Computational Trade-offs in Statistical Learning

by

Alekh Agarwal

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter L. Bartlett (Co-Chair)
Professor Martin J. Wainwright (Co-chair)
Professor Michael I. Jordan
Professor Andrew Lim

Spring 2012

Computational Trade-offs in Statistical Learning

Abstract

Computational Trade-offs in Statistical Learning

by

Alekh Agarwal

Doctor of Philosophy in Computer Science

and the Designated Emphasis

in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Peter L. Bartlett (Co-Chair)
Professor Martin J. Wainwright (Co-chair)

The last several years have seen the emergence of datasets of an unprecedented scale, and solving various inference problems on such datasets has received much focus in the modern machine learning and statistics literature. Understanding the statistical and computational aspects of learning from these large and high-dimensional datasets is the key focus of this thesis. One source of such data is the virtually unbounded amounts of unstructured information on the internet, which can often be compiled into large datasets for training machine learning algorithms using automated techniques, computer vision, natural language tasks, and web search and ranking being prime examples. Computational biology, computational astronomy and collaborative filtering are some other examples of problem domains where vast amounts of training data for learning algorithms are often readily available.

It might be expected that the large number of samples make the statistical problem easier, and subsampling a smaller dataset should be an effective computational strategy. However, this is not always the case. With access to more and more data, the goal is often to understand higher order dependencies in the data, leading to an explosion in the number of parameters to be estimated. Indeed a large body of literature in modern machine learning and statistics is focusing on problems where the number of parameters grows with, and is often larger than the number of samples.

While the number of training samples and parameters seems to grow almost unboundedly, computation is struggling to keep up. The last few years have seen a clear tapering off in the rate of growth of computational power on individual cores. As a result, the algorithmic focus in machine learning has shifted to online and stochastic optimization algorithms, budgeted

learning algorithms and parallel and distributed algorithms fit for multicore, cluster or cloud environments.

Though it might seem that computation is the main bottleneck for such massive data problems, it is not appropriate to consider these problems as purely computational and ignore the underlying statistical structure. In fact, in many cases the statistical nature of the data can make the computational problem easier, and make the goals of computation simpler than needed without this structure. As a result, a fundamental question in modern statistics and machine learning is to understand the error of statistical estimators, as a function of the *sample size*, *number of parameters* and the *computational budget* available.

The goal of this thesis is to study these trade-offs between the computational and statistical aspects of learning problems. This line of research results in several natural questions, some of which are partially addressed in this thesis and others present interesting challenges for future work.

To my parents

Contents

Acknowledgments

The five years of my Ph. D. were an extremely enjoyable experience, largely due to my advisors and peers both within Berkeley and throughout the academic research community, who I had a chance to interact with and learn from during this period. I am grateful to be able to acknowledge their invaluable contributions here.

The quality of one's graduate school experience is greatly influenced by the thesis advisors. In this respect, I was extremely fortunate to have the mentorship and support of Peter Bartlett and Martin Wainwright. When I started as a graduate student at Berkeley, I was already in awe of Peter's research and keen on working with him. However, any intimidation I might have felt due to his academic reputation was quickly dispelled due to his extremely friendly and calm demeanor. The thing that always struck me about Peter was his willingness and patience to give me complete freedom of research topics, and hear any ideas I might have, even as a starting graduate student. It was in a large part to this academic freedom and encouragement to explore from him that I was able to work across a large range of topics during my time at Berkeley. It never ceases to amaze me how I can approach Peter with questions about extremely diverse problems, and he usually has important insights to offer. I can recall many occasions when I went into a meeting feeling I was up against a wall, and came out of it with a concrete direction in my mind. I was also very fortunate to have Peter's only offering of CS281B during my time at Berkeley, which provided crucial foundational material for much of my research.

While I was already hoping to work with Peter when I started at Berkeley, my research with Martin was an outcome of taking his class on Theoretical Statistics in my first year. Indeed it was Martin's clarity of teaching as well as his very approachable personality that gave me the courage to pitch a research problem to him in my first semester at Berkeley. While the nature of problems that I have worked with Martin on has spanned many research areas, the experience of working with and learning from him has always been extremely enjoyable. Martin has an uncanny knack for giving just the right amount of guidance, which is particularly valuable during the early years. Throughout my Ph. D., I have learned a great deal from him not just in technical matters, but also other aspects of research such as technical writing and presentation skills. Not only is Martin a great mentor, but I think every member in the group finds it easy to think of him also as a peer—indeed the Jamón lecture at NIPS 2011 was a testament to this.

When I arrived at Berkeley, I already had some undergraduate research experience and it was a delight to have the mentorship of Prof. Soumen Chakrabarti during my time at IIT Bombay to show me the ropes. His research methodology and his tenacity, beyond just the work I did with him, have made a lasting influence on shaping me as a researcher. I also had a wonderful summer internship experience during my undergraduate studies at the Max Planck Institute for Biological Cybernetics in Tübeingen, Germany in the summer of 2006. I was a member in the group of Prof. Bernhard Schölkopf, working with him along with Bob Williiamson from NICTA and Olivier Chapelle. That internship was my first brush

with theoretical machine learning, and was instrumental in both familiarizing me with the area, and for me to realize my interest in the subject. I was extremely fortunate to have an access to such exceptional researchers, and ask them naïve questions as I was still learning the basics. It also formed the basis for a continued relationship and I have certainly enjoyed interacting with Bernhard and Bob at least annually at various meetings.

I had further opportunities to work with Olivier twice during his stay at Yahoo Research, and doing research with him is always a pleasure. I enjoyed a wonderful summer internship with Lin Xiao at Microsoft Research and with John Langford at Yahoo Research. These experiences were invaluable in experiencing research environments outside Berkeley, and often offered the opportunity to work on a completely different flavor of problems than my thesis research. I am also grateful for Microsoft Research and Google for funding my Ph. D. through fellowships at different stages.

My collaborators in various works have also made key contributions to various elements of this thesis, and I would be remiss to not acknowledge them here. During my initial years, I greatly benefitted from my interactions with two postdoctoral fellows at Berkeley. Sasha Rakhlin was a postdoc in Peter's group with whom I undertook one of my first serious research projects at Berkeley, and have had numerous collaborations with since then. I was quite fortunate for Pradeep Ravikumar to arrive in Berkeley as a postdoc with Martin in the beginning of my first year. I had a wonderful experience doing research in him that led to the first publication of my Ph. D. Having an access to these two fully-trained researchers, often on a daily basis, certainly eased the steep learning curve that initial years of graduate school offer, and made the experience a lot less intimidating than it is to many.

I have also enjoyed extremely productive and enjoyable collaborations with John Duchi and Sahand Negahban. Indeed each of them is responsible in part for a chapter of this thesis (and hence for any drawbacks in it!). It was always wonderful to have such technically skilled and highly motivated peers to consult, and on numerous occasions their enthusiasm rubbed off on me and drove me to work harder. I would also like to thank Jake Abernethy for the many fruitful technical conversations with him (and for the juggling shows!) Having two advisors means being a part of two research groups, and I certainly enjoyed interacting with the many members of Peter's and Martin's groups over the years. I would like to thank Ambuj Tewari, Benjamin Rubinstein, Joe Neeman, Fares Hedayati, Afshin Rostamizadeh, Alan Malek, Arash Amini, Nima Noorshams, Garvesh Raskutti and Miles Lopes; you all made the graduate school experience a lot of fun. A special mention to Po-Ling for the baked deliciousness she frequently brings to group meetings—they will be missed. I have also benefitted from interactions with many researchers across the machine learning community, including Michael Jordan, Bin Yu, Mikael Johansson, Miroslav Dudík, Rob Schapire, Nati Srebro, Sham Kakade, Claudio Gentile and John TsiTsiklis. I apologize in advance to anyone whom I have neglected to mention.

Finally, I cannot find the right words to thank my family who have loved and supported me throughout this Ph. D. If it weren't for their constant motivation and belief, I would have probably never made this far and I owe all my success to them. A very special thanks

to Meghana, I am very fortunate to have your love and support, as well as your patient ear whenever I am feeling down.

Chapter 1

Introduction

Any typical solution to a machine learning problem has two fundamental aspects. The first is a statistical aspect that characterizes how well the solution performs in making future predictions on unseen data, or in recovering the true underlying model. The second is a computational aspect that describes efficient algorithms that indeed compute this solution, and characterizes the computational complexity of these algorithms. Traditionally, statistics has been the main framework to understand the former aspect, while optimization and sampling have served as the principal computational paradigms. The study of the two aspects has largely happened in isolation, in the interest of modularity. In this thesis, we examine how such a two-phased design of learning algorithms can often fail to address the challenges of many modern learning problems involving massive amounts of data. Through a joint analysis of the statistical and computational aspects of our learning problems, we seek to obtain frameworks capable of addressing these challenges.

1.1 Classical statistics, big data and computational constraints

The last several decades of research in statistics, and more recently machine learning have led to remarkable advances in estimation of patterns from data. The cornerstone of these works has been a precise characterization of how well we can estimate the quantities of interest, as the amount of data available increases. A key ingredient for this understanding has been the literature on empirical process theory [166, 128, 163], in particular, uniform laws of large numbers such as the Glivenko-Cantelli theorem [70, 46] and corresponding uniform central limit theorems [62]. In particular, approaches based on VC theory [168, 167], metric entropy [88, 163], Rademacher and Gaussian complexities [17, 99, 154] as well as their sharper localized variants [155, 90, 21] have all played a key role in characterizing three fundamental aspects of statistical problems:

(i) When is it possible for *any possible* estimator to be statistically consistent, as the number of data samples approaches infinity?

(ii) What are sharp upper and lower bounds on the rate at which error of an estimator decays with an increasing number of samples?

(iii) Development of general principles such as (regularized) empirical risk minimization that achieve the above optimal limits for a large class of estimation problems.

An understanding of the minimum number of samples sufficient for a desired level of statistical performance was indeed a critical question in the classical settings. The reason for this is that for a long period in statistics and machine learning, the availability of data samples formed the key constraint. The samples were quite often acquired through painstaking manual data acquisition, or time consuming laboratory experiments. As a result, the key consideration of both the theoretical and algorithmic endeavors in machine learning and statistics was to extract the maximum possible information from the smallest possible number of data samples.

However, the last several years have seen an unprecedented growth in the size of datasets, owing largely due to rapid improvements in automated data acquisition techniques. The Internet has served as a rich data source for machine learning problems, particularly in natural language and computer vision applications, where large corpora can often be created based on the freely available web content. Advancements such as high-throughput sequencing in biology, better sensor technology allowing us to record observations at short intervals, and most recently, the use of crowd-sourcing platforms to compile datasets of machine learning problems have all contributed to this data deluge.

To put the difference of scales in perspective, we observe that the largest dataset in the popular UCI machine learning repository[1] prior to 1990 had about 8000 samples, in 22 dimensions. The largest dataset currently in the repository has 8 million samples, in 100000 dimensions, and this is still dwarfed in comparison to the largest datasets available in the industry which often involve up to billions of samples and millions of features. This apparent abundance of data leads to a very natural question about traditional statistics and machine learning. *Does the emergence of such massive datasets make complex and sophisticated estimators redundant?* Indeed there seems to be a folklore emerging to the extent that a simple estimator computed on enormous amounts of data beats a more complex one based on a smaller subset of the data.

In this thesis, we examine how the questions and concerns of machine learning and statistics take an interesting twist in these massive data problems. In particular, the focus will be on two main sources of complexity that frequently arise in such scenarios. First will be the challenges of high-dimensionality of the data, which often means that the number of data samples while being large, is often small relative to the number of parameters being

[1] http://archive.ics.uci.edu/ml/index.html

estimated. The second crucial challenge is that while the number and dimensionality of the data samples seems to grow without bounds, computation is struggling to keep up. As a result, we might not be able to compute the estimators prescribed by classical theory within feasible computational resources. We discuss these challenges in some more detail next, before moving on to discuss how they can be addressed.

With increasing number of samples, it is often desirable to understand higher-order interactions in the data. For instance, in natural language tasks, so called n-gram features attempt to understand the roles of pairs or triplets of words, and more generally of phrases instead of individual words in determining the syntax and semantics of text. In computer vision, increasingly complicated filters can be used to capture finer aspects of the images and in many biology tasks, the samples often correspond to long genomic strings along with additional features computed on these. Many such datasets challenge a key assumption in much of classical statistics: *the number of samples grows large relative to the number of parameters.* Indeed, in a popular example of the Netflix prize, the goal was to estimate over 8.5 billion parameters from a little more than 100 million samples. While it remains impossible to estimate consistently in such high-dimensional problems in general, a large body of research has investigated various structural assumptions under which statistical recovery is possible even in these seemingly ill-posed scenarios. Examples include a large line of works on sparsity [59, 51, 43, 165, 30], group sparsity [110, 180, 126], low-rank assumptions [136, 1, 150] and more abstract generalizations of these [116, 50]. Understanding the statistical and computational challenges posed by such structural problems in high-dimensions will be a key question explored in this thesis.

In stark contrast to the growth in problem sizes, the last few years have seen a tapering off in the growth of computational power, at least that of a single core. This combination of the large amounts of complex data and limited computational resources brings computational complexity issues to the fore. It might no longer be straightforward to compute the estimators with good statistical properties, prescribed by learning theory. For instance, empirical risk minimization and M-estimation typically involve solving numerical optimization problems, which can be quite challenging with large amounts of high-dimensional data. A natural question to ask in such scenarios is: *how well can we estimate a certain number of parameters, given a fixed amount of data, under reasonable constraints on computational resources?* On the algorithmic side, we are interested in algorithms that can work under specified constraints on computational resources, while having good statistical guarantees on the resulting solution. This interaction of the computational and statistical complexities is pictorially illustrated in Figure 1.1. Furthermore, we want to develop such solutions across a large variety of computational platforms; single cores, multiple cores and distributed systems such as cluster and cloud infrastructures to name a few.

In the next section, we will discuss some of the existing frameworks and approaches that address various facets of the above mentioned problems. We follow that with a detailed discussion of the key questions and contributions of this thesis.

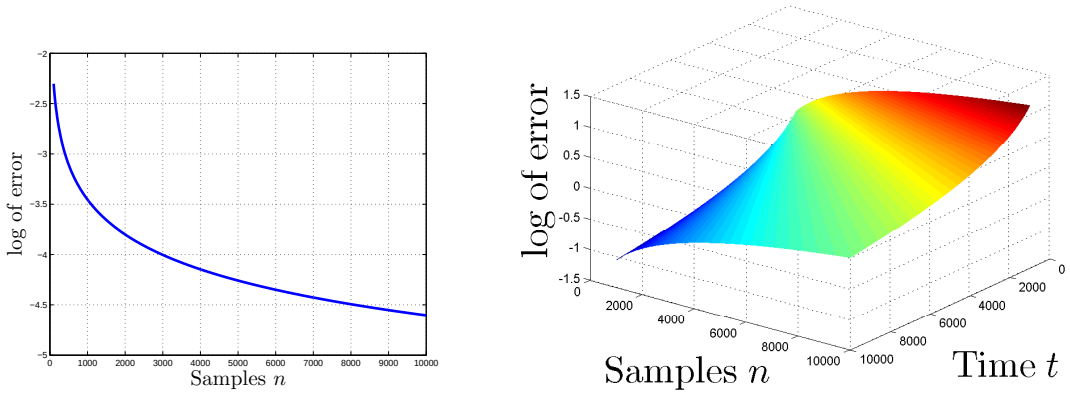

Figure 1.1. An illustration of the key quantity of interest in classical statistics (left) and modern scenarios (right). In the left plot, we study the decay of estimation error as the number of samples is increased. The right panel explores the estimation error as a function of the samples n and the computational time T. Capturing the behavior of the surface in the right panel for different n, T values can be quite challenging, and a primary question of interest in this thesis.

1.2 Connections to existing works

In this section we will survey some of the existing lines of research that explore themes related to the work in this thesis. Since the goals and results of the thesis are rather broad, the related works will also come from a variety of settings and considerations. We start with some approaches aimed at understanding the fundamental interactions between statistical and computational complexities, and then go on to related ideas in high-dimensional and distributed computational scenarios.

1.2.1 PAC Learning

The theory of PAC learning proposed by Valiant [162] aims to ask the precise question at the heart of this thesis: when is it possible to consistently learn a concept, using computation that is polynomial in the number of data samples. The thesis work of Michael Kearns [86] and his book with Vazirani [87] built on Valiant's theory and are influential works in advancing our understanding about when computationally efficient learning is possible. While this line of work is extremely powerful in distinguishing between problems that are solvable in polynomial time from those that are likely to be computationally hard, it provides no way of judging the relative hardness of problems that are solvable in polynomial time. Nevertheless, this remains one of the most active lines of research in understanding the interactions between learning and computation, and has seen some interesting results recently in understanding the trade-offs between sample sizes and training times [146].

1.2.2 Information-Based Complexity

Another natural framework for studying the computational complexity of solving learning problems is Information-Based Complexity (henceforth IBC). This theory aims to understand the computational complexity of problems, where the observed information is partial, noisy and/or comes at a price. This theory has been quite successful in understanding the computation of linear operators, such as numerical integration in high-dimensions and solutions of differential or integral equations. In particular, the results we will develop in Chapter 3 on the complexity of stochastic convex optimization can be seen as understanding the IBC of that problem, as observed in the context of earlier work of Nemirovski and Yudin [119] by Traub and Wozniakowski [157].

1.2.3 Stochastic and online convex optimization

Stochastic convex optimization [119, 93] and the more general online convex optimization [170, 47, 145] frameworks have emerged as the most scalable approaches for large-scale machine learning [32, 147]. These algorithms have the attractive property that they typically make a few passes over the data, doing a simple update based on each sample (or based on a mini-batch). This naturally gives them a computationally budgeted flavor—we run the algorithm till we exhaust the budget and take the predictor at the end. We will survey this line of work in more detail in the next chapter. Also, Chapter 3 will investigate the computational complexity of stochastic convex optimization algorithms.

1.2.4 Budgeted learning algorithms

Online learning algorithms directly equate samples to computational budget in some sense, but this is not always the right way to handle computational restrictions. Other lines of work on budgeted algorithms have explored budget constraints of various kinds, typically tailored to specific problem settings. Some of these works have been in the context of bound on the number of support vectors for kernelized perceptron [55, 153] or SVMs [54], where the number of support vectors imposes both a memory and running time constraint for kernel-based algorithms. A different class of approaches that considers budget on the complexity of prediction during test time is coarse-to-fine learning, where the data samples are passed through a hierarchy of models of increasing complexity. These approaches have found considerable success in natural language processing [127] as well as computer vision [172]. In Chapter 4, we will consider a budgeted framework for the model selection problem.

1.2.5 High-dimensional statistics and compressed sensing

The goal of high-dimensional statistics and its signal processing counterpart of compressed sensing is to obtain structural conditions under which an exceedingly high-dimensional model

can be recovered from a relatively small number of observations. A growing body of research has furthered the understanding of such conditions under structural conditions such as sparsity [59, 51, 43, 165, 30], group sparsity [110, 180, 126], low-rank assumptions [136, 1, 150] and more abstract generalizations of these [116, 50]. The estimators coming out of this theory are often formulated as high-dimensional convex programs, which are considerably challenging computationally. In Chapter 5 of this thesis, we will demonstrate how the structural assumptions made for statistical analysis also improve the computational complexity of these problems.

1.2.6 Distributed optimization algorithms

If convex optimization is a scalable framework for machine learning problems on a single computer, then it is natural to expect distributed convex optimization to be a viable solution for distributed machine learning. Distributed optimization is a classical subject, starting from the early works of Bertsekas [26] and Tsitsiklis[160], with many of the early results described in their seminal book [27]. There has been renewed activity in this area, in part owing to the growing interest from machine learning community [132, 81, 61, 36]. However, these classical approaches often fail to fully capture the structure and assumptions underlying distributed machine learning problems, presenting opportunities for the development of new methods as we will see in Chapter 6.

1.3 Main problems and contributions

The main emphasis in this thesis is on understanding the interactions between statistical and computational complexities of learning algorithms. Such an understanding is fundamental to characterize how we can trade-off the quality of statistical estimation for better computational complexity, or vice versa. As might be expected with an endeavor of this generality, there are many ways of posing the problem and this thesis will examine a few different approaches to understanding such phenomena. More specifically, we will investigate questions such as

(a) Given a computational budget, what are the fundamental limits on the quality of the *best estimator that can be computed*? Conversely, what is the computational cost of computing an estimator with a small error for various classes of learning problems?

(b) Can we design algorithms that work with an explicit constraint on the available *computational budget*?

(c) What are efficient statistical and computational methods for *structured learning problems in high dimensions*?

(d) What are good strategies for *distributed learning and inference*?

In the remainder of this section, we will describe in detail how this thesis attempts to answer each of these questions, introducing novel concepts and frameworks in the process.

1.3.1 Fundamental oracle complexity of stochastic convex optimization

Relative to the large literature on upper bounds on complexity of convex optimization, lesser attention has been paid to the fundamental hardness of these problems. Given the extensive use of convex optimization in machine learning and statistics, gaining an understanding of these complexity-theoretic issues is important for answering the first question mentioned at the start of this section. In Chapter 3, we study the complexity of stochastic convex optimization within the oracle model of complexity introduced by Nemirovski and Yudin [119] (henceforth Nemirovski and Yudin). In this complexity model, at every round the optimization procedure queries an oracle for certain information on the function being optimized. Our work focuses on stochastic first order oracles, where this information consists of noisy gradient evaluations.

Within this setup, we improve upon the work of Nemirovski and Yudin [119] for stochastic convex optimization in three ways. First, our lower bounds have an improved dependence on the dimension of the space. In the context of statistical estimation, these bounds show how the difficulty of the estimation problem increases with the number of parameters. Second, our techniques naturally extend to give sharper results for optimization over simpler function classes. We show that the complexity of optimization for strongly convex losses is smaller than that for convex, Lipschitz losses. Third, we show that for a fixed function class, if the set of optimizers is assumed to have special structure such as sparsity, then the fundamental complexity of optimization can be significantly smaller. All of our proofs exploit a new notion of the discrepancy between two functions that appears to be natural for optimization problems. They involve a reduction from stochastic optimization to a statistical parameter estimation problem, and an application of information-theoretic lower bounds for the estimation problem.

1.3.2 Computationally adaptive model selection

A complementary algorithmic approach regarding fundamental computational complexity of machine learning problems is the development of algorithms that take in explicit constraints on the amount of computation they can perform, and provide statistical guarantees as a function of this budget. A key challenge for such frameworks is setting up a notion of computational budget which admits interesting algorithms, but is also amenable to tractable theoretical analysis. Turing machine complexity is naturally suited to the former concern, but often fails to satisfy the latter. As mentioned previously, online learning provides one way of handling computational restrictions where the budget is directly equated to the number

of samples the algorithm sees. However, this might not always be the right way to impose budget constraints, since it might be preferable to compute a more complex estimator on the existing data samples rather than to repeat the same computation on new samples in different scenarios.

While such a development in a very broad setting is usually quite challenging, specific problems often admit interesting notions of computational budget that can be more intuitive than either of the two extremes of Turing machine or online learning. Based on this philosophy, Chapter 4 looks at a computationally budgeted framework for the problem of model selection. In particular, we analyze general model selection procedures using penalized empirical loss minimization under computational constraints. While classical model selection approaches do not consider computational aspects of performing model selection, we argue that any practical model selection procedure must not only trade off estimation and approximation error, but also the effects of the computational effort required to compute empirical minimizers for different function classes. We provide a framework for analyzing such problems, and we give algorithms for model selection under a computational budget. Our framework is based on the natural postulate that, for a fixed sample size, it is more expensive to estimate a model from a complex class than a simple class. Put inversely, given a computational bound, a simple model class can fit a model to a much larger sample size than a rich model class. So any strategy for model selection under a computational budget constraint should trade off two criteria: (i) the relative training cost of different model classes, which allows simpler classes to receive far more data (thus making them resilient to overfitting), and (ii) lower approximation error in the more complex model classes.

In addressing these computational and statistical issues, this work makes two main contributions. First, we propose a novel computational perspective on the model selection problem, which we believe should be a natural consideration in statistical learning problems. Secondly, within this framework, we provide algorithms for model selection in many different scenarios, and provide oracle inequalities on their estimates under different assumptions. Our first two results address the case where we have a model hierarchy that is ordered by inclusion. The first result provides an inequality that is competitive with an oracle, incurring at most an additional logarithmic penalty in the computational budget. The second result extends our approach to obtaining fast rates for model selection, as demonstrated in computationally unconstrained settings by Bartlett [19] and Koltchinskii [91]. Both of our results carefully refine the existing complexity-regularized risk minimization techniques by a careful consideration of the structure of the problem. Our third result applies to model classes that do not necessarily share any common structure. Here we present a novel algorithm — exploiting algorithms for multi-armed bandit problems—that uses confidence bounds based on concentration inequalities to select a good model under a given computational budget. We also prove a minimax optimal oracle inequality on the performance of the selected model. All of our algorithms are computationally simple and efficient.

1.3.3 Optimization methods for high-dimensional statistical estimation

As mentioned earlier in this introduction, many modern statistical problems tend to be overwhelmingly high-dimensional in nature. A growing body of statistical literature aims to develop conditions under which we can estimate the underlying statistical parameter reliably, given only a few samples, by exploiting the structure of the true parameter as well as regularity assumptions on the data. Computationally, many of these statistical M-estimators are based on convex optimization problems formed by the combination of a data-dependent loss function with a norm-based regularizer. We analyze the convergence rates of projected gradient and composite gradient methods for solving such problems, working within a high-dimensional framework that allows the data dimension d to grow with (and possibly exceed) the sample size n. This high-dimensional structure precludes the usual global assumptions—namely, strong convexity and smoothness conditions—that underlie much of classical optimization analysis.

In Chapter 5, we define appropriately restricted versions of these conditions, and show that they are satisfied with high probability for various statistical models. Under these conditions, our theory guarantees that projected gradient descent has a globally geometric rate of convergence up to the *statistical precision* of the model, meaning the typical distance between the true unknown parameter θ^* and an optimal solution $\widehat{\theta}$. This result is substantially sharper than previous convergence results, which yielded sublinear convergence, or linear convergence only up to the noise level. Our analysis applies to a wide range of M-estimators and statistical models, including sparse linear regression using Lasso (ℓ_1-regularized regression); group Lasso for block sparsity; log-linear models with regularization; low-rank matrix recovery using nuclear norm regularization; and matrix decomposition. Overall, our analysis reveals interesting connections between statistical precision and computational efficiency in high-dimensional estimation.

An interesting aspect of our results is that the global geometric convergence is not guaranteed to an arbitrary numerical precision, but only to an accuracy related to *statistical precision* of the problem. Note that this is very natural from the statistical perspective, since it is the true parameter θ^* itself (as opposed to the solution $\widehat{\theta}$ of the M-estimator) that is of primary interest, and our analysis allows us to approach it as close as is statistically possible. Our analysis shows that we can geometrically converge to a parameter θ such that $\|\theta - \theta^*\| = \left\|\widehat{\theta} - \theta^*\right\| + o\left(\left\|\widehat{\theta} - \theta^*\right\|\right)$, which is the best we can hope for statistically, ignoring lower order terms. Overall, our results reveal an interesting connection between the statistical and computational properties of M-estimators—that is, the properties of the underlying statistical model that make it favorable for estimation also render it more amenable to optimization procedures.

1.3.4 Asymptotically optimal algorithms for distributed learning

A natural computational framework for learning with massive amounts of data is parallel and distributed computation. This presents novel algorithmic challenges, since a vast majority of machine learning algorithms tend to be iterative; stochastic and online optimization algorithms being prime examples. The inherent sequential nature of these algorithms precludes obvious parallelization, which has led to search for new algorithms and abstractions fit for distributed machine learning problems. As mentioned earlier in this introduction, distributed convex optimization algorithms provide a partial answer to this question, but they consider structures more general than typical setups in distributed learning. As a result, they typically incur a slowdown from decentralization—there is a penalty for distributed computation and it is ideal to have a centralized algorithm whenever possible. There have been recent works [57, 137] that demonstrate that these slowdowns can, however, be avoided by leveraging the greater structure present in typical distributed machine learning problems.

In Chapter 6, we analyze the convergence of gradient-based optimization algorithms that base their updates on delayed stochastic gradient information. The main application of our results is to the development of gradient-based distributed optimization algorithms where a master node performs parameter updates while worker nodes compute stochastic gradients based on local information in parallel, which may give rise to delays due to asynchrony. Our main contribution is to show that for smooth stochastic problems, the delays are asymptotically negligible and we can achieve order-optimal convergence results. In application to distributed optimization, we develop procedures that overcome communication bottlenecks and synchronization requirements. We show n-node architectures whose optimization error in stochastic problems—in spite of asynchronous delays—scales asymptotically as $\mathcal{O}(1/\sqrt{nT})$ after T iterations. This rate is known to be optimal for a distributed system with n nodes even in the absence of delays. We additionally complement our theoretical results with numerical experiments on a logistic regression task.

1.4 Thesis Overview

The remainder of this thesis is organized as follows. In Chapter 2, we provide a survey of some of the background material from statistics and learning theory, as well as convex optimization. Chapter 3 presents a minimax framework for understanding the complexity of stochastic convex optimization problems, and gives lower bounds on this complexity for various problem classes of interest. In Chapter 4, we develop a setup for model selection problems, given computational budget constraints. Within the setup, we present new model selection algorithms and statistical oracle inequalities on their performance. Chapter 5 considers how we can exploit the structural assumptions underlying many high-dimensional estimation problems for rapid convergence of optimization algorithms. The results show an intricate interplay between the computational and statistical sample complexities of these

problems. In Chapter 6, we discuss how the computational techniques can be modified to adapt them to parallel and distributed computational infrastructures. Our results demonstrate an asymptotically optimal linear speedup with the number of nodes, under reasonable assumptions on the learning problem. Finally, we conclude in Chapter 7 by summarizing the high-level message of this thesis, and outlining directions for future work. Some of the more technical proofs will be deferred to the Appendices.

Chapter 2

Background

The aim of this chapter is to set up some of the concepts that will be frequently used throughout the thesis. More specialized concepts relevant to particular chapters will be introduced in those chapters. As the title indicates, this thesis studies questions at the intersection of statistics and computation. Consequently, we will need concepts from both the domains to provide the background for this work. Specifically, we will require concepts from convex optimization on the computational side, and basics of decision and learning theory on the statistical side. We start by giving a broad setting that captures a vast majority of machine learning problems, and then go on to discuss these computational and statistical backgrounds.

2.1 Typical problem setup

In a typical learning or statistical estimation problem, we receive a sequence of samples $z_1, z_2, \ldots, z_n$. In this thesis we will restrict attention to scenarios where the samples are drawn i.i.d. according to an unknown distribution. There is an underlying parameter θ of interest that we are trying to estimate from this data. For simplicity, we will restrict our attention to settings where $\theta \in \mathbb{R}^d$ for most of this thesis, although a vast majority of the concepts extend to general functional mappings. The quality of a parameter θ in making a prediction regarding the sample z is measured through a loss function $\ell(\theta; z)$. Examples of some of the loss functions commonly used in the literature are:

- *Negative log likelihood:* Assume an indexed family of distributions $\mathbb{P}_\theta(z)$ with a density p_θ, and suppose that the samples are drawn i.i.d. according to $\mathbb{P}_{\theta^*}(z)$ where θ^* is unknown. Then we define $\ell(\theta; z) = -\log p_\theta(z)$, which results in the maximum likelihood estimation principle.

- *0-1 loss:* Suppose our samples consist of pairs (x, y) where $x \in \mathbb{R}^d$ is the covariate vector and $y \in \{-1, 1\}$ is the associated binary label. Let $\theta : \mathbb{R}^d \mapsto \{-1, +1\}$ be a

mapping that predicts a binary label, given a data vector. A common loss function for such classification problems is the indicator $\ell(\theta; (x, y)) = \mathbb{I}(\theta(x) \neq y)$, where $\theta(x)$ is the prediction on x and $\mathbb{I}(A)$ is the 0-1 valued indicator function for the event A. A typical example is $\theta(x) = \text{sign}(\theta^T x)$.

- *Squared loss:* Once again our samples consist of pairs (x, y) with $x \in \mathbb{R}^d$, but this time we have $y \in \mathbb{R}$. For such problems, least squares regression is defined via the loss function $\ell(\theta; (x, y)) = (y - \theta^T x)^2$ and results in the ordinary least squares estimator.

Letting $\mathbb{E}$ denote expectation with respect to the (unknown) probability distribution underlying our samples, a typical goal in statistical estimation is to recover

$$\theta^* = \arg\min_{\theta \in \Omega} \mathbb{E}\ell(\theta; z). \tag{2.1}$$

Since the distribution is unknown, we cannot compute θ^* directly, and machine learning aims to define estimators for θ^* based on the observed data samples. One particular principle that is rather ubiquitous for defining such estimators has been called (regularized) Empirical Risk Minimization (ERM) in machine learning, (regularized) M-estimation in frequentist statistics and MAP estimation in the Bayesian literature. Based on a regularization function $\mathcal{R}$, we define the estimator as

$$\widehat{\theta}_n = \arg\min_{\theta \in \Omega} f(\theta); \quad \text{where} \quad f(\theta) := \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; z_i) + \lambda_n \mathcal{R}(\theta) \right\}. \tag{2.2}$$

Characterizing the statistical and computational properties of such estimators is a key question of interest in statistics and machine learning. These properties often depend critically on the regularizer $\mathcal{R}$, which serves to prevent overfitting statistically, and often improves the numerical conditioning of the computational problem. Some of the commonly used regularizers in the literature include

- The simplest setting is unregularized ERM, corresponding to $\mathcal{R}(\theta) \equiv 0$.
- For $\theta \in \mathbb{R}^d$, a common choice is the squared ℓ_2-norm, $\mathcal{R}(\theta) = \|\theta\|_2^2$. When θ belongs to a general Hilbert or Banach space, we replace the ℓ_2-norm with the associated norm for the space. This is often called ridge regularization owing to its use in ridge regression.
- For $\theta \in \mathbb{R}^d$, another regularizer of interest is the ℓ_1-norm: $\mathcal{R}(\theta) = \|\theta\|_1 = \sum_{i=1}^d |\theta_i|$. This regularizer is common in applications where the desired estimator is sparse, and various extensions are also considered in the literature.
- The final example we consider is motivated by the maximum entropy principle, and is used when the parameters θ are non-negative. We define the negative entropy regularizer: $\mathcal{R}(\theta) = \sum_{i=1}^d (\theta_i \log \theta_i - \theta_i)$.

Below we present some of the background concepts from convex optimization, crucial to solving the computational problem (2.2) before proceeding to the statistical concepts.

2.2 Background on convex optimization

The estimator (2.2) is naturally defined as the solution of a numerical optimization problem, and in many typical scenarios, the objective $f(\theta)$ and the constraint set Ω are both convex. Clearly f is convex whenever the loss function $\ell(\theta; z)$ and the regularizer $\mathcal{R}(\theta)$ are both convex functions of θ. The negative log likelihood loss is convex when the distribution $\mathbb{P}_\theta$ comes from an exponential family with natural parameter θ, and it is easy to see that the least squares loss is always convex. The 0-1 loss is not convex, and a large body of literature on classification problems studies the computational and statistical properties of convex upper bounds on the 0-1 loss. Two such popular surrogate losses are the hinge loss $\ell(\theta; z) = \max\{0, 1 - y\theta^T x\}$ and the logistic loss $\ell(\theta; z) = \log(1 + \exp(-y\theta^T x))$. We refer the reader to the excellent works [179, 16] for the statistical implications of using these surrogate losses. As for the regularizer $\mathcal{R}$, we see that all the examples discussed earlier are indeed convex, although there has been work on non-convex regularization as well, most prominently for variable selection problems in high dimensions [66, 178].

For the remainder of this thesis, we will restrict our attention to problems where the loss function $\ell(\theta; z)$ and the regularizer $\mathcal{R}(\theta)$ are both convex in θ, and the constraint set Ω is also convex. As a result, the optimization problem (2.2) is convex, and can be solved in polynomial time. Perhaps the simplest algorithm to solve the problem is projected gradient descent which is an iterative algorithm that starts with an arbitrary initialization $\theta^0 \in \Omega$, and successively updates:

$$\theta^{t+1} = \Pi_\Omega \left(\theta^t - \alpha(t) \nabla f(\theta^t) \right). \tag{2.3}$$

Here $\Pi_\Omega(\theta) = \arg\min_{\tilde{\theta} \in \Omega} \left\| \tilde{\theta} - \theta \right\|_2^2$ is the projection of θ on to the constraint set Ω. Some of the other prominent approaches to solve the problem (2.2) include co-ordinate descent methods, quasi-Newton approaches such as L-BFGS [41] and interior point methods [123]. While a detailed discussion of all the relevant optimization algorithms and their properties is beyond the scope of this thesis, we refer the reader to the excellent texts [35, 28] for an in-depth treatment, as well as the recent text [149] that specifically surveys optimization algorithms for machine learning.

To describe some of the complexity aspects of the above mentioned optimization algorithms relevant to this thesis, we will first need to recall some standard definitions about the properties of the objective function f (2.2). We refer the reader to standard texts on convex analysis [76, 140] for a more detailed treatment. For a convex function f, we define the sub-differential set

$$\partial f(\theta) = \left\{ v \in \mathbb{R}^d \ : \ f(\tilde{\theta}) \geq f(\theta) + \langle v, \tilde{\theta} - \theta \rangle \quad \forall \tilde{\theta} \in \mathbb{R}^d \right\}. \tag{2.4}$$

A function f is G-Lipschitz with respect to a norm $\|\cdot\|$ if $|f(\theta) - f(\tilde{\theta})| \leq G \left\|\theta - \tilde{\theta}\right\|$. We say that a convex function f is γ_ℓ-strongly convex with respect to the norm $\|\cdot\|$ if

$$f(\alpha\theta + (1-\alpha)\tilde{\theta}) \leq \alpha f(\theta) + (1-\alpha) f(\tilde{\theta}) - \alpha(1-\alpha)\frac{\gamma_\ell}{2}\left\|\theta - \tilde{\theta}\right\|^2. \tag{2.5}$$

For a norm $\|\cdot\|$ we define the dual norm $\|v\|_* = \sup_{\|\theta\| \leq 1} \langle \theta, v \rangle$. A differentiable function f is γ_u-smooth with respect to the norm $\|\cdot\|$ if the gradients are γ_u-Lipschitz

$$\|\nabla f(\theta) - \nabla f(\tilde{\theta})\|_* \leq \gamma_u \left\|\theta - \tilde{\theta}\right\| \iff f(\tilde{\theta}) \leq f(\theta) + \left\langle \nabla f(\theta), \tilde{\theta} - \theta \right\rangle + \frac{\gamma_u}{2}\left\|\theta - \tilde{\theta}\right\|^2. \tag{2.6}$$

Based on various of the above assumptions, we now mention some complexity results for the commonly used optimization algorithms mentioned above. All the algorithms mentioned above are iterative, and we are interested in understanding the minimum number of iterations T after which $f(\theta^T) - f(\widehat{\theta}_n) \leq \epsilon$. When $\Omega \subseteq \mathbb{R}^d$ and the function f is locally Lipschitz in a neighborhood around $\widehat{\theta}_n$, then the ellipsoid algorithm and interior-point methods require at most $d \log(1/\epsilon)$ iterations to attain this goal [123]. This rate is also known to be optimal for a class of gradient-based algorithms due to matching lower bounds [119]. When the function f is both smooth and strongly convex with respect to the ℓ_2-norm, we define the condition number $\kappa = \gamma_\ell / \gamma_u \in (0, 1)$. Under these conditions, gradient descent converges *linearly*, at a rate $\kappa \log(1/\epsilon)$, while Nesterov's accelerated gradient methods [122, 159, 96] converge at a rate $\sqrt{\kappa} \log(1/\epsilon)$. The latter is known to be optimal for all first order algorithms under these assumptions [119]. We note that L-BFGS also converges linearly with a rate $c \log(1/\epsilon)$, where c is a constant that depends on the line-search conditions that are satisfied, and also enjoys local quadratic convergence.

2.3 Background on stochastic convex optimization

For problems with the structure (2.2), gradient methods have a poor scaling as the number n of the samples grows. This is because the computation of the gradient of $f(\theta)$ involves summing over the $\nabla \ell(\theta; z_i)$ terms, increasing at least linearly in number of samples. Stochastic convex optimization algorithms aim to address this problem by using sampling in place of summation. Stochastic optimization algorithms for the problem (2.2) operate by drawing a sample z_t uniformly at random from the data. This sampling can be with or without replacement, leading to single or multiple pass optimization algorithms. Based on the sample z_t, we can evaluate the single sample (sub-)gradients $\nabla \ell(\theta; z_t) + \lambda_n \nabla \mathcal{R}(\theta)$, which are unbiased for the problems (2.2) or (2.1) depending on whether the sampling is done with or without replacement respectively.

We present below an example of a stochastic first-order algorithm called mirror descent, and describe some of the complexity results for stochastic convex optimization in the context of this algorithm. We note that similar, and somewhat improved results are sometimes possible using other methods such as dual averaging [122, 173] and accelerated gradient methods [122, 159, 96] respectively. For consistency of notation with standard literature, we consider the minimization of a general convex function f, with[1]

$$\theta_f^* \in \arg\min_{\theta \in \Omega} f(\theta). \tag{2.7}$$

Mirror descent is a generalization of (projected) stochastic gradient descent, first introduced by Nemirovski and Yudin [119]; here we follow a more recent presentation of it due to Beck and Teboulle [22]. For a given norm $\|\cdot\|$, let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a differentiable function that is 1-strongly convex with respect to $\|\cdot\|$ (2.5). We assume that ψ is a function of Legendre type [140, 76], which implies that the conjugate dual ψ^* is differentiable on its domain with $\nabla\psi^* = \big(\nabla\psi\big)^{-1}$. For a given proximal function, we let D_ψ be the Bregman divergence induced by ψ, given by

$$D_\psi(\theta, \tilde{\theta}) := \psi(\theta) - \psi(\tilde{\theta}) - \langle \nabla\psi(\tilde{\theta}),\, \theta - \tilde{\theta} \rangle. \tag{2.8}$$

With this set-up, we can now describe the mirror descent algorithm based on the proximal function ψ for minimizing a convex function f over a convex set Ω contained within the domain of ψ. Starting with an arbitrary initial $\theta^0 \in \Omega$, it generates a sequence $\{\theta^t\}_{t=0}^{\infty}$ contained within Ω via the updates

$$\theta^{t+1} = \arg\min_{\theta \in \Omega} \big\{ \alpha(t) \langle \theta,\, \nabla f(\theta^t) \rangle + D_\psi(\theta, \theta^t) \big\}, \tag{2.9}$$

where $\alpha(t) > 0$ is a stepsize. In case of stochastic optimization, $\nabla f(\theta^t)$ is simply replaced by the noisy version $\widehat{v}(\theta^t)$ with $\mathbb{E}[\widehat{v}(\theta^t)|\theta^t] = \nabla f(\theta^t)$. We also note that the gradient can be replaced with an arbitrary element of the subdifferential set (2.4) throughout this discussion.

A special case of this algorithm is obtained by choosing the proximal function $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, which is 1-strongly convex with respect to the Euclidean norm. The associated Bregman divergence $D_\psi(\theta, \tilde{\theta}) = \frac{1}{2}\|\theta - \tilde{\theta}\|_2^2$ is simply the Euclidean norm, so that the updates (2.9) correspond to a standard projected gradient descent method. If one receives only an unbiased estimate of the gradient $\nabla f(\theta^t)$, then this algorithm corresponds to a form of projected stochastic gradient descent. Moreover, other choices of the proximal function lead to different stochastic algorithms, squared-ℓ_p norms and negative entropy function being prime examples.

Explicit convergence rates for this algorithm can be obtained under appropriate convexity and Lipschitz assumptions for f. Following the standard assumptions in the literature, we assume that $\mathbb{E}\left[\|\widehat{v}(\theta^t)\|_*^2 \mid \theta^t\right] \leq G^2$ for all $\theta \in \Omega$. Given stochastic mirror descent

[1] For most of this thesis, our problems will have sufficient regularity that the minimum above is attained, although it suffices to take any θ_f^* such that $f(\theta_f^*) \leq \min_{\theta \in \Omega} f(\theta) + \epsilon$ for ϵ small enough.

based on unbiased estimates of the gradient, it can be showed that (see e.g., Chapter 5.1 of NY [119] or Beck and Teboulle [22]) with the initialization $\theta^1 = \arg\min_{\theta\in\Omega} \psi(\theta)$ and stepsizes $\alpha(t) = 1/\sqrt{t}$, the optimization error of the sequence $\{\theta^t\}$ is bounded as

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[f(\theta^t) - f(\theta_f^*)\big] \leq G\sqrt{\frac{D_\psi(\theta_f^*, \theta^1)}{T}} \;\leq\; G\sqrt{\frac{\psi(\theta_f^*)}{T}} \tag{2.10}$$

Let us define the averaged iterate $\bar{\theta}(T) = \sum_{t=1}^{T} \theta^t/T$. Then Jensen's inequality combined with the convexity of f further yields the expected convergence rate

$$\mathbb{E}\big[f(\bar{\theta}(T)) - f(\theta_f^*)\big] \;\leq\; G\sqrt{\frac{\psi(\theta_f^*)}{T}}$$

The results can also be generalized to hold with high probability [48, 117], and some of the specializations and the optimality of such results will be discussed in Chapter 3. From the point of view of optimization, we see that the sublinear rates (2.10) are exponentially slower than the linear convergence in the noiseless optimization scenarios. However, in statistical applications, two factors drive the preference for stochastic approaches: (i) the complexity of each iteration is typically $\mathcal{O}(1)$ instead of $\mathcal{O}(n)$ for full-gradient algorithms, and (ii) statistical problems typically only require optimization to a moderate and not extremely high precision. A detailed discussion of the latter issue can be found in Chapter 13 of the recent book [149], and these aspects will also be covered in more depth in Chapters 3 and 5 of this thesis. We also refer the reader to the relevant chapters of the recent text [149], as well as the ICML 2010 tutorial [151] for more details on the relevance and use of stochastic optimization methods in machine learning.

2.4 Background on minimax theory in statistics

Once we have an algorithm to compute estimators of the form $\widehat{\theta}_n$, it is natural to ask how well does $\widehat{\theta}_n$ perform statistically in estimating θ^*. Different problems come with different performance criteria, and we discuss the two most frequent notions of risk consistency and parameter consistency. Based on the loss function $\ell(\theta; z)$, one natural performance metric is the excess risk $\mathbb{E}\ell(\widehat{\theta}_n; z) - \mathbb{E}\ell(\theta^*; z)$. Based on a norm $\|\cdot\|$, another commonly used criterion is the mean squared error $\mathbb{E}\left\|\widehat{\theta}_n - \theta^*\right\|^2$.

A large body of literature in statistics and learning theory is devoted to studying how these performance measures decay to 0, as $n \to \infty$, in expectation and with high probability. Such results typically rely on some assumptions, such as the boundedness of the loss function $\ell(\theta; z)$, or on its convexity and Lipschitz properties along with appropriate tail conditions

on the data distribution. When $\widehat{\theta}_n$ is obtained from a stochastic optimization algorithm, that samples the data z_t without replacement for the problem (2.2), then such guarantees are a direct consequence of the so called *online-to-batch conversion* [48, 84]. While the details of how such guarantees are proved are beyond the scope of this brief review, we refer the reader to the excellent texts [58, 11], as well as some more recent papers and survey articles [17, 21, 90, 33, 34, 163] for more details.

A complementary line of works also aims to obtain lower bounds on how well any possible estimator $\widehat{\theta}_n$ can do in estimating θ^*, after observing n samples. A useful concept in decision theory that helps in such a study is that of minimax error

$$\inf_{\widehat{\theta}_n \in \Omega} \sup_{\theta^* \in \Omega} \mathbb{E} \left\| \widehat{\theta}_n - \theta^* \right\|^2,$$

or the corresponding excess risk variant. Here the infimum is taken over all estimators $\widehat{\theta}_n$ that can be computed from n samples. While upper bounds on the minimax error are obtained used the techniques mentioned above, there is also a rich literature on lower bounds. In particular, for many problems matching upper and lower bounds have been obtained through the use of Assouad's lemma, Fano's inequality of Le Cam's method [175, 176, 72]. The work in Chapter 3 on the oracle complexity of stochastic convex optimization will draw upon this minimax risk framework and lower bound techniques from decision theory.

Chapter 3

Lower bounds on oracle complexity of stochastic convex optimization

Convex optimization forms the backbone of many algorithms for statistical learning and estimation. Given that many statistical estimation problems are large-scale in nature—with the problem dimension and/or sample size being large—it is essential to make efficient use of computational resources. Stochastic optimization algorithms are an attractive class of methods, known to yield moderately accurate solutions in a relatively short time [32]. Given the popularity of such stochastic optimization methods, understanding the fundamental computational complexity of stochastic convex optimization is thus a key issue for large-scale learning. A large body of literature is devoted to obtaining rates of convergence of specific procedures for various classes of convex optimization problems. A typical outcome of such analysis is an upper bound on the error—for instance, gap to the optimal cost—as a function of the number of iterations. Such analyses have been performed for many standard optimization algorithms, among them gradient descent, mirror descent, interior point programming, and stochastic gradient descent, to name a few. We refer the reader to various standard texts on optimization (e.g., [35, 28, 120]) for further details on such results.

On the other hand, there has been relatively little study of the inherent complexity of convex optimization problems. To the best of our knowledge, the first formal study in this area was undertaken in the seminal work of Nemirovski and Yudin [119]. One obstacle to a classical complexity-theoretic analysis, as these authors observed, is that of casting convex optimization problems in a Turing Machine model. They avoided this problem by instead considering a natural oracle model of complexity, in which at every round the optimization procedure queries an oracle for certain information on the function being optimized. This information can be either noiseless or noisy, depending on whether the goal is to lower bound the oracle complexity of deterministic or stochastic optimization algorithms. Working within this framework, the authors obtained a series of lower bounds on the computational complexity of convex optimization problems, both in deterministic and stochastic settings. In addition to the original text Nemirovski and Yudin [119], we refer the interested reader to

the book by Nesterov [120], and the lecture notes by Nemirovski [118] for further background.

In this chapter, we consider the computational complexity of stochastic convex optimization within this oracle model. In particular, we improve upon the work of Nemirovski and Yudin [119] for stochastic convex optimization in two ways. First, our lower bounds have an improved dependence on the dimension of the space. In the context of statistical estimation, these bounds show how the difficulty of the estimation problem increases with the number of parameters. Second, our techniques naturally extend to give sharper results for optimization over simpler function classes. We show that the complexity of optimization for strongly convex losses is smaller than that for convex, Lipschitz losses. Third, we show that for a fixed function class, if the set of optimizers is assumed to have special structure such as sparsity, then the fundamental complexity of optimization can be significantly smaller. All of our proofs exploit a new notion of the discrepancy between two functions that appears to be natural for optimization problems. They involve a reduction from a statistical parameter estimation problem to the stochastic optimization problem, and an application of information-theoretic lower bounds for the estimation problem. We note that the results of this chapter appear in the paper [6] and a related study was independently undertaken by Raginsky and Rakhlin [130].

The remainder of this chapter is organized as follows. We begin in Section 3.1 with background on oracle complexity, and a precise formulation of the problems addressed in this chapter. Section 3.2 is devoted to the statement of our main results, and discussion of their consequences. In Section 3.3, we provide the proofs of our main results, which all exploit a common framework of four steps. More technical aspects of these proofs are deferred to the appendices.

Notation: For the convenience of the reader, we collect here some notation used throughout the chapter. For $p \in [1, \infty]$, we use $\|\theta\|_p$ to denote the ℓ_p-norm of a vector $\theta \in \mathbb{R}^d$, and we let q denote the conjugate exponent, satisfying $\frac{1}{p} + \frac{1}{q} = 1$. For two distributions $\mathbb{P}$ and $\mathbb{Q}$, we use $D(\mathbb{P} \| \mathbb{Q})$ to denote the Kullback-Leibler (KL) divergence between the distributions. The notation $\mathbb{I}(A)$ refers to the 0-1 valued indicator random variable of the set A. For two vectors $\alpha, \beta \in \{-1, +1\}^d$, we define the Hamming distance $\Delta_H(\alpha, \beta) := \sum_{i=1}^d \mathbb{I}[\alpha_i \neq \beta_i]$. We also recall the definition of the subdifferential set of a convex function (2.4).

3.1 Background and problem formulation

We begin by introducing background on the oracle model of convex optimization, and then turn to a precise specification of the problem to be studied.

3.1.1 Convex optimization in the oracle model

Convex optimization is the task of minimizing a convex function f over a convex set $\Omega \subseteq \mathbb{R}^d$. Assuming that the minimum is achieved, it corresponds to computing an element θ_f^* that achieves the minimum—that is, an element $\theta_f^* \in \arg\min_{\theta \in \Omega} f(\theta)$. An *optimization method* is any procedure that solves this task, typically by repeatedly selecting values from Ω. For a given class of optimization problems, our primary focus in this chapter is to determine lower bounds on the computational cost, as measured in terms of the number of (noisy) function and subgradient evaluations, required to obtain an ϵ-optimal solution to any optimization problem within the class.

More specifically, we follow the approach of Nemirovski and Yudin [119], and measure computational cost based on the oracle model of optimization. The main components of this model are an *oracle* and an *information set*. An *oracle* is a (possibly random) function $\phi : \Omega \mapsto \mathcal{I}$ that answers any query $\theta \in \Omega$ by returning an element $\phi(\theta)$ in an information set $\mathcal{I}$. The information set varies depending on the oracle; for instance, for an exact oracle of m^{th} order, the answer to a query θ^t consists of θ^t and the first m derivatives of f at θ^t. For the case of stochastic oracles studied in this chapter, these values are corrupted with zero-mean noise with bounded variance. We then measure the computational labor of any optimization method as the number of queries it poses to the oracle.

In particular, given a positive integer T corresponding to the number of iterations, an optimization method $\mathcal{M}$ designed to approximately minimize the convex function f over the convex set Ω proceeds as follows. At any given iteration $t = 1, \ldots, T$, the method $\mathcal{M}$ queries at $\theta^t \in \Omega$, and the oracle reveals the information $\phi(\theta^t, f)$. The method then uses the information $\{\phi(\theta^1, f), \ldots, \phi(\theta^t, f)\}$ to decide at which point θ^{t+1} the next query should be made. For a given oracle function ϕ, let $\mathbb{M}_T$ denote the class of all optimization methods $\mathcal{M}$ that make T queries according to the procedure outlined above. For any method $\mathcal{M} \in \mathbb{M}_T$, we define its error on function f after T steps as

$$\epsilon_T(\mathcal{M}, f, \Omega, \phi) := f(\theta^T) - \min_{\theta \in \Omega} f(\theta) \;=\; f(\theta^T) - f(\theta_f^*), \tag{3.1}$$

where θ^T is the method's query at time T. Note that by definition of θ_f^* as a minimizing argument, this error is a non-negative quantity.

When the oracle is stochastic, the method's query θ^T at time T is itself random, since it depends on the random answers provided by the oracle. In this case, the optimization error $\epsilon_T(\mathcal{M}, f, \Omega, \phi)$ is also a random variable. Accordingly, for the case of stochastic oracles, we measure the accuracy in terms of the expected value $\mathbb{E}_\phi[\epsilon_T(\mathcal{M}, f, \Omega, \phi)]$, where the expectation is taken over the oracle randomness. Given a class of functions $\mathcal{F}$ defined over a convex set Ω and a class $\mathbb{M}_T$ of all optimization methods based on T oracle queries, we define the minimax error

$$\epsilon_T^*(\mathcal{F}, \Omega; \phi) := \inf_{\mathcal{M} \in \mathbb{M}_T} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon_T(\mathcal{M}, f, \Omega, \phi)]. \tag{3.2}$$

In the sequel, we provide results for particular classes of oracles. So as to ease the notation, when the oracle ϕ is clear from the context, we simply write $\epsilon_T^*(\mathcal{F}, \Omega)$.

3.1.2 Stochastic first-order oracles

In this chapter, we study stochastic oracles for which the information set $\mathcal{I} \subset \mathbb{R} \times \mathbb{R}^d$ consists of pairs of noisy function and subgradient evaluations. More precisely, we have:

Definition 3.1. *For a given set Ω and function class $\mathcal{F}$, the class of first-order stochastic oracles consists of random mappings $\phi : S \times \mathcal{F} \to \mathcal{I}$ of the form $\phi(\theta, f) = (\widehat{f}(\theta), \widehat{v}(\theta))$ such that*

$$\mathbb{E}[\widehat{f}(\theta)] = f(\theta), \qquad \mathbb{E}[\widehat{v}(\theta)] \in \partial f(\theta), \qquad \text{and} \quad \mathbb{E}\big[\, \|\widehat{v}(\theta)\|_p^2 \,\big] \leq \sigma^2. \tag{3.3}$$

We use $\mathbb{O}_{p,\sigma}$ to denote the class of all stochastic first-order oracles with parameters (p, σ). Note that the first two conditions imply that $\widehat{f}(\theta)$ is an unbiased estimate of the function value $f(\theta)$, and that $\widehat{v}(\theta)$ is an unbiased estimate of a subgradient $v \in \partial f(\theta)$. When f is actually differentiable, then $\widehat{v}(\theta)$ is an unbiased estimate of the gradient $\nabla f(\theta)$. The third condition in equation (3.3) controls the "noisiness" of the subgradient estimates in terms of the ℓ_p-norm.

Stochastic gradient methods are a widely used class of algorithms that can be understood as operating based on information provided by a stochastic first-order oracle. As a particular example, consider a function of the separable form (2.2). The natural stochastic gradient method for this problem is to choose an index $i \in \{1, 2, \ldots, n\}$ uniformly at random, and then to return the pair $(\ell(\theta; z_i), \nabla \ell(\theta; z_i))$. Taking averages over the randomly chosen index i yields $\frac{1}{n} \sum_{i=1}^n \ell(\theta; z_i) = f(\theta)$, so that $\ell(\theta; z_i)$ is an unbiased estimate of $f(\theta)$, with an analogous unbiased property holding for the gradient of $\ell(\theta; z_i)$.

3.1.3 Function classes of interest

We now turn to the classes $\mathcal{F}$ of convex functions for which we study oracle complexity. In all cases, we consider real-valued convex functions defined over some convex set Ω. We assume without loss of generality that Ω contains an open set around 0, and many of our lower bounds involve the maximum radius $r = r(\Omega) > 0$ such that

$$\Omega \;\supseteq\; \mathbb{B}_\infty(r) \;:=\; \big\{\theta \in \mathbb{R}^d \;\mid\; \|\theta\|_\infty \leq r\big\}. \tag{3.4}$$

Our first class consists of *convex Lipschitz functions:*

Definition 3.2. *For a given convex set $\Omega \subseteq \mathbb{R}^d$ and parameter $p \in [1, \infty]$, the class $\mathcal{F}_{\text{cv}}(\Omega, G, p)$ consists of all convex functions $f : \Omega \to \mathbb{R}$ such that*

$$\big|f(\theta) - f(\tilde{\theta})\big| \leq G \, \|\theta - \tilde{\theta}\|_q \qquad \text{for all } \theta, \tilde{\theta} \in \Omega, \tag{3.5}$$

where $\frac{1}{q} = 1 - \frac{1}{p}$.

We have defined the Lipschitz condition (3.5) in terms of the conjugate exponent $q \in [1, \infty]$, defined by the relation $\frac{1}{q} = 1 - \frac{1}{p}$. To be clear, our motivation in doing so is to maintain consistency with our definition of the stochastic first-order oracle, in which we assumed that $\mathbb{E}\big[\, \|\widehat{v}(\theta)\|_p^2 \,\big] \leq \sigma^2$. We note that the Lipschitz condition (3.5) is equivalent to the condition

$$\|v\|_p \leq G \quad \forall v \in \partial f(\theta), \quad \text{and for all } \theta \in \text{int}(\Omega).$$

If we consider the case of a differentiable function f, the unbiasedness condition in Definition 3.1 implies that

$$\|\nabla f(\theta)\|_p = \|\mathbb{E}[\widehat{v}(\theta)]\|_p \overset{(a)}{\leq} \mathbb{E}\|\widehat{v}(\theta)\|_p \overset{(b)}{\leq} \sqrt{\mathbb{E}\|\widehat{v}(\theta)\|_p^2} \leq \sigma,$$

where inequality (a) follows from the convexity of the ℓ_p-norm and Jensen's inequality, and inequality (b) is a result of Jensen's inequality applied to the concave function $\sqrt{\theta}$. This bound implies that f must be Lipschitz with constant at most σ with respect to the dual ℓ_q-norm. Therefore, we necessarily must have $G \leq \sigma$, in order for the function class from Definition 3.2 to be consistent with the stochastic first-order oracle.

A second function class consists of strongly convex functions, defined as follows:

Definition 3.3. *For a given convex set $\Omega \subseteq \mathbb{R}^d$ and parameter $p \in [1, \infty]$, the class $\mathcal{F}_{\text{scv}}(\Omega, p; G, \gamma_\ell)$ consists of all convex functions $f : \Omega \to \mathbb{R}$ such that the Lipschitz condition (3.5) holds, and such that f satisfies the ℓ_2-strong convexity condition*

$$f(\alpha\theta + (1-\alpha)\tilde{\theta}) \leq \alpha f(\theta) + (1-\alpha) f(\tilde{\theta}) - \alpha(1-\alpha)\frac{\gamma_\ell^2}{2}\|\theta - \tilde{\theta}\|_2^2 \quad \text{for all } \theta, \tilde{\theta} \in \Omega. \tag{3.6}$$

In this chapter, we restrict our attention to the case of strong convexity with respect to the ℓ_2-norm. (Similar results on the oracle complexity for strong convexity with respect to different norms can be obtained by straightforward modifications of the arguments given here). For future reference, it should be noted that the Lipschitz constant G and strong convexity constant γ_ℓ interact with one another. In particular, whenever $\Omega \subset \mathbb{R}^d$ contains the ℓ_∞-ball of radius r, the Lipschitz G and strong convexity γ_ℓ constants must satisfy the inequality

$$\frac{G}{\gamma_\ell^2} \geq \frac{r}{4} d^{1/p}. \tag{3.7}$$

In order to establish this inequality, we note that strong convexity condition with $\alpha = 1/2$ implies that

$$\frac{\gamma_\ell^2}{8} \leq \frac{f(\theta) + f(\tilde{\theta}) - 2f\left(\frac{\theta+\tilde{\theta}}{2}\right)}{2\|\theta - \tilde{\theta}\|_2^2} \leq \frac{G\|\theta - \tilde{\theta}\|_q}{2\|\theta - \tilde{\theta}\|_2^2}$$

We now choose the pair $\theta, \tilde{\theta} \in \Omega$ such that $\|\theta - \tilde{\theta}\|_\infty = r$ and $\|\theta - \tilde{\theta}\|_2 = r\sqrt{d}$. Such a choice is possible whenever Ω contains the ℓ_∞ ball of radius r. Since we have $\|\theta - \tilde{\theta}\|_q \leq d^{1/q}\|\theta - \tilde{\theta}\|_\infty$, this choice yields $\frac{\gamma_\ell^2}{4} \leq \frac{G d^{\frac{1}{q}-1}}{r}$, which establishes the claim (3.7).

As a third example, we study the oracle complexity of optimization over the class of convex functions that have sparse minimizers. This class of functions is well-motivated, since a large body of statistical work has studied the estimation of vectors, matrices and functions under various types of sparsity constraints. A common theme in this line of work is that the ambient dimension d enters only logarithmically, and so has a mild effect. Consequently, it is natural to investigate whether the complexity of optimization methods also enjoys such a mild dependence on ambient dimension under sparsity assumptions.

For a vector $\theta \in \mathbb{R}^d$, we use $\|\theta\|_0$ to denote the number of non-zero elements in θ. Recalling the set $\mathcal{F}_{\mathrm{cv}}(\Omega, G, p)$ from Definition 3.2, we now define a class of Lipschitz functions with sparse minimizers.

Definition 3.4. *For a convex set $\Omega \subset \mathbb{R}^d$ and positive integer $s \leq \lfloor d/2 \rfloor$, let*

$$\mathcal{F}_{\mathrm{sp}}(s; \Omega, G) := \left\{ f \in \mathcal{F}_{\mathrm{cv}}(\Omega, G, \infty) \mid \exists \quad \theta^* \in \arg\min_{\theta \in \Omega} f(\theta) \quad \text{satisfying } \|\theta^*\|_0 \leq s. \right\} \tag{3.8}$$

be the class of all convex functions that are G-Lipschitz in the ℓ_∞-norm, and have at least one s-sparse optimizer.

We frequently use the shorthand notation $\mathcal{F}_{\mathrm{sp}}(s)$ when the set Ω and parameter G are clear from context.

3.2 Main results and their consequences

With the setup of stochastic convex optimization in place, we are now in a position to state the main results of this chapter, and to discuss some of their consequences. As previously mentioned, a subset of our results assume that the set Ω contains an ℓ_∞ ball of radius $r = r(\Omega)$. Our bounds scale with r, thereby reflecting the natural dependence on the size of the set Ω. Also, we set the oracle second moment bound σ to be the same as the Lipschitz constant G in our results.

3.2.1 Oracle complexity for convex Lipschitz functions

We begin by analyzing the minimax oracle complexity of optimization for the class of bounded and convex Lipschitz functions $\mathcal{F}_{\mathrm{cv}}$ from Definition 3.2.

Theorem 3.1. *Let $\Omega \subset \mathbb{R}^d$ be a convex set such that $\Omega \supseteq \mathbb{B}_\infty(r)$ for some $r > 0$. Then for a universal constant $c_0 > 0$, the minimax oracle complexity over the class $\mathcal{F}_{\mathrm{cv}}(\Omega, G, p)$ satisfies the following lower bounds:*

(a) For $1 \leq p \leq 2$,

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \Omega; \phi) \geq \min \left\{ c_0 G\, r \sqrt{\frac{d}{T}}, \frac{Gr}{144} \right\}. \tag{3.9}$$

(b) For $p > 2$,

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \Omega; \phi) \geq \min \left\{ c_0 G\, r\, \frac{d^{1-\frac{1}{p}}}{\sqrt{T}}, \frac{G d^{1-1/p} r}{72} \right\}. \tag{3.10}$$

Remarks: Nemirovski and Yudin [119] proved the lower bound $\Omega\big(\frac{1}{\sqrt{T}}\big)$ for the function class $\mathcal{F}_{\mathrm{cv}}$, in the special case that Ω is the unit ball of a given norm, and the functions are Lipschitz in the corresponding *dual norm*. For $p \geq 2$, they established the minimax optimality of this dimension-independent result by appealing to a matching upper bound achieved by the method of mirror descent. In contrast, here we do not require the two norms—namely, that constraining the set Ω and that for the Lipschitz constraint—to be dual to one other; instead, we give give lower bounds in terms of the largest ℓ_∞ ball contained within the constraint set Ω. As discussed below, our bounds do include the results for the dual setting of past work as a special case, but more generally, by examining the relative geometry of an arbitrary set with respect to the ℓ_∞ ball, we obtain results for arbitrary sets. (We note that the ℓ_∞ constraint is natural in many optimization problems arising in machine learning settings, in which upper and lower bounds on variables are often imposed.) Thus, in contrast to the past work of NY on stochastic optimization, our analysis gives sharper dimension dependence under more general settings. It also highlights the role of the geometry of the set Ω in determining the oracle complexity.

In general, our lower bounds cannot be improved, and hence specify the optimal minimax oracle complexity. We consider here some examples to illustrate their sharpness. Throughout we assume that T is large enough to ensure that the $1/\sqrt{T}$ term attains the lower bound and not the $G/144$ term. (This condition is reasonable given our goal of understanding the rate as T increases, as opposed to the transient behavior over the first few iterations.)

(a) We start from the special case that has been primarily considered in past works. We consider the class $\mathcal{F}_{\mathrm{cv}}(\mathbb{B}_q(1), G, p)$ with $q = 1 - 1/p$ and the stochastic first-order oracles $\mathbb{O}_{p,G}$ for this class. Then the radius r of the largest ℓ_∞ ball inscribed within the $\mathbb{B}_q(1)$ scales as $r = d^{-1/q}$. By inspection of the lower bounds bounds (3.9) and (3.10), we see that

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{B}_q(1); \phi) = \begin{cases} \Omega\left(G\, \frac{d^{1/2-1/q}}{\sqrt{T}}\right) & \text{for } 1 \leq p \leq 2 \\ \Omega\left(\frac{G}{T}\right) & \text{for } p \geq 2. \end{cases} \tag{3.11}$$

As mentioned previously, the dimension-independent lower bound for the case $p \geq 2$ was demonstrated in Chapter 5 of NY, and shown to be optimal[1] since it is achieved using mirror descent with the prox-function $\|\cdot\|_q$. For the case of $1 \leq p < 2$, the lower bounds are also unimprovable, since they are again achieved (up to constant factors) by stochastic gradient descent. See Appendix A.3 for further details on these matching upper bounds.

(b) Let us now consider how our bounds can also make sharp predictions for non-dual geometries, using the special case $\Omega = \mathbb{B}_\infty(1)$. For this choice, we have $r(\Omega) = 1$, and hence Theorem 3.1 implies that for all $p \in [1, 2]$, the minimax oracle complexity is lower bounded as

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{B}_\infty(1); \phi) = \Omega\left(G\sqrt{\frac{d}{T}}\right).$$

This lower bound is sharp for all $p \in [1, 2]$. Indeed, for any convex set Ω, stochastic gradient descent achieves a matching upper bound (see Section 5.2.4, p. 196 of NY [119], as well as Appendix A.3 in this chapter for further discussion).

(c) As another example, suppose that $\Omega = \mathbb{B}_2(1)$. Observe that this ℓ_2-norm unit ball satisfies the relation $\mathbb{B}_2(1) \supset \frac{1}{\sqrt{d}}\mathbb{B}_\infty(1)$, so that we have $r(\mathbb{B}_2(1)) = 1/\sqrt{d}$. Consequently, for this choice, the lower bound (3.9) takes the form

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{B}_2(1); \phi) = \Omega\left(G\frac{1}{\sqrt{T}}\right),$$

which is a dimension-independent lower bound. This lower bound for $\mathbb{B}_2(1)$ is indeed tight for $p \in [1, 2]$, and as before, this rate is achieved by stochastic gradient descent [119].

(d) Turning to the case of $p > 2$, when $\Omega = \mathbb{B}_\infty(1)$, the lower bound (3.10) can be achieved (up to constant factors) using mirror descent with the dual norm $\|\cdot\|_q$; for further discussion, we again refer the reader to Section 5.2.1, p. 190 of NY [119], as well as to Appendix A.3 of this chapter. Also, even though this lower bound requires the oracle to have only bounded variance, our proof actually uses a stochastic oracle based on Bernoulli random variables, for which all moments exist. Consequently, at least in general, our results show that there is no hope of achieving faster rates by restricting to oracles with bounds on higher-order moments. This is an interesting contrast to the case of having *less* than two moments, in which the rates are slower. For instance, as shown in Section 5.3.1 of NY [119], suppose that the gradient estimates in a stochastic oracle satisfy the moment bound $\mathbb{E}\|\widehat{v}(\theta)\|_p^b \leq \sigma^2$ for some $b \in [1, 2)$. In this setting, the

[1] There is an additional logarithmic factor in the upper bounds for $p = \Omega(\log d)$.

oracle complexity is lower bounded by $\Omega\big(T^{-(b-1)/b}\big)$. Since $T^{\frac{b-1}{b}} \ll T^{\frac{1}{2}}$ for all $b \in [1, 2)$, there is a significant penalty in convergence rates for having less than two bounded moments.

(e) Even though the results have been stated in a first-order stochastic oracle model, they actually hold in a stronger sense. Let $\nabla^i f(\theta)$ denote the i_{th}-order derivative of f evaluated at θ, when it exists. With this notation, our results apply to an oracle that responds with a random function $\hat{f}_t$ such that

$$\mathbb{E}[\hat{f}_t(\theta)] = \mathbb{E}[f(\theta)], \text{ and } \mathbb{E}[\nabla^i \hat{f}_t(\theta)] = \nabla^i f(\theta) \text{ for all } \theta \in \Omega \text{ and } i \text{ such that } \nabla^i f(\theta) \text{ exists},$$

along with appropriately bounded second moments of all the derivatives. Consequently, higher-order gradient information cannot improve convergence rates in a worst-case setting. Indeed, the result continues to hold even for the significantly stronger oracle that responds with a random function that is a noisy realization of the true function. In this sense, our result is close in spirit to a statistical sample complexity lower bound. Our proof technique is based on constructing a "packing set" of functions, and thus has some similarity to techniques used in statistical minimax analysis (e.g., [72, 31, 175, 176]) and learning theory (e.g., [169, 65, 148]). A significant difference, as will be shown shortly, is that the metric of interest for optimization is very different than those typically studied in statistical minimax theory.

3.2.2 Oracle complexity for strongly convex Lipschitz functions

We now turn to the statement of lower bounds over the class of Lipschitz and strongly convex functions $\mathcal{F}_{\text{scv}}$ from Definition 3.3. In all these statements, we assume that $\gamma_\ell^2 \leq \frac{4Gd^{-1/p}}{r}$, as is required for the definition of $\mathcal{F}_{\text{scv}}$ to be sensible.

Theorem 3.2. *Let $\Omega = \mathbb{B}_\infty(r)$. Then there exist universal constants $c_1, c_2 > 0$ such that the minimax oracle complexity over the class $\mathcal{F}_{\text{scv}}(\Omega, p; G, \gamma_\ell)$ satisfies the following lower bounds:*

(a) For $p = 1$, we have

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon^*(\mathcal{F}_{\text{scv}}, \phi) \geq \min\left\{ c_1 \frac{G^2}{\gamma_\ell^2 T},\ c_2 Gr\sqrt{\frac{d}{T}},\ \frac{G^2}{1152\gamma_\ell^2 d},\ \frac{Gr}{144} \right\}. \tag{3.12}$$

(b) For $p > 2$, we have:

$$\sup_{\phi \in \mathbb{O}_{p,G}} \epsilon^*(\mathcal{F}_{\text{scv}}, \phi) \geq \min\left(c_1 \frac{G^2 d^{1-2/p}}{\gamma_\ell^2 T},\ c_2 \frac{Grd^{1-1/p}}{\sqrt{T}},\ \frac{G^2 d^{1-2/p}}{1152\gamma_\ell^2},\ \frac{Grd^{1-1/p}}{144} \right). \tag{3.13}$$

As with Theorem 3.1, these lower bounds are sharp. In particular, for $S = \mathbb{B}_\infty(1)$, stochastic gradient descent achieves the rate (3.12) up to logarithmic factors [75], and closely related algorithms proposed in very recent works [74, 82] match the lower bound exactly up to constant factors. It should be noted Theorem 3.2 exhibits an interesting phase transition between two regimes. On one hand, suppose that the strong convexity parameter γ_ℓ^2 is large: then as long as T is sufficiently large, the first term $\Omega(1/T)$ determines the minimax rate, which corresponds to the fast rate possible under strong convexity. In contrast, if we consider a poorly conditioned objective with $\gamma_\ell \approx 0$, then the term involving $\Omega(1/\sqrt{T})$ is dominant, corresponding to the rate for a convex objective. This behavior is natural, since Theorem 3.2 recovers (as a special case) the convex result with $\gamma_\ell = 0$. However, it should be noted that Theorem 3.2 applies only to the set $\mathbb{B}_\infty(r)$, and not to arbitrary sets Ω like Theorem 3.1. Consequently, the generalization of Theorem 3.2 to arbitrary convex, compact sets remains an interesting open question.

3.2.3 Oracle complexity for convex Lipschitz functions with sparse optima

Finally, we turn to the oracle complexity of optimization over the class $\mathcal{F}_{\text{sp}}$ from Definition 3.4.

Theorem 3.3. *Let $\mathcal{F}_{\text{sp}}$ be the class of all convex functions that are G-Lipschitz with respect to the $\|\cdot\|_\infty$ norm and that have a s-sparse optimizer. Let $\Omega \subset \mathbb{R}^d$ be a convex set with $\mathbb{B}_\infty(r) \subseteq \Omega$. Then there exists a universal constant $c_0 > 0$ such that for all $s \leq \lfloor \frac{d}{2} \rfloor$, we have*

$$\sup_{\phi \in \mathbb{O}_{\infty,G}} \epsilon^*(\mathcal{F}_{\text{sp}}, \phi) \geq \min\left(c_0 G r \sqrt{\frac{s^2 \log \frac{d}{s}}{T}}, \frac{Gsr}{432} \right). \tag{3.14}$$

Remark: If $s = \mathcal{O}(d^{1-\delta})$ for some $\delta \in (0,1)$ (so that $\log \frac{d}{s} = \Theta(\log d)$), then this bound is sharp up to constant factors. In particular, suppose that we use mirror descent based on the $\|\cdot\|_{1+\varepsilon}$ norm with $\varepsilon = 2\log d/(2\log d - 1)$. As we discuss in more detail in Appendix A.3, it can be shown that this technique will achieve a solution accurate to $\mathcal{O}\left(\sqrt{\frac{s^2 \log d}{T}}\right)$ within T iterations; this achievable result matches our lower bound (3.14) up to constant factors under the assumed scaling $s = \mathcal{O}(d^{1-\delta})$. To the best of our knowledge, Theorem 3.3 provides the first tight lower bound on the oracle complexity of sparse optimization.

3.3 Proofs of results

We now turn to the proofs of our main results. We begin in Section 3.3.1 by outlining the framework and establishing some basic results on which our proofs are based. Sections 3.3.2

through 3.3.4 are devoted to the proofs of Theorems 3.1 through 3.3 respectively.

3.3.1 Framework and basic results

We begin by establishing a basic set of results that are exploited in the proofs of the main results. At a high-level, our main idea is to show that the problem of convex optimization is at least as hard as estimating the parameters of Bernoulli variables—that is, the biases of d independent coins. In order to perform this embedding, for a given error tolerance ϵ, we start with an appropriately chosen subset of the vertices of a d-dimensional hypercube, each of which corresponds to some values of the d Bernoulli parameters. For a given function class, we then construct a "difficult" subclass of functions that are indexed by these vertices of the hypercube. We then show that being able to optimize any function in this subclass to ϵ-accuracy requires identifying the hypercube vertex. This is a multiway hypothesis test based on the observations provided by T queries to the stochastic oracle, and we apply Fano's inequality [52] or Le Cam's bound [98, 176] to lower bound the probability of error. In the remainder of this section, we provide more detail on each of steps involved in this embedding.

Constructing a difficult subclass of functions

Our first step is to construct a subclass of functions $\mathcal{G} \subseteq \mathcal{F}$ that we use to derive lower bounds. Any such subclass is parametrized by a subset $\mathcal{V} \subseteq \{-1, +1\}^d$ of the hypercube, chosen as follows. Recalling that Δ_H denotes the Hamming metric, we let $\mathcal{V} = \{\alpha^1, \ldots, \alpha^M\}$ be a subset of the vertices of the hypercube such that

$$\Delta_H(\alpha^j, \alpha^k) \geq \frac{d}{4} \quad \text{for all } j \neq k, \tag{3.15}$$

meaning that $\mathcal{V}$ is a $\frac{d}{4}$-packing in the Hamming norm. It is a classical fact (e.g., [109]) that one can construct such a set with cardinality $|\mathcal{V}| \geq (2/\sqrt{e})^{d/2}$.

Now let $\mathcal{G}_{\text{base}} = \{f_i^+, f_i^-, i = 1, \ldots, d\}$ denote some base set of $2d$ functions defined on the convex set Ω, to be chosen appropriately depending on the problem at hand. For a given tolerance $\delta \in (0, \frac{1}{4}]$, we define, for each vertex $\alpha \in \mathcal{V}$, the function

$$g_\alpha(\theta) := \frac{c}{d} \sum_{i=1}^{d} \big\{(1/2 + \alpha_i \delta) f_i^+(\theta) + (1/2 - \alpha_i \delta) f_i^-(\theta)\big\}. \tag{3.16}$$

Depending on the result to be proven, our choice of the base functions $\{f_i^+, f_i^-\}$ and the pre-factor c will ensure that each g_α satisfies the appropriate Lipschitz and/or strong convexity properties over Ω. Moreover, we will ensure that that all minimizers θ_α of each g_α are contained within Ω.

Based on these functions and the packing set $\mathcal{V}$, we define the function class

$$\mathcal{G}(\delta) := \{g_\alpha, \ \alpha \in \mathcal{V}\}. \tag{3.17}$$

Note that $\mathcal{G}(\delta)$ contains a total of $|\mathcal{V}|$ functions by construction, and as mentioned previously, our choices of the base functions etc. will ensure that $\mathcal{G}(\delta) \subseteq \mathcal{F}$. We demonstrate specific choices of the class $\mathcal{G}(\delta)$ in the proofs of Theorems 3.1 through 3.3 to follow.

Optimizing well is equivalent to function identification

We now claim that if a method can optimize over the subclass $\mathcal{G}(\delta)$ up to a certain tolerance, then it must be capable of identifying which function $g_\alpha \in \mathcal{G}(\delta)$ was chosen. We first require a measure for the *closeness* of functions in terms of their behavior near each others' minima. Recall that we use $\theta_f^* \in \mathbb{R}^d$ to denote a minimizing point of the function f. Given a convex set $S \subseteq \mathbb{R}^d$ and two functions f, g, we define

$$\rho(f, g) := \inf_{\theta \in \Omega} \big[f(\theta) + g(\theta) - f(\theta_f^*) - g(\theta_g^*)\big]. \tag{3.18}$$

This discrepancy measure is non-negative, symmetric in its arguments, and satisfies $\rho(f, g) = 0$

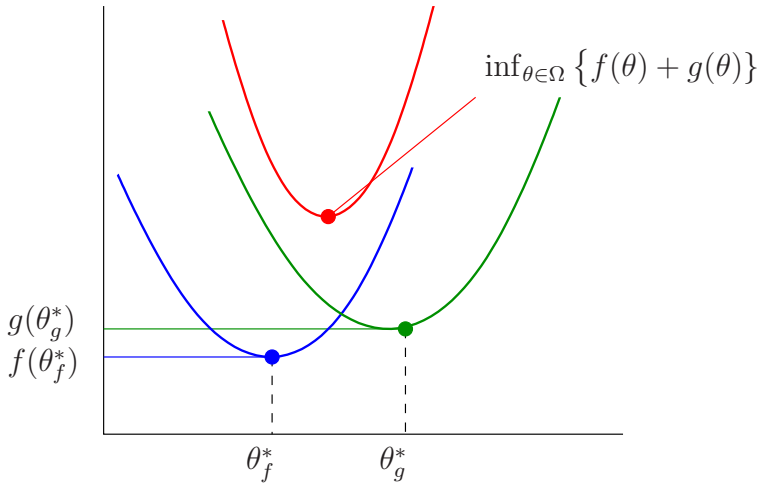

Figure 3.1. Illustration of the discrepancy function $\rho(f, g)$. The functions f and g achieve their minimum values $f(\theta_f^*)$ and $g(\theta_g^*)$ at the points θ_f^* and θ_g^* respectively.

if and only if $\theta_f^* = \theta_g^*$, so that we may refer to it as a premetric. (It does not satisfy the triangle inequality nor the condition that $\rho(f, g) = 0$ if and only if $f = g$, both of which are required for ρ to be a metric.)

Given the subclass $\mathcal{G}(\delta)$, we quantify how densely it is packed with respect to the premetric ρ using the quantity

$$\Phi(\mathcal{G}(\delta)) := \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta). \tag{3.19}$$

We denote this quantity by $\Phi(\delta)$ when the class $\mathcal{G}$ is clear from the context. We now state a simple result that demonstrates the utility of maintaining a separation under ρ among functions in $\mathcal{G}(\delta)$.

Lemma 3.1. *For any $\widetilde{\theta} \in \Omega$, there can be at most one function $g_\alpha \in \mathcal{G}(\delta)$ such that*

$$g_\alpha(\widetilde{\theta}) - \inf_{\theta \in \Omega} g_\alpha(\theta) \leq \frac{\Phi(\delta)}{3}. \tag{3.20}$$

Thus, if we have an element $\widetilde{\theta} \in \Omega$ that approximately minimizes one function in the set $\mathcal{G}(\delta)$ up to tolerance $\Phi(\delta)$, then it cannot approximately minimize any other function in the set.

Proof. For a given $\widetilde{\theta} \in \Omega$, suppose that there exists an $\alpha \in \mathcal{V}$ such that $g_\alpha(\widetilde{\theta}) - g_\alpha(\theta^*_\alpha) \leq \frac{\Phi(\delta)}{3}$. From the definition of $\Phi(\delta)$ in (3.19), for any $\beta \in \mathcal{V}$, $\beta \neq \alpha$, we have

$$\Phi(\delta) \;\leq\; g_\alpha(\widetilde{\theta}) - \inf_{\theta \in \Omega} g_\alpha(\theta) + g_\beta(\widetilde{\theta}) - \inf_{\theta \in \Omega} g_\beta(\theta) \;\leq\; \frac{\Phi(\delta)}{3} + g_\beta(\widetilde{\theta}) - \inf_{\theta \in \Omega} g_\beta(\theta).$$

Re-arranging yields the inequality $g_\beta(\widetilde{\theta}) - g_\beta(\theta^*_\beta) \geq \frac{2}{3}\Phi(\delta)$, from which the claim (3.20) follows.

□

Suppose that for some fixed but unknown function $g_{\alpha^*} \in \mathcal{G}(\delta)$, some method $\mathcal{M}_T$ is allowed to make T queries to an oracle with information function $\phi(\cdot\,; g_{\alpha^*})$, thereby obtaining the information sequence

$$\phi(z_1^T; g^*_\alpha) := \{\phi(\theta^t; g^*_\alpha), t = 1, 2, \ldots, T\}.$$

Our next lemma shows that if the method $\mathcal{M}_T$ achieves a low minimax error over the class $\mathcal{G}(\delta)$, then one can use its output to construct a hypothesis test that returns the true parameter α^* at least 2/3 of the time. (In this statement, we recall the definition (3.2) of the minimax error in optimization.)

Lemma 3.2. *Suppose that based on the data $\phi(z_1^T; g^*_\alpha)$, there exists a method $\mathcal{M}_T$ that achieves a minimax error satisfying*

$$\mathbb{E}\big[\epsilon_T(\mathcal{M}_T, \mathcal{G}(\delta), \Omega, \phi)\big] \leq \frac{\Phi(\delta)}{9}. \tag{3.21}$$

Based on such a method $\mathcal{M}_T$, one can construct a hypothesis test $\widehat{\alpha} : \phi(z_1^T; g^*_\alpha) \to \mathcal{V}$ such that $\max_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi[\widehat{\alpha} \neq \alpha^*] \leq \frac{1}{3}$.

Proof. Given a method $\mathcal{M}_T$ that satisfies the bound (3.21), we construct an estimator $\widehat{\alpha}(\mathcal{M}_T)$ of the true vertex α^* as follows. If there exists some $\alpha \in \mathcal{V}$ such that $g_\alpha(\theta^T) - g_\alpha(\theta_\alpha) \leq \frac{\Phi(\delta)}{3}$ then we set $\widehat{\alpha}(\mathcal{M}_T)$ equal to α. If no such α exists, then we choose $\widehat{\alpha}(\mathcal{M}_T)$ uniformly at random from $\mathcal{V}$. From Lemma 3.1, there can exist only one such $\alpha \in \mathcal{V}$ that satisfies this inequality. Consequently, using Markov's inequality, we have $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \mathbb{P}_\phi\big[\epsilon_T(\mathcal{M}_T, g_{\alpha^*}, \Omega, \phi) \geq \Phi(\delta)/3\big] \leq \frac{1}{3}$. Maximizing over α^* completes the proof. □

We have thus shown that having a low minimax optimization error over $\mathcal{G}(\delta)$ implies that the vertex $\alpha^* \in \mathcal{V}$ can be identified most of the time.

Oracle answers and coin tosses

We now describe stochastic first order oracles ϕ for which the samples $\phi(z_1^T; g_\alpha)$ can be related to coin tosses. In particular, we associate a coin with each dimension $i \in \{1, 2, \ldots, d\}$, and consider the set of coin bias vectors lying in the set

$$\Theta(\delta) = \big\{(1/2 + \alpha_1\delta, \ldots, 1/2 + \alpha_d\delta) \mid \alpha \in \mathcal{V}\big\}, \tag{3.22}$$

Given a particular function $g_\alpha \in \mathcal{G}(\delta)$—or equivalently, vertex $\alpha \in \mathcal{V}$—we consider two different types of stochastic first-order oracles ϕ, defined as follows:

Oracle A: 1-dimensional unbiased gradients

(a) Pick an index $i \in \{1, \ldots, d\}$ uniformly at random.

(b) Draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i\delta$.

(c) For the given input $\theta \in \Omega$, return the value $\widehat{g}_{\alpha,A}(\theta)$ and a sub-gradient $\widehat{v}_{\alpha,A}(\theta) \in \partial\widehat{g}_{\alpha,A}(\theta)$ of the function

$$\widehat{g}_{\alpha,A} \;:=\; c\big[b_i f_i^+ + (1 - b_i) f_i^-\big].$$

By construction, the function value and gradients returned by Oracle A are unbiased estimates of those of g_α. In particular, since each co-ordinate i is chosen with probability $1/d$, we have

$$\mathbb{E}\big[\widehat{g}_{\alpha,A}(\theta)\big] = \frac{c}{d}\sum_{i=1}^{d}\big[\mathbb{E}[b_i] f_i^+(\theta) + \mathbb{E}[1 - b_i] f_i^-(\theta)\big] \;=\; g_\alpha(\theta),$$

with a similar relation for the gradient. Furthermore, as long as the base functions f_i^+ and f_i^- have gradients bounded by 1, we have $\mathbb{E}[\|\widehat{v}_{\alpha,A}(\theta)\|_p] \leq c$ for all $p \in [1, \infty]$.

Parts of proofs are based on an oracle which responds with function values and gradients that are *d-dimensional* in nature.

Oracle B: d-dimensional unbiased gradients

(a) For $i = 1, \ldots, d$, draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i \delta$.

(b) For the given input $\theta \in \Omega$, return the value $\widehat{g}_{\alpha,B}(\theta)$ and a sub-gradient $\widehat{v}_{\alpha,B}(\theta) \in \partial \widehat{g}_{\alpha,B}(\theta)$ of the function

$$\widehat{g}_{\alpha,B} := \frac{c}{d} \sum_{i=1}^{d} \left[b_i f_i^+ + (1 - b_i) f_i^- \right].$$

As with Oracle A, this oracle returns unbiased estimates of the function values and gradients. We frequently work with functions f_i^+, f_i^- that depend only on the i^{th} coordinate $\theta(i)$. In such cases, under the assumptions $|\frac{\partial f_i^+}{\partial \theta(i)}| \leq 1$ and $|\frac{\partial f_i^-}{\partial \theta(i)}| \leq 1$, we have

$$\|\widehat{v}_{\alpha,B}(\theta)\|_p^2 = \frac{c^2}{d^2} \left(\sum_{i=1}^{d} \left| b_i \frac{\partial f_i^+(\theta)}{\partial \theta(i)} + (1 - b_i) \frac{\partial f_i^-(\theta)}{\partial \theta(i)} \right|^p \right)^{2/p} \leq c^2 d^{2/p - 2}. \tag{3.23}$$

In our later uses of Oracles A and B, we choose the pre-factor c appropriately so as to produce the desired Lipschitz constants.

Lower bounds on coin-tossing

Finally, we use information-theoretic methods to lower bound the probability of correctly estimating the true parameter $\alpha^* \in \mathcal{V}$ in our model. At each round of either Oracle A or Oracle B, we can consider a set of d coin tosses, with an associated vector $\theta^* = (\frac{1}{2} + \alpha_1^* \delta, \ldots, \frac{1}{2} + \alpha_d^* \delta)$ of parameters. At any round, the output of Oracle A can (at most) reveal the instantiation $b_i \in \{0, 1\}$ of a randomly chosen index, whereas Oracle B can at most reveal the entire vector $(b_1, b_2, \ldots, b_d)$. Our goal is to lower bound the probability of estimating the true parameter α^*, based on a sequence of length T. As noted previously in remarks following Theorem 3.1, this part of our proof exploits classical techniques from statistical minimax theory, including the use of Fano's inequality (e.g., [72, 31, 175, 176]) and Le Cam's bound (e.g., [98, 176]).

Lemma 3.3. *Suppose that the Bernoulli parameter vector α^* is chosen uniformly at random from the packing set $\mathcal{V}$, and suppose that the outcome of $\ell \leq d$ coins chosen uniformly at random is revealed at each round $t = 1, \ldots, T$. Then for any $\delta \in (0, 1/4]$, any hypothesis test $\widehat{\alpha}$ satisfies*

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{16\ell T\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}, \tag{3.24}$$

where the probability is taken over both randomness in the oracle and the choice of α^*.

Note that we will apply the lower bound (3.24) with $\ell = 1$ in the case of Oracle A, and $\ell = d$ in the case of Oracle B.

Proof. For each time $t = 1, 2, \ldots, T$, let U_t denote the randomly chosen subset of size ℓ, $X_{t,i}$ be the outcome of oracle's coin toss at time t for coordinate i and let $Y_t \in \{-1, 0, 1\}^d$ be a random vector with entries

$$Y_{t,i} = \begin{cases} X_{t,i} & \text{if } i \in U_t, \text{ and} \\ -1 & \text{if } i \notin U_t. \end{cases}$$

By Fano's inequality [52], we have the lower bound

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{I(\{(U_t, Y_t\}_{t=1}^T; \alpha^*) + \log 2}{\log |\mathcal{V}|},$$

where $I(\{(U_t, Y_t\}_{t=1}^T; \alpha^*)$ denotes the mutual information between the sequence $\{(U_t, Y_t)\}_{t=1}^T$ and the random parameter vector α^*. As discussed earlier, we are guaranteed that $\log |\mathcal{V}| \geq \frac{d}{2}\log(2/\sqrt{e})$. Consequently, in order to prove the lower bound (3.24), it suffices to establish the upper bound $I(\{U_t, Y_t\}_{t=1}^T; \alpha^*) \leq 16T\,\ell\,\delta^2$.

By the independent and identically distributed nature of the sampling model, we have

$$I(((U_1, Y_1), \ldots, (U_T, Y_T)); \alpha^*) = \sum_{t=1}^{T} I((U_t, Y_t); \alpha^*) \;=\; T\, I((U_1, Y_1); \alpha^*),$$

so that it suffices to upper bound the mutual information for a single round. To simplify notation, from here onwards we write (Y, U) to mean the pair (Y_1, U_1). With this notation, the remainder of our proof is devoted to establishing that $I(Y; U) \leq 16\,\ell\,\delta^2$,

By chain rule for mutual information [52], we have

$$I((U, Y); \alpha^*) = I(Y; \alpha^* \mid U) + I(\alpha^*; U). \tag{3.25}$$

Since the subset U is chosen independently of α^*, we have $I(\alpha^*; U) = 0$, and so it suffices to upper bound the first term. By definition of conditional mutual information [52], we have

$$I(Y; \alpha^* \mid U) = \mathbb{E}_U\big[D(\mathbb{P}_{Y|\alpha^*,U} \,\|\, \mathbb{P}_{Y|U})\big]$$

Since α has a uniform distribution over $\mathcal{V}$, we have $\mathbb{P}_{Y|U} = \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} \mathbb{P}_{Y|\alpha,U}$, and convexity of the Kullback-Leibler (KL) divergence yields the upper bound

$$D(\mathbb{P}_{Y|\alpha^*,U} \,\|\, \mathbb{P}_{Y|U}) \leq \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D(\mathbb{P}_{Y|\alpha^*,U} \,\|\, \mathbb{P}_{Y|\alpha,U}). \tag{3.26}$$

Now for any pair $\alpha^*, \alpha \in \mathcal{V}$, the KL divergence $D(\mathbb{P}_{Y|\alpha^*,U} \,\|\, \mathbb{P}_{Y|\alpha,U})$ can be at most the KL divergence between ℓ independent pairs of Bernoulli variates with parameters $\frac{1}{2} + \delta$ and $\frac{1}{2} - \delta$. Letting $D(\delta)$ denote the Kullback-Leibler divergence between a single pair of Bernoulli variables with parameters $\frac{1}{2} + \delta$ and $\frac{1}{2} - \delta$, a little calculation yields

$$\begin{aligned} D(\delta) &= \left(\frac{1}{2} + \delta\right) \log \frac{\frac{1}{2} + \delta}{\frac{1}{2} - \delta} + \left(\frac{1}{2} - \delta\right) \log \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta} \\ &= 2\delta \log \left(1 + \frac{4\delta}{1 - 2\delta}\right) \\ &\leq \frac{8\delta^2}{1 - 2\delta}. \end{aligned}$$

Consequently, as long as $\delta \leq 1/4$, we have $D(\delta) \leq 16\delta^2$. Returning to the bound (3.26), we conclude that $D(\mathbb{P}_{Y|\alpha^*,U} \,\|\, \mathbb{P}_{Y|U}) \leq 16\, \ell\, \delta^2$. Taking averages over U, we obtain the bound $I(Y; \alpha^* \mid U) \leq 16\, \ell\, \delta^2$, and applying the decomposition (3.25) yields $I((U, Y); \alpha^*) \leq 16\, \ell\, \delta^2$, thereby completing the proof. □

The reader might have observed that Fano's inequality yields a non-trivial lower bound only when $|\mathcal{V}|$ is large enough. Since $|\mathcal{V}|$ depends on the dimension d for our construction, we can apply the Fano lower bound only for d large enough. Smaller values of d can be lower bounded by reduction to the case $d = 1$; here we state a simple lower bound for estimating the bias of a single coin, which is a straightforward application of Le Cam's bounding technique [98, 176]. In this special case, we have $\mathcal{V} = \{1/2 + \delta, 1/2 - \delta\}$, and we recall that the estimator $\widehat{\alpha}(\mathcal{M}_T)$ takes values in $\mathcal{V}$.

Lemma 3.4. *Given a sample size $T \geq 1$ and a parameter $\alpha^* \in \mathcal{V}$, let $\{X_1, \ldots, X_T\}$ be T i.i.d Bernoulli variables with parameter α^*. Let $\widehat{\alpha}$ be any test function based on these samples and returning an element of $\mathcal{V}$. Then for any $\delta \in (0, 1/4]$, we have the lower bound*

$$\sup_{\alpha^* \in \{\frac{1}{2}+\delta, \frac{1}{2}-\delta\}} \mathbb{P}_{\alpha^*}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \sqrt{8T\delta^2}.$$

Proof. We observe first that for $\widehat{\alpha} \in \mathcal{V}$, $\mathbb{E}_{\alpha^*}[|\widehat{\alpha} - \alpha^*|] = 2\delta \mathbb{P}_{\alpha^*}[\widehat{\alpha} \neq \alpha^*]$, so that it suffices to lower bound the expected error. To ease notation, let $\mathbb{Q}_1$ and $\mathbb{Q}_{-1}$ denote the probability

distributions indexed by $\alpha = \frac{1}{2} + \delta$ and $\alpha = \frac{1}{2} - \delta$ respectively. By Lemma 1 of Yu [176], we have

$$\sup_{\alpha^* \in \mathcal{V}} \mathbb{E}_{\alpha^*}[|\widehat{\alpha} - \alpha^*|] \geq 2\delta \Big\{ 1 - \|\mathbb{Q}_1 - \mathbb{Q}_{-1}\|_1/2 \Big\}.$$

where we use the fact that $|(1/2 + \delta) - (1/2 - \delta)| = 2\delta$. Thus, we need to upper bound the total variation distance $\|\mathbb{Q}_1 - \mathbb{Q}_{-1}\|_1$. From Pinkser's inequality [52], we have

$$\|\mathbb{Q}_1 - \mathbb{Q}_{-1}\|_1 \leq \sqrt{2D(\mathbb{Q}_1 \|\mathbb{Q}_{-1})} \stackrel{(i)}{\leq} \sqrt{32T\delta^2},$$

where inequality (i) follows from the calculation following Equation 3.26 (see proof of Lemma 3.3), and uses our assumption that $\delta \in (0, 1/4]$. Putting together the pieces, we obtain a lower bound on the probability of error

$$\sup_{\alpha^* \in \mathcal{V}} \mathbb{P}[\widehat{\alpha} \neq \alpha^*] \;=\; \sup_{\alpha^* \in \mathcal{V}} \frac{\mathbb{E}|\widehat{\alpha} - \alpha^*|}{2\delta} \;\geq\; 1 - \sqrt{8T\delta^2},$$

as claimed. □

Equipped with these tools, we are now prepared to prove our main results.

3.3.2 Proof of Theorem 3.1

We begin with oracle complexity for bounded Lipschitz functions, as stated in Theorem 3.1. We first prove the result for the set $\Omega = \mathbb{B}_\infty(\frac{1}{2})$.

Part (a)—Proof for $p \in [1, 2]$: Consider Oracle A that returns the quantities $(\widehat{g}_{\alpha,A}(\theta), \widehat{v}_{\alpha,A}(\theta))$. By definition of the oracle, each round reveals only at most one coin flip, meaning that we can apply Lemma 3.3 with $\ell = 1$, thereby obtaining the lower bound

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq 1 - 2\frac{16T\delta^2 + \log 2}{d \log(2/\sqrt{e})}. \tag{3.27}$$

We now seek an upper bound $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha]$ using Lemma 3.2. In order to do so, we need to specify the base functions (f_i^+, f_i^-) involved. For $i = 1, \ldots, d$, we define

$$f_i^+(\theta) := \left|\theta(i) + \frac{1}{2}\right|, \quad \text{and} \quad f_i^-(\theta) := \left|\theta(i) - \frac{1}{2}\right|. \tag{3.28}$$

Given that $\Omega = \mathbb{B}_\infty(\frac{1}{2})$, we see that the minimizers of g_α are contained in S. Also, both the functions are 1-Lipschitz in the ℓ_1-norm. By the construction (3.16), we are guaranteed that for any subgradient of g_α, we have

$$\|\widehat{v}_{\alpha,A}(\theta)\|_p \leq 2c \qquad \text{for all } p \geq 1.$$

Therefore, in order to ensure that g_α is G-Lipschitz in the dual ℓ_q-norm, it suffices to set $c = G/2$.

Let us now lower bound the discrepancy function (3.18). We first observe that each function g_α is minimized over the set $\mathbb{B}_\infty(\frac{1}{2})$ at the vector $\theta_\alpha := -\alpha/2$, at which point it achieves its minimum value

$$\min_{\theta \in \mathbb{B}_\infty(\frac{1}{2})} g_\alpha(\theta) = \frac{c}{2} - c\delta.$$

Furthermore, we note that for any $\alpha \neq \beta$, we have

$$\begin{aligned} g_\alpha(\theta) + g_\beta(\theta) &= \frac{c}{d} \sum_{i=1}^{d} \left[\left(\frac{1}{2} + \alpha_i \delta + \frac{1}{2} + \beta_i \delta \right) f_i^+(\theta) + \left(\frac{1}{2} - \alpha_i \delta + \frac{1}{2} - \beta_i \delta \right) f_i^-(\theta) \right] \\ &= \frac{c}{d} \sum_{i=1}^{d} \left[(1 + \alpha_i \delta + \beta_i \delta)\, f_i^+(\theta) + (1 - \alpha_i \delta - \beta_i \delta)\, f_i^-(\theta) \right] \\ &= \frac{c}{d} \sum_{i=1}^{d} \left[\left(f_i^+(\theta) + f_i^-(\theta) \right) \mathbb{I}(\alpha_i \neq \beta_i) + \left((1 + 2\alpha_i \delta) f_i^+(\theta) + (1 - 2\alpha_i \delta) f_i^-(\theta) \right) \mathbb{I}(\alpha_i = \beta_i) \right]. \end{aligned}$$

When $\alpha_i = \beta_i$ then $\theta_\alpha(i) = \theta_\beta(i) = -\alpha_i/2$, so that this co-ordinate does not make a contribution to the discrepancy function $\rho(g_\alpha, g_\beta)$. On the other hand, when $\alpha_i \neq \beta_i$, we have

$$f_i^+(\theta) + f_i^-(\theta) = \left| \theta(i) + \frac{1}{2} \right| + \left| \theta(i) - \frac{1}{2} \right| \geq 1 \quad \text{for all } \theta \in \mathbb{R}^d.$$

Consequently, any such co-ordinate yields a contribution of $2c\delta/d$ to the discrepancy. Recalling our packing set (3.15) with $d/4$ separation in Hamming norm, we conclude that for any distinct $\alpha \neq \beta$ within our packing set,

$$\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\, \Delta_H(\alpha, \beta) \geq \frac{c\delta}{2},$$

so that by definition of Φ, we have established the lower bound $\Phi(\delta) \geq \frac{c\delta}{2}$.

Setting the target error $\epsilon := \frac{c\delta}{18}$, we observe that this choice ensures that $\epsilon < \frac{\Phi(\delta)}{9}$. Recalling the requirement $\delta < 1/4$, we have $\epsilon < c/72$. In this regime, we may apply Lemma 3.2 to obtain the upper bound $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$. Combining this upper bound with the lower bound (3.27) yields the inequality

$$\frac{1}{3} \geq 1 - 2\frac{16T\delta^2 + \log 2}{d \log(2/\sqrt{e})}.$$

Recalling that $c = \frac{G}{2}$, making the substitution $\delta = \frac{18\epsilon}{c} = \frac{36\epsilon}{G}$, and performing some algebra yields

$$T \geq c_0 \frac{G^2}{\epsilon^2} \left(\frac{d}{3} \log \left(\frac{2}{\sqrt{e}} \right) - \log 2 \right) \geq c_1 \frac{G^2 d}{\epsilon^2} \quad \text{for all } d \geq 11 \text{ and for all } \epsilon \leq \frac{G}{144},$$

where c_0 and c_1 are universal constants. Combined with Theorem 5.3.1 of NY [119] (or by using the lower bound of Lemma 3.4 instead of Lemma 3.3), we conclude that this lower bound holds for all dimensions d.

Part (b)—Proof for $p > 2$: The preceding proof based on Oracle A is also valid for $p > 2$, but yields a relatively weak result. Here we show how the use of Oracle B yields the stronger claim stated in Theorem 3.1(b). When using this oracle, all d coin tosses at each round are revealed, so that Lemma 3.3 with $\ell = d$ yields the lower bound

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq 1 - 2\frac{16\, T\, d\, \delta^2 + \log 2}{d \log(2/\sqrt{e})}. \tag{3.29}$$

We now seek an upper bound on $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha]$. As before, we use the set $\Omega = \mathbb{B}_\infty(\frac{1}{2})$, and the previous definitions (3.28) of $f_i^+(\theta)$ and $f_i^-(\theta)$. From our earlier analysis (in particular, equation (3.23)), the quantity $\|\widehat{v}_{\alpha,B}(\theta)\|_p$ is at most $cd^{1/p-1}$, so that setting $c = Gd^{1-1/p}$ yields functions that are Lipschitz with parameter G.

As before, for any distinct pair $\alpha, \beta \in \mathcal{V}$, we have the lower bound

$$\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\, \Delta_H(\alpha, \beta) \;\geq\; \frac{c\delta}{2},$$

so that $\Phi(\delta) \geq \frac{c\delta}{2}$. Consequently, if we set the target error $\epsilon := \frac{c\delta}{18}$, then we are guaranteed that $\epsilon < \frac{\Phi(\delta)}{9}$, as is required for applying Lemma 3.2. Application of this lemma yields the upper bound $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$. Combined with the lower bound (3.29), we obtain the inequality

$$\frac{1}{3} \geq 1 - 2\frac{16\, d\, T\, \delta^2 + \log 2}{d \log(2/\sqrt{e})}.$$

Substituting $\delta = 18\epsilon/c$ yields the scaling $\epsilon \geq c_0 \frac{c}{\sqrt{T}}$ for all $d \geq 11$, $\epsilon \leq c/72$ and a universal constant c_0. Recalling that $c = Gd^{1-1/p}$, we obtain the bound (3.10). Combining this with Theorem 5.3.1 of NY [119] (or by using the lower bound of Lemma 3.4 instead of Lemma 3.3) gives the claim for all dimensions.

We have thus completed the proof of Theorem 3.1 in the special case $\Omega = \mathbb{B}_\infty(\frac{1}{2})$. In order to prove the general claims, which scale with r when $B_\infty(r) \subseteq \Omega$, we note that our preceding proof required only that $\Omega \supseteq \mathbb{B}_\infty(\frac{1}{2})$ so that the minimizing points $\theta_\alpha = -\alpha/2 \in \Omega$ for all α (in particular, the Lipschitz constant of g_α does not depend on Ω for our construction). In the general case, we define our base functions to be

$$f_i^+(\theta) = \left|\theta(i) + \frac{r}{2}\right|, \quad \text{and} \quad f_i^-(\theta) = \left|\theta(i) - \frac{r}{2}\right|.$$

With this choice, the functions $g_\alpha(\theta)$ are minimized at $\theta_\alpha = -r\alpha/2$, and $\inf_{\theta\in\Omega} g_\alpha(\theta) = cd/2 - cr\delta$. Mimicking the previous steps with $r = 1/2$, we obtain the lower bound

$$\rho(g_\alpha, g_\beta) \geq \frac{cr\delta}{2} \quad \forall \alpha \neq \beta \in \mathcal{V}.$$

The rest of the proof above did not depend on Ω, so that we again obtain the lower bound $T \geq c_0 \frac{d}{\delta^2}$ or $T \geq \frac{c_0}{\delta^2}$ depending on the oracle used, for a universal constant c_0. In this case, the difference in ρ computation means that $\epsilon = \frac{G\delta r}{36} \leq \frac{Gr}{144}$, from which the general claims follow.

3.3.3 Proof of Theorem 3.2

We now turn to the proof of lower bounds on the oracle complexity of the class of strongly convex functions from Definition 3.3. In this case, we work with the following family of base functions, parametrized by a scalar $\varphi \in [0, 1)$:

$$f_i^+(\theta) = r\varphi|\theta(i) + r| + \frac{(1-\varphi)}{4}\left(\theta(i) + r\right)^2, \quad \text{and} \quad f_i^-(\theta) = r\varphi|\theta(i) - r| + \frac{(1-\varphi)}{4}\left(\theta(i) - r\right)^2. \tag{3.30}$$

A key ingredient of the proof is a uniform lower bound on the discrepancy ρ between pairs of these functions:

Lemma 3.5. *Using an ensemble based on the base functions (3.30), we have*

$$\rho(g_\alpha, g_\beta) \geq \begin{cases} \frac{2c\delta^2 r^2}{(1-\varphi)d}\,\Delta_H(\alpha,\beta) & \text{if } 1-\varphi \geq \frac{4\delta}{1+2\delta} \\ \frac{c\delta r^2}{d}\,\Delta_H(\alpha,\beta) & \text{if } 1-\varphi < \frac{4\delta}{1+2\delta}. \end{cases} \tag{3.31}$$

The proof of this lemma is provided in Appendix A.1. Let us now proceed to the proofs of the main theorem claims.

Part (a)—Proof for $p = 1$: We observe that both the functions f_i^+, f_i^- are r-Lipschitz with respect to the $\|\cdot\|_1$ norm by construction. Hence, g_α is cr-Lipschitz and furthermore, by the definition of Oracle A, we have $\mathbb{E}\,\|\widehat{v}_{\alpha,A}(\theta)\|_1^2 \leq c^2r^2$. In addition, the function g_α is $(1-\varphi)c/(4d)$-strongly convex with respect to the Euclidean norm. We now follow the same steps as the proof of Theorem 3.1, but this time exploiting the ensemble formed by the base functions (3.30), and the lower bound on the discrepancy $\rho(g_\alpha, g_\beta)$ from Lemma 3.5. We split our analysis into two sub-cases.

Case 1: First suppose that $1 - \varphi \geq 4\delta/(1+2\delta)$, in which case Lemma 3.5 yields the lower bound

$$\rho(g_\alpha, g_\beta) \;\geq\; \frac{2c\delta^2 r^2}{(1-\varphi)d}\Delta_H(\alpha,\beta) \;\overset{(i)}{\geq}\; \frac{c\delta^2 r^2}{2(1-\varphi)} \quad \forall \alpha \neq \beta \in \mathcal{V},$$

where inequality (i) uses the fact that $\Delta_H(\alpha, \beta) \geq d/4$ by definition of $\mathcal{V}$. Hence by definition of Φ, we have established the lower bound $\Phi(\delta) \geq \frac{c\delta^2 r^2}{2(1-\varphi)}$. Setting the target error $\epsilon := c\delta^2 r^2/(18(1-\varphi))$, we observe that this ensures $\epsilon \leq \Phi(\delta)/9$. Recalling the requirement $\delta < 1/4$, we note that $\epsilon < cr^2/(288(1-\varphi))$. In this regime, we may apply Lemma 3.2 to obtain the upper bound $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$. Combining this upper bound with the lower bound (3.24) yields the inequality

$$\frac{1}{3} \geq 1 - 2\frac{16T\delta^2 + \log 2}{d \log(2/\sqrt{e})} \geq 1 - 2\frac{\frac{288T\epsilon(1-\varphi)}{cr^2} + \log 2}{d \log(2/\sqrt{e})}.$$

Simplifying the above expression yields that for $d \geq 11$, we have the lower bound

$$T \geq cr^2 \left(\frac{\frac{d}{3}\log(2/\sqrt{e}) - \log 2}{288\epsilon(1-\varphi)} \right) \geq cr^2 \frac{d \log(2/\sqrt{e})}{28800\epsilon(1-\varphi)}. \tag{3.32}$$

Finally, we observe that $G = cr$ and $\gamma_\ell^2 = (1-\varphi)c/(4d)$ which gives $1 - \varphi = 4dr\gamma_\ell^2/G$. Substituting the above relations in the lower bound (3.32) gives the first term in the stated result for $d \geq 11$.

To obtain lower bounds for dimensions $d < 11$, we use an argument based on $d = 1$. For this special case, we consider f^+ and f^- to be the two functions of the single coordinate coming out of definition (3.30). The packing set $\mathcal{V}$ consists of only two elements now, corresponding to $\alpha = 1$ and $\alpha = -1$. Specializing the result of Lemma 3.5 to this case, we see that the two functions are $2c\delta^2 r^2/(1-\varphi)$ separated. Now we again apply Lemma 3.2 to get an upper bound on the error probability and Lemma 3.4 to get a lower bound, which gives the result for $d \leq 11$.

Case 2: On the other hand, suppose that $1 - \varphi \leq 4\delta/(1+2\delta)$. In this case, appealing to Lemma 3.5 gives us that $\rho(g_\alpha, \beta) \geq c\delta r^2/4$ for $\alpha \neq \beta \in \mathcal{V}$. Recalling that $G = cr$, we set the desired accuracy $\epsilon := c\delta r^2/36 = G\delta r/36$. From this point onwards, we mimic the proof of Theorem 3.1; doing so yields that for all $\delta \in (0, 1/4)$, we have

$$T \geq c_0 \frac{d}{\delta^2} = c_0 \frac{G^2 d r^2}{\epsilon^2},$$

corresponding to the second term in Theorem 3.1 for a universal constant c_0.

Finally, the third and fourth terms are obtained just like Theorem 3.1 by checking the condition $\delta < 1/4$ in the two cases above. Overall, this completes the proof for the case $p = 1$.

Part (b)—Proof for $p > 2$: As with the proof of Theorem 3.1(b), we use Oracle B that returns d-dimensional values and gradients in this case, with the base functions defined in

equation 3.30. With this choice, we have the upper bound

$$\mathbb{E}\|\widehat{v}_{\alpha,B}(\theta)\|_p^2 \leq c^2 d^{2/p-2} r^2,$$

so that setting the constant $c = Gd^{1-1/p}/r$ ensures that $\mathbb{E}\|\widehat{v}_{\alpha,B}(\theta)\|_p^2 \leq G^2$. As before, we have the strong convexity parameter

$$\gamma_\ell^2 = \frac{c(1-\varphi)}{4d} = \frac{Gd^{-1/p}(1-\varphi)}{4r},$$

Also $\rho(g_\alpha, g_\beta)$ is given by Lemma 3.5. In particular, let us consider the case $1-\varphi \geq 4\delta/(1+2\delta)$ so that $\Phi(\delta) \geq \frac{c\delta^2 r^2}{2(1-\varphi)}$ and we set the desired accuracy $\epsilon := \frac{c\delta^2 r^2}{18(1-\varphi)}$ as before. With this setting of ϵ, we invoke Lemma 3.2 as before to argue that $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$. To lower bound the error probability, we appeal to Lemma 3.3 with $\ell = d$ just like Theorem 3.1(b) and obtain the inequality

$$\frac{1}{3} \geq 1 - 2\frac{16\, d\, T\, \delta^2 + \log 2}{d \log(2/\sqrt{e})}.$$

Rearranging terms and substituting $\epsilon = \frac{c\delta^2 r^2}{18(1-\varphi)}$, we obtain for $d \geq 11$

$$T \geq c_0 \left(\frac{1}{\delta^2}\right) = c_0 \left(\frac{cr^2}{\epsilon(1-\varphi)}\right),$$

for a universal constant c_0. The stated result can now be attained by recalling $c = Gd^{1-1/p}/r$ and $\gamma_\ell^2 = Gd^{-1/p}(1-\varphi)/r$ for $1-\varphi \geq 4\delta/(1+2\delta)$ and $d \geq 11$. For $d < 11$, the cases of $p > 2$ and $p = 1$ are identical up to constant factors in the lower bounds we state. This completes the proof for $1 - \varphi \geq 4\delta/(1 + 2\delta)$.

Finally, the case for $1-\varphi < 4\delta/(1+2\delta)$ involves similar modifications as part(a) by using the different expression for $\rho(g_\alpha, g_\beta)$. Thus we have completed the proof of this theorem.

3.3.4 Proof of Theorem 3.3

We begin by constructing an appropriate subset of $\mathcal{F}_{\text{sp}}(s)$ over which the Fano method can be applied. Let $\mathcal{V}(s) := \{\alpha^1, \ldots, \alpha^M\}$ be a set of vectors, such that each $\alpha^j \in \{-1, 0, +1\}^d$ satisfies

$$\|\alpha^j\|_0 = s \quad \text{for all } j = 1, \ldots, M, \text{ and} \qquad \Delta_H(\alpha^j, \alpha^\ell) \geq \frac{s}{2} \quad \text{for all } j \neq \ell.$$

It can be shown that there exists such a packing set with $|\mathcal{V}(s)| \geq \exp\left(\frac{s}{2}\log\frac{d-s}{s/2}\right)$ elements (e.g., see Lemma 5 in Raskutti et al. [134]).

For any $\alpha \in \mathcal{V}(s)$, we define the function

$$g_\alpha(\theta) := c\left[\sum_{i=1}^{d}\left\{\left(\frac{1}{2}+\alpha_i\delta\right)\Big|\,\theta(i)+r\,\Big|+\left(\frac{1}{2}-\alpha_i\delta\right)|\theta(i)-r|\right\}+\delta\sum_{i=1}^{d}|\theta(i)|\right]. \tag{3.33}$$

In this definition, the quantity $c > 0$ is a pre-factor to be chosen later, and $\delta \in (0, \frac{1}{4}]$ is a given error tolerance. Observe that each function $g_\alpha \in \mathcal{G}(\delta; s)$ is convex, and Lipschitz with parameter c with respect to the $\|\cdot\|_\infty$ norm.

Central to the remainder of the proof is the function class $\mathcal{G}(\delta; s) := \{g_\alpha,\ \alpha \in \mathcal{V}(s)\}$. In particular, we need to control the discrepancy $\Phi(\delta; s) := \Phi(\mathcal{G}(\delta; s))$ for this class. The following result, proven in Appendix A.2, provides a suitable lower bound:

Lemma 3.6. *We have*

$$\Phi(\delta; s) \;=\; \inf_{\alpha\neq\beta\in\mathcal{V}(s)} \rho(g_\alpha, g_\beta) \geq \frac{cs\delta r}{4}. \tag{3.34}$$

Using Lemma 3.6, we may complete the proof of Theorem 3.3. Define the base functions

$$f_i^+(\theta) \;:=\; d\left(|\theta(i)+r| + \delta|\theta(i)|\right), \qquad \text{and} \qquad f_i^-(\theta) := d\left(|\theta(i)-r| + \delta|\theta(i)|\right).$$

Consider Oracle B, which returns d-dimensional gradients based on the function

$$\widehat{g}_{\alpha,B}(\theta) = \frac{c}{d}\sum_{i=1}^{d}\left[b_i f_i^+(\theta) + (1-b_i) f_i^-(\theta)\right],$$

where $\{b_i\}$ are Bernoulli variables. By construction, the function $\widehat{g}_{\alpha,B}$ is at most $3c$-Lipschitz in ℓ_∞ norm (i.e. $\|\widehat{v}_{\alpha,B}(\theta)\|_\infty \leq 3c$), so that setting $c = \frac{G}{3}$ yields a G-Lipschitz function.

Our next step is to use Fano's inequality [52] to lower bound the probability of error in the multiway testing problem associated with this stochastic oracle, following an argument similar to (but somewhat simpler than) the proof of Lemma 3.3. Fano's inequality yields the lower bound

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{\frac{1}{\binom{|\mathcal{V}|}{2}}\sum_{\alpha\neq\beta} D(\mathbb{P}_\alpha \,\|\mathbb{P}_\beta) + \log 2}{\log|\mathcal{V}|}. \tag{3.35}$$

(As in the proof of Lemma 3.3, we have used convexity of mutual information [52] to bound it by the average of the pairwise KL divergences.) By construction, any two parameters $\alpha, \beta \in \mathcal{V}$ differ in at most $2s$ places, and the remaining entries are all zeroes in both vectors. The proof of Lemma 3.3 shows that for $\delta \in [0, \frac{1}{4}]$, each of these $2s$ places makes a contribution of at most $16\delta^2$. Recalling that we have T samples, we conclude that $D(\mathbb{P}_\alpha \,\|\mathbb{P}_\beta) \leq 32sT\delta^2$.

Substituting this upper bound into the Fano lower bound (3.35) and recalling that the cardinality of $\mathcal{V}$ is at least $\exp\left(\frac{s}{2}\log\frac{d-s}{s/2}\right)$, we obtain

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq 1 - 2\left(\frac{32sT\delta^2 + \log 2}{\frac{s}{2}\log\frac{d-s}{s/2}}\right) \tag{3.36}$$

By Lemma 3.6 and our choice $c = G/3$, we have

$$\Phi(\delta) \geq \frac{cs\delta r}{4} = \frac{Gs\delta r}{12},$$

Therefore, if we aim for the target error $\epsilon = \frac{Gs\delta r}{108}$, then we are guaranteed that $\epsilon \leq \frac{\Phi(\delta)}{9}$, as is required for the application of Lemma 3.2. Recalling the requirement $\delta \leq 1/4$ gives $\epsilon \leq Gs\delta r/432$. Now Lemma 3.2 implies that $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq 1/3$, which when combined with the earlier bound (3.36) yields

$$\frac{1}{3} \geq 1 - 2\left(\frac{32sT\delta^2 + \log 2}{\frac{s}{2}\log\frac{d-s}{s/2}}\right).$$

Rearranging yields the lower bound

$$T \geq c_0\left(\frac{\log\frac{d-s}{s/2}}{\delta^2}\right) = c_0\left(G^2r^2\, s^2\,\frac{\log\frac{d-s}{s/2}}{\epsilon^2}\right),$$

for a universal constant c_0, where the second step uses the relation $\delta = \frac{108\epsilon}{Gsr}$ for $k, d \geq 11$. As long as $s \leq \lfloor d/2 \rfloor$, we have $\log\frac{d-s}{s/2} = \Theta\left(\log\frac{d}{s}\right)$, which gives the result for $k, d \geq 11$. The result for $k, d \leq 11$ follows Theorem 3.1(b) applied with $p = \infty$, completing the proof.

3.4 Discussion

In this chapter, we have studied the complexity of convex optimization within the stochastic first-order oracle model. We derived lower bounds for various function classes, including convex functions, strongly convex functions, and convex functions with sparse optima. As we discussed, our lower bounds are sharp in general, since there are matching upper bounds achieved by known algorithms, among them stochastic gradient descent and stochastic mirror descent. Our bounds also reveal various dimension-dependent and geometric aspects of the stochastic oracle complexity of convex optimization. An interesting aspect of our proof technique is the use of tools common in statistical minimax theory. In particular, our proofs are based on constructing packing sets, defined with respect to a pre-metric that measures

how the degree of separation between the optima of different functions. We then leveraged information-theoretic techniques, in particular Fano's inequality and its variants, in order to establish lower bounds.

There are various directions for future research. It would be interesting to consider the effect of memory constraints on the complexity of convex optimization, or to derive lower bounds for problems of distributed optimization. We suspect that the proof techniques developed in this chapter may be useful for studying these related problems.

Chapter 4

Oracle inequalities for computationally adaptive model selection

The setup of this chapter will be in the decision-theoretic framework discussed in Chapter 2. The goal will be to minimize an expected risk criterion (2.1), based on the data distribution $\mathbb{P}$ and loss function ℓ. As opposed to the notation of Chapter 2, we will denote the parameter being estimated by f instead of θ, to reflect that a functional mapping, rather than a finite dimensional parameter is being estimated. Specifically, we will assume that the learner receives samples $\{z_1, \ldots, z_n\} \subseteq \mathcal{Z}$ drawn i.i.d. from some unknown distribution $\mathbb{P}$ over a sample space $\mathcal{Z}$, and given a loss function ℓ, seeks a function f to minimize the risk

$$R(f) := \mathbb{E}[\ell(z, f)]. \tag{4.1}$$

Since $R(f)$ is unknown, the typical approach is to compute estimates based on the empirical risk, $\widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(z_i, f)$, over a function class $\mathcal{F}$. Through this, we seek a function f_n with a risk close to the Bayes risk, the minimal risk over all measurable functions, which is $R_0 := \inf_f R(f)$. There is a natural trade-off based on the class $\mathcal{F}$ one chooses, since

$$R(f_n) - R_0 = \left(R(f_n) - \inf_{f\in\mathcal{F}} R(f)\right) + \left(\inf_{f\in\mathcal{F}} R(f) - R_0\right),$$

which decomposes the excess risk of f_n into estimation error (left) and approximation error (right).

A common approach to addressing this trade-off is to express $\mathcal{F}$ as a union of classes

$$\mathcal{F} = \bigcup_{j\geq 1} \mathcal{F}_j. \tag{4.2}$$

The *model selection problem* is to choose a class $\mathcal{F}_i$ and a function $f \in \mathcal{F}_i$ to give the best trade-off between estimation error and approximation error. This classical problem of model selection will be revisited in this chapter. In keeping with the theme of the thesis, we examine the problem from a fresh viewpoint taking into account both the computational and statistical complexities of our model selection procedure. The result is a new framework, and computationally efficient algorithms with sharp oracle inequalities within that framework. We start by motivating our setup in which we study the problem, before moving on to the technical content.

4.1 Motivation and setup

A common approach to the model selection problem is the now classical idea of *complexity regularization*, which arose out of early works by Mallows [106] and Akaike [9]. The complexity regularization approach balances two competing objectives: the minimum empirical risk of a model class $\mathcal{F}_i$ (approximation error) and a complexity penalty (to control estimation error) for the class. Different choices of the complexity penalty give rise to different model selection criteria and algorithms (for example, see the lecture notes by Massart [108] and the references therein). The complexity regularization approach uses penalties $\gamma_i : \mathbb{N} \to \mathbb{R}_+$ associated with each class $\mathcal{F}_i$ to perform model selection, where $\gamma_i(n)$ is a complexity penalty for class i when n samples are available; usually the functions γ_i decrease to zero in n and increase in the index i. The actual algorithm is as follows: for each i, choose

$$\hat{f}_i \in \operatorname*{argmin}_{f \in \mathcal{F}_i} \widehat{R}_n(f) \quad \text{and select} \quad \widetilde{f}_n = \operatorname*{argmin}_{i=1,2,\ldots} \left\{ \widehat{R}_n(\hat{f}_i) + \gamma_i(n) \right\} \tag{4.3}$$

as the output of the model selection procedure. Results of several authors [18, 104, 108] show that given a dataset of size n, the output $\widetilde{f}_n$ of the procedure roughly satisfies

$$\mathbb{E}R(\widetilde{f}_n) - R_0 \leq \min_i \left[\inf_{f \in \mathcal{F}_i} R(f) - R_0 + \gamma_i(n) \right] + \mathcal{O}\left(\frac{1}{\sqrt{n}} \right). \tag{4.4}$$

Several approaches to complexity regularization are possible, and an incomplete bibliography includes [169, 69, 138, 14, 18, 104].

These oracle inequalities show that, for a given sample size, the model selection procedure gives the best trade-off between the approximation and estimation errors. A drawback with the above mentioned approaches is that we need to be able to optimize over each model in the hierarchy on the entire data, in order to prove guarantees on the result of the model selection procedure. This is natural when the sample size is the key limitation, and it is computationally feasible when the sample size is small and the samples are low-dimensional. However, the cost of fitting a large number of model classes on the entire data sequence can be prohibitive when the datasets become large and high-dimensional as is common in

modern settings. In these cases, it is the computational resources—rather than the sample size—that form the key constraint. In this chapter, we consider model selection from this computational perspective, viewing the amount of computation, rather than the sample size, as the parameter which will enter our oracle inequalities. Specifically, we consider model selection methods that work within a given computational budget.

An interesting and difficult aspect of the problem that we must address is the interaction between model class complexity and computation time. It is natural to assume that for a fixed sample size, it is more expensive to estimate a model from a complex class than a simple class. Put inversely, given a computational bound, a simple model class can fit a model to a much larger sample size than a rich model class. So any strategy for model selection under a computational budget constraint should trade off two criteria: (i) the relative training cost of different model classes, which allows simpler classes to receive far more data (thus making them resilient to overfitting), and (ii) lower approximation error in the more complex model classes.

In addressing these computational and statistical issues, this chapter makes two main contributions. First, we propose a novel computational perspective on the model selection problem, which we believe should be a natural consideration in statistical learning problems. Secondly, within this framework, we provide algorithms for model selection in many different scenarios, and provide oracle inequalities on their estimates under different assumptions. Our first two results address the case where we have a model hierarchy that is ordered by inclusion, that is, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \ldots$. The first result provides an inequality that is competitive with an oracle, incurring at most an additional logarithmic penalty in the computational budget. The second result extends our approach to obtaining fast rates for model selection, as demonstrated in computationally unconstrained settings by Bartlett [19] and Koltchinskii [91]. Both of our results carefully refine the existing complexity-regularized risk minimization techniques by a careful consideration of the structure of the problem. Our third result applies to model classes that do not necessarily share any common structure. Here we present a novel algorithm —exploiting algorithms for multi-armed bandit problems—that uses confidence bounds based on concentration inequalities to select a good model under a given computational budget. We also prove a minimax optimal oracle inequality on the performance of the selected model. All of our algorithms are computationally simple and efficient. We note that the results of this chapter appeared in the paper [5].

The remainder of this chapter is organized as follows. We start in Section 4.2 by presenting our setup, estimator and oracle inequalities for a nested hierarchy of models. In Section 4.3 we refine the estimator and its analysis further to obtain fast rates for model selection in favorable conditions. The setting of unstructured model collections is studied in Section 4.4. Detailed technical arguments and various auxiliary results needed to establish our main theorems and corollaries can be found in Appendices B.1-B.4.

4.2 Model selection over nested hierarchies

In many practical scenarios, the family of models with which one works has some structure. One of the most common model selection settings has the model classes $\mathcal{F}_i$ are ordered by inclusion with increasing complexity. In this section, we study such model selection problems; we begin by formally stating our assumptions and giving a few natural examples, proceeding thereafter to oracle inequalities for a computationally efficient model selection procedure.

4.2.1 Assumptions

Our first main assumption is a natural inclusion assumption, which is perhaps the most common assumption in prior work on model selection (e.g. [18, 104]):

Assumption A. The function classes $\mathcal{F}_i$ satisfy an inclusion hierarchy:

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \dots \tag{4.5}$$

We provide two examples of such problems in the next section. In addition to the inclusion assumption, we make a few assumptions on the computational aspects of the problem. Most algorithms used in the framework of complexity regularization rely on the computation of estimators of the form

$$\widehat{f}_i = \operatorname*{argmin}_{f \in \mathcal{F}_i} \widehat{R}_n(f), \tag{4.6}$$

either exactly or approximately, for each class i. Since the model classes are ordered by inclusion, it is natural to assume that the computational cost of computing an empirical risk minimizer from $\mathcal{F}_i$ is higher than that for a class $\mathcal{F}_j$ when $i > j$. Said differently, given a fixed computational budget T, it may be impossible to use as many samples to compute an estimator from $\mathcal{F}_i$ as it is to compute an estimator from $\mathcal{F}_j$ (again, when $i > j$). We formalize this in the next assumption, which is stated in terms of an (arbitrary) algorithm $\mathcal{A}$ that selects functions $f \in \mathcal{F}_i$ for each index i based on a set of n_i samples.

Assumption B. Given a computational budget T, there is a sequence $\{n_i(T)\}_i \subset \mathbb{N}$ such that

(a) $n_i(T) > n_j(T)$ for $i < j$.

(b) The complexity penalties γ_i satisfy $\gamma_i(n_i(T)) < \gamma_j(n_j(T))$ for $i < j$.

(c) For each class $\mathcal{F}_i$, the computational cost of using the algorithm $\mathcal{A}$ with $n_i(T)$ samples is T. That is, estimation within class $\mathcal{F}_i$ using $n_i(T)$ samples has the same computational complexity for each i.

(d) For all i, the output $\mathcal{A}(i, T)$ of the algorithm $\mathcal{A}$, given a computational budget T, satisfies

$$\widehat{R}_{n_i(T)}(\mathcal{A}(i,T)) - \inf_{f \in \mathcal{F}_i} \widehat{R}_{n_i(T)}(f) \leq \gamma_i(n_i(T)).$$

(e) As $i \uparrow \infty$, $\gamma_i(n) \to \infty$ for any fixed n.

The first two assumptions formalize a natural notion of computational budget in the context of our model selection problem: given equal computation time, a simpler model can be fit using a larger number of samples than a complex model. Assumption B(c) says that the number of samples $n_i(T)$ is chosen to roughly equate the computational complexity of estimation within each class. Assumption B(d) simply states that we compute approximate empirical minimizers for each class $\mathcal{F}_i$. Our choice of the accuracy of computation to be γ_i in part (d) is done mainly for notational convenience in the statements of our results; it can be replaced with an arbitrary accuracy level ϵ in general. Finally part (e) just rules out degenerate cases where the penalty function asymptotes to a finite upper bound, and this assumption is required for our estimator to be well-defined for infinite model hierarchies. In the sequel, we use the shorthand $\gamma_i(T)$ to denote $\gamma_i(n_i(T))$ when the number of samples $n_i(T)$ is clear from context.

Certainly many choices are possible for the penalty functions γ_i, and work studying appropriate penalties is classical [9, 106]. Our focus in this chapter is on complexity estimates derived from concentration inequalities, which have been deeply studied by a variety of researchers [18, 108, 15, 19, 90]. Such complexity estimates are convenient since they ensure that the penalized empirical risk bounds the true risk with high probability. Formally, we have

Assumption C. For each i, there are constants $\kappa_1, \kappa_2 > 0$ such that for any budget T the output $\mathcal{A}(i, T) \in \mathcal{F}_i$ satisfies,

$$\mathbb{P}\left(|\widehat{R}_{n_i(T)}(\mathcal{A}(i,T)) - R(\mathcal{A}(i,T))| > \gamma_i(T) + \kappa_2\epsilon\right) \leq \kappa_1 \exp(-4n_i(T)\epsilon^2). \tag{4.7}$$

In addition, for any fixed function $f \in \mathcal{F}_i$, $\mathbb{P}(|\widehat{R}_{n_i(T)}(f) - R(f)| > \kappa_2\epsilon) \leq \kappa_1 \exp(-4n_i(T)\epsilon^2)$.

4.2.2 Some illustrative examples

We now provide two concrete examples to illustrate Assumptions A–C.

Example 4.1 (Linear classification with nested balls)**.** In a classification problem, each sample z_i consists of a covariate vector $x \in \mathbb{R}^d$ and label $y \in \{-1, +1\}$. In margin-based linear classification, the predictions are the sign of the linear function $f_\theta(x) = \langle \theta, x \rangle$, where $\theta \in \mathbb{R}^d$. A natural sequence of model classes is sets $\{f_\theta\}$ indexed via norm-balls of increasing radii: $\mathcal{F}_i = \{f_\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq r_i\}$, where $0 \leq r_1 < r_2 < \ldots$. By inspection, $\mathcal{F}_i \subset \mathcal{F}_{i+1}$ so that this sequence satisfies Assumption A.

The empirical and expected risks of a function f_θ are often measured using the sample average and expectation, respectively, of a convex upper bound on the 0-1 loss $\mathbb{I}(yf_\theta(x) \leq 0)$. Examples of such losses include the hinge loss, $\ell(yf_\theta(x)) = \max(0, 1 - yf_\theta(x))$, or the logistic

loss, $\ell(yf_\theta(x)) = \log(1+\exp(-yf_\theta(x)))$. Assume that $\mathbb{E}[\|x\|_2^2] \le X^2$ and let σ_i be independent uniform $\{\pm 1\}$-valued random variables. Then we may use a penalty function γ_i based on Rademacher complexity $\mathfrak{R}_n(\mathcal{F}_i)$ of the class i,

$$\mathfrak{R}_n(\mathcal{F}_i) := \left\{ \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_i} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| \right] \right\} \le \frac{2r_i X}{\sqrt{n}}.$$

Setting γ_i to be the Rademacher complexity $\mathfrak{R}_n(\mathcal{F}_i)$ satisfies the conditions of Assumption C [17] for both the logistic and the hinge losses which are 1-Lipschitz. Hence, using the Lipschitz contraction bound from [17, Theorem 12], we may take $\gamma_i(T) = \frac{r_i X}{\sqrt{n_i(T)}}$.

To illustrate Assumption B, we take stochastic gradient descent [139] as an example. Assuming that the computation time to process a sample z is equal to the dimension d, then Nemirovski et al. [117] show that the computation time required by this algorithm to output a function $f = \mathcal{A}(i, T)$ satisfying Assumption B(d) (that is, a γ_i-optimal empirical minimizer) is at most

$$\frac{4r_i^2 X^2}{\gamma_i^2(T)} \cdot d.$$

Substituting the bound on $\gamma_i(T)$ above, we see that the computational time for class i is at most $dn_i(T)$. In other words, given a computational time T, we can satisfy the Assumption B by setting $n_i(T) \propto T/d$ for each class i—the number of samples remains constant across the hierarchy in this example.

Example 4.2 (Linear classification in increasing dimensions)**.** Staying within the linear classification domain, we index the complexity of the model classes $\mathcal{F}_i$ by an increasing sequence of dimensions $\{d_i\} \subset \mathbb{N}$. Formally, we set

$$\mathcal{F}_i = \{f_\theta : \theta_j = 0 \text{ for } j > d_i, \quad \|\theta\|_2 \le r_i\},$$

where $0 \le r_1 < r_2 < \ldots$. This structure captures a variable selection problem where we have a prior ordering on the covariates.

In special scenarios, such as when the design matrix $X = [x_1\ x_2\ \cdots\ x_n]$ satisfies certain incoherence or irrepresentability assumptions [39], variable selection can be performed using ℓ_1-regularization or related methods. However, in general an oracle inequality for variable selection requires some form of exhaustive search over subsets. In the sequel, we show that in this simpler setting of variable selection over nested subsets, we can provide oracle inequalities without computing an estimator for each subset and without any assumptions on the design matrix X.

For this function hierarchy, we consider complexity penalties arising from VC-dimension arguments [168, 17], in which case we may set

$$\gamma_i(T) = \sqrt{\frac{d_i}{n_i(T)}}$$

which satisfies Assumption C. Using arguments similar to those for Example 4.1, we may conclude that the computational assumption B can be satisfied for this hierarchy, where the algorithm $\mathcal{A}$ requires time $d_i n_i(T)$ to select $f \in \mathcal{F}_i$. Thus, given a computational budget T, we set the number of samples $n_i(T)$ for class i to be proportional to T/d_i.

We provide only classification examples above since they demonstrate the essential aspects of our formulation. Similar quantities can also be obtained for a variety of other problems, such as parametric and non-parametric regression, and for a variety of model hierarchies including polynomial or Fourier expansions, wavelets, or Sobolev classes, among others (for more instances, see, e.g. [108, 15, 18]).

4.2.3 The computationally-aware model selection algorithm

Having specified our assumptions and given examples satisfying them, we turn to describing our first computationally-aware model selection algorithm. Let us begin with the simpler scenario where we have only K model classes (we extend this to infinite classes momentarily). Perhaps the most obvious computationally budgeted model selection procedure is the following: allocate a budget of T/K to each model class i. As a result, class i's estimator $\hat{f}_i = \mathcal{A}(i, T/K)$ is computed using $n_i(T/K)$ samples. Let $\widetilde{f}_n$ denote the output of the basic model selection algorithm (4.3) with the choices $n = n_i(T/K)$ and using $n_i(T/K)$ samples to evaluate the empirical risk for class i; very slight modifications of standard arguments [108, 18] yield the oracle inequality

$$R(\widetilde{f}_n) \leq \min_{i=1,\ldots,K} \left(R_i^* + c\gamma_i \left(\frac{T}{K} \right) + \sqrt{\frac{\log i}{n_i(T/K)}} \right),$$

where c is a universal constant. This approach can be quite poor. For instance, in Example 4.2, we have $n_i(T/K) = T/(Kd_i)$, and the above inequality incurs a penalty that grows as $\sqrt{K}$. This is much worse than the logarithmic scaling in K that is typically possible in computationally unconstrained settings [18]. It is thus natural to ask whether we can use the nesting structure of our model hierarchy to allocate computational budget more efficiently.

To answer this question, we introduce the notion of *coarse-grid sets*, which use the growth structure of the complexity penalties γ_i to construct a scheme for allocating the budget across the hierarchy. Recall the constant κ_2 from Assumption C and let $m > 0$ be an arbitrary constant (we will see that m controls the probability of error in our results). Given $s \in \mathbb{N}$ $(s \geq 1)$, we define

$$\overline{\gamma}_i(T, s) := 2\gamma_i \left(\frac{T}{s} \right) + \kappa_2 \sqrt{\frac{2(m + \log s)}{n_i(T/s)}}. \tag{4.8}$$

With the definition (4.8), we now give a definition characterizing the growth characteristics of the penalties and sample sizes.

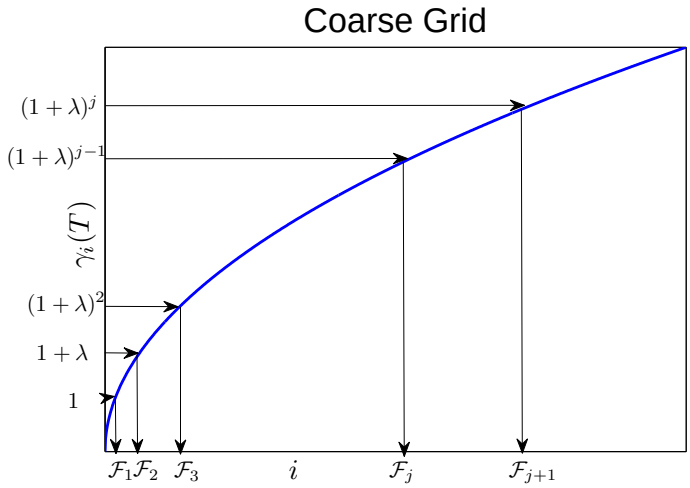

Figure 4.1. Construction of the coarse-grid set S_λ. The X-axis is the class index i, and the Y-axis represents the corresponding complexity $\gamma_i(T)$. When the penalty function grows steeply early on, we include a large number of models. The number of complex models included in S_λ can be significantly smaller as the growth of penalty function tapers out.

Definition 4.1. *Given a budget T, for a set $S \subseteq \mathbb{N}$, we say that S satisfies the* coarse grid condition *with parameters λ, m, and s if $|S| = s$ and for each i there is an index $j \in S$ such that*

$$\overline{\gamma}_i(T, s) \leq \overline{\gamma}_j(T, s) \leq (1+\lambda)\overline{\gamma}_i(T, s). \tag{4.9}$$

This definition of the coarse-grid set is pictorially illustrated in Figure 4.1. If the coarse-grid set is finite and, say, $|S| = s$, then the set S presents a natural collection of indices over which to perform model selection. We simply split the budget uniformly amongst the coarse-grid set S, giving budget T/s to each class in the set. Indeed, the main theorem of this section shows that for a large class of problems, it always suffices to restrict our attention to a finite grid set S, allowing us to present both a computationally tractable estimator and a good oracle inequality for the estimator. In some cases, there may be no finite coarse grid set. Thus we look for way to restrict our selection to finite sets, which we can do with the following assumption (the assumption is unnecessary if the hierarchy is finite).

Assumption D. There is a constant B such that $R(f)$ and $\widehat{R}_n(f)$ are both upper bounded by B for all sample sizes $n \in \mathbb{N}$ and for all $f \in \mathcal{F}_i$, $i = 1, 2, \ldots$.

Under the above assumption, we define the following coarse-grid size. Given $\lambda > 0$ and a constant $m > 0$, let $s(\lambda)$ be any solution to the inequality

$$s(\lambda) \geq \left\lceil \frac{\log\left(1 + \frac{B}{\overline{\gamma}_1(T, s(\lambda))}\right)}{\log(1+\lambda)} \right\rceil + 2. \tag{4.10}$$

Algorithm 1 Computationally budgeted model selection over nested hierarchies

Input: Model hierarchy $\{\mathcal{F}_i\}$ with corresponding penalty functions γ_i, computational budget T, bound B on risk of class 1, and scale factor $\lambda > 0$.

Construction of the coarse-grid set S_λ.

Set $s(\lambda)$ based on inequality (4.10).

for $k = 0$ to $s(\lambda) - 1$ **do**

Set j_{k+1} to be the largest class for which $\overline{\gamma}_j(T/s(\lambda)) \leq (1+\lambda)^k \overline{\gamma}_1(T/s(\lambda))$.

end for

Set $S_\lambda = \{j_k \ : \ k = 1, \ldots, s(\lambda)\}$.

Model selection estimate

Set $\hat{f}_i = \mathcal{A}\left(i, T/s(\lambda)\right)$ for $i \in S_\lambda$.

Select a class $\widehat{i}$ that satisfies

$$\widehat{i} \in \operatorname*{argmin}_{i \in S} \left\{ \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) + \gamma_i\left(T/s(\lambda)\right) + \frac{\kappa_2}{2}\sqrt{\frac{m}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2}\sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} \right\}. \tag{4.11}$$

Output the function $f = \hat{f}_{\widehat{i}} = \mathcal{A}\left(\widehat{i}, T/s(\lambda)\right)$.

To see that such an $s(\lambda)$ exists, note that taking $s(\lambda) \uparrow T$, so that $T/s(\lambda) = \Theta(1)$ yields that $\overline{\gamma}_1(T, s(\lambda)) = \Omega(\sqrt{\log T})$; meaning that the left side is larger (we assume for the remainder that T is suitably large that the inequality (4.10) has a solution). The intuition behind the definition is the following. Under assumption B(e), the complexity penalties continue increase with the class index i. Hence, there is a class $K(\lambda)$ such that the complexity of penalty $\gamma_{K(\lambda)}$ is larger than the penalized risk of the smallest class 1, at which point no class larger than $K(\lambda)$ can be a minimizer in the oracle inequality. As we show in the proof of Theorem 4.1 to follow, our choice of $s(\lambda)$ ensures that there is at least one class $j \in S_\lambda$ such that $j \geq K(\lambda)$, allowing us to restrict our attention only to the function classes $\{\mathcal{F}_i \mid i \in S_\lambda\}$. Using the setting (4.10) of $s(\lambda)$, we provide our computationally budgeted model selection procedure in Algorithm 1.

4.2.4 Main result and some consequences

With the above definitions in place, we can now provide an oracle inequality on the performance of the model selected by Algorithm 1. We start with our main theorem, and then provide corollaries to help explain various aspects of it.

Theorem 4.1. *Let $f = \mathcal{A}(\widehat{i}, T/s(\lambda))$ be the output of the algorithm $\mathcal{A}$ for the class $\widehat{i}$ specified by the procedure (4.11). Let Assumptions A–D be satisfied. With probability at least* $1 -$

$2\kappa_1 \exp(-m)$

$$R(f) \leq \min_{i=1,2,3,\ldots} \left\{ R_i^* + (1+\lambda) \left(2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2m + \log s(\lambda)}{n_i(T/s(\lambda))}} \right) \right\}. \tag{4.12}$$

Remarks: Theorem 4.1 provides a family of results, one for each value of $\lambda > 0$, and the corresponding setting of $s(\lambda)$. In order to obtain the best oracle inequality, one may take the minimum over $\lambda > 0$, though we do not discuss such a choice. We turn to a few additional explanatory remarks.

(a) To better understand the result of Theorem 4.1, we ask what an omniscient oracle with access to the same computational algorithm $\mathcal{A}$ could do. Such an oracle would know the optimal class i^* and allocate the entire budget T to compute $\mathcal{A}(i^*, T)$. By Assumption C, the output f of this oracle satisfies with probability at least $1 - \kappa_1 \exp(-m)$

$$R(f) \leq R_{i^*}^* + \gamma_{i^*}(T) + \kappa_2 \sqrt{\frac{m}{n_{i^*}(T)}} = \min_{i=1,2,3,\ldots} \left\{ R_i^* + \gamma_i(T) + \kappa_2 \sqrt{\frac{m}{n_i(T)}} \right\}. \tag{4.13}$$

Comparing this to the right hand side of Theorem 4.1, we observe that (roughly) we incur a penalty in the computational budget of roughly a factor of $s(\lambda)$, since we do not know the optimal class. As long as $s(\lambda)$ is not too large, the result is not significantly worse than the oracle bound (4.13).

(b) In many interesting cases (such as Examples 4.1 and 4.2), the penalty $\gamma_1(T/s(\lambda))$ is inversely polynomial in T and $n_1(T/s(\lambda))$ is (sub)linear in T. Heuristically, we may write $\gamma_i(T/s) = (s/T)^p$ for some $p \in (0, \infty)$ (and similarly for $n_i(T/s)$), in which case the setting (4.10) for $s(\lambda)$ satisfies

$$\frac{\log\left(1 + \frac{B}{\overline{\gamma}_1(T,s)}\right)}{\log(1+\lambda)} \approx \frac{\log\left(1 + \frac{B}{(s/T)^p}\right)}{\log(1+\lambda)} \approx \frac{p\log(T/s) + \log B}{\log(1+\lambda)}.$$

Thus the setting (4.10) for $s(\lambda)$ is $\mathcal{O}(\log T)$. Comparing inequalities (4.12) and (4.13) in this setting, we see that we match the performance of the oracle up to polylogarithmic factors in the budget T. More generally, we can always use $\overline{\gamma}_1(T, 1)$ in the defining inequality (4.10) in place of $\overline{\gamma}_1(T, s(\lambda))$, which by inspection guarantees that $s(\lambda)$ is large enough to satisfy the inequality.

(c) Algorithm 1 and Theorem 4.1, as stated, require a priori knowledge of the computational budget T. We can address this using a standard doubling argument (see e.g. [47, Sec. 2.3]). Initially we assume $T = 1$ and run Algorithm 1 accordingly. If we do not exhaust the budget, we assume $T = 2$, and rerun Algorithm 1 for another round. If there is more

computational time at our disposal, we update our guess to $T = 4$ and so on. Suppose the real budget is T_0 with $2^k - 1 < T_0 \le 2^{k+1} - 1$. After i rounds of this doubling strategy, we have exhausted a budget of 2^{i-1}, with the last round getting a budget of 2^{i-2} for $i \ge 2$. In particular, the last round with a net budget of T_0 is of length at least $T_0/4$. Since Theorem 4.1 applies to each individual round, we obtain an oracle inequality where we replace T_0 with $T_0/4$; we can be agnostic to the prior knowledge of the budget at the expense of slightly worse constants.

(d) Finally, though we have assumed that the quantity B bounds the risks for all model classes $\mathcal{F}_i$, inspection of the proof of Theorem 4.1 and the setting (4.10) of $s(\lambda)$ require that B only upper bound the risk of the first class $\mathcal{F}_1$; worst-case risks may grow unboundedly for larger function classes $\mathcal{F}_i$.

Now let us turn to a specialization of Theorem 4.1 to the settings outlined in Examples 4.1 and 4.2. The following corollary shows that—roughly—our computational restrictions give oracle inequalities only logarithmically worse than those possible in the computationally unconstrained model selection procedure (4.3).

Corollary 4.1. *Let the conditions of Theorem 4.1 hold. Let $m \ge 0$ and $\lambda > 0$ be (specified) constants, and assume that T is large enough that $BnT \ge 1$ and $nT \ge 2$.*

(a) In the setting of Example 4.1, let nT/d denote the number of samples that can be processed by the inference algorithm $\mathcal{A}$ using T units of computation. Set

$$s(\lambda) = \left\lceil \frac{\log\left(1 + \frac{B\sqrt{Tn}}{\sqrt{d}\left(2r_1 X + \kappa_2\sqrt{2m}\right)}\right)}{\log(1+\lambda)} \right\rceil + 2.$$

Assume that $\kappa_2\sqrt{2md} \ge 1$. With probability at least $1 - 2\kappa_1\exp(-m)$, the output f of Algorithm 1 satisfies

$$R(f) \le \inf_{i=1,2,\ldots} \left\{ R_i^* + (1+\lambda)\sqrt{\frac{d\log(2BTn)}{Tn\log(1+\lambda)}}\left(2r_i X + \sqrt{2}\kappa_2\sqrt{m + \log\frac{\log(2BnT)}{\log(1+\lambda)}}\right)\right\}.$$

(b) In the setting of Example 4.2, let nT/d_i denote the number of samples that can be processed by the inference algorithm $\mathcal{A}$ using T units of computation. Set

$$s(\lambda) = \left\lceil \frac{\log\left(1 + \frac{B\sqrt{nT}}{\sqrt{d_1}(2\sqrt{d_1} + \kappa_2\sqrt{2m})}\right)}{\log(1+\lambda)} \right\rceil + 2$$

With probability at least $1 - 2\kappa_1\exp(-m)$, the output f of Algorithm 1 satisfies

$$R(f) \le \inf_{i=1,2,\ldots} \left\{ R_i^* + (1+\lambda)\sqrt{\frac{d_i\log(2BTn)}{Tn\log(1+\lambda)}}\left(2\sqrt{d_i} + \sqrt{2}\kappa_2\sqrt{m + \log\frac{\log(2BnT)}{\log(1+\lambda)}}\right)\right\}.$$

4.2.5 Proofs

For the proofs of Theorem 4.1 and Corollary 4.1, we require the additional notation

$$K(\lambda) = \max\{j \ : \ j \in S_\lambda\}. \tag{4.14}$$

We begin the proof of Theorem 4.1 by showing that the setting (4.10) of $s(\lambda)$ entails that any class $j > K(\lambda)$ must have penalty too large to be optimal, so we can focus on classes $j \leq K(\lambda)$. We then show that the output f of Algorithm 1 satisfies an oracle inequality for each class in S_λ, which is possible by an adaptation of arguments in prior work [18]. Using the definition (4.1) of our coarse grid set, we can then infer an oracle inequality that applies to each class $j \leq K(\lambda)$, and our earlier reduction to a finite model hierarchy completes the argument.

Proof of Theorem 4.1

The choice (4.10) of $s(\lambda)$ is based on the observation that once the complexity penalty of a class becomes too large, it can never be the minimizer of the penalized risk in the oracle inequality (4.12). Formally, we have the following lemma (see Appendix B.1 for a proof).

Lemma 4.1. *Recall the definition (4.14) of $K(\lambda)$ and let i^* be a class that attains the minimum in the right side of the bound (4.12). For any $\lambda > 0$ and $m > 0$, we have $i^* \leq K(\lambda)$.*

We also require a technical lemma that the selection of the set of j_k in Algorithm 1 satisfies Definition 4.1.

Lemma 4.2. *Let $\{\gamma_i\}$ be a sequence of increasing positive numbers and for each $k \in \{0, \ldots, s\}$ set j_{k+1} to be the largest index j such that $\gamma_j \leq (1+\lambda)^k \gamma_1$. Then for each i such that $i \leq j_k$, there exists a $j \in \{j_1, \ldots, j_k\}$ such that $\gamma_i \leq \gamma_j \leq (1+\lambda)\gamma_i$.*

Proof. Let $i \leq j_k$ and choose the smallest $j \in \{j_1, j_2, \ldots, j_k\}$ such that $\gamma_i \leq \gamma_j$. Assume for the sake of contradiction that $(1+\lambda)\gamma_i < \gamma_j$. There exists some $k \in \{0, \ldots, s\}$ such that $\gamma_j \leq (1+\lambda)^k \gamma_1$ and $\gamma_j > (1+\lambda)^{k-1}\gamma_1$, and thus we obtain

$$\gamma_i < \frac{\gamma_j}{1+\lambda} \leq (1+\lambda)^{k-1}\gamma_1.$$

Then $\gamma_i < (1+\lambda)^{k-1}\gamma_1$ so there is a j' (namely $j' = i$) with $j' < j$ satisfying $\gamma_{j'} \leq (1+\lambda)^{k-1}\gamma_1$; this contradicts the fact that j is the smallest index in $\{j_1, \ldots, j_k\}$ satisfying $\gamma_i \leq \gamma_j$. □

Equipped with the lemmas, we can restrict our attention only to classes $i \leq K(\lambda)$. To that end, the next result will establish an oracle inequality for our algorithm compared to all the classes in S_λ.

Proposition 4.1. Let $f = \mathcal{A}(\widehat{i}, n_{\widehat{i}}(T/s(\lambda)))$ be the output of the algorithm $\mathcal{A}$ for the class $\widehat{i}$ selected by the procedure (4.11). Under the conditions of Theorem 4.1, with probability at least $1 - 2\kappa_1 \exp(-m)$

$$R(f) \leq \min_{i \in S_\lambda} \left\{ R_i^* + 2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2m + \log s(\lambda)}{n_i(T/s(\lambda))}} \right\}.$$

The proof of the proposition follows from an argument similar to that given by Bartlett et al. [18], though we must carefully reason about the different number of independent samples used to estimate within each class $\mathcal{F}_i$. We present a proof in Appendix B.1. We can now complete the proof of Theorem 4.1 using the proposition.

Proof of Theorem 4.1 Let i be any class (not necessarily in S_λ) and $j \in S_\lambda$ be the smallest class satisfying $j \geq i$. Then, by construction of S_λ, we know from Lemma 4.2 that

$$\begin{aligned} 2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} &\leq 2\gamma_j \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_j(T/s(\lambda))}} \\ &\leq (1+\lambda) \left[2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} \right]. \end{aligned}$$

In particular, we can lower bound the penalized risk of class i as

$$R_i^* + 2(1+\lambda) \left[\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} \right] \geq R_j^* + 2\gamma_j \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_j(T/s(\lambda))}},$$

where we used the inclusion assumption A to conclude that $R_j^* \leq R_i^*$. Now applying Proposition 4.1, the above lower bound, and Lemma 4.1 in turn, we see that with probability at least $1 - 2\kappa_1 \exp(-m)$

$$\begin{aligned} R(f) &\leq \min_{i \in S_\lambda} \left\{ R_i^* + 2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} \right\} \\ &\leq \min_{i=1,2,\ldots,K(\lambda)} \left\{ R_i^* + (1+\lambda) \left(2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} \right) \right\} \\ &\leq \inf_{i=1,2,3,\ldots} \left\{ R_i^* + (1+\lambda) \left(2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} \right) \right\}. \end{aligned}$$

This is the desired statement of the theorem. □

Proof of Corollary 4.1

To establish the corollary, we must verify that the condition (4.10) is satisfied for our choice of $s(\lambda)$. For any $s \geq 1$, we have

$$\overline{\gamma}_1(T,1) \leq \overline{\gamma}_1(T,s) \quad \text{so} \quad \log\left(1 + \frac{B}{\overline{\gamma}_1(T,1)}\right) \geq \log\left(1 + \frac{B}{\overline{\gamma}_1(T,s)}\right). \tag{4.15}$$

Following the remarks after Theorem 4.1, we have in the conditions of Example 4.1 that

$$\overline{\gamma}_1(T,s) = 2r_1 X\sqrt{\frac{ds}{nT}} + \kappa_2\sqrt{\frac{2(m+\log s)ds}{nT}} \quad \text{or} \quad \overline{\gamma}_1(T,1) = 2r_1 X\sqrt{\frac{d}{nT}} + \kappa_2\sqrt{\frac{2md}{nT}},$$

whence the inequality (4.15) establishes that taking

$$s(\lambda) = \left\lceil \frac{\log\left(1 + \frac{B\sqrt{nT}}{2r_1 X\sqrt{d} + \kappa_2\sqrt{2md}}\right)}{\log(1+\lambda)} \right\rceil + 2$$

is sufficient to establish the condition (4.10). Similarly, in the context of Example 4.2, we have

$$\overline{\gamma}_1(T,1) = 2\frac{d_1}{\sqrt{nT}} + \kappa_2\sqrt{\frac{2md_1}{nT}} \quad \text{so} \quad s(\lambda) = \left\lceil \frac{\log\left(1 + \frac{B\sqrt{nT}}{\sqrt{d_1}(2\sqrt{d_1} + \kappa_2\sqrt{2m})}\right)}{\log(1+\lambda)} \right\rceil + 2$$

is sufficient for the inequality (4.10) to hold. Since we have assumed that $BnT \geq 1$ and $nT \geq 2$, we see that $1 + B\sqrt{nT} \leq 2BnT$, so that in both Examples 4.1 and 4.2

$$s(\lambda) \leq \frac{\log(2BnT)}{\log(1+\lambda)}.$$

Using this bound on $s(\lambda)$ completes the proof of the corollary.

4.3 Fast rates for model selection

Looking at the result given by Theorem 4.1, we observe that irrespective of the dependence of the penalties γ_i on the sample size, there are terms in the oracle inequality that always decay as $\mathcal{O}(1/\sqrt{n_i(T/s(\lambda))})$. Bartlett [19] notes a similar phenomenon for classical model selection results in computationally unconstrained settings, pointing out that under conditions similar to Assumption C, this inverse-root dependence on the number of samples is the best possible, due to lower bounds on the fluctuations of the empirical process (e.g. [20, Theorem 2.3]).

On the other hand, under suitable low noise conditions [107] or certain curvature properties of the risk functional [21, 91, 16], it is possible to obtain estimation guarantees of the form

$$R(\hat{f}) = R(f^*) + \mathcal{O}_p\left(\frac{1}{n}\right),$$

where $\hat{f}$ (approximately) minimizes the n-sample empirical risk. Bartlett [19] and Koltchinskii [90] demonstrate that under suitable assumptions, complexity regularization can also achieve fast rates for model selection. A natural question is thus whether similar results obtain in computationally constrained inference settings; we demonstrate in this section that under appropriate conditions, this question has an affirmative answer.

4.3.1 Assumptions and example

To study faster rates for model selection in the computationally bounded setting, we begin by modifying our concentration assumption and providing a motivating example.

Assumption E. For each i, let $f_i^* \in \operatorname{argmin}_{f\in\mathcal{F}_i} R(f)$. Then there are constants $\kappa_1, \kappa_2 > 0$ such that for any budget T and the corresponding sample size $n_i(T)$

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_i}\left(R(f) - R(f_i^*) - 2(\widehat{R}_{n_i(T)}(f) - \widehat{R}_{n_i(T)}(f_i^*))\right) > \gamma_i(T) + \kappa_2\epsilon\right] \le \kappa_1 \exp(-n_i(T)\epsilon). \tag{4.16a}$$

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_i}\left(\widehat{R}_{n_i(T)}(f) - \widehat{R}_{n_i(T)}(f_i^*) - 2(R(f) - R(f_i^*))\right) > \gamma_i(T) + \kappa_2\epsilon\right] \le \kappa_1 \exp(-n_i(T)\epsilon). \tag{4.16b}$$

Contrasting with the earlier assumption C, we see that the RHS above has a dependence on ϵ rather than ϵ^2 in the exponent, which leads to faster sub-exponential rates for sample complexity. Concentration inequalities of this form are now well known [21, 91, 16], and the paper [19] uses an identical assumption.

Before continuing, we give an example to elucidate the assumption.

Example 4.3 (Fast rates for classification)**.** We consider the function class hierarchy based on increasing dimensions of Example 4.2. We assume that the risk $R(f_\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$ and that the loss function ℓ is either the squared loss $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$ or the exponential loss from boosting $\ell(y, f_\theta(x)) = \exp(-y f_\theta(x))$. For each of these examples Assumption E is satisfied with

$$\gamma_i(T) = c\,\frac{d_i \log(n_i(T)/d_i)}{n_i(T)},$$

for a universal constant c. This follows from Theorem 3 of [19] (which in turn follows from Theorem 3.3 in [21] combined with an argument based on Dudley's entropy integral [63]). The other parameter settings and computational considerations are identical to those of Example 4.2.

If we define $\hat{f}_i = \mathcal{A}(i, T)$, then using Assumption B(d) (that $\widehat{R}_{n_i(T)}(\hat{f}_i) - \widehat{R}_{n_i(T)}(f_i^*) \leq \gamma_i(T)$) in conjunction with assumption (4.16a), we can conclude that for any time budget T, with probability at least $1 - \kappa_1 \exp(-m)$,

$$R(\hat{f}_i) \leq R(f_i^*) + 3\gamma_i(T) + \frac{\kappa_2 m}{n_i(T)}. \tag{4.17}$$

One might thus expect that by following arguments similar to those in Bartlett [19], it would be possible to show fast rates for model selection based on Algorithm 1. Unfortunately, the results of [19] heavily rely on the fact that the data used for computing the estimators $\hat{f}_i$ is same for each class i, so that the fluctuations of the empirical processes corresponding to the different classes are positively correlated. In our computationally constrained setting, however, each class's estimator is computed on a different set of $n_i(T)$ samples. It is thus more difficult to relate the estimators than in previous work, necessitating a modification of our earlier Algorithm 1 and a new analysis, which follows.

4.3.2 Algorithm and oracle inequality

As in Section 4.2, our approach is based on performing model selection over a coarsened version of the collection $\mathcal{F}_1, \mathcal{F}_2, \ldots$. To construct the coarser collection of indices, we define the composite penalty term (based on Assumption E)

$$\overline{\gamma}_i(T, s) := 20\gamma_i\left(\frac{T}{s}\right) + 8\frac{\kappa_2 m + 2\log s}{n_i(T/s)}. \tag{4.18}$$

Based on the above penalty term, we define our analogue of the coarse grid set (4.9) and the size (4.10) of the coarsening as any solution to the inequality

$$s(\lambda) \geq \left\lceil \frac{\log\left(1 + \frac{B}{s(\lambda)\overline{\gamma}_1(T, s(\lambda))}\right)}{\log(1+\lambda)} \right\rceil + 2. \tag{4.19}$$

We give our modified model selection procedure in Algorithm 2. In the algorithm and in our subsequent analysis, we use the shorthand $\widehat{R}_i(f)$ to denote the empirical risk of the function f on the $n_i(T)$ samples associated with class i. Our main convergence result is the following:

Theorem 4.2. *Let $f = \mathcal{A}(\widehat{i}, T/s(\lambda))$ be the output of the algorithm $\mathcal{A}$ for class $\widehat{i}$ specified by the procedure (4.20). Let Assumptions A, B, D and E be satisfied. With probability at least*

Algorithm 2 Computationally budgeted model selection over hierarchies with fast concentration

Input: Model hierarchy $\{F_i\}$ with corresponding penalty functions γ_i, computational budget T, bound B on the risk of class 1, and scale factor $\lambda > 0$.

Construction of the coarse-grid set S_λ.

Set $s(\lambda)$ satisfying inequality (4.19) with $\overline{\gamma}_i$ defined in (4.18).

for $k = 0$ to $s(\lambda) - 1$ **do**

Set j_{k+1} to be the largest class for which $\overline{\gamma}_j(T/s(\lambda)) \leq (1+\lambda)^k \overline{\gamma}_1(T/s(\lambda))$.

end for

Set $S_\lambda = \{j_k \ : \ k = 1, \ldots, s(\lambda)\}$.

Model selection estimate

Set $\hat{f}_i = \mathcal{A}\left(i, T/s(\lambda)\right)$ for $i \in S_\lambda$.

Select the class $\widehat{i} \in S_\lambda$ to be the largest class that satisfies

$$\widehat{R}_{\widehat{i}}(\hat{f}_{\widehat{i}}) + \frac{17}{2}\gamma_{\widehat{i}}\left(\frac{T}{s(\lambda)}\right) + \frac{7}{2}\kappa_2\left(\frac{m + \log s(\lambda)}{n_{\widehat{i}}(T/s(\lambda))}\right) \leq \widehat{R}_{\widehat{i}}(\hat{f}_j) + \frac{17}{2}\gamma_j\left(\frac{T}{s(\lambda)}\right) \tag{4.20}$$

for all $j \in S_\lambda$ such that $j < \widehat{i}$.

Output the function $\mathcal{A}\left(\widehat{i}, T/s(\lambda)\right)$.

$1 - 2\kappa_1 \exp(-m)$

$$R(f) \leq \inf_{i=1,2,3,\ldots}\left\{R_i^* + (1+\lambda)s(\lambda)\left(20\gamma_i\left(\frac{T}{s(\lambda)}\right) + 8\kappa_2\frac{(m + \log s(\lambda))}{n_i(T/s(\lambda))}\right)\right\}. \tag{4.21}$$

The following corollary shows the application of Theorem 4.2 to the classification problem we discuss in Example 4.3.

Corollary 4.2. *Let the conditions of Theorem 4.2 hold, assume $Bn \geq 1$ and let $m \geq 1$ and $\lambda > 0$ be (specified) constants. There exist universal constants c, c_1, c_2 such that in the setting of Example 4.3, setting*

$$s(\lambda) = \left\lceil \frac{\log\left(1 + \frac{BnT}{cd_1^2 m \log T}\right)}{\log(1+\lambda)} \right\rceil + 2$$

yields that with probability at least $1 - 4\kappa_1 \exp(-m)$, the estimator (4.20) satisfies

$$\begin{aligned} R(f) &\leq \inf_{i=1,2,\ldots}\left\{R_i^* + c_1(1+\lambda)s(\lambda)\left(\frac{d_i^2\, m \log T}{nT}\right)\right\} \\ &\leq \inf_{i=1,2,\ldots}\left\{R_i^* + c_2(1+\lambda)\left(\frac{d_i^2\, m \log^2(2BnT)}{nT\log(1+\lambda)}\right)\right\}. \end{aligned}$$

Corollary 4.2 makes clear that we have indeed achieved the stated goal of our this section: we have an oracle inequality whose dependence on the computational budget (and number of samples available) decreases—modulo logarithmic factors—at a rate of $1/T$.

4.3.3 Proofs of main results

In this section, we provide proofs of Theorem 4.2 and Corollary 4.2. The proof of Theorem 4.2 broadly follows that of Theorem 4.1, in that we establish an analogue of Proposition 4.1, which provides an oracle inequality for each class in the coarse-grid set S_λ. We then extend the proven inequality to apply to each function class $\mathcal{F}_i$ in the hierarchy using the definition (4.9) of the grid set.

Proof of Theorem 4.2 Let n_i be shorthand for $n_i(T/s(\lambda))$, the number of samples available to class i, and let $\widehat{R}_i(f)$ denote the empirical risk of the function f using the n_i samples for class i. In addition, let $\gamma_i(n_i)$ be shorthand for $\gamma_i(n_i(T/s(\lambda)))$, the penalty value for class i using $n_i(T/s(\lambda))$ samples. With these definitions, we adopt the following shorthand for the events in the probability bounds (4.16a) and (4.16b). Let $\epsilon = \{\epsilon_i\}$ be an $s(\lambda)$-dimensional vector with (arbitrary for now) positive entries. For each i define

$$\mathcal{E}_1^i(\epsilon_i) := \left\{ \sup_{f \in \mathcal{F}_i} \left(R(f) - R(f_i^*) - 2\left(\widehat{R}_i(f) - \widehat{R}_i(f_i^*) \right) \right) \leq \gamma_i(n_i) + \kappa_2 \epsilon_i \right\} \tag{4.22a}$$

$$\mathcal{E}_2^i(\epsilon_i) := \left\{ \sup_{f \in \mathcal{F}_i} \left(\widehat{R}_i(f) - \widehat{R}_i(f_i^*) - 2\left(R(f) - R(f_i^*) \right) \right) \leq \gamma_i(n_i) + \kappa_2 \epsilon_i \right\}, \tag{4.22b}$$

and define the joint events

$$\mathcal{E}_1(\epsilon) := \bigcup_{i \in S_\lambda} \mathcal{E}_1^i(\epsilon_i) \quad \text{and} \quad \mathcal{E}_2(\epsilon) := \bigcup_{i \in S_\lambda} \mathcal{E}_2^i(\epsilon_i) \,. \tag{4.23}$$

With the "good" events (4.23) defined, we turn to the two technical lemmas, which relate the risk of the chosen function $\hat{f}_{\widehat{i}}$ to f_i^* for each $i \in S$. To make the proofs of each of the lemmas cleaner and see the appropriate choices of constants, we replace the selection strategy (4.20) with one whose constants have not been specified. Specifically, we select $\widehat{i}$ as the largest class that satisfies

$$\widehat{R}_{\widehat{i}}(\hat{f}_{\widehat{i}}) + c_1 \gamma_{\widehat{i}}\left(\frac{T}{s(\lambda)} \right) + c_2 \kappa_2 \epsilon_{\widehat{i}} \leq \widehat{R}_{\widehat{i}}(\hat{f}_j) + c_1 \gamma_j \left(\frac{T}{s(\lambda)} \right) \tag{4.24}$$

for $j \in S$ with $j \leq \widehat{i}$. The proofs of these lemmas are included in Appendix B.2.

Lemma 4.3. *Let the events (4.22a) and (4.22b) hold for all $j \in S_\lambda$, that is, $\mathcal{E}_1(\epsilon)$ and $\mathcal{E}_2(\epsilon)$ hold. Then using the selection strategy (4.24), for each $j \leq \widehat{i}$ with $j \in S_\lambda$ we have*

$$R(\hat{f}_{\widehat{i}}) \leq R(f_j^*) + \frac{1}{2}\left[\left(\frac{17}{2} - c_1 \right) \gamma_{\widehat{i}}(n_{\widehat{i}}) + (6 + c_1)\gamma_j(n_j) + 2\kappa_2 \epsilon_j + \left(\frac{7}{2} - c_2 \right) \kappa_2 \epsilon_{\widehat{i}} \right].$$

We require a different argument for the case that $j \geq \widehat{i}$, and the constants are somewhat worse.

Lemma 4.4. *Let the events (4.22a) and (4.22b) hold for all $j \in S_\lambda$, that is, $\mathcal{E}_1(\epsilon)$ and $\mathcal{E}_2(\epsilon)$ hold. Assume also that $c_1 \geq 17/2$ and $c_2 \geq 7/2$. Then using the selection strategy (4.24), for each $j \geq \widehat{i}$ with $j \in S_\lambda$ we have*

$$R(\hat{f}_{\widehat{i}}) \leq R(f_j^*) + s(\lambda)\left((2c_1+3)\gamma_j\left(\frac{T}{s(\lambda)}\right) + (2c_2+1)\epsilon_j\right).$$

We now use Lemmas 4.3 and 4.4 to complete the proof of the theorem. When Assumption E holds, the probability that one of the events $\mathcal{E}_1(\epsilon)$ and $\mathcal{E}_2(\epsilon)$ fails to hold, by a union bound, is upper bounded by

$$\mathbb{P}(\mathcal{E}_1(\epsilon)^c \cup \mathcal{E}_2(\epsilon)^c) \leq \sum_{i \in S_\lambda} \mathbb{P}(\mathcal{E}_1^i(\epsilon_i)^c) + \sum_{i \in S_\lambda} \mathbb{P}(\mathcal{E}_2^i(\epsilon_i)^c) \leq 2\kappa_1 \sum_{i \in S_\lambda} \exp(-n_i(T/s(\lambda))\epsilon_i).$$

Thus, we see that if we define the constants

$$\epsilon_i = \frac{m + \log(s(\lambda))}{n_i(T/s(\lambda))},$$

we obtain that all of the events $\mathcal{E}_1^i(\epsilon_i)$ and $\mathcal{E}_2^i(\epsilon_i)$ hold with probability at least $1 - 2\kappa_1 \exp(-m)$. Applying Lemmas 4.3 and 4.4 with the choices $c_1 = \frac{17}{2}$ and $c_2 = \frac{7}{2}$, we obtain that with probability at least $1 - 2\kappa_1 \exp(-m)$

$$R(\hat{f}_{\widehat{i}}) \leq \min_{j \in S_\lambda} \left\{ R(f_j^*) + s(\lambda)\left(20\gamma_j(n_j) + 8\frac{m + \log(s(\lambda))}{n_j(T/s(\lambda))}\right)\right\}. \tag{4.25}$$

The inequality (4.25) is the analogue of Proposition 4.1 in the current setting. Given the inequality, the remainder of the proof of Theorem 4.2 follows the same recipe as that of Theorem 4.1. Recalling the notation (4.14) defining $K(\lambda)$, we apply the inequality (4.25) with the definition of the grid set (Definition 4.1)to obtain an oracle inequality compared to all classes $j \leq K(\lambda)$. Then the setting (4.19) of $s(\lambda)$ ensures that we can transfer the result to the entire model hierarchy as before. □

We complete this section with a proof of Corollary 4.2.

Proof of Corollary 4.2 As in the proof of Corollary 4.1, this proof relies on arguing that our setting of $s(\lambda)$ satisfies the inequality (4.19). Since in the setting of Example 4.3, we have

$$\overline{\gamma}_i(T,s) = c\left(\frac{d_i \log(n_i(T/s)/d_i) + m + 2\log s}{n_i(T/s)}\right),$$

we may take

$$s(\lambda) = \left\lceil \frac{\log\left(1 + \frac{B}{\overline{\gamma}_1(T,1)}\right)}{\log(1+\lambda)} \right\rceil + 2 = \left\lceil \frac{\log\left(1 + \frac{Bn_1(T)}{d_1 \log(n_1(T)/d_1)+m}\right)}{\log(1+\lambda)} \right\rceil + 2.$$

By our assumptions on $n_1(T)$ and B, we can upper bound $s(\lambda)$ by a constant c' so that $s(\lambda) \leq c' \log(BnT)/\log(1+\lambda)$. Recalling that $n_i(T) = nT/d_i$, we can then upper bound $\overline{\gamma}_i$ by

$$\begin{aligned}\overline{\gamma}_i\left(\frac{T}{s(\lambda)}\right) &\leq c_1\left(\frac{d_i^2 s(\lambda)\log(T/(d_i^2/s(\lambda))) + d_i s(\lambda)(m + 2\log s(\lambda))}{nT}\right) \\ &\leq c_2\left(\frac{d_i^2 \log^2(BnT) + 3d_i m \log^2(BnT)}{nT\log(1+\lambda)}\right),\end{aligned}$$

where c_1 and c_2 are constants and we used the assumptions that $m \geq 1$ and $Bn \geq 1$. □

4.4 Oracle inequalities for unstructured models

To this point, our results have addressed the model selection problem in scenarios where we have a nested collection of models. In the most general case, however, the collection of models may be quite heterogeneous, with no relationship between the different model families. In classification, for instance, we may consider generalized linear models with different link functions, decision trees, random forests, or other families among our collection of models. For a non-parametric regression problem, we may want to select across a collection of dictionaries such as wavelets, splines, and polynomials. While this more general setting is obviously more challenging than the structured cases in the prequel, we would like to study the effects that limiting computation has on model selection problems, understanding when it is possible to outperform computation-agnostic strategies.

4.4.1 Problem setting and algorithm

When no structure relates the models under consideration, it is impossible to work with an infinite collection of classes—any estimator must evaluate each class. As a result, we restrict ourselves to finite model collections in this section, so that we have a sequence $\mathcal{F}_1, \ldots, \mathcal{F}_K$ of models from which we wish to select. Our approach to the unstructured case is to incrementally allocate computational quota amongst the function classes, where we trade off receiving samples for classes that have good risk performance against exploring classes for which we have received few data points. More formally, with T available quanta of computation, it is natural to view the model selection problem as a T round game, where

in each round a procedure selects a function class i and allocates it one additional quantum of computation.

With this setup, we turn to stating a few natural assumptions. We assume that the computational complexity of fitting a model grows linearly and incrementally with the number of samples, which means that allocating an additional quantum of training time allows the learning algorithm $\mathcal{A}$ to process an additional n_i samples for class $\mathcal{F}_i$. In the context of Sections 4.2 and 4.3, this means that we assume $n_i(t) = tn_i$ for some fixed number n_i specific to class i. This linear growth assumption is satisfied, for instance, when the loss function ℓ is convex and the black-box learning algorithm $\mathcal{A}$ is a stochastic or online convex optimization procedure [47, 117]. We also require concentration assumptions similar to Assumptions B and C:

Assumption F. Let $\mathcal{A}(i, T) \in \mathcal{F}_i$ denote the output of algorithm $\mathcal{A}$ when executed for class $\mathcal{F}_i$ with a computational budget T.

(a) For each i, there exists an $n_i \in \mathbb{N}$ such that in T units of time, algorithm $\mathcal{A}$ can compute $\mathcal{A}(i, T)$ using $n_i T$ samples.

(b) For each $i \in [K]$, there is a function γ_i and constants $\kappa_1, \kappa_2 > 0$ such that for any $T \in \mathbb{N}$,

$$\mathbb{P}\left(|\widehat{R}_{n_i T}(\mathcal{A}(i, T)) - R(\mathcal{A}(i, T))| > \gamma_i(n_i T) + \kappa_2 \epsilon\right) \leq \kappa_1 \exp(-4 n_i T \epsilon^2). \tag{4.26}$$

(c) The output $\mathcal{A}(i, T)$ is a $\gamma_i(n_i T)$-minimizer of $\widehat{R}_{n_i T}$, that is,

$$\widehat{R}_{n_i T}(\mathcal{A}(i, T)) - \inf_{f \in \mathcal{F}_i} \widehat{R}_{n_i T}(f) \leq \gamma_i(n_i T).$$

(d) For each i, the function γ_i satisfies $\gamma_i(n) \leq c_i n^{-\alpha_i}$ for some $\alpha_i > 0$.

(e) For any fixed function $f \in \mathcal{F}_i$, $\mathbb{P}(|\widehat{R}_n(f) - R(f)| > \kappa_2 \epsilon) \leq \kappa_1 \exp(-4 n \epsilon^2)$.

Comparing to Assumptions B and C, we see that the main difference is in the linear time assumption (a) and growth assumption (d). In addition, the complexity penalties and function classes discussed in our earlier examples satisfy Assumption F. Since it is more natural to keep track of samples received by each class in the setup of this section, we will often use the notation $\mathcal{A}(i, n)$ to denote the output of algorithm $\mathcal{A}$ on class i after receiving n data samples.

We now present our algorithm for successively allocating computational quanta to the function classes. To choose the class i receiving computation at iteration t, the procedure must balance competing goals of *exploration*—evaluating each function class $\mathcal{F}_i$ adequately—and *exploitation*—giving more computation to classes with low empirical risk. To promote

Algorithm 3 Multi-armed bandit algorithm for selection of best class $\widehat{i}$.

For each $i \in [K]$, query n_i examples from class $\mathcal{F}_i$.
for $t = K + 1$ to T **do**
  Let $N_i(t)$ be the total number of examples seen for class i until time t
  Let $i_t = \operatorname{argmin}_{i \in [K]} \overline{R}(j, N_i(t)) - \sqrt{\frac{\log t}{N_i(t)}}$.
  Query n_{i_t} examples for class i_t and set $N_{i_t}(t+1) = N_{i_t}(t) + n_{i_t}$.
end for
Output $\widehat{i}$, the index of the most frequently queried class.

exploration, we use an optimistic selection criterion to choose class i, which—assuming that $\mathcal{F}_i$ has seen n samples and t computational quanta at this point—is

$$\overline{R}(i, n) = \widehat{R}_n(\mathcal{A}(i, n)) - \gamma_i(n) - \sqrt{\frac{\log K}{n}} + \gamma_i(Tn_i). \tag{4.27}$$

The intuition behind the definition of $\overline{R}(i, n)$ is that we would like the algorithm to choose functions f and classes i that minimize $\widehat{R}_n(f) + \gamma_i(Tn_i) \approx R(f) + \gamma_i(Tn_i)$, but the negative $\gamma_i(n)$ and $\sqrt{\log K/n}$ terms lower the criterion significantly when n is small and thus encourage initial exploration. The criterion (4.27) essentially combines a penalized model-selection objective (used, for example in Bartlett et al. [18]) with an optimistic criterion similar to those used in multi-armed bandit algorithms [13]. Algorithm 3 contains the formal description of our bandit procedure for model selection. Algorithm 3 begins by receiving n_i samples for each of the K classes $\mathcal{F}_i$ to form the preliminary empirical estimates (4.27); we then run the optimistic selection criterion for T rounds until the computational budget is exhausted.

4.4.2 Main results and some consequences

The goal of the selection procedure is to find the best penalized class i^*: a class satisfying

$$i^* \in \operatorname*{argmin}_{i \in [K]} \left\{ \inf_{f \in \mathcal{F}_i} R(f) + \gamma_i(Tn_i) \right\} = \operatorname*{argmin}_{i \in [K]} \left\{ R_i^* + \gamma_i(Tn_i) \right\}.$$

To present our main results for Algorithm 3, we define the excess penalized risk Δ_i of class i:

$$\Delta_i := R_i^* + \gamma_i(Tn_i) - R_{i^*}^* - \gamma_{i^*}(Tn_{i^*}) \geq 0. \tag{4.28}$$

Without loss of generality, we assume that the infimum in $R_i^* = \inf_{f \in \mathcal{F}_i} R(f)$ is attained by a function f_i^* (if not, we use a limiting argument, choosing some fixed f_i^* such that $R(f_i^*) \leq \inf_{f \in \mathcal{F}_i} R(f) + \delta$ for an arbitrarily small $\delta > 0$).

The gains of a computationally adaptive strategy over naïve strategies are clearest when the gap (4.28) is non-zero for each i, though in the sequel, we forgo this requirement. Under

this assumption, we can follow the ideas of Auer et al. [13] to show that the fraction of the computational budget allocated to any suboptimal class $i \neq i^*$ goes quickly to zero as T grows. We provide the proof of the following theorem in Section 4.4.3.

Theorem 4.3. *Let Alg. 3 be run for T rounds and let $T_i(t)$ be the number of times class i is queried through round t. Let Δ_i be defined as in (4.28) and Assumption F hold, and assume that $T \geq K$. Define $\beta_i = \max\{1/\alpha_i, 2\}$. There is a constant C such that*

$$\mathbb{E}[T_i(T)] \leq \frac{C}{n_i}\left(\frac{c_i + \kappa_2\sqrt{\log T}}{\Delta_i}\right)^{\beta_i} \quad \text{and} \quad \mathbb{P}\left(T_i(T) > \frac{C}{n_i}\left(\frac{c_i + \kappa_2\sqrt{\log T}}{\Delta_i}\right)^{\beta_i}\right) \leq \frac{\kappa_1}{TK^4},$$

where c_i and α_i are the constants in the definition F(d) of the concentration function γ_i.

At a high level, this result shows that the fraction of budget allocated to any suboptimal class goes to 0 at the rate $\frac{1}{n_i T}\left(\frac{\sqrt{\log T}}{\Delta_i}\right)^{\beta_i}$. Hence, asymptotically in T, the procedure performs almost as if all the computational budget were allocated to class i^*. To see an example of concrete rates that can be concluded from the above result, let $\mathcal{F}_1, \ldots, \mathcal{F}_K$ be model classes with finite VC-dimension,[1] so that Assumption F is satisfied with $\alpha_i = \frac{1}{2}$. Then we have

Corollary 4.3. *Under the conditions of Theorem 4.3, assume $\mathcal{F}_1, \ldots, \mathcal{F}_K$ are model classes of finite VC-dimension, where $\mathcal{F}_i$ has dimension d_i. Then there is a constant C such that*

$$\mathbb{E}[T_i(T)] \leq C\frac{\max\{d_i, \kappa_2^2 \log T\}}{\Delta_i^2 n_i} \quad \text{and} \quad \mathbb{P}\left(T_i(T) > C\frac{\max\{d_i, \kappa_2^2 \log T\}}{\Delta_i^2 n_i}\right) \leq \frac{\kappa_1}{TK^4}.$$

A lower bound by Lai and Robbins [95] for the multi-armed bandit problem shows that Corollary 4.3 is nearly optimal in general. To see the connection, let $\mathcal{F}_i$ correspond to the ith arm in a multi-armed bandit problem and the risk R_i^* be the expected reward of arm i and assume w.l.o.g. that $R_i^* \in [0, 1]$. In this case, the complexity penalty γ_i for each class is 0. Let p_i be a distribution on $\{0, 1\}$, where $p_i(1) = R_i^*$ and $p_i(0) = 1 - R_i^*$ (let $p_i = p_i(1)$ for shorthand). Lai and Robbins give a lower bound that shows that the expected number of pulls of any suboptimal arm is at least $\mathbb{E}[T_i(T)] = \Omega\left(\log T / \mathrm{KL}\left(p_i \| p_{i^*}\right)\right)$, where p_i and p_{i^*} are the reward distributions for the ith and optimal arms, respectively. An asymptotic expansion shows that $\mathrm{KL}\left(p_i \| p_{i^*}\right) = \Delta_i^2/(2p_i(1-p_i))$, plus higher order terms, in this special case. So Corollary 4.3 is essentially tight.

The condition that the gap $\Delta_i > 0$ may not always be satisfied, or Δ_i may be so small as to render the bound in Theorem 4.3 vacuous. Nevertheless, it is intuitive that our algorithm can quickly find a small set of "good" classes—those with small penalized risk—and spend its computational budget to try to distinguish amongst them. In this case, Algorithm 3 does not visit suboptimal classes and so can output a function f satisfying good oracle bounds.

[1] Similar corollaries hold for any model class whose metric entropy grows polynomially in $\log \frac{1}{\epsilon}$.

In order to prove a result quantifying this intuition, we first upper bound the *regret* of Algorithm 3, that is, the average excess risk suffered by the algorithm over all iterations, and then show how to use this bound for obtaining a model with a small risk. For the remainder of the section, we simplify the presentation by assuming that $\alpha_i \equiv \alpha$ and define $\beta = \max\{1/\alpha, 2\}$.

Proposition 4.2. Use the same assumptions as Theorem 4.3, but further assume that $\alpha_i \equiv \alpha$ for all i. With probability at least $1 - \kappa_1/TK^3$, the regret (average excess risk) of Algorithm 3 satisfies

$$\sum_{i=1}^{K} \Delta_i T_i(T) \leq 2eT^{1-1/\beta} \left(C \sum_{i=1}^{K} \frac{(c_i + \kappa_2 \sqrt{\log T})^\beta}{n_i} \right)^{1/\beta}$$

for a constant C dependent on α.

Our final main result builds on Proposition 4.2 to show that when it is possible to average functions across classes $\mathcal{F}_i$, we can aggregate all the "played" functions f_t, one for each iteration t, to obtain a function with small risk. Indeed, setting $f_t = \mathcal{A}(i_t, n_i(t))$, we obtain the following theorem (whose proof, along with that of Proposition 4.2, we provide in Appendix B.4):

Theorem 4.4. *Use the conditions of Proposition 4.2. Let the risk function R be convex on $\mathcal{F}_1 \cup \ldots \cup \mathcal{F}_K$, and let f_t be the function chosen by algorithm $\mathcal{A}$ at round t of Alg. 3. Define the average function $\widehat{f}_T = \frac{1}{T}\sum_{t=1}^{T} f_t$. There are constants C, C' (dependent on α) such that with probability greater than $1 - 2\kappa_2/(TK^3)$,*

$$R(\widehat{f}_T) \leq R^* + \gamma_{i^*}(Tn_{i^*}) + 2e\kappa_2 T^{-\beta}\sqrt{\log T} \left(\sum_{i=1}^{K} \frac{C}{n_i} \right)^{1/\beta}$$
$$+ C' T^{-1/\beta} \left(\sum_{i=1}^{K} \left[c_i n_i^{-\alpha} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log K} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log T} \right]^\beta \right)^{1/\beta}.$$

Let us interpret the above bound and discuss its optimality. When $\alpha = \frac{1}{2}$ (e.g., for VC classes), we have $\beta = 2$; moreover, it is clear that $\sum_{i=1}^{K} \frac{C}{n_i} = \mathcal{O}(K)$. Thus, to within constant factors,

$$R(\widehat{f}_T) = R^*_{i^*} + \gamma_{i^*}(Tn_{i^*}) + \mathcal{O}\left(\frac{\sqrt{K \max\{\log T, \log K\}}}{\sqrt{T}} \right).$$

Ignoring logarithmic factors, the above bound is minimax optimal, which follows by a reduction of our model selection problem to the special case of a multi-armed bandit problem. In this case, Theorem 5.1 of Auer et al. [12] shows that for any set of K, T values, there is

a distribution over the rewards of arms which forces $\Omega(\sqrt{KT})$ regret, that is, the average excess risk of the classes chosen by Alg. 3 must be $\Omega(\sqrt{KT})$, matching Proposition 4.2 and Theorem 4.4.

The scaling $\mathcal{O}(\sqrt{K})$ is essentially as bad as splitting the computational budget T uniformly across each of the K classes, which yields (roughly) an oracle inequality of the form

$$R(f) = R^*_{i^*} + \gamma_{i^*}(Tn_{i^*}/K) + \mathcal{O}\left(\frac{\sqrt{K \log K}}{\sqrt{Tn_{i^*}}}\right).$$

Comparing this bound to Theorem 4.4, we see that the penalty γ_i in the theorem is smaller, and (ignoring logarithmic factors) the difference in quality of results is the difference between

$$\sum_{i=1}^{K} \frac{1}{n_i} \quad \text{and} \quad \frac{K}{n_{i^*}}.$$

When the left quantity is smaller than the right, the bandit-based Algorithm 3 and the extension indicated by Theorem 4.4 give improvements over the naïve strategy of uniformly splitting the budget across classes. However, if each class has similar computational cost n_i, no strategy can outperform the naïve one.

We also observe that we can apply the online procedure of Algorithm 3 to the nested setup of Sections 4.2 and 4.3 as well. In this case, by applying Algorithm 3 only to elements of the coarse-grid set S_λ, we can replace K in the bounds of Theorems 4.3 and 4.4 with $s(\lambda)$, which gives results similar to our earlier Theorems 4.1 and 4.2.

4.4.3 Proof of Theorem 4.3

At a high level, the proof of this theorem involves combining the techniques for analysis of multi-armed bandits developed in [13] with Assumption F. We start by giving a lemma which will be useful to prove the theorem. The lemma states that after a sufficient number of initial iterations τ, the probability that Algorithm 3 chooses to receive samples for a suboptimal function class $i \neq i^*$ is extremely small. Recall also our notational convention that $\beta_i = \max\{1/\alpha_i, 2\}$.

Lemma 4.5. *Let Assumption F hold. For any class i, any $s_i \in [1, T]$ and $s_{i^*} \in [\tau, T]$ where τ satisfies*

$$\tau > \frac{2^{\beta_i}(c_i + \kappa_2\sqrt{\log T} + \kappa_2\sqrt{\log K})^{\beta_i}}{n_i \Delta_i^{\beta_i}},$$

we have

$$\mathbb{P}\left(\overline{R}(i, n_i s_i) - \kappa_2\sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2\sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right) \leq \frac{2\kappa_1}{(TK)^4}.$$

We defer the proof of the lemma to Appendix B.3, though at a high level the proof works as follows. The "bad event" in Lemma 4.5, which corresponds to Algorithm 3 selecting a sub-optimal class $i \neq i^*$, occurs only if one of the following three errors occurs: the empirical risk of class i is much lower than its true risk, the empirical risk of class i^* is higher than its true risk, or s_i is not large enough to actually separate the true penalized risks from one another. The assumptions of the lemma make each of these three sub-events quite unlikely. Now we turn to the proof of Theorem 4.3, assuming the lemma.

Let i_t denote the model class index i chosen by Algorithm 3 at time t, and let $s_i(t)$ denote the number of times class i has been selected at round t of the algorithm. When no time index is needed, s_i will denote the same thing. Note that if $i_t = i$ and the number of times class i is queried exceeds $\tau > 0$, then by the definition of the selection criterion (4.27) and choice of i_t in Alg. 3, for some $s_i \in \{\tau, \dots, t-1\}$ and $s_{i^*} \in \{1, \dots, t-1\}$ we have

$$\overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}.$$

Here we interpret $\overline{R}(i, n_i s_i)$ to mean a random realization of the observed risk consistent with the samples we observe. Using the above implication, we thus have

$$\begin{aligned} T_i(T) = 1 + \sum_{t=K+1}^{T} \mathbb{I}\{i_t = i\} \quad &\leq \quad \tau + \sum_{t=K+1}^{T} \mathbb{I}\{i_t = i, T_i(t-1) \geq \tau\} \\ &\leq \tau + \sum_{t=K+1}^{T} \mathbb{I}\left\{ \min_{\tau \leq s_i < t} \overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \max_{0 < s < t} \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \right\} \\ &\leq \tau + \sum_{t=1}^{T} \sum_{s_{i^*}=1}^{t-1} \sum_{s_i=\tau}^{t-1} \mathbb{I}\left\{ \overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \right\}. \end{aligned} \tag{4.29}$$

To control the last term, we invoke Lemma 4.5 and obtain that

$$\tau > \frac{2^{\beta_i}(c_i + \kappa_2\sqrt{\log T} + \kappa_2\sqrt{\log K})^{\beta_i}}{n_i \Delta_i^{\beta_i}} \quad \Rightarrow \quad \mathbb{E}[T_i(n)] \leq \tau + \sum_{t=1}^{T} \sum_{s=1}^{t-1} \sum_{s_i=\tau}^{t-1} 2\frac{\kappa_1}{(TK)^4} \leq \tau + \frac{\kappa_1}{TK^4}.$$

Hence for any suboptimal class $i \neq i^*$, $\mathbb{E}[T_i(n)] \leq \tau_i + \kappa_1/(TK^4)$, where τ_i satisfies the lower bound of Lemma 4.5 and is thus logarithmic in T. Under the assumption that $T \geq K$, for $i \neq i^*$,

$$\mathbb{E}[T_i(T)] \leq C \frac{(c_i + \kappa_2\sqrt{\log T})^{\max\{1/\alpha_i, 2\}}}{n_i \Delta_i^{\max\{1/\alpha_i, 2\}}} \tag{4.30}$$

for a constant $C \leq 2 \cdot 4^{\max\{1/\alpha_i, 2\}}$. Now we prove the high-probability bound. For this part, we need only concern ourselves with the sum of indicators from (4.29). Markov's inequality

shows that

$$\mathbb{P}\left(\sum_{t=K+1}^{T} \mathbb{I}i_t = i, T_i(t-1) \geq \tau \geq 1\right) \leq \frac{\kappa_1}{TK^4}.$$

Thus we can assert that the bound (4.30) on $T_i(T)$ holds with high probability.

Remark: By examining the proof of Theorem 4.3, it is straightforward to see that if we modify the multipliers on the square root terms in the criterion (4.27) by $m\kappa_2$ instead of κ_2, we get that the probability bound is of the order $T^{3-4m^2}K^{-4m^2}$, while the bound on $T_i(T)$ is scaled by m^{1/α_i}.

4.5 Discussion

In this chapter, we have presented a new framework for model selection with computational constraints. The novelty of our new setting is the idea of using computation—rather than samples—as the quantity against which we measure the performance of our estimators. By carefully capturing the relative computational needs of fitting different models to our data, we are able to formalize the very natural intuition: *Given a computational budget, a simple model can be fitted to a lot more samples than a complex model.* As our main contribution, we have presented algorithms for model selection in several scenarios, and the common thread in each is that we attain good performance by evaluating only a small and intelligently-selected set of models, allocating samples to each model based on the computational cost. For model selection over nested hierarchies, this takes the form of a new estimator based on a coarse gridding of the model space, which is competitive (to logarithmic factors) with an omniscient oracle. A minor extension of this algorithm is adaptive to problem complexity, since it yields fast rates for model selection when the underlying estimation problems have appropriate curvature or low-noise properties. We also presented an exploration-exploitation algorithm for model selection in unstructured cases, showing that it obtains (in some sense) nearly optimal performance.

There are certainly many possible extensions and open questions raised by this work. We address the setting where the complexity penalties are known and can be computed easily in closed form. Often it is desirable to use data-dependent penalties [104, 21, 108], since they adapt to the particular problem instance and data distribution. It appears to be somewhat difficult to extend such penalties to the procedures we have developed in this chapter, but we believe it would be quite interesting. Another natural question to ask is whether there exist intermediate model selection problems between a nested sequence of classes and a completely unstructured collection. Identifying other structures—and obtaining the corresponding oracle inequalities and understanding their dependence on computation—would be an interesting extension of the results presented here.

More broadly, we believe the idea of using computation, in addition to the number of samples available for a statistical inference problem, to measure the performance of statistical procedures is appealing for a much broader class of problems. In large data settings, one would hope that more data would always improve the risk performance of statistical procedures, even with a fixed computational budget. We hope that extending these ideas to other problems, and understanding how computation interacts with and affects the quality of statistical estimation more generally will be quite fruitful.

Chapter 5

Optimization algorithms for statistical estimation in high-dimensions

High-dimensional data sets present challenges that are both statistical and computational in nature. On the statistical side, recent years have witnessed a flurry of results on consistency and rates for various estimators under non-asymptotic high-dimensional scaling, meaning that error bounds are provided for general settings of the sample size n and problem dimension d, allowing for the possibility that $d \gg n$. These results typically involve some assumption regarding the underlying structure of the parameter space, such as sparse vectors, structured covariance matrices, low-rank matrices, or structured regression functions, as well as some regularity conditions on the data-generating process. On the computational side, many estimators for statistical recovery are based on solving convex programs. Examples of such M-estimators include ℓ_1-regularized quadratic programs (also known as the Lasso) for sparse linear regression (e.g., see the papers [156, 51, 177, 111, 30, 40, 164] and references therein), second-order cone programs (SOCP) for the group Lasso (e.g., [180, 103, 79] and references therein), and semidefinite programming relaxations (SDP) for various problems, including sparse PCA and low-rank matrix estimation (e.g., [44, 135, 152, 10, 141, 115, 136] and references therein).

Many of these programs are instances of convex conic programs, and so can (in principle) be solved to ϵ-accuracy in polynomial time using interior point methods, and other standard methods from convex programming (e.g., see the books [28, 35]). However, the complexity of such quasi-Newton methods can be prohibitively expensive for the very large-scale problems that arise from high-dimensional data sets. Accordingly, recent years have witnessed a renewed interest in simpler first-order methods, among them the methods of projected gradient descent and mirror descent. Our aim in this chapter will be to consider such gradient methods, and establish fast linear convergence for them under the typical statistical settings in high-dimensions. These results extend our understanding of these methods beyond cases addressed by the current theory, and highlight interesting interplay between the computational and statistical complexities of such high-dimensional estimation problems.

5.1 Motivation and prior work

There has been a large body of literature growing around the application of existing optimization methods, as well as the development of new ones tailored to the needs of structured optimization problems in high-dimensions resulting from the statistical M-estimators. Several authors (e.g., [24, 80, 23]) have used variants of Nesterov's accelerated gradient method [121] to obtain algorithms for high-dimensional statistical problems with a sublinear rate of convergence. Note that an optimization algorithm, generating a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$, is said to exhibit *sublinear convergence* to an optimum $\widehat{\theta}$ if the optimization error $\|\theta^t - \widehat{\theta}\|$ decays at the rate $1/t^{\kappa}$, for some exponent $\kappa > 0$ and norm $\|\cdot\|$. Although this type of convergence is quite slow, it is the best possible with gradient descent-type methods for convex programs under only Lipschitz conditions [120].

It is known that much faster global rates—in particular, a linear or geometric rate—can be achieved if global regularity conditions like strong convexity and smoothness are imposed [120]. An optimization algorithm is said to exhibit *linear or geometric* convergence if the optimization error $\|\theta^t - \widehat{\theta}\|$ decays at a rate κ^t, for some contraction coefficient $\kappa \in (0, 1)$. Note that such convergence is exponentially faster than sub-linear convergence. For certain classes of problems involving polyhedral constraints and global smoothness, Tseng and Luo [105] have established geometric convergence. However, a challenging aspect of statistical estimation in high dimensions is that the underlying optimization problems can never be strongly convex in a global sense when $d > n$ (since the $d \times d$ Hessian matrix is rank-deficient), and global smoothness conditions cannot hold when $d/n \to +\infty$. Some more recent work has exploited structure specific to the optimization problems that arise in statistical settings. For the special case of sparse linear regression with random isotropic designs (also referred to as compressed sensing), some authors have established fast convergence rates in a local sense, meaning guarantees that apply once the iterates are close enough to the optimum [37, 71]. The intuition underlying these results is that once an algorithm identifies the support set of the optimal solution, the problem is then effectively reduced to a lower-dimensional subspace, and thus fast convergence can be guaranteed in a local sense. Also in the setting of compressed sensing, Tropp and Gilbert [158] studied finite convergence of greedy algorithms based on thresholding techniques, and showed linear convergence up to a certain tolerance. For the same class of problems, Garg and Khandekar [68] showed that a thresholded gradient algorithm converges rapidly up to some tolerance. In both of these results, the convergence tolerance is of the order of the noise variance, and hence substantially larger than the true statistical precision of the problem.

The focus of this chapter is the convergence rate of two simple gradient-based algorithms for solving optimization problems that underlie regularized M-estimators. For a constrained problem with a differentiable objective function, the projected gradient method generates a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ by taking a step in the negative gradient direction, and then projecting the result onto the constraint set. The composite gradient method of Nesterov [121] is

well-suited to solving regularized problems formed by the sum of a differentiable and (potentially) non-differentiable component. The main contribution of this chapter is to establish a form of global geometric convergence for these algorithms that holds for a broad class of high-dimensional statistical problems. In order to provide intuition for this guarantee, Figure 5.1 shows the performance of projected gradient descent for a Lasso problem (ℓ_1-constrained least-squares). In panel (a), we have plotted the logarithm of the optimization error, measured in terms of the Euclidean norm $\|\theta^t - \widehat{\theta}\|$ between the current iterate θ^t and an optimal solution $\widehat{\theta}$, versus the iteration number t. The plot includes three different curves, corresponding to sparse regression problems in dimension $d \in \{5000, 10000, 20000\}$, and a fixed sample size $n = 2500$. Note that all curves are linear (on this logarithmic scale), revealing the geometric convergence predicted by our theory. Such convergence is not predicted by classical optimization theory, since the objective function cannot be strongly convex whenever $n < d$. Moreover, the convergence is geometric even at early iterations, and takes place to a precision far less than the noise level ($\nu^2 = 0.25$ in this example). We also note that the design matrix does not satisfy the restricted isometry property, as assumed in some past work.

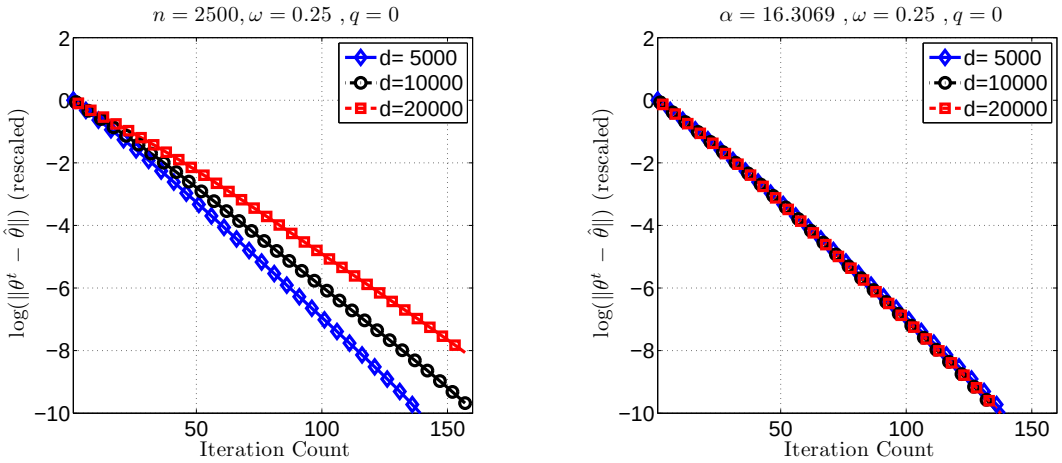

Figure 5.1. Convergence rates of projected gradient descent in application to Lasso programs (ℓ_1-constrained least-squares). Each panel shows the log optimization error $\log \|\theta^t - \widehat{\theta}\|$ versus the iteration number t. Panel (a) shows three curves, corresponding to dimensions $d \in \{5000, 10000, 20000\}$, sparsity $s = \lceil \sqrt{d} \rceil$, and all with the same sample size $n = 2500$. All cases show geometric convergence, but the rate for larger problems becomes progressively slower. (b) For an appropriately rescaled sample size ($\alpha = \frac{n}{s \log d}$), all three convergence rates should be roughly the same, as predicted by the theory.

The results in panel (a) exhibit an interesting property: the convergence rate is *dimension-dependent*, meaning that for a fixed sample size, projected gradient descent converges more

slowly for a large problem than a smaller problem—compare the squares for $d = 20000$ to the diamonds for $d = 5000$. This phenomenon reflects the natural intuition that larger problems are, in some sense, "harder" than smaller problems. A notable aspect of our theory is that in addition to guaranteeing geometric convergence, it makes a quantitative prediction regarding the extent to which a larger problem is harder than a smaller one. In particular, our convergence rates suggest that if the sample size n is re-scaled in a certain way according to the dimension d and also other model parameters such as sparsity, then convergence rates should be roughly similar. Panel (b) provides a confirmation of this prediction: when the sample size is rescaled according to our theory (in particular, see Corollary 5.2 in Section 5.3.2), then all three curves lie essentially on top of another.

Although high-dimensional optimization problems are typically neither strongly convex nor smooth, our work shows that it is fruitful to consider suitably restricted notions of strong convexity and smoothness. Our notion of restricted strong convexity (RSC) is related to but slightly different than that introduced in a recent paper by Negahban et al. [116] for establishing statistical consistency. As we discuss in the sequel, bounding the optimization error introduces new challenges not present when analyzing the statistical error. We also introduce a related notion of restricted smoothness (RSM), not needed for proving statistical rates but essential in the setting of optimization. Our analysis consists of two parts. We first show that for optimization problems underlying many regularized M-estimators, appropriately modified notions of restricted strong convexity (RSC) and smoothness (RSM) are sufficient to guarantee global linear convergence of projected gradient descent. Our second contribution is to prove that for the iterates generated by our first-order method, these RSC/RSM assumptions do indeed hold with high probability for a broad class of statistical models, among them sparse linear models, models with group sparsity constraints, and various classes of matrix estimation problems, including matrix completion and matrix decomposition.

An interesting aspect of our results is that the global geometric convergence is not guaranteed to an arbitrary numerical precision, but only to an accuracy related to *statistical precision* of the problem. For a given error norm $\|\cdot\|$, given by the Euclidean or Frobenius norm for most examples in this chapter, the statistical precision is given by the mean-squared error $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2]$ between the true parameter θ^* and the estimate $\widehat{\theta}$ obtained by solving the optimization problem, where the expectation is taken over randomness in the statistical model. Note that this is very natural from the statistical perspective, since it is the true parameter θ^* itself (as opposed to the solution $\widehat{\theta}$ of the M-estimator) that is of primary interest, and our analysis allows us to approach it as close as is statistically possible. Our analysis shows that we can geometrically converge to a parameter θ such that $\|\theta - \theta^*\| = \left\|\widehat{\theta} - \theta^*\right\| + o\left(\left\|\widehat{\theta} - \theta^*\right\|\right)$, which is the best we can hope for statistically, ignoring lower order terms. Overall, our results reveal an interesting connection between the statistical and computational properties of M-estimators—that is, the properties of the underlying statistical model that make it favorable for estimation also render it more amenable to optimization procedures.

The remainder of this chapter is organized as follows. We begin in Section 5.2 with a precise formulation of the class of convex programs analyzed in this work, along with background on the notions of a decomposable regularizer, and properties of the loss function. Section 5.3 is devoted to the statement of our main convergence result, as well as to the development and discussion of its various corollaries for specific statistical models. In Section 5.4, we provide a number of empirical results that confirm the sharpness of our theoretical predictions. Finally, Section 5.5 contains the proofs, with more technical aspects of the arguments deferred to the Appendix. We note that an extended abstract containing the first main theorem in this chapter and accompanying corollaries appeared in the paper [4].

5.2 Background and problem formulation

In this section, we begin by describing the class of regularized M-estimators to which our analysis applies, as well as the optimization algorithms that we analyze. Finally, we introduce some important notions that underlie our analysis, including the notions of a decomposable regularization, and the properties of restricted strong convexity and smoothness.

5.2.1 Loss functions, regularization and gradient-based methods

Given a random variable $z \sim \mathbb{P}$ taking values in some set $\mathcal{Z}$, let $z_1^n = \{z_1, \ldots, z_n\}$ be a collection of n observations. Here the integer n is the *sample size* of the problem. Assuming that $\mathbb{P}$ lies within some indexed family $\{\mathbb{P}_\theta, \theta \in \Omega\}$, the goal is to recover an estimate of the unknown true parameter $\theta^* \in \Omega$ generating the data. Here Ω is some subset of $\mathbb{R}^d$, and the integer d is known as the *ambient dimension* of the problem. In order to measure the "fit" of any given parameter $\theta \in \Omega$ to a given data set z_1^n, we introduce a loss function $\mathcal{L}_n : \Omega \times \mathcal{Z}^n \to \mathbb{R}_+$. By construction, for any given n-sample data set $z_1^n \in \mathcal{Z}^n$, the loss function assigns a cost $\mathcal{L}_n(\theta; z_1^n) \geq 0$ to the parameter $\theta \in \Omega$. In many (but not all) applications, the loss function has a separable structure across the data set, meaning that $\mathcal{L}_n(\theta; z_1^n) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$ where $\ell : \Omega \times \mathcal{Z} :\to \mathbb{R}_+$ is the loss function associated with a single data point.

Of primary interest in this chapter are estimation problems that are under-determined, meaning that the number of observations n is smaller than the ambient dimension d. In such settings, without further restrictions on the parameter space Ω, there are various impossibility theorems, asserting that consistent estimates of the unknown parameter θ^* cannot be obtained. For this reason, it is necessary to assume that the unknown parameter θ^* either lies within a smaller subset of Ω, or is well-approximated by some member of such a subset. In order to incorporate these types of structural constraints, we introduce a *regularizer* $\mathcal{R} : \Omega \to \mathbb{R}_+$ over the parameter space. With these ingredients, the analysis of in our work

applies to the *constrained M-estimator*

$$\widehat{\theta}_\rho \in \arg\min_{\mathcal{R}(\theta)\leq\rho} \big\{\mathcal{L}_n(\theta; z_1^n)\big\}, \tag{5.1}$$

where $\rho > 0$ is a user-defined radius, as well as to the *regularized M-estimator*

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\mathcal{R}(\theta)\leq\overline{\rho}} \big\{\underbrace{\mathcal{L}_n(\theta; z_1^n) + \lambda_n\mathcal{R}(\theta)}_{\phi_n(\theta)}\big\} \tag{5.2}$$

where the regularization weight $\lambda_n > 0$ is user-defined. Note that the radii ρ and $\overline{\rho}$ may be different in general. Throughout this chapter, we impose the following two conditions:

(a) for any data set z_1^n, the function $\mathcal{L}_n(\cdot; z_1^n)$ is convex and differentiable over Ω, and

(b) the regularizer $\mathcal{R}$ is a norm.

These conditions ensure that the overall problem is convex, so that by Lagrangian duality, the optimization problems (5.1) and (5.2) are equivalent. However, as our analysis will show, solving one or the other can be computationally more preferable depending upon the assumptions made. Some remarks on notation: when the radius ρ or the regularization parameter λ_n is clear from the context, we will drop the subscript on $\widehat{\theta}$ to ease the notation. Similarly, we frequently adopt the shorthand $\mathcal{L}_n(\theta)$, with the dependence of the loss function on the data being implicitly understood. Procedures based on optimization problems of either form are known as M-estimators in the statistics literature.

The focus of this chapter is on two simple algorithms for solving the above optimization problems. The method of *projected gradient descent* applies naturally to the constrained problem (5.1), whereas the *composite gradient descent* method due to Nesterov [121] is suitable for solving the regularized problem (5.2). Each routine generates a sequence $\{\theta^t\}_{t=0}^\infty$ of iterates by first initializing to some parameter $\theta^0 \in \Omega$, and then applying the recursive update

$$\theta^{t+1} = \arg\min_{\theta\in\mathbb{B}_{\mathcal{R}}(\rho)} \big\{\mathcal{L}_n(\theta^t) + \langle\nabla\mathcal{L}_n(\theta^t), \theta - \theta^t\rangle + \frac{\gamma_u}{2}\big\|\theta - \theta^t\big\|^2\big\}, \qquad \text{for } t = 0, 1, 2, \ldots, \tag{5.3}$$

in the case of projected gradient descent, or the update

$$\theta^{t+1} = \arg\min_{\theta\in\mathbb{B}_{\mathcal{R}}(\overline{\rho})} \big\{\mathcal{L}_n(\theta^t) + \langle\nabla\mathcal{L}_n(\theta^t), \theta - \theta^t\rangle + \frac{\gamma_u}{2}\big\|\theta - \theta^t\big\|^2 + \lambda_n\mathcal{R}(\theta)\big\}, \qquad \text{for } t = 0, 1, 2, \ldots, \tag{5.4}$$

for the composite gradient method. Note that the only difference between the two updates is the addition of the regularization term in the objective. These updates have a natural intuition: the next iterate θ^{t+1} is obtained by constrained minimization of a first-order

approximation to the loss function, combined with a smoothing term that controls how far one moves from the current iterate in terms of Euclidean norm. Moreover, it is easily seen that the update (5.3) is equivalent to

$$\theta^{t+1} = \Pi\left(\theta^t - \frac{1}{\gamma_u}\nabla\mathcal{L}_n(\theta^t)\right), \tag{5.5}$$

where $\Pi \equiv \Pi_{\mathbb{B}_{\mathcal{R}}(\rho)}$ denotes Euclidean projection onto the ball $\mathbb{B}_{\mathcal{R}}(\rho) = \{\theta \in \Omega \mid \mathcal{R}(\theta) \leq \rho\}$ of radius ρ. In this formulation, we see that the algorithm takes a step in the gradient direction, using the quantity $1/\gamma_u$ as stepsize parameter, and then projects the resulting vector onto the constraint set. The update (5.4) takes an analogous form, however, the projection will depend on both λ_n and γ_u. As will be illustrated in the examples to follow, for many problems, the updates (5.3) and (5.4), or equivalently (5.5), have a very simple solution. For instance, in the case of ℓ_1-regularization, it can be obtained by an appropriate form of the soft-thresholding operator.

5.2.2 Restricted strong convexity and smoothness

In this section, we define the conditions on the loss function and regularizer that underlie our analysis. Global smoothness and strong convexity assumptions play an important role in the classical analysis of optimization algorithms [28, 35, 120]. In application to a differentiable loss function $\mathcal{L}_n$, both of these properties are defined in terms of a first-order Taylor series expansion around a vector θ' in the direction of θ—namely, the quantity

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') := \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta') - \langle \nabla\mathcal{L}_n(\theta'), \theta - \theta' \rangle. \tag{5.6}$$

By the assumed convexity of $\mathcal{L}_n$, this error is always non-negative, and global strong convexity is equivalent to imposing a stronger condition, namely that for some parameter $\gamma_\ell > 0$, the first-order Taylor error $\mathcal{T}_{\mathcal{L}}(\theta; \theta')$ is lower bounded by a quadratic term $\frac{\gamma_\ell}{2}\|\theta - \theta'\|^2$ for all $\theta, \theta' \in \Omega$. Global smoothness is defined in a similar way, by imposing a quadratic upper bound on the Taylor error. It is known that under global smoothness and strong convexity assumptions, the method of projected gradient descent (5.3) enjoys a *globally geometric convergence rate*, meaning that there is some $\kappa \in (0, 1)$ such that[1]

$$\left\|\theta^t - \widehat{\theta}\right\|^2 \lesssim \kappa^t \left\|\theta^0 - \widehat{\theta}\right\|^2 \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{5.7}$$

We refer the reader to Bertsekas [28, Prop. 1.2.3, p. 145], or Nesterov [120, Thm. 2.2.8, p. 88] for such results on projected gradient descent, and to Nesterov [121] for composite gradient descent.

[1]In this statement (and throughout the chapter), we use $\lesssim$ to mean an inequality that holds with some universal constant c, independent of the problem parameters.

Unfortunately, in the high-dimensional setting ($d > n$), it is usually impossible to guarantee strong convexity of the problem (5.1) in a global sense. For instance, when the data is drawn i.i.d., the loss function consists of a sum of n terms. If the loss is twice differentiable, the resulting $d \times d$ Hessian matrix $\nabla^2 \mathcal{L}(\theta; \phi(z_1^T;))$ is often a sum of n matrices each with rank one, so that the Hessian is rank-degenerate when $n < d$. However, as we show in this chapter, in order to obtain fast convergence rates for the optimization method (5.3), it is sufficient that (a) the objective is strongly convex and smooth in a restricted set of directions, and (b) the algorithm approaches the optimum $\widehat{\theta}$ only along these directions. Let us now formalize these ideas.

Definition 5.1 (Restricted strong convexity (RSC)). The loss function $\mathcal{L}_n$ satisfies restricted strong convexity with respect to $\mathcal{R}$ and with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ over the set Ω' if

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') \geq \frac{\gamma_\ell}{2} \|\theta - \theta'\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega'. \tag{5.8}$$

We refer to the quantity γ_ℓ as the *(lower) curvature parameter*, and to the quantity τ_ℓ as the *tolerance parameter*. The set Ω' corresponds to a suitably chosen subset of the space Ω of all possible parameters.

In order to gain intuition for this definition, first suppose that the condition (5.8) holds with tolerance parameter $\tau_\ell = 0$. In this case, the regularizer plays no role in the definition, and condition (5.8) is equivalent to the usual definition of strong convexity on the optimization set Ω. As discussed previously, this type of global strong convexity typically *fails* to hold for high-dimensional inference problems. In contrast, when tolerance parameter τ_ℓ is strictly positive, the condition (5.8) is much milder, in that it only applies to a *limited set* of vectors. For a given pair $\theta \neq \theta'$, consider the inequality

$$\frac{\mathcal{R}^2(\theta - \theta')}{\|\theta - \theta'\|^2} < \frac{\gamma_\ell}{2\, \tau_\ell(\mathcal{L}_n)}. \tag{5.9}$$

If this inequality is violated, then the right-hand side of the bound (5.8) is non-positive, in which case the RSC constraint (5.8) is vacuous. Thus, restricted strong convexity imposes a non-trivial constraint only on pairs $\theta \neq \theta'$ for which the inequality (5.8) holds, and a central part of our analysis will be to prove that, for the sequence of iterates generated by projected gradient descent, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies a constraint of the form (5.9). We note that since the regularizer $\mathcal{R}$ is convex, strong convexity of the loss function $\mathcal{L}_n$ also implies the strong convexity of the regularized loss ϕ_n as well.

For the least-squares loss, the RSC definition depends purely on the direction (and not the magnitude) of the difference vector $\theta - \theta'$. For other types of loss functions—such as those arising in generalized linear models—it is essential to localize the RSC definition, requiring that it holds only for pairs for which the norm $\|\theta - \theta'\|_2$ is not too large. We refer the reader to Section 5.2.4 for further discussion of this issue.

Finally, we observe that our restricted version of strong convexity can be seen as an instance of the general theory of paraconvexity (e.g., [124]); however, we are not aware of convergence rates for minimizing general paraconvex functions.

We also specify an analogous notion of restricted smoothness:

Definition 5.2 (**Restricted smoothness (RSM)**)**.** We say the loss function $\mathcal{L}_n$ satisfies restricted smoothness with respect to $\mathcal{R}$ and with parameters $(\gamma_u, \tau_u(\mathcal{L}_n))$ over the set Ω' if

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') \leq \frac{\gamma_u}{2} \|\theta - \theta'\|^2 + \tau_u(\mathcal{L}_n)\, \mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega'. \tag{5.10}$$

As with our definition of restricted strong convexity, the additional tolerance $\tau_u(\mathcal{L}_n)$ is not present in analogous smoothness conditions in the optimization literature, but it is essential in our set-up.

5.2.3 Decomposable regularizers

In past work on the statistical properties of regularization, the notion of a decomposable regularizer has been shown to be useful [116]. Although the focus of this chapter is a rather different set of questions—namely, optimization as opposed to statistics—decomposability also plays an important role here. Decomposability is defined with respect to a pair of subspaces defined with respect to the parameter space $\Omega \subseteq \mathbb{R}^d$. The set $\mathcal{M}$ is known as the *model subspace*, whereas the set $\overline{\mathcal{M}}^{\perp}$, referred to as the *perturbation subspace*, captures deviations away from the model subspace.

Definition 5.3. Given a subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ such that $\mathcal{M} \subseteq \overline{\mathcal{M}}$, we say that a norm $\mathcal{R}$ is $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$-*decomposable* if

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta) \qquad \text{for all } \alpha \in \mathcal{M} \text{ and } \beta \in \overline{\mathcal{M}}^{\perp}. \tag{5.11}$$

To gain some intuition for this definition, note that by triangle inequality, we always have the bound $\mathcal{R}(\alpha + \beta) \leq \mathcal{R}(\alpha) + \mathcal{R}(\beta)$. For a decomposable regularizer, this inequality always holds with equality. Thus, given a fixed vector $\alpha \in \mathcal{M}$, the key property of any decomposable regularizer is that it affords the *maximum penalization* of any deviation $\beta \in \overline{\mathcal{M}}^{\perp}$.

For a given error norm $\|\cdot\|$, its interaction with the regularizer $\mathcal{R}$ plays an important role in our results. In particular, we have the following:

Definition 5.4 (Subspace compatibility)**.** Given the regularizer $\mathcal{R}(\cdot)$ and a norm $\|\cdot\|$, the associated *subspace compatibility* is given by

$$\Psi(\overline{\mathcal{M}}) := \sup_{\theta \in \bar{\mathcal{M}} \setminus \{0\}} \frac{\mathcal{R}(\theta)}{\|\theta\|} \qquad \text{when } \overline{\mathcal{M}} \neq \{0\}, \qquad \text{and } \Psi(\{0\}) := 0. \tag{5.12}$$

The quantity $\Psi(\overline{\mathcal{M}})$ corresponds to the Lipschitz constant of the norm $\mathcal{R}$ with respect to $\|\cdot\|$, when restricted to the subspace $\overline{\mathcal{M}}$.

5.2.4 Some illustrative examples

We now describe some particular examples of M-estimators with decomposable regularizers, and discuss the form of the projected gradient updates as well as RSC/RSM conditions. We cover two main families of examples: log-linear models with sparsity constraints and ℓ_1-regularization (Section 5.2.4), and matrix regression problems with nuclear norm regularization (Section 5.2.4).

Sparse log-linear models and ℓ_1-regularization

Suppose that each sample Z_i consists of a scalar-vector pair $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$, corresponding to the scalar response $y_i \in \mathcal{Y}$ associated with a vector of predictors $x_i \in \mathbb{R}^d$. A log-linear model with canonical link function assumes that the response y_i is linked to the covariate vector x_i via a conditional distribution of the form $\mathbb{P}(y_i \mid x_i; \theta^*, \sigma) \propto \exp\left\{\frac{y_i \langle \theta^*, x_i\rangle - \Phi(\langle \theta^*, x_i\rangle)}{c(\sigma)}\right\}$, where $c(\sigma)$ is a known quantity, $\Phi(\cdot)$ is the log-partition function to normalize the density, and $\theta^* \in \mathbb{R}^d$ is an unknown regression vector. In many applications, the regression vector θ^* is relatively sparse, so that it is natural to impose an ℓ_1-constraint. Computing the maximum likelihood estimate subject to such a constraint involves solving the convex program[2]

$$\widehat{\theta} \in \arg\min_{\theta \in \Omega} \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \langle \theta, x_i \rangle - \Phi(\langle \theta, x_i \rangle) \right\} \right\}}_{\mathcal{L}_n(\theta; z_1^n)} \quad \text{such that } \|\theta\|_1 \leq \rho, \tag{5.13}$$

with $x_i \in \mathbb{R}^d$ as its i^{th} row. We refer to this estimator as the log-linear Lasso; it is a special case of the M-estimator (5.1), with the loss function $\mathcal{L}_n(\theta; z_1^n) = \frac{1}{n}\sum_{i=1}^n \left\{ y_i \langle \theta, x_i \rangle - \Phi(\langle \theta, x_i \rangle) \right\}$ and the regularizer $\mathcal{R}(\theta) = \|\theta\|_1 = \sum_{j=1}^d |\theta_j|$.

Ordinary linear regression is the special case of the log-linear setting with $\Phi(t) = t^2/2$ and $\Omega = \mathbb{R}^d$, and in this case, the estimator (5.13) corresponds to ordinary least-squares version of Lasso [51, 156]. Other forms of log-linear Lasso that are of interest include logistic regression, Poisson regression, and multinomial regression.

Projected gradient updates: Computing the gradient of the log-linear loss from equation (5.13) is straightforward: we have $\nabla \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n x_i \big\{ y_i - \Phi'(\langle \theta, x_i \rangle) \big\}$, and the update (5.5) corresponds to the Euclidean projection of the vector $\theta^t - \frac{1}{\gamma_u}\nabla \mathcal{L}_n(\theta^t)$ onto the ℓ_1-ball of radius ρ. It is well-known that this projection can be characterized in terms of

[2]The link function Φ is convex since it is the log-partition function of a canonical exponential family.

soft-thresholding, and that the projected update (5.5) can be computed easily. We refer the reader to Duchi et al. [60] for an efficient implementation requiring $\mathcal{O}(d)$ operations.

Composite gradient updates: The composite gradient update for this problem amounts to solving

$$\theta^{t+1} = \arg \min_{\|\theta\|_1 \leq \overline{\rho}} \left\{ \langle \theta, \nabla \mathcal{L}_n(\theta) \rangle + \frac{\gamma_u}{2} \|\theta - \theta^t\|_2^2 + \lambda_n \|\theta\|_1 \right\}.$$

The update can be computed by two soft-thresholding operations. The first step is soft thresholding the vector $\theta^t - \frac{1}{\gamma_u} \nabla \mathcal{L}_n(\theta^t)$ at a level λ_n. If the resulting vector has ℓ_1-norm greater than $\overline{\rho}$, then we project on to the ℓ_1-ball just like before. Overall, the complexity of the update is still $\mathcal{O}(d)$ as before.

Decomposability of ℓ_1-norm: We now illustrate how the ℓ_1-norm is decomposable with respect to appropriately chosen subspaces. For any subset $S \subseteq \{1, 2, \ldots, d\}$, consider the subspace

$$\mathcal{M}(S) := \big\{ \alpha \in \mathbb{R}^d \mid \alpha_j = 0 \quad \text{for all } j \notin S \big\}, \tag{5.14}$$

corresponding to all vectors supported only on S. Defining $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$, its orthogonal complement (with respect to the usual Euclidean inner product) is given by

$$\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S) = \big\{ \beta \in \mathbb{R}^d \mid \beta_j = 0 \quad \text{for all } j \in S \big\}. \tag{5.15}$$

To establish the decomposability of the ℓ_1-norm with respect to the pair $(\mathcal{M}(S), \overline{\mathcal{M}}^{\perp}(S))$, note that any $\alpha \in \mathcal{M}(S)$ can be written in the partitioned form $\alpha = (\alpha_S, 0_{S^c})$, where $\alpha_S \in \mathbb{R}^s$ and $0_{S^c} \in \mathbb{R}^{d-s}$ is a vector of zeros. Similarly, any vector $\beta \in \overline{\mathcal{M}}^{\perp}(S)$ has the partitioned representation $(0_S, \beta_{S^c})$. With these representations, we have the decomposition

$$\|\alpha + \beta\|_1 = \|(\alpha_S, 0) + (0, \beta_{S^c})\|_1 \;=\; \|\alpha\|_1 + \|\beta\|_1.$$

Consequently, for any subset S, the ℓ_1-norm is decomposable with respect to the pairs $(\mathcal{M}(S), \mathcal{M}^{\perp}(S))$.

In analogy to the ℓ_1-norm, various types of group-sparse norms are also decomposable with respect to non-trivial subspace pairs. We refer the reader to the paper [116] for further discussion and examples of such decomposable norms.

RSC/RSM conditions: A calculation using the mean-value theorem shows that for the loss function (5.13), the error in the first-order Taylor series, as previously defined in equation (5.6), can be written as

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') = \frac{1}{n} \sum_{i=1}^{n} \Phi''\big(\langle \theta_t, x_i \rangle\big) \big(\langle x_i, \theta - \theta' \rangle\big)^2$$

where $\theta_t = t\theta + (1-t)\theta'$ for some $t \in [0,1]$. When $n < d$, then we can always find pairs $\theta \neq \theta'$ such that $\langle x_i, \theta - \theta' \rangle = 0$ for all $i = 1, 2, \ldots, n$, showing that the objective function can never be strongly convex. On the other hand, restricted strong convexity for log-linear models requires only that there exist positive numbers $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \Phi''\big(\langle \theta_t, x_i \rangle\big) \big(\langle x_i, \theta - \theta' \rangle\big)^2 \geq \frac{\gamma_\ell}{2}\|\theta - \theta'\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega', \tag{5.16}$$

where $\Omega' := \Omega \cap \mathbb{B}_2(R)$ is the intersection of the parameter space Ω with a Euclidean ball of some fixed radius R around zero. This restriction is essential because for many generalized linear models, the Hessian function Φ'' approaches zero as its argument diverges. For instance, for the logistic function $\Phi(t) = \log(1+\exp(t))$, we have $\Phi''(t) = \exp(t)/[1+\exp(t)]^2$, which tends to zero as $t \to +\infty$. Restricted smoothness imposes an analogous upper bound on the Taylor error. For a broad class of log-linear models, such bounds hold with with tolerance $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ of the order $\sqrt{\frac{\log d}{n}}$. Further details on such results are provided in the corollaries to follow our main theorem. A detailed discussion of RSC for exponential families in statistical problems can be found in the paper [116].

In order to ensure RSC/RSM conditions on the iterates θ^t of the updates (5.3) or (5.4), we also need to ensure that $\theta^t \in \Omega'$. This can be done by defining $\mathcal{L}_n^{'} = \mathcal{L}_n + \mathbb{I}_{\Omega'}(\theta)$, where $\mathbb{I}_{\Omega'}(\theta)$ is zero when $\theta \in \Omega'$ and ∞ otherwise. This is equivalent to projection on the intersection of ℓ_1-ball with Ω' in the updates (5.3) and (5.4) and can be done efficiently with Dykstra's algorithm [64], for instance, as long as the individual projections are efficient.

In the special case of linear regression, we have $\Phi''(t) = 1$ for all $t \in \mathbb{R}$, so that the lower bound (5.16) involves only the Gram matrix $X^T X/n$. (Here $X \in \mathbb{R}^{n \times d}$ is the usual design matrix, with $x_i \in \mathbb{R}^d$ as its i^{th} row.) For linear regression and ℓ_1-regularization, the RSC condition is equivalent to the lower bound

$$\frac{\|X(\theta - \theta')\|_2^2}{n} \geq \frac{\gamma_\ell}{2}\|\theta - \theta'\|_2^2 - \tau_\ell(\mathcal{L}_n)\, \|\theta - \theta'\|_1^2 \qquad \text{for all } \theta, \theta' \in \Omega. \tag{5.17}$$

Such a condition corresponds to a variant of the restricted eigenvalue (RE) conditions that have been studied in the literature [30, 164]. Such RE conditions are significantly milder than the restricted isometry property; we refer the reader to van de Geer and Buhlmann [164] for an in-depth comparison of different RE conditions. From past work, the condition (5.17) is satisfied with high probability for a broad classes of anisotropic random design matrices [133, 143], and parts of our analysis make use of this fact.

Matrices and nuclear norm regularization

We now discuss a general class of matrix regression problems that falls within our framework. Consider the space of $d_1 \times d_2$ matrices endowed with the trace inner product $\langle\langle A, B \rangle\rangle := \text{trace}(A^T B)$.

In order to ease notation, we define $d := \min\{d_1, d_2\}$. Let $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ be an unknown matrix and suppose that for $i = 1, 2, \ldots, n$, we observe a scalar-matrix pair $Z_i = (y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$ linked to Θ^* via the linear model

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{5.18}$$

where w_i is an additive observation noise. In many contexts, it is natural to assume that Θ^* is exactly low-rank, or approximately so, meaning that it is well-approximated by a matrix of low rank. In such settings, a number of authors (e.g., [67, 141, 115]) have studied the M-estimator

$$\widehat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \langle\langle X_i, \Theta \rangle\rangle \big)^2 \Big\} \quad \text{such that } |||\Theta|||_1 \leq \rho, \tag{5.19}$$

or the corresponding regularized version. Here the *nuclear or trace norm* is given by $|||\Theta|||_1 := \sum_{j=1}^{d} \sigma_j(\Theta)$, corresponding to the sum of the singular values. This optimization problem is an instance of a semidefinite program. As discussed in more detail in Section 5.3.3, there are various applications in which this estimator and variants thereof have proven useful.

Form of projected gradient descent: For the M-estimator (5.19), the projected gradient updates take a very simple form—namely

$$\Theta^{t+1} = \Pi\Big(\Theta^t - \frac{1}{\gamma_u} \frac{\sum_{i=1}^{n} \big(y_i - \langle\langle X_i, \Theta^t \rangle\rangle \big) X_i}{n} \Big), \tag{5.20}$$

where Π denotes Euclidean projection onto the nuclear norm ball $\mathbb{B}_1(\rho) = \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid |||\Theta|||_1 \leq \rho\}$. This nuclear norm projection can be obtained by first computing the singular value decomposition (SVD), and then projecting the vector of singular values onto the ℓ_1-ball. The latter step can be achieved by the fast projection algorithms discussed earlier, and there are various methods for fast computation of SVDs. The composite gradient update also has a simple form, requiring at most two singular value thresholding operations as was the case for linear regression.

Decomposability of nuclear norm: We now define matrix subspaces for which the nuclear norm is decomposable. Given a target matrix Θ^*—that is, a quantity to be estimated—consider its singular value decomposition $\Theta^* = UDV^T$, where the matrix $D \in \mathbb{R}^{d \times d}$ is diagonal, with the ordered singular values of Θ^* along its diagonal, and $d := \min\{d_1, d_2\}$. For an integer $r \in \{1, 2, \ldots, d\}$, let $U^r \in \mathbb{R}^{d \times r}$ denote the matrix formed by the top r left singular vectors of Θ^* in its columns, and we define the matrix V^r in a similar fashion. Using col to

denote the column span of a matrix, we then define the subspaces[3]

$$\mathcal{M}(U^r, V^r) := \big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \mathrm{col}(\Theta^T) \subseteq \mathrm{col}(V^r), \ \mathrm{col}(\Theta) \subseteq \mathrm{col}(U^r)\big\}, \quad \text{and} \tag{5.21a}$$

$$\overline{\mathcal{M}}^{\perp}(U^r, V^r) := \big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \mathrm{col}(\Theta^T) \subseteq (\mathrm{col}(V^r))^{\perp}, \ \mathrm{col}(\Theta) \subseteq (\mathrm{col}(U^r))^{\perp}\big\}. \tag{5.21b}$$

Finally, let us verify the decomposability of the nuclear norm . By construction, any pair of matrices $\Theta \in \mathcal{M}(U^r, V^r)$ and $\Gamma \in \overline{\mathcal{M}}^{\perp}(U^r, V^r)$ have orthogonal row and column spaces, which implies the required decomposability condition—namely $|||\Theta + \Gamma|||_1 = |||\Theta|||_1 + |||\Gamma|||_1$.

In some special cases such as matrix completion or matrix decomposition that we describe in the sequel, Ω' will involve an additional bound on the entries of Θ^* as well as the iterates Θ^t to establish RSC/RSM conditions. This can be done by augmenting the loss with an indicator of the constraint and using cyclic projections for computing the updates as mentioned earlier in Example 5.2.4.

5.3 Main results and some consequences

We are now equipped to state the two main results of this chapter, and discuss some of their consequences. We illustrate its application to several statistical models, including sparse regression (Section 5.3.2), matrix estimation with rank constraints (Section 5.3.3), and matrix decomposition problems (Section 5.3.4).

5.3.1 Geometric convergence

Recall that the projected gradient algorithm (5.3) is well-suited to solving an M-estimation problem in its constrained form, whereas the composite gradient algorithm (5.4) is appropriate for a regularized problem. Accordingly, let $\widehat{\theta}$ be any optimal solution to the constrained problem (5.1), or the regularized problem (5.2), and let $\{\theta^t\}_{t=0}^{\infty}$ be a sequence of iterates generated by generated by the projected gradient updates (5.3), or the the composite gradient updates (5.4), respectively. Of primary interest to us in this work are bounds on the *optimization error*, which can be measured either in terms of the error vector $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, or the difference between the cost of θ^t and the optimal cost defined by $\widehat{\theta}$. In this section, we state two main results —Theorems 5.1 and 5.2—corresponding to the constrained and regularized cases respectively. In addition to the optimization error previously discussed, both of these results involve the *statistical error* $\Delta^* := \widehat{\theta} - \theta^*$ between the optimum $\widehat{\theta}$ and the nominal parameter θ^*. At a high level, these results guarantee that under the RSC/RSM conditions, the optimization error shrinks geometrically, with a contraction coefficient that depends on the the loss function $\mathcal{L}_n$ via the parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$. An interesting feature is that the contraction occurs only up to a certain tolerance parameter ϵ^2

[3] Note that the model space $\mathcal{M}(U^r, V^r)$ is *not equal* to $\overline{\mathcal{M}}(U^r, V^r)$. Nonetheless, as required by Definition 5.3, we do have the inclusion $\mathcal{M}(U^r, V^r) \subseteq \overline{\mathcal{M}}(U^r, V^r)$.

depending on these same parameters, and the statistical error. However, as we discuss, for many statistical problems of interest, we can show that this tolerance parameter is of lower order than the intrinsic statistical error, and hence can be neglected from the statistical point of view. Consequently, our theory gives an explicit upper bound on the number of iterations required to solve an M-estimation problem up to statistical precision.

Convergence rates for projected gradient: We now provide the notation necessary for a precise statement of this claim. Our main result actually involves a family of upper bounds on the optimization error, one for each pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ of $\mathcal{R}$-decomposable subspaces (see Definition 5.3). As will be clarified in the sequel, this subspace choice can be optimized for different models so as to obtain the tightest possible bounds. For a given pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ such that $16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n) < \gamma_u$, let us define the *contraction coefficient*

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\}\Big\{1 - \frac{16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\gamma_u}\Big\}^{-1}. \quad (5.22)$$

In addition, we define the *tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\,\big(2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \Psi(\overline{\mathcal{M}})\,\|\Delta^*\| + 2\mathcal{R}(\Delta^*)\big)^2}{\gamma_u}, \quad (5.23)$$

where $\Delta^* = \widehat{\theta} - \theta^*$ is the statistical error, and $\Pi_{\mathcal{M}^\perp}(\theta^*)$ denotes the Euclidean projection of θ^* onto the subspace $\mathcal{M}^\perp$.

In terms of these two ingredients, we now state our first main result:

Theorem 5.1. *Suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$ respectively. Let $(\mathcal{M}, \overline{\mathcal{M}})$ be any $\mathcal{R}$-decomposable pair of subspaces such that $\mathcal{M} \subseteq \overline{\mathcal{M}}$ and $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. Then for any optimum $\widehat{\theta}$ of the problem (5.1) for which the constraint is active, we have*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t\,\|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{1-\kappa} \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \quad (5.24)$$

Remarks: Theorem 5.1 actually provides a family of upper bounds, one for each $\mathcal{R}$-decomposable pair $(\mathcal{M}, \overline{\mathcal{M}})$ such that $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. This condition is always satisfied by setting $\overline{\mathcal{M}}$ equal to the trivial subspace $\{0\}$: indeed, by definition (5.12) of the subspace compatibility, we have $\Psi(\overline{\mathcal{M}}) = 0$, and hence $\kappa(\mathcal{L}_n; \{0\}) = \big(1 - \frac{\gamma_\ell}{\gamma_u}\big) < 1$. Although this choice of $\overline{\mathcal{M}}$ minimizes the contraction coefficient, it will lead[4] to a very large tolerance

[4]Indeed, the setting $\mathcal{M}^\perp = \mathbb{R}^d$ means that the term $\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) = \mathcal{R}(\theta^*)$ appears in the tolerance; this quantity is far larger than statistical precision.

parameter $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$. A more typical application of Theorem 5.1 involves non-trivial choices of the subspace $\overline{\mathcal{M}}$.

The bound (5.24) guarantees that the optimization error decreases geometrically, with contraction factor $\kappa \in (0, 1)$, up to a certain tolerance proportional to $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as illustrated in Figure 5.2(a). The contraction factor κ approaches the $1 - \gamma_\ell/\gamma_u$ as the number of samples grows. The appearance of the ratio γ_ℓ/γ_u is natural since it measures the conditioning of the objective function; more specifically, it is essentially a restricted condition number of the Hessian matrix. On the other hand, the tolerance parameter ϵ depends on the choice of decomposable subspaces, the parameters of the RSC/RSM conditions, and the statistical error $\Delta^* = \widehat{\theta} - \theta^*$ (see equation (5.23)). In the corollaries of Theorem 5.1 to follow, we show that the subspaces can often be chosen such that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) = o(\|\widehat{\theta} - \theta^*\|^2)$. Consequently, the bound (5.24) guarantees geometric convergence up to a tolerance *smaller than statistical precision*, as illustrated in Figure 5.2(b). This is sensible, since in statistical settings, there is no point to optimizing beyond the statistical precision.

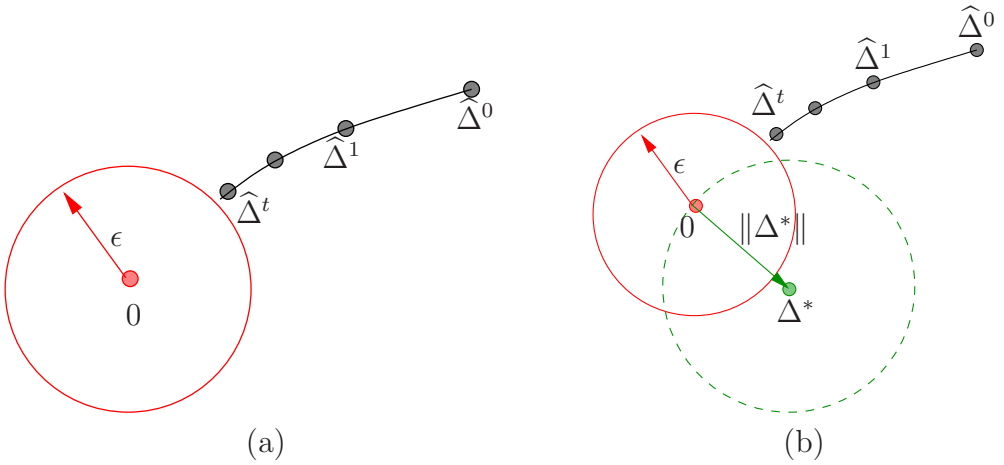

Figure 5.2. (a) Generic illustration of Theorem 5.1. The optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ is guaranteed to decrease geometrically with coefficient $\kappa \in (0, 1)$, up to the tolerance $\epsilon^2 = \epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, represented by the circle. (b) Relation between the optimization tolerance $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ (solid circle) and the statistical precision $\|\Delta^*\| = \|\theta^* - \widehat{\theta}\|$ (dotted circle). In many settings, we have $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \ll \|\Delta^*\|^2$, so that convergence is guaranteed up to a tolerance lower than statistical precision.

The result of Theorem 5.1 takes a simpler form when there is a subspace $\mathcal{M}$ that includes θ^*, and the $\mathcal{R}$-ball radius is chosen such that $\rho \leq \mathcal{R}(\theta^*)$. In this case, by appropriately controlling the error term, we can establish that it is of lower order than the statistical precision —namely, the squared difference $\left\|\widehat{\theta} - \theta^*\right\|^2$ between an optimal solution $\widehat{\theta}$ to the convex program (5.1), and the unknown parameter θ^*.

Corollary 5.1. *In addition to the conditions of Theorem 5.1, suppose that* $\theta^* \in \mathcal{M}$ *and* $\rho \leq \mathcal{R}(\theta^*)$. *Then as long as* $\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big) = o(1)$, *we have*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + o\big(\|\widehat{\theta} - \theta^*\|^2\big) \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{5.25}$$

Thus, Corollary 5.1 guarantees that the optimization error decreases geometrically, with contraction factor κ, up to a tolerance that is of strictly lower order than the statistical precision $\|\widehat{\theta} - \theta^*\|^2$. As will be clarified in several examples to follow, the condition $\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big) = o(1)$ is satisfied for many statistical models, including sparse linear regression and low-rank matrix regression. This result is illustrated in Figure 5.2(b), where the solid circle represents the optimization tolerance, and the dotted circle represents the statistical precision. In the results to follow, we will quantify the term $o\big(\|\widehat{\theta} - \theta^*\|^2\big)$ in a more precise manner for different statistical models.

Convergence rates for composite gradient: We now present our main result for the composite gradient iterates (5.4) that are suitable for the Lagrangian-based estimator (5.2). As before, our analysis yields a range of bounds indexed by subspace pairs $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ that are $\mathcal{R}$-decomposable. For any subspace $\overline{\mathcal{M}}$ such that $64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}) < \gamma_\ell$, we define *effective RSC coefficient* as

$$\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}). \tag{5.26}$$

This coefficient accounts for the residual amount of strong convexity after accounting for the lower tolerance terms. In addition, we define the *compound contraction coefficient* as

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \left\{1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\overline{\gamma_\ell}}\right\} \, \xi(\overline{\mathcal{M}}) \tag{5.27}$$

where $\xi(\overline{\mathcal{M}}) := \big(1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\bar{\mathcal{M}})}{\overline{\gamma_\ell}}\big)^{-1}$, and $\Delta^* = \widehat{\theta}_{\lambda_n} - \theta^*$ is the statistical error vector[5] for a specific choice of $\overline{\rho}$ and λ_n. As before, the coefficient κ measures the geometric rate of convergence for the algorithm. Finally, we define the *compound tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := 8\,\xi(\overline{\mathcal{M}})\,\beta(\overline{\mathcal{M}})\,\big(6\Psi(\overline{\mathcal{M}})\,\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))\big)^2, \tag{5.28}$$

where $\beta(\overline{\mathcal{M}}) := 2\left(\frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{128\tau_u(\mathcal{L}_n)\Psi^2(\bar{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\tau_\ell(\mathcal{L}_n) + 8\tau_u(\mathcal{L}_n) + 2\tau_\ell(\mathcal{L}_n)$. As with our previous result, the tolerance parameter determines the radius up to which geometric convergence can be attained.

Recall that the regularized problem (5.2) involves both a regularization weight λ_n, and a constraint radius $\overline{\rho}$. Our theory requires that the constraint radius is chosen such that

[5] When the context is clear, we remind the reader that we drop the subscript λ_n on the parameter $\widehat{\theta}$.

$\overline{\rho} \geq \mathcal{R}(\theta^*)$, which ensures that θ^* is feasible. In addition, the regularization parameter should be chosen to satisfy the constraint

$$\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*)), \tag{5.29}$$

where $\mathcal{R}^*$ is the dual norm of the regularizer. This constraint is known to play an important role in proving bounds on the statistical error of regularized M-estimators (see the paper [116] and references therein for further details). Recalling the definition (5.2) of the overall objective function $\phi_n(\theta)$, the following result provides bounds on the *excess loss* $\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n})$.

Theorem 5.2. *Consider the optimization problem (5.2) for a radius $\overline{\rho}$ such that θ^* is feasible, and a regularization parameter λ_n satisfying the bound (5.29), and suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$ respectively. Let $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ be any $\mathcal{R}$-decomposable pair such that*

$$\kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) \in [0,1), \quad \text{and} \quad \frac{32\,\overline{\rho}}{1-\kappa(\mathcal{L}_n; \overline{\mathcal{M}})}\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}}) \leq \lambda_n. \tag{5.30}$$

Then for any tolerance parameter $\delta^2 \geq \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{(1-\kappa)}$, we have

$$\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n}) \leq \delta^2 \quad \text{for all} \quad t \geq \frac{2\log\frac{\phi_n(\theta^0)-\phi_n(\widehat{\theta}_{\lambda_n})}{\delta^2}}{\log(1/\kappa)} + \log_2\log_2\left(\frac{\overline{\rho}\lambda_n}{\delta^2}\right)\left(1 + \frac{\log 2}{\log(1/\kappa)}\right). \tag{5.31}$$

Remarks: Note that the bound (5.31) guarantees the excess loss $\phi_n(\theta^t) - \phi_n(\widehat{\theta})$ decays geometrically up to any squared error δ^2 larger than the compound tolerance (5.28). Moreover, the RSC condition also allows us to translate this bound on objective values to a bound on the optimization error $\theta^t - \widehat{\theta}$. In particular, for any iterate θ^t such that $\phi_n(\theta^t) - \phi_n(\widehat{\theta}) \leq \delta^2$, we are guaranteed that

$$\left\|\theta^t - \widehat{\theta}_{\lambda_n}\right\|^2 \leq \frac{2\delta^2}{\overline{\gamma}_\ell} + \frac{16\delta^2\tau_\ell(\mathcal{L}_n)}{\overline{\gamma}_\ell\lambda_n^2} + \frac{4\tau_\ell(\mathcal{L}_n)(6\Psi(\overline{\mathcal{M}}) + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)))^2}{\overline{\gamma}_\ell}. \tag{5.32}$$

In conjunction with Theorem 5.2, we see that it suffices to take a number of steps that is logarithmic in the inverse tolerance $(1/\delta)$, again showing a geometric rate of convergence.

Whereas Theorem 5.1 requires setting the radius so that the constraint is active, Theorem 5.2 has only a very mild constraint on the radius $\overline{\rho}$, namely that it be large enough such that $\overline{\rho} \geq \mathcal{R}(\theta^*)$. The reason for this much milder requirement is that the additive regularization with weight λ_n suffices to constrain the solution, whereas the extra side constraint is only needed to ensure good behavior of the optimization algorithm in the first few iterations.

The regularization parameter λ_n must satisfy the so-called dual norm condition (5.29), which has appeared in past literature on statistical estimation, and is well-characterized for a broad range of statistical models (e.g., see the paper [116] and references therein).

Step-size setting: It seems that the updates (5.3) and (5.4) need to know the smoothness bound γ_u in order to set the step-size for gradient updates. However, we can use the same doubling trick as described in Algorithm (3.1) of Nesterov [121]. At each step, we check if the smoothness upper bound holds at the current iterate relative to the previous one. If the condition does not hold, we double our estimate of γ_u and resume. This guarantees a geometric convergence with a contraction factor worse at most by a factor of 2, compared to the knowledge of γ_u. We refer the reader to Nesterov [121] for details.

The following subsections are devoted to the development of some consequences of Theorems 5.1 and 5.2 and Corollary 5.1 for some specific statistical models, among them sparse linear regression with ℓ_1-regularization, and matrix regression with nuclear norm regularization. In contrast to the entirely deterministic arguments that underlie the Theorems 5.1 and 5.2, these corollaries involve probabilistic arguments, more specifically in order to establish that the RSC and RSM properties hold with high probability.

5.3.2 Sparse vector regression

Recall from Section 5.2.4 the observation model for sparse linear regression. In a variety of applications, it is natural to assume that θ^* is sparse. For a parameter $q \in [0, 1]$ and radius $R_q > 0$, let us define the ℓ_q "ball"

$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\beta_j|^q \leq R_q\big\}. \tag{5.33}$$

Note that $q = 0$ corresponds to the case of "hard sparsity", for which any vector $\beta \in \mathbb{B}_0(R_0)$ is supported on a set of cardinality at most R_0. For $q \in (0, 1]$, membership in the set $\mathbb{B}_q(R_q)$ enforces a decay rate on the ordered coefficients, thereby modeling approximate sparsity. In order to estimate the unknown regression vector $\theta^* \in \mathbb{B}_q(R_q)$, we consider the least-squares Lasso estimator from Section 5.2.4, based on the quadratic loss function $\mathcal{L}(\theta; Z_1^n) := \frac{1}{2n}\|y - X\theta\|_2^2$, where $X \in \mathbb{R}^{n\times d}$ is the design matrix. In order to state a concrete result, we consider a random design matrix X, in which each row $x_i \in \mathbb{R}^d$ is drawn i.i.d. from a $N(0, \Sigma)$ distribution, where Σ is a positive definite covariance matrix. We refer to this as the Σ-*ensemble of random design matrices*, and use $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ to refer the maximum and minimum eigenvalues of Σ respectively, and $\zeta(\Sigma) := \max_{j=1,2,\ldots,d} \Sigma_{jj}$ for the maximum variance. We also assume that the observation noise is zero-mean and sub-Gaussian with parameter ν^2.

Guarantees for constrained Lasso: Our convergence rate on the optimization error $\theta^t - \widehat{\theta}$ is stated in terms of the contraction coefficient

$$\kappa := \left\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\right\}\left\{1 - \chi_n(\Sigma)\right\}^{-1}, \tag{5.34}$$

where we have adopted the shorthand

$$\chi_n(\Sigma) := \begin{cases} \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} R_q \left(\frac{\log d}{n}\right)^{1-q/2} & \text{for } q > 0 \\ \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} s \left(\frac{\log d}{n}\right) & \text{for } q = 0 \end{cases}, \qquad \text{for a numerical constant } c_0, \tag{5.35}$$

We assume that $\chi_n(\Sigma)$ is small enough to ensure that $\kappa \in (0,1)$; in terms of the sample size, this amounts to a condition of the form $n = \Omega(R_q^{1/(1-q/2)} \log d)$. Such a scaling is sensible, since it is known from minimax theory on sparse linear regression [134] to be necessary for any method to be statistically consistent over the ℓ_q-ball.

With this set-up, we have the following consequence of Theorem 5.1:

Corollary 5.2 (Sparse vector recovery)**.** *Under conditions of Theorem 5.1, suppose that we solve the constrained Lasso with $\rho \leq \|\theta^*\|_1$.*

(a) Exact sparsity: *If θ^* is supported on a subset of cardinality s, then with probability at least $1 - \exp(-c_1 \log d)$, the iterates (5.3) with $\gamma_u = 2\sigma_{\max}(\Sigma)$ satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma)\, \|\widehat{\theta} - \theta^*\|_2^2 \qquad \text{for all } t = 0, 1, 2, \ldots. \tag{5.36}$$

(b) Weak sparsity: *Suppose that $\theta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0,1]$. Then with probability at least $1 - \exp(-c_1 \log d)$, the iterates (5.3) with $\gamma_u = 2\sigma_{\max}(\Sigma)$ satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t\, \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma) \left\{R_q\Big(\frac{\log d}{n}\Big)^{1-q/2} + \|\widehat{\theta} - \theta^*\|_2^2\right\}. \tag{5.37}$$

We provide the proof of Corollary 5.2 in Section 5.5.4. Here we compare part (a), which deals with the special case of exactly sparse vectors, to some past work that has established convergence guarantees for optimization algorithms for sparse linear regression. Certain methods are known to converge at sublinear rates (e.g., [24]), more specifically at the rate $\mathcal{O}(1/t^2)$. The geometric rate of convergence guaranteed by Corollary 5.2 is exponentially faster. Other work on sparse regression has provided geometric rates of convergence that hold once the iterates are close to the optimum [37, 71], or geometric convergence up to the noise level ν^2 using various methods, including greedy methods [158] and thresholded gradient methods [68]. In contrast, Corollary 5.2 guarantees geometric convergence for all iterates up to a precision below that of statistical error. For these problems, the statistical

error $\frac{\nu^2 s \log d}{n}$ is typically much smaller than the noise variance ν^2, and decreases as the sample size is increased.

In addition, Corollary 5.2 also applies to the case of approximately sparse vectors, lying within the set $\mathbb{B}_q(R_q)$ for $q \in (0, 1]$. There are some important differences between the case of exact sparsity (Corollary 5.2(a)) and that of approximate sparsity (Corollary 5.2(b)). Part (a) guarantees geometric convergence to a tolerance depending only on the statistical error $\|\widehat{\theta} - \theta^*\|_2$. In contrast, the second result also has the additional term $R_q\big(\frac{\log d}{n}\big)^{1-q/2}$. This second term arises due to the statistical non-identifiability of linear regression over the ℓ_q-ball, and it is no larger than $\left\|\widehat{\theta} - \theta^*\right\|_2^2$ with high probability. This assertion follows from known results [134] about minimax rates for linear regression over ℓ_q-balls; these unimprovable rates include a term of this order.

Guarantees for regularized Lasso: Using similar methods, we can also use Theorem 5.2 to obtain an analogous guarantee for the regularized Lasso estimator. Here focus only on the case of exact sparsity, although the result extends to approximate sparsity in a similar fashion. Letting $c_i, i = 0, 1, 2, 3, 4$ be universal positive constants, we define the modified curvature constant $\overline{\gamma_\ell} := \gamma_\ell - c_0 \frac{s \log d}{n} \zeta(\Sigma)$. Our results assume that $n = \Omega(s \log d)$, a condition known to be necessary for statistical consistency, so that $\overline{\gamma_\ell} > 0$. The contraction factor then takes the form

$$\kappa := \big\{1 - \frac{\sigma_{\min}(\Sigma)}{16\sigma_{\max}(\Sigma)} + c_1\chi_n(\Sigma)\big\}\,\big\{1 - c_2\chi_n(\Sigma)\big\}^{-1}, \quad \text{where} \quad \chi_n(\Sigma) = \frac{\zeta(\Sigma)}{\overline{\gamma_\ell}}\,\frac{s\,\log d}{n}.$$

The tolerance factor in the optimization is given by

$$\epsilon_{\text{tol}}^2 := \frac{5 + c_2\chi_n(\Sigma)}{1 - c_3\chi_n(\Sigma)}\,\frac{\zeta(\Sigma)\; s \log d}{n}\|\theta^* - \widehat{\theta}\|_2^2, \tag{5.38}$$

where $\theta^* \in \mathbb{R}^d$ is the unknown regression vector, and $\widehat{\theta}$ is any optimal solution. With this notation, we have the following corollary.

Corollary 5.3 (Regularized Lasso)**.** *Under conditions of Theorem 5.2, suppose that we solve the regularized Lasso with* $\lambda_n = 6\sqrt{\frac{\nu \log d}{n}}$*, and that* θ^* *is supported on a subset of cardinality at most* s*. Then with probability at least* $1 - \exp(-c_4 \log d)$*, for any* $\delta^2 \geq \epsilon_{tol}^2$*, for any optimum* $\widehat{\theta}_{\lambda_n}$*, we have*

$$\|\theta^t - \widehat{\theta}_{\lambda_n}\|_2^2 \leq \delta^2 \qquad \textit{for all iterations } t \;\geq\; \big(\log \tfrac{\phi_n(\theta^0) - \phi_n(\widehat{\theta}_{\lambda_n})}{\delta^2}\big)/\big(\log \tfrac{1}{\kappa}\big).$$

As with Corollary 5.2(a), this result guarantees that $\mathcal{O}(\log(1/\epsilon_{\text{tol}}^2))$ iterations are sufficient to obtain an iterate θ^t that is within squared error $\mathcal{O}(\epsilon_{\text{tol}}^2)$ of any optimum $\widehat{\theta}_{\lambda_n}$. Moreover, whenever $\frac{s \log d}{n} = o(1)$—a condition that is required for statistical consistency of *any method*—the optimization tolerance ϵ_{tol}^2 is of lower order than the statistical error $\|\theta^* - \theta\|_2^2$.

5.3.3 Matrix regression with rank constraints

We now turn estimation of matrices under various types of "soft" rank constraints. Recall the model of matrix regression from Section 5.2.4, and the M-estimator based on least-squares regularized with the nuclear norm (5.19). So as to reduce notational overhead, here we specialize to square matrices $\Theta^* \in \mathbb{R}^{d \times d}$, so that our observations are of the form

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{5.39}$$

where $X_i \in \mathbb{R}^{d \times d}$ is a matrix of covariates, and $w_i \sim N(0, \nu^2)$ is Gaussian noise. As discussed in Section 5.2.4, the nuclear norm $\mathcal{R}(\Theta) = |||\Theta|||_1 = \sum_{j=1}^{d} \sigma_j(\Theta)$ is decomposable with respect to appropriately chosen matrix subspaces, and we exploit this fact heavily in our analysis.

We model the behavior of both exactly and approximately low-rank matrices by enforcing a sparsity condition on the vector $\sigma(\Theta) = \begin{bmatrix} \sigma_1(\Theta) & \sigma_2(\Theta) & \cdots & \sigma_d(\Theta) \end{bmatrix}$ of singular values. In particular, for a parameter $q \in [0, 1]$, we define the ℓ_q-"ball" of matrices

$$\mathbb{B}_q(R_q) := \Big\{ \Theta \in \mathbb{R}^{d \times d} \mid \sum_{j=1}^{d} |\sigma_j(\Theta)|^q \leq R_q \Big\}. \tag{5.40}$$

Note that if $q = 0$, then $\mathbb{B}_0(R_0)$ consists of the set of all matrices with rank at most $r = R_0$. On the other hand, for $q \in (0, 1]$, the set $\mathbb{B}_q(R_q)$ contains matrices of all ranks, but enforces a relatively fast rate of decay on the singular values.

Bounds for matrix compressed sensing

We begin by considering the compressed sensing version of matrix regression, a model first introduced by Recht et al. [136], and later studied by other authors (e.g., [100, 115]). In this model, the observation matrices $X_i \in \mathbb{R}^{d \times d}$ are dense and drawn from some random ensemble. The simplest example is the standard Gaussian ensemble, in which each entry of X_i is drawn i.i.d. as standard normal $N(0, 1)$. Note that X_i is a dense matrix in general; this in an important contrast with the matrix completion setting to follow shortly.

Here we consider a more general ensemble of random matrices X_i, in which each matrix $X_i \in \mathbb{R}^{d \times d}$ is drawn i.i.d. from a zero-mean normal distribution in $\mathbb{R}^{d^2}$ with covariance matrix $\Sigma \in \mathbb{R}^{d^2 \times d^2}$. The setting $\Sigma = I_{d^2 \times d^2}$ recovers the standard Gaussian ensemble studied in past work. As usual, we let $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ define the maximum and minimum eigenvalues of Σ, and we define $\zeta_{\text{mat}}(\Sigma) = \sup_{\|u\|_2 = 1} \sup_{\|v\|_2 = 1} \operatorname{var}\big(\langle\langle X, uv^T \rangle\rangle\big)$, corresponding to the maximal variance of X when projected onto rank one matrices. For the identity ensemble, we have $\zeta_{\text{mat}}(I) = 1$.

We now state a result on the convergence of the updates (5.20) when applied to a statistical problem involving a matrix $\Theta^* \in \mathbb{B}_q(R_q)$. The convergence rate depends on the contraction coefficient

$$\kappa := \Big\{ 1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma) \Big\} \Big\{ 1 - \chi_n(\Sigma) \Big\}^{-1},$$

where $\chi_n(\Sigma) := \frac{c_1 \zeta_{\text{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} R_q \left(\frac{d}{n}\right)^{1-q/2}$ for some universal constant c_1. In the case $q = 0$, corresponding to matrices with rank at most r, note that we have $R_0 = r$. With this notation, we have the following convergence guarantee:

Corollary 5.4 (Low-rank matrix recovery)**.** *Under conditions of Theorem 5.1, consider the semidefinite program (5.19) with $\rho \leq |||\Theta^*|||_1$, and suppose that we apply the projected gradient updates (5.20) with $\gamma_u = 2\sigma_{\max}(\Sigma)$.*

(a) Exactly low-rank: *In the case $q = 0$, if Θ^* has rank $r < d$, then with probability at least $1 - \exp(-c_0 d)$, the iterates (5.20) satisfy the bound*

$$|||\Theta^t - \widehat{\Theta}|||_F^2 \leq \kappa^t |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \, \chi_n(\Sigma) \, |||\widehat{\Theta} - \Theta^*|||_F^2 \qquad \text{for all } t = 0, 1, 2, \ldots. \quad (5.41)$$

(b) Approximately low-rank: *If $\Theta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0, 1]$, then with probability at least $1 - \exp(-c_0 d)$, the iterates (5.20) satisfy*

$$|||\Theta^t - \widehat{\Theta}|||_F^2 \leq \kappa^t \, |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \chi_n(\Sigma) \left\{ R_q \left(\frac{d}{n} \right)^{1-q/2} + |||\widehat{\Theta} - \Theta^*|||_F^2 \right\}, \quad (5.42)$$

Although quantitative aspects of the rates are different, Corollary 5.4 is analogous to Corollary 5.2. For the case of exactly low rank matrices (part (a)), geometric convergence is guaranteed up to a tolerance involving the statistical error $|||\widehat{\Theta} - \Theta^*|||_F^2$. For the case of approximately low rank matrices (part (b)), the tolerance term involves an additional factor of $R_q \left(\frac{d}{n}\right)^{1-q/2}$. Again, from known results on minimax rates for matrix estimation [141], this term is known to be of comparable or lower order than the quantity $|||\widehat{\Theta} - \Theta^*|||_F^2$. As before, it is also possible to derive an analogous corollary of Theorem 5.2 for estimating low-rank matrices; in the interests of space, we leave such a development to the reader.

Bounds for matrix completion

In this model, observation y_i is a noisy version of a randomly selected entry $\Theta^*_{a(i),b(i)}$ of the unknown matrix Θ^*. Applications of this matrix completion problem include collaborative filtering [152], where the rows of the matrix Θ^* correspond to users, and the columns correspond to items (e.g., movies in the Netflix database), and the entry Θ^*_{ab} corresponds to user's a rating of item b. Given observations of only a subset of the entries of Θ^*, the goal is to fill in, or complete the matrix, thereby making recommendations of movies that a given user has not yet seen.

Matrix completion can be viewed as a particular case of the matrix regression model (5.18), in particular by setting $X_i = E_{a(i)b(i)}$, corresponding to the matrix with a single one in position $(a(i), b(i))$, and zeroes in all other positions. Note that these observation matrices are extremely sparse, in contrast to the compressed sensing model. Nuclear-norm based estimators

for matrix completion are known to have good statistical properties (e.g., [44, 135, 152, 114]). Here we consider the M-estimator

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega} \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \Theta_{a(i)b(i)}\big)^2 \quad \text{such that } |||\Theta|||_1 \leq \rho, \tag{5.43}$$

where $\Omega = \{\Theta \in \mathbb{R}^{d \times d} \mid \|\Theta\|_\infty \leq \frac{\alpha}{d}\}$ is the set of matrices with bounded elementwise ℓ_∞ norm. This constraint eliminates matrices that are overly "spiky" (i.e., concentrate too much of their mass in a single position); as discussed in the paper [114], such spikiness control is necessary in order to bound the non-identifiable component of the matrix completion model.

Corollary 5.5 (Matrix completion)**.** *Under the conditions of Theorem 5.1, suppose that* $\Theta^* \in \mathbb{B}_q(R_q)$*, and that we solve the program (5.43) with* $\rho \leq |||\Theta^*|||_1$*. As long as* $n > c_0 R_q^{1/(1-q/2)}\, d \log d$ *for a sufficiently large constant* c_0*, then with probability at least* $1 - \exp(-c_1 d \log d)$*, there is a contraction coefficient* $\kappa_t \in (0,1)$ *that decreases with* t *such that for all iterations* $t = 0, 1, 2, \ldots$*,*

$$|||\Theta^{t+1} - \widehat{\Theta}|||_F^2 \leq \kappa_t^t\, |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \Big\{ R_q \big(\frac{\alpha^2 d \log d}{n}\big)^{1-q/2} + |||\widehat{\Theta} - \Theta^*|||_F^2 \Big\}. \tag{5.44}$$

In some cases, the bound on $\|\Theta\|_\infty$ in the algorithm (5.43) might be unknown, or undesirable. While this constraint is necessary in general [114], it can be avoided if more information such as the sampling distribution (that is, the distribution of X_i) is known and used to construct the estimator. In this case, Koltchinskii et al. [89] show error bounds on a nuclear norm penalized estimator without requiring ℓ_∞ bound on $\widehat{\Theta}$.

Again a similar corollary of Theorem 5.2 can be derived by combining the proof of Corollary 5.5 with that of Theorem 5.2. An interesting aspect of this problem is that the condition 5.29(b) takes the form $\lambda_n > \frac{c\alpha\sqrt{d \log d/n}}{1-\kappa}$, where α is a bound on $\|\Theta\|_\infty$. This condition is independent of $\overline{\rho}$, and hence, given a sample size as stated in the corollary, the algorithm always converges geometrically for any radius $\overline{\rho} \geq |||\Theta^*|||_1$.

5.3.4 Matrix decomposition problems

In recent years, various researchers have studied methods for solving the problem of matrix decomposition (e.g., [49, 45, 174, 8, 78]). The basic problem has the following form: given a pair of unknown matrices Θ^* and Γ^*, both lying in $\mathbb{R}^{d_1 \times d_2}$, suppose that we observe a third matrix specified by the model $Y = \Theta^* + \Gamma^* + W$, where $W \in \mathbb{R}^{d_1 \times d_2}$ represents observation noise. Typically the matrix Θ^* is assumed to be low-rank, and some low-dimensional structural constraint is assumed on the matrix Γ^*. For example, the papers [49, 45, 78] consider the setting in which Γ^* is sparse, while Xu et al. [174] consider a column-sparse model, in which only a few of the columns of Γ^* have non-zero entries. In order to illustrate the application of our general result to this setting, here we consider the low-rank plus column-sparse

framework [174]. (We note that since the ℓ_1-norm is decomposable, similar results can easily be derived for the low-rank plus entrywise-sparse setting as well.)

Since Θ^* is assumed to be low-rank, as before we use the nuclear norm $|||\Theta|||_1$ as a regularizer (see Section 5.2.4). We assume that the unknown matrix $\Gamma^* \in \mathbb{R}^{d_1 \times d_2}$ is column-sparse, say with at most $s < d_2$ non-zero columns. A suitable convex regularizer for this matrix structure is based on the *columnwise* $(1,2)$-*norm*, given by

$$\|\Gamma\|_{1,2} := \sum_{j=1}^{d_2} \|\Gamma_j\|_2, \tag{5.45}$$

where $\Gamma_j \in \mathbb{R}^{d_1}$ denotes the j^{th} column of Γ. Note also that the dual norm is given by the *elementwise* $(\infty,2)$-*norm* $\|\Gamma\|_{\infty,2} = \max_{j=1,\ldots,d_2} \|\Gamma_j\|_2$, corresponding to the maximum ℓ_2-norm over columns.

In order to estimate the unknown pair (Θ^*, Γ^*), we consider the M-estimator

$$(\widehat{\Theta}, \widehat{\Gamma}) := \arg\min_{\Theta,\Gamma} |||Y - \Theta - \Gamma|||_F^2 \quad \text{such that} \quad |||\Theta|||_1 \leq \rho_\Theta, \quad \|\Gamma\|_{1,2} \leq \rho_\Gamma \text{ and } \|\Theta\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}} \tag{5.46}$$

The first two constraints restrict Θ and Γ to a nuclear norm ball of radius ρ_Θ and a $(1,2)$-norm ball of radius ρ_Γ, respectively. The final constraint controls the "spikiness" of the low-rank component Θ, as measured in the $(\infty,2)$-norm, corresponding to the maximum ℓ_2-norm over the columns. As with the elementwise ℓ_∞-bound for matrix completion, this additional constraint is required in order to limit the non-identifiability in matrix decomposition. (See the paper [8] for more discussion of non-identifiability issues in matrix decomposition.)

With this set-up, consider the projected gradient algorithm when applied to the matrix decomposition problem: it generates a sequence of matrix pairs (Θ^t, Γ^t) for $t = 0, 1, 2, \ldots$, and the optimization error is characterized in terms of the matrices $\widehat{\Delta}^t_\Theta := \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}^t_\Gamma := \Gamma^t - \widehat{\Gamma}$. Finally, we measure the optimization error at time t in terms of the squared Frobenius error $e^2(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma) := |||\widehat{\Delta}^t_\Theta|||_F^2 + |||\widehat{\Delta}^t_\Gamma|||_F^2$, summed across both the low-rank and column-sparse components.

Corollary 5.6 (Matrix decomposition)**.** *Under the conditions of Theorem 5.1, suppose that* $\|\Theta^*\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}}$ *and* Γ^* *has at most* s *non-zero columns. If we solve the convex program (5.46) with* $\rho_\Theta \leq |||\Theta^*|||_1$ *and* $\rho_\Gamma \leq \|\Gamma^*\|_{1,2}$*, then for all iterations* $t = 0, 1, 2, \ldots$*,*

$$e^2(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma) \leq \left(\frac{3}{4}\right)^t e^2(\widehat{\Delta}^0_\Theta, \widehat{\Delta}^0_\Gamma) + c\left(|||\widehat{\Gamma} - \Gamma^*|||_F^2 + \alpha^2 \frac{s}{d_2}\right).$$

This corollary has some unusual aspects, relative to the previous corollaries. First of all, in contrast to the previous results, the guarantee is a deterministic one (as opposed to holding with high probability). More specifically, the RSC/RSM conditions hold deterministic sense, which should be contrasted with the high probability statements given in Corollaries 5.2-5.5. Consequently, the effective conditioning of the problem does not depend on sample size and we are guaranteed geometric convergence at a fixed rate, independent of sample size. The additional tolerance term is completely independent of the rank of Θ^* and only depends on the column-sparsity of Γ^*.

5.4 Simulation results

In this section, we provide some experimental results that confirm the accuracy of our theoretical results, in particular showing excellent agreement with the linear rates predicted by our theory. In addition, the rates of convergence slow down for smaller sample sizes, which lead to problems with relatively poor conditioning. In all the simulations reported below, we plot the log error $\|\theta^t - \widehat{\theta}\|$ between the iterate θ^t at time t versus the final solution $\widehat{\theta}$. Each curve provides the results averaged over five random trials, according to the ensembles which we now describe.

5.4.1 Sparse regression

We begin by considering the linear regression model $y = X\theta^* + w$ where θ^* is the unknown regression vector belonging to the set $\mathbb{B}_q(R_q)$, and i.i.d. observation noise $w_i \sim N(0, 0.25)$. We consider a family of ensembles for the random design matrix $X \in \mathbb{R}^{n \times d}$. In particular, we construct X by generating each row $x_i \in \mathbb{R}^d$ independently according to following procedure. Let $z_1, \ldots, z_n$ be an i.i.d. sequence of $N(0,1)$ variables, and fix some correlation parameter $\omega \in [0, 1)$. We first initialize by setting $x_{i,1} = z_1/\sqrt{1-\omega^2}$, and then generate the remaining entries by applying the recursive update $x_{i,t+1} = \omega x_{i,t} + z_t$ for $t = 1, 2, \ldots, d-1$, so that $x_i \in \mathbb{R}^d$ is a zero-mean Gaussian random vector. It can be verified that all the eigenvalues of $\Sigma = \mathrm{cov}(x_i)$ lie within the interval $[\frac{1}{(1+\omega)^2}, \frac{2}{(1-\omega)^2(1+\omega)}]$, so that Σ has a a finite condition number for all $\omega \in [0, 1)$. At one extreme, for $\omega = 0$, the matrix Σ is the identity, and so has condition number equal to 1. As $\omega \to 1$, the matrix Σ becomes progressively more ill-conditioned, with a condition number that is very large for ω close to one. As a consequence, although incoherence conditions like the restricted isometry property can be satisfied when $\omega = 0$, they will fail to be satisfied (w.h.p.) once ω is large enough.

For this random ensemble of problems, we have investigated convergence rates for a wide range of dimensions d and radii R_q. Since the results are relatively uniform across the choice of these parameters, here we report results for dimension $d = 20,000$, and radius $R_q = \lceil (\log d)^2 \rceil$. In the case $q = 0$, the radius $R_0 = s$ corresponds to the sparsity level. The per iteration cost in this case is $\mathcal{O}(nd)$. In order to reveal dependence of convergence rates

on sample size, we study a range of the form $n = \lceil \alpha \; s \log d \rceil$, where the *order parameter* $\alpha > 0$ is varied.

Our first experiment is based on taking the correlation parameter $\omega = 0$, and the ℓ_q-ball parameter $q = 0$, corresponding to exact sparsity. We then measure convergence rates for sample sizes specified by $\alpha \in \{1, 1.25, 5, 25\}$. As shown by the results plotted in panel (a) of Figure 5.3, projected gradient descent fails to converge for $\alpha = 1$ or $\alpha = 1.25$; in both these cases, the sample size n is too small for the RSC and RSM conditions to hold, so that a constant step size leads to oscillatory behavior in the algorithm. In contrast, once the order parameter α becomes large enough to ensure that the RSC/RSM conditions hold (w.h.p.), we observe a geometric convergence of the error $\left\|\theta^t - \widehat{\theta}\right\|_2$. Moreover the convergence rate is faster for $\alpha = 25$ compared to $\alpha = 5$, since the RSC/RSM constants are better with larger sample size. Such behavior is in agreement with the conclusions of Corollary 5.2, which predicts that the the convergence rate should improve as the number of samples n is increased.

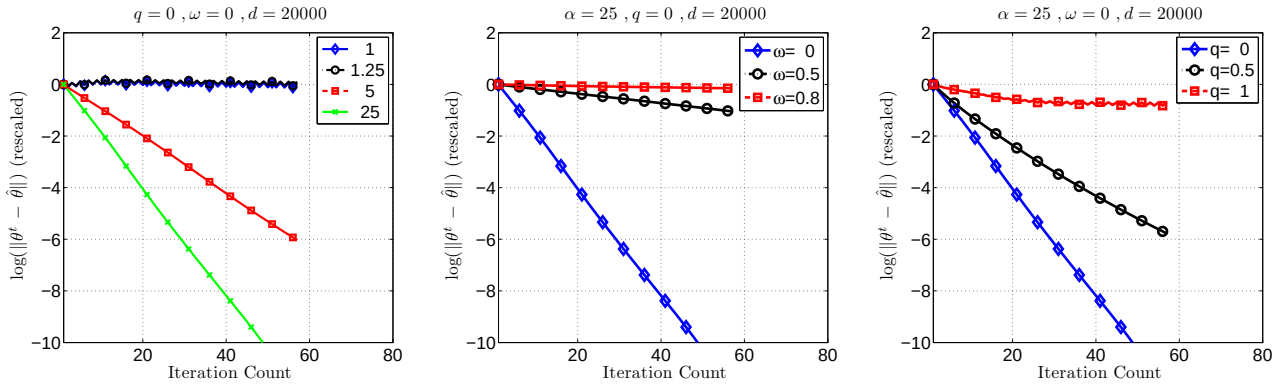

Figure 5.3. Plot of the log of the optimization error $\log\left(\left\|\theta^t - \widehat{\theta}\right\|_2\right)$ in the sparse linear regression problem, rescaled so the plots start at 0. In this problem, $d = 20000$, $s = \lceil \log d \rceil$, $n = \alpha s \log d$. Plot (a) shows convergence for the exact sparse case with $q = 0$ and $\Sigma = I$ (i.e. $\omega = 0$). In panel (b), we observe how convergence rates change as the correlation parameter ω is varied for $q = 0$ and $\alpha = 25$. Plot (c) shows the convergence rates when $\omega = 0$, $\alpha = 25$ and q is varied.

On the other hand, Corollary 5.2 also predicts that convergence rates should be slower when the condition number of Σ is worse. In order to test this prediction, we again studied an exactly sparse problem ($q = 0$), this time with the fixed sample size $n = \lceil 25 s \log d \rceil$, and we varied the correlation parameter $\omega \in \{0, 0.5, 0.8\}$. As shown in panel (b) of Figure 5.3, the convergence rates slow down as the correlation parameter is increased and for the case of extremely high correlation of $\omega = 0.8$, the optimization error curve is almost flat—the method makes very slow progress in this case.

A third prediction of Corollary 5.2 is that the convergence of projected gradient descent should become slower as the sparsity parameter q is varied between exact sparsity ($q = 0$), and the least sparse case ($q = 1$). (In particular, note for $n > \log d$, the quantity χ_n from equation (5.35) is monotonically increasing with q.) Panel (c) of Figure 5.3 shows convergence rates for the fixed sample size $n = 25s \log d$ and correlation parameter $\omega = 0$, and with the sparsity parameter $q \in \{0, 0.5, 1.0\}$. As expected, the convergence rate slows down as q increases from 0 to 1. Corollary 5.2 further captures how the contraction factor changes as the problem parameters (s, d, n) are varied. In particular, it predicts that as we change the triplet simultaneously, while holding the ratio $\alpha = s \log d/n$ constant, the convergence rate should stay the same. We recall that this phenomenon was indeed demonstrated in Figure 5.1 in Section 5.1.

5.4.2 Low-rank matrix estimation

We also performed experiments with two different versions of low-rank matrix regression. Our simulations applied to instances of the observation model $y_i = \langle\langle X_i, \Theta^* \rangle\rangle + w_i$, for $i = 1, 2, \ldots, n$, where $\Theta^* \in \mathbb{R}^{200 \times 200}$ is a fixed unknown matrix, $X_i \in \mathbb{R}^{200 \times 200}$ is a matrix of covariates, and $w_i \sim N(0, 0.25)$ is observation noise. In analogy to the sparse vector problem, we performed simulations with the matrix Θ^* belonging to the set $\mathbb{B}_q(R_q)$ of approximately low-rank matrices, as previously defined in equation (5.40) for $q \in [0, 1]$. The case $q = 0$ corresponds to the set of matrices with rank at most $r = R_0$, whereas the case $q = 1$ corresponds to the ball of matrices with nuclear norm at most R_1.

In our first set of matrix experiments, we considered the matrix version of compressed sensing [135], in which each matrix $X_i \in \mathbb{R}^{200 \times 200}$ is randomly formed with i.i.d. $N(0, 1)$ entries, as described in Section 5.3.3. In the case $q = 0$, we formed a matrix $\Theta^* \in \mathbb{R}^{200 \times 200}$ with rank $R_0 = 5$, and performed simulations over the sample sizes $n = \alpha R_0 \, d$, with the parameter $\alpha \in \{1, 1.25, 5, 25\}$. The per iteration cost in this case is $\mathcal{O}(nd^2)$. As seen in panel (a) of Figure 5.4, the projected gradient descent method exhibits behavior that is qualitatively similar to that for the sparse linear regression problem. More specifically, it fails to converge when the sample size (as reflected by the order parameter α) is too small, and converges geometrically with a progressively faster rate as α is increased. We have also observed similar types of scaling as the matrix sparsity parameter is increased from $q = 0$ to $q = 1$.

In our second set of matrix experiments, we studied the behavior of projected gradient descent for the problem of matrix completion, as described in Section 5.3.3. For this problem, we again studied matrices of dimension $d = 200$ and rank $R_0 = 5$, and we varied the sample size as $n = \alpha \, R_0 \, d \log d$ for $\alpha \in \{1, 2, 5, 25\}$. As shown in panel (b) of Figure 5.4, projected gradient descent for matrix completion also enjoys geometric convergence for α large enough.

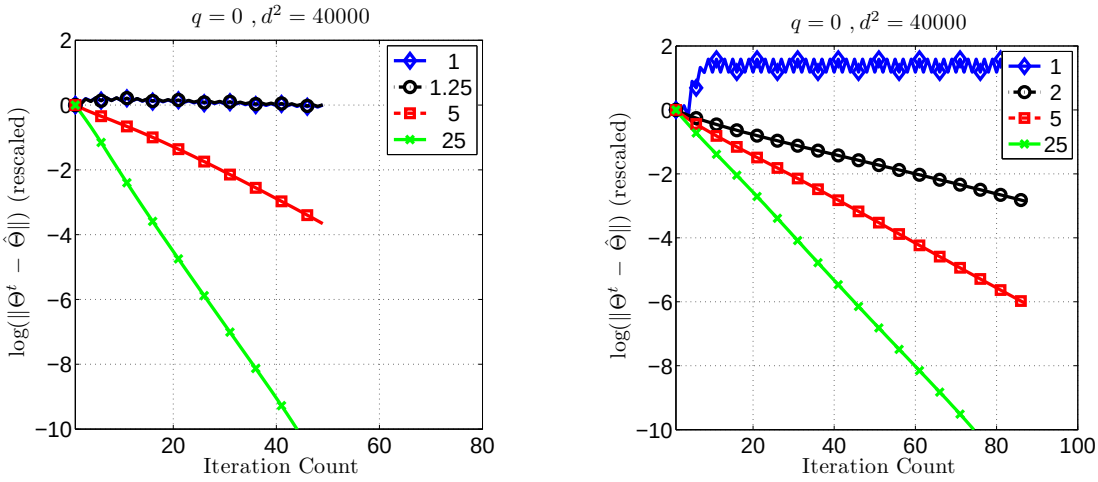

Figure 5.4. (a) Plot of log Frobenius error $\log(|||\Theta^t - \widehat{\Theta}|||_F)$ versus number of iterations in matrix compressed sensing for a matrix size $d = 200$ with rank $R_0 = 5$, and sample sizes $n = \alpha R_0 d$. For $\alpha \in \{1, 1.25\}$, the algorithm oscillates, whereas geometric convergence is obtained for $\alpha \in \{5, 25\}$, consistent with the theoretical prediction. (b) Plot of log Frobenius error $\log(|||\Theta^t - \widehat{\Theta}|||_F)$ versus number of iterations in matrix completion with $d = 200$, $R_0 = 5$, and $n = \alpha R_o d \log(d)$ with $\alpha \in \{1, 2, 5, 25\}$. For $\alpha \in \{2, 5, 25\}$ the algorithm enjoys geometric convergence.

5.5 Proofs

In this section, we provide the proofs of our results. Recall that we use $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ to denote the optimization error, and $\Delta^* = \widehat{\theta} - \theta^*$ to denote the statistical error. For future reference, we point out a slight weakening of restricted strong convexity (RSC), useful for obtaining parts of our results. As the to follow reveals, it is only necessary to enforce an RSC condition of the form

$$\mathcal{T}_{\mathcal{L}}(\theta^t; \widehat{\theta}) \geq \frac{\gamma_\ell}{2} \|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2, \tag{5.47}$$

which is milder than the original RSC condition (5.8), in that it applies only to differences of the form $\theta^t - \widehat{\theta}$, and allows for additional slack δ. We make use of this refined notion in the proofs of various results to follow.

With this relaxed RSC condition and the same RSM condition as before, our proof shows that

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + 2\delta^2/\gamma_u}{1 - \kappa} \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{5.48}$$

Note that this result reduces to the previous statement when $\delta = 0$. This extension of Theorem 5.1 is used in the proofs of Corollaries 5.5 and 5.6.

We will assume without loss of generality that all the iterates lie in the subset Ω' of Ω. This can be ensured by augmenting the loss with the indicator of Ω' or equivalently performing projections on the set $\Omega' \cap \mathbb{B}_{\mathcal{R}}(\rho)$ as mentioned earlier.

5.5.1 Proof of Theorem 5.1

Recall that Theorem 5.1 concerns the constrained problem (5.1). The proof is based on two technical lemmas. The first lemma guarantees that at each iteration $t = 0, 1, 2, \ldots$, the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ belongs to an interesting constraint set defined by the regularizer.

Lemma 5.1. *Let $\widehat{\theta}$ be any optimum of the constrained problem (5.1) for which $\mathcal{R}(\widehat{\theta}) = \rho$. Then for any iteration $t = 1, 2, \ldots$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies the cone bound*

$$\mathcal{R}(\Delta) \leq 2\,\Psi(\overline{\mathcal{M}})\,\|\Delta\| + 2\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) + 2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}})\,\|\Delta^*\|\,. \tag{5.49}$$

The proof of this lemma, provided in Appendix C.1.1, exploits the decomposability of the regularizer in an essential way.

The cone bound (5.49) takes a simpler form in the special case when $\mathcal{M}$ is chosen to contain θ^* and $\overline{\mathcal{M}} = \mathcal{M}$. In this case, we have $\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) = 0$, and hence the optimization error $\widehat{\Delta}^t$ satisfies the inequality

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\mathcal{M})\,\Big\{\,\Big\|\widehat{\Delta}^t\Big\| + \|\Delta^*\|\,\Big\} + 2\mathcal{R}(\Delta^*). \tag{5.50}$$

An inequality of this type, when combined with the definitions of RSC/RSM, allows us to establish the curvature conditions required to prove globally geometric rates of convergence.

We now state a second lemma under the more general RSC condition (5.47):

Lemma 5.2. *Under the RSC condition (5.47) and RSM condition (5.10), for all $t = 0, 1, 2, \ldots$, we have*

$$\begin{aligned}
&\gamma_u\,\langle \theta^t - \theta^{t+1},\, \theta^t - \widehat{\theta}\rangle \\
&\quad\geq \Big\{\frac{\gamma_u}{2}\|\theta^t - \theta^{t+1}\|^2 - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t)\Big\} + \Big\{\frac{\gamma_\ell}{2}\|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2\Big\}.
\end{aligned} \tag{5.51}$$

The proof of this lemma, provided in Appendix C.1.2, follows along the lines of the intermediate result within Theorem 2.2.8 of Nesterov [120], but with some care required to handle the additional terms that arise in our weakened forms of strong convexity and smoothness.

Using these auxiliary results, let us now complete the the proof of Theorem 5.1. We first note the elementary relation

$$\|\theta^{t+1} - \widehat{\theta}\|^2 = \|\theta^t - \widehat{\theta} - \theta^t + \theta^{t+1}\|^2 = \|\theta^t - \widehat{\theta}\|^2 + \|\theta^t - \theta^{t+1}\|^2 - 2\langle \theta^t - \widehat{\theta}, \theta^t - \theta^{t+1}\rangle. \tag{5.52}$$

We now use Lemma 5.2 and the more general form of RSC (5.47) to control the cross-term, thereby obtaining the upper bound

$$\begin{aligned}\|\theta^{t+1} - \widehat{\theta}\|^2 &\leq \|\theta^t - \widehat{\theta}\|^2 - \frac{\gamma_\ell}{\gamma_u}\|\theta^t - \widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^{t+1} - \theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^t - \widehat{\theta}) + \frac{2\delta^2}{\gamma_u} \\ &= \big(1 - \frac{\gamma_\ell}{\gamma_u}\big)\|\theta^t - \widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^{t+1} - \theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^t - \widehat{\theta}) + \frac{2\delta^2}{\gamma_u}.\end{aligned}$$

We now observe that by triangle inequality and the Cauchy-Schwarz inequality,

$$\mathcal{R}^2(\theta^{t+1} - \theta^t) \leq \big(\mathcal{R}(\theta^{t+1} - \widehat{\theta}) + \mathcal{R}(\widehat{\theta} - \theta^t)\big)^2 \leq 2\mathcal{R}^2(\theta^{t+1} - \widehat{\theta}) + 2\mathcal{R}^2(\theta^t - \widehat{\theta}).$$

Recall the definition of the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, we have the upper bound

$$\|\widehat{\Delta}^{t+1}\|^2 \leq \big(1 - \frac{\gamma_\ell}{\gamma_u}\big)\|\widehat{\Delta}^t\|^2 + \frac{4\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \frac{4\tau_u(\mathcal{L}_n) + 2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\widehat{\Delta}^t) + \frac{2\delta^2}{\gamma_u}. \tag{5.53}$$

We now apply Lemma 5.1 to control the terms involving $\mathcal{R}^2$. In terms of squared quantities, the inequality (5.49) implies that

$$\mathcal{R}^2(\widehat{\Delta}^t) \leq 4\,\Psi^2(\overline{\mathcal{M}}^\perp)\,\Big\|\widehat{\Delta}^t\Big\|^2 + 2\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \qquad \text{for all } t = 0, 1, 2, \ldots,$$

where we recall that $\Psi^2(\overline{\mathcal{M}}^\perp)$ is the subspace compatibility (5.12) and $\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ accumulates all the residual terms. Applying this bound twice—once for t and once for $t+1$—and substituting into equation (5.53) yields that $\big\{1 - \frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\tau_u(\mathcal{L}_n)}{\gamma_u}\big\}\|\Delta^{t+1}\|^2$ is upper bounded by

$$\Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\}\|\Delta^t\|^2 + \frac{16\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{\gamma_u} + \frac{2\delta^2}{\gamma_u}.$$

Under the assumptions of Theorem 5.1, we are guaranteed that $\frac{16\Psi^2(\overline{\mathcal{M}^\perp})\tau_u(\mathcal{L}_n)}{\gamma_u} < 1/2$, and so we can re-arrange this inequality into the form

$$\|\Delta^{t+1}\|^2 \leq \kappa \|\Delta^t\|^2 + \epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u} \tag{5.54}$$

where κ and $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ were previously defined in equations (5.22) and (5.23) respectively. Iterating this recursion yields

$$\|\Delta^{t+1}\|^2 \leq \kappa^t \|\Delta^0\|^2 + \left(\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u}\right)\Big(\sum_{j=0}^{t} \kappa^j\Big).$$

The assumptions of Theorem 5.1 guarantee that $\kappa \in (0, 1)$, so that summing the geometric series yields the claim (5.24).

5.5.2 Proof of Theorem 5.2

The Lagrangian version of the optimization program is based on solving the convex program (5.2), with the objective function $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n \mathcal{R}(\theta)$. Our proof is based on analyzing the error $\phi(\theta^t) - \phi(\widehat{\theta})$ as measured in terms of this objective function. It requires two technical lemmas, both of which are stated in terms of a given tolerance $\bar{\eta} > 0$, and an integer $T > 0$ such that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \bar{\eta} \qquad \text{for all } t \geq T. \tag{5.55}$$

Our first technical lemma is analogous to Lemma 5.1, and restricts the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ to a cone-like set.

Lemma 5.3 (Iterated Cone Bound (ICB)). *Let $\widehat{\theta}$ be any optimum of the regularized M-estimator (5.2). Under condition (5.55) with parameters $(T, \bar{\eta})$, for any iteration $t \geq T$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies*

$$\mathcal{R}(\widehat{\Delta}^t) \leq 4\Psi(\overline{\mathcal{M}}) \left\|\widehat{\Delta}^t\right\| + 8\Psi(\overline{\mathcal{M}}) \|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\left(\frac{\bar{\eta}}{\lambda_n}, \overline{\rho}\right) \tag{5.56}$$

Our next lemma guarantees sufficient decrease of the objective value difference $\phi(\theta^t) - \phi(\widehat{\theta})$. Lemma 5.3 plays a crucial role in its proof. Recall the definition (5.27) of the compound contraction coefficient $\kappa(\mathcal{L}_n; \overline{\mathcal{M}})$, defined in terms of the related quantities $\xi(\overline{\mathcal{M}})$ and $\beta(\overline{\mathcal{M}})$. Throughout the proof, we drop the arguments of κ, ξ and β so as to ease notation.

Lemma 5.4. *Under the RSC (5.47) and RSM conditions (5.10), as well as assumption (5.55) with parameters $(\bar{\eta}, T)$, for all $t \geq T$, we have*

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T}(\phi(\theta^T) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi(\mathcal{M})\, \beta(\mathcal{M})(\varepsilon^2 + \bar{\epsilon}_{stat}^2),$$

where $\varepsilon := 2\min(\bar{\eta}/\lambda_n, \overline{\rho})$ and $\bar{\epsilon}_{stat} := 8\Psi(\overline{\mathcal{M}}) \|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$.

We are now in a position to prove our main theorem, in particular via a recursive application of Lemma 5.4. At a high level, we divide the iterations $t = 0, 1, 2, \ldots$ into a series of disjoint epochs $[T_k, T_{k+1})$ with $0 = T_0 \leq T_1 \leq T_2 \leq \cdots$. Moreover, we define an associated sequence of tolerances $\bar{\eta}_0 > \bar{\eta}_1 > \cdots$ such that at the end of epoch $[T_{k-1}, T_k)$, the optimization error has been reduced to $\bar{\eta}_k$. Our analysis guarantees that $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \bar{\eta}_k$ for all $t \geq T_k$, allowing us to apply Lemma 5.4 with smaller and smaller values of $\bar{\eta}$ until it reduces to the statistical error $\bar{\epsilon}_{\text{stat}}$.

At the first iteration, we have no a priori bound on the error $\bar{\eta}_0 = \phi(\theta^0) - \phi(\widehat{\theta})$. However, since Lemma 5.4 involves the quantity $\varepsilon = \min(\bar{\eta}/\lambda_n, \overline{\rho})$, we may still apply it[6] at the first epoch with $\varepsilon_0 = \overline{\rho}$ and $T_0 = 0$. In this way, we conclude that for all $t \geq 0$,

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^t (\phi(\theta^0) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa} \xi\beta (\overline{\rho}^2 + \bar{\epsilon}^2_{\text{stat}}).$$

Now since the contraction coefficient $\kappa \in (0, 1)$, for all iterations $t \geq T_1 := (\lceil \log(2\,\bar{\eta}_0/\bar{\eta}_1)/\log(1/\kappa) \rceil)_+$, we are guaranteed that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \underbrace{\frac{4\,\xi\beta}{1-\kappa}(\overline{\rho}^2 + \bar{\epsilon}^2_{\text{stat}})}_{\bar{\eta}_1} \leq \frac{8\xi\beta}{1-\kappa} \max(\overline{\rho}^2, \bar{\epsilon}^2_{\text{stat}}).$$

This same argument can now be applied in a recursive manner. Suppose that for some $k \geq 1$, we are given a pair $(\bar{\eta}_k, T_k)$ such that condition (5.55) holds. An application of Lemma 5.4 yields the bound

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T_k} (\phi(\theta^{T_k}) - \phi(\widehat{\theta})) + \frac{2\,\xi\beta}{1-\kappa}(\varepsilon_k^2 + \bar{\epsilon}^2_{\text{stat}}) \qquad \text{for all } t \geq T_k.$$

We now define $\bar{\eta}_{k+1} := \frac{4\,\xi\beta}{1-\kappa}(\varepsilon_k^2 + \bar{\epsilon}^2_{\text{stat}})$. Once again, since $\kappa < 1$ by assumption, we can choose $T_{k+1} := \lceil \log(2\bar{\eta}_k/\bar{\eta}_{k+1})/\log(1/\kappa) \rceil + T_k$, thereby ensuring that for all $t \geq T_{k+1}$, we have

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \frac{8\xi\beta}{1-\kappa} \max(\varepsilon_k^2, \bar{\epsilon}^2_{\text{stat}}).$$

In this way, we arrive at recursive inequalities involving the tolerances $\{\bar{\eta}_k\}_{k=0}^{\infty}$ and time steps $\{T_k\}_{k=0}^{\infty}$—namely

$$\bar{\eta}_{k+1} \leq \frac{8\,\xi\beta}{1-\kappa} \max(\varepsilon_k^2, \bar{\epsilon}^2_{\text{stat}}), \qquad \text{where } \varepsilon_k = 2\,\min\{\bar{\eta}_k/\lambda_n, \overline{\rho}\}, \text{ and} \tag{5.57a}$$

$$T_k \leq k + \frac{\log(2^k \bar{\eta}_0/\bar{\eta}_k)}{\log(1/\kappa)}. \tag{5.57b}$$

[6]It is for precisely this reason that our regularized M-estimator includes the additional side-constraint defined in terms of $\overline{\rho}$.

Now we claim that the recursion (5.57a) can be unwrapped so as to show that

$$\bar{\eta}_{k+1} \leq \frac{\bar{\eta}_k}{4^{2^{k-1}}} \quad \text{and} \quad \frac{\bar{\eta}_{k+1}}{\lambda_n} \leq \frac{\overline{\rho}}{4^{2^k}} \qquad \text{for all } k = 1, 2, \ldots. \tag{5.58}$$

Taking these statements as given for the moment, let us now show how they can be used to upper bound the smallest k such that $\bar{\eta}_k \leq \delta^2$. If we are in the first epoch, the claim of the theorem is straightforward from equation (5.57a). If not, we first use the recursion (5.58) to upper bound the number of epochs needed and then use the inequality (5.57b) to obtain the stated result on the total number of iterations needed. Using the second inequality in the recursion (5.58), we see that it is sufficient to ensure that $\frac{\overline{\rho}\lambda_n}{4^{2^{k-1}}} \leq \delta^2$. Rearranging this inequality, we find that the error drops below δ^2 after at most

$$k_\delta \geq \log\left(\log\left(\frac{\overline{\rho}\lambda_n}{\delta^2}\right)/\log(4)\right)/\log(2) + 1 = \log_2\log_2\left(\frac{\overline{\rho}\lambda_n}{\delta^2}\right)$$

epochs. Combining the above bound on k_δ with the recursion 5.57b, we conclude that the inequality $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \delta^2$ is guaranteed to hold for all iterations

$$t \geq k_\delta\left(1 + \frac{\log 2}{\log(1/\kappa)}\right) + \frac{\log\frac{\bar{\eta}_0}{\delta^2}}{\log(1/\kappa)},$$

which is the desired result.

It remains to prove the recursion (5.58), which we do via induction on the index k. We begin with base case $k = 1$. Recalling the setting of $\bar{\eta}_1$ and our assumption on λ_n in the theorem statement (5.30), we are guaranteed that $\bar{\eta}_1/\lambda_n \leq \overline{\rho}/4$, so that $\varepsilon_1 \leq \varepsilon_0 = \overline{\rho}$. By applying equation (5.57a) with $\varepsilon_1 = 2\bar{\eta}_1/\lambda_n$ and assuming $\varepsilon_1 \geq \bar{\epsilon}_{\text{stat}}$, we obtain

$$\bar{\eta}_2 \leq \frac{32\xi\beta\bar{\eta}_1^2}{(1-\kappa)\lambda_n^2} \overset{(i)}{\leq} \frac{32\xi\beta\overline{\rho}\bar{\eta}_1}{(1-\kappa)4\lambda_n} \overset{(ii)}{\leq} \frac{\bar{\eta}_1}{4}, \tag{5.59}$$

where step (i) uses the fact that $\frac{\bar{\eta}_1}{\lambda_n} \leq \frac{\overline{\rho}}{4}$, and step (ii) uses the condition (5.30) on λ_n. We have thus verified the first inequality (5.58) for $k = 1$. Turning to the second inequality in the statement (5.58), using equation 5.59, we have

$$\frac{\bar{\eta}_2}{\lambda_n} \leq \frac{\bar{\eta}_1}{4\lambda_n} \overset{(iii)}{\leq} \frac{\overline{\rho}}{16},$$

where step (iii) follows from the assumption (5.30) on λ_n. Turning to the inductive step, we again assume that $2\bar{\eta}_k/\lambda_n \geq \bar{\epsilon}_{\text{stat}}$ and obtain from inequality (5.57a)

$$\bar{\eta}_{k+1} \leq \frac{32\xi\beta\bar{\eta}_k^2}{(1-\kappa)\lambda_n^2} \overset{(iv)}{\leq} \frac{32\xi\beta\bar{\eta}_k\overline{\rho}}{(1-\kappa)\lambda_n 4^{2^{k-1}}} \overset{(v)}{\leq} \frac{\bar{\eta}_k}{4^{2^{k-1}}}.$$

Here step (iv) uses the second inequality of the inductive hypothesis (5.58) and step (v) is a consequence of the condition on λ_n as before. The second part of the induction is similarly established, completing the proof.

5.5.3 Proof of Corollary 5.1

In order to prove this claim, we must show that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as defined in equation (5.23), is of order lower than $\mathbb{E}[\left\|\widehat{\theta} - \theta^*\right\|^2] = \mathbb{E}[\|\Delta^*\|^2]$. We make use of the following lemma, proved in Appendix C.3:

Lemma 5.5. *If $\rho \leq \mathcal{R}(\theta^*)$, then for any solution $\widehat{\theta}$ of the constrained problem (5.1) and any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, the statistical error $\Delta^* = \widehat{\theta} - \theta^*$ satisfies the inequality*

$$\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^\perp)\,\|\Delta^*\| + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)). \tag{5.60}$$

Using this lemma, we can complete the proof of Corollary 5.1. Recalling the form (5.23), under the condition $\theta^* \in \mathcal{M}$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\,\big(2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}}^\perp)\,\|\Delta^*\|\,\big)^2}{\gamma_u}.$$

Using the assumption $\frac{(\tau_u(\mathcal{L}_n)+\tau_\ell(\mathcal{L}_n))\Psi^2(\overline{\mathcal{M}}^\perp)}{\gamma_u} = o(1)$, it suffices to show that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^\perp)\,\|\Delta^*\|$. Since Corollary 5.1 assumes that $\theta^* \in \mathcal{M}$ and hence that $\Pi_{\mathcal{M}^\perp}(\theta^*) = 0$, Lemma 5.5 implies that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^\perp)\,\|\Delta^*\|$, as required.

5.5.4 Proofs of Corollaries 5.2 and 5.3

The central challenge in proving this result is verifying that suitable forms of the RSC and RSM conditions hold with sufficiently small parameters $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$.

Lemma 5.6. *Define the maximum variance $\zeta(\Sigma) := \max_{j=1,2,\ldots,d} \Sigma_{jj}$. Under the conditions of Corollary 5.2, there are universal positive constants (c_0, c_1) such that for all $\Delta \in \mathbb{R}^d$, we have*

$$\frac{\|X\Delta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\Delta\|_2^2 - c_1\zeta(\Sigma)\frac{\log d}{n}\,\|\Delta\|_1^2, \qquad \textit{and} \tag{5.61a}$$

$$\frac{\|X\Delta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\Delta\|_2^2 + c_1\zeta(\Sigma)\frac{\log d}{n}\,\|\Delta\|_1^2, \tag{5.61b}$$

with probability at least $1 - \exp(-c_0\, n)$.

Note that this lemma implies that the RSC and RSM conditions both hold with high probability, in particular with parameters

$$\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma), \text{ and } \quad \tau_\ell(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n}, \qquad \text{for RSC, and}$$

$$\gamma_u = 2\sigma_{\max}(\Sigma) \text{ and } \quad \tau_u(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n} \qquad \text{for RSM.}$$

This lemma has been proved by Raskutti et al. [133] for obtaining minimax rates in sparse linear regression.

Let us first prove Corollary 5.2 in the special case of hard sparsity ($q = 0$), in which θ^* is supported on a subset S of cardinality s. Let us define the model subspace

$$\mathcal{M} := \big\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\big\},$$

so that $\theta^* \in \mathcal{M}$. Recall from Section 5.2.4 that the ℓ_1-norm is decomposable with respect to $\mathcal{M}$ and $\mathcal{M}^\perp$; as a consequence, we may also set $\overline{\mathcal{M}}^\perp = \mathcal{M}$ in the definitions (5.22) and (5.23). By definition (5.12) of the subspace compatibility between with ℓ_1-norm as the regularizer, and ℓ_2-norm as the error norm, we have $\Psi^2(\mathcal{M}) = s$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 5.6 and substituting into equation (5.22), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\}\Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{5.62}$$

where $\chi_n(\Sigma) := \frac{c_2\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\frac{s\log d}{n}$ for some universal constant c_2. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \le c_3\, \phi(\Sigma; s, d, n)\Big\{\frac{\|\Delta^*\|_1^2}{s} + \|\Delta^*\|_2^2\Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \le \|\theta^*\|_1$, then Lemma 5.5 (as exploited in the proof of Corollary 5.1) shows that $\|\Delta^*\|_1^2 \le 4s\|\Delta^*\|_2^2$, and hence that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \le c_3\ \chi_n(\Sigma)\ \|\Delta^*\|_2^2$. This completes the proof of the claim (5.36) for $q = 0$.

We now turn to the case $q \in (0, 1]$, for which we bound the term $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. For a truncation level $\mu > 0$ to be chosen, define the set $S_\mu := \big\{j \in \{1, 2, \ldots, d\} \mid |\theta_j^*| > \mu\big\}$, and define the associated subspaces $\mathcal{M} = \mathcal{M}(S_\mu)$ and $\overline{\mathcal{M}}^\perp = \mathcal{M}^\perp(S_\mu)$. By combining Lemma 5.5 and the definition (5.23) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \mathcal{M}^\perp(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \le \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\frac{\log d}{n}\big(\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 + \sqrt{|S_\mu|}\,\|\Delta^*\|_2\big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^\perp$ on S_μ. We now choose the threshold μ optimally, so as to trade-off the term $\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1$, which decreases as μ increases, with the term $\sqrt{S_\mu}\|\Delta^*\|_2$, which increases as μ increases.

By definition of $\mathcal{M}^\perp(S_\mu)$, we have

$$\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 = \sum_{j \notin S_\mu} |\theta_j^*| \;=\; \mu \sum_{j \notin S_\mu} \frac{|\theta_j^*|}{\mu} \;\le\; \mu \sum_{j \notin S_\mu} \Big(\frac{|\theta_j^*|}{\mu}\Big)^q,$$

where the inequality holds since $|\theta_j^*| \leq \mu$ for all $j \notin S_\mu$. Now since $\theta^* \in \mathbb{B}_q(R_q)$, we conclude that

$$\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 \leq \mu^{1-q} \sum_{j \notin S_\mu} |\theta_j^*|^q \leq \mu^{1-q} R_q. \tag{5.63}$$

On the other hand, again using the inclusion $\theta^* \in \mathbb{B}_q(R_q)$, we have $R_q \geq \sum_{j \in S_\mu} |\theta_j^*|^q \geq |S_\mu| \, \mu^q$ which implies that $|S_\mu| \leq \mu^{-q} R_q$. By combining this bound with inequality (5.63), we obtain the upper bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q \|\Delta^*\|_2^2\big) = \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \mu^{-q} R_q \big(\mu^{2-q} R_q + \|\Delta^*\|_2^2\big).$$

Setting $\mu^2 = \frac{\log d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \varphi_n(\Sigma) \left\{ R_q \big(\frac{\log d}{n}\big)^{1-q/2} + \|\Delta^*\|_2^2 \right\}, \qquad \text{where } \varphi_n(\Sigma) := \tfrac{c\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} R_q \big(\tfrac{\log d}{n}\big)^{1-q/2}.$$

Finally, let us verify the stated form of the contraction coefficient. For the given subspace $\overline{\mathcal{M}}^\perp = \mathcal{M}(S_\mu)$ and choice of μ, we have $\Psi^2(\overline{\mathcal{M}}^\perp) = |S_\mu| \leq \mu^{-q} R_q$. From Lemma 5.6, we have

$$16 \Psi^2(\overline{\mathcal{M}}^\perp) \frac{\tau_\ell(\mathcal{L}_n) + \tau_u(\mathcal{L}_n)}{\gamma_u} \leq \varphi_n(\Sigma),$$

and hence, by definition (5.22) of the contraction coefficient,

$$\kappa \leq \left\{1 - \frac{\gamma_\ell}{2\gamma_u} + \varphi_n(\Sigma)\right\} \left\{1 - \varphi_n(\Sigma)\right\}^{-1}.$$

For proving Corollary 5.3, we observe that the stated settings $\overline{\gamma_\ell}$, $\chi_n(\Sigma)$ and κ follow directly from Lemma 5.6. The bound for condition 5.2(a) follows from a standard argument about the suprema of d independent Gaussians with variance ν.

5.5.5 Proof of Corollary 5.4

This proof is analogous to that of Corollary 5.2, but appropriately adapted to the matrix setting. We first state a lemma that allows us to establish appropriate forms of the RSC/RSM conditions. Recall that we are studying an instance of matrix regression with random design, where the vectorized form $\text{vec}(X)$ of each matrix is drawn from a $N(0, \Sigma)$ distribution, where $\Sigma \in \mathbb{R}^{d^2 \times d^2}$ is some covariance matrix. In order to state this result, let us define the quantity

$$\zeta_{\text{mat}}(\Sigma) := \sup_{\|u\|_2 = 1, \, \|v\|_2 = 1} \text{var}(u^T X v), \quad \text{where } \text{vec}(X) \sim N(0, \Sigma). \tag{5.64}$$

Lemma 5.7. *Under the conditions of Corollary 5.4, there are universal positive constants* (c_0, c_1) *such that*

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \geq \frac{1}{2}\,\sigma_{\min}(\Sigma)\,\|\Delta\|_F^2 - c_1\zeta_{mat}(\Sigma)\frac{d}{n}\,\|\Delta\|_1^2, \qquad \textit{and} \tag{5.65a}$$

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \leq 2\,\sigma_{\max}(\Sigma)\,\|\Delta\|_F^2 - c_1\,\zeta_{mat}(\Sigma)\,\frac{d}{n}\,\|\Delta\|_1^2, \qquad \textit{for all } \Delta \in \mathbb{R}^{d\times d}. \tag{5.65b}$$

with probability at least $1 - \exp(-c_0\, n)$.

Given the quadratic nature of the least-squares loss, the bound (5.65a) implies that the RSC condition holds with $\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma)$ and $\tau_\ell(\mathcal{L}_n) = c_1\zeta_{\text{mat}}(\Sigma)\frac{d}{n}$, whereas the bound (5.65b) implies that the RSM condition holds with $\gamma_u = 2\sigma_{\max}(\Sigma)$ and $\tau_u(\mathcal{L}_n) = c_1\zeta_{\text{mat}}(\Sigma)\frac{d}{n}$.

We now prove Corollary 5.4 in the special case of exactly low rank matrices ($q = 0$), in which Θ^* has some rank $r \leq d$. Given the singular value decomposition $\Theta^* = UDV^T$, let U^r and V^r be the $d \times r$ matrices whose columns correspond to the r non-zero (left and right, respectively) singular vectors of Θ^*. As in Section 5.2.4, define the subspace of matrices

$$\mathcal{M}(U^r, V^r) := \big\{\Theta \in \mathbb{R}^{d\times d} \mid \text{col}(\Theta) \subseteq U^r \text{ and } \text{row}(\Theta) \subseteq V^r\big\}, \tag{5.66}$$

as well as the associated set $\overline{\mathcal{M}}^\perp(U^r, V^r)$. Note that $\Theta^* \in \mathcal{M}$ by construction, and moreover (as discussed in Section 5.2.4, the nuclear norm is decomposable with respect to the pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$.

By definition (5.12) of the subspace compatibility with nuclear norm as the regularizer and Frobenius norm as the error norm, we have $\Psi^2(\mathcal{M}) = r$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 5.7 and substituting into equation (5.22), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\}\,\Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{5.67}$$

where $\chi_n(\Sigma) := \frac{c_2\zeta_{\text{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)}\,\frac{rd}{n}$ for some universal constant c_2. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\,\phi(\Sigma; r, d, n)\Big\{\frac{\|\Delta^*\|_1^2}{r} + \|\Delta^*\|_F^2\Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq \|\Theta^*\|_1$ by assumption, Lemma 5.5 (as exploited in the proof of Corollary 5.1) shows that $\|\Delta^*\|_1^2 \leq 4r\|\Delta^*\|_F^2$, and hence that

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\,\chi_n(\Sigma)\,\|\Delta^*\|_F^2,$$

which show the claim (5.41) for $q = 0$.

We now turn to the case $q \in (0,1]$; as in the proof of this case for Corollary 5.2, we bound $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair. Recall our notation $\sigma_1(\Theta^*) \geq \sigma_2(\Theta^*) \geq \cdots \geq \sigma_d(\Theta^*) \geq 0$ for the ordered singular values of Θ^*. For a threshold μ to be chosen, define $S_\mu = \big\{j \in \{1,2,\ldots,d\} \mid \sigma_j(\Theta^*) > \mu\big\}$, and $U(S_\mu) \in \mathbb{R}^{d \times |S_\mu|}$ be the matrix of left singular vectors indexed by S_μ, with the matrix $V(S_\mu)$ defined similarly. We then define the subspace $\mathcal{M}(S_\mu) := \mathcal{M}(U(S_\mu), V(S_\mu))$ in an analogous fashion to equation (5.66), as well as the subspace $\overline{\mathcal{M}}^\perp(S_\mu)$.

Now by a combination of Lemma 5.5 and the definition (5.23) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \overline{\mathcal{M}}^\perp(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \frac{c\,\zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{d}{n} \Big(\sum_{j \notin S_\mu} \sigma_j(\Theta^*) + \sqrt{|S_\mu|}\, |\!|\!| \Delta^* |\!|\!|_F \Big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^\perp$ on S_μ. As in the proof of Corollary 5.2, we now choose the threshold μ optimally, so as to trade-off the term $\sum_{j \notin S_\mu} \sigma_j(\Theta^*)$ with its competitor $\sqrt{|S_\mu|}\, |\!|\!| \Delta^* |\!|\!|_F$. Exploiting the fact that $\Theta^* \in \mathbb{B}_q(R_q)$ and following the same steps as the proof of Corollary 5.2 yields the bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \frac{c\,\zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{d}{n} \big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q |\!|\!| \Delta^* |\!|\!|_F^2 \big).$$

Setting $\mu^2 = \frac{d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \varphi_n(\Sigma) \left\{ R_q \big(\frac{d}{n}\big)^{1-q/2} + |\!|\!| \Delta^* |\!|\!|_F^2 \right\},$$

as claimed. The stated form of the contraction coefficient can be verified by a calculation analogous to the proof of Corollary 5.2.

5.5.6 Proof of Corollary 5.5

In this case, we let $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ be the operator defined by the model of random signed matrix sampling [114]. As previously argued, establishing the RSM/RSC property amounts to obtaining a form of uniform control over $\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n}$. More specifically, from the proof of Theorem 5.1, we see that it suffices to have a form of RSC for the difference $\widehat{\Delta}^t = \Theta^t - \widehat{\Theta}$, and a form of RSM for the difference $\Theta^{t+1} - \Theta^t$. The following two lemmas summarize these claims:

Lemma 5.8. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, d-1$, with probability at least $1 - \exp(-d \log d)$,*

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2} |\!|\!| \widehat{\Delta}^t |\!|\!|_F^2 - \underbrace{c\alpha \sqrt{\frac{r\, d \log d}{n}} \left\{ \frac{\sum_{j=r+1}^d \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha \sqrt{\frac{rd \log d}{n}} + |\!|\!| \Delta^* |\!|\!|_F \right\}}_{\delta_\ell(r)}. \quad (5.68)$$

Lemma 5.9. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, d-1$, with probability at least $1 - \exp(-d \log d)$, the difference $\Gamma^t := \Theta^{t+1} - \Theta^t$ satisfies the inequality $\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq 2|||\Gamma^t|||_F^2 + \delta_u(r)$, where*

$$\delta_u(r) := c\alpha\sqrt{\frac{rd\log d}{n}}\Big\{\frac{\sum_{j=r+1}^d \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha\sqrt{\frac{rd\log d}{n}} + |||\Delta^*|||_F + |||\widehat{\Delta}^t|||_F + |||\widehat{\Delta}^{t+1}|||_F\Big\}.$$

We can now complete the proof of Corollary 5.5 by a minor modification of the proof of Theorem 5.1. Recalling the elementary relation (5.52), we have

$$|||\Theta^{t+1} - \widehat{\Theta}|||_F^2 = |||\Theta^t - \widehat{\Theta}|||_F^2 + |||\Theta^t - \Theta^{t+1}|||_F^2 - 2\langle\!\langle \Theta^t - \widehat{\Theta},\ \Theta^t - \Theta^{t+1}\rangle\!\rangle.$$

From the proof of Lemma 5.2, we see that the combination of Lemma 5.8 and 5.9 (with $\gamma_\ell = \frac{1}{2}$ and $\gamma_u = 2$) imply that

$$2\langle\!\langle \Theta^t - \Theta^{t+1},\ \Theta^t - \widehat{\Theta}\rangle\!\rangle \geq |||\Theta^t - \Theta^{t+1}|||_F^2 + \frac{1}{4}|||\Theta^t - \widehat{\Theta}|||_F^2 - \delta_u(r) - \delta_\ell(r)$$

and hence that

$$|||\widehat{\Delta}^{t+1}|||_F^2 \leq \frac{3}{4}|||\widehat{\Delta}^t|||_F^2 + \delta_\ell(r) + \delta_u(r).$$

We substitute the forms of $\delta_\ell(r)$ and $\delta_u(r)$ given in Lemmas 5.8 and 5.9 respectively; performing some algebra then yields

$$\Big\{1 - \frac{c\,\alpha\sqrt{\frac{rd\log d}{n}}}{|||\widehat{\Delta}^{t+1}|||_F}\Big\}|||\widehat{\Delta}^{t+1}|||_F^2 \leq \Big\{\frac{3}{4} + \frac{c\alpha\sqrt{\frac{rd\log d}{n}}}{|||\widehat{\Delta}^t|||_F}\Big\}|||\widehat{\Delta}^t|||_F^2 + c'\,\delta_\ell(r).$$

Consequently, as long as $\min\{|||\widehat{\Delta}^t|||_F^2,\ |||\widehat{\Delta}^{t+1}|||_F^2\} \geq c_3\alpha\frac{rd\log d}{n}$ for a sufficiently large constant c_3, we are guaranteed the existence of some $\kappa \in (0,1)$ such that

$$|||\widehat{\Delta}^{t+1}|||_F^2 \leq \kappa|||\widehat{\Delta}^t|||_F^2 + c'\delta_\ell(r). \tag{5.69}$$

Since $\delta_\ell(r) = \Omega(\frac{rd\log d}{n})$, this inequality (5.69) is valid for all $t = 0, 1, 2, \ldots$ as long as c' is sufficiently large. As in the proof of Theorem 5.1, iterating the inequality (5.69) yields

$$|||\widehat{\Delta}^{t+1}|||_F^2 \leq \kappa^t|||\widehat{\Delta}^0|||_F^2 + \frac{c'}{1-\kappa}\,\delta_\ell(r). \tag{5.70}$$

It remains to choose the cut-off $r \in \{1, 2, \ldots, d-1\}$ so as to minimize the term $\delta_\ell(r)$. In particular, when $\Theta^* \in \mathbb{B}_q(R_q)$, then as shown in the paper [115], the optimal choice is

$r \asymp \alpha^{-q} R_q \big(\frac{n}{d \log d}\big)^{q/2}$. Substituting into the inequality (5.70) and performing some algebra yields that there is a universal constant c_4 such that the bound

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \kappa^t \|\widehat{\Delta}^0\|_F^2 + \frac{c_4}{1-\kappa}\Big\{R_q\big(\frac{\alpha d \log d}{n}\big)^{1-q/2} + \sqrt{R_q\big(\frac{\alpha d \log d}{n}\big)^{1-q/2}}\, \|\Delta^*\|_F\Big\}.$$

holds. Now by the Cauchy-Schwarz inequality we have

$$\sqrt{R_q\big(\frac{\alpha d \log d}{n}\big)^{1-q/2}}\, \|\Delta^*\|_F \leq \frac{1}{2} R_q\big(\frac{\alpha d \log d}{n}\big)^{1-q/2} + \frac{1}{2}\|\Delta^*\|_F^2,$$

and the claimed inequality (5.44) follows.

5.5.7 Proof of Corollary 5.6

Again the main argument in the proof would be to establish the RSM and RSC properties for the decomposition problem. We define $\widehat{\Delta}^t_\Theta = \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}^t_\Gamma = \Gamma^t - \widehat{\Gamma}$. We start with giving a lemma that establishes RSC for the differences $(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma)$. We recall that just like noted in the previous section, it suffices to show RSC only for these differences. Showing RSC/RSM in this example amounts to analyzing $\|\widehat{\Delta}^t_\Theta + \widehat{\Delta}^t_\Gamma\|_F^2$. We recall that this section assumes that Γ^* has only s non-zero columns.

Lemma 5.10. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots,$*

$$\|\widehat{\Delta}^t_\Theta + \widehat{\Delta}^t_\Gamma\|_F^2 \geq \frac{1}{2}\big(\|\widehat{\Delta}^t_\Theta\|_F^2 + \|\widehat{\Delta}^t_\Gamma\|_F^2\big) - c\alpha\sqrt{\frac{s}{d_2}}\bigg(\|\widehat{\Gamma} - \Gamma^*\|_F + \alpha\sqrt{\frac{s}{d_2}}\bigg) \tag{5.71}$$

This proof of this lemma follows by a straightforward modification of analogous results in the paper [8].

Matrix decomposition has the interesting property that the RSC condition holds in a deterministic sense (as opposed to with high probability). The same deterministic guarantee holds for the RSM condition; indeed, we have

$$\|\widehat{\Delta}^t_\Delta + \widehat{\Delta}^t_\Gamma\|_F^2 \leq 2\big(\|\widehat{\Delta}^t_\Theta\|_F^2 + \|\widehat{\Delta}^t_\Gamma\|_F^2\big), \tag{5.72}$$

by Cauchy-Schwartz inequality. Now we appeal to the more general form of Theorem 5.1 as stated in Equation 5.48, which gives

$$\|\widehat{\Delta}^{t+1}_\Theta\|_F^2 + \|\widehat{\Delta}^{t+1}_\Gamma\|_F^2 \leq \bigg(\frac{3}{4}\bigg)^t \big(\|\widehat{\Delta}^0_\Theta\|_F^2 + \|\widehat{\Delta}^0_\Gamma\|_F^2\big) + c\sqrt{\frac{\alpha s}{d_2}}\bigg(\|\widehat{\Gamma} - \Gamma^*\|_F + \frac{\alpha s}{d_2}\bigg).$$

The stated form of the corollary follows by an application of Cauchy-Schwarz inequality.

5.6 Discussion

In this chapter, we have shown that even though high-dimensional M-estimators in statistics are neither strongly convex nor smooth, simple first-order methods can still enjoy global guarantees of geometric convergence. The key insight is that strong convexity and smoothness need only hold in restricted senses, and moreover, these conditions are satisfied with high probability for many statistical models and decomposable regularizers used in practice. Examples include sparse linear regression and ℓ_1-regularization, various statistical models with group-sparse regularization, matrix regression with nuclear norm constraints (including matrix completion and multi-task learning), and matrix decomposition problems. Overall, our results highlight some important connections between computation and statistics: the properties of M-estimators favorable for fast rates in a statistical sense can also be used to establish fast rates for optimization algorithms.

Chapter 6

Asymptotically optimal algorithms for distributed machine learning

In this chapter, we focus on stochastic convex optimization problems of the form

$$\underset{\theta\in\Omega}{\text{minimize}} \; f(\theta) \quad \text{for} \quad f(\theta) := \mathbb{E}_P[F(\theta; z)] = \int_{\mathcal{Z}} F(\theta; z) dP(z), \tag{6.1}$$

where $\Omega \subseteq \mathbb{R}^d$ is a closed convex set, P is a probability distribution over $\mathcal{Z}$, and $F(\cdot\,; z)$ is convex for all $z \in \mathcal{Z}$, so that f is convex. The goal is to find a parameter x that approximately minimizes f over $\theta \in \Omega$. Classical stochastic gradient algorithms [139, 129] iteratively update a parameter $\theta^t \in \Omega$ by sampling $z \sim P$, computing $g(t) = \nabla F(\theta^t; z)$, and performing the update $\theta^{t+1} = \Pi_\Omega(\theta^t - \alpha(t)g(t))$, where Π_Ω denotes projection onto the set Ω and $\alpha(t) \in \mathbb{R}$ is a stepsize. In this chapter, we analyze asynchronous gradient methods, where instead of receiving current information $g(t)$, the procedure receives out of date gradients $g(t - \tau(t)) = \nabla F(\theta^{(t-\tau(t))}, z)$, where $\tau(t)$ is the (potentially random) delay at time t. The central contribution of this chapter is to develop algorithms that—under natural assumptions about the functions F in the objective (6.1)—achieve asymptotically optimal rates for stochastic convex optimization in spite of delays. We start by giving the motivation for this problem in the context of distributed machine learning, and survey some of the past work in this area.

6.1 Motivation and related work

Our model of delayed gradient information is particularly relevant in distributed optimization scenarios, where a master maintains the parameters θ while workers compute stochastic gradients of the objective (6.1). The architectural assumption of a master with several worker nodes is natural for distributed computation, and other researchers have considered models similar to those in this chapter [113, 97]. By allowing delayed and asynchronous updates, we can avoid synchronization issues that commonly handicap distributed systems.

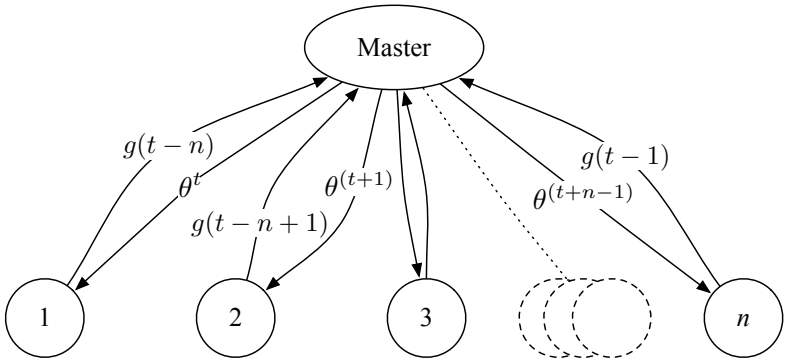

Figure 6.1. Cyclic delayed update architecture. Workers compute gradients cyclically and in parallel, passing out-of-date information to master. Master responds with current parameters. Diagram shows parameters and gradients communicated between rounds t and $t+n-1$.

Certainly distributed optimization has been studied for several decades, tracing back at least to seminal work of Bertsekas and Tsitsiklis (1983, 1984, 1989) on asynchronous computation and minimization of smooth functions where the parameter vector is distributed. More recent work has studied problems in which each processor or node i in a network has a local function f_i, and the goal is to minimize the sum $f(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta)$ [112, 131, 81, 61]. Much prior work in this setting implicitly assumes as a *constraint* that data lies on several different nodes throughout a network. However, as Dekel et al. [57] first noted, in distributed stochastic settings independent realizations of a stochastic gradient can be computed concurrently, and it is thus possible to obtain an aggregated gradient estimate with lower variance. Using modern stochastic optimization algorithms [e.g. 83, 96], Dekel et al. give a series of reductions to show that in an n-node network it is possible to achieve a speedup of $\mathcal{O}(n)$ over a single-processor so long as the objective f is smooth.

Our work is closest to Nedić et al.'s asynchronous incremental subgradient method (2001), which is an incremental gradient procedure in which gradient projection steps are taken using out-of-date gradients. See Figure 6.1 for an illustration. Nedić et al. show that in spite of asynchrony and the fact that f is non-smooth, the asynchronous subgradient method guarantees convergence to a minimum of $\frac{1}{n}\sum_{i=1}^n f_i(\theta)$. In addition, using the results in the above paper one can prove a finite sample convergence rate: if the gradients are computed with a delay of τ, then the optimization error of the procedure after T iterations is at most $\mathcal{O}(\sqrt{\tau/T})$. As in Fig. 6.1, the delay τ can essentially be of order n in an n-node distributed network, giving a convergence rate of $\mathcal{O}(\sqrt{n/T})$. Without delay, a centralized stochastic gradient algorithm attains convergence rate $\mathcal{O}(1/\sqrt{T})$ and Langford et al. [97] also consider general delayed stochastic optimization and attempt to remove the asymptotic delay penalty by considering smooth objective functions, though their approach has a technical error (see

Appendix C of the preprint [2]). They also do not demonstrate any provable benefits of distributed computation. This leads to the two motivating questions of our work: (1) is it possible (perhaps under additional assumptions) to remove the delay penalty and (2) is it possible to demonstrate benefits in convergence rate by leveraging parallel computation, in spite of delays? Analyzing similar asynchronous algorithms, we show that the answer to both the above questions is yes. For smooth stochastic problems the delay is asymptotically negligible—the time τ does not matter—and in fact, with parallelization, delayed updates can give provable performance benefits.

We build on results of Dekel et al. [57], who show that when the objective f has Lipschitz-continuous gradients, then when n processors compute stochastic gradients in parallel using a common parameter θ it is possible to achieve convergence rate $\mathcal{O}(1/\sqrt{Tn})$ so long as the processors are synchronized (under appropriate synchrony conditions, this holds nearly independently of network topology). A variant of their approach is asymptotically robust to asynchrony so long as most processors remain synchronized for most of the time [56]. We show results similar to their initial discovery, but we analyze the effects of asynchronous gradient updates where all the nodes in the network can suffer delays. Application of our main results to the distributed setting provides convergence rates in terms of the number of nodes n in the network and the stochastic process governing the delays. Concretely, we show that under different assumptions on the network and delay process, we achieve convergence rates ranging from $\mathcal{O}(n^3/T + 1/\sqrt{Tn})$ to $\mathcal{O}(n/T + 1/\sqrt{Tn})$, which is $\mathcal{O}(1/\sqrt{nT})$ asymptotically in T. For problems with large n, we demonstrate rates ranging from $\mathcal{O}((n/T)^{2/3} + 1/\sqrt{Tn})$ to $\mathcal{O}(1/T^{2/3} + 1/\sqrt{Tn})$. In either case, the time necessary to achieve ϵ-optimal solution to the problem (6.1) is asymptotically $\mathcal{O}(1/n\epsilon^2)$, a factor of n—the size of the network—better than a centralized procedure in spite of delay.

The remainder of the chapter is organized as follows. We begin by reviewing known algorithms for solving the stochastic optimization problem (6.1) and stating our main assumptions. Then in Section 6.3 we give abstract descriptions of our algorithms and state our main theoretical results, which we make concrete in Section 6.4 by formally placing the analysis in the setting of distributed stochastic optimization. We complement the theory in Section 6.5 with experiments on a real-world dataset, and proofs follow in the remaining sections. We also note that the results of this chapter appear in the paper [3].

Notation We collect our (mostly standard) notation specific to this chapter here. We recall the definitions related to a convex function from Section 2.2. We use the shorthand $\|\partial f(\theta)\|_* := \sup_{g \in \partial f(\theta)} \|g\|_*$. We assume that f is G-Lipschitz, which by convexity, is equivalent to $\|\partial f(\theta)\|_* \le G$ for all $\theta \in \Omega$ [76]. For convex differentiable h, the Bregman divergence [38] between θ and $\tilde{\theta}$ is defined as

$$D_h(\theta, \tilde{\theta}) := h(\theta) - h(\tilde{\theta}) - \left\langle \nabla h(\tilde{\theta}), \theta - \tilde{\theta} \right\rangle. \tag{6.2}$$

We use $[n]$ to denote the set of integers $\{1, \ldots, n\}$.

6.2 Setup and Algorithms

In this section we set up and recall the delay-free algorithms underlying our approach. We then give the appropriate delayed versions of these algorithms, which we analyze in the sequel.

6.2.1 Setup and Delay-free Algorithms

To build intuition for the algorithms we analyze, we first describe two closely related first-order algorithms: the dual averaging algorithm of Nesterov [122] and the mirror descent algorithm of Nemirovski and Yudin [119], which is analyzed further by Beck and Teboulle [22]. We begin by collecting notation and giving useful definitions. Both algorithms are based on a proximal function $\psi(\theta)$, where it is no loss of generality to assume that $\psi(\theta) \geq 0$ for all $\theta \in \Omega$. We assume ψ is 1-strongly convex (by scaling, this is no loss of generality). By definitions (6.2) and (2.5), the divergence D_ψ satisfies $D_\psi(\theta, \tilde{\theta}) \geq \frac{1}{2}\left\|\theta - \tilde{\theta}\right\|^2$.

In the oracle model of stochastic optimization that we assume, at time t both algorithms query an oracle at the point θ^t, and the oracle then samples $z(t)$ i.i.d. from the distribution P and returns $g(t) \in \partial F(\theta^t; z(t))$. The dual averaging algorithm [122] updates a dual vector μ^t and primal vector $\theta^t \in \Omega$ via

$$\mu^{t+1} = \mu^t + g(t) \quad \text{and} \quad \theta^{t+1} = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \left\langle \mu^{t+1}, \theta \right\rangle + \frac{1}{\alpha(t+1)} \psi(\theta) \right\}, \tag{6.3}$$

while mirror descent [119, 22] performs the update

$$\theta^{t+1} = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \left\langle g(t), \theta \right\rangle + \frac{1}{\alpha(t)} D_\psi(\theta, \theta^t) \right\}. \tag{6.4}$$

Both make a linear approximation to the function being minimized—a global approximation in the case of the dual averaging update (6.3) and a more local approximation for mirror descent (6.4)—while using the proximal function ψ to regularize the points θ^t.

We now state the two essentially standard assumptions [83, 96, 173] we most often make about the stochastic optimization problem (6.1), after which we recall the convergence rates of the algorithms (6.3) and (6.4).

Assumption A (Lipschitz Functions)**.** For P-a.e. z, the function $F(\cdot\,; z)$ is convex. Moreover, for any $\theta \in \Omega$, $\mathbb{E}[\|\partial F(\theta; z)\|_*^2] \leq G^2$.

In particular, Assumption A implies that f is G-Lipschitz continuous with respect to the norm $\|\cdot\|$ and that f is convex. Our second assumption has been used to show rates of convergence based on the variance of a gradient estimator for stochastic optimization problems [83, 96].

Assumption B (Smooth Functions)**.** The function f defined in (6.1) has L-Lipschitz continuous gradients, and for all $\theta \in \Omega$ the variance bound $\mathbb{E}[\|\nabla f(\theta) - \nabla F(\theta; z)\|_*^2] \le \sigma^2$ holds.[1]

Several commonly used functions satisfy the above assumptions, and we recall some of the relevant examples from Section 2.1 here:

(i) The *logistic loss*: $F(\theta; z) = \log[1 + \exp(-\langle \theta, z \rangle)]$, the objective for logistic regression [e.g. 73], where $z = xy$ for $y \in \{-1, 1\}$ and $x \in \mathbb{R}^d$. The objective F satisfies Assumptions A and B so long as $\|z\|$ is bounded.

(ii) *Least squares* or *linear regression*: $F(\theta; z) = (y - \langle \theta, x \rangle)^2$ where $z = (x, y)$ for $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, satisfies Assumptions A and B as long as z is bounded and Ω is compact.

We also make a standard compactness assumption on the optimization set Ω.

Assumption C (Compactness)**.** For $\theta^* \in \operatorname{argmin}_{\theta \in \Omega} f(\theta)$ and $\theta \in \Omega$, the bounds $\psi(\theta^*) \le R^2/2$ and $D_\psi(\theta^*, \theta) \le R^2$ both hold.

Under Assumptions A or B in addition to Assumption C, the updates (6.3) and (6.4) have known convergence rates. Define the time averaged vector $\overline{\theta}(T)$ as

$$\overline{\theta}(T) := \frac{1}{T} \sum_{t=1}^{T} \theta^{t+1}. \tag{6.5}$$

Then under Assumption A, both algorithms satisfy

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{RG}{\sqrt{T}}\right) \tag{6.6}$$

for the stepsize choice $\alpha(t) = R/(G\sqrt{t})$ [e.g. 122, 173, 117]. The result (6.6) is sharp to constant factors in general [119, 6], but can be further improved under Assumption B. Building on work of Juditsky et al. [83] and Lan [96], Dekel et al. [57, Appendix A] show that under Assumptions B and C the stepsize choice $\alpha(t)^{-1} = L + \eta(t)$, where $\eta(t)$ is a damping factor set to $\eta(t) = \sigma R\sqrt{t}$, yields for either of the updates (6.3) or (6.4) the convergence rate

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{LR^2}{T} + \frac{\sigma R}{\sqrt{T}}\right). \tag{6.7}$$

[1] If f is differentiable, then $F(\cdot; z)$ is differentiable for P-a.e. z, and conversely, but F need not be smoothly differentiable [25]. Since $\nabla F(\theta; z)$ exists for P-a.e. z, we will write $\nabla F(\theta; z)$ with no loss of generality.

6.2.2 Delayed Optimization Algorithms

We now turn to extending the dual averaging (6.3) and mirror descent (6.4) updates to the setting in which instead of receiving a current gradient $g(t)$ at time t, the procedure receives a gradient $g(t-\tau(t))$, that is, a stochastic gradient of the objective (6.1) computed at the point $\theta^{(t-\tau(t))}$. In the simplest case, the delays are uniform and $\tau(t) \equiv \tau$ for all t, but in general the delays may be a non-i.i.d. stochastic process. Our analysis admits any sequence $\tau(t)$ of delays as long as the mapping $t \mapsto \tau(t)$ satisfies $\mathbb{E}[\tau(t)^2] \leq B^2 < \infty$. We also require that each update happens once, i.e., $t \mapsto t - \tau(t)$ is one-to-one, though this second assumption is easily satisfied.

Recall that the problems we consider are stochastic optimization problems of the form (6.1). Under the assumptions above, we extend the mirror descent and dual averaging algorithms in the simplest way: we replace $g(t)$ with $g(t-\tau(t))$. For dual averaging (cf. the update (6.3)) this yields

$$\mu^{t+1} = \mu^t + g(t - \tau(t)) \quad \text{and} \quad \theta^{t+1} = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \left\langle \mu^{t+1}, \theta \right\rangle + \frac{1}{\alpha(t+1)} \psi(\theta) \right\}, \tag{6.8}$$

while for mirror descent (cf. the update (6.4)) we have

$$\theta^{t+1} = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \left\langle g(t - \tau(t)), \theta \right\rangle + \frac{1}{\alpha(t)} D_\psi(\theta, \theta^t) \right\}. \tag{6.9}$$

An extension of Nedić et al.'s (2001) results by combining their techniques with the convergence proofs of dual averaging [122] and mirror descent [22] is as follows. Under Assumptions A and C, so long as $\mathbb{E}[\tau(t)] \leq B < \infty$ for all t, choosing $\alpha(t) = \frac{R}{G\sqrt{Bt}}$ gives rate

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{RG\sqrt{B}}{\sqrt{T}} \right). \tag{6.10}$$

6.3 Convergence rates for delayed optimization of smooth functions

In this section, we state and discuss several results for asynchronous stochastic gradient methods. We give two sets of theorems. The first are for the asynchronous method when we make updates to the parameter vector x using one stochastic subgradient, according to the update rules (6.8) or (6.9). The second method involves using several stochastic subgradients for every update, each with a potentially different delay, which gives sharper results that we present in Section 6.3.2.

6.3.1 Simple delayed optimization

Intuitively, the $\sqrt{B}$-penalty due to delays for non-smooth optimization arises from the fact that subgradients can change drastically when measured at slightly different locations, so a small delay can introduce significant inaccuracy. To overcome the delay penalty, we now turn to the smoothness assumption B as well as the Lipschitz condition A (we assume both of these conditions along with Assumption C hold for all the theorems). In the smooth case, delays mean that stale gradients are only slightly perturbed, since our stochastic algorithms constrain the variability of the points θ^t. As we show in the proofs of the remaining results, the error from delay essentially becomes a second order term: the penalty is asymptotically negligible. We study both update rules (6.8) and (6.9), and we set $\alpha(t) = \frac{1}{L+\eta(t)}$. Here $\eta(t)$ will be chosen to both control the effects of delays and for errors from stochastic gradient information. We prove the following theorem in Sec. 6.6.1.

Theorem 6.1. *Let the sequence θ^t be defined by the update (6.8). Define the stepsize $\eta(t) \propto \sqrt{t}$ or let $\eta(t) \equiv \eta$ for all t. Then*

$$\mathbb{E}\left[\sum_{t=1}^{T} f(\theta^{t+1})\right] - Tf(\theta^*) \leq \frac{1}{\alpha(T+1)}R^2 + \frac{\sigma^2}{2}\sum_{t=1}^{T}\frac{1}{\eta(t)} + 2LG^2(\tau+1)^2\sum_{t=1}^{T}\frac{1}{\eta(t)^2} + 4\tau GR.$$

The mirror descent update (6.9) exhibits similar convergence properties, and we prove the next theorem in Sec. 6.6.2.

Theorem 6.2. *Use the conditions of Theorem 6.1 but generate θ^t by the update (6.9). Then*

$$\mathbb{E}\left[\sum_{t=1}^{T} f(\theta^{t+1})\right] - Tf(\theta^*) \leq LR^2 + R^2\eta(T) + \frac{\sigma^2}{2}\sum_{t=1}^{T}\frac{1}{\eta(t)} + 2LG^2(\tau+1)^2\sum_{t=1}^{T}\frac{1}{\eta(t)^2} + 4\tau GR.$$

In each of the above theorems, we can set $\eta(t) = \sigma\sqrt{t}/R$. As immediate corollaries, we recall the definition (6.5) of the averaged sequence of θ^t and use convexity to see that

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{LR^2 + \tau GR}{T} + \frac{\sigma R}{\sqrt{T}} + \frac{LG^2\tau^2R^2\log T}{\sigma^2 T}\right)$$

for either update rule. In addition, we can allow the delay $\tau(t)$ to be random:

Corollary 6.1. *Let the conditions of Theorem 6.1 or 6.2 hold, but allow $\tau(t)$ to be a random mapping such that $\mathbb{E}[\tau(t)^2] \leq B^2$ for all t. With the choice $\eta(t) = \sigma\sqrt{T}/R$ the updates (6.8) and (6.9) satisfy*

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{LR^2 + B^2GR}{T} + \frac{\sigma R}{\sqrt{T}} + \frac{LG^2B^2R^2}{\sigma^2 T}\right).$$

We provide the proof of the corollary in Sec. 6.6.3. Even though the corollary is stated only for fixed step size, the conclusion extends to decaying step sizes too, albeit with an additional $\log T$ factor on the last term. The take-home message from the above corollaries, as well as Theorems 6.1 and 6.2, is that the penalty in convergence rate due to the delay $\tau(t)$ is asymptotically negligible. As we discuss in greater depth in the next section, this has favorable implications for robust distributed stochastic optimization algorithms.

6.3.2 Combinations of delays

In some scenarios—including distributed settings similar to those we discuss in the next section—the procedure has access not to only a single delayed gradient but to several with different delays. To abstract away the essential parts of this situation, we assume that the procedure receives n gradients $g_1, \dots, g_n$, where each has a potentially different delay $\tau(i)$. Now let $\lambda = (\lambda_i)_{i=1}^n$ belong to the probability simplex, though we leave λ's values unspecified for now. Then the procedure performs the following updates at time t: for dual averaging,

$$\mu^{t+1} = \mu^t + \sum_{i=1}^n \lambda_i g_i(t - \tau(i)) \quad \text{and} \quad \theta^{t+1} = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \langle \mu^{t+1}, \theta \rangle + \frac{1}{\alpha(t+1)} \psi(\theta) \right\} \tag{6.11}$$

while for mirror descent, the update is

$$g_\lambda(t) = \sum_{i=1}^n \lambda_i g_i(t - \tau(i)) \quad \text{and} \quad \theta^{t+1} = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \langle g_\lambda(t), \theta \rangle + \frac{1}{\alpha(t)} D_\psi(\theta, \theta^t) \right\}. \tag{6.12}$$

The next two theorems build on the proofs of Theorems 6.1 and 6.2, combining several techniques. We provide the proof of Theorem 6.3 in Sec. 6.7, omitting the proof of Theorem 6.4 as it follows in a similar way from Theorem 6.2.

Theorem 6.3. *Let the sequence θ^t be defined by the update (6.11). Under assumptions A, B and C, let $\frac{1}{\alpha(t)} = L + \eta(t)$ and $\eta(t) \propto \sqrt{t}$ or $\eta(t) \equiv \eta$ for all t. Then*

$$\mathbb{E}\left[\sum_{t=1}^T f(\theta^{t+1}) - T f(\theta^*)\right] \le \frac{1}{\alpha(T+1)} R^2 + 4 \sum_{i=1}^n \lambda_i \tau(i) G R + 2 \sum_{i=1}^n \lambda_i L G^2 (\tau(i)+1)^2 \sum_{t=1}^T \frac{1}{\eta(t)^2} + \sum_{t=1}^T \frac{1}{2\eta(t)} \mathbb{E}\left\| \sum_{i=1}^n \lambda_i [\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))] \right\|_*^2.$$

Theorem 6.4. *Use the same conditions as Theorem 6.3, but assume that θ^t is defined by*

the update (6.12) and $D_\psi(\theta^*, \theta) \le R^2$ *for all* $\theta \in \Omega$. *Then*

$$\mathbb{E}\left[\sum_{t=1}^{T} f(\theta^{t+1}) - T f(\theta^*)\right] \le (L + \eta(T))R^2 + 4\sum_{i=1}^{n} \lambda_i \tau(i) GR + 2\sum_{i=1}^{n} \lambda_i L G^2 (\tau(i)+1)^2 \sum_{t=1}^{T} \frac{1}{\eta(t)^2} + \sum_{t=1}^{T} \frac{1}{2\eta(t)} \mathbb{E}\left\| \sum_{i=1}^{n} \lambda_i [\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))] \right\|_*^2.$$

The consequences of Theorems 6.3 and 6.4 are powerful, as we illustrate in the next section.

6.4 Distributed Optimization

We now turn to what we see as the main purpose and application of the above results: developing robust and efficient algorithms for distributed stochastic optimization. Our main motivations here are machine learning and statistical applications where the data is so large that it cannot fit on a single computer. Examples of the form (6.1) include logistic regression, where the task is to learn a linear classifier that assigns labels in $\{-1, +1\}$ to a series of examples, in which case we have the objective $F(\theta; z) = \log[1 + \exp(-\langle z, \theta \rangle)]$ as described in Sec. 6.2.1(i); or linear regression, where $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ and $F(\theta; z) = \frac{1}{2}[y - \langle x, \theta \rangle]^2$ as described in Sec. 6.2.1(ii). Both objectives satisfy assumptions A and B as discussed earlier. We consider both stochastic and online/streaming scenarios for such problems. In the simplest setting, the distribution P in the objective (6.1) is the empirical distribution over an observed dataset, that is,

$$f(\theta) = \frac{1}{N} \sum_{i=1}^{N} F(\theta; z_i).$$

We divide the N samples among n workers so that each worker has an N/n-sized subset of data. In streaming applications, the distribution P is the unknown distribution generating the data, and each worker receives a stream of independent data points $z \sim P$. Worker i uses its subset of the data, or its stream, to compute g_i, an estimate of the gradient ∇f of the global f. We make the simplifying assumption that g_i is an unbiased estimate of $\nabla f(\theta)$, which is satisfied, for example, when each worker receives an independent stream of samples or computes the gradient g_i based on samples picked at random without replacement from its subset of the data.

The architectural assumptions we make are natural and based off of master/worker topologies, but the convergence results in Section 6.3 allow us to give procedures robust to delay and asynchrony. We consider two protocols: in the first, workers compute and communicate asynchronously and independently with the master, and in the second, workers are at different distances from the master and communicate with time lags proportional to

their distances. We show in the latter part of this section that the convergence rates of each protocol are $\mathcal{O}(1/\sqrt{nT})$ for n-node networks (though lower order terms are different for each).

Before describing our architectures, we note that perhaps the simplest master-worker scheme is to have each worker simultaneously compute a stochastic gradient and send it to the master, which takes a gradient step on the averaged gradient. While the n gradients are computed in parallel, accumulating and averaging n gradients at the master takes $\Omega(n)$ time, offsetting the gains of parallelization. Thus we consider alternate architectures that are robust to delay.

Cyclic Delayed Architecture This protocol is the delayed update algorithm mentioned in the introduction, and it parallelizes computation of (estimates of) the gradient $\nabla f(\theta)$. Formally, worker i has parameter θ^t and computes $g_i(t) = \nabla F(\theta^t; z_i(t)) \in \mathbb{R}^d$, where $z_i(t)$ is a random variable sampled at worker i from the distribution P. The master maintains a parameter vector $\theta \in \Omega$. The algorithm proceeds in rounds, cyclically pipelining updates. The algorithm begins by initiating gradient computations at different workers at slightly offset times. At time t, the master receives gradient information at a τ-step delay from some worker, performs a parameter update, and passes the updated central parameter θ^{t+1} back to the worker. Other workers do not see this update and continue their gradient computations on stale parameter vectors. In the simplest case, each node suffers a delay of $\tau = n$, though our earlier analysis applies to random delays throughout the network as well. Recall Fig. 6.1 for a graphic description of the process.

Locally Averaged Delayed Architecture At a high level, the protocol we now describe combines the delayed updates of the cyclic delayed architecture with averaging techniques of previous work [112, 61]. We assume a network $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of n nodes (workers) and $\mathcal{E}$ are the edges between the nodes. We select one of the nodes as the master, which maintains the parameter vector $\theta^t \in \Omega$ over time.

The algorithm works via a series of multicasting and aggregation steps on a spanning tree rooted at the master node. In the first phase, the algorithm broadcasts from the root towards the leaves. At step t the master sends its current parameter vector θ^t to its immediate neighbors. Simultaneously, every other node broadcasts its current parameter vector (which, for a depth d node, is $\theta^{(t-d)}$) to its children in the spanning tree. See Fig. 6.2(a). Every worker receives its new parameter and computes its local gradient at this parameter. The second part of the communication in a given iteration proceeds from leaves toward the root. The leaf nodes communicate their gradients to their parents. The parent takes the gradients of the leaf nodes from the previous round (received at iteration $t-1$) and averages them with its own gradient, passing this averaged gradient back up the tree. Again simultaneously, each node takes the averaged gradient vectors of its children from the previous rounds, averages them with its current gradient vector, and passes the result up the spanning tree. See

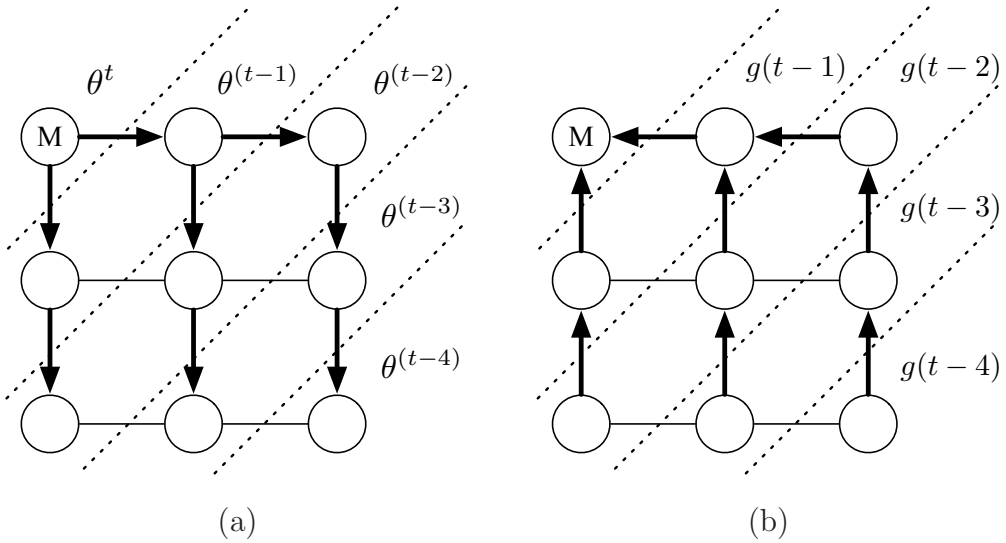

Figure 6.2. Master-worker averaging network. (a): parameters stored at different distances from master node at time t. A node at distance d from master has the parameter $\theta^{(t-d)}$. (b): gradients computed at different nodes. A node at distance d from master computes gradient $g(t-d)$.

Fig. 6.2(b) and Fig. 6.3 for a visual description.

Slightly more formally, associated with each node $i \in \mathcal{V}$ is a delay $\tau(i)$, which is (generally) twice its distance from the master. Fix an iteration t. Each node $i \in \mathcal{V}$ has an out of date parameter vector $\theta^{(t-\tau(i)/2)}$, which it sends further down the tree to its children. So, for example, the master node sends the vector θ^t to its children, which send the parameter vector θ^{t-1} to their children, which in turn send θ^{t-2} to their children, and so on. Each node computes

$$g_i(t-\tau(i)/2) = \nabla F(\theta^{(t-\tau(i)/2)}; z_i(t)),$$

where $z_i(t)$ is a random variable sampled at node i from the distribution P. The communication back up the hierarchy proceeds as follows: the leaf nodes in the tree (say at depth d) send the gradient vectors $g_i(t-d)$ to their immediate parents in the tree. At the previous iteration $t-1$, the parent nodes received $g_i(t-d-1)$ from their children, which they average with their own gradients $g_i(t-d+1)$ and pass to their parents, and so on. The master node at the root receives an average of delayed gradients from the entire tree, with each gradient having a potentially different delay, leading to updates of the form (6.11) or (6.12).

6.4.1 Convergence rates for delayed distributed minimization

Having described our architectures, we can now give corollaries to the theoretical results from the previous sections that show it is possible to achieve asymptotically faster rates (over

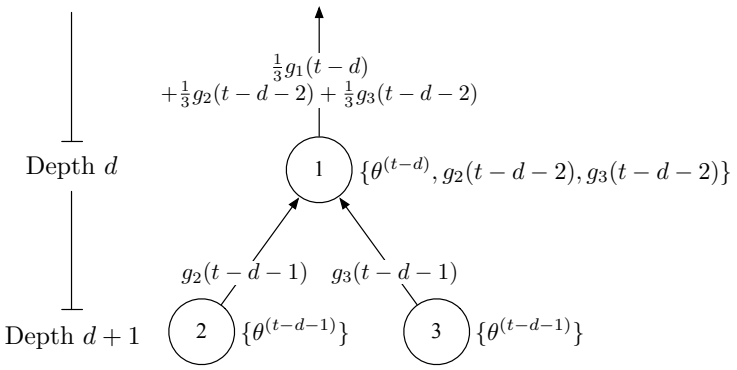

Figure 6.3. Communication of gradient information toward master node at time t from node 1 at distance d from master. Information stored at time t by node i in brackets to right of node.

centralized procedures) using distributed algorithms even without imposing synchronization requirements. We allow workers to pipeline updates by computing asynchronously and in parallel, so each worker can compute low variance estimate of the gradient $\nabla f(\theta)$.

We begin with a simple corollary to the results in Sec. 6.3.1. We ignore the constants L, G, R, and σ, which are not dependent on the characteristics of the network. We also assume that each worker uses m independent samples of $z \sim P$ to compute the stochastic gradient

$$g_i(t) = \frac{1}{m}\sum_{j=1}^{m} \nabla F(\theta^t; z_i(j)).$$

Using the cyclic protocol as in Fig. 6.1, Theorems 6.1 and 6.2 give the following result.

Corollary 6.2. *Let $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, assume the conditions in Corollary 6.1, and assume that each worker uses m samples $z \sim P$ to compute the gradient it communicates to the master. Then with the choice $\eta(t) = \sqrt{T}/\sqrt{m}$ either of the updates (6.8) or (6.9) satisfy*

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{B^2}{T} + \frac{1}{\sqrt{Tm}} + \frac{B^2 m}{T}\right).$$

Proof. The corollary follows straightforwardly from the realization that the variance

$$\sigma^2 = \mathbb{E}[\left\|\nabla f(\theta^t) - g_i(t)\right\|_2^2] = \mathbb{E}[\left\|\nabla f(\theta^t) - \nabla F(\theta^t; z)\right\|_2^2]/m = \mathcal{O}(1/m)$$

when workers use m independent stochastic gradient samples. □

In the above corollary, so long as the bound B on the expected delay satisfies, say, $B = o(T^{1/4})$, then the last term in the bound is asymptotically negligible, and we achieve a convergence rate of $\mathcal{O}(1/\sqrt{Tm})$.

The cyclic delayed architecture has the drawback that information from a worker can take $\mathcal{O}(n)$ time to reach the master. While the algorithm is robust to delay and does not need lock-step coordination of workers, the downside of the architecture is that the essentially n^2m/T term in the bounds above can be quite large. Indeed, if each worker computes its gradient over m samples with $m \approx n$—say to avoid idling of workers—then the cyclic architecture has convergence rate $\mathcal{O}(n^3/T + 1/\sqrt{nT})$. For moderate T or large n, the delay penalty n^3/T may dominate $1/\sqrt{nT}$, offsetting the gains of parallelization.

To address the large n drawback, we turn our attention to the locally averaged architecture described by Figs. 6.2 and 6.3, where delays can be smaller since they depend only on the height of a spanning tree in the network. The algorithm requires more synchronization than the cyclic architecture but still performs limited local communication. Each worker computes $g_i(t-\tau(i)) = \nabla F(\theta^{(t-\tau(i))}; z_i(t))$ where $\tau(i)$ is the delay of worker i from the master and $z_i \sim P$. As a result of the communication procedure, the master receives a convex combination of the stochastic gradients evaluated at each worker i, for which we gave results in Section 6.3.2.

In this architecture, the master receives gradients of the form $g_\lambda(t) = \sum_{i=1}^n \lambda_i g_i(t-\tau(i))$ for some λ in the simplex, which puts us in the setting of Theorems 6.3 and 6.4. We now make the reasonable assumption that the gradient errors $\nabla f(\theta^t) - g_i(t)$ are uncorrelated across the nodes in the network.[2] In statistical applications, for example, each worker may own independent data or receive streaming data from independent sources; more generally, each worker can simply receive independent samples $z_i \sim P$. We also set $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, and observe

$$\mathbb{E}\left\| \sum_{i=1}^n \lambda_i \nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i)) \right\|_2^2 = \sum_{i=1}^n \lambda_i^2 \mathbb{E}\left\| \nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i)) \right\|_2^2.$$

This gives the following corollary to Theorems 6.3 and 6.4.

Corollary 6.3. *Set $\lambda_i = \frac{1}{n}$ for all i, $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, and $\eta(t) = \sigma\sqrt{t+\tau}/R\sqrt{n}$. Let $\bar{\tau}$ and $\overline{\tau^2}$ denote the average of the delays $\tau(i)$ and $\tau(i)^2$, respectively. Under the conditions of Theorem 6.3 or 6.4,*

$$\mathbb{E}\left[\sum_{t=1}^T f(\theta^{t+1}) - T f(\theta^*)\right] = \mathcal{O}\left(LR^2 + \bar{\tau}GR + \frac{LG^2R^2n\overline{\tau^2}}{\sigma^2}\log T + \frac{R\sigma}{\sqrt{n}}\sqrt{T}\right).$$

The $\log T$ multiplier can be reduced to a constant if we set $\eta(t) \equiv \sigma\sqrt{T}/R\sqrt{n}$. By using the averaged sequence $\overline{\theta}(T)$ (6.5), Jensen's inequality gives that asymptotically $\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}(1/\sqrt{Tn})$, which is an optimal dependence on the number of samples z calculated by the method. We also observe in this architecture, the delay τ is bounded by the graph

[2] Similar results continue to hold under weak correlation.

diameter D, giving us the bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} f(\theta^{t+1}) - Tf(\theta^*)\right] = \mathcal{O}\left(LR^2 + DGR + \frac{LG^2R^2nD^2}{\sigma^2}\log T + \frac{R\sigma}{\sqrt{n}}\sqrt{T}\right). \tag{6.13}$$

The above corollaries are general and hold irrespective of the relative costs of communication and computation. However, with knowledge of the costs, we can adapt the stepsizes slightly to give better rates of convergence when n is large or communication to the master node is expensive. For now, we focus on the cyclic architecture (the setting of Corollary 6.2), though the same principles apply to the local averaging scheme. Let C denote the cost of communicating between the master and workers in terms of the time to compute a single gradient sample, and assume that we set $m = Cn$, so that no worker node has idle time. For simplicity, we let the delay be non-random, so $B = \tau = n$. Consider the choice $\eta(t) = \eta\sqrt{T/(Cn)}$ for the damping stepsizes, where $\eta \geq 1$. This setting in Theorem 6.1 gives

$$\mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\frac{Cn^3}{\eta^2 T} + \frac{\eta}{\sqrt{TCn}} + \frac{1}{\eta\sqrt{TCn}}\right) = \mathcal{O}\left(\frac{Cn^3}{\eta^2 T} + \frac{\eta}{\sqrt{TCn}}\right),$$

where the last equality follows since $\eta \geq 1$. Optimizing for η on the right yields

$$\eta = \max\left\{\frac{n^{7/6}C^{1/2}}{T^{1/6}}, 1\right\} \quad \text{and} \quad \mathbb{E}[f(\overline{\theta}(T))] - f(\theta^*) = \mathcal{O}\left(\min\left\{\frac{n^{2/3}}{T^{2/3}}, \frac{n^3}{T}\right\} + \frac{1}{\sqrt{TCn}}\right). \tag{6.14}$$

The convergence rates thus follow two regimes. When $T \leq n^7C^3$, we have convergence rate $\mathcal{O}(n^{2/3}/T^{2/3})$, while once $T > n^7C^3$, we attain $\mathcal{O}(1/\sqrt{TCn})$ convergence. Roughly, in time proportional to TC, we achieve optimization error $1/\sqrt{TCn}$, which is order-optimal given that we can compute a total of TCn stochastic gradients [6]. The scaling of this bound is nicer than that previously: the dependence on network size is at worst $n^{2/3}$, which we obtain by increasing the damping factor $\eta(t)$—and hence decreasing the stepsize $\alpha(t) = 1/(L + \eta(t))$—relative to the setting of Corollary 6.2. We remark that applying the same technique to Corollary 6.3 gives convergence rate scaling as the smaller of $\mathcal{O}((D/T)^{2/3} + 1/\sqrt{TCn})$ and $\mathcal{O}((nCD^2/T + 1/\sqrt{TCn})$. Since the diameter $D \leq n$, this is faster than the cyclic architecture's bound (6.14).

6.4.2 Running-time comparisons

Having derived the rates of convergence of the different distributed procedures above, we now explicitly study the running times of the centralized stochastic gradient algorithms (6.3) and (6.4), the cyclic delayed protocol with the updates (6.8) and (6.9), and the locally averaged architecture with the updates (6.11) and (6.12). To make comparisons more cleanly, we avoid constants, assuming without loss that the variance bound σ^2 on $\mathbb{E}\left\|\nabla f(\theta) - \nabla F(\theta; z)\right\|^2$ is 1, and that sampling $z \sim P$ and evaluating $\nabla F(\theta; z)$ requires one unit of time. Noting

Centralized (6.3, 6.4)	$\mathbb{E}f(\overline{\theta}) - f(\theta^*) = \mathcal{O}\left(\sqrt{\frac{1}{T}}\right)$
Cyclic (6.8, 6.9)	$\mathbb{E}f(\overline{\theta}) - f(\theta^*) = \mathcal{O}\left(\min\left(\frac{n^{2/3}}{T^{2/3}}, \frac{n^3}{T}\right) + \frac{1}{\sqrt{Tn}}\right)$
Local (6.11, 6.12)	$\mathbb{E}f(\overline{\theta}) - f(\theta^*) = \mathcal{O}\left(\min\left(\frac{D^{2/3}}{T^{2/3}}, \frac{n\overline{\tau^2}}{T}\right) + \frac{1}{\sqrt{nT}}\right)$

Table 6.1. Upper bounds on $\mathbb{E}f(\overline{\theta}) - f(\theta^*)$ for three computational architectures, where $\overline{\theta}$ is the output of each algorithm after T units of time. Each algorithm runs for the amount of time it takes a centralized stochastic algorithm to perform T iterations as in (6.15). Here D is the diameter of the network, n is the number of nodes, and $\overline{\tau^2} = \frac{1}{n}\sum_{i=1}^{n}\tau(i)^2$ is the average squared communication delay for the local averaging architecture. Bounds for the cyclic architecture assume delay $\tau = n$.

that $\mathbb{E}[\nabla F(\theta; z)] = \nabla f(\theta)$, it is clear that if we receive m uncorrelated samples of z, the variance $\mathbb{E}\|\nabla f(\theta) - \frac{1}{m}\sum_{j=1}^{m}\nabla F(\theta; z_j)\|_2^2 \leq \frac{1}{m}$.

Now we state our assumptions on the relative times used by each algorithm. Let T be the number of units of time allocated to each algorithm, and let the centralized, cyclic delayed and locally averaged delayed algorithms complete T_{cent}, T_{cycle} and T_{dist} iterations, respectively, in time T. It is clear that $T_{\text{cent}} = T$. We assume that the distributed methods use m_{cycle} and m_{dist} samples of $z \sim P$ to compute stochastic gradients and that the delay τ of the cyclic algorithm is n. For concreteness, we assume that communication is of the same order as computing the gradient of one sample $\nabla F(\theta; z)$ so that $C = 1$. In the cyclic setup of Sec. 6.3.1, it is reasonable to assume that $m_{\text{cycle}} = \Omega(n)$ to avoid idling of workers (Theorems 6.1 and 6.2, as well as the bound (6.14), show it is asymptotically beneficial to have m_{cycle} larger, since $\sigma^2_{\text{cycle}} = 1/m_{\text{cycle}}$). For $m_{\text{cycle}} = \Omega(n)$, the master requires $\frac{m_{\text{cycle}}}{n}$ units of time to receive one gradient update, so $\frac{m_{\text{cycle}}}{n}T_{\text{cycle}} = T$. In the locally delayed framework, if each node uses m_{dist} samples to compute a gradient, the master receives a gradient every m_{dist} units of time, and hence $m_{\text{dist}}T_{\text{dist}} = T$. Further, $\sigma^2_{\text{dist}} = 1/m_{\text{dist}}$. We summarize our assumptions by saying that in T units of time, each algorithm performs the following number of iterations:

$$T_{\text{cent}} = T, \quad T_{\text{cycle}} = \frac{Tn}{m_{\text{cycle}}}, \quad \text{and} \quad T_{\text{dist}} = \frac{T}{m_{\text{dist}}}. \tag{6.15}$$

Plugging the above iteration counts into the earlier bound (6.7) and Corollaries 6.2 and 6.3 via the sharper result (6.14), we can provide upper bounds (to constant factors) on the expected optimization accuracy after T units of time for each of the distributed architectures as in Table 6.1. Asymptotically in the number of units of time T, both the cyclic and locally communicating stochastic optimization schemes have the same convergence rate. However, topological considerations show that the locally communicating method (Figs. 6.2 and 6.3) has better performance than the cyclic architecture, though it requires more worker coordi-

nation. Since the lower order terms matter only for large n or small T, we compare the terms $n^{2/3}/T^{2/3}$ and $D^{2/3}/T^{2/3}$ for the cyclic and locally averaged algorithms, respectively. Since $D \leq n$ for any network, the locally averaged algorithm always guarantees better performance than the cyclic algorithm. For specific graph topologies, however, we can quantify the time improvements:

- n-node cycle or path: $D = n$ so that both methods have the same convergence rate.
- $\sqrt{n}$-by-$\sqrt{n}$ grid: $D = \sqrt{n}$, so the distributed method has a factor of $n^{2/3}/n^{1/3} = n^{1/3}$ improvement over the cyclic architecture.
- Balanced trees and expander graphs: $D = \mathcal{O}(\log n)$, so the distributed method has a factor—ignoring logarithmic terms—of $n^{2/3}$ improvement over cyclic.

Naturally, it is possible to modify our assumptions. In a network in which communication is cheap, or conversely, in a problem for which the computation of $\nabla F(\theta; z)$ is more expensive than communication, then the number of samples $z \sim P$ for which which each worker computes gradients is small. Such problems occur frequently, such as learning conditional random field models for natural language processing, computational biology, or other applications [94]. In this case, it is reasonable to have $m_{\text{cycle}} = \mathcal{O}(1)$, in which case $T_{\text{cycle}} = Tn$ and the cyclic delayed architecture has stronger convergence guarantees of $\mathcal{O}(\min\{n^2/T, 1/T^{2/3}\} + 1/\sqrt{Tn})$. In any case, both non-centralized protocols enjoy significant asymptotically faster convergence rates for stochastic optimization problems in spite of asynchronous delays.

6.5 Numerical Results

Though this work focuses mostly on the theoretical analysis of the methods we have presented, it is important to understand the practical aspects of the above methods in solving real-world tasks and problems with real data. To that end, we use the cyclic delayed method (6.11) to solve a logistic regression problem:

$$\underset{x}{\text{minimize}} \;\; f(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i \langle x_i, \theta \rangle)) \quad \text{subject to} \;\; \|\theta\|_2 \leq R. \tag{6.16}$$

We use the Reuters RCV1 dataset [101], which consists of $N \approx 800000$ news articles, each labeled with some combination of the four labels economics, government, commerce, and medicine. In the above example, the vectors $x_i \in \{0, 1\}^d$, $d \approx 10^5$, are feature vectors representing the words in each article, and the labels b_i are 1 if the article is about government, -1 otherwise.

We simulate the cyclic delayed optimization algorithm (6.8) for the problem (6.16) for several choices of the number of workers n and the number of samples m computed at each

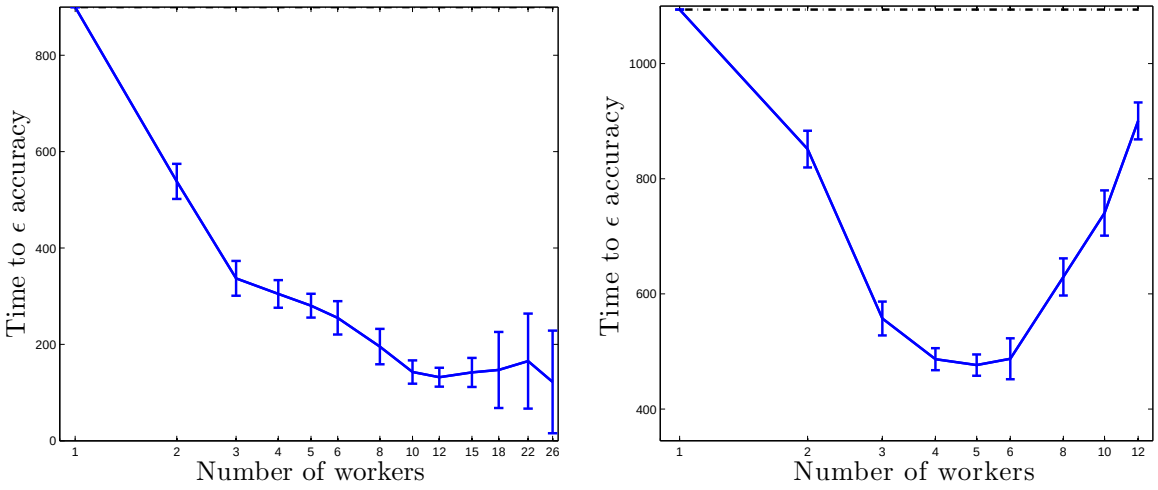

Figure 6.4. Optimization performance of the delayed cyclic method (6.8) for the Reuters RCV1 dataset. Each plot shows the estimated time to compute ϵ-accurate solution to the objective (6.16) as a function of the number of workers n. Plot (a): convergence time assuming the cost of communication to the master is the same as computing the gradient of one term in the objective (6.16). The number of samples m is equal to n for each worker. Plot (b): convergence time assuming the cost of communication to the master is 16 times as expensive as computing the gradient of one term in the objective (6.16). The number of samples m is equal to $16n$ for each worker.

worker. We summarize the results of our experiments in Figure 6.4. To generate the figure, we fix an ϵ (in this case, $\epsilon = .05$), then measure the time it takes the stochastic algorithm (6.8) to output an $\overline{\theta}$ such that $f(\overline{\theta}) \leq \inf_{\theta \in \Omega} f(\theta) + \epsilon$. We perform each experiment ten times. The two plots differ in the amount of time C required to communicate the parameters θ between the master and the workers (relative to the amount of time to compute the gradient of one term in the objective (6.16)). For the experiments exhibited in the left plot in Fig. 6.4(a), we assume that $C = 1$, while for those described in Fig. 6.4(b), we assume that $C = 16$.

We now turn to discussing the individual plots. For Fig. 6.4(a), each worker uses $m = n$ samples to compute a stochastic gradient for the objective (6.16). As mentioned above, we assume the communication cost $C = 1$ so that each worker is continuously performing computation. The plotted results show the delayed update (6.8) enjoys speedup (the ratio of time to ϵ-accuracy for an n-node system versus the centralized procedure) nearly linear in the number n of worker machines until $n \geq 15$ or so. Since we use the stepsize choice $\eta(t) \propto \sqrt{t/n}$, which yields the predicted convergence rate given by Corollary 6.2, the $n^2 m/T \approx n^3/T$ term in the convergence rate presumably becomes non-negligible for larger n. This expands on earlier experimental work with a similar method [97], which experimentally demonstrated linear speedup for small values of n, but did not investigate larger network sizes.

In Fig. 6.4(b), we study the effects of more costly communication by assuming that communication is $C = 16$ times more expensive than gradient computation. As we recommend in our discussion of convergence rates for distributed minimization in Sec. 6.4.1, we thus set the number of samples each worker computes to $m = Cn = 16n$ and correspondingly reduce the damping stepsize $\eta(t) \propto \sqrt{t/(Cn)}$. In the regime of more expensive communication—as our theoretical results predict—small numbers of workers still enjoy significant speedups over a centralized method. For non-asymptotic regimes with moderate to large numbers of workers, the cost of communication and delays mitigate some of the benefits of parallelization. Nevertheless, as our analysis shows, allowing delayed and asynchronous updates still gives significant performance improvements. We remark that the alternate choice of stepsize yielding the rates (6.14) gives qualitatively similar performance.

6.6 Delayed Updates for Smooth Optimization

In this section, we prove Theorems 6.1 and 6.2. We collect in Appendix D a few technical results relevant to our proof; we will refer to results therein without comment. Before proving either theorem, we state the lemma that is the key to our argument. Lemma 6.1 shows that certain gradient-differencing terms are essentially of second order. As a consequence, when we combine the results of the lemma with Lemma D.3, which bounds $\mathbb{E}\left[\left\|\theta^t - \theta^{(t+\tau)}\right\|^2\right]$, the gradient differencing terms become $\mathcal{O}(\log T)$ for step size choice $\eta(t) \propto \sqrt{t}$, or $\mathcal{O}(1)$ for $\eta(t) \equiv \eta\sqrt{T}$.

Lemma 6.1. *Let assumptions A and B on the function f and the compactness assumption C hold. Then for any sequence θ^t*

$$\sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle \le \frac{L}{2} \sum_{t=\tau+1}^{T} \left\|\theta^{(t-\tau)} - \theta^{t+1}\right\|^2 + 4\tau GR.$$

Consequently, if $\mathbb{E}[\|\theta^t - \theta^{t+1}\|^2] \le \kappa(t)^2 G^2$ for a non-increasing sequence $\kappa(t)$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle\right] \le \frac{LG^2(\tau+1)^2}{2} \sum_{t=\tau+1}^{T} \kappa(t-\tau)^2 + 4\tau GR.$$

Proof. The proof follows by using a few Bregman divergence identities to rewrite the left hand side of the above equations, then recognizing that the result is close to a telescoping sum. Recalling the definition of a Bregman divergence (6.2), we note the following well-known four term equality, a consequence of straightforward algebra: for any a, b, c, d,

$$\langle \nabla f(a) - \nabla f(b), c - d \rangle = D_f(d, a) - D_f(d, b) - D_f(c, a) + D_f(c, b). \tag{6.17}$$

Using the equality (6.17), we see that for $t \geq \tau$,

$$\begin{aligned}&\left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle \\ &= D_f(\theta^*, \theta^t) - D_f(\theta^*, \theta^{(t-\tau)}) - D_f(\theta^{t+1}, \theta^t) + D_f(\theta^{t+1}, \theta^{(t-\tau)}). \end{aligned} \tag{6.18}$$

To make the equality (6.18) useful, we note that the Lipschitz continuity of ∇f implies

$$f(\theta^{t+1}) \leq f(\theta^{(t-\tau)}) + \left\langle \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^{(t-\tau)} \right\rangle + \frac{L}{2} \left\| \theta^{(t-\tau)} - \theta^{t+1} \right\|^2$$

so that recalling the definition (6.2) of D_f we have

$$D_f(\theta^{t+1}, \theta^{(t-\tau)}) \leq \frac{L}{2} \left\| \theta^{(t-\tau)} - \theta^{t+1} \right\|^2.$$

In particular, using the non-negativity of $D_f(\theta, \tilde{\theta})$, we can replace (6.18) with the bound

$$\left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle \leq D_f(\theta^*, \theta^t) - D_f(\theta^*, \theta^{(t-\tau)}) + \frac{L}{2} \left\| \theta^{(t-\tau)} - \theta^{t+1} \right\|^2.$$

For $t \leq \tau$, we have $\left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle \leq 2\tau GR$ using the compactness and Lipschitz assumptions. Summing the above two inequalities, we see that

$$\begin{aligned}&\sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle \\ &= \sum_{t=1}^{\tau} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle + \sum_{t=\tau+1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle \\ &\leq 2\tau GR + \sum_{t=T-\tau+1}^{T} D_f(\theta^*, \theta^t) + \frac{L}{2} \sum_{t=\tau+1}^{T} \left\| \theta^{(t-\tau)} - \theta^{t+1} \right\|^2. \end{aligned} \tag{6.19}$$

To bound the Bregman divergence term, we recall that by Assumption C and the strong convexity of ψ, $\|\theta^* - \theta^t\|^2 \leq 2D_\psi(\theta^*, \theta^t) \leq 2R^2$, and hence the optimality of θ^* implies

$$D_f(\theta^*, \theta^t) = f(\theta^*) - f(\theta^t) - \left\langle \nabla f(\theta^t), \theta^* - \theta^t \right\rangle \leq \|\nabla f(\theta^t)\|_* \left\| \theta^* - \theta^t \right\| \leq 2GR.$$

This gives the first bound of the lemma. For the second bound, using convexity, we see that

$$\left\| \theta^{(t-\tau)} - \theta^{t+1} \right\|^2 \leq (\tau+1)^2 \sum_{s=0}^{\tau} \frac{1}{\tau+1} \left\| \theta^{(t-s)} - \theta^{(t-s+1)} \right\|^2,$$

so by taking expectations we have $\mathbb{E}[\left\| \theta^t - \theta^{(t+\tau+1)} \right\|^2] \leq (\tau+1)^2 \kappa(t-\tau)^2 G^2$. Since κ is non-increasing (by the definition of the update scheme) we see that the sum (6.19) is further bounded by $4\tau GR + \frac{L}{2} \sum_{t=\tau+1}^{T} G^2(\tau+1)^2 \kappa(t-\tau)^2$ as desired. □

6.6.1 Proof of Theorem 6.1

The essential idea in this proof is to use convexity and smoothness to bound $f(\theta^t) - f(\theta^*)$, then use the sequence $\{\eta(t)\}$, which decreases the stepsize $\alpha(t)$, to cancel variance terms. We will roughly ignore the first τ terms in the sequence $f(\theta^t) - f(\theta^*)$, which reduce to an $\mathcal{O}(\tau/T)$ term in the final rate. To begin, we define the error $e(t)$

$$e(t) := \nabla f(\theta^t) - g(t-\tau)$$

where $g(t-\tau) = \nabla F(\theta^{(t-\tau)}; z(t))$ for some $z(t) \sim P$. Note that $e(t)$ in general has non-zero expectation, as there is a time delay.

By using the convexity of f and then the L-Lipschitz continuity of ∇f, for any $\theta^* \in \Omega$, we have

$$\begin{aligned} f(\theta^t) - f(\theta^*) &\leq \left\langle \nabla f(\theta^t), \theta^t - \theta^* \right\rangle = \left\langle \nabla f(\theta^t), \theta^{t+1} - \theta^* \right\rangle + \left\langle \nabla f(\theta^t), \theta^t - \theta^{t+1} \right\rangle \\ &\leq \left\langle \nabla f(\theta^t), \theta^{t+1} - \theta^* \right\rangle + f(\theta^t) - f(\theta^{t+1}) + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2, \end{aligned}$$

so that using the definition (6.8) of μ^t,

$$\begin{aligned} &f(\theta^{t+1}) - f(\theta^*) \leq \left\langle \nabla f(\theta^t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 \\ &= \left\langle g(t-\tau), \theta^{t+1} - \theta^* \right\rangle + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 \\ &= \left\langle \mu^{t+1}, \theta^{t+1} - \theta^* \right\rangle - \left\langle \mu^t, \theta^{t+1} - \theta^* \right\rangle + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2. \end{aligned}$$

Applying Lemma D.1 in Appendix D and the definition of the update (6.8), we see that

$$-\left\langle \mu^t, \theta^{t+1} - \theta^* \right\rangle \leq -\left\langle \mu^t, \theta^t - \theta^* \right\rangle + \frac{1}{\alpha(t)} \left[\psi(\theta^{t+1}) - \psi(\theta^t) \right] - \frac{1}{\alpha(t)} D_\psi(\theta^{t+1}, \theta^t),$$

which implies

$$\begin{aligned} &f(\theta^{t+1}) - f(\theta^*) \\ &\leq \left\langle \mu^{t+1}, \theta^{t+1} - \theta^* \right\rangle - \left\langle \mu^t, \theta^t - \theta^* \right\rangle + \frac{1}{\alpha(t)} [\psi(\theta^{t+1}) - \psi(\theta^t)] \\ &\quad - L D_\psi(\theta^{t+1}, \theta^t) - \eta(t) D_\psi(\theta^{t+1}, \theta^t) + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle \\ &\leq \left\langle \mu^{t+1}, \theta^{t+1} - \theta^* \right\rangle - \left\langle \mu^t, \theta^t - \theta^* \right\rangle + \frac{1}{\alpha(t)} [\psi(\theta^{t+1}) - \psi(\theta^t)] \\ &\quad - \eta(t) D_\psi(\theta^{t+1}, \theta^t) + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle. \end{aligned} \tag{6.20}$$

To get the bound (6.20), we substituted $\alpha(t)^{-1} = L + \eta(t)$ and then used the fact that ψ is strongly convex, so $D_\psi(\theta^{t+1}, \theta^t) \geq \frac{1}{2}\|\theta^t - \theta^{t+1}\|^2$. By summing the bound (6.20), we have the following non-probabilistic inequality:

$$\begin{aligned}
&\sum_{t=1}^{T} f(\theta^{t+1}) - f(\theta^*) \\
&\leq \left\langle \mu^{T+1}, \theta^{T+1} - \theta^* \right\rangle + \frac{1}{\alpha(T)}\psi(\theta^{T+1}) + \sum_{t=1}^{T} \psi(\theta^t)\left[\frac{1}{\alpha(t-1)} - \frac{1}{\alpha(t)}\right] \\
&\quad - \sum_{t=1}^{T} \eta(t) D_\psi(\theta^{t+1}, \theta^t) + \sum_{t=1}^{T} \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle \\
&\leq \frac{1}{\alpha(T+1)}\psi(\theta^*) + \sum_{t=1}^{T} \psi(\theta^t)\left[\frac{1}{\alpha(t-1)} - \frac{1}{\alpha(t)}\right] - \sum_{t=1}^{T} \eta(t) D_\psi(\theta^{t+1}, \theta^t) \\
&\quad + \sum_{t=1}^{T} \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle
\end{aligned} \tag{6.21}$$

since $\psi(\theta) \geq 0$ and θ^{T+1} minimizes $\left\langle \mu^{T+1}, \theta \right\rangle + \frac{1}{\alpha(T+1)}\psi(\theta)$. What remains is to control the summed $e(t)$ terms in the bound (6.21). We can do this simply using the second part of Lemma 6.1. Indeed, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle \\
&= \sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle + \sum_{t=1}^{T} \left\langle \nabla f(\theta^{(t-\tau)}) - g(t-\tau), \theta^{t+1} - \theta^* \right\rangle .
\end{aligned} \tag{6.22}$$

We can apply Lemma 6.1 to the first term in (6.22) by bounding $\|\theta^t - \theta^{t+1}\|$ with Lemma D.3. Since $\eta(t) \propto \sqrt{t}$, Lemma D.3 implies $\mathbb{E}[\|\theta^t - \theta^{t+1}\|^2] \leq \frac{4G^2}{\eta(t)^2}$. As a consequence,

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau)}), \theta^{t+1} - \theta^* \right\rangle\right] \leq 4\tau GR + 2L(\tau+1)^2 G^2 \sum_{t=\tau+1}^{T} \frac{1}{\eta(t-\tau)^2}.$$

What remains is to bound the stochastic (second) term in (6.22). This is straightforward:

$$\begin{aligned}
&\left\langle \nabla f(\theta^{(t-\tau)}) - g(t-\tau), \theta^{t+1} - \theta^* \right\rangle \\
&= \left\langle \nabla f(\theta^{(t-\tau)}) - g(t-\tau), \theta^t - \theta^* \right\rangle + \left\langle \nabla f(\theta^{(t-\tau)}) - g(t-\tau), \theta^{t+1} - \theta^t \right\rangle \\
&\leq \left\langle \nabla f(\theta^{(t-\tau)}) - g(t-\tau), \theta^t - \theta^* \right\rangle + \frac{1}{2\eta(t)}\|\nabla f(\theta^{(t-\tau)}) - g(t-\tau)\|_*^2 + \frac{\eta(t)}{2}\left\|\theta^{t+1} - \theta^t\right\|^2
\end{aligned}$$

by the Fenchel-Young inequality applied to the conjugate pair $\frac{1}{2}\|\cdot\|_*^2$ and $\frac{1}{2}\|\cdot\|^2$. In addition, $\nabla f(\theta^{(t-\tau)}) - g(t-\tau)$ is independent of θ^t given the sigma-field containing $g(1), \ldots, g(t-\tau-1)$, since θ^t is a function of gradients to time $t-\tau-1$, so the first term has zero expectation. Also recall that $\mathbb{E}[\|\nabla f(\theta^{(t-\tau)}) - g(t-\tau)\|_*]^2$ is bounded by σ^2 by assumption. Combining the above two bounds into (6.22), we see that

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\langle e(t), \theta^{t+1} - \theta^* \rangle] \\
&\leq \frac{\sigma^2}{2} \sum_{t=1}^{T} \frac{1}{\eta(t)} + \frac{1}{2} \sum_{t=1}^{T} \eta(t) \left\|\theta^{t+1} - \theta^t\right\|^2 + 2LG^2(\tau+1)^2 \sum_{t=\tau+1}^{T} \frac{1}{\eta(t-\tau)^2} + 4\tau GR. \qquad (6.23)
\end{aligned}$$

Since $D_\psi(\theta^{t+1}, \theta^t) \geq \frac{1}{2}\|\theta^t - \theta^{t+1}\|^2$, combining (6.23) with (6.21) and noting the two facts that $\frac{1}{\alpha(t-1)} - \frac{1}{\alpha(t)} \leq 0$ and $\sum_{t=\tau+1}^{T} \eta(t-\tau)^{-2} \leq \sum_{t=1}^{T} \eta(t)^{-2}$ gives

$$\sum_{t=1}^{T} \mathbb{E} f(\theta^{t+1}) - f(\theta^*) \leq \frac{1}{\alpha(T+1)} \psi(\theta^*) + \frac{\sigma^2}{2} \sum_{t=1}^{T} \frac{1}{\eta(t)} + 2LG^2(\tau+1)^2 \sum_{t=1}^{T} \frac{1}{\eta(t)^2} + 4\tau GR.$$

6.6.2 Proof of Theorem 6.2

The proof of Theorem 6.2 is similar to that of Theorem 6.1, so we will be somewhat terse. We define the error $e(t) = \nabla f(\theta^t) - g(t-\tau)$, identically as in the earlier proof, and begin as we did in the proof of Theorem 6.1. Recall that for $t \geq \tau$,

$$f(\theta^{t+1}) - f(\theta^*) \leq \langle g(t-\tau), \theta^{t+1} - \theta^* \rangle + \langle e(t), \theta^{t+1} - \theta^* \rangle + \frac{L}{2} \left\|\theta^t - \theta^{t+1}\right\|^2. \qquad (6.24)$$

Applying the first-order optimality condition to the definition of θ^{t+1} (6.4), we get

$$\langle \alpha(t) g(t-\tau) + \nabla\psi(\theta^{t+1}) - \nabla\psi(\theta^t), \theta - \theta^{t+1} \rangle \geq 0$$

for all $\theta \in \Omega$. In particular, we have

$$\begin{aligned}
\alpha(t) \langle g(t-\tau), \theta^{t+1} - \theta^* \rangle &\leq \langle \nabla\psi(\theta^{t+1}) - \nabla\psi(\theta^t), \theta^* - \theta^{t+1} \rangle \\
&= D_\psi(\theta^*, \theta^t) - D_\psi(\theta^*, \theta^{t+1}) - D_\psi(\theta^{t+1}, \theta^t).
\end{aligned}$$

Applying the above to the inequality (6.24), we see

$$\begin{aligned}
f(\theta^{t+1}) - f(\theta^*) &\leq \frac{1}{\alpha(t)} \left[D_\psi(\theta^*, \theta^t) - D_\psi(\theta^*, \theta^{t+1}) - D_\psi(\theta^{t+1}, \theta^t) \right] \\
&\qquad + \langle e(t), \theta^{t+1} - \theta^* \rangle + \frac{L}{2} \left\|\theta^t - \theta^{t+1}\right\|^2 \\
&\leq \frac{1}{\alpha(t)} \left[D_\psi(\theta^*, \theta^t) - D_\psi(\theta^*, \theta^{t+1}) \right] + \langle e(t), \theta^{t+1} - \theta^* \rangle - \eta(t) D_\psi(\theta^{t+1}, \theta^t) \qquad (6.25)
\end{aligned}$$

where for the last inequality, we use the fact that $D_\psi(\theta^{t+1}, \theta^t) \geq \frac{1}{2}\left\|\theta^t - \theta^{t+1}\right\|^2$, by the strong convexity of ψ, and that $\alpha(t)^{-1} = L + \eta(t)$. By summing the inequality (6.25), we have

$$\sum_{t=1}^{T} f(\theta^{t+1}) - f(\theta^*) \leq \frac{1}{\alpha(1)} D_\psi(\theta^*, \theta^1) + \sum_{t=2}^{T} D_\psi(\theta^*, \theta^t)\left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)}\right] - \sum_{t=1}^{T} \eta(t) D_\psi(\theta^{t+1}, \theta^t) + \sum_{t=1}^{T} \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle. \tag{6.26}$$

Comparing the bound (6.26) with the earlier bound for the dual averaging algorithms (6.21), we see that the only essential difference is the $\alpha(t)^{-1} - \alpha(t-1)^{-1}$ terms. The compactness assumption guarantees that $D_\psi(\theta^*, \theta^t) \leq R^2$, however, so

$$\frac{1}{\alpha(1)} D_\psi(\theta^*, \theta^1) + \sum_{t=2}^{T} D_\psi(\theta^*, \theta^t)\left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)}\right] \leq \frac{R^2}{\alpha(1)} + \sum_{t=2}^{T} R^2\left[\frac{1}{\alpha(t)} - \frac{1}{\alpha(t-1)}\right] = \frac{R^2}{\alpha(T)}.$$

The remainder of the proof uses Lemmas D.3 and 6.1 completely identically to the proof of Theorem 6.1.

6.6.3 Proof of Corollary 6.1

We prove this result only for the mirror descent algorithm (6.9), as the proof for the dual-averaging algorithm (6.8) is similar. We define the error at time t to be $e(t) = \nabla f(\theta^t) - g(t - \tau(t))$, and observe that we only need to control the second term involving $e(t)$ in the bound (6.25) differently. Expanding the error terms above and using Fenchel's inequality as in the proofs of Theorems 6.1 and 6.2, we have

$$\begin{aligned}
&\left\langle e(t), \theta^{t+1} - \theta^* \right\rangle \\
&\leq \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau(t))}), \theta^{t+1} - \theta^* \right\rangle + \left\langle \nabla f(\theta^{(t-\tau(t))}) - g(t - \tau(t)), \theta^t - \theta^* \right\rangle \\
&\quad + \frac{1}{2\eta(t)} \|\nabla f(\theta^{(t-\tau(t))}) - g(t - \tau(t))\|_*^2 + \frac{\eta(t)}{2}\left\|\theta^{t+1} - \theta^t\right\|^2,
\end{aligned}$$

Now we note that conditioned on the delay $\tau(t)$, we have

$$\mathbb{E}[\left\|\theta^{(t-\tau(t))} - \theta^{t+1}\right\|^2 \mid \tau(t)] \leq G^2(\tau(t)+1)^2 \alpha(t - \tau(t))^2.$$

Consequently we apply Lemma 6.1 (specifically, following the bounds (6.18) and (6.19)) and find

$$\begin{aligned}
&\sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau(t))}), \theta^{t+1} - \theta^* \right\rangle \\
&\leq \sum_{t=1}^{T} \left[D_f(\theta^*, \theta^t) - D_f(\theta^*, \theta^{(t-\tau(t))})\right] + G^2 \sum_{t=1}^{T} (\tau(t)+1)^2 \alpha(t - \tau(t))^2.
\end{aligned}$$

The sum of D_f terms telescopes, leaving only terms not received by the gradient procedure within T iterations, and we can use $\alpha(t) \leq \frac{1}{\eta\sqrt{T}}$ for all t to derive the further bound

$$\sum_{t:t+\tau(t)>T} D_f(\theta^*, \theta^t) + \frac{G^2}{\eta^2 T} \sum_{t=1}^{T} (\tau(t)+1)^2. \tag{6.27}$$

To control the quantity (6.27), all we need is to bound the expected cardinality of the set $\{t \in [T] : t+\tau(t) > T\}$. Using Chebyshev's inequality and standard expectation bounds, we have

$$\mathbb{E}\left[\text{card}(\{t \in [T] : t + \tau(t) > T\})\right] = \sum_{t=1}^{T} \mathbb{P}(t + \tau(t) > T) \leq 1 + \sum_{t=1}^{T-1} \frac{\mathbb{E}[\tau(t)^2]}{(T-t)^2} \leq 1 + 2B^2,$$

where the last inequality comes from our assumption that $\mathbb{E}[\tau(t)^2] \leq B^2$. As in Lemma 6.1, we have $D_f(\theta^*, \theta^t) \leq 2GR$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau(t))}), \theta^{t+1} - \theta^* \right\rangle\right] \leq 6GRB^2 + \frac{G^2(B+1)^2}{\eta^2}$$

We can control the remaining terms as in the proofs of Theorems 6.1 and 6.2.

6.7 Proof of Theorem 6.3

The proof of Theorem 6.3 is not too difficult given our previous work—all we need to do is redefine the error $e(t)$ and use $\eta(t)$ to control the variance terms that arise. To that end, we define the gradient error terms that we must control. In this proof, we set

$$e(t) := \nabla f(\theta^t) - \sum_{i=1}^{n} \lambda_i g_i(t - \tau(i)) \tag{6.28}$$

where $g_i(t) = \nabla f(\theta^t; z_i(t))$ is the gradient of node i computed at the parameter θ^t and $\tau(i)$ is the delay associated with node i.

Using Assumption B as in the proofs of previous theorems, then applying Lemma D.1,

we have

$$\begin{aligned}
f(\theta^{t+1}) - f(\theta^*) &\leq \left\langle \nabla f(\theta^t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 \\
&= \left\langle \sum_{i=1}^n \lambda_i g_i(t - \tau(i)), \theta^{t+1} - \theta^* \right\rangle + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 \\
&= \left\langle \mu^{t+1}, \theta^{t+1} - \theta^* \right\rangle - \left\langle \mu^t, \theta^{t+1} - \theta^* \right\rangle + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 \\
&\leq \left\langle \mu^{t+1}, \theta^{t+1} - \theta^* \right\rangle - \left\langle \mu^t, \theta^t - \theta^* \right\rangle + \frac{1}{\alpha(t)} \psi(\theta^{t+1}) - \frac{1}{\alpha(t)} \psi(\theta^t) \\
&\quad - \frac{1}{\alpha(t)} D_\psi(\theta^{t+1}, \theta^t) + \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle + \frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2 .
\end{aligned}$$

We telescope as in the proofs of Theorems 6.1 and 6.2, canceling $\frac{L}{2} \left\| \theta^t - \theta^{t+1} \right\|^2$ with the LD_ψ divergence terms to see that

$$\begin{aligned}
&\sum_{t=1}^T f(\theta^{t+1}) - f(\theta^*) \\
&\leq \left\langle \mu^{T+1}, \theta^{T+1} - \theta^* \right\rangle + \frac{1}{\alpha(T)} \psi(\theta^T) - \sum_{t=1}^T \eta(t) D_\psi(\theta^{t+1}, \theta^t) + \sum_{t=1}^T \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle \\
&\leq \frac{1}{\alpha(T+1)} \psi(\theta^*) - \sum_{t=1}^T \eta(t) D_\psi(\theta^{t+1}, \theta^t) + \sum_{t=1}^T \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle . \qquad (6.29)
\end{aligned}$$

This is exactly as in the non-probabilistic bound (6.21) from the proof of Theorem 6.1, but the definition (6.28) of the error $e(t)$ here is different.

What remains is to control the error term in (6.29). Writing the terms out, we have

$$\begin{aligned}
\sum_{t=1}^T \left\langle e(t), \theta^{t+1} - \theta^* \right\rangle = \sum_{t=1}^T \Bigg\langle \nabla f(\theta^t) - \sum_{i=1}^n \lambda_i \nabla f(\theta^{(t-\tau(i))}), \theta^{t+1} - \theta^* \Bigg\rangle \\
+ \sum_{t=1}^T \Bigg\langle \sum_{i=1}^n \lambda_i \left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t - \tau(i)) \right], \theta^{t+1} - \theta^* \Bigg\rangle \qquad (6.30)
\end{aligned}$$

Bounding the first term above is simple via Lemma 6.1: as in the proof of Theorem 6.1, we

have

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T}\Big\langle \nabla f(\theta^t) - \sum_{i=1}^{n}\lambda_i \nabla f(\theta^{(t-\tau(i))}), \theta^{t+1}-\theta^*\Big\rangle\Big] \\
&= \sum_{i=1}^{n}\lambda_i \sum_{t=1}^{T}\mathbb{E}[\langle \nabla f(\theta^t) - \nabla f(\theta^{(t-\tau(i))}), \theta^{t+1}-\theta^*\rangle] \\
&\le 2\sum_{i=1}^{n}\lambda_i L G^2(\tau(i)+1)^2 \sum_{t=1}^{T}\frac{1}{\eta(t)^2} + 4\sum_{i=1}^{n}\lambda_i \tau(i) GR.
\end{aligned}$$

We use the same technique as the proof of Theorem 6.1 to bound the second term from (6.30). Indeed, the Fenchel-Young inequality gives

$$\begin{aligned}
&\Big\langle \sum_{i=1}^{n}\lambda_i\left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))\right], \theta^{t+1}-\theta^*\Big\rangle \\
&= \Big\langle \sum_{i=1}^{n}\lambda_i\left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))\right], \theta^{t}-\theta^*\Big\rangle \\
&\quad + \Big\langle \sum_{i=1}^{n}\lambda_i\left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))\right], \theta^{t+1}-\theta^t\Big\rangle \\
&\le \Big\langle \sum_{i=1}^{n}\lambda_i\left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))\right], \theta^{t}-\theta^*\Big\rangle \\
&\quad + \frac{1}{2\eta(t)}\Big\|\sum_{i=1}^{n}\lambda_i\left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))\right]\Big\|_*^2 + \frac{\eta(t)}{2}\left\|\theta^{t+1}-\theta^t\right\|^2.
\end{aligned}$$

By assumption, given the information at worker i at time $t-\tau(i)$, $g_i(t-\tau(i)))$ is independent of θ^t, so the first term has zero expectation. More formally, this happens because θ^t is a function of gradients $g_i(1),\dots,g_i(t-\tau(i)-1)$ from each of the nodes i and hence the expectation of the first term conditioned on $\{g_i(1),\dots,g_i(t-\tau(i)-1)\}_{i=1}^n$ is 0. The last term is canceled by the Bregman divergence terms in (6.29), so combining the bound (6.30) with the above two paragraphs yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}f(\theta^{t+1}) - f(\theta^*) \le \frac{1}{\alpha(T+1)}\psi(\theta^*) &+ 2\sum_{i=1}^{n}\lambda_i L G^2(\tau(i)+1)^2\sum_{t=1}^{T}\frac{1}{\eta(t)^2} + 4\sum_{i=1}^{n}\lambda_i\tau(i)GR \\
&+ \sum_{t=1}^{T}\frac{1}{2\eta(t)}\mathbb{E}\Big\|\sum_{i=1}^{n}\lambda_i\left[\nabla f(\theta^{(t-\tau(i))}) - g_i(t-\tau(i))\right]\Big\|_*^2.
\end{aligned}$$

6.8 Conclusion and Discussion

In this chapter, we have studied dual averaging and mirror descent algorithms for smooth and non-smooth stochastic optimization in delayed settings, showing applications of our results to distributed optimization. We showed that for smooth problems, we can preserve the performance benefits of parallelization over centralized stochastic optimization even when we relax synchronization requirements. Specifically, we presented methods that take advantage of distributed computational resources and are robust to node failures, communication latency, and node slowdowns. In addition, by distributing computation for stochastic optimization problems, we were able to exploit asynchronous processing without incurring any asymptotic penalty due to the delays incurred. In addition, though we omit these results for brevity, it is possible to extend all of our expected convergence results to guarantees with high-probability.

We believe several interesting questions remain open after this work. Certainly network topologies other than those we considered are possible, and it would be interesting to compare the effects that topology has on convergence rate of similar distributed procedures. Analyzing the robustness and fail-over capacity of distributed optimization protocols could be quite interesting, as would a careful study of the effect of communication latency. We hope to address these questions in future work.

Chapter 7

Conclusions and future directions

The aim of this section is to reiterate the main themes and contributions of this thesis, and to layout a rough road map of some of the things that could likely follow directly, and not so directly, as future developments on the results presented here. We start with the easy part of summarizing the main ideas in the next section. Section 7.2 outlines some of the themes that are the natural extensions of the ideas in this thesis. Section 7.3 takes a slightly longer-term view of this research agenda, identifying other areas and directions that are relevant to the work presented here, but not necessarily direct consequences.

7.1 Summary and key contributions

With all the technical content and results in place, we are now in a position to revisit the key motivation underlining the bulk of work in this thesis, and understand how the different chapters address different aspects of the quest. As summarized pictorially in Figure 1.1, the aim of this thesis is to understand the interplay between statistical error, and the computational resources available. More concretely, we seek to understand how the statistical error changes, as we vary the number of data samples, and the amount of computation at a learner's disposal. Below, we will discuss how each chapter in this thesis relates to and advances our understanding of this question. Throughout, the notion of computation we will be using will be within the framework of convex optimization which really forms the core of this thesis.

Chapter 3 is a direct attempt at perhaps the most fundamental way of posing this question. It asks *what is the minimum achievable statistical error given a certain computational constraint?* Within the black box model of complexity for stochastic convex optimization, we present lower bounds that answer this question for various problem classes. The notion of computation here really equates samples and computation. Given a desired statistical accuracy target ϵ, we keep drawing samples from the underlying distribution, making a stochastic gradient-style update with every sample till we run out of computation. The re-

sults in Chapter 3 provide lower bounds on the smallest T that is needed by any stochastic optimization algorithm to achieve this error ϵ, for many different problem classes. In doing so, we also shed light on the various properties of a problem that influence its computational complexity, and the corresponding optimal methods. We also provide a general recipe for obtaining such results for other problems.

Chapter 4 considers the natural algorithmic counterpart of the question. Given a computational constraint, how can we design algorithms with good statistical performance that obey the computational constraint. For the problem of model selection, the main results in Chapter 4 provide direct bounds on the statistical quality of the obtained solution as a function of the computational budget. A striking feature of the results is their efficiency in competing with an omniscient oracle, incurring at most an additional logarithmic penalty in many typical applications. The results of both this chapter and the previous one really aim to understand statistical error solely as a function of the computational budget, taking the large data assumption really to an extreme and considering access to an essentially infinite sample pool.

In Chapter 5, we tip the balance against computation even more severely, considering high-dimensional settings where the number of samples is large and the number of parameters is substantially larger. While a general statistical problem with these characteristics is hopeless, of interest are well-behaved problems with structural assumptions on the data generating process. For such problems, the number of samples n and the computational time T combine beautifully to reveal striking interactions between the computational and statistical complexities. Specifically, for the gradient methods examined in this chapter, the sample size n determines the conditioning of the problem, and hence the rate of convergence. The computational time T constrains the number of iterations that the method can perform, and together the two define the statistical quality of the solution we compute. The results also reveal that while the problems might be hard to solve to a numerical precision, we can indeed rapidly solve them to a coarser accuracy related to the statistical precision of the underlying data generating model. Such a delicate interaction between the optimization error and statistical precision was previously unknown to the best of our knowledge.

Chapter 6 considers the natural framework of distributed computation for these large, high-dimensional problems. Given a distributed network of k computers, we have the ability to perform k times as many computations per unit time than a single computer. In Chapter 6, we ask how this computational gain translates into faster machine learning algorithms. The main question to ask is whether the gain of parallel computation is offset by the loss due to decentralization of data, or can we indeed obtain faster algorithms in this framework. The results in Chapter 6 yield a partial answer to the question. They develop partially asynchronous stochastic gradient algorithms, that enjoy linear speedups with the size of the network at least asymptotically. That is, the amount of time to reach a certain statistical error improves to T/k for a k-node network when T is large enough. However, the results also show that the gains might be offset, or even reversed when the communication cost starts to dominate computation. Nevertheless, the results provide an important step in extending our

understanding of computational issues from a centralized setting to distributed networks.

Overall, we see that the various chapters examine different intuitive aspects of an extremely challenging problem. This pursuit has led to the development of techniques and frameworks which are of interest in their own right. The results also naturally lead to further questions which we hope will provide a fertile source of problems in the future years. The next section examines these direct implications for future work in more depth.

7.2 Important open questions and immediate future directions

One of the important contributions of this thesis has been crystallizing some concrete questions that explain different aspects of statistical estimation in computationally constrained settings. While we present interesting results towards answering these different questions, a lot remains to be done before we can claim a complete understanding of any of them.

7.2.1 Better oracle models for time and space complexity

Towards understanding the fundamental computational complexity of statistical estimation, Section 3 takes an approach with deep roots in information theory [52], convex optimization [119] and information-based complexity [157]. The framework captures certain aspects of the complexity of stochastic first-order optimization methods, namely that of gradient computation. Looking at many commonly used stochastic first-order optimization algorithms, the two primary sources of computational complexity are gradient computation and a projection step, the latter also often referred to as computing a prox-mapping. Jointly capturing the cost of gradient computation and projection operations would bring the oracle model of complexity significantly closer to the real computational cost. In some problems, this projection step can be just a simple rescaling or recentering when the projection is on to a simple set such as a norm-ball. However, in many instances the feasible sets are significantly more complex such as a polytope defined by a large number of constraints as in graphical model inference [171], or the cone of positive semidefinite matrices in kernel and metric learning [161]. A natural complexity model for computing the prox-mapping is to assume oracle access to projections onto some simple sets such as linear constraints and norm balls, and formulate the complexity of projections onto more complex sets in terms of the number of oracle calls.

Another interesting line for future work in the setup of Chapter 3 is to impose constraints on both the time and space complexities of optimization methods. This is important, because in terms of oracle information, first-order and Quasi-Newton methods are often identical. However the memory footprint of the Quasi-Newton algorithms is often much larger since they typically store and manipulate some kind of Hessian matrix approximation. A natural

framework for understanding these issues is that of Communication complexity [92], which has been successfully applied for space lower bounds on streaming algorithms.

7.2.2 Computational budget beyond model selection

Chapter 4 introduces a computationally budgeted framework and resulting algorithms for the model selection problem. There are some natural open questions that would be direct extensions to the work, which we discuss in Section 4.5. However the general principle of using computation, rather than data samples, as the unit of allocation to a learning algorithm has appeal and applications beyond just the model selection problem. While model selection is a natural and very broadly used problem, understanding such issues in other common problems such as regression and classification within a fixed model would also be of significant interest. A natural way to formalize this would be close to the online learning framework, but more computation-centric. We consider an algorithm that receives quanta of computation incrementally. The learning algorithm has access to a finite data sample, and it uses the fresh computation to update its model based on the data. Perhaps the most interesting question in this direction is to formalize what it means to receive a quantum of computation. One natural candidate is to define a base set of operations that can be performed in unit time, and a more information-theoretic approach would be allow the computation of estimators with a bounded mutual information with the data. While a typical online learning algorithm would just pick one of more samples and repeat the same form of computation iteratively, we hope that the computational viewpoint will allow for more general algorithms.

7.2.3 Improved computational complexity under structural assumptions

Chapter 5 shows that significantly improved computational complexities are often possible by carefully considering the underlying statistical structure of the problem. This is of course now completely surprising, or hitherto unforeseen; the analysis of the perceptron algorithm[142] for solving the NP-hard problem of minimizing 0-1 loss under a margin assumption being perhaps the most classical example. Two natural questions in the setup of Chapter 5 are what other algorithms and what other problem structures enjoy similar performance gains. In answering the first question, we have recently succeeded in showing that stochastic optimization algorithms also enjoy fast convergence within the setup of Chapter 5, which results in near-linear time algorithms for many problems of interest. The second question of other problem structures remains quite open and challenging. There are some cases such as matrix decompositions [49, 8] where more general structures can be built on top of existing ones and our theory can then be applied. Beyond such cases, manifold-based assumptions have seen a large amount of work on the statistical side although the understanding is still relatively poor. On the computational side, little is known about the implications of these assumptions

beyond nearest-neighbor search [53] and other similar geometric algorithms. Other recent work has explored different tractable restrictions of these assumptions statistically [50], and extending our computational results to these frameworks would be of significant interest.

7.2.4 Communication efficient distributed algorithms with provable speedups

Chapter 6 explores the landscape of distributed algorithms for machine learning. While parallel and distributed computation has a rich tradition in the disciplines of scientific computation such as convex optimization and numerical linear algebra, the application of these to develop better machine learning algorithms is relatively recent. The theory developed in Chapter 6 points out how we can obtain substantially improved complexity for distributed algorithms by taking the statistical nature of machine learning problems into account. However, our current understanding of these issues can still be called only preliminary at best. The algorithms of Chapter 6, while being robust to asynchrony and delays to a certain degree, still require relatively cheap and frequent communication both theoretically and empirically. These assumptions can often be unrealistic in a distributed network, and a study of better algorithms that achieve a more careful balance of their computational and communication needs remains quite important in this area. A key drawback of mini-batch style algorithms explored in Chapter 6 is that the algorithm goes over a large set of examples before the parameter is updated. Online and stochastic optimization algorithms enjoy rapid initial convergence due to the frequent updates they make on the parameters. A natural question to ask is what happens if the nodes continue to update their parameter locally as they compute the stochastic gradients, rather than only updating them after finishing gradient computation on a mini-batch. Another option is a more thorough analysis of hybrid optimization schemes such as those considered in the recent manuscript [7]. It would be also of interest to extend the lower bound techniques of Chapter 3 to distributed optimization scenarios, for an understanding of fundamental limitations in these setups.

7.3 Other suggestions for future work

This thesis completely focuses on the computational framework of convex optimization for understanding trade-offs between statistical and computational complexities. This is quite natural since many machine learning problems are quite naturally cast as convex optimization, and it remains one of the most scalable computational paradigms in machine learning. However, a natural direction for future research to push our frontiers beyond this setting.

A large body of research in combinatorial optimization [144] focuses on convex relaxations of non-convex problems, and there have been successful applications in graphical model inference [171] as well as vector and matrix compressed sensing problems [51, 42, 44], and the classical perceptron algorithm [142, 125] to name a few examples. However, while the

non-convex formulations are often quite natural and abundant, successes in this area are still only numbered. For instance even considering the well-understood problem of binary classification, while the hinge-loss and logistic loss relaxations are typically effective, they can be more susceptible to outliers and missing features than their non-convex capped versions. A better understanding of structural conditions on the problem and the data that allow for efficient solutions of such non-convex problems should be an important consideration in the years to come.

Another important computational paradigm that is of use primarily with Bayesian methods is that of sampling. The chasm between theory and practice here is much wider than convex optimization. While there have been impressive breakthroughs in understanding the mixing rates of MCMC algorithms, the sampling methods that are most widely used in machine learning have little known theoretically on their mixing properties. Convex optimization provides some interesting possibilities here. There are many known connections that go from sampling methods leading to optimization algorithms [85, 29]. This leads to the natural question if convex optimization algorithms also lead to efficient sampling methods. A natural direction here is the following. Suppose there is a simple distribution that we can efficiently and provably sample from. Can we define a convex optimization problem, which takes a random vector from this simple distribution and whose optimal solution is distributed according to our desired distribution of interest?. More importantly, does an approximate optimum also have a distribution close to the desired one? An understanding of such issues would naturally lead to provably efficient sampling algorithms by building on the rich literature of convex optimization. Furthermore, by extending the literature on distributed optimization, such a development would also naturally lead to distributed algorithms for sampling.

On a more statistical side, the thesis assumed throughout that we have access to large amounts of data with no data samples having missing or noisy attributes. In practice, these large datasets are often acquired using automated techniques which combine multiple sources of information. In such settings, not all features are observed for all the examples and even when they are, their values often have varying degrees of reliability. More critically, while it is easy to acquire the covariate vector, the labels in a classification or a regression problem often require manual supervision. In such scenarios, we have access to a massive pool of unlabeled data, but the set of labeled examples is often restricted and involves monetary costs. This naturally shifts the focus to active learning and semi-supervised learning algorithms, since we want to spend our resources on acquiring the minimum number of labels with the maximum statistical impact. Our understanding of these problems has certainly improved significantly over the last decade or so, but it is nowhere as complete as supervised learning settings. Handling of missing and noisy attributes is understood to an even lesser degree, and many of the classical imputation methods [102] come with little in the way of theoretical guarantees both statistically and computationally. Developing computationally and statistically efficient techniques for these more general scenarios would be an important challenge in pushing the frontiers of our understanding.

Appendix A

Technical proofs for Chapter 3

A.1 Proof of Lemma 3.5

Let g_α and g_β be an arbitrary pair of functions in our class, and recall that the constraint set Ω is given by the ball $\mathbb{B}_\infty(r)$. From the definition (3.18) of the discrepancy ρ, we need to compute the single function infimum $\inf_{\theta \in \mathbb{B}_\infty(r)} g_\alpha(\theta)$, as well as the quantity $\inf_{\theta \in \mathbb{B}_\infty(r)} \{g_\alpha(\theta) + g_\beta(\theta)\}$.

Evaluating the single function infimum: Beginning with the former quantity, first observe that for any $\theta \in \mathbb{B}_\infty(r)$, we have

$$|\theta(i) + r| = \theta(i) + r \quad \text{and} \quad |\theta(i) - r| = r - \theta(i). \tag{A.1}$$

Consequently, using the definition (3.30) of the base functions, some algebra yields the relations

$$\begin{aligned} f_i^+(\theta) &= \frac{1-\varphi}{4}\theta(i)^2 + \frac{1+3\varphi}{4}r^2 + \frac{(1+\varphi)}{2}r\theta(i), \qquad \text{and} \\ f_i^-(\theta) &= \frac{1-\varphi}{4}\theta(i)^2 + \frac{1+3\varphi}{4}r^2 - \frac{(1+\varphi)}{2}r\theta(i). \end{aligned}$$

Using these expressions for f_i^+ and f_i^-, we obtain

$$\begin{aligned} \underbrace{\left(\frac{1}{2} + \alpha_i\delta\right) f_i^+(\theta) + \left(\frac{1}{2} - \alpha_i\delta\right) f_i^-(\theta)}_{h_i(\theta)} &= \frac{1}{2}\big(f_i^+(\theta) + f_i^-(\theta)\big) + \alpha_i\delta\big(f_i^+(\theta) - f_i^-(\theta)\big) \\ &= \frac{1-\varphi}{4}\theta(i)^2 + \frac{1+3\varphi}{4}r^2 + (1+\varphi)\alpha_i\delta r\theta(i). \end{aligned}$$

A little calculation shows that constrained minimum of the univariate function h_i over the interval $[-r, r]$ is achieved at

$$\theta^*(i) := \begin{cases} \frac{-2\alpha_i \delta r(1+\varphi)}{1-\varphi} & \text{if } \frac{1-\varphi}{1+\varphi} \geq 2\delta \\ -\alpha_i r & \text{if } \frac{1-\varphi}{1+\varphi} < 2\delta, \end{cases}$$

where we have recalled that α_i takes values in $\{-1, +1\}$. Substituting the minimizing argument $\theta^*(i)$, we find that the minimum value is given by

$$h_i(\theta^*(i)) = \begin{cases} \frac{1+3\varphi}{4} r^2 - \frac{\delta^2 r^2 (1+\varphi)^2}{(1-\varphi)} & \text{if } \frac{1-\varphi}{1+\varphi} \geq 2\delta \\ \frac{1+\varphi}{2} r^2 - (1+\varphi)\delta r^2 & \text{if } \frac{1-\varphi}{1+\varphi} < 2\delta. \end{cases}$$

Summing over all co-ordinates $i \in \{1, 2, \ldots, d\}$, we obtain

$$\inf_{\theta \in \mathbb{B}_\infty(r)} g_\alpha(\theta) = \frac{c}{d} \sum_{i=1}^{d} h_i(\theta^*(i)) \; = \; \begin{cases} -\frac{\delta^2 r^2 c (1+\varphi)^2}{(1-\varphi)} + \frac{c r^2 (1+3\varphi)}{4} & \text{if } \frac{1-\varphi}{1+\varphi} \geq 2\delta \\ \frac{1+\varphi}{2} c r^2 - (1+\varphi) c \delta r^2 & \text{if } \frac{1-\varphi}{1+\varphi} < 2\delta. \end{cases} \tag{A.2}$$

Evaluating the joint infimum: Here we begin by observing that for any two $\alpha, \beta \in \mathcal{V}$, we have

$$g_\alpha(\theta) + g_\beta(\theta) = \frac{c}{d} \sum_{i=1}^{d} \left[\frac{1-\varphi}{2} \theta(i)^2 + \frac{1+3\varphi}{2} r^2 + 2(1+\varphi)\alpha_i \delta r \theta(i) \mathbb{I}(\alpha_i = \beta_i) \right]. \tag{A.3}$$

As in our previous calculation, the only coordinates that contribute to $\rho(g_\alpha, g_\beta)$ are the ones where $\alpha_i \neq \beta_i$, and for such coordinates, the function above is minimized at $\theta^*(i) = 0$. Furthermore, the minimum value for any such coordinate is $(1+3\varphi)cr^2/(2d)$.

We split the remainder of our analysis into two cases: first, if we suppose that $\frac{1-\varphi}{1+\varphi} \geq 2\delta$, or equivalently that $1 - \varphi \geq 4\delta/(1+2\delta)$, then equation (A.3) yields that

$$\inf_{\theta \in \mathbb{B}_\infty(r)} \left\{ g_\alpha(\theta) + g_\beta(\theta) \right\} = \frac{c}{d} \sum_{i=1}^{d} \left[\frac{1+3\varphi}{2} r^2 - \frac{2\delta^2 r^2 (1+\varphi)^2}{1-\varphi} \mathbb{I}(\alpha_i = \beta_i) \right].$$

Combined with our earlier expression (A.2) for the single function infimum, we obtain that the discrepancy is given by

$$\rho(g_\alpha, g_\beta) = \frac{2\delta^2 r^2 c (1+\varphi)^2}{d(1-\varphi)} \Delta_H(\alpha, \beta) \geq \frac{2\delta^2 r^2 c}{d(1-\varphi)} \Delta_H(\alpha, \beta).$$

On the other hand, if we assume that $\frac{1-\varphi}{1+\varphi} < 2\delta$, or equivalently that $1 - \varphi < 4\delta/(1+2\delta)$, then we obtain

$$\inf_{\theta \in \mathbb{B}_\infty(r)} \left\{ g_\alpha(\theta) + g_\beta(\theta) \right\} = \frac{c}{d} \sum_{i=1}^{d} \left[\frac{1+3\varphi}{2} r^2 - \left(2(1+\varphi) r^2 \delta - \frac{1-\varphi}{2} r^2 \right) \mathbb{I}(\alpha_i = \beta_i) \right],$$

Combined with our earlier expression (A.2) for the single function infimum, we obtain

$$\rho(g_\alpha, g_\beta) = \frac{c}{d}\left(2(1+\varphi)r^2\delta - \frac{1-\varphi}{2}r^2\right)\Delta_H(\alpha,\beta) \overset{(i)}{\geq} \frac{c(1+\varphi)r^2\delta}{d}\Delta_H(\alpha,\beta),$$

where step (i) uses the bound $1-\varphi < 2\delta(1+\varphi)$. Noting that $\varphi \geq 0$ completes the proof of the lemma.

A.2 Proof of Lemma 3.6

Recall that the constraint set Ω in this lemma is the ball $\mathbb{B}_\infty(r)$. Thus, recalling the definition (3.18) of the discrepancy ρ, we need to compute the single function infimum $\inf_{\theta\in\mathbb{B}_\infty(r)} g_\alpha(\theta)$, as well as the quantity $\inf_{\theta\in\mathbb{B}_\infty(r)}\{g_\alpha(\theta)+g_\beta(\theta)\}$.

Evaluating the single function infimum: Beginning with the former quantity, first observe that for any $\theta \in \mathbb{B}_\infty(r)$, we have

$$\left[\frac{1}{2}+\alpha_i\delta\right]\,|\theta(i)+r| + \left[\frac{1}{2}-\alpha_i\delta\right]\,|\theta(i)-r| = r + 2\alpha_i\delta\theta(i). \tag{A.4}$$

We now consider one of the individual terms arising in the definition (3.16) of the function g_α. Using the relation (A.4), it can be written as

$$\begin{aligned}\frac{1}{d}\left[\left(\frac{1}{2}+\alpha_i\delta\right)f_i^+(\theta) + \left(\frac{1}{2}-\alpha_i\delta\right)f_i^-(\theta)\right] &= \left(\frac{1}{2}+\alpha_i\delta\right)|\theta(i)+r| + \left(\frac{1}{2}-\alpha_i\delta\right)|\theta(i)-r| + \delta|\theta(i)| \\ &= \begin{cases} r+(2\alpha_i+1)\delta\theta(i) & \text{if } \theta(i)\geq 0 \\ r+(2\alpha_i-1)\delta\theta(i) & \text{if } \theta(i)\leq 0 \end{cases}\end{aligned}$$

From this representation, we see that whenever $\alpha_i \neq 0$, then the i^{th} term in the summation defining g_α minimized at $\theta(i) = -r\alpha_i$, at which point it takes on its minimum value $r(1-\delta)$. On the other hand, for any term with $\alpha_i = 0$, the function is minimized at $\theta(i) = 0$ with associated minimum value of r. Combining these two facts shows that the vector $-\alpha r$ is an element of the set $\arg\min_{\theta\in\Omega} g_\alpha(\theta)$, and moreover that

$$\inf_{\theta\in\Omega} g_\alpha(\theta) = cr\,(d - s\delta)\,. \tag{A.5}$$

Evaluating the joint infimum: We now turn to the computation of $\inf_{\theta\in\mathbb{B}_\infty(r)}\{g_\alpha(\theta)+g_\beta(\theta)\}$. From the relation (A.4) and the definitions of g_α and g_β, some algebra yields

$$\inf_{\theta\in\Omega}\{g_\alpha(\theta)+g_\beta(\theta)\} = c\inf_{\theta\in\Omega}\sum_{i=1}^{d}\{2r + 2\delta\,[(\alpha_i+\beta_i)\theta(i) + |\theta(i)|]\}\,. \tag{A.6}$$

Let us consider the minimizer of the i^{th} term in this summation. First, suppose that $\alpha_i \neq \beta_i$, in which case there are two possibilities.

- If $\alpha_i \neq \beta_i$ and neither α_i nor β_i is zero, then we must have $\alpha_i + \beta_i = 0$, so that the minimum value of $2r$ is achieved at $\theta(i) = 0$.
- Otherwise, suppose that $\alpha_i \neq 0$ and $\beta_i = 0$. In this case, we see from Equation (A.6) that it is equivalent to minimizing $\alpha_i\theta(i) + |\theta(i)|$. Setting $\theta(i) = -\alpha_i$ achieves the minimum value of $2r$.

In the remaining two cases, we have $\alpha_i = \beta_i$.

- If $\alpha_i = \beta_i \neq 0$, then the component is minimized at $\theta(i) = -\alpha_i r$ and the minimum value along the component is $2r(1-\delta)$.
- If $\alpha_i = \beta_i = 0$, then the minimum value is $2r$, achieved at $\theta(i) = 0$.

Consequently, accumulating all of these individual cases into a single expression, we obtain

$$\inf_{\theta\in\Omega}\{g_\alpha(\theta) + g_\beta(\theta)\} = 2cr\left(d - \delta\sum_{i=1}^{d}\mathbb{I}[\alpha_i = \beta_i \neq 0]\right). \tag{A.7}$$

Finally, combining equations (A.5) and (A.7) in the definition of ρ, we find that

$$\begin{aligned}\rho(g_\alpha, g_\beta) &= 2cr\left[d - \delta\sum_{i=1}^{d}\mathbb{I}[\alpha_i = \beta_i \neq 0] - (d - s\delta)\right] \\ &= 2c\delta r\left[s - \sum_{i=1}^{d}\mathbb{I}[\alpha_i = \beta_i \neq 0]\right] \\ &= cr\delta\Delta_H(\alpha, \beta),\end{aligned}$$

where the second equality follows since α and β have exactly s non-zero elements each. Finally, since $\mathcal{V}$ is an $s/2$-packing set in Hamming distance, we have $\Delta_H(\alpha, \beta) \geq s/2$, which completes the proof.

A.3 Upper bounds via mirror descent

This appendix is devoted to background on the family of mirror descent methods. We specialize the known convergence results to specific cases of interest here and show that different forms of mirror descent provide matching upper bounds for several of the lower bounds established in Chapter 3, as discussed in the main text.

A.3.1 Matching upper bounds

Now consider the form of mirror descent obtained by choosing the proximal function

$$\psi_a(\theta) := \frac{1}{(a-1)} \|\theta\|_a^2 \qquad \text{for } 1 < a \leq 2. \tag{A.8}$$

Note that this proximal function is 1-strongly convex with respect to the ℓ_a-norm for $1 < a \leq 2$, meaning that

$$\frac{1}{(a-1)} \|\theta\|_a^2 \geq \frac{1}{(a-1)} \left\|\tilde{\theta}\right\|_a^2 + \left(\nabla \frac{1}{(a-1)} \|\theta\|_a^2\right)^T (\theta - \tilde{\theta}) + \frac{1}{2} \left\|\theta - \tilde{\theta}\right\|_a^2.$$

We also consider a somewhat modified application of mirror descent for consistency with our lower bounds. Specifically, we consider mirror descent for $T-1$ rounds and then set $\theta^T = \bar{\theta}(T-1)$, allowing us to have a convergence error bound on the final iterate θ^T.

Upper bounds for dual setting: Let us start from the case $1 \leq p \leq 2$. In this case we use stochastic gradient descent with , and the choice of p ensures that $\mathbb{E}\|\widehat{v}(\theta)\|_2^2 \leq \mathbb{E}\|\widehat{v}(\theta)\|_p^2 \leq G^2$ (the second inequality is true by assumption of Theorem 3.1). Also a straightforward calculation shows that $\|\theta^*\|_2 \leq \|\theta^*\|_q d^{1/2-1/q}$ so that we get the upper bound:

$$\mathbb{E}\left[f(\theta^T) - f(\theta^*)\right] = \mathcal{O}\left(\frac{G d^{1/2-1/q}}{\sqrt{T}}\right),$$

which matches the lower bound from Equation (3.11) for this case. For $p \geq 2$, we use mirror descent with $a = q = p/(p-1)$. In this case, $\mathbb{E}\|\widehat{v}(\theta)\|_p^2 \leq G^2$ and $\|\theta^*\|_q \leq 1$ for the convex set $\mathbb{B}_q(1)$ and the function class $\mathcal{F}_{\text{cv}}(\mathbb{B}_q(1), G, p)$. Hence in this case, the upper bound from Equation 2.10 is $\mathcal{O}(G/\sqrt{T})$ as long as $p = o(\log d)$, which again matches our lower bound from Equation 3.11. Finally, for $p = \Omega(\log d)$, we use mirror descent with $a = 2\log d/(2\log d - 1)$, which gives an upper bound of $\mathcal{O}(G\sqrt{\log d/T})$ (since $1/(a-1) = \mathcal{O}(\log d)$ in this regime).

Upper bounds for ℓ_∞ ball: For this case, we use mirror descent based on the proximal function ψ_a with $a = q$. Under the condition $\|\theta^*\|_\infty \leq 1$, a condition which holds in our lower bounds, we obtain

$$\|\theta^*\|_q \leq \|\theta^*\|_\infty d^{1/q} \;=\; d^{1/q},$$

which implies that $\Phi_q(\theta^*) = \mathcal{O}(d^{2/q})$. Under the conditions of Theorem 3.1, we have $\mathbb{E}\|\widehat{v}(\theta^t)\|_p^2 \leq G^2$ where $p = q/(q-1)$ defines the dual norm. Note that the condition $1 < q \leq 2$ implies that $p \geq 2$. Based on our setting of θ^T to be $\bar{\theta}(T-1)$, substituting this in the upper bound (2.10) yields

$$\mathbb{E}\big[f(\theta^T) - f(\theta^*)\big] = \mathcal{O}\Big(G\sqrt{d^{2/q}/T}\Big) \;=\; \mathcal{O}\Big(G d^{1-1/p}\sqrt{\frac{1}{T}}\Big),$$

which matches the lower bound from Theorem 3.1(b) (we note that there is an additional log factor here just like the preceding discussion when $p = \mathcal{O}(\log d)$ which we ignore here).

For $1 \leq p \leq 2$, we use stochastic gradient descent with $q = 2$, in which case $\|\theta^*\|_2 \leq \sqrt{d}$ and $\mathbb{E}\,\|\widehat{v}(\theta^t)\|_2^2 \leq \mathbb{E}\,\|\widehat{v}(\theta^t)\|_p^2 \leq G^2$ by assumption. Substituting these in the upper bound for mirror descent yields an upper bound to match the lower bound of Theorem 3.1(a).

Upper bounds for Theorem 3.3: In order to recover matching upper bounds in this case, we use the function ψ_a from Equation (A.8) with $a = \frac{2\log d}{2\log d - 1}$. In this case, the resulting upper bound (2.10) on the convergence rate takes the form

$$\mathbb{E}\left[f(\theta^T) - f(\theta^*)\right] \;=\; \mathcal{O}\left(G\sqrt{\frac{\|\theta^*\|_a^2}{2(a-1)T}}\right) = \mathcal{O}\left(G\sqrt{\frac{\|\theta^*\|_a^2 \log d}{T}}\right), \tag{A.9}$$

since $\frac{1}{a-1} = 2\log d - 1$. Based on the conditions of Theorem 3.3, we are guaranteed that θ^* is s-sparse, with every component bounded by 1 in absolute value, so that $\|\theta^*\|_a^2 \leq s^{2/a} \leq s^2$, where the final inequality follows since $a > 1$. Substituting this upper bound back into Equation (A.9) yields

$$\mathbb{E}\left[f(\theta^T) - f(\theta^*)\right] = \mathcal{O}\left(L\sqrt{\frac{s^2 \log d}{T}}\right).$$

Note that whenever $s = \mathcal{O}(d^{1-\delta})$ for some $\delta > 0$, then we have $\log d = \Theta(\log \frac{d}{s})$, in which case this upper bound matches the lower bound from Theorem 3.3 up to constant factors, as claimed.

Appendix B

Auxiliary results and proofs for Chapter 4

B.1 Auxiliary results for Theorem 4.1 and Corollary 4.1

We start by establishing Lemma 4.1. To prove the lemma, we first need a simple claim.

Lemma B.1. *Let $c_1 > c_2 > 0$, $s > 0$, and define*

$$i_1^* = \operatorname*{argmin}_{i=1,2,3,\ldots} \left\{ R_i^* + c_1 \left(\gamma_i \left(\frac{T}{s} \right) + \kappa_2 \sqrt{\frac{2(m + \log s)}{n_i(T/s)}} \right) \right\},$$

$$i_2^* = \operatorname*{argmin}_{i=1,2,3,\ldots} \left\{ R_i^* + c_2 \left(\gamma_i \left(\frac{T}{s} \right) + \kappa_2 \sqrt{\frac{2(m + \log s)}{n_i(T/s)}} \right) \right\}.$$

Then under the monotonicity assumptions B, we have $i_1^* \leq i_2^*$.

Proof. Recall the shorthand definition (4.8) of $\overline{\gamma}_i$. Under the monotonicity assumptions B(a)–(b), $\overline{\gamma}_i$ is monotone increasing in i. By the definitions of i_1^* and i_2^* we have

$$R_{i_1}^* + c_1 \overline{\gamma}_{i_1^*}(T, s) \leq R_{i_2}^* + c_1 \overline{\gamma}_{i_2^*}(T, s) \quad \text{and} \quad R_{i_2}^* + c_2 \overline{\gamma}_{i_2^*}(T, s) \leq R_{i_1}^* + c_2 \overline{\gamma}_{i_1^*}(T, s).$$

Adding the two inequalities we obtain

$$(c_1 - c_2) \overline{\gamma}_{i_1^*}(T, s) \leq (c_1 - c_2) \overline{\gamma}_{i_2^*}(T, s).$$

Since $c_1 - c_2 > 0$ by assumption, the monotonicity of $\overline{\gamma}_i$ guarantees $i_1^* \leq i_2^*$. □

We now proceed to establish Lemma 4.1.

Proof of Lemma 4.1 Lemma B.1 allows us to establish a simpler version of Lemma 4.1. Since $1+\lambda > 1$, it suffices to establish $i_0 \leq K(\lambda)$, where

$$i_0 = \operatorname*{argmin}_{i=1,2,3,\dots} \left\{ R_i^* + \gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{2(m + \log s(\lambda))}{n_i(T/s(\lambda))}} \right\}.$$

Let $\overline{\gamma}_i$ be shorthand for the quantity (4.8) as usual. Recalling the construction of S in Algorithm 1, we observe that the class $K(\lambda)$ satisfies

$$(1+\lambda)^{s(\lambda)-2} \overline{\gamma}_1 \left(T, s(\lambda)\right) \leq \overline{\gamma}_{K(\lambda)} \left(T, s(\lambda)\right) \leq (1+\lambda)^{s(\lambda)-1} \overline{\gamma}_1 \left(T, s(\lambda)\right).$$

The setting (4.10) of $s(\lambda)$ ensures that

$$(1+\lambda)^{s(\lambda)-2} \geq (1+\lambda)^{\lceil \log(1+B/\overline{\gamma}_1(T,s(\lambda)))/\log(1+\lambda) \rceil} \geq \exp\left(\log\left(1 + \frac{B}{\overline{\gamma}_1(T, s(\lambda))} \right) \right)$$

so that

$$(1+\lambda)^{s(\lambda)-2} \overline{\gamma}_1 \left(T, s(\lambda)\right) \geq B + \overline{\gamma}_1 \left(T, s(\lambda)\right) \geq R_1^* + \overline{\gamma}_1 \left(T, s(\lambda)\right) \geq \inf_i \left\{ R_i^* + \overline{\gamma}_i \left(T, s(\lambda)\right) \right\}.$$

Hence we observe that

$$\begin{aligned} R^*_{K(\lambda)} + \overline{\gamma}_{K(\lambda)} \left(T, s(\lambda)\right) &\geq \overline{\gamma}_{K(\lambda)} \left(T, s(\lambda)\right) \\ &\geq (1+\lambda)^{s(\lambda)-2} \overline{\gamma}_1 \left(T, s(\lambda)\right) \\ &\geq \inf_{i=1,2,3,\dots} \left\{ R_i^* + \overline{\gamma}_i \left(T, s(\lambda)\right) \right\}. \end{aligned}$$

We must thus have $i_0 \leq K(\lambda)$, and Lemma B.1 further implies that $i^* \leq K(\lambda)$. □

We finally provide a proof for Proposition 4.1.

Proof of Proposition 4.1 Since for any $a, b \geq 0$, $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$, it suffices to control the probability of the event

$$R(f) > \min_{i \in S} \left\{ R_i^* + 2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{\log s(\lambda)}{2 n_i(T/s(\lambda))}} + \kappa_2 \sqrt{\frac{m}{n_i(T/s(\lambda))}} \right\}. \tag{B.1}$$

For the event (B.1) to occur, at least one of

$$R(f) > \min_{i \in S} \left\{ \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) + \gamma_i \left(\frac{T}{s(\lambda)} \right) + \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} \right\} \tag{B.2a}$$

or

$$\min_{i \in S} \left\{ \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) + \gamma_i \left(\frac{T}{s(\lambda)} \right) + \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} \right\}$$
$$> \min_{i \in S} \left\{ R_i^* + 2\gamma_i \left(\frac{T}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{\log s(\lambda)}{2n_i(T/s(\lambda))}} + \kappa_2 \sqrt{\frac{m}{n_i(T/s(\lambda))}} \right\} \tag{B.2b}$$

must occur. We bound the probabilities of the events (B.2a) and (B.2b) in turn.

If the event (B.2a) occurs, by definition of the selection strategy (4.11), it must be the case that for some $i \in S$

$$R(\hat{f}_i) > \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) + \gamma_i \left(\frac{T}{s(\lambda)} \right) + \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}}$$

since the choice f minimizes the right side of this display over the classes $\mathcal{F}_i$ for $i \in S$. By a union bound, we see that

$$\mathbb{P}\left[R(f) > \min_{i \in S} \left\{ \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) + \gamma_i \left(\frac{T}{s(\lambda)} \right) + \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} \right\} \right]$$
$$\leq \mathbb{P}\left[\exists\, i \in S \text{ s.t. } R(\hat{f}_i) > \widehat{R}(\hat{f}_i) + \gamma_i \left(\frac{T}{s(\lambda)} \right) + \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} \right]$$
$$\leq \kappa_1 \sum_{i \in S} \exp\left(-m - \log s(\lambda)\right) = \kappa_1 \exp(-m),$$

where the final inequality follows from Assumption C.

Now we bound the probability of the event (B.2b), noting that the event implies that

$$\max_{i \in S} \left\{ \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) - R_i^* - \gamma_i \left(\frac{T}{s(\lambda)} \right) - \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} - \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} \right\} > 0.$$

We can thus apply a union bound to see that the probability of the event (B.2b) is bounded by

$$\mathbb{P}\left[\max_{i \in S} \left\{ \widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) - R_i^* - \gamma_i \left(\frac{T}{s(\lambda)} \right) - \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} - \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} \right\} > 0 \right]$$
$$\leq \sum_{i \in S} \mathbb{P}\left[\widehat{R}_{n_i(T/s(\lambda))}(\hat{f}_i) - R_i^* - \gamma_i \left(\frac{T}{s(\lambda)} \right) - \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} - \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} > 0 \right]$$
$$\leq \sum_{i \in S} \mathbb{P}\left[\widehat{R}(f_i^*) - R_i^* > \frac{\kappa_2}{2} \sqrt{\frac{\log s(\lambda)}{n_i(T/s(\lambda))}} + \frac{\kappa_2}{2} \sqrt{\frac{m}{n_i(T/s(\lambda))}} \right], \tag{B.3}$$

where the final inequality uses Assumption B(d), which states that $\mathcal{A}$ outputs a γ_i-minimizer of the empirical risk. Now we can bound the deviations using the second part of Assumption C, since f_i^* is non-random: the quantity (B.3) is bounded by

$$\sum_{i \in S} \kappa_1 \exp\left(-n_i(T/s(\lambda))\left(\frac{\log s(\lambda)}{n_i(T/s(\lambda))} + \frac{m}{n_i(T/s(\lambda))}\right)\right) \leq \kappa_1 \exp(-m).$$

Combining the two events (B.2a) and (B.2b) completes the proof of the proposition. □

B.2 Auxiliary results for Theorem 4.2

Proof of Lemma 4.3 In the proof of the lemma, assume that both of the events (4.23) hold. Recall that we define $\hat{f}_j = \mathcal{A}(j, n_j)$, so that by the definition (4.22a) and Assumption B that $\hat{f}_j$ is a γ_j-accurate minimizer of the empirical risk, we have

$$R(\hat{f}_j) \leq R(f_j^*) + 3\gamma_j(n_j) + \kappa_2 \epsilon_j \tag{B.4}$$

for any j. By our assumption that the index $j \leq \widehat{i}$, we have $\hat{f}_j \in \mathcal{F}_{\widehat{i}}$, and since the event (4.22b) holds for the class $\widehat{i}$ (i.e. $\mathcal{E}_2^{\widehat{i}}(\epsilon_{\widehat{i}})$ occurs), we further obtain that

$$\widehat{R}_{\widehat{i}}(\hat{f}_j) - \widehat{R}_{\widehat{i}}(f_j^*) \leq 2\left(R(\hat{f}_j) - R(f_j^*)\right) + \gamma_{\widehat{i}}(n_{\widehat{i}}) + \kappa_2 \epsilon_{\widehat{i}}. \tag{B.5}$$

Applying our earlier bound on $R(\hat{f}_j) - R(f_j^*)$ to the inequality (B.5), we see that

$$\widehat{R}_{\widehat{i}}(\hat{f}_j) - \widehat{R}_{\widehat{i}}(f_j^*) \leq 6\gamma_j(n_j) + 2\kappa_2 \epsilon_j + \gamma_{\widehat{i}}(n_{\widehat{i}}) + \kappa_2 \epsilon_{\widehat{i}}. \tag{B.6}$$

Again using the fact that the event (4.22b) holds for the class $\widehat{i}$ and choosing $f = f_j^*$, we see that

$$2\left(R(f_j^*) - R(f_{\widehat{i}}^*)\right) \geq \left(\widehat{R}_{\widehat{i}}(f_j^*) - \widehat{R}_{\widehat{i}}(f_{\widehat{i}}^*)\right) - \gamma_{\widehat{i}}(n_{\widehat{i}}) - \kappa_2 \epsilon_{\widehat{i}}.$$

Now apply the inequality (B.6) to lower bound $\widehat{R}_{\widehat{i}}(f_j^*)$ to see that

$$2\left(R(f_j^*) - R(f_{\widehat{i}}^*)\right) \geq \widehat{R}_{\widehat{i}}(\hat{f}_j) - \widehat{R}_{\widehat{i}}(f_{\widehat{i}}^*) - 6\gamma_j(n_j) - 2\kappa_2 \epsilon_j - 2\gamma_{\widehat{i}}(n_{\widehat{i}}) - 2\kappa_2 \epsilon_{\widehat{i}}.$$

Using the condition (4.24) that defines the selected index $\widehat{i}$, we obtain

$$\begin{aligned} 2\left(R(f_j^*) - R(f_{\widehat{i}}^*)\right) &\geq \widehat{R}_{\widehat{i}}(\hat{f}_{\widehat{i}}) + c_1\gamma_{\widehat{i}}(n_{\widehat{i}}) + c_2\kappa_2\epsilon_{\widehat{i}} - c_1\gamma_j(n_j) - \widehat{R}_{\widehat{i}}(f_{\widehat{i}}^*) - 6\gamma_j(n_j) - 2\kappa_2\epsilon_j - 2\gamma_{\widehat{i}}(n_{\widehat{i}}) - 2\kappa_2\epsilon_{\widehat{i}} \\ &= \widehat{R}_{\widehat{i}}(\hat{f}_{\widehat{i}}) - \widehat{R}_{\widehat{i}}(f_{\widehat{i}}^*) + (c_1 - 2)\gamma_{\widehat{i}}(n_{\widehat{i}}) - (6 + c_1)\gamma_j(n_j) - 2\kappa_2\epsilon_j + (c_2 - 2)\kappa_2\epsilon_{\widehat{i}} \end{aligned}$$

Finally, we note that by the event (4.22a), since $R(f_j^*) - R(f) \leq 0$ for all $f \in \mathcal{F}_j$, we have

$$\widehat{R}_{\widehat{i}}(f_{\widehat{i}}^*) \leq \widehat{R}_{\widehat{i}}(\hat{f}_{\widehat{i}}) + \frac{1}{2}\gamma_{\widehat{i}}(n_{\widehat{i}}) + \frac{1}{2}\kappa_2\epsilon_{\widehat{i}},$$

whence we obtain

$$2\left(R(f_j^*) - R(f_{\widehat{i}}^*)\right) \geq (c_1 - 5/2)\,\gamma_{\widehat{i}}(n_{\widehat{i}}) - (6 + c_1)\gamma_j(n_j) - 2\kappa_2\epsilon_j + (c_2 - 5/2)\kappa_2\epsilon_{\widehat{i}}. \tag{B.7}$$

Applying the inequality (B.4) for the class $\widehat{i}$, we have

$$R(f_j^*) - R(\hat{f}_{\widehat{i}}) \geq R(f_j^*) - R(f_{\widehat{i}}^*) - 3\gamma_{\widehat{i}}(n_{\widehat{i}}) - \kappa_2\epsilon_{\widehat{i}},$$

and combining this inequality with the earlier guarantee (B.7), we find that

$$2\left(R(f_j^*) - R(\hat{f}_{\widehat{i}})\right) \geq (c_1 - 17/2)\gamma_{\widehat{i}}(n_{\widehat{i}}) - (6 + c_1)\gamma_j(n_j) - 2\kappa_2\epsilon_j + (c_2 - 7/2)\kappa_2\epsilon_{\widehat{i}}$$

Rearranging terms, we obtain the statement of the lemma. □

In order to prove the other Lemma 4.4, we need one more simple result. We start by stating and proving a simple lemma, and the establish Lemma 4.4.

Lemma B.2. *Let the events (4.22a) and (4.22b) hold for all $i \in S_\lambda$, that is, $\mathcal{E}_1(\epsilon)$ and $\mathcal{E}_2(\epsilon)$ hold. For any classes $i, j \in S_\lambda$ such that $i \geq j$ and*

$$\widehat{R}_i(\hat{f}_j) + c_1\gamma_j\left(\frac{T}{s(\lambda)}\right) \leq \widehat{R}_i(\hat{f}_i) + c_1\gamma_i\left(\frac{T}{s(\lambda)}\right) + c_2\kappa_2\epsilon_i$$

we have

$$R(\hat{f}_j) \leq R(f_i^*) + (2c_1 + 3)\gamma_i(n_i) + (2c_2 + 1)\kappa_2\epsilon_i.$$

Proof. We begin by noting that since $i \geq j$, we have $\hat{f}_j \in \mathcal{F}_i$, and since the event (4.22a) holds by assumption, we have

$$R(\hat{f}_j) - R(f_i^*) \leq 2\left(\widehat{R}_i(\hat{f}_j) - \widehat{R}_i(f_i^*)\right) + \gamma_i(n_i) + \kappa_2\epsilon_i.$$

Recalling the inequality assumed in the condition of the lemma, we see that

$$R(\hat{f}_j) - R(f_i^*) \leq 2\left(\widehat{R}_i(\hat{f}_i) + c_1\gamma_i(n_i) + c_2\kappa_2\epsilon_i - c_1\gamma_j(n_j) - \widehat{R}_i(f_i^*)\right) + \gamma_i(n_i) + \kappa_2\epsilon_i.$$

Applying Assumption B(d) on the empirical minimizers, we have $\widehat{R}_i(\hat{f}_i) - \widehat{R}_i(f_i^*) \leq \gamma_i(n_i)$, so

$$R(\hat{f}_j) - R(f_i^*) \leq 2\left((c_1 + 1)\gamma_i(n_i) + c_2\kappa_2\epsilon_i - c_1\gamma_j(n_j)\right) + \gamma_i(n_i) + \kappa_2\epsilon_i.$$

Ignoring the negative term $-c_1\gamma_j(n_j)$ yields the lemma. □

Proof of Lemma 4.4 For $j \in S_\lambda$, define $S_\lambda(j)$ to be the position of class j in the coarse-grid set (that is, $S_\lambda(1) = 1$, the next class $j \in S_\lambda$ has $S_\lambda(j) = 2$ and so on). We prove the lemma by induction on the class j for $j \geq \widehat{i}$, $j \in S_\lambda$. Our inductive hypothesis is that

$$R(\hat{f}_{\widehat{i}}) \leq R(f_j^*) + (S_\lambda(j) - S_\lambda(\widehat{i}) + 1)\left((2c_1+3)\gamma_j\left(\frac{T}{s(\lambda)}\right) + (2c_1+1)\kappa_2\epsilon_j\right). \tag{B.8}$$

The base case for $j = \widehat{i}$ is immediate since by assumption, the event (4.22a) holds, so we obtain the inequality (B.4).

For the inductive step, we assume that the claim holds for all $\widehat{i} \leq k \leq j-1$ such that $k \in S_\lambda$ and establish the claim for j. Since $\widehat{i}$ is the largest class in S_λ satisfying the condition (4.24) and $j \geq \widehat{i}$, there must exist a class $k < j$ in S_λ for which

$$\widehat{R}_j(\hat{f}_k) + c_1\gamma_k(n_k) < \widehat{R}_j(\hat{f}_j) + c_1\gamma_j(n_j) + c_2\kappa_2\epsilon_j. \tag{B.9}$$

By inspection, this is precisely the condition of Lemma B.2, so

$$R(f_k^*) \leq R(\hat{f}_k) \leq R(f_j^*) + (2c_1+3)\gamma_j(n_j) + (2c_2+1)\kappa_2\epsilon_j.$$

Now there are two possibilities. If $k \leq \widehat{i}$, Lemma 4.3 applies, and we recall the assumptions on c_1 and c_2, which guarantee $2c_1 + 3 \geq 6 + c_1$ and $2c_2 + 1 \geq 2$. If $j \geq \widehat{i}$, then we can apply our inductive hypothesis since $k < j$. In either case, we conclude that

$$\begin{aligned} R(\hat{f}_{\widehat{i}}) &\leq R(f_k^*) + (S_\lambda(k) - S_\lambda(\widehat{i}) + 1)\left[(2c_1+3)\gamma_k(n_k) + (2c_2+1)\kappa_2\epsilon_k\right] \\ &\leq R(f_k^*) + (S_\lambda(j) - 1 - S_\lambda(\widehat{i}) + 1)\left[(2c_1+3)\gamma_j(n_j) + (2c_2+1)\kappa_2\epsilon_j\right], \end{aligned}$$

where the final inequality uses $S_\lambda(k) \leq S_\lambda(j) - 1$ and the monotonicity assumptions B(a)-(b). Applying the relationship (B.9) of the risk of f_k^* to that of f_j^* shows that the inductive hypothesis (B.8) holds at j. Noting that $s(\lambda) \geq S_\lambda(j) - S_\lambda(\widehat{i}) + 1$ completes the proof. □

B.3 Proof of Lemma 4.5

Following [13], we show that the event in the lemma occurs with very low probability by breaking it up into smaller events more amenable to analysis. Recall that we're interested in controlling the probability of the event

$$\overline{R}(i, n_i s_i) - \kappa_2\sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2\sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \tag{B.10}$$

For this bad event to happen, at least one of the following three events must happen:

$$\widehat{R}_{n_i s_i}(\mathcal{A}(i, n_i s_i)) - \inf_{f \in \mathcal{F}_i} R(f) \leq -\gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \tag{B.11a}$$

$$\widehat{R}_{n_{i^*} s_{i^*}}(\mathcal{A}(i^*, n_{i^*} s_{i^*})) - \inf_{f \in \mathcal{F}_{i^*}} R(f) \geq \gamma_i(n_{i^*} s_{i^*}) + \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} + \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \tag{B.11b}$$

$$R_i^* + \gamma_i(T n_i) \leq R^* + \gamma_{i^*}(T n_{i^*}) + 2 \left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \right). \tag{B.11c}$$

Temporarily use the shorthand $f_i = \mathcal{A}(i, n_i s_i)$ and $f_{i^*} = \mathcal{A}(i^*, n_{i^*} s_{i^*})$. The relationship between Eqs. (B.11a)–(B.11c) and the event in (B.10) follows from the fact that if none of (B.11a)–(B.11c) occur, then

$$\begin{aligned}
&\overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \\
&\quad = \quad \widehat{R}_{n_i s_i}(f_i) + \gamma_i(T n_i) - \gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \\
&\quad \overset{\text{(B.11a)}}{>} \quad \inf_{f \in \mathcal{F}_i} R(f) + \gamma_i(T n_i) - 2 \left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log t}{n_i s_i}} \right) \\
&\quad \overset{\text{(B.11c)}}{>} \quad \inf_{f \in \mathcal{F}_{i^*}} R(f) + \gamma_{i^*}(T n_{i^*}) + 2 \left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \right) \\
&\qquad\qquad - 2 \left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log n}{n_i s_i}} \right) \\
&\quad \overset{\text{(B.11b)}}{>} \quad \widehat{R}_{n_{i^*} s_{i^*}}(f_{i^*}) + \gamma_{i^*}(T n_{i^*}) - \gamma_i(n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} - \kappa_2 \sqrt{\frac{\log t}{n_{i^*} s_{i^*}}} \\
&\quad = \quad \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log t}{n_{i^*} s_{i^*}}}.
\end{aligned}$$

From the above string of inequalities, to show that the event (B.10) has low probability, we need simply show that each of (B.11a), (B.11b), and (B.11c) have low probability.

To prove that each of the bad events have low probability, we note the following consequences of Assumption C. Recall the definition of f_i^* as the minimizer of $R(f)$ over the class $\mathcal{F}_i$. Then by Assumption C(b),

$$R(f_i^*) - \gamma_i(n) - \kappa_2 \epsilon \leq R(\mathcal{A}(i, n)) - \gamma_i(n) - \kappa_2 \epsilon < \widehat{R}_n(\mathcal{A}(i, n)),$$

while Assumptions C(c) and C(e) imply

$$\widehat{R}_n(\mathcal{A}(i, n)) \leq \widehat{R}_n(f_i^*) + \gamma_i(n) \leq R(f_i^*) + \gamma_i(n) + \kappa_2 \epsilon,$$

each with probability at least $1-\kappa_1\exp(-4n\epsilon^2)$. In particular, we see that the events (B.11a) and (B.11b) have low probability:

$$\begin{aligned}
&\mathbb{P}\left[\widehat{R}_{n_is_i}(\mathcal{A}(i,n_is_i)) - R(f_i^*) \le -\gamma_i(n_is_i) - \kappa_2\sqrt{\frac{\log K}{n_is_i}} - \kappa_2\sqrt{\frac{\log T}{n_is_i}}\right] \\
&\qquad \le \kappa_1\exp\left(-4n_is_i\left(\frac{\log K}{n_is_i} + \frac{\log t}{n_is_i}\right)\right) = \frac{\kappa_1}{(tK)^4} \\
&\mathbb{P}\left[\widehat{R}_{n_{i^*}s_{i^*}}(\mathcal{A}(i^*,n_{i^*}s_{i^*})) - R^* \ge \gamma_{i^*}(n_{i^*}s_{i^*}) + \kappa_2\sqrt{\frac{\log K}{n_{i^*}s_{i^*}}} + \kappa_2\sqrt{\frac{\log T}{n_{i^*}s_{i^*}}}\right] \\
&\qquad \le \kappa_1\exp\left(-4n_{i^*}s_{i^*}\left(\frac{\log K}{n_{i^*}s_{i^*}} + \frac{\log T}{n_{i^*}s_{i^*}}\right)\right) = \frac{\kappa_1}{(tK)^4}.
\end{aligned}$$

What remains is to show that for large enough τ, (B.11c) does not happen. Recalling the definition that $R^* + \gamma_{i^*}(Tn_{i^*}) = R_i^* + \gamma_i(Tn_i) - \Delta_i$, we see that for (B.11c) to fail it is sufficient that

$$\Delta_i > 2\gamma_i(\tau n_i) + 2\kappa_2\sqrt{\frac{\log K}{n_i\tau}} + 2\kappa_2\sqrt{\frac{\log T}{n_i\tau}}.$$

Let $x \wedge y := \min\{x,y\}$ and $x \vee y := \max\{x,y\}$. Since $\gamma_i(n) \le c_i n^{-\alpha_i}$, the above is satisfied when

$$\frac{\Delta_i}{2} > c_i(\tau n_i)^{-(\alpha_i\wedge\frac{1}{2})} + \kappa_2\sqrt{\log K}(\tau n_i)^{-(\alpha_i\wedge\frac{1}{2})} + \kappa_2\sqrt{\log T}(\tau n_i)^{-(\alpha_i\wedge\frac{1}{2})} \tag{B.12}$$

We can solve (B.12) above and see immediately that if

$$\tau_i > \frac{2^{1/\alpha_i\vee 2}(c_i + \kappa_2\sqrt{\log T} + \kappa_2\sqrt{\log K})^{1/\alpha_i\vee 2}}{n_i\Delta_i^{1/\alpha_i\vee 2}},$$

then

$$R_i^* > R^* + 2\left(\gamma_i(n_i\tau_i) + \kappa_2\sqrt{\frac{\log K}{n_i\tau_i}} + \kappa_2\sqrt{\frac{\log T}{n_i\tau_i}}\right). \tag{B.13}$$

Thus the event in (B.11c) fails to occur, completing the proof of the lemma.

B.4 Proofs of Proposition 4.2 and Theorem 4.4

In this section we provide proofs for Proposition 4.2 and Theorem 4.4. The proof of the proposition follows by dividing the model classes into two groups: those for which $\Delta_i > \gamma$, and those with small excess risk, i.e. $\Delta_i < \gamma$. Theorem 4.3 provides an upper bound on the fraction of budget allocated to model classes of the first type. For the model classes with

small excess risk, all of them are nearly as good as i^* in the regret criterion of Proposition 4.2. Combining the two arguments gives us the desired result.

Of course, the proposition has the drawback that it does not provide us with a prescription to select a good model or even a model class. This shortcoming is addressed by Theorem 4.4. The theorem relies on an averaging argument used quite frequently to extract a good solution out of online learning or stochastic optimization algorithms [48, 117].

B.4.1 Proof of Proposition 4.2

Define $\beta_i = \max\{1/\alpha_i, 2\}$ as in the conclusion of Theorem 4.3, and let $b_i = c_i + \kappa_2\sqrt{\log T}$. Dividing the regret into classes with high and low excess penalized risk Δ_i, for any threshold $\gamma \geq 0$ we have by a union bound that with probability at least $1 - \kappa_1/TK^3$,

$$\begin{aligned}\sum_{i=1}^{K} \Delta_i T_i(T) &= \sum_{\{i|\Delta_i \geq \gamma\}} \Delta_i T_i(T) + \sum_{\{i|\Delta_i \leq \gamma\}} \Delta_i T_i(T) \\ &\leq C \sum_{\{i|\Delta_i \geq \gamma\}} \Delta_i \frac{b_i^{\beta_i}}{n_i \Delta_i^{\beta_i}} + \gamma T \;\leq\; C \sum_{i=1}^{K} \frac{b_i^{\beta_i}}{n_i \gamma^{\beta_i - 1}} + \gamma T.\end{aligned}$$

To simplify this further, we use the assumption that $\alpha_i \equiv \alpha$ for all i. Hence the complexity penalties of the classes differ only in the sampling rates n_i, that is,

$$\sum_{i=1}^{K} \Delta_i T_i(T) \leq \frac{1}{\gamma^{\beta-1}} \sum_{i=1}^{K} \frac{C b_i^{\beta_i}}{n_i} + \gamma T. \tag{B.14}$$

Minimizing the bound (B.14) over γ by taking derivatives, we get

$$\gamma = T^{-\frac{1}{\beta}} (\beta - 1)^{\frac{1}{\beta}} \left(\sum_{i=1}^{K} \frac{C b_i^{\beta}}{n_i} \right)^{\frac{1}{\beta}},$$

which, when plugged back into (B.14), gives

$$\sum_{i=1}^{K} \Delta_i T_i(T) \leq 2 \left(\sum_{i=1}^{K} \frac{C b_i^{\beta}}{n_i} \right)^{1/\beta} (\beta - 1)^{1/\beta} T^{1 - 1/\beta}.$$

Noting that $\frac{1}{\beta}\log(\beta - 1) \leq \frac{\beta - 2}{\beta} < 1$, we see that $(\beta - 1)^{1/\beta} < \exp(1)$. Plugging the definition of $\beta = \max\{1/\alpha, 2\}$, so that $1/\beta = \min\{\alpha, \frac{1}{2}\}$, gives the result of the proposition.

B.4.2 Proof of Theorem 4.4

Before proving the theorem, we state a technical lemma that makes our argument somewhat simpler.

Lemma B.3. *For $0 < p < 1$ and $a \succ 0$, consider the optimization problem*

$$\max_x \sum_{i=1}^K a_i x_i^p \quad \text{s.t.} \quad \sum_{i=1}^K x_i \le T, \; x_i \ge 0.$$

The solution of the problem is to take $x_i \propto a_i^{1/(1-p)}$, and the optimal value is

$$T^p \left(\sum_{i=1}^K a_i^{\frac{1}{1-p}} \right)^{1-p}.$$

Proof. Reformulating the problem to make it a minimization problem, that is, our objective is $-\sum_{i=1}^K a_i x_i^p$, we have a convex problem. Introducing Lagrange multipliers $\theta \ge 0$ and $\nu \in \mathbb{R}_+^K$ for the inequality constraints, we have Lagrangian

$$\mathcal{L}(x, \theta, \nu) = -\sum_{i=1}^K a_i x_i^p + \theta \left(\sum_{i=1}^K x_i - T \right) - \langle \nu, x \rangle .$$

To find the infimum of the Lagrangian over x, we take derivatives and see that $-a_i p x_i^{p-1} + \theta - \nu_i = 0$, or that $x_i = a_i^{-1/(p-1)} p^{-1/(p-1)} (\theta - \nu_i)^{1/(p-1)}$. Since $a_i > 0$, the complimentary slackness conditions for ν are satisfied with $\nu = 0$, and we see that θ is simply a multiplier to force the sum $\sum_{i=1}^K x_i = T$. That is, $x_i \propto a_i^{1/(1-p)}$, and normalizing appropriately, $x_i = T a_i^{1/(1-p)} / \sum_{j=1}^K a_j^{1/(1-p)}$. By plugging x_i into the objective, we have

$$\sum_{i=1}^K a_i x_i^p = T^p \frac{\sum_{i=1}^K a_i a_i^{p/(1-p)}}{\left(\sum_{j=1}^K a_j^{1/(1-p)} \right)^p} = T^p \frac{\sum_{i=1}^K a_i^{1/(1-p)}}{\left(\sum_{j=1}^K a_j^{1/(1-p)} \right)^p} = T^p \left(\sum_{i=1}^K a_i^{1/(1-p)} \right)^{1-p}$$ □

With the Lemma B.3 in hand, we proceed with the proof of Theorem 4.4. As before, we use the shorthand $\beta = \max\{1/\alpha, 2\}$ throughout the proof to reduce clutter. We also let $s_i(t)$ be the number of times class i was selected by time t. Recalling the definition of the regret from (4.28) and the result of the previous proposition, we have with probability at least $1 - \kappa_1/(TK^3)$

$$\frac{1}{T} \sum_{t=1}^T [R_{i_t}^* + \gamma_{i_t}(T n_{i_t})] \le R^* + \gamma_{i^*}(T n_{i^*}) + 2e\kappa_2 T^{-1/\beta} \sqrt{\log T} \left(\sum_{i=1}^K \frac{C}{n_i} \right)^{1/\beta}.$$

Using the definition of f_i^* as the minimizer of $R(f)$ over $\mathcal{F}_i$, we use Assumptions C(c) and C(e) to see that for fixed s_i, with probability at least $1 - \kappa_1/(TK)^4$,

$$\widehat{R}_{n_i s_i}(\mathcal{A}(i, n_i s_i)) \leq \widehat{R}_{n_i s_i}(f_i^*) + \gamma_i(n_i s_i) \leq R(f_i^*) + \gamma_i(n_i s_i) + \kappa_2\sqrt{\frac{\log K}{n_i s_i}} + \kappa_2\sqrt{\frac{\log T}{n_i s_i}}. \quad \text{(B.15)}$$

Denote by f_t the output of $\mathcal{A}$ on round t, that is, $f_t = \mathcal{A}(i_t, n_{i_t} s_{i_t}(t))$. By the previous equation (B.15), we can use a union bound and the regret bound from Proposition 4.2 to conclude that with probability at least $1 - \kappa_1/(TK^3) - \kappa_1/(T^3K^3)$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \widehat{R}_{n_{i_t} s_{i_t}(t)}(f_t) + \gamma_{i_t}(T n_{i_t}) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left[\gamma_i(n_{i_t} s_{i_t}(t)) + \kappa_2\sqrt{\frac{\log K}{n_{i_t} s_{i_t}(t)}} + \kappa_2\sqrt{\frac{\log T}{n_{i_t} s_{i_t}(t)}}\right] + \frac{1}{T}\sum_{t=1}^{T}\left[R_{i_t}^* + \gamma_{i_t}(T n_{i_t})\right] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left[\gamma_i(n_{i_t} s_{i_t}(t)) + \kappa_2\sqrt{\frac{\log K}{n_{i_t} s_{i_t}(t)}} + \kappa_2\sqrt{\frac{\log T}{n_{i_t} s_{i_t}(t)}}\right] + R(f_i^*) + \gamma_i(n_i s_i) \\
&\qquad + 2e\kappa_2 T^{-1/\beta}\sqrt{\log T}\left(\sum_{i=1}^{K}\frac{C}{n_i}\right)^{1/\beta}. \quad \text{(B.16)}
\end{aligned}$$

Now we again make use of Assumption C(b) to note that with probability at least $1 - \kappa_1/(T^4K^4)$,

$$R(f_t) \leq \widehat{R}_{n_{i_t} s_{i_t}(t)}(f_t) + \gamma_{i_t}(n_{i_t} s_{i_t}(t)) + \kappa_2\sqrt{\frac{\log K}{n_{i_t} s_{i_t}(t)}} + \kappa_2\sqrt{\frac{\log T}{n_{i_t} s_{i_t}(t)}}.$$

Using a union bound and applying the empirical risk bound (B.16), we drop the positive $\gamma_{i_t}(T n_{i_t})$ terms from the left side of the bound and see that with probability at least $1 - \kappa_1/(TK^3) - 2\kappa_1/(T^3K^3)$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} R(f_t) \leq R^* + \gamma_{i^*}(T n_{i^*}) + 2e\kappa_2 T^{-1/\beta}\sqrt{\log T}\left(\sum_{i=1}^{K}\frac{C}{n_i}\right)^{1/\beta} \\
+ \frac{2}{T}\sum_{t=1}^{T}\left[\gamma_i(n_{i_t} s_{i_t}(t)) + \kappa_2\sqrt{\frac{\log K}{n_{i_t} s_{i_t}(t)}} + \kappa_2\sqrt{\frac{\log T}{n_{i_t} s_{i_t}(t)}}\right]. \quad \text{(B.17)}
\end{aligned}$$

Defining $\widehat{f}_T := \frac{1}{T}\sum_{t=1}^{T} f_t$, we use Jensen's inequality to see that $R(\widehat{f}_T) \leq \frac{1}{T}\sum_{t=1}^{T} R(f_t)$. Thus, all that remains is to control the last sum in (B.17). Using the definition of γ_i, we

replace the sum with

$$\sum_{t=1}^{T} c_i n_{i_t}^{-\alpha} s_{i_t}(t)^{-\alpha} + n_{i_t}^{-\frac{1}{2}} s_{i_t}(t)^{-\frac{1}{2}} \kappa_2 \left[\sqrt{\log K} + \sqrt{\log T}\right]$$
$$\leq \sum_{t=1}^{T} \left[c_i n_{i_t}^{-\alpha} + \kappa_2 n_{i_t}^{-\frac{1}{2}} \sqrt{\log K} + \kappa_2 n_{i_t}^{-\frac{1}{2}} \sqrt{\log T}\right] s_{i_t}(t)^{-\min\{\alpha, \frac{1}{2}\}}.$$

Noting that

$$\sum_{t: i_t = i} s_{i_t}(t)^{-\min\{\alpha, \frac{1}{2}\}} = \sum_{t=1}^{T_i(T)} t^{-1/\beta} \leq C' T_i(T)^{1-1/\beta}$$

for some constant C' dependent on α, we can upper bound the last sum in (B.17) by

$$\sum_{t=1}^{T} \left[\gamma_i(n_{i_t} s_{i_t}(t)) + \kappa_2 \sqrt{\frac{\log K}{n_{i_t} s_{i_t}(t)}} + \kappa_2 \sqrt{\frac{\log T}{n_{i_t} s_{i_t}(t)}}\right]$$
$$\leq C' \sum_{i=1}^{K} \left[c_i n_i^{-\alpha} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log K} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log T}\right] T_i(T)^{1-1/\beta}. \tag{B.18}$$

Now that we have a sum of order K with terms $T_i(T)$ that are bounded by T, that is, $\sum_{i=1}^{K} T_i(T) = K$, we can apply Lemma B.3. Indeed, we set $p = 1 - 1/\beta = 1 - \min\{\alpha, \frac{1}{2}\}$ and $a_i = c_i n_i^{-\alpha} + \kappa_2 n_i^{-\frac{1}{2}} [\sqrt{\log K} + \sqrt{\log T}]$ in the lemma, and we see immediately that (B.18) is upper bounded by

$$C' T^{1-\min\{\alpha, \frac{1}{2}\}} \left(\sum_{i=1}^{K} \left[c_i n_i^{-\alpha} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log K} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log T}\right]^{\max\{1/\alpha, 2\}}\right)^{\min\{\alpha, \frac{1}{2}\}}.$$

Dividing by T completes the proof that the average $\widehat{f}_T$ has good risk properties with probability at least $1 - \kappa_1/(TK^3) - 2\kappa_1(T^3K^3) > 1 - 2\kappa_1/(TK^3)$.

Appendix C

Auxiliary results and proofs for Chapter 5

C.1 Auxiliary results for Theorem 5.1

In this appendix, we provide the proofs of various auxiliary lemmas required in the proof of Theorem 5.1.

C.1.1 Proof of Lemma 5.1

Since θ^t and $\widehat{\theta}$ are both feasible and $\widehat{\theta}$ lies on the constraint boundary, we have $\mathcal{R}(\theta^t) \leq \mathcal{R}(\widehat{\theta})$. Since $\mathcal{R}(\widehat{\theta}) \leq \mathcal{R}(\theta^*) + \mathcal{R}(\widehat{\theta} - \theta^*)$ by triangle inequality, we conclude that

$$\mathcal{R}(\theta^t) \leq \mathcal{R}(\theta^*) + \mathcal{R}(\Delta^*).$$

Since $\theta^* = \Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)$, a second application of triangle inequality yields

$$\mathcal{R}(\theta^t) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{C.1}$$

Now define the difference $\Delta^t := \theta^t - \theta^*$. (Note that this is slightly different from $\widehat{\Delta}^t$, which is measured relative to the optimum $\widehat{\theta}$.) With this notation, we have

$$\begin{aligned}
\mathcal{R}(\theta^t) &= \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^t) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) \\
&\overset{(i)}{\geq} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}\big(\Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^t)\big) \\
&\overset{(ii)}{\geq} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)),
\end{aligned}$$

where steps (i) and (ii) each use the triangle inequality. Now by the decomposability condition, we have $\mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t))$, so that we have shown

that

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\theta^t).$$

Combining this inequality with the earlier bound (C.1) yields

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*).$$

Re-arranging yields the inequality

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{C.2}$$

The final step is to translate this inequality into one that applies to the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$. Recalling that $\Delta^* = \widehat{\theta} - \theta^*$, we have $\widehat{\Delta}^t = \Delta^t - \Delta^*$, and hence

$$\mathcal{R}(\widehat{\Delta}^t) \leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*), \qquad \text{by triangle inequality.} \tag{C.3}$$

In addition, we have

$$\begin{aligned}\mathcal{R}(\Delta^t) \;\leq\; \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) &\overset{(i)}{\leq} 2\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\left\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\right\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*),\end{aligned}$$

where inequality (i) uses the bound (C.2), and inequality (ii) uses the definition (5.12) of the subspace compatibility Ψ. Combining with the inequality (C.3) yields

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\left\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\right\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\mathcal{R}(\Delta^*).$$

Since projection onto a subspace is non-expansive, we have $\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| \leq \|\Delta^t\|$, and hence

$$\left\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\right\| \leq \left\|\widehat{\Delta}^t + \Delta^*\right\| \;\leq\; \left\|\widehat{\Delta}^t\right\| + \|\Delta^*\|.$$

Combining the pieces, we obtain the claim (5.49).

C.1.2 Proof of Lemma 5.2

We start by applying the RSC assumption to the pair $\widehat{\theta}$ and θ^t, thereby obtaining the lower bound

$$\begin{aligned}\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 &\geq \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^t\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^t\rangle + \langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^{t+1}\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}).\end{aligned} \tag{C.4}$$

Here the second inequality follows by adding and subtracting terms.

Now for compactness in notation, define

$$\varphi_t(\theta) := \mathcal{L}_n(\theta^t) + \left\langle \nabla \mathcal{L}_n(\theta^t), \theta - \theta^t \right\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2,$$

and note that by definition of the algorithm, the iterate θ^{t+1} minimizes $\varphi_t(\theta)$ over the ball $\mathbb{B}_{\mathcal{R}}(\rho)$. Moreover, since $\widehat{\theta}$ is feasible, the first-order conditions for optimality imply that $\langle \nabla \varphi_t(\theta^{t+1}), \widehat{\theta} - \theta^{t+1} \rangle \geq 0$, or equivalently that $\langle \nabla \mathcal{L}_n(\theta^t) + \gamma_u(\theta^{t+1} - \theta^t), \widehat{\theta} - \theta^{t+1} \rangle \geq 0$. Applying this inequality to the lower bound (C.4), we find that

$$\begin{aligned}
\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 &\geq \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\
&= \varphi_t(\theta^{t+1}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\
&= \varphi_t(\theta^{t+1}) + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^t \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}),
\end{aligned} \tag{C.5}$$

where the last step follows from adding and subtracting θ^{t+1} in the inner product.

Now by the RSM condition, we have

$$\varphi_t(\theta^{t+1}) \geq \mathcal{L}_n(\theta^{t+1}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) \overset{(a)}{\geq} \mathcal{L}_n(\widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t), \tag{C.6}$$

where inequality (a) follows by the optimality of $\widehat{\theta}$, and feasibility of θ^{t+1}. Combining this inequality with the previous bound (C.5) yields that $\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2$ is lower bounded by

$$\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^t \rangle - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),$$

and the claim (5.51) follows after some simple algebraic manipulations.

C.2 Auxiliary results for Theorem 5.2

In this appendix, we prove the two auxiliary lemmas required in the proof of Theorem 5.2.

C.2.1 Proof of Lemma 5.3

This result is a generalization of an analogous result in Negahban et al. [116], with some changes required so as to adapt the statement to the optimization setting. Let θ be any vector, feasible for the problem (5.2), that satisfies the bound

$$\phi(\theta) \leq \phi(\theta^*) + \bar{\eta}, \tag{C.7}$$

and assume that $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*))$. We then claim that the error vector $\Delta := \theta - \theta^*$ satisfies the inequality

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)) \;\leq\; 3\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\}. \tag{C.8}$$

For the moment, we take this claim as given, returning later to verify its validity.

By applying this intermediate claim (C.8) in two different ways, we can complete the proof of Lemma 5.3. First, we observe that when $\theta = \widehat{\theta}$, the optimality of $\widehat{\theta}$ and feasibility of θ^* imply that assumption (C.7) holds with $\bar{\eta} = 0$, and hence the intermediate claim (C.8) implies that the statistical error $\Delta^* = \theta^* - \widehat{\theta}$ satisfies the bound

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \;\leq\; 3\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)). \tag{C.9}$$

Since $\Delta^* = \Pi_{\bar{\mathcal{M}}}(\Delta^*)\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)$, we can write

$$\mathcal{R}(\Delta^*) = \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \;\leq\; 4\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)), \tag{C.10}$$

using the triangle inequality in conjunction with our earlier bound (C.9). Similarly, when $\theta = \theta^t$ for some $t \geq T$, then the given assumptions imply that condition (C.7) holds with $\bar{\eta} > 0$, so that the intermediate claim (followed by the same argument with triangle inequality) implies that the error $\Delta^t = \theta^t - \theta^*$ satisfies the bound

$$\mathcal{R}(\Delta^t) \;\leq\; 4\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\}. \tag{C.11}$$

Now let $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ be the optimization error at time t, and observe that we have the decomposition $\widehat{\Delta}^t = \Delta^t + \Delta^*$. Consequently, by triangle inequality

$$\begin{aligned}
\mathcal{R}(\widehat{\Delta}^t) &\leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*) \\
&\overset{(i)}{\leq} 4\Big\{\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*))\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\} \\
&\overset{(ii)}{\leq} 4\Psi(\overline{\mathcal{M}})\Big\{\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + \|\Pi_{\bar{\mathcal{M}}}(\Delta^*)\|\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\} \\
&\overset{(iii)}{\leq} 4\Psi(\overline{\mathcal{M}})\Big\{\|\Delta^t\| + \|\Delta^*\|\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\},
\end{aligned} \tag{C.12}$$

where step (i) follows by applying both equation (C.10) and (C.11); step (ii) follows from the definition (5.12) of the subspace compatibility that relates the regularizer to the norm $\|\cdot\|$; and step (iii) follows from the fact that projection onto a subspace is non-expansive. Finally, since $\Delta^t = \widehat{\Delta}^t - \Delta^*$, the triangle inequality implies that $\|\Delta^t\| \leq \|\widehat{\Delta}^t\| + \|\Delta^*\|$. Substituting this upper bound into inequality (C.12) completes the proof of Lemma 5.3.

It remains to prove the intermediate claim (C.8). Letting θ be any vector, feasible for the program (5.2), and satisfying the condition (C.7), and let $\Delta = \theta - \theta^*$ be the associated error vector. Re-writing the condition (C.7), we have

$$\mathcal{L}_n(\theta^* + \Delta) + \lambda_n \mathcal{R}(\theta^* + \Delta) \;\leq\; \mathcal{L}_n(\theta^*) + \lambda_n \mathcal{R}(\theta^*) + \bar{\eta}.$$

Subtracting $\big\langle \nabla \mathcal{L}_n(\theta^*), \, \Delta \big\rangle$ from each side and then re-arranging yields the inequality

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \big\langle \nabla \mathcal{L}_n(\theta^*), \, \Delta \big\rangle + \lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \leq \; - \big\langle \nabla \mathcal{L}_n(\theta^*), \, \Delta \big\rangle + \bar{\eta}.$$

The convexity of $\mathcal{L}_n$ then implies that $\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \big\langle \nabla \mathcal{L}_n(\theta^*), \, \Delta \big\rangle \geq 0$, and hence that

$$\lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \;\leq\; - \big\langle \nabla \mathcal{L}_n(\theta^*), \, \Delta \big\rangle + \bar{\eta}.$$

Applying Hölder's inequality to $\big\langle \nabla \mathcal{L}_n(\theta^*), \, \Delta \big\rangle$, as expressed in terms of the dual norms $\mathcal{R}$ and $\mathcal{R}^*$, yields the upper bound

$$\lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \leq \; \mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*)) \, \mathcal{R}(\Delta) \; + \bar{\eta} \; \stackrel{(i)}{\leq} \; \frac{\lambda_n}{2} \, \mathcal{R}(\Delta) + \bar{\eta},$$

where step (i) uses the fact that $\lambda_n \geq 2 \mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*))$ by assumption.

For the remainder of the proof, let us introduce the convenient shorthand $\Delta_{\bar{\mathcal{M}}} := \Pi_{\bar{\mathcal{M}}}(\Delta)$ and $\Delta_{\bar{\mathcal{M}}^\perp} := \Pi_{\bar{\mathcal{M}}^\perp}(\Delta)$, with similar shorthand for projections involving θ^*. Making note of the decomposition $\Delta = \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}$, an application of triangle inequality then yields the upper bound

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \leq \; \frac{1}{2} \Big\{ \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \Big\} + \frac{\bar{\eta}}{\lambda_n}, \tag{C.13}$$

where we have rescaled both sides by $\lambda_n > 0$.

It remains to further lower bound the left-hand side (C.13). By triangle inequality, we have

$$-\mathcal{R}(\theta^*) \geq -\mathcal{R}(\theta^*_{\mathcal{M}}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}). \tag{C.14}$$

Let us now write $\theta^* + \Delta = \theta^*_{\mathcal{M}} + \theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}$. Using this representation and triangle inequality, we have

$$\mathcal{R}(\theta^* + \Delta) \geq \mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}}) \;\geq\; \mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}).$$

Finally, since $\theta^*_{\mathcal{M}} \in \mathcal{M}$ and $\Delta_{\bar{\mathcal{M}}^\perp} \in \overline{\mathcal{M}}^\perp$, the decomposability of $\mathcal{R}$ implies that $\mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) = \mathcal{R}(\theta^*_{\mathcal{M}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp})$, and hence that

$$\mathcal{R}(\theta^* + \Delta) \geq \mathcal{R}(\theta^*_{\mathcal{M}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \tag{C.15}$$

Adding together equations (C.14) and (C.15), we obtain the lower bound

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \tag{C.16}$$

Combining this lower bound with the earlier inequality (C.13), some algebra yields the bound

$$\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) + 2\frac{\eta}{\lambda_n},$$

corresponding to the bound (C.8) when η/λ_n achieves the final minimum. To obtain the final term involving $\overline{\rho}$ in the bound (C.8), two applications of triangle inequality yields

$$\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta) \;\leq\; \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 2\overline{\rho},$$

where we have used the fact that $\mathcal{R}(\Delta) \leq \mathcal{R}(\theta) + \mathcal{R}(\theta^*) \leq 2\overline{\rho}$, since both θ and θ^* are feasible for the program (5.2).

C.2.2 Proof of Lemma 5.4

The proof of this result follows lines similar to the proof of convergence by Nesterov [121]. Recall our notation $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n\mathcal{R}(\theta)$, $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, and that $\eta^t_\phi = \phi(\theta^t) - \phi(\widehat{\theta})$. We begin by proving that under the stated conditions, a useful version of restricted strong convexity (5.47) is in force:

Lemma C.1. *Under the assumptions of Lemma 5.4, we are guaranteed that*

$$\Big\{\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\Big\}\|\widehat{\Delta}^t\|^2 \leq 2\tau_\ell(\mathcal{L}_n)\, v^2 \;+\; \phi(\theta^t) - \phi(\widehat{\theta}), \quad \textit{and} \tag{C.17a}$$

$$\Big\{\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\Big\}\|\widehat{\Delta}^t\|^2 \leq 2\,\tau_\ell(\mathcal{L}_n)\, v^2 \;+\; \mathcal{T}_{\mathcal{L}}(\widehat{\theta}; \theta^t), \tag{C.17b}$$

where $v := \bar{\epsilon}_{stat} + 2\min(\frac{\bar{\eta}}{\lambda_n}, \overline{\rho})$.

See Appendix C.2.3 for the proof of this claim. So as to ease notation in the remainder of the proof, let us introduce the shorthand

$$\phi_t(\theta) := \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \theta - \theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta), \tag{C.18}$$

corresponding to the approximation to the regularized loss function ϕ that is minimized at iteration t of the update (5.4). Since θ^{t+1} minimizes ϕ_t over the set $\mathbb{B}_{\mathcal{R}}(\overline{\rho})$, we are guaranteed that $\phi_t(\theta^{t+1}) \leq \phi_t(\theta)$ for all $\theta \in \mathbb{B}_{\mathcal{R}}(\overline{\rho})$. In particular, for any $\alpha \in (0,1)$, the vector $\theta_\alpha = \alpha\widehat{\theta} + (1-\alpha)\theta^t$ lies in the convex set $\mathbb{B}_{\mathcal{R}}(\overline{\rho})$, so that

$$\begin{aligned}
\phi_t(\theta^{t+1}) \;\leq\; \phi_t(\theta_\alpha) &= \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \theta_\alpha - \theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta_\alpha - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta_\alpha) \\
&\overset{(i)}{=} \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} - \alpha\theta^t\big\rangle + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta_\alpha) \\
&\overset{(ii)}{\leq} \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} - \alpha\theta^t\big\rangle + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \lambda_n\alpha\mathcal{R}(\widehat{\theta}) + \lambda_n(1-\alpha)\mathcal{R}(\theta^t),
\end{aligned}$$

where step (i) follows from substituting the definition of θ_α, and step (ii) uses the convexity of the regularizer $\mathcal{R}$.

Now, the stated conditions of the lemma ensure that $\gamma_\ell/2 - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}) \geq 0$, so that by equation (C.17b), we have $\mathcal{L}_n(\widehat{\theta}) + 2\tau_\ell(\mathcal{L}_n)v^2 \geq \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^t \rangle$. Substituting back into our earlier bound yields

$$\begin{aligned}\phi_t(\theta^{t+1}) &\leq (1-\alpha)\mathcal{L}_n(\theta^t) + \alpha\mathcal{L}_n(\widehat{\theta}) + 2\alpha\tau_\ell(\mathcal{L}_n)v^2 + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \alpha\lambda_n\mathcal{R}(\widehat{\theta}) + (1-\alpha)\lambda_n\mathcal{R}(\theta^t) \\ &\overset{(iii)}{=} \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + 2\tau_\ell(\mathcal{L}_n)v^2 + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2, \end{aligned} \tag{C.19}$$

where we have used the definition of ϕ and $\alpha \leq 1$ in step (iii).

In order to complete the proof, it remains to relate $\phi_t(\theta^{t+1})$ to $\phi(\theta^{t+1})$, which can be performed by exploiting restricted smoothness. In particular, applying the RSM condition at the iterate θ^{t+1} in the direction θ^t yields the upper bound

$$\mathcal{L}_n(\theta^{t+1}) \leq \mathcal{L}_n(\theta^t) + \langle \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),$$

so that

$$\begin{aligned}\phi(\theta^{t+1}) &\leq \mathcal{L}_n(\theta^t) + \langle \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) + \lambda_n\mathcal{R}(\theta^{t+1}) \\ &= \phi_t(\theta^{t+1}) + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t).\end{aligned}$$

Combining the above bound with the inequality (C.19) and recalling the notation $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, we obtain

$$\begin{aligned}\phi(\theta^{t+1}) &\leq \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) + 2\tau_\ell(\mathcal{L}_n)v^2 \\ &\overset{(iv)}{\leq} \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\Delta}^t\|^2 + \tau_u(\mathcal{L}_n)[\mathcal{R}(\widehat{\Delta}^{t+1}) + \mathcal{R}(\widehat{\Delta}^t)]^2 + 2\tau_\ell(\mathcal{L}_n)v^2 \\ &\overset{(v)}{\leq} \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\Delta}^t\|^2 + 2\tau_u(\mathcal{L}_n)(\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \mathcal{R}^2(\widehat{\Delta}^t)) + 2\tau_\ell(\mathcal{L}_n)v^2.\end{aligned} \tag{C.20}$$

Here step (iv) uses the fact that $\theta^t - \theta^{t+1} = \widehat{\Delta}^t - \widehat{\Delta}^{t+1}$ and applies triangle inequality to the norm $\mathcal{R}$, whereas step (v) follows from Cauchy-Schwarz inequality.

Next, combining Lemma 5.3 with the Cauchy-Schwarz inequality inequality yields the upper bound

$$\mathcal{R}^2(\widehat{\Delta}^t) \ \leq 32\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 2v^2 \tag{C.21}$$

where $v = \bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}}) + 2\min(\frac{\bar{\eta}}{\lambda_n}, \overline{\rho})$, is a constant independent of θ^t and $\bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}})$ was previously defined in the lemma statement. Substituting the above bound into inequal-

ity (C.20) yields that $\phi(\theta^{t+1})$ is at most

$$\phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^{t+1}\|^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 8\tau_u(\mathcal{L}_n)v^2 + 2\tau_\ell(\mathcal{L}_n)v^2. \quad \text{(C.22)}$$

The final step is to translate quantities involving $\widehat{\Delta}^t$ to functional values, which may be done using the RSC condition (C.17a) from Lemma C.1. In particular, combining the RSC condition (C.17a) with the inequality (C.22) yields

$$\phi(\theta^{t+1}) \leq \phi(\theta^t) - \alpha\eta_\phi^t + \frac{\left(\gamma_u\alpha^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\right)}{\overline{\gamma_\ell}}(\eta_\phi^t + 2\tau_\ell(\mathcal{L}_n)v^2) + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}(\eta_\phi^{t+1} + 2\tau_\ell(\mathcal{L}_n)v^2) + 8\tau_u(\mathcal{L}_n)v^2 + 2\tau_\ell(\mathcal{L}_n)v^2.$$

where we have introduced the shorthand $\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})$. Recalling the definition of β, adding and subtracting $\phi(\widehat{\theta})$ from both sides, and choosing $\alpha = \frac{\overline{\gamma_\ell}}{2\gamma_u} \in (0,1)$, we obtain

$$\left(1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\eta_\phi^{t+1} \leq \left(1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\eta_\phi^t + \beta(\overline{\mathcal{M}})v^2.$$

Recalling the definition of the contraction factor κ from the statement of Theorem 5.2, the above expression can be rewritten as

$$\eta_\phi^{t+1} \leq \kappa\eta_\phi^t + \beta(\overline{\mathcal{M}})\xi(\overline{\mathcal{M}})v^2, \quad \text{where } \xi(\mathcal{M}) = \left\{1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right\}^{-1}.$$

Finally, iterating the above expression yields $\eta_\phi^t \leq \kappa^{t-T}\eta_\phi^T + \frac{\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}})v^2}{1-\kappa}$, where we have used the condition $\kappa \in (0,1)$ in order to sum the geometric series, thereby completing the proof.

C.2.3 Proof of Lemma C.1

The key idea to prove the lemma is to use the definition of RSC along with the iterated cone bound of Lemma 5.3 for simplifying the error terms in RSC.

Let us first show that condition (C.17a) holds. From the RSC condition assumed in the lemma statement, we have

$$\mathcal{L}_n(\theta^t) - \mathcal{L}_n(\widehat{\theta}) - \langle\nabla\mathcal{L}_n(\widehat{\theta}),\, \theta^t - \widehat{\theta}\rangle \geq \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\widehat{\theta} - \theta^t). \quad \text{(C.23)}$$

From the convexity of $\mathcal{R}$ and definition of the subdifferential $\partial\mathcal{R}(\theta)$, we obtain

$$\mathcal{R}(\theta^t) - \mathcal{R}(\widehat{\theta}) - \left\langle\partial\mathcal{R}(\widehat{\theta}),\, \theta^t - \widehat{\theta}\right\rangle \geq 0.$$

Adding this lower bound with the inequality (C.23) yields

$$\phi(\theta^t) - \phi(\widehat{\theta}) - \langle \nabla \phi(\widehat{\theta}), \, \theta^t - \widehat{\theta} \rangle \geq \frac{\gamma_\ell}{2} \|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n) \, \mathcal{R}^2(\widehat{\theta} - \theta^t),$$

where we recall that $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n \mathcal{R}(\theta)$ is our objective function. By the optimality of $\widehat{\theta}$ and feasibility of θ^t, we are guaranteed that $\langle \nabla \phi(\widehat{\theta}), \, \theta^t - \widehat{\theta} \rangle \geq 0$, and hence

$$\begin{aligned}
\phi(\theta^t) - \phi(\widehat{\theta}) &\geq \frac{\gamma_\ell}{2} \|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n) \, \mathcal{R}^2(\widehat{\theta} - \theta^t) \\
&\stackrel{(i)}{\geq} \frac{\gamma_\ell}{2} \|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n) \;\; \big\{ 32 \Psi^2(\overline{\mathcal{M}}) \|\widehat{\theta} - \theta^t\|^2 + 2v^2 \big\}
\end{aligned}$$

where step (i) follows by applying Lemma 5.3. Some algebra then yields the claim (C.17a).

Finally, let us verify the claim (C.17b). Using the RSC condition, we have

$$\mathcal{L}_n(\widehat{\theta}) - \mathcal{L}_n(\theta^t) - \langle \nabla \mathcal{L}_n(\theta^t), \, \widehat{\theta} - \theta^t \rangle \geq \frac{\gamma_\ell}{2} \|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n) \, \mathcal{R}^2(\widehat{\theta} - \theta^t). \tag{C.24}$$

As before, applying Lemma 5.3 yields

$$\underbrace{\mathcal{L}_n(\widehat{\theta}) - \mathcal{L}_n(\theta^t) - \langle \nabla \mathcal{L}_n(\theta^t), \, \widehat{\theta} - \theta^t \rangle}_{\mathcal{T}_{\mathcal{L}}(\widehat{\theta}; \theta^t)} \geq \frac{\gamma_\ell}{2} \|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n) \;\; \Big(32 \Psi^2(\overline{\mathcal{M}}) \|\widehat{\theta} - \theta^t\|^2 + 2v^2 \Big),$$

and rearranging the terms and establishes the claim (C.17b).

C.3 Proof of Lemma 5.5

Given the condition $\mathcal{R}(\widehat{\theta}) \leq \rho \leq \mathcal{R}(\theta^*)$, we have $\mathcal{R}(\widehat{\theta}) = \mathcal{R}(\theta^* + \Delta^*) \leq \mathcal{R}(\theta^*)$. By triangle inequality, we have

$$\mathcal{R}(\theta^*) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)) \;\; \leq \;\; \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)).$$

We then write

$$\begin{aligned}
\mathcal{R}(\theta^* + \Delta^*) &= \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \\
&\stackrel{(i)}{\geq} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\
&\stackrel{(ii)}{=} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),
\end{aligned}$$

where the bound (i) follows by triangle inequality, and step (ii) uses the decomposability of $\mathcal{R}$ over the pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. By combining this lower bound with the previously established upper bound

$$\mathcal{R}(\theta^* + \Delta^*) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),$$

we conclude that $\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$. Finally, by triangle inequality, we have $\mathcal{R}(\Delta^*) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*))$, and hence

$$\begin{aligned}\mathcal{R}(\Delta^*) &\leq 2\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(i)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\,\|\Pi_{\bar{\mathcal{M}}}(\Delta^*)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\,\|\Delta^*\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where inequality (i) follows from Definition 5.4 of the subspace compatibility Ψ, and the bound (ii) follows from non-expansivity of projection onto a subspace.

C.4 A general result on Gaussian observation operators

In this appendix, we state a general result about a Gaussian random matrices, and show how it can be adapted to prove Lemmas 5.6 and 5.7. Let $X \in \mathbb{R}^{n\times d}$ be a Gaussian random matrix with i.i.d. rows $x_i \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d\times d}$ is a covariance matrix. We refer to X as a sample from the Σ-Gaussian ensemble. In order to state the result, we use $\Sigma^{1/2}$ to denote the symmetric matrix square root.

Proposition C.1. *Given a random matrix X drawn from the Σ-Gaussian ensemble, there are universal constants c_i, $i = 0, 1$ such that*

$$\frac{\|X\theta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\theta\|_2^2 - c_1\frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n}\mathcal{R}^2(\theta) \qquad \text{and} \tag{C.25a}$$

$$\frac{\|X\theta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\theta\|_2^2 + c_1\frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n}\mathcal{R}^2(\theta) \qquad \text{for all } \theta \in \mathbb{R}^d \tag{C.25b}$$

with probability greater than $1 - \exp(-c_0\, n)$.

We omit the proof of this result. The two special instances proved in Lemma 5.6 and 5.7 have been proved in the papers [134] and [115] respectively. We now show how Proposition C.1 can be used to recover various lemmas required in our proofs.

Proof of Lemma 5.6: We begin by establishing this auxiliary result required in the proof of Corollary 5.2. When $\mathcal{R}(\cdot) = \|\cdot\|_1$, we have $\mathcal{R}^*(\cdot) = \|\cdot\|_\infty$. Moreover, the random vector $x_i \sim N(0, \Sigma)$ can be written as $x_i = \Sigma^{1/2}w$, where $w \sim N(0, I_{d\times d})$ is standard normal. Consequently, using properties of Gaussian maxima [99] and defining $\zeta(\Sigma) = \max_{j=1,2,\ldots,d}\Sigma_{jj}$, we have the bound

$$(\mathbb{E}[\|x_i\|_\infty])^2 \;\leq\; \zeta(\Sigma)\,(\mathbb{E}[\|w\|_\infty])^2 \;\leq\; 3\zeta(\Sigma)\,\sqrt{\log d}.$$

Substituting into Proposition C.1 yields the claims (5.61a) and (5.61b).

Proof of Lemma 5.7: In order to prove this claim, we view each random observation matrix $X_i \in \mathbb{R}^{d\times d}$ as a $d = d^2$ vector (namely the quantity $\text{vec}(X_i)$), and apply Proposition C.1 in this vectorized setting. Given the standard Gaussian vector $w \in \mathbb{R}^{d^2}$, we let $W \in \mathbb{R}^{d\times d}$ be the random matrix such that $\text{vec}(W) = w$. With this notation, the term $\mathcal{R}^*(\text{vec}(X_i))$ is equivalent to the operator norm $|||X_i|||_{\text{op}}$. As shown in Negahban and Wainwright [115], $\mathbb{E}[|||X_i|||_{\text{op}}] \leq 24\zeta_{\text{mat}}(\Sigma)\sqrt{d}$, where ζ_{mat} was previously defined (5.64).

C.5 Auxiliary results for Corollary 5.5

In this section, we provide the proofs of Lemmas 5.8 and 5.9 that play a central role in the proof of Corollary 5.5. In order to do so, we require the following result, which is a re-statement of a theorem due to Negahban and Wainwright [114]:

Proposition C.2. *For the matrix completion operator $\mathfrak{X}_n$, there are universal positive constants (c_1, c_2) such that*

$$\left|\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n} - |||\Theta|||_F^2\right| \leq c_1\, d\|\Theta\|_\infty\, |||\Theta|||_1\sqrt{\frac{d\log d}{n}} \;+\; c_2\left(d\|\Theta\|_\infty\sqrt{\frac{d\log d}{n}}\right)^2 \qquad \text{for all } \Theta \in \mathbb{R}^{d\times d} \tag{C.26}$$

with probability at least $1 - \exp(-d\log d)$.

C.5.1 Proof of Lemma 5.8

Applying Proposition C.2 to $\widehat{\Delta}^t$ and using the fact that $d\|\widehat{\Delta}^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq |||\widehat{\Delta}^t|||_F^2 - c_1\alpha|||\widehat{\Delta}^t|||_1\sqrt{\frac{d\log d}{n}} - c_2\,\alpha^2\frac{d\log d}{n}, \tag{C.27}$$

where we recall our convention of allowing the constants to change from line to line. From Lemma 5.1,

$$|||\widehat{\Delta}^t|||_1 \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\,|||\widehat{\Delta}^t|||_F + 2|||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_1 + 2|||\Delta^*|||_1 + \Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F.$$

Since $\rho \leq |||\Theta^*|||_1$, Lemma 5.5 implies that $|||\Delta^*|||_1 \leq 2\Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F + |||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_1$, and hence that

$$|||\widehat{\Delta}^t|||_1 \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\,|||\widehat{\Delta}^t|||_F + 4|||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_1 + 5\Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F. \tag{C.28}$$

Combined with the lower bound, we obtain that $\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n}$ is lower bounded by

$$|||\widehat{\Delta}^t|||_F^2\left\{1 - \frac{2c_1\,\alpha\Psi(\overline{\mathcal{M}}^\perp)\sqrt{\frac{d\log d}{n}}}{|||\widehat{\Delta}^t|||_F}\right\} - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\left\{4|||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_1 + 5\Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F\right\} - c_2\,\alpha^2\frac{d\log d}{n}.$$

Consequently, for all iterations such that $\|\widehat{\Delta}^t\|_F \geq 4c_1\Psi(\overline{\mathcal{M}}^\perp)\sqrt{\frac{d\log d}{n}}$, we have

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 + 5\Psi(\overline{\mathcal{M}}^\perp)\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n}.$$

By subtracting off an additional term, the bound is valid for all $\widehat{\Delta}^t$—viz.

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 + 5\Psi(\overline{\mathcal{M}}^\perp)\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n} - 16c_1^2\alpha^2\Psi^2(\overline{\mathcal{M}}^\perp)\frac{d\log d}{n}.$$

C.5.2 Proof of Lemma 5.9

Applying Proposition C.2 to Γ^t and using the fact that $d\|\Gamma^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq \|\Gamma^t\|_F^2 + c_1\alpha\|\Gamma^t\|_1\sqrt{\frac{d\log d}{n}} + c_2\,\alpha^2\frac{d\log d}{n}, \tag{C.29}$$

where we recall our convention of allowing the constants to change from line to line. By triangle inequality, we have $\|\Gamma^t\|_1 \leq \|\Theta^t - \widehat{\Theta}\|_1 + \|\Theta^{t+1} - \widehat{\Theta}\|_1 = \|\widehat{\Delta}^t\|_1 + \|\widehat{\Delta}^{t+1}\|_1$. Equation C.28 gives us bounds on $\|\widehat{\Delta}^t\|_1$ and $\|\widehat{\Delta}^{t+1}\|_1$. Substituting them into the upper bound (C.29) yields the claim.

Appendix D

Technical proofs for Chapter 6

In this appendix, we collect several useful results about proximal functions and continuity properties of the solutions of proximal operators. We begin with results useful for the dual-averaging updates (6.3) and (6.8). We define the proximal dual function

$$\psi_\alpha^*(\mu) := \sup_{\theta \in \Omega} \left\{ \langle -\mu, \theta \rangle - \frac{1}{\alpha} \psi(\theta) \right\}. \tag{D.1}$$

Since $\nabla \psi_\alpha^*(\mu) = \operatorname{argmax}_{\theta \in \Omega} \{\langle -\mu, \theta \rangle - \alpha^{-1} \psi(\theta)\}$, it is clear that $\theta^t = \nabla \psi_{\alpha(t)}^*(\mu^t)$. Further, by the strong convexity of ψ, we have that $\nabla \psi_\alpha^*(\mu)$ is α-Lipschitz continuous [122, 77, Chapter X], that is, for the norm $\|\cdot\|$ with respect to which ψ is strongly convex and its associated dual norm $\|\cdot\|_*$,

$$\left\| \nabla \psi_\alpha^*(\mu') - \nabla \psi_\alpha^*(\mu) \right\| \le \alpha \|\mu' - \mu\|_*. \tag{D.2}$$

We will find one more result about solutions to the dual averaging update useful. This result has essentially been proven in many contexts [122, 159, 57].

Lemma D.1. *Let θ^+ minimize $\langle \mu, \theta \rangle + A\psi(\theta)$ for all $\theta \in \Omega$. Then for any $\theta \in \Omega$,*

$$\langle \mu, \theta \rangle + A\psi(\theta) \ge \left\langle \mu, \theta^+ \right\rangle + A\psi(\theta^+) + AD_\psi(\theta, \theta^+)$$

Now we turn to describing properties of the mirror-descent step (6.4), which we will also use frequently. The lemma allows us to bound differences between θ^t and θ^{t+1} for the mirror-descent family of algorithms.

Lemma D.2. *Let θ^+ minimize $\langle g, \theta \rangle + \frac{1}{\alpha} D_\psi(\theta, \tilde{\theta})$ over $\theta \in \Omega$. Then $\left\| \theta^+ - \tilde{\theta} \right\| \le \alpha \|g\|_*$.*

Proof. The inequality is clear when $\theta^+ = \tilde{\theta}$, so assume that $\theta^+ \ne \tilde{\theta}$. Since θ^+ minimizes $\langle g, \theta \rangle + \frac{1}{\alpha} D_\psi(\theta, \tilde{\theta})$, the first order conditions for optimality imply

$$\left\langle \alpha g + \nabla \psi(\theta^+) - \nabla \psi(\tilde{\theta}), \theta - \theta^+ \right\rangle \ge 0$$

for any $\theta \in \Omega$. Thus we can choose $\tilde{\theta} = \theta$ and see that

$$\alpha \left\langle g, \tilde{\theta} - \theta \right\rangle \geq \left\langle \nabla\psi(\theta^+) - \nabla\psi(\tilde{\theta}), \theta^+ - \tilde{\theta} \right\rangle \geq \left\| \theta^+ - \tilde{\theta} \right\|^2,$$

where the last inequality follows from the strong convexity of ψ. Using Hölder's inequality gives that $\alpha \|g\|_* \left\| \tilde{\theta} - \theta \right\| \geq \left\| \theta^+ - \tilde{\theta} \right\|^2$, and dividing by $\left\| \tilde{\theta} - \theta \right\|$ completes the proof. □

The last technical lemma we give explicitly bounds the differences between θ^t and $\theta^{(t+\tau)}$, for some $\tau \geq 1$, by using the above continuity lemmas.

Lemma D.3. *Let Assumption A hold. Define θ^t via the dual-averaging updates (6.3), (6.8), or (6.11) or the mirror-descent updates (6.4), (6.9), or (6.12). Let $\alpha(t)^{-1} = L + \eta(t + t_0)^c$ for some $c \in [0, 1]$, $\eta > 0$, $t_0 \geq 0$, and $L \geq 0$. Then for any fixed τ,*

$$\mathbb{E}[\left\| \theta^t - \theta^{(t+\tau)} \right\|^2] \leq \frac{4G^2\tau^2}{\eta^2(t + t_0)^{2c}} \quad \text{and} \quad \mathbb{E}[\left\| \theta^t - \theta^{(t+\tau)} \right\|] \leq \frac{2G\tau}{\eta(t + t_0)^c}.$$

Proof. We first show the lemma for the dual-averaging updates. Recall that $\theta^t = \nabla\psi^*_{\alpha(t)}(\mu^t)$ and $\nabla\psi^*_\alpha$ is α-Lipschitz continuous. Using the triangle inequality,

$$\begin{aligned}
\left\| \theta^t - \theta^{(t+\tau)} \right\| &= \left\| \nabla\psi^*_{\alpha(t)}(\mu^t) - \nabla\psi^*_{\alpha(t+\tau)}(\mu^{(t+\tau)}) \right\| \\
&= \left\| \nabla\psi^*_{\alpha(t)}(\mu^t) - \nabla\psi^*_{\alpha(t+\tau)}(\mu^t) + \nabla\psi^*_{\alpha(t+\tau)}(\mu^t) - \nabla\psi^*_{\alpha(t+\tau)}(\mu^{(t+\tau)}) \right\| \\
&\leq \left\| \nabla\psi^*_{\alpha(t)}(\mu^t) - \nabla\psi^*_{\alpha(t+\tau)}(\mu^t) \right\| + \left\| \nabla\psi^*_{\alpha(t+\tau)}(\mu^t) - \nabla\psi^*_{\alpha(t+\tau)}(\mu^{(t+\tau)}) \right\| \\
&\leq (\alpha(t) - \alpha(t+\tau))\|\mu^t\|_* + \alpha(t+\tau)\|\mu^t - \mu^{(t+\tau)}\|_*. \qquad \text{(D.3)}
\end{aligned}$$

It is easy to check that for $c \in [0, 1]$,

$$\alpha(t) - \alpha(t+\tau) \leq \frac{c\eta\tau}{(L + \eta t^c)^2 t^{1-c}} \leq \frac{c\tau}{\eta t^{1+c}}.$$

By convexity of $\|\cdot\|_*^2$, we can bound $\mathbb{E}[\|\mu^t - \mu^{(t+\tau)}\|_*^2]$:

$$\mathbb{E}[\|\mu^t - \mu^{(t+\tau)}\|_*^2] = \tau^2\mathbb{E}\left[\left\| \frac{1}{\tau}\sum_{s=1}^{\tau} \mu^{(t+s)} - \mu^{(t+s-1)} \right\|_*^2\right] = \tau^2\mathbb{E}\left[\left\| \frac{1}{\tau}\sum_{s=0}^{\tau-1} g(s) \right\|_*^2\right] \leq \tau^2 G^2,$$

since $\mathbb{E}[\|\partial F(\theta; z)\|_*^2] \leq G^2$ by assumption. Thus, bound (D.3) gives

$$\begin{aligned}
\mathbb{E}[\left\| \theta^t - \theta^{(t+\tau)} \right\|^2] &\leq 2(\alpha(t) - \alpha(t+\tau))^2\mathbb{E}[\|\mu^t\|_*^2] + 2\alpha(t+\tau)^2\mathbb{E}[\|\mu^t - \mu^{(t+\tau)}\|_*^2] \\
&\leq \frac{2c^2t^2\tau^2G^2}{\eta^2t^{2+2c}} + 2G^2\tau^2\alpha(t+\tau)^2 = \frac{2c^2\tau^2G^2}{\eta^2t^{2c}} + \frac{2G^2\tau^2}{(L + \eta(t+\tau)^c)^2},
\end{aligned}$$

where we use Cauchy-Schwarz inequality in the first step. Since $c \leq 1$, the last term is clearly bounded by $4G^2\tau^2/\eta^2 t^{2c}$.

To get the slightly tighter bound on the first moment in the statement of the lemma, simply use the triangle inequality from the bound (D.3) and that $\sqrt{\mathbb{E}X^2} \geq \mathbb{E}|X|$.

The proof for the mirror-descent family of updates is similar. We focus on non-delayed update (6.4), as the other updates simply modify the indexing of $g(t+s)$ below. We know from Lemma D.2 and the triangle inequality that

$$\left\|\theta^t - \theta^{(t+\tau)}\right\| \leq \sum_{s=1}^{\tau} \left\|\theta^{(t+s)} - \theta^{(t+s-1)}\right\| \leq \sum_{s=1}^{\tau} \alpha(t+s-1)\|g(t+s)\|_*$$

Squaring the above bound, taking expectations, and recalling that $\alpha(t)$ is non-increasing, we see

$$\begin{aligned}\mathbb{E}[\left\|\theta^t - \theta^{(t+\tau)}\right\|^2] &\leq \sum_{s=1}^{\tau}\sum_{r=1}^{\tau} \alpha(t+s)\alpha(t+r)\mathbb{E}[\|g(t+s)\|_*\|g(t+r)\|_*] \\ &\leq \tau^2\alpha(t)^2 \max_{r,s} \sqrt{\mathbb{E}[\|g(t+s)\|_*^2]}\sqrt{\mathbb{E}[\|g(t+r)\|_*^2]} \leq \tau^2\alpha(t)^2 G^2\end{aligned}$$

by Hölder's inequality. Substituting the appropriate value for $\alpha(t)$ completes the proof. □

Bibliography

[1] J. Abernethy, F. Bach, T. Evgeniou, and J. Stein. Low-rank matrix factorization with attributes. Technical Report Technical Report N-24/06/MM, Ecole des mines de Paris, France, September 2006.

[2] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. URL http://arxiv.org/abs/1104.5525, 2011.

[3] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In *NIPS*, 2011.

[4] A. Agarwal, S. N. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems 23*, pages 37–45. 2010.

[5] A. Agarwal, J. C. Duchi, P. L. Bartlett, and C. Levrard. Oracle inequalities for computationally budgeted model selection. In *Proceedings of the Conference on Learning Theory (COLT2011)*, volume 19, pages 69–86, 2011.

[6] A. Agarwal, P.L. Bartlett, P. Ravikumar, and M.J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 2012. To appear in.

[7] A. Agarwal, O. Chapelle, M. Dudik, and J. Langford. A reliable effective terascale linear learning system. 2012. URL http://arxiv.org/abs/1110.4198.

[8] A. Agarwal, S. N. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 2012. To appear in.

[9] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[10] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semdefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 5B:2877–2921, 2009.

[11] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[12] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

[13] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002. ISSN 0885-6125.

[14] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Kluwer Academic, 1991.

[15] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[16] P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[17] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[18] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[19] Peter L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(2):545–552, 2008.

[20] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[21] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):pp. 1497–1537, 2005.

[22] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[23] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[24] S. Becker, J. Bobin, and E. J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. Technical report, Stanford University, 2009.

[25] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

[26] D. P. Bertsekas. Distributed asynchronous computation of fixed points. *Mathematical Programming*, 27:107–120, 1983.

[27] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

[28] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

[29] Dimitris Bertsimas and Santosh Vempala. Solving convex programs by random walks. *J. ACM*, 51(4):540–556, 2004.

[30] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[31] L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. *Z. Wahrsch. verw. Gebiete*, 65:181–327, 1983.

[32] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. NIPS Foundation (http://books.nips.cc), 2008.

[33] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer Berlin / Heidelberg, 2004.

[34] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

[35] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, volume 3. 2011.

[37] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, 2008.

[38] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[39] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[40] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194, 2007.

[41] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[42] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, December 2005.

[43] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory*, 52(2):489–509, February 2004.

[44] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[45] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? Technical report, Stanford, 2009. URL http://arxiv.org/pdf/0912.3599v1. available at arXiv:0912.3599.

[46] F. P. Cantelli. Sulla determinazione empirica della legge di probabilita. *Giorn. Itnst. Ital. Attuari*, 4:421–424, 1933.

[47] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[48] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Info. Theory*, 50(9):2050–2057, September 2004.

[49] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. Technical report, MIT, June 2009. Available at `arXiv:0906.2220v1`.

[50] V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex algebraic geometry of linear inverse problems. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 699 –703, 2010.

[51] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[52] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[53] Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, pages 537–546, 2008.

[54] O. Dekel and Y. Singer. Support vector machines on a budget. In *NIPS*, 2006.

[55] O. Dekel, S. Shalev-Shwartz, and Y. Singer. The forgetron: A kernel-based perceptron on a budget. *SIAM J. Comput.*, 37(5):1342–1372, January 2008.

[56] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Robust distributed online prediction. URL http://arxiv.org/abs/1012.1370, 2010. URL http://arxiv.org/abs/1012.1370.

[57] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

[58] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[59] D. L. Donoho and I. M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Prob. Theory and Related Fields*, 99:277–303, 1994.

[60] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, 2008.

[61] J. Duchi, A. Agarwal, and M. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592 –606, 2012.

[62] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[63] Richard M. Dudley. *Uniform Central Limit Theorems*. Cambridge Univ. Press, 1999.

[64] R. L. Dykstra. An iterative procedure for obtaining i-projections onto the intersection of convex sets. *Annals of Probability*, 13(3):975–984, 1985.

[65] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. In *Proceedings of the first annual workshop on Computational learning theory*, COLT '88, pages 139–154, 1988.

[66] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

[67] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford, 2002. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

[68] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, New York, NY, USA, 2009. ACM.

[69] S. Geman and C. R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:401–414, 1982.

[70] V. Glivenko. Sulla determinazione empirica della legge di probabilita. *Giorn. Itnst. Ital. Attuari*, 4:92–99, 1933.

[71] E. T. Hale, Y. Wotao, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. on Optimization*, 19(3):1107–1130, 2008.

[72] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[73] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[74] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *COLT*, 2010.

[75] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007. ISSN 0885-6125.

[76] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume I. Springer-Verlag, New York, 1993.

[77] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume II. Springer-Verlag, New York, 1993.

[78] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. Technical Report 11, 2011.

[79] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38 (4):1978–2004, 2010.

[80] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning*, New York, NY, USA, 2009. ACM.

[81] B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009.

[82] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Technical report, Universite Joseph Fourier, 2010. URL http://hal.archives-ouvertes.fr/hal-00508933/.

[83] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with the stochastic mirror-prox algorithm. URL http://arxiv.org/abs/0809.0815, 2008. URL http://arxiv.org/abs/0809.0815.

[84] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, 2009.

[85] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[86] M. J. Kearns. *The computational complexity of machine learning*. PhD thesis, Harvard University, 1989.

[87] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, USA, 1994.

[88] A. Kolmogorov and B. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Uspekhi Mat. Nauk.*, 86:3–86, 1959. Appeared in English as Amer. Math. Soc. Translations, 17:277–364, 1961.

[89] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39:2302–2329, 2011.

[90] Vladimir Koltchinskii. Rejoinder: Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):pp. 2697–2706, 2006.

[91] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):pp. 2593–2656, 2006.

[92] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[93] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York, NY, 1997.

[94] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eightneenth International Conference on Machine Learning*, pages 282–289, 2001.

[95] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[96] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, pages 1–33, 2011.

[97] J. Langford, A. Smola, and M. Zinkevich. Slow learners are fast. In *Advances in Neural Information Processing Systems 22*, pages 2331–2339, 2009.

[98] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1(1):pp. 38–53, 1973.

[99] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[100] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report, UIUC, 2009. Available at arXiv:0903.4742.

[101] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[102] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987.

[103] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

[104] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, 32(4):1679–1697, 2004.

[105] Z. Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46-47:157–178, 1993.

[106] C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

[107] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):pp. 1808–1829, 1999. ISSN 00905364.

[108] P. Massart. *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. Springer, New York, 2003.

[109] J. Matousek. *Lectures on discrete geometry*. Springer-Verlag, New York, 2002.

[110] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.

[111] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[112] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54:48–61, 2009.

[113] A. Nedić, D.P. Bertsekas, and V.S. Borkar. Distributed asynchronous incremental subgradient methods. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, volume 8 of *Studies in Computational Mathematics*, pages 381–407. Elsevier, 2001.

[114] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. Technical report, UC Berkeley, August 2010.

[115] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[116] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS Conference*, Vancouver, Canada, December 2009. Full length version arxiv:1010.2731v1.

[117] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[118] A. S. Nemirovski. Efficient methods in convex programming. Technical report, Georgia Tech, Lecture notes, 2010. URL http://www2.isye.gatech.edu/~nemirovs/OPTI_LectureNotes.pdf.

[119] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. New York, 1983.

[120] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, New York, 2004.

[121] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

[122] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming A*, 120(1):261–283, 2009.

[123] Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics, 1994.

[124] H. V. Ngai and J. P. Penot. Paraconvex functions and paraconvex sets. *Studia Mathematica*, 184:1–29, 2008.

[125] A. B. J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[126] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. *Annals of Statistics*, 2010. To appear.

[127] S. Petrov. *Coarse-to-Fine Natural Language Processing*. PhD thesis, University of California at Berkeley, 2009.

[128] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[129] B. T. Polyak. *Introduction to optimization*. Optimization Software, Inc., 1987.

[130] M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 2011. To appear.

[131] S. S. Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545, 2010.

[132] S. Sundhar Ram, A. Nedic, and V. V. Veeravalli. Distributed subgradient projection algorithm for convex optimization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3653–3656, 2009.

[133] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

[134] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Information Theory*, 57(10): 6976—6994, October 2011.

[135] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 2010. Posted as arXiv:0910.0651v2.

[136] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[137] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, pages 693–701. 2011.

[138] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[139] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[140] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[141] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. Technical Report arXiv:0912.5338v2, Universite de Paris, January 2010.

[142] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958. (Reprinted in *Neurocomputing* (MIT Press, 1988).).

[143] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. Technical report, University of Michigan, July 2011.

[144] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer-Verlag, New York, NY, 2003.

[145] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[146] S. Shalev-Shwartz, O. Shamir, and E. Tromer. Using more data to speed-up training time. In *AISTATS*, 2012.

[147] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. *Math. Programming, Series B*, 127(1):3–30, 2011.

[148] Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM J. Comput.*, 26:751–763, June 1997.

[149] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.

[150] N. Srebro. *Learning with Matrix Factorizations*. PhD thesis, MIT, 2004. Available online: http://ttic.uchicago.edu/ nati/Publications/thesis.pdf.

[151] N. Srebro and A. Tewari. Stochastic optimization for machine learning. ICML 2010 Tutorial, 2010. URL http://ttic.uchicago.edu/~nati/Publications/ICML10tut.pdf.

[152] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

[153] I. Sutskever. A simpler unified analysis of budget perceptrons. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 985–992, 2009.

[154] M. Talagrand. A new look at independence. *The Annals of Probability*, 24(1):pp. 1–34, 1996.

[155] M. Talagrand. *The Generic Chaining*. Springer-Verlag, New York, NY, 2000.

[156] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[157] J. F. Traub and H. Wozniakowski. *A general theory of optimal algorithms*. Academic Press, 1980.

[158] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, December 2007.

[159] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008. URL http://www.math.washington.edu/~tseng/papers/apgm.pdf.

[160] J. Tsitsiklis. *Problems in decentralized decision-making and computation*. PhD thesis, Department of EECS, MIT, 1984.

[161] Koji Tsuda, Gunnar Rätsch, and Manfred K. Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *J. Mach. Learn. Res.*, 6:995–1018, 2005.

[162] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984.

[163] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[164] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[165] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008.

[166] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[167] V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, 1 edition, September 1998.

[168] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, XVI(2):264–280, 1971.

[169] V. N. Vapnik and A. Ya. Chervonenkis. *Theory of pattern recognition*. Nauka, Moscow, 1974. (In Russian).

[170] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[171] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1—305, December 2008.

[172] D. Weiss and B. Taskar. Structured prediction cascades. In *AISTATS*, 2010.

[173] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

[174] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via Outlier Pursuit. Technical report, University of Texas, Austin, 2010. URL http://arxiv.org/pdf/1010.4237v2. available at arXiv:1010.4237.

[175] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[176] B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.

[177] C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[178] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[179] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annal of Statistics*, 53:56–134, 2004.

[180] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.