Constrained Forms of Statistical Minimax: Computation, Communication, and Privacy

Martin J. Wainwright *

Abstract. A fundamental quantity in statistical decision theory is the notion of the minimax risk associated with an estimation problem. It is based on a saddlepoint problem, in which nature plays the role of adversary in choosing the underlying problem instance, and the statistician seeks an estimator with good properties uniformly over a class of problem instances. We argue that in many modern estimation problems arising in the mathematical sciences, the classical notion of minimax risk suffers from a significant deficiency: to wit, it allows for all possible estimators, including those with prohibitive computational costs, unmanageable storage requirements, or other undesirable properties. Accordingly, we introduce some refinements of minimax risk based on imposing additional constraints on the sets of possible estimators. We illustrate this notion of constrained statistical minimax via three vignettes, based on restrictions involving computation, communication, and privacy, respectively.

Mathematics Subject Classification (2010). Primary 62Cxx; Secondary 68W40

Keywords. Statistical minimax; information theory; metric entropy; communication complexity; computational complexity; differential privacy.

1. Introduction

Minimax theory is a cornerstone of statistical decision theory, providing a classical approach to assessing the quality of a statistical estimator in the frequentist sense. It is based on a saddle point problem, in which the adversary chooses a worst-case set of parameters, and the statistician seeks to minimize the worst-case risk via a well-chosen estimator. There is now a rich and well-developed body of theory for bounding and/or computing the minimax risk for various statistical estimation problems (e.g., see the papers [26, 6, 44, 43] and references therein).

In full generality, a statistical estimator of a parameter $\theta \in \Theta$ is a measurable function of the data, taking values in the parameter space Θ. Herein lies a serious deficiency of the classical notion of minimax risk: apart from the measurability requirement, the infimum over estimators is unconstrained. Consequently, the classical notion allows for the use of estimators that may be practically infeasible for

*Research partially supported by National Science Foundation grant CIF-31712-23800, and Office of Naval Research MURI grant N00014-11-1-0688. This lecture is based on pieces of joint work with John Duchi, Michael Jordan and Yuchen Zhang.

various reasons. For instance, it allows for estimators whose computational complexity can scale arbitrarily quickly with the problem dimension and parameters. In practice, it is typically only of interest to consider estimators with polynomial-time complexity, or perhaps even more stringently, with linear or quadratic complexity. In addition, it implicitly assumes that all the data can be aggregated at a central location. For the massive data sets that are generated in many modern scientific and engineering applications, such centralized aggregation is often impossible, and instead, distributed methods should be used. Finally, there are many types of data—including financial records, medical tests, and genetic data— that lead naturally to privacy concerns. Given the prevalence of such data types, another important issue is the study of statistical estimators that have privacy-respecting properties.

Accordingly, with the motivation of addressing these deficiencies of the classical minimax risk, the goal of this overview is to introduce and discuss various constrained forms of minimax risk. We begin in Section 2 by providing a more precise definition of the problem of statistical estimation and the notion of minimax risk. Sections 3, 4, and 5, respectively, are devoted to constrained forms of minimax risk based on communication, privacy, and computation. The results described here are based on joint pieces of work [17, 47, 45] with John Duchi, Michael Jordan, and Yuchen Zhang.

2. Classical minimax risk

In order to set the stage, we begin by describing the problem of statistical estimation in general terms, and then introducing the classical notion of minimax risk. Consider a family of probability distributions $\mathcal{P}$ with support $\mathcal{X}$, and consider a mapping $\theta : \mathcal{P} \to \Theta$. Thus, associated with member $\mathbb{P} \in \mathcal{P}$ is the parameter $\theta(\mathbb{P})$. Given a fixed but unknown distribution $\mathbb{P} \in \mathcal{P}$, suppose that we observe a sequence $X_1^n := (X_1, \ldots, X_n)$ of random variables drawn i.i.d. according to $\mathbb{P}$. Based on observing the sequence X_1^n, our goal is to estimate the *target parameter* $\theta^* := \theta(\mathbb{P})$. More formally, an estimator of θ^* is a measurable function $\widehat{\theta} : \mathcal{X}^n \to \Theta$. In order to assess the quality of any estimator, we let $\rho : \Theta \times \Theta \to [0, \infty)$ be some non-negative measure of error on the parameter space Θ, and consider the associated *risk function*

$$R(\widehat{\theta}, \theta^*) = \mathbb{E}\big[\rho(\widehat{\theta}(X_1^n), \theta^*)\big],$$

where the expectation is taken over the samples. Typical choices of the error function ρ are various metrics, or powers of such metrics.

For any fixed estimator $\widehat{\theta}$, the function $\theta^* \mapsto R(\widehat{\theta}, \theta^*)$ characterizes its performance as the underlying truth θ^* ranges over the parameter space Θ. (Here and throughout the paper, whenever the dependence of $\widehat{\theta}$ on the samples X_1^n is clear from the context, then we simply write $\widehat{\theta}$.) There are various ways in which to "scalarize" the risk function in order to assign a single number to each estimator. In the minimax setting, for each estimator $\widehat{\theta}$, we compute the worst-case risk

$\sup_{\mathbb{P}\in\mathcal{P}} R(\widehat{\theta}, \theta(\mathbb{P}))$, and rank estimators according to this ordering. The estimator that is optimal in this sense defines a quantity known as the *minimax risk*—namely,

$$\mathfrak{M}_n(\theta(\mathcal{P})) := \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} R(\widehat{\theta}, \theta(\mathbb{P})), \tag{1}$$

where the infimum ranges over all possible estimators.

As a prelude to later results, let us consider a few illustrative instances of these definitions.

Location families: For a fixed base density function ϕ and vector $\theta \in \mathbb{R}^d$, consider a distribution $\mathbb{P}_\theta$ specified by a density function (with respect to Lebesgue measure) of the form $f_\theta(x) = \phi(x - \theta)$. Letting Θ be some subset of $\mathbb{R}^d$, the collection of distributions $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ is known as a *location family*, since θ plays the role of a centering quantity. Important examples of location families include the normal location family (specified by the standard normal density $\phi(x) = e^{-\frac{\|x\|_2^2}{2}}/\sqrt{2\pi}$), and for $d = 1$, the uniform location family (specified by the base density $\phi(x) = \mathbb{I}[x \in [0, 1]]$, where $\mathbb{I}$ is a zero-one valued indicator function for set membership). A typical error measure is the squared ℓ_2-norm $\rho(\widehat{\theta}, \theta^*) = \|\widehat{\theta} - \theta^*\|_2^2$. We discuss the role of communication constraints for minimax rates in location families in Section 3.

Density estimation: Parameters need not be limited to vectors, but can be more general infinite-dimensional objects. As one instance, suppose that $\mathcal{P}$ consists of a family of distributions supported on the interval $[0, 1]$, and with densities with respect to Lebesgue measure. Suppose that $f^* = \theta(\mathbb{P})$ is the density of $\mathbb{P}$. In this case, an estimator $\widehat{\theta}$ returns a density function $\widehat{f}$ with support on $[0, 1]$, and a reasonable measure of error is the usual squared $L^2([0, 1])$ norm

$$\rho(\widehat{f}, f^*) = \int_0^1 \big(\widehat{f}(t) - f^*(t)\big)^2 dt. \tag{2}$$

We discuss this example in Section 4.

Linear regression: An instance of linear regression is specified by a known design matrix $X \in \mathbb{R}^{n\times d}$, in which each row corresponds to a vector of d predictors, and an unknown weight vector $\theta^* \in \mathbb{R}^d$. An observed response vector $Y \in \mathbb{R}^n$ is assumed to be generated by the equation

$$Y = X\theta^* + W, \tag{3}$$

where $W \in \mathbb{R}^n$ is a vector of i.i.d. $N(0, \sigma^2)$ variates. Equivalently, the underlying statistical model consists of the family of distributions $\{\mathbb{P}_\theta, \mid \theta \in \Theta\}$, where each $\mathbb{P}_\theta$ is the distribution of a $N(X\theta, \sigma^2 I_{n\times n})$ random vector. (This example is slightly different from our set-up, in that the components of the observed vector Y are

not identically distributed for a fixed X.) One error measure is the in-sample prediction error

$$\rho_X(\widehat{\theta}, \theta^*) := \frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2. \tag{4}$$

The problem of linear regression under this error measure is discussed in detail in Section 5.

3. Estimation under communication constraints

Given the modern "data deluge", it is often the case that centralized methods—in which all the data can be stored on a single computer—are no longer possible to implement. Instead, distributed methods must be used. Given a cluster of m machines, it is natural to consider splitting the full data set into m separate subsets, operating separately on each subset, and then performing some sort of communication in order to agree upon a consensus estimate. In practice, the communication budget is severely limited due to power or bandwidth constraints, and such constraints make the problem mathematically interesting. Various researchers have studied communication-efficient algorithms for statistical estimation (e.g., see the papers [15, 30, 2, 46] and references therein). In this spirit, our first vignette is devoted the role of communication constraints in statistical estimation: we define a communication-constrained version of the minimax risk, and provide sharp bounds for a few examples. See the paper [45] for further details.

Distributed estimation protocols: Recall from Section 1 the general framework of statistical estimation, based on some family of distributions $\mathcal{P}$. Suppose that, for some fixed but unknown member $\mathbb{P}$ of $\mathcal{P}$, there are $m \geq 1$ sets of data, each stored on an individual machines. For $j \in [m] := \{1, \ldots, m\}$, the j^{th} subset $X_{1,j}^n := (X_{1j}, \ldots X_{nj})$ is an i.i.d. sample of size n from the unknown distribution $\mathbb{P}$. Consequently, the *total sample size* across all machines is $N = mn$. Given this distributed collection of local data sets, our goal is to estimate $\theta(\mathbb{P})$ based on the full collection of data $X_1^N = (X_{1,j}^n, \;\; j \in [m])$, but using limited communication. Of particular interest to us is the minimal number of bits that must be exchanged in order for a distributed protocol to match the centralized minimax rate—that is, the optimal performance for an estimator given direct access to all N samples.

We now define a particular class of distributed protocols Π, which operate in a sequence of rounds. At each round $t = 1, 2, \ldots$, machine j sends to a central fusion center a message $Y_{t,j}$ that is a measurable function of the local data $X_{1,j}^n$, and potentially of past messages. We use $\overline{Y}_t = \{Y_{t,j}\}_{j \in [m]}$ denote the collection of all messages sent at round t. Given a total of T rounds, the fusion center collects the sequence $(\overline{Y}_1, \ldots, \overline{Y}_T)$, and constructs an estimator $\widehat{\theta} := \widehat{\theta}(\overline{Y}_1, \ldots, \overline{Y}_T)$.

We refer to the length $L_{t,j}$ of message $Y_{t,j}$ is the minimal number of bits required to encode it, and the total length $L = \sum_{t=1}^T \sum_{j=1}^m L_{t,j}$ 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.

The simplest type of protocol is an *independent* one: it involves only on a single round ($T = 1$) of communication, in which machine j sends message $Y_{1,j}$ to the fusion center. Since there are no past messages, the message $Y_{1,j}$ is a function only of the local data $X_{1,j}^n$. Given a class of distributions $\mathcal{P}$, the class of independent protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\mathrm{ind}}(B, \mathcal{P}) = \Big\{ \text{ independent protocols } \Pi \text{ such that } \quad \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \Big[\sum_{j=1}^{m} L_j \Big] \leq B \Big\}. \tag{5}$$

In the independent case, we use Y_j to indicate the message sent from processor j, and L_j to denote its length.

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 j sends message $Y_{t,j}$ to the fusion center at time t, who then relays it back to all other machines in the system. This type of global broadcast system is reasonable in settings in which the processors have limited power or upstream capacity, but the centralized fusion center can send messages without limit. In the interactive setting, the message $Y_{t,j}$ should be viewed as a measurable function of the local data $X_{1,j}^n$, and the past messages $\overline{Y}_{1:t-1}$. The family of interactive protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\mathrm{inter}}(B, \mathcal{P}) = \Big\{ \text{interactive protocols } \Pi \text{ such that } \quad \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[L] \leq B \Big\}. \tag{6}$$

Distributed minimax risks: We are now equipped to define some distributed analogues of the classical minimax risk (1). Given a class of distributions $\mathcal{P}$, suppose that we are interested in estimating some parameter $\theta : \mathcal{P} \to \Theta$. Given a communication budget B, we apply an independent protocol Π that generates a sequence of messages $Y_1^m = (Y_1, \ldots, Y_m)$, and we use $\widehat{\theta}(Y_1^m)$ to denote an estimator that is a measurable function of these messages. With this set-up, the *minimax risk for independent protocols* under squared ℓ_2-error is given by

$$\mathfrak{M}_{n,m}^{\mathrm{ind}}(\theta(\mathcal{P}); B) := \inf_{\Pi \in \mathcal{A}_{\mathrm{ind}}(B, \mathcal{P})} \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}, \Pi} \big[\|\widehat{\theta}(Y_1^m) - \theta(\mathbb{P})\|_2^2 \big]. \tag{7}$$

Here the double infimum is taken over all independent procotols Π that satisfy the budget constraint B, and over all estimators $\widehat{\theta}(Y_1^m)$. The *minimax risk for interactive protocols*, denoted by $\mathfrak{M}_{n,m}^{\mathrm{inter}}$, is defined analogously, where the infimum is instead taken over the class of interactive protocols.

In either case, of primary interest is the following question: how large a budget B is required so as to ensure that the distributed minimax risk (7) matches the classical minimax risk (1) up to constant factors? In the following subsections, we answer this question precisely for two different classes of statistical estimation problems.

Bounds for uniform location family: We begin by considering a univariate example, in particular the problem of estimating the location parameter in the uniform location family $\mathcal{U} = \{P_\theta,\ \theta \in [-1, 1]\}$, where $\mathbb{P}_\theta$ denotes the uniform distribution on the interval $[\theta - 1,\ \theta + 1]$.

Proposition 1. *Consider the uniform location family $\mathcal{U}$ with n i.i.d. observations per machine:*

(a) There is a universal constant c such that given a budget $B = \log(1/\delta)$ for any $\delta \geq \frac{1}{mn}$, the minimax risk is lower bounded as

$$\mathfrak{M}_{n,m}^{\text{inter}}(\theta(\mathcal{U}); B) \geq \frac{c}{\delta^2}.$$

(b) Conversely, given a budget of $B = \big[2 + 2\ln m\big] \log(mn)$ bits, there is a universal constant c' such that

$$\mathfrak{M}_{n,m}^{\text{inter}}(\theta(\mathcal{U}); B) \leq \frac{c'}{(mn)^2}.$$

If each of m machines receives n observations, we have a total sample size of mn, so the minimax rate over all centralized procedures scales as $1/(mn)^2$. Consequently, Proposition 1 shows that the number of bits required to achieve the centralized rate has only *logarithmic* dependence on the number m of machines and local sampl size n. Part (a) shows that if $B \ll \log(mn)$, then the distributed minimax rate is larger than the centralized optimum, so that this logarithmic scaling is unavoidable.

The proof of Proposition 1 is based on a somewhat more general result, one involving the geometric structure of the parameter space Θ, as captured by its metric entropy [28]. In particular, given a subset $\Theta \subset \mathbb{R}^d$, we say $\{\theta^1, \ldots, \theta^K\}$ are δ-separated if $\|\theta^i - \theta^j\|_2 > \delta$ for $i \neq j$. The *packing entropy* of Θ with respect the Euclidean norm is given by

$$\log M_\Theta(\delta) := \log_2 \big[\max \big\{ K \in \mathbb{N} \mid \{\theta_1, \ldots, \theta^K\} \subset \Theta \text{ are } \delta\text{-separated} \big\} \big]. \quad (8)$$

The function $\theta \mapsto \log M_\Theta(\delta)$ is left-continuous and non-increasing in δ, so we may define the inverse function $\log M_\Theta^{-1}(B) := \sup\{\delta \mid \log M_\Theta(\delta) \geq B\}$. With this notation, we have the following general result:

Theorem 1. *For any family of distributions $\mathcal{P}$ and parameter set $\Theta = \theta(\mathcal{P})$, the interactive minimax risk is lower bounded as*

$$\mathfrak{M}_{n,m}^{\text{inter}}(\theta(\mathcal{P}); B) \geq \Big(\frac{1}{4} \log M_\Theta^{-1}(2B + 2) \Big)^2. \quad (9)$$

Of course, the same lower bound also holds for $\mathfrak{M}_{n,m}^{\text{ind}}(\theta, \mathcal{P}, B)$, since any independent protocol is a special case of an interactive protocol. Theorem 1 is a relatively generic statement, not exploiting any particular structure of the problem; however, there are problems for which it cannot be improved by more than constant factors [45].

Bounds for Gaussian location families: Proposition 1 shows that achieving the minimax risk in the uniform location family requires a budget scaling only logarithmically in the number of machines m. It is natural to wonder whether such logarithmic dependence holds more generally. Here we show that it does not: for the Gaussian location family, the dependence on m must be (nearly) linear.

Consider the *d-dimensional normal location family*

$$\mathcal{N}_d([-1,1]^d) = \{\mathcal{N}(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \Theta = [-1,1]^d\}, \tag{10}$$

and suppose that our goal is to estimate the mean vector $\theta \in \mathbb{R}^d$ under the error measure $\rho(\widehat{\theta}, \theta^*) = \|\widehat{\theta} - \theta^*\|_2^2$. Given a total of $N = mn$ samples, the centralized minimax rate scales as $\frac{\sigma^2}{mn}$, achieved by the sample mean. The following result addresses the minimal budget B required for a distributed protocol to match this centralized minimax rate:

Theorem 2. *There exists a universal (numerical) constant c such that*

$$\mathfrak{M}^{\text{inter}}_{n,m}(\mathcal{N}_d([-1,1]^d); B) \geq c\frac{\sigma^2 d}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log m}, \frac{m}{(B/d+1)\ \log m}\right\}. \tag{11}$$

Consequently, Theorem 2 shows that to match the classical minimax risk up to constant factors, the number of bits communicated must scale with the product of the dimension d and number of machines m—more precisely, we must have $B \asymp dm/\log m$. Apart from the logarithmic factor, 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 using $\mathcal{O}(d\log(n/\sigma^2))$ 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 the optimal order $\sigma^2 d/(mn)$.

4. Minimax theory under privacy constraints

In the modern practice of statistics, privacy concerns are becoming increasingly important. Many forms of data, including financial records, medical records, and genetic tests, have associated privacy concerns. In such settings, it is natural to individuals might request some form of privacy guarantee before allowing their data to be collected. At the same time, there is a great deal of statistical utility associated with the collection of such data, including more efficient allocation of medical resources, and biomedical research into the genetic underpinnings of disease.

There is a very large body of classical research on privacy and statistical inference (e.g., [41, 23, 18, 19]. 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. In a more recent line of work, a formal definition of disclosure risk, known as differential privacy [21, 3, 20], has emerged from the theoretical computer science community, and has been the focus of considerable attention (e.g., see the papers [22, 25, 42, 35, 12, 27, 14, 11] and references

therein). Here we describe how to use the notion of local differential privacy in order to define a constrained version of the minimax risk; see the paper [17] for further details.

Differential privacy: Let us begin by defining the notion of (local) differential privacy. Suppose that X_1^n represents the original data, where each X_i takes values in the space $\mathcal{X}$. As a means of preserving privacy, we release only a "privatized" sequence Z_1^n, where each Z_i takes values in the space $\mathcal{Z}$. In the case of a non-interactive mechanism, the two sequences are related via a conditional distribution $\mathbb{Q}$ that takes the product form

$$\mathbb{Q}_n(Z_1, \ldots, Z_n \mid X_1, \ldots, X_n) = \prod_{i=1}^n \mathbb{Q}(Z_i \mid X_i). \tag{12}$$

We refer to $\mathbb{Q}$ as the *channel distribution*, since it acts as a conduit between the private data X and observed data Z. There are also more complicated, interactive forms of privacy mechanisms, in which the product condition (12) is relaxed, but we restrict attention here to this simpler case.

We now give a precise definition of local differential privacy. Let $\sigma(\mathcal{Z})$ be the σ-field over which the channel distribution $\mathbb{Q}$ is defined. Given a privacy parameter $\alpha \geq 0$, the distribution $\mathbb{Q}$ is said to satisfy *α-local-differential privacy* if

$$\sup_{S \in \sigma(\mathcal{Z})} \sup_{x, x' \in \mathcal{X}} \frac{\mathbb{Q}(Z \in S \mid X = x)}{\mathbb{Q}(Z \in S \mid X = x')} \leq \exp(\alpha), \tag{13}$$

This formulation of local privacy was first proposed by Evfimievski et al. [22]. Since we limit our discussion to local privacy throughout this overview, we typically omit the adjective "local" from here onwards.

The definition (13) has a very natural consequence in terms of disclosure risk: when the privacy parameter α is relatively close to zero, then in a uniform sense over events S, it is impossible to distinguish between two different realizations of the private variable X. Indeed, a simple argument shows that the definition (13) provides a lower bound on the error in a binary hypothesis test between $X = x$ and $X = x'$; see Wasserman and Zhou [42] for more details.

The *Laplace mechanism* is a simple way in which to enforce α-privacy. Given a datum X, suppose that we release the private variable $Z = X + W$, where W follows a Laplace distribution with parameter α—that is, it has density $\phi(w) = \frac{\alpha}{2} \exp\big(-\alpha|w|\big)$. In this case, for any pair $x, x' \in [0, 1]$, we have

$$\frac{\mathbb{Q}(Z = z \mid X = x)}{\mathbb{Q}(Z = z \mid X = x')} = \frac{\frac{\alpha}{2} \exp\big(-\alpha|z - x|\big)}{\frac{\alpha}{2} \exp\big(-\alpha|z - x'|\big)} \leq \exp(\alpha|x - x'|) \leq \exp(\alpha), \tag{14}$$

showing that the Laplace mechanism provides differential privacy on the interval $[0, 1]$. Part of the goal of studying the α-private minimax risk is to determine under what conditions, if any, a specific method for producing α-private variables, such as the Laplace mechanism, is optimal.

The α-private minimax risk: We now turn to a definition of the notion of an α-private minimax risk. As usual, let $\mathcal{P}$ denote a family of probability distributions on the space $\mathcal{X}$, and suppose that our goal is to estimate some parameter $\theta(\mathbb{P})$. Let $\mathcal{Q}_\alpha$ denote the class of all channel distributions $\mathbb{Q}$ satisfying α-local differential privacy (13). In an operational sense, any distribution $\mathbb{Q} \in \mathcal{Q}_\alpha$ can be thought of as a privacy mechanism—namely, one means of generating a privatized data set Z_1^n from the raw data X_1^n. Rather than allowing estimators to depend on the raw data, we consider only estimators $\widetilde{\theta} = \widetilde{\theta}(Z_1^n)$ that are measurable functions of the privatized data Z_1^n. For a fixed channel distribution $\mathbb{Q}$ (and hence fixed distribution over the variables Z_1^n) and a fixed estimator $\widetilde{\theta}$, the usual worst-case risk

$$\sup_{\mathbb{P} \in \mathcal{P}} R(\widetilde{\theta}(Z_1^n, \theta(\mathbb{P})) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}, \mathbb{Q}}\big[\rho(\widetilde{\theta}(Z_1^n, \theta(\mathbb{P})\big] \tag{15}$$

is a measure of the quality of $\widetilde{\theta}$. In addition to finding the optimal estimator $\widetilde{\theta}$, we also seek an *optimal privacy mechanism*—namely, a member of $\mathcal{Q}_\alpha$ for which the minimax risk is minimized. More formally, we define the *α-private minimax risk* as

$$\mathfrak{M}_n(\theta(\mathcal{P}); \alpha) := \inf_{\mathbb{Q} \in \mathcal{Q}(\alpha)} \inf_{\widetilde{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}, \mathbb{Q}}\big[\rho(\widetilde{\theta}(Z_1^n), \theta(\mathbb{P}))\big]. \tag{16}$$

When $\alpha = \infty$, it reduces to the usual notion of minimax risk, but of primary interest are value of α relatively close to zero. The private minimax risk (16) allows us to study the tradeoff between the privacy, as measured by the differential privacy parameter α, and the statistical utility, as measured by the minimax risk of all estimators that make use only of the privatized data Z_1^n.

Density estimation under α-local-privacy: We now turn to an example that demonstrates some striking differences between the ordinary and α-private minimax risks. Recall the problem of density estimation introduced in Section 1. Given n i.i.d. samples X_1^n drawn from an univariate distribution with density f^* supported on $[0, 1]$, and the goal is to return an estimate $\widehat{f}$ of the unknown density function, and we evaluate its quality using the squared $L^2([0, 1])$ error previously defined in equation (2). In this section, we state a result that demonstrates how the minimax rate for estimating density functions with Sobolev classes changes with the addition of a privacy constraint.

We begin by defining Sobolev classes in terms of elliptical subsets of the sequence space $\ell^2(\mathbb{N})$. Consider a sequence of functions $\{\phi_j\}_{j=1}^\infty$ that form an orthonormal basis for $L^2([0, 1])$, so that any function $f \in L^2([0, 1])$ can be expanded as a sum $\sum_{j=1}^\infty \theta_j \phi_j$ in terms of the basis coefficients $\theta_j := \int f(x)\phi_j(x)dx$, and we are guaranteed that $\{\theta_j\}_{j=1}^\infty \in \ell^2(\mathbb{N})$. Sobolev classes are obtained by enforcing a particular decay rate on the coefficients θ. In particular, given a parameter $s \geq 1$,

the generalized Sobolev class $\mathcal{F}_s([0,1])$ is given by

$$\mathcal{F}_s([0,1]) := \big\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \in L^2([0,1]) \quad \text{for a sequence } \{\theta_j\}_{j=1}^{\infty} \text{ s.t. } \sum_{j=1}^{\infty} j^{2s} \theta_j^2 \leq 1 \big\}. \tag{17}$$

If we choose the trigonometric basis as our orthonormal basis, then membership in the classical Sobolev class $\mathcal{F}_s([0,1])$ corresponds to certain smoothness constraints on the derivatives of f (e.g., see the book [38] for details).

In the classical (non-private) setting, the density estimate $\widehat{f}$ is constructed based on direct observation of the original samples X_1^n, where each $X_i \sim \mathbb{P}$. In this setting, it is known [36, 38] that the minimax risk for non-private estimation of densities in the class $\mathcal{F}_s([0,1])$ scales as

$$\mathfrak{M}_n\big(\mathcal{F}_s([0,1])\big) \asymp \big(\frac{1}{n}\big)^{\frac{2s}{2s+1}}. \tag{18}$$

For instance, when $s = 1$, corresponding to Lipschitz functions when using the trigonometric basis, then the minimax rate scales as $n^{-\frac{2}{3}}$. Naturally, the minimax rate increases towards the parametric rate n^{-1} as the smoothness parameter s tends to infinity. The minimax rate (18) can be achieved by various methods, with one of the simplest being the orthogonal series estimator. Given the samples X_1^n, this method is based on computing the empirical basis coefficients $\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^n \phi_j(X_i)$, and then setting

$$\widehat{f} = \sum_{j=1}^{T} \widehat{\theta}_j \phi_j, \qquad \text{where } T = n^{\frac{1}{2s+1}}. \tag{19}$$

The specified choice of truncation level T provides the optimal trade-off between the bias and variance of the estimator, and some calculations show that it achieves the minimax rate (18), assuming that the smoothness level s is known to the method.

Now consider the case of α-private density estimation, in which we only observe a privatized version Z_1^n of the raw data X_1^n. The following theorem [17] characterizes the minimax rate when the α-private channel is chosen in an optimal way:

Theorem 3. *Consider the Sobolev class $\mathcal{F}_s([0,1])$ of densities for some $s \geq 1$. Then there are universal constants $0 < c_\ell \leq c_u < \infty$ such that for all $\alpha \in (0, 1/4]$, the (non-interactive) α-private minimax rate* (16) *is sandwiched as*

$$c_\ell \big(\frac{1}{n\alpha^2}\big)^{-\frac{2s}{2s+2}} \;\leq\; \mathfrak{M}_n\left(\mathcal{F}_s([0,1]); \alpha\right) \;\leq\; c_u \big(\frac{1}{n\alpha^2}\big)^{-\frac{2s}{2s+2}}. \tag{20}$$

The private minimax rate (20) differs from the classical rate (18) in two key ways. The effective sample size is reduced from n to $\alpha^2 n$, and more importantly, the

exponent is reduced from $\frac{2s}{2s+1}$ to $\frac{2s}{2s+2}$. Thus, in the case of Lipschitz densities ($s = 1$), the minimax rate changes from $n^{-\frac{2}{3}}$ to $n^{-\frac{1}{2}}$. Consequently, Theorem 3 reveals a fundamental tradeoff between privacy and statistical utility for density estimation.

How is the α-private minimax rate (20) achieved? In order to answer this question, two choices must be made: a choice of the α-private channel that generates the privatized samples Z_1^n, and an estimator that takes the private data as input. It is natural to wonder whether the α-private Laplace mechanism (14)—namely, forming the samples $Z_i = X_i + W_i$ where W_i is α-Laplace noise—combined with the orthogonal series estimate might achieve the optimal private rate. Interestingly, this approach turns out to be highly sub-optimal, as can be established by recourse to known results on nonparametric deconvolution. Given the observation $Z = X + W$, the density of Z is a convolution of the densities of X and W. In their study of the deconvolution problem, Carroll and Hall [10] show that if the additive noise has a characteristic function ϕ_W with tails behaving as $|\phi_W(t)| = \mathcal{O}(|t|^{-a})$ for some $a > 0$, then no method can estimate the s-smooth density of X to accuracy greater than $n^{-2s/(2s+2a+1)}$. Note that the Laplace distribution has a characteristic function with tails decaying as t^{-2}; consequently, as a special case of this result, no estimator based on applying the Laplace mechanism directly to the samples can attain a rate of convergence better than $n^{-2s/(2s+5)}$.

An optimal mechanism for α-private density estimation: This cautionary calculation motivates consideration of privacy mechanisms that are not simply based on direct perturbation of the samples X_1^n, and here we describe one such mechanism that achieves the α-private minimax rate (20). Recall the truncation level $T = n^{\frac{1}{2s+1}}$ from our earlier discussion of the orthogonal series estimator (19). Now consider the T-dimensional vectors

$$\phi(X_i) = \begin{bmatrix} \phi_1(X_i) & \phi_2(X_i) & \cdots & \phi_T(X_i) \end{bmatrix}, \tag{21}$$

defined for each $i = 1, \ldots, n$. These vectors are sufficient statistics for computing the orthogonal series estimator. Accordingly, our goal is to construct a channel $\mathbb{Q}$ with output space $\mathcal{Z} = \mathbb{R}^T$ such that

$$\mathbb{E}[Z_i \mid X_i] = \phi(X_i) \quad \text{for each } i = 1, \ldots, n. \tag{22}$$

Our construction assumes that the orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ is b_0-uniformly bounded, meaning that $\sup_x |\phi_j(x)| \leq b_0 < \infty$ for all $j = 1, 2, \ldots$. Note that many standard bases, among them the trigonometric basis and the Walsh basis, satisfy this type of boundedness condition. For some fixed $b > b_0$ to be specified, the following privacy mechanism takes as input any T-dimensional vector of the form $\tau = \phi(X_i)$ for $i = 1, \ldots, n$, as previously defined in equation (21). It consists of three steps, and returns a vector $Z_i \in \mathbb{R}^T$ that is α-private, and such that the unbiasedness condition (22) holds.

- Given a vector τ in the box $[-b_0, b_0]^T$, form a random vector $\widetilde{\tau} \in \{-b_0, b_0\}^T$ with independently sampled coordinates

$$\widetilde{\tau}_j = \begin{cases} +b_0 & \text{with probability } \frac{1}{2} + \frac{\tau_j}{2b_0}. \\ -b_0 & \text{otherwise.} \end{cases}$$

- Draw a Bernoulli random variable V equal to 1 with probability $e^\alpha/(e^\alpha + 1)$, and then draw $Z_i \in \{-b, b\}^T$ according to

$$Z_i \sim \begin{cases} \text{Uniform on } \left\{z \in \{-b, b\}^T \mid \langle z, \widetilde{\tau} \rangle > 0\right\} & \text{if } V = 1 \\ \text{Uniform on } \left\{z \in \{-b, b\}^T \mid \langle z, \widetilde{\tau} \rangle \leq 0\right\} & \text{if } V = 0. \end{cases} \tag{23}$$

It can be shown that the random vector Z_i is α-differentially private for any initial vector in the box $[-b_0, b_0]^T$, and moreover, the samples (23) are efficiently computable, say by rejection sampling. Iteration of expectation yields

$$\mathbb{E}[Z_i \mid X = x] = c_T \frac{b}{b_0\sqrt{T}} \left(\frac{e^\alpha}{e^\alpha + 1} - \frac{1}{e^\alpha + 1} \right) \phi(x) = c_T \frac{b}{b_0\sqrt{T}} \frac{e^\alpha - 1}{e^\alpha + 1} \phi(x), \tag{24}$$

for a constant c_T bounded away from zero. Consequently, setting $b = \frac{b_0\sqrt{T}}{c_T} \frac{e^\alpha + 1}{e^\alpha - 1}$ ensures that the unbiasedness condition (22) is satisfied.

Based on this α-private mechanism, we can compute the T-dimensional random vector $\widetilde{\theta} := \frac{1}{n} \sum_{i=1}^n Z_i$, which is guaranteed to be an unbiased estimate of the vector $\widehat{\theta} = \frac{1}{n} \sum_{i=1}^n \phi(X_i)$ of empirical basis coefficients. Using this unbiased estimate, we can then form the density estimate $\widetilde{f} = \sum_{j=1}^T \widetilde{\theta}_j \phi_j$. As shown in the paper [17], this estimator achieves the α-private minimax rate (20).

5. Computationally-constrained minimax rates

In the classical definition of minimax risk (1), the infimum is allowed to range over all measurable functions $\widehat{\theta} : \mathcal{X}^n \to \mathbb{R}^n$. In practice, however, one is limited to estimators with computational complexity that scales polynomially in the problem parameters. For this reason, it is natural to consider more refined notions of minimax rate, in which constraints are imposed on the computational complexity of the underlying estimators. For many problems, at least up to constant factors, the classical minimax risk can be achieved by computationally efficient estimators. In these cases, the computationally constrained minimax risk is no different than the classical minimax risk. Thus, such refinements of minimax rates are interesting only when it is possible to establish a gap between the performance of optimal procedures, and that of computationally constrained methods. A recent line of work [4, 29] has established such gaps for testing problems involving sparse and low-rank matrices, working under a conjecture in average-case complexity theory. Here we describe how, under a standard assumption in worst-case complexity theory, such a gap exists for the problem of high-dimensional sparse regression [47].

High-dimensional sparse regression: We begin by describing the problem of sparse regression and discussing some possible estimators, both computationally efficient and inefficient ones. As previously discussed, linear regression is a canonical problem in statistics, in which a response vector $Y \in \mathbb{R}^n$ is related to matrix $X \in \mathbb{R}^{n \times d}$ of covariates via the observation model (3). Given the goal of estimating θ^*, the quality of an estimate $\widehat{\theta}$ can be assessed in various ways. In this discussion, we model the matrix X as a fixed quantity, known as the case of deterministic design, and consider the (in-sample) *prediction error*, as previously defined in equation (4).

Recent years have witnessed intense study of the sparse form of linear regression, in which the unknown regression vector $\theta^* \in \mathbb{R}^d$ is assumed to have at most $k \ll d$ non-zero entries (e.g., see the papers [24, 16, 9, 5, 31, 34, 40, 39] and references therein). The most direct approach to solving a k-sparse instance of the linear regression model (3) is to seek a k-sparse minimizer to the least-squares cost $\|Y - X\theta\|_2^2$. Doing so leads to the ℓ_0-based estimator

$$\widehat{\theta}_{\ell_0} := \arg \min_{\theta \in \mathbb{B}_0(k)} \|Y - X\theta\|_2^2. \tag{25}$$

Note that this estimator returns an estimate that belongs to the ℓ_0-"ball"

$$\mathbb{B}_0(k) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} \mathbb{I}[\theta_j \neq 0] \leq k\big\} \tag{26}$$

of k-sparse vectors. More generally, given an estimator $\widetilde{\theta}$, we say that it belongs to class $\mathcal{A}(k)$ if its output always belongs to $\mathbb{B}_0(k)$.

The following result [8, 34] provides an upper bound on the prediction error performance of the ℓ_0-based estimator:

Proposition 2 (Prediction error for $\widehat{\theta}_{\ell_0}$). *There is a universal constant c_0 such that for any design matrix X, the ℓ_0-based estimator $\widehat{\theta}_{\ell_0}$ satisfies*

$$\max_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{1}{n}\|X(\widehat{\theta}_{\ell_0} - \theta^*)\|_2^2\Big] \leq c_0 \frac{\sigma^2 k \log d}{n}. \tag{27}$$

Moreover, Raskutti et al. [34] establish a lower bound that is matching up to constant factors, showing that this bound is unimprovable when $k \ll d$. A notable aspect of the upper bound (27) is that it holds for any design matrix $X \in \mathbb{R}^{n \times d}$.

Thus, in terms of the classical minimax risk (1), the ℓ_0-based estimator is an optimal method. However, it is unattractive from a computational point of view. A brute force approach requires iterating over all $\binom{d}{k}$ subsets of size k, and Natarajan [32] shows that computing a sparse solution to a set of linear equations is an NP-hard problem. Given this intractability, it is natural to consider the performance of computationally efficient methods.

Prediction guarantees for ℓ_1-based methods: Convex relaxation is a standard method for replacing a combinatorial constraint—in this case, the requirement

that θ have at most k non-zero entries—with a looser but convex constraint. A familiar relaxation of the ℓ_0-constraint is to replace it with an ℓ_1-norm. For concreteness, we consider a Lagrangian form of the ℓ_1-relaxation, which leads to the *Lasso estimator* [37, 13]

$$\widehat{\theta}_{\ell_1} = \arg\min_{\theta \in \mathbb{R}^d} \big\{ \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \big\}. \tag{28}$$

In contrast to the ℓ_0-based estimator (25), it is easy to compute the Lasso estimate. Indeed, a quadratic program of the form (28) can be solved to δ-accuracy in time polynomial in the problem parameters, and $\log(1/\delta)$, by various standard optimization methods (e.g., see the books [7, 33]).

Is the Lasso estimator (28) is an optimal method? For some error metrics, including the ℓ_2-norm error $\|\widehat{\theta} - \theta^*\|_2$, it can be shown that the Lasso is essentially an optimal method, in terms of matching the classical minimax risk [34]. However, for the prediction error (4), the best known results on the Lasso fail to match the ℓ_0-guarantee. In particular, in contrast to the ℓ_0-estimate (25), the best known results on either the Lasso [5], or the closely related Dantzig selector [9], all involve constraints known as (sparse) restricted eigenvalue (RE) conditions, which we define next.

Restricted eigenvalues are defined in terms of subsets S of the index set $\{1, 2, \ldots, d\}$, and a cone associated with any such subset. In particular, letting S^c denote the complement of S, we define the cone $\mathbb{C}(S) := \big\{\theta \in \mathbb{R}^d \mid \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1\big\}$. Here $\|\theta_{S^c}\|_1 := \sum_{j \in S^c} |\theta_j|$ corresponds to the ℓ_1-norm of the coefficients indexed by S^c, with $\|\theta_S\|_1$ defined similarly.

Definition 4 (Sparse restricted eigenvalue). The matrix $X \in \mathbb{R}^{n \times d}$ is said to satisfy a *uniform γ-RE* condition if

$$\frac{1}{n}\|X\theta\|_2^2 \geq \gamma \|\theta\|_2^2 \qquad \text{for all } \theta \in \bigcup_{|S|=k} \mathbb{C}(S). \tag{29}$$

The restricted eigenvalue constant of X, denoted by $\gamma(X)$, is the greatest γ such that X satisfies the condition (29). The RE condition (29) and related quantities have been studied extensively in past work on basis pursuit and the Lasso (e.g., [5, 9, 31, 34]). van de Geer and Buhlmann [39] provide an overview of the different types of RE parameters, and their relationships. The following result [5] provides an upper bound on the Lasso prediction error under a sparse RE condition:

Proposition 3 (Prediction error for Lasso). *There is a universal constant c_1 such that for any column-normalized design matrix X with a RE constant $\gamma(X) > 0$, the Lasso estimate $\widehat{\theta}_{\ell_1}$ satisfies*

$$\max_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{1}{n}\|X(\widehat{\theta}_{\ell_1} - \theta^*)\|_2^2\Big] \leq \frac{c_1}{\gamma^2(X)} \frac{\sigma^2 k \log d}{n}. \tag{30}$$

Apart from the difference in universal constants, the key difference between the ℓ_1-guarantee and the ℓ_0-guarantee is that the RE constant $\gamma^2(X)$ appears in the Lasso bound (30), but is absent from the ℓ_0-bound (27). It is natural to wonder whether it might be possible to prove a sharper bound on the Lasso, not involving the RE constant. From a fundamental point of view, given the goal of minimizing the prediction risk (4), the restricted eigenvalues of X should not be relevant. For instance, duplicating two rows of X would force the RE constant to zero, but would not make the underlying prediction problem any more difficult. We are thus left with two possibilities:

- either the analysis leading to the bound (30) is not sharp, and could be improved;
- or the appearance of the RE constant is unavoidable for an ℓ_1-based method.

Our recent work shows that in fact, the second option is correct, and even more generally, the appearance of the RE constant is intrinsic to the class of polynomial-time estimators.

Computationally-constrained minimax risk: In order to state our main result, we need to make precise a particular notion of a polynomial-efficient estimator. Since the observation (X, Y) consists of real numbers, any efficient algorithm can only take a finite-length representation of the input. Consequently, we begin by introducing an appropriate notion of discretization. For any input value x and integer τ, the operator

$$\lfloor x \rfloor_\tau := 2^{-\tau} \lfloor 2^\tau x \rfloor \tag{31}$$

represents a $2^{-\tau}$-precise quantization of x. (Here $\lfloor u \rfloor$ denotes the largest integer smaller than or equal to u.) Given a real value x, an efficient estimator is allowed to take $\lfloor x \rfloor_\tau$ as its input for some finite choice τ. We denote by $\text{size}(x; \tau)$ the length of binary representation of $\lfloor x \rfloor_\tau$, and denote by $\text{size}(X, Y; \tau)$ the total length of the discretized matrix vector pair (X, Y).

The following definition of computational efficiency is parameterized in terms of three quantities: (i) a positive integer b, corresponding to the number of bits required to implement the estimator as a computer program; (ii) a polynomial function G of the triplet (n, d, k), corresponding to the discretization accuracy of the input, and (iii) a polynomial function H of input size, corresponding to the runtime of the program.

Definition 5 (Polynomial-efficient estimators). Given a pair of polynomial functions $G : (\mathbb{Z}_+)^3 \to \mathbb{R}_+$, $H : \mathbb{Z}_+ \to \mathbb{R}_+$ and a positive integer $b \in \mathbb{Z}_+$, an estimator $(Y, X) \mapsto \widehat{\theta}(Y, X)$ is said to be (b, G, H)-*efficient* if:

- It can be represented by a computer program that is encoded in b bits.
- For every problem of scale (n, d, k), it accepts inputs quantized to accuracy $\lfloor \cdot \rfloor_\tau$ where the quantization level is bounded as $\tau \leq G(n, d, k)$.

- For every input (X, Y), it is guaranteed to terminate in time $H(\text{size}(X, Y; \tau))$.

In computational complexity theory, the class **POLY** corresponds to problems that are solvable in polynomial time by a Turing machine. A closely related class denoted by **PPOLY**, corresponds to all problems solvable in polynomial time by a Turing machine with a so-called advice string—meaning a side-input to the machine—that is of polynomial length. The class **PPOLY** is strictly bigger than the class **POLY** (e.g, [1]); however, it is widely believed that **NP** $\not\subset$ **PPOLY**, and the following result is stated using this inclusion as an assumption. Moreover, we use $c_j, j = 2, 3$ to denote universal constants independent of the scaling parameters (n, d, k), polynomials (F, G, H) and constants (γ, σ, δ).

Theorem 6. *If* **NP** $\not\subset$ **PPOLY**, *then for any positive integer* b, *any scalar* $\delta \in (0, 1)$, *any polynomial functions* $G : (\mathbb{Z}_+)^3 \to \mathbb{R}_+$ *and* $F, H : \mathbb{Z}_+ \to \mathbb{R}_+$, *there is a sparsity level* $k \geq 1$ *such that the following holds:*

For any dimension $d \in [4k, F(k)]$, *any sample size* n *in the interval* $[c_2 k \log d, F(k)]$, *and any scalar* $\gamma \in [2^{-G(n,d,k)}, \frac{1}{24\sqrt{2}})$, *there is a matrix* $X \in \mathbb{R}^{n \times d}$ *such that:*

(a) It has an RE constant $\gamma(X)$ *that is bounded as* $|\gamma(X) - \gamma| \leq 2^{-G(n,d,k)}$.

(b) For any (b, G, H)*-efficient estimator* $\widehat{\theta} \in \mathcal{A}(k)$, *the mean-squared prediction error is lower bounded as*

$$\max_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{\|X(\widehat{\theta} - \theta^*)\|_2^2}{n}\Big] \geq \frac{c_3}{\gamma^2} \frac{\sigma^2 k^{1-\delta} \log d}{n}. \tag{32}$$

Disregarding technical aspects regarding quantization, the essential part of the theorem is that the lower bound grows inversely with the squared RE constant γ^2. Consequently, within the class of polynomial-time methods, the Lasso is an optimal method, but faster rates can be obtained using algorithms with exponential-time complexity. We note that Theorem 6 is restricted to methods that return k-sparse estimates—that is, belong to the class $\mathcal{A}(k)$. It is an open question as to whether analogous lower bounds can be established without this requirement.

References

[1] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[2] M. F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity and privacy. http://arxiv.org/abs/1204.3514, 2012.

[3] 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.

[4] Q. Berthet and P. Rigollet. Computational lower bounds for sparse PCA. Technical report, Princeton University, April 2013. arxiv1304.0828.

[5] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[6] L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. *Z. Wahrsch. verw. Gebiete*, 65:181–327, 1983.

[7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[8] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

[9] E. J. Candés and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[10] R. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

[11] K. Chaudhuri and D. Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[12] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[13] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[14] A. De. Lower bounds in differential privacy. In *Proceedings of the Ninth Theory of Cryptography Conference*, 2012.

[15] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

[16] D. L. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[17] J. C. Duchi, M. J. Wainwright, and M. I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. Technical report, UC Berkeley, 2013.

[18] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18, 1986.

[19] G. T. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7(2):207–217, 1989.

[20] 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.

[21] 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.

[22] 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.

[23] I. P. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1972.

[24] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

[25] 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.

[26] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[27] 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.

[28] A. Kolmogorov and B. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Uspekhi Mat. Nauk.*, 86:3–86, 1959. Appeared in English as Amer. Math. Soc. Translations, 17:277–364, 1961.

[29] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *arXiv preprint arXiv:1309.5914*, 2013.

[30] 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.

[31] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

[32] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 24(2):227–234, 1995.

[33] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, New York, 2004.

[34] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Information Theory*, 57(10):6976—6994, October 2011.

[35] 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.

[36] C. J. Stone. Optimal global rates of convergence for non-parametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982.

[37] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[38] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

[39] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[40] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

[41] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[42] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[43] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[44] B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.

[45] Y. Zhang, J. C. Duchi, , M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. Technical report, UC Berkeley, 2013. Presented at the NIPS Conference 2013.

[46] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, November 2013.

[47] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. Technical report, UC Berkeley, 2014. arXiv:1402.1918.

Martin J. Wainwright

E-mail: wainwrig@berkeley.edu