In this example we consider the error $\mathrm{err}(x, f) = \inf_{y \in \mathcal{X}_f^*} |v^T(x - y)|$ for some vector v, and let f be an arbitrary convex function satisfying $\nabla^2 f(x_f^*) \succ 0$ with Hessian continuous around x_f^*. Thus the optimizer x_f^* is unique. If we define $g_w(x) = f(x) - w^T \nabla^2 f(x_f^*) x$, then $g_w(x)$ is a convex function with unique minimizer and

$$\kappa(f, g_w) = \sup_x \left\{ \left\| \nabla f(x) - (\nabla f(x) - \nabla^2 f(x_f^*) w) \right\| \right\} = \left\| \nabla^2 f(x_f^*) w \right\|. \tag{5.14}$$

Thus, defining $\delta(w) = x_f^* - x_{g_w}^*$,

$$\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \ge \sup_w \{|v^T \delta(w)| : \left\| \nabla^2 f(x_f^*) w \right\| \le \sigma/\sqrt{T}\} \ge \sup_u \left| v^T \delta\left(\frac{\sigma}{\sqrt{T}} \nabla^2 f(x_f^*)^{-1} u\right) \right|. \tag{5.15}$$

By the convexity of g_w, we know that $x_{g_w}^*$ satisfies $\nabla f(x_{g_w}^*) - \nabla^2 f(x_f^*)^{-1} w = 0$, and therefore by the implicit function theorem, $x_{g_w}^* = x_f^* + w + o(\|w\|)$ as $w \to 0$. Thus,

$$\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \ge \frac{\sigma}{\sqrt{T}} \left\| \nabla^2 f(x_f^*)^{-1} v \right\| + o\left(\frac{\sigma}{\sqrt{T}}\right) \quad \text{as } T \to \infty. \tag{5.16}$$

In particular, we have the local minimax lower bound

$$\liminf_{T \to \infty} \sqrt{T} R_T(f; \mathcal{F}) \ge C_1 \sigma \left\| \nabla^2 f(x_f^*)^{-1} v \right\| \tag{5.17}$$

where C_1 is the same constant appearing in Theorem 5.2.1. This shows that the local minimax complexity captures the function-specific dependence on the constant in the strongly

convex case. Stochastic gradient descent with averaging is known to adapt to this strong convexity constant (Ruppert, 1988; Polyak and Juditsky, 1992; Moulines and Bach, 2011).

5.2.1 Superefficiency

Having characterized the local minimax complexity in terms of a computational modulus of continuity, we would now like to show that there are consequences to outperforming it at some function. This will strengthen the case that the local minimax complexity serves as a meaningful benchmark to quantify the difficulty of optimizing any particular convex function.

Suppose that f is any one-dimensional function such that $\mathcal{X}_f^* = [x_l, x_r]$, which has as asymptotic expansion around $\{x_l, x_r\}$ of the form

$$f(x_l - \delta) = f(x_l) + \lambda_l \delta^{k_l} + o(\delta^{k_l}) \quad \text{and} \quad f(x_r + \delta) = f(x_r) + \lambda_r \delta^{k_r} + o(\delta^{k_r}) \tag{5.18}$$

for $\delta > 0$, some powers $k_l, k_r > 1$, and constants $\lambda_l, \lambda_r > 0$. The following result shows that if any algorithm significantly outperforms the local modulus of continuity on such a function, then it underperforms the modulus on a nearby function.

Proposition 5.2.2. *Let f be any convex function satisfying the asymptotic expansion* (5.21) *around its optimum. Suppose that $A \in \mathcal{A}_T$ is any algorithm that satisfies*

$$\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f) \leq \sqrt{\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f)^2} \leq \delta_T \, \omega_f\left(\frac{\sigma}{\sqrt{T}}\right), \tag{5.19}$$

where $\delta_T < C_1$. Define $g_{-1}(x) = f(x) - \epsilon_T x$ and $g_1(x) = f(x) + \epsilon_T x$, where ϵ_T is given by $\epsilon_T = \sqrt{\sigma^2 \log\left(\frac{C_1}{\delta_T}\right)/T}$. Then for some $g \in \{g_{-1}, g_1\}$, there exists T_0 such that $T \geq T_0$ implies

$$\mathbb{E}_g \operatorname{err}(\widehat{x}_A, g) \geq C \, \omega_g\left(\sqrt{\frac{\sigma^2 \log\left(C_1/\delta_T\right)}{T}}\right) \tag{5.20}$$

for some constant C that only depends on $k = k_l \vee k_r$.

A proof of this result is given in the appendix, where it is derived as a consequence of a more general statement. We remark that while condition (5.19) involves the squared error $\sqrt{\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f)^2}$, we expect that the result holds with only the weaker inequality on the absolute error $\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f)$.

It follows from this proposition that if an algorithm A significantly outperforms the local minimax complexity in the sense that (5.19) holds for some sequence $\delta_T \to 0$ with $\liminf_T e^T \delta_T = \infty$, then there exists a sequence of convex functions g_T with $\kappa(f, g_T) \to 0$, such that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{g_T} \operatorname{err}(\widehat{x}_A, g_T)}{\omega_{g_T}\left(\sqrt{\sigma^2 \log\left(\frac{C_1}{\delta_T}\right)/T}\right)} > 0. \tag{5.21}$$

This is analogous to the phenomenon of superefficiency in classical parametric estimation problems, where outperforming the asymptotically optimal rate given by the Fisher information implies worse performance at some other point in the parameter space. In this sense, ω_f can be viewed as a computational analogue of Fisher information in the setting of convex optimization. We note that superefficiency has also been studied in nonparametric settings (Brown and Low, 1996a), and a similar result was shown by Cai and Low (2015) for local minimax estimation of convex functions.

5.3 An adaptive optimization algorithm

In this section, we show that a simple stochastic binary search algorithm achieves the local minimax complexity in the one-dimensional case.

The general idea of the algorithm is as follows. Suppose that we are given a budget of T queries to the oracle. We divide this budget into $T_0 = \lfloor T/E \rfloor$ queries over each of $E = \lfloor r \log T \rfloor$ many rounds, where $r > 0$ is a constant to be specified later. In each round, we query the oracle T_0 times for the derivative at the mid-point of the current interval. Estimating the derivative by averaging over the queries, we proceed to the left half of the

interval if the estimated sign is positive, and to the right half of the interval of the estimated sign is negative. The details are given in Algorithm 1.

Algorithm 1 Sign testing binary search

Input: T, r.
Initialize: (a_0, b_0), $E = \lfloor r \log T \rfloor$, $T_0 = \lfloor T/E \rfloor$.
for $e = 1, \ldots, E$ **do**
  Query $x_e = (a_e + b_e)/2$ for T_0 times to get $Z_t^{(e)}$ for $t = 1, \ldots, T_0$.
  Calculate the average $\bar{Z}_{T_0}^{(e)} = \frac{1}{T_0} \sum_{t=1}^{T_0} Z_t^{(e)}$.
  If $\bar{Z}_{T_0}^{(e)} > 0$, set $(a_{e+1}, b_{e+1}) = (a_e, x_e)$.
  If $\bar{Z}_{T_0}^{(e)} \leq 0$, set $(a_{e+1}, b_{e+1}) = (x_e, b_e)$.
end for
Output: x_E.

We will show that this algorithm adapts to the local minimax complexity up to a logarithmic factor. First, the following result shows that the algorithm gets us close to the "flat set" of the function.

Proposition 5.3.1. *For $\delta \in (0, 1)$, let $C_\delta = \sigma\sqrt{2\log(E/\delta)}$. Define*

$$\mathcal{I}_\delta = \left\{ y \in \mathrm{dom}(f) : |f'(y)| < \frac{C_\delta}{\sqrt{T_0}} \right\}. \tag{5.22}$$

Suppose that $(a_0, b_0) \cap \mathcal{I}_\delta \neq \emptyset$. Then

$$\mathit{dist}(x_E, \mathcal{I}_\delta) \leq 2^{-E}(b_0 - a_0) \tag{5.23}$$

with probability at least $1 - \delta$.

This proposition tells us that after E rounds of bisection, we are at most a distance $2^{-E}(b_0 - a_0)$ from the flat set $\mathcal{I}_\delta$. In terms of the distance to the minimum point, we have

$$\inf_{x \in \mathcal{X}_f^*} |x_E - x| \leq 2^{-E}(b_0 - a_0) + \sup\Big\{ \inf_{x \in \mathcal{X}_f^*} |x - y| : y \in \mathcal{I}_\delta \Big\}. \tag{5.24}$$

If the modulus of continuity satisfies the polynomial growth condition, we then obtain the following.

Corollary 5.3.2. *Let $\alpha_0 > 0$. Suppose ω_f satisfies the polynomial growth condition* (5.12) *with constant $\alpha \leq \alpha_0$. Let $r = \frac{1}{2}\alpha_0$. Then with probability at least $1-\delta$ and for large enough T,*

$$\inf_{x \in \mathcal{X}_f^*} |x_E - x| \leq \tilde{C}\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \tag{5.25}$$

where the term $\tilde{C}$ hides a dependence on $\log T$ and $\log(1/\delta)$.

The proofs of these results are given in the appendix.

Simulations showing adaptation to the benchmark

We now demonstrate the performance of the stochastic binary search algorithm, making a comparision to stochastic gradient descent. For the stochastic gradient descent algorithm, we perform T steps of update

$$x_{t+1} = x_t - \eta(t) \cdot \widehat{g}(x_t) \tag{5.26}$$

where $\eta(t)$ is a stepsize function, chosen as either $\eta(t) = \frac{1}{t}$ or $\eta(t) = \frac{1}{\sqrt{t}}$. We first consider the following setup with symmetric functions f:

1. The function to optimize is $f_k(x) = \frac{1}{k}|x - x^*|^k$ for $k = \frac{3}{2}$, 2 or 3.

2. The minimum point $x^* \sim \text{Unif}(-1, 1)$ is selected uniformaly at random over the interval.

3. The oracle returns the derivative at the query point with additive $N(0, \sigma^2)$ noise, $\sigma = 0.1$.

4. The optimization algorithms know *a priori* that the minimum point is inside the interval $(-2, 2)$. Therefore, the binary search starts with interval $(-2, 2)$ and the stochastic

gradient descent starts at $x_0 \sim \text{Unif}(-2, 2)$ and project the query points to the interval $(-2, 2)$.

5. We carry out the simulation for values of T on a logarithmic grid between 100 and 10,000. For each setup, we average the error $|\widehat{x} - x^*|$ over 1,000 runs.

The simulation results are shown in the top 3 panels of Figure 5.2. Several properties predicted by our theory are apparent from the simulations. First, the risk curves for the stochastic binary search algorithm parallel the gray curves. This indicates that the optimal rate of convergence is achieved. Thus, the stochastic binary search adapts to the curvature of different functions and yields the optimal local minimax complexity, as given by our benchmark. Second, the stochastic gradient descent algorithms with stepsize $1/t$ achieve the optimal rate when $k = 2$, but not when $k = 3$; with stepsize $1/\sqrt{t}$ SGD gets close to the optimal rate when $k = 3$, but not when $k = 2$. Neither leads to the faster rate when $k = \frac{3}{2}$. This is as expected, since the stepsize needs to be adapted to the curvature at the optimum in order to achieve the optimal rate.

Next, we consider a set of asymmetric functions. Using the same setup as in the symmetric case, we consider the functions of the form $f(x) = \frac{1}{k_l}|x-x^*|^{k_l} I(x-x^* \leq 0) + \frac{1}{k_r}|x-x^*|^{k_r} I(x-x^* > 0)$, for exponent pairs (k_1, k_2) chosen to be $(\frac{3}{2}, 2)$, $(\frac{3}{2}, 3)$ and $(2, 3)$. The simulation results are shown in the bottom three panels of Figure 5.2. We observe that the stochastic binary search once again achieves the optimal rate, which is determined by the flatter side of the function, that is, the larger of k_l and k_r.

5.4 Related work and future directions

In related recent work, Ramdas and Singh (2013b) study minimax complexity for the class of Lipschitz convex functions that satisfy $f(x) - f(x_f^*) \geq \frac{\lambda}{2}\|x - x_f^*\|^k$. They show that the minimax complexity under the function value error is of the order $T^{-\frac{k}{2(k-1)}}$. Iouditski and Nesterov (2014) also consider minimax complexity for the class of k-uniformly convex

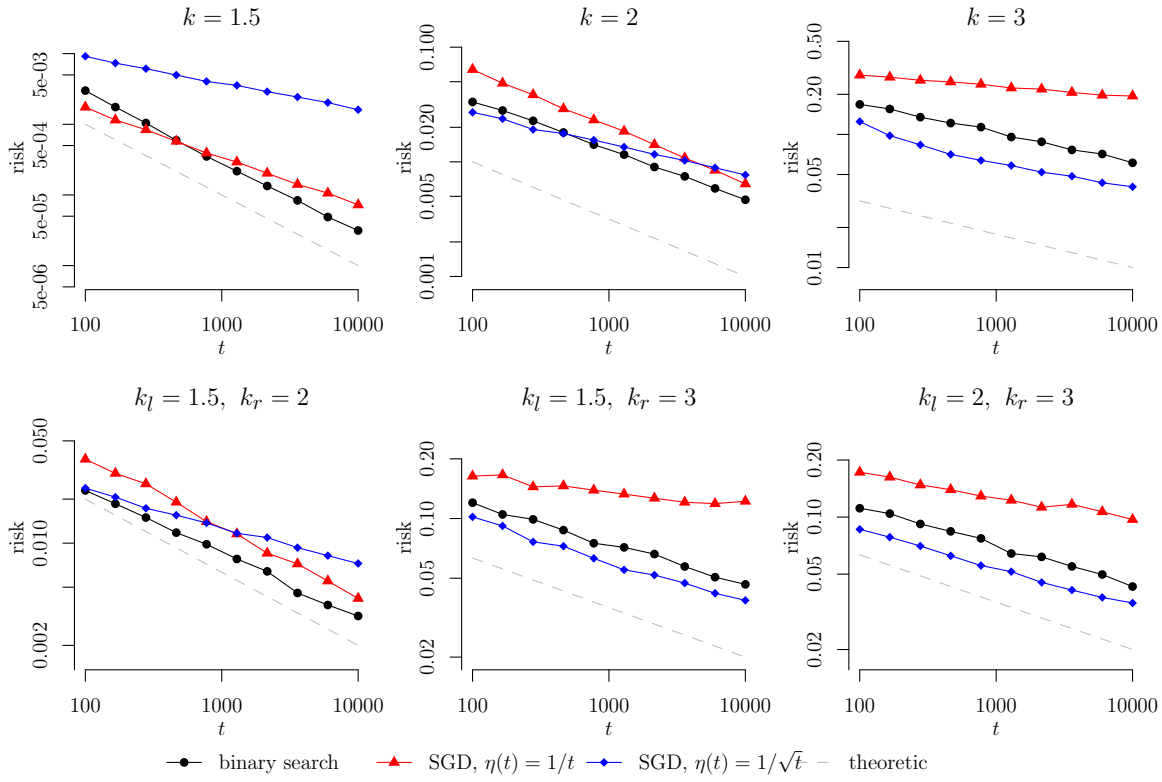

Figure 5.2: Simulation results: Averaged risk versus number of queries T. The black curves correspond to the risk of the stochastic binary search algorithm. The red and blue curves are for the stochastic gradient descent methods, red for stepsize $1/t$ and blue for $1/\sqrt{t}$. The dashed gray lines indicate the optimal convergence rate. Note that the plots are on a log-log scale. The plots on the top panels are for the symmetric cases $f(x) = \frac{1}{k}|x - x^*|^k$; the lower plots are for the asymmetric cases.

functions for $k > 2$. They give an adaptive algorithm based on stochastic gradient descent that achieves the minimax complexity up to a logarithmic factor. Connections with active learning are developed in Ramdas and Singh (2013a), with related ideas appearing in Castro and Nowak (2008). Adaptivity in this line of work corresponds to the standard notion in statistical estimation, which seeks to adapt to a large subclass of a parameter space. In contrast, the results in the current paper quantify the difficulty of stochastic convex optimization at a much finer scale, as the benchmark is determined by the specific function to be optimized.

The stochastic binary search algorithm presented in Section 5.3, despite being adaptive, has a few drawbacks. It requires the modulus of continuity of the function to satisfy polynomial growth, with a parameter α bounded away from 0. This rules out cases such as $f(x) = |x|$, which should have an error that decays exponentially in T; it is of interest to handle this case as well. It would also be of interest to construct adaptive optimization procedures tuned to a fixed numerical precision. Such procedures should have different running times depending on the hardness of the problem. Progress on both problems has been made, and will be reported elsewhere.

Another challenge is to remove the logarithmic factors appearing in the binary search algorithm developed in Section 5.3. In one dimension, stochastic convex optimization is intimately related to a noisy root finding problem for a monotone function taking values in $[-a, a]$ for some $a > 0$. Karp and Kleinberg (2007) study optimal algorithms for such root finding problems in a discrete setting. A binary search algorithm that allows backtracking is proposed, which saves log factors in the running time. It would be interesting to study the use of such techniques in our setting.

Other areas that warrant study involve the dependence on dimension. The scaling with dimension of the local minimax complexity and modulus of continuity is not fully revealed by the current analysis. Moreover, the superefficiency result and the adaptive algorithm presented here are only for the one-dimensional case. We note that a form of adaptive

stochastic gradient algorithm for the class of uniformly convex functions in general, fixed dimension is developed in Iouditski and Nesterov (2014).

Finally, a more open-ended direction is to consider larger classes of stochastic optimization problems. For instance, minimax results are known for functions of the form $f(x) := \mathbb{E}\, F(x;\xi)$ where ξ is a random variable and $x \mapsto F(x;\xi)$ is convex for any ξ, when f is twice continuously differentiable around the minimum point with positive definite Hessian. However, the role of the local geometry is not well understood. It would be interesting to further develop the local complexity techniques introduced in the current paper, to gain insight into the geometric structure of more general stochastic optimization problems.

5.5 Proofs of technical results

5.5.1 Proof of Theorem 5.2.1

Lower bound

For a function $f \in \mathcal{F}$, let P_f denote the distribution of stochastic gradients observable by an estimation scheme $\widehat{x}$, and let P_f^T denote the distribution of T sequentially queried stochastic gradients for f. We define the pairwise minimax risk for optimization of a pair of function f and g by

$$R_T(f,g) := \inf_{A \in \mathcal{A}_T} \max\left\{\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f), \mathbb{E}_g \operatorname{err}(\widehat{x}_A, g)\right\}, \tag{5.27}$$

and the local minimax lower bound can be written as

$$R_T(f;\mathcal{F}) := \sup_{g \in \mathcal{F}} R_T(f,g). \tag{5.28}$$

Let us show how the modulus of continuity gives a lower bound. We first state a lemma.

Lemma 5.5.1. *Let f, g be arbitrary convex functions and d satisfy the exclusion inequal-*

ity (5.5). *Then*

$$R_T(f,g) \geq \frac{d(f,g)}{4}\left(1 - \left\|P_f^T - P_g^T\right\|_{\mathrm{TV}}\right). \tag{5.29}$$

Proof. Temporarily hiding the number of iterations T for simplicity, we have by Markov's inequality that

$$\max\left\{\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f), \mathbb{E}_g \operatorname{err}(\widehat{x}_A, g)\right\} \tag{5.30}$$

$$\geq \frac{1}{2}d(f,g)\max\left\{P_f(\operatorname{err}(\widehat{x}_A, f) \geq \frac{1}{2}d(f,g)), P_g(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g))\right\}. \tag{5.31}$$

Now, we apply an essentially standard reduction of estimation to testing, because we have

$$2\max\left\{P_f(\operatorname{err}(\widehat{x}_A, f) \geq \frac{1}{2}d(f,g)), P_g(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g))\right\} \tag{5.32}$$

$$\geq P_f(\operatorname{err}(\widehat{x}_A, f) \geq \frac{1}{2}d(f,g)) + P_g(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g)) \tag{5.33}$$

$$= 1 - P_f(\operatorname{err}(\widehat{x}_A, f) < \frac{1}{2}d(f,g)) + P_g(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g)) \tag{5.34}$$

$$\geq 1 - P_f(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g)) + P_g(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g)), \tag{5.35}$$

where in the last line we have used the exclusion inequality to see that $\operatorname{err}(\widehat{x}_A, f) < \frac{1}{2}d(f,g)$ implies $\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g)$ so that

$$P_f(\operatorname{err}(\widehat{x}_A, f) < \frac{1}{2}d(f,g)) \leq P_f(\operatorname{err}(\widehat{x}_A, g) \geq \frac{1}{2}d(f,g)). \tag{5.36}$$

Thus, we find that

$$\frac{4}{d(f,g)}\max\left\{\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f), \mathbb{E}_g \operatorname{err}(\widehat{x}_A, g)\right\} \geq \inf_S\left\{1 - P_f^T(S) + P_g^T(S)\right\} = 1 - \left\|P_f^T - P_g^T\right\|_{\mathrm{TV}}, \tag{5.37}$$

which yields the lemma. □

Now we can prove a minimax lower bound. Let Y_i be the ith observed gradient, where $P_f(Y_i \mid Y_{1:i-1})$ denotes the conditional distribution of Y_i under the oracle for function f.

We have by the chain rule that

$$D_{\mathrm{kl}}\left(P_f^T \| P_g^T\right) = \sum_{i=1}^{T} \mathbb{E}_f\left[D_{\mathrm{kl}}\left(P_f(Y_i \mid Y_{1:i-1}) \| P_g(Y_i \mid Y_{1:i-1})\right)\right]. \tag{5.38}$$

It is no loss of generality to assume that the ith gradient query point x_i is measureable with respect to $Y_{1:i-1}$ (this follows because if a randomized algorithm does well in expectation, there is at least one realization of its randomness that has small risk, so we can just take that realization and assume the procedure is deterministic). Using that we have a Gaussian oracle, we have

$$\begin{aligned} D_{\mathrm{kl}}\left(P_f(Y_i \mid Y_{1:i-1}) \| P_g(Y_i \mid Y_{1:i-1})\right) &= D_{\mathrm{kl}}\left(\mathcal{N}(f'(x_i), \sigma^2 I_{d\times d}) \| \mathcal{N}(g'(x_i), \sigma^2 I_{d\times d})\right) & (5.39) \\ &= \frac{1}{2\sigma^2}\left\|f'(x_i) - g'(x_i)\right\|^2 \leq \frac{1}{2\sigma^2}\kappa(f,g)^2. & (5.40) \end{aligned}$$

Noting the not completely standard upper bound

$$\left\|P_f^T - P_g^T\right\|_{\mathrm{TV}} \leq 1 - \exp\left(-\frac{1}{2}D_{\mathrm{kl}}\left(P_f^T \| P_g^T\right)\right) \tag{5.41}$$

on the variation distance (see Tsybakov (2008, Lemma 2.6)), we also have by Lemma 5.5.1 that

$$R_T(f,g) \geq \frac{d(f,g)}{4}\exp\left(-\frac{T}{4\sigma^2}\kappa(f,g)^2\right). \tag{5.42}$$

Consider the collection of functions

$$\mathcal{F}_T := \left\{g \in \mathcal{F} : \kappa(f,g)^2 \leq \frac{\sigma^2}{T}\right\}. \tag{5.43}$$

Certainly this collection is non-empty (it includes f). For any $\epsilon > 0$, there must exist some $g \in \mathcal{F}_T$ such that $d(f,g) \geq (1-\epsilon)\omega_f(1/\sqrt{T})$. Let g_T denote such a g. Then we have

$$R_T(f) \geq R_T(f, g_T) \geq \frac{d(f,g_T)}{4}e^{-\frac{1}{4}} \geq \frac{1-\epsilon}{4}e^{-\frac{1}{4}}\omega_f\left(\frac{\sigma}{\sqrt{T}}\right). \tag{5.44}$$

We have

$$R_T(f) \geq \frac{1}{4e^{1/4}}\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \geq \frac{3}{16}\omega_f\left(\frac{\sigma}{\sqrt{T}}\right). \tag{5.45}$$

Upper bound

Suppose that we have two functions $f_{-1}, f_1 \in \mathcal{F}$. Let

$$x^\dagger = \arg\max_{x\in\mathcal{C}} \left\{\|f'_{-1}(x) - f'_1(x)\|\right\} \tag{5.46}$$

be the point at which the two functions differ the most in terms of the subgradients. Let $\theta \in \{-1, 1\}$ be the parameter. Consider an algorithm that queries the oracle with $x^\dagger$ for T times. Let Z_t be the response from the oracle at time t. Let

$$W = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t - \frac{\sqrt{T}}{2}(f'_1(x^\dagger) + f'_{-1}(x^\dagger)) \tag{5.47}$$

With the normality assumption on the noise, we have

$$W \sim N(\theta\gamma_T, \sigma^2 I) \tag{5.48}$$

where

$$\gamma_T = \frac{\sqrt{T}}{2}\left(f'_1(x^\dagger) - f'_{-1}(x^\dagger)\right). \tag{5.49}$$

Then we construct

$$\overline{W} = \|\gamma_T\|^{-1}\gamma_T^{\mathrm{T}} W \sim N(\theta\|\gamma_T\|, \sigma^2), \tag{5.50}$$

which is a sufficient statistic for the problem of estimating θ. Based on $\overline{W}$ we can obtain an estimate $\widehat{\theta}$ of θ, and let the output of our algorithm be

$$\widehat{x}_T = \frac{x_1^* + x_{-1}^*}{2} + \widehat{\theta}\frac{x_1^* - x_{-1}^*}{2} \tag{5.51}$$

where $x_1^* \in \mathcal{X}_{f_1}^*$ and $x_{-1}^* \in \mathcal{X}_{f_{-1}}^*$ satisfy $\|x_1 - x_{-1}\| = \inf_{x \in \mathcal{X}_{f_1}^*} \inf_{y \in \mathcal{X}_{f_{-1}}^*} \|x - y\|$. It then follows

$$\inf_{A \in \mathcal{A}_T} \max_{\theta = \pm 1} \mathbb{E}_\theta \|\widehat{x}_A - x_\theta^*\| \leq \max_{\theta = \pm 1} \mathbb{E}_\theta \|\widehat{x}_T - x_\theta^*\| \tag{5.52}$$

$$\leq \frac{1}{2} \|x_1^* - x_{-1}^*\| \inf_{\widehat{\theta}} \sup_{\theta = \pm 1} \mathbb{E}_\theta |\widehat{\theta} - \theta| \tag{5.53}$$

$$= \frac{1}{2} \|x_1^* - x_{-1}^*\| \|\gamma_T\|^{-1} \lambda(\|\gamma_T\|, \sigma) \tag{5.54}$$

where $\lambda(\tau, \sigma) = \inf_{\widehat{\mu}} \sup_{\mu = \pm \tau} \mathbb{E}_\mu |\widehat{\mu} - \mu|$ is the minimax (ℓ_1) risk of estimating the mean of $Z \sim N(\tau, \sigma^2)$ for the class $\mu \in \{-\tau, \tau\}$.

Now take $f_{-1} = f$ and $f_1 = g$. Note that $\|\gamma_T\| = \frac{\sqrt{T}}{2} \kappa(f, g)$. From (5.54) we have

$$R_T(f; \mathcal{F}) = \sup_{g \in \mathcal{F}} \inf_{A \in \mathcal{A}_T} \max \left\{ \mathbb{E}_f \|\widehat{x}_T - x_f^*\|, \mathbb{E}_g \|\widehat{x}_T - x_g^*\| \right\} \tag{5.55}$$

$$\leq \frac{1}{2} \sup_{\|\gamma_T\|} \sup_{g \in \mathcal{F} : \kappa(f,g) = \frac{2\|\gamma_T\|}{\sqrt{T}}} \|x_f^* - x_g^*\| \|\gamma_T\|^{-1} \lambda(\|\gamma_T\|, \sigma) \tag{5.56}$$

$$\leq \frac{1}{2} \sup_{\tau} \omega_f \left(\frac{2\tau}{\sqrt{T}} \right) \tau^{-1} \lambda(\tau, \sigma). \tag{5.57}$$

We have the following bound derived from Donoho (1994)

$$\lambda(\tau, \sigma) \leq \tau \exp\left(-\frac{\tau^2}{4\sigma^2} \right), \tag{5.58}$$

which yields

$$R_T(f; \mathcal{F}) \leq \frac{1}{2} \sup_{\tau} \omega \left(\frac{2\tau}{\sqrt{T}} \right) \exp\left(-\frac{\tau^2}{4\sigma^2} \right). \tag{5.59}$$

To upper bound the last quantity, we write

$$\sup_{\tau} \omega \left(\frac{2\tau}{\sqrt{T}} \right) \exp\left(-\frac{\tau^2}{4\sigma^2} \right) \leq \max \left\{ \sup_{\tau \leq r} \psi(\tau), \sup_{r < \tau \leq \frac{1}{2}\epsilon_0 \sqrt{T}} \psi(\tau), \sup_{\tau > \frac{1}{2}\epsilon_0 \sqrt{T}} \psi(\tau) \right\} \tag{5.60}$$

for some $r > 0$, where $\psi(\tau) = \omega\left(\frac{2\tau}{\sqrt{T}} \right) \exp\left(-\frac{\tau^2}{4\sigma^2} \right)$. We bound the three terms separately

by

$$\sup_{\tau \leq r} \omega\left(\frac{2\tau}{\sqrt{T}}\right) \exp\left(-\frac{\tau^2}{4\sigma^2}\right) \leq \omega\left(\frac{2r}{\sqrt{T}}\right), \tag{5.61}$$

and

$$\sup_{r<\tau\leq \frac{1}{2}\epsilon_0\sqrt{T}} \omega\left(\frac{2\tau}{\sqrt{T}}\right) \exp\left(-\frac{\tau^2}{4\sigma^2}\right) \tag{5.62}$$

$$= \sup_{s\geq 1 \,\&\, \frac{2sr}{\sqrt{T}}\leq \epsilon_0} \omega\left(\frac{2sr}{\sqrt{T}}\right) \exp\left(-\frac{s^2r^2}{4\sigma^2}\right) \tag{5.63}$$

$$\leq \sup_{s\geq 1} s^\alpha \omega\left(\frac{2r}{\sqrt{T}}\right) \exp\left(-\frac{s^2r^2}{4\sigma^2}\right) \tag{5.64}$$

$$\leq \left(\frac{\sqrt{2\alpha}\sigma}{r}\right)^\alpha \omega\left(\frac{2r}{\sqrt{T}}\right) \tag{5.65}$$

since ω_f satisfies $\omega_f(c\epsilon) \leq c^\alpha \omega_f(\epsilon)$ for $c > 1$, $c\epsilon \leq \epsilon_0$ and some $\alpha > 0$, and

$$\sup_{\tau > \frac{1}{2}\epsilon_0\sqrt{T}} \omega\left(\frac{2\tau}{\sqrt{T}}\right) \exp\left(-\frac{\tau^2}{4\sigma^2}\right) \leq \operatorname{diam}(\mathcal{C}) \exp\left(-\frac{\epsilon_0^2 T}{16\sigma^2}\right) \tag{5.66}$$

Setting $r = \sigma/2$ and noting that $\omega_f(\epsilon) \geq \epsilon^\alpha \frac{\omega_f(\epsilon_0)}{\epsilon_0^\alpha}$, we have that there exists $T_0 > 0$ such that for all $T \geq T_0$

$$R_T(f; \mathcal{F}) \leq C\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \tag{5.67}$$

where $C = \frac{1}{2}\max\{1, (8\alpha)^{\frac{\alpha}{2}}\}$.

5.5.2 Proofs for superefficiency results

We begin by recalling the following results about properties of the subdifferential of a convex function f and its Fenchel conjugate

$$f^*(y) := \sup_x \left\{y^T x - f(x)\right\}, \tag{5.68}$$

including duality between the subdifferential sets ∂f and ∂f^*, increasing gradients, and continuous differentiability.

Lemma 5.5.2 (Hiriart-Urruty and Lemaréchal (1993))**.** *Let f be a closed convex function. Then*

$$x \in \partial f^*(y) \quad \textit{if and only if} \quad y \in \partial f(x). \tag{5.69}$$

Additionally, subgradient sets are increasing in the sense that

$$s_1 \in \partial f(x_1) \textit{ and } s_2 \in \partial f(x_2) \quad \textit{implies} \quad \langle s_1 - s_2, x_1 - x_2 \rangle \geq 0. \tag{5.70}$$

Lastly, if $f : \mathbb{R} \to \mathbb{R}$ is strictly convex on an interval $[x_l, x_r]$, then f^* is continuously differentiable on the interval $[\inf\{s : s \in \partial f(x_l)\}, \sup\{s : s \in \partial f(x_r)\}]$.

Moduli of continuity

Lemma 5.5.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be a subdifferentiable convex function. Define $f_\epsilon(x) = f(x) + \epsilon x$. Then*

$$\arg\min_x f_\epsilon(x) = \partial f^*(-\epsilon) \tag{5.71}$$

Moreover,

$$\operatorname{dist}(\partial f^*(0), \partial f^*(\epsilon)) \vee \operatorname{dist}(\partial f^*(0), \partial f^*(-\epsilon)) \leq \omega_f(\epsilon) \tag{5.72}$$

$$\omega_f(\epsilon) \leq \sup_x\{\operatorname{dist}(x, \partial f^*(0)) : x \in \partial f^*(\epsilon)\} \vee \sup_x\{\operatorname{dist}(x, \partial f^*(0)) : x \in \partial f^*(-\epsilon)\} \tag{5.73}$$

In particular, if $x_0 = \arg\min_x f(x)$ is unique and f is strictly convex in a neighborhood of x_0, then there exists an $\epsilon_0 > 0$ such that $\epsilon \leq \epsilon_0$ implies that

$$\omega_f(\epsilon) = \max\left\{|f^{*\prime}(\epsilon) - x_0|, |f^{*\prime}(-\epsilon) - x_0|\right\}. \tag{5.74}$$

Proof. Let $x_0 \in \arg\min_x f(x)$. Using Lemma 5.5.2, it is clear that $\arg\min_x f(x) = \partial f^*(0)$,

and more generally, that

$$\partial f^*(y) = \arg\max_x \left\{ y^T x - f(x) \right\} = \arg\min_x \left\{ f(x) - y^T x \right\}. \tag{5.75}$$

We begin by providing the lower bound on ω_f. For $\epsilon > 0$, define the function $f_\epsilon(x) = f(x) + \epsilon x$. Then certainly $\kappa(f, f_\epsilon) \leq \epsilon$. Moreover, we have

$$f_\epsilon^*(y) = \sup_x \{yx - f(x) - \epsilon x\} = \sup_x \{(y - \epsilon)x - f(x)\} = f^*(y - \epsilon), \tag{5.76}$$

so that $\arg\min_x f_\epsilon(x) = \partial f^*(-\epsilon)$. Noting that $x_0 \in \partial f^*(0)$ and that subgradients are increasing by Lemma 5.5.2, we have that

$$\arg\min_x f_\epsilon(x) = \partial f^*(-\epsilon) \leq \partial f^*(0) = \arg\min_x f(x). \tag{5.77}$$

That is, we have $\sup\{x_\epsilon \in \arg\min_x f_\epsilon(x)\} \leq \inf\{x_0 \in \arg\min_x f(x)\}$ and

$$\omega_f(\epsilon) \geq \inf \left\{ |s_\epsilon - s_0| : s_\epsilon \in \partial f^*(-\epsilon), s_0 \in \partial f^*(0) \right\}. \tag{5.78}$$

An identical argument with $f_{-\epsilon}$ gives the lower bound.

For the upper bound on the modulus of continuity, we note that if g is a convex function with $\kappa(f, g) \leq \epsilon$, and $x_g \in \arg\min_x g(x)$, then there must be some $s \in \partial f(x_g)$ with $\epsilon \geq s \geq -\epsilon$, because $0 \in \partial g(x_g)$, where we have used the definition of the Hausdorff distance. Now, for this particular s, by Lemma 5.5.2 we have that

$$x_g \in \partial f^*(s). \tag{5.79}$$

Again using the increasing behavior of subgradients, we obtain that

$$\inf \partial f^*(-\epsilon) \leq x_g \leq \sup \partial f^*(\epsilon), \tag{5.80}$$

which gives the claimed upper bound in the lemma once we recognize that $x_0 \in \partial f^*(0)$, and the definition of distance for ω_f is $d(f,g) = \inf\{|x_0 - x_g^\star| : x_0 \in \arg\min_x f(x), x_g^\star \in \arg\min_x g(x)\}$.

The final result, with the uniqueness, is an immediate consequence of the differentiability properties in Lemma 5.5.2. $\square$

Now we calculate bounds for a few example moduli of contiuity using Lemma 5.5.3. Roughly, we focus on non-pathological convex functions to allow us to give explicit calculations. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function satisfying $\partial f^*(0) = \arg\min_x f(x) = [x_l, x_r]$. In addition, assume that for $\delta > 0$, we have for some powers $k_l, k_r \geq 1$ and constants $\lambda_l > 0$ and $\lambda_r > 0$ that

$$f(x_l - \delta) = f(x_l) + \lambda_l \delta^{k_l} + o(\delta^{k_l}) \quad \text{and} \quad f(x_r + \delta) = f(x_r) + \lambda_r \delta^{k_r} + o(\delta^{k_r}). \tag{5.81}$$

That is, in a neighborhood of the optimal region, the function f grows like a polynomial. The condition (5.81) is not too restrictive, but does rule out functions such as $f(x) = e^{-\frac{1}{x^2}}$.

Lemma 5.5.4. *Let f satisfy the condition* (5.81)*. For any $c > 1$, there exists some $\epsilon_0 > 0$ such that for $\epsilon \in (0, \epsilon_0)$*

$$x_r + \left(\frac{\epsilon}{C\lambda_r k_r}\right)^{\frac{1}{k_r - 1}} \leq \inf \partial f^*(\epsilon) \leq \sup \partial f^*(\epsilon) \leq x_r + \left(\frac{C\epsilon}{\lambda_r}\right)^{\frac{1}{k_r - 1}} \tag{5.82a}$$

and

$$x_l - \left(\frac{\epsilon}{C\lambda_l k_l}\right)^{\frac{1}{k_l - 1}} \geq \sup \partial f^*(-\epsilon) \geq \inf \partial f^*(-\epsilon) \geq x_l - \left(\frac{C\epsilon}{\lambda_l}\right)^{\frac{1}{k_l - 1}}. \tag{5.82b}$$

Moreover, setting $k = \max\{k_r, k_l\}$ *and letting*

$$\lambda = \begin{cases} \lambda_l & \text{if } k_l > k_r, \\ \lambda_r & \text{if } k_r > k_l, \\ \max\{\lambda_r, \lambda_l\} & \text{otherwise,} \end{cases} \tag{5.83}$$

we have for all $\epsilon \in (0, \epsilon_0)$ *that*

$$\left(\frac{\epsilon}{C\lambda k}\right)^{\frac{1}{k-1}} \le \omega_f(\epsilon) \le \left(\frac{C\epsilon}{\lambda}\right)^{\frac{1}{k-1}}. \tag{5.84}$$

Proof. We focus on the right side bound (5.82a), as the proof of the left bound (5.82b) is similar. We also let the constant be $c = 2$ for simplicity.

For notational simplicity, let $\lambda = \lambda_r$ and $k = k_r$. By the fact that subgradients are increasing, we have for any $\delta > 0$ that

$$\inf \partial f(x_r + \delta) \ge \frac{f(x_r + \delta) - f(x_r)}{\delta} = \frac{\lambda\delta^k + o(\delta^k)}{\delta} = \lambda(1 - o_\delta(1))\delta^{k-1} \tag{5.85}$$

as $\delta \downarrow 0$. Similarly, $\delta > 0$ we have

$$\begin{aligned} \sup \partial f(x_r + \delta) &\le \frac{f(x_r + 2\delta) - f(x_r + \delta)}{\delta} = \frac{\lambda(2\delta)^k - \lambda\delta^k + o(\delta^k)}{\delta} \\ &= \frac{\lambda k\delta^{k-1}\delta + o(\delta^k)}{\delta} = (1 + o_\delta(1))\lambda k\delta^{k-1}. \end{aligned} \tag{5.86}$$

Combining inequalities (5.85) and (5.86), we thus see that there exists some $\delta_0 > 0$ such that for $\delta \in (0, \delta_0)$ we have

$$\frac{\lambda}{2}\delta^{k-1} \le \inf \partial f(x_r + \delta) \le \sup \partial f(x_r + \delta) \le 2\lambda k\delta^{k-1}. \tag{5.87}$$

Noting that $x_r + \delta \in \partial f^*(\epsilon)$ if and only if $\epsilon \in \partial f(x_r + \delta)$ by standard subgradient calculus

(recall Lemma 5.5.2), we solve for $\epsilon = \frac{\lambda}{2}\delta^{k-1}$ and $\epsilon = 2\lambda k\delta^{k-1}$ to attain inequality (5.82a). The bound (5.82b) is similar. □

Lemma 5.5.4 shows that, as $\epsilon \to 0$, we have $\omega_f(\epsilon) \asymp \epsilon^{\frac{1}{k-1}}$, where $k = \max\{k_r, k_l\}$. Finally, we show a type of continuity property with the modulus of continuity.

Lemma 5.5.5. *Assume that f has expansion (5.81), and that either (i) $k_r > k_l$ or (ii) $k_r \geq k_l$ and $\lambda_r \geq \lambda_l$. Define $g(x) = f(x) - \epsilon x$. Then for any constants $c < 1 < C$, we have*

$$\omega_g(c\epsilon) \leq (2C)^{\frac{1}{k_r-1}} \left(\frac{\epsilon}{\lambda_r}\right)^{\frac{1}{k_r-1}} \leq (2C^2)^{\frac{1}{k_r-1}} e\, \omega_f(\epsilon) \tag{5.88}$$

for all ϵ suitably close to 0.

Proof. We know by the increasing properties of the subgradient set and Lemma 5.5.3 that for any $c < 1$

$$\omega_g(c\epsilon) \leq \max\{\operatorname{dist}(\partial g^*(\epsilon), \partial g^*(0)), \operatorname{dist}(\partial g^*(-\epsilon), \partial g^*(0))\} \tag{5.89}$$

$$= \max\{\operatorname{dist}(\partial f^*(2\epsilon), \partial f^*(\epsilon)), \operatorname{dist}(\partial f^*(0), \partial f^*(\epsilon))\}, \tag{5.90}$$

where we have used that $g^*(y) = \sup_x\{(y+\epsilon)x - f(x)\} = f^*(y+\epsilon)$. For small enough $\epsilon > 0$, we have by Lemma 5.5.4 that

$$\sup \partial f^*(2\epsilon) \leq \left(\frac{2C\epsilon}{\lambda_r}\right)^{\frac{1}{k_r-1}}, \tag{5.91}$$

which gives the first inequality.

For the second inequality, we use that $\omega_f(\epsilon) \geq (\epsilon/(C\lambda_r k_r))^{\frac{1}{k_r-1}}$ to obtain

$$\left(\frac{2C\epsilon}{\lambda_r}\right)^{\frac{1}{k_r-1}} = k_r^{\frac{1}{k_r-1}}(2C^2)^{\frac{1}{k_r-1}} \left(\frac{\epsilon}{C\lambda_r k_r}\right)^{\frac{1}{k_r-1}} \leq k_r^{\frac{1}{k_r-1}}(2C^2)^{\frac{1}{k_r-1}}\omega_f(\epsilon) \leq e(2C^2)^{\frac{1}{k_r-1}}\omega_f(\epsilon) \tag{5.92}$$

as desired. □

Superefficiency

For distributions P_0 and P_1 define the χ-divergence by

$$D_\chi\left(P_1 \| P_0\right) := \int \left(\frac{dP_1}{dP_0} - 1\right) dP_1 = \int \left(\frac{dP_1}{dP_0}\right) dP_1 - 1. \tag{5.93}$$

The following lemma, which is a stronger version of a result due to Brown and Low (1996b), gives a result on superefficiency.

Lemma 5.5.6. *Let $\widehat{x}$ be any function of a sample ξ, and let X_0 and X_1 be compact convex sets (associated with distributions P_0 and P_1). Let $\operatorname{dist}(x, X) = \inf_{y \in X} |y - x|$ and $\operatorname{dist}(X_0, X_1) = \inf_{x_0 \in X_0} \operatorname{dist}(x_0, X_1)$. Then*

$$\mathbb{E}_{P_1}[\operatorname{dist}(\widehat{x}, X_1)] \geq \left[\operatorname{dist}(X_0, X_1) - \sqrt{\mathbb{E}_{P_0}[\operatorname{dist}(\widehat{x}, X_0)^2](D_\chi\left(P_1 \| P_0\right) + 1)}\right]_+ \tag{5.94}$$

$$\geq \operatorname{dist}(X_0, X_1)\left[1 - \frac{\sqrt{\mathbb{E}_{P_0}[\operatorname{dist}(\widehat{x}, X_0)^2](D_\chi\left(P_1 \| P_0\right) + 1)}}{\operatorname{dist}(X_0, X_1)}\right]_+. \tag{5.95}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}_{P_1}[\operatorname{dist}(\widehat{x}, X_1)] &\overset{(i)}{\geq} \operatorname{dist}(X_0, X_1) - \mathbb{E}_{P_1}[\operatorname{dist}(\widehat{x}, X_0)] \\
&\overset{(ii)}{\geq} \operatorname{dist}(X_0, X_1) - \sqrt{\mathbb{E}_{P_0}[\operatorname{dist}(\widehat{x}, X_0)^2] \cdot \int \left(\frac{dP_1}{dP_0}\right) dP_1} \\
&= \operatorname{dist}(X_0, X_1) - \sqrt{\mathbb{E}_{P_0}[\operatorname{dist}(\widehat{x}, X_0)^2](D_\chi\left(P_1 \| P_0\right) + 1)}
\end{aligned}$$

where inequality (i) uses the triangle inequality and inequality (ii) uses Cauchy-Schwarz. □

We now present two lemmas on χ-divergence that will be useful. The first is a standard algebraic calculation.

Lemma 5.5.7. *Let P_0 and P_1 be normal distributions with means μ_0 and μ_1, respectively,*

and variances σ^2*. Then*

$$D_\chi\left(P_0 \| P_1\right) = D_\chi\left(P_1 \| P_0\right) = \exp\left(\frac{\left(\mu_0 - \mu_1\right)^2}{\sigma^2}\right) - 1. \tag{5.96}$$

For the second lemma, we assume that $\widehat{x}$ is constructed based on noisy subgradient information from a subgradient oracle, which upon being queried at a point x, returns

$$f'(x) + \varepsilon, \quad \text{where } \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{and} \quad f'(x) = \operatorname*{arg\,min}_{s \in \partial f(x)} \{|s|\}. \tag{5.97}$$

The latter condition simply specifies the subgradient the oracle chooses; any specified choice of subgradient is sufficient. Because $\partial f(x)$ is a closed convex set for any x, we see that if f and g are convex functions with $\kappa(f, g) \le \epsilon$, then $|f'(x) - g'(x)| \le \epsilon$ with the construction (5.97) of subgradient oracle.

Lemma 5.5.8. *Let the subgradient oracle be given by* (5.97)*, and let* P_f^T *and* P_g^T *be the distributions (respectively) of the observed stochastic sub-gradients*

$$s_i = f'(x_i) + \varepsilon_i \quad \textit{or} \quad s_i = g'(x_i) + \varepsilon_i, \tag{5.98}$$

where x_i *is a measurable function of an independent noise variable* ξ_0 *and the preceding sequence of stochastic gradients* $\{s_1, \ldots, s_{i-1}\}$*. Let* $\kappa(f, g) \le \epsilon$*. Then*

$$D_\chi\left(P_f^T \| P_g^T\right) \le \exp\left(\frac{T\epsilon^2}{\sigma^2}\right) - 1. \tag{5.99}$$

Proof. Let s_i be the ith observed stochastic subgradient in the sequence, and let the σ-field

of the observed sequence through time i be $\mathcal{F}_i = \sigma(\xi_0, s_1, \ldots, s_i)$. Then we have

$$D_\chi\left(P_f^T \| P_g^T\right) + 1 = \int \frac{dP_f^T(s_{1:n})}{dP_g^T(s_{1:n})} dP_f^T(s_{1:n}) \tag{5.100}$$

$$= \int \prod_{i=1}^{T} \left[\frac{dP_f(s_i \mid s_{1:i-1})}{dP_g(s_i \mid s_{1:i-1})} dP_f(s_i \mid s_{1:i-1})\right] \tag{5.101}$$

$$= \mathbb{E}\left[\prod_{i=1}^{T} \mathbb{E}_{P_f}\left[\frac{dP_f(S_i \mid \mathcal{F}_{i-1})}{dP_g(S_i \mid \mathcal{F}_{i-1})} \mid \mathcal{F}_{i-1}\right]\right]. \tag{5.102}$$

By the measurability assumption on x_i, that is, $x_i \in \mathcal{F}_{i-1}$, the inner expectation is simply one plus the χ^2 distance between two distributions $\mathcal{N}(f'(x_i), \sigma^2)$ and $\mathcal{N}(g'(x_i), \sigma^2)$, which we know satisfies

$$\mathbb{E}_{P_f}\left[\frac{dP_f(S_i \mid \mathcal{F}_{i-1})}{dP_g(S_i \mid \mathcal{F}_{i-1})} \mid \mathcal{F}_{i-1}\right] = \exp\left(\frac{(f'(x_i) - g'(x_i))^2}{\sigma^2}\right) \leq \exp\left(\frac{\epsilon^2}{\sigma^2}\right). \tag{5.103}$$

Taking the product over all T terms yields the lemma. □

Lemma 5.5.9. *Let f be a closed convex function. Define the function*

$$\begin{aligned} H(\epsilon) &:= \inf\{|x - x_0| : x \in \partial f^*(\epsilon), x_0 \in \partial f^*(0)\} \vee \inf\{|x - x_0| : x \in \partial f^*(-\epsilon), x_0 \in \partial f^*(0)\} \\ &= \operatorname{dist}(\partial f^*(\epsilon), \partial f^*(0)) \vee \operatorname{dist}(\partial f^*(-\epsilon), \partial f^*(0)). \end{aligned} \tag{5.104}$$

For any $0 \leq c_l < 1$ and $1 < c_u < \infty$,

$$\omega_f(c_u \epsilon) \geq H(\epsilon) \geq \omega_f(c_l \epsilon). \tag{5.105}$$

Proposition 5.5.10. *Define H to be the function* (5.104) *and assume additionally that $\delta < \sqrt{\frac{1}{8e}}$. If $\widehat{x}$ is any estimator such that*

$$\sqrt{\mathbb{E}_{P_f^T}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_f^*)^2\right]} \leq \delta \omega_f(\sigma/\sqrt{T}), \tag{5.106}$$

then taking $f_1(x) = f(x) + \sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}x$ *and* $f_{-1}(x) = f(x) - \sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}x$*, we have*

$$\max_{g \in \{f_1, f_{-1}\}} \mathbb{E}_{P_g^T}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_g^*)\right] \geq \sup_{0<c<\log \frac{1}{8\delta^2}} \omega_f\left(\sqrt{\frac{c\sigma^2}{T}}\right)\left(1 - \frac{\omega_f(\sigma/\sqrt{T})}{2\sqrt{2}\omega_f(\sqrt{c\sigma^2/T})}\right) \tag{5.107}$$

$$\geq \frac{4-\sqrt{2}}{4} H\left(\sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}\right). \tag{5.108}$$

Proof. Without loss of generality, we assume that $0 \in \arg\min_x f(x) = \partial f^*(0)$, and set $x_0 = 0$ for simplicity in the derivation. For any $\epsilon \in \mathbb{R}$, we may construct the function $f_\epsilon(x) = f(x) + \epsilon x$. Lemma 5.5.6 and Lemma 5.5.8 thus yield that that for $\mathcal{X}_\epsilon = \arg\min_x f_\epsilon(x)$, we have

$$\mathbb{E}_{P_{f_\epsilon}^T}[\operatorname{dist}(\widehat{x}, \mathcal{X}_\epsilon)] \geq \operatorname{dist}(\partial f^*(-\epsilon), \partial f^*(0))\left[1 - \frac{\omega_f(\sigma/\sqrt{T})\sqrt{\delta \exp(\frac{T\epsilon^2}{\sigma^2})}}{\operatorname{dist}(\partial f^*(-\epsilon), \partial f^*(0))}\right]_+ \tag{5.109}$$

and

$$\mathbb{E}_{P_{f_{-\epsilon}}^T}[\operatorname{dist}(\widehat{x}, \mathcal{X}_{-\epsilon})] \geq \operatorname{dist}(\partial f^*(\epsilon), \partial f^*(0))\left[1 - \frac{\omega_f(\sigma/\sqrt{T})\sqrt{\delta \exp(\frac{T\epsilon^2}{\sigma^2})}}{\operatorname{dist}(\partial f^*(\epsilon), \partial f^*(0))}\right]_+. \tag{5.110}$$

In particular, with $H(\epsilon) = \operatorname{dist}(\partial f^*(\epsilon), \partial f^*(0)) \vee \operatorname{dist}(\partial f^*(\epsilon), \partial f^*(0))$, we have

$$\max_{g \in f_\epsilon, f_{-\epsilon}} \mathbb{E}_{P_g^T}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_g^*)\right] \geq H(\epsilon)\left[1 - \frac{\omega_f(\sigma/\sqrt{T})\sqrt{\delta \exp(\frac{n\epsilon^2}{\sigma^2})}}{H(\epsilon)}\right]_+. \tag{5.111}$$

Take $\epsilon^2 = \frac{\sigma^2}{T} \log \frac{1}{8\delta^2}$ to obtain

$$\max_{g \in f_\epsilon, f_{-\epsilon}} \mathbb{E}_{P_g^T}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_g^*)\right] \geq H\left(\sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}\right)\left[1 - \frac{\omega_f(\sigma/\sqrt{T})}{2\sqrt{2}H(\sigma \log^{\frac{1}{2}} \frac{1}{8\delta^2}/\sqrt{T})}\right]_+. \tag{5.112}$$

Notably, by Lemma 5.5.3, our w.l.o.g. assumption and the fact that subgradients are increas-

ing, we have that for any constant $(\log \frac{1}{8\delta^2})^{-\frac{1}{2}} \leq c < 1$ that

$$\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \leq \omega_f\left(c\sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}\right) \tag{5.113}$$

$$\leq \sup\left\{\operatorname{dist}(x, X_0) : x \in \partial f^*\left(c\frac{\sigma \log^{\frac{1}{2}} \frac{1}{8\delta^2}}{\sqrt{T}}\right)\right\} \vee \sup\left\{\operatorname{dist}(x, X_0) : x \in \partial f^*\left(-c\frac{\sigma \log^{\frac{1}{2}} \frac{1}{8\delta^2}}{\sqrt{T}}\right)\right\} \tag{5.114}$$

$$\leq \sup\left\{\operatorname{dist}(x, X_0) : x \in \partial f^*\left(\frac{\sigma \log^{\frac{1}{2}} \frac{1}{8\delta^2}}{\sqrt{T}}\right)\right\} \vee \sup\left\{\operatorname{dist}(x, X_0) : x \in \partial f^*\left(-\frac{\sigma \log^{\frac{1}{2}} \frac{1}{8\delta^2}}{\sqrt{T}}\right)\right\} \tag{5.115}$$

$$= H\left(\frac{\sigma \log^{\frac{1}{2}} \frac{1}{8\delta^2}}{\sqrt{T}}\right). \tag{5.116}$$

In particular, we have the lower bound

$$\max_{g \in f_\epsilon, f_{-\epsilon}} \mathbb{E}_{P_g^T}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_g^*)\right] \geq H\left(\sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}\right)\frac{4-\sqrt{2}}{4}. \tag{5.117}$$

This is the desired result. □

Proposition 5.5.10 is a basic result on superefficiency that we may specialize to obtain more concrete results. We would like give a result that holds when f^* is differentiable in a neighborhood of 0, which is equivalent to f being strictly convex in a neighborhood of $x_0 = \arg\min_x f(x)$, by Lemma 5.5.2. This would mean that the function H defined in Proposition 5.5.10 satisfies

$$H(\epsilon) = \max\{|f^{*\prime}(\epsilon) - x_0|, |f^{*\prime}(-\epsilon) - x_0|\} = \omega_f(\epsilon) \tag{5.118}$$

for all small enough $\epsilon > 0$. In this setting, we obtain

Corollary 5.5.11. *Let the conditions of Proposition 5.5.10 hold, and let f be strictly convex*

in a neighborhood of $x_0 = \arg\min_x f(x)$. *Assume that* $\widehat{x}$ *is any estimator satisfying*

$$\sqrt{\mathbb{E}_{P_f^T}\left[(\widehat{x} - x_0)^2\right]} \leq \delta\omega_f(\sigma/\sqrt{T}), \tag{5.119}$$

where $\delta < \sqrt{\frac{1}{8e}}$. *Define* $f_{\pm 1}(x) = f(x) \pm \sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}x$. *Then for large enough* T,

$$\max_{g \in \{f_1, f_{-1}\}} \mathbb{E}_{P_g^T}|\widehat{x} - x_g^\star| \geq \frac{4 - \sqrt{2}}{4}\omega_f\left(\sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}\right). \tag{5.120}$$

This corollary has a striking weakness, however—the right hand side depends on ω_f, rather than ω_g, which is what we would prefer. We can, however, state a simpler result that is achievable.

Corollary 5.5.12. *Let* f *be any convex function satisfying the asymptotic expansion* (5.81) *around its optimum. Suppose that* $\widehat{x}$ *is any estimator such that*

$$\sqrt{\mathbb{E}_{P_f^T}[\mathrm{dist}(\widehat{x}, \mathcal{X}_f^*)^2]} \leq \delta\omega_f\left(\frac{\sigma}{\sqrt{T}}\right), \tag{5.121}$$

where $\delta < \sqrt{\frac{1}{8e}}$. *Define* $g_{-1}(x) = f(x) - \epsilon_T x$ *and* $g_1(x) = f(x) + \epsilon_T x$, *where* $\epsilon_T = \sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}$, *and let* $k = k_r \vee k_l$. *Let* $C > 1$ *and* $0 < c < 1$ *be otherwise arbitrary numerical constants. Then for one of* $g \in \{g_{-1}, g_1\}$, *there exists* T_0 *such that* $T \geq T_0$ *implies*

$$\mathbb{E}_{P_g}\left[\mathrm{dist}(\widehat{x}, \mathcal{X}_g^*)\right] \geq \frac{4 - \sqrt{2}}{4(2C^2)^{\frac{1}{k-1}}e}\,\omega_g\left(c\sqrt{\frac{\sigma^2 \log \frac{1}{8\delta^2}}{T}}\right). \tag{5.122}$$

Proof. Without loss of generality, we assume that $k_r \geq k_l$, and if $k_l = k_r$ then $\lambda_r \geq \lambda_l$. By inspection of the proof of Proposition 5.5.10, we have that

$$\mathbb{E}_{P_{g_{-1}}^T}[\mathrm{dist}(\widehat{x}, \partial g_{-1}^*(0))] \geq \frac{1}{2}\,\mathrm{dist}(\partial f^*(\epsilon_T), \partial f^*(0)). \tag{5.123}$$

Moreover, we know that for suitably large n, we have by Lemma 5.5.4

$$\begin{aligned}\operatorname{dist}(\partial f^*(\epsilon_T), \partial f^*(0)) &= \operatorname{dist}(\partial f^*(\epsilon_T), \partial f^*(0)) \vee \operatorname{dist}(\partial f^*(-\epsilon_T), \partial f^*(0)) && (5.124)\\ &\geq \omega_f(c\epsilon_T) && (5.125)\end{aligned}$$

for any $c < 1$. Then Lemma 5.5.5 implies that for any $C > 1$, there exists T_0 such that $T \geq T_0$ implies

$$\omega_f(c\epsilon_T) \geq \frac{1}{(2C^2)^{\frac{1}{k-1}} e} \omega_{g-1}(c^2 \epsilon_T). \tag{5.126}$$

This gives the desired result. □

As an immediate consequence of Corollary 5.5.12, we see that if there exists any sequence $\delta_T \to 0$ with $\liminf_T e^T \delta_T = \infty$ such that

$$\sqrt{\mathbb{E}_{P_f}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_f^*)^2\right]} \leq \delta_T \omega_f\left(\frac{\sigma}{\sqrt{T}}\right), \tag{5.127}$$

then there exists a sequence of convex functions g_T, with $\kappa(f, g_T) \to 0$, such that

$$\liminf_T \frac{\mathbb{E}_{P_{g_T}}\left[\operatorname{dist}(\widehat{x}, \mathcal{X}_{g_T})\right]}{\omega_{g_T}\left(\sqrt{\frac{\sigma^2 \log {\delta_T}^{-1}}{T}}\right)} > 0. \tag{5.128}$$

5.5.3 Algorithm

Proof of Proposition 5.3.1

First, by the monotonicity of the derivative f', note that the interval $\mathcal{I}_\delta$ is such that $x \in \mathcal{I}_\delta$ holds if and only if $|f'(x)| < C_\delta/\sqrt{T_0}$. Now suppose that at round e, $(a_e, b_e) \cap \mathcal{I}_\delta \neq \emptyset$. For the next round, if $x_e = (a_e + b_e)/2 \in \mathcal{I}_\delta$, then $(a_{e+1}, b_{e+1}) \cap \mathcal{I}_\delta \neq \emptyset$. Otherwise, if $x_e \notin \mathcal{I}_\delta$, we know that $|f'(x_e)| \geq C_\delta/\sqrt{T_0}$, and without loss of generality, we assume that it

is positive. Then, we have

$$\mathbb{P}\left((a_{e+1}, b_{e+1}) \cap \mathcal{I}_\delta \neq \emptyset\right) = \mathbb{P}\left(\mathcal{N}\left(f'(x_e), \frac{\sigma^2}{T_0}\right) < 0\right) = \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\sqrt{T_0} f'(x_e)}{\sigma}\right) \tag{5.129}$$

$$\leq \mathbb{P}\left(\mathcal{N}(0,1) > \frac{C_\delta}{\sigma}\right) \leq \frac{\sigma}{C_\delta \sqrt{2\pi}} \exp\left(-\frac{C_\delta^2}{2\sigma^2}\right) \tag{5.130}$$

Therefore,

$$\mathbb{P}\left((a_{e+1}, b_{e+1}) \cap \mathcal{I}_\delta \neq \emptyset \middle| (a_e, b_e) \cap \mathcal{I}_\delta \neq \emptyset\right) \geq 1 - \frac{\sigma}{C_\delta \sqrt{2\pi}} \exp\left(-\frac{C_\delta^2}{2\sigma^2}\right) \tag{5.131}$$

It then follows that

$$\mathbb{P}\left((a_E, b_E) \cap \mathcal{I}_\delta \neq \emptyset\right) = \mathbb{P}\left((a_e, b_e) \cap \mathcal{I}_\delta \neq \emptyset \text{ for } e = 1, \ldots, E\right) \tag{5.132}$$

$$= \prod_{e=0}^{E-1} \mathbb{P}\left((a_{e+1}, b_{e+1}) \cap \mathcal{I}_\delta \neq \emptyset \middle| (a_e, b_e) \cap \mathcal{I}_\delta \neq \emptyset\right) \tag{5.133}$$

$$\geq \left(1 - \frac{\sigma}{C_\delta \sqrt{2\pi}} \exp\left(-\frac{C_\delta^2}{2\sigma^2}\right)\right)^E \tag{5.134}$$

$$\geq 1 - \frac{E\sigma}{C_\delta \sqrt{2\pi}} \exp\left(-\frac{C_\delta^2}{2\sigma^2}\right) \tag{5.135}$$

$$\geq 1 - \delta \tag{5.136}$$

by the choice of C_δ.

Proof of Corollary 5.3.2

By the polynomial growth condition, we have for $T > \sigma^2/\epsilon_0$,

$$\omega_f(\epsilon_0) \leq \left(\frac{\epsilon_0 \sqrt{T}}{\sigma}\right)^\alpha \omega_f\left(\frac{\sigma}{\sqrt{T}}\right). \tag{5.137}$$

Since $r = \frac{1}{2}\alpha_0 \geq \frac{1}{2}\alpha$ and $E = \lfloor r \log T \rfloor$,

$$2^{-E}(b_0 - a_0) \leq 2(b_0 - a_0)T^{-r} \leq 2(b_0 - a_0)T^{-\frac{1}{2}\alpha} \leq \frac{2(b_0 - a_0)\epsilon_0^{\alpha}}{\omega_f(\epsilon_0)\sigma^{\alpha}}\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \tag{5.138}$$

By the expression we obtained in Example 5.2.1,

$$\sup\{\inf_{x \in \mathcal{X}_f^*} |x - y| : y \in \mathcal{I}_\delta\} \tag{5.139}$$

$$= \omega_f\left(\frac{C_\delta}{\sqrt{T_0}}\right) \leq \left(\sqrt{2r\left(\log(r \log T) + \log\frac{1}{\delta}\right)\log T}\right)^{\alpha} \omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \tag{5.140}$$

for T large enough. Therefore, we obtain that there exist $T' > 0$ such that for $T > T'$,

$$\inf_{x \in \mathcal{X}_f^*} |x_E - x| \leq \tilde{C}\omega_f\left(\frac{1}{\sqrt{T}}\right) \tag{5.141}$$

where

$$\tilde{C} = \frac{2(b_0 - a_0)\epsilon_0^{\alpha}}{\omega_f(\epsilon_0)\sigma^{\alpha}} + \left(\sqrt{2r\left(\log(r \log T) + \log\frac{1}{\delta}\right)\log T}\right)^{\alpha}. \tag{5.142}$$

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

We have considered two variants of traditional minimax theory to accommodate modern settings of data analysis tasks, and to provide more realistic and more customized evaluations of the hardness of the statistical tasks.

The first variant we considered is a set of computationally constrained minimax risks, in order to address the computational issue present in statistical estimation problems. Such constrained forms of statistical minimax risks, when possible to be calculated, quantify the tradeoff between statistical accuracy and computational efficiency, and set up guidelines for designing statistical procedures under certain computational budgets. As an illustrating example for such constrained forms of statistical minimax risk, we studied the problem of nonparametric estimation with storage constraints. We showed that the convergence rate and leading constant can be sharply characterized for the case of Sobolev spaces, and identifies three regimes where statistical error is primarily due to estimation or quantization, or both of them.

In addition to what has already been mentioned in Chapter 3, there are plenty of other problems which remain to be solved and understood in this area. For example, it is of interest and importance to understand what roles storage plays for machine learning procedures with a large amount of parameters, such as tree-based models, ensembles, and deep neural network. How much does performance degrade, if at all, when the fitted model has to be compressed?

The examples described above are concerned with space constraints. Can we say anything meaningful about time constraints? A lower bound on the running time of an algorithm is the time required to read the data, or the number of "queries" of the data values. Suppose an algorithm is judicious in how it selects which data values to read. What is the minimax optimal number of values required to achieve a given level or risk? One future direction is to study minimax risks with such computational time constraints measured by the number

of queries allowed.

The second variant of minimax theory we considered is a localized form formulated by examining the hardest two-point problem. We showed that this formulation gives a meaningful benchmark in the setting of convex optimization. The benchmark quantifies the hardness of optimizing a particular function, without putting it into a rich class of functions. It essentially ties the hardness to a geometric quantity of each particular function, a computational analogue of modulus of continuity. Two properties, in particular, make the benchmark interesting–the superefficiency and the achievability.

There are some interesting questions left unsolved for the case of convex optimization, as mentioned in Chapter 5. In addition to those, it would be interesting to consider such formulation applied on other problems where the hardness of each instance is known to be different. Moreover, formulations of the minimax risks other than the hardest two-point is also worth consideration. For example, one could consider a local ball surrounding the target function, instead of a whole general parameter space; the choices of the metric and the radius are then crucial in order for the formulation to be interesting and meaningful.

To sum up, we view the study of the variants of minimax theory as theoretic problems of fundamental interest, helping researchers and practitioners understand basic limits of statistical tasks, and providing useful guidelines in designing efficient and adaptive algorithms and methods.

REFERENCES

Ahlswede, R. and Burnashev, M. (1990). On minimax estimation in the presence of side information about remote data. *The Annals of Statistics*, 18(1):141–171.

Ahlswede, R. and Csiszár, I. (1986). Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, 32:533–542.

Berger, T., Zhang, Z., and Viswanathan, H. (1996). The CEO problem. *IEEE Transactions on Information Theory*, 42(3):887–902.

Berthet, Q. and Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066.

Brown, L. D. and Low, M. G. (1996a). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398.

Brown, L. D. and Low, M. G. (1996b). A constrained risk inequality with applications to nonparametric functional estimation. *Annals of Statistics*, 24(6):2524–2535.

Bruer, J. J., Tropp, J. A., Cevher, V., and Becker, S. (2014). Time–data tradeoffs by aggressive smoothing. In *Advances in Neural Information Processing Systems*, pages 1664–1672.

Cai, T. and Low, M. (2015). A framework for estimation of convex functions. *Statistica Sinica*, pages 423–456.

Castro, R. M. and Nowak, R. D. (2008). Minimax bounds for active learning. *Information Theory, IEEE Transactions on*, 54(5):2339–2353.

Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190.

Chattamvelli, R. and Jones, M. (1995). Recurrence relations for noncentral density, distribution functions and inverse moments. *Journal of Statistical Computation and Simulation*, 52(3):289–299.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

Donoho, D. and Liu, R. C. (1987). Geometrizing rates of convergence, I. Technical report, University of California, Berkeley. Department of Statistics, Technical Report 137.

Donoho, D. and Liu, R. C. (1991). Geometrizing rates of convergence, II. *Annals of Statistics*, 19:633–667.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270.

Donoho, D. L. (2000). Wald lecture I: Counting bits with Kolmogorov and Shannon.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369.

Draper, S. C. and Wornell, G. W. (2004). Side information aware coding strategies for sensor networks. *IEEE Journal on Selected Areas in Communications*, 22(6):966–976.

Duchi, J., Wainwright, M. J., and Jordan, M. I. (2013). Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems*, pages 1529–1537.

Galal, S. and Horowitz, M. (2011). Energy-efficient floating-point unit design. *IEEE Trans. Computers*, 60(7):913–922.

Gallager, R. G. (1968). *Information Theory and Reliable Communication*. John Wiley & Sons.

Gao, C., Ma, Z., and Zhou, H. H. (2014). Sparse cca: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*.

Garg, A., Ma, T., and Nguyen, H. (2014). On communication cost of distributed statistical estimation and dimensionality. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 2726–2734. Curran Associates, Inc.

Han, T. S. and Amari, S. (1998). Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324.

Hiriart-Urruty, J. and Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms I & II*. Springer, New York.

Iouditski, A. and Nesterov, Y. (2014). Primal-dual subgradient methods for minimizing uniformly convex functions. arXiv:1401.1792.

Johnstone, I. (2015). Gaussian estimation: sequence and wavelet models. Unpublished manuscript.

Johnstone, I. M. and Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*.

Karp, R. M. and Kleinberg, R. (2007). Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 881–890. Society for Industrial and Applied Mathematics.

Lucic, M., Ohannessian, M. I., Karbasi, A., and Krause, A. (2015). Tradeoffs for space, time, data and risk in unsupervised learning. In *AISTATS*.

Ma, Z., Wu, Y., et al. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116.

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 451–459.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons.

Nussbaum, M. (1999). Minimax risk: Pinsker bound. *Encyclopedia of Statistical Sciences*, 3:451–460.

Nussbaum, M. et al. (1985). Spline smoothing in regression models and asymptotic efficiency in l_2. *The Annals of Statistics*, 13(3):984–997.

Pinsker, M. S. (1980). Optimal filtering of square-integrable signals in gaussian noise. *Problemy Peredachi Informatsii*, 16(2):52–68.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Raginsky, M. (2007). Learning from compressed observations. In *Proceedings of the IEEE Information Theory Workshop*. IEEE.

Ramdas, A. and Singh, A. (2013a). Algorithmic connections between active learning and stochastic convex optimization. In *Algorithmic Learning Theory*, pages 339–353. Springer.

Ramdas, A. and Singh, A. (2013b). Optimal rates for stochastic convex optimization under Tsybakov noise condition. In *Proceedings of The 30th International Conference on Machine Learning*, pages 365–373.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Report 781, Cornell University Operations Research and Industrial Engineering.

Sakrison, D. J. (1968). A geometric treatment of the source encoding of a Gaussian random variable. *IEEE Transactions on Information Theory*, 14(3):481–486.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.

Venkataramanan, R., Sarkar, T., and Tatikonda, S. (2013). Lossy compression via sparse linear regression: Computationally efficient encoding and decoding. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1182–1186. IEEE.

Wainwright, M. J. (2014). Constrained forms of statistical minimax: Computation, communication and privacy. In *Proceedings of the International Congress of Mathematicians*.

Wang, T., Berthet, Q., and Samworth, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. *arXiv preprint arXiv:1408.5369*.

Zhang, Y., Duchi, J., Jordan, M., and Wainwright, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918*.

Zhang, Z. and Berger, T. (1988). Estimation via compressed information. *IEEE Transactions on Information Theory*, 34(2):198–211.

Zhou, S., Lafferty, J. D., and Wasserman, L. A. (2009). Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866.

Zhu, Y. and Lafferty, J. (2014). Quantized estimation of Gaussian sequence models in Euclidean balls. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 3662–3670. Curran Associates, Inc.

Zhu, Y. and Lafferty, J. (2015). Quantized nonparametric estimation. *arXiv preprint arXiv:1503.07368*.