Multiple Optimality Guarantees in Statistical Learning

John Duchi

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-79
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-79.html

May 15, 2014

Multiple Optimality Guarantees in Statistical Learning

by

John C Duchi

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 Michael I. Jordan, Co-chair
Professor Martin J. Wainwright, Co-chair
Professor Peter Bickel
Professor Laurent El Ghaoui

Spring 2014

Multiple Optimality Guarantees in Statistical Learning

Abstract

Multiple Optimality Guarantees in Statistical Learning

by

John C Duchi

Doctor of Philosophy in Computer Science

and the Designated Emphasis

in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Michael I. Jordan, Co-chair

Professor Martin J. Wainwright, Co-chair

Classically, the performance of estimators in statistical learning problems is measured in terms of their predictive ability or estimation error as the sample size n grows. In modern statistical and machine learning applications, however, computer scientists, statisticians, and analysts have a variety of additional criteria they must balance: estimators must be efficiently computable, data providers may wish to maintain anonymity, large datasets must be stored and accessed. In this thesis, we consider the fundamental questions that arise when trading between multiple such criteria—computation, communication, privacy—while maintaining statistical performance. Can we develop lower bounds that show there must be tradeoffs? Can we develop new procedures that are both theoretically optimal and practically useful?

To answer these questions, we explore examples from optimization, confidentiality preserving statistical inference, and distributed estimation under communication constraints. Viewing our examples through a general lens of constrained minimax theory, we prove fundamental lower bounds on the statistical performance of any algorithm subject to the constraints—computational, confidentiality, or communication—specified. These lower bounds allow us to guarantee the optimality of the new algorithms we develop addressing the additional criteria we consider, and additionally, we show some of the practical benefits that a focus on multiple optimality criteria brings.

In somewhat more detail, the central contributions of this thesis include the following: we

- develop several new stochastic optimization algorithms, applicable to general classes of stochastic convex optimization problems, including methods that are automatically

adaptive to the structure of the underlying problem, parallelize naturally to attain linear speedup in the number of processors available, and may be used asynchronously,

- prove lower bounds demonstrating the optimality of these methods,
- provide a variety of information-theoretic tools—strong data processing inequalities—useful for proving lower bounds in privacy-preserving statistical inference, communication-constrained estimation, and optimization,
- develop new algorithms for private learning and estimation, guaranteeing their optimality, and
- give simple distributed estimation algorithms and prove fundamental limits showing that they (nearly) optimally trade off between communication (in terms of the number of bits distributed processors may send) and statistical risk.

To Emily

Contents

List of Figures

Acknowledgments

There are so many people to whom I owe credit for this thesis that I must begin with an apology: I will probably forget to mention several of you in the coming paragraphs. If I do, please forgive me, and let me know and I will be happy to buy you a beer.

My acknowledgments must begin with my advisors, my two official advisors at Berkeley, and my one surrogate advisor down at Google: Michael Jordan, Martin Wainwright, and Yoram Singer. It has become clear to me that having three advisors was a necessary thing, if only for their sakes, because it kept me out of the hair of the other two while I bothered the third. More seriously, Mike and Martin have pushed me into contact with a multitude of disciplines, encouraging and exemplifying a fearlessness and healthy disrespect for academic boundaries. Without them, I could not have fallen in love with as many subjects—statistics, optimization, computing—as I have, and their guidance about how to approach research, write, give talks, and cooling down my neuroses has been invaluable. They have also provided phenomenal environments for doing research, and (because of them) I have been fortunate to be surrounded constantly by other wonderful students and colleagues. Yoram has been a great mentor and friend, going on runs and bike rides with me, making sure I do not lose touch with practicality, and (hopefully) helping me develop a taste for problems at the border of theory and practice, where one informs the other and vice versa. I hope I can maintain the balance the three of them have modeled for me.

There are a number of other faculty who have been important to my PhD: Pieter Abbeel has shown me what it is like to go from a graduate student to a professor and been a friend since I was just learning to do research as an undergraduate, and his pointers helped me navigate the academic job market without panic. Ben Recht, Chris Ré, and Steve Wright have all been wonderfully encouraging, giving feedback on papers, talks, and trying to force me into the thinking position. I also would like to thank Daphne Koller and Gal Elidan, who opened my eyes to the fun research gives and hard work it takes when I was an undergraduate at Stanford, and Stephen Boyd, who piqued my interest in optimization and has always been a refreshing smart aleck and kept an open door for my harassment whenever I ran into him. Fernando Pereira, with his nightly wanderings around the office in search of researchers to talk with, was a source of interesting and probing questions for all my work. Peter Bickel and Laurent El Ghaoui, who both helped by being on my thesis committee, have provided great feedback on several of my ideas and given good perspective.

As I wrote above, Berkeley has been an awesome environment during my PhD. The collaborators I have had have been unequivocally phenomenal. Alekh Agarwal's quick insights and sharp thinking got us through a number of papers, classes, and into all sorts of fun new research areas. Lester Mackey's patient listening and deep thinking made for wonderful collaborations on ranking algorithms as well as interesting conversations across a variety of topics, and his eating was a never-ending source of entertainment. I have also been honored to collaborate with younger students, Yuchen Zhang and Andre Wibisono, who were great colleagues after Alekh and Lester graduated. Andre's mathematical insights are impressive,

and I must thank Yuchen for his amazing dedication, ability to simply bulldoze any problem in front of him, and his other-worldly demeanor.

There have been a number of other great folks around the lab while I have been here: Sahand Negahban, who helped me feel not alone in my spaziness and came on some fun bike rides with me (as well as providing a good sounding board to a whole pile of research questions), Po-Ling Loh, whose research questions and solutions were always enlightening, and Percy Liang, who showed how to really be dedicated to good work. Mike, Martin, and Peter Bartlett's groups and the AMPLab at Berkeley, with the slew of stellar students and post-docs coming through—Jake Abernethy (always willing to crash in a room I have at a conference), Arash Amini, Tamara Broderick, Ariel Kleiner, Simon Lacoste-Julien, Garvesh Raskutti, Pradeep Ravikumar, Ameet Talwalkar (always up for a beer or some ultimate)—have been stellar. Jon Kuroda, our AMPLab jack of all trades, computer lender, and make-things-happener extraordinaire, was indispensable. And of course the EECS department IM ultimate team Floppy DisKs, who remind me that in spite of our astounding nerdiness, there are some seriously athletic folks in this department, so I should never stop running.

And now I must go back to the beginning, where everything really started, with my family. My parents, Mary and Joe, who encouraged me and gave me every learning opportunity they could think of, set me on a trajectory of simple hunger for knowing more. With art classes, Legos, music, they seeded any creativity I might have, and (Mom) helping out with school and (Dad) making sure I busted my tail in the swimming pool, tennis court, water polo, they showed how important a great work ethic is. I want to thank my brother, Andrew, for still being my friend after about 12 years of abuse (I imagine your first two years were pretty good, and I think things got better when I jumped across the country when you began eighth grade). Also, Andrew, thanks for your stories about poison oak. I'll make sure to stay away from it.

Finally, I must thank my love and wonderful wife, Emily Roberts. Marrying her, I gained a partner with an extraordinary drive to learn and listen. Emily has put up with my conference travel, years of odd sleeping schedules and doing research until morning, and picked up cycling (perhaps with a bit of pushing from me) so we could spend more time together. She taught me to backpack and will never let us spend a nice weekend inside. Emily, thank you for your love and making me happy to be with you.

Part I

Introduction and background

Chapter 1

Introduction

Modern techniques for data gathering—arising from medicine and bioinformatics [120], internet applications such as web-search [86], physics and astronomical experiments [2], mobile data gathering platforms—have yielded an explosion in the mass and diversity of data. Yet the amount of data is no panacea; even in medium scale settings, it is challenging to identify the best ways to analyze and make inferences from information we collect. As a consequence of these difficulties, it is important to develop procedures that intelligently use the available data and trade among scarce resources: how can we balance multiple criteria—computation, communication, privacy—while maintaining statistical performance? Can we be sure we do not ask too much or make incorrect inferences?

In this thesis, we develop theoretically-motivated procedures address problems like these. Broadly, we identify desiderata that—in addition to classical metrics of statistical efficiency—we would like methods to have. In particular, we consider computational, privacy, and communication-based considerations as axes along which to evaluate procedures. Development along these axes has, for the most part, happened independently, as each of the considerations presents challenges in isolation. Given the understanding built by the substantial research in optimization, information theory, computer science, and statistics and the current challenges we face in statistical learning and data analysis, however, it is important to bring together insights from multiple fields to develop methods that trade amongst several criteria for improved (or optimal) performance. Progress in this direction is the goal of this thesis.

1.1 Evaluating statistical learning procedures

The classical estimation problem in statistical decision theory [118, 175] is, given an unknown distribution P, to estimate an (unknown) parameter $\theta(P)$ of the distribution given a sample X drawn from the distribution. We measure the performance of a method $\widehat{\theta}$ according to its expected loss, or risk,

$$R(\widehat{\theta}) := \mathbb{E}_P \left[\ell(\widehat{\theta}(X), \theta(P)) \right], \tag{1.1}$$

where $\ell(\theta, \theta^*)$ denotes a loss incurred for taking the action (or making the prediction) θ when the true state of the world is θ^*. The simplicity of the formulation (1.1) hides its complexity. In statistical learning scenarios [176, 50, 94], the loss may measure the expected prediction error of a binary classifier under the distribution P. In classical statistical problems [23, 118], we may wish to measure error $\|\widehat{\theta} - \theta^*\|_2^2$ made by $\widehat{\theta}$ in recovering a parameter θ^*. In this thesis, we generally use the minimax principle, originally suggested by Wald [178], to evaluate a procedure: for a given family $\mathcal{P}$ of distributions, one chooses the procedure $\widehat{\theta}$ minimizing

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\ell(\widehat{\theta}(X), \theta(P)) \right]. \tag{1.2}$$

There are, of course, a variety of approaches for evaluation of procedures, including Bayesian approaches [23], but we choose to focus on minimax as a unifying principle.

The risk functional (1.1) necessarily focuses attention on performance relative to only one metric: how well does the procedure $\widehat{\theta}$ minimize the loss ℓ? This central question has driven much of the theoretical work in statistics (and machine learning) over the last several decades. But in the absence of other considerations, the risk (1.1) and minimax principle (1.2) may lose practicality; there are often real-world considerations that drive development of statistical procedures, such as the costs of collecting additional observations of the process P or establishing a controlled trial rather than simply observing the process P, questions of privacy [179, 75, 76], or the difficulty of computing particular estimators $\widehat{\theta}$ that may have good risk performance [174, 107, 42]. The central question of this thesis, then, is this: how does the worst-case risk (1.2) changes when we constrain the estimator $\widehat{\theta}$ to belong to some restricted class of procedures? Below, we outline several practical criteria that, when considered against notions of statistical efficiency for minimizing the risk (1.1), can lead to the development of exciting new procedures and new techniques for attacking statistical learning problems.

1.1.1 Computation

"What can be computed and how can we compute it? What is the best way to solve it?" As Jeannette Wing [181] describes it, these questions—and others like them—are the central tenets of what she terms *computational thinking*. Statistical learning problems present a new flavor of problem different from standard computational tasks: there is noise inherent in the problem (1.1), and even given infinite computational power, there often is no perfect solution, just one that attains some type of optimal error. Thus we ask a more subtle question: for a given "computational" budget, what is the best possible error of any method? The challenge in such a question is the nebulous notion of computation [10]: the Turing machine model is often too powerful for identifying subtle differences in computation, and other definitions of computation are often more specialized.

We are not the first to ask this type of question, and several authors have attacked similar problems, beginning with the work of Valiant [174] on Probably Approximately Correct

(PAC) learning, which separates concepts that can be learned with polynomial sample size from those that cannot. More recent work building out of this has often arrived at the sometimes counterintuitive result that more data (larger problems) yields *faster* algorithms. Bottou and Bousquet [30] and Shalev-Shwartz et al. [162, 161] describe settings in which increased data set sizes allow faster learning of classifiers. Berthet and Rigollet [24] study computation time/data tradeoffs in sparse principal component analysis, and work by Chandrasekaran and Jordan [42] show how the complexity of convex sets relates to estimation performance in Gaussian sequence models. In another line of work, Agarwal, Bartlett, and Duchi [5] study model selection under computational constraints, where they force procedures to operate within a given abstract computational budget.

To give ourselves a model of computation that is both practically relevant and theoretically tractable, in this thesis, we identify optimization complexity with computational complexity, using the tools of *information-based complexity* developed by Nemirovski and Yudin [134] and Traub et al. [169] (see also the introductory survey by Traub and Werschulz [168], the book of Plaskota [144], and the paper of Agarwal et al. [6], which is particularly relevant to this thesis). In these settings, we usually assume we are minimizing a function

$$f(\theta) := \mathbb{E}[F(\theta; X)]$$

over $\theta \in \Theta \subset \mathbb{R}^d$, and we treat the computation of an instantaneous loss $F(\theta; X)$ or gradient evaluation $\nabla_\theta F(\theta; X)$ as our computational unit. Given the prevalence of optimization algorithms based on first-order (gradient) information [136, 32, 98, 134], this focus is both natural and practically motivated. We may then study—for different classes of domains Θ, types of losses F, and families of probability distributions P according to which the data X are generated—the complexity of optimization.

In addition to asking about the fundamental complexity of optimization in terms of zeroth or first-order information, given the modern development of multi-core and distributed computing infrastructure, it is also important to study aspects of parallel, asynchronous, and distributed computation [26, 79, 14, 33, 49]. To what extent can we maintain optimality of optimization procedures while leveraging parallel computation? We address such questions in this thesis as well, giving algorithms that maintain both optimal complexity—in terms of their number of gradient or function evaluations—while running in parallel or asynchronously.

1.1.2 Privacy

As the scale of datasets—with the concomitant amount of information we collect about individuals—increases, maintaining anonymity and privacy has become more important. While the maintaining privacy is an old issue, dating at least to Warner's 1960s work on randomized response and survey sampling [179], it has become clear that modern data collection poses new risks of disclosure. For example, Homer et al. [99] recently showed it is possible to identify the presence of individual genotypes in high-density SNP arrays, leading

to the removal of some publicly available genomics data [83]. A major challenge in statistical inference has thus become that of characterizing and balancing statistical utility with the privacy of individuals from whom we obtain data [63, 64, 76].

In the large body of research on privacy and statistical inference [e.g., 179, 75, 63, 64, 76], a major focus has been on the problem of reducing disclosure risk: the probability that a member of a dataset can be identified given released statistics of the dataset. The literature has stopped short, however, of providing a formal treatment of disclosure risk that would permit decision-theoretic tools to be used in characterizing tradeoffs between the utility of achieving privacy and the utility associated with an inferential goal. Recently, a formal treatment of disclosure risk known as "differential privacy" has been proposed by Dwork and colleagues and studied in the cryptography, database and theoretical computer science literatures [68, 65]. Differential privacy has strong semantic privacy guarantees that make it a good candidate for declaring a statistical procedure private, and it has been the focus of a growing body of recent work [65, 74, 91, 180, 164, 44, 105]. Direct connections between statistical risk (1.1) and privacy, however, have been somewhat more challenging to make; with modern issues in data collection, however, it is becoming more important to understand quantitative tradeoffs between privacy and statistical efficiency.

1.1.3 Communication

While computational considerations are important for the development of estimation and inferential procedures, the scale of modern datasets often necessitates distributed storage and computation [86]. Perhaps even more saliently, computer processing speeds are beginning to hit fundamental physical limits [79], and, as Fuller and Millett [79] point out in a survey for the National Academy of Sciences, our only "known hope" for continued improvement in computational performance is to leverage parallel and distributed computing. The relative expense of communication with respect to computation, however, makes inter-processor or inter-machine communication an especially important resource and measure of the performance of algorithms [14, 79]. Moreover, the rates at which communication costs improve are slower than those of other computing tasks, so communication is becoming both more important—due to the rise in parallelism and large datasets—and its relative costs are also increasing. Additionally, connecting the tools of information theory [47], which allow us to describe the fundamental limits of communication and storage, with statistical inference problems has been somewhat challenging [89]. It is thus important to understand fundamental limits in distributed statistical estimation problems and to discover new procedures that attain these limits.

1.2 Thesis goals and contributions

The focus of this thesis is to develop, via examples in stochastic approximation, privacy-preserving inference, and communication-constrained estimation, approaches for designing

and analyzing methods whose performance is measured along multiple axes. Using classical statistical minimax theory as our starting point, we introduce a notion of constrained minimax risk, and we using this measure of performance, we develop fundamental lower bounds and procedures attaining them for a variety of problems. This development requires a two-pronged approach, where we show the fundamental hardness of problems—giving lower bounds, leveraging ideas from optimization, information theory, and statistics—and derive efficient algorithms achieving these bounds. By attacking problems from both sides, it is possible to gain deeper insights into the underlying difficulties, relaxations and circumventions of those difficulties, and essential structures of the problems being solved. Building on these insights, we can derive algorithms that trade amongst a multitude of criteria for improved performance, yielding more efficient procedures for real large-scale statistical, learning, and optimization problems.

In particular, the goals of this thesis are to

(1) Introduce a notion of minimax risk for estimators constrained by particular resource (or other) requirements and to develop tools for proving fundamental lower bounds on these notions of risk, and

(2) Develop new procedures for different types of constraints, focusing on computational (via optimization), confidentiality, and communication-based constraints.

More specifically, the central contributions of the thesis are the following:

- We review and extend several techniques for proving minimax lower bounds, developing a few finer-grained information-theoretic inequalities that allow easier proofs of many lower bounds
- We show new ways the performance of stochastic optimization algorithms depends on the geometry underlying the problem, and we show how to give algorithms that are optimal—and adaptive—to the underlying problem structure
- We show how the use of dual averaging algorithms allows (nearly) completely asynchronous optimization schemes whose performance improves linearly with the number of processors in parallel computing environments, as long as data obeys certain sparsity restrictions
- We develop randomized smoothing techniques that (i) yield optimal algorithms for nonsmooth (sub)gradient-based optimization, even in parallel computing environments, and (ii) extend these to zeroth order optimization schemes, providing new optimal algorithms (as well as guarantees of their optimality)
- We develop quantitative data processing inequalities that allow the application of our information-theoretic techniques to privacy-preserving data analyses, providing new fundamental lower bounds on the performance of procedures that maintain privacy in estimation problems

- We provide new algorithms that attain the fundamental limits for privacy-preserving estimation in "local-privacy" settings where data providers do not even trust the data collector
- We review recent low-communication optimization and estimation schemes, and we adapt our information-theoretic tools (again based on new data-processing inequalities) to prove fundamental limits on communication-constrained estimation procedures.

A common theme in all of our algorithms is that they exploit problem structure—the statistical properties and the noise inherent to the problem—for more efficient methods. In the stochastic optimization case, this comes in the form of adding additional noise that is of lower order than that already in the problem, either via randomized smoothing or asynchrony, and enjoying consequent speedups. In the privacy case, this comes in the form of directly adding noise to data to protect confidentiality, while using the statistical structure of the data to avoid adding more noise than necessary. In the low-communication case, this consists of observing that averaging independent distributed solutions is nearly sufficient for optimal solution of many statistical problems, because each processor's local solution is simply a noisy approximation to the truth, rather than the solution to an adversarially chosen problem. In a sense, the thesis simply studies what good a little noise can do.

1.3 Organization of the thesis and previously published work

Several portions of this thesis are based on joint work of mine with collaborators, which I describe (briefly) here, in addition to outlining the rest of the thesis. Part I of this thesis provides some background on minimax theory, setting up (at an abstract level) the constrained minimax problem by which we evaluate our procedures in Chapter 2. Much of the material in the chapter is classical, though some of it is based on joint work with Martin Wainwright and Michael Jordan [51, 59].

In Part II of the thesis, we focus on stochastic optimization problems, investigating computational limits (via information-based complexity) as well as distributed and asynchronous optimization techniques. Chapters 3 and 4 study adaptive optimization schemes and single-processor optimality guarantees (Chapter 3) and characteristics of data that allow asynchronous parallel algorithms (Chapter 4). They contain some new material and some based on joint work with Michael Jordan and Brendan McMahan [58], which builds off of earlier research performed jointly with Elad Hazan and Yoram Singer [53]. Chapter 5 studies randomized smoothing techniques to develop optimal optimization schemes for non-smooth problems and is based on work with Peter Bartlett and Martin Wainwright [55], while Chapter 6 extends these randomized smoothing ideas to attack optimization problems where only zeroth order (function value) information is available, providing new (optimal) schemes and

fundamental lower bounds for such problems. It is based on work with Michael Jordan, Andre Wibisono, and Martin Wainwright [61].

Part III of the thesis is on tradeoffs between privacy and statistical utility, studying the effects of imposing local privacy on convergence rates for statistical estimators, and builds out of joint work with Michael Jordan and Martin Wainwright [60, 59].

Finally, Part IV of the thesis studies some effects of communication on distributed estimators. Chapter 9 reviews simple distributed estimation algorithms developed jointly with Yuchen Zhang and Martin Wainwright [189], and in Chapter 10, we develop information theoretic tools to exhibit the fundamental limits and tradeoffs between statistical efficiency and inter-machine communication. This final chapter is based off of joint work with Yuchen Zhang, Michael Jordan, and Martin Wainwright [62].

1.4 Notation

Before proceeding to the thesis proper, we define notation and terminology that we commonly use; our settings are essentially standard. Throughout, we use $\mathbb{R}$ to denote the real numbers and $\mathbb{N} = \{1, 2, \ldots\}$ to denote the counting numbers.

Asymptotic notation We use standard asymptotic notation throughout the thesis. In particular, we use $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, and $o(\cdot)$. Formally, for real-valued sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, we say that $a_n = \mathcal{O}(b_n)$ if there exists a constant $c < \infty$ and an $N \in \mathbb{N}$ such that $a_n \le cb_n$ for all $n \ge N$. Similarly, we say $a_n = \Omega(b_n)$ if $b_n = \mathcal{O}(a_n)$, that is, there exists a constant $c > 0$ and $N \in \mathbb{N}$ such that $a_n \ge cb_n$ for $n \ge N$, and $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. We say $a_n = o(b_n)$ if $|a_n|/|b_n| \to 0$ as $n \to \infty$. We use the notation $a_n \lesssim b_n$ to denote $a_n = \mathcal{O}(b_n)$, and $a_n \ll b_n$ to denote $a_n = o(b_n)$, meaning that "a_n is at most on the order of b_n" and "a_n is significantly smaller than b_n." In general, unless otherwise specified, our asymptotic notation will hide only numerical constants that do not depend on problem parameters and will apply with $N = 1$.

Statistical notation We require standard statistical notation throughout the thesis. Given a sequence of random variables $\{X_n\}$ and another random variable (or constant) Y all taking values in a metric space $\mathcal{X}$ with distance ρ, we say that X_n converges in probability to Y, meaning that $X_n \xrightarrow{p} Y$, if for all $\epsilon > 0$ we have $\mathbb{P}(\rho(X_n, Y) > \epsilon) \to 0$ as $n \to \infty$. We say that X_n converges in distribution to Y (or, if Y is distributed according to P, denoted $Y \sim P$, that X_n converges in distribution to P) if for all bounded and continuous functions $f : \mathcal{X} \to \mathbb{R}$ we have $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(Y)]$ as $n \to \infty$. See, for example, the Portmanteau Theorem [e.g. 175] for equivalent definitions. We use statistical big-O notation as well (see, for example, Lehmann and Casella [118]). Given two sequences of random variables or vectors $\{X_n\}$ and $\{Y_n\}$ on spaces $\mathcal{X}$ and $\mathcal{Y}$ with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, we say that $X_n = \mathcal{O}_P(Y_n)$ if for each $\epsilon > 0$, there exists a constant $C(\epsilon)$ and $N \in \mathbb{N}$ such that $n \ge N$

implies $\mathbb{P}(\|X_n\|_{\mathcal{X}} \geq C(\epsilon) \|Y_n\|_{\mathcal{Y}}) \leq \epsilon$. A random variable X is sub-Gaussian [e.g. 36] with parameter σ^2 if for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$.

Given two probability distributions P and Q on a space $\mathcal{X}$, each assumed absolutely continuous with respect to an underlying measure μ with densities p and q respectively,[1] the KL-divergence between P and Q is

$$D_{\rm kl}\left(P\|Q\right) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x).$$

Given a discrete random variable X defined on a space $\mathcal{X}$ with probability mass function (p.m.f.) p, its (Shannon) entropy [47] is $H(X) := -\sum_{x\in\mathcal{X}} p(x) \log p(x)$. The conditional entropy of X given Y, where X and Y have joint p.m.f. $p(x, y)$, is $H(X \mid Y) := -\sum_{x,y} p(x, y) \log p(x \mid y)$. Given random variables X and Y with marginal distributions P_X and P_Y, respectively, and joint $P_{X,Y}$, the mutual information between X and Y is

$$\begin{aligned} I(X;Y) &:= \int_{\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)} d\mu(x, y) \\ &= \mathbb{E}_{P_X}[D_{\rm kl}\left(P_Y(\cdot \mid X)\|P_Y(\cdot)\right)] = \int_{\mathcal{X}} D_{\rm kl}\left(P_Y(\cdot \mid X = x)\|P_Y(\cdot)\right) dP_X(x), \end{aligned}$$

where μ is a measure assumed to dominate $P_{X,Y}$ and p is the density of P with respect to μ. Throughout we use log base-e for our entropy and information theoretic calculations. If $\sigma(\mathcal{X})$ denotes the σ-field on $\mathcal{X}$, the total variation distance between two distributions P and Q defined on $(\mathcal{X}, \sigma(\mathcal{X}))$ is

$$\|P - Q\|_{\rm TV} := \sup_{S\in\sigma(\mathcal{X})} |P(S) - Q(S)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| d\mu(x).$$

We use $\mathsf{N}(\theta, \Sigma)$ to denote the normal distribution with mean θ and covariance matrix Σ and Laplace(κ) to denote the Laplace distribution with inverse shape parameter κ, that is, density $p(x) \propto \exp(-\kappa|x|)$.

Analytic, matrix, and vector notation For vectors $x \in \mathbb{R}^d$, we use ℓ_p to denote the usual p-norms $\|x\|_p = (\sum_{j=1}^d |x_j|^p)^{\frac{1}{p}}$, where $\|x\|_\infty = \max_j |x_j|$. The ℓ_2-operator norm of a matrix $A \in \mathbb{R}^{d_1\times d_2}$ is its maximum singular value, defined by

$$|\!|\!| A |\!|\!| = |\!|\!| A |\!|\!|_2 := \sup_{v\in\mathbb{R}^{d_2}, \|v\|_2\leq 1} \|Av\|_2 .$$

We use $\gamma_i(A)$ to denote the ith singular value of A, and $\|A\|_{\rm Fr}$ to denote its Frobenius norm. We let $\langle\cdot, \cdot\rangle$ denote the standard inner product on $\mathbb{R}^d$ (or whatever space is being used), and given a norm $\|\cdot\|$ on $\mathbb{R}^d$, the dual norm $\|\cdot\|_*$ is given by

$$\|y\|_* = \sup_{x\in\mathbb{R}^d} \left\{\langle x, y\rangle : \|x\| \leq 1\right\}.$$

[1] This is no loss of generality, as we may take $\mu = \frac{1}{2}P + \frac{1}{2}Q$

A set $\mathcal{X} \subset \mathbb{R}^d$ is convex if $x, y \in \mathcal{X}$ implies that $\lambda x + (1-\lambda)y \in \mathcal{X}$ for all $\lambda \in [0,1]$. A function f is convex if its domain $\operatorname{dom} f$ is convex, and $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all $x, y \in \operatorname{dom} f$ and $\lambda \in [0,1]$. We denote the subgradient set of f at a point x by

$$\partial f(x) := \{g \in \mathbb{R}^d : f(y) \geq f(x) + \langle g, y - x \rangle \quad \text{for all } y \in \mathbb{R}^d\}.$$

For shorthand, we let $\|\partial f(x)\| = \sup_{g \in \partial f(x)} \|g\|$. We make the standard assumption [98, 152] that $f(y) = +\infty$ for all $y \notin \operatorname{dom} f$ for convex f. To avoid pathologies, any convex function f in this thesis is assumed to be sub-differentiable over all of $\operatorname{dom} f$. The Euclidean projection of a point y onto a closed convex set C is

$$\Pi_C(y) := \operatorname*{argmin}_{x \in C} \left\{ \|x - y\|_2^2 \right\}.$$

For any function f and a norm $\|\cdot\|$, we say that f is M-Lipschitz continuous with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ if

$$|f(x) - f(y)| \leq M \|x - y\| \quad \text{for all } x, y \in \mathcal{X}.$$

Similarly, for f differentiable on a set $\mathcal{X}$, we say that ∇f is L-Lipschitz continuous with respect to a norm $\|\cdot\|$ (with associated dual norm $\|\cdot\|_*$) if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \quad \text{for all } x, y \in \mathcal{X}.$$

We let $\otimes$ denote the Kronecker product, and for a pair of vectors u, v, we define the outer product $u \otimes v = uv^\top$. For a three-times differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we denote the third derivative tensor by $\nabla^3 f$, so that for each $x \in \operatorname{dom} f$ the operator $\nabla^3 f(x) : \mathbb{R}^{d \times d} \to \mathbb{R}^d$ is linear and satisfies the relation

$$\left[\nabla^3 f(x)(v \otimes v)\right]_i = \sum_{j,k=1}^{d} \left(\frac{\partial^3}{\partial x_i \partial x_j \partial x_k} f(x) \right) v_j v_k.$$

Miscellaneous notation We denote the indicator function of an event E by $\mathbf{1}\{E\}$, which is 1 if E occurs (or is true) and 0 otherwise. For an integer n, the notation $[n]$ denotes the set of integers $\{1, \ldots, n\}$. We let $\vee$ and $\wedge$ denote maximum and minimum, respectively, so that $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

Chapter 2

Minimax rates of convergence

Understanding the fundamental limits of estimation and optimization procedures is important for a multitude of reasons. Indeed, developing bounds on the performance of procedures can give complementary insights. By exhibiting fundamental limits of performance (perhaps over restricted classes of estimators), it is possible to guarantee that an algorithm we have developed is optimal, so that searching for estimators with better statistical performance will have limited returns, though searching for estimators with better performance in other metrics may be interesting. Moreover, exhibiting refined lower bounds on the performance of estimators can also suggest avenues for developing alternative, new optimal estimators; lower bounds need not be a fully pessimistic exercise.

In this chapter, we define and then discuss techniques for lower-bounding the minimax risk, giving three standard techniques for deriving minimax lower bounds that have proven fruitful in a variety of statistical learning problems [188]. In addition to reviewing these standard techniques—the Fano, Assouad, and Le Cam methods—we also present a few simplifications and extensions that may make them more "user friendly."

2.1 Basic framework and minimax risk

Our first step here is to establish the minimax framework we use throughout the thesis. Depending on the problem we study, we use either minimax risk or what is known as minimax *excess* risk to evaluate optimality of our estimation procedures. Our setting is essentially standard, and we refer to references [188, 185, 173] for further background. Let us begin by defining the standard minimax risk, deferring temporarily our discussion of minimax excess risk. Throughout, we let $\mathcal{P}$ denote a class of distributions on a sample space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \Theta$ denote a function defined on $\mathcal{P}$, that is, a mapping $P \mapsto \theta(P)$. The goal is to estimate the parameter $\theta(P)$ based on observations X_i drawn from the (unknown) distribution P.

The space Θ in which the parameter $\theta(P)$ takes values depends on the underlying statistical problem; as an example, if the goal is to estimate the univariate mean $\theta(P) = \mathbb{E}_P[X]$,

we have $\Theta \subset \mathbb{R}$. To evaluate the quality of an estimator $\widehat{\theta}$, we let $\rho : \Theta \times \Theta$ denote a (semi)metric on the space Θ, which we use to measure the error of an estimator for the parameter θ, and let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (for example, $\Phi(t) = t^2$).

In classical settings, the statistician is given direct access to i.i.d. observations X_i drawn according to some $P \in \mathcal{P}$. Based on these $\{X_i\}$, the goal is to estimate the unknown parameter $\theta(P) \in \Theta$, and an estimator $\widehat{\theta}$ is a measurable function $\widehat{\theta} : \mathcal{X}^n \to \Theta$. We then assess the quality of the estimate $\widehat{\theta}(X_1, \ldots, X_n)$ in terms of the risk

$$\mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1 \ldots, X_n), \theta(P))\big) \right].$$

For instance, for a univariate mean problem with $\rho(\theta, \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$, this risk is the mean-squared error. Of course, for any fixed distribution P, it is easy to estimate $\theta(P)$: simply return $\theta(P)$, which will have minimal risk. It is thus important to ask for a more uniform notion of risk, which leads to the minimax principle, first suggested by Wald [178], which is to choose the estimator (measurable function) $\widehat{\theta}$ minimizing the maximum risk

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1 \ldots, X_n), \theta(P))\big) \right].$$

An optimal estimator for this metric then gives the *minimax risk*, which is defined as

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big) \right], \tag{2.1}$$

where we take the supremum (worst-case) over distributions $P \in \mathcal{P}$, and the infimum is taken over all estimators $\widehat{\theta}$.

In some scenarios, we study a slightly different notion of risk, which is more appropriate for some learning and optimization problems. In these settings, we assume there exists some loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$, where for an observation $x \in \mathcal{X}$, the value $\ell(\theta; x)$ measures the instantaneous loss associated with using θ as a predictor. In this case, we define the risk

$$R_P(\theta) := \mathbb{E}_P[\ell(\theta; X)] = \int_{\mathcal{X}} \ell(\theta; x) dP(x) \tag{2.2}$$

as the expected loss of the parameter vector θ. For a (potentially random) estimator $\widehat{\theta} : \mathcal{X}^n \to \Theta$ given access to a sample $X_1, \ldots, X_n$, we may define the associated maximum *excess risk* for the family $\mathcal{P}$ by

$$\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[R_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} R(\theta) \right\},$$

where the expectation is taken over X_i and any randomness in the procedure $\widehat{\theta}$. This expression captures the difference between the (expected) risk performance of the procedure $\widehat{\theta}$ and

the best possible risk, available if the distribution P were known ahead of time. The *minimax excess risk*, defined with respect to the loss ℓ, domain Θ, and family $\mathcal{P}$ of distributions, is then defined by the best possible maximum excess risk,

$$\mathfrak{M}_n(\Theta, \mathcal{P}, \ell) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[R_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} R_P(\theta) \right\}, \tag{2.3}$$

where the infimum is taken over all estimators $\widehat{\theta} : \mathcal{X}^n \to \Theta$ and the risk R_P is implicitly defined in terms of the loss ℓ. The techniques for providing lower bounds for the minimax risk (2.1) or the excess risk (2.3) are essentially identical; we focus for the remainder of this section on techniques for providing lower bounds on the minimax risk.

The minimax risk (2.1) is well-studied, beginning with work of Wald [178] and continuing through a multitude of researchers; important references include the books by Le Cam [115] and Ibragimov and Has'minskii [101], the papers of Birgé [27], Yu [188], and Yang and Barron [185], and the recent introductory survey of Tsybakov [173] provides an overview of minimax techniques in non-parametric estimation. In this thesis, however, we study a variant of the minimax risk where we constrain our estimators $\widehat{\theta}$ to belong to a particular class $\mathcal{C}$ of estimators. In particular, letting $\mathcal{C}$ denote a subset of the (measurable) functions $\widehat{\theta} : \mathcal{X}^n \to \Theta$, we define the *constrained* minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) := \inf_{\widehat{\theta} \in \mathcal{C}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big) \right]. \tag{2.4}$$

In this thesis, we take a few steps—via examples in optimization, communication-constrained estimation, and privacy-preserving statistical inference—toward the study of the object (2.4).

While the change to a constrained class of estimators may appear at first glance to be superficial, it will become clear that study of such constrained estimators is both challenging and can yield new, interesting procedures. The constrained minimax principle (2.4) is thus, essentially, the driving force of this research. Indeed, in Part II of the thesis, we show how such constraints and an understanding of minimax lower bounds can give rise to new and efficient algorithms for optimization. In Part III, we develop techniques for proving minimax lower bounds for statistical estimation when providers of the data wish to guarantee that their data is private, and we give corresponding new (optimal) procedures for private estimation. Finally, in Part IV, we analyze a few simple procedures for distributed statistical estimation, showing that they in fact enjoy optimality guarantees both in statistical and communication-theoretic senses.

2.2 Methods for lower bounds: Le Cam, Assouad, and Fano

There are a variety of techniques for providing lower bounds on the minimax risk (2.1) (and, by extension, (2.4)). Each of them transforms the maximum risk by lower bounding it via a

Bayesian problem (e.g. [101, 115, 118]), then proving a lower bound on the performance of all possible estimators for the Bayesian problem. In particular, let $\{P_v\} \subset \mathcal{P}$ be a collection of distributions in $\mathcal{P}$ indexed by v and π be any probability mass function over v. Then for any estimator $\widehat{\theta}$, the maximum risk has lower bound

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))) \right] \geq \sum_v \pi(v) \mathbb{E}_{P_v} \left[\Phi(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P_v))) \right].$$

While trivial, this lower bound serves as the departure point for each of the subsequent techniques for lower bounding the minimax risk.

2.2.1 From estimation to testing

A standard first step in proving minimax bounds is to reduce the estimation problem to a testing problem [188, 185, 173]. We use two types of testing problems: one a multiple hypothesis test, the second based on multiple binary hypothesis tests. We begin with the simpler of the two. Given an index set $\mathcal{V}$ of finite cardinality, consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ contained within $\mathcal{P}$. This family induces a collection of parameters $\{\theta(P_v)\}_{v \in \mathcal{V}}$; it is a 2δ-packing in the ρ-semimetric if

$$\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta \quad \text{for all } v \neq v'.$$

We use this family to define the *canonical hypothesis testing problem*:

- first, nature chooses V according to the uniform distribution over $\mathcal{V}$;
- second, conditioned on the choice $V = v$, the random sample $X = (X_1, \ldots, X_n)$ is drawn from the n-fold product distribution P_v^n.

Given the observed sample X, the goal is to determine the value of the underlying index v. We refer to any measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$ as a test function. Its associated error probability is $\mathbb{P}(\Psi(X_1, \ldots, X_n) \neq V)$, where $\mathbb{P}$ denotes the joint distribution over the random index V and X. The classical reduction from estimation to testing [e.g., 173, Section 2.2] guarantees that the minimax error (2.1) has lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(X_1, \ldots, X_n) \neq V), \tag{2.5}$$

where the infimum ranges over all testing functions.

To see this result, fix an arbitrary estimator $\widehat{\theta}$. Suppressing dependence on X throughout the derivation, first note that it is clear that for any fixed θ, we have

$$\mathbb{E}[\Phi(\rho(\widehat{\theta}, \theta))] \geq \mathbb{E}\left[\Phi(\delta) \mathbf{1}\left\{ \rho(\widehat{\theta}, \theta) \geq \delta \right\} \right] \geq \Phi(\delta) \mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta),$$

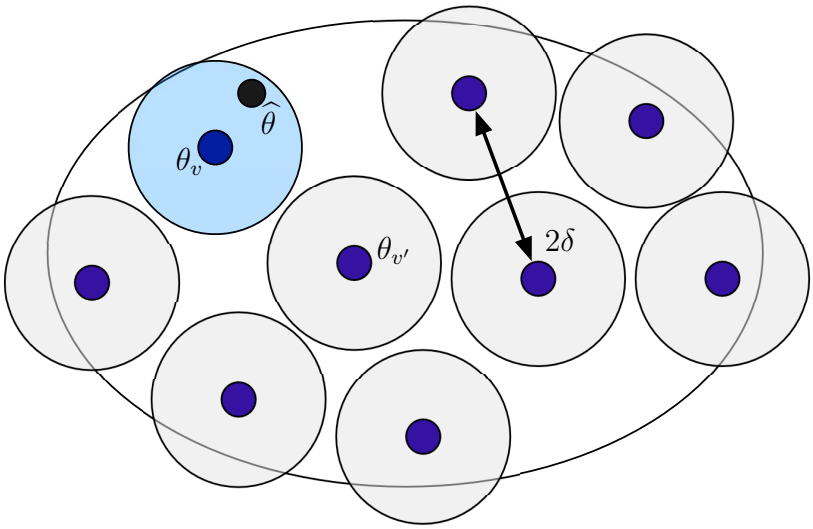

Figure 2.1. Example of a 2δ-packing of a set. The estimate $\widehat{\theta}$ is contained in at most one of the δ-balls around the points θ_v.

where the final inequality follows because Φ is non-decreasing. Now, let us define $\theta_v = \theta(P_v)$, so that $\rho(\theta_v, \theta_{v'}) \geq 2\delta$ for $v \neq v'$. By defining the testing function

$$\Psi(\widehat{\theta}) := \operatorname*{argmin}_{v \in \mathcal{V}} \{\rho(\widehat{\theta}, \theta_v)\},$$

breaking ties arbitrarily, we have that $\rho(\widehat{\theta}, \theta_v) < \delta$ implies that $\Psi(\widehat{\theta}) = v$ because of the triangle inequality and 2δ-separation of the set $\{\theta_v\}_{v \in \mathcal{V}}$. Equivalently, for $v \in \mathcal{V}$, the inequality $\Psi(\widehat{\theta}) \neq v$ implies $\rho(\widehat{\theta}, \theta_v) \geq \delta$. (See Figure 2.1.) By averaging over $\mathcal{V}$, we find that

$$\sup_P \mathbb{P}(\rho(\widehat{\theta}, \theta(P)) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\widehat{\theta}, \theta(P_v)) \geq \delta \mid V = v) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\Psi(\widehat{\theta}) \neq v \mid V = v).$$

Taking an infimum over all tests $\Psi : \mathcal{X}^n \to V$ gives inequality (2.5).

The remaining challenge is to lower bound the probability of error in the underlying multi-way hypothesis testing problem, which we do by choosing the separation δ to trade off between the loss $\Phi(\delta)$ (large δ increases the loss) and the probability of error (small δ, and hence separation, makes the hypothesis test harder). Usually, one attempts to choose the largest separation δ that guarantees a constant probability of error. There are a variety of techniques for this, and we present three: Le Cam's method, Fano's method, and Assouad's method, including extensions of the latter two to enhance their applicability.

2.2.2 Le Cam's method

Le Cam's method, in its simplest form, provides lower bounds on the error in simple binary hypothesis testing testing problems. That is, it is applicable when there are two values v, v'

in $\mathcal{V}$. It is a standard result [115, 188, Lemma 1] that the total variation distance has the variational representation

$$\inf_{f\geq 0, g\geq 0: f+g\geq 1} \{\mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(X)]\} = \inf_{\Psi} \{P(\Psi(X) \neq 0) + Q(\Psi(X) \neq 1)\} \\ = 1 - \|P - Q\|_{\mathrm{TV}} \tag{2.6}$$

for any two distributions P, Q, where the first infimum is taken over all non-negative measurable functions and the second over all tests. Thus, when $V = v$ with probability $\frac{1}{2}$ and v' with probability $\frac{1}{2}$, we have

$$\inf_{\Psi} \mathbb{P}\left(\Psi(X_1, \ldots, X_n) \neq V\right) = \frac{1}{2} - \frac{1}{2}\left\|P_v^n - P_{v'}^n\right\|_{\mathrm{TV}}. \tag{2.7}$$

In particular, this lower bound implies that for any pair of family $\mathcal{P}$ of distributions for which there exists a pair $P_1, P_2 \in \mathcal{P}$ satisfying $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, then

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2}\left\|P_1^n - P_2^n\right\|_{\mathrm{TV}}\right]. \tag{2.8}$$

Example: Bernoulli mean estimation As an illustrative application of Le Cam's method, consider the problem of estimating the mean $\theta \in [-1, 1]$ of a $\{\pm 1\}$-valued Bernoulli distribution under the squared error loss, where $X_i \in \{-1, 1\}$. In this case, by fixing some $\delta > 0$, we set $\mathcal{V} = \{-1, 1\}$, and we define P_v so that

$$P_v(X = 1) = \frac{1 + v\delta}{2} \quad \text{and} \quad P_v(X = -1) = \frac{1 - v\delta}{2},$$

whence we see that the mean $\theta(P_v) = \delta v$. Using the metric $\rho(\theta, \theta') = |\theta - \theta'|$ and loss $\Phi(\delta) = \delta^2$, we have separation 2δ of $\theta(P_{-1})$ and $\theta(P_1)$. Thus, via Le Cam's method (2.8), we have that

$$\mathfrak{M}_n(\text{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2}\delta^2\left(1 - \left\|P_{-1}^n - P_1^n\right\|_{\mathrm{TV}}\right).$$

We would thus like to upper bound $\|P_{-1}^n - P_1^n\|_{\mathrm{TV}}$ as a function of the separation δ and sample size n; we do this using Pinsker's inequality [e.g. 47]. Indeed, we have

$$\left\|P_{-1}^n - P_1^n\right\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}\left(P_{-1}^n \| P_1^n\right) = \frac{n}{2} D_{\mathrm{kl}}\left(P_{-1} \| P_1\right) = \frac{n}{2}\delta \log \frac{1 + \delta}{1 - \delta}.$$

Noting that $\delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, 1/2]$, we obtain that $\|P_{-1}^n - P_1^n\|_{\mathrm{TV}} \leq \delta\sqrt{3n/2}$ for $\delta \leq 1/2$. In particular, we can guarantee a high probability of error in the associated hypothesis testing problem (recall inequality (2.7)) by taking $\delta = 1/\sqrt{6n}$; this guarantees $\|P_{-1}^n - P_1^n\|_{\mathrm{TV}} \leq \frac{1}{2}$. We thus have the minimax lower bound

$$\mathfrak{M}_n(\text{Bernoulli}([-1, 1]), (\cdot)^2) \geq \frac{1}{2}\delta^2\left(1 - \frac{1}{2}\right) = \frac{1}{24n},$$

which is sharp to within constant factors.

2.2.3 Fano's method

Fano's method, originally proposed by Has'minskii [92] for providing lower bounds in non-parametric estimation problems, gives a somewhat more general technique than Le Cam's method, and it applies when the packing set $\mathcal{V}$ has cardinality larger than two. The method has played a central role in minimax theory, beginning with the pioneering work of Has'minskii and Ibragimov [92, 101]. More recent work following this initial push continues to the present day (e.g. [27, 188, 185, 28, 148, 85, 37]).

We begin by stating Fano's inequality, which provides a lower bound on the error in a multi-way hypothesis testing problem. Let V be a random variable taking values in a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$. If we define the binary entropy function $h_2(p) = -p \log p - (1-p) \log(1-p)$, Fano's inequality takes the following form [e.g. 47, Chapter 2]:

Lemma 2.1 (Fano)**.** *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(\mathbb{P}(\widehat{V} \neq V)) + \mathbb{P}(\widehat{V} \neq V) \log(|\mathcal{V}| - 1) \geq H(V \mid \widehat{V}). \tag{2.9}$$

A standard simplification of Lemma 2.1 is to note that $h_2(p) \leq \log 2$ for any $p \in [0, 1]$, so that if V is uniform on the set $\mathcal{V}$ and hence $H(V) = \log |\mathcal{V}|$, then for a sample X, we have the testing lower bound

$$\inf_{\Psi} \mathbb{P}(\Psi(X) \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log |\mathcal{V}|} \tag{2.10}$$

in the canonical hypothesis testing problem from Section 2.2.1.

While the testing lower bound (2.10) is sufficient for proving lower bounds for many estimation problems, for the sharpest results it sometimes requires a somewhat delicate construction of a well-separated packing (e.g. [37, 60]). To that end, we also provide extensions of inequalities (2.9) and (2.10) that more directly yield bounds on estimation error, allowing more direct and simpler proofs of a variety of minimax lower bounds (see also reference [51]).

More specifically, suppose that the distance function $\rho_{\mathcal{V}}$ is defined on $\mathcal{V}$, and we are interested in bounding the estimation error $\rho_{\mathcal{V}}(\widehat{V}, V)$. We begin by providing analogues of the lower bounds (2.9) and (2.10) that replace the testing error with the tail probability $\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$. By Markov's inequality, such control directly yields bounds on the expectation $\mathbb{E}[\rho_{\mathcal{V}}(\widehat{V}, V)]$. As we show in the sequel and in chapters to come, these distance-based Fano inequalities allow more direct proofs of a variety of minimax bounds without the need for careful construction of packing sets or metric entropy calculations as in other arguments.

We begin with the distance-based analogue of the usual discrete Fano inequality in Lemma 2.1. Let V be a random variable supported on a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$, and let $\rho : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a function defined on $\mathcal{V} \times \mathcal{V}$. In the usual setting, the function ρ is a metric on the space $\mathcal{V}$, but our theory applies to general functions. For a given scalar $t \geq 0$, the maximum and minimum *neighborhood sizes at radius t* are given by

$$N_t^{\max} := \max_{v \in \mathcal{V}} \left\{ \operatorname{card}\{v' \in \mathcal{V} \mid \rho(v, v') \leq t\} \right\} \quad \text{and} \quad N_t^{\min} := \min_{v \in \mathcal{V}} \left\{ \operatorname{card}\{v' \in \mathcal{V} \mid \rho(v, v') \leq t\} \right\}. \tag{2.11}$$

Defining the error probability $P_t = \mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$, we then have the following generalization of Fano's inequality:

Proposition 2.1. *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(P_t) + P_t \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log N_t^{\max} \geq H(V \mid \widehat{V}). \tag{2.12}$$

Before proving the proposition, which we do in Section 2.4.1, it is informative to note that it reduces to the standard form of Fano's inequality (2.9) in a special case. Suppose that we take $\rho_{\mathcal{V}}$ to be the 0-1 metric, meaning that $\rho_{\mathcal{V}}(v, v') = 0$ if $v = v'$ and 1 otherwise. Setting $t = 0$ in Proposition 2.1, we have $P_0 = \mathbb{P}[\widehat{V} \neq V]$ and $N_0^{\min} = N_0^{\max} = 1$, whence inequality (2.12) reduces to inequality (2.9). Other weakenings allow somewhat clearer statements (see Section 2.4.2 for a proof):

Corollary 2.1. *If V is uniform on $\mathcal{V}$ and $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$, then*

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}. \tag{2.13}$$

Inequality (2.13) is the natural analogue of the classical mutual-information based form of Fano's inequality (2.10), and it provides a qualitatively similar bound. The main difference is that the usual cardinality $|\mathcal{V}|$ is replaced by the ratio $|\mathcal{V}|/N_t^{\max}$. This quantity serves as a rough measure of the number of possible "regions" in the space $\mathcal{V}$ that are distinguishable—that is, the number of subsets of $\mathcal{V}$ for which $\rho_{\mathcal{V}}(v, v') > t$ when v and v' belong to different regions. While this construction is similar in spirit to the usual construction of packing sets in the standard reduction from testing to estimation (cf. Section 2.2.1), our bound allows us to skip the packing set construction. We can directly compute $I(V; X)$ where V takes values over the full space, as opposed to computing the mutual information $I(V'; X)$ for a random variable V' uniformly distributed over a packing set contained within $\mathcal{V}$. In some cases, the former calculation can be much simpler, as illustrated in examples and chapters to follow.

We now turn to providing a few consequences of Proposition 2.1 and Corollary 2.1, showing how they can be used to derive lower bounds on the minimax risk. Proposition 2.1 is a generalization of the classical Fano inequality (2.9), so it leads naturally to a generalization of the classical Fano lower bound on minimax risk, which we describe here. This reduction from estimation to testing is somewhat more general than the classical reductions, since we do not map the original estimation problem to a strict test, but rather a test that allows errors. Consider as in the standard reduction of estimation to testing in Section 2.2.1 a family of distributions $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by a finite set $\mathcal{V}$. This family induces an associated collection of parameters $\{\theta_v := \theta(P_v)\}_{v \in \mathcal{V}}$. Given a function $\rho_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and a scalar t, we define the separation $\delta(t)$ of this set relative to the metric ρ on Θ via

$$\delta(t) := \sup \left\{ \delta \mid \rho(\theta_v, \theta_{v'}) \geq \delta \text{ for all } v, v' \in \mathcal{V} \text{ such that } \rho_{\mathcal{V}}(v, v') > t \right\}. \tag{2.14}$$

As a special case, when $t = 0$ and $\rho_{\mathcal{V}}$ is the discrete metric, this definition reduces to that of a packing set: we are guaranteed that $\rho(\theta_v, \theta_{v'}) \geq \delta(0)$ for all distinct pairs $v \neq v'$, as in the classical approach to minimax lower bounds. On the other hand, allowing for $t > 0$ lends greater flexibility to the construction, since only certain pairs θ_v and $\theta_{v'}$ are required to be well-separated.

Given a set $\mathcal{V}$ and associated separation function (2.14), we assume the canonical estimation setting: nature chooses $V \in \mathcal{V}$ uniformly at random, and conditioned on this choice $V = v$, a sample X is drawn from the distribution P_v. We then have the following corollary of Proposition 2.1, whose argument is completely identical to that for inequality (2.5):

Corollary 2.2. *Given V uniformly distributed over $\mathcal{V}$ with separation function $\delta(t)$, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi\Big(\frac{\delta(t)}{2}\Big) \left[1 - \frac{I(X; V) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}\right] \qquad \text{for all } t. \tag{2.15}$$

Notably, using the discrete metric $\rho_{\mathcal{V}}(v, v') = \mathbf{1}\{v \neq v'\}$ and taking $t = 0$ in the lower bound (2.15) gives the classical Fano lower bound on the minimax risk based on constructing a packing [101, 188, 185]. We now turn to an example illustrating the use of Corollary 2.2 in providing a minimax lower bound on the performance of regression estimators.

Example: Normal regression model Consider the d-dimensional linear regression model $Y = X\theta + \varepsilon$, where $\varepsilon \in \mathbb{R}^n$ is i.i.d. $\mathsf{N}(0, \sigma^2)$ and $X \in \mathbb{R}^{n \times d}$ is known, but θ is not. In this case, our family of distributions is

$$\mathcal{P}_X := \left\{Y \sim \mathsf{N}(X\theta, \sigma^2 I_{n\times n}) \mid \theta \in \mathbb{R}^d\right\} = \left\{Y = X\theta + \varepsilon \mid \varepsilon \sim \mathsf{N}(0, \sigma^2 I_{n\times n}), \theta \in \mathbb{R}^d\right\}.$$

We then obtain the following minimax lower bound on the minimax error in squared ℓ_2-norm: there is a universal (numerical) constant $c > 0$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X, \|\cdot\|_2^2) \geq c\frac{\sigma^2 d^2}{\|X\|_{\mathrm{Fr}}^2} \geq \frac{c}{\gamma_{\max}(X/\sqrt{n})^2} \cdot \frac{\sigma^2 d}{n}, \tag{2.16}$$

where $\gamma_{\max}$ denotes the maximum singular value. Notably, this inequality is nearly the sharpest known bound proved via Fano inequality-based methods [37], but our technique is essentially direct and straightforward.

To see inequality (2.16), let the set $\mathcal{V} = \{-1, 1\}^d$ be the d-dimensional hypercube, and define $\theta_v = \delta v$ for some fixed $\delta > 0$. Then letting $\rho_{\mathcal{V}}$ be the Hamming metric on $\mathcal{V}$ and ρ be the usual ℓ_2-norm, the associated separation function (2.14) satisfies $\delta(t) > \max\{\sqrt{t}, 1\}\delta$. Now, for any $t \leq \lceil d/3 \rceil$, the neighborhood size satisfies

$$N_t^{\max} = \sum_{\tau=0}^{t} \binom{d}{\tau} \leq 2\binom{d}{t} \leq 2\left(\frac{de}{t}\right)^t.$$

Consequently, for $t \leq d/6$, the ratio $|\mathcal{V}|/N_t^{\max}$ satisfies

$$\log \frac{|\mathcal{V}|}{N_t^{\max}} \geq d \log 2 - \log 2 \binom{d}{t} \geq d \log 2 - \frac{d}{6} \log(6e) - \log 2 = d \log \frac{2}{2^{1/d} \sqrt[6]{6e}} > \max \left\{ \frac{d}{6}, \log 4 \right\}$$

for $d \geq 12$. (The case $2 \leq d < 12$ can be checked directly). In particular, by taking $t = \lfloor d/6 \rfloor$ we obtain via Corollary 2.2 that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \geq \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4} \left(1 - \frac{I(Y;V) + \log 2}{\max\{d/6, 2 \log 2\}}\right).$$

But of course, for V uniform on $\mathcal{V}$, we have $\mathbb{E}[VV^\top] = I_{d \times d}$, and thus for V, V' independent and uniform on $\mathcal{V}$,

$$\begin{aligned} I(Y;V) &\leq n \frac{1}{|\mathcal{V}|^2} \sum_{v \in \mathcal{V}} \sum_{v' \in \mathcal{V}} D_{\mathrm{kl}} \left(\mathsf{N}(X\theta_v, \sigma^2 I_{n \times n}) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_{n \times n}) \right) \\ &= \frac{\delta^2}{2\sigma^2} \mathbb{E}\left[\|XV - XV'\|_2^2 \right] = \frac{\delta^2}{\sigma^2} \|X\|_{\mathrm{Fr}}^2 . \end{aligned}$$

Substituting this into the preceding minimax bound, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \geq \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4} \left(1 - \frac{\delta^2 \|X\|_{\mathrm{Fr}}^2 / \sigma^2 + \log 2}{\max\{d/6, 2 \log 2\}}\right).$$

Choosing $\delta^2 \asymp d\sigma^2 / \|X\|_{\mathrm{Fr}}^2$ gives the result (2.16).

2.2.4 Assouad's method

Assouad's method provides a somewhat different technique for proving lower bounds. Instead of reducing the estimation problem to a multiple hypothesis test or simpler estimation problem, as with Le Cam's method and Fano's method from the preceding sections, here we transform the original estimation problem into multiple binary hypothesis testing problems, using the structure of the problem in an essential way. For some $d \in \mathbb{N}$, let $\mathcal{V} = \{-1, 1\}^d$, and let us consider a family $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by the hypercube. We say that the the family P_v induces a 2δ-*Hamming separation* for the loss $\Phi \circ \rho$ if there exists a function $\mathsf{v} : \theta(\mathcal{P}) \to \{-1, 1\}^d$ satisfying

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{ [\mathsf{v}(\theta)]_j \neq v_j \right\}. \tag{2.17}$$

As in the standard reduction from estimation to testing, we consider the following random process: nature chooses a vector $V \in \{-1, 1\}^d$ uniformly at random, after which the sample $X_1, \ldots, X_n$ is drawn from the distribution P_v conditional on $V = v$. Then, if we let $\mathbb{P}_{\pm j}$ denote the joint distribution over the random index V and X conditional on the jth coordinate $V_j = \pm 1$, we obtain the following sharper version of Assouad's lemma [11] (see also the paper [9]; we provide a proof in Section 2.4.3).

Lemma 2.2. *Under the conditions of the previous paragraph, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \inf_{\Psi} \left[\mathbb{P}_{+j}(\Psi(X_{1:n}) \neq +1) + \mathbb{P}_{-j}(\Psi(X_{1:n}) \neq -1) \right].$$

While Lemma 2.2 requires conditions on the loss Φ and metric ρ for the separation condition (2.17) to hold, it is sometimes easier to apply than Fano's method, and it appears to allow easier application in so-called "interactive" settings: those for which the sampling of the X_i may not be precisely i.i.d. It is closely related to Le Cam's method, discussed previously, as we see that if we define $P^n_{+j} = 2^{1-d} \sum_{v:v_j=1} P^n_v$ (and similarly for $-j$), Lemma 2.2 is equivalent to

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \left[1 - \left\| P^n_{+j} - P^n_{-j} \right\|_{\mathrm{TV}} \right]. \tag{2.18}$$

We conclude this section with an example application of Assouad's lemma to a minimax lower bound for a normal mean estimation problem.

Example: Normal mean estimation For some $\sigma^2 > 0$ and $d \in \mathbb{N}$, we consider estimation of mean parameter for the normal location family

$$\mathcal{N} := \left\{ \mathsf{N}(\theta, \sigma^2 I_{d\times d}) : \theta \in \mathbb{R}^d \right\}$$

in squared Euclidean distance. We now show how for this family, the sharp Assouad's method implies the lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{d\sigma^2}{8n}. \tag{2.19}$$

Up to constant factors, this bound is sharp; the sample mean has mean squared error $d\sigma^2/n$.

We proceed in (essentially) the usual way we have set up. Fix some $\delta > 0$ and define $\theta_v = \delta v$, taking $P_v = \mathsf{N}(\theta_v, \sigma^2 I_{d\times d})$ to be the normal distribution with mean θ_v. In this case, we see that the hypercube structure is natural, as our loss function decomposes on coordinates: we have $\|\theta_v - \theta_{v'}\|_2^2 = 4\delta^2 \sum_{j=1}^d \mathbf{1}\left\{v_j \neq v'_j\right\}$. The family P_v thus induces a $4\delta^2$-Hamming separation for the loss $\|\cdot\|_2^2$, and by Assouad's method (2.18), we have

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq 2\delta^2 \sum_{j=1}^{d} \left[1 - \left\| P^n_{+j} - P^n_{-j} \right\|_{\mathrm{TV}} \right].$$

It remains to provide upper bounds on $\|P^n_{+j} - P^n_{-j}\|_{\mathrm{TV}}$. By the convexity of $\|\cdot\|^2_{\mathrm{TV}}$ and Pinsker's inequality, we have

$$\left\| P^n_{+j} - P^n_{-j} \right\|^2_{\mathrm{TV}} \leq \max_{d_{ham}(v,v')\leq 1} \|P^n_v - P^n_{v'}\|^2_{\mathrm{TV}} \leq \frac{1}{2} \max_{d_{ham}(v,v')\leq 1} D_{\mathrm{kl}}\left(P^n_v \| P^n_{v'}\right).$$

But of course, for any v and v' differing in only 1 coordinate,

$$D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) = \frac{n}{2\sigma^2} \left\|\theta_v - \theta_{v'}\right\|_2^2 = \frac{2n}{\sigma^2}\delta^2,$$

giving the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq 2\delta^2 \sum_{j=1}^{d} \left[1 - \sqrt{2n\delta^2/\sigma^2}\right].$$

Choosing $\delta^2 = \sigma^2/8n$ gives the claimed lower bound (2.19).

2.3 Summary

We have seen reductions to testing and the error bounds from Le Cam's method (2.7), the Fano method (2.10) and (2.15), and Assouad's method (2.18). Consequently, to obtain bounds on the minimax risk (2.1), we control divergences between probability distributions of many forms: by controlling variation distances of the form $\|P_1^n - P_2^n\|_{\rm TV}$, mutual information quantities between random parameter indices V and the sequence of random variables $X_1, \ldots, X_n$, or other distances between mixtures of distributions. In addition to these (essentially standard) techniques for providing minimax lower bounds, we also develop techniques in this thesis for providing lower bounds for the more complicated *constrained* minimax risk (2.4). In short, it is often the case that as a consequence of constraining our statistical learning procedures $\widehat{\theta}$ to belong to some class $\mathcal{C}$, we see a sequence $Z_1, \ldots, Z_n$ related (but not identical) to the original observations $X_1, \ldots, X_n$. To provide minimax bounds, we thus must (i) develop an understanding of how these constraints give rise to the Z_i and (ii) see precisely how observing Z rather than X effects the divergence measures, such as $I(V; Z)$ in Fano's method (2.10), and the other associated probability distributions in our minimax lower bounds. Developing an understanding of the probabilistic structure of the observed variables Z becuase of the constraints place on the method $\widehat{\theta}$ leads to several new challenges, and we devote the remaining chapters to these tasks.

2.4 Proofs of results

2.4.1 Proof of Proposition 2.1

Our argument for proving the proposition parallels that of the classical Fano inequality by Cover and Thomas [47]. Letting E be a $\{0, 1\}$-valued indicator variable for the event $\rho(\widehat{V}, V) \leq t$, we compute the entropy $H(E, V \mid \widehat{V})$ in two different ways. On one hand, by the chain rule for entropy, we have

$$H(E, V \mid \widehat{V}) = H(V \mid \widehat{V}) + \underbrace{H(E \mid V, \widehat{V})}_{=0}, \tag{2.20}$$

where the final term vanishes since E is $(V, \widehat{V})$-measurable. On the other hand, we also have

$$H(E, V \mid \widehat{V}) = H(E \mid \widehat{V}) + H(V \mid E, \widehat{V}) \leq H(E) + H(V \mid E, \widehat{V}),$$

using the fact that conditioning reduces entropy. Applying the definition of conditional entropy yields

$$H(V \mid E, \widehat{V}) = \mathbb{P}(E = 0) H(V \mid E = 0, \widehat{V}) + \mathbb{P}(E = 1) H(V \mid E = 1, \widehat{V}),$$

and we upper bound each of these terms separately. For the first term, we have

$$H(V \mid E = 0, \widehat{V}) \leq \log(|\mathcal{V}| - N_t^{\min}),$$

since conditioned on the event $E = 0$, the random variable V may take values in a set of size at most $|\mathcal{V}| - N_t^{\min}$. For the second, we have

$$H(V \mid E = 1, \widehat{V}) \leq \log N_t^{\max},$$

since conditioned on $E = 1$, or equivalently on the event that $\rho(\widehat{V}, V) \leq t$, we are guaranteed that V belongs to a set of cardinality at most $N_t^{\max}$.

Combining the pieces and and noting $\mathbb{P}(E = 0) = P_t$, we have proved that

$$H(E, V \mid \widehat{V}) \leq H(E) + P_t \log \left(|\mathcal{V}| - N^{\min} \right) + (1 - P_t) \log N_t^{\max}.$$

Combining this inequality with our earlier equality (2.20), we see that

$$H(V \mid \widehat{V}) \leq H(E) + P_t \log(|\mathcal{V}| - N_t^{\min}) + (1 - P_t) \log N_t^{\max}.$$

Since $H(E) = h_2(P_t)$, the claim (2.12) follows.

2.4.2 Proof of Corollary 2.1

First, by the information-processing inequality [e.g. 47, Chapter 2], we have $I(V; \widehat{V}) \leq I(V; X)$, and hence $H(V \mid X) \leq H(V \mid \widehat{V})$. Since $h_2(P_t) \leq \log 2$, inequality (2.12) implies that

$$H(V \mid X) - \log N_t^{\max} \leq H(V \mid \widehat{V}) - \log N_t^{\max} \leq \mathbb{P}(\rho(\widehat{V}, V) > t) \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log 2.$$

Rearranging the preceding equations yields

$$\mathbb{P}(\rho(\widehat{V}, V) > t) \geq \frac{H(V \mid X) - \log N_t^{\max} - \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}}. \tag{2.21}$$

Note that his bound holds without any assumptions on the distribution of V.

By definition, we have $I(V; X) = H(V) - H(V \mid X)$. When V is uniform on $\mathcal{V}$, we have $H(V) = \log |\mathcal{V}|$, and hence $H(V \mid X) = \log |\mathcal{V}| - I(V; X)$. Substituting this relation into the bound (2.21) yields the inequality

$$\mathbb{P}(\rho(\widehat{V}, V) > t) \geq \frac{\log \frac{|\mathcal{V}|}{N_t^{\max}}}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}} - \frac{I(V; X) + \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}} \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}.$$

2.4.3 Proof of Lemma 2.2

Fix an (arbitrary) estimator $\widehat{\theta}$. By assumption (2.17), we have

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\mathsf{v}(\theta)]_j \neq v_j\right\}.$$

Taking expectations, we see that

$$\begin{aligned}\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\Phi(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P)))\right] &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\Phi(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta_v))\right] \\ &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} 2\delta \sum_{j=1}^{d} \mathbb{E}_{P_v}\left[\mathbf{1}\left\{[\psi(\widehat{\theta})]_j \neq v_j\right\}\right]\end{aligned}$$

as the average is smaller than the maximum of a set and using the separation assumption (2.17). Recalling the definition of the mixtures $\mathbb{P}_{\pm j}$ as the joint distribution of V and X conditional on $V_j = \pm 1$, we swap the summation orders to see that

$$\begin{aligned}\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v\left([\mathsf{v}(\widehat{\theta})]_j \neq v_j\right) &= \frac{1}{|\mathcal{V}|} \sum_{v : v_j = 1} P_v\left([\mathsf{v}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{|\mathcal{V}|} \sum_{v : v_j = -1} P_v\left([\mathsf{v}(\widehat{\theta})]_j \neq v_j\right) \\ &= \frac{1}{2}\mathbb{P}_{+j}\left([\mathsf{v}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{2}\mathbb{P}_{-j}\left([\mathsf{v}(\widehat{\theta})]_j \neq v_j\right).\end{aligned}$$

This gives the statement claimed in the lemma, while taking an infimum over all testing procedures $\Psi : \mathcal{X}^n \to \{-1, +1\}$ gives the claim (2.18).

Part II

Optimization

Chapter 3

Stochastic optimization and adaptive gradient methods

In this part of the thesis, we consider a variety of stochastic convex optimization problems and associated algorithms for solving them. Throughout this and the next several chapters, we focus on a single mathematical program. For a fixed closed convex subset $\Theta \subset \mathbb{R}^d$ of $\mathbb{R}^d$, consider the following optimization problem:

$$\text{minimize } f(\theta) := \mathbb{E}\left[F(\theta; X)\right] = \int_{\mathcal{X}} F(\theta; x) dP(x), \quad \text{subject to } \theta \in \Theta, \tag{3.1}$$

where for P-almost every $x \in \mathcal{X}$, the function $\theta \mapsto F(\theta; x)$ is convex. The problem (3.1) is challenging for many reasons, though we focus mainly on one throughout: in many cases, f cannot actually be evaluated. When x is high-dimensional, the integral (3.1) cannot be efficiently computed, and in statistical learning problems we usually do not even know what the distribution P is; indeed, inferring properties of the distribution P—such as the minimizer of the objective (3.1)—is the main goal of statistical estimation. The applications of such stochastic optimization problems are numerous: much of modern machine learning relies on minimization of objectives of the form (3.1), and there has been a huge literature on stochastic optimization (and closely related online learning) techniques off of which we build (a partial list of references includes work by Nemirovski and Yudin [134], Zinkevich [190], Cesa-Bianchi and Lugosi [39], Nemirovski et al. [135], Lan [114], Shalev-Shwartz [159], Nesterov [138], and Xiao [182]).

It is of great interest to develop algorithms for solving the problem (3.1). In addition, due to the potential complexity of such minimization algorithms, it is also essential to understand precisely the complexity of such schemes: are there notions under which different schemes may be called optimal? Can we develop such optimal schemes? Can we leverage the structure of the problem (3.1) to solve problems even more quickly, perhaps by using parallel computation or randomization? In the coming chapters, we illustrate a few answers to these questions.

As part of our desire to guarantee optimality, we require a notion of the computational complexity of optimization procedures as described in the introduction to this thesis. A natural notion of complexity for numerical and optimization problems is *information based complexity*, studied in depth by Nemirovski and Yudin [134], as well as by Traub et al. [169] and Plaskota [144], which is often simpler to work with than Turing Machine or other models of computation. Given the prevalence of function- and gradient-based optimization schemes, it is natural in our setting to assume access to an oracle that, when queried with a point $\theta \in \mathbb{R}^d$, draws a random sample $X \sim P$ and returns one (or both) of an instantaneous function evaluation $F(\theta; X)$ or gradient $\nabla F(\theta; X)$ of the loss F with respect to θ. The computational question is then as follows: given a pre-specified $\epsilon > 0$, how few function (or gradient) evaluations do we require to solve the problem (3.1) to within accuracy ϵ? Casting this in the minimax framework of Chapter 2, we let $\mathcal{C}_n$ denote the class of estimation procedures using at most n function (or gradient) evaluations. In this case, the minimax rate (2.3) (also (2.4)) associated with the problem (3.1) becomes

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F, \mathcal{C}_n) := \inf_{\widehat{\theta} \in \mathcal{C}_n} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}[f_P(\widehat{\theta}(X_1, \ldots, X_n))] - \inf_{\theta \in \Theta} f_P(\theta) \right\},$$

where the risk functional $f_P(\theta) = \int F(\theta; x) dP(x)$ and the infimum is taken over all estimation procedures in the class $\mathcal{C}_n$. While this is a theoretical object, we will see how an understanding of its properties prompts us to develop and evaluate new algorithms with good practical and theoretical performance. We note in passing that this minimax risk is a somewhat more fine-grained object than some similar quantities studied previously [134, 6]. In particular, we provide lower bounds on this minimax risk for a *fixed* instantaneous loss function F rather than considering an entire class of such loss functions; our upper bounds, on the other hand, apply uniformly over classes of certain types of loss functions.

3.1 Stochastic optimization algorithms

We begin this chapter by reviewing several algorithms developed for stochastic optimization, focusing on stochastic gradient-based algorithms for solving problem (3.1). The classical gradient algorithm for minimization (e.g. [32]) is as follows: starting from a point θ^1, we repeatedly iterate

$$g^k = \nabla f(\theta^k), \quad \theta^{k+1} = \theta^k - \alpha g^k,$$

where $\alpha > 0$ is a stepsize. As noted previously, this algorithmic scheme is either expensive—because the integral in expression (3.1) is difficult to compute—or impossible, because the distribution P is not even known. To address these issues, several authors [134, 190, 40, 39, 162, 135] have suggested and analyzed stochastic and online gradient-based methods. In this case, the iterative scheme is as follows: at iteration k, we draw a random $X_k \sim P$, then compute

$$g^k \in \partial F(\theta^k; X_k) \quad \text{and update} \quad \theta^{k+1} = \Pi_\Theta \left(\theta^k - \alpha_k g^k \right), \tag{3.2}$$

where we recall that Π_Θ denotes the (Euclidean) projection onto the domain Θ and α_k is a non-increasing stepsize sequence (where α_k may depend on the sequences $\theta^1, \ldots, \theta^k$ and $g^1, \ldots, g^k$). More rigorously, we assume there exists a measurable subgradient selection

$$\mathsf{g} : \Theta \times \mathcal{X} \to \mathbb{R}^d \text{ such that } \mathsf{g}(\theta; x) \in \partial F(\theta; x) \text{ for all } \theta \text{ and } P\text{-a.e. } x \in \mathcal{X}, \tag{3.3}$$

and we take $g^k = \mathsf{g}(\theta^k; X_k)$ for all k. When F is convex, this measureability implies the containment $\mathbb{E}[\mathsf{g}(\theta; X)] \in \partial f(\theta)$ for all θ (e.g. Bertsekas [25], Rockafellar and Wets [153]).

The convergence behavior of the method (3.2) is not challenging to analyze; for example, see the lecture notes of Boyd and Mutapcic [31] for an elegant and standard proof based on expanding the squared distance $\|\theta^k - \theta^*\|_2^2$ as a function of θ^{k-1}. The proposition is also a consequence of more general results on mirror descent methods (also known as nonlinear projected subgradient methods) due to Nemirovski and Yudin [134] and Beck and Teboulle [18], which we present shortly.

Proposition 3.1. *Let the gradient method* (3.2) *be run for n iterations and assume there exists a finite radius r_2 such that $\left\|\theta^k - \theta^*\right\|_2 \le \mathsf{r}_2$ for all iterations k. Then*

$$\frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n f(\theta^k) - f(\theta^*)\right] \le \mathbb{E}\left[\frac{1}{2\alpha_n}\mathsf{r}_2^2 + \frac{1}{2}\sum_{k=1}^n \alpha_k \left\|g^k\right\|_2^2\right].$$

In passing, we note that if $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \le M^2$ for all θ, then taking $\alpha_k \equiv \alpha = \mathsf{r}_2/M\sqrt{n}$ in Proposition 3.1 gives

$$\mathbb{E}[f(\widehat{\theta}(n)) - f(\theta^*)] \le \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n f(\theta^k) - f(\theta^*)\right] \le \frac{\mathsf{r}_2 M}{\sqrt{n}}, \tag{3.4}$$

where $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$ is the average parameter. This is the well-known $\mathcal{O}(n^{-\frac{1}{2}})$ convergence rate for stochastic gradient descent [134, 18, 190].

We now present the (stochastic) mirror descent method (following the presentation of Beck and Teboulle [18]) and work through a few of its consequences, but we first rewrite the method (3.2) in a manner more amenable to the coming generalization. First, the update (3.2) is, via algebraic manipulation, equivalent to

$$\theta^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{\left\langle g^k, \theta \right\rangle + \frac{1}{2\alpha}\left\|\theta - \theta^k\right\|_2^2\right\}.$$

Now, instead of using the Euclidean distance in the update (3.2), we can replace it with another distance-like function to obtain mirror descent [134, 18]. For this, we require a few definitions to develop our new distance-like functions.

Definition 3.1. *A function f is c-*strongly convex *with respect to the norm $\|\cdot\|$ over a domain Θ if for all $\theta, \theta' \in \Theta$ and any $g \in \partial f(\theta)$,*

$$f(\theta') \ge f(\theta) + \langle g, \theta' - \theta \rangle + \frac{c}{2}\left\|\theta - \theta'\right\|^2.$$

With this definition, we can define a proximal function:

Definition 3.2. *A function ψ is a* prox-function *for the set Θ if ψ is differentiable and 1-strongly convex with respect to a norm $\|\cdot\|$ over Θ.*

With a proximal function in hand, we may define the associated *Bregman divergence*

$$D_\psi(\theta, \theta') := \psi(\theta) - \psi(\theta') - \left\langle \nabla\psi(\theta'), \theta - \theta' \right\rangle. \tag{3.5}$$

For any proximal function ψ, the divergence D_ψ is always non-negative, convex in its first argument, and satisfies $D_\psi(\theta, \theta') \geq \frac{1}{2}\left\|\theta - \theta'\right\|^2$.

The mirror descent (MD) method generates a series of iterates $\{\theta^k\}_{k=1}^\infty$ contained in Θ using (stochastic) gradient information to perform the update from iterate to iterate. The algorithm is initialized at some point $\theta^1 \in \Theta$. At iterations $k = 1, 2, 3, \ldots$, the mirror descent method receives a (subgradient) vector $g^k \in \mathbb{R}^d$, which it uses to compute the next iterate via the Bregman divergence-based update

$$\theta^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \left\langle g^k, \theta \right\rangle + \frac{1}{\alpha_k} D_\psi(\theta, \theta^k) \right\}. \tag{3.6}$$

In the standard stochastic mirror descent method, the vectors g^k are stochastic (sub)gradients satisfying $g^k = \mathsf{g}(\theta^k; X_k) \in \partial F(\theta^k; X_k)$ for $X_k \stackrel{\text{i.i.d.}}{\sim} P$ as in the standard (projected) stochastic gradient descent method.

With this update scheme, we obtain the following proposition, whose proof (essentially due to Beck and Teboulle [18], with some extensions by Nemirovski et al. [135]) we provide for completeness in Section 3.5.1.

Proposition 3.2. *Let θ^k be generated according to the stochastic mirror descent method* (3.6) *and let $\theta^* \in \Theta$. Additionally, assume that there is a radius $\mathsf{r}_\psi < \infty$ such that $D_\psi(\theta^*, \theta) \leq \mathsf{r}_\psi^2$ for all $\theta \in \Theta$. Then*

$$\frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n f(\theta^k) - f(\theta^*)\right] \leq \mathbb{E}\left[\frac{1}{n\alpha_n}\mathsf{r}_\psi^2 + \frac{1}{2n}\sum_{k=1}^n \alpha_k \left\|g^k\right\|_*^2\right].$$

To see how Proposition 3.2 implies 3.1, we take $\psi(\theta) = \frac{1}{2}\left\|\theta\right\|_2^2$, in which case the divergence $D_\psi(\theta, \theta') = \frac{1}{2}\|\theta - \theta'\|_2^2$, and we recover the results for stochastic gradient descent. Now let us assume that there exists a constant $M < \infty$ such that $\mathbb{E}[\|\partial F(\theta; X)\|_*^2] \leq M^2$ for all $\theta \in \Theta$. Taking $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$ and using the convexity of f in Proposition 3.2 implies

$$\mathbb{E}[f(\widehat{\theta}(n))] - f(\theta^*) \leq \mathbb{E}\left[\frac{1}{n\alpha_n}\mathsf{r}_\psi^2 + \frac{M^2}{2n}\sum_{k=1}^n \alpha_k\right],$$

and if we choose $\alpha_k = \sqrt{2}\mathsf{r}_\psi / M\sqrt{n}$ to minimize the preceding bound we obtain

$$\mathbb{E}[f(\widehat{\theta}(n))] - f(\theta^*) \leq \frac{\sqrt{2}M\mathsf{r}_\psi}{\sqrt{n}}. \tag{3.7}$$

In Section 3.3, we specialize this result to give a few more concrete bounds and associated optimality guarantees.

We also present one final algorithm, a variant of Nesterov's dual averaging algorithm [138], that provides similar convergence guarantees, but often proves more natural for certain distributed and asynchronous algorithms (e.g. [54]). In this case, we assume there exists a sequence $\{\psi_k\}$ of proximal functions, each strongly convex over the domain Θ with respect to a norm $\|\cdot\|_{\psi_k}$, whose dual norm we denote by $\|\cdot\|_{\psi_k^*}$. In the dual averaging algorithm, one iteratively constructs a sequence of points θ^k via the following iteration: at iteration k, we sample $g^k = \mathsf{g}(\theta^k; X_k) \in \partial F(\theta^k; X_k)$, where $X_k \sim P$, but we additionally maintain a dual vector z, defined via

$$z^k = \sum_{i=1}^{k} g^i.$$

After computing this dual vector, the dual averaging update (we present a slightly more general update that allows the proximal functions to change over time) then sets

$$\theta^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \left\langle z^k, \theta \right\rangle + \frac{1}{\alpha} \psi_{k+1}(\theta) \right\}. \tag{3.8}$$

While this method is perhaps not as intuitive as the simpler stochastic gradient methods (3.2) or (3.6), its convergence behavior is similar. Indeed, if for all iterations k we have $\psi_k(\cdot) = \frac{1}{2} \|\cdot\|_2^2$ and the domain $\Theta = \mathbb{R}^d$, then dual averaging (3.8) and stochastic gradient descent (3.2) are identical with fixed stepsize α.

Because dual averaging is essential to our further arguments, we present a few "raw" convergence results for the method here. We begin with a lemma that captures a regret bound (see, e.g. Cesa-Bianchi and Lugosi [39] for definitions of regret) for the method (3.8), but the result itself is new: for one, it allows the method to use non-standard vectors z^k at each iteration, and secondly, the method allows the proximal function to change between iterations. For our theoretical development, we define the conjugate to ψ_k and associated dual norm

$$\psi_k^*(z) := \sup_{\theta \in \Theta} \left\{ \langle z, \theta \rangle - \psi_k(\theta) \right\} \quad \text{and} \quad \|z\|_{\psi_k^*} := \sup_x \left\{ \langle z, \theta \rangle \mid \|\theta\|_{\psi_k} \le 1 \right\}.$$

In the lemma, we set $\theta^0 = \operatorname{argmin}_{\theta \in \Theta} \psi_0(\theta)$. With these definitions, we have

Lemma 3.1. *Let θ^k be generated via the update* (3.8) *for all k, where z^k is an arbitrary sequence of vectors. In addition, let $\{x_k\} \subset \mathcal{X}$ be an arbitrary sequence in $\mathcal{X}$ and assume that $g^k \in \partial F(\theta^k; x_k)$. Define the "corrected" point sequence*

$$\widetilde{\theta}^k := \operatorname*{argmin}_{\theta \in \Theta} \left\{ \sum_{i=1}^{k-1} \left\langle g^i, \theta \right\rangle + \frac{1}{\alpha} \psi_k(\theta) \right\}.$$

For any sequence of observations x_k and any $\theta^* \in \Theta$,

$$\sum_{k=1}^n \left[F(\theta^k; x_k) - F(\theta^*; x_k)\right] \le \sum_{k=1}^n \alpha^{-1}\left[\psi_k^*\left(-\sum_{i=1}^{k-1} g^i\right) - \psi_{k-1}^*\left(-\sum_{i=1}^{k-1} g^i\right)\right] + \frac{\alpha}{2}\sum_{k=1}^n \left\|g^k\right\|_{\psi_k^*}^2 + \sum_{k=1}^n \left\langle g^k, \theta^k - \widetilde{\theta}^k\right\rangle + \frac{1}{\alpha}\left[\psi_n(\theta^*) - \psi_0(\theta^0)\right].$$

See Section 3.5.2 for a proof of Lemma 3.1.

As an immediate consequence of Lemma 3.1, we note that if we take $\psi_k \equiv \psi$ for a fixed proximal function ψ, which we assume is strongly convex with respect to the norm $\|\cdot\|$ over Θ (with dual norm $\|\cdot\|_*$), and we let $z^k = \sum_{i=1}^k g^i$ be computed properly, then for any sequence $\{x_k\}$ and $\theta^* \in \Theta$, we have the regret bound

$$\sum_{k=1}^n \left[F(\theta^k; x_k) - F(\theta^*; x_k)\right] \le \frac{1}{\alpha}\left[\psi(\theta^*) - \psi(\theta^0)\right] + \frac{\alpha}{2}\sum_{k=1}^n \left\|g^k\right\|_*^2.$$

In the stochastic optimization setting, we have $x_k = X_k \stackrel{\text{i.i.d.}}{\sim} P$, and this implies the next proposition, which holds when the proximal functions are fixed as $\psi_k \equiv \psi$.

Proposition 3.3. *For any $\theta^* \in \Theta$, dual averaging* (3.8) *has convergence guarantee*

$$\frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n f(\theta^k) - f(\theta^*)\right] \le \frac{1}{\alpha n}\mathbb{E}[\psi(\theta^*) - \psi(\theta^0)] + \frac{\alpha}{2n}\mathbb{E}\left[\sum_{k=1}^n \left\|g^k\right\|_*^2\right].$$

Proof Let $\mathcal{F}_k$ denote the σ-field containing $X_1, \ldots, X_k$ and any additional randomness used to construct θ^{k+1}. By construction, we have $\theta^k \in \mathcal{F}_{k-1}$ and $\psi_k \in \mathcal{F}_{k-1}$, and by definition of the risk functional (3.1), we have

$$\mathbb{E}[F(\theta^k; X_k)] = \mathbb{E}\left[\mathbb{E}[F(\theta^k; X_k) \mid \mathcal{F}_{k-1}]\right] = \mathbb{E}[f(\theta^k)].$$

Applying Lemma 3.1 and noting that $\psi_k^* = \psi_{k-1}^*$ completes the proof. □

We remark that if we redefine $\mathsf{r}_\psi^2 = \psi(\theta^*) - \psi(\theta^0)$, then an argument paralleling inequality (3.7) guarantees that taking $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$ with stepsize choice $\alpha = \sqrt{2}\mathsf{r}_\psi/M\sqrt{n}$ yields the same convergence rate, $\mathcal{O}(1)\mathsf{r}_\psi M/\sqrt{n}$. Moreover, with additional restrictions on the it is possible to convert the results of Propositions 3.1, 3.2, and 3.3 into convergence guarantees with high probability, for example, under the sub-Gaussian type assumption

$$\mathbb{E}\left[\exp\left(\frac{\left\|\partial F(\theta; X)\right\|_*^2}{M^2}\right)\right] \le \exp(1) \quad \text{for any } \theta \in \Theta,$$

where the expectation is taken over X. For results of this type, see Nemirovski et al. [135].

3.2 Adaptive optimization

Standard stochastic subgradient methods largely follow a predetermined procedural scheme that is oblivious to the characteristics of the data being observed. Often, this can lead to non-robust optimization schemes; for example, a pre-specified stepsize schedule may not take advantage of the sizes of the observed gradient norms $\|g^k\|_*^2$ in mirror descent or dual averaging methods, yielding consequent oscillatory behavior or making too too little progress. In this section, we show that it is possible to design algorithms that attain convergence rates that are (nearly) optimal for a fixed, known domain Θ, but where no upper bound M is known *a-priori* on the size of the gradient norms $\mathbb{E}[\|\partial F(\theta; X)\|_*^2]$. In addition, we review some results due to Duchi, Hazan, and Singer [53] and McMahan and Streeter [130] on finer grained adaptivity.

3.2.1 Adaptivity and robustness to gradient magnitude

To state and understand these results, we begin with a few preliminary justifications. Looking at the convergence guarantee in Proposition 3.2 (also inequality (3.7)), assuming we use a fixed stepsize α for all iterations k, the convergence rate is governed by the quantity

$$\mathbb{E}\left[\frac{\mathsf{r}_\psi^2}{\alpha n} + \frac{\alpha}{2n}\sum_{k=1}^n \left\|g^k\right\|_*^2\right].$$

Notably, if it were possible to take the infimum over all $\alpha > 0$ in the preceding expression, we would choose $\alpha = \sqrt{2}\mathsf{r}_\psi / (\sum_{k=1}^n \left\|g^k\right\|_*^2)^{\frac{1}{2}}$, yielding convergence guarantee

$$\mathbb{E}[f(\widehat{\theta}) - f(\theta^*)] \le \frac{\sqrt{2}\mathsf{r}_\psi}{n}\mathbb{E}\left[\left(\sum_{k=1}^n \left\|g^k\right\|_*^2\right)^{\frac{1}{2}}\right].$$

By Jensen's inequality, this is always at least as good as the bound (3.7). Of course, it is impossible to select such a stepsize, but it *is* possible to achieve convergence rates that are qualitatively similar. With this in mind, let us assume that at each step k of the mirror descent method (3.6) we choose α_k as though we were optimizing the associated bound on convergence: we choose

$$\alpha_k := \alpha \frac{\mathsf{r}_\psi}{\left(\delta^2 + \sum_{i=1}^k \left\|g^i\right\|_*^2\right)^{\frac{1}{2}}}, \quad \text{where } \alpha > 0 \text{ and } \delta \ge 0 \text{ are fixed.} \tag{3.9a}$$

In dual averaging, we can accomplish a similar type of adaptivity in the update (3.8) using a non-decreasing sequence of proximal functions, where we choose

$$\psi_{k+1}(\cdot) := \left(\delta^2 + \sum_{i=1}^k \left\|g^i\right\|_*^2\right)^{\frac{1}{2}} \psi(\cdot). \tag{3.9b}$$

The analysis of the stepsize choices (3.9) is made possible by the following lemma, which shows that it is possible to nearly minimize the bound in Proposition 3.2 without knowing the norms $\|g^k\|_*$ ahead of time:

Lemma 3.2 (Auer and Gentile [12], Duchi et al. [53] Lemma 4, McMahan and Streeter [130])**.** *For any non-negative sequence $\{a_k\}_k$, where we define $0/\sqrt{0} = 0$, we have*

$$\sum_{k=1}^n \frac{a_k}{\sqrt{\sum_{i=1}^k a_i}} \le 2\left(\sum_{k=1}^n a_k\right)^{\frac{1}{2}}.$$

Proof The proof is by induction. For $n = 1$, the result is obvious, so assume it holds for $n-1$. Define $b_k = \sum_{i=1}^k a_k$. Then

$$\sum_{k=1}^n \frac{a_k}{\sqrt{\sum_{i=1}^k a_i}} = \sum_{k=1}^{n-1} \frac{a_k}{\sqrt{\sum_{i=1}^k a_i}} + \frac{a_n}{\sqrt{b_n}} \le 2\sqrt{b_n - a_n} + \frac{a_n}{\sqrt{b_n}},$$

the inequality following from the inductive hypothesis. The concavity of $x \mapsto \sqrt{x}$ implies (via the first-order concavity inequality) that $\sqrt{y} \le \sqrt{x} + (2\sqrt{x})^{-1}(y-x)$, and setting $y = b_n - a_n$ and $x = b_n$ gives $2\sqrt{b_n - a_n} \le 2\sqrt{b_n} - a_n/\sqrt{b_n}$, implying the lemma. □

Using Lemma 3.2, we obtain the following corollary to Propositions 3.2 and 3.3. See Section 3.5.3 for a proof of the corollary.

Corollary 3.1. *Define $\widehat{\theta} = \frac{1}{n}\sum_{k=1}^n \theta^k$. For the mirror descent update* (3.6)*, under the conditions of Proposition 3.2, the stepsize choice* (3.9a) *with $\delta^2 = 0$ yields*

$$\mathbb{E}[f(\widehat{\theta}) - f(\theta^*)] \le \frac{2\max\{\alpha, \alpha^{-1}\}\mathsf{r}_\psi}{n}\mathbb{E}\left[\left(\sum_{k=1}^n \left\|g^k\right\|_*^2\right)^{\frac{1}{2}}\right]. \tag{3.10a}$$

For the dual averaging update (3.8)*, under the conditions of Proposition 3.3, the proximal choice* (3.9b) *and any choice of δ such that $\delta^2 \ge \mathbb{E}[\|\partial F(\theta; X)\|_*^2]$ for all $\theta \in \Theta$ yields*

$$\mathbb{E}[f(\widehat{\theta}) - f(\theta^*)] \le \frac{2\max\{\alpha, \alpha^{-1}\}\mathsf{r}_\psi}{n}\mathbb{E}\left[\left(\sum_{k=1}^n \left\|g^k\right\|_*^2\right)^{\frac{1}{2}}\right] + \frac{\delta \mathsf{r}_\psi}{\alpha n}. \tag{3.10b}$$

The corollary shows that it is possible to (essentially) achieve the optimal convergence guarantee—to within a numerical constant factor of $\sqrt{2}$—of the "best" fixed stepsize sequence. This is a heuristic statement, as an adaptive choice of stepsizes may change the observed gradient norm terms, but the corollary does show how the stepsize choice (3.9a) is robust: so long as the radius of the optimization domain Θ is known, the mirror descent or dual averaging methods do not need to know anything about the norms of the gradients. The method adapts to these gradient sizes.

3.2.2 Adaptive gradient (AdaGrad) algorithms and sparse data

The methods of the preceding section offer a limited type of adaptation to problem instances: they look only at the sizes of the gradient norms $\|g\|$. In many applications of online and stochastic optimization, however, different dimensions may exhibit fairly heterogeneous behavior. For a motivating set of problems, consider statistical learning problems for which the input instances are of very high dimension, yet within any particular instance only a few features are non-zero. It is often the case, however, that infrequently occurring features are highly informative and discriminative. The informativeness of rare features has led practitioners to craft domain-specific feature weightings, such as TF-IDF [157], which pre-emphasize infrequently occurring features. As one example, consider a text classification problem: data $x \in \mathbb{R}^d$ represents words appearing in a document, and we wish to minimize a logistic loss $F(\theta; x) = \log(1 + \exp(\langle x, \theta \rangle))$ on the data (we encode the label implicitly with the sign of x). While instances may be very high dimensional, in any given instance, very few entries of x are non-zero [126].

From a modelling perspective, it thus makes sense to allow a *dense* predictor θ: any non-zero entry of x is potentially relevant and important. In a sense, this is dual to the standard approaches to high-dimensional problems; one usually assumes that the data x may be dense, but there are only a few relevant features, and thus a parsimonious model θ is desirous [35]. So while such sparse data problems are prevalent—natural language processing, information retrieval, and other large data settings all have significant data sparsity—they do not appear to have attracted as much study as their high-dimensional "duals" of dense data and sparse predictors.

Such problems have led us [53] and McMahan and Streeter [130] to develop algorithms that dynamically incorporate knowledge of the geometry of the data observed in earlier iterations to perform more informative gradient-based learning. Informally, these procedures give frequently occurring features very low learning rates and infrequent features high learning rates, where the intuition is that each time an infrequent feature is seen, the learner should "take notice." Thus, the adaptation facilitates finding and identifying very predictive but comparatively rare features.

The AdaGrad algorithm [53, 130] is a slightly more complicated extension of the preceding stochastic gradient methods. It maintains a diagonal matrix S, initialized as $\delta^2 I_{d \times d}$, where upon receiving a new data point x, AdaGrad performs the following: it computes $g^k = \mathsf{g}(\theta^k; x) \in \partial F(\theta^k; x)$, then updates

$$S_j^{k+1} = S_j^k + (g_j^k)^2 \quad \text{for } j \in [d],$$

where S_j denotes the jth entry of the diagonal of S. We let $G = S^{\frac{1}{2}}$ denote the square root of the diagonal matrix S (so that $G^k = (S^k)^{\frac{1}{2}}$). Depending on whether the dual averaging or stochastic gradient descent (SGD) variant is being used, AdaGrad performs one of two updates. In the dual averaging case, it maintains the dual vector z^k, which is updated by $z^k = z^{k-1} + g^k$; in the SGD case, the parameter θ^k is maintained. The updates for the two

cases are then

$$\theta^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \left\langle g^k, \theta \right\rangle + \frac{1}{2\alpha} \left\langle \theta - \theta^k, G^k(\theta - \theta^k) \right\rangle \right\}$$

for stochastic gradient descent and

$$\theta^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \left\langle z^k, \theta \right\rangle + \frac{1}{2\alpha} \left\langle \theta, G^k \theta \right\rangle \right\}$$

for dual averaging, where α is a stepsize.

Showing the convergence of AdaGrad using our prior results, specifically Propositions 3.2 and 3.3 and Lemma 3.2, is not terribly difficult. In particular, letting $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} f(\theta)$, if we have $\mathsf{r}_\infty \geq \sup_{\theta \in \Theta} \|\theta - \theta^*\|_\infty$, then choosing $\alpha = \mathsf{r}_\infty$ yields the following result. After n samples X_k, the averaged parameter vector $\widehat{\theta}(n) = \frac{1}{n} \sum_{k=1}^n \theta^k$ of AdaGrad satisfies

$$\mathbb{E}[f(\widehat{\theta}(n))] - \inf_{\theta \in \Theta} f(\theta) \leq \frac{3}{2} \frac{\mathsf{r}_\infty \mathbb{E}[\operatorname{tr}(G^n)]}{n}. \tag{3.11}$$

For a full proof for both dual averaging and the standard stochastic gradient variants, see, for example, Sections 1.3 and Theorem 5 of Duchi et al. [53]. For completeness, we include a proof of inequality (3.11) for the SGD case in Section 3.5.4. Specializing this bound to the case of sparse data, we arrive at a bound that we will show presently is sharp. Let $\mathsf{g}(\theta; X) \in \partial F(\theta; X)$ be a measurable (sub)gradient selection, and let us assume that for all $\theta \in \Theta$, we have $P(\mathsf{g}_j(\theta; X) \neq 0) \leq p_j$ and $|\mathsf{g}_j(\theta; X)| \leq M$ with probability 1. Then inequality (3.11) specializes to

$$\mathbb{E}[f(\widehat{\theta}(n))] - \inf_{\theta \in \Theta} f(\theta) \leq \frac{3}{2} \frac{\mathsf{r}_\infty M}{\sqrt{n}} \sum_{j=1}^d \sqrt{p_j}. \tag{3.12}$$

In the next section we see that this rate is optimal.

3.3 A few optimality guarantees

Having given, in the preceding sections, a general exposition of some stochastic gradient-based procedures for optimization, we now investigate some of their optimality properties. We use the minimax excess risk measure (2.3), which we recall is

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[f_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} f_P(\theta) \right\}$$

where $f_P = \mathbb{E}_P[F(\theta; X)]$, as our evaluation metric. Our techniques in this section build off of those developed by Agarwal et al. [6], who in turn were inspired by Nemirovski and Yudin [134]. Using our coming results, we provide several optimality guarantees. Specifically, stochastic gradient descent, mirror descent, and dual averaging—including their adaptive

stepsize variants—are optimal, as is AdaGrad: to within numerical constant factors, their rates of convergence are unimprovable. We show problems for which each attains the best possible rate of convergence, for AdaGrad showing that it enjoys optimality properties in those situations in which the data is sparse. These optimality guarantees can provide provide some guidance in choosing which of stochastic gradient descent, mirror descent, and dual averaging is likely to be effective for a given problem.

Let us give a more precise characterization of the set of optimization problems we consider to provide the first of the two lower bounds we give. For the next proposition, we let $\mathcal{P}$ consist distributions supported on $\mathcal{X} = \{-1, 0, 1\}^d$, and we let $p_j := P(X_j \neq 0)$ be the marginal probability of appearance of feature j ($j \in \{1, \ldots, d\}$). Assume that $\Theta \supset [-\mathsf{r}_\infty, \mathsf{r}_\infty]^d$, that is, Θ contains the ℓ_∞ ball of radius r_∞. Now given $x \in \{-1, 0, 1\}^d$, define the loss

$$F(\theta; x) := \sum_{j=1}^{d} M_j |x_j| |\theta_j - \mathsf{r}_\infty x_j|.$$

This is essentially a multi-dimensional median, where one suffers a loss only when component j of the vector x is non-zero. With this loss, we obtain the following proposition, whose proof we provide in Section 3.5.5.

Proposition 3.4. *Let the conditions of the preceding paragraph hold. Let r_∞ be a constant such that $\Theta \supset [-\mathsf{r}_\infty, \mathsf{r}_\infty]^d$. Then*

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F) \geq \frac{1}{8} \mathsf{r}_\infty \sum_{j=1}^{d} M_j \min \left\{ p_j, \frac{\sqrt{p_j}}{\sqrt{n \log 3}} \right\}.$$

We provide a few remarks here. First, this minimax lower bound essentially matches the AdaGrad rate of convergence (3.12), showing the optimality of AdaGrad (we discuss this more subsequently). Second, an inspection of the proof shows that we may assume the coordinates X_j are independent of one another in Proposition 3.4.

Third, Proposition 3.4 implies Theorem 1 of Agarwal et al. [6] as a special case, giving their result with somewhat sharper constants. Indeed, let $p_j = 1/d$ and $M_j = M$ for all j, let the coordinates of X be independent, and assume for simplicity that $d \leq n$. Then we have the minimax lower bound

$$\frac{\mathsf{r}_\infty M \sqrt{d}}{8\sqrt{n \log 3}},$$

while the gradient mapping $\mathsf{g}(\theta; x) \in \partial F(\theta; x)$ satisfies $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_1^2] \leq 2M^2$ for all $\theta \in \mathbb{R}^d$. By inspection, this is a sharper version of the bound (9) of Agarwal et al. [6], and it implies optimality guarantees for several stochastic gradient methods. For example, if Θ contains an ℓ_2-ball, that is, $\Theta \supset \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq \mathsf{r}_2\}$, then the set of optimization problems satisfying the conditions of Proposition 3.1 has minimax lower bound scaling as $\mathsf{r}_2 M / \sqrt{n}$; the convergence rate (3.4) is sharp to within a numerical constant factor of $8\sqrt{\log 3}$. We recover Agarwal et al.'s bound (10) by taking $p_j = 1$.

Our fourth remark is to give a corollary to Proposition 3.4 that follows when the data x obeys a type of power law: let $p_0 \in [0,1]$, and assume that $P(X_j \neq 0) = p_0 j^{-\alpha}$. We have

Corollary 3.2. *Let $\alpha \geq 0$. Let the conditions of Proposition 3.4 hold with $M_j \equiv M$ for all j, and assume the power law condition $P(X_j \neq 0) = p_0 j^{-\alpha}$ on coordinate appearance probabilities. Then*

(1) If $d > (p_0 n)^{1/\alpha}$,

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F) \geq \frac{M \mathsf{r}_\infty}{8} \left[\frac{2}{2-\alpha} \sqrt{\frac{p_0}{n}} \left((p_0 n)^{\frac{2-\alpha}{2\alpha}} - 1 \right) + \frac{p_0}{1-\alpha} \left(d^{1-\alpha} - (p_0 n)^{\frac{1-\alpha}{\alpha}} \right) \right].$$

(2) If $d \leq (p_0 n)^{1/\alpha}$,

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F) \geq \frac{M \mathsf{r}_\infty}{8} \sqrt{\frac{p_0}{n}} \left(\frac{1}{1-\alpha/2} d^{1-\frac{\alpha}{2}} - \frac{1}{1-\alpha/2} \right).$$

The proof of the corollary follows by an (omitted) integration argument.

By inspection, the AdaGrad rate (3.12) matches the lower bound in Proposition 3.4 and is thus optimal. It is interesting to note, though, that in the power law setting of Corollary 3.2, a calculation shows that the multiplier for the SGD guarantee (3.4) becomes $\mathsf{r}_\infty \sqrt{d} \max\{d^{(1-\alpha)/2}, 1\}$, while AdaGrad attains rate at worst $\mathsf{r}_\infty \max\{d^{1-\alpha/2}, \log d\}$ (by evaluation of $\sum_j \sqrt{p_j}$). Thus for $\alpha \in [0,1]$, the AdaGrad rate is no worse, for $\alpha > 1$, the AdaGrad rate strictly improves, and for $\alpha \geq 2$, is more than $\sqrt{d}/\log d$ better than SGD—an exponential improvement in the dimension. In general, the difference between the two rates is most apparent when the Cauchy-Schwarz inequality is loose: indeed, assume that $\Theta = [-\mathsf{r}_\infty, \mathsf{r}_\infty]^d$ is a scaled ℓ_∞-ball. In the setting of Proposition 3.4, we find the convergence rates of AdaGrad and stochastic gradient descent are

$$\frac{\mathsf{r}_\infty M \sum_{j=1}^d \sqrt{p_j}}{\sqrt{n}} \text{ [AdaGrad]} \quad \text{and} \quad \frac{\mathsf{r}_\infty M \sqrt{d} \sqrt{\sum_{j=1}^d p_j}}{\sqrt{n}} \text{ [SGD]},$$

as the radius of the set Θ in ℓ_2 norm is $\mathsf{r}_\infty \sqrt{d}$. The Cauchy-Schwarz inequality implies that the first bound is always tighter than the second, and may be as much as a factor of $\sqrt{d}$ tighter if $\sum_{j=1}^d \sqrt{p_j} = \mathcal{O}(1)$.

Lastly, we state without proof a minimax lower bound for high-dimensional sparse—or nearly sparse—optimization. It is important to understand such bounds for mirror descent methods, as they often exhibit small but non-constant dimension dependence. (We prove similar results in the sequel.) Fix $M > 0$ and for $x \in \{-1, 1\}^d$, define the linear loss $F(\theta; x) = \langle \theta, x \rangle$. Let $\mathcal{P}$ denote the family of distributions supported on $\{\pm 1\}^d$, and assume that $\Theta = \{\theta \in \Theta : \|\theta\|_1 \leq \mathsf{r}_1\}$. We obtain the following result:

Corollary 3.3 (Theorem 1 of Duchi, Jordan, and Wainwright [57]). *Under the conditions of the previous paragraph, we have minimax lower bound*

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F) \geq \frac{1}{8}\mathsf{r}_1 M \min\left\{1, \frac{\sqrt{\log(2d)}}{2\sqrt{n}}\right\}.$$

This bound is sharp to within constant factors; consider the mirror descent algorithm (3.6) or the dual averaging procedure (3.8) and assume without loss of generality that $d \geq 2$. The proximal function

$$\psi(\theta) = \frac{1}{2(p-1)} \|\theta\|_p^2 \quad \text{for } p = 1 + \frac{1}{\log d}$$

is 1-strongly convex with respect to the norm $\|\cdot\|_p$ (e.g. Ben-Tal et al. [22]), and by setting $q = 1 + \log d$ so that $p^{-1} + q^{-1} = 1$, Hölder's inequality implies

$$\|\theta\|_1 \leq \|\theta\|_p \|\mathbb{1}\|_q = \|\theta\|_p d^{\frac{1}{1+\log d}} \leq e \|\theta\|_p.$$

Thus ψ is e^{-2}-strongly convex with respect to the ℓ_1-norm, and using ψ as the proximal function in mirror descent (3.6) or dual averaging (3.8) gives a rate of convergence in inequality (3.7) identical (up to constant factors) to that in the Corollary 3.3. (See also Beck and Teboulle [18] and Nemirovski et al. [135]).

3.4 Summary

In this chapter, we have briefly reviewed several stochastic optimization algorithms and associated techniques for proving their convergence. Additionally, we have developed techniques based on Assouad's method (cf. Lemma 2.2) for proving lower bounds on the performance of gradient-based optimization methods, exhibiting specific losses that are difficult for any method to optimize. We have also provided insights into the AdaGrad method, giving some of its optimality properties. In the coming chapters, we show that in spite of our lower bounds on optimization complexity in terms of the number of gradient evaluations (or sample observations X_k), it is possible to develop faster optimization schemes by taking advantage of specific structures of the problem at hand. More specifically, we show that by allowing access to parallel computation, there are scenarios in which we can attain the same "computational" complexity in terms of gradient evaluations, but we can evaluate gradient information substantially faster, reaping real performance benefits. In addition, in Chapter 6, we extend the techniques in this section to show that when there are restrictions on the amount of information available to the method, such as observing only function values, there may be degradation in performance, but we can still develop procedures with optimality guarantees.

3.5 Proofs of convergence and minimax bounds

3.5.1 Proof of Proposition 3.2

The proof of this proposition hinges on the following lemma, which capture the behavior of a single step of the mirror descent method.

Lemma 3.3. *Let the sequence $\{\theta^k\}$ be generated by the mirror descent update* (3.6) *for any (arbitrary) g^k. Then for any $\theta^* \in \Theta$,*

$$\left\langle g^k, \theta^k - \theta^* \right\rangle \leq \frac{1}{\alpha_k} \left[D_\psi(\theta^*, \theta^k) - D_\psi(\theta^*, \theta^{k+1}) \right] + \frac{\alpha_k}{2} \left\| g^k \right\|_*^2 .$$

Proof Recall (e.g. [98, 151, 32]) that for any sub-differentiable convex function h defined on a set C, the point $x \in C$ minimizes h over C if and only if there exists some $g \in \partial h(x)$ such that

$$\langle g, y - x \rangle \geq 0 \quad \text{for all } y \in C. \tag{3.13}$$

Applying this to the Bregman-divergence based update (3.6), we see that θ^{k+1} satisfies

$$\left\langle g^k + \frac{1}{\alpha_k} \left[\nabla\psi(\theta^{k+1}) - \nabla\psi(\theta^k) \right], \theta - \theta^{k+1} \right\rangle \geq 0 \quad \text{for all } \theta \in \Theta.$$

In particular, by choosing $\theta = \theta^*$, we obtain

$$\left\langle g^k, \theta^{k+1} - \theta^* \right\rangle \leq \frac{1}{\alpha_k} \left\langle \nabla\psi(\theta^{k+1}) - \nabla\psi(\theta^k), \theta^* - \theta^k \right\rangle .$$

Via a few algebraic manipulations, we have that

$$\left\langle \nabla\psi(\theta^{k+1}) - \nabla\psi(\theta^k), \theta^* - \theta^k \right\rangle = D_\psi(\theta^*, \theta^k) - D_\psi(\theta^*, \theta^{k+1}) - D_\psi(\theta^{k+1}, \theta^k). \tag{3.14}$$

As a consequence, we have

$$\begin{aligned} \left\langle g^k, \theta^k - \theta^* \right\rangle &= \left\langle g^k, \theta^{k+1} - \theta^* \right\rangle + \left\langle g^k, \theta^k - \theta^{k+1} \right\rangle \\ &\leq \frac{1}{\alpha_k} \left[D_\psi(\theta^*, \theta^k) - D_\psi(\theta^*, \theta^{k+1}) - D_\psi(\theta^{k+1}, \theta^k) \right] + \left\langle g^k, \theta^k - \theta^{k+1} \right\rangle \end{aligned}$$

Applying the Fenchel-Young inequality, we have

$$\left\langle g^k, \theta^k - \theta^{k+1} \right\rangle \leq \frac{\alpha_k}{2} \left\| g^k \right\|_*^2 + \frac{1}{2\alpha_k} \left\| \theta^k - \theta^{k+1} \right\|^2 ,$$

and noting that $D_\psi(\theta^{k+1}, \theta^k) \geq \frac{1}{2}\|\theta^k - \theta^{k+1}\|^2$ gives the result. □

Let e^k denote the error in the subgradient estimate used in the mirror descent update (3.6), so that (at the risk of some abuse of notation) setting $\nabla f(\theta^k) = \mathbb{E}[\mathsf{g}(\theta^k; X_k) \mid \theta^k]$ to be the expected subgradient, we have $e^k = \nabla f(\theta^k) - g^k = \nabla f(\theta^k) - \mathsf{g}(\theta^k; X_k)$. Then by definition of the subgradient of the risk f, we have

$$f(\theta^k) - f(\theta^*) \le \left\langle \nabla f(\theta^k), \theta^k - \theta^* \right\rangle = \left\langle g^k, \theta^k - \theta^* \right\rangle + \left\langle e^k, \theta^k - \theta^* \right\rangle. \tag{3.15}$$

As a consequence, applying Lemma 3.3 gives

$$f(\theta^k) - f(\theta^*) \le \frac{1}{\alpha_k} \left[D_\psi(\theta^*, \theta^k) - D_\psi(\theta^*, \theta^{k+1}) \right] + \frac{\alpha_k}{2} \left\| g^k \right\|_*^2 + \left\langle e^k, \theta^k - \theta^* \right\rangle.$$

By definition of the subgradient g^k, the selection of the subgradient $\nabla f(\theta)$, and the σ-fields $\mathcal{F}_k$, we have

$$\mathbb{E}[\left\langle e^k, \theta^k - \theta^* \right\rangle] = \mathbb{E}\left[\mathbb{E}[\left\langle e^k, \theta^k - \theta^* \right\rangle \mid \mathcal{F}_{k-1}] \right] = \mathbb{E}\left[\left\langle \mathbb{E}[e^k \mid \mathcal{F}_{k-1}], \theta^k - \theta^* \right\rangle \right] = 0.$$

Thus, summing inequality (3.15) and taking expectations yields

$$\sum_{k=1}^n \mathbb{E}\left[f(\theta^k) - f(\theta^*) \right] \le \mathbb{E}\left[\sum_{k=1}^n \frac{1}{\alpha_k} \left[D_\psi(\theta^*, \theta^k) - D_\psi(\theta^*, \theta^{k+1}) \right] \right] + \mathbb{E}\left[\sum_{k=1}^n \frac{\alpha_k}{2} \left\| g^k \right\|_*^2 \right].$$

Rearranging the first summed divergences, we have

$$\begin{aligned}
&\sum_{k=1}^n \frac{1}{\alpha_k} \left[D_\psi(\theta^*, \theta^k) - D_\psi(\theta^*, \theta^{k+1}) \right] \\
&= \sum_{k=2}^n \left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}} \right) D_\psi(\theta^*, \theta^k) + \frac{1}{\alpha_1} D_\psi(\theta^*, \theta^1) - \frac{1}{\alpha_{n+1}} D_\psi(\theta^*, \theta^{n+1}) \\
&\le \sum_{k=2}^n \left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}} \right) \mathsf{r}_\psi^2 + \frac{1}{\alpha_1} \mathsf{r}_\psi^2 = \frac{1}{\alpha_n} \mathsf{r}_\psi^2,
\end{aligned}$$

where for the last inequality we used the compactness assumption of the proposition and the fact that $\alpha_{k-1} \ge \alpha_k$.

3.5.2 Proof of Lemma 3.1

To prove this lemma, we recall a definition state an auxiliary result, which is essentially standard. The result says that as ψ_k is strongly convex with respect to the norm $\|\cdot\|_{\psi_k}$, its dual is smoothly differentiable and, more strongly, has Lipschitz derivative with respect to the dual norm $\|\cdot\|_{\psi_k^*}$. For a proof of this type of standard result, see, for example, the book of Hiriart-Urruty and Lemaréchal [98, Chapter X]; the result follows by algebraic manipulations of the first-order optimality conditions for the update (3.8).

Lemma 3.4. *The function ψ_k^* is 1-strongly smooth with respect to $\|\cdot\|_{\psi_k^*}$, meaning that*

$$\|\nabla\psi_k^*(z) - \nabla\psi_k^*(z')\|_{\psi_k} \le \|z - z'\|_{\psi_k^*},$$

and moreover $\nabla\psi_k^*(-z^{k-1}) = \theta^k$.

We also recall the standard fact [e.g. 98] that if a function h has Lipschitz continuous derivative with respect to a norm $\|\cdot\|$, then $h(\theta') \le h(\theta) + \langle\nabla h(\theta), \theta' - \theta\rangle + \frac{1}{2}\|\theta - \theta'\|^2$ for all $\theta, \theta' \in \operatorname{dom} h$.

Our proof is similar to other analyses of dual averaging (e.g. [138, 53]), but we track the changing time indices. We also assume without loss of generality that $\alpha = 1$; indeed, the conjugate of $\theta \mapsto \alpha^{-1}\psi(\theta)$ is $\alpha\psi^*(\cdot)$. For shorthand throughout this proof, we define the running sum $g^{1:k} := \sum_{i=1}^k g^i$. First, we note by convexity and the definition of $g^k \in \partial F(\theta; x_k)$ that

$$\sum_{k=1}^n \left[F(\theta^k; x_k) - F(\theta^*; x_k)\right] \le \sum_{k=1}^n \left\langle g^k, \theta^k - \theta^*\right\rangle. \tag{3.16}$$

By definition of ψ_n and the conjugate $\psi_k^*(z) = \sup_{\theta\in\Theta}\{\langle z, \theta\rangle - \psi_k(\theta)\}$, we find that

$$\begin{aligned}\sum_{k=1}^n \left\langle g^k, \theta^k - \theta^*\right\rangle &= \sum_{k=1}^n \left\langle g^k, \theta^k\right\rangle + \sum_{k=1}^n \left\langle -g^k, \theta^*\right\rangle + \psi_n(\theta^*) - \psi_n(\theta^*) \\ &\le \psi_n(\theta^*) + \psi_n^*(-g^{1:n}) + \sum_{k=1}^n \left\langle g^k, \theta^k\right\rangle. \end{aligned} \tag{3.17}$$

Now, by applying Lemma 3.4 and the definition of 1-strongly-smooth, we have that

$$\psi_k^*(-g^{1:k}) \le \psi_k^*(-g^{1:k-1}) + \left\langle -g^k, \nabla\psi_k^*(-g^{1:k-1})\right\rangle + \frac{1}{2}\left\|g^k\right\|_{\psi_k^*}^2.$$

By construction of θ and $\widetilde{\theta}$, we have $\theta^k = \nabla\psi_k^*(-z^{k-1})$ and $\widetilde{\theta}^k = \nabla\psi_k^*(-g^{1:k-1})$. Thus, rearranging the preceding display, we have

$$0 \le \left\langle -g^k, \widetilde{\theta}^k\right\rangle - \psi_k^*(-g^{1:k})) + \psi_k^*(-g^{1:k-1}) + \frac{1}{2}\left\|g^k\right\|_{\psi_k^*}^2,$$

and adding $\left\langle g^k, \theta^k\right\rangle$ to both sides of the above expression gives

$$\left\langle g^k, \theta^k\right\rangle \le \left\langle g^k, \theta^k - \widetilde{\theta}^k\right\rangle - \psi_k^*(-g^{1:k}) + \psi_k^*(-g^{1:k-1}) + \frac{1}{2}\left\|g^k\right\|_{\psi_k^*}^2. \tag{3.18}$$

Thus we obtain the inequalities

$$\begin{aligned}
&\sum_{k=1}^{n} \left\langle g^k, \theta^k - \theta^* \right\rangle \\
&\stackrel{(i)}{\leq} \psi_k(\theta^*) + \psi_n^*(-g^{1:n})) + \sum_{k=1}^{n} \left\langle g^k, \theta^k \right\rangle \\
&\stackrel{(ii)}{\leq} \psi_n(\theta^*) + \psi_n^*(-g^{1:n}) + \sum_{k=1}^{n} \Big[\left\langle g^k, \theta^k - \widetilde{\theta}^k \right\rangle - \psi_k^*(-g^{1:k}) + \psi_k^*(-g^{1:k-1}) + \frac{1}{2} \left\| g^k \right\|_{\psi_k^*}^2 \Big] \\
&= \psi_n(\theta^*) + \sum_{k=1}^{n} \Big[\left\langle g^k, \theta^k - \widetilde{\theta}^k \right\rangle + \psi_k^*(-g^{1:k-1}) - \psi_{k-1}^*(-g^{1:k-1}) + \frac{1}{2} \left\| g^k \right\|_{\psi_k^*}^2 \Big] + \psi_0^*(0),
\end{aligned}$$

where for step (i) we have applied inequality (3.17), step (ii) follows from the bound (3.18), and the last equality follows by re-indexing terms in the sum. Combining the above sum with the first-order convexity inequality (3.16) proves the lemma.

3.5.3 Proof of Corollary 3.1

The proof of inequality (3.10a) is nearly immediate. The choice (3.9a), when applied in Proposition 3.2, yields

$$\mathbb{E}[f(\widehat{\theta}) - f(\theta^*)] \leq \mathbb{E}\left[\frac{\mathsf{r}_\psi}{n\alpha} \left(\sum_{k=1}^{n} \left\| g^k \right\|_*^2 \right)^{\frac{1}{2}} + \frac{\alpha \mathsf{r}_\psi}{2n} \sum_{k=1}^{n} \frac{\left\| g^k \right\|_*^2}{(\sum_{i=1}^{k} \left\| g^i \right\|_*^2)^{\frac{1}{2}}} \right].$$

Applying Lemma 3.2 and noting that $a + b \leq 2\max\{a, b\}$ gives the result.

For inequality (3.10b), we require a bit more work. Without loss of generality, we assume that $\psi(\theta^0) = 0$. By the adaptive choice of the sequence of proximal functions, we have $\psi_k \geq \psi_{k-1}$ for all k, and consequently, $\psi_k^* \leq \psi_{k-1}^*$ for all k. Inspecting Lemma 3.1, we thus obtain

$$\sum_{k=1}^{n} \left[F(\theta^k; X_k) - F(\theta^*; X_k) \right] \leq \frac{\alpha}{2} \sum_{k=1}^{n} \left\| g^k \right\|_{\psi_k^*}^2 + \frac{1}{\alpha} \psi_n(\theta^*)$$

In addition, as proximal functions ψ_k are only scaled multiples of ψ, the dual norms $\|\cdot\|_{\psi_k^*}$ are similarly scaled: we have $\|z\|_{\psi_k^*} = \mathsf{r}_\psi(\delta^2 + \sum_{i=1}^{k-1} \|g^i\|_*^2)^{-\frac{1}{2}} \|z\|_*$. Thus, the preceding display becomes

$$\begin{aligned}
\sum_{k=1}^{n} \left[F(\theta^k; X_k) - F(\theta^*; X_k) \right] &\leq \frac{\alpha \mathsf{r}_\psi}{2} \sum_{k=1}^{n} \frac{\left\| g^k \right\|_*^2}{(\delta^2 + \sum_{i=1}^{k-1} \left\| g^i \right\|_*^2)^{\frac{1}{2}}} + \frac{1}{\alpha \mathsf{r}_\psi} \left(\delta^2 + \sum_{k=1}^{n} \left\| g^k \right\|_*^2 \right)^{\frac{1}{2}} \psi(\theta^*) \\
&\leq \alpha \mathsf{r}_\psi \left(\sum_{k=1}^{n} \left\| g^k \right\|_*^2 \right)^{\frac{1}{2}} + \frac{1}{\alpha \mathsf{r}_\psi} \mathsf{r}_\psi^2 \left(\sum_{k=1}^{n} \left\| g^k \right\|_*^2 \right)^{\frac{1}{2}} + \frac{\delta \mathsf{r}_\psi^2}{\alpha \mathsf{r}_\psi},
\end{aligned}$$

where we applied the adaptivity Lemma 3.2. Taking expectations, averaging, and using the convexity of the risk f gives the result.

3.5.4 Derivation of inequality (3.11)

We focus on the mirror descent case; see Duchi et al. [53] for the full arguments. For a positive definite matrix A, we recall the definition of the Mahalanobis norm $\|x\|_A$ via $\|x\|_A^2 = \langle x, Ax \rangle$. Now, let $\psi_k(\theta) = \frac{1}{2}\|\theta\|_{G^k}^2$. Then the norm $\|\cdot\|_{\psi_k}$ is the Mahalanobis norm $\|\cdot\|_{G^k}$, with dual norm defined by $\|g\|_{\psi_k^*} = \|g\|_{(G^k)^{-1}}$. Then Lemma 3.3 implies

$$F(\theta^k; X_k) - F(\theta^*; X_k) \le \frac{1}{2\alpha}\left[\left\|\theta^k - \theta^*\right\|_{G^k}^2 - \left\|\theta^{k+1} - \theta^*\right\|_{G^k}^2\right] + \frac{\alpha}{2}\left\|g^k\right\|_{(G^k)^{-1}}^2$$

Summing this inequality gives

$$\begin{aligned}\sum_{k=1}^n [F(\theta^k; X_k) - F(\theta^*; X_k)] &\le \frac{1}{2\alpha}\sum_{k=1}^n \left[\left\|\theta^k - \theta^*\right\|_{G^k}^2 - \left\|\theta^{k+1} - \theta^*\right\|_{G^k}^2\right] + \frac{\alpha}{2}\sum_{k=1}^n \left\|g^k\right\|_{(G^k)^{-1}}^2 \\ &\le \frac{1}{2\alpha}\sum_{k=1}^n \left[\left\|\theta^k - \theta^*\right\|_{G^k}^2 - \left\|\theta^{k+1} - \theta^*\right\|_{G^k}^2\right] + \alpha \sum_{j=1}^d \left(\sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}},\end{aligned}$$

where we apply Lemma 3.2. For the first sum in the preceding bound, we note that since G is a diagonal matrix and $G^k \succeq G^{k-1}$,

$$\begin{aligned}&\sum_{k=1}^n \left[\left\|\theta^k - \theta^*\right\|_{G^k}^2 - \left\|\theta^{k+1} - \theta^*\right\|_{G^k}^2\right] \\ &= \sum_{k=2}^n \left[\left\|\theta^k - \theta^*\right\|_{G^k}^2 - \left\|\theta^k - \theta^*\right\|_{G^{k-1}}^2\right] + \left\|\theta^1 - \theta^*\right\|_{G^1}^2 - \left\|\theta^{n+1} - \theta^*\right\|_{G^n}^2 \\ &\le \sum_{k=2}^n \left\|\theta^k - \theta^*\right\|_\infty^2 \operatorname{tr}(G^k - G^{k-1}) + \left\|\theta^1 - \theta^*\right\|_\infty^2 \operatorname{tr}(G^1) \le \mathsf{r}_\infty^2 \operatorname{tr}(G^n).\end{aligned}$$

In particular, we have the convergence guarantee

$$\sum_{k=1}^n [F(\theta^k; X_k) - F(\theta^*; X_k)] \le \frac{\mathsf{r}_\infty^2}{2\alpha}\operatorname{tr}(G^n) + \alpha \operatorname{tr}(G^k).$$

Set $\alpha = \mathsf{r}_\infty$ and take expectations to complete the proof.

3.5.5 Proof of Proposition 3.4

Our proof proceeds in a few steps: we first define our family of loss functions, after which we perform an essentially standard reduction of the estimation (optimization) problem to testing. Following this step, we carefully lower bound the probabilities of error in our multiple hypothesis testing problem (in a manner similar to Assouad's method in Chapter 2.2.4) to obtain the desired statement of the proposition. For simplicity in the proof, we assume that

$M_j = M$ for all j and use the shorthand r for r_∞. We also note that any subscript j denotes a coordinate subscript of a vector and subscripting by k denotes the subscript of an observation, so $X_j \in \{-1, 0, 1\}$ is the jth coordinate of $X \in \{-1, 0, 1\}^d$, while $X_k \in \{-1, 0, 1\}^d$ denotes the kth observation.

From estimation to testing Given $x \in \{-1, 0, 1\}^d$, recall that the loss is defined as

$$F(\theta; x) := M \sum_{j=1}^{d} |x_j| |\theta_j - \mathsf{r} x_j|.$$

Letting $p_j = P(X_j \neq 0)$, $p_j^+ = P(X_j = 1)$, and $p_j^- = P(X_j = 1)$, we obtain that such a P, the associated risk f_P is

$$f_P(\theta) := \mathbb{E}_P[F(\theta; X)] = M \sum_{j=1}^{d} \left(p_j^+ |\theta_j - \mathsf{r}| + p_j^- |\theta_j + \mathsf{r}| \right),$$

so the objective f behaves like a weighted 1-norm type of quantity and its minimizer is a multi-dimensional median.

Now we proceed through a reduction of estimation to testing. Fix $\delta_j > 0$ for $j \in \{1, \ldots, d\}$ (we optimize these choices later). Let $\mathcal{V} = \{-1, 1\}^d$, and for a fixed $v \in \mathcal{V}$ let P_v be the distribution supported on $\{-1, 0, 1\}^d$ whose (independent) coordinate marginals are specified by

$$P_v(X_j = 1) = p_j \frac{1 + \delta_j v_j}{2} \quad \text{and} \quad P_v(X_j = -1) = p_j \frac{1 - \delta_j v_j}{2}. \tag{3.19}$$

Now, we claim that for any estimator $\widehat{\theta}$, we have the following analogue of the 2δ-Hamming separation (2.17) that underlies Assouad's method (cf. Lemma 2.2): for any estimator $\widehat{\theta}$,

$$\begin{aligned} \sup_{P \in \mathcal{P}} \mathbb{E}_P[f_P(\widehat{\theta}) - \inf_{\theta \in \Theta} f_P(\theta)] &\geq \max_{v \in \mathcal{V}} \mathbb{E}_{P_v}[f_{P_v}(\widehat{\theta}) - \inf_{\theta \in \Theta} f_{P_v}(\theta)] \\ &\geq M \mathsf{r} \max_{v \in \mathcal{V}} \sum_{j=1}^{d} p_j \delta_j P_v \left(\operatorname{sign}\left(\widehat{\theta}_j(X_1, \ldots, X_n) \right) \neq v_j \right), \end{aligned} \tag{3.20}$$

where the last probability distribution is the product P_v^n over the sample $X_1, \ldots, X_n$ of size n. To see that inequality (3.20) holds, define

$$\theta_v^* = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{P_v}[F(\theta; X)] = \mathsf{r} v,$$

the last inequality following by inspection of the loss. We then have

$$\begin{aligned} &f_{P_v}(\widehat{\theta}) - \inf_{\theta \in \Theta} f_{P_v}(\theta) \\ &= M \sum_{j=1}^{d} p_j \left[\frac{1 + \delta_j}{2} \left| \widehat{\theta}_j - \mathsf{r} v_j \right| - \frac{1 - \delta_j}{2} \left| \widehat{\theta}_j + \mathsf{r} v_j \right| \frac{1 + \delta_j}{2} \left| \theta_{v,j}^* - \mathsf{r} v_j \right| - \frac{1 - \delta_j}{2} \left| \theta_{v,j}^* + \mathsf{r} v_j \right| \right]. \end{aligned}$$

By inspecting the cases for the possible values of $\sign(\widehat{\theta}_j)$, we have

$$\frac{1+\delta_j}{2}\left|\widehat{\theta}_j - \mathsf{r}v_j\right| - \frac{1-\delta_j}{2}\left|\widehat{\theta}_j + \mathsf{r}v_j\right| + \frac{1+\delta_j}{2}\left|\theta^*_{v,j} - \mathsf{r}v_j\right| - \frac{1-\delta_j}{2}\left|\theta^*_{v,j} + \mathsf{r}v_j\right| \\ \geq \mathsf{r}\delta_j \mathbf{1}\left\{\sign(\widehat{\theta}_j) \neq v_j\right\}.$$

Taking expectations of this quantity gives the result (3.20).

Bounding the test error Recalling the stronger variant of Assouad's Lemma 2.2, we see that if we let $P^n_{\pm j} = 2^{1-d}\sum_{v:v_j=\pm 1} P^n_v$ be the mixture of several n-fold product distributions, then inequality (3.20) implies the bound

$$\max_{v\in\mathcal{V}} \sum_{j=1}^d p_j\delta_j P_v\left(\sign(\widehat{\theta}_j(X_1,\ldots,X_n)) \neq v_j\right) \geq \frac{1}{2}\sum_{j=1}^d p_j\delta_j \sum_{v\in\mathcal{V}}\left(1 - \left\|P^n_{+j} - P^n_{-j}\right\|_{\TV}\right).$$

Using that the total variation distance is convex, if we define $P_{v,j}$ to be the distrbution (3.19) with v_j constrained to be $+1$ (and similarly for $P_{v,-j}$), this bound implies the following (fairly weak) lower bound:

$$\max_{v\in\mathcal{V}} \sum_{j=1}^d p_j\delta_j P_v\left(\sign(\widehat{\theta}_j(X_{1:n})) \neq v_j\right) \geq \sum_{j=1}^d p_j\delta_j \frac{1}{2|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left(1 - \left\|P^n_{v,j} - P^n_{v,-j}\right\|_{\TV}\right). \quad (3.21)$$

Simple hypothesis tests For the majority of the remainder of the proof, we derive bounds on $\|P^n_{v,j} - P^n_{v,-j}\|_{\TV}$ to apply inequalities (3.20) and (3.21). Using Pinsker's inequality, we have

$$\left\|P^n_{v,j} - P^n_{v,-j}\right\|^2_{\TV} \leq \frac{1}{2}D_{\kl}\left(P^n_{v,j}\|P^n_{v,-j}\right) \leq \frac{n}{2}D_{\kl}\left(P_{v,j}\|P_{v,-j}\right).$$

Noting that P_v is a product distribution over the coordinates of the samples x (recall the construction (3.19)), we have the equality

$$D_{\kl}\left(P_{v,j}\|P_{v,-j}\right) = p_j\left[\frac{1+\delta_j}{2}\log\frac{1+\delta_j}{1-\delta_j} + \frac{1-\delta_j}{2}\log\frac{1-\delta_j}{1+\delta_j}\right] = p_j\left[\delta_j\log\frac{1+\delta_j}{1-\delta_j}\right].$$

Now we use the fact that $\delta\log\frac{1+\delta}{1-\delta} \leq 2\log(3)\delta^2$ for $\delta \leq 1/2$, so

$$\left\|P^n_{v,j} - P^n_{v,-j}\right\|^2_{\TV} \leq np_j\delta_j^2\log(3) \;\; \text{for } \delta_j \in [0,1/2]. \quad (3.22)$$

Combining inequalities (3.20), (3.21) and (3.22), using the fact that $\widehat{\theta}$ was an arbitrary estimator, we find the minimax lower bound

$$\frac{1}{M\mathsf{r}}\mathfrak{M}_n(\Theta,\mathcal{P},F) \geq \frac{1}{2}\sum_{j=1}^d p_j\delta_j\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left[1 - \delta_j\sqrt{np_j\log(3)}\right] \\ = \frac{1}{2}\sum_{j=1}^d p_j\delta_j\left[1 - \delta_j\sqrt{np_j\log(3)}\right]. \quad (3.23)$$

Inequality (3.23) holds for all $\delta_j \in [0, 1/2]$, so we may maximize over such δ_j. By setting

$$\delta_j = \min\left\{\frac{1}{2}, \frac{1}{2\sqrt{np_j \log(3)}}\right\},$$

we have

$$p_j\delta_j\left[1 - \delta_j\sqrt{np_j\log(3)}\right] \geq p_j \min\left\{\frac{1}{4}, \frac{1}{4\sqrt{\log 3}}\frac{1}{\sqrt{np_j}}\right\}.$$

In particular, our simplified Assouad analogue (3.23) implies

$$\sup_P \mathbb{E}_P[f_P(\widehat{\theta}) - \inf_{\theta\in\Theta} f_P(\theta)] \geq \frac{M\mathsf{r}}{8}\sum_{j=1}^{d} \min\left\{p_j, \frac{\sqrt{p_j}}{\sqrt{n\log 3}}\right\}$$

for any estimator $\widehat{\theta}$ based on the n observations X_k.

Chapter 4

Data sparsity, asynchrony, and faster stochastic optimization

In this chapter, we investigate a particular structure of optimization problems that allows faster solution of stochastic optimization problems of the form (3.1) outlined in the previous chapter. In particular, we study stochastic optimization problems when the *data* is sparse, which is in a sense dual to the current understanding of high-dimensional statistical learning and optimization. We highlight both the difficulties—in terms of increased sample complexity that sparse data necessitates, as demonstrated in the previous chapter—and the potential benefits, in terms of allowing parallelism and asynchrony in the design of algorithms. Leveraging sparsity allows us to develop (minimax optimal in terms of the number of gradient evaluations) parallel and asynchronous algorithms that enjoy a linear speedup in the amount of parallel computation available. We also provide experimental evidence complementing our theoretical results on several medium to large-scale learning tasks.

4.1 Problem setting

First, we recall that we wish to solve the following optimization problem:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; f(\theta) := \mathbb{E}[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x) dP(x), \tag{4.1}$$

where $\{F(\cdot; x)\}_{x \in \mathcal{X}}$ is a collection of real-valued convex functions, each of whose domains contains the closed convex set $\Theta \subset \mathbb{R}^d$. As before, we assume that we have access to a measurable (sub)gradient oracle

$$\mathsf{g} : \Theta \times \mathcal{X} \to \mathbb{R}^d \;\; \text{satisfying} \;\; \mathsf{g}(\theta; x) \in \partial F(\theta; x)$$

for each $\theta \in \Theta, x \in \mathcal{X}$.

In this chapter, we investigate the consequences of *data sparsity*, where the sampled data x is sparse. In the settings considered here, this means we assume the observations x are in

$\mathbb{R}^d$, and if we define the support $\operatorname{supp}(x)$ of a vector x to the set of indices of its non-zero components (and the support $\operatorname{supp}(C)$ of a set $C \subset \mathbb{R}^d$ to be the union $\cup_{x \in C} \operatorname{supp}(x)$), we assume that

$$\operatorname{supp} \mathsf{g}(\theta; x) \subset \operatorname{supp} \partial F(\theta; x) \subset \operatorname{supp} x. \tag{4.2}$$

The sparsity condition (4.2) means that $F(\theta; x)$ does not "depend" on the values of θ_j for indices j such that $x_j = 0$.[1] This type of data sparsity is prevalent in statistical optimization problems and machine learning applications, though in spite of its prevalence, study of such problems has been somewhat limited.

In this chapter, we investigate algorithms and their inherent limitations for solving problem (4.1) under natural conditions on the data generating distribution. Recent work in the optimization and machine learning communities has shown that data sparsity can be leveraged to develop parallel optimization algorithms [141, 149, 166], but the authors do not study the statistical effects of data sparsity, and there are no notions of optimality in their work. Moreover, each of the previous works requires the objective (4.1) to be smooth (have Lipschitz continuous gradient), which in some scenarios is a limitation. In the previous chapter, we showed how the AdaGrad algorithms of Duchi et al. [53] and McMahan and Streeter [130] adapt to data geometry to address problems in sparse data regimes, such as those satisfying (4.2), and have certain (theoretical) optimality guarantees. Whether they can leverage parallel computing, as in the papers [141, 166], has not been as clear.

To that end, in this chapter we study how sparsity may be leveraged in parallel computing frameworks to give substantially faster algorithms still achieving optimal sample complexity in terms of the number of observations x used. We develop two new algorithms, asynchronous dual averaging (AsyncDA) and asynchronous AdaGrad (AsyncAdaGrad), which allow asynchronous parallel solution of the problem (4.1) for general convex losses F and Θ. Combining insights of Niu et al.'s Hogwild! [141] with a new analysis, we prove our algorithms can achieve linear speedup in the number of processors while maintaining optimal statistical guarantees. We also give experiments on text-classification and web-advertising tasks to illustrate the benefits of the new algorithms.

4.2 Parallel and asynchronous optimization with sparsity

As we note in the previous section, recent work, for example, that by Niu et al. [141] and Takáč et al. [166], has suggested that sparsity can yield benefits in our ability to *parallelize* stochastic gradient-type algorithms. Given the optimality of AdaGrad-type algorithms, (recall Chapter 3.3) it is natural to focus on their parallelization in the hope that we can

[1] Formally, if we define π_x as the coordinate projection that zeros all indices j of its argument where $x_j = 0$, then $F(\pi_x(\theta); x) = F(\theta; x)$ for all θ, x. This is implied by standard first order conditions for convexity [98, Chapter VI.2]

leverage their ability to "adapt" to sparsity in the data. To provide the setting for our further algorithms, we first revisit Niu et al.'s HOGWILD!.

Niu et al.'s HOGWILD! algorithm [141] is an asynchronous (parallelized) stochastic gradient algorithm that proceeds as follows. To apply HOGWILD!, we must assume the domain Θ in problem (4.1) is a product space, that it decomposes as $\Theta = \Theta_1 \times \cdots \times \Theta_d$, where $\Theta_j \subset \mathbb{R}$. Fix a stepsize $\alpha > 0$. Then a pool of processors, each running independently, performs the following updates asynchronously to a centralized vector θ:

1. Sample $X \sim P$
2. Read θ and compute $g = \mathsf{g}(\theta; X) \in \partial_\theta F(\theta; X)$
3. For each j s.t. $g_j \neq 0$, update $\theta_j \leftarrow \Pi_{\Theta_j}(\theta_j - \alpha g_j)$

Here Π_{Θ_j} denotes projection onto the jth coordinate of the domain Θ, and we use $\leftarrow$ to denote the update as it may be done asynchronously—there is no true time index. The difficulty in HOGWILD! is that in step 2, the parameter θ at which g is calculated may be somewhat inconsistent—it may have received partial gradient updates from many processors—though for appropriate problems, this inconsistency is negligible. Indeed, Niu et al. [141] show a linear speedup in optimization time as the number of independent processors grow; they show this empirically in many scenarios, and they provide a proof under the somewhat restrictive assumptions that there is at most one non-zero entry in any gradient g, the risk f is strongly convex, and that f has Lipschitz continuous gradient.

4.2.1 Asynchronous dual averaging

One of the weaknesses of HOGWILD! is that, as written it appears to only be applicable to problems for which the domain Θ is a product space, and the known analysis assumes that $\|g\|_0 = 1$ for all gradients g. In effort to alleviate these difficulties, we now develop and present our asynchronous dual averaging algorithm (recall the update (3.8)), ASYNCDA. In ASYNCDA, instead of asynchronously updating a centralized parameter vector θ, we maintain a centralized dual vector z. A pool of processors performs asynchronous additive updates to z, where each processor repeatedly and independently performs the following updates:

1. Read z and compute $\theta := \operatorname{argmin}_{\theta \in \Theta} \left\{ \langle z, \theta \rangle + \frac{1}{\alpha} \psi(\theta) \right\}$ *// Implicitly increment "time" counter k and let $\theta^k = \theta$*
2. Sample $X \sim P$ and let $g = \mathsf{g}(\theta; X) \in \partial_\theta F(\theta; X)$ *// Let $g^k = g$.*
3. For $j \in [d]$ such that $g_j \neq 0$, update $z_j \leftarrow z_j + g_j$

The actual computation of the vector θ in asynchronous dual averaging (ASYNCDA) is performed locally on each processor in step 1 of the algorithm, so the algorithm can be *executed* with any proximal function ψ and domain Θ. The only communication point

between any of the processors is the addition operation in step 3. As noted by Niu et al. [141], this operation can often be performed atomically on modern processors.

In our analysis of AsyncDA, and in our subsequent analysis of the adaptive methods, we require a measurement of time elapsed. With that in mind, we let k denote an implicitly existing time index, so that θ^k denotes the vector $\theta \in \Theta$ computed in the "kth step" 1 of the AsyncDA algorithm, that is, whichever is the kth θ actually computed by any of the processors. We note that this quantity exists and is recoverable from the algorithm, and it is also possible to track the running sum $\sum_{i=1}^{k} \theta^i$.

Additionally, we require two assumptions that underly our analysis.

Assumption 4A. *There is an upper bound m on the delay of any processor. In addition, for each $j \in [d]$ there is a constant $p_j \in [0,1]$ such that $P(X_j \neq 0) \leq p_j$.*

We also require an assumption about the continuity (Lipschitzian) properties of the loss functions being minimized; the assumption amounts to a second moment constraint on the sub-gradients of the instantaneous F along with a rough measure of the sparsity of the gradients.

Assumption 4B. *There exist constants M and $(M_j)_{j=1}^d$ such that the following bounds hold for all $\theta \in \Theta$: $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \leq \mathsf{M}^2$, and for each $j \in [d]$ we have $\mathbb{E}[|\mathsf{g}_j(\theta; X)|] \leq p_j M_j$.*

With these definitions, we have the following theorem, which captures the convergence behavior of AsyncDA under the assumption that Θ is a Cartesian product, meaning that $\Theta = \Theta_1 \times \cdots \times \Theta_d$, where $\Theta_j \subset \mathbb{R}$, and that $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$. Note the algorithm itself can still be efficiently parallelized for more general convex Θ, even if the theorem does not apply. In the theorem, we superscript the observations X^k, as the coordinates X_j^k are important throughout the analyses.

Theorem 4.1. *Let Assumptions 4A and 4B and the conditions in the preceding paragraph hold. Then for any $\theta^* \in \Theta$,*

$$\mathbb{E}\left[\sum_{k=1}^{n} F(\theta^k; X^k) - F(\theta^*; X^k)\right] \leq \frac{1}{2\alpha}\|\theta^*\|_2^2 + \frac{\alpha}{2} T \mathsf{M}^2 + \alpha T m \sum_{j=1}^{d} p_j^2 M_j^2.$$

We provide the proof of Theorem 4.1 in Section 4.4.1.

As stated, the theorem is somewhat unwieldy, so we provide a corollary and a few remarks to explain and simplify the result. Under a more stringent condition that $|\mathsf{g}_j(\theta; x)| \leq M_j$, Assumption 4A implies $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \leq \sum_{j=1}^{d} p_j M_j^2$. Thus, without loss of generality for the remainder of this section we take $\mathsf{M}^2 = \sum_{j=1}^{d} p_j M_j^2$, which serves as an upper bound on the Lipschitz continuity constant of the objective function f. We then obtain the following corollary.

Corollary 4.1. *Define* $\widehat{\theta}(n) = \frac{1}{n}\sum_k^n \theta^k$ *and set* $\alpha = \|\theta^*\|_2 / \mathsf{M}\sqrt{n}$. *Then*

$$\mathbb{E}[f(\widehat{\theta}(n)) - f(\theta^*)] \le \frac{\mathsf{M}\|\theta^*\|_2}{\sqrt{n}} + m\frac{\|\theta^*\|_2}{2\mathsf{M}\sqrt{n}}\sum_{j=1}^d p_j^2 M_j^2$$

Corollary 4.1 is almost immediate. To see the result, note that since X^k is independent of θ^k, we have $\mathbb{E}[F(\theta^k; X^k) \mid \theta^k] = f(\theta^k)$; applying Jensen's inequality to $f(\widehat{\theta})$ and performing an algebraic manipulation give the corollary.

If the data is suitably "sparse," meaning that $p_j \le 1/m$ (which may also occur if the data is of relatively high variance in Assumption 4B) the bound in Corollary 4.1 simplifies to

$$\mathbb{E}[f(\widehat{\theta})(n) - f(\theta^*)] \le \frac{3}{2}\frac{\mathsf{M}\|\theta^*\|_2}{\sqrt{n}} = \frac{3}{2}\frac{\sqrt{\sum_{j=1}^d p_j M_j^2}\,\|\theta^*\|_2}{\sqrt{n}} \tag{4.3}$$

which is the convergence rate of stochastic gradient descent (and dual averaging) even in non-asynchronous situations (3.4). (More generally, if $\sum_{j=1}^d p_j^2 M_j^2 \le \frac{1}{m}\mathsf{M}^2$, we obtain the same inequality (4.3)). In non-sparse cases, setting $\alpha \propto \|\theta^*\|_2 / \sqrt{m\mathsf{M}^2 T}$ in Theorem 4.1 recovers the bound

$$\mathbb{E}[f(\widehat{\theta}(n)) - f(\theta^*)] \le \mathcal{O}(1)\sqrt{m} \cdot \frac{\mathsf{M}\|\theta^*\|_2}{\sqrt{n}}.$$

The convergence guarantee (4.3) shows that after n gradient updates, we have error scaling $1/\sqrt{n}$; however, if we have N processors, then updates can occur roughly N times as quickly, as all updates are asynchronous. Thus, in time scaling as n/N, we can evaluate n gradients: a *linear* speedup.

4.2.2 Asynchronous AdaGrad

We now turn to extending AdaGrad to asynchronous settings, developing AsyncAdaGrad (asynchronous AdaGrad). As in the AsyncDA algorithm, AsyncAdaGrad maintains a shared dual vector z among the processors, which is the sum of gradients observed; AsyncAdaGrad also maintains the matrix S, which is the diagonal sum of squares of gradient entries (recall Section 3.2.2). The matrix S is initialized as $\mathrm{diag}(\delta^2)$, where $\delta_j \ge 0$ is an initial value. Each processor asynchronously performs the following iterations:

1. Read S and z and set $G = S^{\frac{1}{2}}$. Compute $\theta := \mathrm{argmin}_{\theta\in\Theta}\left\{\langle z, \theta\rangle + \frac{1}{2\alpha}\langle \theta, G\theta\rangle\right\}$
   // Implicitly increment "time" counter k and let $\theta^k = \theta$, $S^k = S$
2. Sample $X \sim P$ and let $g = \mathsf{g}(\theta; X) \in \partial F(\theta; X)$
3. For $j \in [d]$ such that $g_j \ne 0$, update $S_j \leftarrow S_j + g_j^2$ and $z_j \leftarrow z_j + g_j$

As in the description of AsyncDA, we note that θ^k is the vector $\theta \in \Theta$ computed in the kth "step" of the algorithm (step 1), and similarly associate X^k with θ^k.

To analyze AsyncAdaGrad, we make a somewhat stronger assumption on the sparsity properties of the instantaneous losses F than Assumption 4B.

Assumption 4C. *There exist constants $(M_j)_{j=1}^d$ such that for any $\theta \in \Theta$ and $j \in [d]$, we have $\mathbb{E}[\mathsf{g}_j(\theta; X)^2 \mid X_j \neq 0] \le M_j^2$.*

Taking $\mathsf{M}^2 = \sum_j p_j M_j^2$ shows that Assumption 4C implies Assumption 4B with specific constants. We then have the following convergence result, whose proof we provide in Section 4.4.2.

Theorem 4.2. *In addition to the conditions of Theorem 4.1, let Assumption 4C hold. Assume that $\delta^2 \ge M_j^2 m$ for all j and that $\Theta \subset [-\mathsf{r}_\infty, \mathsf{r}_\infty]^d$. Then*

$$\sum_{k=1}^n \mathbb{E}\left[F(\theta^k; X^k) - F(\theta^*; X^k)\right] \le \sum_{j=1}^d \min\left\{\frac{1}{\alpha}\mathsf{r}_\infty^2 \mathbb{E}\left[\left(\delta^2 + \sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}}\right] + \alpha \mathbb{E}\left[\left(\sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}}\right](1 + p_j m), M_j \mathsf{r}_\infty p_j n\right\}.$$

At the expense of some additional notational overhead, we can also relax the condition $\delta^2 \ge M_j^2 m$ on the initial constant diagonal term δ slightly. This gives a qualitatively similar result while allowing us to only require (roughly) that the initial matrix value δ^2 be large enough to overwhelm $p_j m$ updates rather than m of them. (See Section 4.4.5 for a proof.)

Corollary 4.2. *Under the conditions of Theorem 4.2, assume additionally that for all j we have $\delta^2 \ge M_j^2 \min\{m, 6\max\{\log T, m p_j\}\}$. Then*

$$\sum_{k=1}^n \mathbb{E}\left[F(\theta^k; X^k) - F(\theta^*; X^k)\right] \le \sum_{j=1}^d \min\left\{\frac{1}{\alpha}\mathsf{r}_\infty^2 \mathbb{E}\left[\left(\delta^2 + \sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}}\right] + \frac{3}{2}\alpha \mathbb{E}\left[\sum_{k=1}^n (g_j^k)^2\right]^{\frac{1}{2}}(1 + p_j m), M_j \mathsf{r}_\infty p_j n\right\}.$$

It is natural to ask in which situations the bound Theorem 4.2 and Corollary 4.2 provides is optimal. We note that, as in the case with Theorem 4.1, we may take an expectation with respect to X^k and obtain a convergence rate for $f(\widehat{\theta}(n)) - f(\theta^*)$, where $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$. By Jensen's inequality, we have for any δ that

$$\mathbb{E}\left[\left(\delta^2 + \sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}}\right] \le \left(\delta^2 + \sum_{k=1}^n \mathbb{E}[(g_j^k)^2]\right)^{\frac{1}{2}} \le \sqrt{\delta^2 + n p_j M_j^2}.$$

For interpretation, let us now make a few assumptions on the probabilities p_j. If we assume that $p_j \le c/m$ for a universal (numerical) constant c, then Theorem 4.2 guarantees that

$$\mathbb{E}[f(\widehat{\theta}(n)) - f(\theta^*)] \le \mathcal{O}(1)\left[\frac{1}{\alpha}\mathsf{r}_\infty^2 + \alpha\right]\sum_{j=1}^d M_j \min\left\{\frac{\sqrt{n^{-1}\log n + p_j}}{\sqrt{n}}, p_j\right\}, \tag{4.4}$$

which is the convergence rate of AdaGrad except for a small factor of $\min\{\sqrt{\log n}/n, p_j\}$ in addition to the usual $\sqrt{p_j/n}$ rate. In particular, optimizing by choosing $\alpha = \mathsf{r}_\infty$, and assuming $p_j \gtrsim \frac{1}{n}\log n$, we have convergence guarantee

$$\mathbb{E}[f(\widehat{\theta}(n)) - f(\theta^*)] \le \mathcal{O}(1)\mathsf{r}_\infty \sum_{j=1}^{d} M_j \min\left\{\frac{\sqrt{p_j}}{\sqrt{n}}, p_j\right\},$$

which is minimax-optimal by Proposition 3.4.

In fact, however, the bounds of Theorem 4.2 and Corollary 4.2 are somewhat stronger: they provide bounds using the *expectation* of the squared gradients g_j^k rather than the maximal value M_j, though the bounds are perhaps clearer in the form (4.4). We note also that our analysis applies to more adversarial settings than stochastic optimization (e.g. to online convex optimization [95]). Specifically, an adversary may choose an arbitrary sequence of functions subject to the data sparsity constraint (4.2) and feature appearance constraints in Assumptions 4A–4C, and our results provide an expected regret bound, which is strictly stronger than the stochastic convergence guarantees provided (and guarantees high-probability convergence in stochastic settings [40]). Moreover, our comments in Chapter 3.3 about the relative optimality of AdaGrad versus standard gradient methods apply. When the data is sparse, we indeed should use asynchronous algorithms, but using adaptive methods yields even more improvement than simple gradient-based methods.

4.3 Experiments

In this section, we give experimental validation of our theoretical results on AsyncAdaGrad and AsyncDA, giving results on two datasets selected for their high-dimensional sparsity.[2]

4.3.1 Malicious URL detection

For our first set of experiments, we consider the speedup attainable by applying AsyncAdaGrad and AsyncDA, investigating the performance of each algorithm on a malicious URL prediction task [124]. The dataset in this case consists of an anonymized collection of URLs labeled as malicious (e.g. spam, phishing, etc.) or benign over a span of 120 days. The data in this case consists of $2.4 \cdot 10^6$ examples with dimension $d = 3.2 \cdot 10^6$ (sparse) features. We perform several experiments, randomly dividing the dataset into $1.2 \cdot 10^6$ training and test samples for each experiment.

In Figure 4.1 and we compare the performance of AsyncAdaGrad and AsyncDA after doing after single pass through the training dataset. (For each algorithm, we choose the stepsize α for optimal training set performance.) We perform the experiments on a single

[2]We also performed experiments using Hogwild! instead of AsyncDA; the results are similar.

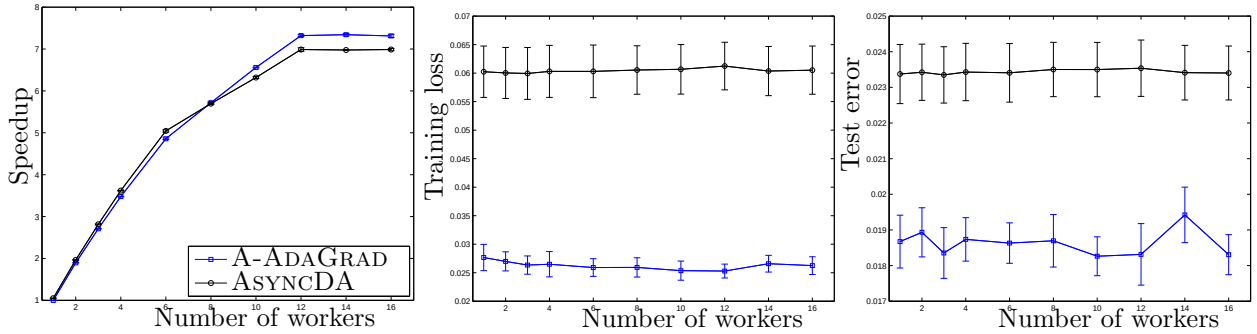

Figure 4.1. Experiments with URL data. Left: speedup relative to 1 processor. Middle: training dataset loss versus number of processors. Right: test set error rate versus number of processors. A-AdaGrad abbreviates AsyncAdaGrad.

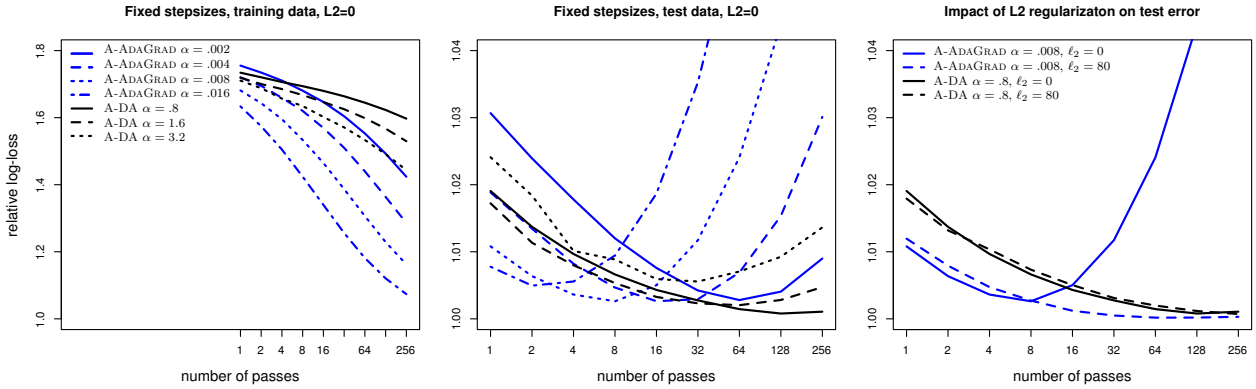

Figure 4.2. Relative accuracy for various stepsize choices on an click-through rate prediction dataset. A-AdaGrad abbreviates AsyncAdaGrad and A-DA abbreviates AsyncDA.

machine running Ubuntu Linux with 6 cores (with two-way hyperthreading) and 32Gb of RAM. From the left-most plot in Fig. 4.1, we see that up to 6 processors, both AsyncDA and AsyncAdaGrad enjoy the expected linear speedup, and from 6 to 12, they continue to enjoy a speedup that is linear in the number of processors though at a lesser slope (this is the effect of hyperthreading). For more than 12 processors, there is no further benefit to parallelism on this machine.

The two right plots in Figure 4.1 plot performance of the different methods (with standard errors) versus the number of worker threads used. Both are essentially flat; increasing the amount of parallelism does nothing to the average training loss or the test error rate for either method. It is clear, however, that for this dataset, the adaptive AsyncAdaGrad algorithm provides substantial performance benefits over AsyncDA.

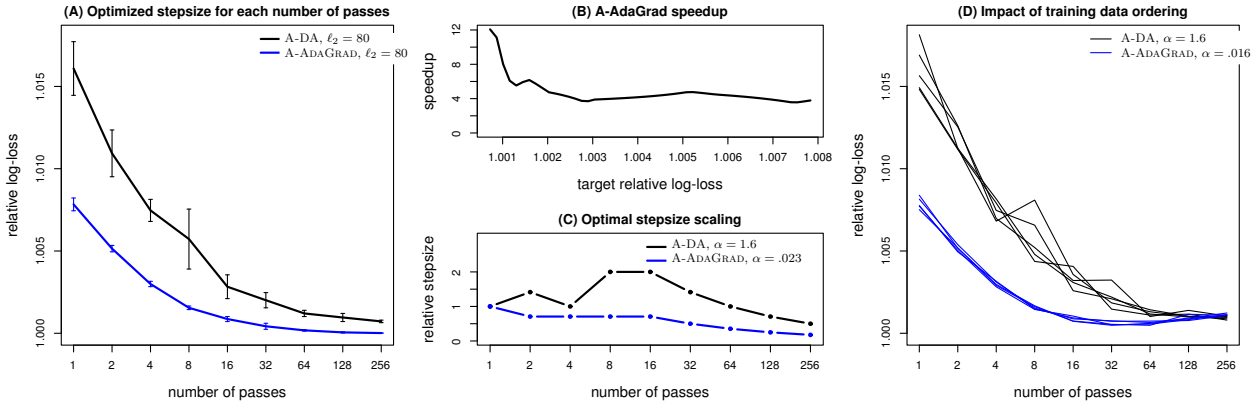

Figure 4.3. (A) Relative test-set log-loss for AsyncDA and AsyncAdaGrad, choosing the best stepsize (within a factor of about 1.4×) individually for each number of passes. (B) Effective speedup for AsyncAdaGrad. (C) The best stepsize η, expressed as a scaling factor on the stepsize used for one pass. (D) Five runs with different random seeds for each algorithm (with $L_2 = 80$).

4.3.2 Click-through-rate prediction experiments

We also experimented on a proprietary datasets consisting of search ad impressions. Each example corresponds to showing a search-engine user a particular text ad in response to a query string. From this, we construct a very sparse feature vector based on the text of the ad displayed and the query string (no user-specific data was used). The target label is 1 if the user clicked the ad, and -1 otherwise. We fit logistic regression models using both AsyncDA and AsyncAdaGrad. Rather than running few experiments on a large dataset, we ran extensive experiments on a moderate-sized dataset (about 10^7 examples, split evenly between training and testing). This allowed us to thoroughly investigate the impact of the stepsize η, the number of training passes,[3] and L_2 regularization on accuracy. Section 4.3.1 shows that AsyncAdaGrad achieves a similar speedup to AsyncDA, so for these experiments we used 32 threads on 16 core machines for each run.

On this dataset, AsyncAdaGrad typically achieves an effective *additional* speedup over AsyncDA of 4× or more. That is, to reach a given level of accuracy, AsyncDA generally needs four times as many effective passes over the dataset. We measure accuracy with log-loss (the logistic loss) averaged over 5 runs using different random seeds (which control the order in which the algorithms sample examples during training). We report relative values in Figures 4.2 and 4.3, that is, the ratio of the mean loss for the given datapoint to the lowest (best) mean loss obtained. Our results are not particularly sensitive to the choice of relative

[3]Here "number of passes" more precisely means the expected number of times each example in the dataset is trained on. That is, each worker thread randomly selects a training example from the dataset for each update, and we continued making updates until (dataset size) × (number of passes) updates have been processed.

log-loss as the metric of interest; we also considered AUC (the area under the ROC curve) and observed similar results.

Figure 4.2 (A–B) shows relative log-loss as a function of the number of training passes for various stepsizes. Without regularization, we see that AsyncAdaGrad is prone to overfitting: it achieves significantly higher accuracy on the training data 4.2(A), but unless the step-size is tuned carefully to the number of passes, it will overfit and predict poorly on test data 4.2(B). Fortunately, the addition of L_2 regularization largely solves this problem. Figure 4.2(C) shows that adding an L_2 penalty of 80 has very little impact on Hogwild!, but effectively prevents the overfitting of AsyncAdaGrad.[4]

Fixing $L_2 = 80$, for each number of passes and for each algorithm, we varied the stepsize α over a multiplicative grid with resolution $\sqrt{2}$. Figure 4.3 reports the results obtained by selecting the best stepsize in terms of test set log-loss for each number of passes. Figure 4.3(A) shows relative log-loss of the best stepsize for each algorithm; 4.3(B) is based on the same data, but considers on the x-axis relative losses between the 256-pass AsyncDA loss (about 1.001) and the 1-pass AsyncAdaGrad loss (about 1.008). For these values, we can take the linear interpolation shown in 4.3(A), and look at the ratio of the number of passes the two algorithms needed to achieve a fixed relative log-loss. This gives an estimate of the relative speedup obtained by using AsyncAdaGrad over a range of different target accuracies; speedups range from 3.6× to 12×. Figure 4.3(C) shows the optimal stepsizes as a function of the best setting for one pass. The optimal stepsize decreases moderately for AsyncAdaGrad, but are somewhat noisy for Hogwild!.

It is interesting to note that AsyncAdaGrad's accuracy is largely independent of the ordering of the training data, while Hogwild! shows significant variability. This can be seen both in the error bars on Figure 4.3(A), and explicitly in Figure 4.3(D), where we plot one line for each of the 5 random seeds used. Thus, while on the one hand Hogwild! requires somewhat less tuning of the stepsize and L_2 parameter to control overfitting, tuning AsyncAdaGrad is much easier because of the predictable response.

4.4 Proofs of convergence

4.4.1 Proof of Theorem 4.1

Our proof begins by recalling Lemma 3.1, which has immediate implications in the context of Theorem 4.1. Since $\psi_k(\theta) = \frac{1}{2\alpha}\|\theta\|_2^2$ in this case, Lemma 3.1 immediately implies

$$\sum_{k=1}^{n}\left[F(\theta^k;X^k) - F(\theta^*;X^k)\right] \le \frac{1}{2\alpha}\|\theta^*\|_2^2 + \frac{\alpha}{2}\sum_{k=1}^{n}\left\|g^k\right\|_2^2 + \sum_{k=1}^{n}\left\langle g^k, \theta^k - \widetilde{\theta}^k\right\rangle \tag{4.5}$$

[4]For both algorithms, this is accomplished by adding the term $\eta 80\|x\|_2^2$ to the ψ function. We could have achieved slightly better results for AsyncAdaGrad by varying the L_2 penalty with the number of passes (with more passes benefiting from more regularization).

as $\psi_k^* = \psi_{k-1}^*$ and $\psi^*(0) \le 0$, and for any v

$$\|v\|_\psi^2 = \frac{1}{\alpha}\|v\|_2^2 \quad \text{and} \quad \|v\|_{\psi^*}^2 = \alpha\|v\|_2^2.$$

Now we return to the proof of Theorem 4.1. Each of the terms present in Theorem 4.1 is present in Eq. (4.5) except for the last, because

$$\mathbb{E}[\left\|g^k\right\|_2^2] = \mathbb{E}[\left\|\mathsf{g}(\theta^k; X_k)\right\|_2^2] \le \mathsf{M}^2.$$

For the final term in the bound in the theorem, we note that by assumption that Θ is a product domain,

$$\left\langle g^k, \theta^k - \widetilde{\theta}^k \right\rangle \le \sum_{j=1}^d |g_j^k||\theta_j^k - \widetilde{\theta}_j^k| \le \sum_{j=1}^d \alpha|g_j^k| \left| \sum_{i=1}^{k-1} g_j^i - z_j^k \right|.$$

For the final inequality we have used that by the definition of the θ update (recall Lemma 3.4),

$$|\widetilde{\theta}_j^k - \theta_j^k| = \left| \nabla_j \psi^* \Big(-\sum_{i=1}^{k-1} g^i \Big) - \nabla_j \psi^*(-z^k) \right| \le \alpha \left| \sum_{i=1}^{k-1} g_j^i - z_j^i \right|.$$

Conditioned on the σ-field $\mathcal{F}_{k-1}$ of $\{X^i\}_{i=1}^{k-1}$, we have $\mathbb{E}[|g_j^k| \mid \mathcal{F}_{k-1}] \le p_j M_j$ by assumption (since X^k is independent of X^i for $i < t$). Moreover, we have $\mathbb{E}[|\sum_{i=1}^{k-1} g_j^i - z_j^k|] \le m p_j M_j$ because the delay in each processor is assumed to be at most m and $\mathbb{E}[|g_j^i|] \le p_j M_j$. Thus we find

$$\mathbb{E}[\left\langle g^k, \theta^k - \widetilde{\theta}^k \right\rangle] \le \alpha \sum_{j=1}^d \mathbb{E}\left[\mathbb{E}[|g_j^k| \mid \mathcal{F}_{k-1}] \left| \sum_{i=1}^{k-1} g_j^i - z_j^k \right| \right] \le \alpha \sum_{j=1}^d p_j^2 M_j^2 m.$$

This completes the proof.

4.4.2 Proof of Theorem 4.2

Before beginning, we establish a bit of notation. Throughout this proof, as the coordinates x_j of the vectors $x \in \mathcal{X}$ will be important, we index the individual observations by superscript, so X^k is the kth data point. We recall the definitions of z^k and S^k to be the values *read* in the computation of step 1 of the algorithm to construct the vector θ^k in the definition of AsyncAdaGrad. In addition, we define the two temporal inequalities $\prec_{S_j}$ and $\prec_{z_j}$ to capture the order in which the updates are applied in the AsyncAdaGrad algorithm. We say that $i \prec_{S_j} k$ if the gradient term $(g_j^i)^2$ has been incorporated into the matrix coordinate S_j at the instant S_j is read in step (1) of the AsyncAdaGrad algorithm to compute θ^k, and similarly, we say $i \prec_{z_j} k$ if the gradient term g_j^i has been incorporated in the dual vector coordinate z_j.

The proof of this theorem follows from the general bound of Lemma 3.1 applied with a particular choice of the proximal functions ψ_k. Before actually applying the general bound of Lemma 3.1, we note that by convexity,

$$\sum_{k=1}^{n} \left[F(\theta^k; x^k) - F(\theta^*; x^k)\right] \le \sum_{k=1}^{n} \left\langle g^k, \theta^k - \theta^* \right\rangle.$$

Considering a particular coordinate j, we have

$$\sum_{k=1}^{n} \mathbb{E}\left[g_j^k(\theta_j^k - \theta_j^*)\right] \le \mathsf{r}_\infty \sum_{k=1}^{n} \mathbb{E}[|g_j^k|] \le \mathsf{r}_\infty T M_j p_j, \tag{4.6}$$

where we have used the compactness assumption on Θ. The remainder of our proof bounds the regret-like term $\sum_{k=1}^{n} \left\langle g^k, \theta^k - \theta^* \right\rangle$ in a per-coordinate way, and thus for each coordinate we always have the bound (4.6), giving the $\min\{\cdot, p_j\}$ terms in the theorem statement. It remains to show the bound that applies when p_j is large.

We now re-state the general bound of Lemma 3.1 with some minor modifications in notation. AsyncAdaGrad is dual averaging with the choice $\psi_k(\theta) := \frac{1}{2\alpha}\left\langle \theta, G^k \theta \right\rangle$ for the proximal function. With this choice, the norm and dual norm $\|\cdot\|_{\psi_k}$ and $\|\cdot\|_{\psi_k^*}$ defined for vectors $v \in \mathbb{R}^d$ are

$$\|v\|_{\psi_k}^2 := \frac{1}{\alpha}\|v\|_{G^k}^2 \quad \text{and} \quad \|v\|_{\psi_k^*}^2 := \alpha \|v\|_{(G^k)^{-1}}^2.$$

Rewriting Lemma 3.1, we thus have

$$\begin{aligned}\sum_{k=1}^{n} \left[F(\theta^k; x^k) - F(\theta^*; x^k)\right] \le \sum_{k=1}^{n} \left[\psi_k^*\Big(-\sum_{i=1}^{k-1} g^i\Big) - \psi_{k-1}^*\Big(-\sum_{i=1}^{k-1} g^i\Big)\right] + \frac{\alpha}{2}\sum_{k=1}^{n} \left\|g^k\right\|_{(G^k)^{-1}}^2 \\ + \sum_{k=1}^{n} \left\langle g^k, \theta^k - \widetilde{\theta}^k \right\rangle + \frac{1}{2\alpha}\|\theta^*\|_{G^k}^2 \end{aligned} \tag{4.7}$$

for any sequence x^k, where as in Lemma 3.1 we define the "corrected" sequences $\widetilde{\theta}^k = \nabla\psi_k^*(-g^{1:k-1})$ where $g^{1:k} = \sum_{i=1}^{k} g^i$. Note that the corrected sequences still use the proximal functions ψ_k^* from the *actual* run of the algorithm.

We focus on bounding each of the terms in the sums (4.7) in turn, beginning with the summed conjugate differences.

Lemma 4.1. *Define the matrix G^{n+} to be diagonal with jth diagonal entry $(\delta^2 + \sum_{k=1}^{n}(g_j^k)^2)^{\frac{1}{2}}$. For any sequence x^k*

$$\sum_{k=1}^{n} \left[\psi_k^*\Big(-\sum_{i=1}^{k-1} g^i\Big) - \psi_{k-1}^*\Big(-\sum_{i=1}^{k-1} g^i\Big)\right] \le \frac{\mathsf{r}_\infty^2}{2\alpha}\operatorname{tr}(G^{n+}).$$

We defer the proof of Lemma 4.1 to Section 4.4.3, noting that the proof follows by carefully considering the conditions under which $\psi_k^* \geq \psi_{k-1}^*$, which may only occur when updates to S (and hence G) are out of order, a rearrangement of the sum to put the updates to S in the correct order, and an application of the AdaGrad Lemma 3.2.

To complete the proof of the theorem, we must bound the two summed gradient quantities in expression (4.7). For shorthand, let us define

$$\mathcal{T}_1 := \sum_{k=1}^{n} \left\|g^k\right\|_{(G^k)^{-1}}^2 \quad \text{and} \quad \mathcal{T}_2 := \sum_{k=1}^{n} \left\langle g^k, \theta^k - \widetilde{\theta}^k \right\rangle \tag{4.8}$$

We provide the proof under the assumption that $\delta^2 \geq mM_j^2$ for all j. At the end of the proof, we show how to weaken this assumption while retaining the main conclusions of the theorem.

Recalling the temporal ordering notation $\prec_{S_j}$, we see that

$$\mathcal{T}_1 = \sum_{j=1}^{d} \sum_{k=1}^{n} \frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}}.$$

Now, by our assumption that processors are *at most* m steps out of date and $\delta^2 \geq mM_j^2$, we have

$$\frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \leq \frac{(g_j^k)^2}{\sqrt{\sum_{i=1}^{k} (g_j^i)^2}},$$

and thus the standard AdaGrad summation result (Lemma 3.2) implies

$$\mathcal{T}_1 \leq \sum_{j=1}^{d} \sum_{k=1}^{n} \frac{(g_j^k)^2}{\sqrt{\sum_{i=1}^{k} (g_j^i)^2}} \leq \sum_{j=1}^{d} 2 \left(\sum_{k=1}^{n} (g_j^k)^2 \right)^{\frac{1}{2}}. \tag{4.9}$$

Thus we turn to $\mathcal{T}_2$ as defined in expression (4.8). We focus on a per-coordinate version of $\mathcal{T}_2$, stating the following lemma, whose technical proof we defer to Section 4.4.4:

Lemma 4.2. *Under the conditions of Theorem 4.2,*

$$\frac{1}{\alpha} \sum_{k=1}^{n} \mathbb{E}[g_j^k (\theta_j^k - \widetilde{\theta}_j^k)] \leq 2 p_j m \mathbb{E} \left[\left(\sum_{k=1}^{n} (g_j^k)^2 \right)^{\frac{1}{2}} \right] \leq 2 p_j m M_j \sqrt{p_j n}.$$

Applying the result of Lemma 4.2, we obtain the following bound on $\mathcal{T}_2$:

$$\mathbb{E}[\mathcal{T}_2] \leq 2\alpha \sum_{j=1}^{d} p_j m \mathbb{E} \left[\left(\sum_{k=1}^{n} (g_j^k)^2 \right)^{\frac{1}{2}} \right].$$

Combining Lemma 4.1 with our bounds (4.9) on $\mathcal{T}_1$ and the preceding bound on $\mathcal{T}_2$, the basic inequality (4.7) implies

$$\mathbb{E}\Big[\sum_{k=1}^{n}\big(F(\theta^k;X^k)-F(\theta^*;X^k)\big)\Big]$$
$$\leq \frac{1}{2\alpha}\mathbb{E}\left[\|\theta^*\|_{G^k}^2+\mathsf{r}_\infty^2\operatorname{tr}(G^{n_+})\right]+\alpha\sum_{j=1}^{d}\mathbb{E}\left[\Big(\sum_{k=1}^{n}(g_j^k)^2\Big)^{\frac{1}{2}}\right](1+2p_jm).$$

Noting that

$$\|\theta^*\|_{G^k}^2\leq\mathsf{r}_\infty^2\operatorname{tr}(G^k)\leq\mathsf{r}_\infty^2\sum_{j=1}^{d}\left(\delta^2+\sum_{k=1}^{n}(g_j^k)^2\right)^{\frac{1}{2}}$$

completes the proof of Theorem 4.2 under the assumption that $\delta^2\geq mM_j^2$ for all j. □

4.4.3 Proof of Lemma 4.1

Since the domain $\Theta=\Theta_1\times\ldots\times\Theta_d$ is assumed Cartesian and the matrices S and $G=S^{\frac{1}{2}}$ are diagonal, we focus on the individual coordinate terms of ψ_k^*. With that in mind, consider the difference

$$\sup_{\theta_j\in\Theta_j}\left\{-\sum_{i=1}^{k-1}g_j^i\theta_j-\frac{1}{2\alpha}\theta_jG_j^k\theta_j\right\}-\sup_{\theta_j\in\Theta_j}\left\{-\sum_{i=1}^{k-1}g_j^i\theta_j-\frac{1}{2\alpha}\theta_jG_j^{k-1}\theta_j\right\}. \tag{4.10}$$

To understand the difference of the terms (4.10), we recall the temporal ordering $\prec_{S_j}$ defined in the beginning of the proof of Theorem 4.2 (we say $i\succeq_{S_j}k$ if and only if $i\not\prec_{S_j}k$). Though throughout the algorithm, the matrix S (and the matrix G) is always increasing—only positive updates are applied to S—when indexed by update time, we may have $G_j^{k-1}\leqq G_j^k$. The term (4.10), however, may be positive only when $G_j^k<G_j^{k-1}$, and this is possible only if

$$\left\{i\in\mathbb{N}\mid i\prec_{S_j}k-1\right\}\supsetneq\left\{i\in\mathbb{N}\mid i\prec_{S_j}k\right\}.$$

Finally, we note that for any matrices A,B and vector z, that if we define

$$\theta(A):=\operatorname*{argmax}_{\theta\in\Theta}\left\{\langle z,\theta\rangle-\langle\theta,A\theta\rangle\right\}$$

then

$$\sup_{\theta\in\Theta}\left\{\langle z,\theta\rangle-\langle\theta,A\theta\rangle\right\}-\sup_{\theta\in\Theta}\left\{\langle z,\theta\rangle-\langle\theta,B\theta\rangle\right\}$$
$$\leq\langle z,\theta(A)\rangle-\langle\theta(A),A\theta(A)\rangle-\langle z,\theta(A)\rangle+\langle\theta(A),B\theta(A)\rangle\leq\sup_{\theta\in\Theta}\left\{\langle\theta,(B-A)\theta\rangle\right\}.$$

By considering expression (4.10), we have

$$\psi_k^*\Big(-\sum_{i=1}^{k-1} g^i\Big) - \psi_{k-1}^*\Big(-\sum_{i=1}^{k-1} g^i\Big) \le \frac{1}{2\alpha}\sum_{j=1}^{d} \sup_{\theta_j \in \Theta_j} \left\{\theta_j^2 (G_j^{k-1} - G_j^k)\right\} \tag{4.11}$$
$$\le \frac{\mathsf{r}_\infty^2}{2\alpha}\sum_{j=1}^{d} \left|G_j^k - G_j^{k-1}\right| \mathbf{1}\left\{\{i \prec_{S_j} k-1\} \supsetneq \{i \prec_{S_j} k\}\right\}.$$

It thus remains to bound the sum, over all k, of the terms (4.11). To that end, we note by concavity of $\sqrt{\cdot}$ that for any $a, b \ge 0$, we have $\sqrt{a+b} - \sqrt{a} \le b/2\sqrt{a}$. Thus we find that

$$\begin{aligned}
&\left|G_j^k - G_j^{k-1}\right| \mathbf{1}\left\{\{i \prec_{S_j} k-1\} \supsetneq \{i \prec_{S_j} k\}\right\} \\
&= \left|\left(\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2\right)^{\frac{1}{2}} - \left(\delta^2 + \sum_{i \prec_{S_j} k-1} (g_j^i)^2\right)^{\frac{1}{2}}\right| \mathbf{1}\left\{\{i \prec_{S_j} k-1\} \supsetneq \{i \prec_{S_j} k\}\right\} \\
&\le \frac{\sum_{i \prec_{S_j} k-1, i \succeq_{S_j} k} (g_j^i)^2}{2\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}}.
\end{aligned}$$

We note the following: the sequence of update sets $\Delta_k := \{i \in \mathbb{N} : i \prec_{S_j} k-1, i \succeq_{S_j} k\}$ satisfies $\cup_{k=1}^n \Delta_k \subset [n]$, and since the incremental updates to S occur only once, we have $\Delta_k \cap \Delta_{k'} = \emptyset$ for all $k \neq k'$. That is, if $i \in \Delta_k$ for some k, then $i \notin \Delta_{k'}$ for any $k' \neq k$. Using the assumption that updates may be off by at most m time steps, we thus see that there must exist some permutation $\{u_k\}_{k=1}^n$ of $[n]$ such that

$$\sum_{k=1}^{n} \frac{\sum_{i \in \Delta_k} (g_j^i)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \le \sum_{k=1}^{n} \frac{(g_j^{u_k})^2}{\sqrt{\delta^2 + \sum_{i \le k-m} (g_j^{u_i})^2}}. \tag{4.12}$$

For our last step, we use our assumption that $\delta^2 \ge m M_j^2$ and the standard AdaGrad result (Lemma 3.2) to obtain

$$\sum_{k=1}^{n} \frac{\sum_{i \prec_{S_j} k-1, i \succeq_{S_j} k} (g_j^i)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \le \sum_{k=1}^{n} \frac{(g_j^{u_k})^2}{\sqrt{\sum_{i=1}^{k} (g_j^{u_i})^2}} \le 2\Big(\sum_{k=1}^{n} (g_j^k)^2\Big)^{\frac{1}{2}}.$$

Recalling inequality (4.11), we have

$$\sum_{k=1}^{n} \left[\psi_k^*\Big(-\sum_{i=1}^{k-1} g^i\Big) - \psi_{k-1}^*\Big(-\sum_{i=1}^{k-1} g^i\Big)\right] \le \frac{\mathsf{r}_\infty^2}{2\alpha} \sum_{j=1}^{d} \Big(\sum_{k=1}^{n} (g_j^k)^2\Big)^{\frac{1}{2}},$$

which gives the statement of the lemma. □

4.4.4 Proof of Lemma 4.2

Let us provide a bit of notation before proving the lemma. We define the batch of "outstanding" updates for coordinate j at time k as $\mathcal{B}_j^k := \{i : k-1 \geq i \succeq_{z_j} k\}$, and we define the quantity that we wish to bound in expectation in Lemma 4.2 as

$$\mathcal{T}^j := \frac{1}{\alpha} \sum_{k=1}^{n} g_j^k (\theta_j^k - \widetilde{\theta}_j^k).$$

Turning to the proof of the lemma proper, we first note that z^k does not include any gradient terms g^i for any $i \geq k$ by the definition of the AsyncAdaGrad algorithm. Thus

$$\sum_{i=1}^{k-1} g_j^i - z_j^k = \sum_{i \in \mathcal{B}_j^k} g_j^i.$$

For brief notational convenience define $\kappa_k = \alpha(\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2)^{\frac{1}{2}}$. Applying the definition of the AsyncAdaGrad updates and Young's inequality, we see that

$$\begin{aligned} g_j^k \cdot (x_j^k - \widetilde{x}_j^k) &\leq \kappa_k |g_j^k| \left| g_j^{1:k-1} - z_j^k \right| \\ &= \kappa_k |g_j^k| \left| \sum_{i \in \mathcal{B}_j^k} g_j^i \right| \leq \frac{1}{2} \kappa_k |g_j^k|^2 \mathbf{1}\left\{x_j^k \neq 0\right\} + \frac{1}{2} \kappa_k \mathbf{1}\left\{x_j^k \neq 0\right\} \left(\sum_{i \in \mathcal{B}_j^k} g_j^i \right)^2. \end{aligned}$$

As a consequence, we find that

$$2\mathbb{E}[\mathcal{T}^j] \leq \sum_{k=1}^{n} \mathbb{E}\left[\frac{\mathrm{card}(\{i \in \mathcal{B}_j^k : X_j^i \neq 0\})(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} + \frac{\mathbf{1}\left\{X_j^k \neq 0\right\} \sum_{i \in \mathcal{B}_j^k} (g_j^i)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \right]. \tag{4.13}$$

Looking at the first term in the bound (4.13), we note that $\mathcal{B}_j^k$ consists of time indices i such that $k \preceq_{z_j} i \leq k-1$, which consequently have not been incorporated into any vectors used in the computation of g_j^k. Thus, if we let $\mathcal{F}_{k,j}$ denote the σ-field containing X_j^k and X_j^i for $i \prec_{z_j} k$, we have $g_j^i \in \mathcal{F}_{k,j}$ for any $i \prec_{S_j} k$, the inclusion $g_j^k \in \mathcal{F}_{k,j}$, and we also have that X_j^i is independent of $\mathcal{F}_{k,j}$ for $i \in \mathcal{B}_j^k$. Thus, iterating expectations, we find

$$\begin{aligned} \mathbb{E}\left[\frac{\mathrm{card}(\{i \in \mathcal{B}_j^k : X_j^i \neq 0\})(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \right] &= \mathbb{E}\left[\frac{\mathbb{E}[\mathrm{card}(\{i \in \mathcal{B}_j^k : X_j^i \neq 0\})(g_j^k)^2 \mid \mathcal{F}_{k,j}]}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \right] \\ &\leq p_j m \mathbb{E}\left[\frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \right], \end{aligned}$$

since $\mathbb{E}[\text{card}(\{i \in \mathcal{B}_j^k : X_j^i \neq 0\})] \leq p_j m$ because $|\mathcal{B}_j^k| \leq m$ by assumption. A similar iteration of expectation—since X_j^k is independent of any g_j^i for $i \in \mathcal{B}_j^k$—yields

$$\mathbb{E}\left[\mathbf{1}\left\{X_j^k \neq 0\right\} \sum_{i \in \mathcal{B}_j^k} (g_j^i)^2\right] \leq p_j \mathbb{E}\left[\sum_{i \in \mathcal{B}_j^k} (g_j^i)^2\right].$$

We replace the relevant terms in the expectation (4.13) with the preceding bounds to obtain

$$2\mathbb{E}[\mathcal{T}^j] \leq p_j m \sum_{k=1}^n \mathbb{E}\left[\frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}}\right] + p_j \sum_{k=1}^n \mathbb{E}\left[\frac{\sum_{i \in \mathcal{B}_j^k} (g_j^i)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}}\right].$$

For the second term, note each g_j^i can occur in at most m of the sets $\mathcal{B}_j^k$, and the maximum delay is also at most m. Thus, following the same argument as (4.12), there must exist a permutation $\{u_k\}$ of the indices $[n]$ such that

$$\begin{aligned}\sum_{k=1}^n \frac{\sum_{i \in \mathcal{B}_j^k} (g_j^i)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} &\leq \sum_{k=1}^n \frac{m(g_j^{u_k})^2}{\sqrt{\delta^2 + \sum_{i=1}^{k-m} (g_j^{u_i})^2}} \\ &\leq \sum_{k=1}^n \frac{m(g_j^{u_k})^2}{\sqrt{\sum_{i=1}^{k} (g_j^{u_i})^2}} \leq 2m\left(\sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}},\end{aligned}$$

where we have used the fact that $\delta^2 \geq m M_j^2$ and Lemma 3.2. With this, we immediately find that

$$2\mathbb{E}[\mathcal{T}^j] \leq p_j m \sum_{k=1}^n \mathbb{E}\left[\frac{(g_j^k)^2}{\sqrt{\sum_{i=1}^k (g_j^i)^2}}\right] + p_j \sum_{k=1}^n \mathbb{E}\left[\frac{\sum_{i=k-m}^{k-1} (g_j^i)^2}{\sqrt{\sum_{i=1}^k (g_j^i)^2}}\right] \leq 4p_j m \mathbb{E}\left[\left(\sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}}\right].$$

By inspection, this completes the proof of the lemma.

4.4.5 Sharpening the analysis (proof of Corollary 4.2)

We now demonstrate how to sharpen the analysis in the proof of Theorem 4.2 to allow the initial matrix δ^2 to be smaller than $m M_j^2$. Roughly, we argue that for a smaller setting of δ^2, we can have $\delta^2 \geq \sum_{i=k-m+1}^k (g_j^i)^2$ for all t with high probability, in which case all the previous arguments go through verbatim. In particular, we show how the terms $\mathcal{T}_1$ and $\mathcal{T}_2$ defined in expression (4.8) may be bounded under the weaker assumptions on δ^2 specified in Corollary 4.2.

For this argument, we focus on $\mathcal{T}_1$, as the argument for $\mathcal{T}_2$ is identical. We begin by defining the event $\mathcal{E}$ to occur if $\delta^2 \geq \sum_{i=k-m+1}^k (g_j^i)^2$ for all k. We then have

$$\mathbf{1}\{\mathcal{E}\} \sum_{k=1}^n \frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{Sj} k} (g_j^i)^2}} \leq \mathbf{1}\{\mathcal{E}\} \sum_{k=1}^n \frac{(g_j^k)^2}{\sum_{i=1}^k (g_j^i)^2} \leq 2\left(\sum_{k=1}^n (g_j^k)^2\right)^{\frac{1}{2}}$$

by Lemma 3.2. On the other hand, on $\mathcal{E}^c$, we have by our assumption that $\delta^2 \geq M_j^2$ that

$$\mathbf{1}\left\{\mathcal{E}^c\right\} \sum_{k=1}^{n} \frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}} \leq \mathbf{1}\left\{\mathcal{E}^c\right\} \sum_{k=1}^{n} |g_j^k|,$$

so if we can show that $\mathcal{E}^c$ has sufficiently low probability, then we still obtain our desired results. Indeed, by Hölder's inequality we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{n} \frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}}\right] &\leq 2\mathbb{E}\left[\left(\sum_{k=1}^{n} (g_j^k)^2\right)^{\frac{1}{2}}\right] + \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}^c\right\} \sum_{k=1}^{n} \frac{(g_j^k)^2}{\sqrt{\delta^2 + \sum_{i \prec_{S_j} k} (g_j^i)^2}}\right] \\
&\leq 2\mathbb{E}\left[\left(\sum_{k=1}^{n} (g_j^k)^2\right)^{\frac{1}{2}}\right] + \mathbb{E}[\mathbf{1}\left\{\mathcal{E}^c\right\}]^{\frac{1}{2}} \mathbb{E}\left[\left(\sum_{k=1}^{n} |g_j^k|\right)^2\right]^{\frac{1}{2}}. \quad (4.14)
\end{aligned}$$

It thus remains to argue that $\mathbb{P}(\mathcal{E}^c)$ is very small, since

$$\mathbb{E}\left[\left(\sum_{k=1}^{n} |g_j^k|\right)^2\right] \leq n\mathbb{E}\left[\sum_{k=1}^{n} (g_j^k)^2\right]$$

by Jensen's inequality. Now, we note that $(g_j^k)^2 \leq M_j^2$ and that the X_j^k are i.i.d., so we can define the sequence $Y_k = \mathbf{1}\left\{X_j^k \neq 0\right\}$ and we have

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\exists\, k \in [n] : \sum_{i=k}^{k+m-1} (g_j^i)^2 > \delta^2\right) \leq \mathbb{P}\left(\exists\, k \in [n] : \sum_{i=k}^{k+m-1} Y_k > \delta^2/M_j^2\right).$$

Definine $\gamma = \delta^2/M_j^2$, and let $p = p_j$ for shorthand. Since $Y_k \leq 1$, $\mathbb{E}[Y_k] \leq p$, and $\mathrm{Var}(Y_k) \leq p(1-p)$, Bernstein's inequality implies that for any fixed t and any $\epsilon \geq 0$

$$\mathbb{P}\left(\sum_{i=k}^{k+m-1} Y_k \geq pm + \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2mp(1-p) + 2\epsilon/3}\right). \quad (4.15)$$

By solving a quadratic, we find that if

$$\epsilon \geq \frac{1}{3}\log\frac{1}{\delta} + \sqrt{\frac{1}{9}\log^2\frac{1}{\delta} + 2mp(1-p)\log\frac{1}{\delta}}$$

then the quantity (4.15) is bounded by δ. By a union bound (and minor simplification), we find

$$\epsilon \geq \frac{2}{3}\log\frac{1}{\delta} + \sqrt{2mp(1-p)\log\frac{1}{\delta}} \quad \text{implies} \quad \mathbb{P}(\mathcal{E}^c) \leq n\delta.$$

Setting $\delta = n^{-2}$ means that $\mathbb{P}(\mathcal{E}^c) \leq 1/n$, which in turn implies that

$$\mathbb{E}[\mathbf{1}\{\mathcal{E}^c\}]^{\frac{1}{2}} \mathbb{E}\left[\left(\sum_{k=1}^{n} |g_j^k|\right)^2\right]^{\frac{1}{2}} \leq \frac{1}{\sqrt{n}} \sqrt{n} \mathbb{E}\left[\sum_{k=1}^{n} (g_j^k)^2\right]^{\frac{1}{2}}.$$

Combining the preceding display with inequality (4.14), we find that the term $\mathcal{T}_1$ from expression (4.8) is bounded by

$$\mathbb{E}[\mathcal{T}_1] \leq \sum_{j=1}^{d} \left(2\mathbb{E}\left[\left(\sum_{k=1}^{n} (g_j^k)^2\right)^{\frac{1}{2}}\right] + \mathbb{E}\left[\sum_{k=1}^{n} (g_j^k)^2\right]^{\frac{1}{2}}\right)$$

whenever $\delta^2 \geq \frac{4}{3}\log n + 2\sqrt{mp_j(1-p_j)\log n}$ for all $j \in [d]$. This completes the sharper proof for the bound on $\mathcal{T}_1$. To provide a similar bound for $\mathcal{T}_2$ in analogy to Lemma 4.2, we recall the bound (4.13). Then following the above steps, *mutatis mutandis*, gives the desired result.

Chapter 5

Randomized smoothing for stochastic optimization

In this chapter of the thesis, we continue our study of efficient stochastic optimization algorithms and structures we may leverage to derive faster algorithms. In particular, we analyze convergence rates of stochastic optimization algorithms for non-smooth convex optimization problems. By combining randomized smoothing techniques with accelerated gradient methods, we obtain convergence rates that have optimal dependence on the *variance* of the gradient estimates rather than their maximum magnitude. To the best of our knowledge, these are the first variance-based rates for non-smooth optimization. We give several applications of our results to statistical estimation problems, and provide experimental results that demonstrate the effectiveness of the proposed algorithms. We also describe how a combination of our algorithm with recent work on decentralized optimization yields an order-optimal distributed stochastic optimization algorithm. For our randomized smoothing techniques to guarantee sufficient speedup over other standard methods, it is essential that we solve stochastic optimization problems (3.1); the noise already inherent in the problem means that adding a bit of additional randomness does not hurt.

5.1 Introduction

In this chapter, we develop and analyze randomized smoothing procedures for solving the class of stochastic optimization problems introduced in Chapter 3 and described by the problem (3.1). Recalling this family of problems for the sake of the exposition of the chapter, we begin with the usual risk functional

$$f(\theta) := \mathbb{E}\big[F(\theta; X)\big] = \int_{\mathcal{X}} F(\theta; x) dP(x).$$

We focus here on potentially non-smooth stochastic optimization problems of the form

$$\underset{\theta \in \Theta}{\text{minimize}} \ \big\{ f(\theta) + \varphi(\theta) \big\}, \tag{5.1}$$

where $\varphi : \Theta \to \mathbb{R}$ is a known regularizing function. We assume that φ is closed and convex, but we allow for non-differentiability so that the framework includes the ℓ_1-norm and related regularizers.

While we do consider effects of the regularizer φ on our optimization procedures, our primary focus is on the properties of the stochastic function f. The problem (5.1) is challenging mainly for two reasons. First, as mentioned in introducing the problem in Chapter 3, in many cases f cannot actually be evaluated, either because the associated integral is computationally intractable or P is not known. Thus, as usual, we assume only that we have access to a stochastic oracle that allows us to obtain i.i.d. observations $X \sim P$ and may compute (sub)gradients $\mathsf{g}(\theta; X) \in \partial F(\theta; X)$ as in expression (3.3). Second, in many cases, the f function is non-smooth, that is, it is non-differentiable.

In order to address difficulties associated with non-smooth objective functions, several researchers have considered techniques for smoothing the objective. Such approaches for deterministic non-smooth problems are by now well-known, and include Moreau-Yosida regularization (e.g. [119]), methods based on recession functions [21]; and Nesterov's approach using conjugacy and proximal regularization [137]. Several works study methods to smooth exact penalties of the form $\max\{0, f(\theta)\}$ in convex problems, where smoothing is applied to the $\max\{0, \cdot\}$ operator (for instance, see the paper [45] and references therein). The difficulty of such approaches is that most require quite detailed knowledge of the structure of the function f to be minimized and are thus impractical in stochastic settings.

Because the convex objective (5.1) cannot actually be evaluated except through stochastic realization of f and its (sub)gradients, we develop an algorithm for solving problem (5.1) based on stochastic subgradient methods. Such methods are classical [150, 73]; in recent work, Juditsky et al. [104] and Lan [114] have shown that if f is smooth, meaning that its gradients are Lipschitz continuous, and if the variance of the stochastic gradient estimator is at most σ^2, then the resulting stochastic optimization procedure has convergence rate $\mathcal{O}(\sigma/\sqrt{n})$. Of particular relevance to our study is the following fact: if the gradient oracle—instead of returning just a single estimate—returns m unbiased estimates of the gradient, the variance of the gradient estimator is reduced by a factor of m. Indeed, Dekel et al. [49] exploit this fact to develop asymptotically order-optimal distributed optimization algorithms, as we discuss in the sequel.

To the best of our knowledge, there is no work on *non-smooth* stochastic problems for which a reduction in the variance of the stochastic estimate of the true subgradient gives an improvement in convergence rates. For non-smooth stochastic optimization, known convergence rates depend only on the Lipschitz constant of the functions $F(\cdot; x)$ and the number of actual updates performed (recall Chapter 3.1). Within the oracle model of convex optimization [134], the optimizer has access to a black-box oracle that, given a point $\theta \in \Theta$, returns an unbiased estimate of a (sub)gradient of f at the point θ. In most stochastic optimization procedures, an algorithm updates a parameter θ^k after each query of the oracle; we consider the natural extension to the case when the optimizer issues several queries to the stochastic oracle at every iteration.

The starting point for our approach is a convolution-based smoothing technique amenable

to non-smooth stochastic optimization problems. A number of authors (e.g., Katkovnik and Kulchitsky [106], Rubinstein [156], Lakshmanan and de Farias [113] and Yousefian et al. [187]) have noted that random perturbation of the variable θ can be used to transform f into a smooth function. The intuition underlying such approaches is that the convolution of two functions is at least as smooth as the smoothest of the two original functions. In particular, letting μ denote the density of a random variable with respect to Lebesgue measure, consider the smoothed objective function

$$f_u(\theta) := \mathbb{E}_\mu[f(\theta + uZ)] = \int_{\mathbb{R}^d} f(\theta + uz)\mu(z)dz, \tag{5.2}$$

where Z is a random variable with density μ. Clearly, the function f_u is convex when f is convex; moreover, since μ is a density with respect to Lebesgue measure, the function f_u is also guaranteed to be differentiable (e.g. Bertsekas [25]).

We analyze minimization procedures that solve the non-smooth problem (5.1) by using stochastic gradient samples from the smoothed function (5.2) with appropriate choice of smoothing density μ. The main contribution of this chapter is to show that the ability to issue several queries to the stochastic oracle for the original objective can give faster rates of convergence than a simple stochastic oracle. Our main theorem quantifies the above statement in terms of expected values (Theorem 5.1). Under an additional reasonable tail condition, it is possible to provide high-probability guarantees on convergence rate, but to keep this chapter relatively compact and focused on the essential ideas, we leave such statements to the paper off of which this chapter is based [55]. One consequence of our results is that a procedure that queries the non-smooth stochastic oracle for m subgradients at iteration k achieves rate of convergence $\mathcal{O}(\mathsf{r}_2 M/\sqrt{nm})$ in expectation and with high probability. (Here M is the Lipschitz constant of the function f and r_2 is the ℓ_2-radius of the domain Θ.) As we discuss in Section 5.2.4, this convergence rate is optimal up to constant factors. Moreover, this fast rate of convergence has implications for applications in statistical problems, distributed optimization, and other areas, as discussed in Section 5.3.

The remainder of the chapter is organized as follows. In Section 5.2, we begin by providing background on some standard techniques for stochastic optimization, noting a few of their deficiencies for our setting. We then describe an algorithm based on the randomized smoothing technique (5.2), and we state our main theorems guaranteeing faster rates of convergence for non-smooth stochastic problems. In proving these claims, we make frequent use of the analytic properties of randomized smoothing, many of which we collect in Section 5.6. In Section 5.3, we discuss applications of our methods and provide experimental results illustrating the merits of our approach. Finally, we provide the proofs of our results in Section 5.5, with certain more technical aspects deferred.

5.2 Main results and some consequences

We begin by motivating the algorithm studied in this paper, and we then state our main results on its convergence.

5.2.1 Some background

We focus on stochastic gradient descent methods[1] based on dual averaging schemes (3.8) (due to Nesterov [138], whose composite version, which incorporates the regularizer φ, Xiao [183] develops) for solving the stochastic problem (5.1). We recall that, for the regularized objective (5.1), the composite dual averaging update based on the strongly convex proximal function ψ is as follows. Given a point $\theta^k \in \Theta$, the algorithm queries the stochastic oracle $g^k = \mathsf{g}(\theta^k; X_k) \in \partial F(\theta^k; X_k)$ where $X_k \stackrel{\text{i.i.d.}}{\sim} P$, and the algorithm then performs the update

$$\theta^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \sum_{i=1}^{k} \left\langle g^i, \theta \right\rangle + k\varphi(\theta) + \frac{1}{\alpha_k} \psi(\theta) \right\}, \tag{5.3}$$

where $\alpha_k > 0$ is a sequence of stepsizes. Recalling our results from Chapter 3.1, we note if ψ is strongly convex with respect to the norm $\|\cdot\|$ and $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_*^2] \le M^2$ for all $\theta \in \Theta$, then with stepsize $\alpha_k \propto \sqrt{\psi(\theta^*)}/M\sqrt{k}$ and $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^{n} \theta^k$ we have

$$\mathbb{E}\left[f(\widehat{\theta}(n)) + \varphi(\widehat{\theta}(n)) \right] - f(\theta^*) - \varphi(\theta^*) \le \mathcal{O}(1) \frac{M\sqrt{\psi(\theta^*)}}{\sqrt{n}}. \tag{5.4}$$

We refer to the papers by Nesterov [138] and Xiao [183] for results of this type.

An unsatisfying aspect of the bound (5.4) is the absence of any role for the variance of the (sub)gradient estimator g^k. Even if an algorithm is able to obtain $m > 1$ samples of the (sub)gradient of f at θ^k—giving a more accurate gradient estimate—this result fails to capture the potential improvement of the method. We address this problem by stochastically smoothing the non-smooth objective f and then adapt recent work on so-called accelerated gradient methods [114, 170, 183], which apply only to smooth functions, to achieve variance-based improvements. With this motivation in mind, we now turn to developing the tools necessary for stochastic smoothing of the non-smooth objective function (5.1).

5.2.2 Description of algorithm

Our algorithm is based on observations of stochastically perturbed gradient information at each iteration, where we slowly decrease the perturbation as the algorithm proceeds. Consider the following scheme. Let $\{u_k\} \subset \mathbb{R}_+$ be a non-increasing sequence of positive real numbers; these quantities control the perturbation size. At iteration k, rather than query the stochastic oracle at a fixed query point w^k, the algorithm queries the oracle at m points drawn randomly from some neighborhood around w^k. Specifically, it performs the following three steps:

(1) Draws random variables $\{Z_{k,t}\}_{t=1}^m$ i.i.d. according to the distribution μ.

[1] We note in passing that essentially identical results can also be obtained for methods based on mirror descent [134, 170], though we omit these to avoid overburdening the reader.

(2) Queries the oracle at the m points $w^k + u_k Z_{k,t}$ for $t = 1, 2, \ldots, m$, yielding the stochastic (sub)gradients

$$g_t^k = \mathsf{g}(w^k + u_k Z_{k,t}; X_{k,t}) \in \partial F(w^k + u_k Z_{k,t}; X_{k,t}), \quad \text{where } X_{k,t} \overset{\text{i.i.d.}}{\sim} P \text{ for } t \in [m]. \tag{5.5}$$

(3) Computes the average $g^k = \frac{1}{m} \sum_{t=1}^m g_t^k$.

Here and throughout we denote the distribution of the random variable $u_k Z$ by μ_k, and we note that this procedure ensures $\mathbb{E}[g^k \mid w^k] = \nabla f_{u_t}(w^k) = \nabla \mathbb{E}[F(w^k + u_k Z; X) \mid w^k]$, where

$$f_u(\theta) := \mathbb{E}[f(\theta + uZ)] = \int_{\mathbb{R}^d} f(\theta + uz) \mu(z) dz$$

is the smoothed function (5.2) indexed by amount of smoothing u.

We combine the sampling scheme (5.5) with extensions of Tseng's recent work on accelerated gradient methods [170] and propose an update that is essentially a smoothed version of the simpler method (5.3). The method uses three series of points, denoted $\{\theta^k, w^k, v^k\} \in \Theta^3$. We use w^k as a "query point", so that at iteration k, the algorithm receives a vector g^k as described in the sampling scheme (5.5). The three sequences evolve according to a dual-averaging algorithm, which in our case involves three scalars $(L_k, \nu_k, \eta_k) \in \mathbb{R}_+ \times [0, 1] \times \mathbb{R}_+$ to control step sizes. The recursions are as follows:

$$w^k = (1 - \nu_k)\theta^k + \nu_k v^k \tag{5.6a}$$

$$v^{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \sum_{i=0}^k \frac{1}{\nu_i} \left\langle g^i, \theta \right\rangle + \sum_{i=0}^k \frac{1}{\nu_i} \varphi(\theta) + L_{k+1} \psi(\theta) + \frac{\eta_{k+1}}{\nu_{k+1}} \psi(\theta) \right\} \tag{5.6b}$$

$$\theta^{k+1} = (1 - \nu_k)\theta^k + \nu_k v^{k+1}. \tag{5.6c}$$

In prior work on accelerated schemes for stochastic and non-stochastic optimization [170, 114, 183], the term L_k is set equal to the Lipschitz constant of ∇f; in contrast, our choice of varying L_k allows our smoothing schemes to be oblivious to the number of iterations n. The extra damping term η_k / ν_k provides control over the fluctuations induced by using the random vector g^k as opposed to deterministic subgradient information. As in Tseng's work [170], we assume that $\nu_0 = 1$ and $(1 - \nu_k)/\nu_k^2 = 1/\nu_{k-1}^2$; the latter equality is ensured by setting $\nu_k = 2/(1 + \sqrt{1 + 4/\nu_{k-1}^2})$.

5.2.3 Convergence rates

We now state our two main results on the convergence rate of the randomized smoothing procedure (5.5) with accelerated dual averaging updates (5.6a)–(5.6c). To avoid cluttering the theorem statements, we begin by stating our main assumptions. Whenever we state that a function f is Lipschitz continuous, we mean with respect to the norm $\|\cdot\|$, and we assume

that ψ is nonnegative and is strongly convex with respect to the same norm $\|\cdot\|$. Our main assumption ensures that the smoothing operator and smoothed function f_u are relatively well-behaved.

Assumption 5A (Smoothing)**.** *The random variable Z is zero-mean and has density μ (with respect to Lebesgue measure on the affine hull $\operatorname{aff}(\Theta)$ of Θ). There are constants M and L such that for $u > 0$, $\mathbb{E}[f(\theta + uZ)] \leq f(\theta) + Mu$, and $\mathbb{E}[f(\theta + uZ)]$ has $\frac{L}{u}$-Lipschitz continuous gradient with respect to the norm $\|\cdot\|$. Additionally, for P-almost every $x \in \mathcal{X}$, the set $\operatorname{dom} F(\cdot; x) \supseteq u_0 \operatorname{supp} \mu + \Theta$.*

Recall our definition of the smoothed function $f_{u_k}(\theta) = \int f(\theta + u_k z) d\mu(z)$. The function f_{u_k} is guaranteed to be smooth whenever μ is a density with respect to Lebesgue measure, so Assumption 5A ensures that f_{u_k} is uniformly close to f and not too "jagged." For Lipschitz f, many smoothing distributions, including Gaussians and uniform distributions on norm balls, satisfy Assumption 5A (see Section 5.6); we use such examples in the corollaries to follow. The containment of $u_0 \operatorname{supp} \mu + \Theta$ in $\operatorname{dom} F(\cdot; x)$ guarantees that the subdifferential $\partial F(\cdot; x)$ is non-empty at all sampled points $w^k + u_k Z$. Indeed, since μ is a density with respect to Lebesgue measure on $\operatorname{aff}(\Theta)$, with probability one $w^k + u_k Z \in \operatorname{relint} \operatorname{dom} F(\cdot; x)$ and thus [97] the subdifferential $\partial F(w^k + u_k Z; x) \neq \emptyset$.

In the algorithm (5.6a)–(5.6c), we set L_k to be an upper bound on the Lipschitz constant $\frac{L}{u_k}$ of the gradient of $\mathbb{E}[f(\theta + u_k Z)]$; this choice ensures good convergence properties of the algorithm. The following is our main theorem.

Theorem 5.1. *Define $u_k = \nu_k u$, use the scalar sequence $L_k = L/u_k$, and assume that η_k is non-decreasing. Under Assumption 5A, for any $\theta^* \in \Theta$ and $n \geq 4$,*

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{6L\psi(\theta^*)}{nu} + \frac{2\eta_k \psi(\theta^*)}{n} + \frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{\eta_k} \mathbb{E}[\|e^k\|_*^2] + \frac{4Mu}{n}, \quad (5.7)$$

where $e^k := \nabla f_{u_k}(w^k) - g^k$ is the error in the gradient estimate.

Remarks The convergence rate (5.7) involves the variance $\mathbb{E}[\|e^k\|_*^2]$ explicitly, which we exploit in the corollaries to be stated shortly. In addition, Theorem 5.1 does not require a priori knowledge of the number of iterations n to be performed, thereby rendering it suitable to online and streaming applications. If n is known, a similar result holds for constant smoothing parameter u, as formalized by Corollary 5.1, which uses a fixed setting of the smoothing parameter u_k:

Corollary 5.1. *Suppose that $u_k \equiv u$ for all k and set $L_k \equiv L/u$. With the remaining conditions as in Theorem 5.1, then for any $\theta^* \in \Theta$, we have*

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{4L\psi(\theta^*)}{n^2 u} + \frac{2\eta_n \psi(\theta^*)}{n} + \frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{\eta_k} \mathbb{E}\big[\|e^k\|_*^2\big] + Mu.$$

It is clear that by setting $u \propto 1/n$, the rates achieved by Theorem 5.1 and Corollary 5.1 are identical to constant factors.

5.2.4 Some consequences

We now turn to corollaries of the above theorems and the consequential optimality guarantees of the algorithm. More precisely, we establish concrete convergence bounds for algorithms using different choices of the smoothing distribution μ. For each corollary, we impose the assumptions that the point $\theta^* \in \Theta$ satisfies $\psi(\theta^*) \leq \mathsf{r}_\psi^2$, the iteration number $n \geq 4$, and $u_k = u\nu_k$.

We begin with a corollary that provides bounds when the smoothing distribution μ is uniform on the ℓ_2-ball. The conditions on F in the corollary hold, for example, when $F(\cdot; x)$ is M-Lipschitz with respect to the ℓ_2-norm for P-a.e. sample of x.

Corollary 5.2. *Let μ be uniform on the ℓ_2-ball B_2 of radius 1, use the proximal function $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$ and assume $\mathsf{r}_2^2 \geq \psi(\theta^*)$. Also assume $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \leq M^2$ for $\theta \in \Theta + uB_2$, where we set $u = \mathsf{r}_2 d^{1/4}$. With step sizes $\eta_k = M\sqrt{k+1}/\mathsf{r}_2\sqrt{m}$ and $L_k = M\sqrt{d}/u_k$,*

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{10M\mathsf{r}_2 d^{1/4}}{n} + \frac{5M\mathsf{r}_2}{\sqrt{nm}}.$$

The following corollary shows that similar convergence rates are attained when smoothing with the normal distribution using the same proximal function as that in Corollary 5.2.

Corollary 5.3. *Let μ be the d-dimensional normal distribution with zero mean and identity covariance $I_{d\times d}$ and assume $F(\cdot; x)$ is M-Lipschitz with respect to the ℓ_2-norm for P-a.e. x. With smoothing parameter $u = \mathsf{r}_2 d^{-1/4}$ and step sizes $\eta_k = M\sqrt{k+1}/\mathsf{r}_2\sqrt{m}$ and $L_k = M/u_k$, we have*

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{10M\mathsf{r}_2 d^{1/4}}{n} + \frac{5M\mathsf{r}_2}{\sqrt{nm}}.$$

We note here (deferring deeper discussion to Lemma 5.6) that the dimension dependence of $d^{1/4}$ on the $1/n$ term in the previous corollaries cannot be improved by more than a constant factor. Essentially, functions f exist whose smoothed version f_u cannot have both Lipschitz continuous gradient and be uniformly close to f in a dimension-independent sense, at least for the uniform or normal distributions.

The advantage of using normal random variables—as opposed to Z uniform on the ℓ_2-ball B_2—is that no normalization of Z is necessary, though there are more stringent requirements on f. The lack of normalization is a useful property in very high dimensional scenarios, such as statistical natural language processing (NLP) [127]. Similarly, we can sample Z from an ℓ_∞ ball, which, like B_2, is still compact, but gives slightly looser bounds than sampling from B_2. Nonetheless, it is much easier to sample from B_∞ in high dimensional settings, especially sparse data scenarios such as NLP where only a few coordinates of the random variable Z are needed.

There are several objectives $f + \varphi$ with domains Θ for which the natural geometry is non-Euclidean, which motivates the mirror descent family of algorithms [134]. Here we give an example that is quite useful for problems in which the optimizer θ^* is sparse; for example, the optimization set Θ may be a simplex or ℓ_1-ball, or $\varphi(\theta) = \lambda \left\|\theta\right\|_1$. The point here is that we achieve a pair of dual norms that may give better optimization performance than the ℓ_2-ℓ_2 pair above.

Corollary 5.4. *Let μ be uniform on the ℓ_∞-ball B_∞ and assume that $F(\cdot; x)$ is M-Lipschitz continuous with respect to the ℓ_1-norm over $\Theta + uB_\infty$ for $x \in \mathcal{X}$, where we set $u = \mathsf{r}_\psi \sqrt{d \log d}$. Use the proximal function $\psi(\theta) = \frac{1}{2(p-1)} \left\|\theta\right\|_p^2$ for $p = 1 + 1/\log d$ and set $\eta_k = \sqrt{k+1}/\mathsf{r}_\psi \sqrt{m \log d}$ and $L_k = M/u_k$. There is a universal constant C such that*

$$\begin{aligned}\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] &\le C \frac{M \mathsf{r}_\psi \sqrt{d}}{n} + C \frac{M \mathsf{r}_\psi \sqrt{\log d}}{\sqrt{nm}} \\ &= \mathcal{O}(1) \left[\frac{M \left\|\theta^*\right\|_1 \sqrt{d \log d}}{n} + \frac{M \left\|\theta^*\right\|_1 \log d}{\sqrt{nm}} \right].\end{aligned}$$

The dimension dependence of $\sqrt{d \log d}$ on the leading $1/n$ term in the corollary is weaker than the $d^{1/4}$ dependence in the earlier corollaries, so for very large m the corollary is not as strong as one might desire when applied to non-Euclidean geometries. Nonetheless, for large n the $1/\sqrt{nm}$ terms dominate the convergence rates, and Corollary 5.4 can be an improvement.

Remarks Let us pause to make some remarks concerning the corollaries given above. First, if one abandons the requirement that the optimization procedure be an "any time" algorithm, meaning that it is able to return a result at any iteration, it is possible to obtain essentially the same results as Corollaries 5.2 through 5.4 by choosing a fixed setting $u_k = u/n$ (recall Corollary 5.1). As a side benefit, it is then easier to satisfy the Lipschitz condition that $\mathbb{E}[\|\mathsf{g}(\theta; X)\|^2] \le M^2$ for $\theta \in \Theta + u_0 \operatorname{supp} \mu$. Our second observation is that Theorem 5.1 and the corollaries appear to require a very specific setting of the constant L_k to achieve fast rates. However, the algorithm is robust to mis-specification of L_k since the instantaneous smoothness constant L_k is dominated by the stochastic damping term η_k in the algorithm. Indeed, since η_k grows proportionally to $\sqrt{k}$ for each corollary, we have $L_k = L/u_k = L/\nu_k u = \mathcal{O}(\eta_k/\sqrt{k}\nu_k)$; that is, L_k is order $\sqrt{k}$ smaller than η_k/ν_k, so setting L_k incorrectly up to order $\sqrt{k}$ has essentially negligible effect. (See also the experimental section of [56].)

We can show the bounds in the theorems above are tight, meaning unimprovable up to constant factors, by exploiting the lower bounds we presented in Chapter 3.3 for stochastic optimization problems (see also Nemirovski and Yudin [134] and Agarwal et al. [6]). For instance, let us set $\Theta = \{\theta \in \mathbb{R}^d \mid \left\|\theta\right\|_2 \le \mathsf{r}_2\}$, and consider the class of all convex functions f that are $M_{0,2}$-Lipschitz with respect to the ℓ_2-norm. Assume that the stochastic

(sub)gradient oracle (3.3), for any fixed θ, satisfies $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \le M_{0,2}^2$. Then for *any* method that outputs a point $\theta^n \in \Theta$ after n queries of the oracle, we have the minimax lower bound

$$\Omega(1)\frac{M_{0,2}\mathsf{r}_2}{\sqrt{n}}$$

(see Chapter 3.3, Proposition 3.4, or Section 3.1 of Agarwal et al. [6]). Moreover, similar bounds hold for problems with non-Euclidean geometry. For instance, let us consider loss functions F that are $M_{0,\infty}$-Lipschitz with respect to the ℓ_1-norm, meaning that $|F(\theta; x) - F(\theta'; x)| \le M_{0,\infty} \|\theta - \theta'\|_1$. If we define $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le \mathsf{r}_1\}$, we have the minimax lower bound

$$\frac{1}{8}M_{0,\infty}\mathsf{r}_1 \min\left\{1, \frac{\sqrt{\log(2d)}}{2\sqrt{n}}\right\},$$

as given in Corollary 3.3. In either geometry, no method can have optimization error smaller than $\mathcal{O}(M\mathsf{r}_\psi/\sqrt{n})$ after at most n queries of the stochastic oracle.

Let us compare the above lower bounds to the convergence rates in Corollaries 5.2 through 5.4. Examining the bound in Corollaries 5.2 and 5.3, we see that the dominant terms are on the order of $M\mathsf{r}_\psi/\sqrt{nm}$ so long as $m \le n/\sqrt{d}$. Since our method issues nm queries to the oracle, this is optimal. Similarly, the strategy of sampling uniformly from the ℓ_∞-ball in Corollary 5.4 is optimal for large enough n. In contrast to other optimization procedures, however, our algorithm performs an update to the parameter θ^k only after every m queries to the oracle; as we show in the next section, this is beneficial in several applications.

5.3 Applications and experimental results

In this section, we describe applications of our results and give experiments that illustrate our theoretical predictions.

5.3.1 Some applications

The first application of our results is to parallel computation and distributed optimization. Imagine that instead of querying the stochastic oracle serially, we can issue queries and aggregate the resulting stochastic gradients in parallel. In particular, assume that we have a procedure in which the m queries of the stochastic oracle occur concurrently. Then Corollaries 5.2 through 5.4 imply that in the same amount of time required to perform n queries and updates of the stochastic gradient oracle serially, achieving an optimization error of $\mathcal{O}(1/\sqrt{n})$, the parallel implementation can process nm queries and consequently has optimization error $\mathcal{O}(1/\sqrt{nm})$.

We now briefly describe two possibilities for a distributed implementation of the above. The simplest architecture is a master-worker architecture, in which one master maintains the parameters (θ^k, w^k, v^k), and each of m workers has access to an uncorrelated stochastic

oracle for P and the smoothing distribution μ. The master broadcasts the point w^k to the workers $t \in [m]$, each of which independently sample $X_t \sim P$ and $Z_t \sim \mu$, returning sample gradients to the master. In a tree-structured network, broadcast and aggregation require at most $\mathcal{O}(\log m)$ steps; the relative speedup over a serial implementation is $\mathcal{O}(m/\log m)$. In recent work, Dekel et al. [49] give a series of reductions showing how to distribute variance-based stochastic algorithms and achieve an asymptotically optimal convergence rate. The algorithm given here, as specified by equations (5.5) and (5.6a)–(5.6c), can be exploited within their framework to achieve an $\mathcal{O}(m)$ improvement in convergence rate over a serial implementation. More precisely, whereas achieving optimization error ϵ requires $\mathcal{O}(1/\epsilon^2)$ iterations for a centralized algorithm, the distributed adaptation requires only $\mathcal{O}(1/(m\epsilon^2))$ iterations. Such an improvement is possible as a consequence of the variance reduction techniques we have described.

A second application of interest involves problems where the set Θ and/or the function φ are complicated, so that calculating the proximal update (5.6b) becomes expensive. The proximal update may be distilled to computing

$$\min_{\theta \in \Theta} \left\{ \langle g, \theta \rangle + \psi(\theta) \right\} \quad \text{or} \quad \min_{\theta \in \Theta} \left\{ \langle g, \theta \rangle + \psi(\theta) + \varphi(\theta) \right\}. \tag{5.8}$$

In such cases, it may be beneficial to accumulate gradients by querying the stochastic oracle several times in each iteration, using the averaged subgradient in the update (5.6b), and thus solve only one proximal sub-problem for a collection of samples.

Let us consider some concrete examples. In statistical applications involving the estimation of covariance matrices, the domain Θ is constrained in the positive semidefinite cone $\{\theta \in \mathbb{S}_d \mid \theta \succeq 0\}$; other applications involve additional nuclear-norm constraints of the form $\Theta = \{\theta \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j=1}^{\min\{d_1, d_2\}} \gamma_j(\theta) \leq C\}$. Examples of such problems include covariance matrix estimation, matrix completion, and model identification in vector autoregressive processes (see the paper [132] and references therein for further discussion). Another example is the problem of metric learning [184, 160], in which the learner is given a set of n points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and a matrix $Y \in \mathbb{R}^{n \times n}$ indicating which points are close together in an unknown metric. The goal is to estimate a positive semidefinite matrix $\theta \succeq 0$ such that $\langle (x_i - x_j), \theta(x_i - x_j) \rangle$ is small when x_i and x_j belong to the same class or are close, while $\langle (x_i - x_j), \theta(x_i - x_j) \rangle$ is large when x_i and x_j belong to different classes. It is desirable that the matrix θ have low rank, which allows the statistician to discover structure or guarantee performance on unseen data. As a concrete illustration, suppose that we are given a matrix $Y \in \{-1, 1\}^{n \times n}$, where $y_{ij} = 1$ if x_i and x_j belong to the same class, and $y_{ij} = -1$ otherwise. In this case, one possible optimization-based estimator involves solving the non-smooth program

$$\min_{\theta, \theta_0} \frac{1}{\binom{n}{2}} \sum_{i<j} \left[1 + y_{ij}(\operatorname{tr}(\theta(x_i - x_j)(x_i - x_j)^\top) + \theta_0)\right]_+ \text{ s.t. } \theta \succeq 0, \ \operatorname{tr}(\theta) \leq C. \tag{5.9}$$

Now let us consider the cost of computing the projection update (5.8) for the metric learning problem (5.9). When $\psi(\theta) = \frac{1}{2} \|\theta\|_{\text{Fr}}^2$, the update (5.8) reduces for appropriate choice of V

to

$$\min_{\theta} \frac{1}{2} \|\theta - V\|_{\text{Fr}}^2 \quad \text{subject to} \quad \theta \succeq 0, \;\; \text{tr}(\theta) \leq C.$$

(As a side-note, it is possible to generalize this update to Schatten p-norms [52].) This problem is equivalent to projecting the eigenvalues of V to the simplex $\{x \in \mathbb{R}^d \mid x \succeq 0, \langle \mathbb{1}, x \rangle \leq C\}$, which after an $\mathcal{O}(d^3)$ eigen-decomposition requires time $\mathcal{O}(d)$ [34]. To see the benefits of the randomized perturbation and averaging technique (5.5) over standard stochastic gradient descent (5.3), consider that the cost of querying a stochastic oracle for the objective (5.9) for one sample pair (i, j) requires time $\mathcal{O}(d^2)$. Thus, m queries require $\mathcal{O}(md^2)$ computation, and each update requires $\mathcal{O}(d^3)$. So we see that after $nmd^2 + nd^3$ units of computation, our randomized perturbation method has optimization error $\mathcal{O}(1/\sqrt{nm})$, while standard stochastic gradient requires nmd^3 units of computation to attain the same error. In short, for $m \approx d$ the randomized smoothing technique (5.5) uses a factor $\mathcal{O}(d)$ less computation than standard stochastic gradient; we give experiments corroborating this in Section 5.3.2.2.

5.3.2 Experimental results

We now describe experimental results that confirm the sharpness of our theoretical predictions. The first experiment explores the benefit of using multiple samples m when estimating the gradient $\nabla f(w^k)$ as in the averaging step (5.5). The second experiment studies the actual amount of time required to solve a statistical metric learning problem, as described in the objective (5.9) above.

5.3.2.1 Iteration Complexity of Reduced Variance Estimators

In this experiment, we consider the number of iterations of the accelerated method (5.6a)–(5.6c) necessary to achieve an ϵ-optimal solution to the problem (5.1). To understand how the iteration scales with the number m of gradient samples, we consider our results in terms of the number of iterations

$$T(\epsilon, m) := \inf \left\{ n \in \{1, 2, \ldots\} \mid f(\theta^n) - \min_{\theta^* \in \Theta} f(\theta^*) \leq \epsilon \right\}$$

required to achieve optimization error ϵ when using m gradient samples in the averaging step (5.5). We focus on the algorithm analyzed in Corollary 5.2, which uses uniform sampling of the ℓ_2-ball. The corollary implies there should be two regimes of convergence: one where the $M\mathsf{r}_2/\sqrt{nm}$ term is dominant, and the other when the number of samples m is so large that the $M\mathsf{r}d^{1/4}/n$ term is dominant. By inverting the first term, we see that for small m, $T(\epsilon, m) = \mathcal{O}(M^2\mathsf{r}_2^2/m\epsilon^2)$, while the second gives $T(\epsilon, m) = \mathcal{O}(M\mathsf{r}_2 d^{1/4}/\epsilon)$. In particular, our theory predicts that

$$T(\epsilon, m) = \mathcal{O}\left(\max\left\{ \frac{M^2\mathsf{r}_2^2}{m\epsilon^2}, \frac{M\mathsf{r}_2 d^{1/4}}{\epsilon} \right\} \right). \tag{5.10}$$

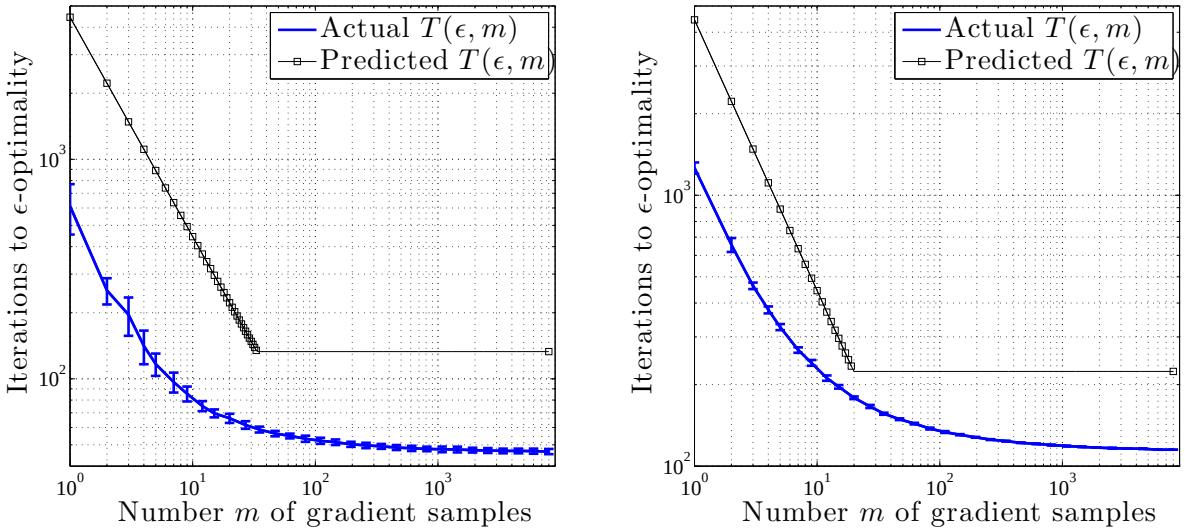

Figure 5.1. The number of iterations $T(\epsilon, m)$ to achieve ϵ-optimal solution for the problem (5.11) as a function of the number of samples m used in the gradient estimate (5.5). The prediction (5.10) is the square black line in each plot; plot (a) shows results for dimension $d = 50$, (b) for $d = 400$.

In order to assess the accuracy of this prediction, we consider a robust linear regression problem, commonly studied in system identification and robust statistics [146, 100]. Specifically, given a matrix $X \in \mathbb{R}^{n \times d}$ and vector $y \in \mathbb{R}^n$, the goal is to minimize the non-smooth objective function

$$f(x) = \frac{1}{n} \|X\theta - y\|_1 = \frac{1}{n} \sum_{i=1}^{n} |\langle x_i, \theta \rangle - y_i|, \tag{5.11}$$

where $x_i \in \mathbb{R}^d$ denotes a transposed row of X. The stochastic oracle in this experiment, when queried at a point θ, chooses an index $i \in [n]$ uniformly at random and returns the vector $\text{sign}(\langle x_i, \theta \rangle - y_i) x_i$.

In our experiments, we generated $n = 1000$ points with $d \in \{50, 100, 200, 400, 800, 1600\}$ dimensions, each with fixed norm $\|x_i\|_2 = M$, and then assigned values y_i by computing $\langle x_i, w \rangle$ for a random vector w (adding normally distributed noise with variance 0.1). We estimated the quantity $T(\epsilon, m)$ for solving the robust regression problem (5.11) for several values of m and d. Figure 5.1 shows results for dimensions $d \in \{50, 400\}$, averaged over 20 experiments for each choice of dimension d. (Other settings of d exhibited similar behavior.) Each panel in the figure shows—on a log-log scale—the experimental average $T(\epsilon, m)$ and the theoretical prediction (5.10). The decrease in $T(\epsilon, m)$ is nearly linear for smaller numbers of samples m; for larger m, the effect is quite diminished. We present numerical results in Table 5.1 that allow us to roughly estimate the number m at which increasing the batch size in the gradient estimate (5.5) gives no further improvement. For each dimension d, Table 5.1

	m	1	2	3	5	20	100	1000	10000
$d = 50$	MEAN	612.2	252.7	195.9	116.7	66.1	52.2	47.7	46.6
	STD	158.29	34.67	38.87	13.63	3.18	1.66	1.42	1.28
$d = 100$	MEAN	762.5	388.3	272.4	193.6	108.6	83.3	75.3	73.3
	STD	56.70	19.50	17.59	10.65	1.91	1.27	0.78	0.78
$d = 200$	MEAN	1002.7	537.8	371.1	267.2	146.8	109.8	97.9	95.0
	STD	68.64	26.94	13.75	12.70	1.66	1.25	0.54	0.45
$d = 400$	MEAN	1261.9	656.2	463.2	326.1	178.8	133.6	118.6	115.0
	STD	60.17	38.59	12.97	8.36	2.04	1.02	0.49	0.00
$d = 800$	MEAN	1477.1	783.9	557.9	388.3	215.3	160.6	142.0	137.4
	STD	44.29	24.87	12.30	9.49	2.90	0.66	0.00	0.49
$d = 1600$	MEAN	1609.5	862.5	632.0	448.9	251.5	191.1	169.4	164.0
	STD	42.83	30.55	12.73	8.17	2.73	0.30	0.49	0.00

Table 5.1. The number of iterations $T(\epsilon, m)$ to achieve ϵ-accuracy for the regression problem (5.11) as a function of number of gradient samples m used in the gradient estimate (5.5) and the dimension d. Each box in the table shows the mean and standard deviation of $T(\epsilon, m)$ measured over 20 trials.

indeed shows that from $m = 1$ to 5, the iteration count $T(\epsilon, m)$ decreases linearly, halving again when we reach $m \approx 20$, but between $m = 100$ and 1000 there is at most an 11% difference in $T(\epsilon, m)$, while between $m = 1000$ and $m = 10000$ the decrease amounts to at most 3%. The good qualitative match between the iteration complexity predicted by our theory and the actual performance of the methods is clear.

5.3.2.2 Metric Learning

Our second set of experiments were based on instances of the metric learning problem. For each $i, j = 1, \ldots, n$, we are given a vector $x_i \in \mathbb{R}^d$, and a measure $y_{ij} \geq 0$ of the similarity between the vectors x_i and x_j. (Here $y_{ij} = 0$ means that x_i and x_j are the same). The statistical goal is to learn a matrix θ—inducing a pseudo-norm via $\|x\|_\theta^2 := \langle x, \theta x \rangle$—such that $\langle (x_i - x_j), \theta(x_i - x_j) \rangle \approx y_{ij}$. One method for doing so is to minimize the objective

$$f(\theta) = \frac{1}{\binom{n}{2}} \sum_{i<j} \left| \operatorname{tr}\left(\theta(x_i - x_j)(x_i - x_j)^\top\right) - y_{ij} \right| \quad \text{subject to} \quad \operatorname{tr}(\theta) \leq C, \ \theta \succeq 0.$$

The stochastic oracle for this problem is simple: given a query matrix θ, the oracle chooses a pair (i, j) uniformly at random, then returns the subgradient

$$\operatorname{sign}\left[\langle (x_i - x_j), \theta(x_i - x_j) \rangle - y_{ij}\right] (x_i - x_j)(x_i - x_j)^\top.$$

We solve ten random problems with dimension $d = 100$ and $n = 2000$, giving an objective with $4 \cdot 10^6$ terms and 5050 parameters. Performing stochastic optimization is more viable

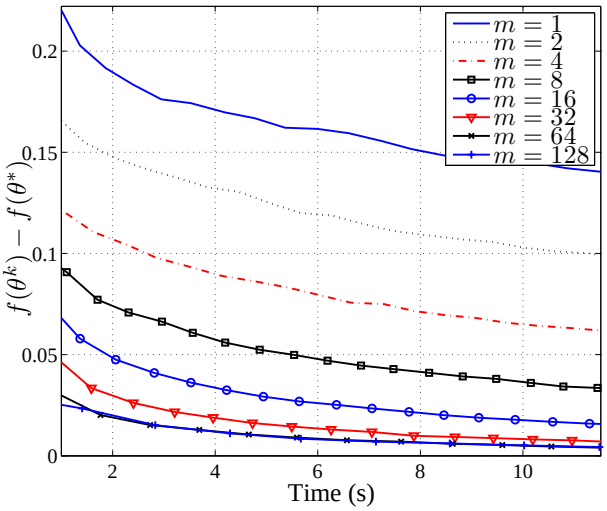

Figure 5.2. Optimization error $f(\theta^k) - \inf_{\theta^* \in \Theta} f(\theta^*)$ in the metric learning problem of Section 5.3.2.2 as a function of time in seconds. Each line indicates optimization error over time for a particular number of samples m in the gradient estimate (5.5); we set $m = 2^i$ for $i = \{1, \ldots, 7\}$.

for this problem than a non-stochastic method, as even computing the objective requires $\mathcal{O}(n^2 d^2)$ operations. We plot experimental results in Figure 5.2 showing the optimality gap $f(\theta^k) - \inf_{\theta^* \in \Theta} f(\theta^*)$ as a function of computation time. We plot several lines, each of which captures the performance of the algorithm using a different number m of samples in the smoothing step (5.5). As predicted by our theory and discussion in Section 5.3, receiving more samples m gives improvements in convergence rate as a function of time. Our theory also predicts that for $m \geq d$, there should be no improvement in actual time taken to minimize the objective; the plot in Figure 5.2 suggests that this too is correct, as the plots for $m = 64$ and $m = 128$ are essentially indistinguishable.

5.3.2.3 Necessity of randomized smoothing

A reasonable question is whether the extra sophistication of the random smoothing (5.5) is necessary. Can receiving more samples m from the stochastic oracle—all evaluated at the same point—give the same benefit to the simple dual averaging method (5.3)? We do not know the full answer to this question, though we give an experiment here that suggests that the answer is negative, in that smoothing does give demonstrable improvement.

For this experiment, we use the objective

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \|\theta - x_i\|_1 , \tag{5.12}$$

where the $x_i \in \{-1, +1\}^d$, and each component j of the vector x_i is sampled independently

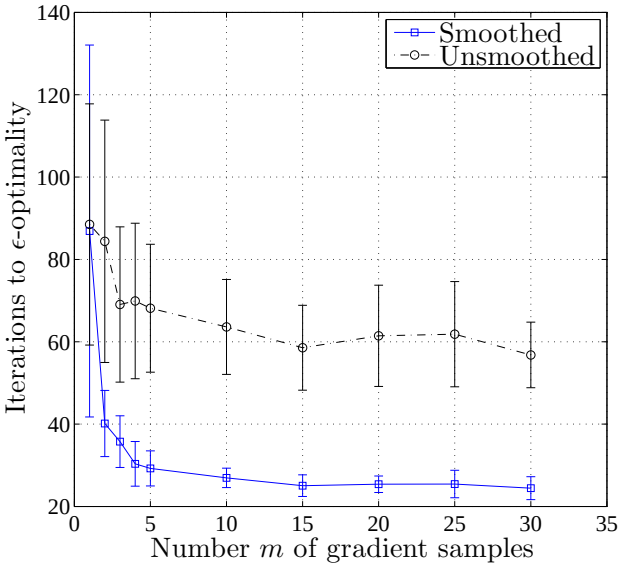

Figure 5.3. The number of iterations $T(\epsilon, m)$ to achieve an ϵ-optimal solution to (5.12) for simple mirror descent and the smoothed gradient method.

from $\{-1, 1\}$ and equal to 1 with probability $1/\sqrt{j}$. Even as $n \uparrow \infty$, the function f remains non-smooth, since the x_i belong to a discrete set and each value of x_i occurs with positive probability. As in Section 5.3.2.1, we compute $T(\epsilon, m)$ to be the number of iterations required to achieve an ϵ-optimal solution to the objective (5.12). We compare two algorithms that use m queries of the stochastic gradient oracle, which when queried at a point x chooses an index $i \in [n]$ uniformly at random and returns $\mathbf{g}(\theta; x_i) = \operatorname{sign}(\theta - x_i) \in \partial \|\theta - x_i\|_1$. The first algorithm is the dual averaging algorithm (5.3), where g^k is the average of m queries to the stochastic oracle at the current iterate θ^k. The second is the accelerated method (5.6a)–(5.6c) with the randomized averaging (5.5). We plot the results in Figure 5.3. We plot the best stepsize sequence α_k for the update (5.3) of several we tested to make comparison as favorable as possible for simple mirror descent. It is clear that while there is moderate improvement for the non-smooth method when the number of samples m grows, and both methods are (unsurprisingly) essentially indistinguishable for $m = 1$, the smoothed sampling strategy has much better iteration complexity as m grows.

5.4 Summary

In this chapter, we have developed and analyzed smoothing strategies for stochastic non-smooth optimization that are provably optimal in the stochastic oracle model of optimization complexity and given—to our knowledge—the first variance reduction techniques for non-smooth stochastic optimization. We think that at least two obvious questions remain. The first is whether the randomized smoothing is necessary to achieve such optimal rates of

convergence; it is clearly not when the data obeys certain nice characteristics, such as data sparsity, as outlined in the previous chapter. The second question is whether dimension-independent smoothing techniques are possible, that is, whether the d-dependent factors in the bounds in Corollaries 5.2–5.4 are necessary. Answering this question would require study of different smoothing distributions, as the dimension dependence for our choices of μ is tight. We have outlined several applications for which smoothing techniques give provable improvement over standard methods. Our experiments also show qualitatively good agreement with the theoretical predictions we have developed.

5.5 Proofs of convergence

In this section, we provide the proofs of Theorem 5.1 as well as Corollaries 5.1 through through 5.4. We begin with the proofs of the corollaries, after which we give the full proofs of the theorems. In both cases, we defer some of the more technical lemmas to appendices.

The general technique for the proof of each corollary is as follows. First, we note that the randomly smoothed function $f_u(\theta) = \mathbb{E}[f(\theta + uZ)]$ has Lipschitz continuous gradient, and it is uniformly close to the original non-smooth function f. This fact allows us to apply Theorem 5.1. The second step is to realize that with the sampling procedure (5.5), the variance $\mathbb{E}[\|e^k\|_*^2]$ decreases by a factor of approximately m, the number of gradient samples. Choosing the stepsizes appropriately in the theorems then completes the proofs. Proofs of these corollaries require relatively tight control of the smoothness properties of the smoothing convolution (5.2), so we refer frequently to lemmas stated in Section 5.6.

5.5.1 Proof of Corollaries 5.2 and 5.3

We begin by proving Corollary 5.2. Recall the averaged quantity $g^k = \frac{1}{m}\sum_{t=1}^m g_t^k$ defined in expression (5.5), and that $g_t^k \in \partial F(w^k + u_t Z_{k,t}; X_{k,t})$, where the random variables $Z_{k,t}$ are distributed uniformly on the ℓ_2-ball B_2. From Lemma 5.4 in Section 5.6, the variance of g_t as an estimate of $\nabla f_{u_t}(w^k)$ satisfies

$$\sigma^2 := \mathbb{E}\left[\left\|e^k\right\|_2^2\right] = \mathbb{E}\left[\left\|g^k - \nabla f_{u_k}(w^k)\right\|_2^2\right] \leq \frac{M^2}{m}. \tag{5.13}$$

Further, for Z distributed uniformly on B_2, we have the bound

$$f(\theta) \leq \mathbb{E}[f(\theta + uZ)] \leq f(\theta) + Mu,$$

and moreover, the function $\theta \mapsto \mathbb{E}[f(\theta + uZ)]$ has $M\sqrt{d}/u$-Lipschitz continuous gradient. Thus, applying Lemma 5.4 and Theorem 5.1 with the setting $L_k = M\sqrt{d}/u\nu_k$, we obtain

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{6M\mathsf{r}^2\sqrt{d}}{nu} + \frac{2\eta_n \mathsf{r}^2}{n} + \frac{1}{n}\sum_{k=0}^{n-1}\frac{1}{\eta_k}\cdot\frac{M^2}{m} + \frac{4Mu}{n},$$

where we have used the bound (5.13).

Recall that $\eta_k = M\sqrt{k+1}/\mathsf{r}\sqrt{m}$ by construction. Coupled with the inequality

$$\sum_{k=1}^{n} \frac{1}{\sqrt{k}} \leq 1 + \int_1^n \frac{1}{\sqrt{t}} dt = 1 + 2(\sqrt{n} - 1) \leq 2\sqrt{n}, \tag{5.14}$$

we use that $2\sqrt{n+1}/n + 2/\sqrt{n} \leq 5/\sqrt{n}$ to obtain

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{6M\mathsf{r}_2^2\sqrt{d}}{nu} + \frac{5M\mathsf{r}_2}{\sqrt{nm}} + \frac{4Mu}{n}.$$

Substituting the specified setting of $u = \mathsf{r}_2 d^{1/4}$ completes the proof.

The proof of Corollary 5.3 is essentially identical, differing only in the setting of $u = \mathsf{r}_2 d^{-1/4}$ and the application of Lemma 5.5 instead of Lemma 5.4 in Section 5.6.

5.5.2 Proof of Corollary 5.4

Under the conditions of the corollary, Lemma 5.3 implies that when μ is uniform on B_∞), then the function $f_u(\theta) := \mathbb{E}[f(\theta + uZ)]$ has M/u-Lipschitz continuous gradient with respect to the ℓ_1-norm, and moreover it satisfies the upper bound $f_u(\theta) \leq f(\theta) + \frac{Mdu}{2}$. Fix $\theta \in \Theta$ and let $g_t = \mathsf{g}(\theta + Z_t; X_t) \in \partial F(x + Z_t; X_t)$, with $g = \frac{1}{m}\sum_{t=1}^m g_t$. We claim that for any u, the error satisfies

$$\mathbb{E}\big[\|g - \nabla f_u(\theta)\|_\infty^2\big] \leq C\, \frac{M^2 \log d}{m} \tag{5.15}$$

for some universal constant C. Indeed, Lemma 5.3 shows that $\mathbb{E}[g] = \nabla f_u(\theta)$; moreover, component j of the random vector g_t is an unbiased estimator of the jth component of $\nabla f_u(\theta)$. Since $\|g_t\|_\infty \leq M$ and $\|\nabla f_u(\theta)\|_\infty \leq M$, the vector $g_t - \nabla f_u(\theta)$ is a d-dimensional random vector whose components are sub-Gaussian with sub-Gaussian parameter $4M^2$. Conditional on θ, the g_t are independent, so $g - \nabla f_u(\theta)$ has sub-Gaussian components with sub-Gaussian parameter at most $4M^2/m$ (cf. Buldygin and Kozachenko [36]). By standard concentration results [36], this immediately yields the claim (5.15).

Now, as in the proof of Corollary 5.2, we can apply Theorem 5.1. Recall as in our discussion following Corollary 3.3 that $\frac{1}{2(p-1)}\|\theta\|_p^2$ is strongly convex over $\mathbb{R}^d$ with respect to the ℓ_p-norm for $p \in (1, 2]$ (e.g. [134, 22]). Thus, with the choice $\psi(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$ for $p = 1 + 1/\log d$, it is clear that the squared radius r_ψ^2 of the set Θ is order $\|\theta^*\|_p^2 \log d \leq \|\theta^*\|_1^2 \log d$. All that remains is to relate the Lipschitz constant M with respect to the ℓ_1 norm to that for the ℓ_p norm. Let q be conjugate to p, that is, $1/q + 1/p = 1$. Under the assumptions of the theorem, we have $q = 1 + \log d$. For any $g \in \mathbb{R}^d$, we have $\|g\|_q \leq d^{1/q}\|g\|_\infty$. Of course, $d^{1/(\log d + 1)} \leq d^{1/(\log d)} = \exp(1)$, so $\|g\|_q \leq e\|g\|_\infty$.

Having shown that the Lipschitz constant L for the ℓ_p norm satisfies $L \leq Me$, where M is the Lipschitz constant with respect to the ℓ_1 norm, we apply Theorem 5.1 and the variance bound (5.15) to obtain the result. Specifically, Theorem 5.1 implies

$$\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \leq \frac{6M\mathsf{r}_\psi^2}{nu} + \frac{2\eta_n \mathsf{r}_\psi^2}{n} + \frac{C}{n}\sum_{k=0}^{n-1} \frac{1}{\eta_k} \cdot \frac{M^2 \log d}{m} + \frac{4Mdu}{2n}.$$

Substituting u, η_k, and $\mathsf{r}_\psi \leq \|\theta^*\|_1 \sqrt{\log d}$ and using bound (5.14) completes the proof.

5.5.3 Proof of Theorem 5.1

This proof is more involved than that of the above corollaries. In particular, we build on techniques used in the work of Tseng [170], Lan [114], and Xiao [183]. The changing smoothness of the stochastic objective—which comes from changing the shape parameter of the sampling distribution Z in the averaging step (5.5)—adds some challenge. The proof begins by defining $f_{u_k}(\theta) := \mathbb{E}[f(\theta + u_k Z)]$, where u_k is the non-increasing sequence of shape parameters in the averaging scheme (5.5). We show via Jensen's inequality that $f(\theta) \leq f_{u_k}(\theta) \leq f_{u_{k-1}}(\theta)$ for all k, which is intuitive because the variance of the sampling scheme is decreasing. Then we apply a suitable modification of the accelerated gradient method [170] to the sequence of functions f_{u_k} decreasing to f, and by allowing u_k to decrease appropriately we achieve our result. At the end of this section, we prove Corollary 5.1, which gives an alternative setting for u given a priori knowledge of the number of iterations.

We begin by stating two technical lemmas:

Lemma 5.1. *Let f_{u_k} be a sequence of functions such that f_{u_k} has L_k-Lipschitz continuous gradients with respect to the norm $\|\cdot\|$ and assume that $f_{u_k}(\theta) \leq f_{u_{k-1}}(\theta)$ for any $x \in \Theta$. Let the sequence $\{\theta^k, w^k, v^k\}$ be generated according to the updates* (5.6a)–(5.6c)*, and define the error term $e^k = \nabla f_{u_k}(w^k) - g^k$. Then for any $\theta^* \in \Theta$,*

$$\frac{1}{\nu_k^2}[f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1})] \leq \sum_{i=0}^{k} \frac{1}{\nu_i}[f_{u_i}(\theta^*) + \varphi(\theta^*)] + \left(L_{k+1} + \frac{\eta_{k+1}}{\nu_{k+1}}\right)\psi(\theta^*) + \sum_{i=0}^{k} \frac{1}{2\nu_i\eta_i}\left\|e^k\right\|_*^2 + \sum_{i=0}^{k} \frac{1}{\nu_i}\left\langle e^i, v^i - \theta^*\right\rangle.$$

See Section 5.5.4 for the proof of this claim.

Lemma 5.2. *Let the sequence ν_k satisfy $\frac{1-\nu_k}{\nu_k^2} = \frac{1}{\nu_{k-1}^2}$ and $\nu_0 = 1$. Then*

$$\nu_k \leq \frac{2}{k+2} \quad \textit{and} \quad \sum_{i=0}^{k} \frac{1}{\nu_i} = \frac{1}{\nu_k^2}.$$

Tseng [170] proves the second statement; the first follows by induction.

We now proceed with the proof. Recalling $f_{u_k}(\theta) = \mathbb{E}[f(\theta + u_k Z)]$, let us verify that $f_{u_k}(\theta) \leq f_{u_{k-1}}(\theta)$ for any x and t so we can apply Lemma 5.1. Since $u_k \leq u_{k-1}$, we may define a random variable $U \in \{0, 1\}$ such that $\mathbb{P}(U = 1) = \frac{u_k}{u_{k-1}} \in [0, 1]$. Then

$$\begin{aligned} f_{u_k}(\theta) = \mathbb{E}[f(\theta + u_k Z)] &= \mathbb{E}\big[f\big(\theta + u_{k-1} Z \mathbb{E}[U]\big)\big] \\ &\leq \mathbb{P}[U = 1]\ \mathbb{E}[f(\theta + u_{k-1} Z)] + \mathbb{P}[U = 0]\ f(\theta), \end{aligned}$$

where the inequality follows from Jensen's inequality. By a second application of Jensen's inequality, we have $f(\theta) = f(\theta + u_{k-1}\mathbb{E}[Z]) \leq \mathbb{E}[f(\theta + u_{k-1} Z)] = f_{u_{k-1}}(\theta)$. Combined with the previous inequality, we conclude that $f_{u_k}(\theta) \leq \mathbb{E}[f(\theta + u_{k-1} Z)] = f_{u_{k-1}}(\theta)$ as claimed. Consequently, we have verified that the function f_{u_k} satisfies the assumptions of Lemma 5.1, where ∇f_{u_k} has Lipschitz parameter $L_k = L/u_k$ and error term $e^k = \nabla f_{u_k}(w^k) - g^k$. We apply the lemma momentarily.

Using Assumption 5A that $f(\theta) \geq \mathbb{E}[f(\theta + u_k Z)] - M u_k = f_{u_k}(\theta) - M u_k$ for all $x \in \Theta$, Lemma 5.2 implies

$$\begin{aligned} &\frac{1}{\nu_{n-1}^2}[f(\theta^n) + \varphi(\theta^n)] - \frac{1}{\nu_{n-1}^2}[f(\theta^*) + \varphi(\theta^*)] \\ &\quad = \frac{1}{\nu_{n-1}^2}[f(\theta^n) + \varphi(\theta^n)] - \sum_{k=0}^{n-1} \frac{1}{\nu_k}[f(\theta^*) + \varphi(\theta^*)] \\ &\quad \leq \frac{1}{\nu_{n-1}^2}[f_{u_{n-1}}(\theta^n) + \varphi(\theta^n)] - \sum_{k=0}^{n-1} \frac{1}{\nu_k}[f_{u_k}(\theta^*) + \varphi(\theta^*)] + \sum_{k=0}^{n-1} \frac{M u_k}{\nu_k}, \end{aligned}$$

which by the definition of $u_k = \nu_k u$ is in turn bounded by

$$\frac{1}{\nu_{n-1}^2}[f_{u_{n-1}}(\theta^n) + \varphi(\theta^n)] - \sum_{k=0}^{n-1} \frac{1}{\nu_k}[f_{u_k}(\theta^*) + \varphi(\theta^*)] + nMu. \tag{5.16}$$

Now we apply Lemma 5.1 to the bound (5.16), which gives us

$$\begin{aligned} &\frac{1}{\nu_{n-1}^2}\left[f(\theta^n) + \varphi(\theta^n) - f(\theta^*) - \varphi(\theta^*)\right] \\ &\leq L_n \psi(\theta^*) + \frac{\eta_n}{\nu_n}\psi(\theta^*) + \sum_{k=0}^{n-1} \frac{1}{2\nu_k \eta_k} \left\|e^k\right\|_*^2 + \sum_{k=0}^{n-1} \frac{1}{\nu_k} \left\langle e^k, v^k - \theta^* \right\rangle + nMu. \end{aligned} \tag{5.17}$$

The non-probabilistic bound (5.17) is the key to the remainder of this proof, as well as the high probability guarantees presented in the paper off of which this chapter is based [55]. What remains here is to take expectations in the bound (5.17).

Recall the filtration of σ-fields $\mathcal{F}_k$, which satisfy $\theta^k, w^k, v^k \in \mathcal{F}_{k-1}$, that is, $\mathcal{F}_k$ contains the randomness in the stochastic oracle to time t. Since g^k is an unbiased estimator of $\nabla f_{u_k}(w^k)$ by construction, we have $\mathbb{E}[g^k \mid \mathcal{F}_{k-1}] = \nabla f_{u_k}(w^k)$ and

$$\mathbb{E}[\langle e^k, v^k - \theta^* \rangle] = \mathbb{E}\big[\mathbb{E}[\langle e^k, v^k - \theta^* \rangle \mid \mathcal{F}_{k-1}]\big] \; = \; \mathbb{E}\big[\, \langle \mathbb{E}[e^k \mid \mathcal{F}_{k-1}], v^k - \theta^* \rangle \,\big] \; = \; 0,$$

where we have used the fact that v^k are measurable with respect to $\mathcal{F}_{k-1}$. Now, recall from Lemma 5.2 that $\nu_k \leq \frac{2}{2+k}$ and that $(1-\nu_k)/\nu_k^2 = 1/\nu_{k-1}^2$. Thus

$$\frac{\nu_{k-1}^2}{\nu_k^2} = \frac{1}{1-\nu_k} \leq \frac{1}{1-\frac{2}{2+k}} = \frac{2+k}{k} \leq \frac{3}{2} \qquad \text{for } k \geq 4.$$

Furthermore, we have $\nu_{k+1} \leq \nu_k$, so by multiplying both sides of our bound (5.17) by ν_{n-1}^2 and taking expectations over the random vectors g^k,

$$\begin{aligned}
&\mathbb{E}[f(\theta^n) + \varphi(\theta^n)] - [f(\theta^*) + \varphi(\theta^*)] \\
&\leq \nu_{n-1}^2 L_n \psi(\theta^*) + \nu_{n-1}\eta_n \psi(\theta^*) + \nu_{n-1}^2 nMu \\
&\qquad + \nu_{n-1} \sum_{k=0}^{n-1} \frac{1}{2\eta_k} \mathbb{E}[\left\|e^k\right\|_*^2] + \nu_{n-1} \sum_{k=0}^{n-1} \mathbb{E}[\langle e^k, v^k - \theta^* \rangle] \\
&\leq \frac{6L\psi(\theta^*)}{nu} + \frac{2\eta_n \psi(\theta^*)}{n} + \frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{\eta_k} \mathbb{E}[\left\|e^k\right\|_*^2] + \frac{4Mu}{n},
\end{aligned}$$

where we used that $L_n = L/u_n = L/\nu_n u$. This completes the proof of Theorem 5.1. □

We conclude this section by proving Corollary 5.1, which uses a fixed setting of the smoothing parameter u_k. It is clear that by setting $u \propto 1/n$, the rates achieved by Theorem 5.1 and Corollary 5.1 are identical to constant factors. If we fix $u_k \equiv u$ for all k, then the bound (5.17) holds with the last term nMu replaced by $\nu_{n-1}^2 Mu$, which we see by invoking Lemma 5.2. The remainder of the proof follows unchanged, with $L_k \equiv L$ for all k.

5.5.4 Proof of Lemma 5.1

Define the linearized version of the cumulative objective

$$\mathsf{L}_k(\theta) := \sum_{i=0}^{k} \frac{1}{\nu_i} [f_{u_i}(w^i) + \langle g^i, \theta - w^i \rangle + \varphi(\theta)], \tag{5.18}$$

and let $\mathsf{L}_{-1}(z)$ denote the indicator function of the set Θ. For conciseness, we temporarily adopt the shorthand notation

$$\alpha_k^{-1} = L_k + \eta_k/\nu_k \qquad \text{and} \qquad \phi_k(\theta) = f_{u_k}(\theta) + \varphi(\theta).$$

By the smoothness of f_{u_k}, we have

$$\underbrace{f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1})}_{\phi_k(\theta^{k+1})} \leq f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), \theta^{k+1} - w^k \right\rangle + \frac{L_k}{2} \left\| \theta^{k+1} - w^k \right\|^2 + \varphi(\theta^{k+1}).$$

From the definition (5.6a)–(5.6c) of the triple (θ^k, w^k, v^k), we obtain

$$\begin{aligned}\phi_k(\theta^{k+1}) \leq f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), \nu_k v^{k+1} + (1-\nu_k)\theta^k \right\rangle + \frac{L_k}{2} \left\| \nu_k v^{k+1} - \nu_k v^k \right\|^2 \\ + \varphi(\nu_k v^{k+1} + (1-\nu_k)\theta^k).\end{aligned}$$

Finally, by convexity of the regularizer φ, we conclude

$$\begin{aligned}\phi_k(\theta^{k+1}) \leq \nu_k \left[f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), v^{k+1} - w^k \right\rangle + \varphi(v^{k+1}) + \frac{L_k \nu_k}{2} \left\| v^{k+1} - v^k \right\|^2 \right] \\ + (1-\nu_k)[f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), \theta^k - w^k \right\rangle + \varphi(\theta^k)]. \qquad (5.19)\end{aligned}$$

By the strong convexity of ψ, we know that we have the lower bound

$$D_\psi(\theta, \theta') = \psi(\theta) - \psi(\theta') - \langle \nabla\psi(\theta'), \theta - \theta' \rangle \;\geq\; \frac{1}{2} \left\| \theta - \theta' \right\|^2. \qquad (5.20)$$

On the other hand, by the convexity of f_{u_k}, we have

$$f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), \theta^k - w^k \right\rangle \leq f_{u_k}(\theta^k). \qquad (5.21)$$

Substituting inequalities (5.20) and (5.21) into the bound (5.19) and simplifying yields

$$\begin{aligned}\phi_k(\theta^{k+1}) \leq \nu_k \left[f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), v^{k+1} - w^k \right\rangle + \varphi(v^{k+1}) + L_k \nu_k D_\psi(v^{k+1}, v^k) \right] \\ + (1-\nu_k)[f_{u_k}(\theta^k) + \varphi(\theta^k)].\end{aligned}$$

We now re-write this upper bound in terms of the error $e^k = \nabla f_{u_k}(w^k) - g^k$:

$$\begin{aligned}\phi_k(\theta^{k+1}) &\leq \nu_k \left[f_{u_k}(w^k) + \left\langle g^k, v^{k+1} - w^k \right\rangle + \varphi(v^{k+1}) + L_k \nu_k D_\psi(v^{k+1}, v^k) \right] \\ &\quad + (1-\nu_k)[f_{u_k}(\theta^k) + \varphi(\theta^k)] + \nu_k \left\langle e^k, v^{k+1} - w^k \right\rangle \\ &= \nu_k^2 \left[\mathsf{L}_k(v^{k+1}) - \mathsf{L}_{t-1}(v^{k+1}) + L_k D_\psi(v^{k+1}, v^k) \right] \\ &\quad + (1-\nu_k)[f_{u_k}(\theta^k) + \varphi(\theta^k)] + \nu_k \left\langle e^k, v^{k+1} - w^k \right\rangle. \qquad (5.22)\end{aligned}$$

The first order convexity conditions for optimality imply that for some $g \in \partial \mathsf{L}_{k-1}(v^k)$ and all $\theta \in \Theta$, we have $\left\langle g + \frac{1}{\alpha_k} \nabla\psi(v^k), \theta - v^k \right\rangle \geq 0$ since v^k minimizes $\mathsf{L}_{k-1}(\theta) + \frac{1}{\alpha_k}\psi(\theta)$. Thus, first-order convexity gives

$$\begin{aligned}\mathsf{L}_{k-1}(\theta) - \mathsf{L}_{k-1}(v^k) \geq \left\langle g, \theta - v^k \right\rangle &\geq -\frac{1}{\alpha_k} \left\langle \nabla\psi(v^k), \theta - v^k \right\rangle \\ &= \frac{1}{\alpha_k}\psi(v^k) - \frac{1}{\alpha_k}\psi(\theta) + \frac{1}{\alpha_k} D_\psi(\theta, v^k).\end{aligned}$$

Adding $\mathsf{L}_k(v^{k+1})$ to both sides of the above and substituting $\theta = v^{k+1}$, we conclude

$$\mathsf{L}_k(v^{k+1}) - \mathsf{L}_{k-1}(v^{k+1}) \leq \mathsf{L}_k(v^{k+1}) - \mathsf{L}_{k-1}(v^k) - \frac{1}{\alpha_k}\psi(v^k) + \frac{1}{\alpha_k}\psi(v^{k+1}) - \frac{1}{\alpha_k}D_\psi(v^{k+1}, v^k).$$

Combining this inequality with the bound (5.22) and the definition $\alpha_k^{-1} = L_k + \eta_k/\nu_k$,

$$\begin{aligned} f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1}) &\leq \nu_k^2 \left[\mathsf{L}_k(v^{k+1}) - \mathsf{L}_k(v^k) - \frac{1}{\alpha_k}\psi(v^k) + \frac{1}{\alpha_k}\psi(v^{k+1}) - \frac{\eta_k}{\nu_k}D_\psi(v^{k+1}, v^k)\right] \\ &\quad + (1-\nu_k)[f_{u_k}(\theta^k) + \varphi(\theta^k)] + \nu_k \left\langle e^k, v^{k+1} - w^k \right\rangle \\ &\leq \nu_k^2 \left[\mathsf{L}_k(v^{k+1}) - \mathsf{L}_k(v^k) - \frac{1}{\alpha_k}\psi(v^k) + \frac{1}{\alpha_{k+1}}\psi(v^{k+1}) - \frac{\eta_k}{\nu_k}D_\psi(v^{k+1}, v^k)\right] \\ &\quad + (1-\nu_k)[f_{u_k}(\theta^k) + \varphi(\theta^k)] + \nu_k \left\langle e^k, v^{k+1} - w^k \right\rangle \end{aligned}$$

since α_k^{-1} is non-decreasing. We now divide both sides by ν_k^2, and unwrap the recursion. By construction $(1-\nu_k)/\nu_k^2 = 1/\nu_{k-1}^2$ and $f_{u_k} \leq f_{u_{k-1}}$, so we obtain

$$\begin{aligned} \frac{1}{\nu_k^2}[f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1})] &\leq \frac{1-\nu_k}{\nu_k^2}[f_{u_k}(\theta^k) + \varphi(\theta^k)] - \frac{1}{\alpha_k}\psi(v^k) + \frac{1}{\alpha_{k+1}}\psi(v^{k+1}) \\ &\quad + \mathsf{L}_k(v^{k+1}) - \mathsf{L}_k(v^k) - \frac{\eta_k}{\nu_k}D_\psi(v^{k+1}, v^k) + \frac{1}{\nu_k}\left\langle e_k, v^{k+1} - w^k \right\rangle \\ &\leq \frac{1}{\nu_{k-1}^2}[f_{u_{k-1}}(\theta^k) + \varphi(\theta^k)] - \frac{1}{\alpha_k}\psi(v^k) + \frac{1}{\alpha_{k+1}}\psi(v^{k+1}) \\ &\quad + \mathsf{L}_k(v^{k+1}) - \mathsf{L}_k(v^k) - \frac{\eta_k}{\nu_k}D_\psi(v^{k+1}, v^k) + \frac{1}{\nu_k}\left\langle e_k, v^{k+1} - w^k \right\rangle. \end{aligned}$$

The second inequality follows by combination of the facts that $(1-\nu_k)/\nu_k^2 = 1/\nu_{k-1}^2$ and $f_{u_k} \leq f_{u_{k-1}}$. By applying the two steps above successively to $[f_{u_{k-1}}(\theta^k) + \varphi(\theta^k)]/\nu_{k-1}^2$, then to $[f_{u_{k-2}}(\theta^{k-1}) + \varphi(\theta^{k-1})]/\nu_{k-2}^2$, and so on until $k = 0$, we find

$$\begin{aligned} \frac{1}{\nu_k^2}[f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1})] &\leq \frac{1-\nu_0}{\nu_0^2}[f_{u_0}(\theta^0) + \varphi(\theta^0)] + \mathsf{L}_k(v^{k+1}) + \frac{1}{\alpha_{k+1}}\psi(v^{k+1}) \\ &\quad - \sum_{i=0}^{k}\frac{\eta_i}{\nu_i}D_\psi(v^{i+1}, v^i) + \sum_{i=0}^{k}\frac{1}{\nu_i}\left\langle e^i, v^{i+1} - w^i \right\rangle - \mathsf{L}_{-1}(v^0) - \frac{1}{\alpha_0}\psi(v^0) \end{aligned}$$

By construction, $\nu_0 = 1$, we have $\mathsf{L}_{-1}(v^0) = 0$, and v^{k+1} minimizes $\mathsf{L}_k(\theta) + \frac{1}{\alpha_{k+1}}\psi(\theta)$ over Θ. Thus, for any $\theta^* \in \Theta$, we have

$$\begin{aligned} &\frac{1}{\nu_k^2}[f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1})] \\ &\leq \mathsf{L}_k(\theta^*) + \frac{1}{\alpha_{k+1}}\psi(\theta^*) - \sum_{i=0}^{k}\frac{\eta_i}{\nu_i}D_\psi(v^{i+1}, v^i) + \sum_{i=0}^{k}\frac{1}{\nu_i}\left\langle e^i, v^{i+1} - w^i \right\rangle. \end{aligned}$$

Recalling the definition (5.18) of L_k and noting that the first-order conditions for convexity imply that $f_{u_k}(w^k) + \left\langle \nabla f_{u_k}(w^k), x - w^k \right\rangle \le f_{u_k}(\theta)$, we expand L_k and have

$$\begin{aligned}
\frac{1}{\nu_k^2}[f_{u_k}(\theta^{k+1}) + \varphi(\theta^{k+1})] &\le \sum_{i=0}^{k} \frac{1}{\nu_i}[f_{u_i}(w^i) + \left\langle g^i, \theta^* - w^i \right\rangle + \varphi(\theta^*)] + \frac{1}{\alpha_{k+1}}\psi(\theta^*) \\
&\quad - \sum_{i=0}^{k} \frac{\eta_i}{\nu_i} D_\psi(v^{i+1}, v^i) + \sum_{i=0}^{k} \frac{1}{\nu_i} \left\langle e^i, v^{i+1} - w^k \right\rangle \\
&= \sum_{i=0}^{k} \frac{1}{\nu_i}[f_{u_i}(w^i) + \left\langle \nabla f_{u_i}(w^i), \theta^* - w^i \right\rangle + \varphi(\theta^*)] + \frac{1}{\alpha_{k+1}}\psi(\theta^*) \\
&\quad - \sum_{i=0}^{k} \frac{\eta_i}{\nu_i} D_\psi(v^{i+1}, v^i) + \sum_{i=0}^{k} \frac{1}{\nu_i} \left\langle e^i, v^{i+1} - \theta^* \right\rangle \\
&\le \sum_{i=0}^{k} \frac{1}{\nu_i}[f_{u_i}(\theta^*) + \varphi(\theta^*)] + \frac{1}{\alpha_{k+1}}\psi(\theta^*) \\
&\quad - \sum_{i=0}^{k} \frac{\eta_i}{\nu_i} D_\psi(v^{i+1}, v^i) + \sum_{i=0}^{k} \frac{1}{\nu_i} \left\langle e^i, v^{i+1} - \theta^* \right\rangle . \qquad (5.23)
\end{aligned}$$

Now we apply the Fenchel-Young inequality to the conjugates $\frac{1}{2}\|\cdot\|^2$ and $\frac{1}{2}\|\cdot\|_*^2$, yielding

$$\begin{aligned}
\left\langle e^k, v^{k+1} - \theta^* \right\rangle &= \left\langle e^k, v^k - \theta^* \right\rangle + \left\langle e^k, v^{k+1} - v^k \right\rangle \\
&\le \left\langle e^k, v^k - \theta^* \right\rangle + \frac{1}{2\eta_k} \left\| e^k \right\|_*^2 + \frac{\eta_k}{2} \left\| v^k - v^{k+1} \right\|^2 .
\end{aligned}$$

In particular,

$$-\frac{\eta_k}{\nu_k} D_\psi(v^{k+1}, v^k) + \frac{1}{\nu_k} \left\langle e^k, v^{k+1} - \theta^* \right\rangle \le \frac{1}{2\eta_k \nu_k} \left\| e^k \right\|_*^2 + \frac{1}{\nu_k} \left\langle e^k, v^k - \theta^* \right\rangle .$$

Using this inequality and rearranging (5.23) proves the lemma.

5.6 Properties of randomized smoothing

In this section, we discuss the analytic properties of the smoothed function f_u from the convolution (5.2). We assume throughout that functions are sufficiently integrable without bothering with measurability conditions (since $F(\cdot; x)$ is convex, this is no real loss of generality [25, 153, 154]). By Fubini's theorem, we have

$$f_u(\theta) = \int_{\mathcal{X}} \int_{\mathbb{R}^d} F(\theta + uz; x)\mu(y)dydP(x) = \int_{\mathcal{X}} F_u(\theta; x)dP(x).$$

Here $F_u(\theta; x) = (F(\cdot; x) * \mu(-\cdot))(\theta)$. We begin with the observation that since μ is a density with respect to Lebesgue measure, the function f_u is in fact differentiable [25]. So we have already made our problem somewhat smoother, as it is now differentiable; for the remainder, we consider finer properties of the smoothing operation. In particular, we will show that under suitable conditions on μ, $F(\cdot; x)$, and P, the function f_u is uniformly close to f over Θ and ∇f_u is Lipschitz continuous.

The next lemmas apply to general (possibly non-smooth) convex functions f, where we let

$$f_u(x) = \mathbb{E}[f(x + uZ)] = \int_{\mathbb{R}^d} f(x + uz)\mu(z)dz$$

denote the function f smoothed by the scaled distribution μ. Because f is almost-everywhere differentiable [98], we may without loss of generality compute $\nabla f(x + uZ)$ whenever Z has a density (see also Bertsekas [25] and Rockafellar and Wets [153]). We give proofs of the lemmas in the subsections to follow, for each we use the notation $B_p = \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$ to denote ℓ_p-ball of radius 1 and $B_p(x, u) = \{x + y \in \mathbb{R}^d : \|y\|_p \le u\}$ to denote the ℓ_p-ball of radius u centered at x.

Lemma 5.3. *Let μ be the uniform density on the ℓ_∞-ball B_∞. Assume that f is convex and M-Lipschitz with respect to the ℓ_1-norm on* $\operatorname{int}(\operatorname{dom} f + uB_\infty)$. *Then*

(i) $f(x) \le f_u(x) \le f(x) + \frac{Md}{2}u$

(ii) f_u *is M-Lipschitz with respect to the ℓ_1-norm over* $\operatorname{dom} f$.

(iii) f_u *is continuously differentiable; moreover, its gradient is $\frac{M}{u}$-Lipschitz continuous with respect to the ℓ_1-norm.*

There exists a function f for which each of the estimates (i)–(iii) are tight simultaneously.

A similar lemma can be proved when μ is the density of the uniform distribution on B_2. In this case, Yousefian et al. give (i)–(iii) of the following lemma [187] (though the tightness of the bounds is new).

Lemma 5.4 (Yousefian, Nedić, Shanbhag)**.** *Let f_u be defined as in* (5.2) *where μ is the uniform density on the ℓ_2-ball B_2. Assume that f is convex and M-Lipschitz with respect to the ℓ_2-norm on* $\operatorname{int}(\operatorname{dom} f + uB_2)$. *Then*

(i) $f(x) \le f_u(x) \le f(x) + Mu$

(ii) f_u *is M-Lipschitz over* $\operatorname{dom} f$.

(iii) f_u *is continuously differentiable; moreover, its gradient is $\frac{M\sqrt{d}}{u}$-Lipschitz continuous.*

In addition, there exists a function f for which each of the bounds (i)–(iii) is tight—cannot be improved by more than a constant factor—simultaneously.

For situations in which f is M-Lipschitz with respect to the ℓ_2-norm over all of $\mathbb{R}^d$ and for, we can use the normal distribution to perform smoothing. The following lemma is similar to a result of Lakshmanan and de Farias [113, Lemma 3.3], but they consider functions Lipschitz-continuous with respect to the ℓ_∞-norm, i.e. $|f(x) - f(y)| \leq L \|x - y\|_\infty$, which is too stringent for our purposes, and we carefully quantify the dependence on the dimension of the underlying problem.

Lemma 5.5. *Let μ be the $\mathsf{N}(0, u^2 I_{d\times d})$ distribution. Assume that f is M-Lipschitz with respect to the ℓ_2-norm. The following properties hold:*

(i) $f(x) \leq f_u(x) \leq f(x) + Mu\sqrt{d}$

(ii) *f_u is M-Lipschitz with respect to the ℓ_2 norm*

(iii) *f_u is continuously differentiable; moreover, its gradient is $\frac{M}{u}$-Lipschitz continuous with respect to the ℓ_2-norm.*

In addition, there exists a function f for which each of the bounds (i)–(iii) is tight (to within a constant factor) simultaneously.

Our final lemma illustrates the sharpness of the bounds we have proved for functions that are Lipschitz with respect to the ℓ_2-norm. Specifically, we show that at least for the normal and uniform distributions, it is impossible to obtain more favorable tradeoffs between the uniform approximation error of the smoothed function f_u and the Lipschitz continuity of ∇f_u. We begin with the following definition of our two types of error (uniform and gradient), then give the lemma:

$$E_U(f) := \inf\left\{L \in \mathbb{R} \mid \sup_{x \in \operatorname{dom} f} |f(x) - f_u(x)| \leq L\right\} \tag{5.24}$$

$$E_\nabla(f) := \inf\left\{L \in \mathbb{R} \mid \|\nabla f_u(x) - \nabla f_u(y)\|_2 \leq L \|x - y\|_2 \ \forall\, x, y \in \operatorname{dom} f\right\} \tag{5.25}$$

Lemma 5.6. *There exists a universal (numerical) constant $c > 0$ such that the following holds. If μ equal to either the uniform distribution on uB_2 or $\mathsf{N}(0, u^2 I_{d\times d})$, there exists an M-Lipschitz continuous function f such that*

$$E_U(f) E_\nabla(f) \geq cM^2\sqrt{d}.$$

Remarks Inspecting the convergence guarantee of Theorem 5.1 makes the importance of the above bound clear. The terms L and M in the bound (5.7) can be replaced with $E_\nabla(f)$ and $E_U(f)$, respectively. Minimizing over u, we see that the leading term in the convergence guarantee (5.7) is of order $\frac{\sqrt{E_\nabla(f) E_U(f) \psi(x^*)}}{n} \geq \frac{cMd^{1/4}\sqrt{\psi(\theta^*)}}{n}$. In particular, this result shows that our analysis of the dimension dependence of the randomized smoothing in Lemmas 5.4 and 5.5 is sharp and cannot be improved by more than a constant factor (see also Corollaries 5.2 and 5.3).

5.6.1 Proofs of smoothing lemmas

The following technical lemma is a building block for our results; we provide a proof in Sec. 5.6.1.4.

Lemma 5.7. *Let f be convex and M-Lipschitz continuous with respect to a norm $\|\cdot\|$ over the domain* $\operatorname{supp}\mu + \operatorname{dom} f$. *Let Z be distributed according to the distribution μ. Then*

$$\|\nabla f_u(x) - \nabla f_u(y)\|_* = \|\mathbb{E}\left[\nabla f(x+Z) + \nabla f(y+Z)\right]\|_* \leq M \int |\mu(z-x) - \mu(z-y))| dz. \tag{5.26}$$

If the norm $\|\cdot\|$ is the ℓ_2-norm and the density $\mu(z)$ is rotationally symmetric and non-increasing as a function of $\|z\|_2$, the bound (5.26) *holds with equality for the function*

$$f(x) = M \left| \left\langle \frac{y}{\|y\|_2}, x \right\rangle - \frac{1}{2} \right|.$$

5.6.1.1 Proof of Lemma 5.3

To simplify notation, we redefine $Z \sim \mu$ so that μ is the uniform density on $B_\infty(0,u)$. Let $h_u(x)$ denote the (shifted) Huber loss

$$h_u(x) = \begin{cases} \frac{x^2}{2u} + \frac{u}{2} & \text{for } x \in [-u,u] \\ |x| & \text{otherwise.} \end{cases} \tag{5.27}$$

Now we prove each of the parts of the lemma in turn.

(i) Since $\mathbb{E}[Z] = 0$, Jensen's inequality shows $f(x) = f(x + u\mathbb{E}[Z]) \leq \mathbb{E}[f(x+uZ)] = f_u(x)$, by definition of f_u. To get the upper uniform bound, note that by assumption, f is M-Lipschitz continuous over $\operatorname{dom} f + uB_\infty$ with respect to the ℓ_1-norm, so

$$f_u(x) = \mathbb{E}[f(x+uZ)] \leq \mathbb{E}[f(x)] + uM\mathbb{E}[\|Z\|_1] = f(x) + \frac{dMu}{2}.$$

To see that the estimate is tight, note that for $f(x) = \|x\|_1$, we have $f_u(x) = \sum_{j=1}^d h_u(x_j)$, where h_u is the shifted Huber loss (5.27), and $f_u(0) = du/2$, while $f(0) = 0$.

(ii) We now prove that f_u is M-Lipschitz with respect to $\|\cdot\|_1$. Under the stated conditions, we have $\|\partial f(x)\|_\infty \leq M$ for all $x \in \operatorname{dom} f + \operatorname{supp}\mu$, whence

$$\|\nabla f_u(x)\|_\infty = \|\mathbb{E}[\nabla f(x+Z)]\|_\infty \leq \mathbb{E}[\|\nabla f(x+Z)\|_\infty] \leq M.$$

Tightness follows again by considering $f(x) = \|x\|_1$, where $M = 1$.

(iii) Recall that differentiability is directly implied by earlier work of Bertsekas [25]. Since f is a.e.-differentiable, we have $\nabla f_u(x) = \mathbb{E}[\nabla f(x+Z)]$ for Z uniform on $[-u,u]^d$. We now establish Lipschitz continuity of $\nabla f_u(x)$.

For a fixed pair $x, y \in \operatorname{dom} f + B_\infty(0,u)$, we have from Lemma 5.7

$$\|\mathbb{E}[\nabla f(x+Z)] - \mathbb{E}[\nabla f(y+Z)]\|_\infty \leq M \cdot \frac{1}{(2u)^d} \lambda\big(B_\infty(x,u) \Delta B_\infty(y,u)\big),$$

where λ denotes Lebesgue measure and Δ denotes the symmetric set-difference. By a straightforward geometric calculation, we see that

$$\lambda\left(B_\infty(x,u)\Delta B_\infty(y,u)\right) = 2\left((2u)^d - \prod_{j=1}^d \left[2u - |x_j - y_j|\right]_+\right). \tag{5.28}$$

To control the volume term (5.28) and complete the proof, we need an auxiliary lemma (which we prove at the end of this subsection).

Lemma 5.8. *Let $a \in \mathbb{R}^d_+$ and $u \in \mathbb{R}_+$. Then $\prod_{j=1}^d [u - a_j]_+ \geq u^d - \|a\|_1 u^{d-1}$.*

The volume (5.28) is easy to control using Lemma 5.8. Indeed, we have

$$\frac{1}{2}\lambda\left(B_\infty(x,u)\Delta B_\infty(y,u)\right) \leq (2u)^d - (2u)^d + \|x-y\|_1 (2u)^{d-1},$$

which implies the desired result, that is, that

$$\|\mathbb{E}[\nabla f(x+Z)] - \mathbb{E}\nabla[f(y+Z)]\|_\infty \leq \frac{M\|x-y\|_1}{u}.$$

To see the tightness claimed in the proposition, consider as usual $f(x) = \|x\|_1$ and let e_j denote the jth standard basis vector. Then $M = 1$, $\nabla f_u(0) = 0$, $\nabla f_u(ue_j) = e_j$, and $\|\nabla f_u(0) - \nabla f_u(ue_j)\|_\infty = 1 = \frac{M}{u}\|0 - ue_j\|_1$.

Proof of Lemma 5.8 We begin by noting that the statement of the lemma trivially holds whenever $\|a\|_1 \geq u$, as the right hand side of the inequality is then non-positive. Now, fix some $c < u$, and consider the problem

$$\min_a \prod_{j=1}^d (u - a_i)_+ \quad \text{s.t.} \quad a \succeq 0, \|a\|_1 \leq c. \tag{5.29}$$

We show that the minimum is achieved when one index is set to $a_i = c$ and the rest to 0. Indeed, suppose for the sake of contradiction that $\tilde{a}$ is the solution to (5.29) but that there are indices i, j with $a_i \geq a_j > 0$, that is, at least two non-zero indices. By taking a logarithm,

it is clear that minimizing the objective (5.29) is equivalent to minimizing $\sum_{j=1}^{d} \log(u - a_j)$. Taking the derivative of $a \mapsto \log(u - a)$ for a_i and a_j, we see that

$$\frac{\partial}{\partial a_i} \log(u - a_i) = \frac{-1}{u - a_i} \leq \frac{-1}{u - a_j} = \frac{\partial}{\partial a_j} \log(u - a_j).$$

Since $\frac{-1}{u-a}$ is decreasing function of a, increasing a_i slightly and decreasing a_j slightly causes $\log(u - a_i)$ to decrease faster than $\log(u - a_j)$ increases, thus decreasing the overall objective. This is the desired contradiction. □

5.6.1.2 Proof of Lemma 5.5

Throughout this proof, we use Z to denote a random variable distributed as $\mathsf{N}(0, u^2 I_{d\times d})$.

(i) As in the previous lemma, Jensen's inequality gives $f(x) = f(x + \mathbb{E}Z) \leq \mathbb{E}f(x + Z) = f_u(x)$. By assumption, f is M-Lipschitz, so

$$f_u(x) = \mathbb{E}[f(x + Z)] \leq \mathbb{E}[f(x)] + M\mathbb{E}[\|Z\|_2] \leq f(x) + M\sqrt{\mathbb{E}[\|Z\|_2^2]} = f(x) + Mu\sqrt{d}.$$

(ii) This proof is analogous to that of part (ii) of Lemma 5.3. The tightness of the Lipschitz constant can be verified by taking $f(x) = \langle v, x\rangle$ for $v \in \mathbb{R}^d$, in which case $f_u(x) = f(x)$, and both have gradient v.

(iii) Now we show that ∇f_u is Lipschitz continuous. Indeed, applying Lemma 5.7 we have

$$\|\nabla f_u(x) - \nabla f_u(y)\|_2 \leq M \underbrace{\int |\mu(z - x) - \mu(z - y)| dz}_{I_2}. \tag{5.30}$$

What remains is to control the integral term (5.30), denoted I_2.

In order to do so, we follow a technique used by Lakshmanan and Pucci de Farias [113]. Since μ satisfies $\mu(z - x) \geq \mu(z - y)$ if and only if $\|z - x\|_2 \geq \|z - y\|_2$, we have

$$I_2 = \int |\mu(z - x) - \mu(z - y)| dz = 2 \int_{z:\|z-x\|_2 \leq \|z-y\|_2} (\mu(z - x) - \mu(z - y)) dz.$$

By making the change of variable $w = z - x$ for the $\mu(z - x)$ term in I_2 and $w = z - y$ for $\mu(z - y)$, we rewrite I_2 as

$$\begin{aligned} I_2 &= 2 \int_{w:\|w\|_2 \leq \|w-(x-y)\|_2} \mu(w) dw - 2 \int_{w:\|w\|_2 \geq \|w-(x-y)\|_2} \mu(w) dw \\ &= 2\mathbb{P}_\mu(\|Z\|_2 \leq \|Z - (x - y)\|_2) - 2\mathbb{P}_\mu(\|Z\|_2 \geq \|Z - (x - y)\|_2) \end{aligned}$$

where $\mathbb{P}_\mu$ denotes probability according to the density μ. Squaring the terms inside the probability bounds, we note that

$$\begin{aligned}\mathbb{P}_\mu\left(\|Z\|_2^2 \leq \|Z-(x-y)\|_2^2\right) &= \mathbb{P}_\mu\left(2\langle Z, x-y\rangle \leq \|x-y\|_2^2\right) \\ &= \mathbb{P}_\mu\left(2\left\langle Z, \frac{x-y}{\|x-y\|_2}\right\rangle \leq \|x-y\|_2\right)\end{aligned}$$

Since $(x-y)/\|x-y\|_2$ has norm 1 and $Z \sim \mathsf{N}(0, u^2 I)$ is rotationally invariant, the random variable $W = \left\langle Z, \frac{x-y}{\|x-y\|_2}\right\rangle$ has distribution $\mathsf{N}(0, u^2)$. Consequently, we have

$$\begin{aligned}\frac{I_2}{2} &= \mathbb{P}\left(W \leq \|x-y\|_2/2\right) - \mathbb{P}\left(W \geq \|x-y\|_2/2\right) \\ &= \int_{-\infty}^{\|x-y\|_2/2} \frac{1}{\sqrt{2\pi u^2}} \exp(-w^2/(2u^2))dw - \int_{\|x-y\|_2/2}^{\infty} \frac{1}{\sqrt{2\pi u^2}} \exp(-w^2/(2u^2))dw \\ &\leq \frac{1}{u\sqrt{2\pi}} \|x-y\|_2,\end{aligned}$$

where we have exploited symmetry and the inequality $\exp(-w^2) \leq 1$. Combining this bound with the earlier inequality (5.30), we have

$$\|\nabla f_u(x) - \nabla f_u(y)\|_2 \leq \frac{2M}{u\sqrt{2\pi}} \|x-y\|_2 \leq \frac{M}{u} \|x-y\|_2.$$

That each of the bounds above is tight is a consequence of Lemma 5.6.

5.6.1.3 Proof of Lemma 5.6

Throughout this proof, c will denote a dimension independent constant and may change from line to line and inequality to inequality. We will show the result holds by considering a convex combination of the "difficult" functions $f_1(x) = M\|x\|_2$ and $f_2(x) = M|\langle x, y/\|y\|_2\rangle - 1/2|$, and choosing $f = \frac{1}{2}f_1 + \frac{1}{2}f_2$. Our first step in the proof will be to control E_U.

By definition (5.24) of the constant E_U, we have $E_U(\frac{1}{2}f_1 + \frac{1}{2}f_2) \geq \frac{1}{2}\max\{E_U(f_1), E_U(f_2)\}$ for any convex f_1 and f_2. Thus for $Z \sim \mathsf{N}(0, u^2 I_{d\times d})$ we have $\mathbb{E}[f_1(Z)] \geq cMu\sqrt{d}$, i.e. $E_U(f) \geq cMu\sqrt{d}$, and for Z uniform on $B_2(0,u)$, we have $\mathbb{E}[f_1(Z)] \geq cMu$, implying $E_U(f) \geq cMu$.

Turning to control of E_∇, we note that for any random variable Z rotationally symmetric about the origin, symmetry implies that

$$\mathbb{E}[\nabla f_1(Z+y)] = M\mathbb{E}\left[\frac{Z+y}{\|Z+y\|_2}\right] = a_z y$$

where $a_z > 0$ is a constant dependent on Z. Thus we have

$$\mathbb{E}[\nabla f_1(Z)] - \mathbb{E}[\nabla f_1(Z+y)] + \mathbb{E}[\nabla f_2(Z)] - \mathbb{E}[\nabla f_2(Z+y)] = 0 - a_z y - M\frac{y}{\|y\|_2}\int |\mu(z) - \mu(z-y)|dz$$

from Lemma 5.7. As a consequence (since $a_z y$ is parallel to $y/\left\|y\right\|_2$), we see that

$$E_\nabla\left(\frac{1}{2}f_1+\frac{1}{2}f_2\right)\geq\frac{1}{2}M\int|\mu(z)-\mu(z-y)|dz.$$

So what remains is to lower bound $\int|\mu(z)-\mu(z-y)|dz$ for the uniform and normal distributions. As we saw in the proof of Lemma 5.5, for the normal distribution

$$\int|\mu(z)-\mu(z-y)|dz=\frac{1}{u\sqrt{2\pi}}\int_{-\left\|y\right\|_2/2}^{\left\|y\right\|_2/2}\exp(-w^2/(2u^2))dw=\frac{1}{u\sqrt{2\pi}}\left\|y\right\|_2+\mathcal{O}\left(\frac{\left\|y\right\|_2^2}{u}\right).$$

By taking small enough $\left\|y\right\|_2$, we achieve the inequality $E_\nabla\left(\frac{1}{2}f_1+\frac{1}{2}f_2\right)\geq c\frac{M}{u}$ when $Z\sim\mathsf{N}(0,u^2I_{d\times d})$.

To show that the bound in the lemma is sharp for the case of the uniform distribution on $B_2(0,u)$, we slightly modify the proof of Lemma 2 in [187]. In particular, by using a Taylor expansion instead of first-order convexity in inequality (11) of [187], it is not difficult to show that

$$\int|\mu(z)-\mu(z-y)|dz=\kappa\frac{d!!}{(d-1)!!}\frac{\left\|y\right\|_2}{u}+\mathcal{O}\left(\frac{d\left\|y\right\|_2^2}{u^2}\right),$$

where $\kappa=2/\pi$ if d is even and 1 otherwise. Since $d!!/(d-1)!!=\Theta(\sqrt{d})$, we have proved that for small enough $\left\|y\right\|_2$, there is a constant c such that $\int|\mu(z)-\mu(z-y)|dz\geq c\sqrt{d}\left\|y\right\|_2/u$.

5.6.1.4 Proof of Lemma 5.7

Without loss of generality, we assume that $x=0$ (a linear change of variables allows this). Let $g:\mathbb{R}^d\to\mathbb{R}^d$ be a vector-valued function such that $\left\|g(z)\right\|_*\leq M$ for all $z\in\{y\}+\operatorname{supp}\mu$. Then

$$\begin{aligned}\mathbb{E}[g(Z)-g(y+Z)]&=\int g(z)\mu(z)dz-\int g(y+z)\mu(z)dz\\&=\int g(z)\mu(z)dz-\int g(z)\mu(z-y)dz\\&=\int_{I_>}g(z)[\mu(z)-\mu(z-y)]dz-\int_{I_<}g(z)[\mu(z-y)-\mu(z)]dz\end{aligned}\tag{5.31}$$

where $I_>=\{z\in\mathbb{R}^d\mid\mu(z)>\mu(z-y)\}$ and $I_<=\{z\in\mathbb{R}^d\mid\mu(z)<\mu(z-y)\}$. It is now clear that when we take norms we have

$$\begin{aligned}\left\|\mathbb{E}g(Z)-g(y+Z)\right\|_*&\leq\sup_{z\in I_>\cup I_<}\left\|g(z)\right\|_*\left|\int_{I_>}[u(z)-u(z-y)]dz+\int_{I_<}[u(z-y)-u(z)]dz\right|\\&\leq M\left|\int_{I_>}\mu(z)-\mu(z-y)dz+\int_{I_<}\mu(z-y)-\mu(z)dz\right|\\&=M\int|\mu(z)-\mu(z-y)|dz.\end{aligned}$$

Taking $g(z)$ to be an arbitrary element of $\partial f(z)$ completes the proof of the bound (5.26).

To see that the result is tight when μ is rotationally symmetric and the norm $\|\cdot\| = \|\cdot\|_2$, we note the following. From the equality (5.31), we see that $\|\mathbb{E}[g(Z) - g(y+Z)]\|_2$ is maximized by choosing $g(z) = v$ for $z \in I_>$ and $g(z) = -v$ for $z \in I_<$ for any v such that $\|v\|_2 = M$. Since μ is rotationally symmetric and non-increasing in $\|z\|_2$,

$$I_> = \left\{z \in \mathbb{R}^d \mid \mu(z) > \mu(z-y)\right\} = \left\{z \in \mathbb{R}^d \mid \|z\|_2^2 < \|z-y\|_2^2\right\} = \left\{z \mid \langle z, y\rangle < \frac{1}{2}\|y\|_2^2\right\}$$

$$I_< = \left\{z \in \mathbb{R}^d \mid \mu(z) < \mu(z-y)\right\} = \left\{z \in \mathbb{R}^d \mid \|z\|_2^2 > \|z-y\|_2^2\right\} = \left\{z \mid \langle z, y\rangle > \frac{1}{2}\|y\|_2^2\right\}.$$

So all we need do is find a function f for which there exists v with $\|v\|_2 = M$, and such that $\partial f(x) = \{v\}$ for $x \in I_>$ and $\partial f(x) = \{-v\}$ for $x \in I_<$. By inspection, the function f defined in the statement of the lemma satisfies these two desiderata for $v = M\frac{y}{\|y\|_2}$.

Chapter 6

Zero-order optimization: the power of two function evaluations

In this chapter, we consider derivative-free algorithms for stochastic and non-stochastic optimization problems that use only function values rather than gradients. It is of interest to study such scenarios, as a variety of black-box optimization problems—for example, simulation-based objectives—can only provide function evaluations. Focusing on non-asymptotic bounds on convergence rates, we show that if pairs of function values are available, algorithms for d-dimensional optimization that use gradient estimates based on random perturbations suffer a factor of at most $\sqrt{d}$ in convergence rate over traditional stochastic gradient methods. We establish such results for both smooth and non-smooth cases, sharpening previous analyses that suggested a worse dimension dependence. We complement our algorithmic development with information-theoretic lower bounds on the minimax convergence rate of such problems, establishing the sharpness of our achievable results up to constant factors. That is, when we are faced with an informational constraint that only allows access to function values—a constraint in the notion of Chapter 2.1—we identify new (and re-analyze old) algorithms, building off of the randomized smoothing tools developed in the previous chapter, and prove their optimality in terms of minimax excess risk (2.3).

6.1 Introduction

Derivative-free optimization schemes have a long history in optimization; for instance, see the book by Spall [165] for an overview. Such schemes are desirable in settings in which explicit gradient calculations may be computationally infeasible, expensive, or impossible. Classical techniques in stochastic and non-stochastic optimization, including Kiefer-Wolfowitz-type procedures [e.g. 112], use function difference information to approximate gradients of the function to be minimized rather than calculating gradients. There has been renewed interest in optimization problems with only functional (zero-order) information available—rather than first-order gradient information—in optimization, machine learning, and statistics.

In machine learning and statistics, this interest has centered around the bandit convex optimization setting, where a player and adversary compete, with the player choosing points θ in some domain Θ and an adversary choosing a point x, forcing the player to suffer a loss $F(\theta; x)$, where $F(\cdot; x) : \Theta \to \mathbb{R}$ is a convex function [78, 17, 4]. The goal is to choose an optimal point $\theta \in \Theta$ based only on possibly noisy observations of function values $F(\theta; x)$. Applications of such bandit problems include online auctions and advertisement selection for search engines. Similarly, the field of simulation-based optimization provides many examples of problems in which optimization is performed based only on function values [165, 46, 139]. Finally, in many problems in statistics—including graphical model inference problems [177] and structured-prediction problems [167]—the objective is defined variationally (as the maximum of a family of functions), so explicit differentiation may be difficult.

Despite the long history and recent renewed interest in such procedures, an understanding of their finite-sample convergence rates remains elusive. In this chapter, we study algorithms for solving stochastic convex risk minimization problems of the usual form (3.1), that is,

$$\underset{\theta \in \Theta}{\text{minimize}} \, f(\theta) := \mathbb{E}_P[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x) dP(x),$$

where $\Theta \subseteq \mathbb{R}^d$ is a compact convex set, P is a distribution over the space $\mathcal{X}$, and for P-almost every $x \in \mathcal{X}$, the function $F(\cdot; x)$ is closed and convex. Our focus is on the convergence rates of algorithms that observe only stochastic realizations of the function values $f(\theta)$, though our algorithms naturally apply in the non-stochastic case as well.

One body of work focuses on problems where, for a given value $x \in \mathcal{X}$, it is only possible to observe a noisy versions of $F(\theta; x)$ at a single location θ. Nemirovski and Yudin [134, Chapter 9.3] develop a randomized sampling strategy that estimates the gradient $\nabla F(\theta; x)$ via randomized evaluations of function values at samples from the surface of the ℓ_2-sphere. Flaxman et al. [78] further build on this approach, and establish some implications for bandit convex optimization problems. The convergence rates given in these early papers are sub-optimal, as shown by more recent work [139, 7]. For instance, Agarwal et al. [7] provide algorithms that achieve convergence rates of $\mathcal{O}(\text{poly}(d)/\sqrt{n})$, where $\text{poly}(d)$ is a polynomial in the dimension d; however, as the authors themselves note, the algorithms are quite complicated. Jamieson et al. [102] present somewhat simpler comparison-based algorithms for solving such problems, and Shamir [163] gives optimal algorithms for quadratic objectives, as well as providing some lower bounds on optimization error when only single function values are available.

Some of the difficulties inherent in optimization using only a single function evaluation can be alleviated when the function $F(\cdot; x)$ can be evaluated at *two* points, as noted independently by Agarwal et al. [4] and Nesterov [139]. The insight is that for a small non-zero scalar u and a vector $Z \in \mathbb{R}^d$, the quantity $(F(\theta + uZ; x) - F(\theta; x))/u$ approximates a directional derivative of $F(\theta; x)$ in the direction Z. Such an approximation can be exploited by first-order optimization schemes. Relative to schemes based on only a single function evaluation at each iteration, such two-sample-based gradient estimators exhibit faster convergence

rates [4, 139, 82]. In the current chapter, we take this line of work further, in particular by characterizing the *optimal* rate of convergence over all iterative procedures based on noisy function evaluations. Moreover, adopting the two-point perspective, we present simple randomization-based algorithms that achieve these optimal rates.

More formally, we study algorithms that receive a vector of paired observations, $Y(\theta, w) \in \mathbb{R}^2$, where θ and w are points selected by the algorithm. The kth observation takes the form

$$Y^k(\theta^k, w^k) := \begin{bmatrix} F(\theta^k; X_k) \\ F(w^k; X_k) \end{bmatrix}, \tag{6.1}$$

where X_k is an independent sample drawn from the distribution P. After n iterations, the algorithm returns a vector $\widehat{\theta}(n) \in \Theta$. In this setting, we analyze stochastic gradient and mirror-descent procedures [190, 134, 18, 135] that construct gradient estimators using the two-point observations Y^k. By a careful analysis of the dimension dependence of certain random perturbation schemes, we show that the convergence rate attained by our stochastic gradient methods is roughly a factor of $\sqrt{d}$ worse than that attained by stochastic methods that observe the full gradient $\nabla F(\theta; X)$. Under appropriate conditions, our convergence rates are a factor of $\sqrt{d}$ better than those attained in past work [4, 139]. For smooth problems, Ghadimi and Lan [82] provide results sharper than those in the papers [4, 139], but do not show optimality of their methods nor consider high-dimensional (non-Euclidean) problems. In addition, although we present our results in the framework of stochastic optimization, our analysis also applies to (two-point) bandit online convex optimization problems [78, 17, 4] and non-stochastic problems [134, 139]; in these settings, we obtain the sharpest rates derived to date. Our algorithms apply in both smooth and non-smooth cases. In sharp contrast to gradient-based methods, we show that there is no difference—apart from a logarithmic factor in the dimension—in the attainable convergence rates for the smooth versus non-smooth settings. Finally, we establish that our achievable rates are sharp up to constant factors, in particular by using information-theoretic techniques for proving lower bounds in statistical estimation.

The remainder of this chapter is organized as follows: in the next section, we present our two-point gradient estimators and their associated convergence rates, providing results in Section 6.2.1 and 6.2.2 for smooth and non-smooth objectives F, respectively. In Section 6.3, we provide information-theoretic minimax lower bounds on the best possible convergence rates, uniformly over all schemes based on function evaluations. We devote Sections 6.5 and Section 6.6 to proofs of the achievable convergence rates and the lower bounds, respectively, deferring proofs of more technical results.

6.2 Algorithms

In this chapter, we use (variants of) stochastic mirror descent methods for solving the stochastic convex optimization problem (3.1); recall Section 3.1 in Chapter 3. We recall that they are based on a strongly convex proximal function ψ and its associated Bregman divergence

$D_\psi(\theta, w) = \psi(\theta) - \psi(w) - \langle \nabla\psi(w), \theta - w\rangle$, with stochastic (sub)gradient updates

$$\theta^{k+1} = \operatorname*{argmin}_{\theta\in\Theta} \left\{ \left\langle g^k, \theta \right\rangle + \frac{1}{\alpha_k} D_\psi(\theta, \theta^k) \right\},$$

for a non-increasing sequence $\{\alpha_k\}_{k=1}^\infty$ of positive stepsizes.

Throughout this chapter, we impose two assumptions that are standard in analysis of mirror descent methods (cf. Section 3.1 and references [134, 18, 135]). Letting θ^* denote a minimizer of the problem (3.1), the first assumption concerns properties of the proximal function ψ and the optimizaton domain Θ.

Assumption 6A. *The proximal function ψ is 1-strongly convex with respect to the norm $\|\cdot\|$. The domain Θ is compact, and there exists $\mathsf{r}_\psi < \infty$ such that $D_\psi(\theta^*, \theta) \le \frac{1}{2}\mathsf{r}_\psi^2$ for $\theta \in \Theta$.*

Our second assumption is standard for almost all first-order stochastic gradient methods [135, 183, 139], and it holds whenever the functions $F(\cdot; x)$ are M-Lipschitz with respect to the norm $\|\cdot\|$. We use $\|\cdot\|_*$ to denote the dual norm to $\|\cdot\|$, and let $\mathsf{g} : \Theta \times \mathcal{X} \to \mathbb{R}^d$ denote a measurable subgradient selection for the functions F; that is, $\mathsf{g}(\theta; x) \in \partial F(\theta; x)$ with $\mathbb{E}[\mathsf{g}(\theta; X)] \in \partial f(\theta)$.

Assumption 6B. *There is a constant $M < \infty$ such that the (sub)gradient selection g satisfies $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_*^2] \le M^2$ for $\theta \in \Theta$.*

When Assumptions 6A and 6B hold, as in Section 3.1, the behavior of stochastic mirror descent methods is well understood [18, 135]. As noted in Proposition 3.2 and the subsequent inquality (3.7), the stepsize choice $\alpha_k = \alpha\mathsf{r}_\psi/M\sqrt{k}$ implies that the running average $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$ satisfies

$$\mathbb{E}[f(\widehat{\theta}(n))] - f(\theta^*) \le \mathcal{O}(1) \max\{\alpha, \alpha^{-1}\} \frac{\mathsf{r}_\psi M}{\sqrt{n}}.$$

For the remainder of this section, we explore the use of function difference information to obtain subgradient estimates that can be used in mirror descent methods to achieve statements similar to the convergence guarantee (3.7). We begin by analyzing the smooth case—when the instantaneous functions $F(\cdot; x)$ have Lipschitz gradients—and proceed to the more general (non-smooth) case in the subsequent section.

6.2.1 Two-point gradient estimates and convergence rates: smooth case

Our first step is to show how to use two function values to construct nearly unbiased estimators of the gradient of the objective function f, under a smoothness condition. Using analytic methods different from those from past work [4, 139], we are able to obtain optimal dependence with the problem dimension d. In more detail, our procedure is based on a non-increasing sequence of positive smoothing parameters $\{u_k\}_{k=1}^\infty$, and a distribution μ on

$\mathbb{R}^d$, to be specified, satisfying $\mathbb{E}_\mu[ZZ^\top] = I_{d\times d}$. Given a smoothing constant u, vector z, and observation x, we define the directional gradient estimate at the point θ as

$$\mathsf{g}_{\mathrm{sm}}(\theta; u, z, x) := \frac{F(\theta + uz; x) - F(\theta; x)}{u} z. \tag{6.2}$$

Using the estimator (6.2), we then perform the following two steps. First, upon receiving the point $X_k \in \mathcal{X}$, we sample an independent vector Z^k and set

$$g^k = \mathsf{g}_{\mathrm{sm}}(\theta^k; u_k, Z^k, X_k) = \frac{F(\theta^k + u_k Z^k; X_k) - F(\theta^k; X_k)}{u_k} Z^k. \tag{6.3}$$

In the second step, we apply the standard mirror descent update (3.6) to the quantity g^k to obtain the next parameter θ^{k+1}.

A consideration of directional derivatives may give intuition for the estimator (6.2). The directional derivative $f'(\theta, z)$ of the function f at the point θ in the direction z is

$$f'(\theta, z) := \lim_{u \downarrow 0} \frac{1}{u}(f(\theta + uz) - f(\theta)).$$

This limit always exists when f is convex [98, Chapter VI], and if f is differentiable at θ, then $f'(\theta, z) = \langle \nabla f(\theta), z\rangle$. With this background, the estimate (6.2) is motivated by the following standard fact [139, equation (32)]: whenever $\nabla f(\theta)$ exists, we have

$$\mathbb{E}[f'(\theta, Z)Z] = \mathbb{E}[\langle \nabla f(\theta), Z\rangle\, Z] = \mathbb{E}[ZZ^\top \nabla f(\theta)] = \nabla f(\theta),$$

where the final equality uses our assumption that $\mathbb{E}[ZZ^\top] = I_{d\times d}$. Consequently, given sufficiently small choices of u_k, the vector (6.3) should be a nearly unbiased estimator of the gradient $\nabla f(\theta^k)$ of the risk.

In addition to the condition $\mathbb{E}_\mu[ZZ^\top] = I$, we require that

$$\operatorname{dom} F(\cdot; x) \supset \Theta + u_{1,1} \operatorname{supp} \mu \quad \text{for } x \in \mathcal{X} \tag{6.4}$$

to ensure that the estimator g^k of (6.3) is well-defined. If we apply smoothing with Gaussian perturbation, the containment (6.4) implies $\operatorname{dom} F(\cdot; x) = \mathbb{R}^d$, though we still optimize over the compact set Θ in the update (3.6). We also impose the following properties on the smoothing distribution:

Assumption 6C. *For $Z \sim \mu$, the quantity $M(\mu) := \sqrt{\mathbb{E}[\|Z\|^4 \|Z\|_*^2]}$ is finite, and moreover, there is a function $s : \mathbb{N} \to \mathbb{R}_+$ such that*

$$\mathbb{E}[\|\langle g, Z\rangle\, Z\|_*^2] \le s(d) \|g\|_*^2 \quad \textit{for any vector } g \in \mathbb{R}^d. \tag{6.5}$$

Although the quantity $M(\mu)$ is required to be finite, its value does not appear explicitly in our theorem statements. On the other hand, the dimension-dependent quantity $s(d)$ from condition (6.5) appears explicitly in our convergence rates. As an example of these two quantities, suppose that we take μ to be the distribution of the standard normal $\mathsf{N}(0, I_{d\times d})$, and use the ℓ_2-norm $\|\cdot\| = \|\cdot\|_2$. In this case, a straightfoward calculation shows that $M(\mu)^2 \lesssim d^3$ and $s(d) \lesssim d$.

Finally, as previously stated, the analysis of this section requires a smoothness assumption:

Assumption 6D. *There is a function $L : \mathcal{X} \to \mathbb{R}_+$ such that for P-almost every $x \in \mathcal{X}$, the function $F(\cdot; x)$ has $L(x)$-Lipschitz continuous gradient with respect to the norm $\|\cdot\|$, and moreover the quantity $L(P) := \sqrt{\mathbb{E}[(L(X))^2]}$ is finite.*

As we have seen in Chapter 3, essential to stochastic gradient procedures is that the gradient estimator g^k be nearly unbiased and have small norm. Accordingly, the following lemma provides quantitative guarantees on the error associated with the gradient estimator (6.2).

Lemma 6.1. *Under Assumptions 6C and 6D, the gradient estimate* (6.2) *has expectation*

$$\mathbb{E}[\mathsf{g}_{\mathrm{sm}}(\theta; u, Z, X)] = \nabla f(\theta) + uL(P)v \tag{6.6}$$

for some vector v such that $\|v\|_* \leq \frac{1}{2}\mathbb{E}[\|Z\|^2 \|Z\|_*]$. Moreover, its expected squared norm is bounded as

$$\mathbb{E}[\|\mathsf{g}_{\mathrm{sm}}(\theta; u, Z, X)\|_*^2] \leq 2s(d)\mathbb{E}\left[\|\mathsf{g}(\theta; X)\|_*^2\right] + \frac{1}{2}u^2 L(P)^2 M(\mu)^2. \tag{6.7}$$

See Section 6.5.2 for the proof. The bound (6.6) shows that the estimator g^k is unbiased for the gradient up to a correction term of order u_k, while the second inequality (6.7) shows that the second moment is—up to an order u_k^2 correction—within a factor $s(d)$ of the standard second moment $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_*^2]$.

Our main result in this section is the following theorem on the convergence rate of the mirror descent method using the gradient estimator (6.3).

Theorem 6.1. *Under Assumptions 6A, 6B, 6C, and 6D, consider a sequence $\{\theta^k\}_{k=1}^\infty$ generated according to the mirror descent update* (3.6) *using the gradient estimator* (6.3)*, with step and perturbation sizes*

$$\alpha_k = \alpha \frac{\mathsf{r}_\psi}{2M\sqrt{s(d)}\sqrt{k}} \quad \text{and} \quad u_k = u \frac{M\sqrt{s(d)}}{L(P)M(\mu)} \cdot \frac{1}{k} \qquad \text{for } k = 1, 2, \ldots.$$

Then for all n,

$$\mathbb{E}\left[f(\widehat{\theta}(n)) - f(\theta^*)\right] \leq 2\frac{\mathsf{r}_\psi M\sqrt{s(d)}}{\sqrt{n}} \max\left\{\alpha, \alpha^{-1}\right\} + \alpha u^2 \frac{\mathsf{r}_\psi M\sqrt{s(d)}}{n} + u\frac{\mathsf{r}_\psi M\sqrt{s(d)}\log(2n)}{n}, \tag{6.8}$$

where $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$, and the expectation is taken with respect to the samples X and Z.

The proof of Theorem 6.1 builds on convergence proofs developed in the analysis of online and stochastic convex optimization [190, 135, 4, 139], but requires additional technical care, since we never truly receive unbiased gradients. We provide the proof in Section 6.5.1.

Before continuing, we make a few remarks. First, the method is reasonably robust to the selection of the step-size multiplier α; Nemirovski et al. [135] previously noted this robustness for gradient-based MD methods. As long as $\alpha_k \propto 1/\sqrt{k}$, mis-specifying the multiplier α results in a scaling at worst linear in $\max\{\alpha, \alpha^{-1}\}$. In addition, the convergence rate of the method is independent of the Lipschitz continuity constant $L(P)$ of the instantaneous gradients $\nabla F(\cdot; X)$, suggesting that similar results might hold for non-differentiable functions. Indeed, as we show in the next section, a slightly more complicated construction of the estimator g^k leads to analogous guarantees for general non-smooth functions.

Although we have provided only bounds on the expected convergence rate, it is possible to give high-probability convergence guarantees [cf. 40, 135] under additional tail conditions on g—for example, under a condition of the form $\mathbb{E}[\exp(\|\mathsf{g}(\theta; X)\|_*^2 / M^2)] \leq \exp(1)$. Additionally, though we have presented our results as convergence guarantees for stochastic optimization problems, an inspection of our analysis in Section 6.5.1 shows that we obtain (expected) regret bounds for bandit online convex optimization problems [cf. 78, 17, 4].

6.2.1.1 Examples and corollaries

We now provide examples of random sampling strategies that lead to concrete bounds for the mirror descent algorithm based on the subgradient estimator (6.3). For each corollary, we specify the norm $\|\cdot\|$, proximal function ψ, and distribution μ. We then compute the values that the distribution μ implies in Assumption 6D and apply Theorem 6.1 to obtain a convergence rate.

We begin with a corollary that characterizes the convergence rate of our algorithm with the proximal function $\psi(\theta) := \frac{1}{2}\|\theta\|_2^2$ under a Lipschitz continuity condition:

Corollary 6.1. *Given an optimization domain $\Theta \subseteq \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq \mathsf{r}_2\}$, suppose that μ is uniform on the surface of the ℓ_2-ball of radius $\sqrt{d}$, and that $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \leq M^2$. Then*

$$\mathbb{E}\left[f(\widehat{\theta}(n)) - f(\theta^*)\right] \leq 2\frac{\mathsf{r}_2 M\sqrt{d}}{\sqrt{n}} \max\{\alpha, \alpha^{-1}\} + \alpha u^2 \frac{\mathsf{r}_2 M\sqrt{d}}{n} + u\frac{\mathsf{r}_2 M\sqrt{d}\log n}{n}.$$

Proof Since $\|Z\|_2 = \sqrt{d}$, we have $M(\mu) = \sqrt{\mathbb{E}[\|Z\|_2^6]} = d^{3/2}$. Since $\mathbb{E}[ZZ^\top] = I$ by assumption, we see that

$$\mathbb{E}[\|\langle g, Z\rangle Z\|_2^2] = d\mathbb{E}[\langle g, Z\rangle^2] = d\mathbb{E}[g^\top ZZ^\top g], \qquad \text{valid for any } g \in \mathbb{R}^d,$$

showing that Assumption 6C holds with $s(d) = d$. The claim follows from Theorem 6.1. □

The rate provided by Corollary 6.1 is the fastest derived to date for zero-order stochastic optimization using two function evaluations; both Agarwal et al. [4] and Nesterov [139] achieve

rates of convergence of order $\mathsf{r}_2 Md/\sqrt{n}$. In concurrent work, Ghadimi and Lan [82] provide a result (their Corollary 3.3) that achieves a similar rate to that above, but their primary focus is on non-convex problems. Moreover, we show in the sequel that this convergence rate is actually optimal.

In high-dimensional scenarios, appropriate choices for the proximal function ψ yield better scaling on the norm of the gradients [134, 81, 135]. In the setting of online learning or stochastic optimization, suppose that one observes gradients $\mathsf{g}(\theta; X)$. If the domain Θ is the simplex, then exponentiated gradient algorithms [110, 18] using the proximal function $\psi(\theta) = \sum_j \theta_j \log \theta_j$ obtain rates of convergence dependent on the ℓ_∞-norm of the gradients $\|\mathsf{g}(\theta; X)\|_\infty$. This scaling is more palatable than bounds that depend on Euclidean norms applied to the gradient vectors, which may be a factor of $\sqrt{d}$ larger. Similar results apply using proximal functions based on ℓ_p-norms [22, 18]. Concretely, if we make the choice $p = 1 + \frac{1}{\log(2d)}$ and $\psi(\theta) = \frac{1}{2(p-1)} \|\theta\|_p^2$, we obtain the following corollary, which holds under the conditions of Theorem 6.1.

Corollary 6.2. *Suppose that $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_\infty^2] \le M^2$, the optimization domain Θ is contained in the ℓ_1-ball $\{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le \mathsf{r}_1\}$, and μ is uniform on the hypercube $\{-1, 1\}^d$. There is a universal constant $C \le 2e$ such that*

$$\mathbb{E}\left[f(\widehat{\theta}(n)) - f(\theta^*)\right] \le C \frac{\mathsf{r}_1 M \sqrt{d \log(2d)}}{\sqrt{n}} \max\left\{\alpha, \alpha^{-1}\right\} + C \frac{\mathsf{r}_1 M \sqrt{d \log(2d)}}{n} \left(\alpha u^2 + u \log n\right).$$

Proof Recall that from the discussion following Corollary 3.3 that the stated choice of proximal function ψ is strongly convex with respect to the norm $\|\cdot\|_p$ (see also [134, Appendix 1] or [22]). In addition, if we define $q = 1 + \log(2d)$, then we have $1/p + 1/q = 1$, and $\|v\|_q \le e \|v\|_\infty$ for any $v \in \mathbb{R}^d$. Consequently, we have $\mathbb{E}[\|\langle g, Z\rangle Z\|_q^2] \le e^2 \mathbb{E}[\|\langle g, Z\rangle Z\|_\infty^2]$, which allows us to apply Theorem 6.1, with the norm $\|\cdot\| = \|\cdot\|_1$ and the dual norm $\|\cdot\|_* = \|\cdot\|_\infty$.

We claim that Assumption 6C is satisfied with $s(d) \le d$. Since $Z \sim \text{Uniform}(\{-1, 1\}^d)$, we have

$$\mathbb{E}\left[\|\langle g, Z\rangle Z\|_\infty^2\right] = \mathbb{E}\left[\langle g, Z\rangle^2\right] = g^\top \mathbb{E}[ZZ^\top] g = \|g\|_2^2 \le d \|g\|_\infty^2 \quad \text{for any } g \in \mathbb{R}^d.$$

Finally, we have $M(\mu) = \sqrt{\mathbb{E}[\|Z\|_1^4 \|Z\|_\infty^2]} = d^2$, which is finite as needed. By the inclusion of Θ in the ℓ_1-ball of radius R and our choice of proximal function, we have

$$(p-1) D_\psi(\theta, w) \le \frac{1}{2} \|\theta\|_p^2 + \frac{1}{2} \|w\|_p^2 + \|\theta\|_p \|w\|_p.$$

(For instance, see Lemma 3 in the paper [81].) We thus find that $D_\psi(\theta, w) \le 2\mathsf{r}_1^2 \log(2d)$ for any $\theta, w \in \Theta$, and using the step size choices of Theorem 6.1 gives the result. □

Corollary 6.2 attains a convergence rate that scales with dimension as $\sqrt{d \log d}$. This dependence on dimension is much worse than that of (stochastic) mirror descent using full

gradient information [134, 135]. The additional dependence on d suggests that while $\mathcal{O}(1/\epsilon^2)$ iterations are required to achieve ϵ-optimization accuracy for mirror descent methods, the two-point method requires $\mathcal{O}(d/\epsilon^2)$ iterations to obtain the same accuracy. A similar statement holds for the results of Corollary 6.1. In Section 6.3 we show that this dependence is sharp: apart from logarithmic factors, no algorithm can attain better convergence rates, including the problem-dependent constants r_ψ and M.

6.2.2 Two-point gradient estimates and convergence rates: general case

We now turn to the general setting, in which the function $F(\cdot; x)$, rather than having a Lipschitz continuous gradient, satisfies only the milder condition of Lipschitz continuity. The difficulty in this non-smooth case is that the simple gradient estimator (6.3) may have overly large norm. For instance, a naive calculation using only the M-Lipschitz continuity of the function f gives the bound

$$\mathbb{E}\left[\|(f(\theta + uZ) - f(\theta))Z/u\|_2^2\right] \leq M^2 \mathbb{E}\left[\|u \|Z\|_2 Z/u\|_2^2\right] = M^2 \mathbb{E}[\|Z\|_2^4]. \tag{6.9}$$

This upper bound always scales at least quadratically in the dimension, since we have the lower bound $\mathbb{E}[\|Z\|_2^4] \geq (\mathbb{E}[\|Z\|_2^2])^2 = d^2$, where the final equality uses the fact that $\mathbb{E}[ZZ^\top] = I_{d\times d}$ by assumption. This quadratic dependence on dimension leads to a sub-optimal convergence rate. Moreover, this scaling appears to be unavoidable using a single perturbing random vector: taking $f(\theta) = M\|\theta\|_2$ and setting $\theta = 0$ shows that the bound (6.9) may hold with equality.

Nevertheless, the convergence rate in Theorem 6.1 shows that *near* non-smoothness is effectively the same as being smooth. This suggests that if we can smooth the objective f slightly, we may achieve a rate of convergence even in the non-smooth case that is roughly the same as that in Theorem 6.1. We have already seen in previous chapters how smoothing the objective can yield faster convergence rates in stochastic optimization; Nesterov [137] has also shown how such ideas can yield better performance for certain deterministic problems. In the stochastic setting, of course, we can readily use convolution, as it is a smoothing operation, and adding a bit of additional noise has essentially negligible effect on performance. As noted in the previous chapter, the smoothed function

$$f_u(\theta) := \mathbb{E}[f(\theta + uZ)] = \int f(\theta + uz) d\mu(z), \tag{6.10}$$

where $Z \in \mathbb{R}^d$ has density with respect to Lebesgue measure, is always differentiable; moreover, if f is Lipschitz, then ∇f_u is Lipschitz under mild conditions.

The smoothed function (6.10) leads us to a *two-point* strategy: we use a random direction as in the smooth case (6.3) to estimate the gradient, but we introduce an extra step of randomization for the point at which we evaluate the function difference. Roughly speaking, this randomness has the effect of making it unlikely that the perturbation vector Z is near

a point of non-smoothness, which allows us to apply results similar to those in the smooth case.

More precisely, our construction uses two non-increasing sequences of positive parameters $\{u_{1,k}\}_{k=1}^{\infty}$ and $\{u_{2,k}\}_{k=1}^{\infty}$ with $u_{2,k} \le u_{1,k}/2$, and two smoothing distributions μ_1, μ_2 on $\mathbb{R}^d$. Given smoothing constants u_1, u_2, vectors z_1, z_2, and observation x, we define the (non-smooth) directional gradient estimate at the point θ as

$$\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, z_1, z_2, x) := \frac{F(\theta + u_1 z_1 + u_2 z_2; x) - F(\theta + u_1 z_1; x)}{u_2} z_2. \tag{6.11}$$

Using g_{ns} we may define our gradient estimator, which follows the same intuition as our construction of the stochastic gradient (6.3) from the smooth estimator (6.2). Now, upon receiving the point X_k, we sample independent vectors $Z_1^k \sim \mu_1$ and $Z_2^k \sim \mu_2$, and set

$$g^k = \mathsf{g}_{\text{ns}}(\theta^k; u_{1,k}, u_{2,k}, Z_1^k, Z_2^k, X_k) = \frac{F(\theta^k + u_{1,k} Z_1^k + u_{2,k} Z_2^k; X_k) - F(\theta^k + u_{1,k} Z_1^k; X_k)}{u_{2,k}} Z_2^k. \tag{6.12}$$

We then proceed as in the preceding section, using this estimator in the standard mirror descent method.

To demonstrate the convergence of gradient-based schemes with gradient estimator (6.12), we require a few additional assumptions. For simplicity, in this section we focus on results for the Euclidean norm $\|\cdot\|_2$. We impose the following condition on the Lipschitzian properties of $F(\cdot; x)$, which is a slight strengthening of Assumption 6B.

Assumption 6B′. *There is a function $M \colon \mathcal{X} \to \mathbb{R}_+$ such that for P-a.e. $x \in \mathcal{X}$, the function $F(\cdot; x)$ is $M(x)$-Lipschitz with respect to the ℓ_2-norm $\|\cdot\|_2$, and the quantity $M(P) := \sqrt{\mathbb{E}[M(X)^2]}$ is finite.*

We also impose the following assumption on the smoothing distributions μ_1 and μ_2.

Assumption 6E. *The smoothing distributions are one of the following pairs: (1) both μ_1 and μ_2 are standard normal in $\mathbb{R}^d$ with identity covariance, (2) both μ_1 and μ_2 are uniform on the ℓ_2-ball of radius $\sqrt{d+2}$, or (3) the distribution μ_1 is uniform on the ℓ_2-ball of radius $\sqrt{d+2}$, whereas the distribution μ_2 is uniform on the ℓ_2-sphere of radius $\sqrt{d}$.*

In all cases, we assume the domain containment condition

$$\operatorname{dom} F(\cdot; x) \supset \Theta + u_{1,1} \operatorname{supp} \mu_1 + u_{2,1} \operatorname{supp} \mu_2 \quad \text{for } x \in \mathcal{X}.$$

Under this condition, we have the following analog of Lemma 6.1:

Lemma 6.2. *Under Assumptions 6B′ and 6E, the gradient estimator* (6.11) *has expectation*

$$\mathbb{E}[\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, Z_1, Z_2, X)] = \nabla f_{u_1}(\theta) + \frac{u_2}{u_1} M v, \tag{6.13}$$

where v is a vector bounded as $\|v\|_2 \leq \frac{1}{2}\mathbb{E}[\|Z_2\|_2^3]$. Moreover, there exists a numerical constant c (independent of u_1 and u_2) such that

$$\mathbb{E}\left[\|\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, Z_1, Z_2, X)\|_2^2\right] \leq c\, M^2 d \left(\sqrt{\frac{u_2}{u_1}}\, d + 1 + \log d\right). \tag{6.14}$$

See Section 6.5.4 for the proof of this lemma.

Comparing Lemma 6.2 to Lemma 6.1, both show that one can obtain nearly unbiased gradient of the function f using two function evaluations, but additionally, they show that the squared norm of the gradient estimator is *at most* d times larger than the expected norm of the subgradients $\partial F(\theta; x)$, as captured by the quantity M^2 from Assumption 6B or 6B$'$. In our approach, non-smoothness introduces an additional logarithmic penalty in the dimension; it may be possible to remove this factor, but we do not know how at this time. The key is that taking the second smoothing parameter u_2 to be small enough means that, aside from the dimension penalty, the gradient estimator g^k is essentially unbiased for $\nabla f_{u_{1,k}}(\theta^k)$ and has squared norm at most $M^2 d \log d$. This bound on size is essential for our main result, which we now state.

Theorem 6.2. *Under Assumptions 6A, 6B$'$, and 6E, consider a sequence $\{\theta^k\}_{k=1}^\infty$ generated according to the mirror descent update* (3.6) *using the gradient estimator* (6.12) *with step and perturbation sizes*

$$\alpha_k = \alpha \frac{\mathsf{r}_\psi}{M\sqrt{d\log(2d)}\sqrt{k}}, \qquad u_{1,k} = u\frac{\mathsf{r}_\psi}{k}, \qquad \text{and} \qquad u_{2,k} = u\frac{\mathsf{r}_\psi}{d^2k^2}.$$

Then there exists a universal (numerical) constant c such that for all n,

$$\mathbb{E}\left[f(\widehat{\theta}(n)) - f(\theta^*)\right] \leq c\max\{\alpha, \alpha^{-1}\}\frac{\mathsf{r}_\psi M\sqrt{d\log(2d)}}{\sqrt{n}} + cu\mathsf{r}_\psi M\sqrt{d}\,\frac{\log(2n)}{n}, \tag{6.15}$$

where $\widehat{\theta}(n) = \frac{1}{n}\sum_{k=1}^n \theta^k$, and the expectation is taken with respect to the samples X and Z.

The proof of Theorem 6.2 roughly follows that of Theorem 6.1, except that we prove that the sequence θ^k approximately minimizes the sequence of smoothed functions $f_{u_{1,k}}$ rather than f. However, for small $u_{1,k}$, these two functions are quite close, which combined with the estimates from Lemma 6.2 gives the result. We give the full argument in Section 6.5.3.

Remarks Theorem 6.2 shows that the convergence rate of our two-point stochastic gradient algorithm for general non-smooth functions is (at worst) a factor of $\sqrt{\log d}$ worse than the rate for smooth functions in Corollary 6.1. Notably, the rate of convergence here has substantially better dimension dependence than previously known results [4, 139, 82]. It is interesting to note, additionally, that the difference between smooth and non-smooth optimization with only functional evaluations as feedback appears to be (nearly) negligible. Using carefully constructed random perturbations, we can achieve rates of convergence of $\mathsf{r}_\psi M\sqrt{d}/\sqrt{n}$ in both cases, up to logarithmic factors in d.

6.3 Lower bounds on zero-order optimization

Thus far in the chapter, we have presented two main results (Theorems 6.1 and 6.2) that provide achievable rates for perturbation-based gradient procedures. It is natural to wonder whether or not these rates are sharp. In this section, we show that our results are unimprovable up to either a constant factor (in most cases), or a logarithmic factor in dimension in the remaining cases. These results show that *no* algorithm exists that can achieve a faster convergence rate than those we have presented under the oracle model (6.1), that is, when we constrain all procedures to use at most two function evaluations per observation X_k.

We begin by describing the constrained notion of minimax excess risk we consider here (recall the standard definition (2.3)). Let $\mathcal{C}_n^{\mathsf{zo}}$ denote the collection of all zeroth-order (optimization) algorithms that observe a sequence of data points $(Y^1, \ldots, Y^n) \subset \mathbb{R}^2$ with $Y^k = [F(\theta^k, X_k) \; F(w^k, X_k)]$ and return an estimate in Θ. Given an algorithm $\widehat{\theta} \in \mathcal{C}_n^{\mathsf{zo}}$, loss F, and data distribution P, we measure error via the optimality gap

$$f_P(\widehat{\theta}(n)) - \inf_{\theta \in \Theta} f_P(\theta) \quad \text{where} \quad f_P(\theta) = \mathbb{E}_P\big[F(\theta; X)\big]$$

and $\widehat{\theta}(n)$ is the output of algorithm $\widehat{\theta}$ on the sequence of observed function values. Taking the expectation of the above quantity, we arrive at the *constrained minimax excess risk*

$$\mathfrak{M}_n(\Theta, \mathcal{P}, F, \mathcal{C}_n^{\mathsf{zo}}) := \inf_{\widehat{\theta} \in \mathcal{C}_n^{\mathsf{zo}}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[f_P(\widehat{\theta}) - \inf_{\theta \in \Theta} f_P(\theta)], \tag{6.16}$$

where the expectation is taken over the observations $(Y^1, \ldots, Y^n)$ and any randomness in $\widehat{\theta}$. This quantity measures the performance of the best algorithm in the restricted zero-order class $\mathcal{C}_n^{\mathsf{zo}}$, where performance is required to be uniformly good for all distributions $P \in \mathcal{P}$.

We now turn to the statement of our lower bounds, which are based on relatively simple choices of the classes $\mathcal{P}$ and loss functions F. We always use the linear functional

$$F(\theta; x) = \langle \theta, x \rangle$$

as our instantaneous loss, and given an ℓ_p-norm $\|\cdot\|_p$, we consider the class of probability distributions

$$\mathcal{P}_{M,p} := \big\{P \mid \mathbb{E}_P[\|X\|_p^2] \le M^2\big\}.$$

Note any paired loss F and distribution $P \in \mathcal{P}$ satisfies Assumption 6B$'$ by construction (by taking $p = 2$), and moreover, $\nabla F(\cdot; x)$ has Lipschitz constant 0 for all x. We state each of our lower bounds assuming that the domain Θ is equal to some ℓ_q-ball of radius r_q, that is, $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_q \le \mathsf{r}_q\}$. Our first result considers the case $p = 2$ with domain Θ an arbitrary ℓ_q-ball with $q \ge 1$, so we measure gradients in the ℓ_2-norm.

Proposition 6.1. *For the class $\mathcal{P}_{M,2}$ and $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_q \le \mathsf{r}_q\}$ and any $d_0 \le d$, we have*

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,2}, F, \mathcal{C}_n^{\mathsf{zo}}) \ge \frac{1}{12}\left(1 - \frac{1}{q}\right) \frac{d_0^{1-1/q} M \mathsf{r}_q}{\sqrt{n}} \min\left\{1, \sqrt{n/d_0}\right\}. \tag{6.17}$$

Combining the lower bound (6.17) with our algorithmic schemes in Section 6.2 shows that they are optimal up to constant factors. More specifically, for $q \geq 2$, the ℓ_2-ball of radius $d^{1/2-1/q}\mathsf{r}_q$ contains the ℓ_q-ball of radius r_q, so Corollary 6.1 provides an upper bound on the minimax rate of convergence of order $\mathsf{r}_q M\sqrt{d}d^{1/2-1/q}/\sqrt{n} = \mathsf{r}_q M d^{1-1/q}/\sqrt{n}$ in the smooth case, while for $n \geq d$, Proposition 6.1 provides the lower bound $\mathsf{r}_q M d^{1-1/q}/\sqrt{n}$. Theorem 6.2, providing a rate of $\mathsf{r}_q M\sqrt{d\log d}/\sqrt{n}$ in the general (non-smooth) case, is also tight to within logarithmic factors. Consequently, the stochastic mirror descent algorithm (3.6) (even the simple stochastic gradient descent scheme (3.2)) coupled with the sampling strategies (6.3) and (6.12) is optimal for stochastic problems with two-point feedback.

For our second lower bound, we investigate the minimax rates at which it is possible to solve stochastic convex optimization problems in which the objective is Lipschitz continuous in the ℓ_1-norm, or equivalently, in which the gradients are bounded in ℓ_∞-norm. As noted earlier, such scenarios are suitable for high-dimensional problems [e.g. 135].

Proposition 6.2. *For the class $\mathcal{P}_{M,\infty}$ with $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq \mathsf{r}_1\}$, we have*

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,\infty}, F, \mathcal{C}_n^{\mathsf{zo}}) \geq \frac{1}{12\sqrt{2}} \frac{M\mathsf{r}_1}{\sqrt{n}} \min\left\{\frac{\sqrt{n}}{\sqrt{3+\log n}}, \frac{\sqrt{d}}{\sqrt{3+\log d}}\right\}.$$

This result also demonstrates the optimality of our mirror descent algorithms up to logarithmic factors. Recalling Corollary 6.2, the MD algorithm (3.6) with prox $\psi(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$, where $p = 1+1/\log(2d)$, has convergence guarantee $M\mathsf{r}_1\sqrt{d\log(2d)}/\sqrt{n}$. On the other hand, Proposition 6.2 provides the lower bound $\mathfrak{M}_n(\Theta, \mathcal{P}_{M,\infty}, F, \mathcal{C}_n^{\mathsf{zo}}) \gtrsim M\mathsf{r}_1\sqrt{d}/\sqrt{n\log d}$. These upper and lower bounds are matching up to logarithmic factors in the dimension.

It is worth comparing these lower bounds to the achievable rates when full gradient information is available—that is, when one has access to the subgradient selection $\mathsf{g}(\theta; X)$. Each of Propositions 6.1 and 6.2 has an additional $\sqrt{d}$ factor as compared to analogous lower bounds [6] applicable to the case of full gradient information. Similarly, the $\sqrt{d}$ factors disappear from the achievable convergence rates in Corollaries 6.1 and 6.2 when one uses $g^k = \mathsf{g}(\theta; X)$ in the mirror descent updates (3.6) (e.g. [18, 135]). Consequently, our analysis shows that in the zero-order setting—in addition to dependence on the radius r_ψ and second moment M^2—any algorithm must suffer at least an additional $\mathcal{O}(\sqrt{d})$ penalty in convergence rate, and optimal algorithms suffer precisely this penalty. This suggests that for high-dimensional problems, it is preferable to use full gradient information if possible, even when the cost of obtaining the gradients is somewhat nontrivial.

6.4 Summary

We have analyzed algorithms for optimization problems that use only random function values—as opposed to gradient computations—to minimize an objective function. The algorithms we present are optimal: their convergence rates cannot be improved (in the sense

of minimax optimality (2.4) for procedures constrained to use only function evaluations) by more than numerical constant factors. In addition to showing the optimality of several algorithms for smooth convex optimization without gradient information, we have also shown that the non-smooth case is no more difficult from an iteration complexity standpoint, though it requires more carefully constructed randomization schemes. As a consequence of our results, we note in passing that we have additionally attained sharp rates for bandit online convex optimization problems with multi-point feedback.

In addition, our results show that constraining estimators to use only two-point feedback (in the form of function evaluations) is a fairly stringent constraint: there is a necessary transition in convergence rates between gradient-based algorithms and those that compute only function values. Broadly, when (sub)gradient information is available, attaining ϵ-accurate solution to an optimization problem requires $\mathcal{O}(1/\epsilon^2)$ gradient observations, while at least $\Omega(d/\epsilon^2)$ observations—but no more—are necessary using paired function evaluations. An interesting open question is to understand optimization problems for which only a single stochastic function evaluation is available per sample: what is the optimal iteration complexity in this case?

6.5 Convergence proofs

We provide the proofs of the convergence results from Section 6.2 in this section, deferring more technical arguments to subsequent sections.

6.5.1 Proof of Theorem 6.1

Before giving the proof of Theorem 6.1, we state a standard lemma on the mirror descent iterates (recall Section 3.5.1 and Lemma 3.3, or see, for example, Nemirovski et al. [135, Section 2.3] or Beck and Teboulle [18, Eq. (4.21)]).

Lemma 6.3. *Let $\{g^k\}_{k=1}^n \subset \mathbb{R}^d$ be a sequence of vectors, and let θ^k be generated by the mirror descent iteration* (3.6)*. If Assumption 6A holds, then for any $\theta^* \in \Theta$ we have*

$$\sum_{k=1}^n \left\langle g^k, \theta^k - \theta^* \right\rangle \leq \frac{1}{2\alpha_n} \mathsf{r}_\psi^2 + \sum_{k=1}^n \frac{\alpha_k}{2} \left\| g^k \right\|_*^2.$$

Defining the error vector $e^k := \nabla f(\theta^k) - g^k$, Lemma 6.3 implies

$$\begin{aligned} \sum_{k=1}^n \left(f(\theta^k) - f(\theta^*) \right) \leq \sum_{k=1}^n \left\langle \nabla f(\theta^k), \theta^k - \theta^* \right\rangle &= \sum_{k=1}^n \left\langle g^k, \theta^k - \theta^* \right\rangle + \sum_{k=1}^n \left\langle e^k, \theta^k - \theta^* \right\rangle \\ &\leq \frac{1}{2\alpha_n} \mathsf{r}_\psi^2 + \sum_{k=1}^n \frac{\alpha_k}{2} \left\| g^k \right\|_*^2 + \sum_{k=1}^n \left\langle e^k, \theta^k - \theta^* \right\rangle. \quad (6.18) \end{aligned}$$

For each iteration $k = 2, 3, \ldots$, let $\mathcal{F}_{k-1}$ denote the σ-field of $X_1, \ldots, X_{k-1}$ and $Z^1, \ldots, Z^{k-1}$. Then Lemma 6.1 implies $\mathbb{E}[e^k \mid \mathcal{F}_{k-1}] = u_k L(P) v_k$, where $v_k \equiv v(\theta^k, u_k)$ satisfies $\|v_k\|_* \leq \frac{1}{2}M(\mu)$. Since $\theta^k \in \mathcal{F}_{k-1}$, we can first take an expectation conditioned on $\mathcal{F}_{k-1}$ to obtain

$$\sum_{k=1}^{n} \mathbb{E}[\langle e^k, \theta^k - \theta^* \rangle] \leq L(P) \sum_{k=1}^{n} u_k \mathbb{E}[\|v_k\|_* \left\|\theta^k - \theta^*\right\|] \leq \frac{1}{2} M(\mu) \mathsf{r}_\psi L(P) \sum_{k=1}^{n} u_k,$$

where in the last step above we have used the relation $\|\theta^k - \theta^*\| \leq \sqrt{2D_\psi(\theta^*, \theta)} \leq \mathsf{r}_\psi$. Statement (6.7) of Lemma 6.1 coupled with the assumption that $\mathbb{E}[\|\mathsf{g}(\theta^k; X)\|_*^2 \mid \mathcal{F}_{k-1}] \leq M^2$ yields

$$\mathbb{E}\left[\left\|g^k\right\|_*^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|g^k\right\|_*^2 \mid \mathcal{F}_{k-1}\right]\right] \leq 2s(d)M^2 + \frac{1}{2}u_k^2 L(P)^2 M(\mu)^2.$$

Applying the two estimates above to our initial bound (6.18), $\sum_{k=1}^{n} \mathbb{E}\big[f(\theta^k) - f(\theta^*)\big]$ is upper bounded by

$$\frac{1}{2\alpha_n}\mathsf{r}_\psi^2 + s(d)M^2 \sum_{k=1}^{n} \alpha_k + \frac{1}{4}L(P)^2 M(\mu)^2 \sum_{k=1}^{n} u_k^2 \alpha_k + \frac{1}{2}M(\mu)\mathsf{r}_\psi L(P) \sum_{k=1}^{n} u_k. \tag{6.19}$$

Now we use our choices of the sample size α_k and u_k to complete the proof. For the former, we have $\alpha_k = \alpha \mathsf{r}_\psi / (2M\sqrt{s(d)}\sqrt{k})$. Since $\sum_{k=1}^{n} k^{-\frac{1}{2}} \leq \int_0^n t^{-\frac{1}{2}} dt = 2\sqrt{n}$, we have

$$\frac{1}{2\alpha_n}\mathsf{r}_\psi^2 + s(d)M^2 \sum_{k=1}^{n} \alpha_k \leq \frac{\mathsf{r}_\psi M\sqrt{s(d)}}{\alpha}\sqrt{n} + \alpha \mathsf{r}_\psi M\sqrt{s(d)}\sqrt{n} \leq 2\mathsf{r}_\psi M\sqrt{s(d)}\sqrt{n}\max\{\alpha, \alpha^{-1}\}.$$

For the second summation in the quantity (6.19), we have the bound

$$\alpha u^2 \left(\frac{M^2 s(d)}{L(P)^2 M(\mu)^2}\right) \frac{\mathsf{r}_\psi L(P)^2 M(\mu)^2}{4M\sqrt{s(d)}} \sum_{k=1}^{n} \frac{1}{k^{5/2}} \leq \alpha u^2 \mathsf{r}_\psi M\sqrt{s(d)}$$

since $\sum_{k=1}^{n} k^{-5/2} \leq 4$. The final term in the inequality (6.19) is similarly bounded by

$$u\left(\frac{M\sqrt{s(d)}}{L(P)M(\mu)}\right)\frac{\mathsf{r}_\psi L(P)M(\mu)}{2}(\log n + 1) = u\frac{\mathsf{r}_\psi M\sqrt{s(d)}}{2}(\log n + 1) \leq u\mathsf{r}_\psi M\sqrt{s(d)}\log(2n).$$

Combining the preceding inequalities with Jensen's inequality yields the claim (6.8).

6.5.2 Proof of Lemma 6.1

Let h be an arbitrary convex function with L_h-Lipschitz continuous gradient with respect to the norm $\|\cdot\|$. Using the tangent plane lower bound for a convex function and the L_h-Lipschitz continuity of the gradient, for any $u > 0$ we have

$$\begin{aligned} h'(\theta, z) = \frac{\langle \nabla h(\theta), uz\rangle}{u} &\leq \frac{h(\theta + uz) - h(\theta)}{u} \\ &\leq \frac{\langle \nabla h(\theta), uz\rangle + (L_h/2)\left\|uz\right\|^2}{u} = h'(\theta, z) + \frac{L_h u}{2}\left\|z\right\|^2. \end{aligned}$$

Consequently, for any point $\theta \in \operatorname{relint} \operatorname{dom} h$ and for any $z \in \mathbb{R}^d$, we have

$$\frac{h(\theta + uz) - h(\theta)}{u} z = h'(\theta, z)z + \frac{L_h u}{2} \|z\|^2 \gamma(u, \theta, z)z, \tag{6.20}$$

where γ is some function with range contained in $[0, 1]$. Since $\mathbb{E}[ZZ^\top] = I_{d\times d}$ by assumption, equality (6.20) implies

$$\mathbb{E}\left[\frac{h(\theta + uZ) - h(\theta)}{u} Z\right] = \mathbb{E}\left[h'(\theta, Z)Z + \frac{L_h u}{2} \|Z\|^2 \gamma(u, \theta, Z)Z\right] = \nabla h(\theta) + uL_h v(\theta, u), \tag{6.21}$$

where $v(\theta, u) \in \mathbb{R}^d$ is an error vector with $\|v(\theta, u)\|_* \leq \frac{1}{2}\mathbb{E}[\|Z\|^2 \|Z\|_*]$.

We now turn to proving the statements of the lemma. Recalling the definition (6.2) of the gradient estimator, we see that for P-almost every $x \in \mathcal{X}$, expression (6.21) implies that

$$\mathbb{E}[\mathsf{g}_{\text{sm}}(\theta; u, Z, x)] = \nabla F(\theta; x) + uL(x)v$$

for some vector v with $2\|v\|_* \leq \mathbb{E}[\|Z\|^2 \|Z\|_*]$. We have $\mathbb{E}[\nabla F(\theta; X)] = \nabla f(\theta^k)$, and independence implies that

$$\mathbb{E}[L(X) \|v\|_*] \leq \sqrt{\mathbb{E}[L(X)^2]}\sqrt{\mathbb{E}[\|v\|_*^2]} \leq \frac{1}{2}L(P)\mathbb{E}[\|Z\|^2 \|Z\|_*],$$

from which the bound (6.6) follows.

For the second statement (6.7) of the lemma, apply equality (6.20) to $F(\cdot; X)$, obtaining

$$\mathsf{g}_{\text{sm}}(\theta; u, Z, X) = \langle \mathsf{g}(\theta, X), Z\rangle Z + \frac{L(\theta)u}{2} \|Z\|^2 \gamma Z$$

for some function $\gamma \equiv \gamma(u, \theta, Z, X) \in [0, 1]$. The relation $(a + b)^2 \leq 2a^2 + 2b^2$ then gives

$$\begin{aligned}
\mathbb{E}[\|\mathsf{g}_{\text{sm}}(\theta; u, Z, X)\|_*^2] &\leq \mathbb{E}\left[\left(\|\langle \mathsf{g}(\theta, X), Z\rangle Z\|_* + \frac{1}{2}\left\|L(X)u \|Z\|^2 \gamma Z\right\|_*\right)^2\right] \\
&\leq 2\mathbb{E}\left[\|\langle \mathsf{g}(\theta, X), Z\rangle Z\|_*^2\right] + \frac{u^2}{2}\mathbb{E}\left[L(X)^2 \|Z\|^4 \|Z\|_*^2\right].
\end{aligned}$$

Finally, Assumption 6C coupled with the independence of X and Z gives the bound (6.7).

6.5.3 Proof of Theorem 6.2

The proof of Theorem 6.2 is similar to that of Theorem 6.1. To simplify our proof, we first state a lemma bounding the moments of vectors that satisfy Assumption 6E.

Lemma 6.4. *Let the random vector Z be distributed as $\mathsf{N}(0, I_{d\times d})$, uniformly on the ℓ_2-ball of radius $\sqrt{d+2}$, or uniformly on the ℓ_2-sphere of radius $\sqrt{d}$. For any $k \in \mathbb{N}$, there is a constant c_k (dependent only on k) such that*

$$\mathbb{E}\left[\|Z\|_2^k\right] \leq c_k d^{\frac{k}{2}}.$$

In all cases we have $\mathbb{E}[ZZ^\top] = I_{d\times d}$, and $c_k \leq 3$ for $k = 4$ and $c_k \leq \sqrt{3}$ for $k = 3$.

See Section 6.7.1 for the proof.

We now turn to the proof proper. From Lemmas 5.4 and 5.5 from the previous chapter, the function f_u defined in (6.10) satisfies $f(\theta) \leq f_u(\theta) \leq f(\theta) + uM\sqrt{d+2}$ for $\theta \in \Theta$. Defining the error vector $e^k := \nabla f_{u_{1,k}}(\theta^k) - g^k$ and noting that $\sqrt{d+2} \leq \sqrt{3d}$, we thus have

$$\begin{aligned}
\sum_{k=1}^n \left(f(\theta^k) - f(\theta^*)\right) &\leq \sum_{k=1}^n \left(f_{u_{1,k}}(\theta^k) - f_{u_{1,k}}(\theta^*)\right) + \sqrt{3}M\sqrt{d}\sum_{k=1}^n u_{1,k} \\
&\leq \sum_{k=1}^n \left\langle \nabla f_{u_{1,k}}(\theta^k), \theta^k - \theta^* \right\rangle + \sqrt{3}M\sqrt{d}\sum_{k=1}^n u_{1,k} \\
&= \sum_{k=1}^n \left\langle g^k, \theta^k - \theta^* \right\rangle + \sum_{k=1}^n \left\langle e^k, \theta^k - \theta^* \right\rangle + \sqrt{3}M\sqrt{d}\sum_{k=1}^n u_{1,k},
\end{aligned}$$

where we have used the convexity of f_u and the definition of e^k. Applying Lemma 6.3 to the summed $\left\langle g^k, \theta^k - \theta^* \right\rangle$ terms as in the proof of Theorem 6.1, we obtain

$$\sum_{k=1}^n \left(f(\theta^k) - f(\theta^*)\right) \leq \frac{\mathsf{r}_\psi^2}{2\alpha_n} + \frac{1}{2}\sum_{k=1}^n \alpha_k \left\|g^k\right\|_2^2 + \sum_{k=1}^n \left\langle e^k, \theta^k - \theta^* \right\rangle + \sqrt{3}M\sqrt{d}\sum_{k=1}^n u_{1,k}. \quad (6.22)$$

The proof from this point is similar to the proof of Theorem 6.1 (cf. inequality (6.18)). Specifically, we bound the squared gradient $\|g^k\|_2^2$ terms, the error $\left\langle e^k, \theta^k - \theta^* \right\rangle$ terms, and then control the summed u_k terms. For the remainder of the proof, we let $\mathcal{F}_{k-1}$ denote the σ-field generated by the random variables $X_1, \ldots, X_{k-1}$, $Z_1^1, \ldots, Z_1^{k-1}$, and $Z_2^1, \ldots, Z_2^{k-1}$.

Bounding $\left\langle e^k, \theta^k - \theta^* \right\rangle$: Our first step is note that Lemma 6.2 implies $\mathbb{E}[e^k \mid \mathcal{F}_{k-1}] = \frac{u_{2,k}}{u_{1,k}} M v_k$, where the vector $v_k \equiv v(\theta^k, u_{1,k}, u_{2,k})$ satisfies $\|v_k\|_2 \leq \frac{1}{2}\mathbb{E}[\|Z_2\|_2^3]$. As in the proof of Theorem 6.1, this gives

$$\sum_{k=1}^n \mathbb{E}[\left\langle e^k, \theta^k - \theta^* \right\rangle] \leq M \sum_{k=1}^n \frac{u_{2,k}}{u_{1,k}} \mathbb{E}[\|v_k\|_2 \left\|\theta^k - \theta^*\right\|_2] \leq \frac{1}{2}\mathbb{E}[\|Z_2\|_2^3]\, \mathsf{r}_\psi M \sum_{k=1}^n \frac{u_{2,k}}{u_{1,k}}.$$

When Assumption 6E holds, Lemma 6.4 implies the expectation bound $\mathbb{E}[\|Z_2\|_2^3] \leq \sqrt{3}d^{3/2}$. Thus

$$\sum_{k=1}^n \mathbb{E}[\left\langle e^k, \theta^k - \theta^* \right\rangle] \leq \frac{\sqrt{3}d\sqrt{d}}{2} \mathsf{r}_\psi M \sum_{k=1}^n \frac{u_{2,k}}{u_{1,k}}.$$

Bounding $\|g^k\|_2^2$: Turning to the squared gradient terms from the bound (6.22), Lemma 6.2 gives

$$\mathbb{E}\left[\left\|g^k\right\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|g^k\right\|_2^2 \mid \mathcal{F}_{k-1}\right]\right] \leq c\, M^2 d\left(\sqrt{\frac{u_{2,k}}{u_{1,k}}}\, d + 1 + \log d\right)$$
$$\leq c'\, M^2 d\left(\sqrt{\frac{u_{2,k}}{u_{1,k}}}\, d + \log(2d)\right),$$

where $c, c' > 0$ are numerical constants independent of $\{u_{1,k}\}$, $\{u_{2,k}\}$.

Summing out the smoothing penalties: Applying the preceding estimates to our earlier bound (6.22), we get that for a numerical constant c,

$$\sum_{k=1}^n \mathbb{E}\big[f(\theta^k) - f(\theta^*)\big] \leq \frac{\mathsf{r}_\psi^2}{2\alpha_n} + cM^2 d\log(2d)\sum_{k=1}^n \alpha_k \tag{6.23}$$
$$+ cM^2 d^2 \sum_{k=1}^n \sqrt{\frac{u_{2,k}}{u_{1,k}}}\,\alpha_k + \frac{\sqrt{3}}{2}\mathsf{r}_\psi M d\sqrt{d}\sum_{k=1}^n \frac{u_{2,k}}{u_{1,k}} + \sqrt{3}M\sqrt{d}\sum_{k=1}^n u_{1,k}.$$

We bound the right hand side above using our choices of α_k, $u_{1,k}$, and $u_{2,k}$. We also use the relations $\sum_{k=1}^n k^{-\frac{1}{2}} \leq 2\sqrt{n}$ and $\sum_{k=1}^n k^{-1} \leq 1 + \log n \leq 2\log n$ for $n \geq 3$. With the setting $\alpha_k = \alpha\mathsf{r}_\psi/(M\sqrt{d\log(2d)}\sqrt{k})$, the first two terms in (6.23) become

$$\frac{\mathsf{r}_\psi^2}{2\alpha_n} + cM^2 d\log(2d)\sum_{k=1}^n \alpha_k \leq \frac{\mathsf{r}_\psi M\sqrt{d\log(2d)}}{2\alpha}\sqrt{n} + 2c\alpha\mathsf{r}_\psi M\sqrt{d\log(2d)}\sqrt{n}$$
$$\leq c'\max\{\alpha, \alpha^{-1}\}\mathsf{r}_\psi M\sqrt{d\log(2d)}\sqrt{n}$$

for a universal constant c'. Since we have chosen $u_{2,k}/u_{1,k} = 1/(d^2 k)$, we may bound the third term in expression (6.23) by

$$cM^2 d^2\sum_{k=1}^n \sqrt{\frac{u_{2,k}}{u_{1,k}}}\,\alpha_k = cM^2 d^2\left(\frac{\alpha\mathsf{r}_\psi}{M\sqrt{d\log(2d)}}\right)\frac{1}{d}\sum_{k=1}^n \frac{1}{k} \leq \frac{c'\alpha\mathsf{r}_\psi M\sqrt{d}}{\sqrt{\log(2d)}}\log(2n).$$

Similarly, the fourth term in the bound (6.23) becomes

$$\frac{\sqrt{3}}{2}\mathsf{r}_\psi M d\sqrt{d}\sum_{k=1}^n \frac{u_{2,k}}{u_{1,k}} = \frac{\sqrt{3}}{2}\mathsf{r}_\psi M d\sqrt{d}\,\frac{1}{d^2}\sum_{k=1}^n \frac{1}{k} \leq \frac{\sqrt{3}\mathsf{r}_\psi M}{\sqrt{d}}\log(2n).$$

Finally, since $u_{1,k} = u\mathsf{r}_\psi/k$, we may bound the last term in expression (6.23) with

$$\sqrt{3}M\sqrt{d}\sum_{k=1}^n u_{1,k} = \sqrt{3}M\sqrt{d}\,u\mathsf{r}_\psi\sum_{k=1}^n \frac{1}{k} \leq 2\sqrt{3}u\mathsf{r}_\psi M\sqrt{d}\log(2n).$$

Using Jensen's inequality to note that $\mathbb{E}[f(\widehat{\theta}(n)) - f(\theta^*)] \leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[f(\theta^k) - f(\theta^*)\right]$ and eliminating lower-order terms, we obtain the claim (6.15).

6.5.4 Proof of Lemma 6.2

The proof of Lemma 6.2 relies on the following key technical result:

Lemma 6.5. *Let $k \geq 1$ and $u \geq 0$. Let $Z_1 \sim \mu_1$ and $Z_2 \sim \mu_2$ be independent random variables in $\mathbb{R}^d$, where μ_1 and μ_2 satisfy Assumption 6E. There exists a constant c_k, depending only on k, such that for every 1-Lipschitz convex function h,*

$$\mathbb{E}\left[|h(Z_1 + uZ_2) - h(Z_1)|^k\right] \leq c_k u^k \left[ud^{\frac{k}{2}} + 1 + \log^{\frac{k}{2}}(d + 2k)\right].$$

The proof is fairly technical, so we defer it to section 6.7.2. It is based on the dimension-free concentration of Lipschitz functions of standard Gaussian vectors and vectors uniform on the ℓ_2 ball.

We return now to the proof of Lemma 6.2 proper, providing arguments for inequalities (6.13) and (6.14). For convenience we recall the definition $M(x)$ as the Lipschitz constant of $F(\cdot; x)$ (Assumption 6B′) and the definition (6.11) of the non-smooth directional gradient

$$\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, z_1, z_2, x) = \frac{F(\theta + u_1 z_1 + u_2 z_2; x) - F(\theta + u_1 z_1; x)}{u_2} z_2.$$

We begin with the second statement (6.14) of Lemma 6.2. By applying Lemma 6.5 to the 1-Lipschitz convex function $h(w) = \frac{1}{u_1 M(X)} F(\theta + u_1 w; X)$ and setting $u = u_2/u_1$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, Z_1, Z_2, x)\|_2^2\right] &= \frac{u_1^2 M(x)^2}{u_2^2} \mathbb{E}\left[\left(h(Z_1 + (u_2/u_1)Z_2) - h(Z_1)\right)^2 \|Z_2\|_2^2\right] \\
&\leq \frac{M(x)^2}{u^2} \mathbb{E}\left[\left(h(Z_1 + uZ_2) - h(Z_1)\right)^4\right]^{\frac{1}{2}} \mathbb{E}\left[\|Z_2\|_2^4\right]^{\frac{1}{2}}. \qquad (6.24)
\end{aligned}$$

Lemma 6.4 implies that $\mathbb{E}[\|Z_2\|_2^4]^{\frac{1}{2}} \leq \sqrt{3}d$ for smoothing distributions satisfying Assumption 6E.

It thus remains to bound the first expectation in the product (6.24). By Lemma 6.5,

$$\mathbb{E}\left[\left(h(Z_1 + uZ_2) - h(Z_1)\right)^4\right] \leq cu^4 \left[ud^2 + 1 + \log^2 d\right]$$

for a numerical constant $c > 0$. Taking the square root of both sides of the preceding display, then applying inequality (6.24), yields

$$\mathbb{E}\left[\|\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, Z_1, Z_2, x)\|_2^2\right] \leq c \frac{M(x)^2}{u^2} u^2 d \left[\sqrt{u}d + 1 + \log d\right].$$

Integrating over x using the Lipschitz Assumption 6B′ proves the inequality (6.14).

For the first statement of the lemma, we define the shorthand $F_u(\theta; x) = \mathbb{E}[F(\theta + uZ_1; x)]$, where the expectation is over $Z_1 \sim \mu_1$, and note that by Fubini's theorem, $\mathbb{E}[F_u(\theta; X)] = f_u(\theta)$. By taking the expectation of g_{ns} with respect to Z_1 only, we get

$$\mathbb{E}\left[\mathsf{g}_{\text{ns}}(\theta; u_1, u_2, Z_1, z_2, x)\right] = \frac{F_{u_1}(\theta + u_2 z_2; x) - F_{u_1}(\theta; x)}{u_2} z_2.$$

Since $\theta \mapsto F(\theta; x)$ is $M(x)$-Lipschitz, Lemmas 5.4(iii) and 5.5(iii) of the previous chapter imply $F_u(\cdot; x)$ is $M(x)$-Lipschitz, has $M(x)/u$-Lipschitz continuous gradient, and satisfies the unbiasedness condition $\mathbb{E}[\nabla F_u(\theta; X)] = \nabla f_u(\theta)$. Therefore, the same argument bounding the bias (6.6) in the proof of Lemma 6.1 (recall inequalities (6.20) and (6.21)) yields the claim (6.13).

6.6 Proofs of lower bounds

We now present the proofs for our lower bounds on the minimax error (6.16). Our lower bounds follow the techniques outlined in Chapter 2, specifically Section 2.2.4 on Assouad's method, where we reduce the optimization problem to several binary hypothesis testing problems. Specifically, as described in Section 2.2.4, we choose a finite set of functions, show that optimizing well implies that one can solve each of the hypothesis tests, and then, as in statistical minimax theory [185, 188, 173, 9], apply divergence-based lower bounds for the probability of error in hypothesis testing problems. Our proofs are similar and somewhat inspired by recent work of Arias-Castro et al. [9] and Shamir [163].

6.6.1 Proof of Proposition 6.1

The basic outline of both of our proofs is similar to the proof of Proposition 3.4 in Section 3.5.5, which builds off of the strengthened version of Assouad's method (Lemma 2.2).

In detail, we proceed as follows, giving a separation lower bound of the form (2.17) to be able to apply the techniques of the sharper Assouad's method developed in Lemma 2.2 via the canonical multiple binary hypothesis testing problem. Consider (instantaneous) objective functions of the form $F(\theta; x) = \langle \theta, x \rangle$. Let $\mathcal{V} = \{-1, 1\}^d$ denote the Boolean hypercube, and for each $v \in \mathcal{V}$, let P_v denote the Gaussian distribution $\mathsf{N}(\delta v, \sigma^2 I_{d \times d})$, where $\delta > 0$ is a parameter to be chosen. Then the risk functionals defined as

$$f_v(\theta) := \mathbb{E}_{P_v}[F(\theta; X)] = \delta \langle \theta, v \rangle$$

are "well-separated" enough to apply Assouad's method, as we formalize presently. For each $v \in \mathcal{V}$, we define $\theta^v = \operatorname{argmin}_{\theta \in \Theta} f_v(\theta)$, where $\Theta := \{\theta \in \mathbb{R}^d \mid \|\theta\|_q \le \mathsf{r}_q\}$. A calculation shows that $\theta^v = -\mathsf{r}_q \, d^{1/q} \, v$, so that $\operatorname{sign}(\theta^v_j) = -v_j$. Next we claim that, for any vector $\widehat{\theta} \in \mathbb{R}^d$,

$$f_v(\widehat{\theta}) - f_v(\theta^v) \ge \frac{1 - 1/q}{d^{1/q}} \delta \mathsf{r}_q \sum_{j=1}^{d} \mathbf{1} \left\{ \operatorname{sign}(\widehat{\theta}_j) \ne \operatorname{sign}(\theta^v_j) \right\}. \tag{6.25}$$

Inequality (6.25) shows that if it is possible to optimize well—that is, to find a vector $\widehat{\theta}$ with a relatively small optimality gap—then it is also possible to estimate the signs of v, and it is our analogue of the risk separation (2.17) necessary for our applications of the sharpened Assouad method. To establish inequality (6.25), we first state a simple lemma:

Lemma 6.6. *For a given integer $i \in [d]$, consider the two optimization problems (over $\theta \in \mathbb{R}^d$)*

$$(A) \quad \begin{array}{ll} \text{minimize} & \theta^\top \mathbb{1} \\ \text{subject to} & \|\theta\|_q \le 1 \end{array} \quad \text{and} \quad (B) \quad \begin{array}{ll} \text{minimize} & \theta^\top \mathbb{1} \\ \text{subject to} & \|\theta\|_q \le 1, \ \theta_j \ge 0 \text{ for } j \in [i], \end{array}$$

with optimal solutions θ^A and θ^B, respectively. Then $\langle \mathbb{1}, \theta^A \rangle \le \langle \mathbb{1}, \theta^B \rangle - (1 - 1/q) i / d^{1/q}$.

See Section 6.8.1 for a proof. Returning to inequality (6.25), we note that $f_v(\widehat{\theta}) - f_v(\theta^v) = \delta \langle v, \widehat{\theta} - \theta^v \rangle$. By symmetry, Lemma 6.6 implies that for every coordinate j such that $\operatorname{sign}(\widehat{\theta}_j) \neq \operatorname{sign}(\theta^v_j)$, the objective value $f_v(\widehat{\theta})$ must be at least a quantity $(1 - 1/q)\delta \mathsf{r}_q / d^{1/q}$ larger than the optimal value $f_v(\theta^v)$, which yields inequality (6.25).

Now we use inequality (6.25) to give a probabilistic lower bound. Consider the mixture distribution $\mathbb{P} := (1/|\mathcal{V}|) \sum_{v \in \mathcal{V}} P_v$. For any estimator $\widehat{\theta}$, we have

$$\max_v \mathbb{E}_{P_v}[f_v(\widehat{\theta}) - f_v(\theta^v)] \ge \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v}[f_v(\widehat{\theta}) - f_v(\theta^v)] \ge \frac{1 - 1/q}{d^{1/q}} \delta \mathsf{r}_q \sum_{j=1}^d \mathbb{P}(\operatorname{sign}(\widehat{\theta}_j) \neq -V_j).$$

Consequently, in parallel to Lemma 2.2, the minimax error is lower bounded as

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,2}, F, \mathcal{C}_n^{\mathsf{zo}}) \ge \frac{1 - 1/q}{d^{1/q}} \delta \, \mathsf{r}_q \Big\{ \inf_\Psi \sum_{j=1}^d \mathbb{P}(\Psi_j(Y^1, \ldots, Y^n) \neq V_j) \Big\}, \tag{6.26}$$

where Ψ denotes any testing function mapping from the observations $\{Y^k\}_{k=1}^n$ to $\{-1, 1\}^d$.

Next we lower bound the testing error by a total variation distance. By Le Cam's inequality (2.6), for any set A and distributions P, Q, we have $P(A) + Q(A^c) \ge 1 - \|P - Q\|_{\mathrm{TV}}$. We apply this inequality to the "positive jth coordinate" and "negative jth coordinate" sampling distributions

$$P_{+j} := \frac{1}{2^{d-1}} \sum_{v \in \mathcal{V} : v_j = 1} P_v \quad \text{and} \quad P_{-j} := \frac{1}{2^{d-1}} \sum_{v \in \mathcal{V} : v_j = -1} P_v,$$

corresponding to conditional distributions over Y^k given the events $\{v_j = 1\}$ or $\{v_j = -1\}$. Applying Le Cam's inequality yields

$$\mathbb{P}(\Psi_j(Y^{1:n}) \neq V_j) = \frac{1}{2} P_{+j}(\Psi_j(Y^{1:n}) \neq 1) + \frac{1}{2} P_{-j}(\Psi_j(Y^{1:n}) \neq -1) \ge \frac{1}{2} \left(1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}}\right).$$

Combined with the upper bound $\sum_{j=1}^d \|P_{+j} - P_{-j}\|_{\mathrm{TV}} \le \sqrt{d} (\sum_{j=1}^d \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2)^{\frac{1}{2}}$ via the Cauchy-Schwartz inequality, we obtain (recall inequality (2.18))

$$\begin{aligned} \mathfrak{M}_n(\Theta, \mathcal{P}_{M,2}, F, \mathcal{C}_n^{\mathsf{zo}}) &\ge \left(1 - \frac{1}{q}\right) \frac{\delta \mathsf{r}_q}{2 d^{1/q}} \sum_{j=1}^d \left(1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}}\right) \\ &\ge \left(1 - \frac{1}{q}\right) \frac{d^{1 - 1/q} \delta \mathsf{r}_q}{2} \left(1 - \frac{1}{\sqrt{d}} \left(\sum_{j=1}^d \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 \right)^{\frac{1}{2}} \right). \end{aligned} \tag{6.27}$$

The remainder of the proof provides sharp enough bounds on $\sum_j \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2$ to leverage inequality (6.27). Define the covariance matrix

$$\Sigma := \sigma^2 \begin{bmatrix} \|\theta\|_2^2 & \langle \theta, w \rangle \\ \langle \theta, w \rangle & \|w\|_2^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} \theta & w \end{bmatrix}^\top \begin{bmatrix} \theta & w \end{bmatrix}, \tag{6.28}$$

with the corresponding shorthand Σ^k for the covariance computed for the kth pair (θ^k, w^k). We have:

Lemma 6.7. *For each $j \in \{1, \ldots, d\}$, the total variation norm is bounded as*

$$\|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 \le \delta^2 \sum_{k=1}^n \mathbb{E}\left[\begin{bmatrix} \theta_j^k \\ w_j^k \end{bmatrix}^\top (\Sigma^k)^{-1} \begin{bmatrix} \theta_j^k \\ w_j^k \end{bmatrix} \right]. \tag{6.29}$$

See Section 6.8.2 for a proof of this lemma.

Now we use the bound (6.29) to provide a further lower bound on inequality (6.27). We first note the identity

$$\sum_{j=1}^d \begin{bmatrix} \theta_j \\ w_j \end{bmatrix} \begin{bmatrix} \theta_j \\ w_j \end{bmatrix}^\top = \begin{bmatrix} \|\theta\|_2^2 & \langle \theta, w \rangle \\ \langle \theta, w \rangle & \|w\|_2^2 \end{bmatrix}.$$

Recalling the definition (6.28) of the covariance matrix Σ, Lemma 6.7 implies that

$$\begin{aligned} \sum_{j=1}^d \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 &\le \delta^2 \sum_{k=1}^n \mathbb{E}\Big[\sum_{j=1}^d \mathrm{tr}\left((\Sigma^k)^{-1} \begin{bmatrix} \theta_j^k \\ w_j^k \end{bmatrix} \begin{bmatrix} \theta_j^k \\ w_j^k \end{bmatrix}^\top \right) \Big] \\ &= \frac{\delta^2}{\sigma^2} \sum_{k=1}^n \mathbb{E}\left[\mathrm{tr}\left((\Sigma^k)^{-1} \Sigma^k \right) \right] = 2 \frac{n\delta^2}{\sigma^2}. \end{aligned} \tag{6.30}$$

Returning to the estimation lower bound (6.27), we thus find the nearly final lower bound

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,2}, F, \mathcal{C}_n^{\mathsf{zo}}) \ge \left(1 - \frac{1}{q}\right) \frac{d^{1-1/q} \delta \mathsf{r}_q}{2} \left(1 - \left(\frac{2n\delta^2}{d\sigma^2}\right)^{\frac{1}{2}}\right). \tag{6.31}$$

The last thing we must do is enforce that $P \in \mathcal{P}_{M,2}$, which amounts to choosing the parameters σ^2 and δ^2 so that $\mathbb{E}[\|X\|_2^2] \le M^2$ for $X \sim \mathsf{N}(\delta v, \sigma^2 I_{d \times d})$, after which we may use inequality (6.31) to complete the proof of the lower bound. By construction, we have $\mathbb{E}[\|X\|_2^2] = (\delta^2 + \sigma^2)d$, so choosing $\sigma^2 = 8M^2/9d$ and $\delta^2 = (M^2/9) \min\{1/n, 1/d\}$ guarantees that

$$1 - \left(\frac{2n\delta^2}{d\sigma^2}\right)^{\frac{1}{2}} \ge 1 - \left(\frac{18}{72}\right)^{\frac{1}{2}} = \frac{1}{2} \quad \text{and} \quad \mathbb{E}[\|X\|_2^2] = \frac{8M^2}{9} + \frac{M^2 d}{9} \min\left\{\frac{1}{n}, \frac{1}{d}\right\} \le M^2.$$

Substituting these choices of δ and σ^2 in inequality (6.31) gives the lower bound

$$\begin{aligned}\mathfrak{M}_n(\Theta, \mathcal{P}_{M,2}, F, \mathcal{C}_n^{\mathsf{zo}}) &\geq \frac{1}{12}\left(1 - \frac{1}{q}\right) d^{1-1/q} \mathsf{r}_q M \min\left\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{d}}\right\} \\ &= \frac{1}{12}\left(1 - \frac{1}{q}\right) \frac{d^{1-1/q}\mathsf{r}_q M}{\sqrt{n}} \min\left\{1, \sqrt{n/d}\right\}.\end{aligned}$$

To complete the proof of the claim (6.17), we note that the above lower bound also applies to any d_0-dimensional problem for $d_0 \leq d$. More rigorously, we choose $\mathcal{V} = \{-1, 1\}^{d_0} \times \{0\}^{d-d_0}$, and define the sampling distribution P_v on X so that given $v \in \mathcal{V}$, the coordinate distributions of X are independent with $X_j \sim \mathsf{N}(\delta v_j, \sigma^2)$ for $j \leq d_0$ and $X_j = 0$ for $j > d_0$. A reproduction of the preceding proof, substituting $d_0 \leq d$ for each appearance of the dimension d, then yields the claimed bound (6.17).

6.6.2 Proof of Proposition 6.2

The proof is similar to that of Proposition 6.1, except instead of using the set $\mathcal{V} = \{-1, 1\}^d$, we use the $2d$ standard basis vectors and their negatives, that is, $\mathcal{V} = \{\pm e_j\}_{j=1}^d$. We use the same sampling distributions as in the proof of Proposition 6.1, so under P_v the random vectors $X \sim \mathsf{N}(\delta v, \sigma^2 I_{d\times d})$, and we have $f_v = \mathbb{E}_{P_v}[F(\theta; X)] = \delta \langle \theta, v \rangle$. Let us define P_j to be the distribution P_v for $v = e_j$ and similarly for P_{-j}, and let

$$\theta^v = \operatorname*{argmin}_\theta \{f_v(\theta) \mid \|\theta\|_1 \leq \mathsf{r}_1\} = -\mathsf{r}_1 v.$$

We now provide the reduction from optimization to testing, which is similar to our previous uses of Assouad's method, but somewhat different as we use $\mathcal{V} = \{\pm e_j\}_{j=1}^d$. First, if $v = \pm e_j$, then any estimator $\widehat{\theta}$ satisfying $\operatorname{sign}(\widehat{\theta}_j) \neq \operatorname{sign}(\theta_j^v)$ must have $f_v(\widehat{\theta}) - f_v(\theta^v) \geq \delta \mathsf{r}_1$. Defining the coordinate sign function $\operatorname{sgn}_j(x) := \operatorname{sign}(x_j)$, we see that for $v \in \{\pm e_j\}$,

$$f_v(\widehat{\theta}) - f_v(\theta^v) \geq \delta\, \mathsf{r}_1\, \mathbf{1}\left\{\operatorname{sgn}_j(\widehat{\theta}) \neq \operatorname{sgn}_j(\theta^v)\right\}.$$

Consequently, we obtain the multiple binary hypothesis testing lower bound

$$\begin{aligned}\max_v \mathbb{E}_{P_v}[f_v(\widehat{\theta}) - f_v(\theta^v)] &\geq \frac{1}{2d}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}[f_v(\widehat{\theta}) - f_v(\theta^v)] \\ &\geq \frac{\delta \mathsf{r}_1}{2d}\sum_{j=1}^d \left[P_j(\operatorname{sgn}_j(\widehat{\theta}) \neq -1) + P_{-j}(\operatorname{sgn}_j(\widehat{\theta}) \neq 1)\right] \overset{(i)}{\geq} \frac{\delta \mathsf{r}_1}{2d}\sum_{j=1}^d \left[1 - \|P_j - P_{-j}\|_{\mathrm{TV}}\right].\end{aligned}$$

For the final inequality (i), we applied Le Cam's inequality as in the proof of Proposition 6.1. Thus, as in the derivation of inequality (6.27) from the Cauchy-Schwarz inequality, this yields

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,\infty}, F, \mathcal{C}_n^{\mathsf{zo}}) \geq \frac{\delta \mathsf{r}_1}{2}\left(1 - \frac{1}{\sqrt{d}}\left(\sum_{j=1}^d \|P_j - P_{-j}\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}}\right). \tag{6.32}$$

We now turn to providing a bound on $\sum_{j=1}^{d} \|P_j - P_{-j}\|_{\mathrm{TV}}^2$ analogous to that in the proof of Proposition 6.1. We claim that

$$\sum_{j=1}^{d} \|P_j - P_{-j}\|_{\mathrm{TV}}^2 \leq 2\frac{n\delta^2}{\sigma^2}. \tag{6.33}$$

Inequality (6.33) is nearly immediate from Lemma 6.7. Indeed, given the pair $W = [\theta \; w] \in \mathbb{R}^{d\times 2}$, the observation $Y = W^\top X$ is distributed (conditional on v and W) as $\mathsf{N}(\delta W^\top v, \Sigma)$ where $\Sigma = \sigma^2 W^\top W$ is the covariance (6.28). For $v = e_j$ and $v' = -e_j$, we know that $\langle \theta, v - v' \rangle = 2\theta_j$ and so

$$D_{\mathrm{kl}}\left(\mathsf{N}(\delta W^\top v, \Sigma) \| \mathsf{N}(\delta W^\top v', \Sigma)\right) = 2\delta^2 \begin{bmatrix} \theta_j \\ w_j \end{bmatrix}^\top \Sigma^{-1} \begin{bmatrix} \theta_j \\ w_j \end{bmatrix}.$$

By analogy with the proof of Lemma 6.7, we may repeat the derivation of inequalities (6.29) and (6.30) *mutatis mutandis* to obtain inequality (6.33). Combining inequalities (6.32) and (6.33) then gives the lower bound

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,\infty}, F, \mathcal{C}_n^{\mathsf{zo}}) \geq \frac{\delta \mathsf{r}_1}{2}\left(1 - \left(\frac{2\delta^2 n}{d\sigma^2}\right)^{\frac{1}{2}}\right).$$

It thus remains to choose δ and σ^2 to guarantee the containment $P \in \mathcal{P}_{M,\infty}$. Equivalently, we must establish the gradient bound $\mathbb{E}[\|X\|_\infty^2] \leq M^2$.

Lemma 6.8. *Given any vector with $\|v\|_\infty \leq 1$, and the random vector $X \sim \mathsf{N}(\delta v, \sigma^2 I_{d\times d})$, we have*

$$\mathbb{E}[\|X\|_\infty^2] \leq 2\sigma^2(3 + \log d) + 2\delta^2.$$

Proof The vector $Z = X - \delta v$ has $\mathsf{N}(0, \sigma^2 I_{d\times d})$ distribution. Letting $(X_1, \ldots, X_d)$ and $(Z_1, \ldots, Z_d)$ denote the components of X and Z, respectively, we see that $X_j^2 \leq 2Z_j^2 + 2\delta^2 v_j^2$, so

$$\|X\|_\infty^2 \leq 2\max\{Z_1^2, \ldots, Z_d^2\} + 2\delta^2 \max\{v_1^2, \ldots, v_d^2\} \leq 2\|Z\|_\infty^2 + 2\delta^2.$$

Each Z_j is a random variable with $\mathsf{N}(0, \sigma^2)$ distribution, and standard results [36, Chapter 2] imply that $\mathbb{E}[\|Z\|_\infty^2] \leq \sigma^2(1 + \log(3\sqrt{3}d))$, from which the lemma follows. $\square$

As a consequence of Lemma 6.8, by taking

$$\sigma^2 = \frac{4M^2}{9(3 + \log d)} \quad \text{and} \quad \delta^2 = \frac{M^2}{18(3 + \log d)} \min\left\{1, \frac{d}{n}\right\}$$

we obtain the bounds

$$\mathbb{E}[\|X\|_\infty^2] \leq \frac{8M^2}{9} + \frac{2M^2}{18} = M^2 \quad \text{and} \quad 1 - \left(\frac{2\delta^2 n}{d\sigma^2}\right)^{\frac{1}{2}} \geq 1 - \left(\frac{18}{72}\right)^{\frac{1}{2}} = \frac{1}{2}.$$

Noting that $\sqrt{18} = 3\sqrt{2}$ and substituting into the lower bound on $\mathfrak{M}_n$ yields

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,\infty}, F, \mathcal{C}_n^{\mathsf{zo}}) \geq \frac{1}{12\sqrt{2}\sqrt{3 + \log d}} \frac{M\mathsf{r}_1}{\sqrt{n}} \min\left\{\sqrt{n}, \sqrt{d}\right\}.$$

Modulo this lower bound holding for each dimension $d_0 \leq d$, this completes the proof.

To complete the proof, we note that as in the proof of Proposition 6.1, we may provide a lower bound on the optimization error for any $d_0 \leq d$-dimensional problem. In particular, fix $d_0 \leq d$ and let $\mathcal{V} = \{\pm e_j\}_{j=1}^{d_0} \subset \mathbb{R}^d$. Now, conditional on $v \in \mathcal{V}$, let P_v denote the distribution on X with independent coordinates whose distributions are $X_j \sim \mathsf{N}(\delta v_j, \sigma^2)$ for $j \leq d_0$ and $X_j = 0$ for $j > d_0$. As in the proof Proposition 6.1, we may reproduce the preceding arguments by substituting $d_0 \leq d$ for every appearance of the dimension d, giving that for all $d_0 \leq d$,

$$\mathfrak{M}_n(\Theta, \mathcal{P}_{M,\infty}, F, \mathcal{C}_n^{\mathsf{zo}}) \geq \frac{1}{12\sqrt{2}\sqrt{3 + \log d_0}} \frac{M\mathsf{r}_1}{\sqrt{n}} \min\left\{\sqrt{n}, \sqrt{d_0}\right\}.$$

Choosing $d_0 = \min\{d, n\}$ completes the proof of Proposition 6.2.

6.7 Technical results for convergence arguments

In this section, we collect the proofs of the various lemmas used in our convergence arguments. Throughout the section, we recall that the notation B_2 denotes the ℓ_2-ball of radius 1, and $B_2(x, u) = x + uB_2$ denotes the ℓ_2-ball of radius u centered at x. (We let B_2^d denote the d-dimensional ball if we wish to make the dimension explicit.)

6.7.1 Proof of Lemma 6.4

We consider each of the distributions in turn. When Z has $\mathsf{N}(0, I_{d\times d})$ distribution, standard χ^2-distributed random variable calculations imply

$$\mathbb{E}\left[\|Z\|_2^k\right] = 2^{\frac{k}{2}} \frac{\Gamma(\frac{k}{2} + \frac{d}{2})}{\Gamma(\frac{d}{2})}.$$

That $\mathbb{E}[ZZ^\top] = I_{d\times d}$ is immediate, and the constant values c_k for $k \leq 4$ follow from direct calculations. For samples Z from the ℓ_2-sphere, it is clear that $\|Z\|_2 = \sqrt{d}$, so we may take $c_k = 1$ in the statement of the lemma. When $Z \sim \text{Uniform}(B_2^d)$, the density $p(t)$ of $\|Z\|_2$ is given by $d \cdot t^{d-1}$; consequently, for any $k > -d$ we have

$$\mathbb{E}[\|Z\|_2^k] = \int_0^1 t^k p(t)\, dt = d \int_0^1 t^{d+k-1}\, dt = \frac{d}{d+k}. \tag{6.34}$$

Thus for $Z \sim \text{Uniform}(\sqrt{d+2}\, B_2^d)$ we have $\mathbb{E}[ZZ^\top] = I_{d\times d}$, and $\mathbb{E}[\|Z\|_2^k] = (d+2)^{k/2} d/(d+k)$.

6.7.2 Proof of Lemma 6.5

The proof of Lemma 6.5 is based on a sequence of auxiliary results. Since the Lipschitz continuity of h implies the result for $d = 1$ directly, we focus on the case $d \geq 2$. First, we have the following standard result on the dimension-independent concentration of rotationally symmetric sub-Gaussian random vectors. We use this to prove that the perturbed h is close to the unperturbed h with high probability.

Lemma 6.9 (Rotationally invariant concentration)**.** *Let Z be a random variable in $\mathbb{R}^d$ having one of the following distributions: $\mathsf{N}(0, I_{d\times d})$, Uniform$(\sqrt{d+2}\, B_2^d)$, or Uniform$(\sqrt{d}\, \mathbb{S}^{d-1})$. There is a universal (numerical) constant $c > 0$ such that for any M-Lipschitz continuous function h,*

$$\mathbb{P}\left(|h(\theta + uZ) - \mathbb{E}[h(\theta + uZ)]| > \epsilon\right) \leq 2\exp\left(-\frac{c\,\epsilon^2}{M^2}\right).$$

In the case of the normal distribution, we may take $c = \frac{1}{2}$.

These results are standard (e.g., see Propositions 1.10 and 2.9 of Ledoux [117]).

Our next result shows that integrating out Z_2 leaves us with a smoother deviation problem, at the expense of terms of order at most $u^k \log^{k/2}(d)$. To state the lemma, we define the difference function $\Delta_u(\theta) = \mathbb{E}[h(\theta + uZ_2)] - h(\theta)$. Note that since h is convex and $\mathbb{E}[Z_2] = 0$, Jensen's inequality implies $\Delta_u(\theta) \geq 0$.

Lemma 6.10. *Under the conditions of Lemma 6.5, we have*

$$\mathbb{E}\left[|h(Z_1 + uZ_2) - h(Z_1)|^k\right] \leq 2^{k-1}\mathbb{E}[\Delta_u(Z_1)^k] + c^{-\frac{k}{2}}2^{k-1}k^{\frac{k}{2}}u^k \log^{\frac{k}{2}}(d+2k) + \sqrt{2}u^k$$

for any $k \geq 1$. Here c is the same constant in Lemma 6.9.

Proof For each $\theta \in \Theta$, the function $w \mapsto h(\theta + uw)$ is u-Lipschitz, so that Lemma 6.9 implies that

$$\mathbb{P}\left(\left|h(\theta + uZ_2) - \mathbb{E}[h(\theta + uZ_2)]\right| > \epsilon\right) \leq 2\exp\left(-\frac{c\,\epsilon^2}{u^2}\right).$$

On the event $A_\theta(\epsilon) := \{|h(\theta + uZ_2) - \mathbb{E}[h(\theta + uZ_2)]| \leq \epsilon\}$, we have

$$|h(\theta + uZ_2) - h(\theta)|^k \leq 2^{k-1}\,|h(\theta + uZ_2) - \mathbb{E}[h(\theta + uZ_2)|^k + 2^{k-1}\Delta_u(\theta)^k \leq 2^{k-1}\epsilon^k + 2^{k-1}\Delta_u(\theta)^k,$$

which implies

$$\mathbb{E}\left[|h(\theta + uZ_2) - h(\theta)|^k \cdot \mathbf{1}\left\{A_\theta(\epsilon)\right\}\right] \leq 2^{k-1}\Delta_u(\theta)^k + 2^{k-1}\epsilon^k. \tag{6.35a}$$

On the complement $A_\theta^c(\epsilon)$, which occurs with probability at most $2\exp(-c\epsilon^2/u^2)$, we use the Lipschitz continuity of h and Cauchy-Schwarz inequality to obtain

$$\mathbb{E}\left[|h(\theta + uZ_2) - h(\theta)|^k \cdot \mathbf{1}\left\{A_\theta(\epsilon)^c\right\}\right] \leq \mathbb{E}\left[u^k\,\|Z_2\|_2^k \cdot \mathbf{1}\left\{A_\theta(\epsilon)^c\right\}\right] \leq u^k\mathbb{E}[\|Z_2\|_2^{2k}]^{\frac{1}{2}} \cdot \mathbb{P}\left(A_\theta(\epsilon)^c\right)^{\frac{1}{2}}.$$

By direct calculations, Assumption 6E implies that $\mathbb{E}[\|Z_2\|_2^{2k}] \le (d+2k)^k$. Thus,

$$\mathbb{E}\left[|h(\theta+uZ_2)-h(\theta)|^k \cdot \mathbf{1}\left\{A_\theta(\epsilon)^c\right\}\right] \le u^k(d+2k)^{\frac{k}{2}} \cdot \sqrt{2}\exp\left(-\frac{c\,\epsilon^2}{2u^2}\right). \tag{6.35b}$$

Combining the estimates (6.35a) and (6.35b) gives

$$\mathbb{E}\left[|h(\theta+uZ_2)-h(\theta)|^k\right] \le 2^{k-1}\Delta_u(\theta)^k + 2^{k-1}\epsilon^k + \sqrt{2}u^k(d+2k)^{\frac{k}{2}}\exp\left(-\frac{c\,\epsilon^2}{2u^2}\right).$$

Setting $\epsilon^2 = \frac{k}{c}u^2\log(d+2k)$ and taking expectations over $Z_1 \sim \mu_1$ gives Lemma 6.10. □

By Lemma 6.10, it suffices to control the bias $\mathbb{E}[\Delta_u(Z_1)] = \mathbb{E}[h(Z_1+uZ_2)-h(Z_1)]$. The following result allows us to reduce this problem to one of bounding a certain one-dimensional expectation.

Lemma 6.11. *Let Z and W be random variables in $\mathbb{R}^d$ with rotationally invariant distributions and finite first moments. Let $\mathcal{H}$ denote the set of 1-Lipschitz convex functions $h\colon \mathbb{R}^d \to \mathbb{R}$, and for $h \in \mathcal{H}$, define $V(h) = \mathbb{E}[h(W)-h(Z)]$. Then*

$$\sup_{h\in\mathcal{H}} V(h) = \sup_{a\in\mathbb{R}_+} \mathbb{E}\left[\,|\,\|W\|_2 - a| - |\,\|Z\|_2 - a|\right].$$

Proof First, we note that $V(h) = V(h\circ U)$ for any unitary transformation U; since V is linear, if we define $\hat{h}$ as the average of $h\circ U$ over all unitary U then $V(h) = V(\hat{h})$. Moreover, for $h \in \mathcal{H}$, we have $\hat{h}(\theta) = \hat{h}_1(\|\theta\|_2)$ for some $\hat{h}_1 : \mathbb{R}_+ \to \mathbb{R}$, which is necessarily 1-Lipschitz and convex.

Letting $\mathcal{H}_1$ denote the 1-Lipschitz convex $h : \mathbb{R} \to \mathbb{R}$ satisfying $h(0) = 0$, we thus have $\sup_{h\in\mathcal{H}} V(h) = \sup_{h\in\mathcal{H}_1} \mathbb{E}[h(\|W\|_2) - h(\|Z\|_2)]$. Now, we define $\mathcal{G}_1$ to be the set of measurable non-decreasing functions bounded in $[-1,1]$. Then by known properties of convex functions [98], for any $h \in \mathcal{H}_1$, we can write $h(t) = \int_0^t g(s)ds$ for some $g \in \mathcal{G}_1$. Using this representation, we have

$$\begin{aligned}\sup_{h\in\mathcal{H}} V(h) &= \sup_{h\in\mathcal{H}_1}\left\{\mathbb{E}[h(\|W\|_2) - h(\|Z\|_2)]\right\} \\ &= \sup_{g\in\mathcal{G}_1}\left\{\mathbb{E}[h(\|W\|_2) - h(\|Z\|_2)],\ \text{where } h(t) = \int_0^t g(s)ds\right\}. \end{aligned}\tag{6.36}$$

Let g_a denote the $\{-1,1\}$-valued function with step at a, that is, $g_a(t) = -\mathbf{1}\{t \le a\} + \mathbf{1}\{t > a\}$. We define $\mathcal{G}_1^{(n)}$ to be the set of non-decreasing step functions bounded in $[-1,1]$ with at most n steps, that is, functions of the form $g(t) = \sum_{i=1}^n b_i g_{a_i}(t)$, where $|g(t)| \le 1$ for all $t \in \mathbb{R}$. We may then further simplify the expression (6.36) by replacing $\mathcal{G}_1$ with $\mathcal{G}_1^{(n)}$, that is,

$$\sup_{h\in\mathcal{H}} V(h) = \sup_{n\in\mathbb{N}}\sup_{g\in\mathcal{G}_1^{(n)}}\left\{\mathbb{E}[h(\|W\|_2) - h(\|Z\|_2)],\ \text{where } h(t) = \int_0^t g(s)ds\right\}.$$

The extremal points of $\mathcal{G}_1^{(n)}$ are the step functions $\{g_a \mid a \in \mathbb{R}\}$, and since the supremum (6.36) is linear in g, it may be taken over such g_a. Lemma 6.11 then follows by noting the integral equality $\int_0^t g_a(s)ds = |t-a| - |a|$. The restriction to $a \geq 0$ in the lemma follows since $\|v\|_2 \geq 0$ for all $v \in \mathbb{R}^d$. □

By Lemma 6.11, for any 1-Lipschitz h, the associated difference function has expectation bounded as

$$\mathbb{E}[\Delta_u(Z_1)] = \mathbb{E}[h(Z_1 + uZ_2) - h(Z_1)] \leq \sup_{a \in \mathbb{R}_+} \mathbb{E}\left[|\|Z_1 + uZ_2\|_2 - a| - |\|Z_1\|_2 - a|\right].$$

For the distributions identified by Assumption 6E, we can in fact show that the preceding supremum is attained at $a = 0$.

Lemma 6.12. *Let $Z_1 \sim \mu_1$ and $Z_2 \sim \mu_2$ be independent, where μ_1 and μ_2 satisfy Assumption 6E. For any $u \geq 0$, the function*

$$a \mapsto \zeta(a) := \mathbb{E}\left[|\,\|Z_1 + uZ_2\|_2 - a| - |\,\|Z_1\|_2 - a|\right]$$

is non-increasing in $a \geq 0$.

We return to prove this lemma at the end of the section.

With the intermediate results above, we can complete our proof of Lemma 6.5. In view of Lemma 6.10, we only need to bound $\mathbb{E}[\Delta_u(Z_1)^k]$, where $\Delta_u(\theta) = \mathbb{E}[h(\theta + uZ_2)] - h(\theta)$. Recall that $\Delta_u(\theta) \geq 0$ since h is convex. Moreover, since h is 1-Lipschitz,

$$\Delta_u(\theta) \leq \mathbb{E}\left[\left|h(\theta + uZ_2) - h(\theta)\right|\right] \leq \mathbb{E}\left[\|uZ_2\|_2\right] \leq u\mathbb{E}\left[\|Z_2\|_2^2\right]^{1/2} = u\sqrt{d},$$

where the last equality follows from the choices of Z_2 in Assumption 6E. Therefore, we have the crude but useful bound

$$\mathbb{E}[\Delta_u(Z_1)^k] \leq u^{k-1} d^{\frac{k-1}{2}}\, \mathbb{E}[\Delta_u(Z_1)] = u^{k-1} d^{\frac{k-1}{2}}\, \mathbb{E}[h(Z_1 + uZ_2) - h(Z_1)], \tag{6.37}$$

where the last expectation is over both Z_1 and Z_2. Since Z_1 and Z_2 both have rotationally invariant distributions, Lemmas 6.11 and 6.12 imply that the expectation in expression (6.37) is bounded by

$$\mathbb{E}[h(Z_1 + uZ_2) - h(Z_1)] \leq \mathbb{E}\left[\|Z_1 + uZ_2\|_2 - \|Z_1\|_2\right].$$

Lemma 6.5 then follows by bounding the norm difference in the preceding display for each choice of the smoothing distributions in Assumption 6E. We claim that

$$\mathbb{E}\left[\|Z_1 + uZ_2\|_2 - \|Z_1\|_2\right] \leq \frac{1}{\sqrt{2}} u^2 \sqrt{d}. \tag{6.38}$$

To see this inequality, we consider the possible distributions for the pair Z_1, Z_2 under Assumption 6E.

1. Let T_d have χ^2-distribution with d degrees of freedom. Then for Z_1, Z_2 independent and $\mathsf{N}(0, I_{d\times d})$-distributed, we have the distributional identities $\|Z_1 + uZ_2\|_2 \stackrel{d}{=} \sqrt{1+u^2}\sqrt{T_d}$ and $\|Z_1\|_2 \stackrel{d}{=} \sqrt{T_d}$. Using the inequalities $\sqrt{1+u^2} \leq 1 + \frac{1}{2}u^2$ and $\mathbb{E}[\sqrt{T_d}] \leq \mathbb{E}[T_d]^{\frac{1}{2}} = \sqrt{d}$, we obtain

$$\mathbb{E}\left[\|Z_1 + uZ_2\|_2 - \|Z_1\|_2\right] = \left(\sqrt{1+u^2} - 1\right)\mathbb{E}[\sqrt{T_d}] \leq \frac{1}{2}u^2\sqrt{d}.$$

2. By assumption, if Z_1 is uniform on $\sqrt{d+2}\,B_2^d$ then Z_2 has either Uniform($\sqrt{d+2}\,B_2^d$) or Uniform($\sqrt{d}\,\mathbb{S}^{d-1}$) distribution. Using the inequality $\sqrt{a+b} - \sqrt{a} \leq b/(2\sqrt{a})$, valid for $a \geq 0$ and $b \geq -a$, we may write

$$\begin{aligned}\|Z_1 + uZ_2\|_2 - \|Z_1\|_2 &= \sqrt{\|Z_1\|_2^2 + 2u\langle Z_1, Z_2\rangle + u^2\|Z_2\|_2^2} - \sqrt{\|Z_1\|_2^2} \\ &\leq \frac{2u\langle Z_1, Z_2\rangle + u^2\|Z_2\|_2^2}{2\|Z_1\|_2} = u\left\langle \frac{Z_1}{\|Z_1\|_2}, Z_2\right\rangle + \frac{1}{2}u^2\frac{\|Z_2\|_2^2}{\|Z_1\|_2}.\end{aligned}$$

   Since Z_1 and Z_2 are independent and $\mathbb{E}[Z_2] = 0$, the expectation of the first term on the right hand side above vanishes. For the second term, the independence of Z_1 and Z_2 and moment calculation (6.34) imply

$$\begin{aligned}\mathbb{E}\left[\|Z_1 + uZ_2\|_2 - \|Z_1\|_2\right] &\leq \frac{1}{2}u^2\,\mathbb{E}\left[\frac{1}{\|Z_1\|_2}\right]\,\mathbb{E}\left[\|Z_2\|_2^2\right] \\ &= \frac{1}{2}u^2 \cdot \frac{1}{\sqrt{d+2}}\,\frac{d}{(d-1)} \cdot d \leq \frac{1}{\sqrt{2}}u^2\sqrt{d},\end{aligned}$$

   where the last inequality holds for $d \geq 2$.

We thus obtain the claim (6.38), and applying inequality (6.38) to our earlier computation (6.37) yields

$$\mathbb{E}[\Delta_u(Z_1)^k] \leq \frac{1}{\sqrt{2}}u^{k+1}d^{\frac{k}{2}}.$$

Plugging in this bound on Δ_u to Lemma 6.10, we obtain the result

$$\begin{aligned}\mathbb{E}\left[\left|h(Z_1 + uZ_2) - h(Z_1)\right|^k\right] &\leq 2^{k-\frac{3}{2}}u^{k+1}d^{\frac{k}{2}} + c^{-\frac{k}{2}}2^{k-1}k^{\frac{k}{2}}u^k\log^{\frac{k}{2}}(d+2k) + \sqrt{2}u^k \\ &\leq c_k u^k\left[ud^{\frac{k}{2}} + 1 + \log^{\frac{k}{2}}(d+2k)\right],\end{aligned}$$

where c_k is a numerical constant that only depends on k. This is the desired statement of Lemma 6.5.

We now return to prove the remaining intermediate lemma.

Proof of Lemma 6.12 Since the quantity $\|Z_1 + uZ_2\|_2$ has a density with respect to Lebesgue measure, standard results on differentiating through an expectation [e.g., 25] imply

$$\begin{aligned}\frac{d}{da}\mathbb{E}\left[|\|Z_1 + uZ_2\|_2 - a|\right] &= \mathbb{E}[\text{sign}(a - \|Z_1 + uZ_2\|_2)] \\ &= \mathbb{P}(\|Z_1 + uZ_2\|_2 \leq a) - \mathbb{P}(\|Z_1 + uZ_2\|_2 > a),\end{aligned}$$

where we used that the subdifferential of $a \mapsto |v - a|$ is $\text{sign}(a - v)$. As a consequence, we find that

$$\begin{aligned}\frac{d}{da}\zeta(a) &= \mathbb{P}(\|Z_1 + uZ_2\|_2 \leq a) - \mathbb{P}(\|Z_1 + uZ_2\|_2 > a) - \mathbb{P}(\|Z_1\|_2 \leq a) + \mathbb{P}(\|Z_1\|_2 > a) \\ &= 2\left[\mathbb{P}\left(\|Z_1 + uZ_2\|_2 \leq a\right) - \mathbb{P}\left(\|Z_1\|_2 \leq a\right)\right]. \qquad (6.39)\end{aligned}$$

If we can show the quantity (6.39) is non-positive for all a, we obtain our desired result. It thus remains to prove that $\|Z_1 + uZ_2\|_2$ stochastically dominates $\|Z_1\|_2$ for each choice of μ_1, μ_2 satisfying Assumption 6E. We enumerate each of the cases below.

1. Let T_d have χ^2-distribution with d degrees of freedom and $Z_1, Z_2 \sim \mathsf{N}(0, I_{d\times d})$. Then by definition we have $\|Z_1 + uZ_2\|_2 \stackrel{d}{=} \sqrt{1+u^2}\sqrt{T_d}$ and $\|Z_1\|_2 \stackrel{d}{=} \sqrt{T_d}$, and

$$\mathbb{P}\left(\|Z_1 + uZ_2\|_2 \leq a\right) = \mathbb{P}\left(\sqrt{T_d} \leq \frac{a}{\sqrt{1+u^2}}\right) \leq \mathbb{P}\left(\sqrt{T_d} \leq a\right) = \mathbb{P}\left(\|Z_1\|_2 \leq a\right)$$

   as desired.

2. Now suppose Z_1, Z_2 are independent and distributed as Uniform$(r\,B_2^d)$; our desired result will follow by setting $r = \sqrt{d+2}$. Let $p_0(t)$ and $p_u(t)$ denote the densities of $\|Z_1\|_2$ and $\|Z_1 + uZ_2\|_2$, respectively, with respect to Lebesgue measure on $\mathbb{R}$. We now compute them explicitly. For p_0, for $0 \leq t \leq r$ we have

$$p_0(t) = \frac{d}{dt}\mathbb{P}(\|Z_1\|_2 \leq t) = \frac{d}{dt}\left(\frac{t}{r}\right)^d = \frac{d\,t^{d-1}}{r^d},$$

   and $p_0(t) = 0$ otherwise. For p_u, let λ denote the Lebesgue measure in $\mathbb{R}^d$ and σ denote the $(d-1)$-dimensional surface area in $\mathbb{R}^d$. The random variables Z_1 and uZ_2 have densities, respectively,

$$q_1(x) = \frac{1}{\lambda(r\,B_2^d)} = \frac{1}{r^d\lambda(B_2^d)} \quad \text{for } x \in rB_2^d$$

   and

$$q_u(x) = \frac{1}{\lambda(ur\,B_2^d)} = \frac{1}{u^d r^d \lambda(B_2^d)} \quad \text{for } x \in urB_2^d,$$

and $q_1(x) = q_u(x) = 0$ otherwise. Then the density of $Z_1 + uZ_2$ is given by the convolution

$$\tilde{q}(z) = \int_{\mathbb{R}^d} q_1(x) q_u(z-x)\, \lambda(dx) = \int_{E(z)} \frac{1}{r^d \lambda(B_2^d)} \cdot \frac{1}{u^d r^d \lambda(B_2^d)}\, \lambda(dx) = \frac{\lambda(E(z))}{u^d\, r^{2d} \lambda(B_2^d)^2}.$$

Here $E(z) := B_2^d(0,r) \cap B_2^d(z,ur)$ is the domain of integration, in which the densities $q_1(x)$ and $q_u(z-x)$ are nonzero. The volume $\lambda(E(z))$—and hence also $\tilde{q}(z)$—depend on z only via its norm $\|z\|_2$. Therefore, the density $p_u(t)$ of $\|Z_1 + uZ_2\|_2$ can be expressed as

$$p_u(t) = \tilde{q}(te_1)\, \sigma(t\mathbb{S}^{d-1}) = \frac{\lambda(E(te_1))\, t^{d-1}\, \sigma(\mathbb{S}^{d-1})}{u^d\, r^{2d}\, \lambda(B_2^d)^2} = d\, \frac{\lambda(E(te_1))\, t^{d-1}}{u^d\, r^{2d}\, \lambda(B_2^d)},$$

where the last equality above follows from the relation $\sigma(\mathbb{S}^{d-1}) = d\lambda(B_2^d)$. Since $E(te_1) \subseteq B_2^d(te_1, ur)$ by definition,

$$\lambda(E(te_1)) \le \lambda\left(B_2^d(te_1, ur)\right) = u^d r^d\, \lambda(B_2^d),$$

so for all $0 \le t \le (1+u)r$ we have

$$p_u(t) = d\, \frac{\lambda(E(te_1))\, t^{d-1}}{u^d\, r^{2d}\, \lambda(B_2^d)} \le \frac{d\, t^{d-1}}{r^d},$$

and clearly $p_u(t) = 0$ for $t > (1+u)r$. In particular, $p_u(t) \le p_1(t)$ for $0 \le t \le r$, which gives us our desired stochastic dominance inequality (6.39): for $a \in [0, r]$,

$$\mathbb{P}(\|Z_1 + uZ_2\|_2 \le a) = \int_0^a p_u(t)\, dt \le \int_0^a p_0(t)\, dt = \mathbb{P}(\|Z_1\|_2 \le a),$$

and for $a > r$ we have $\mathbb{P}(\|Z_1 + uZ_2\|_2 \le a) \le 1 = \mathbb{P}(\|Z_1\|_2 \le a)$.

3. Finally, consider the case when $Z_1 \sim \text{Uniform}(\sqrt{d+2}\, B_2^d)$ and $Z_2 \sim \text{Uniform}(\sqrt{d}\, \mathbb{S}^{d-1})$. As in the previous case, we will show that $p_0(t) \le p_u(t)$ for $0 \le t \le \sqrt{d+2}$, where $p_0(t)$ and $p_u(t)$ are the densities of $\|Z_1\|_2$ and $\|Z_1 + uZ_2\|_2$, respectively. We know that the density of $\|Z_1\|_2$ is

$$p_0(t) = \frac{d\, t^{d-1}}{(d+2)^{\frac{d}{2}}} \quad \text{for } 0 \le t \le \sqrt{d+2},$$

and $p_0(t) = 0$ otherwise. To compute p_u, we first determine the density $\tilde{q}(z)$ of the random variable $Z_1 + uZ_2$ with respect to the Lebesgue measure λ on $\mathbb{R}^d$. The usual convolution formula does not directly apply as Z_1 and Z_2 have densities with respect to different base measures (λ and σ, respectively). However, as Z_1 and Z_2 are both

uniform, we can argue as follows. Integrating over the surface $u\sqrt{d}\mathbb{S}^{d-1}$ (essentially performing a convolution), each point $uy \in u\sqrt{d}\,\mathbb{S}^{d-1}$ contributes the amount

$$\frac{1}{\sigma(u\sqrt{d}\,\mathbb{S}^{d-1})} \cdot \frac{1}{\lambda(\sqrt{d+2}\,B_2^d)} = \frac{1}{u^{d-1}\,d^{\frac{d-1}{2}}\,(d+2)^{\frac{d}{2}}\,\sigma(\mathbb{S}^{d-1})\,\lambda(B_2^d)}$$

to the density $\tilde{q}(z)$, provided $\|z - uy\|_2 \leq \sqrt{d+2}$. For fixed $z \in (\sqrt{d+2} + u\sqrt{d})B_2^d$, the set of such contributing points uy can be written as $E(z) = B_2^d(z, \sqrt{d+2}) \cap \mathbb{S}^{d-1}(0, u\sqrt{d})$. Therefore, the density of $Z_1 + uZ_2$ is given by

$$\tilde{q}(z) = \frac{\sigma(E(z))}{u^{d-1}\,d^{\frac{d-1}{2}}\,(d+2)^{\frac{d}{2}}\,\sigma(\mathbb{S}^{d-1})\,\lambda(B_2^d)}.$$

Since $\tilde{q}(z)$ only depends on z via its norm $\|z\|_2$, the formula above also gives us the density $p_u(t)$ of $\|Z_1 + uZ_2\|_2$:

$$p_u(t) = \tilde{q}(te_1)\,\sigma(t\mathbb{S}^{d-1}) = \frac{\sigma(E(z))\,t^{d-1}}{u^{d-1}\,d^{\frac{d-1}{2}}\,(d+2)^{\frac{d}{2}}\,\lambda(B_2^d)}.$$

Noting that $E(z) \subseteq \mathbb{S}^{d-1}(0, u\sqrt{d})$ gives us

$$p_u(t) \leq \frac{\sigma(u\sqrt{d}\,\mathbb{S}^{d-1})\,t^{d-1}}{u^{d-1}\,d^{\frac{d-1}{2}}\,(d+2)^{\frac{d}{2}}\,\lambda(B_2^d)} = \frac{d\,t^{d-1}}{(d+2)^{\frac{d}{2}}}.$$

In particular, we have $p_u(t) \leq p_0(t)$ for $0 \leq t \leq \sqrt{d+2}$, which, as we saw in the previous case, gives us the desired stochastic dominance inequality (6.39).

6.8 Technical proofs associated with lower bounds

In this section, we prove the technical results necessary for the proofs of Propositions 6.1 and 6.2.

6.8.1 Proof of Lemma 6.6

First, note that the optimal vector $\theta^A = -d^{-1/q}\mathbb{1}$ with optimal value $-d^{1-1/q}$, and $\theta^B = -(d-i)^{-1/q}\mathbb{1}_{i+1:d}$, where $\mathbb{1}_{i+1:d}$ denotes the vector with 0 entries in its first i coordinates and 1 elsewhere. As a consequence, we have $\left\langle \theta^B, \mathbb{1} \right\rangle = -(d-i)^{1-1/q}$. Now we use the fact that by convexity of the function $x \mapsto -x^{1-1/q}$ for $q \in [1, \infty]$,

$$-d^{1-1/q} \leq -(d-i)^{1-1/q} - \frac{1-1/q}{d^{1/q}}i,$$

since the derivative of $x \mapsto -x^{1-1/q}$ at $x = d$ is given by $-(1-1/q)/d^{1/q}$ and the quantity $-x^{1-1/q}$ is non-increasing in x for $q \in [1, \infty]$.

6.8.2 Proof of Lemma 6.7

For notational convenience, let the distribution $P_{v,+j}$ be identical to the distribution P_v but with the jth coordinate v_j forced to be $+1$ and similarly for $P_{v,-j}$. Using Pinsker's inequality and the joint convexity of the KL-divergence, we have

$$\begin{aligned}\|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 &\le \frac{1}{4}\left[D_{\mathrm{kl}}\left(P_{+j}\|P_{-j}\right) + D_{\mathrm{kl}}\left(P_{-j}\|P_{+j}\right)\right] \\ &\le \frac{1}{2^{d+2}}\sum_{v\in\mathcal{V}}\left[D_{\mathrm{kl}}\left(P_{v,+j}\|P_{v,-j}\right) + D_{\mathrm{kl}}\left(P_{v,-j}\|P_{v,+j}\right)\right].\end{aligned}$$

By the chain-rule for KL-divergences [47], if we define $P_v^k(\cdot \mid Y^{1:k-1})$ to be the distribution of the kth observation Y^k conditional on v and $Y^{1:k-1}$, then we have

$$D_{\mathrm{kl}}\left(P_{v,+j}\|P_{v,-j}\right) = \sum_{k=1}^n \int_{\mathcal{Y}^{k-1}} D_{\mathrm{kl}}\left(P_{v,+j}^k(\cdot \mid Y^{1:k-1} = y)\|P_{v,-j}^k(\cdot \mid Y^{1:k-1} = y)\right) dP_{v,+j}(y).$$

We show how to bound the preceding sequence of KL-divergences for the observational scheme based on function-evaluations we allow. Let $W = [\theta\ w] \in \mathbb{R}^{d\times 2}$ denote the pair of query points, so we have by construction that the observation $Y = W^\top X$ where $X \mid V = v \sim \mathsf{N}(\delta v, \sigma^2 I_{d\times d})$. In particular, given v and the pair W, the vector $Y \in \mathbb{R}^d$ is normally distributed with mean $\delta W^\top v$ and covariance $\sigma^2 W^\top W = \Sigma$, where the covariance Σ is defined in equation (6.28). The KL divergence between normal distributions is $D_{\mathrm{kl}}\left(\mathsf{N}(\mu_1,\Sigma)\|\mathsf{N}(\mu_2,\Sigma)\right) = \frac{1}{2}(\mu_1-\mu_2)^\top\Sigma^{-1}(\mu_1-\mu_2)$. Note that if v and v' differ in only coordinate j, then $\langle v - v', \theta\rangle = (v_j - v'_j)\theta_j$. We thus obtain

$$D_{\mathrm{kl}}\left(P_{v,+j}^k(\cdot \mid y^{1:k-1})\|P_{v,-j}^k(\cdot \mid y^{1:k-1})\right) \le 2\delta^2\mathbb{E}\left[\begin{bmatrix}\theta_j^k\\ w_j^k\end{bmatrix}^\top (\Sigma^k)^{-1}\begin{bmatrix}\theta_j^k\\ w_j^k\end{bmatrix} \mid y^{1:k-1}\right]$$

where the expectation is taken with respect to any additional randomness in the construction of the pair (θ^k, w^k) (as, aside from this randomness, they are measureable $Y^{1:n-1}$). Combining the sequence of inequalities from the preceding paragraph, we obtain

$$\begin{aligned}&\|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 \\ &\quad\le \frac{\delta^2}{2^{d+1}}\sum_{k=1}^n\sum_{v\in\mathcal{V}}\int_{\mathcal{Y}^{k-1}}\mathbb{E}\left[\begin{bmatrix}\theta_j^k\\ w_j^k\end{bmatrix}^\top (\Sigma^k)^{-1}\begin{bmatrix}\theta_j^k\\ w_j^k\end{bmatrix} \mid y^{1:k-1}\right]\left(dP_{v,+j}(y^{1:k-1}) + dP_{v,-j}(y^{1:k-1})\right) \\ &\quad= \frac{\delta^2}{2}\sum_{k=1}^n\int_{\mathcal{Y}^{k-1}}\mathbb{E}\left[\begin{bmatrix}\theta_j^k\\ w_j^k\end{bmatrix}^\top (\Sigma^k)^{-1}\begin{bmatrix}\theta_j^k\\ w_j^k\end{bmatrix} \mid y^{1:k-1}\right]\left(dP_{+j}(y^{1:k-1}) + dP_{-j}(y^{1:k-1})\right),\end{aligned}$$

where for the equality we used the definitions of the distributions $P_{v,\pm j}$ and $P_{\pm j}$. Integrating over the observations y proves the claimed inequality (6.29).

Part III

Privacy

Chapter 7

Privacy, minimax rates of convergence, and data processing inequalities

In this chapter, we study another type of constraint on inference and estimation procedures that has growing importance: we would like our estimators to remain confidential. With this in mind, working under a model of privacy in which data remains private even from the statistician, we study the tradeoff between privacy guarantees and the utility of the resulting statistical estimators. To do this, our first step is to define a notion of (minimax) optimality for private estimation. To control these new constrained minimax risks, we prove bounds on information-theoretic quantities, including mutual information and Kullback-Leibler divergence, that depend on the privacy guarantees. When combined with standard minimax techniques, including the Le Cam, Fano, and Assouad methods of outline in Chapter 2, these inequalities allow for a precise characterization of statistical rates under local privacy constraints. We provide a treatment of several canonical families of problems: mean estimation, parameter estimation in fixed-design regression, multinomial probability estimation, and nonparametric density estimation. For all of these families, we provide lower and upper bounds that match up to constant factors, and exhibit new (optimal) privacy-preserving mechanisms and computationally efficient estimators that achieve the bounds.

7.1 Introduction

A major challenge in statistical inference is that of characterizing and balancing statistical utility with the privacy of individuals from whom data is obtained [63, 64, 76]. Such a characterization requires a formal definition of privacy, and *differential privacy* has been put forth as one such formalization [e.g., 68, 29, 69, 90, 91]. In the database and cryptography literatures from which differential privacy arose, early research was mainly algorithmic in focus, and researchers have used differential privacy to evaluate privacy-retaining mechanisms

for transporting, indexing, and querying data. More recent work aims to link differential privacy to statistical concerns [66, 180, 88, 164, 44, 155]; in particular, researchers have developed algorithms for private robust statistical estimators, point and histogram estimation, and principal components analysis, among others. Guarantees of optimality in this line of work have typically been with respect to estimators, where the goal is to approximate an estimator itself under privacy-respecting transformations of the data. There has also been recent work within the context of classification problems and the "probably approximately correct" framework of statistical learning theory [e.g. 105, 20] that treats the data as random and aims to recover aspects of the underlying population.

In this chapter, we take a fully inferential point of view on privacy by bringing differential privacy into contact with statistical decision theory. Our focus is on the fundamental limits of differentially-private estimation. By treating differential privacy as an abstract constraint on estimators, we obtain independence from specific estimation procedures and privacy-preserving mechanisms. Within this framework, we derive both lower bounds and matching upper bounds on minimax risk. We obtain our lower bounds by integrating differential privacy into the classical paradigms for bounding minimax risk via the inequalities of Le Cam, Fano, and Assouad, while we obtain matching upper bounds by proposing and analyzing specific private procedures.

We study the setting of *local privacy*, in which providers do not even the statistician collecting the data. Although local privacy is a relatively stringent requirement, we view this setting as a natural step in identifying minimax risk bounds under privacy constraints. Indeed, local privacy is one of the oldest forms of privacy: its essential form dates to Warner [179], who proposed it as a remedy for what he termed "evasive answer bias" in survey sampling. We hope that we can leverage deeper understanding of this classical setting to treat other privacy-preserving approaches to data analysis.

More formally, let $X_1, \ldots, X_n \in \mathcal{X}$ be observations drawn according to a distribution P, and let $\theta = \theta(P)$ be a parameter of this unknown distribution. We wish to estimate θ based on access to obscured views $Z_1, \ldots, Z_n \in \mathcal{Z}$ of the original data. The original random variables $\{X_i\}_{i=1}^n$ and the privatized observations $\{Z_i\}_{i=1}^n$ are linked via a family of conditional distributions $Q_i(Z_i \mid X_i = x, Z_{1:i-1} = z_{1:i-1})$. To simplify notation, we typically omit the subscript in Q_i. We refer to Q as a *channel distribution*, as it acts as a conduit from the original to the privatized data, and we assume it is *sequentially interactive*, meaning the channel has the conditional independence structure

$$\{X_i, Z_1, \ldots, Z_{i-1}\} \to Z_i \quad \text{and} \quad Z_i \perp X_j \mid \{X_i, Z_1, \ldots, Z_{i-1}\} \text{ for } j \neq i,$$

illustrated on the left of Figure 7.1. A special case of a such a channel is the *non-interactive* case, in which each Z_i depends only on X_i (Fig. 7.1, right).

Our work is based on the following definition of privacy. For a given privacy parameter $\alpha \geq 0$, we say that Z_i is an *α-differentially locally private* view of X_i if for all $z_1, \ldots, z_{i-1}$ and $x, x' \in \mathcal{X}$ we have

$$\sup_{S \in \sigma(\mathcal{Z})} \frac{Q_i(Z_i \in S \mid X_i = x, Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1})}{Q_i(Z_i \in S \mid X_i = x', Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1})} \leq \exp(\alpha), \tag{7.1}$$

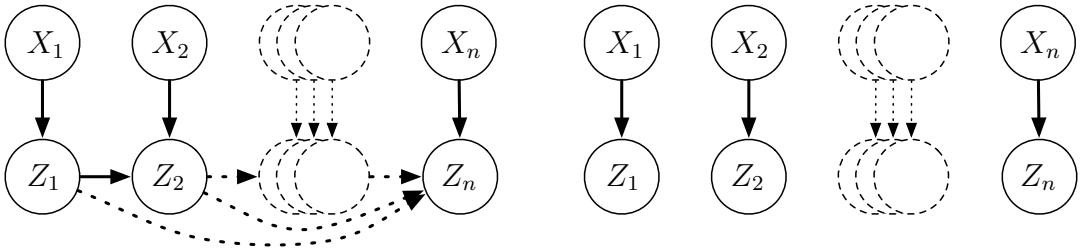

Figure 7.1. Left: graphical structure of private Z_i and non-private data X_i in interactive case. Right: graphical structure of channel in non-interactive case.

where $\sigma(\mathcal{Z})$ denotes an appropriate σ-field on $\mathcal{Z}$. Definition (7.1) does not constrain Z_i to be a release of data based on exclusively on X_i: the channel Q_i may be *interactive* [68], changing based on prior private observations Z_j. We also consider the non-interactive case [179, 74] where Z_i depends only on X_i (see the right side of Figure 7.1); here the bound (7.1) reduces to

$$\sup_{S \in \sigma(\mathcal{Z})} \sup_{x, x' \in \mathcal{X}} \frac{Q(Z_i \in S \mid X_i = x)}{Q(Z_i \in S \mid X_i = x')} \leq \exp(\alpha). \tag{7.2}$$

These definitions capture a type of plausible-deniability: no matter what data Z is released, it is nearly equally as likely to have come from one point $x \in \mathcal{X}$ as any other. It is also possible to interpret differential privacy within a hypothesis testing framework, where α controls the error rate in tests for the presence or absence of individual data points in a dataset [180]. Such guarantees against discovery, together with the treatment of issues of side information or adversarial strength that are problematic for other formalisms, have been used to make the case for differential privacy within the computer science literature; see, for example, the papers [74, 68, 16, 80].

Although differential privacy provides an elegant formalism for limiting disclosure and protecting against many forms of privacy breach, it is a stringent measure of privacy, and it is conceivably overly stringent for statistical practice. Indeed, Fienberg et al. [77] criticize the use of differential privacy in releasing contingency tables, arguing that known mechanisms for differentially private data release can give unacceptably poor performance. As a consequence, they advocate—in some cases—recourse to weaker privacy guarantees to maintain the utility and usability of released data. There are results that are more favorable for differential privacy; for example, Smith [164] shows that the non-local form of differential privacy [68] can be satisfied while yielding asymptotically optimal parametric rates of convergence for some point estimators. Resolving such differing perspectives requires investigation into whether particular methods have optimality properties that would allow a general criticism of the framework, and characterizing the trade-offs between privacy and statistical efficiency. Such are the goals of this part of the thesis.

7.1.1 Our contributions

The main contribution of this chapter is to provide general techniques for deriving minimax bounds under local privacy constraints and to illustrate these techniques by computing minimax rates for several canonical problems: (a) mean estimation; (b) parameter estimation in fixed design regression; (c) multinomial probability estimation; and (d) density estimation. We now outline our main contributions. Because a deeper comparison of the current work with prior research requires formally defining our minimax framework and presentation of our main results, we defer more expansive discussion of related work to Section 7.6. We emphasize, however, that our minimax rates are for estimation of *population* quantities, in accordance with our connections to statistical decision theory; most prior work in the privacy literature focuses on accurate approximation of estimators in a conditional analysis in which the data are treated as fixed.

Many methods for obtaining minimax bounds involve information-theoretic quantities relating distributions that may have generated the data [188, 185, 173]. In particular, let P_1 and P_2 denote two distributions on the observations X_i, and for $v \in \{1, 2\}$, define the marginal distribution M_v^n on $\mathcal{Z}^n$ by

$$M_v^n(S) := \int Q^n(S \mid x_1, \ldots, x_n) dP_v(x_1, \ldots, x_n) \quad \text{for } S \in \sigma(\mathcal{Z}^n). \tag{7.3}$$

Here $Q^n(\cdot \mid x_1, \ldots, x_n)$ denotes the joint distribution on $\mathcal{Z}^n$ of the private sample $Z_{1:n}$, conditioned on $X_{1:n} = x_{1:n}$. The mutual information of samples drawn according to distributions of the form (7.3) and the KL divergence between such distributions are key objects in statistical discriminability and minimax rates [92, 27, 188, 185, 173], where they are often applied in one of three lower-bounding techniques: Le Cam's, Fano's, and Assouad's methods.

Keeping in mind the centrality of these information-theoretic quantities, we summarize our main results at a high-level as follows. Theorem 7.1 bounds the KL divergence between distributions M_1^n and M_2^n, as defined by the marginal (7.3), by a quantity dependent on the differential privacy parameter α and the total variation distance between P_1 and P_2. The essence of Theorem 7.1 is that

$$D_{\rm kl}\left(M_1^n \| M_2^n\right) \lesssim \alpha^2 n \left\|P_1 - P_2\right\|_{\rm TV}^2,$$

where $\lesssim$ denotes inequality up to numerical constants. When $\alpha^2 < 1$, which is the usual region of interest, this result shows that for statistical procedures whose minimax rate of convergence can be determined by classical information-theoretic methods, the additional requirement of α-local differential privacy causes the *effective sample size* of *any* statistical procedure to be reduced from n to at most $\alpha^2 n$. Section 7.3.1 contains the formal statement of this theorem, while Section 7.3.2 provides corollaries showing its application to minimax risk bounds. We follow this in Section 7.3.3 with applications of these results to estimation of one-dimensional means and fixed-design regression problems, providing corresponding upper bounds on the minimax risk. In addition to our general analysis, we exhibit some striking difficulties of locally private estimation in non-compact spaces: if we wish to estimate the

mean of a random variable X satisfying $\mathrm{Var}(X) \le 1$, the minimax rate of estimation of $\mathbb{E}[X]$ decreases from the parametric $1/n$ rate to $1/\sqrt{n\alpha^2}$.

Theorem 7.1 is appropriate for many one dimensional problems, but it does not address difficulties inherent in higher dimensional problems. With this motivation, our next two main results (Theorems 7.2 and 7.3) generalize Theorem 7.1 and incorporate dimensionality in an essential way: each provides bounds on information-theoretic quantities by dimension-dependent analogues of total-variation. Somewhat more specifically, Theorem 7.2 provides bounds on mutual information quantities essential in information theoretic techniques such as Fano's method [188, 185], while Theorem 7.3 provides analogous bounds on summed pairs of KL-divergences useful in applications of Assouad's method [11, 188, 9].

As a consequence of Theorems 7.2 and 7.3, we obtain that for many d-dimensional estimation problems the effective sample size is reduced from n to $n\alpha^2/d$; as our examples illustrate, this dimension-dependent reduction in sample size can have dramatic consequences. We provide the main statement and consequences of Theorem 7.2 in Section 7.4, showing its application to obtaining minimax rates for mean estimation in both classical and high-dimensional settings. In Section 2.2.4, we present Theorem 7.3, showing how it provides (sharp) minimax lower bounds for multinomial and probability density estimation. Our results enable us to derive (often new) optimal mechanisms for these problems. One interesting consequence of our results is that Warner's randomized response procedure [179] from the 1960s is an optimal mechanism for multinomial estimation.

7.2 Background and problem formulation

We first recall the minimax framework established in Chapter 2, in use throughout the thesis, putting the general constrained minimax quantities (2.4) in the setting of private estimation. As previously, we let $\mathcal{P}$ denote a class of distributions on the sample space $\mathcal{X}$, let $\theta(P) \in \Theta$ denote a function defined on $\mathcal{P}$, the function $\rho : \Theta \times \Theta \to \mathbb{R}_+$ denote a semi-metric on the space Θ, which we use to measure the error of an estimator for the parameter θ, and let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (for example, $\Phi(t) = t^2$).

In the classical setting, the statistician is given direct access to i.i.d. observations X_i drawn according to some $P \in \mathcal{P}$. The local privacy setting involves an additional ingredient, namely, a conditional distribution Q that transforms the sample $\{X_i\}_{i=1}^n$ into the private sample $\{Z_i\}_{i=1}^n$ taking values in $\mathcal{Z}$. Based on these Z_i, our goal is to estimate the unknown parameter $\theta(P) \in \Theta$. An estimator $\widehat{\theta}$ in the locally private setting is a measurable function $\widehat{\theta} : \mathcal{Z}^n \to \Theta$, and we assess the quality of the estimate $\widehat{\theta}(Z_1, \ldots, Z_n)$ in terms of the risk

$$\mathbb{E}_{P,Q}\left[\Phi\big(\rho(\widehat{\theta}(Z_1, \ldots, Z_n), \theta(P))\big)\right].$$

For any fixed conditional distribution Q, the minimax rate is

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[\Phi\big(\rho(\widehat{\theta}(Z_1, \ldots, Z_n), \theta(P))\big)\right], \tag{7.4}$$

where we take the supremum (worst-case) over distributions $P \in \mathcal{P}$, and the infimum is taken over all estimators $\widehat{\theta}$. This is identical to the classical minimax risk (2.1), but the data passes through the channel Q before being observed by the estimator $\widehat{\theta}$.

For $\alpha > 0$, let $\mathcal{Q}_\alpha$ denote the set of all conditional distributions guaranteeing α-local privacy (7.1). By minimizing the minimax risk (7.4) over all $Q \in \mathcal{Q}_\alpha$, we obtain the central object of study for this chapter, which characterizes the optimal rate of estimation in terms of the privacy parameter α.

Definition 7.1. *Given a family of distributions $\theta(\mathcal{P})$ and a privacy parameter $\alpha > 0$, the* α-minimax rate *in the metric ρ is*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q} \left[\Phi(\rho(\widehat{\theta}(Z_1, \ldots, Z_n), \theta(P))) \right]. \tag{7.5}$$

Notably, the quantity (7.5) is simply a variant of the constrained minimax risk (2.4), the central theoretical object of study in this thesis.

From estimation to testing in private settings We now show how to adapt the minimax framework of Chapter 2, especially Section 2.2, to the private setting, which essentially amounts to tracking the appearance of the channel distribution Q in the standard reductions from estimation to testing.

We begin by recalling the settings of Section 2.2. Given an index set $\mathcal{V}$ of finite cardinality, consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ contained within $\mathcal{P}$ and the induced collection of parameters $\{\theta(P_v)\}_{v \in \mathcal{V}}$. In the classical setting, the statistician directly observes the sample X, while the local privacy constraint means that a new random sample $Z = (Z_1, \ldots, Z_n)$ is generated by sampling Z_i from the distribution $Q(\cdot \mid X_{1:n})$. By construction, if the data $X_{1:n}$ is generated according to the distribution P_v, the private sample Z is distributed according to the marginal measure M_v^n defined in equation (7.3).

Recalling the canonical hypothesis testing setting of Section 2.2.1, we consider determining the value of the underlying index v given the observed vector Z. Then, if V is drawn uniformly at random from $\mathcal{V}$, whenever the set $\theta(P_v)$ forms a 2δ-packing in the ρ-semimetric, we have the following analogue of the classical lower bound (2.5): the minimax error (7.4) has lower bound

$$\mathfrak{M}_n(\Theta, \Phi \circ \rho, Q) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(Z_1, \ldots, Z_n) \neq V),$$

where the infimum ranges over all testing functions, and $\mathbb{P}$ denotes the joint distribution over the random index V and Z.

We can then use Le Cam's or Fano's methods to lower bound the probability of error in the private hypothesis testing problem. In particular, by applying Le Cam's method (2.7), we obtain the analogue of the minimax lower bound (2.8): for any pair $P_1, P_2 \in \mathcal{P}$ satisfying $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, then

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q) \geq \Phi(\delta) \left[\frac{1}{2} - \frac{1}{2} \left\| M_1^n - M_2^n \right\|_{\mathrm{TV}} \right],$$

where the marginal M_v is defined as in expression (7.3). We can also extend the non-private Fano method from Section 2.2.3: given the separation function $\delta(t)$ associated with the set $\mathcal{V}$ and parameter θ (as defined by (2.14)), Corollary 2.2 implies

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q) \geq \Phi\Big(\frac{\delta(t)}{2}\Big) \left(1 - \frac{I(Z_1, \ldots, Z_n; V) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}\right) \qquad \text{for all } t, \tag{7.6}$$

where we recall from the definition (2.11) that $N_t^{\max} = \max_{v \in \mathcal{V}} \operatorname{card}\{v' \in \mathcal{V} : \rho(v, v') \leq t\}$.

In addition, Assouad's method from Section 2.2.4 applies. Assume that the index set $\mathcal{V} = \{-1, 1\}^d$ and the family P induces a 2δ-Hamming separation (2.17), that is, there exists a function v satisfying $\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^d \mathbf{1}\left\{[\mathsf{v}(\theta)]_j \neq v_j\right\}$ for all $\theta \in \Theta$. Then if we define the marginals $M_{\pm j}^n = 2^{-d+1} \sum_{v : v_j = \pm 1} M_v^n$, Lemma 2.2 and its equivalent minimax lower bound (2.18) become

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q) \geq \delta \sum_{j=1}^d \left[1 - \left\|M_{+j}^n - M_{-j}^n\right\|_{\mathrm{TV}}\right]. \tag{7.7}$$

As a consequence of the reductions to testing from Chapter 2 and the error bounds above, we obtain bounds on the private minimax rate (7.5) by controlling variation distances of the form $\|M_1^n - M_2^n\|_{\mathrm{TV}}$ or the mutual information between the random parameter index V and the sequence of random variables $Z_1, \ldots, Z_n$. We devote the following sections to these tasks.

7.3 Pairwise bounds under privacy: Le Cam and local Fano methods

We begin with results upper bounding symmetrized Kullback-Leibler divergence under a privacy constraint, developing consequences of this result for both Le Cam's method and a local form of Fano's method. Using these methods, we derive sharp minimax rates under local privacy for estimating 1-dimensional means and for d-dimensional fixed design regression.

7.3.1 Pairwise upper bounds on Kullback-Leibler divergences

Many statistical problems depend on comparisons between a pair of distributions P_1 and P_2 defined on a common space $\mathcal{X}$. Any channel Q transforms such a pair of distributions into a new pair (M_1, M_2) via the marginalization (7.3), that is, $M_v(S) = \int_{\mathcal{X}} Q(S \mid x) dP_v(x)$ for $v = 1, 2$. Our first main result bounds the symmetrized Kullback-Leibler (KL) divergence between these induced marginals as a function of the privacy parameter $\alpha > 0$ associated with the conditional distribution Q and the total variation distance between P_1 and P_2.

Theorem 7.1. *For any $\alpha \geq 0$, let Q be a conditional distribution that guarantees α-differential privacy. Then for any pair of distributions P_1 and P_2, the induced marginals M_1 and M_2 satisfy the bound*

$$D_{\rm kl}\left(M_1 \| M_2\right) + D_{\rm kl}\left(M_2 \| M_1\right) \leq \min\{4, e^{2\alpha}\}(e^\alpha - 1)^2 \left\|P_1 - P_2\right\|_{\rm TV}^2 . \tag{7.8}$$

Remarks Theorem 7.1 is a type of *strong data processing* inequality [8], providing a quantitative relationship from the divergence $\|P_1 - P_2\|_{\rm TV}$ to the KL-divergence $D_{\rm kl}\left(M_1 \| M_2\right)$ that arises after applying the channel Q. The result of Theorem 7.1 is similar to a result due to Dwork et al. [69, Lemma III.2], who show that $D_{\rm kl}\left(Q(\cdot \mid x) \| Q(\cdot \mid x')\right) \leq \alpha(e^\alpha - 1)$ for any $x, x' \in \mathcal{X}$, which implies $D_{\rm kl}\left(M_1 \| M_2\right) \leq \alpha(e^\alpha - 1)$ by convexity. This upper bound is weaker than Theorem 7.1 since it lacks the term $\|P_1 - P_2\|_{\rm TV}^2$. This total variation term is essential to our minimax lower bounds: more than providing a bound on KL divergence, Theorem 7.1 shows that differential privacy acts as a contraction on the space of probability measures. This contractivity holds in a strong sense: indeed, the bound (7.8) shows that even if we start with a pair of distributions P_1 and P_2 whose KL divergence is infinite, the induced marginals M_1 and M_2 always have finite KL divergence.

We provide the proof of Theorem 7.1 in Section 8.1. Here we develop a corollary that has useful consequences for minimax theory under local privacy constraints. Suppose that conditionally on $V = v$, we draw a sample $X_1, \ldots, X_n$ from the product measure $\prod_{i=1}^n P_{v,i}$, and that we draw the α-locally private sample $Z_1, \ldots, Z_n$ according to the channel $Q(\cdot \mid X_{1:n})$. Conditioned on $V = v$, the private sample is distributed according to the measure M_v^n defined previously (7.3). Because we allow interactive protocols, the distribution M_v^n need not be a product distribution in general. Given this set-up, we have the following:

Corollary 7.1. *For any α-locally differentially private* (7.1) *conditional distribution Q and any paired sequences of distributions $\{P_{v,i}\}$ and $\{P_{v',i}\}$,*

$$D_{\rm kl}\left(M_v^n \| M_{v'}^n\right) + D_{\rm kl}\left(M_{v'}^n \| M_v^n\right) \leq 4(e^\alpha - 1)^2 \sum_{i=1}^n \left\|P_{v,i} - P_{v',i}\right\|_{\rm TV}^2 . \tag{7.9}$$

See Section 8.1.2 for the proof, which requires a few intermediate steps to obtain the additive inequality. Inequality (7.9) also immediately implies a mutual information bound, which may be useful in applications of Fano's inequality. In particular, if we define the mean distribution $\overline{M}^n = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} M_v^n$, then by the definition of mutual information, we have

$$\begin{aligned} I(Z_1, \ldots, Z_n; V) &= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(M_v^n \| \overline{M}^n\right) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v, v'} D_{\rm kl}\left(M_v^n \| M_{v'}^n\right) \\ &\leq 4(e^\alpha - 1)^2 \sum_{i=1}^n \frac{1}{|\mathcal{V}|^2} \sum_{v, v' \in \mathcal{V}} \left\|P_{v,i} - P_{v',i}\right\|_{\rm TV}^2 , \end{aligned} \tag{7.10}$$

the first inequality following from the joint convexity of the KL divergence and the final inequality from Corollary 7.1.

Remarks Mutual information bounds under local privacy have appeared previously. McGregor et al. [129] study relationships between communication complexity and differential privacy, showing that differentially private schemes allow low communication. They provide a result [129, Prop. 7] guaranteeing $I(X_{1:n}; Z_{1:n}) \leq 3\alpha n$; they strengthen this bound to $I(X_{1:n}; Z_{1:n}) \leq (3/2)\alpha^2 n$ when the X_i are i.i.d. uniform Bernoulli variables. Since the total variation distance is at most 1, our result also implies this scaling (for arbitrary X_i), but it is stronger since it involves the total variation terms $\|P_{v,i} - P_{v',i}\|_{\mathrm{TV}}$, which are essential in our minimax results. In addition, Corollary 7.1 allows for *any* (sequentially) interactive channel Q; each Z_i may depend on the private answers $Z_{1:i-1}$ of other data providers.

7.3.2 Consequences for minimax theory under local privacy constraints

We now turn to some consequences of Theorem 7.1 for minimax theory under local privacy constraints. For ease of presentation, we analyze the case of independent and identically distributed (i.i.d.) samples, meaning that $P_{v,i} \equiv P_v$ for $i = 1, \ldots, n$. We show that in both Le Cam's inequality and the local version of Fano's method, the constraint of α-local differential privacy reduces the effective sample size (at least) from n to $4\alpha^2 n$.

Consequence for Le Cam's method We have seen in Section 2.2.2 how Le Cam's method provides lower bounds on the classical minimax risk via a binary hypothesis test. By applying Pinsker's inequality, one version of Le Cam's method (2.8) asserts that, for any pair of distributions $\{P_1, P_2\}$ such that $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, we have

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left\{\frac{1}{2} - \frac{1}{2\sqrt{2}} \sqrt{n D_{\mathrm{kl}}\left(P_1 \| P_2\right)}\right\}. \tag{7.11}$$

Returning to the α-locally private setting, in which the estimator $\widehat{\theta}$ depends only on the private variables $(Z_1, \ldots, Z_n)$, we measure the α-private minimax risk (7.5). By applying Le Cam's method to the pair (M_1, M_2) along with Corollary 7.1 in the form of inequality (7.9), we find:

Corollary 7.2 (Private form of Le Cam bound)**.** *Given observations from an α-locally differential private channel for some $\alpha \in [0, \frac{22}{35}]$, the α-private minimax risk has lower bound*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \alpha) \geq \Phi(\delta) \left\{\frac{1}{2} - \frac{1}{2\sqrt{2}} \sqrt{8n\alpha^2 \left\|P_1 - P_2\right\|_{\mathrm{TV}}^2}\right\}. \tag{7.12}$$

Using the fact that $\|P_1 - P_2\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}\left(P_1 \| P_2\right)$, comparison with the original Le Cam bound (7.11) shows that for $\alpha \in [0, \frac{22}{35}]$, the effect of α-local differential privacy is to reduce the *effective sample size* from n to $4\alpha^2 n$. We illustrate use of this private version of Le Cam's bound in our analysis of the one-dimensional mean problem to follow.

Consequences for local Fano's method We now turn to consequences for the so-called local form of Fano's method. This method is based on constructing a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ that defines a 2δ-packing, meaning $\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta$ for all $v \neq v'$, satisfying

$$D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) \leq \kappa^2 \delta^2 \quad \text{for some fixed } \kappa > 0. \tag{7.13}$$

We refer to any such construction as a (δ, κ) *local packing*. Recalling Fano's inequality (2.10), the pairwise upper bounds (7.13) imply $I(X_1, \ldots, X_n; V) \leq n\kappa^2\delta^2$ by a convexity argument. We thus obtain the local Fano lower bound [92, 27] on the classical minimax risk:

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left\{1 - \frac{n\kappa^2\delta^2 + \log 2}{\log |\mathcal{V}|}\right\}. \tag{7.14}$$

We now state the extension of this bound to the α-locally private setting.

Corollary 7.3 (Private form of local Fano inequality)**.** *Consider observations from an α-locally differential private channel for some $\alpha \in [0, \frac{22}{35}]$. Given any (δ, κ) local packing, the α-private minimax risk has lower bound*

$$\mathfrak{M}_n(\Theta, \Phi \circ \rho, \alpha) \geq \Phi(\delta)\Big\{1 - \frac{4n\alpha^2\kappa^2\delta^2 + \log 2}{\log |\mathcal{V}|}\Big\}. \tag{7.15}$$

Once again, by comparison to the classical version (7.14), we see that, for all $\alpha \in [0, \frac{22}{35}]$, the price for privacy is a reduction in the effective sample size from n to $4\alpha^2 n$. The proof is again straightfoward using Theorem 7.1. By Pinsker's inequality, the pairwise bound (7.13) implies that

$$\|P_v - P_{v'}\|_{\mathrm{TV}}^2 \leq \frac{1}{2}\kappa^2\delta^2 \quad \text{for all } v \neq v'.$$

We find that $I(Z_1, \ldots, Z_n; V) \leq 4n\alpha^2\kappa^2\delta^2$ for all $\alpha \in [0, \frac{22}{35}]$ by combining this inequality with the upper bound (7.10) from Corollary 7.1. The claim (7.15) follows by combining this upper bound with the usual local Fano bound (7.14).

7.3.3 Some applications of Theorem 7.1

In this section, we illustrate the use of the α-private versions of Le Cam's and Fano's inequalities, established in the previous section as Corollaries 7.2 and 7.3 of Theorem 7.1. First, we study the problem of one-dimensional mean estimation. In addition to demonstrating how the minimax rate changes as a function of α, we also reveal some interesting (and perhaps disturbing) effects of enforcing α-local differential privacy: the effective sample size may be even polynomially smaller than $\alpha^2 n$. Our second example studies fixed design linear regression, where we again see the reduction in effective sample size from n to $\alpha^2 n$. We state each of our bounds assuming $\alpha \in [0, 1]$; the bounds hold (with different numerical constants) whenever $\alpha \in [0, C]$ for some universal constant C.

7.3.3.1 One-dimensional mean estimation

For some $k > 1$, consider the family

$$\mathcal{P}_k := \big\{\text{distributions } P \text{ such that } \mathbb{E}_P[X] \in [-1, 1] \text{ and } \mathbb{E}_P[|X|^k] \leq 1\big\},$$

and suppose that our goal is to estimate the mean $\theta(P) = \mathbb{E}_P[X]$. The next proposition characterizes the α-private minimax risk in squared ℓ_2-error

$$\mathfrak{M}_n(\theta(\mathcal{P}_k), (\cdot)^2, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}_k} \mathbb{E}\left[\left(\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)\right)^2\right].$$

Proposition 7.1. *There exist universal constants $0 < c_\ell \leq c_u < \infty$ such that for all $k > 1$ and $\alpha \in [0, 1]$, the minimax error $\mathfrak{M}_n(\theta(\mathcal{P}_k, (\cdot)^2, \alpha)$ is bounded as*

$$c_\ell \min\left\{1, \left(n\alpha^2\right)^{-\frac{k-1}{k}}\right\} \leq \mathfrak{M}_n(\theta(\mathcal{P}_k), (\cdot)^2, \alpha) \leq c_u \min\left\{1, u_k \left(n\alpha^2\right)^{-\frac{k-1}{k}}\right\}, \tag{7.16}$$

where $u_k = \max\{1, (k-1)^{-2}\}$.

We prove this result using the α-private version (7.12) of Le Cam's inequality, as stated in Corollary 7.2. See Section 8.1.3 for the details.

To understand the bounds (7.16), it is worthwhile considering some special cases, beginning with the usual setting of random variables with finite variance ($k = 2$). In the non-private setting in which the original sample $(X_1, \ldots, X_n)$ is observed, the sample mean $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$ has mean-squared error at most $1/n$. When we require α-local differential privacy, Proposition 7.1 shows that the minimax rate worsens to $1/\sqrt{n\alpha^2}$. More generally, for any $k > 1$, the minimax rate scales as $\mathfrak{M}_n(\theta(\mathcal{P}_k), (\cdot)^2, \alpha) \asymp (n\alpha^2)^{-\frac{k-1}{k}}$, ignoring k-dependent pre-factors. As $k \uparrow \infty$, the moment condition $\mathbb{E}[|X|^k] \leq 1$ becomes equivalent to the boundedness constraint $|X| \leq 1$ a.s., and we obtain the more standard parametric rate $(n\alpha^2)^{-1}$, where there is no reduction in the exponent.

More generally, the behavior of the α-private minimax rates (7.16) helps demarcate situations in which local differential privacy may or may not be acceptable. In particular, for bounded domains—where we may take $k \uparrow \infty$—local differential privacy may be quite reasonable. However, in situations in which the sample takes values in an unbounded space, local differential privacy provides much stricter constraints.

7.3.3.2 Linear regression with fixed design

We turn now to the problem of linear regression. Concretely, for a given design matrix $X \in \mathbb{R}^{n \times d}$, consider the standard linear model

$$Y = X\theta^* + \varepsilon, \tag{7.17}$$

where $\varepsilon \in \mathbb{R}^n$ is a sequence of independent, zero-mean noise variables. By rescaling as needed, we may assume that $\theta^* \in \Theta = B_2$, the Euclidean ball of radius one. Moreover, we assume that a scaling constant $\sigma < \infty$ such that the noise sequence $|\varepsilon_i| \leq \sigma$ for all i. Given the challenges of non-compactness exhibited by the location family estimation problems (cf. Proposition 7.1), this type of assumption is required for non-trivial results. We also assume that X has rank d; otherwise, the design matrix X has a non-trivial nullspace and θ^* cannot be estimated even when $\sigma = 0$.

With the model (7.17) in place, let us consider estimation of θ^* in the squared ℓ_2-error, where we provide α-locally differentially private views of the response $Y = \{Y_i\}_{i=1}^n$. By following the outline established in Section 7.3.2, we provide a sharp characterization of the α-private minimax rate. In stating the result, we let $\gamma_j(A)$ denote the jth singular value of a matrix A. (See Section 8.1.4 for the proof.)

Proposition 7.2. *In the fixed design regression model where the variables Y_i and are α-locally differentially private for some $\alpha \in [0, 1]$,*

$$\min\left\{1, \frac{\sigma^2 d}{n\alpha^2 \gamma_{\max}^2(X/\sqrt{n})}\right\} \lesssim \mathfrak{M}_n\left(\Theta, \|\cdot\|_2^2, \alpha\right) \lesssim \min\left\{1, \frac{\sigma^2 d}{\alpha^2 n \gamma_{\min}^2(X/\sqrt{n})}\right\}. \tag{7.18}$$

To interpret the bounds (7.18), it is helpful to consider some special cases. First consider the case of an orthonormal design, meaning that $\frac{1}{n}X^\top X = I_{d\times d}$. The bounds (7.18) imply that $\mathfrak{M}_n(\Theta, \|\cdot\|_2^2, \alpha) \asymp \sigma^2 d/(n\alpha^2)$, so that the α-private minimax rate is fully determined (up to constant pre-factors). Standard minimax rates for linear regression problems scale as $\sigma^2 d/n$; thus, by comparison, we see that requiring differential privacy indeed causes an effective sample size reduction from n to $n\alpha^2$. More generally, up to the difference between the maximum and minimum singular values of the design X, Proposition 7.2 provides a sharp characterization of the α-private rate for fixed-design linear regression. As the proof makes clear, the upper bounds are attained by adding Laplacian noise to the response variables Y_i and solving the resulting normal equations as in standard linear regression. In this case, the standard Laplacian mechanism [68] is optimal.

7.4 Mutual information under local privacy: Fano's method

As we have previously noted, Theorem 7.1 provides indirect upper bounds on the mutual information. However, since the resulting bounds involve pairwise distances only, as in Corollary 7.1, they must be used with local packings. Exploiting Fano's inequality in its full generality requires a more sophisticated upper bound on the mutual information under local privacy, which is the main topic of this section. We illustrate this more powerful technique by deriving lower bounds for mean estimation problems in both classical as well as high-dimensional settings under the non-interactive privacy model (7.2).

7.4.1 Variational bounds on mutual information

We begin by introducing some definitions needed to state the result. Let V be a discrete random variable uniformly distributed over some finite set $\mathcal{V}$. Given a family of distributions $\{P_v, v \in \mathcal{V}\}$, we define the *mixture distribution*

$$\overline{P} := \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v.$$

A sample $X \sim \overline{P}$ can be obtained by first drawing V from the uniform distribution over $\mathcal{V}$, and then conditionally on $V = v$, drawing X from the distribution P_v. By definition, the mutual information between the random index V and the sample X is

$$I(X; V) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\mathrm{kl}} \left(P_v \| \overline{P} \right),$$

a representation that plays an important role in our theory. As in the definition (7.3), any conditional distribution Q induces the family of marginal distributions $\{M_v, v \in \mathcal{V}\}$ and the associated mixture $\overline{M} := \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} M_v$. Our goal is to upper bound the mutual information $I(Z_1, \ldots, Z_n; V)$, where conditioned on $V = v$, the random variables Z_i are drawn according to M_v.

Our upper bound is variational in nature: it involves optimization over a subset of the space $L^\infty(\mathcal{X}) := \left\{ f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\infty < \infty \right\}$ of uniformly bounded functions, equipped with the usual norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. We define the 1-ball of the supremum norm

$$\mathcal{B}_\infty(\mathcal{X}) := \{ \gamma \in L^\infty(\mathcal{X}) \mid \|\gamma\|_\infty \leq 1 \}. \tag{7.19}$$

We show that this set describes the maximal amount of perturbation allowed in the conditional Q. Since the set $\mathcal{X}$ is generally clear from context, we typically omit this dependence. For each $v \in \mathcal{V}$, we define the linear functional $\varphi_v : L^\infty(\mathcal{X}) \to \mathbb{R}$ by

$$\varphi_v(\gamma) = \int_{\mathcal{X}} \gamma(x) (dP_v(x) - d\overline{P}(x)).$$

With these definitions, we have the following result:

Theorem 7.2. *Let $\{P_v\}_{v \in \mathcal{V}}$ be an arbitrary collection of probability measures on $\mathcal{X}$, and let $\{M_v\}_{v \in \mathcal{V}}$ be the set of marginal distributions induced by an α-differentially private distribution Q. Then*

$$\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left[D_{\mathrm{kl}} \left(M_v \| \overline{M} \right) + D_{\mathrm{kl}} \left(\overline{M} \| M_v \right) \right] \leq \frac{(e^\alpha - 1)^2}{|\mathcal{V}|} \sup_{\gamma \in \mathcal{B}_\infty(\mathcal{X})} \sum_{v \in \mathcal{V}} \left(\varphi_v(\gamma) \right)^2. \tag{7.20}$$

It is important to note that, at least up to constant factors, Theorem 7.2 is never weaker than the results provided by Theorem 7.1, including the bounds of Corollary 7.1. By definition of the linear functional φ_v, we have

$$\sup_{\gamma \in \mathcal{B}_\infty(\mathcal{X})} \sum_{v \in \mathcal{V}} \left(\varphi_v(\gamma)\right)^2 \stackrel{(i)}{\leq} \sum_{v \in \mathcal{V}} \sup_{\gamma \in \mathcal{B}_\infty(\mathcal{X})} \left(\varphi_v(\gamma)\right)^2 = 4 \sum_{v \in \mathcal{V}} \left\| P_v - \overline{P} \right\|_{\mathrm{TV}}^2,$$

where inequality (i) follows by interchanging the summation and supremum. Overall, we have

$$I(Z; V) \leq 4(e^\alpha - 1)^2 \frac{1}{|\mathcal{V}|^2} \sum_{v, v' \in \mathcal{V}} \|P_v - P_{v'}\|_{\mathrm{TV}}^2.$$

The strength of Theorem 7.2 arises from the fact that inequality (i)—the interchange of the order of supremum and summation—may be quite loose.

We now present a corollary that extends Theorem 7.2 to the setting of repeated sampling, providing a tensorization inequality analogous to Corollary 7.1. Let V be distributed uniformly at random in $\mathcal{V}$, and assume that given $V = v$, the observations X_i are sampled independently according to the distribution P_v for $i = 1, \ldots, n$. For this corollary, we require the non-interactive setting (7.2) of local privacy, where each private variable Z_i depends only on X_i.

Corollary 7.4. *Suppose that the distributions $\{Q_i\}_{i=1}^n$ are α-locally differentially private in the non-interactive setting* (7.2)*. Then*

$$I(Z_1, \ldots, Z_n; V) \leq n(e^\alpha - 1)^2 \frac{1}{|\mathcal{V}|} \sup_{\gamma \in \mathcal{B}_\infty} \sum_{v \in \mathcal{V}} \left(\varphi_v(\gamma)\right)^2. \tag{7.21}$$

We provide the proof of Corollary 7.4 in Section 8.2.2. We conjecture that the bound (7.21) also holds in the fully interactive setting, but given well-known difficulties of characterizing multiple channel capacities with feedback [47, Chapter 15], it may be challenging to show.

Theorem 7.2 and Corollary 7.4 relate the amount of mutual information between the random perturbed views Z of the data to geometric or variational properties of the underlying packing $\mathcal{V}$ of the parameter space Θ. In particular, Theorem 7.2 and Corollary 7.4 show that if we can find a packing set $\mathcal{V}$ that yields linear functionals φ_v whose sum has good "spectral" properties—meaning a small operator norm when taking suprema over L^∞-type spaces—we can provide sharper results. This requirement of nice "spectral" properties helps to exhibit the use of the generalized Fano construction in Corollary 2.1: it is often easy to find sets $\mathcal{V}$, for example $\{-1, 1\}^d$, for which randomly sampled vectors have nice independence properties—making for easier mutual information calculations—but individual vectors may not be well separated.

7.4.2 Applications of Theorem 7.2 to mean estimation

In this section, we show how Theorem 7.2, coupled with Corollary 7.4, leads to sharp characterizations of the α-private minimax rates for classical and high-dimensional mean estimation problems. Our results show that for in d-dimensional mean-estimation problems, the requirement of α-local differential privacy causes a reduction in effective sample size from n to $n\alpha^2/d$. Throughout this section, we assume that the channel Q is *non-interactive*, meaning that the random variable Z_i depends only on X_i, and so that local privacy takes the simpler form (7.2). We also state each of our results for privacy parameter $\alpha \in [0, 1]$, but note that all of our bounds hold for any constant α, with appropriate changes in the numerical pre-factors.

Before proceeding, we describe two sampling mechanisms for enforcing α-local differential privacy. Our methods for achieving the upper bounds in our minimax rates are based on unbiased estimators of a data vector, often the observation X. Let us assume we wish to construct an α-private unbiased estimate Z for the vector $v \in \mathbb{R}^d$. The following sampling strategies are based on a radius $r > 0$ and a bound $B > 0$ specified for each problem, and they require the Bernoulli random variable

$$T \sim \text{Bernoulli}(\pi_\alpha), \quad \text{where} \quad \pi_\alpha := e^\alpha/(e^\alpha + 1).$$

Strategy A: Given a vector v with $\|v\|_2 \le \mathsf{r}$, set $\widetilde{v} = \mathsf{r}v/\|v\|_2$ with probability $\frac{1}{2} + \|v\|_2/2\mathsf{r}$ and $\widetilde{v} = -\mathsf{r}v/\|v\|_2$ with probability $\frac{1}{2} - \|v\|_2/2\mathsf{r}$. Then sample $T \sim \text{Bernoulli}(\pi_\alpha)$ and set

$$Z \sim \begin{cases} \text{Uniform}(z \in \mathbb{R}^d : \langle z, \widetilde{v}\rangle > 0, \|z\|_2 = B) & \text{if } T = 1 \\ \text{Uniform}(z \in \mathbb{R}^d : \langle z, \widetilde{v}\rangle \le 0, \|z\|_2 = B) & \text{if } T = 0. \end{cases} \tag{7.22a}$$

Strategy B: Given a vector v with $\|v\|_\infty \le \mathsf{r}$, construct $\widetilde{v} \in \mathbb{R}^d$ with coordinates $\widetilde{v}_j$ sampled independently from $\{-\mathsf{r}, \mathsf{r}\}$ with probabilities $1/2 - v_j/(2\mathsf{r})$ and $1/2 + v_j/(2\mathsf{r})$. Then sample $T \sim \text{Bernoulli}(\pi_\alpha)$ and set

$$Z \sim \begin{cases} \text{Uniform}(z \in \{-B, B\}^d : \langle z, \widetilde{v}\rangle > 0) & \text{if } T = 1 \\ \text{Uniform}(z \in \{-B, B\}^d : \langle z, \widetilde{v}\rangle \le 0) & \text{if } T = 0. \end{cases} \tag{7.22b}$$

See Figure 7.2 for visualizations of these sampling strategies. By inspection, each is α-differentially private for any vector satisfying $\|v\|_2 \le \mathsf{r}$ or $\|v\|_\infty \le \mathsf{r}$ for Strategy A or B, respectively. Moreover, each strategy is efficiently implementable: Strategy A by normalizing a sample from the $\mathsf{N}(0, I_{d\times d})$ distribution, and Strategy B by rejection sampling over the scaled hypercube $\{-B, B\}^d$.

Our sampling strategies specified, we study the d-dimensional problem of estimating the mean $\theta(P) := \mathbb{E}_P[X]$ of a random vector. We consider a few different metrics for the error of a mean estimator to flesh out the testing reduction in Section 7.2. Due to the difficulties associated with differential privacy on non-compact spaces (recall Section 7.3.3.1), we focus on distributions with compact support. We defer all proofs to Section 8.4; they use a combination of Theorem 7.2 with Fano's method.

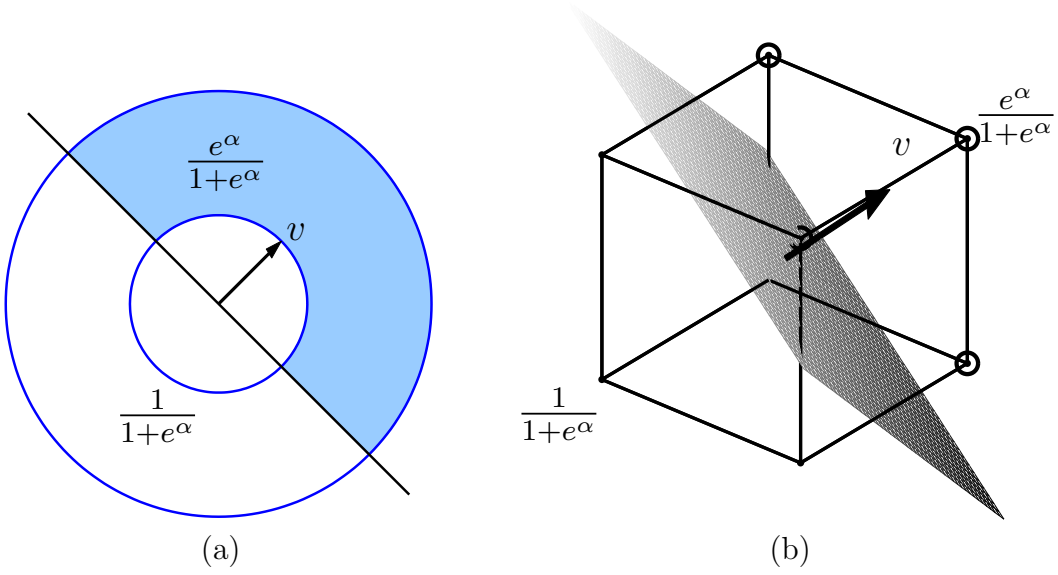

Figure 7.2. Private sampling strategies. (a) Strategy (7.22a) for the ℓ_2-ball. Outer boundary of highlighted region sampled uniformly with probability $e^\alpha/(e^\alpha+1)$. (b) Strategy (7.22b) for the ℓ_∞-ball. Circled point set sampled uniformly with probability $e^\alpha/(e^\alpha+1)$.

7.4.2.1 Minimax rates

We begin by bounding the minimax rate in the squared ℓ_2-metric. For a parameter $p \in [1,2]$ and radius $\mathsf{r} < \infty$, consider the family

$$\mathcal{P}_{p,\mathsf{r}} := \left\{\text{distributions } P \text{ supported on } B_p(\mathsf{r}) \subset \mathbb{R}^d\right\}. \tag{7.23}$$

where $B_p(\mathsf{r}) = \{x \in \mathbb{R}^d \mid \|x\|_p \leq \mathsf{r}\}$ is the ℓ_p-ball of radius r.

Proposition 7.3. *For the mean estimation problem, for all $p \in [1,2]$ and privacy levels $\alpha \in [0,1]$,*

$$\frac{1}{40}\mathsf{r}^2 \min\left\{1, \frac{1}{3\sqrt{n\alpha^2}}, \frac{d}{9n\alpha^2}\right\} \leq \mathfrak{M}_n(\theta(\mathcal{P}_{p,\mathsf{r}}), \|\cdot\|_2^2, \alpha) \lesssim \mathsf{r}^2 \min\left\{\frac{d}{n\alpha^2}, 1\right\}.$$

This bound does not depend on the norm p bounding X so long as $p \in [1,2]$, which is consistent with the classical mean estimation problem. Proposition 7.3 demonstrates the substantial difference between d-dimensional mean estimation in private and non-private settings: more precisely, the privacy constraint leads to a multiplicative penalty of d/α^2 in terms of mean-squared error. Indeed, in the non-private setting, the standard mean estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$ has mean-squared error at most r^2/n, since $\|X\|_2 \leq \|X\|_p \leq \mathsf{r}$ by assumption. Thus, Proposition 7.3 exhibits an effective sample size reduction of $n \mapsto n\alpha^2/d$.

To show the applicability of the general metric construction in Section 7.2, we now consider estimation in ℓ_∞-norm; estimation in this metric is natural in scenarios where one

wishes only to guarantee that the maximum error of any particular component in the vector θ is small. We focus in this scenario on the family $\mathcal{P}_{\infty,\mathsf{r}}$ of distributions P supported on $B_\infty(\mathsf{r}) \subset \mathbb{R}^d$.

Proposition 7.4. *For the mean estimation problem, for all $\alpha \in [0,1]$,*

$$\frac{\mathsf{r}}{12} \min\left\{1, \frac{\sqrt{d\log(2d)}}{2\sqrt{3n\alpha^2}}\right\} \le \mathfrak{M}_n(\theta(\mathcal{P}_{\infty,\mathsf{r}}), \|\cdot\|_\infty, \alpha) \lesssim \min\left\{\mathsf{r}, \frac{\mathsf{r}\sqrt{d\log(2d)}}{\sqrt{n\alpha^2}}\right\}.$$

Proposition 7.4 provides a similar message to Proposition 7.3 on the loss of statistical efficiency. This is clearest from an example: let X_i be random vectors bounded by 1 in ℓ_∞-norm. Then classical results on sub-Gaussian random variables [e.g., 36]) immediately imply that the standard non-private mean $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies $\mathbb{E}[\|\widehat{\theta} - \mathbb{E}[X]\|_\infty] \le \sqrt{\log(2d)/n}$. Comparing this result to the rate $\sqrt{d\log(2d)/n}$ of Proposition 7.4, we again see the effective sample size reduction $n \mapsto n\alpha^2/d$.

Recently, there has been substantial interest in in high-dimensional problems, in which the dimension d is larger than the sample size n, but there is a low-dimensional latent structure that makes inference possible. (See the paper by Negahban et al. [133] for a general overview.) Accordingly, let us consider an idealized version of the high-dimensional mean estimation problem, in which we assume that $\theta(P) = \mathbb{E}[X] \in \mathbb{R}^d$ has (at most) one non-zero entry, so $\|\mathbb{E}[X]\|_0 \le 1$. In the non-private case, estimation of such a s-sparse predictor in the squared ℓ_2-norm is possible at rate $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \le s\log(d/s)/n$, so that the dimension d can be exponentially larger than the sample size n. With this context, the next result exhibits that privacy can have a dramatic impact in the high-dimensional setting. Consider the family

$$\mathcal{P}^s_{\infty,\mathsf{r}} := \left\{\text{distributions } P \text{ supported on } B_\infty(\mathsf{r}) \subset \mathbb{R}^d \text{ with } \|\mathbb{E}_P[X]\|_0 \le s\right\}.$$

Proposition 7.5. *For the 1-sparse means problem, for all $\alpha \in [0,1]$,*

$$\mathsf{r}^2 \min\left\{1, \frac{d\log(2d)}{n\alpha^2}\right\} \lesssim \mathfrak{M}_n\left(\theta(\mathcal{P}^1_{\infty,\mathsf{r}}), \|\cdot\|_2^2, \alpha\right) \lesssim \mathsf{r}^2 \min\left\{1, \frac{d\log(2d)}{n\alpha^2}\right\}.$$

See Section 8.4.3 for a proof. From Proposition 7.5, it becomes clear that in locally private but non-interactive (7.2) settings, high dimensional estimation is effectively impossible.

7.4.2.2 Optimal mechanisms: attainability for mean estimation

In this section, we describe how to achieve matching upper bounds in Propositions 7.3 and 7.4 using simple and practical algorithms—namely, the "right" type of stochastic perturbation of the observations X_i coupled with a standard mean estimator. We show the optimality of privatizing via the sampling strategies (7.22a) and (7.22b); interestingly, we also show that privatizing via Laplace perturbation is strictly sub-optimal. To give a private mechanism,

we must specify the conditional distribution Q satisfying α-local differential privacy used to construct Z. In this case, given an observation X_i, we construct Z_i by perturbing X_i in such a way that $\mathbb{E}[Z_i \mid X_i = x] = x$. Each of the strategies (7.22a) and (7.22b) also requires a constant B, and we show how to choose B for each strategy to satisfy the unbiasedness condition $\mathbb{E}[Z \mid X = x] = x$.

We begin with the mean estimation problem for distributions $\mathcal{P}_{p,\mathsf{r}}$ in Proposition 7.3, for which we use the sampling scheme (7.22a). That is, let $X = x \in \mathbb{R}^d$ satisfy $\|x\|_2 \le \|x\|_p \le \mathsf{r}$. Then we construct the random vector Z according to strategy (7.22a), where we set the initial vector $v = x$ in the sampling scheme. To achieve the unbiasedness condition $\mathbb{E}[Z \mid x] = x$, we set the bound

$$B = \mathsf{r}\frac{e^\alpha + 1}{e^\alpha - 1}\frac{d\sqrt{\pi}\Gamma(\frac{d-1}{2} + 1)}{\Gamma(\frac{d}{2} + 1)} \tag{7.24}$$

(see Section 8.4.5 for a proof that $\mathbb{E}[Z \mid x] = x$ with this choice of B). Notably, the choice (7.24) implies $B \le c\mathsf{r}\sqrt{d}/\alpha$ for a universal constant $c < \infty$, since $d\Gamma(\frac{d-1}{2} + 1)/\Gamma(\frac{d}{2} + 1) \lesssim \sqrt{d}$ and $e^\alpha - 1 = \alpha + \mathcal{O}(\alpha^2)$. As a consequence, generating each Z_i by this perturbation strategy and using the mean estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n Z_i$, the estimator $\widehat{\theta}$ is unbiased for $\mathbb{E}[X]$ and satisfies

$$\mathbb{E}\left[\|\widehat{\theta} - \mathbb{E}[X]\|_2^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Z_i) \le \frac{B^2}{n} \le c\frac{\mathsf{r}^2 d}{n\alpha^2}$$

for a universal constant c.

In Proposition 7.4, we consider the family $\mathcal{P}_{\infty,\mathsf{r}}$ of distributions supported on the ℓ_∞-ball of radius r. In our mechanism for attaining the upper bound, we use the sampling scheme (7.22b) to generate the private Z_i, so that for an observation $X = x \in \mathbb{R}^d$ with $\|x\|_\infty \le \mathsf{r}$, we resample Z (from the initial vector $v = x$) according to strategy (7.22b). Again, we would like to guarantee the unbiasedness condition $\mathbb{E}[Z \mid X = x] = x$, for which we use an earlier result of ours [57]. In that paper, we show that taking

$$B = c\frac{\mathsf{r}\sqrt{d}}{\alpha} \tag{7.25}$$

for a (particular) universal constant c, we obtain the desired unbiasedness [57, Corollary 3]. Since the random variable Z satisfies $Z \in B_\infty(\mathsf{r})$ with probability 1, each coordinate $[Z]_j$ of Z is sub-Gaussian. As a consequence, we obtain via standard bounds [36] that

$$\mathbb{E}[\|\widehat{\theta} - \theta\|_\infty^2] \le \frac{B^2 \log(2d)}{n} = c^2\frac{\mathsf{r}^2 d \log(2d)}{n\alpha^2}$$

for a universal constant c, proving the upper bound in Proposition 7.4.

To conclude this section, we note that the strategy of adding Laplacian noise to the vectors X is sub-optimal. Indeed, consider the the family $\mathcal{P}_{2,1}$ of distributions supported on $B_2(1) \subset \mathbb{R}^d$ as in Proposition 7.3. To guarantee α-differential privacy using independent Laplace noise vectors for $x \in B_2(1)$, we take $Z = x + W$ where $W \in \mathbb{R}^d$ has components W_j

that are independent and distributed as Laplace($\alpha/\sqrt{d}$). We have the following information-theoretic result: if the Z_i are constructed via the Laplace noise mechanism,

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\widehat{\theta}(Z_1, \ldots, Z_n) - \mathbb{E}_P[X]\|_2^2 \right] \gtrsim \min \left\{ \frac{d^2}{n\alpha^2}, 1 \right\}. \tag{7.26}$$

See Section 8.4.4 for the proof of this claim. The poorer dimension dependence exhibted by the Laplace mechanism (7.26) in comparison to Proposition 7.3 demonstrates that sampling mechanisms must be chosen carefully, as in the strategies (7.22a)—(7.22b), in order to obtain statistically optimal rates.

7.5 Bounds on multiple pairwise divergences: Assouad's method

Thus far, we have seen how Le Cam's method and Fano's method, in the form of Theorem 7.2 and Corollary 7.4, can give sharp minimax rates for various problems. However, their application appears to be limited to simpler models: either problems whose minimax rates can be controlled via reductions to binary hypothesis tests (Le Cam's method) or for non-interactive channels satisfying the simpler definition (7.2) of local privacy (Fano's method). In this section, we show that a privatized form of Assouad's method (in the form of Lemma 2.2 via inequality (7.7)) can be used to obtain sharp minimax rates in interactive settings. In particular, it can be applied when the loss is sufficiently "decomposable", so that the coordinate-wise nature of the Assouad construction can be brought to bear. Concretely, we show an upper bound on a sum of paired KL-divergences, which combined with Assouad's method provides sharp lower bounds for several problems, including multinomial probability estimation and nonparametric density estimation. Each of these problems can be characterized in terms of an effective dimension d, and our results (paralleling those of Section 7.4) show that the requirement of α-local differential privacy causes a reduction in effective sample size from n to $n\alpha^2/d$.

7.5.1 Variational bounds on paired divergences

For a fixed $d \in \mathbb{N}$, we consider collections of distributions indexed using the Boolean hypercube $\mathcal{V} = \{-1, 1\}^d$. For each $i \in [n]$ and $v \in \mathcal{V}$, we let the distribution $P_{v,i}$ be supported on the fixed set $\mathcal{X}$, and we define the product distribution $P_v^n = \prod_{i=1}^n P_{v,i}$. Then for $j \in [d]$ we define the paired mixtures

$$P_{+j}^n = \frac{1}{2^{d-1}} \sum_{v : v_j = 1} P_v^n, \quad P_{-j}^n = \frac{1}{2^{d-1}} \sum_{v : v_j = -1} P_v^n, \quad \text{and} \quad P_{\pm j, i} = \frac{1}{2^{d-1}} \sum_{v : v_j = \pm 1} P_{v,i}. \tag{7.27}$$

(Note that P_{+j}^n is generally not a product distribution.) Recalling the marginal channel (7.3), we may then define the marginal mixtures

$$M_{+j}^n(S) := \frac{1}{2^{d-1}} \sum_{v:v_j=1} M_v^n(S) = \int Q^n(S \mid x_{1:n}) dP_{+j}^n(x_{1:n}) \quad \text{for } j = 1, \ldots, d,$$

with the distributions M_{-j}^n defined analogously. For a given pair of distributions (M, M'), we let $D_{\rm kl}^{\rm sy}(M \| M') = D_{\rm kl}(M \| M') + D_{\rm kl}(M' \| M)$ denote the symmetrized KL-divergence. Recalling the 1-ball of the supremum norm (7.19), with these definitions we have the following theorem:

Theorem 7.3. *Under the conditions of the previous paragraph, for any α-locally differentially private (7.1) channel Q, we have*

$$\sum_{j=1}^d D_{\rm kl}^{\rm sy}\left(M_{+j}^n \| M_{-j}^n\right) \le 2(e^\alpha - 1)^2 \sum_{i=1}^n \sup_{\gamma \in \mathcal{B}_\infty(\mathcal{X})} \sum_{j=1}^d \left(\int_{\mathcal{X}} \gamma(x) dP_{+j,i}(x) - dP_{-j,i}(x) \right)^2.$$

Theorem 7.3 generalizes Theorem 7.1, which corresponds to the special case $d = 1$, though has parallels with Theorem 7.2, as taking the supremum outside the summation is essential to obtain sharp results. We provide the proof of Theorem 7.2 in Section 8.3.

Theorem 7.3 allows us to prove sharper lower bounds on the minimax risk. As in the proof of Proposition 6.1 in Chapter 6 (recall inequality (6.27)), a combination of Pinsker's and the Cauchy-Schwarz inequalities implies

$$\sum_{j=1}^d \left\| M_{+j}^n - M_{-j}^n \right\|_{\rm TV} \le \frac{1}{2} \sqrt{d} \left(\sum_{j=1}^d D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\rm kl}\left(M_{-j}^n \| M_{+j}^n\right) \right)^{\frac{1}{2}}.$$

Thus, in combination with the sharper Assouad inequality (7.7), whenever P_v induces a 2δ-Hamming separation for $\Phi \circ \rho$ we have

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \ge d\delta \left[1 - \left(\frac{1}{4d} \sum_{j=1}^d D_{\rm kl}^{\rm sy}\left(M_{+j}^n \| M_{-j}^n\right) \right)^{\frac{1}{2}} \right]. \tag{7.28}$$

The combination of inequality (7.28) with Theorem 7.3 is the foundation for the remainder of this section: multinomial estimation in Section 7.5.2, and density estimation in Section 7.5.3.

7.5.2 Multinomial estimation under local privacy

For our first illustrative application of Theorem 7.3, we return to the original motivation for local privacy [179]: avoiding survey answer bias. Consider the probability simplex

$$\Delta_d := \left\{ \theta \in \mathbb{R}^d \mid \theta \ge 0 \text{ and } \sum_{j=1}^d \theta_j = 1 \right\}.$$

Any vector $\theta \in \Delta_d$ specifies a multinomial random variable taking d states, in particular with probabilities $P_\theta(X = j) = \theta_j$ for $j \in \{1, \ldots, d\}$. Given a sample from this distribution, our goal is to estimate the probability vector θ. Warner [179] studied the Bernoulli variant of this problem (corresponding to $d = 2$), proposing a mechanism known as *randomized response*: for a given survey question, respondents answer truthfully with probability $p > 1/2$ and a lie with probability $1 - p$. Here we show that an extension of this mechanism is optimal for α-locally differentially private (7.1) multinomial estimation.

7.5.2.1 Minimax rates of convergence for multinomial estimation

Our first result provides bounds on the minimax error measured in either the squared ℓ_2-norm or the ℓ_1-norm for (sequentially) interactive channels. The ℓ_1-norm norm is sometimes more appropriate for probability estimation due to its connections with total variation distance and testing.

Proposition 7.6. *For the multinomial estimation problem, for any α-locally differentially private channel* (7.1)*, there exist universal constants $0 < c_\ell \leq c_u < 5$ such that for all $\alpha \in [0, 1]$,*

$$c_\ell \min\left\{1, \frac{1}{\sqrt{n\alpha^2}}, \frac{d}{n\alpha^2}\right\} \leq \mathfrak{M}_n\left(\Delta_d, \|\cdot\|_2^2, \alpha\right) \leq c_u \min\left\{1, \frac{d}{n\alpha^2}\right\}, \tag{7.29}$$

and

$$c_\ell \min\left\{1, \frac{d}{\sqrt{n\alpha^2}}\right\} \leq \mathfrak{M}_n\left(\Delta_d, \|\cdot\|_1, \alpha\right) \leq c_u \min\left\{1, \frac{d}{\sqrt{n\alpha^2}}\right\}. \tag{7.30}$$

See Section 8.5 for the proofs of the lower bounds. We provide simple estimation strategies achieving the upper bounds in the next section.

As in the previous section, let us compare the private rates to the classical rate in which there is no privacy. The maximum likelihood estimate $\widehat{\theta}$ sets $\widehat{\theta}_j$ as the proportion of samples taking value j; it has mean-squared error

$$\mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] = \sum_{j=1}^d \mathbb{E}\left[(\widehat{\theta}_j - \theta_j)^2\right] = \frac{1}{n}\sum_{j=1}^d \theta_j(1 - \theta_j) \leq \frac{1}{n}\left(1 - \frac{1}{d}\right) < \frac{1}{n}.$$

An analogous calculation for the ℓ_1-norm yields

$$\mathbb{E}[\|\widehat{\theta} - \theta\|_1] \leq \sum_{j=1}^d \mathbb{E}[|\widehat{\theta}_j - \theta_j|] \leq \sum_{j=1}^d \sqrt{\mathrm{Var}(\widehat{\theta}_j)} \leq \frac{1}{\sqrt{n}}\sum_{j=1}^d \sqrt{\theta_j(1 - \theta_j)} < \frac{\sqrt{d}}{\sqrt{n}}.$$

Consequently, for estimation in ℓ_1 or ℓ_2-norm, the effect of providing α-differential privacy causes the effective sample size to decrease as $n \mapsto n\alpha^2/d$.

7.5.2.2 Optimal mechanisms: attainability for multinomial estimation

An interesting consequence of the lower bound (7.29) is the following: a minor variant of Warner's randomized response strategy is an optimal mechanism. There are also other relatively simple estimation strategies that achieve convergence rate $d/n\alpha^2$; the Laplace perturbation approach [68] is another. Nonetheless, its ease of use, coupled with our optimality results, provide support for randomized response as a desirable probability estimation method.

Let us demonstrate that these strategies attain the optimal rate of convergence. Since there is a bijection between multinomial observations $x \in \{1, \ldots, d\}$ and the d standard basis vectors $e_1, \ldots, e_d \in \mathbb{R}^d$, we abuse notation and represent observations x as either when designing estimation strategies. In randomized response, we construct the private vector $Z \in \{0,1\}^d$ from a multinomial observation $x \in \{e_1, \ldots, e_d\}$ by sampling d coordinates independently via the procedure

$$[Z]_j = \begin{cases} x_j & \text{with probability } \frac{\exp(\alpha/2)}{1+\exp(\alpha/2)} \\ 1 - x_j & \text{with probability } \frac{1}{1+\exp(\alpha/2)}. \end{cases} \tag{7.31}$$

The distribution (7.31) is α-differentially private: indeed, for $x, x' \in \Delta_d$ and any $z \in \{0,1\}^d$, we have

$$\frac{Q(Z = z \mid x)}{Q(Z = z \mid x')} = \exp\left(\frac{\alpha}{2}\left(\|z - x\|_1 - \|z - x'\|_1\right)\right) \in \left[\exp(-\alpha), \exp(\alpha)\right],$$

where the triangle inequality guarantees $\left| \|z - x\|_1 - \|z - x'\|_1 \right| \le 2$. We now compute the expected value and variance of the random variables Z. Using the definition (7.31), we have

$$\mathbb{E}[Z \mid x] = \frac{e^{\alpha/2}}{1 + e^{\alpha/2}} x + \frac{1}{1 + e^{\alpha/2}}(\mathbb{1} - x) = \frac{e^{\alpha/2} - 1}{e^{\alpha/2} + 1} x + \frac{1}{1 + e^{\alpha/2}} \mathbb{1}.$$

Since the random variables Z are Bernoulli, we have the variance bound $\mathbb{E}[\|Z\|_2^2] \le d$. Letting Π_{Δ_d} denote the projection operator onto the simplex, we arrive at the natural estimator

$$\widehat{\theta}_{\text{part}} := \frac{1}{n} \sum_{i=1}^{n} \left(Z_i - \mathbb{1}/(1 + e^{\alpha/2})\right) \frac{e^{\alpha/2} + 1}{e^{\alpha/2} - 1} \quad \text{and} \quad \widehat{\theta} := \Pi_{\Delta_d}\left(\widehat{\theta}_{\text{part}}\right). \tag{7.32}$$

The projection of $\widehat{\theta}_{\text{part}}$ onto the probability simplex can be done in time linear in the dimension d of the problem [34], so the estimator (7.32) is efficiently computable. Since projections onto convex sets are non-expansive, any pair of vectors in the simplex are at most ℓ_2-distance $\sqrt{2}$ apart, and $\mathbb{E}_\theta[\widehat{\theta}_{\text{part}}] = \theta$ by construction, we have

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] &\le \min\left\{2, \mathbb{E}\left[\|\widehat{\theta}_{\text{part}} - \theta\|_2^2\right]\right\} \\ &\le \min\left\{2, \frac{d}{n}\left(\frac{e^{\alpha/2} + 1}{e^{\alpha/2} - 1}\right)^2\right\} \lesssim \min\left\{1, \frac{d}{n\alpha^2}\right\}. \end{aligned}$$

Similar results hold for the ℓ_1-norm: using the same estimator, since Euclidean projections to the simplex are non-expansive for the ℓ_1 distance,

$$\mathbb{E}\left[\|\widehat{\theta} - \theta\|_1\right] \leq \min\left\{1, \sum_{j=1}^{d} \mathbb{E}\left[|\widehat{\theta}_{\text{part},j} - \theta_j|\right]\right\} \lesssim \min\left\{1, \frac{d}{\sqrt{n\alpha^2}}\right\}.$$

7.5.3 Density estimation under local privacy

In this section, we show that the effects of local differential privacy are more severe for nonparametric density estimation: instead of just a multiplicative loss in the effective sample size as in previous sections, imposing local differential privacy leads to a different convergence rate. This result holds even though we solve a problem in which the function estimated and the observations themselves belong to compact spaces.

A probability density with respect to Lebesgue measure on the interval $[0, 1]$ is a non-negative integrable function $f : [0, 1] \to \mathbb{R}_+$ that is normalized ($\int_0^1 f(x)dx = 1$). The Sobolev classes [e.g., 173, 70] are subsets of densities that satisfy certain generalized smoothness conditions. More precisely, let $\{\varphi_j\}_{j=1}^{\infty}$ be any orthonormal basis for $L^2([0, 1])$. Then any function $f \in L^2([0, 1])$ can be expanded as a sum $\sum_{j=1}^{\infty} \theta_j \varphi_j$ in terms of the basis coefficients $\theta_j := \int f(x)\varphi_j(x)dx$. By Parseval's theorem, we are guaranteed that $\{\theta_j\}_{j=1}^{\infty} \in \ell^2(\mathbb{N})$. The *Sobolev space* $\mathcal{F}_\beta[C]$ is obtained by enforcing a particular decay rate on the basis coefficients, as formalized in the following:

Definition 7.2 (Elliptical Sobolev space)**.** *For a given orthonormal basis $\{\varphi_j\}$ of $L^2([0, 1])$, smoothness parameter $\beta > 1/2$ and radius C, the Sobolev class of order β is given by*

$$\mathcal{F}_\beta[C] := \left\{ f \in L^2([0, 1]) \mid f = \sum_{j=1}^{\infty} \theta_j \varphi_j \text{ such that } \sum_{j=1}^{\infty} j^{2\beta}\theta_j^2 \leq C^2 \right\}.$$

If we choose the trignometric basis as our orthonormal basis, membership in the class $\mathcal{F}_\beta[C]$ corresponds to smoothness constraints on the derivatives of f. More precisely, for $j \in \mathbb{N}$, consider the orthonormal basis for $L^2([0, 1])$ of trigonometric functions:

$$\varphi_0(t) = 1, \quad \varphi_{2j}(t) = \sqrt{2}\cos(2\pi jt), \quad \varphi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt). \tag{7.33}$$

Let f be a β-times almost everywhere differentiable function for which $|f^{(\beta)}(x)| \leq C$ for almost every $x \in [0, 1]$ satisfying $f^{(k)}(0) = f^{(k)}(1)$ for $k \leq \beta - 1$. Then, uniformly over all such f, there is a universal constant $c \leq 2$ such that that $f \in \mathcal{F}_\beta[cC]$ (see, for instance, [173, Lemma A.3]).

Suppose our goal is to estimate a density function $f \in \mathcal{F}_\beta[C]$ and that quality is measured in terms of the squared error (squared $L^2[0, 1]$-norm)

$$\|\widehat{f} - f\|_2^2 := \int_0^1 (\widehat{f}(x) - f(x))^2 dx.$$

The well-known [188, 185, 173] (non-private) minimax squared risk scales as

$$\mathfrak{M}_n\left(\mathcal{F}_\beta, \|\cdot\|_2^2, \infty\right) \asymp n^{-\frac{2\beta}{2\beta+1}}. \tag{7.34}$$

The goal of this section is to understand how this minimax rate changes when we add an α-privacy constraint to the problem. Our main result is to demonstrate that the classical rate (7.34) is no longer attainable when we require α-local differential privacy. In particular, we prove a lower bound that is substantially larger. In Sections 7.5.3.2 and 7.5.3.3, we show how to achieve this lower bound using histogram and orthogonal series estimators.

7.5.3.1 Lower bounds on density estimation

We begin by giving our main lower bound on the minimax rate of estimation of densities when observations from the density are differentially private. We provide the proof of the following proposition in Section 8.6.1.

Proposition 7.7. *Consider the class of densities $\mathcal{F}_\beta$ defined using the trigonometric basis* (7.33)*. There exists a constant $c_\beta > 0$ such that for any α-locally differentially private channel* (7.1) *with $\alpha \in [0, 1]$, the private minimax risk has lower bound*

$$\mathfrak{M}_n\left(\mathcal{F}_\beta[1], \|\cdot\|_2^2, \alpha\right) \geq c_\beta \left(n\alpha^2\right)^{-\frac{2\beta}{2\beta+2}}. \tag{7.35}$$

The most important feature of the lower bound (7.35) is that it involves a *different polynomial exponent* than the classical minimax rate (7.34). Whereas the exponent in classical case (7.34) is $2\beta/(2\beta + 1)$, it reduces to $2\beta/(2\beta + 2)$ in the locally private setting. For example, when we estimate Lipschitz densities ($\beta = 1$), the rate degrades from $n^{-2/3}$ to $n^{-1/2}$.

Interestingly, no estimator based on Laplace (or exponential) perturbation of the observations X_i themselves can attain the rate of convergence (7.35). This fact follows from results of Carroll and Hall [38] on nonparametric deconvolution. They show that if observations X_i are perturbed by additive noise W, where the characteristic function ϕ_W of the additive noise has tails behaving as $|\phi_W(t)| = \mathcal{O}(|t|^{-a})$ for some $a > 0$, then no estimator can deconvolve $X + W$ and attain a rate of convergence better than $n^{-2\beta/(2\beta+2a+1)}$. Since the Laplace distribution's characteristic function has tails decaying as t^{-2}, no estimator based on the Laplace mechanism (applied directly to the observations) can attain rate of convergence better than $n^{-2\beta/(2\beta+5)}$. In order to attain the lower bound (7.35), we must thus study alternative privacy mechanisms.

7.5.3.2 Achievability by histogram estimators

We now turn to the mean-squared errors achieved by specific practical schemes, beginning with the special case of Lipschitz density functions ($\beta = 1$). In this special case, it suffices

to consider a private version of a classical histogram estimate. For a fixed positive integer $k \in \mathbb{N}$, let $\{\mathcal{X}_j\}_{j=1}^k$ denote the partition of $\mathcal{X} = [0,1]$ into the intervals

$$\mathcal{X}_j = [(j-1)/k, j/k) \quad \text{for } j = 1, 2, \ldots, k-1, \text{ and } \mathcal{X}_k = [(k-1)/k, 1].$$

Any histogram estimate of the density based on these k bins can be specified by a vector $\theta \in k\Delta_k$, where we recall $\Delta_k \subset \mathbb{R}_+^k$ is the probability simplex. Letting $\mathbf{1}_E$ denote the characteristic (indicator) function of the set E, any such vector $\theta \in \mathbb{R}^k$ defines a density estimate via the sum

$$f_\theta := \sum_{j=1}^k \theta_j \mathbf{1}_{\mathcal{X}_j}.$$

Let us now describe a mechanism that guarantees α-local differential privacy. Given a sample $\{X_1, \ldots, X_n\}$ from the distribution f, consider vectors

$$Z_i := \mathsf{e}_k(X_i) + W_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{7.36}$$

where $\mathsf{e}_k(X_i) \in \Delta_k$ is a k-vector with j^{th} entry equal to one if $X_i \in \mathcal{X}_j$ and zeroes in all other entries, and W_i is a random vector with i.i.d. Laplace($\alpha/2$) entries. The variables $\{Z_i\}_{i=1}^n$ so-defined are α-locally differentially private for $\{X_i\}_{i=1}^n$. Using these private variables, we form the density estimate $\widehat{f} := f_{\widehat{\theta}} = \sum_{j=1}^k \widehat{\theta}_j \mathbf{1}_{\mathcal{X}_j}$ based on the vector $\widehat{\theta} := \Pi_k\left(\frac{k}{n}\sum_{i=1}^n Z_i\right)$, where Π_k denotes the Euclidean projection operator onto the set $k\Delta_k$. By construction, we have $\widehat{f} \geq 0$ and $\int_0^1 \widehat{f}(x)dx = 1$, so $\widehat{f}$ is a valid density estimate. The following result characterizes its mean-squared estimation error:

Proposition 7.8. *Consider the estimate $\widehat{f}$ based on $k = (n\alpha^2)^{1/4}$ bins in the histogram. For any 1-Lipschitz density $f : [0,1] \to \mathbb{R}_+$, the MSE is upper bounded as*

$$\mathbb{E}_f\left[\left\|\widehat{f} - f\right\|_2^2\right] \leq 5(\alpha^2 n)^{-\frac{1}{2}} + \sqrt{\alpha} n^{-3/4}. \tag{7.37}$$

For any fixed $\alpha > 0$, the first term in the bound (7.37) dominates, and the $\mathcal{O}((\alpha^2 n)^{-\frac{1}{2}})$ rate matches the minimax lower bound (7.35) in the case $\beta = 1$. Consequently, the privatized histogram estimator is minimax-optimal for Lipschitz densities, providing a private analog of the classical result that histogram estimators are minimax-optimal for Lipshitz densities. See Section 8.6.2 for a proof of Proposition 7.8. We remark that a randomized response scheme parallel to that of Section 7.5.2.2 achieves the same rate of convergence, showing that this classical mechanism is again an optimal scheme.

7.5.3.3 Achievability by orthogonal projection estimators

For higher degrees of smoothness ($\beta > 1$), standard histogram estimators no longer achieve optimal rates in the classical setting [158]. Accordingly, we now turn to developing estimators based on orthogonal series expansion, and show that even in the setting of local privacy, they can achieve the lower bound (7.35) for all orders of smoothness $\beta \geq 1$.

Recall the elliptical Sobolev space (Definition 7.2), in which a function f is represented in terms of its basis expansion $f = \sum_{j=1}^{\infty} \theta_j \varphi_j$. This representation underlies the orthonormal series estimator as follows. Given a sample $X_{1:n}$ drawn i.i.d. according to a density $f \in L^2([0,1])$, compute the empirical basis coefficients

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i) \quad \text{for } j \in \{1, \ldots, k\}, \tag{7.38}$$

where the value $k \in \mathbb{N}$ is chosen either a priori based on known properties of the estimation problem or adaptively, for example, using cross-validation [70, 173]. Using these empirical coefficients, the density estimate is $\widehat{f} = \sum_{j=1}^{k} \widehat{\theta}_j \varphi_j$.

In our local privacy setting, we consider a mechanism that, instead of releasing the vector of coefficients $\big(\varphi_1(X_i), \ldots, \varphi_k(X_i)\big)$ for each data point, employs a random vector $Z_i = (Z_{i,1}, \ldots, Z_{i,k})$ satisfying $\mathbb{E}[Z_{i,j} \mid X_i] = \varphi_j(X_i)$ for each $j \in [k]$. We assume the basis functions are B_0-uniformly bounded, that is, $\sup_j \sup_x |\varphi_j(x)| \le B_0 < \infty$. This boundedness condition holds for many standard bases, including the trigonometric basis (7.33) that underlies the classical Sobolev classes and the Walsh basis. We generate the random variables from the vector $v \in \mathbb{R}^k$ defined by $v_j = \varphi_j(X)$ in the hypercube-based sampling scheme (7.22b), where we assume that the outer bound $B > B_0$. With this sampling strategy, iteration of expectation yields

$$\mathbb{E}[[Z]_j \mid X = x] = c_k \frac{B}{B_0 \sqrt{k}} \left(\frac{e^\alpha}{e^\alpha + 1} - \frac{1}{e^\alpha + 1} \right) \varphi_j(x), \tag{7.39}$$

where $c_k > 0$ is a constant (which is bounded independently of k). Consequently, it suffices to take $B = \mathcal{O}(B_0 \sqrt{k}/\alpha)$ to guarantee the unbiasedness condition $\mathbb{E}[[Z_i]_j \mid X_i] = \varphi_j(X_i)$.

Overall, the privacy mechanism and estimator perform the following steps:

- given a data point X_i, set the vector $v = [\varphi_j(X_i)]_{j=1}^k$
- sample Z_i according to the strategy (7.22b), starting from the vector v and using the bound $B = B_0 \sqrt{k}(e^\alpha + 1)/c_k(e^\alpha - 1)$.
- compute the density estimate

$$\widehat{f} := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{i,j} \varphi_j. \tag{7.40}$$

The resulting estimate enjoys the following guarantee, which (along with Proposition 7.8) makes clear that the private minimax lower bound (7.35) is sharp, providing a variant of the classical rates with a polynomially worse sample complexity. (See Section 8.6.3 for a proof.)

Proposition 7.9. *Let $\{\varphi_j\}$ be a B_0-uniformly bounded orthonormal basis for $L^2([0,1])$. There exists a constant c (depending only on C and B_0) such that, for any f in the Sobolev space $\mathcal{F}_\beta[C]$, the estimator (7.40) with $k = (n\alpha^2)^{1/(2\beta+2)}$ has MSE upper bounded as*

$$\mathbb{E}_f \left[\|f - \widehat{f}\|_2^2 \right] \leq c \left(n\alpha^2\right)^{-\frac{2\beta}{2\beta+2}}. \tag{7.41}$$

Before concluding our exposition, we make a few remarks on other potential density estimators. Our orthogonal series estimator (7.40) and sampling scheme (7.39), while similar in spirit to that proposed by Wasserman and Zhou [180, Sec. 6], is different in that it is locally private and requires a different noise strategy to obtain both α-local privacy and the optimal convergence rate. Lastly, similarly to our remarks on the insufficiency of standard Laplace noise addition for mean estimation, it is worth noting that density estimators that are based on orthogonal series and Laplace perturbation are sub-optimal: they can achieve (at best) rates of $(n\alpha^2)^{-\frac{2\beta}{2\beta+3}}$. See Section 8.6.4 for this result. This rate is polynomially worse than the sharp result provided by Proposition 7.9. Again, we see that appropriately chosen noise mechanisms are crucial for obtaining optimal results.

7.6 Comparison to related work

There has been a substantial amount of work in developing differentially private mechanisms, both in local and non-local settings, and a number of authors have attempted to characterize optimal mechanisms. For example, Kasiviswanathan et al. [105], working within a local differential privacy setting, study Probably-Approximately-Correct (PAC) learning problems and show that the statistical query model [108] and local learning are equivalent up to polynomial changes in the sample size. In our work, we are concerned with a finer-grained assessment of inferential procedures—that of rates of convergence of procedures and their optimality. In the remainder of this section, we discuss further connections of our work to previous research on optimality, global (non-local) differential privacy, as well as error-in-variables models.

7.6.1 Sample versus population estimation

The standard definition of differential privacy, due to Dwork et al. [68], is somewhat less restrictive than the local privacy considered here. In particular, a conditional distribution Q with output space $\mathcal{Z}$ is α-differentially private if

$$\sup \left\{ \frac{Q(S \mid x_{1:n})}{Q(S \mid x'_{1:n})} \mid x_i, x'_i \in \mathcal{X}, S \in \sigma(\mathcal{Z}), d_{\text{ham}}(x_{1:n}, x'_{1:n}) \leq 1 \right\} \leq \exp(\alpha), \tag{7.42}$$

where d_{ham} denotes the Hamming distance between sets. Local privacy as previously defined (7.1) is more stringent.

Several researchers have considered quantities similar to our minimax criteria under local (7.2) or non-local (7.42) differential privacy [19, 91, 88, 48]. However, the objective has often been substantially different from ours: instead of bounding errors based on population-based quantities, they provide bounds in which the data are assumed to be held fixed. More precisely, let $\theta : \mathcal{X}^n \to \Theta$ denote an estimator, and let $\theta(x_{1:n})$ be a sample quantity based on $x_{1:n}$. Prior work is based on *conditional minimax* risks of the form

$$\mathfrak{M}_n^{\mathsf{cond}}(\theta(\mathcal{X}), \Phi \circ \rho, \alpha) := \inf_Q \sup_{x_{1:n} \in \mathcal{X}^n} \mathbb{E}_Q \left[\Phi\big(\rho\big(\theta(x_{1:n}), \widehat{\theta}\big)\big) \mid X_{1:n} = x_{1:n} \right], \tag{7.43}$$

where $\widehat{\theta}$ is drawn according to $Q(\cdot \mid x_{1:n})$, the infimum is taken over all α-differentially private channels Q, and the supremum is taken over all possible samples of size n. The only randomness in this conditional minimax risk is provided by the channel; the data are held fixed, so there is no randomness from an underlying population distribution. A partial list of papers that use definitions of this type include Beimel et al. [19, Section 2.4], Hardt and Talwar [91, Definition 2.4], Hall et al. [88, Section 3], and De [48].

The conditional (7.43) and population minimax risk (7.5) can differ substantially, and such differences are precisely those addressed by the theory of statistical inference. The goal of inference is to draw conclusions about the *population-based quantity* $\theta(P)$ based on the sample. Moreover, lower bounds on the conditional minimax risk (7.43) do not imply bounds on the rate of estimation for the population $\theta(P)$. In fact, the conditional minimax risk (7.43) involves a supremum over *all possible samples* $x \in \mathcal{X}$, so the opposite is usually true: population risks provide lower bounds on the conditional minimax risk, as we show presently.

An illustrative example is useful to understand the differences. Consider estimation of the mean of a normal distribution with known standard deviation σ^2, in which the mean $\theta = \mathbb{E}[X] \in [-1, 1]$ is assumed to belong to the unit interval. As our Proposition 7.1 shows, it is possible to estimate the mean of a normally-distributed random variable even under α-local differential privacy (7.1). In sharp contrast, the following result shows that the conditional minimax risk is infinite for this problem:

Lemma 7.1. *Consider the normal location family* $\{\mathsf{N}(\theta, \sigma^2) \mid \theta \in [-1, 1]\}$ *under* α*-differential privacy* (7.42)*. The conditional minimax risk of the mean is* $\mathfrak{M}_n^{\mathsf{cond}}(\theta(\mathbb{R}), (\cdot)^2, \alpha) = \infty$.

Proof Assume for sake of contradiction that $\delta > 0$ satisfies

$$Q(|\widehat{\theta} - \theta(x_{1:n})| > \delta \mid x_{1:n}) \le \frac{1}{2} \quad \text{for all samples } x_{1:n} \in \mathbb{R}^n.$$

Fix $N(\delta) \in \mathbb{N}$ and choose points 2δ-separated points θ_v, $v \in [N(\delta)]$, that is, $|\theta_v - \theta_{v'}| \ge 2\delta$ for $v \neq v'$. Then the sets $\{\theta \in \mathbb{R} \mid |\theta - \theta_v| \le \delta\}$ are all disjoint, so for any pair of samples

$x_{1:n}$ and $x^v_{1:n}$ with $d_{\mathrm{ham}}(x_{1:n}, x^v_{1:n}) \le 1$,

$$Q(\exists v \in \mathcal{V} \text{ s.t. } |\widehat{\theta} - \theta_v| \le \delta \mid x_{1:n}) = \sum_{v=1}^{N(\delta)} Q(|\widehat{\theta} - \theta_v| \le \delta \mid x_{1:n}) \\ \ge e^{-\alpha} \sum_{v=1}^{N(\delta)} Q(|\widehat{\theta} - \theta_v| \le \delta \mid x^v_{1:n}).$$

We may take each sample $x^v_{1:n}$ such that $\theta(x^v_{1:n}) = \frac{1}{n}\sum_{i=1}^n x_i^v = \theta_v$ (for example, for each $v \in [N(\delta)]$ set $x_1^v = n\theta_v - \sum_{i=2}^n x_i$) and by assumption,

$$1 \ge Q(\exists v \in \mathcal{V} \text{ s.t. } |\widehat{\theta} - \theta_v| \le \delta \mid x_{1:n}) \ge e^{-\alpha} N(\delta) \frac{1}{2}.$$

Taking $N(\delta) > 2e^{\alpha}$ yields a contradiction. Our argument applies to an arbitrary $\delta > 0$, so the claim follows. □

There are variations on this result. For instance, even if the output of the mean estimator is restricted to $[-1, 1]$, the conditional minimax risk remains constant. Similar arguments apply to weakenings of differential privacy (e.g., δ-approximate α-differential privacy [67]). Conditional and population risks are very different quantities.

More generally, the population minimax risk usually lower bounds the conditional minimax risk. Suppose we measure minimax risks in some given metric ρ (so the loss $\Phi(t) = t$). Let $\widetilde{\theta}$ be any estimator based on the original sample $X_{1:n}$, and let $\widehat{\theta}$ be any estimator based on the privatized sample. We then have the following series of inequalities:

$$\mathbb{E}_{Q,P}[\rho(\theta(P), \widehat{\theta})] \le \mathbb{E}_{Q,P}[\rho(\theta(P), \widetilde{\theta})] + \mathbb{E}_{Q,P}[\rho(\widetilde{\theta}, \widehat{\theta})] \\ \le \mathbb{E}_P[\rho(\theta(P), \widetilde{\theta})] + \sup_{x_{1:n} \in \mathcal{X}^n} \mathbb{E}_{Q,P}[\rho(\widetilde{\theta}(x_{1:n}), \widehat{\theta}) \mid X_{1:n} = x_{1:n}]. \tag{7.44}$$

The population minimax risk (7.5) thus lower bounds the conditional minimax risk (7.43) via $\mathfrak{M}_n^{\mathsf{cond}}(\widetilde{\theta}(\mathcal{X}), \rho, \alpha) \ge \mathfrak{M}_n(\theta(\mathcal{P}), \rho, \alpha) - \mathbb{E}_P[\rho(\theta(P), \widetilde{\theta})]$. In particular, if there exists an estimator $\widetilde{\theta}$ based on the original (non-private data) such that $\mathbb{E}_P[\rho(\theta(P), \widetilde{\theta})] \le \frac{1}{2}\mathfrak{M}_n(\theta(\mathcal{P}), \rho, \alpha)$ we are guaranteed that

$$\mathfrak{M}_n^{\mathsf{cond}}(\widetilde{\theta}(\mathcal{X}), \rho, \alpha) \ge \frac{1}{2}\, \mathfrak{M}_n(\theta(\mathcal{P}), \rho, \alpha),$$

so the conditional minimax risk is lower bounded by a constant multiple of the population minimax risk. This lower bound holds for each of the examples in Sections 7.3–7.5; lower bounds on the α-private population minimax risk (7.5) are stronger than lower bounds on the conditional minimax risk.

To illustrate one application of the lower bound (7.44), consider the estimation of the sample mean of a data set $x_{1:n} \in \{0, 1\}^n$ under α-local privacy. This problem has been

considered before; for instance, Beimel et al. [19] study distributed protocols for this problem. In Theorem 2 of their work, they show that if a protocol has ℓ rounds of communication, the squared error in estimating the sample mean $(1/n)\sum_{i=1}^n x_i$ is $\Omega(1/(n\alpha^2\ell^2))$. The standard mean estimator $\widetilde{\theta}(x_{1:n}) = (1/n)\sum_{i=1}^n x_i$ has error $\mathbb{E}[|\widetilde{\theta}(x_{1:n}) - \theta|] \le n^{-\frac{1}{2}}$. Consequently, the lower bound (7.44) with combined with Proposition 7.1 implies

$$c\frac{1}{\sqrt{n\alpha^2}} - \frac{1}{\sqrt{n}} \le \mathfrak{M}_n(\theta(\mathcal{P}), |\cdot|, \alpha) - \sup_{\theta \in [-1,1]} \mathbb{E}[|\widetilde{\theta}(x_{1:n}) - \theta|] \le \mathfrak{M}_n^{\mathsf{cond}}(\theta(\{-1,1\}), |\cdot|, \alpha).$$

for some numerical constant $c > 0$. A corollary of our results is thus such an $\Omega(1/(n\alpha^2))$ lower bound on the conditional minimax risk for mean estimation, allowing for sequential interactivity but not multiple "rounds." An inspection of Beimel et al.'s proof technique [19, Section 4.2] shows that their lower bound also implies a lower bound of $1/n\alpha^2$ for estimation of the population mean $\mathbb{E}[X]$ in one dimension in *non-interactive* (7.2) settings; it is, however, unclear how to extend their technique to other settings.

7.6.2 Local versus non-local privacy

It is also worthwhile to make some comparisons to work on non-local forms of differential privacy, mainly to understand the differences between local and global forms of privacy. Chaudhuri and Hsu [43] provide lower bounds for estimation of certain one dimensional statistics based on a two-point family of problems. Their techniques differ from those of the current paper, and do not appear to provide bounds on the statistic being estimated, but rather one that is near to it. Beimel et al. [20] provide some bounds on sample complexity in the "probably approximate correct" (PAC) framework of learning theory, though extensions to other inferential tasks are unclear. Other work on non-local privacy [e.g., 88, 44, 164] shows that for various types of estimation problems, adding Laplacian noise leads to degraded convergence rates in at most lower-order terms. In contrast, our work shows that the Laplace mechanism may be highly sub-optimal in local privacy.

To understand convergence rates for non-local privacy, let us return to estimation of a multinomial distribution in Δ_d, based on observations $X_i \in \{e_j\}_{j=1}^d$. In this case, adding a noise vector $W \in \mathbb{R}^d$ with i.i.d. entries distributed as Laplace(αn) provides differential privacy [67]; the associated mean-squared error is at most

$$\mathbb{E}_\theta\left[\left\|\frac{1}{n}\sum_{i=1}^n X_i + W - \theta\right\|_2^2\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n X_i - \theta\right\|_2^2\right] + \mathbb{E}[\|W\|_2^2] \le \frac{1}{n} + \frac{d}{n^2\alpha^2}.$$

In particular, in the asymptotic regime $n \gg d$, there is no penalty from providing differential privacy except in higher-order terms. Similar results hold for histogram estimation [88], classification problems [44], and classical point estimation problems [164]; in this sense, local and global forms of differential privacy can be rather different.

7.7 Summary

In this chapter, we have linked minimax analysis from statistical decision theory with differential privacy, bringing some of their respective foundational principles into close contact. Our main technique, in the form of the divergence inequalities in Theorems 7.1 and 7.2, and their Corollaries 7.1–7.4, shows that applying differentially private sampling schemes essentially acts as a contraction on distributions. These contractive inequalities allow us to give sharp minimax rates for estimation in locally private settings. As we see in Chapter 10 to come, these types of results—strong data processing inequalities for certain restricted observation models—are more generally applicable, for example, to distributed estimation. With our examples in Sections 7.4.2, 7.5.2, and 7.5.3, we have developed a set of techniques that show that roughly, if one can construct a family of distributions $\{P_v\}$ on the sample space $\mathcal{X}$ that is not well "correlated" with any member of $f \in L^\infty(\mathcal{X})$ for which $f(x) \in \{-1, 1\}$, then providing privacy is costly: the contraction Theorems 7.2 and 7.3 provide is strong.

By providing sharp convergence rates for many standard statistical estimation procedures under local differential privacy, we have developed and explored some tools that may be used to better understand privacy-preserving statistical inference. We have identified a fundamental continuum along which privacy may be traded for utility in the form of accurate statistical estimates, providing a way to adjust statistical procedures to meet the privacy or utility needs of the statistician and the population being sampled.

There are a number of open questions raised by our work. It is natural to wonder whether it is possible to obtain tensorized inequalities of the form of Corollary 7.4 even for interactive mechanisms. Another important question is whether the results we have provided can be extended to settings in which standard (non-local) differential privacy holds. Such extensions could yield insights into optimal mechanisms for differentially private procedures.

Chapter 8

Technical arguments for private estimation

In this chapter, we collect proofs of all the unproven results from the previous chapter. As the chapter is entirely technical, it may be skipped by the uninterested reader.

8.1 Proof of Theorem 7.1 and related results

We now turn to the proofs of our results, beginning with Theorem 7.1 and related results. In all cases, we defer the proofs of more technical lemmas to subsequent sections.

8.1.1 Proof of Theorem 7.1

Observe that M_1 and M_2 are absolutely continuous with respect to one another, and there is a measure μ with respect to which they have densities m_1 and m_2, respectively. The channel probabilities $Q(\cdot \mid x)$ and $Q(\cdot \mid x')$ are likewise absolutely continuous, so that we may assume they have densities $q(\cdot \mid x)$ and write $m_i(z) = \int q(z \mid x) dP_i(x)$. In terms of these densities, we have

$$\begin{aligned} D_{\rm kl}\left(M_1 \| M_2\right) + D_{\rm kl}\left(M_2 \| M_1\right) &= \int m_1(z) \log \frac{m_1(z)}{m_2(z)} d\mu(z) + \int m_2(z) \log \frac{m_2(z)}{m_1(z)} d\mu(z) \\ &= \int \left(m_1(z) - m_2(z)\right) \log \frac{m_1(z)}{m_2(z)} d\mu(z). \end{aligned}$$

Consequently, we must bound both the difference $m_1 - m_2$ and the log ratio of the marginal densities. The following two auxiliary lemmas are useful:

Lemma 8.1. *For any α-locally differentially private conditional, we have*

$$|m_1(z) - m_2(z)| \le c_\alpha \inf_x q(z \mid x) \left(e^\alpha - 1\right) \left\|P_1 - P_2\right\|_{\rm TV}, \tag{8.1}$$

where $c_\alpha = \min\{2, e^\alpha\}$.

Lemma 8.2. *Let $a, b \in \mathbb{R}_+$. Then* $\left|\log \frac{a}{b}\right| \leq \frac{|a-b|}{\min\{a,b\}}$.

We prove these two results at the end of this section.

With the lemmas in hand, let us now complete the proof of the theorem. From Lemma 8.2, the log ratio is bounded as

$$\left|\log \frac{m_1(z)}{m_2(z)}\right| \leq \frac{|m_1(z) - m_2(z)|}{\min\{m_1(z), m_2(z)\}}.$$

Applying Lemma 8.1 to the numerator yields

$$\begin{aligned}\left|\log \frac{m_1(z)}{m_2(z)}\right| &\leq \frac{c_\alpha \left(e^\alpha - 1\right) \|P_1 - P_2\|_{\mathrm{TV}} \inf_x q(z \mid x)}{\min\{m_1(z), m_2(z)\}} \\ &\leq \frac{c_\alpha \left(e^\alpha - 1\right) \|P_1 - P_2\|_{\mathrm{TV}} \inf_x q(z \mid x)}{\inf_x q(z \mid x)},\end{aligned}$$

where the final step uses the inequality $\min\{m_1(z), m_2(z)\} \geq \inf_x q(z \mid x)$. Putting together the pieces leads to the bound

$$\left|\log \frac{m_1(z)}{m_2(z)}\right| \leq c_\alpha (e^\alpha - 1) \|P_1 - P_2\|_{\mathrm{TV}}.$$

Combining with inequality (8.1) yields

$$D_{\mathrm{kl}}\left(M_1 \| M_2\right) + D_{\mathrm{kl}}\left(M_2 \| M_1\right) \leq c_\alpha^2 \left(e^\alpha - 1\right)^2 \|P_1 - P_2\|_{\mathrm{TV}}^2 \int \inf_x q(z \mid x) d\mu(z).$$

The final integral is at most one, which completes the proof of the theorem.

It remains to prove Lemmas 8.1 and 8.2. We begin with the former. For any $z \in \mathcal{Z}$, we have

$$\begin{aligned}m_1(z) - m_2(z) &= \int_{\mathcal{X}} q(z \mid x) \left[dP_1(x) - dP_2(x)\right] \\ &= \int_{\mathcal{X}} q(z \mid x) \left[dP_1(x) - dP_2(x)\right]_+ + \int_{\mathcal{X}} q(z \mid x) \left[dP_1(x) - dP_2(x)\right]_- \\ &\leq \sup_{x \in \mathcal{X}} q(z \mid x) \int_{\mathcal{X}} \left[dP_1(x) - dP_2(x)\right]_+ + \inf_{x \in \mathcal{X}} q(z \mid x) \int_{\mathcal{X}} \left[dP_1(x) - dP_2(x)\right]_- \\ &= \left(\sup_{x \in \mathcal{X}} q(z \mid x) - \inf_{x \in \mathcal{X}} q(z \mid x)\right) \int_{\mathcal{X}} \left[dP_1(x) - dP_2(x)\right]_+.\end{aligned}$$

By definition of the total variation norm, we have $\int \left[dP_1 - dP_2\right]_+ = \|P_1 - P_2\|_{\mathrm{TV}}$, and hence

$$|m_1(z) - m_2(z)| \leq \sup_{x, x'} |q(z \mid x) - q(z \mid x')| \, \|P_1 - P_2\|_{\mathrm{TV}}. \tag{8.2}$$

For any $\hat{x} \in \mathcal{X}$, we may add and subtract $q(z \mid \hat{x})$ from the quantity inside the supremum, which implies that

$$\begin{aligned}\sup_{x,x'} |q(z \mid x) - q(z \mid x')| &= \inf_{\hat{x}} \sup_{x,x'} |q(z \mid x) - q(z \mid \hat{x}) + q(z \mid \hat{x}) - q(z \mid x')| \\ &\leq 2 \inf_{\hat{x}} \sup_{x} |q(z \mid x) - q(z \mid \hat{x})| \\ &= 2 \inf_{\hat{x}} q(z \mid \hat{x}) \sup_{x} \left| \frac{q(z \mid x)}{q(z \mid \hat{x})} - 1 \right|.\end{aligned}$$

Similarly, we have for any x, x'

$$|q(z \mid x) - q(z \mid x')| = q(z \mid x') \left| \frac{q(z \mid x)}{q(z \mid x')} - 1 \right| \leq e^{\alpha} \inf_{\widehat{x}} q(z \mid \widehat{x}) \left| \frac{q(z \mid x)}{q(z \mid x')} - 1 \right|.$$

Since for any choice of $x, \hat{x}$, we have $q(z \mid x)/q(z \mid \hat{x}) \in [e^{-\alpha}, e^{\alpha}]$, we find that (since $e^{\alpha} - 1 \geq 1 - e^{-\alpha}$)

$$\sup_{x,x'} |q(z \mid x) - q(z \mid x')| \leq \min\{2, e^{\alpha}\} \inf_{x} q(z \mid x) \left(e^{\alpha} - 1\right).$$

Combining with the earlier inequality (8.2) yields the claim (8.1).

To see Lemma 8.2, note that for any $x > 0$, the concavity of the logarithm implies that

$$\log(x) \leq x - 1.$$

Setting alternatively $x = a/b$ and $x = b/a$, we obtain the inequalities

$$\log \frac{a}{b} \leq \frac{a}{b} - 1 = \frac{a - b}{b} \quad \text{and} \quad \log \frac{b}{a} \leq \frac{b}{a} - 1 = \frac{b - a}{a}.$$

Using the first inequality for $a \geq b$ and the second for $a < b$ completes the proof.

8.1.2 Proof of Corollary 7.1

Let us recall the definition of the induced marginal distribution (7.3), given by

$$M_v(S) = \int_{\mathcal{X}} Q(S \mid x_{1:n}) dP_v^n(x_{1:n}) \quad \text{for } S \in \sigma(\mathcal{Z}^n).$$

For each $i = 2, \ldots, n$, we let $M_{v,i}(\cdot \mid Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}) = M_{v,i}(\cdot \mid z_{1:i-1})$ denote the (marginal over X_i) distribution of the variable Z_i conditioned on $Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}$. In addition, use the shorthand notation

$$D_{\mathrm{kl}}\left(M_{v,i} \| M_{v',i}\right) := \int_{\mathcal{Z}^{i-1}} D_{\mathrm{kl}}\left(M_{v,i}(\cdot \mid z_{1:i-1}) \| M_{v',i}(\cdot \mid z_{1:i-1})\right) dM_v^{i-1}(z_1, \ldots, z_{i-1})$$

to denote the integrated KL divergence of the conditional distributions on the Z_i. By the chain-rule for KL divergences [84, Chapter 5.3], we obtain

$$D_{\mathrm{kl}}\left(M_v^n \| M_{v'}^n\right) = \sum_{i=1}^n D_{\mathrm{kl}}\left(M_{v,i} \| M_{v',i}\right).$$

By assumption (7.1), the distribution $Q_i(\cdot \mid X_i, Z_{1:i-1})$ on Z_i is α-differentially private for the sample X_i. As a consequence, if we let $P_{v,i}(\cdot \mid Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1})$ denote the conditional distribution of X_i given the first $i-1$ values $Z_1, \ldots, Z_{i-1}$ and the packing index $V = v$, then from the chain rule and Theorem 7.1 we obtain

$$\begin{aligned} D_{\mathrm{kl}}\left(M_v^n \| M_{v'}^n\right) &= \sum_{i=1}^n \int_{\mathcal{Z}^{i-1}} D_{\mathrm{kl}}\left(M_{v,i}(\cdot \mid z_{1:i-1}) \| M_{v',i}(\cdot \mid z_{1:i-1})\right) dM_v^{i-1}(z_{1:i-1}) \\ &\leq \sum_{i=1}^n 4(e^\alpha - 1)^2 \int_{\mathcal{Z}^{i-1}} \|P_{v,i}(\cdot \mid z_{1:i-1}) - P_{v',i}(\cdot \mid z_{1:i-1})\|_{\mathrm{TV}}^2 \, dM_v^{i-1}(z_1, \ldots, z_{i-1}). \end{aligned}$$

By the construction of our sampling scheme, the random variables X_i are conditionally independent given $V = v$; thus the distribution $P_{v,i}(\cdot \mid z_{1:i-1}) = P_{v,i}$, where $P_{v,i}$ denotes the distribution of X_i conditioned on $V = v$. Consequently, we have

$$\|P_{v,i}(\cdot \mid z_{1:i-1}) - P_{v',i}(\cdot \mid z_{1:i-1})\|_{\mathrm{TV}} = \|P_{v,i} - P_{v',i}\|_{\mathrm{TV}},$$

which gives the claimed result.

8.1.3 Proof of Proposition 7.1

The minimax rate characterized by equation (7.16) involves both a lower and an upper bound, and we divide our proof accordingly. We provide the proof for $\alpha \in (0, 1]$, but note that a similar result (modulo different constants) holds for any finite value of α.

Lower bound We use Le Cam's method to prove the lower bound in equation (7.16). Fix a given constant $\delta \in (0, 1]$, with a precise value to be specified later. For $v \in \mathcal{V} \in \{-1, 1\}$, define the distribution P_v with support $\{-\delta^{-1/k}, 0, \delta^{1/k}\}$ by

$$P_v(X = \delta^{-1/k}) = \frac{\delta(1+v)}{2}, \quad P_v(X = 0) = 1 - \delta, \quad \text{and} \quad P_v(X = -\delta^{-1/k}) = \frac{\delta(1-v)}{2}.$$

By construction, we have $\mathbb{E}[|X|^k] = \delta(\delta^{-1/k})^k = 1$ and $\theta_v = \mathbb{E}_v[X] = \delta^{\frac{k-1}{k}} v$, whence the mean difference is given by $\theta_1 - \theta_{-1} = 2\delta^{\frac{k-1}{k}}$. Applying Le Cam's method (2.7) and the minimax bound (2.5) yields

$$\mathfrak{M}_n(\Theta, (\cdot)^2, Q) \geq \left(\delta^{\frac{k-1}{k}}\right)^2 \left(\frac{1}{2} - \frac{1}{2}\left\|M_1^n - M_{-1}^n\right\|_{\mathrm{TV}}\right),$$

where M_v^n denotes the marginal distribution of the samples $Z_1, \ldots, Z_n$ conditioned on $\theta = \theta_v$. Now Pinsker's inequality implies that $\left\|M_1^n - M_{-1}^n\right\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}\left(M_1^n \| M_{-1}^n\right)$, and Corollary 7.1 yields

$$D_{\mathrm{kl}}\left(M_1^n \| M_{-1}^n\right) \leq 4(e^\alpha - 1)^2 n \left\|P_1 - P_{-1}\right\|_{\mathrm{TV}}^2 = 4(e^\alpha - 1)^2 n \delta^2.$$

Putting together the pieces yields $\left\|M_1^n - M_{-1}^n\right\|_{\mathrm{TV}} \leq (e^\alpha - 1)\delta\sqrt{2n}$. For $\alpha \in (0, 1]$, we have $e^\alpha - 1 \leq 2\alpha$, and thus our earlier application of Le Cam's method implies

$$\mathfrak{M}_n(\Theta, (\cdot)^2, \alpha) \geq \left(\delta^{\frac{k-1}{k}}\right)^2 \left(\frac{1}{2} - \alpha\delta\sqrt{2n}\right).$$

Substituting $\delta = \min\{1, 1/\sqrt{32n\alpha^2}\}$ yields the claim (7.16).

Upper bound We must demonstrate an α-locally private conditional distribution Q and an estimator that achieves the upper bound in equation (7.16). We do so via a combination of truncation and addition of Laplacian noise. Define the truncation function $[\cdot]_T : \mathbb{R} \to [-T, T]$ by

$$[x]_T := \max\{-T, \min\{x, T\}\},$$

where the truncation level T is to be chosen. Let W_i be independent Laplace$(\alpha/(2T))$ random variables, and for each index $i = 1, \ldots, n$, define $Z_i := [X_i]_T + W_i$. By construction, the random variable Z_i is α-differentially private for X_i. For the mean estimator $\widehat{\theta} := \frac{1}{n}\sum_{i=1}^n Z_i$, we have

$$\mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathrm{Var}(\widehat{\theta}) + \left(\mathbb{E}[\widehat{\theta}] - \theta\right)^2 = \frac{4T^2}{n\alpha^2} + \frac{1}{n}\mathrm{Var}([X_1]_T) + \left(\mathbb{E}[Z_1] - \theta\right)^2. \tag{8.3}$$

We claim that

$$\mathbb{E}[Z] = \mathbb{E}\left[[X]_T\right] \in \left[\mathbb{E}[X] - \frac{1}{(k-1)T^{k-1}}, \mathbb{E}[X] + \frac{1}{(k-1)T^{k-1}}\right]. \tag{8.4}$$

Indeed, by the assumption that $\mathbb{E}[|X|^k] \leq 1$, we have by a change of variables that

$$\int_T^\infty x dP(x) = \int_T^\infty P(X \geq x) dx \leq \int_T^\infty \frac{1}{x^k} dx = \frac{1}{(k-1)T^{k-1}}.$$

Thus

$$\begin{aligned}\mathbb{E}[[X]_T] \geq \mathbb{E}[\min\{X, T\}] &= \mathbb{E}[\min\{X, T\} + [X - T]_+ - [X - T]_+] \\ &= \mathbb{E}[X] - \int_T^\infty (x - T) dP(x) \geq \mathbb{E}[X] - \frac{1}{(k-1)T^{k-1}}.\end{aligned}$$

A similar argument yields the upper bound in equation (8.4).

From the bound (8.3) and the inequalities that since $[X]_T \in [-T, T]$ and $\alpha^2 \leq 1$, we have

$$\mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] \leq \frac{5T^2}{n\alpha^2} + \frac{1}{(k-1)^2 T^{2k-2}} \quad \text{valid for any } T > 0.$$

Choosing $T = (5(k-1))^{-\frac{1}{2k}}(n\alpha^2)^{1/(2k)}$ yields

$$\begin{aligned}\mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] &\leq \frac{5(5(k-1))^{\frac{-1}{k}}(n\alpha^2)^{\frac{1}{k}}}{n\alpha^2} + \frac{1}{(k-1)^2(5(k-1))^{-1+1/k}(n\alpha^2)^{1-1/k}} \\ &= 5^{1-1/k}\left(1 + \frac{1}{k-1}\right)\frac{1}{(k-1)^{\frac{1}{k}}(n\alpha^2)^{1-\frac{1}{k}}}.\end{aligned}$$

Since $(1 + (k-1)^{-1})(k-1)^{-\frac{1}{k}} < (k-1)^{-1} + (k-1)^{-2}$ for $k \in (1, 2)$ and is bounded by $1 + (k-1)^{-1} \leq 2$ for $k \in [2, \infty]$, the upper bound (7.16) follows.

8.1.4 Proof of Proposition 7.2

We now turn to the proof of minimax rates for fixed design linear regression.

Lower bound We use a slight generalization of the α-private form (7.15) of the local Fano inequality from Corollary 7.3. For concreteness, we assume throughout that $\alpha \in [0, \frac{23}{35}]$, but analogous arguments hold for any bounded α with changes only in the constant pre-factors. Consider an instance of the linear regression model (7.17) in which the noise variables $\{\varepsilon_i\}_{i=1}^n$ are drawn i.i.d. from the uniform distribution on $[-\sigma, +\sigma]$. Our first step is to construct a suitable packing of the unit sphere $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ in ℓ_2-norm:

Lemma 8.3. *There exists a 1-packing $\mathcal{V} = \{v^1, \ldots, v^N\}$ of the unit sphere $\mathbb{S}^{d-1}$ with $N \geq \exp(d/8)$.*

Proof By the Varshamov-Gilbert bound [e.g., 188, Lemma 4], there is a packing $\mathcal{H}_d$ of the d-dimensional hypercube $\{-1, 1\}^d$ of size $|\mathcal{H}_d| \geq \exp(d/8)$ satisfying $\|a - a'\|_1 \geq d/2$ for all $a, a' \in \mathcal{H}_d$ with $a \neq a'$. For each $a \in \mathcal{H}_d$, set $v_a = a/\sqrt{d}$, so that $\|v_a\|_2 = 1$ and $\|v_a - v_{a'}\|_2^2 \geq d/d = 1$ for $a \neq a' \in \mathcal{H}_d$. Setting $\mathcal{V} = \{v_a \mid a \in \mathcal{H}_d\}$ gives the desired result. □

For a fixed $\delta \in (0, 1]$ to be chosen shortly, define the family of vectors $\{\theta_v, v \in \mathcal{V}\}$ with $\theta_v := \delta v$. Since $\|v\|_2 \leq 1$, we have $\|\theta_v - \theta_{v'}\|_2 \leq 2\delta$. Let $P_{v,i}$ denote the distribution of Y_i conditioned on $\theta^* = \theta_v$. By the form of the linear regression model (7.17) and our assumption on the noise variable ε_i, $P_{v,i}$ is uniform on the interval $[\langle\theta_v, x_i\rangle - \sigma, \langle\theta_v, x_i\rangle + \sigma]$. Consequently,

for $v \neq v' \in \mathcal{V}$, we have

$$\begin{aligned}\|P_{v,i} - P_{v',i}\|_{\mathrm{TV}} &= \frac{1}{2}\int |p_{v,i}(y) - p_{v',i}(y)|dy \\ &\leq \frac{1}{2}\left[\frac{1}{2\sigma}|\langle\theta_v, x_i\rangle - \langle\theta_{v'}, x_i\rangle| + \frac{1}{2\sigma}|\langle\theta_v, x_i\rangle - \langle\theta_{v'}, x_i\rangle|\right] = \frac{1}{2\sigma}|\langle\theta_v - \theta_{v'}, x_i\rangle|.\end{aligned}$$

Letting V denote a random sample from the uniform distribution on $\mathcal{V}$, Corollary 7.1 implies that the mutual information is upper bounded as

$$\begin{aligned}I(Z_1, \ldots, Z_n; V) &\leq 4(e^\alpha - 1)^2 \sum_{i=1}^n \frac{1}{|\mathcal{V}|^2}\sum_{v,v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{\mathrm{TV}}^2 \\ &\leq \frac{(e^\alpha - 1)^2}{\sigma^2}\sum_{i=1}^n \frac{1}{|\mathcal{V}|^2}\sum_{v,v' \in \mathcal{V}} (\langle\theta_v - \theta_{v'}, x_i\rangle)^2 \\ &= \frac{(e^\alpha - 1)^2}{\sigma^2}\frac{1}{|\mathcal{V}|^2}\sum_{v,v' \in \mathcal{V}} (\theta_v - \theta_{v'})^\top X^\top X(\theta_v - \theta_{v'}).\end{aligned}$$

Since $\theta_v = \delta v$, we have by definition of the maximum singular value that

$$(\theta_v - \theta_{v'})^\top X^\top X(\theta_v - \theta_{v'}) \leq \delta^2 \|v - v'\|_2^2 \gamma_{\max}(X^\top X) \leq 4\delta^2\gamma_{\max}^2(X) = 4n\delta^2\gamma_{\max}^2(X/\sqrt{n}).$$

Putting together the pieces, we find that

$$I(Z_1, \ldots, Z_n; V) \leq \frac{4n\delta^2(e^\alpha - 1)^2}{\sigma^2}\gamma_{\max}^2(X/\sqrt{n}) \leq \frac{8n\alpha^2\delta^2}{\sigma^2}\gamma_{\max}^2(X/\sqrt{n}),$$

where the second inequality is valid for $\alpha \in [0, \frac{23}{35}]$. Consequently, Fano's inequality combined with the packing set $\mathcal{V}$ from Lemma 8.3 implies that

$$\mathfrak{M}_n\left(\theta, \|\cdot\|_2^2, \alpha\right) \geq \frac{\delta^2}{4}\left(1 - \frac{8n\delta^2\alpha^2\gamma_{\max}^2(X/\sqrt{n})/\sigma^2 + \log 2}{d/8}\right).$$

We split the remainder of the analysis into cases.
Case 1: First suppose that $d \geq 16$. Then setting $\delta^2 = \min\{1, \frac{d\sigma^2}{128n\gamma_{\max}^2(X/\sqrt{n})}\}$ implies that

$$\frac{8n\delta^2\alpha^2\gamma_{\max}^2(X/\sqrt{n})/\sigma^2 + \log 2}{d/8} \leq 8\left[\frac{\log 2}{d} + \frac{64}{128}\right] < \frac{7}{8}.$$

As a consequence, we have the lower bound

$$\mathfrak{M}_n\left(\theta, \|\cdot\|_2^2, \alpha\right) \geq \frac{1}{4}\min\left\{1, \frac{d\sigma^2}{128n\gamma_{\max}^2(X/\sqrt{n})}\right\} \cdot \frac{1}{8},$$

which yields the claim for $d \geq 16$.

Case 2: Otherwise, we may assume that $d < 16$. In this case, a lower bound for the case $d = 1$ is sufficient, since apart from constant factors, the same bound holds for all $d < 16$. We use the Le Cam method based on a two point comparison. Indeed, let $\theta_1 = \delta$ and $\theta_2 = -\delta$ so that the total variation distance is at upper bounded $\|P_{1,i} - P_{2,i}\|_{\mathrm{TV}} \leq \frac{\delta}{\sigma}|x_i|$. By Corollary 7.2, we have

$$\mathfrak{M}_n\left(\theta, (\cdot)^2, \alpha\right) \geq \delta^2 \left(\frac{1}{2} - \delta \frac{(e^\alpha - 1)}{\sigma} \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}}\right).$$

Letting $x = (x_1, \ldots, x_n)$ and setting $\delta^2 = \min\{1, \sigma^2/(16(e^\alpha - 1)^2 \|x\|_2^2)\}$ gives the desired result.

Upper bound We now turn to the upper bound, for which we need to specify a private conditional Q and an estimator $\widehat{\theta}$ that achieves the stated upper bound on the mean-squared error. Let W_i be independent Laplace$(\alpha/(2\sigma))$ random variables. Then the additively perturbed random variable $Z_i = Y_i + W_i$ is α-differentially private for Y_i, since by assumption the response $Y_i \in [\langle\theta, x_i\rangle - \sigma, \langle\theta, x_i\rangle + \sigma]$. We now claim that the standard least-squares estimator of θ^* achieves the stated upper bound. Indeed, the least-squares estimate is given by

$$\widehat{\theta} = (X^\top X)^{-1} X^\top Y = (X^\top X)^{-1} X^\top (X\theta^* + \varepsilon + W).$$

Moreover, from the independence of W and ε, we have

$$\mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_2^2\right] = \mathbb{E}\left[\|(X^\top X)^{-1} X^\top (\varepsilon + W)\|_2^2\right] = \mathbb{E}\left[\|(X^\top X)^{-1} X^\top \varepsilon\|_2^2\right] + \mathbb{E}\left[\|(X^\top X)^{-1} X^\top W)\|_2^2\right].$$

Since $\varepsilon \in [-\sigma, \sigma]^n$, we know that $\mathbb{E}[\varepsilon\varepsilon^\top] \preceq \sigma^2 I_{n\times n}$, and for the given choice of W, we have $\mathbb{E}[WW^\top] = (4\sigma^2/\alpha^2) I_{n\times n}$. Since $\alpha \leq 1$, we thus find

$$\mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_2^2\right] \leq \frac{5\sigma^2}{\alpha^2} \operatorname{tr}\left(X(X^\top X)^{-2} X^\top\right) = \frac{5\sigma^2}{\alpha^2} \operatorname{tr}\left((X^\top X)^{-1}\right).$$

Noting that $\operatorname{tr}((X^\top X)^{-1}) \leq d/\gamma_{\min}^2(X) = d/n\gamma_{\min}^2(X/\sqrt{n})$ gives the claimed upper bound.

8.2 Proof of Theorem 7.2 and related results

In this section, we collect together the proof of Theorem 7.2 and related corollaries.

8.2.1 Proof of Theorem 7.2

Let $\mathcal{Z}$ denote the domain of the random variable Z. We begin by reducing the problem to the case when $\mathcal{Z} = \{1, 2, \ldots, k\}$ for an arbitrary positive integer k. Indeed, in the general

setting, we let $\mathcal{K} = \{K_i\}_{i=1}^k$ be any (measurable) finite partition of $\mathcal{Z}$, where for $z \in \mathcal{Z}$ we let $[z]_\mathcal{K} = K_i$ for the K_i such that $z \in K_i$. The KL divergence $D_{\rm kl}\left(M_v \| \overline{M}\right)$ can be defined as the supremum of the (discrete) KL divergences between the random variables $[Z]_\mathcal{K}$ sampled according to M_v and $\overline{M}$ over all partitions $\mathcal{K}$ of $\mathcal{Z}$; for instance, see Gray [84, Chapter 5]. Consequently, we can prove the claim for $\mathcal{Z} = \{1, 2, \ldots, k\}$, and then take the supremum over k to recover the general case. Accordingly, we can work with the probability mass functions $m(z \mid v) = M_v(Z = z)$ and $\overline{m}(z) = \overline{M}(Z = z)$, and we may write

$$D_{\rm kl}\left(M_v \| \overline{M}\right) + D_{\rm kl}\left(\overline{M} \| M_v\right) = \sum_{z=1}^k \left(m(z \mid v) - \overline{m}(z)\right) \log \frac{m(z \mid v)}{\overline{m}(z)}. \tag{8.5}$$

Throughout, we will also use (without loss of generality) the probability mass functions $q(z \mid x) = Q(Z = z \mid X = x)$, where we note that $m(z \mid v) = \int q(z \mid x) dP_v(x)$.

Now we use Lemma 8.2 from the proof of Theorem 7.1 to complete the proof of Theorem 7.2. Starting with equality (8.5), we have

$$\begin{aligned}
\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left[D_{\rm kl}\left(M_v \| \overline{M}\right) + D_{\rm kl}\left(\overline{M} \| M_v\right)\right] &\le \sum_{v \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{z=1}^k \left|m(z \mid v) - \overline{m}(z)\right| \left|\log \frac{m(z \mid v)}{\overline{m}(z)}\right| \\
&\le \sum_{v \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{z=1}^k \left|m(z \mid v) - \overline{m}(z)\right| \frac{\left|m(z \mid v) - \overline{m}(z)\right|}{\min\left\{\overline{m}(z), m(z \mid v)\right\}}.
\end{aligned}$$

Now, we define the measure m^0 on $\mathcal{Z} = \{1, \ldots, k\}$ by $m^0(z) := \inf_{x \in \mathcal{X}} q(z \mid x)$. It is clear that $\min\left\{\overline{m}(z), m(z \mid v)\right\} \ge m^0(z)$, whence we find

$$\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left[D_{\rm kl}\left(M_v \| \overline{M}\right) + D_{\rm kl}\left(\overline{M} \| M_v\right)\right] \le \sum_{v \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{z=1}^k \frac{\left(m(z \mid v) - \overline{m}(z)\right)^2}{m^0(z)}.$$

It remains to bound the final sum. For any constant $c \in \mathbb{R}$, we have

$$m(z \mid v) - \overline{m}(z) = \int_\mathcal{X} \left(q(z \mid x) - c\right) \left(dP_v(x) - d\overline{P}(x)\right).$$

We define a set of functions $f : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$ (depending implicitly on q) by

$$\mathcal{F}_\alpha := \left\{f \mid f(z, x) \in [1, e^\alpha] m^0(z) \text{ for all } z \in \mathcal{Z} \text{ and } x \in \mathcal{X}\right\}.$$

By the definition of differential privacy, when viewed as a joint mapping from $\mathcal{Z} \times \mathcal{X} \to \mathbb{R}$, the conditional p.m.f. q satisfies $\{(z, x) \mapsto q(z \mid x)\} \in \mathcal{F}_\alpha$. Since constant (with respect to x) shifts do not change the above integral, we can modify the range of functions in $\mathcal{F}_\alpha$ by subtracting $m^0(z)$ from each, yielding the set

$$\mathcal{F}'_\alpha := \left\{f \mid f(z, x) \in [0, e^\alpha - 1]\, m^0(z) \text{ for all } z \in \mathcal{Z} \text{ and } x \in \mathcal{X}\right\}.$$

As a consequence, we find that

$$\sum_{v\in\mathcal{V}}\left(m(z\mid v)-\overline{m}(z)\right)^2 \le \sup_{f\in\mathcal{F}_\alpha}\left\{\sum_{v\in\mathcal{V}}\left(\int_{\mathcal{X}} f(z,x)\left(dP_v(x)-d\overline{P}(x)\right)\right)^2\right\}$$
$$= \sup_{f\in\mathcal{F}'_\alpha}\left\{\sum_{v\in\mathcal{V}}\left(\int_{\mathcal{X}} \left(f(z,x)-m^0(z)\right)\left(dP_v(x)-d\overline{P}(x)\right)\right)^2\right\}.$$

By inspection, when we divide by $m^0(z)$ and recall the definition of the set $\mathcal{B}_\infty \subset L^\infty(\mathcal{X})$ in the statement of Theorem 7.2, we obtain

$$\sum_{v\in\mathcal{V}}\left(m(z\mid v)-\overline{m}(z)\right)^2 \le \left(m^0(z)\right)^2(e^\alpha-1)^2 \sup_{\gamma\in\mathcal{B}_\infty}\sum_{v\in\mathcal{V}}\left(\int_{\mathcal{X}}\gamma(x)\left(dP_v(x)-d\overline{P}(x)\right)\right)^2.$$

Putting together our bounds, we have

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left[D_{\rm kl}\left(M_v\|\overline{M}\right)+D_{\rm kl}\left(\overline{M}\|M_v\right)\right]$$
$$\le (e^\alpha-1)^2\sum_{z=1}^{k}\frac{1}{|\mathcal{V}|}\frac{\left(m^0(z)\right)^2}{m^0(z)}\sup_{\gamma\in\mathcal{B}_\infty}\sum_{v\in\mathcal{V}}\left(\int_{\mathcal{X}}\gamma(x)\left(dP_v(x)-d\overline{P}(x)\right)\right)^2$$
$$\le (e^\alpha-1)^2\frac{1}{|\mathcal{V}|}\sup_{\gamma\in\mathcal{B}_\infty}\sum_{v\in\mathcal{V}}\left(\int_{\mathcal{X}}\gamma(x)\left(dP_v(x)-d\overline{P}(x)\right)\right)^2,$$

since $\sum_z m^0(z)\le 1$, which is the statement of the theorem.

8.2.2 Proof of Corollary 7.4

In the non-interactive setting (7.2), the marginal distribution M_v^n is a product measure and Z_i is conditionally independent of $Z_{1:i-1}$ given V. Thus by the chain rule for mutual information [84, Chapter 5] and the fact (as in the proof of Theorem 7.2) that we may assume w.l.o.g. that Z has finite range

$$I(Z_1,\ldots,Z_n;V)=\sum_{i=1}^n I(Z_i;V\mid Z_{1:i-1})=\sum_{i=1}^n\left[H(Z_i\mid Z_{1:i-1})-H(Z_i\mid V,Z_{1:i-1})\right].$$

Since conditioning reduces entropy and $Z_{1:i-1}$ is conditionally independent of Z_i given V, we have $H(Z_i\mid Z_{1:i-1})\le H(Z_i)$ and $H(Z_i\mid V,Z_{1:i-1})=H(Z_i\mid V)$. In particular, we have

$$I(Z_1,\ldots,Z_n;V)\le\sum_{i=1}^n I(Z_i;V)=\sum_{i=1}^n\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}D_{\rm kl}\left(M_{v,i}\|\overline{M}_i\right).$$

Applying Theorem 7.2 completes the proof.

8.3 Proof of Theorem 7.3

The proof of this theorem combines the techniques we used in the proofs of Theorems 7.1 and 7.2; the first handles interactivity, while the techniques to derive the variational bounds are reminiscent of those used in Theorem 7.2. Our first step is to note a consequence of the independence structure in Fig. 7.1 that is essential to our tensorization steps. More precisely, we claim that for any set $S \in \sigma(\mathcal{Z})$,

$$M_{\pm j}(Z_i \in S \mid z_{1:i-1}) = \int Q(Z_i \in S \mid Z_{1:i-1} = z_{1:i-1}, X_i = x) dP_{\pm j,i}(x). \tag{8.6}$$

We postpone the proof of this intermediate claim to the end of this section.

Now consider the summed KL-divergences. Let $M_{\pm j,i}(\cdot \mid z_{1:i-1})$ denote the conditional distribution of Z_i under $P_{\pm j}$, conditional on $Z_{1:i-1} = z_{1:i-1}$. As in the proof of Corollary 7.1, the chain-rule for KL-divergences [e.g. 84, Chapter 5] implies

$$D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right) = \sum_{i=1}^n \int_{\mathcal{Z}^{i-1}} D_{\rm kl}\left(M_{+j}(\cdot \mid z_{1:i-1}) \| M_{-j}(\cdot \mid z_{1:i-1})\right) dM_{+j}^{i-1}(z_{1:i-1}).$$

For notational convenience in the remainder of the proof, we recall that the symmetrized KL divergence between measures M and M' as $D_{\rm kl}^{\rm sy}\left(M \| M'\right) = D_{\rm kl}\left(M \| M'\right) + D_{\rm kl}\left(M' \| M\right)$.

Defining $\overline{P} := 2^{-d} \sum_{v \in \mathcal{V}} P_v^n$, we have $2\overline{P} = P_{+j} + P_{-j}$ for each j simultaneously, We also introduce $\overline{M}(S) = \int Q(S \mid x_{1:n}) d\overline{M}(x_{1:n})$, and let $\mathbb{E}_{\pm j}$ denote the expectation taken under the marginals $M_{\pm j}$. We then have

$$\begin{aligned}
&D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\rm kl}\left(M_{-j}^n \| M_{+j}^n\right) \\
&= \sum_{i=1}^n \left(\mathbb{E}_{+j}[D_{\rm kl}\left(M_{+j,i}(\cdot \mid Z_{1:i-1}) \| M_{-j,i}(\cdot \mid Z_{1:i-1})\right)] + \mathbb{E}_{-j}[D_{\rm kl}\left(M_{-j,i}(\cdot \mid Z_{1:i-1}) \| M_{+j,i}(\cdot \mid Z_{1:i-1})\right)] \right) \\
&\le \sum_{i=1}^n \left(\mathbb{E}_{+j}[D_{\rm kl}^{\rm sy}\left(M_{+j,i}(\cdot \mid Z_{1:i-1}) \| M_{-j,i}(\cdot \mid Z_{1:i-1})\right)] + \mathbb{E}_{-j}[D_{\rm kl}^{\rm sy}\left(M_{+j,i}(\cdot \mid Z_{1:i-1}) \| M_{-j,i}(\cdot \mid Z_{1:i-1})\right)] \right) \\
&= 2 \sum_{i=1}^n \int_{\mathcal{Z}^{i-1}} D_{\rm kl}^{\rm sy}\left(M_{+j,i}(\cdot \mid z_{1:i-1}) \| M_{-j,i}(\cdot \mid z_{1:i-1})\right) d\overline{M}^{i-1}(z_{1:i-1}),
\end{aligned}$$

where we have used the definition of $\overline{M}$ and that $2\overline{P} = P_{+j} + P_{-j}$ for all j. Summing over $j \in [d]$ yields

$$\sum_{j=1}^d D_{\rm kl}^{\rm sy}\left(M_{+j}^n \| M_{-j}^n\right) \le 2 \sum_{i=1}^n \int_{\mathcal{Z}^{i-1}} \sum_{j=1}^d \underbrace{D_{\rm kl}^{\rm sy}\left(M_{+j,i}(\cdot \mid z_{1:i-1}) \| M_{-j,i}(\cdot \mid z_{1:i-1})\right)}_{=: \mathcal{T}_{j,i}} d\overline{M}^{i-1}(z_{1:i-1}). \tag{8.7}$$

We bound the underlined expression in inequality (8.7), whose elements we denote by $\mathcal{T}_{j,i}$.

Without loss of generality (as in the proof of Theorem 7.2), we may assume $\mathcal{Z}$ is finite, and that $\mathcal{Z} = \{1, 2, \ldots, k\}$ for some positive integer k. Using the probability mass functions $m_{\pm j,i}$ and omitting the index i when it is clear from context, Lemma 8.2 implies

$$\begin{aligned}\mathcal{T}_{j,i} &= \sum_{z=1}^{k} \left(m_{+j}(z \mid z_{1:i-1}) - m_{+j}(z \mid z_{1:i-1})\right) \log \frac{m_{+j}(z \mid z_{1:i-1})}{m_{-j}(z \mid z_{1:i-1})} \\ &\leq \sum_{z=1}^{k} \left(m_{+j}(z \mid z_{1:i-1}) - m_{+j}(z \mid z_{1:i-1})\right)^2 \frac{1}{\min\{m_{+j}(z \mid z_{1:i-1}), m_{-j}(z \mid z_{1:i-1})\}}.\end{aligned}$$

For each fixed $z_{1:i-1}$, define the infimal measure $m^0(z \mid z_{1:i-1}) := \inf_{x \in \mathcal{X}} q(z \mid X_i = x, z_{1:i-1})$. By construction, we have $\min\{m_{+j}(z \mid z_{1:i-1}), m_{-j}(z \mid z_{1:i-1})\} \geq m^0(z \mid z_{1:i-1})$, and hence

$$\mathcal{T}_{j,i} \leq \sum_{z=1}^{k} \left(m_{+j}(z \mid z_{1:i-1}) - m_{+j}(z \mid z_{1:i-1})\right)^2 \frac{1}{m^0(z \mid z_{1:i-1})}.$$

Recalling equality (8.6), we have

$$\begin{aligned}m_{+j}(z \mid z_{1:i-1}) - m_{+j}(z \mid z_{1:i-1}) &= \int_{\mathcal{X}} q(z \mid x, z_{1:i-1})(dP_{+j,i}(x) - dP_{-j,i}(x)) \\ &= m^0(z \mid z_{1:i-1}) \int_{\mathcal{X}} \left(\frac{q(z \mid x, z_{1:i-1})}{m^0(z \mid z_{1:i-1})} - 1\right)(dP_{+j,i}(x) - dP_{-j,i}(x)).\end{aligned}$$

From this point, the proof is similar to that of Theorem 7.2. Define the collection of functions

$$\mathcal{F}_\alpha := \{f : \mathcal{X} \times \mathcal{Z}^i \to [0, e^\alpha - 1]\}.$$

Using the definition of differential privacy, we have $\frac{q(z|x,z_{1:i-1})}{m^0(z|z_{1:i-1})} \in [1, e^\alpha]$, so there exists $f \in \mathcal{F}_\alpha$ such that

$$\begin{aligned}\sum_{j=1}^{d} \mathcal{T}_{j,i} &\leq \sum_{j=1}^{d} \sum_{z=1}^{k} \frac{\left(m^0(z \mid z_{1:i-1})\right)^2}{m^0(z \mid z_{1:i-1})} \left(\int_{\mathcal{X}} f(x, z, z_{1:i-1})(dP_{+j,i}(x) - dP_{-j,i}(x))\right)^2 \\ &= \sum_{z=1}^{k} m^0(z \mid z_{1:i-1}) \sum_{j=1}^{d} \left(\int_{\mathcal{X}} f(x, z, z_{1:i-1})(dP_{+j,i}(x) - dP_{-j,i}(x))\right)^2.\end{aligned}$$

Taking a supremum over $\mathcal{F}_\alpha$, we find the further upper bound

$$\sum_{j=1}^{d} \mathcal{T}_{j,i} \leq \sum_{z=1}^{k} m^0(z \mid z_{1:i-1}) \sup_{f \in \mathcal{F}_\alpha} \sum_{j=1}^{d} \left(\int_{\mathcal{X}} f(x, z, z_{1:i-1})(dP_{+j,i}(x) - dP_{-j,i}(x))\right)^2.$$

The inner supremum may be taken independently of z and $z_{1:i-1}$, so we rescale by $(e^\alpha - 1)$ to obtain our penultimate inequality

$$\sum_{j=1}^{d} D_{\mathrm{kl}}^{\mathrm{sy}}\left(M_{+j,i}(\cdot \mid z_{1:i-1}) \| M_{-j,i}(\cdot \mid z_{1:i-1})\right)$$
$$\leq (e^\alpha - 1)^2 \sum_{z=1}^{k} m^0(z \mid z_{1:i-1}) \sup_{\gamma \in \mathcal{B}_\infty(\mathcal{X})} \sum_{j=1}^{d} \left(\int_{\mathcal{X}} \gamma(x)(dP_{+j,i}(x) - dP_{-j,i}(x)) \right)^2.$$

Noting that m^0 sums to a quantity ≤ 1 and substituting the preceding expression in inequality (8.7) completes the proof.

Finally, we return to prove our intermediate marginalization claim (8.6). We have that

$$\begin{aligned} M_{\pm j}(Z_i \in S \mid z_{1:i-1}) &= \int Q(Z_i \in S \mid z_{1:i-1}, x_{1:n}) dP_{\pm j}(x_{1:n} \mid z_{1:i-1}) \\ &\stackrel{(i)}{=} \int Q(Z_i \in S \mid z_{1:i-1}, x_i) dP_{\pm j}(x_{1:n} \mid z_{1:i-1}) \\ &\stackrel{(ii)}{=} \int Q(Z_i \in S \mid Z_{1:i-1} = z_{1:i-1}, X_i = x) dP_{\pm j,i}(x), \end{aligned}$$

where equality (i) follows by the assumed conditional independence structure of Q (recall Figure 7.1) and equality (ii) is a consequence of the independence of X_i and $Z_{1:i-1}$ under $P_{\pm j}$. That is, we have $P_{+j}(X_i \in S \mid Z_{1:i-1} = z_{1:i-1}) = P_{+j,i}(S)$ by the definition of P_v^n as a product and that $P_{\pm j}$ are a mixture of the products P_v^n.

8.4 Proofs of multi-dimensional mean-estimation results

In this section, we prove the main results in Section 7.4.2 from the previous chapter. At a high level, our proofs of these results consist of three steps, the first of which is relatively standard, while the second two exploit specific aspects of the local privacy setting. We outline them here:

(1) The first step is a standard reduction, based on inequalities (2.5)–(2.15) in Section 2.2, from an estimation problem to a multi-way testing problem that involves discriminating between indices v contained within some subset $\mathcal{V}$ of $\mathbb{R}^d$. (Recall also inequalities (7.6) and (7.7) in Section 7.2.)

(2) The second step is the selection of the set $\mathcal{V}$, then choosing the appropriate separation radius δ to apply inequality (7.6); essentially, we require the existence of a well-separated set: one for which ratio of the packing set size $|\mathcal{V}|$ to neighborhood size $N_t^{\max}$ is large enough relative to the separation $\delta(t)$ defined by expression (2.14).

(3) The final step is to apply Theorem 7.2 in order to control the mutual information associated with the testing problem. Doing so requires bounding the supremum in Corollary 7.4 via the operator norm of $\mathrm{Cov}(V)$, which is easy to control because of the uniformity of the sampling scheme allowed by our extension (2.15) of the Fano method.

The estimation to testing reduction of Step 1 was previously described in Sections 2.2.1 and 7.2. Accordingly, the proofs to follow are devoted to the second and third steps in each case.

8.4.1 Proof of Proposition 7.3

We provide a proof of the lower bound, as we provided the argument for the upper bound in Section 7.4.2.2.

Constructing a well-separated set Let k be an arbitrary integer in $\{1, 2, \ldots, d\}$, and let $\mathcal{V}_k = \{-1, 1\}^k$ denote the k-dimensional hypercube. We extend the set $\mathcal{V}_k \subseteq \mathbb{R}^k$ to a subset of $\mathbb{R}^d$ by setting $\mathcal{V} = \mathcal{V}_k \times \{0\}^{d-k}$. For a parameter $\delta \in (0, 1/2]$ to be chosen, we define a family of probability distributions $\{P_v\}_{v \in \mathcal{V}}$ constructively. In particular, the random vector $X \sim P_v$ (a single observation) is formed by the following procedure:

$$\text{Choose index } j \in \{1, \ldots, k\} \text{ uniformly at random and set } X = \begin{cases} \mathsf{r} e_j & \text{w.p. } \frac{1+\delta v_j}{2} \\ -\mathsf{r} e_j & \text{w.p. } \frac{1-\delta v_j}{2}. \end{cases} \tag{8.8}$$

By construction, these distributions have mean vectors

$$\theta_v := \mathbb{E}_{P_v}[X] = \frac{\delta \mathsf{r}}{k} v.$$

Consequently, given the properties of the packing $\mathcal{V}$, we have $X \in B_1(\mathsf{r})$ with probability 1, and fixing $t \leq k/3$, we have that the associated separation function (2.14) satisfies

$$\delta^2(t) \geq \min \left\{ \|\theta_v - \theta_{v'}\|_2^2 \mid \|v - v'\|_1 \geq t \right\} \geq \frac{\mathsf{r}^2 \delta^2}{k^2} 2t.$$

Moreover, as in the derivation of inequality (2.16) in Section 2.2.3, we have that so long as $t \leq k/3$ and $k \geq 3$, then

$$\log \frac{|\mathcal{V}|}{N_t^{\max}} > \max \left\{ \frac{k}{6}, 2 \right\}.$$

Thus we see that the mean vectors $\{\theta_v\}_{v \in \mathcal{V}}$ provide us with an $\mathsf{r}\delta\sqrt{2t}/k$-separated set (in ℓ_2-norm) with log ratio of its size at least $\max\{k/6, 2\}$.

Upper bounding the mutual information Our next step is to bound the mutual information $I(Z_1, \ldots, Z_n; V)$ when the observations X come from the distribution (8.8) and V is uniform in the set $\mathcal{V}$. We have the following lemma, which applies so long as the channel Q is non-interactive and α-locally private (7.2). See Section 8.7.1 for the proof.

Lemma 8.4. *Fix $k \in \{1, \ldots, d\}$. Let Z_i be α-locally differentially private for X_i, and let X be sampled according to the distribution* (8.8) *conditional on $V = v$. Then*

$$I(Z_1, \ldots, Z_n; V) \le n \frac{\delta^2}{4k} (e^\alpha - 1)^2.$$

Applying testing inequalities We now show how a combination the sampling scheme (8.8) and Lemma 8.4 give us our desired lower bound. Fix $k \le d$ and let $\mathcal{V} = \{-1, 1\}^k \times \{0\}^{d-k}$. Combining Lemma 8.4 and the fact that the vectors θ_v provide a $\mathsf{r}\delta/\sqrt{2t}/k$-separated set of log-cardinality at least $\max\{k/6, 2\}$, the generalized minimax Fano bound (2.15) (and its private version (7.6)) imply that for any $k \in \{1, \ldots, d\}$ and $t \le k/3$, we have

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2, \alpha) \ge \frac{\mathsf{r}^2 \delta^2 t}{2k^2} \left(1 - \frac{n\delta^2 (e^\alpha - 1)^2/(4k) + \log 2}{\max\{k/6, 2\}}\right).$$

Because of the 1-dimensional mean-estimation lower bounds provided in Section 7.3.3.1, we may assume w.l.o.g. that $k \ge 12$. Setting $t = k/3$ and $\delta_{n,\alpha,k}^2 = \min\{1, k^2/(3n(e^\alpha - 1)^2)\}$, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2, \alpha) \ge \frac{\mathsf{r}^2 \delta_{n,\alpha,k}^2}{6k} \left(1 - \frac{1}{2} - \frac{\log 2}{2}\right) \ge \frac{1}{40} \mathsf{r}^2 \min\left\{\frac{1}{k}, \frac{k}{3n(e^\alpha - 1)^2}\right\}$$

for a universal (numerical) constant c. Since $(e^\alpha - 1)^2 < 3\alpha^2$ for $\alpha \in [0, 1]$, we obtain the lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2, \alpha) \ge \frac{1}{40} \mathsf{r}^2 \max_{k \in [d]} \left\{\min\left\{\frac{1}{k}, \frac{k}{9n\alpha^2}\right\}\right\}$$

for $\alpha \in [0, 1]$. Setting k in the preceding display to be the integer in $\{1, \ldots, d\}$ nearest $\sqrt{n\alpha^2}$ gives the result of the proposition.

8.4.2 Proof of Proposition 7.4

Since the upper bound was established in Section 7.4.2.2, we focus on the lower bound.

Constructing a well-separated set In this case, the packing set is very simple: set $\mathcal{V} = \{\pm e_j\}_{j=1}^d$ so that $|\mathcal{V}| = 2d$. Fix some $\delta \in [0, 1]$, and for $v \in \mathcal{V}$, define a distribution P_v supported on $\mathcal{X} = \{-\mathsf{r}, \mathsf{r}\}^d$ via

$$P_v(X = x) = (1 + \delta v^\top x/\mathsf{r})/2^d.$$

In words, for $v = e_j$, the coordinates of X are independent uniform on $\{-\mathsf{r}, \mathsf{r}\}$ except for the coordinate j, for which $X_j = \mathsf{r}$ with probability $1/2 + \delta v_j$ and $X_j = -\mathsf{r}$ with probability $1/2 - \delta v_j$. With this scheme, we have $\theta(P_v) = \mathsf{r}\delta v$, and since $\|\delta \mathsf{r} v - \delta \mathsf{r} v'\|_\infty \geq \delta \mathsf{r}$, we have constructed a $\delta \mathsf{r}$ packing in ℓ_∞-norm.

Upper bounding the mutual information Let V be drawn uniformly from the packing set $\mathcal{V} = \{\pm e_j\}_{j=1}^d$. With the sampling scheme in the previous paragraph, we may provide the following upper bound on the mutual information $I(Z_1, \ldots, Z_n; V)$ for any non-interactive private distribution (7.2):

Lemma 8.5. *For any non-interactive α-differentially private distribution Q, we have*

$$I(Z_1, \ldots, Z_n; V) \leq \frac{2n}{d} \left(e^\alpha - 1\right)^2 \delta^2.$$

See Section 8.7.2 for a proof.

Applying testing inequalities Finally, we turn to application of the testing inequalities. Lemma 8.5, in conjunction with the standard testing reduction and Fano's inequality (2.9), implies that

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_\infty, \alpha) \geq \frac{\mathsf{r}\delta}{2} \left(1 - \frac{2\delta^2 n (e^\alpha - 1)^2/d + \log 2}{\log(2d)}\right).$$

There is no loss of generality in assuming that $d \geq 4$, in which case the choice

$$\delta^2 = \min\left\{1, \frac{d \log(2d)}{4(e^\alpha - 1)^2 n}\right\}$$

yields the proposition.

8.4.3 Proof of Proposition 7.5

For this proposition, the construction of the packing and lower bound used in the proof of Proposition 7.4 also apply. Under these packing and sampling procedures, note that the separation of points $\theta(P_v) = \mathsf{r}\delta v$ in ℓ_2-norm is $\mathsf{r}\delta$. It thus remains to provide the upper bound. In this case, we use the sampling strategy (7.22b), as in Proposition 7.4 and Section 7.4.2.2, noting that we may take the bound B on $\|Z\|_\infty$ to be $B = c\sqrt{d}\mathsf{r}/\alpha$ for a constant c. Let θ^* denote the true mean, assumed to be s-sparse. Now consider estimating θ^* by the ℓ_1-regularized optimization problem

$$\widehat{\theta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \left\| \sum_{i=1}^n (Z_i - \theta) \right\|_2^2 + \lambda \|\theta\|_1 \right\},$$

Defining the error vector $W = \theta^* - \frac{1}{n}\sum_{i=1}^n Z_i$, we claim that

$$\lambda \geq 2\|W\|_\infty \quad \text{implies that} \quad \|\widehat{\theta} - \theta\|_2 \leq 3\lambda\sqrt{s}. \tag{8.9}$$

This result is a consequence of standard results on sparse estimation (e.g., Negahban et al. [133, Theorem 1 and Corollary 1]).

Now we note if $W_i = \theta^* - Z_i$, then $W = \frac{1}{n}\sum_{i=1}^n W_i$, and by construction of the sampling mechanism (7.22b) we have $\|W_i\|_\infty \le c\sqrt{d}\mathsf{r}/\alpha$ for a constant c. By Hoeffding's inequality and a union bound, we thus have for some (different) universal constant c that

$$\mathbb{P}(\|W\|_\infty \ge t) \le 2d \exp\left(-c\frac{n\alpha^2 t^2}{\mathsf{r}^2 d}\right) \quad \text{for } t \ge 0.$$

By taking $t^2 = \mathsf{r}^2 d(\log(2d) + \epsilon^2)/(cn\alpha^2)$, we find that $\|W\|_\infty^2 \le \mathsf{r}^2 d(\log(2d) + \epsilon^2)/(cn\alpha^2)$ with probability at least $1 - \exp(-\epsilon^2)$, which gives the claimed minimax upper bound by appropriate choice of $\lambda = c\sqrt{d \log d / n\alpha^2}$ in inequality (8.9).

8.4.4 Proof of inequality (7.26)

We prove the bound by an argument using the private form of Fano's inequality from Corollary 7.3. The proof uses the classical Varshamov-Gilbert bound (e.g. [188, Lemma 4]):

Lemma 8.6 (Varshamov-Gilbert)**.** *There is a packing $\mathcal{V}$ of the d-dimensional hypercube $\{-1, 1\}^d$ of size $|\mathcal{V}| \ge \exp(d/8)$ such that*

$$\|v - v'\|_1 \ge d/2 \quad \textit{for all distinct pairs } v, v' \in \mathcal{V}.$$

Now, let $\delta \in [0, 1]$ and the distribution P_v be a point mass at $\delta v/\sqrt{d}$. Then $\theta(P_v) = \delta v/\sqrt{d}$ and $\|\theta(P_v) - \theta(P_{v'})\|_2^2 \ge \delta^2$. In addition, a calculation implies that if M_1 and M_2 are d-dimensional Laplace(κ) distributions with means θ_1 and θ_2, respectively, then

$$D_{\rm kl}\left(M_1 \| M_2\right) = \sum_{j=1}^d \left(\exp(-\kappa|\theta_{1,j} - \theta_{2,j}|) + \kappa|\theta_{1,j} - \theta_{2,j}| - 1\right) \le \frac{\kappa^2}{2}\left\|\theta_1 - \theta_2\right\|_2^2.$$

As a consequence, we have that under our Laplacian sampling scheme for the Z and with V chosen uniformly from $\mathcal{V}$,

$$I(Z_1, \ldots, Z_n; V) \le \frac{1}{|\mathcal{V}|^2} n \sum_{v,v' \in \mathcal{V}} D_{\rm kl}\left(M_v \| M_{v'}\right) \le \frac{n\alpha^2}{2d|\mathcal{V}|^2} \sum_{v,v' \in \mathcal{V}} \left\|(\delta/\sqrt{d})(v - v')\right\|_2^2 \le \frac{2n\alpha^2\delta^2}{d}.$$

Now, applying Fano's inequality (2.9) in the context of the testing inequality (2.5), we find that

$$\inf_{\widehat{\theta}} \sup_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\|\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P_v)\|_2^2\right] \ge \frac{\delta^2}{4}\left(1 - \frac{2n\alpha^2\delta^2/d + \log 2}{d/8}\right).$$

We may assume (based on our one-dimensional results in Proposition 7.1) w.l.o.g. that $d \ge 10$. Taking $\delta^2 = d^2/(48n\alpha^2)$ then implies the result (7.26).

8.4.5 Proof of unbiasedness for sampling strategy (7.22a)

We compute the expectation of a random variable Z sampled according to the strategy (7.22a), i.e. we compute $\mathbb{E}[Z \mid v]$ for a vector $v \in \mathbb{R}^d$. By scaling, it is no loss of generality to assume that $\|v\|_2 = 1$, and using the rotational symmetry of the ℓ_2-ball, we see it is no loss of generality to assume that $v = e_1$, the first standard basis vector.

Let the function s_d denote the surface area of the sphere in $\mathbb{R}^d$, so that

$$s_d(r) = \frac{d\pi^{d/2}}{\Gamma(d/2+1)} r^{d-1}$$

is the surface area of the sphere of radius r. (We use s_d as a shorthand for $s_d(1)$ when convenient.) Then for a random variable W sampled uniformly from the half of the ℓ_2-ball with first coordinate $W_1 \geq 0$, symmetry implies that by integrating over the radii of the ball,

$$\mathbb{E}[W] = e_1 \frac{2}{s_d} \int_0^1 s_{d-1}\left(\sqrt{1-r^2}\right) r dr.$$

Making the change of variables to spherical coordinates (we use ϕ as the angle), we have

$$\frac{2}{s_d} \int_0^1 s_{d-1}\left(\sqrt{1-r^2}\right) r dr = \frac{2}{s_d} \int_0^{\pi/2} s_{d-1}\left(\cos\phi\right) \sin\phi \, d\phi = \frac{2 s_{d-1}}{s_d} \int_0^{\pi/2} \cos^{d-2}(\phi) \sin(\phi) \, d\phi.$$

Noting that $\frac{d}{d\phi} \cos^{d-1}(\phi) = -(d-1)\cos^{d-2}(\phi)\sin(\phi)$, we obtain

$$\frac{2 s_{d-1}}{s_d} \int_0^{\pi/2} \cos^{d-2}(\phi) \sin(\phi) \, d\phi = -\frac{\cos^{d-1}(\phi)}{d-1}\Big|_0^{\pi/2} = \frac{1}{d-1},$$

or that

$$\mathbb{E}[W] = e_1 \frac{(d-1)\pi^{\frac{d-1}{2}}\Gamma(\frac{d}{2}+1)}{d\pi^{\frac{d}{2}}\Gamma(\frac{d-1}{2}+1)} \frac{1}{d-1} = e_1 \underbrace{\frac{\Gamma(\frac{d}{2}+1)}{\sqrt{\pi} d \Gamma(\frac{d-1}{2}+1)}}_{=:c_d}, \tag{8.10}$$

where we define the constant c_d to be the final ratio.

Allowing again $\|v\|_2 \leq \mathsf{r}$, with the expression (8.10), we see that for our sampling strategy for Z, we have

$$\mathbb{E}[Z \mid v] = v \frac{B}{\mathsf{r}} c_d \left(\frac{e^\alpha}{e^\alpha + 1} - \frac{1}{e^\alpha + 1}\right) = \frac{B}{\mathsf{r}} c_d \frac{e^\alpha - 1}{e^\alpha + 1}.$$

Consequently, the choice

$$B = \frac{e^\alpha + 1}{e^\alpha - 1} \frac{\mathsf{r}}{c_d} = \frac{e^\alpha + 1}{e^\alpha - 1} \frac{\mathsf{r}\sqrt{\pi} d \Gamma(\frac{d-1}{2}+1)}{\Gamma(\frac{d}{2}+1)}$$

yields $\mathbb{E}[Z \mid v] = v$. Moreover, we have

$$\|Z\|_2 = B \leq \mathsf{r} \frac{e^\alpha + 1}{e^\alpha - 1} \frac{3\sqrt{\pi}\sqrt{d}}{2}$$

by Stirling's approximation to the Γ-function. By noting that $(e^\alpha + 1)/(e^\alpha - 1) \leq 3/\alpha$ for $\alpha \leq 1$, we see that $\|Z\|_2 \leq 8\mathsf{r}\sqrt{d}/\alpha$.

8.5 Proofs of multinomial estimation results

In this section, we prove the lower bounds in Proposition 7.6. Before proving the bounds, however, we outline our technique, which borrows from that in Section 8.4, and which we also use to prove the lower bounds on density estimation. The outline is as follows:

(1) As in step (1) of Section 8.4, our first step is a standard reduction using the sharper version of Assouad's method (Lemma 2.2 and inequality (7.7)) from estimation to a multiple binary hypothesis testing problem. Specifically, we perform a (essentially standard) reduction of the form (2.17).

(2) Having constructed appropriately separated binary hypothesis tests, we use apply Theorem 7.3 via inequality (7.28) to control the testing error in the binary testing problem. Applying the theorem requires bounding certain suprema related to the covariance structure of randomly selected elements of $\mathcal{V} = \{-1, 1\}^d$, as in the arguments in Section 8.4. In this case, though, the symmetry of the binary hypothesis testing problems eliminates the need for carefully constructed packings of step 8.4(2).

With this outline in mind, we turn to the proofs of inequalities (7.29) and (7.30). As we proved the upper bounds in Section 7.5.2.2, this section focuses on the argument for the lower bound. We provide the full proof for the mean-squared Euclidean error, after which we show how the result for the ℓ_1-error follows.

Our first step is to provide a lower bound of the form (2.17), giving a Hamming separation for the squared error. To that end, fix $\delta \in [0, 1]$, and for simplicity, let us assume that d is even. In this case, we set $\mathcal{V} = \{-1, 1\}^{d/2}$, and for $v \in \mathcal{V}$ let P_v be the multinomial distribution with parameter

$$\theta_v := \frac{1}{d}\mathbb{1} + \delta\frac{1}{d}\begin{bmatrix} v \\ -v \end{bmatrix} \in \Delta_d.$$

For any estimator $\widehat{\theta}$, by defining $\widehat{v}_j = \operatorname{sign}(\widehat{\theta}_j - 1/d)$ for $j \in [d/2]$ we have the lower bound

$$\|\widehat{\theta} - \theta_v\|_2^2 \geq \frac{\delta^2}{d^2}\sum_{j=1}^{d/2} \mathbf{1}\left\{\widehat{v}_j \neq v_j\right\},$$

so that by the sharper variant (7.28) of Assouad's Lemma, we obtain

$$\max_{v\in\mathcal{V}} \mathbb{E}_{P_v}[\|\widehat{\theta} - \theta_v\|_2^2] \geq \frac{\delta^2}{4d}\left[1 - \left(\frac{1}{2d}\sum_{j=1}^{d/2} D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\rm kl}\left(M_{-j}^n \| M_{+j}^n\right)\right)^{\frac{1}{2}}\right]. \tag{8.11}$$

Now we apply Theorem 7.3, which requires bounding sums of integrals $\int \gamma(dP_{+j} - dP_{-j})$, where P_{+j} is defined in expression (7.27). We claim the following inequality:

$$\sup_{\gamma\in\mathcal{B}_\infty(\mathcal{X})} \sum_{j=1}^{d/2}\left(\int_{\mathcal{X}} \gamma(x)dP_{+j}(x) - dP_{-j}(x)\right)^2 \leq \frac{8\delta^2}{d}. \tag{8.12}$$

Indeed, by construction P_{+j} is the multinomial with parameter $(1/d)\mathbb{1}+(\delta/d)[e_j^\top \; -e_j^\top]^\top \in \Delta_d$ and similarly for P_{-j}, where $e_j \in \{0,1\}^{d/2}$ denotes the jth standard basis vector. Abusing notation and identifying γ with vectors $\gamma \in [-1,1]^d$, we have

$$\int_{\mathcal{X}} \gamma(x) dP_{+j}(x) - dP_{-j}(x) = \frac{2\delta}{d} \gamma^\top \begin{bmatrix} e_j \\ -e_j \end{bmatrix},$$

whence we find

$$\sum_{j=1}^{d/2} \left(\int_{\mathcal{X}} \gamma(x) dP_{+j}(x) - dP_{-j}(x) \right)^2 = \frac{4\delta^2}{d^2} \gamma^\top \sum_{j=1}^{d/2} \begin{bmatrix} e_j \\ -e_j \end{bmatrix} \begin{bmatrix} e_j \\ -e_j \end{bmatrix}^\top \gamma = \frac{4\delta^2}{d^2} \gamma^\top \begin{bmatrix} I & -I \\ -I & I \end{bmatrix} \gamma \le \frac{8\delta^2}{d},$$

because the operator norm of the matrix is bounded by 2. This gives the claim (8.12).

Substituting the bound (8.12) into the bound (8.11) via Theorem 7.3, we obtain

$$\max_{v \in \mathcal{V}} \mathbb{E}_{P_v}[\|\widehat{\theta} - \theta_v\|_2^2] \ge \frac{\delta^2}{4d} \left[1 - \left(4n(e^\alpha - 1)^2 \delta^2 / d^2 \right)^{\frac{1}{2}} \right].$$

Choosing $\delta^2 = \min\{1, d^2/(16n(e^\alpha - 1)^2)\}$ gives the lower bound

$$\mathfrak{M}_n(\Delta_d, \|\cdot\|_2^2, \alpha) \ge \min \left\{ \frac{1}{4d}, \frac{d}{64n(e^\alpha - 1)^2} \right\}.$$

To complete the proof, we note that we can prove the preceding upper bound for any even $d_0 \in \{2, \ldots, d\}$; this requires choosing $v \in \mathcal{V} = \{-1,1\}^{d_0/2}$ and constructing the multinomial vectors

$$\theta_v = \frac{1}{d_0} \begin{bmatrix} \mathbb{1}_{d_0} \\ 0_{d-d_0} \end{bmatrix} + \frac{\delta}{d_0} \begin{bmatrix} v \\ -v \\ 0_{d-d_0} \end{bmatrix} \in \Delta_d, \quad \text{where } \; \mathbb{1}_{d_0} = [1 \; 1 \; \cdots \; 1]^\top \in \mathbb{R}^{d_0}.$$

Repeating the proof *mutatis mutandis* gives the bound

$$\mathfrak{M}_n(\Delta_d, \|\cdot\|_2^2, \alpha) \ge \max_{d_0 \in \{2,4,\ldots,2\lfloor d/2 \rfloor\}} \min \left\{ \frac{1}{4d_0}, \frac{d_0}{64n(e^\alpha - 1)^2} \right\}.$$

Choosing d_0 to be the even integer closest to $\sqrt{n\alpha^2}$ in $\{1, \ldots, d\}$ and noting that $(e^\alpha - 1)^2 \le 3\alpha^2$ for $\alpha \in [0,1]$ gives the claimed result (7.29).

In the case of measuring error in the ℓ_1-norm, we provide a completely identical proof, except that we have the separation $\|\widehat{\theta} - \theta_v\|_1 \ge (\delta/d) \sum_{j=1}^{d/2} \mathbf{1}\{\widehat{v}_j \ne v_j\}$, and thus inequality (8.11) holds with the initial multiplier $\delta^2/(4d)$ replaced by $\delta/(4d)$. Parallel reasoning to the ℓ_2^2 case then gives the minimax lower bound

$$\mathfrak{M}_n(\Delta_d, \|\cdot\|_1, \alpha) \ge \frac{\delta}{4d_0} \left[1 - (4n(e^\alpha - 1)^2 \delta^2 / d_0^2)^{\frac{1}{2}} \right]$$

for any even $d_0 \in \{2, \ldots, d\}$. Choosing $\delta = \min\{1, d_0^2/(16n(e^\alpha - 1)^2)\}$ gives the claim (7.30).

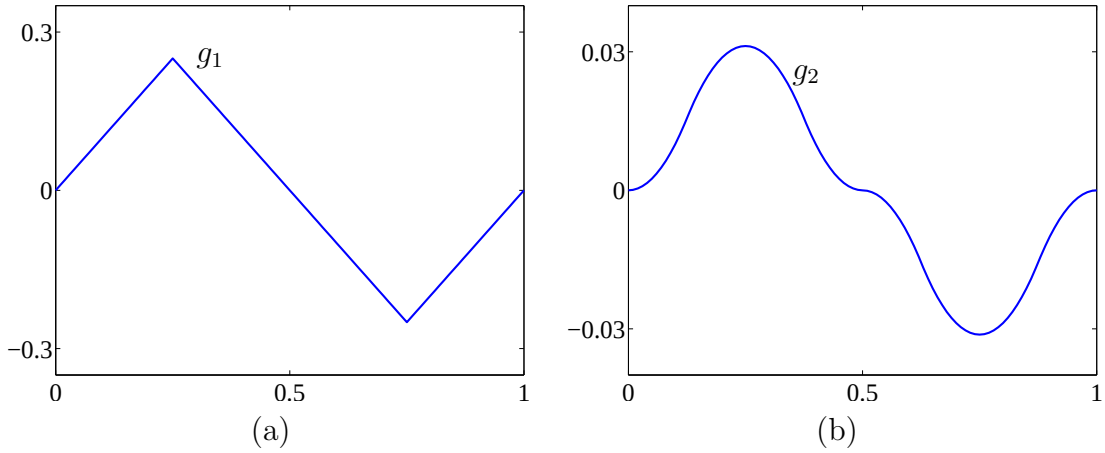

Figure 8.1. Panel (a): illustration of 1-Lipschitz continuous bump function g_1 used to pack $\mathcal{F}_\beta$ when $\beta = 1$. Panel (b): bump function g_2 with $|g_2''(x)| \leq 1$ used to pack $\mathcal{F}_\beta$ when $\beta = 2$.

8.6 Proofs of density estimation results

In this section, we provide the proofs of the results stated in Section 7.5.3 on density estimation. We defer the proofs of more technical results to later appendices. Throughout all proofs, we use c to denote a universal constant whose value may change from line to line.

8.6.1 Proof of Proposition 7.7

As with our proof for multinomial estimation, the argument follows the general outline described at the beginning of Section 8.5. We remark that our proof is based on an explicit construction of densities identified with corners of the hypercube, a more classical approach than the global metric entropy approach of Yang and Barron [185] (cf. [188]). We use the local packing approach since it is better suited to the privacy constraints and information contractions that we have developed. In comparison with our proofs of previous propositions, the construction of a suitable packing of $\mathcal{F}_\beta$ is somewhat more challenging: the identification of densities with finite-dimensional vectors, which we require for our application of Theorem 7.3, is not immediately obvious. In all cases, we guarantee that our density functions f belong to the trigonometric Sobolev space, so we may work directly with smooth density functions f.

Constructing well-separated densities We begin by describing a standard framework for defining local packings of density functions. Let $g_\beta : [0, 1] \to \mathbb{R}$ be a function satisfying the following properties:

(a) The function g_β is β-times differentiable with

$$0 = g_\beta^{(i)}(0) = g_\beta^{(i)}(1/2) = g_\beta^{(i)}(1) \text{ for all } i < \beta.$$

(b) The function g_β is centered with $\int_0^1 g_\beta(x)dx = 0$, and there exist constants $c, c_{1/2} > 0$ such that

$$\int_0^{1/2} g_\beta(x)dx = -\int_{1/2}^1 g_\beta(x)dx = c_{1/2} \quad \text{and} \quad \int_0^1 \left(g_\beta^{(i)}(x)\right)^2 dx \geq c \text{ for all } i < \beta.$$

(c) The function g_β is non-negative on $[0, 1/2]$ and non-positive on $[1/2, 1]$, and Lebesgue measure is absolutely continuous with respect to the measures $G_j, j = 1, 2$, given by

$$G_1(A) = \int_{A \cap [0,1/2]} g_\beta(x)dx \quad \text{and} \quad G_2(A) = -\int_{A \cap [1/2,1]} g_\beta(x)dx. \tag{8.13}$$

(d) Lastly, for almost every $x \in [0, 1]$, we have $|g_\beta^{(\beta)}(x)| \leq 1$ and $|g_\beta(x)| \leq 1$.

As illustrated in Figure 8.1, the functions g_β are smooth "bump" functions.

Fix a positive integer k (to be specified in the sequel). Our first step is to construct a family of "well-separated" densities for which we can reduce the density estimation problem to one of identifying corners of a hypercube, which allows application of Lemma 2.2. Specifically, we must exhibit a condition similar to the separation condition (2.17). For each $j \in \{1, \ldots, k$ define the function

$$g_{\beta,j}(x) := \frac{1}{k^\beta} g_\beta\left(k\Big(x - \frac{j-1}{k}\Big)\right) \mathbf{1}\left\{x \in \left[\tfrac{j-1}{k}, \tfrac{j}{k}\right]\right\}.$$

Based on this definition, we define the family of densities

$$\left\{f_v := 1 + \sum_{j=1}^k v_j g_{\beta,j} \quad \text{for } v \in \mathcal{V}\right\} \subseteq \mathcal{F}_\beta. \tag{8.14}$$

It is a standard fact [188, 173] that for any $v \in \mathcal{V}$, the function f_v is β-times differentiable, satisfies $|f^{(\beta)}(x)| \leq 1$ for all x. Now, based on some density $f \in \mathcal{F}_\beta$, let us define the sign vector $\mathsf{v}(f) \in \{-1, 1\}^k$ to have entries

$$\mathsf{v}_j(f) := \operatorname*{argmin}_{s \in \{-1,1\}} \int_{[\frac{j-1}{k}, \frac{j}{k}]} \left(f(x) - s g_{\beta,j}(x)\right)^2 dx.$$

Then by construction of the g_β and v, we have for a numerical constant c (whose value may depend on β) that

$$\|f - f_v\|_2^2 \geq c \sum_{j=1}^k \mathbf{1}\left\{\mathsf{v}_j(f) \neq v_j\right\} \int_{[\frac{j-1}{k}, \frac{j}{k}]} (g_{\beta,j}(x))^2 dx = \frac{c}{k^{2\beta+1}} \sum_{j=1}^k \mathbf{1}\left\{\mathsf{v}_j(f) \neq v_j\right\}.$$

By inspection, this is the Hamming separation required in inequality (2.17), whence the sharper version (7.28) of Assouad's Lemma 2.2 gives the result

$$\mathfrak{M}_n\left(\mathcal{F}_\beta[1], \|\cdot\|_2^2, \alpha\right) \geq \frac{c}{k^{2\beta}}\left[1 - \left(\frac{1}{4k}\sum_{j=1}^{k}(D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\rm kl}\left(M_{-j}^n \| M_{+j}^n\right))\right)^{\frac{1}{2}}\right], \quad (8.15)$$

where we have defined $P_{\pm j}$ to be the probability distribution associated with the averaged densities $f_{\pm j} = 2^{1-k}\sum_{v:v_j=\pm 1} f_v$.

Applying divergence inequalities Now we must control the summed KL-divergences. To do so, we note that by the construction (8.14), symmetry implies that

$$f_{+j} = 1 + g_{\beta,j} \quad \text{and} \quad f_{-j} = 1 - g_{\beta,j} \quad \text{for each} \;\; j \in [k]. \quad (8.16)$$

We then obtain the following result, which bounds the averaged KL-divergences.

Lemma 8.7. *For any α-locally private conditional distribution Q, the summed KL-divergences are bounded as*

$$\sum_{j=1}^{k}\left(D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\rm kl}\left(M_{+j}^n \| M_{-j}^n\right)\right) \leq 4c_{1/2}^2\, n\frac{(e^\alpha - 1)^2}{k^{2\beta+1}}.$$

The proof of this lemma is fairly involved, so we defer it to Section 8.7.3. We note that, for $\alpha \leq 1$, we have $(e^\alpha - 1)^2 \leq 3\alpha^2$, so we may replace the bound in Lemma 8.7 with the quantity $cn\alpha^2/k^{2\beta+1}$ for a constant c. We remark that standard divergence bounds using Assouad's lemma [188, 173] provide a bound of roughly $n/k^{2\beta}$; our bound is thus essentially a factor of the "dimension" k tighter.

The remainder of the proof is an application of inequality (8.15). In particular, by applying Lemma 8.7, we find that for any α-locally private channel Q, there are constants c_0, c_1 (whose values may depend on β) such that

$$\mathfrak{M}_n\left(\mathcal{F}_\beta, \|\cdot\|_2^2, Q\right) \geq \frac{c_0}{k^{2\beta}}\left[1 - \left(\frac{c_1 n\alpha^2}{k^{2\beta+2}}\right)^{\frac{1}{2}}\right].$$

Choosing $k_{n,\alpha,\beta} = (4c_1 n\alpha^2)^{\frac{1}{2\beta+2}}$ ensures that the quantity inside the parentheses is at least $1/2$. Substituting for k in the preceding display proves the proposition.

8.6.2 Proof of Proposition 7.8

Note that the operator Π_k performs a Euclidean projection of the vector $(k/n)\sum_{i=1}^n Z_i$ onto the scaled probability simplex, thus projecting $\widehat{f}$ onto the set of probability densities. Given the non-expansivity of Euclidean projection, this operation can only decrease the

error $\|\widehat{f} - f\|_2^2$. Consequently, it suffices to bound the error of the unprojected estimator; to reduce notational overhead we retain our previous notation of $\widehat{\theta}$ for the unprojected version. Using this notation, we have

$$\mathbb{E}\left[\left\|\widehat{f} - f\right\|_2^2\right] \leq \sum_{j=1}^{k} \mathbb{E}_f\left[\int_{\frac{j-1}{k}}^{\frac{j}{k}} (f(x) - \widehat{\theta}_j)^2 dx\right].$$

Expanding this expression and noting that the independent noise variables $W_{ij} \sim \text{Laplace}(\alpha/2)$ have zero mean, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f} - f\right\|_2^2\right] &\leq \sum_{j=1}^{k} \mathbb{E}_f\left[\int_{\frac{j-1}{k}}^{\frac{j}{k}} \left(f(x) - \frac{k}{n}\sum_{i=1}^{n}[\mathsf{e}_k(X_i)]_j\right)^2 dx\right] + \sum_{j=1}^{k}\int_{\frac{j-1}{k}}^{\frac{j}{k}} \mathbb{E}\left[\left(\frac{k}{n}\sum_{i=1}^{n} W_{ij}\right)^2\right] \\
&= \sum_{j=1}^{k}\int_{\frac{j-1}{k}}^{\frac{j}{k}} \mathbb{E}_f\left[\left(f(x) - \frac{k}{n}\sum_{i=1}^{n}[\mathsf{e}_k(X_i)]_j\right)^2\right] dx + k\,\frac{1}{k}\,\frac{4k^2}{n\alpha^2}. \qquad (8.17)
\end{aligned}$$

We now bound the error term inside the expectation (8.17). Defining the probabilities $p_j := \mathbb{P}_f(X \in \mathcal{X}_j) = \int_{\mathcal{X}_j} f(x)dx$, we have

$$k\mathbb{E}_f\left[[\mathsf{e}_k(X)]_j\right] = kp_j = k\int_{\mathcal{X}_j} f(x)dx \in \left[f(x) - \frac{1}{k}, f(x) + \frac{1}{k}\right] \quad \text{for any } x \in \mathcal{X}_j,$$

by the Lipschitz continuity of f. Thus, expanding the bias and variance of the integrated expectation above, we find that

$$\begin{aligned}
\mathbb{E}_f\left[\left(f(x) - \frac{k}{n}\sum_{i=1}^{n}[\mathsf{e}_k(X_i)]_j\right)^2\right] &\leq \frac{1}{k^2} + \text{Var}\left(\frac{k}{n}\sum_{i=1}^{n}\left[\mathsf{e}_k(X_i)\right]_j\right) \\
&= \frac{1}{k^2} + \frac{k^2}{n}\,\text{Var}([\mathsf{e}_k(X)]_j) = \frac{1}{k^2} + \frac{k^2}{n}p_j(1-p_j).
\end{aligned}$$

Recalling the inequality (8.17), we obtain

$$\mathbb{E}_f\left[\left\|\widehat{f} - f\right\|_2^2\right] \leq \sum_{j=1}^{k}\int_{\frac{j-1}{k}}^{\frac{j}{k}}\left(\frac{1}{k^2} + \frac{k^2}{n}p_j(1-p_j)\right)dx + \frac{4k^2}{n\alpha^2} = \frac{1}{k^2} + \frac{4k^2}{n\alpha^2} + \frac{k}{n}\sum_{j=1}^{k} p_j(1-p_j).$$

Since $\sum_{j=1}^{k} p_j = 1$, we find that

$$\mathbb{E}_f\left[\left\|\widehat{f} - f\right\|_2^2\right] \leq \frac{1}{k^2} + \frac{4k^2}{n\alpha^2} + \frac{k}{n},$$

and choosing $k = (n\alpha^2)^{\frac{1}{4}}$ yields the claim.

8.6.3 Proof of Proposition 7.9

We begin by fixing $k \in \mathbb{N}$; we will optimize the choice of k shortly. Recall that, since $f \in \mathcal{F}_\beta[C]$, we have $f = \sum_{j=1}^\infty \theta_j \varphi_j$ for $\theta_j = \int f \varphi_j$. Thus we may define $\overline{Z}_j = \frac{1}{n} \sum_{i=1}^n Z_{i,j}$ for each $j \in \{1, \ldots, k\}$, and we have

$$\|\widehat{f} - f\|_2^2 = \sum_{j=1}^k (\theta_j - \overline{Z}_j)^2 + \sum_{j=k+1}^\infty \theta_j^2.$$

Since $f \in \mathcal{F}_\beta[C]$, we are guaranteed that $\sum_{j=1}^\infty j^{2\beta} \theta_j^2 \leq C^2$, and hence

$$\sum_{j>k} \theta_j^2 = \sum_{j>k} j^{2\beta} \frac{\theta_j^2}{j^{2\beta}} \leq \frac{1}{k^{2\beta}} \sum_{j>k} j^{2\beta} \theta_j^2 \leq \frac{1}{k^{2\beta}} C^2.$$

For the indices $j \leq k$, we note that by assumption, $\mathbb{E}[Z_{i,j}] = \int \varphi_j f = \theta_j$, and since $|Z_{i,j}| \leq B$, we have

$$\mathbb{E}\left[(\theta_j - \overline{Z}_j)^2\right] = \frac{1}{n} \operatorname{Var}(Z_{1,j}) \leq \frac{B^2}{n} = \frac{B_0^2}{c_k} \frac{k}{n} \left(\frac{e^\alpha + 1}{e^\alpha - 1}\right)^2,$$

where $c_k = \Omega(1)$ is the constant in expression (7.39). Putting together the pieces, the mean-squared L^2-error is upper bounded as

$$\mathbb{E}_f \left[\|\widehat{f} - f\|_2^2\right] \leq c \left(\frac{k^2}{n\alpha^2} + \frac{1}{k^{2\beta}}\right),$$

where c is a constant depending on B_0, c_k, and C. Choose $k = (n\alpha^2)^{1/(2\beta+2)}$ to complete the proof.

8.6.4 Insufficiency of Laplace noise for density estimation

Finally, we consider the insufficiency of standard Laplace noise addition for estimation in the setting of this section. Consider the vector $[\varphi_j(X_i)]_{j=1}^k \in [-B_0, B_0]^k$. To make this vector α-differentially private by adding an independent Laplace noise vector $W \in \mathbb{R}^k$, we must take $W_j \sim \text{Laplace}(\alpha/(B_0 k))$. The natural orthogonal series estimator [e.g., 180] is to take $Z_i = [\varphi_j(X_i)]_{j=1}^k + W_i$, where $W_i \in \mathbb{R}^k$ are independent Laplace noise vectors. We then use the density estimator (7.40), except that we use the Laplacian perturbed Z_i. However, this estimator suffers the following drawback:

Observation 8.1. *Let $\widehat{f} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k Z_{i,j} \varphi_j$, where the Z_i are the Laplace-perturbed vectors of the previous paragraph. Assume the orthonormal basis $\{\varphi_j\}$ of $L^2([0,1])$ contains the constant function. There is a constant c such that for any $k \in \mathbb{N}$, there is an $f \in \mathcal{F}_\beta[2]$ such that*

$$\mathbb{E}_f \left[\|f - \widehat{f}\|_2^2\right] \geq c(n\alpha^2)^{-\frac{2\beta}{2\beta+3}}.$$

Proof We begin by noting that for $f = \sum_j \theta_j \varphi_j$, by definition of $\widehat{f} = \sum_j \widehat{\theta}_j \varphi_j$ we have

$$\mathbb{E}\left[\|f - \widehat{f}\|_2^2\right] = \sum_{j=1}^{k} \mathbb{E}\left[(\theta_j - \widehat{\theta}_j)^2\right] + \sum_{j \geq k+1} \theta_j^2 = \sum_{j=1}^{k} \frac{B_0^2 k^2}{n\alpha^2} + \sum_{j \geq k+1} \theta_j^2 = \frac{B_0^2 k^3}{n\alpha^2} + \sum_{j \geq k+1} \theta_j^2.$$

Without loss of generality, let us assume $\varphi_1 = 1$ is the constant function. Then $\int \varphi_j = 0$ for all $j > 1$, and by defining the true function $f = \varphi_1 + (k+1)^{-\beta} \varphi_{k+1}$, we have $f \in \mathcal{F}_\beta[2]$ and $\int f = 1$, and moreover,

$$\mathbb{E}\left[\|f - \widehat{f}\|_2^2\right] \geq \frac{B_0^2 k^3}{n\alpha^2} + \frac{1}{(k+1)^{-2\beta}} \geq C_{\beta, B_0} (n\alpha^2)^{-\frac{2\beta}{2\beta+3}},$$

where C_{β, B_0} is a constant depending on β and B_0. This final lower bound comes by minimizing over all k. (If $(k+1)^{-\beta} B_0 > 1$, we can rescale φ_{k+1} by B_0 to achieve the same result and guarantee that $f \geq 0$.) $\square$

This lower bound shows that standard estimators based on adding Laplace noise to appropriate basis expansions of the data fail: there is a degradation in rate from $n^{-\frac{2\beta}{2\beta+2}}$ to $n^{-\frac{2\beta}{2\beta+3}}$. While this is not a formal proof that no approach based on Laplace perturbation can provide optimal convergence rates in our setting, it does suggest that finding such an estimator is non-trivial.

8.7 Information bounds

In this section, we collect the proofs of lemmas providing mutual information and KL-divergence bounds.

8.7.1 Proof of Lemma 8.4

Our strategy is to apply Theorem 7.2 to bound the mutual information. Without loss of generality, we may assume that $\mathsf{r} = 1$ so the set $\mathcal{X} = \{\pm e_j\}_{j=1}^k$, where $e_j \in \mathbb{R}^d$. Thus, under the notation of Theorem 7.2, we may identify vectors $\gamma \in L^\infty(\mathcal{X})$ by vectors $\gamma \in \mathbb{R}^{2k}$. Noting that $\overline{v} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} v = 0$ is the mean element of the "packing" set by our construction, the linear functional φ_v defined in Theorem 7.2 is

$$\varphi_v(\gamma) = \frac{1}{2k} \sum_{j=1}^{k} \left[\frac{\delta}{2}\gamma(e_j) v_j - \frac{\delta}{2}\gamma(-e_j) v_j\right] = \frac{\delta}{4k} \gamma^\top \begin{bmatrix} I_{k \times k} & 0_{k \times d-k} \\ -I_{k \times k} & 0_{k \times d-k} \end{bmatrix} v.$$

Define the matrix

$$A := \begin{bmatrix} I_{k \times k} & 0_{k \times d-k} \\ -I_{k \times k} & 0_{k \times d-k} \end{bmatrix} \in \{-1, 0, 1\}^{2k \times d}.$$

Then we have that

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\varphi_v(\gamma)^2 = \frac{\delta^2}{(4k)^2}\gamma^\top A\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}vv^\top A^\top\gamma = \frac{\delta^2}{(4k)^2}\gamma^\top A\operatorname{Cov}(V)A^\top\gamma$$
$$= \frac{\delta^2}{(4k)^2}\gamma^\top AA^\top\gamma = \left(\frac{\delta}{4k}\right)^2\gamma^\top\begin{bmatrix} I_{k\times k} & -I_{k\times k} \\ -I_{k\times k} & I_{k\times k}\end{bmatrix}\gamma. \tag{8.18}$$

Here we have used that $A\operatorname{Cov}(V)A^\top = AI_{d\times d}A^\top$ by the fact that $\mathcal{V} = \{-1,1\}^k\times\{0\}^{d-k}$.

We complete our proof using the bound (8.18). The operator norm of the matrix specified in (8.18) is 2. As a consequence, since we have the containment

$$\mathcal{B}_\infty = \left\{\gamma\in\mathbb{R}^{2k} : \|\gamma\|_\infty \le 1\right\} \subset \left\{\gamma\in\mathbb{R}^{2k} : \|\gamma\|_2^2 \le 2k\right\}$$

we have the inequality

$$\sup_{\gamma\in\mathcal{B}_\infty}\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\varphi_v(\gamma)^2 \le \frac{\delta^2}{16k^2}\cdot 2\cdot 2k = \frac{1}{4}\frac{\delta^2}{k}.$$

Applying Theorem 7.2 completes the proof.

8.7.2 Proof of Lemma 8.5

It is no loss of generality to assume the radius $\mathsf{r} = 1$. We use the notation of Theorem 7.2, recalling the linear functionals $\varphi_v : L^\infty(\mathcal{X})\to\mathbb{R}$. Because the set $\mathcal{X} = \{-1,1\}^d$, we can identify vectors $\gamma\in L^\infty(\mathcal{X})$ with vectors $\gamma\in\mathbb{R}^{2^d}$. Moreover, we have (by construction) that

$$\varphi_v(\gamma) = \sum_{x\in\{-1,1\}^d}\gamma(x)p_v(x) - \sum_{x\in\{-1,1\}^d}\gamma(x)\overline{p}(x)$$
$$= \frac{1}{2^d}\sum_{x\in\mathcal{X}}\gamma(x)(1+\delta v^\top x - 1) = \frac{\delta}{2^d}\sum_{x\in\mathcal{X}}\gamma(x)v^\top x.$$

For each $v\in\mathcal{V}$, we may construct a vector $u_v\in\{-1,1\}^{2^d}$, indexed by $x\in\{-1,1\}^d$, with

$$u_v(x) = v^\top x = \begin{cases} 1 & \text{if } v = \pm e_j \text{ and } \operatorname{sign}(v_j) = \operatorname{sign}(x_j) \\ -1 & \text{if } v = \pm e_j \text{ and } \operatorname{sign}(v_j)\neq\operatorname{sign}(x_j).\end{cases}$$

For $v = e_j$, we see that $u_{e_1},\ldots,u_{e_d}$ are the first d columns of the standard Hadamard transform matrix (and u_{-e_j} are their negatives). Then we have that $\sum_{x\in\mathcal{X}}\gamma(x)v^\top x = \gamma^\top u_v$, and

$$\varphi_v(\gamma) = \gamma^\top u_v u_v^\top\gamma.$$

Note also that $u_vu_v^\top = u_{-v}u_{-v}^\top$, and as a consequence we have

$$\sum_{v\in\mathcal{V}}\varphi_v(\gamma)^2 = \frac{\delta^2}{4^d}\gamma^\top\sum_{v\in\mathcal{V}}u_vu_v^\top\gamma = \frac{2\delta^2}{4^d}\gamma^\top\sum_{j=1}^d u_{e_j}u_{e_j}^\top\gamma. \tag{8.19}$$

But now, studying the quadratic form (8.19), we note that the vectors u_{e_j} are orthogonal. As a consequence, the vectors (up to scaling) u_{e_j} are the only eigenvectors corresponding to positive eigenvalues of the positive semidefinite matrix $\sum_{j=1}^{d} u_{e_j} u_{e_j}^\top$. Thus, since the set

$$\mathcal{B}_\infty = \left\{\gamma \in \mathbb{R}^{2^d} : \|\gamma\|_\infty \leq 1\right\} \subset \left\{\gamma \in \mathbb{R}^{2^d} : \|\gamma\|_2^2 \leq 2^d\right\},$$

we have via an eigenvalue calculation that

$$\begin{aligned}\sup_{\gamma \in \mathcal{B}_\infty} \sum_{v \in \mathcal{V}} \varphi_v(\gamma)^2 &\leq \frac{2\delta^2}{4^d} \sup_{\gamma : \|\gamma\|_2^2 \leq 2^d} \gamma^\top \sum_{j=1}^{d} u_{e_j} u_{e_j}^\top \gamma \\ &= \frac{2\delta^2}{4^d} \|u_{e_1}\|_2^4 = 2\delta^2\end{aligned}$$

since $\|u_{e_j}\|_2^2 = 2^d$ for each j. Applying Theorem 7.2 and Corollary 7.4 completes the proof.

8.7.3 Proof of Lemma 8.7

This result relies on Theorem 7.3, along with a careful argument to understand the extreme points of $\gamma \in L^\infty([0,1])$ that we use when applying the result. First, we take the packing $\mathcal{V} = \{-1, 1\}^\beta$ and densities f_v for $v \in \mathcal{V}$ as in the construction (8.14). Overall, our first step is to show for the purposes of applying Theorem 7.3, it is no loss of generality to identify $\gamma \in L^\infty([0,1])$ with vectors $\gamma \in \mathbb{R}^{2k}$, where γ is constant on intervals of the form $[i/2k, (i+1)/2k]$. With this identification complete, we can then provide a bound on the correlation of any $\gamma \in \mathcal{B}_\infty$ with the densities $f_{\pm j}$ defined in (8.16), which completes the proof.

With this outline in mind, let the sets D_i, $i \in \{1, 2, \ldots, 2k\}$, be defined as $D_i = [(i-1)/2k, i/2k)$ except that $D_{2k} = [(2k-1)/2k, 1]$, so the collection $\{D_i\}_{i=1}^{2k}$ forms a partition of the unit interval $[0,1]$. By construction of the densities f_v, the sign of $f_v - 1$ remains constant on each D_i. Let us define (for shorthand) the linear functionals $\varphi_j : L^\infty([0,1]) \to \mathbb{R}$ for each $j \in \{1, \ldots, k\}$ via

$$\varphi_j(\gamma) := \int \gamma(dP_{+j} - dP_{-j}) = \sum_{i=1}^{2k} \int_{D_i} \gamma(x)(f_{+j}(x) - f_{-j}(x))dx = 2 \int_{D_{2j-1} \cup D_{2j}} \gamma(x) g_{\beta,j}(x) dx,$$

where we recall the definitions (8.16) of the mixture densities $f_{\pm j} = 1 \pm g_{\beta,j}$. Since the set $\mathcal{B}_\infty$ from Theorem 7.3 is compact, convex, and Hausdorff, the Krein-Milman theorem [143, Proposition 1.2] guarantees that it is equal to the convex hull of its extreme points; moreover, since the functionals $\gamma \mapsto \varphi_j^2(\gamma)$ are convex, the supremum in Theorem 7.3 must be attained at the extreme points of $\mathcal{B}_\infty([0,1])$. As a consequence, when applying the divergence bound

$$\sum_{j=1}^{k} \left(D_{\mathrm{kl}}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\mathrm{kl}}\left(M_{-j}^n \| M_{+j}^n\right)\right) \leq 2n(e^\alpha - 1)^2 \sup_{\gamma \in \mathcal{B}_\infty} \sum_{j=1}^{k} \varphi_j^2(\gamma), \tag{8.20}$$

we can restrict our attention to $\gamma \in \mathcal{B}_\infty$ for which $\gamma(x) \in \{-1, 1\}$.

Now we argue that it is no loss of generality to assume that γ, when restricted to D_i, is a constant (apart from a measure zero set). Fix $i \in [2k]$, and assume for the sake of contradiction that there exist sets $B_i, C_i \subset D_i$ such that $\gamma(B_i) = \{1\}$ and $\gamma(C_i) = \{-1\}$, while $\lambda(B_i) > 0$ and $\lambda(C_i) > 0$ where λ denotes Lebesgue measure.[1] We will construct vectors γ_1 and $\gamma_2 \in \mathcal{B}_\infty$ and a value $\lambda \in (0, 1)$ such that

$$\int_{D_i} \gamma(x) g_{\beta,j}(x) dx = \lambda \int_{D_i} \gamma_1(x) g_{\beta,j}(x) dx + (1 - \lambda) \int_{D_i} \gamma_2(x) g_{\beta,j}(x) dx$$

simultaneously for all $j \in [k]$, while on $D_i^c = [0, 1] \setminus D_i$, we will have the equivalence

$$\gamma_1|_{D_i^c} \equiv \gamma_2|_{D_i^c} \equiv \gamma|_{D_i^c}.$$

Indeed, set $\gamma_1(D_i) = \{1\}$ and $\gamma_2(D_i) = \{-1\}$, otherwise setting $\gamma_1(x) = \gamma_2(x) = \gamma(x)$ for $x \notin D_i$. For the unique index $j \in [k]$ such that $[(j-1)/k, j/k] \supset D_i$, we define

$$\lambda := \frac{\int_{B_i} g_{\beta,j}(x) dx}{\int_{D_i} g_{\beta,j}(x) dx} \quad \text{so} \quad 1 - \lambda = \frac{\int_{C_i} g_{\beta,j}(x) dx}{\int_{D_i} g_{\beta,j}(x) dx}.$$

By the construction of the function g_β, the functions $g_{\beta,j}$ do not change signs on D_i, and the absolute continuity conditions on g_β specified in equation (8.13) guarantee $1 > \lambda > 0$, since $\lambda(B_i) > 0$ and $\lambda(C_i) > 0$. We thus find that for any $j \in [k]$,

$$\begin{aligned}
\int_{D_i} \gamma(x) g_{\beta,j}(x) dx &= \int_{B_i} \gamma_1(x) g_{\beta,j}(x) dx + \int_{C_i} \gamma_2(x) g_{\beta,j}(x) dx \\
&= \int_{B_i} g_{\beta,j}(x) dx - \int_{C_i} g_{\beta,j}(x) dx = \lambda \int_{D_i} g_{\beta,j}(x) dx - (1 - \lambda) \int_{D_i} g_{\beta,j}(x) dx \\
&= \lambda \int \gamma_1(x) g_{\beta,j}(x) dx + (1 - \lambda) \int \gamma_2(x) g_{\beta,j}(x) dx.
\end{aligned}$$

(Notably, for j such that $g_{\beta,j}$ is identically 0 on D_i, this equality is trivial.) By linearity and the strong convexity of the function $x \mapsto x^2$, then, we find that for sets $E_j := D_{2j-1} \cup D_{2j}$,

$$\begin{aligned}
\sum_{j=1}^k \varphi_j^2(\gamma) &= \sum_{j=1}^k \left(\int_{E_j} \gamma(x) g_{\beta,j}(x) dx \right)^2 \\
&< \lambda \sum_{j=1}^k \left(\int_{E_j} \gamma_1(x) g_{\beta,j}(x) dx \right)^2 + (1 - \lambda) \sum_{v \in \mathcal{V}} \left(\int_{E_j} \gamma_2(x) g_{\beta,j}(x) dx \right)^2.
\end{aligned}$$

Thus one of the densities γ_1 or γ_2 must have a larger objective value than γ. This is our desired contradiction, which shows that (up to measure zero sets) any γ attaining the supremum in the information bound (8.20) must be constant on each of the D_i.

[1] For a function f and set A, the notation $f(A)$ denotes the image $f(A) = \{f(x) \mid x \in A\}$.

Having shown that γ is constant on each of the intervals D_i, we conclude that the supremum (8.20) can be reduced to a finite-dimensional problem over the subset

$$\mathcal{B}_{1,2k} := \left\{ u \in \mathbb{R}^{2k} \mid \|u\|_\infty \leq 1 \right\}$$

of $\mathbb{R}^{2k}$. In terms of this subset, the supremum (8.20) can be rewritten as the the upper bound

$$\sup_{\gamma \in \mathcal{B}_\infty} \sum_{j=1}^{k} \varphi_j(\gamma)^2 \leq \sup_{\gamma \in \mathcal{B}_{1,2k}} \sum_{j=1}^{k} \left(\gamma_{2j-1} \int_{D_{2j-1}} g_{\beta,j}(x)dx + \gamma_{2j} \int_{D_{2j}} g_{\beta,j}(x)dx \right)^2$$

By construction of the function g_β, we have the equality

$$\int_{D_{2j-1}} g_{\beta,j}(x)dx = -\int_{D_{2j}} g_{\beta,j}(x)dx = \int_0^{\frac{1}{2k}} g_{\beta,1}(x)dx = \int_0^{\frac{1}{2k}} \frac{1}{k^\beta} g_\beta(kx)dx = \frac{c_{1/2}}{k^{\beta+1}}.$$

This implies that

$$\begin{aligned} &\frac{1}{2e^\alpha(e^\alpha-1)^2 n} \sum_{j=1}^{k} \left(D_{\mathrm{kl}}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\mathrm{kl}}\left(M_{+j}^n \| M_{-j}^n\right)\right) \leq \sup_{\gamma \in \mathcal{B}_\infty} \sum_{j=1}^{k} \varphi_j(\gamma)^2 \\ &\leq \sup_{\gamma \in \mathcal{B}_{1,2k}} \sum_{j=1}^{k} \left(\frac{c_{1/2}}{k^{\beta+1}} \gamma^\top (e_{2j-1} - e_{2j}) \right)^2 = \frac{c_{1/2}^2}{k^{2\beta+2}} \sup_{\gamma \in \mathcal{B}_{1,2k}} \gamma^\top \sum_{j=1}^{k} (e_{2j-1} - e_{2j})(e_{2j-1} - e_{2j})^\top \gamma, \end{aligned} \tag{8.21}$$

where $e_j \in \mathbb{R}^{2k}$ denotes the jth standard basis vector. Rewriting this using the Kronecker product $\otimes$, we have

$$\sum_{j=1}^{k} (e_{2j-1} - e_{2j})(e_{2j-1} - e_{2j})^\top = I_{k\times k} \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \preceq 2 I_{2k \times 2k}.$$

Combining this bound with our inequality (8.21), we obtain

$$\sum_{j=1}^{k} \left(D_{\mathrm{kl}}\left(M_{+j}^n \| M_{-j}^n\right) + D_{\mathrm{kl}}\left(M_{+j}^n \| M_{-j}^n\right)\right) \leq 4n(e^\alpha - 1)^2 \frac{c_{1/2}^2}{k^{2\beta+2}} \sup_{\gamma \in \mathcal{B}_{1,2k}} \|\gamma\|_2^2 = 4c_{1/2}^2 \frac{n(e^\alpha-1)^2}{k^{2\beta+1}}.$$

Part IV

Communication

Chapter 9

Communication efficient algorithms

The broad question in this part of the thesis is the extent to which it is possible to avoid communication in solving distributed estimation problems; such problems arise in settings involving large-scale data sets. As we show in this chapter, for suitable (classical) statistical problems, it is possible to have extremely low communication: for d-dimensional problems distributed across m processors, it is possible for each of the m processors to communicate only a single (quantized) vector in $\mathbb{R}^d$.

In this chapter, which is based off of a paper by Zhang, Duchi, and Wainwright [189], we present three communication-efficient procedures for distributed statistical optimization. These set the stage for several optimality guarantees we give in Chapter 10 for procedures constrained to communicate small numbers of bits; the current chapter shows that our coming lower bounds are in fact sharp. This chapter's purpose is mainly illustrative, motivating Chapter 10, so we present a proof only of the first major theorem, as the proof is gentler than previous arguments [189] while suggesting the main techniques. We refer otherwise to the paper [189], as we simply wish to give intuition for the potential successes (and drawbacks) of low-communication algorithms.

The basic problem in the chapter is the following: we have N observations distributed across m machines, and we wish to construct estimates as statistically efficient as those with access to a full sample of size N. The first algorithm is a standard averaging method that distributes the N data observations evenly to m machines, performs separate minimization on each subset, and then averages the estimates. We provide a sharp analysis of this average mixture algorithm, showing that under a reasonable set of conditions, the combined parameter is asymptotically normal with variance decreasing as $\mathcal{O}_P(N^{-1} + (N/m)^{-2})$. Whenever $m \ll \sqrt{N}$, this guarantee matches the best possible rate achievable by a centralized algorithm having access to all N samples; indeed, the estimator is even locally asymptotically minimax [115, 116].

In addition, we also review a novel method due to Zhang et al. [189] based on an appropriate form of bootstrap subsampling, known as the *subsampled average mixture* (Savgm) algorithm. Requiring only a single round of communication, it has mean-squared error that decays as $\mathcal{O}(N^{-1} + (N/m)^{-3})$, and so is more robust to the amount of parallelization. We

also describe a stochastic gradient-based method that attains mean-squared error decaying as $\mathcal{O}(N^{-1} + (N/m)^{-3/2})$, easing computation at the expense of a potentially slower mean-squared-error (MSE) rate.

As this chapter is meant mostly as motivation to show how low-communication schemes may be effective, we omit experimental results complementing these theoretical results, referring to Zhang et al.'s paper [189]. We note in passing, however, that our paper [189] investigates the performance of these methods both on simulated data and on a large-scale regression problem from the internet search domain. In particular, we show that the methods can be used to efficiently solve an advertisement prediction problem from the Chinese SoSo Search Engine, which involves logistic regression with $N \approx 2.4 \times 10^8$ samples and $d \approx 740{,}000$ covariates, and moreover, the experiments show how the SAVGM can offer improved performance over more naive approaches.

9.1 Introduction

Many procedures for statistical estimation are based on a form of (regularized) empirical risk minimization, meaning that a parameter of interest is estimated by minimizing an objective function defined by the average of a loss function over the data. Given the current explosion in the size and amount of data available in statistical studies, a central challenge is to design efficient algorithms for solving large-scale problem instances. In a centralized setting, there are many procedures for solving empirical risk minimization problems, among them standard convex programming approaches [e.g. 32] as well as stochastic approximation and optimization algorithms [150, 96, 135]. When the size of the dataset becomes extremely large, however, it may be infeasible to store all of the data on a single computer, or at least to keep the data in memory. Accordingly, the focus of this chapter is the study of some distributed and communication-efficient procedures for empirical risk minimization.

Recent years have witnessed a flurry of research on distributed approaches to solving very large-scale statistical optimization problems. Although we cannot survey the literature adequately—the papers Nedić and Ozdaglar [131], Ram et al. [147], Johansson et al. [103], Duchi et al. [54], Dekel et al. [49], Agarwal and Duchi [3], Niu et al. [141] and references therein contain a sample of relevant work—we touch on a few important themes here. It can be difficult within a purely optimization-theoretic setting to show explicit benefits arising from distributed computation. In statistical settings, however, distributed computation can lead to gains in computational efficiency, as shown by a number of authors [3, 49, 141, 55]. Within the family of distributed algorithms, there can be significant differences in communication complexity: different computers must be synchronized, and when the dimensionality of the data is high, communication can be prohibitively expensive. It is thus interesting to study distributed estimation algorithms that require fairly limited synchronization and communication while still enjoying the greater statistical accuracy that is usually associated with a larger dataset.

With this context, perhaps the simplest algorithm for distributed statistical estimation

is what we term the *average mixture* (AVGM) algorithm. It is an appealingly simple method: given m different machines and a dataset of size N, first assign to each machine a (distinct) dataset of size $n = N/m$, then have each machine i compute the empirical minimizer θ_i on its fraction of the data, and finally average all the parameter estimates θ_i across the machines. This approach has been studied for some classification and estimation problems by Mann et al. [125] and McDonald, Hall, and Mann [128], as well as for certain stochastic approximation methods by Zinkevich et al. [191]. Given an empirical risk minimization algorithm that works on one machine, the procedure is straightforward to implement and is extremely communication efficient, requiring only a single round of communication. It is also relatively robust to possible failures in a subset of machines and/or differences in speeds, since there is no repeated synchronization. When the local estimators are all unbiased, it is clear that the the AVGM procedure will yield an estimate that is essentially as good as that of an estimator based on all N samples. Many estimators used in practice are biased, so it is natural to ask whether the method has any guarantees in a more general setting.

This chapter reviews several natural one-shot—requiring one round of communication—distributed algorithms. First, in Section 9.3, we provide a sharp analysis of the AVGM algorithm, showing that under a reasonable set of conditions on the population risk, the AVGM procedure is asymptotically normal with optimal covariance, with additional error terms scaling as $\mathcal{O}_P(\sqrt{m/n})$. Whenever the number of machines m is less than the number of samples n per machine, this guarantee matches the best possible rate achievable by a centralized algorithm having access to all $N = nm$ observations. We also present results showing that the result extends to stochastic programming approaches, exhibiting a stochastic gradient-descent based procedure that also attains convergence rates scaling as $\mathcal{O}((nm)^{-1})$, but with slightly worse dependence on different problem-specific parameters.

We also study a novel extension of simple averaging based on an appropriate form of resampling [71, 87, 145], which we refer to as the *subsampled average mixture* (SAVGM) approach. At a high level, the SAVGM algorithm distributes samples evenly among m processors or computers as before, but instead of simply returning the empirical minimizer, each processor further subsamples its own dataset in order to estimate its estimate's bias and returns a subsample-corrected estimate. Under appropriate conditions, which we provide, the SAVGM algorithm has mean-squared error decaying as $\mathcal{O}(m^{-1}n^{-1} + n^{-3})$. As long as $m < n^2$, the subsampled method again matches the centralized gold standard in the first-order term, and has a second-order term smaller than the standard averaging approach.

9.2 Background and Problem Set-up

We begin by establishing our framework for risk minimization, which closely follows that studied in Part II of this thesis on stochastic optimization problems. After this, we describe our algorithms and provide a few assumptions we require for our main theoretical results.

Empirical Risk Minimization

Let $\{\ell(\cdot; x),\ x \in \mathcal{X}\}$ be a collection of real-valued and convex loss functions, each defined on a set containing the convex set $\Theta \subseteq \mathbb{R}^d$. Let P be a probability distribution over the sample space $\mathcal{X}$. Assuming that each function $x \mapsto \ell(\theta; x)$ is P-integrable, the *population risk* $R : \Theta \to \mathbb{R}$ is given by the standard formula (2.2), $R(\theta) := \mathbb{E}_P[\ell(\theta; X)] = \int_{\mathcal{X}} \ell(\theta; x) dP(x)$. Our goal is to estimate the parameter vector minimizing the population risk,

$$\theta^* := \operatorname*{argmin}_{\theta \in \Theta} R(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \int_{\mathcal{X}} \ell(\theta; x) dP(x),$$

which we assume to be unique. In practice, the population distribution P is unknown to us, but we have access to a collection $S = \{x_1, \ldots, x_N\}$ of observations from the distribution P. Given centralized access to the entire sample S, a natural procedure is empirical risk minimization [118, 175, 176] which estimates θ^* by solving the optimization problem

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; x_i) \right\}. \tag{9.1}$$

Averaging Methods

Consider a data set consisting of $N = mn$ observations, drawn i.i.d. according to the distribution P. In the distributed setting, we divide this N-observation data set evenly and uniformly at random among a total of m processors. (For simplicity, we have assumed the total number of observations is a multiple of m.) For $i = 1, \ldots, m$, we let $S_{1,i}$ denote the data set assigned to processor i; by construction, it is a collection of n observations drawn i.i.d. according to P, and the observations in subsets $S_{1,i}$ and $S_{1,j}$ are independent for $i \neq j$. In addition, for each processor i we define the (local) empirical objective $\widehat{R}_{1,i}$ via $\widehat{R}_{1,i}(\theta) := \frac{1}{|S_{1,i}|} \sum_{x \in S_{1,i}} \ell(\theta; x)$. With this notation, the AVGM algorithm is simple to describe.

Average mixture algorithm:

(1) For each $i \in \{1, \ldots, m\}$, processor i uses its local dataset $S_{1,i}$ to compute a local empirical minimizer

$$\theta_{1,i} \in \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_{1,i}(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \frac{1}{|S_{1,i}|} \sum_{x \in S_{1,i}} \ell(\theta; x) \right\}. \tag{9.2}$$

(2) These m local estimates are then averaged together—that is, we compute

$$\overline{\theta}_{\text{AVGM}} = \frac{1}{m} \sum_{i=1}^{m} \theta_{1,i}. \tag{9.3}$$

The subsampled average mixture (SAVGM) algorithm is based on an additional level of sampling on top of the first, involving a fixed subsampling rate $r \in [0, 1]$. It consists of the following additional steps:

Subsampled average mixture algorithm:

(1) Each processor i draws a subset $S_{2,i}$ of size $\lceil rn \rceil$ by sampling uniformly at random without replacement from its local data set $S_{1,i}$.

(2) Each processor i computes both the local empirical minimizers $\theta_{1,i}$ from equation (9.2) and the empirical minimizer

$$\theta_{2,i} \in \operatorname*{argmin}_{\theta \in \Theta} \Big\{ \frac{1}{|S_{2,i}|} \sum_{x \in S_{2,i}} \ell(\theta; x) \Big\}.$$

(3) In addition to the previous average (9.3), the SAVGM algorithm computes the bootstrap average $\overline{\theta}_2 := \frac{1}{m} \sum_{i=1}^m \theta_{2,i}$, and then returns the weighted combination

$$\overline{\theta}_{\text{SAVGM}} := \frac{\overline{\theta}_{\text{AVGM}} - r\overline{\theta}_2}{1 - r}. \tag{9.4}$$

The intuition for the weighted estimator (9.4) is similar to that for standard bias correction procedures using the bootstrap or subsampling [71, 87, 145]. Roughly speaking, if $b_0 = \theta^* - \overline{\theta}_{\text{AVGM}}$ is the bias of the first estimator, then we may approximate b_0 by the subsampled estimate of bias $b_1 = \theta^* - \overline{\theta}_2$. But because $b_1 \approx b_0/r$, it is possible to argue that $\theta^* \approx (\overline{\theta}_{\text{AVGM}} - r\overline{\theta}_2)/(1 - r)$. The re-normalization enforces that the relative "weights" of $\overline{\theta}_{\text{AVGM}}$ and $\overline{\theta}_2$ sum to 1.

Our goal is to understand under what conditions—and in what sense—the estimators (9.3) and (9.4) approach the *oracle performance*, by which we mean the error of a centralized risk minimization procedure that is given access to all $N = nm$ observations. When is it possible to achieve the performance of the empirical risk minimizer (9.1)?

9.3 Theoretical Results

Having described the AVGM and SAVGM algorithms, we now turn to statements of our main theorems on their statistical properties, along with some consequences and comparison to past work.

9.3.1 Smoothness Conditions

Throughout the paper, we impose some regularity conditions on the parameter space, the risk function R, and the instantaneous loss functions $\ell(\cdot; x) : \Theta \to \mathbb{R}$. These conditions are standard in classical statistical analysis of M-estimators [e.g. 118, 109]; our first set of assumptions is the weakest and is required for all the results, while subsequent assumptions appear to be necessary only for stronger theoretical guarantees. Throughout, without further notice, we assume that the parameter space $\Theta \subset \mathbb{R}^d$ is convex, and we also require that $\theta^* \in \operatorname{int} \Theta$. In addition, the risk function is required to have some amount of curvature. We formalize this notion in terms of the Hessian of the risk R:

Assumption 9A (Local strong convexity)**.** *The population risk is twice differentiable, and there exists a parameter $\lambda > 0$ such that $\nabla^2 R(\theta^*) \succeq \lambda I_{d\times d}$.*

This local condition is milder than a global strong convexity condition and is required to hold only for the population risk R evaluated at θ^*. Of course, some type of curvature of the risk is required for any method to consistently estimate the parameters θ^*.

In addition, we require regularity conditions on the empirical risk functions. It is simplest to state these in terms of the functions $\theta \mapsto \ell(\theta; x)$; it is possible to obtain convergence guarantees for AVGM and SAVGM while requiring this assumption to hold only locally around the optimal point θ^*, but we opt for simpler statements.

Assumption 9B (Smoothness)**.** *For any $x \in \mathcal{X}$, the function $\theta \mapsto \ell(x; \theta)$ has $H(x)$-Lipschitz continuous Hessian with respect to the operator norm on matrices, meaning that*

$$\left|\left|\left|\nabla^2 \ell(\theta'; x) - \nabla^2 \ell(\theta; x)\right|\right|\right|_2 \leq H(x) \left\|\theta' - \theta\right\|_2, \tag{9.5}$$

where $\mathbb{E}[H(X)^2] \leq H^2$. Around θ^*, the gradients have finite variance: there exists a constant $M < \infty$ such that $\mathbb{E}[\|\nabla \ell(\theta^*; X)\|_2^2] \leq M^2$.

It important to note that some type of smoothness condition on the Hessian matrix, as in the Lipschitz condition (9.5), is *essential* in order for simple averaging methods to work. This necessity is illustrated by the following example:

Example 9.1 (Necessity of Hessian conditions)*.* Let X be a Bernoulli variable with parameter $\frac{1}{2}$, and consider the loss function

$$\ell(\theta; x) = \begin{cases} \theta^2 - \theta & \text{if } x = 0 \\ \theta^2 \mathbf{1}\left\{\theta \leq 0\right\} + \theta & \text{if } x = 1, \end{cases} \tag{9.6}$$

where $\mathbf{1}\left\{\theta \leq 0\right\}$ is the indicator of the event $\left\{\theta \leq 0\right\}$. The associated population risk is $R(\theta) = \frac{1}{2}(\theta^2 + \theta^2 \mathbf{1}\left\{\theta \leq 0\right\})$. Since $|R'(w) - R'(v)| \leq 2|w - v|$, the population risk is strongly convex and smooth, but it has discontinuous second derivative. The unique minimizer of the population risk is $\theta^* = 0$, and by an asymptotic expansion (see [189, Appendix A]), we have $\mathbb{E}[\theta_{1,i}] = \Omega(n^{-\frac{1}{2}})$. Consequently, the bias of $\overline{\theta}_{\text{AVGM}}$ is $\Omega(n^{-\frac{1}{2}})$, and the AVGM algorithm using $N = mn$ observations must suffer mean squared error $\mathbb{E}[(\overline{\theta}_{\text{AVGM}} - \theta^*)^2] = \Omega(n^{-1})$.

The previous example establishes the necessity of a smoothness condition. However, in a certain sense, it is a pathological case: both the smoothness condition given in Assumption 9B and the local strong convexity condition given in Assumption 9A are relatively innocuous for practical problems. For instance, both conditions will hold for standard forms of regression, such as linear and logistic, as long as the *population* data covariance matrix is not rank deficient and the data has suitable moments.

9.3.2 Bounds for Simple Averaging

We now turn to our first theorem that provides guarantees on the statistical error associated with the AVGM procedure. We recall that θ^* denotes the minimizer of the population risk function R, and that for each $i \in \{1, \ldots, m\}$, we use S_i to denote a dataset of n independent samples. For each i, we use $\theta_i \in \operatorname{argmin}_{\theta \in \Theta}\{\frac{1}{n}\sum_{x \in S_i} \ell(\theta; x)\}$ to denote a minimizer of the empirical risk for the dataset S_i, and we define the averaged vector $\overline{\theta}_{\text{AVGM}} = \frac{1}{m}\sum_{i=1}^m \theta_i$. The following simple result provides an asymptotic expansion of the averaged vector $\overline{\theta}_{\text{AVGM}}$ in terms of θ^*, which we can use to show asymptotic normality of $\overline{\theta}_{\text{AVGM}}$. In the theorem, we let X_j^i denote the jth observation in subsampled data set i. This is the only theorem we prove in this chapter, as its proof is substantially simpler than the mean-squared error proofs presented in the original work [189] off of which this chapter is based. (See Section 9.5.)

Theorem 9.1. *With the definitions and assumptions above, we have*

$$\overline{\theta}_{\text{AVGM}} - \theta^* = -[\nabla^2 R(\theta^*)]^{-1} \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^n \nabla \ell(\theta^*; X_j^i) + \mathcal{O}_P\left(\frac{1}{n}\right). \tag{9.7}$$

As an immediate consequence of Theorem 9.1, we obtain the following corollary.

Corollary 9.1. *Define the matrix*

$$\Sigma = [\nabla^2 R(\theta^*)]^{-1} \mathbb{E}[\nabla \ell(\theta^*; X) \nabla \ell(\theta^*; X)^\top][\nabla^2 R(\theta^*)]^{-1}.$$

Then so long as $m/n \to 0$ as $n \to \infty$, we have

$$\sqrt{N}\left(\overline{\theta}_{\text{AVGM}} - \theta^*\right) \xrightarrow{d} \mathsf{N}\left(0, \Sigma\right).$$

Proof Multiplying both sides of the equality (9.7) by $\sqrt{N} = \sqrt{nm}$, we obtain

$$\sqrt{N}(\overline{\theta} - \theta^*) = -\frac{1}{\sqrt{N}}[\nabla^2 R(\theta^*)]^{-1} \sum_{i=1}^m \sum_{j=1}^n \nabla \ell(\theta^*; X_j^i) + \mathcal{O}_P\left(\frac{\sqrt{N}}{n}\right). \tag{9.8}$$

Because $\sqrt{N}/n = \sqrt{m/n}$, the remainder term is $\mathcal{O}_P(\sqrt{m/n}) \to 0$ if $m/n \to 0$, and Slutsky's theorem (see, e.g. [175]) guarantees that as long as the first term converges in distribution, the $\mathcal{O}_P$ term is negligible. The first term in the preceding display is asymptotically normal with mean 0 and covariance Σ, because $\mathbb{E}[\nabla \ell(\theta^*; X)] = 0$. □

Under stronger conditions explored in our paper [189], it is possible to give mean-squared error convergence guarantees for the average mixture parameter $\overline{\theta}_{\text{AVGM}}$. In particular, under

some additional moment conditions on the Hessian smoothness constant H and a compactness assumption on Θ, Theorem 1 of Zhang, Duchi, and Wainwright [189] states the following (after a bit of inspection of the proof): for any $\epsilon > 0$,

$$\mathbb{E}\left[\left\|\overline{\theta}_{\text{AVGM}} - \theta^*\right\|_2^2\right] \leq \frac{(1+\epsilon)}{N}\mathbb{E}\left[\left\|\nabla^2 R(\theta^*)^{-1}\nabla\ell(\theta^*; X)\right\|_2^2\right] + C\left(1 + \frac{1}{\epsilon}\right)\frac{m^2}{\lambda^2 N^2}, \tag{9.9}$$

where the constant C hides several problem-dependent constants. This upper bound shows that the leading term decays proportionally to $(nm)^{-1}$, with the pre-factor depending inversely on the strong convexity constant λ and growing proportionally with the bound M on the loss gradient, and is what one would heuristically expect from the expansion (9.8).

The leading term in the upper bound (9.9) involves the product of the gradient $\nabla\ell(\theta^*; X)$ with the inverse Hessian. In many statistical settings, including linear regression, the effect of this matrix-vector multiplication is to perform some type of standardization. When the loss $\ell(\cdot; x) : \Theta \to \mathbb{R}$ is actually the negative log-likelihood $\ell_{\log}(x \mid \theta)$ for a parametric family of models $\{P_\theta\}$, we can make this intuition precise. In particular, under suitable regularity conditions [e.g. 118, Chapter 6], we can define the Fisher information matrix

$$I(\theta^*) := \mathbb{E}\left[\nabla\ell_{\log}(X \mid \theta^*)\nabla\ell_{\log}(X \mid \theta^*)^\top\right] = \mathbb{E}[\nabla^2\ell_{\log}(X \mid \theta^*)].$$

Recalling that $N = mn$ is the total sample size, let us define the neighborhood $B_2(\theta, t) := \{\theta' \in \mathbb{R}^d : \|\theta' - \theta\|_2 \leq t\}$. Under our assumptions, the Hájek-Le Cam minimax theorem [116, 175, Theorem 8.11] guarantees for *any estimator* $\widehat{\theta}_N$ based on N observations that

$$\lim_{\delta\to\infty}\liminf_{N\to\infty}\sup_{\theta\in B_2(\theta^*, \delta/\sqrt{N})} N\,\mathbb{E}_\theta\left[\left\|\widehat{\theta}_N - \theta\right\|_2^2\right] \geq \operatorname{tr}(I(\theta^*)^{-1}).$$

In connection with Theorem 9.1, we obtain the following result under the conditions of the theorem, whenever the loss functions are negative log-likelihoods (a mean-squared error bound based on inequality (9.9) is also possible).

Corollary 9.2. *If $m/n \to 0$ as $n \to \infty$, then*

$$\sqrt{mn}\left(\overline{\theta}_{\text{AVGM}} - \theta^*\right) \xrightarrow{d} \mathsf{N}\left(0, I(\theta^*)^{-1}\right).$$

Proof In the notation of Theorem 9.1, we have $\nabla\ell_{\log}(x \mid \theta^*) = \nabla\ell(\theta^*; x)$, and

$$\begin{aligned} I(\theta^*)^{-1} &= \mathbb{E}\left[I(\theta^*)^{-1}\nabla\ell_{\log}(X \mid \theta^*)\nabla\ell_{\log}(X \mid \theta^*)^\top I(\theta^*)^{-1}\right] \\ &= \mathbb{E}\Big[\left(\nabla^2 R(\theta^*)^{-1}\nabla\ell(\theta^*; X)\right)\left(\nabla^2 R(\theta^*)^{-1}\nabla\ell(\theta^*; X)\right)^\top\Big] \end{aligned}$$

by the definition of Fisher information. □

Corollary 9.2 and inequality (9.9) show that under appropriate conditions, the AVGM algorithm essentially achieves the best possible result. The important aspect of our bound, however, is that AVGM obtains this convergence guarantee without calculating an estimate on all $N = mn$ observations: instead, we calculate m independent estimators, and then average them to attain the convergence guarantee.

As noted in the introduction, these bounds are certainly to be expected for unbiased estimators, since in such cases averaging m independent solutions reduces the variance by $1/m$. In this sense, our results are similar to classical distributional convergence results in M-estimation: as Theorem 9.1 shows, for smooth enough problems, M-estimators behave asymptotically like averages [175, 118], and averaging multiple independent realizations reduces their variance. However, it is often desirable to use biased estimators, and such bias introduces difficulty in the analysis, which we explore more in the next section. The finite sample mean-squared error results, summarized in inequality (9.9), of our work [189] are also sharper than classical analyses, applicable to finite samples, and give explicit upper bounds. Lastly, our results are not tied to a specific model, which allows for fairly general sampling distributions.

9.3.3 Bounds for Subsampled Mixture Averaging

When the number of machines m is relatively small, Theorem 9.1, inequality (9.9) and Corollary 9.2 show that the convergence rate of the AVGM algorithm is mainly determined by the first term in the bound (9.9), which is at most $\frac{M^2}{\lambda^2 mn}$. In contrast, when the number of processors m grows, the second term in the bound (9.9), in spite of being $\mathcal{O}(n^{-2})$, may have non-negligible effect (that is, the constants hidden in the $\mathcal{O}_P(\sqrt{N}/n)$ in expression (9.8) may be large). This issue is exacerbated when the local strong convexity parameter λ of the risk R is close to zero or the Lipschitz continuity constant L of $\nabla\ell$ is large. This concern motivated our development of the subsampled average mixture (SAVGM) algorithm; we now review a few theoretical results available for the method (see [189]).

We begin by explicitly codifying our assumptions. First, we assume that the parameter space Θ is compact (in addition to being convex). In addition to our previously stated assumptions, we require a few additional regularity conditions on the empirical risk functions, which are necessary due to the additional randomness introduced by the subsampling in SAVGM (and because we provide mean-squared-error bounds). It is simplest to state these in terms of the functions $\theta \mapsto \ell(\theta; x)$, and we note that, as with Assumption 9A, we require these to hold only locally around the optimal point θ^*, in particular within some Euclidean ball $U = \{\theta \in \mathbb{R}^d \mid \|\theta^* - \theta\|_2 \le \rho\} \subseteq \Theta$ of radius $\rho > 0$.

Assumption 9C (Smoothness)**.** *There are finite constants M, L such that the first and the second partial derivatives of f exist and satisfy the bounds*

$$\mathbb{E}[\|\nabla\ell(\theta; X)\|_2^8] \le M^8 \quad \textit{and} \quad \mathbb{E}[\left\|\nabla^2\ell(\theta; X) - \nabla^2 R(\theta)\right\|_2^8] \le L^8 \quad \textit{for all } \theta \in U.$$

In addition, for any $x \in \mathcal{X}$, the Hessian matrix $\nabla^2 \ell(\theta; x)$ is $H(x)$-Lipschitz continuous, meaning that

$$\left\|\left\|\nabla^2 \ell(\theta'; x) - \nabla^2 \ell(\theta; x)\right\|\right\|_2 \leq H(x) \left\|\theta' - \theta\right\|_2 \quad \text{for all } \theta, \theta' \in U.$$

We require that $\mathbb{E}[H(X)^8] \leq H^8$ and $\mathbb{E}[(H(X) - \mathbb{E}[H(X)])^8] \leq H^8$ for a constant $H < \infty$. Lastly, for each $x \in \mathcal{X}$, the third derivatives of ℓ are $G(x)$-Lipschitz continuous,

$$\left\|\left(\nabla^3 \ell(\theta; x) - \nabla^3 \ell(\theta'; x)\right)(u \otimes u)\right\|_2 \leq G(x) \left\|\theta - \theta'\right\|_2 \left\|u\right\|_2^2 \quad \text{for all } \theta, \theta' \in U, \text{ and } u \in \mathbb{R}^d,$$

where $\mathbb{E}[G^8(X)] \leq G^8$ for some constant $G < \infty$.

It is easy to verify that Assumption 9C holds for least-squares regression with $G = 0$. It also holds for various types of non-linear regression problems (e.g., logistic, multinomial) as long as the covariates have finite eighth moments. With this set-up, the SAVGM method—averaging with bootstrap resampling—enjoys improved performance [189, Theorem 4, sharpened]:

Theorem 9.2. *Under Assumptions 9A and 9C, the output $\overline{\theta}_{\text{SAVGM}} = (\overline{\theta}_{\text{AVGM}} - r\overline{\theta}_2)/(1-r)$ of the bootstrap* SAVGM *algorithm has mean-squared error bounded, for any $\epsilon > 0$, as*

$$\mathbb{E}\left[\left\|\overline{\theta}_{\text{SAVGM}} - \theta^*\right\|_2^2\right] \leq \frac{1 + \epsilon + 3r}{(1-r)^2} \cdot \frac{1}{nm} \mathbb{E}\left[\left\|\nabla^2 R(\theta^*)^{-1} \nabla \ell(\theta^*; X)\right\|_2^2\right] \tag{9.10}$$
$$+ c\left(\frac{G^2 M^6}{\lambda^6} + \frac{M^4 H^2 d \log d}{\lambda^4}\right)\left(\frac{1 + \frac{1}{\epsilon}}{r(1-r)^2}\right) n^{-3} + \mathcal{O}\left(\frac{1}{(1-r)^2} m^{-1} n^{-2}\right)$$

for a numerical constant c.

Inspecting the conclusions of Theorem 9.2, we see that the the $\mathcal{O}(n^{-2})$ term in the bound (9.9) has been eliminated. The reason for this elimination is that subsampling at a rate r reduces the bias of the SAVGM algorithm to $\mathcal{O}(n^{-3})$, whereas in contrast, the bias of the AVGM algorithm induces terms of order n^{-2}. Theorem 9.2 suggests that the performance of the SAVGM algorithm is affected by the subsampling rate r; in order to minimize the upper bound (9.10) in the regime $m < N^{2/3}$, the optimal choice is of the form $r \propto C\sqrt{m}/n = Cm^{3/2}/N$. Roughly, as the number of machines m becomes larger, we may increase r, since we enjoy averaging affects from the SAVGM algorithm.

Let us consider the relative effects of having larger numbers of machines m for both the AVGM and SAVGM algorithms, which provides some guidance to selecting m in practice. We define $\sigma^2 = \mathbb{E}[\|\nabla^2 R(\theta^*)^{-1} \nabla \ell(\theta^*; X)\|_2^2]$ to be the asymptotic variance. Then to obtain the optimal convergence rate of σ^2/N, we must have

$$m \ll N^{\frac{1}{2}}, \quad \text{or} \quad m \ll n \tag{9.11}$$

in Theorem 9.1 and expression (9.9). Applying the bound of Theorem 9.2, we find that to obtain the same rate after setting $r = Cm^{3/2}/N$ as in the previous paragraph, that

$$m \ll N^{\frac{2}{3}} \quad \text{or} \quad m \ll n^2. \tag{9.12}$$

Comparing inequalities (9.11) and (9.12), we see that in both cases m may grow polynomially with the global sample size N while still guaranteeing optimal convergence rates, and this asymptotic growth may be faster in the subsampled case (9.12). Averaging methods are, of course, not a panacea: the allowed number of partitions m does not grow linearly in either case, so blindly increasing the number of machines proportionally to the total sample size N will not lead to a useful estimate.

9.3.4 Stochastic Gradient Descent with Averaging

The previous strategy involved a combination of stochastic gradient descent and standard gradient descent. In many settings, it may be appealing to use only a stochastic gradient algorithm, due to their ease of their implementation and limited computational requirements. In this section, we describe an extension of the AVGM algorithm to the case in which each machine computes an approximate minimizer using only stochastic gradient descent, which we presented and reviewed in Chapter 3, Section 3.1.

More precisely, the averaged stochastic gradient algorithm (SGDAVGM) is performs the following two steps:

(1) Given some constant $c > 1$, each machine performs n iterations of stochastic gradient descent (3.2) on its local dataset of n samples using the stepsize $\alpha_k = \frac{c}{\lambda k}$, then outputs the resulting local parameter θ_i'.

(2) The algorithm computes the average $\overline{\theta}^n = \frac{1}{m}\sum_{i=1}^m \theta_i'$.

To prove convergence of our stochastic gradient-based averaging algorithms, we require the following smoothness and strong convexity condition, which is an alternative to the assumptions used previously.

Assumption 9D (Smoothness and Strong Convexity II)**.** *There exists a function $H : \mathcal{X} \to \mathbb{R}_+$ such that*

$$\left|\left|\left|\nabla^2 \ell(\theta; x) - \nabla^2 \ell(\theta^*; x)\right|\right|\right|_2 \leq H(x) \left\|\theta - \theta^*\right\|_2 \quad \textit{for all } x \in \mathcal{X},$$

and $\mathbb{E}[H^2(X)] \leq H^2 < \infty$. There are finite constants M and L such that

$$\mathbb{E}[\|\nabla \ell(\theta; X)\|_2^4] \leq M^4, \quad \textit{and} \quad \mathbb{E}[\left|\left|\left|\nabla^2 \ell(\theta^*; X)\right|\right|\right|_2^4] \leq L^4 \quad \textit{for each fixed } \theta \in \Theta.$$

In addition, the population function R is λ-strongly convex over the space Θ, that is,

$$\nabla^2 R(\theta) \succeq \lambda I_{d \times d} \quad \textit{for all } \theta \in \Theta.$$

Assumption 9D does not require as many moments as does Assumption 9C, but it does require each moment bound to hold globally, that is, over the entire space Θ, rather than only in a neighborhood of the optimal point θ^*. Similarly, the necessary curvature—in the form

of the lower bound on the Hessian matrix $\nabla^2 R$—is also required to hold globally, rather than only locally. Nonetheless, Assumption 9D holds for many common problems; for instance, it holds for any linear regression problem in which the covariates have finite fourth moments and the domain Θ is compact.

The following result [189, Theorem 5] characterizes the mean-squared error of this procedure in terms of the constants

$$\alpha := 4c^2 \quad \text{and} \quad \beta := \max\left\{\left\lceil \frac{cL}{\lambda} \right\rceil, \frac{c\alpha^{3/4} M^{3/2}}{(c-1)\lambda^{5/2}} \left(\frac{\alpha^{1/4} H M^{1/2}}{\lambda^{1/2}} + \frac{4M + L\mathsf{r}_2}{\rho^{3/2}} \right) \right\}.$$

Theorem 9.3. *Under Assumption 9D, the output $\overline{\theta}^n$ of the* Savgm *algorithm has mean-squared error upper bounded as*

$$\mathbb{E}\left[\left\| \overline{\theta}^n - \theta^* \right\|_2^2 \right] \leq \frac{\alpha M^2}{\lambda^2 mn} + \frac{\beta^2}{n^{3/2}}. \tag{9.13}$$

Theorem 9.3 shows that the averaged stochastic gradient descent procedure attains the optimal convergence rate $\mathcal{O}(N^{-1})$ as a function of the total number of observations $N = mn$. The constant and problem-dependent factors are certainly worse than those in the earlier results we presented in Theorems 9.1 and 9.2, but the practical implementability of such a procedure may in some circumstances outweigh those differences.

9.4 Summary

Large scale statistical inference problems are challenging, and the difficulty of solving them will only grow as data becomes more abundant: the amount of data we collect is growing much faster than the speed or storage capabilities of our computers. The Avgm, Savgm, and SGDavgm methods provide strategies for efficiently solving such large-scale risk minimization problems, enjoying performance comparable to an oracle method that is able to access the entire large dataset. That said, these methods may not always have good practical performance; it may be that they provide good initialization for further efficient optimization. An understanding of the interplay between statistical efficiency and communication is of general interest as well: given the expense of communication in modern systems, minimizing it is of increasing importance [14, 79]. The algorithms in this chapter have shown that, in some scenarios, it is possible to perform *very little* communication; in the next, we show that these procedures are (essentially) communication optimal.

9.5 Proof of Theorem 9.1

The proof of this theorem follows those of similar standard distributional convergence results; see, for example, Lehmann and Casella [118]. We begin by considering a somewhat simpler

problem: we assume that $m = 1$ and study the convergence of

$$\widehat{\theta} := \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i)$$

for X_i sampled i.i.d. according to P. By assumption, ℓ is twice continuously differentiable, and as a consequence, we have by a Taylor expansion that

$$0 = \nabla \widehat{R}(\widehat{\theta}) = \nabla \widehat{R}(\theta^*) + \nabla^2 \widehat{R}(\widetilde{\theta})(\widehat{\theta} - \theta^*),$$

where $\widetilde{\theta} = t\theta^* + (1-t)\widehat{\theta}$ for some $t \in [0, 1]$. Expanding this expression by adding and subtracting $\nabla^2 R(\theta^*)(\widehat{\theta} - \theta^*)$, we have

$$0 = \nabla \widehat{R}(\theta^*) + \nabla^2 R(\theta^*)(\widehat{\theta} - \theta^*) + (\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 R(\theta^*))(\widehat{\theta} - \theta^*).$$

In particular, we find that since $\nabla^2 R(\theta^*) \succ 0$ by assumption,

$$\widehat{\theta} - \theta^* = -[\nabla^2 R(\theta^*)]^{-1} \nabla \widehat{R}(\theta^*) - \underbrace{[\nabla^2 R(\theta^*)]^{-1} (\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 R(\theta^*))(\widehat{\theta} - \theta^*)}_{=: \mathcal{T}}. \tag{9.14}$$

To complete the proof, it remains to show that $\mathcal{T} = \mathcal{O}_P(1/n)$, because the average of m independent terms $\mathcal{T}$, each with $\mathcal{T} = \mathcal{O}_P(n^{-1})$, will still be $\mathcal{O}_P(n^{-1})$. We show the result in three steps. First, we assume that $\widehat{\theta} - \theta^* = \mathcal{O}_P(n^{-\frac{1}{2}})$; we will show that this implies the result. After this, we will assume simply the consistency guarantee that $\widehat{\theta} - \theta^* = o_P(1)$, which we show will imply $\widehat{\theta} - \theta^* = \mathcal{O}_P(n^{-\frac{1}{2}})$. After this, we simply cite standard results guaranteeing consistency of M-estimators [118, 175].

Beginning under the assumption that $\widehat{\theta} - \theta^* = \mathcal{O}_P(n^{-\frac{1}{2}})$, we always have that

$$\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 R(\theta^*) = \underbrace{\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 \widehat{R}(\theta^*)}_{=: \mathcal{T}_1} + \underbrace{\nabla^2 \widehat{R}(\theta^*) - \nabla^2 R(\theta^*)}_{=: \mathcal{T}_2}.$$

For the first term $\mathcal{T}_1$, we have by the Lipschitz Assumption 9B on $\nabla^2 \ell$ that

$$\left\| \nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 \widehat{R}(\theta^*) \right\| \le \frac{1}{n} \sum_{i=1}^{n} H(X_i) \left\| \widetilde{\theta} - \theta^* \right\|_2 \le \frac{1}{n} \sum_{i=1}^{n} H(X_i) \left\| \widehat{\theta} - \theta^* \right\|_2.$$

For any $\epsilon > 0$, there exists a $C(\epsilon)$ such that

$$P(\|\widehat{\theta} - \theta^*\|_2 \ge C(\epsilon)/\sqrt{n}) \le \epsilon,$$

and similarly we have that $\frac{1}{n} \sum_{i=1}^{n} H(X_i) = \mathcal{O}_P(1)$ by Assumption 9B. Now we show that $\mathcal{T}_1 = \mathcal{O}_P(n^{-\frac{1}{2}})$. Indeed, for fixed $t > 0$, to have $\frac{1}{n} \sum_{i=1}^{n} H(X_i) \|\widehat{\theta} - \theta^*\|_2 \ge t/\sqrt{n}$ requires that at least one of $\frac{1}{n} \sum_{i=1}^{n} H(X_i) \ge \sqrt{t}$ or $\|\widehat{\theta} - \theta^*\|_2 \ge \sqrt{t/n}$, and consequently, we see

that $\|\mathcal{T}_1\| = \|\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 \widehat{R}(\theta^*)\| = \mathcal{O}_P(n^{-\frac{1}{2}})$. The central limit theorem implies that $\mathcal{T}_2 = \mathcal{O}_P(n^{-\frac{1}{2}})$, and revisiting the equality (9.14), we have that

$$\mathcal{T} = V(\widehat{\theta} - \theta^*) \quad \text{for some random } V \in \mathbb{R}^{d \times d} \text{ with } V = \mathcal{O}_P(n^{-\frac{1}{2}}).$$

Since $\widehat{\theta} - \theta^* = \mathcal{O}_P(n^{-\frac{1}{2}})$ by assumption, we see that inequality (9.7) holds.

Now, we show that under the consistency condition that $\widehat{\theta} - \theta^* \xrightarrow{p} 0$, we have $\widehat{\theta} - \theta^* = \mathcal{O}_P(n^{-\frac{1}{2}})$. Indeed, recalling expression (9.14), we have

$$\left(I_{d \times d} + [\nabla^2 R(\theta^*)]^{-1}(\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 R(\theta^*))\right)(\widehat{\theta} - \theta^*) = -[\nabla^2 R(\theta^*)]^{-1} \nabla \widehat{R}(\theta^*).$$

Using reasoning identical to the previous paragraph, we have that $\nabla^2 \widehat{R}(\widetilde{\theta}) - \nabla^2 R(\theta^*) = o_P(1)$ under our assumptions, so that

$$(I_{d \times d} + V)(\widehat{\theta} - \theta^*) = -[\nabla^2 R(\theta^*)]^{-1} \nabla \widehat{R}(\theta^*) \quad \text{for some } V = o_P(1).$$

By the central limit theorem, it is clear that $\nabla \widehat{R}(\theta^*) = \mathcal{O}_P(n^{-\frac{1}{2}})$, and since for sufficiently large n we have $\|V\| < \epsilon$ with probability at least $1 - \epsilon$, we have $\widehat{\theta} - \theta^* = \mathcal{O}_P(n^{-\frac{1}{2}})$. But now we are simply in the first case, in which case our previous reasoning implies the desired result.

Lastly, we must argue that under the assumptions of the theorem, the empirical risk minimizer $\widehat{\theta}$ is consistent for the population minimizer θ^*. But by the positive definiteness of $\nabla^2 R(\theta^*)$, we know that θ^* is unique, and the smoothness of $\theta \mapsto \nabla^2 R(\theta)$ guarantees consistency (cf. [175, 118, Chapter 6.3]).

Chapter 10

Optimality guarantees for distributed estimation

In this final chapter of the thesis, we complement the results in Chapter 9 by establishing lower bounds on minimax risk for distributed statistical estimation under communication constraints. In the language of Chapter 2, we formulate a question of constrained minimax risk (2.4), asking for (and establishing) lower bounds on the best possible rates of convergence for estimation procedures constrained to use a limited communication budget. Such lower bounds reveal the minimum amount of communication required by any procedure to achieve the centralized minimax-optimal rates of convergence for statistical estimation problems. We study two classes of protocols: one in which machines send messages independently (over channels without feedback), and a second allowing for interactive communication (specifically, protocols in which machines may freely broadcast any messages they send to a central server to all other machines). We establish lower bounds for a variety of problems, including several types of location models and for parameter estimation in regression models.

10.1 Introduction

Rapid growth in the size and scale of datasets has fueled increasing interest in statistical estimation in distributed settings (a highly incomplete list includes the works [33, 49, 128, 140, 54, 166], as well as some of our own work in the previous chapters). As noted in Chapter 9, modern data sets are often too large to be stored on a single machine, and so it is natural to consider methods that involve multiple machines, each assigned a smaller subset of the full dataset. Yet communication between machines or processors is often expensive, slow, or power-intensive; as noted by Fuller and Millett [79] in a survey of the future of computing, "there is no known alternative to parallel systems for sustaining growth in computing performance," yet the power consumption and latency of communication is often relatively high. Indeed, bandwidth limitations on network and inter-chip communication often impose significant bottlenecks on algorithmic efficiency. It is thus important to study the amount of

communication required between machines or chips in algorithmic development, especially as we scale to larger and larger datasets.

Building off of the low-communication algorithms of the previous chapter, the focus of this chapter is the communication complexity of a few classes of statistical estimation problems. Suppose we are interested in estimating some parameter $\theta(P)$ of an unknown distribution P, based on a dataset of N i.i.d. observations. In the classical setting, one considers *centralized estimators* that have access to all N observations. In contrast, in the distributed setting, one is given m different machines, and each machine is assigned a subset of the sample of size $n = \lfloor \frac{N}{m} \rfloor$. Each machine may perform arbitrary operations on its own subset of data, and it then communicates results of these intermediate computations to the other processors or to a central fusion node. In this chapter, we try to answer the following question: what is the minimal number of bits that must be exchanged in order to achieve the optimal estimation error achievable by centralized schemes?

More precisely, we study problems of the following form (recall Chapter 2 and the constrained minimax risk (2.4)): given a budget B of the total number of bits that may be communicated from the m distributed datasets, what is the minimax risk of any estimator based on the communicated messages? While there is a rich literature connecting information-theoretic techniques with the risk of statistical estimators (e.g. [101, 188, 185, 173]), little of it characterizes the effects of limiting communication. In this chapter, we present minimax lower bounds for distributed statistical estimation. For some problems, we show an exponential gap between the number of bits required to describe a problem (and solutions to the problem to optimal statistical precision) and the amount of communication required to solve the problem (see Theorems 10.1 and 10.2). By comparing our lower bounds with recent results in statistical estimation, we can identify the minimal communication cost that a distributed estimator must pay to have performance comparable to classical centralized estimators. Moreover, the results of Chapter 9 show that these fundamental limits are, to within logarithmic factors, achievable; it is possible to provide estimators that are optimal both from statistical and communication-focused perspectives.

10.2 Problem setting

We begin with a formal description of the statistical estimation problems considered here. As we have done throughout the thesis, let $\mathcal{P}$ denote a family of distributions and let $\theta : \mathcal{P} \to \Theta \subseteq \mathbb{R}^d$ denote a function defined on $\mathcal{P}$. A canonical example throughout the chapter is the problem of mean estimation, in which $\theta(P) = \mathbb{E}_P[X]$. Suppose that, for some fixed but unknown member P of $\mathcal{P}$, there are m sets of data stored on individual machines, where each subset $X^{(i)}$ is an i.i.d. sample of size n from the unknown distribution P.[1] Given this distributed collection of local data sets, our goal is to estimate $\theta(P)$ based on the m samples $X^{(1)}, \ldots, X^{(m)}$, but using limited communication.

[1] Although we assume in this chapter that every machine has the same amount of data, our technique generalizes to prove tight lower bounds for distinct data sizes on different machines.

We consider a class of distributed protocols Π, in which at each round $t = 1, 2, \ldots$, machine i sends a message $Y_{t,i}$ that is a measurable function of the local data $X^{(i)}$, and potentially of past messages. It is convenient to model this message as being sent to a central fusion center. Let $\overline{Y}_t = \{Y_{t,i}\}_{i \in [m]}$ denote the collection of all messages sent at round t. Given a total of T rounds, the protocol Π collects the sequence $(\overline{Y}_1, \ldots, \overline{Y}_T)$, and constructs an estimator $\widehat{\theta} := \widehat{\theta}(\overline{Y}_1, \ldots, \overline{Y}_T)$. The length $L_{t,i}$ of message $Y_{t,i}$ is the minimal number of bits required to encode it, and the total $L = \sum_{t=1}^{T} \sum_{i=1}^{m} L_{t,i}$ of all messages sent corresponds to the *total communication cost* of the protocol. Note that the communication cost is a random variable, since the length of the messages may depend on the data, and the protocol may introduce auxiliary randomness.

It is useful to distinguish two different classes, namely *independent* versus *interactive* protocols. An independent protocol Π is based on a single round ($T = 1$) of communication, in which machine i sends message $Y_{1,i}$ to the fusion center. Since there are no past messages, the message $Y_{1,i}$ can depend only on the local sample $X^{(i)}$. Given a family $\mathcal{P}$, the class of independent protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\text{ind}}(B, \mathcal{P}) = \left\{ \text{ independent protocols } \Pi \text{ s.t. } \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\sum_{i=1}^{m} L_i \right] \leq B \right\}. \tag{10.1}$$

(For simplicity, we use Y_i to indicate the message sent from processor i and L_i to denote its length in the independent case.) It can be useful in some situations to have more granular control on the amount of communication, in particular by enforcing budgets on a per-machine basis. In such cases, we introduce the shorthand $B_{1:m} = (B_1, \ldots, B_m)$ and define

$$\mathcal{A}_{\text{ind}}(B_{1:m}, \mathcal{P}) = \{\text{independent protocols } \Pi \text{ s.t. } \mathbb{E}_P[L_i] \leq B_i \text{ for } i \in [m] \text{ and } P \in \mathcal{P}\}. \tag{10.2}$$

In contrast to independent protocols, the class of interactive protocols allows for interaction at different stages of the message passing process. In particular, suppose that machine i sends message $Y_{t,i}$ to the fusion center at time t, who then posts it on a "public blackboard," where all machines can read $Y_{t,i}$. We think of this as a global broadcast system, which may be natural in settings in which processors have limited power or upstream capacity, but the centralized fusion center can send messages without limit. In the interactive setting, the message $Y_{t,i}$ should be viewed as a measurable function of the local data $X^{(i)}$, and the past messages $\overline{Y}_{1:t-1}$. The family of interactive protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\text{inter}}(B, \mathcal{P}) = \left\{ \text{interactive protocols } \Pi \text{ such that } \quad \sup_{P \in \mathcal{P}} \mathbb{E}_P[L] \leq B \right\}. \tag{10.3}$$

We conclude this section by specializing the general minimax framework of Chapter 2 to that used throughout this chapter. We wish to characterize the best achievable performance of estimators $\widehat{\theta}$ that are functions of only the messages $(\overline{Y}_1, \ldots, \overline{Y}_T)$. We measure the quality of a protocol and estimator $\widehat{\theta}$ by the mean-squared error

$$\mathbb{E}_{P,\Pi} \left[\|\widehat{\theta}(\overline{Y}_1, \ldots, \overline{Y}_T) - \theta(P)\|_2^2 \right],$$

where the expectation is taken with respect to the protocol Π and the m i.i.d. samples $X^{(i)}$ of size n from distribution P. Now we cast the constrained minimax risk (2.4) outlined in Chapter 2 in the framework of this chapter. Given a class of distributions $\mathcal{P}$, parameter $\theta : \mathcal{P} \to \Theta$, and communication budget B, the *minimax risk for independent protocols* is

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B) := \inf_{\Pi \in \mathcal{A}_{\text{ind}}(B,\mathcal{P})} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,\Pi} \left[\left\| \widehat{\theta}(Y_1, \ldots, Y_m) - \theta(P) \right\|_2^2 \right]. \tag{10.4}$$

Here, the infimum is taken jointly over all independent procotols Π that satisfy the budget constraint B, and over all estimators $\widehat{\theta}$ that are measurable functions of the messages in the protocol. This minimax risk should also be understood to depend on both the number of machines m and the individual sample size n. The *minimax risk for interactive protocols*, denoted by $\mathfrak{M}^{\text{inter}}$, is defined analogously, where the infimum is instead taken over the class of interactive protocols. These communication-dependent minimax risks are the central objects in this chapter: they provide a sharp characterization of the optimal rate of statistical estimation as a function of the communication budget B.

10.3 Related Work

There is of course a substantial literature on communication complexity in many areas, ranging from theoretical computer science (beginning with the work of Yao [186] and Abelson [1]) to decentralized detection and estimation (e.g. in work by Tsitsiklis and Luo [171, 123]) and information theory (see, for example, Han and Amari [89] and El Gamal and Kim [72]). In addition, our work builds from the long literature on minimax rates of convergence in statistics (recall Chapter 2, or see, e.g. Ibragimov and Has'minskii [101], Yu [188], and Yang and Barron [185]). We review a few of these and highlight their main results in the coming paragraphs.

In the computer science literature, Yao [186] and Abelson [1] initiated the study of communication complexity (see also the survey by Kushilevitz and Nisan [111]). Using our notation, the prototypical problem in this setting is as follows. Consider two sets $\mathcal{X}$ and $\mathcal{Y}$ and a function $\theta : \mathcal{X} \times \mathcal{Y} \to \Theta$ with range Θ. We assume there are two parties (usually given the names Alice and Bob), one of which holds a point $x \in \mathcal{X}$ and the other $y \in \mathcal{Y}$, and we wish to compute the value $\theta(x, y)$. The (classical) communication complexity problem is to find the protocol using the fewest bits that guarantees that $\theta(x, y)$ is computed correctly for all possible settings of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. More recent work studies randomization and introduces information-theoretic measures for communication complexity (e.g. Chakrabarti et al. [41] and Bar-Yossef et al. [15]), where the problem is to guarantee that $\theta(x, y)$ is computed correctly with high probability under a given (known) distribution P on x and y. In contrast, our goal—indeed, the goal of most of statistical inference and estimation—is to recover characteristics of the distribution P (which we assume to be unknown) based on observations X drawn from P. Though this difference is somewhat subtle, it makes work on communication

complexity difficult to apply in our settings. However, lower bounds on the estimation of population quantities $\theta(P)$ based on communication-constrained observations—including those we present here—do imply lower bounds in classical communication complexity settings. (For related issues, see also the discussion in the introduction to Chapter 7 and Section 7.6.1 on privacy.) We also prove our results assuming only an expected bound on communication.

Work in decentralized detection and estimation also studies limits of communication. For example, Tsitsiklis and Luo [172] provide lower bounds on the difficulty of distributed convex optimization, and in subsequent work also study limits on certain distributed algebraic computations [122, 123]. In these problems, as in other early work in communication complexity, data held by the distributed parties may be chosen adversarially, which precludes conclusions about statistical estimation. Other work in distributed control provides lower bounds on consensus and averaging, but in settings where messages sent are restricted to be of particular smooth forms [142]. Study of communication complexity has also given rise to interesting algorithmic schemes; for example, Luo [121] considers architectures in which machines may send only a single bit to a centralized processor; for certain problems, he shows that if each machine receives a single one-dimensional sample, it is possible to achieve the optimal centralized rate to within constant factors.

Han and Amari [89] provide a survey of distributed estimation problems from an information theoretic perspective. In particular, they focus on the problem of testing a hypothesis or estimating a parameter from samples $\{(x_i, y_i)\}_{i=1}^n$ where $\{(x_i)\}_{i=1}^n$ and $\{(y_i)\}_{i=1}^n$ are correlated but stored separately in two machines. Han and Amari study estimation error for fixed encoding rates $R > 0$, meaning 2^{nR} messages may be sent. In all the settings we study, however, this setting is essentially trivial: any non-zero rate allows distributed estimation at the statistically optimal mean-squared error (i.e. that attainable with no communication constraints). They also address some zero-rate statistical inference problems, that is, those for which they have a sequence of rates R_n and $R_n \to 0$ as $n \to \infty$. Even these are too lenient for the distributed statistical estimation settings we consider. As an example, assume that m machines have n i.i.d. observations according to a Bernoulli(θ) distribution. Then each can send a message capturing perfectly the number of non-zero observations using at most $\lceil \log_2 n \rceil$ bits—so that the rate is $R_n = \frac{1}{n} \log_2 n$—and attain (statistically) optimal estimation. In our settings, we are interested in more quantitative results, such as understanding at what rates R_n may go to zero or the consequences of setting $R_n \le t/n$ for specific values $t > 0$, while still attaining optimal statistical estimation; these are somewhat more stringent conditions.

10.4 Main results

With our setup in place, we now turn to the statement of our main results, along with some discussion of their consequences. Our first set of results applies in essentially all situations by providing bounds exclusively based on metric entropy, which implies (somewhat trivially) that any procedure must communicate at least as many bits as are required to describe

a problem solution. Subsequently, we extend these results for interactive communication schemes, showing that these bounds are (essentially) tight for some problems, but can be made considerably stronger for some types of mean estimation problems. We conclude the section by giving our sharpest results for non-interactive communication, outlining a few open questions.

10.4.1 Lower bound based on metric entropy

We begin with a general but relatively naive lower bound that depends only on the geometric structure of the parameter space, as captured by its metric entropy. As in Chapter 2, given a subset $\Theta \subset \mathbb{R}^d$, we say $\{\theta^1, \ldots, \theta^K\}$ are δ-separated if $\|\theta^i - \theta^j\|_2 \geq \delta$ for $i \neq j$. We then define the *packing number* of Θ as

$$M_\Theta(\delta) := \max\left\{K \in \mathbb{N} \mid \{\theta_1, \ldots, \theta^K\} \subset \Theta \text{ are } \delta\text{-separated}\right\}. \tag{10.5}$$

The *packing entropy* of Θ is simply the logarithm of the packing number, $\log_2 M_\Theta(\delta)$. The function $\delta \mapsto \log_2 M_\Theta(\delta)$ is continuous from the right and non-increasing in δ, so we may define the inverse function $\log_2 M_\Theta^{-1}(B) := \sup\{\delta \mid \log_2 M_\Theta(\delta) \geq B\}$, and if $\delta = \log_2 M_\Theta^{-1}(B)$, then $\log_2 M_\Theta(\delta) \geq B$. With this definition, we have the following (essentially standard) proposition.

Proposition 10.1. *For any family of distributions $\mathcal{P}$ and parameter set $\theta = \theta(\mathcal{P})$, the interactive minimax risk is lower bounded as*

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \geq \frac{1}{8}\left(\log_2 M_\Theta^{-1}(2B+2)\right)^2. \tag{10.6}$$

Proof We prove the lower bound via a standard information-theoretic argument. Fix $\delta > 0$, and let $\mathcal{V} = [M_\Theta(2\delta)]$ index a maximal 2δ-packing of Θ, which we identify by $\{\theta_v\}_{v \in \mathcal{V}} \subset \Theta$. Fix an (arbitrary) protocol Π for communication.

Following the standard reduction from estimation to testing and using Fano's method as in Chapter 2, Section 2.2.3, let V be sampled uniformly from $\mathcal{V}$. Then for any messages $Y = (Y_1, \ldots, Y_T)$ sent by the protocol Π, Fano's inequality implies

$$\max_{v \in \mathcal{V}} \mathbb{E}\left[\|\widehat{\theta}(Y) - \theta_v\|_2^2\right] \geq \delta^2\left(1 - \frac{I(V;Y)+1}{\log_2 M_\Theta(2\delta)}\right).$$

Because $I(V;Y) \leq H(Y)$, Shannon's source coding theorem [47, Chapter 5] guarantees the lower bound $I(V;Y) \leq H(Y) \leq B$. Since the protocol Π was arbitrary, we have as an immediate consequence of the previous display that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \geq \delta^2\left(1 - \frac{B+1}{\log_2 M_\Theta(2\delta)}\right) \quad \text{for any } \delta \geq 0. \tag{10.7}$$

Using inequality (10.7), the remainder of the proof is straightforward. Indeed, we have

$$1 - \frac{B+1}{\log_2 M_\Theta(2\delta)} \geq \frac{1}{2} \text{ if and only if } \frac{\log_2 M_\Theta(2\delta)}{B+1} \geq 2,$$

which is implied by $2\delta \leq \log_2 M_\Theta^{-1}(2B+2)$. Setting $\delta = \frac{1}{2}\log_2 M_\Theta^{-1}(2B+2)$ thus gives the result. □

Of course, the same lower bound also holds for $\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B)$, since any independent protocol is a special case of an interactive protocol. Although Proposition 10.1 is a relatively generic statement, not exploiting any particular structure of the problem, it is in general unimprovable by more than constant factors, as the following example illustrates.

Example: Bounded mean estimation. Suppose that our goal is to estimate the mean $\theta = \theta(P)$ of a class of distributions $\mathcal{P}$ supported on the interval $[0,1]$, so that $\Theta = \theta(\mathcal{P}) = [0,1]$. Suppose that a single machine ($m = 1$) receives n i.i.d. observations X_i according to P. Since the packing entropy is lower bounded as $\log_2 M_\Theta(\delta) \geq \log_2(1/\delta)$, the lower bound (10.6) implies

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B) \geq \mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \geq \frac{1}{8}\left(\frac{1}{4}\, 2^{-2B}\right)^2.$$

Thus, setting $B = \frac{1}{4}\log_2 n$ yields the lower bound $\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}([0,1]), B) \geq \frac{1}{128n}$. This lower bound is sharp up to the constant pre-factor, since it can be achieved by a simple method. Given its n observations, the single machine can compute the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Since the sample mean lies in the interval $[0,1]$, it can be quantized to accuracy $1/n$ using $\log_2 n$ bits, and this quantized version $\widehat{\theta}$ can be transmitted. A straightforward calculation shows that $\mathbb{E}[(\widehat{\theta} - \theta)^2] \leq \frac{2}{n}$, so Proposition 10.1 yields an order-optimal bound in this case.

10.4.2 Independent protocols in multi-machine settings

We now turn to the more interesting multi-machine setting ($m > 1$). We would like to study how the *budget* B—the of bits required to achieve the minimax rate—scales with the number of machines m. For our first set of results in this setting, we consider the non-interactive case, where each machine i sends messages Y_i independently of all the other machines. We can obtain our most precise results in this setting, and the results here serve as pre-cursors to the results in the next section, where we allow feedback.

We first provide lower bounds for the problem of mean estimation in the parameter for the *d-dimensional normal location family model*

$$\mathcal{N}_d = \{\mathsf{N}(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \Theta = [-1,1]^d\}. \tag{10.8}$$

Here each machine receives an i.i.d. sample of size n from a normal distribution $\mathsf{N}(\theta, \sigma^2 I_{d\times d})$ with unknown mean θ. In this case—with independent communication—we obtain the following result on estimating the unknown mean θ, whose proof we provide in Section 10.9.3.

Theorem 10.1. *For $i = 1, \ldots, m$, assume that each machine has communication budget B_i, and receives an i.i.d. sample of size n from a distribution $P \in \mathcal{N}_d$. There exists a universal (numerical) constant c such that*

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{N}_d, B_{1:m}) \geq c\frac{\sigma^2 d}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log m}, \frac{m}{(\sum_{i=1}^m \min\{1, \frac{B_i}{d}\}) \log m} \vee 1\right\}. \tag{10.9}$$

Given centralized access to the full mn-sized sample, a reasonable procedure would be to compute the sample mean, leading to an estimate with mean-squared error $\frac{\sigma^2 d}{mn}$, which is minimax optimal [118]. Consequently, the lower bound (10.9) shows that each machine individually must communicate at least $d/\log m$ bits for a decentralized procedure to match the centralized rate. If we ignore logarithmic factors, this lower bound is achievable by a simple procedure: each machine computes the sample mean of its local data and quantizes each coordinate to precision σ^2/n (truncating if the sample mean is outside the region $[-1 - \sigma/\sqrt{n}, 1 + \sigma/\sqrt{n}]$), which requires $\mathcal{O}(d \log(n/\sigma^2))$ bits. These quantized sample averages are communicated to the fusion center using $B = \mathcal{O}(dm \log(n/\sigma^2))$ total bits. The fusion center averages them, obtaining an estimate with mean-squared error of optimal order $\sigma^2 d/(mn)$ as required.

The techniques we develop also apply to other families of probability distributions, and we finish this section by presenting a result that gives lower bounds that are sharp to numerical constant prefactors. In particular, we consider mean estimation for the family $\mathcal{P}_d$ of distributions supported on the compact set $[-1, 1]^d$, which include (for example) Bernoulli $\{\pm 1\}$-valued random variables, among others.

Proposition 10.2. *Assume that each of m machines receives a single observation ($n = 1$) from a distribution in $\mathcal{P}_d$. There exists a universal (numerical) constant c such that*

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}_d, B_{1:m}) \geq c\,\frac{d}{m} \min\left\{m, \frac{m}{\sum_{i=1}^m \min\{1, \frac{B_i}{d}\}}\right\}, \tag{10.10}$$

where B_i is the budget for machine i.

See Section 10.9.1 for a proof.

The standard minimax rate for d-dimensional mean estimation on $\mathcal{P}_d$ scales as d/m, which is achieved by the sample mean. The lower bound (10.10) shows that to achieve this scaling, we must have $\sum_{i=1}^m \min\{1, \frac{B_i}{d}\} \gtrsim m$, showing that each machine must send $B_i \gtrsim d$ bits. In addition, a simple scheme achieves this lower bound, so we describe it here. Suppose that machine i receives a d-dimensional vector $X_i \in [-1, 1]^d$. Based on X_i, it generates a Bernoulli random vector $Z_i = (Z_{i1}, \ldots, Z_{id})$ with $Z_{ij} \in \{0, 1\}$ taking the value 1 with

probability $(1 + X_{ij})/2$, independently across coordinates. Machine i uses d bits to send the vector $Z_i \in \{0,1\}^d$ to the fusion center. The fusion center then computes the average $\widehat{\theta} = \frac{1}{m}\sum_{i=1}^m (2Z_i - 1)$. This average is unbiased, and its expected squared error is bounded by d/m. We note in passing that for both the normal location family of Theorem 10.1 and the simpler bounded single observation model in Proposition 10.2, there is an exponential gap between the information required to describe the problem to the minimax mean squared error of $\frac{d}{mn}$—scaling as as $\mathcal{O}(d\log(mn))$—and the number of bits that must be communicated, which scales nearly linearly in m. See also our discussion following Theorem 10.2.

10.4.3 Interactive protocols in multi-machine settings

Having provided results on mean estimation in the non-interactive setting, we now turn to the substantially harder setting of distributed statistical inference where feedback is allowed on the channels. As described in the introduction problem setup, we allow a substantial amount of communication: there exists a public blackboard upon which every message sent to the fusion is stored (i.e. freely broadcast to all other nodes in the network). This makes providing lower bounds on communication substantially more challenging, but also (in some cases) allows somewhat more powerful algorithms.

We begin by considering the uniform location family $\mathcal{U}_d = \{P_\theta,\ \theta \in [-1,1]^d\}$, where P_θ is the uniform distribution on the rectangle $[\theta_1 - 1, \theta_1 + 1] \times \cdots \times [\theta_d - 1, \theta_d + 1]$. For this problem, a direct application of Proposition 10.1 gives a nearly sharp result.

Proposition 10.3. *Consider the uniform location family $\mathcal{U}_d$ with n i.i.d. observations per machine:*

(a) There are universal (numerical) constants $c_1, c_2 > 0$ such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{U}, B) \geq c_1 \max\left\{\exp\left(-c_2 \frac{B}{d}\right), \frac{d}{(mn)^2}\right\}.$$

(b) Conversely, given a budget of $B = d\left[2\log_2(2mn) + \log(m)(\lceil \log_2 d \rceil + 2\log_2(2mn))\right]$ bits, there is a universal constant c such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{U}, B) \leq c\frac{d}{(mn)^2}.$$

If each of the m machines receives n observations, we have a total sample size of mn, so the minimax rate over all centralized procedures scales as $d/(mn)^2$ (for instance, see [118]). Consequently, Proposition 10.3(b) shows that the number of bits required to achieve the centralized rate has only *logarithmic* dependence on the number m of machines. Part (a) shows that this logarithmic dependence on m is unavoidable: at least $B \gtrsim d\log(mn)$ bits are necessary to attain the optimal rate of $d/(mn)^2$.

Proof We prove Proposition 10.3 in two parts: the upper bound (part (b)) by exhibiting an interactive protocol Π^* and the lower bound (part (a)) by applying Proposition 10.1.

Upper bound on the minimax risk: We consider the following communication protocol $\Pi^* \in \mathcal{A}_{\text{inter}}(B, \mathcal{P})$:

1. Machine $i \in \{1, \ldots, m\}$ computes its local minimum $a_j^{(i)} = \min\{X_j^{(i,k)} : k = 1, \ldots, n\}$ for each coordinate $j \in [d]$.

2. Machine 1 broadcasts the vector $a^{(1)}$, where each of its components is quantized to accuracy $(mn)^{-2}$ in $[-2, 2]$, using $2d \log(2mn)$ bits. Upon receiving the broadcast, all machines initialize global minimum variables $s_j \leftarrow a_j^{(1)}$ for $j = 1, \ldots, d$.

3. In the order $i = 2, 3, \ldots, m$, machine i performs the following operations:

   (i) Find all indices j such that $a_j^{(i)} < s_j$, call them J_i. For each $j \in J_i$, machine i updates $s_j \leftarrow a_j^{(i)}$, and then broadcasts the list of indices J_i (which requires $|J_i| \lceil \log d \rceil$ bits) and the associated values s_j, using a total of $|J_i| \lceil \log d \rceil + 2|J_i| \log(2mn)$ bits.

   (ii) All other machines update their local vectors s after receiving machine i's update.

4. One machine outputs $\widehat{\theta} = s + 1$.

Using the protocol Π^* above, it is clear that for each $j \in [d]$ we have computed a global minimum

$$s_j = \min\left\{X_j^{(i,k)} \mid i \in [m], k \in [n]\right\}$$

to within accuracy $1/(mn)^2$ (because of quantization). As a consequence, classical convergence analyses (e.g. [118]) yield that the estimator $\widehat{\theta} = s + 1$ achieves the minimax optimal convergence rate $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \leq Cd/(mn)^2$, where C is a numerical constant.

It remains to understand the communication complexity of the protocol Π^*. To do so, we study steps 2 and 3. In Step 2, machine 1 sends a $2d \log(2mn)$-bit message as Y_1. In Step 3, machine i sends $|J_i|(\lceil \log d \rceil + 2 \log(2mn))$ bits, that is,

$$\sum_{j=1}^{d} \mathbf{1}\left\{a_j^{(i)} < \min\{a_j^{(1)}, \ldots, a_j^{(i-1)}\}\right\} (\lceil \log d \rceil + 2 \log(2mn))$$

bits, as no message is sent for index j if $a_j^{(i)} \geq \min\{a_j^{(1)}, \ldots, a_j^{(i-1)}\}$. This event happens with probability bounded by $1/i$, so we find that the expected length of message Y_i is

$$\mathbb{E}[L_i] \leq \frac{d(\lceil \log d \rceil + 2 \log(2mn))}{i}.$$

Putting all pieces together, we obtain that

$$\begin{aligned}\mathbb{E}[L] = \sum_{i=1}^{m} \mathbb{E}[L_i] &\leq 2d \log(2mn) + \sum_{i=2}^{m} \frac{d(\lceil \log d \rceil + 2 \log(2mn))}{i} \\ &\leq d\left[2 \log(2mn) + \ln(m)(\lceil \log d \rceil + 2 \log(2mn))\right].\end{aligned}$$

Lower bound on the minimax risk: To prove the lower bound, we simply evaluate packing entropies by using a volume argument [13]. Since $\Theta = [-1,1]^d$, the size of a maximal 2δ-packing can be lower bounded by

$$M_\Theta(2\delta) \geq \frac{\text{Volume}(\Theta)}{\text{Volume}(\{x \in \mathbb{R}^d : \|x\|_2 \leq 2\delta\})} \geq \left(\frac{1}{2\delta}\right)^d. \tag{10.11}$$

Taking logarithms and inverting $B = \log_2 M_\Theta(1/(mn))$ yields the lower bound. □

It is natural to wonder whether such logarithmic dependence holds more generally. The following result shows that it does not: for some problems, the dependence on m must be (nearly) linear. In particular, we reconsider estimation in the normal location family model (10.8), showing a lower bound that is nearly identical to that of Theorem 10.1. We prove Theorem 10.1 in Section 10.10.

Theorem 10.2. *For $i = 1, \ldots, m$, assume that each machine receives an i.i.d. sample of size n from a normal location model* (10.8) *and that there is a total communication budget B. Then there exists a universal (numerical) constant c such that*

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{N}_d, B) \geq c\frac{\sigma^2 d}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{(B/d+1)\log m} \vee 1\right\}. \tag{10.12}$$

Theorem 10.2 provides a somewhat weaker lower bound than the non-interactive case we present in Theorem 10.1. In particular, the lower bound (10.12) shows that at least $B = \Omega\big(\frac{dm}{\log m}\big)$ bits are required for any decentralized procedure—even allowing fully interactive communication—to attain the (centralized) minimax rate of convergence $\frac{\sigma^2 d}{mn}$. That is, to achieve an order-optimal mean-squared error, the total number of bits communicated must (nearly) scale with the product of the dimension d and number of machines m. This is somewhat weaker than the bound in Theorem 10.1, which shows that each machine individually must communicate at least $d/\log m$ bits, while the present bound requires only that the total number of bits be $md/\log m$.

Theorems 10.1 and 10.2 show that there is an *exponential* gap between the "information" content of the estimation problem and what must be communicated. More specifically, assuming (for simplicity) that $\sigma^2 = 1$, describing a solution of the normal mean estimation problem to accuracy $d/(mn)$ in squared ℓ_2-error requires at most $\mathcal{O}(d\log(mn))$ bits; Theorems 10.1 and 10.2 show that nearly dm bits must be communicated. This type of scaling—that the amount of communication must grow linearly in m—is dramatically different than the logarithmic scaling for the uniform family. This scaling is disctinct from other familiar source coding scenarios; in Slepian-Wolf coding, for example, it is possible to have a communication rate at the joint entropy rate of the sequences being communicated, while here, this is impossible (admittedly, we are working in a fairly different type of one-shot regime). Establishing sharp communication-based lower bounds thus requires careful study of the underlying family of distributions.

For both Theorems 10.2 and 10.1, there are logarithmic gaps in the amount of communication the minimax lower bound requires and that of the procedures we propose (quantize and communicate). It will be quite interesting if it is possible to make these gaps tighter, though we leave such questions for further work. It would also be interesting if the interactive setting for the Gaussian location family, while requiring the nearly linear $\Omega(\frac{dm}{\log m})$ bits of communication, was still asymptotically smaller than the non-interactive case presented in Theorem 10.1.

10.5 Consequences for regression

Having identified (to within logarithmic factors) the minimax rates of convergence for several mean estimation problems, we now show how they imply lower bounds on the communication-constrained minimax rates for other, more complex estimation problems. In particular, we focus on two standard, but important, linear models [93]: linear regression and probit regression.

10.5.1 Linear regression

We consider a distributed instantiation of linear regression with fixed design matrices. Concretely, suppose that each of m machines has stored a fixed design matrix $A^{(i)} \in \mathbb{R}^{n \times d}$ and then observes a response vector $b^{(i)} \in \mathbb{R}^d$ from the standard linear regression model

$$b^{(i)} = A^{(i)}\theta + \varepsilon^{(i)}, \tag{10.13}$$

where $\varepsilon^{(i)} \sim \mathsf{N}(0, \sigma^2 I_{n \times n})$ is a noise vector. Our goal is to estimate the unknown regression vector $\theta \in \theta = [-1, 1]^d$, identical for each machine, in a distributed manner. To state our result, we assume uniform upper and lower bounds on the eigenvalues of the rescaled design matrices, namely

$$0 < \lambda_{\min}^2 \leq \min_{i \in \{1,\ldots,m\}} \frac{\gamma_{\min}((A^{(i)})^\top A^{(i)})}{n} \quad \text{and} \quad \max_{i \in \{1,\ldots,m\}} \frac{\gamma_{\max}((A^{(i)})^\top A^{(i)})}{n} \leq \lambda_{\max}^2. \tag{10.14}$$

Corollary 10.1. *Consider an instance of the linear regression model* (10.13) *under condition* (10.14).

(a) Then there is a universal positive constant c such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B_{1:m}) \geq c \frac{\sigma^2 d}{\lambda_{\max}^2 mn} \min\left\{ \frac{\lambda_{\max}^2 mn}{\sigma^2}, \frac{m}{(B/d + 1)\log m} \right\}.$$

(b) Conversely, given total budget $B \geq dm \log(mn)$, there is a universal constant c' such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B_{1:m}) \leq \frac{c'}{\lambda_{\min}^2} \frac{\sigma^2 d}{mn}.$$

It is a classical fact (e.g. [118]) that the minimax rate for d-dimensional linear regression scales as $d\sigma^2/(nm)$. Part (a) of Corollary 10.1 shows this optimal rate is attainable only if the total budget B grows as $\frac{dm}{\log m}$. Part (b) of the corollary shows that the minimax rate is achievable with budgets that match the lower bound up to logarithmic factors.

Proof The proof of part (b) follows from Chapter 9 (the convergence guarantee (9.9)), which shows that solving each regression problem separately, quantizing the (local) solution vectors $\widehat{\theta}^{(i)} \in [-1,1]^d$ to accuracy $1/(mn)$ using $B_i = d\log(mn)$ bits, and then performing averaging achieves the minimax rate up to constant prefactors.

To prove part (a), we show that solving an arbitrary Gaussian mean estimation problem can be reduced to solving a specially constructed linear regression problem. This reduction allows us to apply the lower bound from Theorem 10.1. Given $\theta \in \Theta$, consider the Gaussian mean model

$$X^{(i)} = \theta + w^{(i)}, \quad \text{where} \quad w^{(i)} \sim \mathsf{N}\left(0, \frac{\sigma^2}{\lambda_{\max}^2 n} I_{d\times d}\right).$$

Each machine i has its own design matrix $A^{(i)}$, and we use it to construct a response vector $b^{(i)} \in \mathbb{R}^n$. Since $\gamma_{\max}(A^{(i)}/\sqrt{n}) \le \lambda_{\max}$, the matrix $\Sigma^{(i)} := \sigma^2 I_{n\times n} - \frac{\sigma^2}{\lambda_{\max}^2 n} A^{(i)}(A^{(i)})^\top$ is positive semidefinite. Consequently, we may form a response vector via

$$b^{(i)} = A^{(i)}X^{(i)} + z^{(i)} = A^{(i)}\theta + A^{(i)}w^{(i)} + z^{(i)}, \qquad z^{(i)} \sim \mathsf{N}\left(0, \Sigma^{(i)}\right) \text{ independent of } w^{(i)}. \tag{10.15}$$

The independence of $w^{(i)}$ and $z^{(i)}$ guarantees that $b^{(i)} \sim \mathsf{N}(A^{(i)}\theta, \sigma^2 I_{n\times n})$, so the pair $(b^{(i)}, A^{(i)})$ is faithful to the regression model (10.13).

Now consider any protocol $\Pi \in \mathcal{A}_{\text{inter}}(B, \mathcal{P})$ that can solve any regression problem to within accuracy δ, so that $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \le \delta^2$. By the previously described reduction, the protocol Π can also solve the mean estimation problem to accuracy δ, in particular via the pair $(A^{(i)}, b^{(i)})$ described in expression (10.15). Combined with this reduction, the corollary thus follows from Theorem 10.1. □

10.5.2 Probit regression

We now turn to the problem of binary classification, in particular considering the probit regression model. As in the previous section, each of m machines has a fixed design matrix $A^{(i)} \in \mathbb{R}^{n\times d}$, where $A^{(i,k)}$ denotes the kth row of $A^{(i)}$. Machine i receives n binary responses $Z^{(i)} = (Z^{(i,1)}, \ldots, Z^{(i,n)})$, drawn from the conditional distribution

$$\mathbb{P}(Z^{(i,k)} = 1 \mid A^{(i,k)}, \theta) = \Phi(A^{(i,k)}\theta) \quad \text{for some fixed } \theta \in \theta = [-1,1]^d, \tag{10.16}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. The log-likelihood of the probit model (10.16) is concave [32, Exercise 3.54]. Under condition (10.14) on the design matrices, we have:

Corollary 10.2. *Consider the probit model* (10.16) *under condition* (10.14)*. Then*

(a) There is a universal constant c such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B_{1:m}) \geq c \frac{d}{\lambda_{\max}^2 mn} \min \left\{ \lambda_{\max}^2 mn, \frac{m}{(B/d + 1) \log m} \right\}.$$

(b) Conversely, given total budget $B \geq dm \log(mn)$, there is a universal constant c' such that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \leq \frac{c'}{\lambda_{\min}^2} \frac{d}{mn}.$$

Proof As in the previous case with linear regression, the results of Chapter 9 give part (b): each machine solves the local probit regression separately, quantizes its local solution to accuracy $1/mn$ using $B_i = d \log(mn)$ bits, after which the fusion center averages all the quantized local solutions.

To prove part (a), we show that linear regression problems can be solved via estimation in a specially constructed probit model. Consider an arbitrary $\theta \in \Theta$; assume we have a regression problem of the form (10.13) with noise variance $\sigma^2 = 1$. We construct the binary responses for our probit regression $(Z^{(i,1)}, \ldots, Z^{(i,n)})$ by

$$Z^{(i,k)} = \begin{cases} 1 & \text{if } b^{(i,k)} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10.17}$$

By construction, we have $\mathbb{P}(Z^{(i,k)} = 1 \mid A^{(i)}, \theta) = \Phi(A^{(i,k)}\theta)$ as desired for our model (10.16). By inspection, any protocol $\Pi \in \mathcal{A}_{\text{inter}}(B, \mathcal{P})$ solving the probit regression problem provides an estimator with the same mean-squared error as the original linear regression problem via the construction (10.17). Corollary 10.1 provides the desired lower bound. □

10.6 Summary

In this chapter, we have developed several results showing fundamental bounds on the amount of communication required for several statistical estimation problems. In particular, we have shown that—even when broadcasts from a fusion center to all nodes in a network are free—estimation of the mean in a d-dimensional normal location model with data across m machines requires communicating at least $\Omega(dm/\log m)$ bits. Several open questions remain. First, our arguments are somewhat complex; simplifying them could lead to much wider applicability of results of this form. Second, our data processing inequalities, those inequalities of the form (10.20), build off of likelihood ratio bounds similar to those of Chapter 7, but we require strong independence assumptions in the chains $V \to X \to Y$.

In fact, these necessitate use of the variant Fano inequalities developed in Section 2.2.3. In particular, our random vectors V must have independent coordinates, and we similarly require the vectors X to have independent coordinates. In standard "packing" constructions for lower bounds, however (e.g. [185, 148, 9, 37]), it seems difficult to construct vectors with independent coordinates, for example, in high-dimensional settings in which the "true" mean vectors θ are sparse [148, 9]. If we could obtain data processing inequalities that were in some way less dependent on the particular structure of the problems we solve, this might yield broader insights into the interaction of communication, computational, statistical, and geometric conditions underlying distributed inference problems.

10.7 Proof outline of major results

Having stated each of our main results, in this section we outline the major steps in developing the lower bounds—converse inequalities—we establish for distributed estimation problems. Our lower bounds follow the basic strategy introduced in Chapter 2: we reduce the estimation problem to a testing problem, and following this reduction, we use the distance-based Fano method described in Corollary 2.1 to relate the probability of error in the test to the number of bits contained in the messages Y_i sent from each machine. Establishing these links is the most technically challenging aspect of our results.

We now describe the setting for our reduction. Let $\mathcal{V}$ denote an index set of finite cardinality, where $v \in \mathcal{V}$ indexes a family of probability distributions $\{P(\cdot \mid v)\}_{v \in \mathcal{V}}$. To each member of this family we associate a parameter $\theta_v := \theta(P(\cdot \mid v)) \in \Theta$, where Θ denotes the parameter space. In our proofs applicable to d-dimensional problems, we set $\mathcal{V} = \{-1, 1\}^d$, and we index vectors θ_v by $v \in \mathcal{V}$. Now, we sample V uniformly at random from $\mathcal{V}$. Conditional on $V = v$, we then sample X from a distribution $P_X(\cdot \mid V = v)$ satisfying $\theta_v := \theta(P_X(\cdot \mid v)) = \delta v$, where $\delta > 0$ is a fixed quantity that we control. We define $d_{ham}(v, v')$ to be the Hamming distance between $v, v' \in \mathcal{V}$. This construction gives

$$\|\theta_v - \theta_{v'}\|_2 = 2\delta \sqrt{d_{\rm ham}(v, v')}.$$

Then for fixed $t \in \mathbb{R}$, Corollary 2.2 (via the separation function (2.14)) implies that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\widehat{\theta} - \theta(P)\|_2^2 \right] \geq \delta^2 \left(\lfloor t \rfloor + 1 \right) \left[1 - \frac{I(V; Y) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}} \right], \tag{10.18}$$

where $N_t^{\max} = \max_{v \in \mathcal{V}} |\{v' \in \mathcal{V} : d_{\rm ham}(v, v') \leq t\}|$ is the size of the largest t-neighborhood in $\mathcal{V}$. The lower bound involves the information $I(V; Y)$ because our distributed protocol enforces that the estimator $\widehat{\theta}$ may observe only Y rather than the sample X off of which it is based, and hence we have the Markov chain $V \to X \to Y$. As noted in our discussion of Corollaries 2.1 and 2.2 in our explanation of Fano-type methods for minimax lower bounds in Section 2.2.3, inequality (10.18) allows flexibility in its application. If there is a large set

$\mathcal{V}$ for which it is easy to control $I(V; X)$ while neighborhoods in $\mathcal{V}$ are relatively small (i.e., $N_t^{\max}$ is small), we can obtain sharp lower bounds.

Now we show how to apply inequality (10.18) in our minimax bounds; these calculations parallel those for the lower bound for the normal regression model (2.16) in Section 2.2.3. First, with the choice $\mathcal{V} = \{-1, 1\}^d$ and the Hamming metric $d_{\hamm}$, for $0 \le t \le \lceil d/3 \rceil$, we have $N_t^{\max} = \sum_{\tau=0}^{t} \binom{d}{\tau} \le 2\binom{d}{t}$. Since $\binom{d}{t} \le (de/t)^t$, for $t \le d/6$ we have

$$\log \frac{|\mathcal{V}|}{N_t^{\max}} \ge d \log 2 - \log 2 \binom{d}{t} \ge d \log 2 - \frac{d}{6} \log(6e) - \log 2 = d \log \frac{2}{2^{1/d} \sqrt[6]{6e}} > \frac{d}{6}$$

for $d \ge 12$ (the case $d < 12$ can be checked directly). Substituting this into inequality (10.18), we find that for $t = \lfloor d/6 \rfloor$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[\|\widehat{\theta}(Y) - \theta(P)\|_2^2\right] \ge \delta^2(\lfloor d/6 \rfloor + 1) \left(1 - \frac{I(Y; V) + \log 2}{d/6}\right). \tag{10.19}$$

Inequality (10.19) is the essential point of departure for the proofs of our major results. Using the inequality, it remains to upper bound the mutual information $I(Y; V)$, which is our main technical difficulty. At a very high level, our results give sharper characterizations of the mutual information between the random variable V and each machine's message Y_i. For most scenarios, we show (roughly) that there exists a problem-dependent constant κ such that

$$I(V; Y_i) \le \kappa \delta^2 I(X^{(i)}; Y_i). \tag{10.20}$$

We prove such quantitative data processing inequalities using techniques similar to those of Chapters 7 and 8, where we provide similar data-processing inequalities based on likelihood ratio bounds.

Because the random variable Y_i takes discrete values, we have $I(X^{(i)}; Y_i) \le H(Y_i) \le B_i$ by Shannon's source coding theorem [47] (recall that B_i is the communication budget on machine i). In particular, inequality (10.20) establishes the inequality $I(V; Y_i) \le \kappa \delta^2 B_i$. For independent communication schemes, $I(V; Y_{1:m}) \le \sum_{i=1}^m I(V; Y_i)$, whence we have the simplification

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \ge \delta^2(\lfloor d/6 \rfloor + 1) \left(1 - \frac{\kappa \delta^2 \sum_{i=1}^m B_i + \log 2}{d/6}\right).$$

Thus, by choosing $\delta^2 = c \min\{1, d/(\kappa \sum_{i=1}^m B_i)\}$ for an appropriate numerical constant, we see that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \ge c' \delta^2(\lfloor d/6 \rfloor + 1) = cc' d \min\left\{1, \frac{d}{\kappa \sum_{i=1}^m B_i}\right\}$$

for numerical constants c, c'. This then implies that the sum of the communication budgets B_i must be sufficiently large to allow small estimation error. We make these calculations more explicit in the sections to follow.

Outline of proofs In the coming sections, we provide the proofs of all our major results. Before presenting our results, however, we give a brief outline of the remainder of the chapter, as we do not prove the results completely in their order of presentation in the text: they build on one another, so we present them in (rough) order of most basic to most complex. In the first section, Section 10.8, we provide a few techniques that are useful throughout our results. Section 10.9 begins the proofs of our major (multi-machine) lower bounds by proving our results on independent protocols, which lay the groundwork and develop most of our major techniques, which also prove useful in the interactive case. Section 10.9.1 contains the proof of Proposition 10.2, the simplest of our major (multi-machine) lower bounds, while we prove Theorem 10.1 in Section 10.9.3. We prove Theorem 10.2 in Section 10.10.

Notation For our proofs, we require a bit of additional notation. For a random variable X, we let P_X denote the probability measure on X, so that $P_X(S) = P(X \in S)$, and we abuse notation by writing p_X for the probability mass function or density of X, depending on the situation, so that $p_X(x) = P(X = x)$ in the discrete case and denotes the density of X at x when p_X is a density.

10.8 Techniques, tools, and setup for proofs

In this section, we provide a bit more setup for the proofs of Proposition 10.2 and Theorems 10.1 and 10.2. We begin by reviewing a few of the basic techniques for minimax bounds from Chapter 2 that are essential for our results, and we also state an important technical lemma, paralleling Lemma 8.1 (and similar inequalities from the proofs of Theorems 7.1, 7.2, and 7.3 from Chapter 8).

10.8.1 Common techniques

Le Cam's method

In low-dimensional settings (when the dimension d is small), it is difficult to apply the incarnation Fano's inequality we outline in Section 10.7. In such settings, we use the two-point lower bound technique of Le Cam's method from Section 2.2.2. By the basic minimax bound (2.8), we see that if $\mathcal{V} = \{-1, 1\}$ and $\theta_v = \theta(P_v)$, then if the pair $\{\theta_v\}$ is 2δ-separated,

$$\max_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\|\widehat{\theta}(Y) - \theta_v\|_2^2\right] \geq \delta^2 \left(\frac{1}{2} - \frac{1}{2}\left\|P_Y(\cdot \mid V = 1) - P_Y(\cdot \mid V = -1)\right\|_{\mathrm{TV}}\right).$$

Here, as usual, we assume that V is uniform on $\mathcal{V}$ and we have the Markov chain $V \to X \to Y$, where Y is the message available to the estimator $\widehat{\theta}$. We claim this inequality implies

$$\max_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\|\widehat{\theta}(Y) - \theta_v\|_2^2\right] \geq \delta^2 \left(\frac{1}{2} - \frac{1}{\sqrt{2}}\sqrt{I(Y; V)}\right). \tag{10.21}$$

It is clear that inequality (10.21) will hold if we can show the following: for any pair of random variables $V \to Y$, if V is chosen uniformly in a set $\mathcal{V} = \{v, v'\}$, then

$$\|P_Y(\cdot \mid V = v) - P_Y(\cdot \mid V = v')\|_{\mathrm{TV}}^2 \le 2I(Y, V). \tag{10.22}$$

To see inequality (10.22), let P_v be shorthand for $P_Y(\cdot \mid V = v)$. The triangle inequality implies that

$$\|P_v - P_{v'}\|_{\mathrm{TV}} \le \|P_v - (1/2)(P_v + P_{v'})\|_{\mathrm{TV}} + \frac{1}{2}\|P_v - P_{v'}\|_{\mathrm{TV}},$$

and by swapping the roles of v' and v, we obtain

$$\|P_v - P_{v'}\|_{\mathrm{TV}} \le 2\min\{\|P_v - (1/2)(P_{v'} + P_v)\|_{\mathrm{TV}}, \|P_{v'} - (1/2)(P_{v'} + P_v)\|_{\mathrm{TV}}\}.$$

By Pinsker's inequality, we thus have the upper bound

$$\begin{aligned}\|P_v - P_{v'}\|_{\mathrm{TV}}^2 &\le 2\min\{D_{\mathrm{kl}}\left(P_v \| (1/2)(P_v + P_{v'})\right), D_{\mathrm{kl}}\left(P_{v'} \| (1/2)(P_v + P_{v'})\right)\} \\ &\le D_{\mathrm{kl}}\left(P_v \| (1/2)(P_v + P_{v'})\right) + D_{\mathrm{kl}}\left(P_{v'} \| (1/2)(P_v + P_{v'})\right) = 2I(Y; V).\end{aligned}$$

Tensorization of information

We also require a type of tensorization inequality in each of our proofs for independent protocols. When Y_i is constructed based only on $X^{(i)}$, we have

$$\begin{aligned}I(V; Y_{1:m}) = \sum_{i=1}^m I(V; Y_i \mid Y_{1:i-1}) &= \sum_{i=1}^m H(Y_i \mid Y_{1:i-1}) - H(Y_i \mid V, Y_{1:i-1}) \\ &\le \sum_{i=1}^m H(Y_i) - H(Y_i \mid V, Y_{1:i-1}) \\ &= \sum_{i=1}^m H(Y_i) - H(Y_i \mid V) = \sum_{i=1}^m I(V; Y_i)\end{aligned} \tag{10.23}$$

where we have used that conditioning reduces entropy and Y_i is conditionally independent of $Y_{1:i-1}$ given V.

10.8.2 Total variation contraction

Our results rely on certain data processing inequalities—contractions of mutual information and other divergences—inspired by results on information contraction under privacy constraints we developed in Chapters 7 and 8. Consider four random variables A, B, C, D, of which we assume that A, C, and D have discrete distributions. We denote the conditional distribution of A given B by $P_{A|B}$ and their full joint distribution by $P_{A,B,C,D}$. We assume

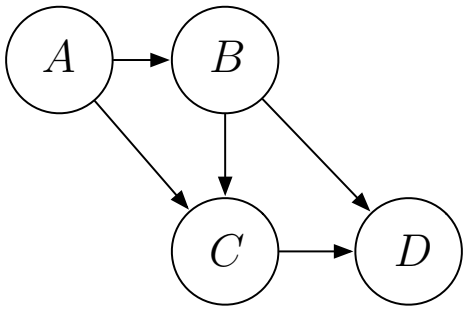

Figure 10.1: Graphical model for Lemma 10.1

that the random variables have conditional indpendence structure specified by the graphical model in Figure 10.1, that is, that we can write the joint distribution as the product

$$P_{A,B,C,D} = P_A P_{B|A} P_{C|A,B} P_{D|B,C}. \tag{10.24}$$

We denote the domain of a random variable by the identical calligraphic letter, so $A \in \mathcal{A}$, $B \in \mathcal{B}$, and so on. We write $\sigma(\mathcal{A})$ for the sigma-field on $\mathcal{A}$ with respect to which our measures are defined. Sometimes we write $P_A(\cdot \mid B)$ for the conditional distribution of A given B. In addition to the conditional independence assumption (10.24), we assume that the conditional distribution of C given A, B factorizes in the following specific form. There exist functions $\Psi_1 : \mathcal{A} \times \sigma(\mathcal{C}) \to \mathbb{R}_+$ and $\Psi_2 : \mathcal{B} \times \sigma(\mathcal{C}) \to \mathbb{R}_+$ such that for any (measureable) set S in the range $\mathcal{C}$ of C, we have

$$P_C(S \mid A, B) = \Psi_1(A, S)\Psi_2(B, S). \tag{10.25}$$

Since C is assumed discrete, we abuse notation and write $P(C = c \mid A, B) = \Psi_1(A, c)\Psi_2(B, c)$. Lastly, we assume that for any $a, a' \in \mathcal{A}$, we have the following likelihood ratio bound:

$$\sup_{S \in \sigma(\mathcal{B})} \frac{P_B(S \mid A = a)}{P_B(S \mid A = a')} \leq \exp(\alpha). \tag{10.26}$$

Lemma 10.1. *Under assumptions* (10.24)*,* (10.25)*, and* (10.26)*, the following inequality holds:*

$$\begin{aligned} &|P(A = a \mid C, D) - P(A = a \mid C)| \\ &\qquad \leq 2\left(e^{2\alpha} - 1\right) \min\left\{P(A = a \mid C), P(A = a \mid C, D)\right\} \left\|P_B(\cdot \mid C, D) - P_B(\cdot \mid C)\right\|_{\mathrm{TV}}. \end{aligned}$$

Proof By assumption, A is independent of D given $\{B, C\}$. Thus we may write

$$P(A = a \mid C, D) - P(A = a \mid C) = \int P(A = a \mid B = b, C) \left(dP_B(b \mid C, D) - dP_B(b \mid C)\right)$$

Combining this equation with the inequality

$$\int P(A = a \mid C) \left(dP_B(b \mid C, D) - dP_B(b \mid C) \right) = 0$$

we find that

$$\begin{aligned} & P(A = a \mid C, D) - P(A = a \mid C) \\ & = \int \left(P(A = a \mid B = b, C) - P(A = a \mid C) \right) \left(dP_B(b \mid C, D) - dP_B(b \mid C) \right). \end{aligned}$$

Using the fact that $|\int f(b) d\mu(b)| \leq \sup_b\{|f(b)|\} \int |d\mu(b)|$ for any signed measure μ on $\mathcal{B}$, we conclude from the previous equality that for *any* version $P_A(\cdot \mid B, C)$ of the conditional probability of A given $\{B, C\}$ that since $\int |d\mu| = \|\mu\|_{\mathrm{TV}}$,

$$\begin{aligned} & |P(A = a \mid C, D) - P(A = a \mid C)| \\ & \quad \leq 2 \sup_{b \in \mathcal{B}} \{ |P(A = a \mid B = b, C) - P(A = a \mid C)| \} \left\| P_B(\cdot \mid C, D) - P_B(\cdot \mid C) \right\|_{\mathrm{TV}}. \end{aligned}$$

Thus, to prove the lemma, it is sufficient to show (for some version of the conditional distribution[2] $P_A(\cdot \mid B, C)$) that for any $b \in \mathcal{B}$

$$|P(A = a \mid B = b, C) - P(A = a \mid C)| \leq (e^{2\alpha} - 1) \min\{P(A = a \mid C), P(A = a \mid C, D)\}. \tag{10.27}$$

To prove this upper bound, we consider the joint distribution (10.24) and likelihood ratio bound (10.26). The distributions $\{P_B(\cdot \mid A = a)\}_{a \in \mathcal{A}}$ are all absolutely continuous with respect to one another by assumption (10.26), so it is no loss of generality to assume that there exists a density $p_B(\cdot \mid A = a)$ for which $P(B \in S \mid A = a) = \int p_B(b \mid A = a) d\mu(b)$, for some fixed measure μ, and for which the ratio $p_B(b \mid A = a)/p_B(b \mid A = a') \in [e^{-\alpha}, e^{\alpha}]$ for all b. By elementary conditioning we have for any $S_b \in \sigma(\mathcal{B})$ and $c \in \mathcal{C}$

$$\begin{aligned} & P(A = a \mid B \in S_b, C = c) \\ & \quad = \frac{P(A = a, B \in S_b, C = c)}{P(B \in S_b, C = c)} \\ & \quad = \frac{P(B \in S_b, C = c \mid A = a) P(A = a)}{\sum_{a' \in \mathcal{A}} P(A = a') P(B \in S_b, C = c \mid A = a')} \\ & \quad = \frac{P(A = a) \int_{S_b} P(C = c \mid B = b, A = a) p_B(b \mid A = a) d\mu(b)}{\sum_{a' \in \mathcal{A}} P(A = a') \int_{S_b} P(C = c \mid B = b, A = a') p_B(b \mid A = a') d\mu(b)}, \end{aligned}$$

where for the last equality we used the conditional independence assumptions (10.24). But now we recall the decomposition formula (10.25), and we can express the likelihood functions

[2] If $P(A = a \mid C)$ is undefined, we simply set it to have value 1 and assign $P(A = a \mid B, C) = 1$ as well.

by

$$P(A = a \mid B \in S_b, C = c) = \frac{P(A = a) \int_{S_b} \Psi_1(a, c)\Psi_2(b, c) p_B(b \mid A = a) d\mu(b)}{\sum_{a'} P(A = a') \int_{S_b} \Psi_1(a', c)\Psi_2(b, c) p_B(b \mid A = a') d\mu(b)}.$$

As a consequence, there is a version of the conditional distribution of A given B and C such that

$$P(A = a \mid B = b, C = c) = \frac{P(A = a)\Psi_1(a, c) p_B(b \mid A = a)}{\sum_{a'} P(A = a')\Psi_1(a', c) p_B(b \mid A = a')}. \tag{10.28}$$

Define the shorthand

$$\beta = \frac{P(A = a)\Psi_1(a, c)}{\sum_{a' \in \mathcal{A}} P(A = a')\Psi_1(a', c)}.$$

We claim that

$$e^{-\alpha}\beta \leq P(A = a \mid B = b, C = c) \leq e^{\alpha}\beta. \tag{10.29}$$

Assuming the correctness of bound (10.29), we establish inequality (10.27). Indeed, $P(A = a \mid C = c)$ is a weighted average of $P(A = a \mid B = b, C = c)$, so we also have the same upper and lower bound for $P(A = a \mid C)$, that is

$$e^{-\alpha}\beta \leq P(A = a \mid C) \leq e^{\alpha}\beta,$$

while the conditional independence assumption that A is independent of D given B, C (recall Figure 10.1 and the product (10.24)) implies

$$\begin{aligned} P(A = a \mid C = c, D = d) &= \int_{\mathcal{B}} P(A = a \mid B = b, C = c, D = d) dP_B(b \mid C = c, D = d) \\ &= \int_{\mathcal{B}} P(A = a \mid B = b, C = c) dP_B(b \mid C = c, D = d), \end{aligned}$$

and the final integrand belongs to $\beta[e^{-\alpha}, e^{\alpha}]$. Combining the preceding three displayed expressions, we find that

$$\begin{aligned} |P(A = a \mid B = b, C) - P(A = a \mid C)| &\leq \left(e^{\alpha} - e^{-\alpha}\right) \beta \\ &\leq \left(e^{\alpha} - e^{-\alpha}\right) e^{\alpha} \min\left\{P(A = a \mid C), P(A = a \mid C, D)\right\}. \end{aligned}$$

This completes the proof of the upper bound (10.27).

It remains to prove inequality (10.29). We observe from expression (10.28) that

$$P(A = a \mid B = b, C) = \frac{P(A = a)\Psi_1(a, C)}{\sum_{a' \in \mathcal{A}} P(A = a')\Psi_1(a', C) \frac{p_B(b \mid A = a')}{p_B(b \mid A = a)}}.$$

By the likelihood ratio bound (10.26), we have $p_B(b \mid A = a')/p_B(b \mid A = a) \in [e^{-\alpha}, e^{\alpha}]$, and combining this with the above equation yields inequality (10.29). □

10.9 Proofs of lower bounds for independent protocols

10.9.1 Proof of Proposition 10.2

The proof of this proposition follows the basic outline described in Section 10.7.

We first describe the distribution of the step $V \to X$. Given $v \in \mathcal{V}$, we assume that each machine i receives a d-dimensional sample $X^{(i)}$ with coordinates independently sampled according to

$$P(X_j = v_j \mid v) = \frac{1 + \delta v_j}{2} \quad \text{and} \quad P(X_j = -v_j \mid v) = \frac{1 - \delta v_j}{2}.$$

Then $\theta_v = \mathbb{E}_v[X]$; to apply Lemma 10.1, we require the the likelihood ratio bound

$$\frac{P(X_j \in S \mid v)}{P(X_j \in S \mid v')} \le \frac{1+\delta}{1-\delta} = \exp\left(\log \frac{1+\delta}{1-\delta}\right)$$

We now present a lemma that relates this ratio bound to a quantitative data processing inequality. The lemma is somewhat more general than what we require, and we prove it in Section 10.9.2. The result is similar to the results in Theorems 7.1, 7.2, and 7.3 in the preceding chapters, which show similar strong data processing inequalities in the context of privacy-preserving data analysis. The current proof, however is different, as we have the Markov chain $V \to X \to Y$, and instead of a likelihood ratio bound on the channel $X \to Y$, we place a likelihood ratio bound on $V \to X$.

Lemma 10.2. *Let V be sampled uniformly at random from $\{-1, 1\}^d$. For any (i, j), assume that $X_j^{(i)}$ is independent of $\{X_{j'}^{(i)} : j' \neq j\} \cup \{V_{j'} : j' \neq j\}$ given V_j. Let $\mathbb{P}_{X_j}$ be the probability measure of $X_j^{(i)}$ and assume in addition that*

$$\sup_{S \in \sigma(X_j)} \frac{\mathbb{P}_{X_j}(S \mid V = v)}{\mathbb{P}_{X_j}(S \mid V = v')} \le \exp(\alpha).$$

Then

$$I(V; Y_i) \le 2(e^{2\alpha} - 1)^2 I(X^{(i)}; Y_i).$$

Lemma 10.2 provides a quantitative data processing inequality relating the mutual information in the channel $X^{(i)} \to Y_i$ to that in $V \to Y_i$. In particular, we find that

$$I(V; Y_i) \le 2\left(e^{2\log\frac{1+\delta}{1-\delta}} - 1\right)^2 I(X^{(i)}; Y_i) = 2\left(\frac{(1+\delta)^2}{(1-\delta)^2} - 1\right)^2 \le 80\delta^2 I(X^{(i)}; Y_i)$$

for $\delta \in [0, 1/5]$. Recalling our outline from Section 10.7, this is the claimed strong data processing inequality (10.20). Recalling the tensorization inequality (10.23), we also have

$$I(V; Y_{1:m}) \le \sum_{i=1}^m I(V; Y_i) \le 80\delta^2 \sum_{i=1}^m I(Y_i; X^{(i)}). \tag{10.30}$$

The remainder of the proof we break into two cases: when $d \geq 10$ and when $d < 10$. In either case, we have $\theta(P(\cdot \mid V = v)) = \mathbb{E}[X \mid v] = \delta v$, which controls the separation of the points θ_v. For the case $d \geq 10$, our proof sketch in Section 10.7, beginning from inequality (10.20) with $\kappa = 80$, essentially completes the proof. Since Y_i satisfies $H(Y_i) \leq B_i$ by Shannon's source coding theorem [47] and $H(X^{(i)}) \leq d$, we have $I(Y_i; X^{(i)}) \leq \min\{H(Y_i), H(X^{(i)})\} \leq \min\{B_i, d\}$. Thus we have the inequality

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \geq \delta^2(\lfloor d/6 \rfloor + 1)\left(1 - \frac{80\delta^2 \sum_{i=1}^m \min\{B_i, d\} + \log 2}{d/6}\right).$$

The choice $\delta^2 = \min\{1/25, d/960 \sum_{i=1}^m \min\{B_i, d\}\}$ guarantees that the expression inside parentheses in the previous display is lower bounded by $2/25$, which gives the proposition for $d \geq 10$.

When $d < 10$, we use a slightly different argument. By a reduction to a smaller dimensional problem, we may assume without loss of generality that $d = 1$, and we set $\mathcal{V} = \{-1, 1\}$. In this case, Le Cam's method (10.21) coupled with the subsequent information implies

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \geq \delta^2\left(\frac{1}{2} - \frac{1}{2}\sqrt{2I(V; Y_{1:m})}\right). \tag{10.31}$$

Applying the bound (10.30), that $I(V; Y_{1:m}) \leq 80\delta^2 \sum_{i=1}^m I(Y_i; X^{(i)})$, and noting that $I(X^{(i)}; Y_i) \leq \min\{1, H(Y_i)\}$ as $X^{(i)} \in \{-1, 1\}$, we obtain

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \geq \delta^2\left(\frac{1}{2} - 7\left(\delta^2 \sum_{i=1}^m \min\{1, H(Y_i)\}\right)^{\frac{1}{2}}\right).$$

Because $H(Y_i) \leq B_i$, setting

$$\delta^2 = \min\left\{1, \frac{1}{400 \sum_{i=1}^m \min\{1, B_i\}}\right\}$$

completes the proof.

10.9.2 Proof of Lemma 10.2

Let $Y = Y_i$; we suppress the dependence on the index i (and similarly let $X = X^{(i)}$ denote a single fixed sample). We begin with the observation that by the chain rule for mutual information,

$$I(V; Y) = \sum_{j=1}^d I(V_j; Y \mid V_{1:j-1}).$$

Using the definition of mutual information and non-negativity of the KL-divergence, we have

$$\begin{aligned} I(V_j; Y \mid V_{1:j-1}) &= \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\text{kl}}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right) \mid V_{1:j-1}\right]\right] \\ &\leq \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\text{kl}}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right)\right.\right. \\ &\quad \left.\left. + D_{\text{kl}}\left(P_{V_j}(\cdot \mid V_{1:j-1}) \| P_{V_j}(\cdot \mid Y, V_{1:j-1})\right) \mid V_{1:j-1}\right]\right]. \end{aligned}$$

Now, we require an argument that builds off of our technical Lemma 10.1. We claim that Lemma 10.1 implies that

$$\begin{aligned}
&|P(V_j = v_j \mid V_{1:j-1}, Y) - P(V_j = v_j \mid V_{1:j-1})| \\
&\quad \le 2(e^{2\alpha} - 1) \min \left\{P(V_j = v_j \mid V_{1:j-1}, Y), P(V_j = v_j \mid V_{1:j-1})\right\} \\
&\qquad \times \left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}. \qquad (10.32)
\end{aligned}$$

Indeed, making the identification

$$V_j \equiv A, \quad X_j \equiv B, \quad V_{1:j-1} \equiv C, \quad Y \equiv D$$

satisfies the condition (10.24) clearly, condition (10.25) because $V_{1:j-1}$ is independent of V_j and X_j, and condition (10.26) by construction. This gives inequality (10.32) by our independence assumptions. Expanding our KL divergence bound, we have

$$\begin{aligned}
&D_{\mathrm{kl}}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right) \\
&\quad \le \sum_{v_j} \left(P_{V_j}(v_j \mid Y, V_{1:j-1}) - P_{V_j}(v_j \mid V_{1:j-1})\right) \log \frac{P_{V_j}(v_j \mid Y, V_{1:j-1})}{P_{V_j}(v_j \mid V_{1:j-1})}.
\end{aligned}$$

Now, using the elementary inequality for $a, b \ge 0$ that

$$\left|\log \frac{a}{b}\right| \le \frac{|a - b|}{\min\{a, b\}},$$

inequality (10.32) implies that

$$\begin{aligned}
&\left(P_{V_j}(v_j \mid Y, V_{1:j-1}) - P_{V_j}(v_j \mid V_{1:j-1})\right) \log \frac{P_{V_j}(v_j \mid Y, V_{1:j-1})}{P_{V_j}(v_j \mid V_{1:j-1})} \\
&\le \frac{(P_{V_j}(v_j \mid Y, V_{1:j-1}) - P_{V_j}(v_j \mid V_{1:j-1}))^2}{\min\{P_{V_j}(v_j \mid Y, V_{1:j-1}), P_{V_j}(v_j \mid V_{1:j-1})\}} \\
&\le 4(e^{2\alpha} - 1)^2 \min\left\{P_{V_j}(v_j \mid Y, V_{1:j-1}), P_{V_j}(v_j \mid V_{1:j-1})\right\} \left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}^2.
\end{aligned}$$

Substituting this into our bound on KL-divergence, we obtain

$$\begin{aligned}
&I(V_j; Y \mid V_{1:j-1}) \\
&\quad = \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\mathrm{kl}}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right) \mid V_{1:j-1}\right]\right] \\
&\quad \le 4(e^{2\alpha} - 1)^2 \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[\left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}^2 \mid V_{1:j-1}\right]\right].
\end{aligned}$$

Using Pinsker's inequality, we then find that

$$\begin{aligned}
&\mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[\left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}^2 \mid V_{1:j-1}\right]\right] \\
&\le \frac{1}{2}\mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\mathrm{kl}}\left(P_{X_j}(\cdot \mid Y, V_{1:j-1}) \| P_{X_j}(\cdot \mid V_{1:j-1})\right) \mid V_{1:j-1}\right]\right] = \frac{1}{2} I(X_j; Y \mid V_{1:j-1}).
\end{aligned}$$

In particular, we have

$$I(V_j; Y \mid V_{1:j-1}) \le 2\left(e^{2\alpha}-1\right)^2 I(X_j; Y \mid V_{1:j-1}). \tag{10.33}$$

Lastly, we argue that $I(X_j; Y \mid V_{1:j-1}) \le I(X_j; Y \mid X_{1:j-1})$. Indeed, we have by definition[3] that

$$\begin{aligned} I(X_j; Y \mid V_{1:j-1}) &\stackrel{(i)}{=} H(X_j) - H(X_j \mid Y, V_{1:j-1}) \\ &\stackrel{(ii)}{\le} H(X_j) - H(X_j \mid Y, V_{1:j-1}, X_{1:j-1}) \\ &\stackrel{(iii)}{=} H(X_j \mid X_{1:j-1}) - H(X_j \mid Y, X_{1:j-1}) = I(X_j; Y \mid X_{1:j-1}). \end{aligned}$$

Here, equality (i) follows since X_j is independent of $V_{1:j-1}$, inequality (ii) because conditioning reduces entropy, and equality (iii) because X_j is independent of $X_{1:j-1}$. Thus

$$I(V; Y) = \sum_{j=1}^{d} I(V_j; Y \mid V_{1:j-1}) \le 2(e^{2\alpha}-1)^2 \sum_{j=1}^{d} I(X_j; Y \mid X_{1:j-1}) = 2(e^{2\alpha}-1)^2 I(X; Y),$$

which completes the proof.

10.9.3 Proof of Theorem 10.1

In this section, we represent the ith sample by an $d \times n_i$ sample matrix $X^{(i)} \in \mathbb{R}^{d \times n_i}$, where we denote the kth column of $X^{(i)}$ by $X^{(i,k)}$ and the jth row of $X^{(i)}$ by $X_j^{(i)}$. As we describe in our proof outline in Section 10.7, we assume the testing Markov chain $V \to X^{(i)} \to Y_i$. Throughout this argument, we assume that $m \ge 5$; otherwise the interactive lower bound Proposition 10.1 provides a stronger result.

Our first result is a quantitative data processing inequality, analogous to Lemma 10.2 in Section 10.9.1. For the lemma, we do not need to assume normality of the sample X; the full conditions on X are specified in the conditions in the lemma.

Lemma 10.3. *Let V be uniformly random on $\{-1, 1\}^d$. For any (i, j), assume that $X_j^{(i)}$ is independent of $\{X_{j'}^{(i)} : j' \ne j\} \cup \{V_{j'} : j' \ne j\}$ given V_j. Let P_{X_j} be the probability measure of $X_j^{(i)}$ and assume in addition that there exist (measureable) sets $B_j \subset \mathrm{range}(X_j^{(i)})$ such that*

$$\sup_{S \in \sigma(B_j)} \frac{P_{X_j}(S \mid V = v)}{P_{X_j}(S \mid V = v')} \le \exp(\alpha).$$

Define the random variable E_j to be 1 if $X_j^{(i)} \in B_j$ and 0 otherwise. Then

$$I(V; Y_i) \le 2\left(e^{4\alpha}-1\right)^2 I(X^{(i)}; Y_i) + \sum_{j=1}^{d} H(E_j) + \sum_{j=1}^{d} P(E_j = 0).$$

[3] We assume for simplicity and with no loss of generality that X is discrete or has a density with respect to Lebesgue measure.

Now, we provide concrete bounds on each of the terms in the conclusion of Lemma 10.3. Fixing $\delta \geq 0$, for each $v \in \{-1, 1\}^d$ define $\theta_v = \delta v$, and conditional on $V = v \in \{-1, 1\}^d$, let $X^{(i,k)}$, $k = 1, \ldots, n_i$, be drawn i.i.d. from a $\mathsf{N}(\theta_v, \sigma^2 I_{d \times d})$ distribution. That is, each machine has a sample of size n_i from a normal distribution with mean $\theta_v = \delta v$. Under the preceding assumptions, we obtain

Lemma 10.4. *Let $a > 0$ and $\delta > 0$ be chosen such that for all $i \in \{1, \ldots, m\}$, $\frac{\sqrt{n_i} a \delta}{\sigma^2} \leq \frac{1.2564}{4}$ and $a \geq \delta \sqrt{n_i}$. Let $h_2(p) = -p \log(p) - (1-p) \log(1-p)$ denote binary entropy. Then*

$$I(V; Y_i) \leq \frac{d n_i \delta^2}{\sigma^2}, \qquad \text{and} \tag{10.34a}$$

$$\begin{aligned} I(V; Y_i) \leq 128 \frac{\delta^2 a^2}{\sigma^4} n_i H(Y_i) \\ + d \, h_2 \left(\min \left\{ 2 \exp \left(- \frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2} \right), \frac{1}{2} \right\} \right) + 2d \exp \left(- \frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2} \right). \end{aligned} \tag{10.34b}$$

With the bounds (10.34a) and (10.34b) on the mutual information $I(Y_i; V)$, we may now divide our proof into two cases: when $d \geq 10$ and $d < 10$. Let us being with $d \geq 10$. We claim that by combining inequalities (10.34a), (10.34b), and our basic information-theoretic minimax bound (10.19), we have

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \geq \delta^2 \left(\lfloor d/6 \rfloor + 1 \right) \left(\frac{1}{3} - \frac{6\delta^2 \sum_{i=1}^m n_i \min\{128 \cdot 16 \log m \cdot H(Y_i), d\}}{d\sigma^2} \right) \tag{10.35}$$

for all $0 \leq \delta \leq \sigma / 16 \sqrt{\max_i n_i \log m}$. Deferring the proof of inequality (10.35), we show how our desired minimax bound follows essentially immediately. Indeed, by Shannon's source coding theorem we have $H(Y_i) \leq B_i$, whence the minimax bound (10.35) becomes

$$\delta^2 \left(\lfloor d/6 \rfloor + 1 \right) \left(\frac{1}{3} - \frac{6\delta^2 \sum_{i=1}^m n_i \min\{128 \cdot 16 B_i \log m, d\}}{d\sigma^2} \right).$$

In particular, if we choose

$$\delta^2 = \min \left\{ 1, \frac{\sigma^2}{16^2 \max_i n_i \log m}, \frac{d\sigma^2}{36 \sum_{i=1}^m n_i \min\{128 \cdot 16 B_i \log m, d\}} \right\},$$

we obtain

$$\frac{1}{3} - \delta^2 \frac{6 \sum_{i=1}^m n_i \min\{128 \cdot 16 B_i \log m, d\}}{d\sigma^2} \geq \frac{1}{6},$$

which yields the minimax lower bound

$$\begin{aligned} &\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \\ &\geq \frac{1}{6} \left(\lfloor d/6 \rfloor + 1 \right) \min \left\{ 1, \frac{\sigma^2}{16^2 \max_i n_i \log m}, \frac{d\sigma^2}{36 \sum_{i=1}^m n_i \min\{128 \cdot 16 B_i \log m, d\}} \right\}. \end{aligned}$$

To obtain inequality (10.9), we simplify by assuming that $n_i \equiv n$ for all i and perform simple algebraic manipulations, noting that the minimax lower bound $d\sigma^2/(nm)$ holds independently of any communication budget.

Finally, we return to the case when $d \leq 10$, an appeal to Le Cam's method (2.7), as in the proof of Proposition 10.2 (recall inequality (10.31)), and an identical series of steps to bound the mutual information using inequality (10.36) (i.e. applying the same sequence of steps following definition (10.37)) completes the proof.

Showing inequality (10.35) We return to proving the lower bound (10.35), which requires careful data-processing inequalities. First, by inequalities (10.34a) and (10.34b), we have the mutual information bound

$$I(V; Y_i) \leq \frac{n_i \delta^2}{\sigma^2} \min \left\{ 128 \frac{a^2}{\sigma^2} H(Y_i), d \right\} + d\, h_2 \left(\min \left\{ 2 \exp \left(- \frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2} \right), \frac{1}{2} \right\} \right) + 2d \exp \left(- \frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2} \right), \tag{10.36}$$

true for all $a, \delta \geq 0$ and n_i, σ^2 such that $\sqrt{n_i} a \delta \leq 1.2564 \sigma^2/4$ and $a \geq \delta \sqrt{n_i}$.

Now, we consider each of the terms in the bound in inequality (10.36) in turn, finding settings of δ and a so that each term is small. Let us set $a = 4\sigma\sqrt{\log m}$. We begin with the third term in the bound (10.36), where we note that by definining

$$\delta_3^2 := \frac{\sigma^2}{16 \cdot 16 \log(m) \max_i n_i} \tag{10.37}$$

then for $\delta^2 \leq \delta_3^2$ the conditions $\frac{\sqrt{n_i} a \delta}{\sigma^2} \leq \frac{1.2564}{4}$ and $\sqrt{n_i}\delta \leq a$ in Lemma 10.4 are satisfied. In addition, we have $(a - \sqrt{n_i}\delta)^2 \geq (4 - 1/256)^2 \sigma^2 \log m \geq 15\sigma^2 \log m$ for $|\delta| \leq |\delta_3|$, so for such δ

$$\sum_{i=1}^m 2 \exp \left(- \frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2} \right) \leq 2m \exp\left(-(15/2) \log m\right) = \frac{2}{m^{15/2}} < 2 \cdot 10^{-5}.$$

Secondly, we have $h_2(q) \leq (6/5)\sqrt{q}$ for $q \geq 0$. As a consequence, we see that for δ_2^2 chosen identically to the choice (10.37) for δ_3, we have

$$\sum_{i=1}^m 2 h_2 \left(2 \exp \left(- \frac{(a - \sqrt{n_i}\delta_2)^2}{2\sigma^2} \right) \right) \leq \frac{12m}{5} \sqrt{2} \exp\left(-(15/4) \log m\right) < \frac{2}{49}.$$

In particular, with the choice $a = 4\sigma\sqrt{\log m}$ and for all $|\delta| \leq |\delta_3|$, inequality (10.36) implies that

$$\sum_{i=1}^m I(V; Y_i) \leq \delta^2 \sum_{i=1}^m \frac{n_i}{\sigma^2} \min\left\{128 \cdot 16 \log m \cdot H(Y_i), d\right\} + d \left(\frac{2}{49} + 2 \cdot 10^{-5} \right).$$

Substituting this upper bound into the minimax lower bound (10.19), then noting that for $d \geq 10$, we have $6(2/49 + 2 \cdot 10^{-5}) + 6 \log 2/d \leq 2/3$, gives inequality (10.35).

10.9.4 Proof of Lemma 10.3

The proof of this lemma is similar to that of Lemma 10.2, but we must be careful when conditioning on events of the form $X_j^{(i)} \in B_j$. For notational simplicity, we again suppress all dependence of X and Y on the machine index i.

We begin by noting that given E_j, the variable V_j is independent of $V_{1:j-1}$, $X_{1:j-1}$, $V_{j+1:d}$, and $X_{j+1:d}$. Moreover, by the assumption in the lemma we have for any $S \in \sigma(B_j)$ that

$$\frac{P_{X_j}(S \mid V = v, E_j = 1)}{P_{X_j}(S \mid V = v', E_j = 1)} = \frac{P_{X_j}(S \mid V = v)}{P_{X_j}(X_j \in B_j \mid V = v)} \frac{P_{X_j}(X_j \in B_j \mid V = v')}{P_{X_j}(X_j \in S \mid V = v')} \le \exp(2\alpha).$$

We thus obtain the following analogue of the bound (10.32): by Lemma 10.1, we have

$$\begin{aligned} &P(V_j = v_j \mid V_{1:j-1}, Y, E_j = 1) - P(V_j = v_j \mid V_{1:j-1}, E_j = 1) \\ &\le 2\left(e^{4\alpha} - 1\right) \left\| P_{X_j}(\cdot \mid V_{1:j-1}, Y, E_j = 1) - P_{X_j}(\cdot \mid V_{1:j-1}, E_j = 1) \right\|_{\mathrm{TV}} \cdot \ldots \\ &\qquad \min\left\{P(V_j = v_j \mid V_{1:j-1}, Y, E_j = 1), P(V_j = v_j \mid V_{1:j-1}, E_j = 1)\right\}. \end{aligned} \tag{10.38}$$

Proceeding as in the proof of Lemma 10.2 (applying the argument preceding inequality (10.33)), the expression (10.38) implies

$$I(V_j; Y \mid V_{1:j-1}, E_j = 1) \le 2\left(e^{4\alpha} - 1\right)^2 I(X_j; Y \mid V_{1:j-1}, E_j = 1). \tag{10.39}$$

The bound (10.39) as stated conditions on E_j, which makes it somewhat unwieldy. We turn to removing this conditioning. By the definition of (conditional) mutual information, we have

$$\begin{aligned} &P(E_j = 1) I(V_j; Y \mid V_{1:j-1}, E_j = 1) \\ &\quad = I(V_j; Y \mid V_{1:j-1}, E_j) - I(V_j; Y \mid V_{1:j-1}, E_j = 0) P(E_j = 0) \\ &\quad = I(V_j; E_j, Y \mid V_{1:j-1}) - I(V_j; E_j \mid V_{1:j-1}) - I(V_j; Y \mid V_{1:j-1}, E_j = 0) P(E_j = 0). \end{aligned}$$

Conditioning reduces entropy, so

$$\begin{aligned} I(V_j; E_j, Y \mid V_{1:j-1}) &= H(V_j \mid V_{1:j-1}) - H(V_j \mid E_j, Y, V_{1:j-1}) \\ &\ge H(V_j \mid V_{1:j-1}) - H(V_j \mid Y, V_{1:j-1}) = I(V_j; Y \mid V_{1:j-1}), \end{aligned}$$

and noting that $I(V_j; Y \mid V_{1:j-1}, E_j = 0) \le H(V_j) \le 1$ and $I(V_j; E_j \mid V_{1:j-1}) \le H(E_j)$ gives

$$P(E_j = 1) I(V_j; Y \mid V_{1:j-1}, E_j = 1) \ge I(V_j; Y \mid V_{1:j-1}) - H(E_j) - P(E_j = 0). \tag{10.40}$$

We now combine inequalities (10.40) and (10.39) to complete the proof of the lemma. By the definition of conditional mutual information,

$$I(X_j; Y \mid V_{1:j-1}, E_j = 1) \le \frac{I(X_j; Y \mid V_{1:j-1}, E_j)}{P(E_j = 1)} \le \frac{I(X_j; Y \mid V_{1:j-1})}{P(E_j = 1)}.$$

Combining this with inequalities (10.40) and (10.39) yields

$$I(V_j; Y \mid V_{1:j-1}) \le H(E_j) + P(E_j = 0) + 2\left(e^{4\alpha} - 1\right)^2 I(X_j; Y \mid V_{1:j-1}).$$

Up to the additive terms, this is equivalent to the earlier bound (10.33) in the proof of Lemma 10.2; proceeding *mutatis mudandis* we complete the proof.

10.9.5 Proof of Lemma 10.4

To prove inequality (10.34a), we note that $V \to X^{(i)} \to Y_i$ forms a Markov chain. Thus, the data-processing inequality [47] implies that

$$I(V; Y_i) \leq I(V; X^{(i)}) \leq \sum_{k=1}^{n_i} I(V; X^{(i,k)}).$$

Let P_v denote the conditional distribution of $X^{(i,k)}$ given $V = v$. Then the convexity of the KL-divergence establishes inequality (10.34a) via

$$I(V; X^{(i,k)}) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v' \in \mathcal{V}} D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) = \frac{\delta^2}{2\sigma^2} \frac{1}{|\mathcal{V}|^2} \sum_{v,v' \in \mathcal{V}} \|v - v'\|_2^2 = \frac{d\delta^2}{\sigma^2}.$$

To prove inequality (10.34b), we apply Lemma 10.3. First, consider two one-dimensional normal distributions, each with n_i independent observations and variance σ^2, but where one has mean δ and the other mean $-\delta$. For fixed $a \geq 0$, the ratio of their densities is

$$\frac{\exp(-\frac{1}{2\sigma^2}\sum_{l=1}^{n_i}(x_l - \delta)^2)}{\exp(-\frac{1}{2\sigma^2}\sum_{l=1}^{n_i}(x_l + \delta)^2)} = \exp\left(\frac{\delta}{\sigma^2}\sum_{l=1}^{n_i} x_l\right) \leq \exp\left(\frac{\sqrt{n_i}\delta a}{\sigma^2}\right)$$

whenever $|\sum_l x_l| \leq \sqrt{n_i} a$. As a consequence, we see that by taking the sets

$$B_j = \left\{x \in \mathbb{R}^{n_i} : \left|\sum_{l=1}^{n_i} x_l\right| \leq \sqrt{n_i} a\right\},$$

we satisfy the conditions of Lemma 10.3 with the quantity α defined as $\alpha = \sqrt{n_i}\delta a/\sigma^2$. In addition, when $\alpha \leq 1.2564$, we have $\exp(\alpha) - 1 \leq 2\alpha$, so under the conditions of the lemma, $\exp(4\alpha) - 1 = \exp(4\sqrt{n_i}\delta a/\sigma^2) - 1 \leq 8\sqrt{n_i}\delta a/\sigma^2$. Recalling the definition of the event $E_j = \{X_j^{(i)} \in B_j\}$ from Lemma 10.3, we obtain

$$I(V; Y_i) \leq 128 \frac{\delta^2 a^2}{\sigma^4} n_i I(X^{(i)}; Y_i) + \sum_{j=1}^{d} H(E_j) + \sum_{j=1}^{d} P(E_j = 0). \tag{10.41}$$

Comparing this inequality with inequality (10.34b), we see that we must bound the probability of the event $E_j = 0$.

Bounding $P(E_j = 0)$ is not challenging, however. From standard Gaussian tail bounds, we have for Z_l distributed i.i.d. according to $\mathsf{N}(\delta, \sigma^2)$ that

$$\begin{aligned} P(E_j = 0) &= P\left(\left|\sum_{l=1}^{n_i} Z_l\right| \geq \sqrt{n_i} a\right) \\ &= P\left(\sum_{l=1}^{n_i}(Z_l - \delta) \geq \sqrt{n_i} a - n\delta\right) + P\left(\sum_{l=1}^{n_i}(Z_l - \delta) \leq -\sqrt{n_i} a - n\delta\right) \\ &\leq 2\exp\left(-\frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2}\right). \end{aligned} \tag{10.42}$$

Since $h_2(p) \le h_2(\frac{1}{2})$, this provides the bounds on the entropy and probability terms in inequality (10.41) to yield the result (10.34b).

10.10 Proofs of interactive lower bounds for Gaussian observations

In this section, we prove Theorem 10.2 as well as a few auxiliary lemmas on (essentially) data-processing inequalities in interactive settings.

10.10.1 Proof of Theorem 10.2

As in the proof of Theorem 10.1, we choose $V \in \{-1, 1\}^d$ uniformly at random, defining $\theta := \delta V$ for some $\delta > 0$, and we assume machine i draws a sample $X^{(i)} \in \mathbb{R}^{d \times n}$ of size n i.i.d. according to $\mathsf{N}(\theta, \sigma^2 I_{d\times d})$. We denote the full sample—across all machines—along dimension j by X_j. In addition, for each $j \in [d]$, we let $V_{\setminus j}$ denote the coordinates of $V \in \{-1, 1\}^d$ except the jth coordinate.

However, in this situation, while the local samples are independent, the messages are not: the sequence of random variables $Y = (Y_1, \ldots, Y_T)$ is generated in such a way that the distribution of Y_t is $(X^{(i_t)}, Y_{1:t-1})$-measurable, where $i_t \in \{1, \ldots, m\}$ is the machine index upon which Y_t is based (i.e. the machine sending message Y_t). We assume without loss of generality that the sequence $\{i_1, i_2, \ldots, \}$ is fixed in advance: if the choice of index i_t is not fixed but chosen based on $Y_{1:t-1}$ and X, we simply say there exists a default value (say no communication or $Y_t = \perp$) that indicates "nothing" and is 0 bits.

We begin with a lemma that parallels Lemma 10.3 in the proof of Theorem 10.1, though the lemma's conditions are a bit more stringent.

Lemma 10.5. *Assume that $|\mathcal{V}| = 2$ and let V be uniformly random on $\mathcal{V}$. Let $P_{X^{(i)}}$ denote the probability measure of the ith sample $X^{(i)}$. In addition, assume that there is a (measurable) set B such that for any $v, v' \in \mathcal{V}$ we have*

$$\sup \left\{ \frac{P_{X^{(i)}}(S \mid v)}{P_{X^{(i)}}(S \mid v')} \mid S \in \sigma(B), \; v, v' \in \mathcal{V} \right\} \le e^{\alpha}. \tag{10.43}$$

Define the random variable E to be 1 if $X^{(i)} \in B$ for all i and 0 otherwise. Then

$$I(V; Y) \le 2 \left(e^{4\alpha} - 1\right)^2 I(X; Y) + H(E) + P(E = 0).$$

See Section 10.10.2 for a proof of Lemma 10.5.

Now we can provide a concrete bound on mutual information that parallels that of Lemma 10.4. Under the conditions in the preceding paragraphs, we obtain the following lemma. See Section 10.10.3 for a proof of the lemma.

Lemma 10.6. *Let $a > 0$ and $\delta > 0$ be chosen such that $\frac{\sqrt{n}a\delta}{\sigma^2} \leq \frac{1.2564}{4}$ and $a \geq \delta\sqrt{n}$. Let $h_2(p) = -p\log(p) - (1-p)\log(1-p)$ denote binary entropy. Then*

$$I(V_j; Y \mid V_{\setminus j}) \leq 128\frac{\delta^2 n a^2}{\sigma^4} I(X_j; Y \mid V_{\setminus j}) \tag{10.44}$$
$$+ mh_2\left(\min\left\{2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right), \frac{1}{2}\right\}\right) + 2m\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right).$$

To apply the result of inequality (10.44), we need two intermediate inequalities. By construction, V_j is independent of $V_{\setminus j}$, so we have

$$I(V;Y) = \sum_{j=1}^d I(V_j; Y \mid V_{1:j-1}) = \sum_{j=1}^d [H(V_j \mid V_{1:j-1}) - H(V_j \mid Y, V_{1:j-1})]$$
$$\leq \sum_{j=1}^d [H(V_j \mid V_{\setminus j}) - H(V_j \mid Y, V_{\setminus j})] = \sum_{j=1}^d I(V_j; Y \mid V_{\setminus j}) \tag{10.45}$$

because conditioning reduces entropy. Similarly, as X_j is independent of $V_{\setminus j}$ and the $\{X_j\}_{j=1}^d$ are mutually independent, we have the upper bound

$$\sum_{j=1}^d I(X_j; Y \mid V_{-j}) = \sum_{j=1}^d [H(X_j \mid V_{\setminus j}) - H(X_j \mid Y, V_{\setminus j})] \stackrel{(i)}{=} H(X) - \sum_{j=1}^d H(X_j \mid Y, V_{\setminus j})$$
$$\stackrel{(ii)}{\leq} H(X) - \sum_{j=1}^d H(X_j \mid Y, V) \stackrel{(iii)}{\leq} H(X) - H(X|Y,V) = I(X; Y, V),$$

where equality (i) follows by the independence of X_j and $V_{\setminus j}$, inequality (ii) because conditioning reduces entropy, and inequality (iii) because $H(X \mid Y, V) \leq \sum_j H(X_j \mid Y, V)$. Noting that $I(X; V, Y) \leq H(V, Y) \leq H(Y) + d$, we see that

$$\sum_{j=1}^d I(X_j; Y \mid V_{\setminus j}) \leq I(X; V, Y) \leq H(Y) + d. \tag{10.46}$$

Beginning with our original (strong) data-processing bound (10.44), we may combine inequalities (10.45) and (10.46) to obtain

$$I(V;Y) \leq 128\frac{\delta^2 n a^2}{\sigma^4}(H(Y) + d) \tag{10.47}$$
$$+ mdh_2\left(\min\left\{2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right), \frac{1}{2}\right\}\right) + 2md\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right).$$

Inequality (10.47) parallels inequality (10.34b) in Lemma 10.4, whence we may follow the proof of Theorem 10.1 to complete our proof. We now outline the proof for completeness—there are a few minor differences—focusing on the case $d \geq 10$ (the proof in the case that

$d < 10$ is completely parallel to the previous proof). By choosing $a = 4\sigma\sqrt{\log m}$ and $0 \le \delta < \sigma/16\sqrt{n \log m}$, we have

$$I(V;Y) \le \delta^2 \frac{128 \cdot 16 n \log m}{\sigma^2}(H(Y) + d) + d\left(\frac{2}{49} + 2 \cdot 10^{-5}\right).$$

By the minimax lower bound (10.19), we obtain

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \ge \delta^2(\lfloor d/6 \rfloor + 1)\left(\frac{1}{3} - (128 \cdot 16 \cdot 6)\delta^2 \frac{(H(Y) + d) n \log m}{d\sigma^2}\right).$$

By Shannon's source-coding theorem, we have $H(Y) \le B$, and consequently, by setting

$$\delta^2 = \min\left\{1, \frac{\sigma^2}{256 n \log m}, \frac{d\sigma^2}{2048 \cdot 36 \cdot n(B+d)\log m}\right\} = \min\left\{1, \frac{d\sigma^2}{2048 \cdot 36 \cdot n(B+d)\log m}\right\}$$

we obtain $\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \ge \delta^2(\lfloor d/6 \rfloor + 1)/6$. Combining with the above assignment to δ^2, and noting that $\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, \infty) \gtrsim \sigma^2 d/(nm)$ gives the result.

10.10.2 Proof of Lemma 10.5

We state an intermediate claim from which Lemma 10.5 follows quickly. Let us temporarily assume that the set B in the statement of the lemma is $B = \text{range}(X^{(i)})$, so that there is no restriction on the distributions $P_{X^{(i)}}$, that is, the likelihood ratio bound (10.43) holds for all measurable sets S. We claim that in this case,

$$I(V;Y) \le 2\left(e^{2\alpha} - 1\right)^2 I(X;Y). \tag{10.48}$$

Assuming that we have established inequality (10.48), the proof of Lemma 10.5 follows, *mutatis mutandis*, as in the proof of Lemma 10.3 from Lemma 10.2. Thus, it only remains to prove our claim (10.48).

Proof of the data processing inequality (10.48) By the chain-rule for mutual information, we have that

$$I(V;Y) = \sum_{t=1}^{T} I(V; Y_t \mid Y_{1:t-1}).$$

Let $P_{Y_t}(\cdot \mid Y_{1:t-1})$ denote the (marginal) distribution of Y_t given $Y_{1:t-1}$ and define $P_V(\cdot \mid Y_{1:t})$ to be the distribution of V conditional on $Y_{1:t}$. Then we have by marginalization that

$$P_V(\cdot \mid Y_{1:t-1}) = \int P_V(\cdot \mid Y_{1:t-1}, y_t) dP_{Y_t}(y_t \mid Y_{1:t-1})$$

and thus

$$I(V; Y_t \mid Y_{1:t-1}) = \mathbb{E}_{Y_{1:t-1}}\left[\mathbb{E}_{Y_t}\left[D_{\text{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) \mid Y_{1:t-1}\right]\right]. \tag{10.49}$$

We now bound the above KL divergence using the assumed likelihood ratio bound on P_X in the lemma (when $B = \mathcal{X}$, the entire sample space).

By the nonnegativity of the KL divergence, we have

$$\begin{aligned}
&D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) \\
&\quad \le D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) + D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t-1}) \| P_V(\cdot \mid Y_{1:t})\right) \\
&\quad = \sum_{v \in \mathcal{V}} \left(p_V(v \mid Y_{1:t-1}) - p_V(v \mid Y_{1:t})\right) \log \frac{p_V(v \mid Y_{1:t-1})}{p_V(v \mid Y_{1:t})}
\end{aligned}$$

where p_V denotes the p.m.f. of V. We claim that Lemma 10.1 implies that

$$\begin{aligned}
&|p_V(v \mid Y_{1:t-1}) - p_V(v \mid Y_{1:t})| \\
&\le 2\left(e^{2n\alpha} - 1\right) \min\left\{p_V(v \mid Y_{1:t-1}), p_V(v \mid Y_{1:t})\right\} \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}.
\end{aligned} \tag{10.50}$$

Deferring the proof of inequality (10.50) to the end of this section, we give the remainder of the proof. First, by a first-order convexity argument, we have that for any $a, b > 0$

$$\log \frac{a}{b} \le \frac{|a - b|}{\min\{a, b\}}.$$

As a consequence, we find

$$\begin{aligned}
&\left(p_V(v \mid Y_{1:t-1}) - p_V(v \mid Y_{1:t})\right) \log \frac{p_V(v \mid Y_{1:t-1})}{p_V(v \mid Y_{1:t})} \le \frac{\left(p_V(v \mid Y_{1:t-1}) - p_V(v \mid Y_{1:t})\right)^2}{\min\{p_V(v \mid Y_{1:t-1}), p_V(v \mid Y_{1:t})\}} \\
&\quad \le 4\left(e^{2n\alpha} - 1\right)^2 \min\left\{p_V(v \mid Y_{1:t-1}), p_V(v \mid Y_{1:t})\right\} \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2
\end{aligned}$$

by using inequality (10.50). Using the fact that p_V is a p.m.f., we thus have

$$\begin{aligned}
&D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) + D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t-1}) \| P_V(\cdot \mid Y_{1:t})\right) \\
&\le 4\left(e^{2n\alpha} - 1\right)^2 \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2 \sum_{v \in \mathcal{V}} \min\left\{p_V(v \mid Y_{1:t-1}), p_V(v \mid Y_{1:t})\right\} \\
&\le 4\left(e^{2n\alpha} - 1\right)^2 \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2.
\end{aligned}$$

Using Pinsker's inequality, we then find that

$$\begin{aligned}
&\mathbb{E}_{Y_{1:t-1}}\left[\mathbb{E}_{Y_t}\left[\left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2 \mid Y_{1:t-1}\right]\right] \\
&\le \frac{1}{2}\mathbb{E}_{Y_{1:t-1}}\left[\mathbb{E}_{Y_t}\left[D_{\mathrm{kl}}\left(P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) \| P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right) \mid Y_{1:t-1}\right]\right] = \frac{1}{2} I(X^{(i_t)}; Y_t \mid Y_{1:t-1}).
\end{aligned}$$

Since conditioning reduces entropy and Y is discrete, we have

$$\begin{aligned}
I(X^{(i_t)}; Y_t \mid Y_{1:t-1}) &= H(Y_t \mid Y_{1:t-1}) - H(Y_t \mid X^{(i_t)}, Y_{1:t-1}) \\
&\le H(Y_t \mid Y_{1:t-1}) - H(Y_t \mid X, Y_{1:t-1}) = I(X; Y_t \mid Y_{1:t-1}).
\end{aligned}$$

This completes the proof of the lemma, since $\sum_{t=1}^{T} I(X; Y_t \mid Y_{1:t-1}) = I(X; Y)$ by the chain rule for information.

Proof of inequality (10.50) To establish the inequality, we give a one-to-one correspondence between the variables in Lemma 10.1 and the variables in inequality (10.50). We make the following identifications:

$$V \leftrightarrow A \qquad X^{(i_t)} \leftrightarrow B \qquad Y_{1:t-1} \leftrightarrow C \qquad Y_t \leftrightarrow D.$$

For Lemma 10.1 to hold, we must verify conditions (10.24), (10.25), and (10.26). For condition (10.24) to hold, Y_t must be independent of V given $\{Y_{1:t-1}, X^{(i_t)}\}$. Since the distribution of $P_{Y_t}(\cdot \mid Y_{1:t-1}, X^{(i_t)})$ is measurable-$\{Y_{1:t-1}, X^{(i_t)}\}$, Condition (10.26) is satisfied by the assumption in the lemma.

Finally, for condition (10.25) to hold, we must be able to factor the conditional probability of $Y_{1:t-1}$ given $\{V, X^{(i_t)}\}$ as

$$P(Y_{1:t-1} = y_{1:t-1} \mid V, X^{(i_t)}) = \Psi_1(V, y_{1:t-1})\Psi_2(X^{(i_t)}, y_{1:t-1}). \tag{10.51}$$

To prove this decomposition, notice that

$$P(Y_{1:t-1} = y_{1:t-1} \mid V, X^{(i_t)}) = \prod_{k=1}^{t-1} P(Y_k = y_k \mid Y_{1:k-1}, V, X^{(i_t)}).$$

For any $k \in \{1, \ldots, t-1\}$, if $i_k = i_t$—that is, the message Y_k is generated based on sample $X^{(i_t)} = X^{(i_k)}$—then Y_k is independent of V given $\{X^{(i_t)}, Y_{1:k-1}\}$. Thus, $P_{Y_k}(\cdot \mid Y_{1:k-1}, V, X^{(i_t)})$ is measurable-$\{X^{(i_t)}, Y_{1:k-1}\}$. If the kth index $i_k \neq i_t$, then Y_k is independent of $X^{(i_t)}$ given $\{Y_{1:k-1}, V\}$ by construction, which means $P_{Y_k}(\cdot \mid Y_{1:k-1}, V, X^{(i_t)}) = P_{Y_k}(\cdot \mid Y_{1:k-1}, V)$. The decomposition (10.51) thus holds, and we have verified that each of the conditions of Lemma 10.1 holds. We thus establish inequality (10.50).

10.10.3 Proof of Lemma 10.6

To prove inequality (10.44), fix an arbitrary realization $v_{\setminus j} \in \{-1, 1\}^{d-1}$ of $V_{\setminus j}$. Conditioning on $V_{\setminus j} = v_{\setminus j}$, note that $v_j \in \{-1, 1\}$, and consider the distributions of the jth coordinate of each (local) sample $X_j^{(i)} \in \mathbb{R}^n$,

$$P_{X_j^{(i)}}(\cdot \mid V_j = v_j, V_{\setminus j} = v_{\setminus j}) \quad \text{and} \quad P_{X_j^{(i)}}(\cdot \mid V_j = -v_j, V_{\setminus j} = v_{\setminus j}).$$

We claim that these distributions—with appropriate constants—satisfy the conditions of Lemma 10.5. Indeed, fix $a \geq 0$, take the set $B = \{x \in \mathbb{R}^n \mid \|x\|_1 \leq \sqrt{n}a\}$, and set the log-likelihood ratio parameter $\alpha = \sqrt{n}\delta a/\sigma^2$. Then the random variable $E_j = 1$ if $X_j^{(i)} \in B$ for all $i = 1, \ldots, m$, and the proof of Lemma 10.5 proceeds immediately (we still obtain the factorization (10.51) by conditioning everything on $V_{\setminus j} = v_{\setminus j}$). Thus we obtain

$$\begin{aligned} I(V_j; Y \mid V_{\setminus j} = v_{\setminus j}) \leq 2\left(e^{4\alpha} - 1\right)^2 I(X_j; Y \mid V_{\setminus j} = v_{\setminus j}) \\ + H(E_j \mid V_{\setminus j} = v_{\setminus j}) + P(E_j = 0 \mid V_{\setminus j} = v_{\setminus j}). \end{aligned} \tag{10.52}$$

Of course, the event E_j is independent of $V_{\setminus j}$ by construction, so that $P(E_j = 0 \mid V_{\setminus j}) = P(E_j = 0)$, and $H(E_j \mid V_{\setminus j} = v_{\setminus j}) = H(E_j)$, and standard Gaussian tail bounds (cf. the proof of Lemma 10.4 and inequality (10.42)) imply that

$$H(E_j) \leq m h_2 \left(2 \exp\left(-\frac{(a - \sqrt{n}\delta)^2}{2\sigma^2} \right) \right) \quad \text{and} \quad P(E_j = 0) \leq 2m \exp\left(-\frac{(a - \sqrt{n}\delta)^2}{2\sigma^2} \right).$$

Thus by integrating over $V_{\setminus j} = v_{\setminus j}$, inequality (10.52) implies the lemma.

Bibliography

[1] H. Abelson. Lower bounds on information transfer in distributed computations. *Journal of the Association for Computing Machinery*, 27(2):384–392, 1980.

[2] J. Adelman-McCarthy *et al.* The sixth data release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 175(2):297–313, 2008. doi: 10.1086/524984.

[3] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems 24*, 2011.

[4] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[5] A. Agarwal, P. L. Bartlett, and J. Duchi. Oracle inequalities for computationally adaptive model selection. *arXiv:1208.0129 [stat.ML]*, 2012. URL `http://arxiv.org/abs/1208.0129`.

[6] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[7] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

[8] V. Anantharam, A. Gohari, S. Kamath, and C. Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *arXiv:1304.6133 [cs.IT]*, 2013. URL `http://arxiv.org/abs/1304.6133`.

[9] E. Arias-Castro, E. Candés, and M. Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[10] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[11] P. Assouad. Deux remarques sur l'estimation. *C. R. Academy Scientifique Paris Séries I Mathematics*, 296(23):1021–1024, 1983.

[12] P. Auer and C. Gentile. Adaptive and self-confident online learning algorithms. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[13] K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. MSRI Publications, 1997.

[14] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in numerical linear algebra. *SIAM Journal on Matrix Analysis and Applications*, 32(3): 866–901, 2011.

[15] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[16] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the 26th ACM Symposium on Principles of Database Systems*, 2007.

[17] P. L. Bartlett, V. Dani, T. P. Hayes, S. M. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008.

[18] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[19] A. Beimel, K. Nissim, and E. Omri. Distributed private data analysis: Simultaneously solving how and what. In *Advances in Cryptology*, volume 5157 of *Lecture Notes in Computer Science*, pages 451–468. Springer, 2008.

[20] A. Beimel, S. P. Kasiviswanathan, and K. Nissim. Bounds on the sample complexity for private learning and private data release. In *Proceedings of the 7th Theory of Cryptography Conference*, pages 437–454, 2010.

[21] A. Ben-Tal and M. Teboulle. A smoothing technique for nondifferentiable optimization problems. In *Optimization*, Lecture Notes in Mathematics 1405, pages 1–11. Springer Verlag, 1989.

[22] A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12:79–108, 2001.

[23] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985.

[24] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, 2013.

[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 and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

[27] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet*, 65:181–238, 1983.

[28] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1614, 2005.

[29] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fourtieth Annual ACM Symposium on the Theory of Computing*, 2008.

[30] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20*, 2007.

[31] S. Boyd and A. Mutapcic. Stochastic subgradient methods. Course notes for EE364b at Stanford, available at http://www.stanford.edu/class/ee364b/notes/stoch_subgrad_notes.pdf, 2007.

[32] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 2011.

[34] P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984.

[35] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[36] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[37] E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector. *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[38] R. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

[39] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[40] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

[41] A. Chakrabarti, Y. Shi, A. Wirth, and A. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *42nd Annual Symposium on Foundations of Computer Science*, pages 270–278, 2001.

[42] V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):1181–1190, 2013.

[43] K. Chaudhuri and D. Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[44] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[45] C. Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5: 97–138, 1996.

[46] A. Conn, K. Scheinberg, and L. Vicente. *Introduction to Derivative-Free Optimization*, volume 8 of *MPS-SIAM Series on Optimization*. SIAM, 2009.

[47] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[48] A. De. Lower bounds in differential privacy. In *Proceedings of the Ninth Theory of Cryptography Conference*, 2012. URL `http://arxiv.org/abs/1107.2183`.

[49] 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.

[50] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[51] J. C. Duchi and M. J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv:1311.2669 [cs.IT]*, 2013. URL `http://arxiv.org/abs/1311.2669`.

[52] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[53] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[54] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

[55] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[56] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for (parallel) stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[57] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. *arXiv:1210.2085 [stat.ML]*, 2012. URL `http://arxiv.org/abs/1210.2085`.

[58] J. C. Duchi, M. I. Jordan, and H. B. McMahan. Estimation, optimization, and parallelism when data is sparse. In *Advances in Neural Information Processing Systems 26*, 2013.

[59] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy, data processing inequalities, and minimax rates. *arXiv:1302.3203 [math.ST]*, 2013. URL `http://arxiv.org/abs/1302.3203`.

[60] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, 2013.

[61] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order optimization: the power of two function evaluations. *arXiv:1312.2139 [math.OC]*, 2013. URL `http://arxiv.org/abs/1312.2139`.

[62] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *arXiv:1405.0782 [cs.IT]*, 2014. URL `http://arxiv.org/abs/1405.0782`.

[63] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18, 1986.

[64] G. T. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7(2):207–217, 1989.

[65] C. Dwork. Differential privacy: a survey of results. In *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[66] C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the Fourty-First Annual ACM Symposium on the Theory of Computing*, 2009.

[67] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006.

[68] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.

[69] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

[70] S. Efromovich. *Nonparametric Curve Estimation: Methods, Theory, and Applications.* Springer-Verlag, 1999.

[71] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, 1993.

[72] A. El Gamal and Y.-H. Kim. *Network Information Theory.* Cambridge University Press, 2011.

[73] Y. M. Ermoliev. On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences. *Kibernetika*, 2:72–83, 1969.

[74] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 211–222, 2003.

[75] I. P. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1972.

[76] S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14(4):485–502, 1998.

[77] S. E. Fienberg, A. Rinaldo, and X. Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *The International Conference on Privacy in Statistical Databases*, 2010.

[78] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[79] S. Fuller and L. Millett. *The Future of Computing Performance: Game Over or Next Level?* National Academies Press, 2011.

[80] S. R. Ganta, S. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD Conference on Knowledge and Data Discovery (KDD)*, 2008.

[81] C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3), 2002.

[82] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. Technical report, Department of Industrial and Systems Engineering, University of Florida, 2013.

[83] N. Gilbert. Researchers criticize genetic data restrictions. *Nature News*, September 2008. doi: 10.1038/news.2008.1083.

[84] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

[85] A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

[86] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, March–April 2009.

[87] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer, 1992.

[88] R. Hall, A. Rinaldo, and L. Wasserman. Random differential privacy. *arXiv:1112.2680 [stat.ME]*, 2011. URL `http://arxiv.org/abs/1112.2680`.

[89] S. Han and S. Amari. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, 1998.

[90] M. Hardt and G. N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *51st Annual Symposium on Foundations of Computer Science*, 2010.

[91] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the Fourty-Second Annual ACM Symposium on the Theory of Computing*, pages 705–714, 2010. URL `http://arxiv.org/abs/0907.3754`.

[92] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and Applications*, 23:794–798, 1978.

[93] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall, 1995.

[94] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[95] E. Hazan. The convex optimization approach to regret minimization. In *Optimization for Machine Learning*, chapter 10. MIT Press, 2012.

[96] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

[97] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1996.

[98] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1996.

[99] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics*, 4(8):e1000167, 2008.

[100] P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[101] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, 1981.

[102] K. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems 25*, 2012.

[103] 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.

[104] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with the stochastic mirror-prox algorithm. URL `http://arxiv.org/abs/0809.0815`, 2008.

[105] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[106] V. Katkovnik and Y. Kulchitsky. Convergence of a class of random search algorithms. *Automation and Remote Control*, 33(8):1321–1326, 1972.

[107] M. Kearns. *The Computational Complexity of Machine Learning*. PhD thesis, Harvard University, May 1989.

[108] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association for Computing Machinery*, 45(6):983–1006, 1998.

[109] R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

[110] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, Jan. 1997.

[111] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[112] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, Second edition, 2003.

[113] H. Lakshmanan and D. P. de Farias. Decentralized resource allocation in dynamic networks of agents. *SIAM Journal on Optimization*, 19(2):911–940, 2008.

[114] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2010. Online first. URL `http://www.ise.ufl.edu/glan/papers/OPT_SA4.pdf`.

[115] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[116] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[117] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[118] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

[119] C. Lemaréchal and C. Sagastizábal. Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

[120] H. Li and N. Homer. A survey of sequence alignment algorithms for next generation sequencing. *Briefings in Bionformatics*, 11(5):473–483, 2010.

[121] Z.-Q. Luo. Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on Information Theory*, 51(6):2210–2219, 2005.

[122] Z.-Q. Luo and J. N. Tsitsiklis. On the communication complexity of distributed algebraic computation. *Journal of the Association for Computing Machinery*, 40(5): 1019–1047, 1993.

[123] Z.-Q. Luo and J. N. Tsitsiklis. Data fusion with minimal communication. *IEEE Transactions on Information Theory*, 40(5):1551–1563, 1994.

[124] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Identifying malicious urls: An application of large-scale online learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[125] G. Mann, R. McDonald, M. Mohri, N. Silberman, and D. Walker. Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models. In *Advances in Neural Information Processing Systems 22*, pages 1231–1239, 2009.

[126] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[127] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[128] R. McDonald, K. Hall, and G. Mann. Distributed training strategies for the structured perceptron. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.

[129] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. The limits of two-party differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, 2010.

[130] B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[131] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54:48–61, 2009.

[132] 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.

[133] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[134] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[135] 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.

[136] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[137] Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[138] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):261–283, 2009.

[139] Y. Nesterov. Random gradient-free minimization of convex functions. URL `http://www.ecore.be/DPs/dp_1297333890.pdf`, 2011.

[140] F. Niu, B. Recht, C. Re, and S. Wright. Hogwild: a lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, 2011.

[141] F. Niu, B. Recht, C. Ré, and S. Wright. Hogwild: a lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, 2011.

[142] A. Olshevsky and J. N. Tsitsiklis. Convergence speed in distributed consensus and averaging. *SIAM Journal on Control and Optimization*, 48(1):33–55, 2009.

[143] R. R. Phelps. *Lectures on Choquet's Theorem, Second Edition*. Springer, 2001.

[144] L. Plaskota. *Noisy Information and Computational Complexity*. Cambridge University Press, 1996.

[145] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

[146] B. T. Polyak and J. Tsypkin. Robust identification. *Automatica*, 16:53–63, 1980. doi: 10.1016/0005-1098(80)90086-2. URL `http://dx.doi.org/10.1016/0005-1098(80)90086-2`.

[147] 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.

[148] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976—6994, 2011.

[149] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *arXiv:1212.0873 [math.OC]*, 2012. URL `http://arxiv.org/abs/1212.0873`.

[150] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[151] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[152] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[153] R. T. Rockafellar and R. J. B. Wets. On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics: An International Journal of Probability and Stochastic Processes*, 7:173–182, 1982.

[154] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

[155] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

[156] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley, 1981.

[157] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 1988.

[158] D. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[159] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

[160] S. Shalev-Shwartz, Y. Singer, and A. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[161] S. Shalev-Shwartz, O. Shamir, and E. Tromer. Using more data to speed-up training time. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, 2012.

[162] S. Shalev-Swhartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[163] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, 2013.

[164] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing*, 2011.

[165] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, 2003.

[166] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro. Mini-batch primal and dual methods for SVMs. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[167] B. Taskar. *Learning Structured Prediction Models: A Large Margin Approach*. PhD thesis, Stanford University, 2005.

[168] J. Traub and A. Werschulz. *Complexity and Information*. Cambridge University Press, 1999.

[169] J. Traub, H. Wasilkowski, and H. Wozniakowski. *Information-Based Complexity*. Academic Press, 1988.

[170] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008. URL `http://www.math.washington.edu/~tseng/papers/apgm.pdf`.

[171] J. N. Tsitsiklis. Decentralized detection. In *Advances in Signal Processing, Vol. 2*, pages 297–344. JAI Press, 1993.

[172] J. N. Tsitsiklis and Z.-Q. Luo. Communication complexity of convex optimization. *Journal of Complexity*, 3:231–243, 1987.

[173] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[174] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, Nov. 1984.

[175] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

[176] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[177] 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, 2008.

[178] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[179] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[180] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[181] J. M. Wing. Computational thinking. *Communications of the ACM*, 49(3):33–35, 2006.

[182] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems 22*, 2009.

[183] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

[184] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, 2003.

[185] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[186] A. C.-C. Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, pages 209–213. ACM, 1979.

[187] F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48:56–67, 2012.

[188] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[189] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

[190] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[191] M. A. Zinkevich, A. Smola, M. Weimer, and L. Li. Parallelized Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 23*, 2010.