Privacy Aware Learning

JOHN C. DUCHI, MICHAEL I. JORDAN, and MARTIN J. WAINWRIGHT,
University of California, Berkeley

We study statistical risk minimization problems under a privacy model in which the data is kept confidential even from the learner. In this local privacy framework, we establish sharp upper and lower bounds on the convergence rates of statistical estimation procedures. As a consequence, we exhibit a precise tradeoff between the amount of privacy the data preserves and the utility, as measured by convergence rate, of any statistical estimator or learning procedure.

Categories and Subject Descriptors: G.1.6 [**Numerical Analysis**]: Optimization—*Convex programming; gradient methods*; G.3 [**Probability and Statistics**]: *Nonparametric statistics*; H.1.1 [**Models and Principles**]: Systems and Information Theory—*Information theory*; I.2.6 [**Artificial Intelligence**]: Learning—*Parameter learning*; K.4.1 [**Computers and Society**]: Public Policy Issue—*Privacy*

General Terms: Algorithms, Theory, Security

Additional Key Words and Phrases: Differential privacy, lower bounds, machine learning, minimax convergence rates, saddle points

ACM Reference Format:
John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2014. Privacy aware learning. J. ACM 61, 6, Article 38 (November 2014), 57 pages.
DOI: http://dx.doi.org/10.1145/2666468

38

1. INTRODUCTION

Natural tensions between learning and privacy arise whenever a learner must aggregate data across multiple individuals. The learner wishes to make optimal use of each data point, whereas the providers of the data may wish to limit detailed exposure, either to the learner or to other individuals. A characterization of such tensions in the form of quantitative tradeoffs is of great utility: it can inform public discourse surrounding the design of systems that learn from data, and the tradeoffs can be exploited as controllable degrees of freedom whenever such a system is deployed.

In this article, we approach this problem from the point of view of statistical decision theory. The decision-theoretic perspective offers a number of advantages. First, the use of loss functions and risk functions provides a compelling formal foundation for defining "learning", one that dates back to Wald [1939], and that has seen continued development in the context of research on machine learning over the past two decades. Second, by formulating the goals of a learning system in terms of loss functions, we make it

---

A preliminary version of this article appeared in *Neural Information Processing Systems 2012*.
This material supported in part by ONR MURI grant N00014-11-1-0688 and the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-11-1-0391. J. C. Duchi was supported by a fellowship from the National Defense Science and Engineering Graduate Fellowship Program (NDSEG) and additionally by a Facebook Ph.D. fellowship.
Authors' present addresses: J. C. Duchi, Stanford University, 450 Serra Mall, Stanford, CA 94305; email: jduchi@stanford.edu; M. I. Jordan and M. J. Wainwright, Departments of EECS and Statistics, University of California, Berkeley, Berkeley, CA, email: jordan@cs.barkeley.edu; wainwrig@eecs.barkeley.edu.

DOI: http://dx.doi.org/10.1145/2666468

possible for individuals to assess whether the goals of a learning system align with their own personal utility, and thereby determine the extent to which they are willing to sacrifice some privacy. Third, an appeal to decision theory permits abstraction over the details of specific learning procedures, allowing for the derivation of minimax lower bounds that apply to any specific procedure. Fourth, the use of loss functions—and more specifically, convex loss functions—in the design of a learning system allows the powerful tools of optimization theory to be brought to bear. Not only are optimization-based learning systems often successful in practice, but they are also often amenable to theoretical analysis. Finally, the decision-theoretic framework is a probabilistic framework, with probabilistic models defining the transformation from losses to risks. This connection provides a natural mechanism for the use of randomization to provide control over privacy.

In more formal detail, the analysis of this paper takes place within the following framework. Given a compact convex set $\Theta \subset \mathbb{R}^d$, we wish to find a parameter value $\theta \in \Theta$ achieving good average performance under a loss function $\ell : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}_+$. Here the value $\ell(X, \theta)$ measures the performance of the parameter vector $\theta \in \Theta$ on the sample $X \in \mathcal{X}$, and $\ell(x, \cdot) : \mathbb{R}^d \to \mathbb{R}_+$ is convex for $x \in \mathcal{X}$. We measure the expected performance of $\theta \in \Theta$ via the risk function

$$\theta \mapsto R(\theta) := \mathbb{E}_P[\ell(X, \theta)], \tag{1}$$

where the expectation is taken over some unknown distribution P over the space $\mathcal{X}$.

In the standard formulation of statistical risk minimization, a method $\mathcal{M}$ is given n samples $X_1, \ldots, X_n$, each drawn independently from P, and its goal to to output an estimate $\widehat{\theta}_n$ that approximately minimizes the risk function R. In this article, instead of providing the method $\mathcal{M}$ with access to the samples $X_1, \ldots, X_n$, however, we study the effect of giving only some disguised view Z_i of each datum X_i. With $\widehat{\theta}_n$ now denoting an estimator based on the perturbed samples Z_i, we explicitly quantify the rate of convergence of $R(\widehat{\theta}_n)$ to $\inf_{\theta \in \Theta} R(\theta)$ as a function of the number of samples n and the amount of privacy provided by Z_i.

1.1. Prior Work

There is a long history of research at the intersection of privacy and statistics, going back at least to the 1960s, when Warner [1965] suggested privacy-preserving methods for survey sampling, and to later work related to census taking and presentation of tabular data (e.g., Fellegi [1972]). More recently, there has been a large amount of computationally oriented work on privacy [Dwork et al. 2006; Dwork 2008; Zhou et al. 2009a; Wasserman and Zhou 2010; Hall et al. 2011; Dinur and Nissim 2003; Blum et al. 2008; Chaudhuri et al. 2011; Rubinstein et al. 2012]. We overview some of the key ideas in this section, but cannot hope to do justice to the large body of relevant work, referring the reader to the comprehensive survey by Dwork [2008] and the statistical treatment by Wasserman and Zhou [2010] for background and references.

Most work on privacy attempts to limit disclosure risk: the probability that some adversary can link a released record to a particular member of the population or identify that someone belongs to a dataset that generates a statistic [Duncan and Lambert 1986, 1989; Reiter 2005; Karr et al. 2006]. In the statistical literature, work on disclosure limitation and so-called linkage risk, for example as in the framework of Duncan and Lambert [1986], has yielded several techniques for maintaining privacy, such as aggregation, swapping features or responses among different datums, or perturbation of data. Other authors have proposed measures for measuring utility of released data (e.g., Karr et al. [2006] and Cox et al. [2011]). The currently standard measure of privacy is differential privacy, due to Dwork et al. [2006], which roughly states that $\widehat{\theta}_n$ must not depend too much on the n samples, and it should be difficult to ascertain whether a vector x belongs to the set $\{X_1, \ldots, X_n\}$ given $\widehat{\theta}_n$. Formally, paraphrasing the definition of Wasserman and Zhou [2010], the method $\mathcal{M}$ has α-differential privacy if

$$\sup_{S \in \sigma(\Theta)} \sup_{x_1,\ldots,x_n} \sup_{x'_1,\ldots,x'_n} \frac{Q(S \mid X_1 = x_1, \ldots, X_n = x_n)}{Q(S \mid X_1 = x'_1, \ldots, X_n = x'_n)} \le \exp(\alpha). \tag{2}$$

where the sets $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ differ in at most one element, $Q(\cdot \mid X_1, \ldots, X_n)$ is (a version of) the conditional probability of the estimator $\widehat{\theta}$ constructed by the method $\mathcal{M}$ using the n samples, and $\sigma(\Theta)$ is a suitable σ-algebra on Θ.

Differentially private algorithms enjoy many desirable properties [Dwork et al. 2006; Dwork 2008; Ganta et al. 2008] and essentially guarantee that even if an adversary knows all the entries in a dataset but the nth, it is difficult to discern whether a vector x is equal to X_n given the output of the method $\mathcal{M}$. Indeed, differential privacy protects against side information and many adversarial attacks that break previous definitions of privacy, such as k-anonymity [Ganta et al. 2008]. Several researchers have studied differentially private algorithms for empirical risk minimization, providing guarantees on the excess risk of differentially private estimators $\widehat{\theta}$. Chaudhuri et al. [2011] use the stability of the output of regularized empirical risk minimization algorithms to show that by adding Laplace-distributed noise to an empirical estimator θ or by adding an additional random term to the empirical risk $\frac{1}{n}\sum_{i=1}^n \ell(X_i, \theta)$, it is possible to obtain differential privacy and consistency of $\widehat{\theta}$. Dwork and Lei [2009] obtain similar results using robust statistical estimators, and Smith [2011] shows that if one has suitably unbiased estimators, then differential privacy is possible without compromising asymptotic rates of convergence. Rubinstein et al. [2012] use similar stability and perturbation techniques to demonstrate that it is possible to obtain differential privacy when solving support vector machine problems, and also show that if the desired privacy level α in the definition (2) is too small, it is actually impossible to obtain a parameter $\widehat{\theta}_n$ minimizing the risk R.

Our goal is to understand the fundamental tradeoffs between maintaining privacy while still providing a useful output from the statistical learning procedure $\mathcal{M}$. Though intuitively there must be some tradeoff, quantifying it precisely has been difficult. As alluded to above, Rubinstein et al. [2012] are able to show that it is impossible to obtain what they call an (ϵ, δ)-useful parameter vector θ that enjoys any differential privacy guarantees; however, it is unknown whether or not their guarantees might be improvable. Hall et al. [2011] show that if a given histogram, based on a sample $\{x_i\}_{i=1}^n$, has d bins and we must guarantee α-differential privacy (2), then the (expected) L^1-distance between the sample and released histograms must be at least $d/(n\alpha)$, and Hardt and Talwar [2010] give similar lower bounds on the amount of noise necessary to answer linear database queries. Nikolov et al. [2013] followed this work with extensions to relaxed notions (so called (α, δ)-approximate differential privacy) of privacy and providing higher-dimensional settings, while Kasiviswanathan et al. [2013] give bounds on amounts of additive noise to protect against blatant failures of privacy in similar linear settings. Blum et al. [2008] also give lower bounds on the closeness of certain statistical quantities computed from the dataset, though their upper and lower bounds do not match. Sankar et al. [2010] provide rate-distortion theorems for utility models involving information-theoretic quantities, which has some similarity to our risk-based framework, but it appears somewhat challenging to explicitly map their setting onto ours. With the goal of characterizing what it means to be both useful and private, Ghosh et al. [2009] show that for a one-time computation of counts on a dataset $X_1, \ldots, X_n$ (i.e., the number of variables satisfying $X_i \in C$ for some set C), perturbing the output of a counting function using geometrically distributed noise is the unique optimal way to guarantee differential privacy while maximizing a natural notion of utility.

Much of the work providing sharp lower bounds, however, focuses on showing that if one wishes to accurately report a statistic $\widehat{\theta}(x_{1:n})$ computed on a sample $\{x_i\}_{i=1}^n$, then there must be some worst-case sample such that the error is large (see, e.g., Hardt and Talwar [2010], Hall et al. [2011], Nikolov et al. [2013], and Ghosh et al. [2009]). In contrast, we focus on *population* quantities—which are substantially different—in that we wish to return a private estimator $\widehat{\theta}_n$ approximately minimizing the population risk $R(\theta) = \mathbb{E}[\ell(X, \theta)]$ rather than the sample risk $\frac{1}{n}\sum_{i=1}^n \ell(X_i, \theta)$. Providing guarantees on the population risk performance of $\widehat{\theta}_n$, rather than on the observed sample, has been a driving force behind much of the theoretical work in statistics and machine learning, and thus provides a natural focus for our work.

1.2. Our Setting

In contrast to this work, we study a more local notion of privacy [Evfimievski et al. 2003; Kasiviswanathan et al. 2011], in which each datum X_i is kept private from the method $\mathcal{M}$. The goal of many types of privacy is to guarantee that the output $\widehat{\theta}_n$ of the method $\mathcal{M}$ based on the data cannot be used to discover information about the individual samples $X_1, \ldots, X_n$, but *locally private* algorithms only access disguised views of each datum X_i. Local algorithms are among the most classical approaches to privacy, tracing back to work on randomized response in the statistical literature [Warner 1965], and rely on communication only of some disguised view Z_i of each true sample X_i. In this setting, for example, the natural variant of α-differential privacy (2) is the noninteractive (in the sense that Z_i depends only on X_i and not on any other private variables Z_j) local privacy guarantee

$$\sup_S \sup_{x,x'} \frac{Q(Z_i \in S \mid X_i = x)}{Q(Z_i \in S \mid X_i = x')} \le \exp(\alpha). \tag{3}$$

Locally private algorithms are natural when the providers of the data—the population sampled to give $X_1, \ldots, X_n$—do not even trust the statistician or statistical method $\mathcal{M}$, but the providers are interested in the parameter vector θ^* that minimizes the risk function. For example, in medical applications, a participant may be embarrassed about his use of drugs, or perhaps about his marital status, but if the loss ℓ is able to measure the likelihood of developing cancer, then the participant has high utility for access to the optimal parameters θ^*. Internet applications, where a user's activity is logged across multiple websites or searches, provide another example: the user has a utility for a search engine to have a ranking function θ that returns relevant results for web searches, yet may not wish to reveal his or her search data. In essence, we would like the statistical procedure $\mathcal{M}$ to learn *from* the data $X_1, \ldots, X_n$ but not *about* it.

The work most related to ours seems to be that of Kasiviswanathan et al. [2011], who show that that (in some settings) locally private algorithms coincide with concepts that can be learned with polynomial sample complexity in Kearns's statistical query (SQ) model [Kearns 1998]. This result is powerful, but has some limitations, as the statistical query model relies exclusively on count queries, and we are interested in measures more precise than polynomial sample complexity to quantity convergence rates. In contrast, our analysis applies to estimators deriving from a broad class of convex risks (1), and it provides sharp rates of convergence.

We develop our approach to local privacy in the setting of three related privacy measures. The first is a worst-case measure of mutual information, where we view privacy preservation as a game between the providers of the data, who wish to preserve privacy, and nature. The second is based on differential privacy, where the provider of each datum communicates—subject to some constraints we make explicit later—the *most* differentially private view Z_i of his or her datum X_i. In this general setting we allow interactivity (i.e., the mapping between Z_i and X_i may depend on other Z_j for $j \neq i$). The third setting is a noninteractive version of local differential privacy.

Turning first to the information-theoretic formulation, and recalling that the method $\mathcal{M}$ sees only the perturbed version Z_i of X_i, we use a uniform variant of mutual information $I(Z_i; X_i)$ between the random variables X_i and Z_i as a measure for privacy. Using mutual information and related information-theoretic ideas in the privacy and security context is by no means original; see, for example, the survey by Liang et al. [2008]. It is important to note, however, that standard mutual information has deficiencies as a measure of privacy (e.g., Evfimievski et al. [2003]). Accordingly, our uniform notion of mutual information is as follows: we say that the distribution Q generating Z from X is private only if $I(X; Z)$ is small for all possible distributions P on X, possibly subject to some constraints.

In this setting, we design procedures that allow consistent estimation of the parameter θ^* minimizing $R(\theta) = \mathbb{E}_P[\ell(X, \theta)]$, for any convex loss ℓ and distribution P on the data X. One central consequence of our analysis is a sharp characterization of the *excess risk*,

$$\Delta_n(\widehat{\theta}; \ell, \Theta) := \mathbb{E}[R(\widehat{\theta}(Z_1, \ldots, Z_n))] - \inf_{\theta \in \Theta} R(\theta), \tag{4}$$

associated with any estimator $\widehat{\theta}$ that satisfies a pre-specified privacy constraint. For particular collections $\mathfrak{L}$ of loss functions $\ell \in \mathfrak{L}$, we bound the minimax convergence rate of all estimation procedures. More precisely, let us focus on d-dimensional problems, that is, those for which the domain $\Theta \subset \mathbb{R}^d$ and observations $X_i \in \mathbb{R}^d$. If one wishes to uniformly guarantee a level of privacy $I(X_i; Z_i) \leq I^*$, then we show that there exists a constant $a(\mathfrak{L}, \Theta) \in \mathbb{R}_+$—dependent only on the properties of the collection $\mathfrak{L}$ and domain Θ—such that *for any* estimator $\widehat{\theta}$ for the family $\mathfrak{L}$, the excess risk is lower bounded as

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\widehat{\theta}; \ell, \Theta) \geq \frac{\sqrt{d}}{\sqrt{I^*}} \frac{a(\mathfrak{L}, \Theta)}{\sqrt{n}}, \tag{5a}$$

where $a(\mathfrak{L}, \Theta)$ is a constant characterizing the non-private minimax rate of estimation (see, e.g., Agarwal et al. [2012] for such constants). Moreover, we also prove that there exists another constant $b(\mathfrak{L}, \Theta) \geq a(\mathfrak{L}, \Theta)$ and provide explicit estimators $\widehat{\theta}$ with privacy guarantee I^* such that

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\widehat{\theta}; \ell, \Theta) \leq \frac{\sqrt{d}}{\sqrt{I^*}} \frac{b(\mathfrak{L}, \Theta)}{\sqrt{n}}. \tag{5b}$$

Turning to the setting of differential privacy, we are able to show similar results to the bounds (5a) and (5b). Namely, there exist constants $b'(\mathfrak{L}, \Theta) \geq a'(\mathfrak{L}, \Theta)$ such that if we wish to guarantee α-differential privacy, then for any estimator $\widehat{\theta}$, the risk is lower bounded by

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\widehat{\theta}; \ell, \Theta) \geq \frac{\sqrt{d}}{\alpha} \frac{a'(\mathfrak{L}, \Theta)}{\sqrt{n}}, \tag{6a}$$

while there exist estimators $\widehat{\theta}$ such that

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\widehat{\theta}; \ell, \Theta) \leq \frac{\sqrt{d}}{\alpha} \frac{b'(\mathfrak{L}, \Theta)}{\sqrt{n}}. \tag{6b}$$

Here again, the constant $a'(\mathfrak{L}, \Theta)$ controls the nonprivate minimax rate of estimation.

Table I. Main Sample Complexity Results

	Optimal local privacy at level I^* (Def. 2.1)	Optimal local differential privacy at level α (Def. 2.2)	Noninteractive local α-differential privacy (inequality (3))
Necessary sample size relative to n	$\frac{d}{I^*}n$	$\frac{d}{\alpha^2}n$	$\frac{d}{\alpha^2}n$

Summary of sample complexities: Each entry gives the effective sample size for optimization under the specified privacy constraint. That is, if in the nonprivate case, n observations are required to achieve a fixed excess risk ϵ (see Definition (4)), the table entry shows the number of observations necessary to achieve excess risk ϵ under the specified privacy constraint.

Finally, we show that stochastic gradient descent is one procedure that achieves the above upper bounds, and moreover, that the ratios $b(\mathfrak{L}, \Theta)/a(\mathfrak{L}, \Theta)$ and $b'(\mathfrak{L}, \Theta)/a'(\mathfrak{L}, \Theta)$ are bounded above by a universal (numerical) constant. The bounds (5) and (6) thus establish and quantify explicitly the sharp tradeoff between learning and statistical estimation and the amount of privacy provided to the population. More concretely, we can evaluate the effective sample size of learning procedures receiving private observations. In the case of information-based privacy, the sample size of any learning procedure receiving maximally privatized observations from the data providers is decreased from n to roughly nI^*/d, while in differentially private settings, we see that the effective sample size decreases from n to $n\alpha^2/d$. The first of these is perhaps intuitive: in rough terms, a d-dimensional observation X_i contains about d-bits of information, and so we expect a loss in (statistical) efficiency by a factor I^*/d. For the second, recent results suggest scalings of $I^* \approx \alpha^2$ (e.g., Dwork et al. [2010, Lemma 3.2]), so the loss $n \mapsto n\alpha^2/d$ is also perhaps intuitive. Table I summarizes these sample complexity tradeoffs.

Our subsequent analysis will build on this favorable property of gradient-based methods. Indeed, in the remainder of this article, we will assume that the communication protocol—except in the noninteractive α-differentially private case, which allows any protocol—by which data is conveyed to the learner $\mathcal{M}$ is based on (sub)gradients of the loss. As further motivation for this choice, note that the subgradient (more generally, a score function) of the loss ℓ is asymptotically sufficient in the sense of Le Cam [1956]. A bit more precisely, gradients (in an asymptotic sense) contain all of the statistical information for risk minimization problems. Second, estimation procedures based on stochastic gradient information are asymptotically efficient [Polyak and Juditsky 1992], in the sense of both Bahadur and minimax efficiency [van der Vaart 1998, Chapter 8], and are thus essentially sample optimal; they also have minimax-optimality guarantees in finite-sample settings [Agarwal et al. 2012]. Moreover, many estimation procedures are gradient-based [Nemirovski and Yudin 1983; Boyd and Vandenberghe 2004], and distributed optimization procedures that send subgradient information across a network to a centralized procedure $\mathcal{M}$ are natural (e.g., Bertsekas and Tsitsiklis [1989]). Our arguments also show that in many settings, disguising subgradients is equivalent to disguising the data X itself. Thus, as an additional consequence of our gradient-based focus, our algorithmic bounds also apply in streaming and online settings, requiring only a fixed-size memory footprint.

1.3. Outline and Techniques

We spend the remainder of the article deriving the bounds (5) and (6). Our route to obtaining these bounds is based on a two-part analysis. First, we consider saddle points of the mutual information $I(X; Z)$, when viewed as a function of the distribution P of X and the conditional distribution $Q(\cdot \mid X)$ of Z, under natural constraints that still allow estimation. We consider related saddle points for differentially private conditional distributions. Having computed these saddle points, we can apply information-theoretic techniques for obtaining lower bounds on estimation and optimization [Yang and Barron 1999; Agarwal et al. 2012] to prove the results of the form (5a) or (6a). Our upper bounds then follow by application of known convergence rates for computationally efficient methods, such as the stochastic gradient and mirror descent algorithms [Nemirovski and Yudin 1983; Nemirovski et al. 2009]. We provide full proofs—except for technical results deferred to appendices—and a more complete outline of our technique in Section 5.

The remainder of this article is organized as follows. We give a precise definition of our notions of local privacy in Section 2. Section 3 is devoted to information-theoretic lower bounds on the convergence rate of any statistical method $\mathcal{M}$ in terms of the mutual information I^* between what the method $\mathcal{M}$ observes and each sample X_i. We characterize the unique privacy guaranteeing distributions in Section 4, which provides a constructive mechanism for trading off privacy and learning. We present our conclusions in Section 6.

Notation. Before continuing, we give our notation and a few standard definitions. The *Kullback-Leibler (KL)* divergence between distributions P and Q defined on a set S, where P and Q are assumed to have densities p and q with respect to a base measure ν[1] is given by

$$D_{\mathrm{kl}}(P \| Q) := \int_S p(s) \log \frac{p(s)}{q(s)} d\nu(s).$$

Similarly, the *total-variation distance* between the distributions P and Q is defined as

$$\|P - Q\|_{\mathrm{TV}} := \sup_{A \subset S} |P(A) - Q(A)| = \frac{1}{2} \int_S |p(s) - q(s)| d\nu(s).$$

For a convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the subgradient set $\partial f(\theta)$ of f at the point θ is

$$\partial f(\theta) := \left\{ g \in \mathbb{R}^d : f(\theta') \geq f(\theta) + g^\top(\theta' - \theta), \text{ for all } \theta' \in \mathbb{R}^d \right\}.$$

We use $\partial \ell(x, \theta)$ to denote the subgradient set of the function $\theta \mapsto \ell(x, \theta)$, and for a convex function, $\nabla \ell(x, \theta)$ denotes an arbitrary element of $\partial \ell(x, \theta)$. We say that a function f is L-Lipschitz with respect to the norm $\|\cdot\|$ over the set Θ if

$$|f(\theta) - f(\theta')| \leq L\|\theta - \theta'\| \quad \text{for all } \theta, \theta' \in \Theta.$$

The notation $\|\cdot\|_p$ denotes a standard ℓ_p-norm. We use the abbreviation r.c.d. throughout for regular conditional distribution [Billingsley 1986]. The extreme points of a set $C \subset \mathbb{R}^d$ are denoted by $\mathrm{Ext}(C)$, the convex hull of C is denoted by $\mathrm{Conv}(C)$, and the support of a distribution P is denoted $\mathrm{supp}\, P$. We say values $a_n \asymp b_n$ if $\lim_n (a_n/b_n) = 1$. The symbol e_i denotes the ith standard basis vector in $\mathbb{R}^d$. Lastly, the symbol $\rightrightarrows$ denotes a set-valued mapping [Hiriart-Urruty and Lemaréchal 1996].

2. PROBLEM FORMULATION

We begin with a formal description of the communication protocol by which information about the random variables X is communicated to the procedure $\mathcal{M}$. We then define the notion of *optimal local privacy* studied in this article and the minimax framework in which we state our main results.

[1]This is no loss of generality, as P and Q are absolutely continuous with respect to $\nu = \frac{1}{2}(P + Q)$.

2.1. Communication Protocol

In this article, we focus on statistical learning procedures that have access to data through the subgradients $\partial\ell(X, \theta)$ of the loss functions. More formally, at each round, the method $\mathcal{M}$ is given access to a random vector Z_i such that

$$\mathbb{E}[Z_i \mid X_i, \theta] \in \partial\ell(X_i, \theta), \tag{7}$$

where $\theta \in \Theta$ is a parameter chosen by the method. In Appendix A, we present an argument that shows that the unbiasedness of the subgradient inclusion (7) is not only intuitively appealing but is, in a certain sense, necessary.

In detail, our communication protocol consists of the following three steps:

—the method $\mathcal{M}$ sends the parameter vector θ to the owner of the ith sample X_i;
—owner i computes a subgradient vector $g \in \partial\ell(X_i, \theta)$ to be communicated privately;
—the vector Z_i is communicated to $\mathcal{M}$ under the constraint that

$$\mathbb{E}[Z_i \mid X_i, \theta] = g \in \partial\ell(X_i, \theta).$$

We assume throughout that there is a compact set $C \subset \mathbb{R}^d$ such that $\partial\ell(x, \theta) \subseteq C$ for all pairs $(\theta, x) \in \Theta \times \mathcal{X}$. Our goal is "disguise" the subgradient information with a random variable Z satisfying $Z \in D$, for some compact set D such that $C \subset \operatorname{int} D \subset \mathbb{R}^d$. For instance, a common choice of these sets are norm balls, say of the form

$$C = \{g \in \mathbb{R}^d : \|g\| \leq L\}, \quad \text{and} \quad D = \{g \in \mathbb{R}^d : \|g\| \leq M\},$$

where $\|\cdot\|$ is a given norm on $\mathbb{R}^d$, and the radius choice $M > L$ ensures that $C \subset \operatorname{int} D$. This choice covers a variety of online optimization and stochastic approximation algorithms [Zinkevich 2003; Beck and Teboulle 2003; Nemirovski et al. 2009; Agarwal et al. 2012], for which it is assumed that for any $x \in \mathcal{X}$ and $\theta \in \Theta$, if $g \in \partial\ell(x, \theta)$ then $\|g\| \leq L$ for some norm $\|\cdot\|$. We may obtain privacy by allowing perturbation of the subgradient g, which is then required to live in a (larger) norm ball of radius $M > L$.

2.2. Optimal Local Privacy

Suppose that X has distribution P, and for each $x \in \mathcal{X}$, let $Q(\cdot \mid x)$ denote the regular conditional probability measure of Z given that $X = x$. This pair defines the marginal distribution $Q(\cdot)$ via $Q(A) = \mathbb{E}[Q(A \mid X)]$, where the expectation taken with respect to $X \sim P$. The mutual information between X and Z is the expected Kullback-Leibler (KL) divergence between $Q(\cdot \mid X)$ and $Q(\cdot)$:

$$I(P, Q) = I(X; Z) := \mathbb{E}_P[D_{\text{kl}}(Q(\cdot \mid X) \| Q(\cdot))]. \tag{8}$$

We view the problem of privacy as a game between the adversary controlling P and the data owners, who use Q to obscure the samples X. In particular, we say a distribution Q guarantees a level of privacy I^* if and only if $\sup_P I(P, Q) \leq I^*$. Note that this guarantee is worst-case, ensuring that for any choice of distribution P, the publicly available random variable Z provides at most mutual information I^* about the sample X.

Our goal is to find a saddle point P^*, Q^* such that

$$\sup_P I(P, Q^*) \leq I(P^*, Q^*) \leq \inf_Q I(P^*, Q), \tag{9}$$

where the first supremum is taken over all distributions P on X such that $\nabla\ell(X, \theta) \in C$ with P-probability 1, and the infimum is taken over all regular conditional distributions Q such that if $Z \sim Q(\cdot \mid X)$ (meaning that Z is drawn from Q conditional on X), then $Z \in D$ and $\mathbb{E}_Q[Z \mid X, \theta] = \nabla\ell(X, \theta)$. Indeed, if we can find P^* and Q^* satisfying the saddle point (9), then combination with the trivial direction of the max-min inequality yields

$$\sup_P \inf_Q I(P, Q) = I(P^*, Q^*) = \inf_Q \sup_P I(P, Q).$$

To fully formalize this idea and our notions of privacy, we define two collections of probability measures and associated losses. For sets $C \subset D \subset \mathbb{R}^d$, we define the source set

$$\mathcal{P}(C) := \{\text{Distributions } P \text{ such that } \operatorname{supp} P \subset C\} \tag{10a}$$

and the set of communicating distributions as the following regular conditional distributions (r.c.d.'s):

$$\mathcal{Q}(C, D) := \left\{ \text{r.c.d.'s } Q \text{ s.t. } \operatorname{supp} Q(\cdot \mid c) \subset D \text{ and } \int_D z dQ(z \mid c) = c \text{ for } c \in C \right\}. \tag{10b}$$

The definitions (10a) and (10b) formally define the sets over which we may take infima and suprema in the saddle point calculations, and they capture what may be communicated. The conditional distributions $Q \in \mathcal{Q}(C, D)$ are defined so that for any loss ℓ with $\nabla \ell(x, \theta) \in C$, we have

$$\mathbb{E}_Q[Z \mid X = x, \theta] := \int_D z dQ\big(z \mid \nabla \ell(x, \theta)\big) = \nabla \ell(x, \theta).$$

We now make the following key definition:

Definition 2.1. The conditional distribution Q^* satisfies *optimal local privacy* for the sets $C \subset D \subset \mathbb{R}^d$ if

$$\sup_P I(P, Q^*) = \inf_Q \sup_P I(P, Q),$$

where the supremum is taken over distributions $P \in \mathcal{P}(C)$ and the infimum is taken over regular conditional distributions $Q \in \mathcal{Q}(C, D)$. We say Q^* satisfies *optimal local privacy at level* I^* if in addition $\sup_P I(P, Q^*) = I^*$.

We also formulate a corresponding notion of local optimality in the differentially private setting. For given sets $C \subset D$, define the differential privacy measure

$$\alpha^\star(C, D) := \inf_Q \log \left[\sup_{S \in \sigma(D)} \sup_{x, x' \in C} \frac{Q(S \mid X = x)}{Q(S \mid X = x')} \right], \tag{11}$$

where the infimum is taken over all regular conditional distributions $Q \in \mathcal{Q}(C, D)$ such that $\mathbb{E}_Q[Z \mid X = x] = x$. We define optimal local differential privacy as follows:

Definition 2.2. The conditional distribution Q^* satisfies *optimal local differential privacy* for the sets $C \subset D \subset \mathbb{R}^d$ if $Q^* \in \mathcal{Q}(C, D)$,

(1) the distribution Q^* is $\alpha^\star(C, D)$-differentially private; and
(2) we have $\sup_P I(P, Q^*) \le \sup_P I(P, Q)$, for all $\alpha^\star(C, D)$-differentially private $Q \in \mathcal{Q}(C, D)$, where the supremum is taken over all distributions $P \in \mathcal{P}(C)$.

If a distribution Q^* satisfies optimal local privacy or optimal local differential privacy, then it guarantees that even for the worst possible distribution on X, the information communicated about X is limited. (Part of our results consist in showing that for suitable sets $C \subset D$, it is possible to attain $\alpha^\star(C, D)$, so it is sensible to, in addition, choose the distribution that minimizes mutual information.)

In a sense, Definitions 2.1 and 2.2 capture the natural competition between privacy and learnability. The method $\mathcal{M}$ specifies the set D to which the data Z it receives must belong; the "teachers," or owners of the data X, choose the distribution Q to guarantee as much privacy as possible subject to this constraint. Using these mechanisms, if we can characterize a unique distribution Q^* attaining the infimum (9) for P^* (and by extension, for any P), then we may study the effects of requiring a bounded amount of information to be communicated to the method $\mathcal{M}$ about X, which we do in Section 3.

2.3. Minimax Error

Given an estimate $\widehat{\theta}$ based on n samples X from a distribution P, we assess its quality in terms of the risk function (1), that is, $R(\theta) = \mathbb{E}[\ell(X, \theta)]$. In this section, we describe the minimax framework for obtaining bounds uniformly over all possible estimators. Let $\mathcal{M}$ denote any statistical procedure or method that operates on stochastic gradient samples, and let $\widehat{\theta}_n$ denote the output of $\mathcal{M}$ after receiving n such samples. The excess risk of the method $\mathcal{M}$ on the risk $R(\theta)$ after receiving n sample gradients is

$$\epsilon_n(\mathcal{M}, \ell, \Theta, P) := R(\widehat{\theta}_n) - \inf_{\theta \in \Theta} R(\theta) = \mathbb{E}_P[\ell(X, \widehat{\theta}_n)] - \inf_{\theta \in \Theta} \mathbb{E}_P[\ell(X, \theta)]. \tag{12}$$

The excess risk is a random variable, since the output $\widehat{\theta}_n$ of the method is random.

In our settings, in addition to the randomness in the sampling distribution P, there is additional randomness from the perturbation applied to stochastic gradients of the objective $\ell(X, \cdot)$ to mask X from the statistitician or method $\mathcal{M}$. Let Q denote the regular conditional probability—the channel distribution—whose conditional part is defined on the range of the (set-valued) subgradient mapping $\partial \ell(X, \cdot) : \Theta \rightrightarrows \mathbb{R}^d$. Since the output $\widehat{\theta}_n$ of the statistical procedure $\mathcal{M}$ is a random function of both P and Q, we take the expectation and measure the expected suboptimality of the risk according to P and Q. We let $\mathfrak{L}$ denote a collection of loss functions, where for a distribution P on $\mathcal{X}$, the set $\mathfrak{L}(P)$ denotes the losses $\ell : \operatorname{supp} P \times \Theta \to \mathbb{R}_+$ belonging to $\mathfrak{L}$. The *minimax error* is then given by

$$\epsilon_n^*(\mathfrak{L}, \Theta) := \inf_{\mathcal{M}} \sup_{P} \sup_{\ell \in \mathfrak{L}(P)} \mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)], \tag{13}$$

where the expectation is taken over the random samples $X \sim P$ and $Z \sim Q(\cdot \mid X, \theta)$. In this article, we provide characterizations of the minimax error (13) for several classes of loss functions $\mathfrak{L}(P)$, giving sharp results when the privacy distribution Q satisfies optimal local privacy for any loss function $\ell \in \mathfrak{L}(P)$ and distribution P.

3. OPTIMAL LEARNING RATES AND TRADEOFFS

With the basic framework in place, we now turn to statements of our main results. We begin by imposing certain (weak) conditions on the families of loss functions that we consider, and subsequently turn to the main results of this section (Theorems 3.4 and 3.5, which apply to information-based privacy, and Theorems 3.8–3.11, which apply to α-differential privacy) as well as some of their consequences (Corollaries 3.6, 3.7, and 3.9). After describing the optimal privacy-preserving distributions in Section 4, we provide proofs of the results in this section in Section 5.

3.1. Families of Loss Functions and Stochastic Gradient Methods

We assume that our collection of loss functions obey certain natural smoothness conditions. For each $p \in [1, \infty]$, we use $\|\cdot\|_p$ to denote the usual ℓ_p-norm, and we use $q = \frac{p}{p-1}$ to denote the conjugate exponent satisfying the relation $1/p + 1/q = 1$. With this notation, we have the following definition.

Definition 3.1. For parameters $L > 0$ and $p \geq 1$, an (L, p)*-loss function* is a measurable function $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ such that for $x \in \mathcal{X}$, the function $\theta \mapsto \ell(x, \theta)$ is convex and L-Lipschitz continuous with respect to the norm $\|\cdot\|_q$.

A convex loss ℓ satisfies Definition 3.1 if and only if for all $\theta \in \Theta$, we have the inequality $\|g\|_p \le L$ for any subgradient $g \in \partial\ell(x, \theta)$ (e.g., Hiriart-Urruty and Lemaréchal [1996]).

To illustrate this definition, let us consider a few examples.

Example 3.2. Consider the problem of finding a multidimensional median, in which case each sample $x \in \mathbb{R}^d$, and the loss function takes the form

$$\ell(x, \theta) = L \|\theta - x\|_1 .$$

This loss is L-Lipschitz with respect to the ℓ_1-norm, subgradients belonging to $[-L, L]^d$, and hence it belongs to the class of (L, ∞)-loss functions.

Example 3.3 (*Classification*). We may also consider classification based on either the hinge loss or logistic regression loss. In this setting, the data comes in pairs $x = (a, b)$, where $a \in \mathbb{R}^d$ is the set of regressors or predictors and $b \in \{-1, 1\}$ is the label; the losses are

$$\ell(x, \theta) = \left[1 - b \langle a, \theta\rangle\right]_+ \quad \text{and} \quad \ell(x, \theta) = \log(1 + \exp(-b \langle a, \theta\rangle)).$$

By computing (sub)gradients, we may verify that each of these is an (L, p)-loss if and only if the covariate vector $a \in \mathbb{R}^d$ satisfies $\|a\|_p \le L$, which is a common assumption [Chaudhuri et al. 2011; Rubinstein et al. 2012].

Definition 3.1 is natural, given the communication strategy we outline in Section 2.1. Since our loss functions satisfy $\|\partial\ell(X, \theta)\| \le L$, the channel distribution Q amounts to perturbing subgradients to larger norm balls while maintaining the appropriate expectations.

Before proceeding, we briefly review standard algorithms for solving problems of the forms outlined previously, since they are essential to our results: for each of our main results, the optimal convergence rate is attained by (a variant of) mirror descent [Nemirovski and Yudin 1983; Beck and Teboulle 2003; Nemirovski et al. 2009], which is a non-Euclidean generalization of the stochastic gradient method [Nemirovski and Yudin 1983; Polyak and Juditsky 1992; Zinkevich 2003]. Stochastic gradient methods are iterative methods that update a parameter θ^t over iterations t of an algorithm using stochastic gradient information. At iteration t, the algorithm receives a vector $g_t \in \mathbb{R}^d$ with conditional expectation $\mathbb{E}[g_t \mid \theta^t] \in \partial R(\theta^t)$, then performs the update

$$\theta^{t+1} = \operatorname*{argmin}_{\theta\in\Theta}\{\eta \langle g_t, \theta\rangle + \Psi(\theta, \theta^t)\}.$$

Here η is a step-size and Ψ is a Bregman divergence, which keeps θ^{t+1} relatively close to θ^t. (See Beck and Teboulle [2003] and Nemirovski et al. [2009] for further details.) With appropriate choice of Ψ, the mirror descent algorithm enjoys the following convergence guarantees. Define $\widehat{\theta}_n = \frac{1}{n}\sum_{t=1}^n \theta^t$. If $\mathbb{E}[\|g_t\|_\infty^2 \mid \theta^t] \le M_\infty^2$ for all t and Θ is contained in the ℓ_1-ball of radius r_1, then with appropriate choice of Ψ and η

$$\mathbb{E}[R(\widehat{\theta}_n)] - R(\theta^*) = \mathcal{O}\left(\frac{M_\infty r_1\sqrt{\log d}}{\sqrt{n}}\right). \tag{14a}$$

See, for example, Beck and Teboulle [2003, Section 5] or Nemirovski et al. [2009, Section 2.3]. Similarly, with the choice $\Psi(\theta, \theta') = \|\theta - \theta'\|_2^2$, if $\mathbb{E}[\|g_t\|_2^2 \mid \theta^t] \le M_2^2$ and Θ is contained in the ℓ_2-ball of radius r_2, then

$$\mathbb{E}[R(\widehat{\theta}_n)] - R(\theta^*) = \mathcal{O}\left(\frac{M_2 r_2}{\sqrt{n}}\right). \tag{14b}$$

For instance, see Zinkevich [2003] and Nemirovski et al. [2009] for results of this type.

3.2. Minimax Error Bounds under Privacy

We now state our main theorems, and discuss some of their consequences. All proofs are deferred to Section 5.

3.2.1. Minimax Errors with Mutual Information-Based Privacy. Our first two main results consider privacy mechanisms Q satisfying optimal local privacy, Definition 2.1. We state the theorems first focusing on their dependence on the geometry of the subdifferential sets (in which the subgradients live); in the corollaries to follow we show how these choices correspond to particular mutual information guarantees on privacy.

Our first theorem applies to the class of (L, ∞) loss functions as given in Definition 3.1. For this theorem, we assume that the set to which the perturbed data Z must belong is $[-M_\infty, M_\infty]^d$, where $M_\infty \geq L$. In the notation of Definition 2.1, this corresponds to taking $C = [-L, L]^d$ and $D = [-M_\infty, M_\infty]^d$. We state two variants of the first theorem, as one version gives slightly sharper results for an important special case.

THEOREM 3.4. *Let $\mathfrak{L}$ be the collection of (L, ∞) loss functions, assume the conditions of the preceding paragraph, and let Q be optimally locally private (Definition 2.1) for $\mathfrak{L}$. Then*

(*a*) *if Θ contains the ℓ_∞ ball of radius r, then*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{20} \min\left\{rLd, \frac{M_\infty rd}{9\sqrt{n}}\right\}.$$

(*b*) *if $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$, then*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{8} \min\left\{rL, \frac{M_\infty r\sqrt{\log(2d)}}{2\sqrt{n}}\right\}.$$

Our second main theorem applies to loss functions and objectives with a different geometry. Now we assume that the loss functions $\mathfrak{L}$ consist of $(L, 1)$ losses, and that the perturbed data must belong to the ℓ_1 ball of radius M_1, that is, $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \leq M_1\}$. Thus, in the notation of Definition 2.1, we have $D = (M_1/L)C$, where $C = \{g \in \mathbb{R}^d : \|g\|_1 \leq L\}$. If we define $M = M_1/L$, we may define the constants

$$\gamma := \log\left(\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2-1)}}{2(M-1)}\right) \quad \text{and} \quad \Delta(\gamma) := \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d-1)}, \tag{15}$$

which are related to the unique distribution achieving optimal local privacy for the $(L, 1)$ losses and the larger ℓ_1-ball (see Eq. (20) and Proposition 4.4). We have the following theorem.

THEOREM 3.5. *Let $\mathfrak{L}$ be the collection of $(L, 1)$ loss functions, assume the conditions of the preceding paragraph, and let Q be optimally private for the collection $\mathfrak{L}$. If Θ contains the ℓ_∞-ball of radius r,*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{20} \min\left\{rL, \frac{rL\sqrt{d}}{9\sqrt{n}\Delta(\gamma)}\right\}.$$

Remarks. We make a few remarks on Theorems 3.4 and 3.5. First, we note that, when reduced to the special case of having no random distribution Q, Theorems 3.4 and 3.5 each yield a minimax rate for stochastic optimization problems. Indeed, in Theorem 3.4, we may take $M_\infty = L$, in which case (focusing on the second statement of the theorem) we obtain that, for $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{rL}{16}\sqrt{\frac{\log(2d)}{n}}.$$

Mirror descent algorithms [Nemirovski and Yudin 1983; Nemirovski et al. 2009] can be used to minimize this class of loss functions, and their convergence rate matches this lower bound up to constant factors (also see our results in the sequel, as well as the explanation of Agarwal et al. [2012]). When specialized to this setting, our result is thus unimprovable. Moreover, our analysis is sharper than previous analyses: none of the existing lower bounds recover the logarithmic dependence on the dimension d, which is evidently necessary.

Our second remark is that while our results appear to require disguising only gradient information, based on our communication formulation in Section 2.1, this restriction is not actually substantial. Indeed, when the domain Θ is a norm ball, we can establish each of our lower bounds using the loss function $\ell(x, \theta) = \langle x, \theta \rangle$. In this case, $\nabla \ell(x, \theta) = x$, so that the communication scheme explicitly disguises *exactly* the individual data X_i.

We now turn to some consequences of Theorems 3.4 and 3.5, where we exhibit the tradeoffs between rates of convergence for any statistical procedure and the desired privacy of a user. We present two corollaries that characterize this tradeoff. Looking ahead to Section 4, we may use Propositions 4.3 and 4.4 in that section to derive a bijection between the sizes M_∞ and M_1 of the perturbation sets and the amount of privacy as measured by the worst case mutual information I^*. We can then combine the lower bounds of Theorems 3.4 and 3.5 with results on stochastic approximation (the mirror descent convergence rates (14)) to obtain the following tradeoffs. We provide the full proofs in Sections 5.7 and 5.8, respectively.

COROLLARY 3.6. *Under the conditions of Theorem 3.4(b), assume moreover that $M_\infty \geq 2L$, and that Q^* satisfies optimal local privacy at information level I^* in the sense of Definition 2.1. Then, for universal constants $0 < c_\ell \leq c_u < \infty$, the minimax error is sandwiched as*

$$c_\ell \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{\log(2d)}}{\sqrt{n}} \leq \epsilon_n^*(\mathfrak{L}, \Theta) \leq c_u \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{\log(2d)}}{\sqrt{n}}.$$

Similar upper and lower bounds can be obtained under the conditions of part (a) of Theorem 3.4, again by using mirror descent, but we lose a factor of $\sqrt{\log d}$ in the lower bound. (There is an additional factor of d in the statement (a), and $\Theta \supseteq \{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq r/d\}$.) In this case, we would not need to assume that Θ is an ℓ_1-ball for the lower bound.

We now turn to an analogous result based on an application of Theorem 3.5 and Proposition 4.4.

COROLLARY 3.7. *Under the conditions of Theorem 3.5, assume that $M_1 \geq 2L$ and Q^* satisfies optimal local privacy at information level I^*. Moreover, suppose that Θ contains an ℓ_∞-ball of radius $c_1 r$ and is contained in an ℓ_∞-ball of radius $c_2 r$, where $0 < c_1 \leq c_2$ are constants. Then, for universal constants $0 < c_\ell \leq c_u < \infty$, the minimax error is sandwiched as*

$$c_\ell \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{d}}{\sqrt{n}} \leq \epsilon_n^*(\mathfrak{L}, \Theta) \leq c_u \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{d}}{\sqrt{n}}.$$

As a final remark, we have stated results that depend on specific geometric properties of the loss functions $\mathfrak{L}$. While these geometric properties are natural, as illustrated by the example Section 3.1, it is also possible to use our techniques to derive alternative results. Such extensions require computing the optimal distribution attaining local privacy in accordance with Definitions 2.1 or 2.2, then applying the lower-bounding techniques to developed in Section 5.

3.2.2. Minimax Errors under Differential Privacy. We now turn to the setting of differentially private algorithms. We focus on two settings for differential privacy: in the first (Theorem 3.8), we assume that communication respects optimal local differential privacy, as given by Definition 2.2. For the second two results, Theorems 3.10 and 3.11, we change the setting slightly, assuming only that the mechanism by which the private quantity Z_i is communicated to the method $\mathcal{M}$ is α-differentially private and noninteractive (recall Eq. (3)).

Optimal Local Differential Privacy. We begin with the result assuming optimal local differential privacy. We use the same collection of loss functions $\mathfrak{L}$ as in Theorem 3.4, that is, (L, ∞)-loss functions. We also assume that the set to which the perturbed data Z belong is $[-M_\infty, M_\infty]^d$, though the specific value of M_∞ is not important for the statement of the theorem.

THEOREM 3.8. *Let $\mathfrak{L}$ be the collection of (L, ∞) loss functions, and assume that Z is optimally locally differentially private (Definition 2.2), attaining α-differential privacy for the set $\mathfrak{L}$. Let $d \geq 2$ and assume $\alpha \leq 5/4$. Then*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{8} \min\left\{rL, \frac{\sqrt{d}}{\alpha} \frac{rL\sqrt{\log(2d)}}{4\sqrt{n}}\right\}.$$

As a corollary to this result, we can show an upper bound on the necessary magnitude of the gradient bound M_∞ to allow α-differential privacy, again applying the mirror descent result (14a). See Section 5.9 for a proof.

COROLLARY 3.9. *Under the conditions of Theorem 3.8, assume that Q^* satisfies Definition 2.2, attaining α-differential privacy. Then, for universal constants $0 < c_\ell \leq c_u$, the minimax error is sandwiched as*

$$c_\ell \frac{\sqrt{d}}{\alpha} \cdot \frac{rL\sqrt{\log(2d)}}{\sqrt{n}} \leq \epsilon_n^*(\mathfrak{L}, \Theta) \leq c_u \frac{\sqrt{d}}{\alpha} \cdot \frac{rL\sqrt{\log(2d)}}{\sqrt{n}}.$$

Noninteractive Local Differential Privacy. We turn to our two results under noninteractive differential privacy, where we no longer assume that the channel is optimally private (the data provider simply guarantees α-differential privacy). In this setting, we give a minor refinement of the definition of minimax error (13). We let $\mathcal{Q}_\alpha$ denote the family of α-differentially private distributions where the channel Q is α-differentially private and non-interactive, meaning that the private variable Z_i is conditionally independent of Z_j for $j \neq i$ given X_i; recall the definition (3). With this, the minimax error is defined as

$$\epsilon_n^*(\mathfrak{L}, \Theta, \alpha) := \inf_{\mathcal{M}, Q \in \mathcal{Q}_\alpha} \sup_P \sup_{\ell \in \mathfrak{L}(P)} \mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)],$$

where now the infimum is taken over all α-private, noninteractive local mechanisms Q, as well as all methods $\mathcal{M}$. Thus, the channel Q and $\mathcal{M}$ work together to find the best possible estimator, subject to the differential privacy constraint.

Our first lower bound applies to a class of functions that are Lipschitz with respect to the ℓ_1-norm, where the optimization takes place over the ball $\mathbb{B}_1(r) := \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq r\}$.We define the set $\mathfrak{L}(\mathbb{B}_1(r); L)$ to be the collection of convex (L, ∞)-loss functions defined on $\mathbb{B}_1(r)$. By Example 3.2, this loss class covers the problem of the multidimensional median. In stating our minimax bounds, we use a more restrictive (i.e., simpler to optimize) class, the collection of (L, ∞)-*linear* losses:

$$\mathfrak{L}_{\text{lin}}(L, \infty) := \left\{ \ell : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R} \mid \exists\, \phi : \mathcal{X} \to \mathbb{R}^d \text{ s.t. } \ell(x, \theta) = \langle \phi(x), \theta \rangle, \sup_x \|\phi(x)\|_\infty \le L \right\}.$$

For this class, we have the following minimax rate (see Section 5.5 for a proof).

THEOREM 3.10. *For the loss class $\mathfrak{L} = \mathfrak{L}_{\text{lin}}(L, \infty)$ and privacy parameter $\alpha = \mathcal{O}(1)$, assuming that the channel Q is noninteractive and α-differentially private, there are universal constants $0 < c_\ell \le c_u < \infty$ such that*

$$c_\ell \min\left\{ \frac{\sqrt{d}}{\alpha} \frac{rL\sqrt{\log(2d)}}{\sqrt{n}}, rL \right\} \;\le\; \epsilon_n^*(\mathfrak{L}, \mathbb{B}_1(r), \alpha) \;\le\; c_u \min\left\{ \frac{\sqrt{d}}{\alpha} \frac{rL\sqrt{\log(2d)}}{\sqrt{n}}, rL \right\}. \quad (16)$$

We can also give a result for a larger class of domains and related optimization functions. In particular, consider the loss class

$$\mathfrak{L}(\Theta; L, p) := \{\ell : \mathcal{X} \times \Theta \to \mathbb{R} \mid \ell \text{ is a convex } (L, p)\text{-loss function}\} \quad (17)$$

for some $p \in [1, 2]$. Restricting the set (17) to the smaller collection of linear functionals, we define

$$\mathfrak{L}_{\text{lin}}(\Theta; L, p) := \left\{ \ell : \mathcal{X} \times \Theta \to \mathbb{R} \mid \exists\, \phi : \mathcal{X} \to \mathbb{R}^d \text{ s.t. } \ell(x, \theta) = \langle \phi(x), \theta \rangle, \sup_x \left\| \phi(x) \right\|_p \le L \right\}.$$

We then have the following result, which captures rates of convergence for optimization of linear functionals over ℓ_q-norm balls of the form

$$\mathbb{B}_q(r_q) := \{\theta \in \mathbb{R}^d : \|\theta\|_q \le r_q\}, \quad \text{where } q \in [2, \infty].$$

THEOREM 3.11. *For the loss class $\mathfrak{L} = \mathfrak{L}_{\text{lin}}(\mathbb{B}_q(r_q); L, p)$ with $q \in [2, \infty]$ and noninteractive α-differentially private channel Q with $\alpha = \mathcal{O}(1)$, there exist universal constants $0 < c_\ell \le c_u < \infty$ such that*

$$c_\ell r_q L \min\left\{ \frac{\sqrt{d}}{\alpha} \frac{d^{\frac{1}{2} - \frac{1}{q}}}{\sqrt{n}}, (\sqrt{n\alpha^2})^{-\frac{1}{q}}, 1 \right\} \;\le\; \epsilon_n^*(\mathfrak{L}, \mathbb{B}_q(r_q), \alpha) \;\le\; c_u r_q L \min\left\{ \frac{\sqrt{d}}{\alpha} \frac{d^{\frac{1}{2} - \frac{1}{q}}}{\sqrt{n}}, 1 \right\}. \quad (18)$$

For the loss class $\mathfrak{L} = \mathfrak{L}(\Theta; L, p)$ from Eq. (17)*, if $\Theta \supset \mathbb{B}_q(r_q)$, there exists a universal (numerical) constant $0 < c_\ell$ such that*

$$c_\ell \min\left\{ \frac{\sqrt{d}}{\alpha} \frac{r_q L d^{\frac{1}{2} - \frac{1}{q}}}{\sqrt{n}}, r_q L \right\} \;\le\; \epsilon_n^*(\mathfrak{L}, \Theta, \alpha). \quad (19)$$

See Section 5.6 for a proof.

Remarks. Each of our theorems and corollaries provide sharp characterizations of the minimax rate of estimation up to the constant factors (c_ℓ, c_u). As noted in the previous section, the nonprivate minimax rate for the class $\mathfrak{L}(\mathbb{B}_1(r); L)$ is $rL\sqrt{\log(2d)}/\sqrt{n}$. We may compare this with the rate in Theorems 3.8 and 3.10, as well as Corollary 3.9. We see that α-local differential privacy has a dimension-dependent effect on the minimax rate: the effective sample size is reduced from n to $\alpha^2 n/d$. This is a substantial reduction,as instead of a logarithmic dependence on the dimension d—which one hopes for in high-dimensional settings such as those specified by the theorem—we have a linear dependence, which is unavoidable under the conditions of the theorems.

In Theorem 3.11 as well, the inequalities (18) provide a characterization of the α-private minimax rate that is tight up to constant factors. Again, it is worthwhile to relate this minimax rate to the non-private setting: from Theorem 1 and Eq. (11) of Agarwal et al. [2012], the nonprivate minimax rate for the function class $\mathfrak{L}_{\text{lin}}(\Theta; L, p)$ is lower bounded by $r_q L d^{\frac{1}{2}-\frac{1}{q}}/\sqrt{n}$. Consequently, the price for α-privacy is again a reduction in effective sample size by the dimension-dependent factor α^2/d.

In general, stochastic gradient descent methods require interactivity—they iteratively process the data and query for gradients at points θ depending on the data observed—except in linear settings. We do, however, obtain matching upper bounds for the general convex case in both Theorems 3.10 and 3.11 using stochastic gradient methods (which is unsurprising, as the linear setting is, in a sense, the hardest [Nemirovski and Yudin 1983; Agarwal et al. 2012]). This leads to the intriguing open question of whether interactivity can sharpen the results of Theorems 3.10 and 3.11. (For Theorems 3.4–3.8, the optimal privacy game played by the data providers allows interactivity, and hence the results cannot be improved.) It is also interesting to note that in Theorems 3.8–3.11, in the α-differentially private setting, adding Laplace noise—the most common mechanism for achieving privacy [Dwork 2008]—appears to be substantially suboptimal: the magnitude of noise necessary to privatize the user's data is $\Omega(d)$ larger than that provided by the optimal sampling mechanisms we develop in the sequel.

Summarizing, each of the preceding results indicates that—no matter the type of privacy—there is a dimension-dependent increase in sample complexity. From Corollaries 3.6 and 3.7, we see that incorporating privacy induces a penalty of roughly $\sqrt{d}/\sqrt{I^*}$ in convergence rate, or an effective sample size reduction from n to nI^*/d; in the differential privacy case we have $n \mapsto n\alpha^2/d$. While we do not know of an explicit comparison between these two bounds, work by Dwork et al. [2010, Lemma 3.2] shows that KL divergence between α-differentially private distributions scales as α^2, which implies roughly that $I^* \approx \alpha^2$ (though this is informal). We see roughly similar results, though there does not appear to be a simple mapping between information-theoretic notions of privacy and differential privacy.

4. OPTIMAL PRIVACY-PRESERVING DISTRIBUTIONS

In this section, we explore conditions for a distribution Q^* to satisfy optimal local privacy as given by Definitions 2.1 and 2.2. We give a few characterizations of necessary (and sometimes sufficient) conditions based on the compact sets $C \subset D$ for distributions P^* and Q^* to achieve the saddle point (9). Our results can be viewed as rate distortion theorems [Gray 1990; Cover and Thomas 2006; Csiszár and Körner 1981] (with source P and channel Q) for certain compact alphabets, though as far as we know, they are all new. Thus, we refer to the conditional distribution Q, which is designed to maintain the privacy of the data X by communication of Z, interchangeably as the privacy-preserving distribution or the channel distribution.

Note that since we wish to bound $I(X; Z)$ for general losses ℓ, as captured in the definitions of the source $\mathcal{P}(C)$ and communication set $\mathcal{Q}(C, D)$ in Eqs. (10a) and (10b), we must address the case when $\ell(X, \theta) = \langle \theta, X \rangle$, in which case $\nabla \ell(X, \theta) = X$; this shows (by the data-processing inequality [Gray 1990, Chapter 5]) that it is no loss of generality to assume that $X \in C$ with probability 1 and that we must have $\mathbb{E}[Z \mid X] = X$. Thus, we present each of our results assuming that $\ell(X, \theta) = \langle \theta, X \rangle$, since a distribution Q^* is optimally locally private or optimally differentially locally private if and only if it attains the saddle point with this choice of loss.

4.1. General Saddle Point Characterizations

We begin with a general characterization, first defining the types of sets C and D that we use in our characterization of privacy. Such sets are reasonable for many applications (recall Section 3.1). We focus on the case when the compact sets C and D are (suitably symmetric) norm balls.

Definition 4.1. Let $C \subset \mathbb{R}^d$ be a compact convex set with extreme points $u_i \in \mathbb{R}^d$, $i \in I$ for some index set I. Then, C is a *rotationally invariant through its extreme points* if $\|u_i\|_2 = \|u_j\|_2$ for each i, j, and for any unitary matrix U such that $Uu_i = u_j$ for some $i \neq j$, then $UC = C$.

Some examples of convex sets rotationally invariant through their extreme points include ℓ_p-norm balls for $p = 1, 2, \infty$, though ℓ_p-balls for $p \notin \{1, 2, \infty\}$ are not.

The following theorem gives a general characterization of the minimax mutual information for such rotationally invariant sets by providing saddle point distributions P^* and Q^*. We provide the proof of Theorem 4.2 in Section E.1 of Appendix E.

THEOREM 4.2. *Let C be a compact convex polytope rotationally invariant through its $m < \infty$ extreme points $\{u_i\}_{i=1}^m$ and $D = (1+\kappa)C$ for some $\kappa > 0$. Let Q^* be the conditional distribution of $Z \mid X$ that maximizes the entropy $H(Z \mid X = x)$ subject to the constraints that*

$$\mathbb{E}_Q[Z \mid X = x] = x$$

for $x \in C$ and that Z is supported on $(1+\kappa)u_i$ for $i = 1, \ldots, m$. Then, Q^* satisfies Definition 2.1, optimal local privacy, and Q^* is (up to measure zero sets) unique. Moreover, the distribution P^* that is uniform on $\{u_i\}_{i=1}^m$ attains the saddle point (9).

Remarks. We make a few brief remarks here, deferring a somewhat deeper discussion of the implications of Theorem 4.2 to Section E.1, of Appendix E as an understanding of the proof helps. The theorem requires that for Q^* to attain the saddle point guaranteeing optimal local privacy, $Q^*(\cdot \mid X = x)$ should maximize the entropy of Z for each $x \in C$, but this is not essential. If $x \notin \{u_i\}_{i=1}^m$, a two-phase approach still obtains optimal local privacy. We construct a Markov chain $X \to X' \to Z$, where X' is supported on the extreme points $\{u_i\}_{i=1}^m$ of C. The distribution $X \to X'$ may then be any distribution satisfying $\mathbb{E}[X' \mid X] = X$; we then take the conditional distribution $Q^*(\cdot \mid u_i)$ defined for $X' \to Z$ to be the maximum entropy distribution $Q^*(\cdot \mid u_i)$ defined in the theorem. By the data processing inequality [Gray 1990, Chapter 5], this Markov chain $X \to X' \to Z$ guarantees the minimax information bound $I(X; Z) \leq \inf_Q \sup_P I(P, Q)$.

4.2. Specific Saddle Point Computations

With Theorem 4.2 in place, we can explicitly characterize the minimax mutual information for ℓ_1 and ℓ_∞ balls by computing maximum entropy distributions. That is, we compute the unique distributions that attain optimal local privacy—the distributions that guarantee as much (of our definition of) privacy as possible subject to certain constraints. We present two propositions in this regard, providing some discussion and giving proofs in Sections E.2 and E.3 of Appendix E.

First, consider the case where $X \in [-L, L]^d$ and $Z \in [-M, M]^d$, where $M \geq L$. For notational convenience, we define the binary entropy $h(p) = -p \log p - (1-p) \log(1-p)$. We have the following proposition.

PROPOSITION 4.3. *For constants $M \geq L > 0$, let $X \in [-L, L]^d$ and $Z \in [-M, M]^d$ be random variables such that $\mathbb{E}[Z \mid X] = X$ almost surely. Define Q^* to be the conditional distribution on $Z \mid X$ such that the coordinates of Z are independent, have range*

$\{-M, M\}$, *and satisfy*

$$Q^*(Z_i = M \mid X) = \frac{1}{2} + \frac{X_i}{2M} \quad \textit{and} \quad Q^*(Z_i = -M \mid X) = \frac{1}{2} - \frac{X_i}{2M}.$$

Then, Q^* satisfies Definition 2.1, optimal local privacy, and moreover,

$$\sup_P I(P, Q^*) = d - d \cdot h\left(\frac{1}{2} + \frac{L}{2M}\right).$$

Before continuing, we give a slightly more intuitive understanding of Proposition 4.3. Let $L = 1$ for simplicity (this is no loss of generality by scaling). Concavity implies that for $a, b > 0$, $\log(a) \le \log b + b^{-1}(a - b)$, or $-\log(a) \ge -\log(b) + b^{-1}(b - a)$, so

$$-\log\left(\frac{1}{2} - \frac{1}{2M}\right) \ge -\log\frac{1}{2} + 2 \cdot \frac{1}{2M} \quad \text{and} \quad -\log\left(\frac{1}{2} + \frac{1}{2M}\right) \ge -\log\frac{1}{2} - 2 \cdot \frac{1}{2M}.$$

In particular, we see that

$$h\left(\frac{1}{2} + \frac{1}{2M}\right) \ge -\left(\frac{1}{2} + \frac{1}{2M}\right)\left(-\log 2 - \frac{1}{M}\right) - \left(\frac{1}{2} - \frac{1}{2M}\right)\left(-\log 2 + \frac{1}{M}\right) = \log 2 - \frac{1}{M^2}.$$

That is, we have for any distribution P on X, where $X \in [-L, L]^d$, that (in natural logarithms)

$$I(P, Q^*) \le \frac{dL^2}{M^2},$$

and this bound is tight to $\mathcal{O}((M/L)^{-3})$.

We now consider the case when $X \in \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$ and $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \le M\}$. Here the arguments are slightly more complicated, as the coordinates of the random variables are no longer independent, but Theorem 4.2 still allows us to explicitly characterize the saddle point of the mutual information. Before stating the proposition, we recall that if $e_i \in \mathbb{R}^d$ are the standard basis vectors, then the extreme points of the ℓ_1-ball of radius 1 are the $2d$ vectors $\{\pm e_i\}_{i=1}^d$.

PROPOSITION 4.4. *For a constant $M > 1$, let $X \in \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$ and $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \le M\}$ be random variables. Define the parameter*

$$\gamma := \log\left(\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2 - 1)}}{2(M-1)}\right), \tag{20}$$

and let Q^* be the conditional distribution on $Z \mid X$ such that Z is supported on $\{\pm Me_i\}_{i=1}^d$, and

$$Q^*(Z = Me_i \mid X = e_i) = \frac{e^\gamma}{e^\gamma + e^{-\gamma} + (2d-2)}, \tag{21a}$$

$$Q^*(Z = -Me_i \mid X = e_i) = \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + (2d-2)}, \tag{21b}$$

$$Q^*(Z = \pm Me_j \mid X = e_i, j \ne i) = \frac{1}{e^\gamma + e^{-\gamma} + (2d-2)}. \tag{21c}$$

(For $X \notin \{\pm e_i\}$, define X' to be randomly selected in any way from among $\{\pm e_i\}$ such that $\mathbb{E}[X' \mid X] = X$, then sample Z from X' in accordance with (21a)–(21c).*) Then, Q^* satisfies*

Definition 2.1, optimal local privacy, and

$$\sup_P I(P, Q^*) = \log(2d) - \log\left(e^\gamma + e^{-\gamma} + 2d - 2\right) + \gamma \frac{e^\gamma}{e^\gamma + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2}.$$

By scaling, if we have $X \in \{x \in \mathbb{R}^d : \|x\|_1 \leq L\}$ and $Z \in \{z \in \mathbb{R}^d : \|Z\|_1 \leq M_1\}$, then the theorem holds with $M = M_1/L$ in Eq. (20) and with $X = e_i$ replaced by $X = Le_i$ in Eq. (21).

Proposition 4.4 is somewhat more complex than the ℓ_∞ case. We remark that the additional sampling to guarantee that $X' \in \{\pm e_i\}$ (where the conditional distribution Q^* is defined) can be accomplished simply: define the random variable X' so that $X' = e_i \operatorname{sign}(x_i)$ with probability $|x_i| / \|x\|_1$. Evidently, $\mathbb{E}[X' \mid X] = x$, and $X \to X' \to Z$ for Z distributed according to Q^* defines a Markov chain as in our remarks following Theorem 4.2. An asymptotic expansion allows us to gain a somewhat clearer picture of the values of the mutual information, though we do not derive upper bounds as we did for Proposition 4.3. We have the following corollary, proved in Section E.5, of Appendix E.

COROLLARY 4.5. *Let Q^* denote the conditional distribution in Proposition 4.4, where X and Z lie in ℓ_1-balls of radii L and M, respectively. Then*

$$\sup_P I(P, Q^*) = \frac{dL^2}{2M^2} + \Theta\left(\min\left\{\frac{d^3L^4}{M^4}, \frac{\log^4(d)}{d}\right\}\right).$$

4.3. Saddle Points for Differentially Private Communication

Our final result in this section characterizes saddle points for distributions satisfying Definition 2.2. Such calculations are, in general, nontrivial, so we restrict our attention to results necessary for the setting of Theorem 3.8. To that end, we focus on the case where C and D are ℓ_∞ balls, which is relevant for high-dimensional statistical and optimization settings. Without loss of generality (by scaling), we may take $C = [-1, 1]^d$ and $D = [-M, M]^d$. We have the following proposition.

PROPOSITION 4.6. *For a constant $M \geq 1$, let $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$ be random variables such that $\mathbb{E}[Z \mid X] = X$ almost surely. Fix any $x \in \{-1, 1\}^d$ and for $k = \{0, 2, 4, \ldots, 2\lceil d/2\rceil - 2\}$ define the constants $q_k^+ \geq 0$ and $q_k^- \geq 0$ to satisfy the linear equations*

$$Mq_k^+ \sum_{z \in \{-1,1\}^d : \langle z, x\rangle > k} z + Mq_k^- \sum_{z \in \{-1,1\}^d : \langle z, x\rangle \leq k} z = x \quad \textit{and} \quad q_k^+ + q_k^- = 1.$$

Set $k^* \in \operatorname{argmin}_k\{q_k^+/q_k^-\}$. The optimal differential privacy (11) *for the sets $C = [-1, 1]^d$ and $D = [-M, M]^d$ is*

$$\alpha^\star(C, D) = \log \frac{q_{k^*}^+}{q_{k^*}^-} = \min_{k \in \{0,2,\ldots,2\lceil d/2\rceil - 2\}} \log \frac{q_k^+}{q_k^-}.$$

The distributions attaining optimal local differential privacy are characterized as follows. Define Q_k^* to be the distribution supported on $\{-M, M\}^d$ with probability mass function defined by

$$Q_k^*(Z = Mz \mid X = x) = \begin{cases} q_k^+ & \textit{if } \langle z, x\rangle > k \\ q_k^- & \textit{if } \langle z, x\rangle \leq k \end{cases} \tag{22}$$

for $z, x \in \{-1, 1\}^d$. (For $X \notin \{-1, 1\}^d$, define X' to be randomly chosen from $\{-1, 1\}^d$ such that $\mathbb{E}[X' \mid X] = X$, then sample Z according to this p.m.f.) A distribution Q^* satisfies Definition 2.2, optimal local differential privacy, if and only if it can be written as a convex combination of those Q^*_k for which $k \in \mathrm{argmin}_k\{q_k^+, q_k^-\}$, that is,

$$Q^* = \sum_{k \in \mathrm{argmin}_k\{q_k^+, q_k^-\}} \beta_k Q^*_k, \quad \textit{where} \;\; \beta_k \geq 0 \;\; \textit{and} \sum_{k \in \mathrm{argmin}_k\{q_k^+, q_k^-\}} \beta_k = 1.$$

The proof of Proposition 4.6 is technical, and we defer it to Section E.4. We make a few remarks, however. First, we provide a simplified explanation of the linear equations in the proposition. By symmetry, no matter the value of $x \in \{-1, 1\}^d$ chosen, the same q_k^+ and q_k^- solve the linear equations. Proposition 4.6 shows the *structure* of the distribution attaining optimal local differential privacy. That is, the proposition shows that the distribution $Q^*(\cdot \mid x)$ assigns mass only on the points $z \in \{-M, M\}^d$, and moreover, it assigns one of two masses: either q^+ or q^-. To sample from this distribution given an initial point x, one simply flips a coin with bias probabilities $\{q_k^+, q_k^-\}$, and depending on the result of the coint flip, samples $Z \in \{-M, M\}^d$ uniformly from one of the sets $\{z : \langle z, x\rangle > k/M\}$ or $\{z : \langle z, x\rangle \leq k/M\}$.

5. PROOFS OF STATISTICAL RATES

In this section, we prove Theorems 3.4–3.11 as well as Corollaries 3.6, 3.7, and 3.9. Our proofs are based on classical information-theoretic techniques from statistical minimax theory [Yang and Barron 1999; Yu 1997], and also exploit some additional results due to Agarwal et al. [2012]. At a high level, our approach consists of the following steps. Beginning with an appropriately chosen finite set $\mathcal{V}$, we assign a risk function R_v to each member $v \in \mathcal{V}$. The resulting collection $\{R_v\}_{v \in \mathcal{V}}$ of risk functions is chosen so that they "separate" points in the set $\mathcal{V}$, meaning that if $\theta \in \Theta$ is a point that approximately minimizes the function R_v, then for any $w \neq v$, the point θ cannot also be an approximate minimizer of R_w. This separation property allows us to deduce that statistical estimation implies the existence of a testing procedure that distinguishes v from w for $w \neq v$. We then use lower bounds on the error probability in tests, such as Fano's and Le Cam's inequalities [Yu 1997], to obtain a lower bound on the testing error. These inequalities depend on the mutual information between the random variable X_i and the vector Z_i communicated, so the final step is to obtain good upper bounds on this mutual information. In the next section, we describe in more detail this reduction, finishing with an outline of the proofs to follow.

5.1. Reduction to Testing

We begin by describing the reduction that lower bounds the minimax error by the error of a testing problem. It assumes a given collection of risk functions $\{R_v\}_{v \in \mathcal{V}}$ indexed by a finite set $\mathcal{V}$; see the individual theorem proofs to follow for constructions of the particular collections used in our analysis. For each $v \in \mathcal{V}$, we choose some representative $\theta_v^* \in \mathrm{argmin}_{\theta \in \Theta} R_v(\theta)$ of the set of all minimizing vectors. Our reduction is based on a discrepancy measure between pairs of risk functions, first introduced by Agarwal et al. [2012], defined as

$$\rho(R_v, R_w) := \inf_{\theta \in \Theta} \left[R_v(\theta) + R_w(\theta) - R_v(\theta_v^*) - R_w(\theta_w^*)\right].$$

The ρ-separation of the set $\mathcal{V}$ is defined as

$$\rho^\star(\mathcal{V}) := \min\{\rho(R_v, R_w) : v, w \in \mathcal{V}, v \neq w\}. \tag{23}$$

When the set $\mathcal{V}$ is clear from context, we use $\rho^\star$ as shorthand for this separation.

The key to the definition (23) is that the separation allows us to lower bound the expected optimality gap of a statistical method $\mathcal{M}$ by the probability of error in a hypothesis test. That is, with the family $\{R_v\}_{v\in\mathcal{V}}$, we can define the *canonical hypothesis testing* problem: nature chooses a uniformly random $V \in \mathcal{V}$, and conditional on $V = v$, the n observations X_i, $i = 1, \ldots, n$, are drawn independently from a distribution P_v satisfying $R_v(\theta) = \mathbb{E}_{P_v}[\ell(X, \theta)]$. In the private setting, instead of observing X_i, the learner observes the privatized vector Z_i, and given the set $\{Z_i\}_{i=1}^n$, the goal is to determine the underlying index v.

Recall the definition (12) of the excess risk ϵ_n. The previously described hypothesis testing problem can be used to establish lower bounds on the estimation error, as demonstrated in the following.

LEMMA 5.1. *Let P be a joint distribution over $X \in \mathbb{R}^d$ and $V \in \mathcal{V}$ such that X are i.i.d. given V and*

$$\mathbb{E}_P[\ell(X, \theta) \mid V = v] = R_v(\theta).$$

Let Q be the conditional distribution of Z given the observations X. Then, for any minimization procedure $\mathcal{M}$, there exists a hypothesis test $\widehat{v}_{\mathcal{M}} : (Z_1, \ldots, Z_n) \to \mathcal{V}$ such that

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{\rho^\star(\mathcal{V})}{2} \mathbb{P}_{P,Q}[\widehat{v}_{\mathcal{M}}(Z_1, \ldots, Z_n) \neq V].$$

This result is a variant of Lemma 2 due to Agarwal et al. [2012]. It shows that if we can bound the probability of error of any hypothesis test for identifying V based on the sample $Z_1, \ldots, Z_n$, we have lower bounded the rate at which it is possible to minimize the risk R.

The remaining challenge is to provide a lower bound on the error of hypothesis testing problems. To do so, we apply one of two well-known inequalities: Fano's inequality [Cover and Thomas 2006], which applies when $|\mathcal{V}| > 2$, or Le Cam's method [Le Cam 1973; Yu 1997], which we apply when $|\mathcal{V}| = 2$. Let $V \in \mathcal{V}$ be chosen uniformly at random from $\mathcal{V}$. If a procedure observes random variables $Z_1, \ldots, Z_n$, Fano's inequality ensures that for any estimate $\widehat{v}$ of V—that is, any measurable function $\widehat{v}$ of $Z_1, \ldots, Z_n$—the test error probability satisfies the lower bound

$$\mathbb{P}(\widehat{v}(Z_1, \ldots, Z_n) \neq V) \geq 1 - \frac{I(Z_1, \ldots, Z_n; V) + \log 2}{\log |\mathcal{V}|}. \tag{24}$$

By contrast, Le Cam's method provides lower bounds on the probability of error in binary hypothesis testing problems. In this setting, assume that $\mathcal{V} = \{-1, 1\}$ has two elements, and let $V \in \mathcal{V}$ be chosen uniformly at random from $\mathcal{V}$. If a procedure observes random variables $Z_1, \ldots, Z_n$ distributed according to Q_1^n if $V = 1$ and Q_{-1}^n if $V = -1$, then any estimate $\widehat{v}$ of V satisfies the lower bound

$$\mathbb{P}(\widehat{v}(Z_1, \ldots, Z_n) \neq V) \geq \frac{1}{2} - \frac{1}{2} \left\| Q_1^n - Q_{-1}^n \right\|_{\mathrm{TV}}. \tag{25}$$

See, for example, Le Cam [1973, Section 2] or Yu [1997, Lemma 1].

Using the lower bound provided by Lemma 5.1 and Fano's inequality (24) or Le Cam's inequality (25), the structure of our remaining proofs becomes more apparent. Each lower bound argument proceeds in three steps.

(1) We construct a collection of loss functions satisfying Definition 3.1, computing the minimal separation (23) so that we may apply Lemma 5.1.
(2) The second step is to provide an upper bound on the appropriate information theoretic quantity in order to apply Fano's inequality (24), in which case we bound

$I(Z_1, \ldots, Z_n; V)$, or Le Cam's inequality (25), where we bound $\|Q_1^n - Q_{-1}^n\|_{\mathrm{TV}}$. This step requires the most work and constitutes the major arguments in this section. We provide these bounds using one of two techniques.

(a) In the proofs of Theorems 3.4, 3.5, and 3.8, we use a distribution Q that satisfies Definition 2.1 of optimal local privacy (Theorems 3.4 and 3.5) or Definition 2.2 of optimal local differential privacy (Theorem 3.8). We can then explicitly upper bound the mutual information $I(Z_{1:n}; V)$ using the definition of Q and the losses ℓ from Step (1). (See Lemmas 5.4, 5.5, 5.7, and 5.8 in the subsequent sections.)

(b) In the proofs of Theorems 3.10 and 3.11, we use a bound on mutual information in noninteractive locally differentially private schemes, which we recently presented [Duchi et al. 2013]. This requires a careful packing construction in conjunction with the loss choice from Step (1).

(3) The final step is to use the results of Steps (1) and (2) in the application of Lemma 5.1 and Fano's inequality (24) (when the dimension d is low, we use Le Cam's inequality (25)). This then yields the theorems.

We now turn to the proofs of the theorems. We provide each proof in turn, following the steps in the preceding outline.

5.2. Proof of Theorem 3.4

We provide the most detail in the proof of this theorem, as it closely exhibits the blueprint by which we prove the other results.

5.2.1. Constructing Well-Separated Losses. The first step in proving our minimax lower bounds is to construct a family of well-separated risks. For Theorem 3.4, we use one of two families of loss functions: linear losses and median-based losses. Each of these gives a well-separated family with subgradients bounded in ℓ_∞-norm.

Linear Losses. Our first collection of risk functionals is relatively simple, based on families of linear loss functions; we describe the sampling scheme for X to generate them. Let $\mathcal{V} = \{\pm e_i\}_{i=1}^d$, where the vectors e_i are the standard basis vectors in $\mathbb{R}^d$, whence $|\mathcal{V}| = 2d$. Fix $\delta \in (0, 1]$, which we specify later. We choose the distribution P on X to be nearly uniform on $X \in \{-1, 1\}^d$. Conditional on the parameter $v \in \mathcal{V}$, we use the following sampling distribution for X:

$$\text{Choose } X \in \{-1, 1\}^d \quad \text{with independent coordinates,} \\ \text{where } X_j = \begin{cases} 1 & \text{w.p. } \frac{1+\delta v_j}{2} \\ -1 & \text{w.p. } \frac{1-\delta v_j}{2}. \end{cases} \tag{26}$$

Now, for $L \in \mathbb{R}_+$, we may define the linear loss functions

$$\ell(X, \theta) := L \langle X, \theta \rangle = L \sum_{j=1}^d X_j \theta_j. \tag{27}$$

By inspection, the final risk is $R_v(\theta) = \mathbb{E}_P[\langle \theta, X \rangle] = L\delta \langle v, \theta \rangle$. We obtain the following result on the separation of the risks.

LEMMA 5.2. *Given the sampling scheme* (26),

(*a*) *the loss* (27) *is L-Lipschitz with respect to the ℓ_1-norm.*

(*b*) *for the optimization domain $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$, the ρ-separation of the set $\mathcal{V} = \{\pm e_i\}_{i=1}^d$ is $\rho^\star(\mathcal{V}) = Lr\delta$.*

PROOF. The first statement of the lemma is immediate, since $\nabla \ell(X, \theta) = LX$ and $\|X\|_\infty \le 1$ (cf. Hiriart-Urruty and Lemaréchal [1996]). For the second, we verify that

$\inf_{\theta\in\Theta}[R_v(\theta) + R_w(\theta)] - R_v(\theta_v^*) - R_w(\theta_w^*) \geq Lr\delta$. To do so, we compute the minimizers of R_v: since ℓ_1 and ℓ_∞ are dual norms, we see that for $v \in \mathcal{V}$,

$$\inf_{\|\theta\|_1\leq r} R_v(\theta) = \inf_{\|\theta\|_1\leq r} L\delta \langle v, \theta\rangle = -L\delta r \|v\|_\infty = -L\delta r,$$

and the minimizer is uniquely attained at $\theta_v^* = -rv$. Then, we have for any $w \neq v$ that

$$\inf_{\|\theta\|_1\leq 1} [\langle v + w, \theta\rangle] + \|v\|_\infty + \|w\|_\infty = -\|v + w\|_\infty + \|v\|_\infty + \|w\|_\infty \geq -1 + 1 + 1 = 1,$$

since any nonzero coefficients of v and w have differing signs. Multiplying the result by $Lr\delta$ completes the proof. □

Median-Type Losses. We now describe a class of median-type losses, one with more general applicability than the linear losses of Section 5.2.1. Let $\mathcal{V} \subset \{-1, 1\}^d$ be a subset of the binary hypercube such that for all distinct pairs $v \neq v'$, we have $\|v - v'\|_1 \geq d/2$, or equivalently $\|v - v'\|_0 \geq d/4$. From the Gilbert-Varshamov bound [Yu 1997, Lemma 4] there are sets of this form with cardinality at least $\text{card}(\mathcal{V}) \geq \exp(d/8)$. We define the distribution on X, conditional on $v \in \mathcal{V}$, as follows:

$$\text{Choose } X \in \{-1, 1\}^d \quad \text{with independent coordinates,} \\ \text{where } X_j = \begin{cases} 1 & \text{w.p. } \frac{1+\delta v_j}{2} \\ -1 & \text{w.p. } \frac{1-\delta v_j}{2} \end{cases}. \tag{28}$$

For $L > 0$, we then define the median-type loss function

$$\ell(X, \theta) = L \|rX - \theta\|_1, \tag{29}$$

which under the sampling scheme (28) gives rise to the risk functional

$$R_v(\theta) = L \sum_{j=1}^{d} \frac{1 + \delta v_j}{2} |\theta_j - r| + \frac{1 - \delta v_j}{2} |\theta_j + r| = L \left(\frac{1+\delta}{2} \|\theta - rv\|_1 + \frac{1-\delta}{2} \|\theta + rv\|_1 \right).$$

By construction, whenever Θ contains the ℓ_∞ ball of radius r, this risk function has the unique minimizer

$$\theta_v^* := \operatorname*{argmin}_{\theta\in\Theta} R_v(\theta) = rv \in r\{-1, 1\}^d \subset \Theta.$$

The following lemma, due to Agarwal et al. [2012], captures the separation properties of the collection $\{R_v\}_{v\in\mathcal{V}}$ of risk functionals.

LEMMA 5.3. *Assume that Θ contains $[-r, r]^d$ and let R_v be defined as in the preceding paragraph. If $v, w \in \mathcal{V}$ with $v \neq w$, the discrepancy $\rho(R_v, R_w) \geq rLd\delta/2$.*

As a final remark, for random variables $X \in \mathbb{R}^d$, the loss function (29) is Lipschitz continuous (for appropriate choice of L) for any distribution P on X. Specifically, defining the sign(·) function coordinate-wise, we have the subgradient equality $\partial\ell(x, \theta) = L\,\text{sign}(\theta - rx)$. Thus, for any $p \in [1, \infty]$ and $L_p \geq 0$, setting $L = L_p d^{-1/p}$ yields a member of the collection of (L_p, p)-loss functions.

5.2.2. Bounding the Mutual Information. As outlined in Section 5.1, the second step in our lower bound proofs is to bound the mutual information $I(Z_1, \ldots, Z_n; V)$, where Z_i are the private views available to the learning method. Here, we provide mutual information bounds for the family of linear losses (Lemma 5.4) and median-based losses (Lemma 5.5). Each of these mutual information bounds—and our subsequent bounds on mutual information—proceed by using independence to reduce the problem to estimating the mutual information $I(Z; V \mid \theta)$ for a single randomized gradient sample

Z. Then, careful calculation of the distribution of $Z \mid V$ yields the final inequalities. As the proofs are somewhat long and technical, we defer them to Appendix B.

LEMMA 5.4. *Let V be drawn uniformly at random from $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Let X have the distribution* (26) *conditional on $V = v$ and assume $\ell(X, \theta) = L\langle X, \theta\rangle$. Let Z be constructed according to the conditional distribution specified by Proposition 4.3 given a subgradient $\partial\ell(X_i;\theta)$ with $Z \in [-M_\infty, M_\infty]^d$, where $M_\infty \geq L$. Then*

$$I(Z_1, \ldots, Z_n; V) \leq n\frac{\delta^2 L^2}{M_\infty^2}.$$

See Section B.1 of Appendix B for a proof of Lemma 5.4.

LEMMA 5.5. *Let V be drawn uniformly at random from a set $\mathcal{V} \subset \{-1, 1\}^d$. Let X have the distribution* (28) *conditional on $V = v$ and assume $\ell(X, \theta) = L\|rX - \theta\|_1$, where $r > 0$ is a constant. Let Z be constructed according to the distribution specified by Proposition 4.3 conditional on a subgradient $\partial\ell(X_i;\theta)$, where $Z \in [-M_\infty, M_\infty]^d$ and $M_\infty \geq L$. Then*

$$I(Z_1, \ldots, Z_n; V) \leq n\frac{\delta^2 L^2 d}{M_\infty^2}.$$

See Section B.2 of Appendix B for a proof of Lemma 5.5.

5.2.3. Applying Testing Inequalities. Having established the two families of loss functions we consider and the resultant mutual information bounds, it remains to apply Lemma 5.1 and a testing inequality. We begin by proving part (a) of the theorem.

PROOF OF THEOREM 3.4(a). We divide the proof of part (a) of the theorem into two parts: one assuming the dimension $d \geq 9$ and the other assuming $d < 9$. For the first, we use Fano's inequality (24), while for the second, an application of Le Cam's method (25) completes the result. For both results, we use the median-type loss $\ell(X, \theta) = L\|rX - \theta\|_1$. We first recall the beginning of the previous section, stating the following application of Lemma 5.1 and Fano's inequality (24):

$$\frac{2}{\rho^\star(\mathcal{V})}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \mathbb{P}_{P,Q}(\widehat{v}(\mathcal{M}) \neq V) \geq 1 - \frac{I(Z_1, \ldots, Z_n; V) + \log 2}{\log|\mathcal{V}|}. \tag{30}$$

Now we give the proof of the first statement of the theorem in the case that $d \geq 9$. Applying Lemmas 5.3 and 5.5, we immediately have the following specialization of the inequality (30):

$$\frac{4}{rLd\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log|\mathcal{V}|} - n\frac{\delta^2 L^2 d}{M_\infty^2 \log|\mathcal{V}|}.$$

Taking the set $\mathcal{V} \subset \{-1, 1\}^d$ to be a $d/4$ packing of the hypercube $\{-1, 1\}^d$ satisfying $|\mathcal{V}| \geq \exp(d/8)$, as in our construction of median-type losses in Section 5.2.1, we see that

$$\frac{4}{rLd\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{8\log 2}{d} - n\frac{8\delta^2 L^2}{M_\infty^2}.$$

The numerical inequality $8\log 2 < 6$ coupled with the preceding bound implies

$$\frac{4}{rdL\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] > 1 - \frac{6}{d} - 8n\frac{\delta^2 L^2}{M_\infty^2}.$$

By our assumption that $d \geq 9$, if we choose $\delta = \min\{M_\infty/8L\sqrt{n}, 1\}$, we are guaranteed the lower bound $\frac{4}{rdL\delta}\mathbb{E}_{P,Q}\left[\epsilon_n(\mathcal{M}, \ell, \Theta, P)\right] > \frac{1}{5}$, or equivalently

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] > \frac{rdL\delta}{20} = \frac{1}{20} \cdot \min\left\{rdL, \frac{M_\infty rd}{8\sqrt{n}}\right\}.$$

When $d < 9$, we may reduce to the case that $d = 1$, since a lower bound in this setting extends to higher dimensions (though we may lose dimension dependence). For this case, we use the packing set $\mathcal{V} = \{-1, 1\}$ with the linear loss function from Lemma 5.2, which has $\rho^\star(\mathcal{V}) = Lr\delta$. In this case, the marginal distribution $Q(\cdot \mid V)$ is given by

$$Q(Z = z \mid V = 1) = \frac{1}{2} + \begin{cases} \frac{\delta L}{2M} & \text{if } z = M \\ -\frac{\delta L}{2M} & \text{otherwise, that is, if } z = -M. \end{cases}$$

Now, let $Q^n(\cdot \mid V)$ denote the distribution of $Z_1, \ldots, Z_n$ conditional on V. Then, applying Lemma 5.1 and Le Cam's lower bound (25), we obtain the inequality

$$\frac{2}{rL\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \mathbb{P}_{P,Q}\left(\widehat{v}(\mathcal{M}) \neq V\right) \geq \frac{1}{2} - \frac{1}{2}\left\|Q^n(\cdot \mid V{=}1) - Q^n(\cdot \mid V{=}{-1})\right\|_{\rm TV}.$$

A standard result on the total variation distance of Bernoulli distributions (see Lemma B.2 in B.5 of Appendix B) implies that

$$\|Q^n(\cdot \mid V = 1) - Q^n(\cdot \mid V = -1)\|_{\rm TV} \leq \frac{\delta L}{M}\sqrt{(3/2)n}$$

if $\delta \leq M/(3L)$. Thus, we have the bound

$$\frac{2}{rL\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{2} - \frac{\sqrt{3n}}{2\sqrt{2}} \cdot \frac{\delta L}{M}. \tag{31}$$

Multiplying both sides by $rL\delta$, then setting $\delta = \min\{M/(3L\sqrt{n}), 1\} \leq M/(3L)$, we have

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \left(\frac{1}{2} - \frac{1}{2\sqrt{6}}\right)\frac{rL\delta}{2} \geq \frac{\sqrt{6}-1}{4\sqrt{6}} rL \min\left\{\frac{M}{3L\sqrt{n}}, 1\right\}.$$

In turn, for any $d \leq 8$, we immediately find that $(\sqrt{6}-1)/4\sqrt{6} \geq d/(9 \cdot 20)$, which completes the proof of Theorem 3.4(a).

Proof of Theorem 3.4(b). For the second statement of the theorem, we use the linear losses of Section 5.2.1 and apply Lemmas 5.2 and 5.4 with the choice $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. In this case, the lower bound (30) and Lemma 5.2's separation guarantee imply that

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log(2d)} - \frac{I(Z_1, \ldots, Z_n; V)}{\log(2d)}.$$

We may assume that $d \geq 2$ (using the result of part (a) for $d = 1$), and we have $\log 2/\log(2d) \leq 1/2$, which, after an application of Lemma 5.4, yields

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{2} - n\frac{\delta^2 L^2}{M_\infty^2 \log(2d)}.$$

If we choose $\delta = \min\{M_\infty\sqrt{\log(2d)}/2L\sqrt{n}, 1\}$, we see that we have

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{4},$$

which is equivalent in this case to

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{rL\delta}{8} = \frac{1}{8}\min\left\{Lr, \frac{M_\infty r\sqrt{\log(2d)}}{2\sqrt{n}}\right\}.$$

5.3. Proof of Theorem 3.5

The proof of Theorem 3.5 is quite similar to that of Theorem 3.4, again following our outline from Section 5.1. In this section, however, we construct a different family of loss functions, necessitating a new mutual information bound.

5.3.1. Constructing Well-Separated Losses. We construct families of losses that are useful for analyzing the case of stochastic subgradients bounded in ℓ_1-norm. As was the case with median losses (recall Section 5.2.1), let $\mathcal{V} \subset \{-1, 1\}^d$ be a $d/4$-packing of the hypercube in ℓ_0-norm; we know there is such a set with cardinality $|\mathcal{V}| \geq \exp(d/8)$. As our sampling process for the data, we choose X from among the $2d$ positive and negative standard basis vectors $\pm e_j$, that is:

$$\text{Choose index } j \in \{1, \ldots, d\} \text{ uniformly at random,}$$
$$\text{and set } X = \begin{cases} e_j & \text{w.p. } \frac{1+\delta v_j}{2} \\ -e_j & \text{w.p. } \frac{1-\delta v_j}{2}, \end{cases} \tag{32}$$

where $\delta \in (0, 1]$ is fixed. For a fixed $L > 0$, we define the hinge loss, common in classification problems,

$$\ell(x, \theta) = L\,[r - \langle x, \theta\rangle]_+ \,. \tag{33}$$

The combination of hinge loss (33) and sampling strategy (32) yields the risk functional

$$R_v(\theta) = \frac{L}{d}\sum_{j=1}^{d}\frac{1+\delta v_j}{2}[r - \langle e_j, \theta\rangle]_+ + \frac{L}{d}\sum_{j=1}^{d}\frac{1-\delta v_j}{2}[r + \langle e_j, \theta\rangle]_+.$$

Assuming that Θ contains the ℓ_∞ ball of radius r, the (unique) minimizer of the risk over Θ is

$$\theta_v^* := \operatorname*{argmin}_{\theta\in\Theta} R_v(\theta) = rv \in r\{-1, 1\}^d \subset \Theta.$$

Moreover, this risk has the following properties.

LEMMA 5.6. *For any set $\Theta \supseteq [-r, r]^d$, we have*

(a) *for P with support* supp $P \subseteq \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$, *the loss function* (33) *is L-Lipschitz with respect to the ℓ_∞-norm;*
(b) *if $v, w \in \mathcal{V}$ with $v \neq w$, the discrepancy $\rho(R_v, R_w) \geq rL\delta/2$.*

PROOF. The first claim is immediate (cf. Hiriart-Urruty and Lemaréchal [1996]), since $\|\partial\ell(x, \theta)\|_1 \leq L\|x\|_1 \leq L$. For the second statement of the lemma, as in the proof of Lemma 5.2, we verify the separation condition directly by computing the minimizers of R_v. The minimum of

$$\begin{aligned} R_v(\theta) + R_w(\theta) \;=\; & \frac{L}{d}\sum_{j=1}^{d}\big([r - \langle e_j, \theta\rangle] + [r + \langle e_j, \theta\rangle]\big) \\ & + \frac{L\delta}{d}\sum_{j: v_j = w_j} v_j\big([r - \langle e_j, \theta\rangle - [r + \langle e_j, \theta\rangle]_+\big) \end{aligned}$$

is attained by any $\theta \in \mathbb{R}^d$ with $\theta_j \in [-r, r]$ for j such that $v_j \neq w_j$ and $\theta_j = rv_j$ for j such that $v_j = w_j$; a minimizer of R_v is $\theta_v^* = rv$. Thus we have

$$\inf_{\theta \in \Theta}\{R_v(\theta) + R_w(\theta)\} - R_v(\theta_v^*) - R_w(\theta_w^*) = \frac{L}{d}\sum_{j=1}^{d} 2r - \frac{2L}{d}\sum_{j: v_j = w_j} r\delta - 2Lr(1-\delta)$$

$$= 2Lr - 2Lr + 2Lr\delta - \frac{2Lr\delta}{d}(d - \|v - w\|_0) = \frac{2Lr\delta}{d}\|v - w\|_0 .$$

Since $\|v - w\|_0 \geq d/4$ by construction, we have $\rho(R_v, R_w) \geq rL\delta/2$, as desired. □

5.3.2. Bounding the Mutual Information. For Theorem 3.5, we require a somewhat careful bound on the mutual information between the subgradients and the unknown index. We have the following lemma, whose proof we provide in Section B.3 in Appendix B.

LEMMA 5.7. *Let V be drawn uniformly at random from a set $\mathcal{V} \subset \{-1, 1\}^d$. Define the distribution $P(\cdot \mid A)$ on X as in the random sampling scheme* (32) *and use the loss* (33). *Let Z be constructed according to the conditional distribution specified by Proposition 4.4, where $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \leq M_1\}$, and define $M = M_1/L$. Then*

$$I(Z_1, \ldots, Z_n; V) \leq n\delta^2\Delta(\gamma)^2,$$

where

$$\gamma := \log\left(\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2 - 1)}}{2(M-1)}\right) \quad \textit{and} \quad \Delta(\gamma) := \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d-1)}.$$

5.3.3. Applying Testing Inequalities. The remainder of the proof is similar to that of Theorem 3.4, except that we apply Lemma 5.7 in place of Lemmas 5.4 or 5.5. Indeed, following identical steps to those in the proof of Theorem 3.4, we see that with the specified packing $\mathcal{V} \subset \{-1, 1\}^d$ of size $|\mathcal{V}| \geq \exp(d/8)$, we have (recall Eq. (30))

$$\frac{4}{rL\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log|\mathcal{V}|} - n\frac{\delta^2\Delta(\gamma)^2}{\log|\mathcal{V}|} \geq 1 - \frac{6}{d} - 8n\frac{\delta^2\Delta(\gamma)^2}{d}.$$

Consequently, if we choose $\delta = \min\{\sqrt{d}/(8\Delta(\gamma)\sqrt{n}), 1\}$, then for all $d \geq 9$, we have the lower bound $\frac{4}{rL\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{5}$, or equivalently

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{rL\delta}{20} = \frac{1}{20}\min\left\{rL, \frac{rL\sqrt{d}}{8\sqrt{n}\Delta(\gamma)}\right\},$$

which completes the proof (the case $d \leq 8$ is identical to that in Theorem 3.4).

5.4. Proof of Theorem 3.8

We are somewhat more terse in our proof of Theorem 3.8 than the previous two, though we repeat the same steps to emphasize our technique.

5.4.1. Constructing Well-Separated Losses. We begin by choosing the family of loss functions we require: since our optimization domain $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$, we use the linear losses of Section 5.2.1 with the sampling scheme (26) as in Theorem 3.4. Thus, using the packing set $\mathcal{V} = \{\pm e_i\}_{i=1}^d$, we find that $\rho^\star(\mathcal{V}) = Lr\delta$, and consequently

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log(2d)} - \frac{I(Z_1, \ldots, Z_n; V)}{\log(2d)}$$

as earlier.

5.4.2. Bounding the Mutual Information. The mutual information bound in this theorem is somewhat more complicated than the previous bounds, as the optimal privacy-preserving distribution (recall Proposition 4.6) is more complex. We begin by stating a lemma.

LEMMA 5.8. *Let V be drawn uniformly at random from $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Let $X \mid V$ be sampled according to the distribution* (26)*, and let $Z \mid X = x$ have support on $\{-1, 1\}^d$ and have p.m.f.*

$$q(z \mid x) \propto \begin{cases} \exp(\alpha) & \text{if } z^\top x > k \\ 1 & \text{if } z^\top x \le k \end{cases}$$

for some $k \ge 0$. Define the constants $C_d(k)$ and $\Delta(\delta, \alpha, d, k)$ by

$$C_d(k) := \mathrm{card}\{z \in \{-1, 1\}^d : \langle z, x\rangle > k\} = \sum_{i=0}^{\lceil (d-k)/2\rceil - 1} \binom{d}{i},$$

and

$$\Delta(\delta, \alpha, d, k) := \delta \cdot \frac{e^\alpha - 1}{(e^\alpha + 1)C_d(k) + 2^d} \binom{d-1}{\lceil (d-k)/2\rceil - 1}.$$

Then

$$I(Z; V) \le \Delta(\delta, \alpha, d, k)^2.$$

We provide the proof of the lemma in Section B.4 of Appendix B.

For any $\alpha \le 5/4$, we have $e^\alpha - 1 \le 2\alpha$, and by properties of binomial coefficients and Stirling's approximation we have

$$\frac{1}{2^d}\binom{d-1}{\lceil (d-k)/2\rceil - 1} \le \frac{1}{2^d}\binom{d-1}{\lceil d/2\rceil - 1} \le \frac{1}{\sqrt{d}}$$

for any k. For any distribution Q satisfying optimal local differential privacy at a differential privacy level α, Proposition 4.6 implies Q is a convex combination of distributions with p.m.f.s of the form in Lemma 5.8. That is, we sample with a channel Q whose p.m.f. is a convex combination of p.m.f.s of the form

$$q_k(z \mid x) = \frac{1}{e^\alpha C_d(k) + (2^d - C_d(k))} \cdot \begin{cases} \exp(\alpha) & \text{if } z^\top x/M > k \\ 1 & \text{if } z^\top x/M \le k \end{cases} \quad \text{for } z \in \{-M, M\}^d,$$

so $Q(Z = z \mid x) = \sum_k \beta_k q_k(z \mid x)$ for some $\beta_k \ge 0$ with $\sum_k \beta_k = 1$. Applying the convexity of mutual information—taking a convex combination of channel distributions Q can only reduce mutual information—and Lemma 5.8, we thus obtain

$$\begin{aligned} I(Z_1, \ldots, Z_n; V) &\le n \max_{k \ge 0} \Delta(\delta, \alpha, d, k)^2 \\ &\le n\delta^2 (e^\alpha - 1)^2 \max_k \left(\frac{1}{(e^\alpha + 1)C_d(k) + 2^d} \binom{d-1}{\lceil (d-k)/2\rceil - 1} \right)^2 \\ &\le 4n\delta^2\alpha^2 \cdot \frac{1}{d}. \end{aligned} \tag{34}$$

5.4.3. Applying Testing Inequalities. As a consequence of the display (34), we have the lower bound

$$\frac{2}{Lr\delta} \mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \ge \frac{1}{2} - \max_k \frac{n\Delta(\delta, \alpha, d, k)^2}{\log(2d)} \ge \frac{1}{2} - \frac{4n\delta^2\alpha^2}{d \log(2d)}.$$

By choosing $\delta = \min\{\sqrt{d\log(2d)}/4\alpha\sqrt{n}, 1\}$, we find that

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{4},$$

which is equivalent to the bound given in the theorem.

5.5. Proof of Theorem 3.10

In our proofs of Theorems 3.10 and 3.11, we exploit some of our own recent results [Duchi et al. 2013] on the contractive properties of mutual information and KL-divergence under local differential privacy. A few definitions are required in order to state these results. As usual, we have an indexed set of probability measures $\{P_v\}_{v\in\mathcal{V}}$, and we let $Q^n(\cdot \mid x_1, \ldots, x_n)$ denote the joint probability of the n released private random variables $Z_1, \ldots, Z_n$. For each $v \in \mathcal{V}$, we then define the *marginal* distribution

$$M_v^n(A) := \int Q^n(A \mid x_1, \ldots, x_n) dP_v^n(x_1, \ldots, x_n) \quad \text{for } A \in \sigma(\mathcal{Z}^n). \tag{35}$$

Duchi et al. [2013] establish two results that provide bounds on $\|M_v^n - M_w^n\|_{\text{TV}}$ and $I(Z_1, \ldots, Z_n; V)$ as a function of the amount of privacy provided and the distances between the underlying distributions P_v.

The bounds apply to any channel distribution Q that is α-locally differentially private (for the first result) and to any noninteractive α-locally differentially private channel (for the second result; recall Definition (3)). Let P_v be the distribution of X conditional on the random packing element $V = v$, and let M_v^n be the marginal distribution (35) induced by passing X_i through Q. Define the mixture distribution $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v$, and let $\sigma(\mathcal{X})$ denote the σ-field on $\mathcal{X}$ over which P_v are defined. With this notation we can state the following proposition, which summarizes the results we need from Duchi et al. [2013, Theorems 1–2, Corollaries 1–2].

PROPOSITION 5.9 (INFORMATION BOUNDS). *Let the conditions of the previous paragraph hold and assume that Q is α-locally differentially private.*

(*a*) *For all $\alpha \geq 0$,*

$$D_{\text{kl}}\left(M_v^n \| M_w^n\right) \leq 4n(e^\alpha - 1)^2 \left\|P_v - P_w\right\|_{\text{TV}}^2. \tag{36}$$

(*b*) *If Q is noninteractive, then for all $\alpha \geq 0$,*

$$I(Z_1, \ldots, Z_n; V) \leq e^\alpha n(e^\alpha - e^{-\alpha})^2 \sup_{S\in\sigma(\mathcal{X})} \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}(P_v(S) - \overline{P}(S))^2. \tag{37}$$

With Proposition 5.9 in place, we proceed with the proof of Theorem 3.10. To establish the lower bound, we follow the outline of Section 5.1. Establishing the upper bound requires a few additional steps. We defer the formal proofs of achievability to Appendix C (see C.1), as the proofs are similar to Corollaries 3.6–3.9.

5.5.1. Constructing Well-Separated Losses. Our lower bound uses an identical construction to that in Section 5.2.1. We let the loss function $\ell(x, \theta) = L\langle x, \theta\rangle$, and we use the distribution (26) on x; that is, we have $\mathcal{V} = \{\pm e_j\}_{j=1}^d$ and for $\delta \in [0, 1/2]$, we sample vectors from $\mathcal{X} = \{-1, 1\}^d$ with probability $P_v(X = x) = (1 + \delta v^\top x)/2^d$. We then have $\rho^\star(\mathcal{V}) = Lr\delta$ and $|\mathcal{V}| = 2d$ (recall Lemma 5.2).

5.5.2. Bounding the Mutual Information. With the construction of the (nearly) uniform sampling scheme, we have the following lemma [Duchi et al. 2013, Lemma 7].

LEMMA 5.10. *Under the conditions of the previous paragraph, let* $\delta \le 1$ *and* V *be sampled uniformly from* $\{\pm e_j\}_{j=1}^d$. *For any noninteractive* α*-differentially private channel* Q,

$$I(Z_1, \ldots, Z_n; V) \le n\frac{e^\alpha}{4d}(e^\alpha - e^{-\alpha})^2\delta^2.$$

5.5.3. Applying Testing Inequalities. Using Lemma 5.10, we can give an almost immediate proof of the lower bound in Theorem 3.10. Indeed, using Fano's inequality (24), Lemmas 5.1 and 5.10, and the separation $\rho^\star(\mathcal{V}) = Lr\delta$ from Lemma 5.2(b), we obtain

$$\epsilon_n^*(\mathfrak{L}, \Theta, \alpha) \ge \frac{Lr\delta}{2}\left(1 - \frac{ne^\alpha(e^\alpha - e^{-\alpha})^2\delta^2/4d + \log 2}{\log(2d)}\right).$$

So long as $d \ge 2$, setting

$$\delta = \min\left\{\frac{\sqrt{d\log(2d)}}{\sqrt{e^\alpha n}(e^\alpha - e^{-\alpha})}, 1\right\}$$

and noting that $e^\alpha = \mathcal{O}(1)$ and $e^\alpha - e^{-\alpha} \le c\alpha$ for a universal constant c if $\alpha = \mathcal{O}(1)$, completes the proof.

When $d = 1$, an argument via Le Cam's method (25) yields an identical result. Given Proposition 5.9, the argument is quite similar to that used in the proof of Theorem 3.4. We use the packing set $\mathcal{V} = \{\pm 1\}$ and conditional on $V = v$, set $X = 1$ with probability $(1 + v\delta)/2$ and $X = -1$ with probability $(1 - v\delta)/2$, which yields separation $\rho^\star(\mathcal{V}) = Lr\delta$ by Lemma 5.2. We also have the marginal contraction

$$\left\|M_1^n - M_{-1}^n\right\|_{\mathrm{TV}}^2 \le \frac{1}{2}D_{\mathrm{kl}}\left(M_1^n \| M_{-1}^n\right) \le 2(e^\alpha - 1)^2 n \left\|P_1 - P_{-1}\right\|_{\mathrm{TV}}^2$$

by Pinsker's inequality and Proposition 5.9. By construction, the total variation $\|P_1 - P_{-1}\|_{\mathrm{TV}} = \delta$, whence we find that $\|M_1^n - M_{-1}^n\|_{\mathrm{TV}}^2 \le 2(e^\alpha - 1)^2 n\delta^2$. Applying Le Cam's method (25) and Lemma 5.1, we obtain

$$\epsilon_n^*(\mathfrak{L}, \Theta, \alpha) \ge \frac{Lr\delta}{2}\left(\frac{1}{2} - \frac{\sqrt{n}(e^\alpha - 1)\delta}{\sqrt{2}}\right).$$

Take $\delta = \min\{(2\sqrt{2}\sqrt{n}(e^\alpha - 1))^{-1}, 1\}$ to complete the proof in this case.

5.6. Proof of Theorem 3.11

The proof of this theorem follows the outline established in Section 5.1, as did the previous results. We defer the attainability results to C.2 of Appendix C.

5.6.1. Constructing Well-Separated Losses. Before constructing the well-separated loss functions, we exhibit the set $\mathcal{V}$ that underlies our sampling distributions. The following lemma exhibits the existence of a special packing of the Boolean hypercube [Duchi et al. 2013, Lemma 5].

LEMMA 5.11. *There exists a packing* $\mathcal{V}$ *of the* d*-dimensional hypercube* $\{-1, 1\}^d$ *with* $\|v - w\|_1 \ge d/2$ *for each* $v, w \in \mathcal{V}$ *with* $v \ne w$ *such that the cardinality of* $\mathcal{V}$ *is at least* $\lceil\exp(d/16)\rceil$ *and*

$$\frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} vv^\top \preceq 25 I_{d\times d}.$$

With this packing $\mathcal{V}$, we as usual let $V \in \mathcal{V}$, and conditional on $V = v$, we sample $X \in \{\pm e_j\}_{j=1}^d$ according the sampling scheme (32).

Linear Losses (*for the Bound* (18)). We first consider the case that the loss functions are linear functionals that are L-Lipschitz with respect to the $\ell_{p'}$-norm for some $p' \geq 2$, and we optimize over ℓ_q balls $\mathbb{B}_q(r_q)$. In this case, we define the loss $\ell(x, \theta) = L\langle x, \theta\rangle$, and we note that since $\|x\|_1 \leq 1$ for $x \in \mathcal{X}$, the function $\ell(x, \cdot)$ is L-Lipschitz with respect to the ℓ_∞-norm. With the sampling scheme (32), $R_v(\theta) = L\delta\langle v, \theta\rangle/d$. Since $\inf_{\theta\in\mathbb{B}_q(r_q)}\langle v, \theta\rangle = -\|v\|_p$ when $p = (1 - 1/q)^{-1}$ is the conjugate of q, we have the pairwise separation

$$\rho(R_v, R_w) = -\frac{r_q L\delta}{d}(\|v + w\|_p - \|v\|_p - \|w\|_p).$$

With the packing set $\mathcal{V}$ exhibited by Lemma 5.11, we have

$$\|v + w\|_p^p = \sum_{j=1}^{d}(v_j + w_j)^p = \sum_{j:v_j=w_j} 2^p \leq \frac{3d}{4}2^p,$$

and $\|v\|_p = d^{1/p}$ for each $v \in \mathcal{V}$. This implies the following lower bound on the discrepancy:

$$\rho^\star(\mathcal{V}) \geq \max_{v\neq w}\rho(R_v, R_w) \geq \frac{rL\delta}{d}(2d^{1/p}(1 - (3/4)^{1/p})) \geq \frac{3}{5}rL\delta d^{\frac{1}{p}-1}. \tag{38}$$

General Loss Functions (*for the Bound* (19)). For the general lower bound of the theorem, we use the hinge loss $\ell(x, \theta) = L[r - \langle x, \theta\rangle]_+$ as our loss function. In this case, as in Theorem 3.5 (recall Lemma 5.6), our sampling strategy yields that the loss $\ell(x, \theta)$ is an $(L, 1)$ loss, since $\|x\|_1 \leq 1$, and we have the discrepancy bound $\rho^\star(\mathcal{V}) \geq \frac{rL\delta}{2}$, since the separation depends only on the distance $\|v - w\|_1$, which Lemma 5.11 lower bounds.

5.6.2. Bounding the Mutual Information. Using Lemma 5.11, we may bound the mutual information between samples Z from a particular distribution and a random sample V from a set $\mathcal{V}$ of the form in the lemma. Indeed, let $\mathcal{V}$ be a packing of the d-dimensional hypercube specified in Lemma 5.11. Conditional on $V = v \in \{-1, 1\}^d$, we sample the random vector $X \in \{\pm e_j\}_{j=1}^d$ according to the sampling scheme (32). Then, we have the following lemma [Duchi et al. 2013, Lemma 6], which applies as long as the channel Q is non-interactive and α-locally differentially private.

LEMMA 5.12. *Let Z_i be α-locally differentially private for X_i and the conditions of the previous paragraph hold. Then*

$$I(Z_1, \ldots, Z_n; V) \leq n\frac{25e^\alpha}{16}\frac{\delta^2}{d}(e^\alpha - e^{-\alpha})^2.$$

5.6.3. Applying Testing Inequalities. Our last step is to apply the usual testing inequalities. We first prove the lower bound in inequality (18). Let $\mathfrak{L}_{\text{lin}} = \mathfrak{L}_{\text{lin}}(\mathbb{B}_q(r_q); L, p)$. Then, by applying Lemma 5.12 and Fano's inequality (24) to Lemma 5.1—using the separation (38)—we obtain

$$\epsilon_n^*(\mathfrak{L}_{\text{lin}}, \Theta, \alpha) \geq \frac{3r_q L\delta d^{\frac{1}{p}-1}}{10}\left(1 - \frac{25ne^\alpha\delta^2(e^\alpha - e^{-\alpha})^2/16d + \log 2}{d/16}\right).$$

So long as $d \geq 16$, we have $16\log 2/d \leq \log 2 < 7/10$. Thus, choosing

$$\delta = \min\left\{\frac{d}{10\sqrt{nC_\alpha}(e^\alpha - e^{-\alpha})}, 1\right\}$$

and noting that $e^\alpha - e^{-\alpha} = \mathcal{O}(\alpha)$ and $e^\alpha = \mathcal{O}(1)$ for $\alpha = \mathcal{O}(1)$, we obtain

$$\epsilon_n^*(\mathfrak{L}, \Theta, \alpha) \geq \frac{3r_q L\delta d^{\frac{1}{p}-1}}{10} \cdot \frac{1}{20} \geq c \min\left\{\frac{r_q L d^{\frac{1}{p}}}{\sqrt{n\alpha^2}}, r_q L d^{\frac{1}{p}-1}\right\} = c r_q L \min\left\{\frac{d^{1-\frac{1}{q}}}{\sqrt{n\alpha^2}}, d^{\frac{1}{p}-1}\right\}$$

for a universal constant c. As in the proof of Theorem 3.10, when $d < 16$, we apply an essentially similar argument but with Le Cam's method (25), which gives the desired result. (The proof of the case $d = 1$ from Theorem 3.10 applies here as well.) Finally, we remark that we may repeat a completely identical proof to the previous proof, replacing d with any $k \leq d$, which *mutatis mutandis* implies the lower bound

$$\epsilon_n^*(\mathfrak{L}, \Theta, \alpha) \geq c L r_q \max_{k \in [d]} \min\left\{\frac{k^{1-\frac{1}{q}}}{\sqrt{n\alpha^2}}, k^{\frac{1}{p}-1}\right\} \geq c L r_q \min\left\{\frac{d^{1-\frac{1}{q}}}{\sqrt{n\alpha^2}}, \frac{(n\alpha^2)^{\frac{1}{2p}}}{\sqrt{n\alpha^2}}, 1\right\},$$

where, for the rightmost bound, we have taken $k = \max\{1, \min\{\sqrt{n\alpha^2}, d\}\}$. Noting that $1/p = 1 - 1/q$ completes the proof of the lower bound (18).

To prove inequality (19), we apply the same reasoning as in the proof of the inequality (18). Use Lemma 5.12 and Fano's inequality (24) (via Lemma 5.1 and the separation from Lemma 5.6) to obtain

$$\epsilon_n^*(\mathfrak{L}, \Theta, \alpha) \geq \frac{rL\delta}{4}\left(1 - \frac{25ne^\alpha(e^\alpha - e^{-\alpha})^2\delta^2/16d + \log 2}{d/16}\right).$$

Then choosing $\delta \asymp d/(\sqrt{n}(e^\alpha - e^{-\alpha}))$ as in the proof of inequality (18) gives the desired result.

5.7. Proof of Corollary 3.6

Since $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$, the bound (14a) guarantees that mirror descent obtains convergence rate $\mathcal{O}(M_\infty r\sqrt{\log(2d)}/\sqrt{n})$. This matches the second statement of Theorem 3.4. Now fix our desired amount of mutual information I^*. From the remarks following Proposition 4.3, if we must guarantee that $I^* \geq \sup_P I(P, Q)$ for any distribution P and loss function ℓ whose gradients are bounded in ℓ_∞-norm by L, we must (because of the uniqueness of the optimal privacy distribution Q) have

$$I^* \asymp \frac{dL^2}{M_\infty^2}. \tag{39}$$

Up to higher order terms, to guarantee a level of privacy with mutual information I^*, we must allow gradient noise up to a level $M_\infty = L\sqrt{d/I^*}$. The equality (39) establishes that for a given level of allowed mutual information I^*, if optimal local privacy holds, then we must have $M_\infty \asymp L\sqrt{d}/\sqrt{I^*}$. That is, we have a bijection between I^* and M_∞ whenever optimal local privacy holds, so substituting $M_\infty = L\sqrt{d}/\sqrt{I^*}$ into our upper and lower bounds yields the claim.

5.8. Proof of Corollary 3.7

According to the conditions of optimal local privacy, if we must guarantee that $I^* \geq \sup_P I(P, Q)$ for any loss function ℓ whose gradients are bounded in ℓ_1-norm by L, we must have

$$I^* \asymp \frac{dL^2}{2M_1^2},$$

using Corollary 4.5 after the statement of Proposition 4.4. Rewriting this, we see that we must have $M_1 = L\sqrt{d/2I^*}$ (to higher-order terms) to be able to guarantee an amount of privacy I^*. As in the ℓ_∞ case, we have a bijection between the multiplier M_1 and the amount of information I^* and can apply similar techniques. Now recall the convergence guarantee (14b) provided by stochastic gradient descent. Since the ℓ_∞-ball of radius r is contained in the ℓ_2-ball of radius $r_2 = r\sqrt{d}$, and $\|g\|_1 \le \|g\|_2$ for all $g \in \mathbb{R}^d$, stochastic gradient descent guarantees that $\epsilon_n^*(\mathfrak{L}, \Theta) \le CM_1 r\sqrt{d}/\sqrt{n}$. Applying the lower bound provided by Theorem 3.5 and substituting for M_1 completes the proof.

5.9. Proof of Corollary 3.9

Without loss of generality (by scaling), we assume that $L = 1$. Now we consider Proposition 4.6, which characterizes the distributions satisfying optimal local differential privacy. We use the proposition to find an upper bound on M_∞ in terms of the differential privacy level α, which in turn allows us to apply the bound from mirror descent (14a). Instead of directly using Proposition 4.6, it is simpler to use the linear program (61) in its proof, and note that finding a lower bound on t (in the LP) as a function of α provides an upper bound on M_∞ since $M_\infty = 1/t$. Now, in the linear program (61), we choose the values for $q(z)$ specified by Lemma E.2. Let q_+ and q_- denote the larger and smaller probabilities, respectively. Fix an $x \in \{-1, 1\}^d$, and let z range over $\{-1, 1\}^d$. With those choices, we note that for d odd,

$$\sum_{z:\langle z,x\rangle>0} z = \sum_{z:\langle z,x\rangle=1} z + \sum_{z:\langle z,x\rangle=3} z + \cdots + \sum_{z:\langle z,x\rangle=d} z$$
$$= \left[\binom{d-1}{\frac{d-1}{2}} - \binom{d-1}{\frac{d+1}{2}}\right] x + \left[\binom{d-1}{\frac{d+1}{2}} - \binom{d-1}{\frac{d+3}{2}}\right] x + \cdots = \binom{d-1}{\frac{d-1}{2}} x.$$

For d even, a similar calculation yields $\sum_{z:\langle z,x\rangle>0} z = \binom{d-1}{d/2} x$. As a consequence, we find that

$$\sum_z zq(z \mid x) = q_+ \sum_{z:\langle z,x\rangle>0} z + q_- \sum_{z:\langle z,x\rangle\le 0} z = x(q_+ - q_-) \cdot \begin{cases} \binom{d-1}{\frac{d-1}{2}} & d \text{ odd} \\ \binom{d-1}{d/2} & d \text{ even.} \end{cases}$$

Focusing on the odd case for simplicity—identical bounds hold in the even case—we have for a universal constant $c > 0$ that

$$(q_+ - q_-)\binom{d-1}{\frac{d-1}{2}} = \frac{e^\alpha - 1}{2^{d-1}(e^\alpha + 1)}\binom{d-1}{\frac{d-1}{2}} \ge c\frac{e^\alpha - 1}{e^\alpha + 1}\frac{1}{\sqrt{d}} \ge c\frac{\alpha}{\sqrt{d}},$$

the first inequality following from Stirling's approximation and the second from convexity of the function $\alpha \mapsto e^\alpha$. In particular, we see that the minimizing value t in the linear program (61) will satisfy $t \ge c\alpha/\sqrt{d}$, which in turn yields $M_\infty = 1/t \le \sqrt{d}/(c\alpha)$. Noting that the lower bound in the corollary is given by Theorem 3.8, applying the convergence guarantee (14a) of mirror descent based on M_∞ completes the proof.

6. DISCUSSION

We have studied methods for protecting privacy in general statistical risk minimization problems, and have described general techniques for obtaining sharp tradeoffs between privacy protection and estimation rates. The latter are a natural measure of utility for statistical problems.

We believe that there are a number of interesting open issues and areas for future work. First, we studied procedures that access each datum only once, and through a perturbed view Z_i of the subgradient $\partial\ell(X_i, \theta)$, which is natural in the context of convex risk minimization. A natural question is whether there are restrictions of the class of loss functions so that a transformed version $(Z_1, \ldots, Z_n)$ of the data are sufficient for inference. For instance, other researchers [Zhou et al. 2009a, 2009b] have studied applications in which a data matrix $X = [X_1 \ \cdots \ X_n]^\top \in \mathbb{R}^{n\times d}$ is premultiplied by a normal matrix $\Phi \in \mathbb{R}^{m\times n}$, where $m \ll n$, and statistical inference is performed using ΦX. For problems such as linear regression and PCA, the resulting estimators enjoy good statistical properties. This transformation, however, cannot be computed without the entire dataset at one's disposal. Nonparametric data releases, such as those studied by Hall et al. [2011], could provide insights here, though again, current approaches require the data to be aggregated by a trusted curator before release.

Our constraints on the privacy-inducing channel distribution Q require that its support lie in some compact set. We find this restriction useful, but perhaps it possible to achieve faster estimation rates if all we require are moment conditions, for example, $\mathbb{E}_Q[\|Z - X\|_p^2 \mid X] \le M^2$. A better understanding of general privacy-preserving channels Q for alternative constraints to those we have proposed is also desirable. Moreover, one might consider attempting only to guarantee that $\phi(X)$ is private, where ϕ is some (known) function. For example, members of a dataset may not care if their genders are known, but more personal features of X may be more sensitive.

These questions do not appear to have easy answers, especially when we wish to allow each provider of a single datum to be able to guarantee his or her own privacy. Nevertheless, we hope that our view of privacy and the techniques we have developed in this article prove fruitful, and we hope to investigate some of these issues in future work.

APPENDIXES

A. UNBIASEDNESS

In this appendix, we show that if an optimization procedure receives biased subgradients it is possible to be arbitrarily wrong. We do so by constructing a simple problem instance. Fix a bias $b > 0$ and consider the following one-dimensional problem:

$$\text{minimize} \ \ f(\theta) := \frac{b\theta}{2} \ \ \text{subject to} \ \ \theta \in [-c, c].$$

If a gradient oracle returns biased gradients of the form $-b/2$ at each point $\theta \in [-c, c]$, it is impossible to distinguish the objective from $-b\theta/2$. The minimizer of this objective is $\theta_{\text{bias}} = \text{sign}(b)c$. The true optimal point is $\theta^* = -\text{sign}(b)c$, yielding the worst possible error

$$f(\theta_{\text{bias}}) - f(\theta^*) = \sup_{\theta\in[-c,c]} f(\theta) - \inf_{\theta\in[-c,c]} f(\theta).$$

We can show this more formally using an information theoretic derivation similar to that in Section 5. Omitting details, the argument is as follows. In the notation of Section 5, if a bias is chosen independently of the parameters $v \in \mathcal{V}$ of the risk R_v, then there is a *bounded* amount of mutual information that can be communicated to any optimization procedure. Consequently, Fano's inequality (24) guarantees that the estimation accuracy of any procedure must be bounded away from zero.

B. CALCULATION OF THE MUTUAL INFORMATION FOR SAMPLING STRATEGIES

This appendix is devoted to the proofs of our bounds on mutual information: Lemma 5.4, Lemma 5.5, Lemma 5.7, and Lemma 5.8. Before proving the lemmas, we make an observation that allows us to tensorize the mutual information, making our arguments simpler (we need only compute the single observation information $I(Z_1; V)$). For each of Lemmas 5.4, 5.5, 5.7, and 5.8, recall that $Z_1, \ldots, Z_n$ are constructed based on an evaluation of the subgradient set $\partial_\theta \ell(X_i, \theta)$, where X_i are independent samples according to a distribution $P(\cdot \mid V)$. Then, the samples Z_i are conditionally independent of V

given X_i and the parameters θ, since Z is a random function of $\partial\ell(X_i, \theta)$. Our goal is to upper bound the mutual information between the sequence $Z_1, \ldots, Z_n$ of observed (stochastic) gradients and the random element $V \in \mathcal{V}$.

By the general definition of mutual information [Gray 1990, Chapter 5], it is no loss of generality to assume (temporarily) that the random variable Z_i are supported on finite sets. Thus (using the chain rule for mutual information [Cover and Thomas 2006; Gray 1990, Chapter 5]), we have the decomposition

$$I(Z_1, \ldots, Z_n; V) = \sum_{i=1}^{n} [H(Z_i \mid Z_1, \ldots, Z_{i-1}) - H(Z_i \mid V, Z_1, \ldots, Z_{i-1})].$$

Let θ_i denote the point at which the ith gradient is computed. Then, by inspection, we must have $\theta_i \in \sigma(Z_1, \ldots, Z_{i-1})$. Since Z_i is conditionally independent of $Z_1, \ldots, Z_{i-1}$ given V and θ_i and conditioning decreases entropy, we have

$$\begin{aligned} H(Z_i \mid Z_1, \ldots, Z_{i-1}) - H(Z_i \mid V, Z_1, \ldots, Z_{i-1}) &= H(Z_i \mid Z_1, \ldots, Z_{i-1}) - H(Z_i \mid V, \theta_i) \\ &\le H(Z_i \mid \theta_i) - H(Z_i \mid V, \theta_i) \\ &= I(Z_i; V \mid \theta_i). \end{aligned}$$

In particular, letting F_i denote the distribution of θ_i, we have

$$I(Z_1, \ldots, Z_n; V) \le \sum_{i=1}^{n} \int_{\Theta} I(Z_i; V \mid \theta) dF_i(\theta) \le \sum_{i=1}^{n} \sup_{\theta \in \Theta} I(Z_i; V \mid \theta). \tag{40}$$

The representation (40) is the key to our calculations in this appendix.

In addition, the proofs of Lemmas 5.5 and 5.7 require a minor lemma, which we present here before giving the proofs proper.

LEMMA B.1. *Let* $1 > p > \delta > 0$ *and* $p + \delta \le 1$. *Then*

$$(p + \delta)\log(p + \delta) + (p - \delta)\log(p - \delta) > 2p \log p.$$

PROOF. Since the function $p \mapsto f(p) = p \log p$ is strictly convex over $[0, \infty)$, we may apply convexity. Indeed, $p = \frac{1}{2}(p + \delta) + \frac{1}{2}(p - \delta)$, so

$$p \log p = f\left(\frac{1}{2}(p + \delta) + \frac{1}{2}(p - \delta)\right) < \frac{1}{2} f(p + \delta) + \frac{1}{2} f(p - \delta),$$

which is the desired result. □

B.1. Proof of Lemma 5.4

The subgradient set $\partial\ell(X_i; \theta)$ is independent of θ, so we may use the inequality (40) to bound the mutual information of V and a single sample Z while ignoring the dependence on θ. Define $M = M_\infty/L$. Since the sampling scheme (26) is independent per-coordinate, we see immediately that if Z_j denotes the jth coordinate of Z, then

$$I(Z; V) = H(Z) - H(Z \mid V) \le d \log(2) - \sum_{j=1}^{d} H(Z_j \mid V).$$

Since V is uniformly chosen from one of $2d$ vectors, we additionally find that

$$I(Z; V) \le d \left[\log 2 - \frac{1}{2d} \sum_{v \in \mathcal{V}} H(Z \mid V = v) \right].$$

By the choice of our sampling scheme for X and Z, we see that $H(Z \mid V = v)$ is identical for each $v \in \mathcal{V}$, and we have

$$Q(Z_j = M_\infty \mid V_j = v_j = 0) = Q(Z_j = -M_\infty \mid V_j = v_j = 0) = \frac{1}{2}.$$

On the other hand, by our choice of sampling scheme, for the "on" index in V, we have

$$\begin{aligned} Q(Z_j = -M_\infty \mid V_j = v_j = -1) &= Q(Z_j = M_\infty \mid V_j = v_j = 1) \\ &= Q(Z_j = M_\infty \mid X_j = 1)P(X_j = 1 \mid V_j = v_j = 1) \\ &\quad + Q(Z_j = M_\infty \mid X_j = -1)P(X_j = -1 \mid V_j = v_j = 1) \\ &= \left(\frac{M+1}{2M}\right)\left(\frac{1+\delta}{2}\right) + \left(\frac{M-1}{2M}\right)\left(\frac{1-\delta}{2}\right) = \frac{1}{2} + \frac{\delta}{2M}. \end{aligned}$$

Consequently, defining the Bernoulli entropy $h(p) = -p\log p - (1-p)\log(1-p)$, then

$$\begin{aligned} I(Z; V) &\le d\left[\log 2 - \frac{1}{2d}\left((2d-2)\log 2 + 2h\left(\frac{1}{2} + \frac{\delta}{2M}\right)\right)\right] \\ &= \log 2 + \left(\frac{1}{2} + \frac{\delta}{2M}\right)\log\left(\frac{1}{2} + \frac{\delta}{2M}\right) + \left(\frac{1}{2} - \frac{\delta}{2M}\right)\log\left(\frac{1}{2} - \frac{\delta}{2M}\right). \end{aligned}$$

The concavity of the function $p \mapsto \log(p)$ yields that $\log(1/2 + p) \le \log(1/2) + 2p$, so

$$I(Z; V) \le \log 2 + \left(\frac{1}{2} + \frac{\delta}{2M}\right)\left(-\log 2 + \frac{\delta}{M}\right) + \left(\frac{1}{2} - \frac{\delta}{2M}\right)\left(-\log 2 - \frac{\delta}{M}\right) = \frac{\delta^2}{M^2}.$$

Making the substitution $M = M_\infty / L$ completes the proof.

B.2. Proof of Lemma 5.5

By using the inequality (40), a bound on the mutual information $I(Z; V \mid \theta)$ implies a bound on the joint information in the statement of the lemma, so we focus on bounding the mutual information of a single sample Z. In addition, it is no loss of generality to assume that $r = 1$.

Define $M = M_\infty / L$ to be the multiple of the ℓ_∞-norm of the subgradients that we take, and let Z_j denote the jth coordinate of Z. Using the coordinate-wise independence of the sampling, we have

$$I(Z; V \mid \theta) = H(Z \mid \theta) - H(Z \mid V, \theta) \le d\log(2) - \sum_{j=1}^{d} H(Z_j \mid V_j, \theta_j).$$

Now consider the distribution of Z_j given V_j and θ_j. By symmetry, the distribution has identical entropy for any value of V_j, so we may fix $V = v$ and assume $v_j =$ without loss of generality. Then, for $\theta_j \in (-1, 1)$, the jth component of the subgradient $\partial\ell(X; \theta)$ is $-X_j$, whence we see that

$$\begin{aligned} Q(Z_j = M_\infty \mid v_j = 1, \theta_j) &= Q(Z_j = M_\infty \mid X_j = 1, \theta_j)P(X_j = 1 \mid v_j = 1) \\ &\quad + Q(Z_j = M_\infty \mid X_j = -1, \theta_j)P(X_j = -1 \mid v_j = 1) \\ &= \left(\frac{M-1}{2M}\right)\left(\frac{1+\delta}{2}\right) + \left(\frac{M+1}{2M}\right)\left(\frac{1-\delta}{2}\right) \\ &= \frac{2M - 2\delta}{4M} = \frac{1}{2} - \frac{\delta}{2M}. \end{aligned}$$

Similarly, $Q(Z_j = -M_\infty \mid v_j = 1, \theta_j) = \frac{1}{2} + \frac{\delta}{2M}$. If $\theta_j \geq 1$, then we have that the subgradient $\partial|\theta_j - X_j| = 1$ with probability 1, and thus

$$Q(Z_j = M_\infty \mid v_j = 1, \theta_j) = \left(\frac{M+1}{2M}\right)\left(\frac{1+\delta}{2}\right) + \left(\frac{M+1}{2M}\right)\left(\frac{1-\delta}{2}\right) = \frac{1}{2},$$

which increases the entropy $H(Z_j \mid V_j, \theta_j)$ by Lemma B.1. Thus we see that the value θ_j minimizing the entropy $H(Z_j \mid V_j, \theta_j)$ is given by any $\theta_j \in (-1, 1)$, yielding Bernoulli marginal $(\frac{1}{2} + \delta/2M, \frac{1}{2} - \delta/2M)$ on $Z_j \mid V_j$. Summarizing, we have

$$I(Z; V \mid \theta) \leq d\log(2) + d\left[\left(\frac{1}{2} + \frac{\delta}{2M}\right)\log\left(\frac{1}{2} + \frac{\delta}{2M}\right) + \left(\frac{1}{2} - \frac{\delta}{2M}\right)\log\left(\frac{1}{2} - \frac{\delta}{2M}\right)\right].$$

As in the proof of Lemma 5.4, we use the concavity of log to see that

$$\begin{aligned} I(Z; V \mid \theta) &\leq d\log(2) + d\left[\left(\frac{1}{2} + \frac{\delta}{2M}\right)(-\log(2) + \delta/M) + \left(\frac{1}{2} - \frac{\delta}{2M}\right)(-\log(2) - \delta/M)\right] \\ &= d\left(\frac{1}{2} + \frac{\delta}{2M}\right)\left(\frac{\delta}{M}\right) + d\left(\frac{1}{2} - \frac{\delta}{2M}\right)\left(-\frac{\delta}{M}\right) = \frac{d\delta^2}{M^2}. \end{aligned}$$

Applying the bound (40) and replacing $M = M_\infty/L$ completes the proof.

B.3. Proof of Lemma 5.7

Letting Z denote a single subgradient sample using the conditional distribution Q specified by Proposition 4.4, we prove that

$$I(Z; V \mid \theta) \leq \delta^2 \Delta(\gamma)^2 \quad \text{for any } \theta \in \mathbb{R}^d, \tag{41}$$

which implies the lemma by the representation (40). Recall the SVM risk defined using the individual hinge losses (33): by construction, whenever $X = e_i$, then the loss is equal to $L\,[r - \theta_i]_+$. We have

$$\partial\ell(e_i, \theta) = L\begin{cases} 0 & \text{if } \theta_i > r \\ -e_i & \text{otherwise} \end{cases} \quad \text{and} \quad \partial\ell(-e_i, \theta) = L\begin{cases} 0 & \text{if } \theta_i < -r \\ e_i & \text{otherwise.} \end{cases}$$

For the remainder of this proof, we use the shorthand

$$D_\gamma := e^\gamma + e^{-\gamma} + 2(d-2)$$

for the denominator in many of our expressions. If $X = e_i$, then $\partial\ell(e_i, \theta) = Le_i$ or 0 as $\theta_i \leq r$ or $\theta_i > r$. Therefore, as we wish to communicate Le_i or 0, the construction in Proposition 4.4 implies

$$Q(Z = M_1 e_i \mid X = e_i, \theta) = \begin{cases} \frac{e^{-\gamma}}{D_\gamma} & \text{if } \theta_i \leq r \\ \frac{1}{2d} & \text{if } \theta_i > r, \end{cases} \tag{42}$$

and similarly we have for $j \neq i$ that

$$Q(Z = M_1 e_j \mid X = e_i, \theta) = \begin{cases} \frac{1}{D_\gamma} & \text{if } \theta_i \leq r \\ \frac{1}{2d} & \text{if } \theta_i > r. \end{cases} \tag{43}$$

For $X = -e_i$, we have the conditional distribution parallel to (42):

$$Q(Z = M_1 e_i \mid X = -e_i, \theta) = \begin{cases} \frac{e^{\gamma}}{D_\gamma} & \text{if } \theta_i \geq -r \\ \frac{1}{D_\gamma} & \text{if } \theta_i < -r. \end{cases}$$

For any given θ, we have that

$$I(Z; V \mid \theta) = H(Z \mid \theta) - H(Z \mid V, \theta) \leq \log(2d) - \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} H(Z \mid \theta, V = v) \tag{44}$$

since the choice of V is uniform and Z takes on at most $2d$ values. We thus use the conditional distributions (42) and (43) to compute the entropy $H(Z \mid \theta, V)$ (specifically, the minimal such entropy across all values of θ). To do this, we compute the marginal distribution $Q(z \mid v)$, arguing that $H(Z \mid \theta, V)$ is minimal for $\theta \in \text{int}[-r, r]^d$. When $\theta_j \in (-r, r)$ for all j, we have

$$\begin{aligned} Q(Z = M_1 e_i \mid V = v, \theta) = & \sum_{j=1}^{d} Q(Z = M_1 e_i \mid X = e_j, \theta) P(X = e_j \mid V = v) \\ & + \sum_{j=1}^{d} Q(Z = M_1 e_i \mid X = -e_j, \theta) P(X = -e_j \mid V = v). \end{aligned}$$

When $v_i = 1$, we thus have that

$$\begin{aligned} Q(Z = M_1 e_i \mid V = v, \theta) &= \frac{1+\delta}{2d} \frac{e^{-\gamma}}{D_\gamma} + \frac{1-\delta}{2d} \frac{e^{\gamma}}{D_\gamma} + \sum_{j \neq i} \frac{1}{D_\gamma} \left(\frac{1+\delta v_j}{2d} + \frac{1-\delta v_j}{2d} \right) \\ &= \frac{e^{\gamma} + e^{-\gamma} + \delta(e^{-\gamma} - e^{\gamma})}{2dD_\gamma} + \frac{d-1}{dD_\gamma} = \frac{1}{2d} + \frac{\delta(e^{-\gamma} - e^{\gamma})}{2dD_\gamma}, \end{aligned} \tag{45a}$$

and under the same condition,

$$Q(Z = -M_1 e_i \mid A = v, \theta) = \frac{e^{\gamma} + e^{-\gamma} + \delta(e^{\gamma} + e^{-\gamma})}{2dD_\gamma} + \frac{d-1}{dD_\gamma} = \frac{1}{2d} + \frac{\delta(e^{\gamma} - e^{-\gamma})}{2dD_\gamma}. \tag{45b}$$

If for any (possibly multiple) indices j we have $\theta_j \notin (-r, r)$, then via a bit of algebra and the conditional distributions (42) and (43), we see that there exists an $\epsilon \in (0, 1)$ such that

$$Q(Z = M_1 e_i \mid V = v, \theta) = \epsilon \frac{1}{2d} + (1 - \epsilon) \left(\frac{1}{2d} + \frac{\delta(e^{-\gamma} - e^{\gamma})}{2dD_\gamma} \right).$$

Lemma B.1 then implies that if $\theta \in \text{int}[-r, r]^d$ while $\theta' \notin \text{int}[-r, r]^d$, then

$$H(Z \mid \theta, V = v) < H(Z \mid \theta', V = v).$$

Since we seek an upper bound on the mutual information, we may thus assume without loss of generality that $\theta \in \text{int}[-r, r]^d$.

Now we compute the entropy $H(Z \mid \theta, v)$ using the marginal conditional distributions (45a) and (45b), which describe $Z \mid V$ when $\theta \in \text{int}[-r, r]^d$. Indeed, recall the definition in the statement of the lemma of the difference $\Delta(\gamma)$. For $z \in \{\pm M_1 e_j\}_{j=1}^d$, define the relation $z \sim v$ to mean that if $z = M_1 e_i$, then $v_i = 1$, and if $z = -M_1 e_i$, then $v_i = -1$. We then see that the entropy is

$$\begin{aligned} H(Z \mid \theta, V = v) \; &= \; -\sum_{z \sim v} Q(z \mid v, \theta) \log Q(z \mid v, \theta) - \sum_{z \not\sim v} Q(z \mid v, \theta) \log Q(z \mid v, \theta) \\ &= \; -d \left(\frac{1}{2d} + \frac{\delta \Delta(\gamma)}{2d} \right) \log \left(\frac{1}{2d} + \frac{\delta \Delta(\gamma)}{2d} \right) \\ & \quad\; -d \left(\frac{1}{2d} - \frac{\delta \Delta(\gamma)}{2d} \right) \log \left(\frac{1}{2d} - \frac{\delta \Delta(\gamma)}{2d} \right). \end{aligned}$$

As in the proofs of Lemmas 5.4 and 5.5, we use the concavity of $\log(\cdot)$ to see that

$$\begin{aligned}
-H(Z \mid \theta, V = v) &= \left(\frac{1}{2} + \frac{\delta\Delta(\gamma)}{2}\right)\log\left(\frac{1}{2d} + \frac{\delta\Delta(\gamma)}{2d}\right) + \left(\frac{1}{2} - \frac{\delta\Delta(\gamma)}{2}\right)\log\left(\frac{1}{2d} - \frac{\delta\Delta(\gamma)}{2d}\right) \\
&\le \left(\frac{1}{2} + \frac{\delta\Delta(\gamma)}{2}\right)(-\log(2d) + \delta\Delta(\gamma)) \\
&\quad + \left(\frac{1}{2} - \frac{\delta\Delta(\gamma)}{2}\right)(-\log(2d) - \delta\Delta(\gamma)) \\
&= -\log(2d) + \delta^2\Delta(\gamma)^2.
\end{aligned}$$

Invoking the earlier bound (44) and adding $\log(2d)$ to this expression completes the proof of the claim (41).

B.4. Proof of Lemma 5.8

Let Z_j denote the jth coordinate of Z. We first argue that conditional on V, the random variable Z has independent coordinates. Indeed, let $q^+ = q(z \mid x)$ for z such that $z^\top x > k$ and $q^- = e^{-\alpha}q^+$. Without loss of generality, we may take $V = e_1$, the first basis vector, and hence

$$\begin{aligned}
Q(Z = z \mid V = e_1) &= \sum_{x \in \{-1,1\}^d} Q(Z = z \mid X = x)P(X = x \mid V = e_1) \\
&= \frac{1}{2^{d-1}} \sum_{x \in \{-1,1\}^d} Q(Z = z \mid X = x) \cdot \frac{1 + x_1\delta}{2} \\
&= \frac{1}{2^{d-1}} \left[\sum_{x:\langle z,x\rangle > k} q^+ \frac{1 + x_1\delta}{2} + \sum_{x:\langle z,x\rangle \le k} q^- \frac{1 + x_1\delta}{2}\right]. \qquad (46)
\end{aligned}$$

Now, if $z_1 = 1$, then

$$\begin{aligned}
\sum_{x:\langle x,z\rangle > k} \frac{1 + x_1\delta}{2} &= \sum_{x:\langle x,z\rangle > k, x_1 = 1} \frac{1 + \delta}{2} + \sum_{x:\langle x,z\rangle > k, x_1 = -1} \frac{1 - \delta}{2} \\
&= \frac{1 + \delta}{2} C_{d-1}(k - 1) + \frac{1 - \delta}{2} C_{d-1}(k + 1),
\end{aligned}$$

and similarly

$$\sum_{x:\langle x,z\rangle \le k} \frac{1 + x_1\delta}{2} = \frac{1 + \delta}{2}(2^{d-1} - C_{d-1}(k - 1)) + \frac{1 - \delta}{2}(2^{d-1} - C_{d-1}(k + 1)).$$

On the other hand, we find that if $z_1 = -1$, then similar equalities hold, but with the counters $C_{d-1}(k - 1)$ and $C_{d-1}(k + 1)$ flipped:

$$\begin{aligned}
\sum_{x:\langle x,z\rangle > k} \frac{1 + x_1\delta}{2} &= \frac{1 + \delta}{2} C_{d-1}(k + 1) + \frac{1 - \delta}{2} C_{d-1}(k - 1) \\
\sum_{x:\langle x,z\rangle \le k} \frac{1 + x_1\delta}{2} &= \frac{1 + \delta}{2}(2^{d-1} - C_{d-1}(k + 1)) + \frac{1 - \delta}{2}(2^{d-1} - C_{d-1}(k - 1)).
\end{aligned}$$

In particular, we find that so long as the first coordinate $z_1 = z_1'$ of z remains constant, then $Q(Z = z \mid V = e_1) = Q(Z = z' \mid V = e_1)$, and that we thus have $Z_2, \ldots, Z_d$ are distributed uniformly at random in $\{-1, 1\}^d$.

We now determine q^+ and compute the marginal value $Q(Z_1 = 1 \mid V = e_1)$. For the first, we note that

$$C_d(k)q^+ + (2^d - C_d(k))q^- = 1, \quad \text{or} \quad C_d(k)q^+ + e^{-\alpha}(2^d - C_d(k))q^+ = 1,$$

which yields the expressions

$$q^+ = \frac{e^\alpha}{(e^\alpha - 1)C_d(k) + 2^d} \quad \text{and} \quad q^- = \frac{1}{(e^\alpha - 1)C_d(k) + 2^d}.$$

By the expression (46) and the calculations that follow, we thus find that, when $z_1 = 1$, we have

$$\begin{aligned} q(z \mid e_1) &= \frac{1}{2^{d-1}} \cdot \Bigg[q^+ \left(\frac{1+\delta}{2} C_{d-1}(k-1) + \frac{1-\delta}{2} C_{d-1}(k+1) \right) \\ &\quad + q^- \left(\frac{1+\delta}{2}(2^{d-1} - C_{d-1}(k-1)) + \frac{1-\delta}{2}(2^{d-1} - C_{d-1}(k+1)) \right) \Bigg] \\ &= \frac{1}{2^{d-1}} \cdot \Bigg[2^{d-1} q^- + \frac{1}{2}(q^+ - q^-)(C_{d-1}(k-1) + C_{d-1}(k+1)) \\ &\quad + \frac{\delta}{2}(q^+ - q^-)(C_{d-1}(k-1) - C_{d-1}(k+1)) \Bigg], \end{aligned} \tag{47a}$$

and similarly, when $z_1 = -1$, we have

$$\begin{aligned} q(z \mid e_1) = \frac{1}{2^{d-1}} \cdot \Bigg[& 2^{d-1} q^- + \frac{1}{2}(q^+ - q^-)(C_{d-1}(k-1) + C_{d-1}(k+1)) \\ & - \frac{\delta}{2}(q^+ - q^-)(C_{d-1}(k-1) - C_{d-1}(k+1)) \Bigg]. \end{aligned} \tag{47b}$$

Now note that

$$C_{d-1}(k-1) - C_{d-1}(k+1) = \sum_{i=0}^{\lceil (d-k)/2 \rceil - 1} \binom{d-1}{i} - \sum_{i=0}^{\lceil (d-k)/2 \rceil - 2} \binom{d-1}{i} = \binom{d-1}{\lceil (d-k)/2 \rceil - 1}$$

and that the difference

$$q^+ - q^- = \frac{e^\alpha - 1}{(e^\alpha - 1)C_d(k) + 2^d}.$$

Recalling the definition of the constant Δ, we thus find from the expansions (47a) and (47b)—since they must sum to 1—that

$$Q(Z = z \mid V = e_1) = \frac{1}{2^{d-1}} \cdot \begin{cases} \frac{1}{2} + \frac{\Delta(\delta,\alpha,d,k)}{2} & \text{if } z_1 = 1 \\ \frac{1}{2} - \frac{\Delta(\delta,\alpha,d,k)}{2} & \text{if } z_1 = -1. \end{cases} \tag{48}$$

It is clear that similar statements hold in the other symmetric cases (i.e., if $V = -e_2$, then the probabilities depend on $z_2 = -1$ or 1).

It remains to use the marginalized representation (48) to compute the bound on the mutual information in the statement of the lemma. Given $V = v$, $Z \in \{-M, M\}^d$ is uniform except on the coordinate j for which $v_j \neq 0$, by symmetry. (Marginally, Z is uniform on $\{-M, M\}^d$.) By a direct calculation, we have $H(Z) = d \log 2$ and

$$H(Z \mid V = e_1) = \sum_{j=1}^{d} H(Z_j \mid Z_{1:j-1}, V = e_1) = H(Z_1 \mid V = e_1) + (d-1)\log 2,$$

and similarly for the other possible values of V. Therefore, using the probabilities (48), we have the mutual information bound

$$\begin{aligned} I(Z;V) &= H(Z) - H(Z \mid V) \le d\log 2 - \frac{1}{2d}\sum_v H(Z \mid V = v) \\ &= d\log 2 - (d-1)\log 2 \\ &\quad + \left(\frac{1}{2} + \frac{\Delta(\delta,\alpha,d,k)}{2}\right)\log\left(\frac{1}{2} + \frac{\Delta(\delta,\alpha,d,k)}{2}\right) \\ &\quad + \left(\frac{1}{2} - \frac{\Delta(\delta,\alpha,d,k)}{2}\right)\log\left(\frac{1}{2} - \frac{\Delta(\delta,\alpha,d,k)}{2}\right) \\ &\le \log 2 + \left(\frac{1}{2} + \frac{\Delta(\delta,\alpha,d,k)}{2}\right)\left[\log\frac{1}{2} + \Delta(\delta,\alpha,d,k)\right] \\ &\quad + \left(\frac{1}{2} - \frac{\Delta(\delta,\alpha,d,k)}{2}\right)\left[\log\frac{1}{2} - \Delta(\delta,\alpha,d,k)\right] \\ &= \Delta(\delta,\alpha,d,k)^2, \end{aligned}$$

where the inequality follows from the concavity of $p \mapsto \log(p)$.

B.5. Bounds on Total Variation Norm

LEMMA B.2. *Let Q_1 and Q_{-1} be distributions on $\{-1, 1\}$, where*

$$Q_1(Z=z) = \frac{1}{2} + \frac{1}{2}\cdot\begin{cases} \delta & \text{if } z = 1 \\ -\delta & \text{otherwise} \end{cases} \quad \text{and} \quad Q_{-1}(Z=z) = \frac{1}{2} + \frac{1}{2}\cdot\begin{cases} -\delta & \text{if } z = 1 \\ \delta & \text{otherwise.} \end{cases}$$

Let Q_i^n denote the n-fold product distribution of Q_i. Then for $\delta \in [0, 1/3]$,

$$\left\| Q_1^n - Q_{-1}^n \right\|_{\rm TV} \le \delta\sqrt{(3/2)n}.$$

PROOF. For any two probability distributions P, Q, Pinsker's inequality [Cover and Thomas 2006] asserts that the total variation norm is bounded as $\|P - Q\|_{\rm TV} \le \sqrt{D_{\rm kl}(P\|Q)/2}$. Applying this inequality in our setting, we find that

$$\left\| Q_1^n - Q_{-1}^n \right\|_{\rm TV} \le \sqrt{\frac{1}{2}D_{\rm kl}\left(Q_1^n \| Q_{-1}^n\right)} = \frac{1}{\sqrt{2}}\sqrt{nD_{\rm kl}(Q_1\|Q_{-1})},$$

where we have exploited the product nature of Q_i^n. Now we note that by the concavity of the log, we have (via the first-order inequality) that $\log\frac{1+\delta}{1-\delta} \le 2\delta/(1-\delta)$, so

$$\frac{1+\delta}{2}\log\frac{\frac{1+\delta}{2}}{\frac{1-\delta}{2}} + \frac{1-\delta}{2}\log\frac{\frac{1-\delta}{2}}{\frac{1+\delta}{2}} = \frac{1+\delta}{2}\log\frac{1+\delta}{1-\delta} + \frac{1-\delta}{2}\log\frac{1-\delta}{1+\delta} = \delta\log\frac{1+\delta}{1-\delta} \le \frac{2\delta^2}{1-\delta}.$$

Assuming that $\delta \le 1/3$, the final term is upper bounded by $3\delta^2$. But, of course, by definition of Q_1 and Q_{-1}, we have

$$D_{\rm kl}(Q_1\|Q_{-1}) = \frac{1+\delta}{2}\log\frac{\frac{1+\delta}{2}}{\frac{1-\delta}{2}} + \frac{1-\delta}{2}\log\frac{\frac{1-\delta}{2}}{\frac{1+\delta}{2}} \le 3\delta^2,$$

which completes the proof. □

C. ACHIEVABILITY BY STOCHASTIC MIRROR DESCENT

In this appendix, we provide further details on the algorithm used to achieve the upper bounds in Theorems 3.10 and 3.11. Both of our achievability results rely on stochastic gradient mechanisms, and their most important ingredient is a conditional distribution Q that satisfies α-local differential privacy. In particular, if $g \in \mathbb{R}^d$ is a (sub)gradient of the loss $\ell(x, \theta)$, we construct $Z \in \mathbb{R}^d$ by perturbing g in such a way that $\mathbb{E}[Z \mid g] = g$. Thus, each of the achievability guarantees consists of describing an α-differentially private sampling distribution, then bounding the expected norm of Z and applying one of the convergence guarantees (14).

C.1. Achievability in Theorem 3.10

The sampling strategy we use is essentially identical to that used in Corollary 3.9 (and the optimal α-private scheme of Proposition 4.6; see also Strategy B in our paper [Duchi et al. 2013]). Let $\pi_\alpha := e^\alpha/(e^\alpha + 1)$ and T be a Bernoulli(π_α)-random variable, and let $B \geq L$ be a fixed constant (to be specified). Then, given a vector $g \in \mathbb{R}^d$ with $\|g\|_\infty \leq L$, construct $\widetilde{g} \in \mathbb{R}^d$ with coordinates $\widetilde{g}_j$ sampled independently from $\{-L, L\}$ with probabilities $1/2 - g_j/(2L)$ and $1/2 + g_j/(2L)$. Then, sample T and set

$$Z \sim \begin{cases} \text{Uniform}(z \in \{-B, B\}^d : \langle z, \widetilde{g}\rangle > 0) & \text{if } T = 1 \\ \text{Uniform}(z \in \{-B, B\}^d : \langle z, \widetilde{g}\rangle \leq 0) & \text{if } T = 0. \end{cases} \tag{49}$$

By inspection, the scheme (49) is α-differentially private. Moreover, we have by the calculations in the proof of Corollary 3.9 (see Section 5.9) that by the sampling strategy (49)

$$\mathbb{E}[Z \mid g] = \mathbb{E}[\mathbb{E}[Z \mid \widetilde{g}] \mid g] = g \frac{B}{2^{d-1}L} \frac{e^\alpha - 1}{e^\alpha + 1} \cdot \begin{cases} \binom{d-1}{\frac{d-1}{2}} & d \text{ odd} \\ \binom{d-1}{d/2} & d \text{ even.} \end{cases}$$

Thus, without loss of generality assuming d is odd, choosing

$$B = 2^{d-1} L \frac{e^\alpha + 1}{e^\alpha - 1} \binom{d-1}{\frac{d-1}{2}}^{-1} \quad \text{implies} \quad \mathbb{E}[Z \mid g] = g \quad \text{and} \quad \|Z\|_\infty = B = \mathcal{O}(1)(L/\alpha)\sqrt{d}.$$

Applying the mirror descent method to the gradients provided from the sampling strategy (49), we obtain the bound (14a) with $M_\infty = B = \mathcal{O}(1)(L/\alpha)\sqrt{d}$, which is our desired result.

C.2. Achievability in Theorem 3.11

The achievability result for Theorem 3.11 is similar to that for Theorem 3.10, but we use a modified sampling distribution. Now, using the same notation as that for the strategy (49), we use the following. Given a vector g with $\|g\|_2 \leq L$, set $\widetilde{g} = Lg/\|g\|_2$ with probability $\frac{1}{2} + \|g\|_2/2L$ and $\widetilde{g} = -Lg/\|g\|_2$ with probability $\frac{1}{2} - \|g\|_2/2L$. Then, sample $T \sim$ Bernoulli(π_α) and set

$$Z \sim \begin{cases} \text{Uniform}(z \in \mathbb{R}^d : \langle z, \widetilde{g}\rangle > 0, \|z\|_2 = B) & \text{if } T = 1 \\ \text{Uniform}(z \in \mathbb{R}^d : \langle z, \widetilde{g}\rangle \leq 0, \|z\|_2 = B) & \text{if } T = 0. \end{cases} \tag{50}$$

(This is Strategy A in our paper Duchi et al. [2013].) Then we have [Duchi et al. 2013, Appendix D.1] that

$$\mathbb{E}[Z \mid g] = \frac{B}{L} \frac{e^\alpha - 1}{e^\alpha + 1} \frac{\Gamma(\frac{d}{2} + 1)}{\sqrt{\pi} d \Gamma(\frac{d-1}{2} + 1)},$$

so choosing

$$B = L \frac{e^\alpha + 1}{e^\alpha - 1} \frac{\sqrt{\pi} d \Gamma(\frac{d-1}{2} + 1)}{\Gamma(\frac{d}{2} + 1)} \leq L \frac{e^\alpha + 1}{e^\alpha - 1} \frac{3\sqrt{\pi}\sqrt{d}}{2}$$

implies that $\mathbb{E}[Z \mid g] = g$ and $\|Z\|_2 \le B = \mathcal{O}(1)(L/\alpha)\sqrt{d}$. Applying the stochastic gradient descent method to the gradients provided by the sampling scheme (50), we obtain the bound (14b) with $M_2 = B$, which implies that if $\Theta \subset \mathbb{B}_2(r_2)$, then

$$\mathbb{E}[R(\widehat{\theta}_n)] - R(\theta^*) = \mathcal{O}\left(\frac{\sqrt{d}}{\alpha}\frac{Lr_2}{\sqrt{n}}\right).$$

Noting that $\mathbb{B}_q(r_q) \subset d^{\frac{1}{2}-\frac{1}{q}}\mathbb{B}_2(r_q)$ completes the proof of the achievability result.

D. BACKGROUND ON CONDITIONAL PROBABILITIES

In this appendix, we present some basic lemmas on conditional independence and regular conditional probabilities that will be useful in Appendix E.

We first recall the following classical data-processing inequality, which holds for essentially arbitrary random variables [Gray 1990, Chapter 5].

LEMMA D.1 (DATA PROCESSING). *Let $X \to Z \to Y$ be a Markov chain. Then, $I(X;Y) \le I(X;Z)$, with equality if and only if X is conditionally independent of Y given Z.*

This inequality, in conjunction with with Carathéodory and Minkowski's finite-dimensional version of the Krein-Milman theorem (e.g. Hiriart-Urruty and Lemaréchal [1996]), allows us to argue any Q minimizing $I(P, Q)$ must be supported on the extreme points of D. To make this point precise, however, we need to address certain measurability issues involved in the choice of the extreme points.

We begin with a precise definition of a regular conditional probability.

Definition D.2. Let $(\Omega, \mathcal{F})$ and $(T, \sigma(T))$ be measurable spaces. A *regular conditional probability*, also known as a Markov kernel or transition probability, is a function $\nu : T \times \mathcal{F} \to [0, 1]$ such that

$$t \mapsto \nu(t, A) \text{ is measurable for all } A \in \mathcal{F}$$
$$\nu(t, \cdot) : \mathcal{F} \to [0, 1] \text{ is a probability measure for all } t \in T.$$

Any Markov chain has a transition probability; conversely, any set of consistent transition probabilities define a Markov chain (see, e.g., Chapter 5 of Kallenberg [1997]).

Some difficulties with measurability arise in constructing the appropriate Markov chain for our setting. To deal with them, we use results from Choquet theory, which extend Krein-Milman theorems to integral representations [Phelps 2001]. We begin our proof by stating a measurable selection theorem [Phelps 2001, Theorem 11.4], though we restrict the theorem's statement to subsets of finite dimensional space.

PROPOSITION D.3. *Let $D \subset \mathbb{R}^d$ be a compact convex set. For each x, there exists a probability measure μ_x supported on $\operatorname{Ext}(D)$ such that $\int_D y d\mu_x(y) = x$. Moreover, the mapping $x \mapsto \mu_x$ can be taken to be measurable.*

In the statement of this result, measurability is taken with respect to the σ-field generated by the topology of weak convergence. As a consequence of the proposition, however, it is clear that since for any continuous function f the mapping $x \mapsto \int f d\mu_x$ is measurable, we have that for relatively open sets $A \subset C$ the mapping $x \mapsto \mu_x(A)$ is measurable, whence for any measurable set $A \subset C$ the mapping $x \mapsto \mu_x(A)$ is measurable. That is, we can define the Markov kernel $\nu : \mathbb{R}^d \times \sigma(C) \to [0, 1]$ according to the mapping specified by Proposition D.3 (we take $\nu(x, \cdot) = \mu_x$) with the additional properties that

$$\int_D y\nu(x, dy) = x \quad \text{and} \quad \nu(x, D \setminus \operatorname{Ext}(D)) = 0 \ \text{ for all } x \in D.$$

In finite dimensions, a trivial extension of Proposition D.3 allows us to drop the assumption that D is convex. Indeed, we have that since D is compact, then $\text{Ext}(D) = \text{Ext}(\text{Conv}(D))$ [Hiriart-Urruty and Lemaréchal 1996, Chapter III.2].

Given this measure-theoretic background, we turn to a key lemma that we will need in Appendix E. In this lemma, we assume as usual that $C \subset D \subset \mathbb{R}^d$ are compact sets, and that $Q \in \mathcal{Q}(C, D)$ (recall Definition (10b)).

LEMMA D.4. *Let P be a distribution supported on C. If there exists a set $A \subset C$ with $P(A) > 0$ and a set $B \subset D \setminus \text{Ext}(D)$ with $Q(B \mid X = x) > 0$ for $x \in A$, there exists a regular conditional probability distribution $Q' \in \mathcal{Q}(C, D)$ where $Q'(\cdot \mid x)$ has support contained in* $\text{Ext}(D)$ *and*

$$I(P, Q) > I(P, Q').$$

Paraphrasing the lemma slightly, we have that any conditional distribution Q minimizing $I(P, Q)$ must (outside of a set of measure zero) be completely supported on the extreme points $\text{Ext}(D)$.

PROOF. For any $y \in D$, Proposition D.3 guarantees that we can represent y as the (regular conditional) measure $\nu(y, \cdot)$. Thus, we can define a random variable Z_y distributed according to $\nu(y, \cdot)$, whose existence we are guaranteed by standard constructions [Billingsley 1986; Kallenberg 1997] with regular conditional probability. Then, $\mathbb{E}[Z_y] = \int_D z\nu(y, dz) = y$, and moreover, we can define the measurable version of the conditional expectation $\mathbb{E}[Z_Y \mid Y]$ via

$$\mathbb{E}[Z_Y \mid Y] = \int_D z\nu(Y, dz) = Y,$$

so we have the (almost sure) chain of equalities

$$\begin{aligned}\mathbb{E}[Z_Y \mid X = x] &= \mathbb{E}[\mathbb{E}[Z_Y \mid Y] \mid X = x] = \int_D \mathbb{E}[Z_Y \mid Y = y] dQ(y \mid X = x) \\ &= \int_D \int_D z\nu(y, dz) dQ(y \mid X = x) = \int_D y dQ(y \mid X = x) = x.\end{aligned}$$

By construction, $X \to Y \to Z$ is a valid Markov chain, and since the sets A and B satisfy $P(A) > 0$ and $\int_A Q(B \mid X = x) dP(x) > 0$, we see that $I(X; Y) > I(X; Z)$ by Lemma D.1. □

We turn to an analogue of Lemma D.4 in the differentially private setting.

LEMMA D.5. *Let the conditions of Lemma D.4 hold, and let P be a distribution supported on C. If there exists a set $A \subset C$ with $P(A) > 0$ and a set $B \subset D \setminus \text{Ext}(D)$ with $Q(B \mid X = x) > 0$ for $x \in A$, there exists a regular conditional probability distribution $Q' \in \mathcal{Q}(C, D)$ where $Q'(\cdot \mid x)$ has support contained in* $\text{Ext}(D)$, *satisfies*

$$I(P, Q) > I(P, Q'),$$

and has no worse differential privacy than Q:

$$\sup_{S \in \sigma(D)} \sup_{x, x' \in C} \frac{Q'(S \mid X = x)}{Q'(S \mid X = x')} \le \sup_{S \in \sigma(D)} \sup_{x, x' \in C} \frac{Q(S \mid X = x)}{Q(S \mid X = x')}.$$

PROOF. Let $\nu : \mathbb{R}^d \times \sigma(C) \to [0, 1]$ be the Markov kernel defined in the proof of Lemma D.4, and, without loss of generality, assume that $Q(\cdot \mid X = x)$ and $Q(\cdot \mid X = x')$ have density q with respect to an underlying measure $\mu_{x,x'}$. Define the distribution

$$Q'(S \mid X = x) := \int_D \int_D \nu(y, dz) q(y \mid x) d\mu_{x,x'}(y).$$

By assumption, if Q is α-differentially private, then for μ-almost all $y \in D$, we have $q(y \mid x) \le e^{\alpha} q(y \mid x')$. We find that

$$\begin{aligned} Q'(S \mid X = x) &= \int_D \int_D \nu(y, dz) q(y \mid x) d\mu_{x,x'}(y) \\ &\le \int_D \int_D \nu(y, dz) e^{\alpha} q(y \mid x') d\mu_{x,x'}(y) = e^{\alpha} Q'(S \mid X = x'), \end{aligned}$$

so Q' is at least as differentially private as Q. □

Finally, we will need the following standard maximum entropy result. Let z denote a discrete random variable and let $q(z \mid x)$ denote the conditional probability mass function of $Z \mid X = x$. Consider the finite dimensional entropy maximization problem

$$\begin{aligned} &\underset{q}{\text{minimize}} \;\; \sum_z q(z \mid x) \log q(z \mid x) \\ &\text{subject to} \;\; \sum_z z q(z \mid x) = x, \quad \sum_z q(z \mid x) = 1, \quad q(z \mid x) \ge 0 \text{ for all } z. \end{aligned} \tag{51}$$

We have the following lemma, which establishes the form of the solution to the problem (51). We include a proof for completeness.

LEMMA D.6. *The p.m.f. $q(\cdot \mid x)$ solving problem* (51) *is given by*

$$q(z \mid x) = \frac{\exp(-\mu^{\top} z)}{\sum_{z'} \exp(-\mu^{\top} z')}, \tag{52}$$

where $\mu \in \mathbb{R}^d$ is any vector chosen to satisfy the constraint $\sum_z z q(z \mid x) = x$. Such a $\mu \in \mathbb{R}^d$ exists.

PROOF. We may write the Lagrangian with dual variables $\mu \in \mathbb{R}^d$, $\lambda(z) \ge 0$, and $\theta \in \mathbb{R}$,

$$\begin{aligned} &\mathcal{L}(q, \mu, \lambda, \theta) \\ &= \sum_z q(z \mid x) \log q(z \mid x) + \mu^{\top} \left(\sum_z z q(z \mid x) - x \right) + \theta \left(\sum_z q(z \mid x) - 1 \right) - \sum_z \lambda(z) q(z \mid x). \end{aligned}$$

Since the problem (51) has convex cost, linear constraints, and nonempty domain, strong duality obtains [Boyd and Vandenberghe 2004, Chapter 5], and the KKT conditions hold for the problem. Thus, minimizing q out of $\mathcal{L}$ to find the dual, we take derivatives with respect to the m variables $q(z \mid x)$ for $z = (1 + \alpha) u_i$ and find the optimal conditional p.m.f. q must satisfy

$$\log q(z \mid x) + 1 + \mu^{\top} z + \theta - \lambda(z) = 0, \quad \text{or} \quad q(z \mid x) = \exp(\lambda(z) - 1 - \theta) \exp(-\mu^{\top} z).$$

In particular, we see that since $q(z \mid x) > 0$, we must have $\lambda(z) = 0$ by complementarity, and (satisfying the summability constraint $\sum_z q(z \mid x) = 1$) we see that

$$q(z \mid x) = \frac{\exp(-\mu^{\top} z)}{\sum_{z'} \exp(-\mu^{\top} z')},$$

where $\mu \in \mathbb{R}^d$ is any vector chosen to satisfy the constraint $\sum_z z q(z \mid x) = x$. The existence of such a μ is guaranteed by the attainment of the KKT conditions. □

E. PROOFS OF MINIMAX MUTUAL INFORMATION CHARACTERIZATIONS

In this section, we provide the proofs of the results stated in Section 4, all of which follow a broadly similar outline. We make use of Lemma D.4 to guarantee that any conditional distribution Q minimizing the mutual information $I(P, Q)$ must be supported on the extreme points of the set D. This allows us to reduce computing maximal entropies and minimal mutual information values to finite dimensional convex programs, whose optimality we can check using results from convex analysis and optimization.

E.1. Proof of Theorem 4.2

We begin by considering $\sup_P$, where Q^* is defined as in the statement of the theorem. Since the support of Q^* is finite (there are m extreme points of D), we have

$$I(P, Q^*) = I(X; Z) = H(Z) - H(Z \mid X) \leq \log(m) - H(Z \mid X)$$
$$= \log(m) - \int H(Z \mid X = x) dP(x).$$

Now, for any distribution P on the set C and for any $x \in \operatorname{supp} P$, we can write x as $x = \sum_i \beta_i(x) u_i$, where u_i are the extreme points of C, and where $\beta_i(x) \geq 0$ and $\sum_i \beta_i(x) = 1$ (using the Krein-Milman theorem). Define the individual probability mass functions q^i to be the maximum entropy p.m.f. (52) for each of the extreme points u_i. Then, we can define the conditional probability mass function by

$$q(\cdot \mid x) = \sum_i \beta_i(x) q^i(\cdot).$$

(Without loss of generality, we may assume the β_i are continuous, since the set of extreme points is finite, and thus $q(\cdot \mid x)$ can be viewed as a regular conditional probability. We can make this formal using the techniques in the proof of Lemma D.4.) Denoting $H(q(\cdot \mid x)) := H(Z \mid X = x)$, we can use the convexity of the negative entropy to see that

$$I(P, Q^*) \leq \log(m) - \int \sum_i \beta_i(x) H(q^i(\cdot)) dP(x). \tag{53}$$

By symmetry, the entropy $H(q^i(\cdot)) = H(Q^*(\cdot \mid X = u_i))$ is a constant determined by the maximum entropy distribution (52), and thus

$$I(P, Q^*) \leq \log(m) - H(Q^*(\cdot \mid X = u_i)). \tag{54}$$

Equality in the upper bound (54) is attained by taking P^* to be the uniform distribution on the extreme points $\{u_i\}$ of C.

It remains to establish an identical lower bound for $I(P^*, Q)$ over all conditional distributions Q satisfying the constraints of the theorem statement. We know from Lemma D.4 that Q must be supported on $(1 + \kappa)u_i$ for $i = 1, \ldots, m$. Denoting by $q(z \mid x)$ the p.m.f. of Q conditional on x (for x in the finite set of extreme points of C that make up the support $\operatorname{supp} P^*$), we can write minimizing the mutual information as the parametric convex optimization problem

$$\underset{q}{\text{minimize}} \;\; \sum_z \left(\sum_x q(z \mid x) p(x) \right) \log \left(\sum_x q(z \mid x) p(x) \right) - \sum_x p(x) \sum_z q(z \mid x) \log q(z \mid x) \tag{55}$$

$$\text{subject to} \;\; \sum_z q(z \mid x) = 1 \text{ for all } x, \quad \sum_z z q(z \mid x) = x \text{ for all } x, \quad q(z \mid x) \geq 0 \text{ for all } x, z.$$

In the problem (55), the sums over x and z are over the extreme points of C and D, respectively, and p is the uniform distribution with $p(x) = 1/m$. Mutual information is convex in the conditional distribution q; moreover, it is strictly convex except when $q(z \mid x) = \sum_{x'} q(z \mid x')p(x')$ for all x, z. (This can be seen by an inspection of the proof of Theorem 2.7.4 by Cover and Thomas [2006].) In our case, since Q^* does not satisfy this equality, the uniqueness of Q^* as the minimizer of $I(P^*, Q^*)$ will follow if we show that Q^* is a minimizer at all.

We proceed to solve the problem (55). Writing $I(p, q)$ as a shorthand for the mutual information, we introduce Lagrange multipliers $\theta(x) \in \mathbb{R}$ for the normalization constraints, $\mu(x) \in \mathbb{R}^d$ for the conditional expectation constraints, and $\lambda(x, z) \geq 0$ for the nonnegativity constraints. This yields the Lagrangian

$$\mathcal{L}(q, \mu, \lambda, \theta)$$
$$= I(p, q) - \sum_{x,z} \lambda(x, z) q(z \mid x) + \sum_x \mu(x)^\top \left(\sum_z z q(z \mid x) - x \right) + \sum_x \theta(x) \left(\sum_z q(z \mid x) - 1 \right).$$

If we can satisfy the Karush-Kuhn-Tucker (KKT) conditions (see, e.g., Boyd and Vandenberghe [2004]) for optimality of the problem (55), we will be done. Taking derivatives with respect to $q(z \mid x)$, we see

$$\begin{aligned} \frac{\partial}{\partial q(z \mid x)} \mathcal{L}(q, \mu, \lambda, \theta) &= p(x)[\log(q(z \mid x)) + 1] - p(x) \log \left(\sum_{x'} q(z \mid x')p(x') \right) \\ &\quad - q(z) \cdot \frac{1}{q(z)} p(x) - \lambda(z, x) + \theta(x) + \mu(x)^\top z \\ &= p(x) \log q(z \mid x) - p(x) \log \left(\sum_{x'} q(z \mid x')p(x') \right) \\ &\quad - \lambda(z, x) + \theta(x) + \mu(x)^\top z, \end{aligned}$$

where we set $q(z) = \sum_{x'} q(z \mid x')p(x')$ for shorthand. Now, we use symmetry to note that since we have chosen q to be the maximum entropy distribution (52) for each x in the extreme points $\{u_i\}$ of C, the marginal $q(z) = \sum_{x'} q(z \mid x')p(x') = 1/m$ is uniform by the symmetry of the set D and since p is uniform. In addition, since $q(z \mid x) > 0$ strictly, we have $\lambda(z, x) = 0$ by complementarity. Thus, at q chosen to be the maximum entropy distribution, we can rewrite the derivative of the Lagrangian

$$\frac{\partial}{\partial q(z \mid x)} \mathcal{L}(q, \mu, \lambda, \theta) = \frac{1}{m} \log q(z \mid x) - \frac{1}{m} \log \frac{1}{m} + \theta(x) + \mu(x)^\top z.$$

Recalling the definition (52) of $q(z \mid x)$, and denoting the maximum entropy parameters μ there by $\mu^*(x)$, we have

$$\begin{aligned} \frac{\partial}{\partial q(z \mid x)} \mathcal{L}(q, \mu, \lambda, \theta) &= -\frac{1}{m} \mu^*(x)^\top z + \frac{1}{m} \log \left(\sum_{z'} \exp(-\mu^*(x)^\top z') \right) \\ &\quad - \frac{1}{m} \log \frac{1}{m} + \theta(x) + \mu(x)^\top z. \end{aligned}$$

Now, by inspection, we may set

$$\theta(x) = \frac{1}{m} \log \frac{1}{m} - \frac{1}{m} \log \left(\sum_{z'} \exp(-\mu^*(x)^\top z') \right) \quad \text{and} \quad \mu(x) = \frac{1}{m} \mu^*(x),$$

and we satisfy the KKT conditions for the mutual information minimization problem (55).

Summarizing, the conditional distribution Q^* specified in the statement of the theorem as the maximum entropy distribution (52) satisfies

$$\inf_Q I(P^*, Q) \geq I(P^*, Q^*),$$

which, when combined with the first part of the proof, gives the saddle point inequality

$$\sup_P I(P, Q^*) \leq \log(m) - H(q(\cdot \mid X = u_i)) = I(P^*, Q^*) \leq \inf_Q I(P^*, Q),$$

as claimed.

Remarks. In the proof of the theorem, we have defined $Q^*(\cdot \mid x)$ as a conditional distribution only for $x \in \text{Ext}(C)$, the extreme points of C. This can easily be remedied: take $Q^*(\cdot \mid x)$ to be the distribution maximizing the entropy $H(Z \mid X = x)$ for each $x \in C$ under the constraint that the support of Z be contained in $\text{Ext}(D)$. This is equivalent to—for each $x \in C$—choosing $Z = z_i$ for $z_i \in \text{Ext}(D)$, $i = 1, \ldots, m$, with probability q_i, where $q \in \mathbb{R}^m$ solves the entropy maximization problem

$$\underset{q \in \mathbb{R}^m}{\text{maximize}} \quad -\sum_i q_i \log q_i \quad \text{subject to} \sum_i z_i q_i = x, \;\; \sum_i q_i = 1, \;\; q_i \geq 0.$$

Inspecting the proof of Theorem 4.2 (see the bound (53)) shows that this choice can only decrease the mutual information $I(X; Z)$. Additionally, the strong convexity of the entropy over the simplex guarantees that the solutions to this optimization problem are continuous in x (see Chapter X of Hiriart-Urruty and Lemaréchal [1996]) so this distribution $q(\cdot \mid x)$ defines a measurable random variable as desired.

E.2. Proof of Proposition 4.3

By scaling, we may assume without loss of generality that $L = 1$ and $M \geq 1$. Using Theorem 4.2 (and the remarks immediately following its proof), we can focus on maximizing the entropy of the random variable Z conditional on $X = x$ for each fixed $x \in [-1, 1]^d$. Let Z_i denote the ith coordinate of the random vector Z; we take the conditional distribution of Z_i to be independent of Z_j and let Z be distributed as

$$Z_i \mid X = \begin{cases} M & \text{w.p. } \frac{1}{2} + \frac{X_i}{2M} \\ -M & \text{w.p. } \frac{1}{2} - \frac{X_i}{2M}. \end{cases} \tag{56}$$

Let us now verify that the distribution (56) maximizes the entropy $H(Z \mid X = x)$. Indeed, we may fix x (leaving it implicit in the vector $[q(z)]_z := [q(z \mid x)]_z$), and we solve the entropy maximization problem

$$\underset{q}{\text{minimize}} \quad -H(q) \quad \text{subject to} \;\; \sum_z q(z) = 1, \;\; q(z) \geq 0, \;\; \sum_z z q(z) = x, \tag{57}$$

where all sums are taken over $z \in \text{Ext}([-M, M]^d) = \{-M, M\}^d$. Introducing the Lagrange multipliers $\mu \in \mathbb{R}^d$, $\lambda(z) \geq 0$, and $\theta \in \mathbb{R}$, we find that problem (57) has the Lagrangian

$$\mathcal{L}(q, \mu, \lambda, \theta) = -H(q) - \sum_z \lambda(z) q(z) + \mu^\top \left(\sum_z z q(z) - x \right) + \theta \left(\sum_z q(z) - 1 \right).$$

To find the infimum of the Lagrangian with respect to q, we take derivatives (since we make the identification $q \in \mathbb{R}^{2^d}$). We see that

$$\frac{\partial}{\partial q(z)}\mathcal{L}(q, \mu, \lambda, \theta) = \log(q(z)) + 1 - \lambda(z) + \theta + \mu^\top z.$$

With the definition (56) of the probability mass function q (that z_i are independent Bernoulli random variables with parameters $\frac{1}{2} + x_i/2M$), the coordinate conditional distributions are

$$q(z_i \mid x_i) = \left(\frac{1}{2} + \frac{1}{2M}\right)^{\frac{1}{2}+\frac{x_i z_i}{2M}} \left(\frac{1}{2} - \frac{1}{2M}\right)^{\frac{1}{2}-\frac{x_i z_i}{2M}}.$$

Theorem 4.2 says that, without loss of generality, we may assume that $x \in \{-1, 1\}^d$, the full probability mass function q can be written

$$q(z) = \left(\frac{1}{2} + \frac{1}{2M}\right)^{\frac{d}{2}+\frac{x^\top z}{2M}} \left(\frac{1}{2} - \frac{1}{2M}\right)^{\frac{d}{2}-\frac{x^\top z}{2M}}. \tag{58}$$

Plugging the conditional (58) results in

$$\begin{aligned}\frac{\partial}{\partial q(z)}\mathcal{L}(q, \mu, \lambda, \theta) &= \left(\frac{d}{2} + \frac{x^\top z}{2M}\right)\log\left(\frac{1}{2} + \frac{1}{2M}\right) \\ &\quad + \left(\frac{d}{2} - \frac{x^\top z}{2M}\right)\log\left(\frac{1}{2} - \frac{1}{2M}\right) + 1 - \lambda(z) + \theta + \mu^\top z \\ &= \frac{d}{2}\left[\log\left(\frac{1}{2} + \frac{1}{2M}\right) + \log\left(\frac{1}{2} - \frac{1}{2M}\right)\right] \\ &\quad + \frac{x^\top z}{2M}\left[\log\left(\frac{1}{2} + \frac{1}{2M}\right) - \log\left(\frac{1}{2} - \frac{1}{2M}\right)\right] + 1 - \lambda(z) + \theta + \mu^\top z.\end{aligned}$$

Performing a few algebraic manipulations with the logarithmic terms, the final equality becomes

$$d\log\left(\frac{\sqrt{(M+1)(M-1)}}{M}\right) + \frac{x^\top z}{M}\log\left(\sqrt{\frac{M+1}{M-1}}\right) + 1 - \lambda(z) + \theta + \mu^\top z.$$

The complementarity conditions for optimality [Boyd and Vandenberghe 2004] imply that $\lambda(z) = 0$, and since the equality constraints in the problem (57) are satisfied, we can choose θ and μ arbitrarily. Taking

$$\theta = -d\log\left(\frac{\sqrt{(M+1)(M-1)}}{M}\right) - 1 \quad \text{and} \quad \mu = -x\frac{1}{M}\log\left(\sqrt{\frac{M+1}{M-1}}\right)$$

yields that the partial derivatives of $\mathcal{L}$ are 0, which shows that indeed our choice of Q^* is optimal.

E.3. Proof of Proposition 4.4

The proof follows along lines similar to the ℓ_∞ case: we compute the maximum entropy distribution subject to the constraint that $\mathbb{E}[Z] = x$ for some $x \in \mathbb{R}^d$ with $\|x\|_1 \leq 1$, and Z must be supported on the extreme points $\pm Me_i$ of the ℓ_1-ball of radius M. (Recall that $e_i \in \mathbb{R}^d$ are the standard basis vectors.) Based on Theorem 4.2, in order to find the minimax mutual information, we need only consider the cases where $x = \pm e_i$ for some $i \in \{1, \dots, d\}$.

Following this plan, we recall the entropy maximization problem (57), where now $x = \pm e_i$ and the sums are over $z \in M\{\pm e_i\}_{i=1}^d$. As in the proof of Proposition 4.3, we can write the Lagrangian and take its derivatives, finding that for $z = \pm Me_i$ we have

$$\frac{\partial}{\partial q(z)}\mathcal{L}(q, \mu, \lambda, \theta) = \log(q(z)) + 1 - \lambda(z) + \theta - \mu^\top z.$$

Solving for $q(z)$, we find that

$$q(z) = \exp(\lambda(z) - 1 - \theta)\exp(\mu^\top z),$$

but complementarity [Boyd and Vandenberghe 2004] guarantees that $\lambda(z) = 0$ since $q(z) > 0$, and normalizing we may write $q(z) = \exp(-\mu^\top z)/\exp(-\mu^\top \sum_{z'} z')$, where the sum is over the extreme points of the ℓ_1-ball of radius M. In particular, $q(Me_i) \propto e^{-\mu_i}$ and $q(-Me_i) \propto e^{\mu_i}$. Without loss of generality, let $x = e_i$. Symmetry suggests we take (and we verify this to be true)

$$q(z) = \exp(-1-\theta)\begin{cases}\exp(\mu_i) & \text{if } z = Me_i \\ \exp(-\mu_i) & \text{if } z = -Me_i \\ \exp(0) & \text{otherwise.}\end{cases} \tag{59}$$

Indeed, with the choice (59) of q, we have $q(Me_j) - q(-Me_j) = 0$ for $j \neq i$, while (setting $\gamma = \mu_i$ and normalizing appropriately)

$$q(Me_i) - q(-Me_i) = \frac{e^\gamma}{e^{-\gamma} + e^\gamma + 2(d-1)} - \frac{e^{-\gamma}}{e^{-\gamma} + e^\gamma + 2(d-1)}.$$

Thus, if we can solve the equation $Mq(Me_i) - Mq(-Me_i) = 1$, we will be nearly done. To that end, we write

$$\frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d-1)} = \frac{1}{M} \quad \text{or} \quad \beta - \beta^{-1} = \frac{1}{M}(\beta + \beta^{-1} + 2(d-1)),$$

where we identified $\beta = e^\gamma$. Multiplying both sides by β, we have a quadratic equation in β:

$$\beta^2 - 1 = \frac{1}{M}(\beta^2 + 2\beta(d-1) + 1) \quad \text{or} \quad (M-1)\beta^2 - 2(d-1)\beta - (M+1) = 0,$$

whose solution is the positive root of

$$\beta = \frac{2d - 2 \pm \sqrt{(2d-2)^2 + 4(M^2-1)}}{2(M-1)} \quad \text{or} \quad \gamma = \log\left(\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2-1)}}{2(M-1)}\right).$$

By our construction, with γ so defined, we satisfy the constraints that $M[q(Me_i) - q(-Me_i)] = 1$ and $q(Me_j) - q(-Me_j) = 0$ for $j \neq i$. Since q belongs to the exponential family and satisfies the constraints, it maximizes the entropy $H(Z)$ as desired [Cover and Thomas 2006].

Algebraic manipulations and the computation of the conditional entropy $H(Z \mid X = e_i)$ give the remainder of the statement of the proposition.

E.4. Proof of Proposition 4.6

The outline of the proof of Proposition 4.6 is as follows. Lemma D.5 implies that any distribution satisfying optimal local differential privacy must be supported on the extreme points of the outer set D (as in the proof of Theorem 4.2). Given this result, we reduce the problem of finding an optimally private distribution to a linear program, using symmetry arguments to simplify the LP. Finally, we show that the solution to the linear program is unique, which means that we have found the unique distribution satisfying optimal local differential privacy.

We begin by developing a reduction of the problem of finding a distribution with optimal local differential privacy to a linear program. Note that there is a nonincreasing mapping between M—the radius of the larger ℓ_∞ ball—and $\alpha^\star$. Indeed, whenever M increases, the set of distributions Q from which to choose a privacy channel increases, so $\alpha^\star$ decreases. Put inversely, for a given differential privacy level α, we can find the smallest M such that it is possible to construct an α-differentially private channel Q mapping from $[-1, 1]^d$ to $[-M, M]^d$. (Lemma E.2 shows that the mapping from M to $\alpha^\star$ is implicitly invertible.)

Thus, rather than solving for α as a function of M, we take the converse view of finding the largest M such that an α-differentially private distribution exists. Fix $d \in \mathbb{N}$ and (with some abuse of notation) let $Z \in \{-1, 1\}^{d\times 2^d}$ be the matrix whose columns are the edges of the hypercube $\{-1, 1\}^d$. For each $z, x \in \{-1, 1\}^d$, define the variables $q(z \mid x) \geq 0$ to represent the conditional probability of observing Mz given x. Let $q(\cdot \mid x) = [q(z \mid x)]_{z\in\{-1,1\}^d}$ denote the vector version of $q(z \mid x)$. Then, we have that a α-differentially private channel providing an unbiased perturbtation of vectors in $[-1, 1]^d$ to $[-M, M]^d$, exists only if we can find settings of $q(z \mid x)$ such that

$$Zq(\cdot \mid x) - \frac{1}{M}x = 0 \text{ for all } x \in \{-1, 1\}^d,$$

while additionally $q(z \mid x) \leq e^\alpha q(z \mid x')$ and $\sum_z q(z \mid x) = 1$, $q(z \mid x) \geq 0$ for all z, x, x'. Thus, if we make the change of variables $t = 1/M$, we see that finding the smallest possible M—which corresponds to the least perturbation possible for a given privacy level α—can be cast as solving the linear program

$$\begin{aligned} \text{minimize} \quad & -t \\ \text{subject to} \quad & Zq(\cdot \mid x) - tx = 0 \text{ for all } x \in \{-1, 1\}^d \\ & q(z \mid x) \leq e^\alpha q(z \mid x') \text{ for all } x, x', z \in \{-1, 1\}^d \\ & \sum_z q(z \mid x) = 1,\ q(\cdot \mid x) \succeq 0 \text{ for all } x \in \{-1, 1\}^d. \end{aligned} \tag{60}$$

The solution vectors $q(\cdot \mid x)$, $x \in \{-1, 1\}^d$, give the probability mass function for an α-differentially private channel perturbing from $[-1, 1]^d$ to $[-M, M]^d$, where $M = 1/t^*$ and t^* denotes the solution to the LP. This p.m.f. is then optimally locally differentially private with $\alpha = \alpha^\star([-1, 1]^d, [-M, M]^d)$.

It is possible to calculate the solution of the LP (60) by hand, but it is tedious. We thus use the structure of optimal local differential privacy to reduce the problem to a single minimization problem over a vector $q \in \mathbb{R}^{2^d}$ (rather than a matrix $[q(z \mid x)] \in \mathbb{R}^{2^d\times 2^d}$). We have the following lemma.

LEMMA E.1. *A distribution satisfying optimal local differential privacy must, for each $x \in \{-1, 1\}^d$, satisfy $q(\cdot \mid x) = \Pi(x)q$, where $\Pi(x) \in \{0, 1\}^{2^d\times 2^d}$ is a permutation matrix and q is a fixed vector.*

PROOF. Suppose for the sake of contradiction that this is not the case, but the vectors $q(x)$ and t solve the linear program (60). Let Q_1 denote the matrix of the vectors $q(\cdot \mid x)$. Choose vectors $q(\cdot \mid x)$ and $q(\cdot \mid x')$ such that $q(\cdot \mid x) \neq \Pi q(\cdot \mid x')$ for any permutation matrix Π. Now construct vectors $q_2(\cdot \mid x)$ and $q_2(\cdot \mid x')$ such that $q_2(z \mid x) = q(z' \mid x')$, where z' is chosen so that $z'_i x'_i = z_i x_i$, and similarly choose q_2 so that $q_2(z \mid x') = q(z' \mid x)$, where $z_i x_i' = z_i' x_i$. Let Q_2 denote the matrix of vectors q, but where $q_2(\cdot \mid x)$ and $q_2(\cdot \mid x')$ replace $q(\cdot \mid x), q(\cdot \mid x')$. Then, by construction, all the constraints of the original linear program (60) are satisfied. By symmetry and the strict convexity of the mutual information in the channel distribution Q, however, we see that

$$I(P, Q_1) = I(P, Q_2) = \frac{1}{2}\left(I(P, Q_1) + I(P, Q_2)\right) > I\left(P, \frac{1}{2}(Q_1 + Q_2)\right).$$

The decrease in mutual information gives the necessary contradiction. □

With Lemma E.1 in hand, we can now turn to the smaller linear program—in a single vector q and for a single vector $x \in \{-1, 1\}^d$—that will give us the locally optimal differentially private channel. Indeed, we consider the linear program in the variables $t \in \mathbb{R}$ and $q \in \mathbb{R}^{2^d}$, where we let $q(z)$ denote the entry of q corresponding to column z of Z:

$$\begin{aligned} &\text{minimize} \quad -t \\ &\text{subject to} \quad Zq - tx = 0, \;\; q(z) \le e^{\alpha} q(z') \text{ for all } z, z', \;\; \sum_z q(z) = 1, \; q \ge 0. \end{aligned} \tag{61}$$

Define the constants

$$K_d = \sum_{i=0}^{\lfloor d/2 \rfloor} (d - 2i)\binom{d}{i} \;\text{ and }\; C_d = \text{card}\{z \in \{-1, 1\}^d : z^\top x > 0\} = \begin{cases} 2^{d-1} & d \text{ odd} \\ 2^{d-1} - \frac{1}{2}\binom{d}{d/2} & d \text{ even.} \end{cases}$$

We have the following lemma, which characterizes the structure of the solution vector q.

LEMMA E.2. *Define* $\alpha^* = \log \frac{K_d + 2^d - C_d}{K_d - C_d}$. *For any* $\alpha < \alpha^*$, *the unique solution to the linear program* (61) *is given by*

$$q(z) = \begin{cases} \frac{e^{\alpha}}{e^{\alpha} C_d + 2^d - C_d} & \text{if } \langle z, x \rangle > 0 \\ \frac{1}{e^{\alpha} C_d + 2^d - C_d} & \text{otherwise.} \end{cases}$$

PROOF. First, problem (61) is clearly equivalent to the linear program

$$\begin{aligned} &\text{minimize} \quad -t \\ &\text{subject to} \quad Zq - tx = 0, \quad \max_z\{q(z)\} + e^{\alpha} \max_z\{-q(z)\} \le 0, \quad \sum_z q(z) = 1, \; q \ge 0. \end{aligned} \tag{62}$$

Our proof proceeds in two large steps: first, we argue that a q of the form specified in the lemma is indeed the solution to the problem (62), then we use results on uniqueness of solutions to linear programs due to Mangasarian [1979].

For the first step, we begin by writing the Lagrangian to the problem (62). We introduce dual variables $\theta \in \mathbb{R}^{2^d}$ for the constraint $Zq - tx = 0$, $\lambda \ge 0$ for the first inequality, $\tau \in \mathbb{R}$ for the sum constraint, and $\beta \in \mathbb{R}^{2^d}_+$ for the nonnegativity of q. With this, we have Lagrangian

$$\begin{aligned} \mathcal{L}(q, t, \theta, \lambda, \tau, \beta) \; = \; & -t + \theta^\top \left(\sum_z q(z) - tx \right) \\ & + \lambda \max_z\{q(z)\} + e^{\alpha} \max_z\{-q(z)\} + \tau(\mathbb{1}^\top q - 1) - \beta^\top q. \end{aligned} \tag{63}$$

Recall the generalized subgradient KKT conditions for optimality of the solution to an optimization problem [Hiriart-Urruty and Lemaréchal 1996, Chapter VII]. A vector

$q > 0$ is optimal for the problem (62) if the constraints $\max_i\{q_i\} \le e^\alpha \min_i\{q_i\}$ and $\sum_i q_i = 1$ hold, there is a $t \ge 0$ such that $Zq - tx = 0$, and we can find θ, λ, and τ such that

$$Z^\top\theta + \lambda[v_+ - e^\alpha v_-] + \tau\mathbb{1} = 0, \quad \beta = 0, \quad \text{and} \;\; \theta^\top x = -1, \tag{64}$$

where v_+ and v_- are vectors satisfying

$$v_+ \in \operatorname{Conv}\left\{e_i : q_i = \max_j\{q_j\}\right\} \quad \text{and} \quad v_- \in \operatorname{Conv}\left\{e_i : q_i = \min_j\{q_j\}\right\}.$$

That $\beta = 0$ follows by complementarity (recall that $q > 0$ is assumed).

If we can find settings for the vectors θ, λ, τ, and $v_\pm$ satisfying the KKT conditions (64), we are done. To that end, set $\theta = -x/d$. Then, by inspection, $\theta^\top x = -\|x\|_2^2/d = -1$, and we can rewrite the remaining KKT condition by noting that we must find vectors v_+, v_-, and $\tau \in \mathbb{R}$ such that

$$\begin{aligned} &-\frac{1}{d}Z^\top x + v_+ - e^\alpha v_- + \tau\mathbb{1} = 0, \quad v_+^\top\mathbb{1} = v_-^\top\mathbb{1}, \quad v_+ \ge 0, v_- \ge 0, \\ &v_+(z) = 0 \text{ if } q(z) < \max_z\{q(z)\}, \quad \text{and} \quad v_-(z) = 0 \text{ if } q(z) > \min_z\{q(z)\}. \end{aligned} \tag{65}$$

Note that we have eliminated λ as it is a nonnegative homogeneous scaling term on v_+ and v_-. We choose values q^+, q^- with $0 < q^- < q^+$ and set $q(z) = q^+$ when $z^\top x > 0$ and $q(z) = q^-$ when $z^\top x \le 0$, where q^+, q^- are chosen so that $\sum_z q(z) = 1$. We now choose the values of v_+, v_-, and τ satisfying the KKT conditions in the expression (65) based on the values q^+, q^-. Indeed, set

$$v_+(z) = \begin{cases} \frac{z^\top x}{d} - \tau & \text{if } z^\top x > 0 \\ 0 & \text{otherwise;} \end{cases} \quad \text{and} \quad v_-(z) = \begin{cases} -e^{-\alpha}\frac{z^\top x}{d} + e^{-\alpha}\tau & \text{if } z^\top x \le 0 \\ 0 & \text{otherwise.} \end{cases} \tag{66}$$

By inspection, we see that $-Z^\top x/d + v_+ - e^\alpha v_- + \tau = 0$, so the only question remaining is whether we can choose τ such that $v_\pm \ge 0$ and $v_+^\top\mathbb{1} = v_-^\top\mathbb{1}$.

To that end, we recall the definition of the constant K_d, and we seek τ such that

$$\sum_z v_+(z) = \frac{1}{d}K_d - \tau C_d = e^{-\alpha}\frac{1}{d}K_d + e^{-\alpha}\tau(2^d - C_d) = \sum_z v_-(z)$$

by the symmetry in the sums. Rewriting the equation, we find that, for equality, we must have

$$\tau(e^{-\alpha}(2^d - C_d) + C_d) = \frac{1}{d}K_d(1 - e^{-\alpha}) \;\; \text{or} \;\; \tau = \frac{K_d}{d}\cdot\frac{e^\alpha - 1}{e^\alpha C_d + 2^d - C_d} = \frac{K_d}{dC_d}\cdot\frac{e^\alpha - 1}{e^\alpha + 2^d/C_d - 1}.$$

Thus, we find that if α is such that

$$\frac{K_d}{dC_d}\frac{e^\alpha - 1}{e^\alpha + 2^d/C_d - 1} < \frac{1}{d}, \tag{67}$$

then by our choice (66) of the vectors v_+ and v_-, we have $v_+(z) > 0$ whenever $z^\top x > 0$, and $v_-(z) > 0$ whenever $z^\top x \le 0$. Noting that by our setting of $q(z)$, we have by symmetry of Z that there exists a $t > 0$ such that $Zq = tx$, we find that our choice of q is optimal (since the KKT conditions (65) hold).

We have two arguments remaining in the proof. The first is to show that for $\alpha < \alpha^*$ defined in the statement of the lemma, the inequality (67) holds. Rewriting the inequality, we solve

$$e^{\alpha} - 1 = \frac{C_d}{K_d}\left(e^{\alpha} + 2^d/C_d - 1\right) \;\text{ or }\; e^{\alpha}\left(1 - \frac{C_d}{K_d}\right) = \frac{2^d - C_d}{K_d} + 1,$$
$$\text{i.e. } \alpha^* = \log\frac{K_d + 2^d - C_d}{K_d - C_d}.$$

For any $\alpha < \alpha^*$, the strict inequality (67) holds, so the setting (66) of v_+ and v_- satisfy the KKT conditions.

Our last argument regards the uniqueness of the two-valued solution vector q. For that, we apply Mangasarian's result [Mangasarian 1979, Theorem 1] that if there exists an $\epsilon > 0$ such that for any vector $u \in \mathbb{R}^{2^d}$ with $\|u\|_2 = 1$, q is a solution of the linear program (61) when the objective is $-t + \epsilon u^\top q$, then q is unique. Luckily, this is not difficult given our previous work. The Lagrangian (63) for the modified linear program becomes

$$\epsilon u^\top q - t + \theta^\top\left(\sum_z q(z) - tx\right) + \lambda \max_z\{q(z)\} + e^{\alpha}\max_z\{-q(z)\} + \tau(\mathbb{1}^\top q - 1) - \beta^\top q.$$

The only modification in our KKT conditions (64) is that the first equality becomes

$$\epsilon u + Z^\top\theta + \lambda[v_+ - e^{\alpha}v_-] + \tau\mathbb{1} = 0.$$

By the strictness of the inequalities $v_+(z) > 0$ for z such that $z^\top x > 0$ (and similarly for v_-) in the definitions (66) whenever $\alpha < \alpha^*$, we see that for suitably small $\epsilon > 0$, the vectors v_+ and v_- can be perturbed so that the KKT conditions are still satisfied. This proves the uniqueness of the two-valued solution vector q. □

Remarks. Following an argument with completely the same structure as the proof, we see that for any $d \in \mathbb{N}$ (say $d \geq 3$), there are different "regimes" of α, that is, there exists a sequence $\alpha_0^*, \alpha_2^*, \ldots, \alpha_{d-1}^*$ (or α_{d-2}^* if d is even) such that for $\alpha \in (\alpha_{2i}^*, \alpha_{2i+2}^*)$, the unique optimal solution to the linear program (61) is given by taking

$$q(z) \propto \begin{cases} \exp(\alpha) & \text{for } z \text{ s.t. } \langle z, x\rangle > 2(i+1) \\ 1 & \text{for } z \text{ s.t. } \langle z, x\rangle \leq 2(i+1) \end{cases}$$

(for $\alpha < \alpha_0^*$, we say $i = -1$ as before). For $\alpha = \alpha_{2i}^*$, the set of solutions is given by the convex combinations of the solution vectors

$$q_<(z) \propto \begin{cases} \exp(\alpha) & \text{for } z \text{ s.t. } \langle z, x\rangle > 2i \\ 1 & \text{for } z \text{ s.t. } \langle z, x\rangle \leq 2i \end{cases}$$

and

$$q_>(z) \propto \begin{cases} \exp(\alpha) & \text{for } z \text{ s.t. } \langle z, x\rangle > 2(i+1) \\ 1 & \text{for } z \text{ s.t. } \langle z, x\rangle \leq 2(i+1), \end{cases}$$

which follows from arguments similar to our application of Mangasarian's results [Mangasarian 1979].

Now we may complete the proof of Proposition 4.6. Indeed, we see from Lemma E.2 that the distribution satisfying optimal local differential privacy must assign probability masses at two levels—at least when the point being perturbed comes from $\{-1, 1\}^d$. Now let Q be a distribution specified in the lemma. An argument identical to that in our proof of Proposition 4.3—by symmetry—shows that the distribution P maximizing the mutual information $I(P, Q)$ is uniform on $\{-1, 1\}^d$. The uniqueness of Q

then follows from Lemmas E.1 and E.2, which show that such Q is the only distribution that minimizes the radius M of the ball $[-M, M]^d$; inverting this bound gives the proposition. □

E.5. Proof of Corollary 4.5

By scaling, we may assume that $M \geq L = 1$ in the proof of the corollary. First, we claim that as $\gamma \to 0$, the following expansion holds:

$$\log(2d) - \log\left(e^{\gamma} + e^{-\gamma} + 2d - 2\right) + \gamma \frac{e^{\gamma}}{e^{\gamma} + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^{\gamma} + e^{-\gamma} + 2d - 2} = \frac{\gamma^2}{2d} + \Theta\left(\frac{\gamma^4}{d}\right). \quad (68)$$

Before proving this, we use the expansion (68) to prove Corollary 4.5. Noting that

$$\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2 - 1)}}{2(M-1)} = \sqrt{\frac{M+1}{M-1}} + \frac{d-1}{M-1} + \Theta(d^2/M^2),$$

we see that since $\log(1+x) = x - x^2/2 + \Theta(x^3)$, we have $\gamma = \frac{d}{M} + \Theta(\frac{d^2}{M^2})$. Thus, the mutual information in Proposition 4.4 is

$$\begin{aligned} I(P^*, Q^*) &= \frac{\log^2(\sqrt{(M+1)/(M-1)} + d/M + \Theta(d^2/M^2))}{2d} + \Theta\left(\frac{\log^4(1 + d/M)}{d}\right) \\ &= \frac{d}{2M^2} + \Theta\left(\min\left\{\frac{d^3}{M^4}, \frac{\log^4(d)}{d}\right\}\right). \end{aligned}$$

Now we return to showing the claim (68). Indeed, define $f(\gamma) = \log(e^{\gamma} + e^{-\gamma} + 2d - 2)$. Letting $f^{(i)}$ denote the ith derivative of f, we have

$$f^{(1)}(\gamma) = \frac{e^{\gamma} - e^{-\gamma}}{e^{\gamma} + e^{-\gamma} + 2d - 2}, \qquad f^{(2)}(\gamma) = \frac{(e^{\gamma} + e^{-\gamma})(2d-2) + 4}{(e^{\gamma} + e^{-\gamma} + 2d - 2)^2},$$

and

$$f^{(3)}(\gamma) = \frac{-(e^{2\gamma} - e^{-2\gamma})(2d-2) - 8(e^{\gamma} - e^{-\gamma}) + (2d-2)^2(e^{\gamma} - e^{-\gamma})}{(e^{\gamma} + e^{-\gamma} + 2d - 2)^3}.$$

Via a Taylor expansion, we have $f(0) = f(\gamma) - \gamma f^{(1)}(\gamma) + \frac{\gamma^2}{2} f^{(2)}(\gamma) + \mathcal{O}(f^{(3)}(\gamma)\gamma^3)$, and so substituting values for $f(\gamma)$ and $f^{(1)}(\gamma)$, we have

$$\begin{aligned} &\log(2d) - \log(e^{\gamma} + e^{-\gamma} + 2d - 2) + \gamma \frac{e^{\gamma}}{e^{\gamma} + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^{\gamma} + e^{-\gamma} + 2d - 2} \\ &= \frac{(e^{\gamma} + e^{-\gamma})(2d-2) + 4}{(e^{\gamma} + e^{-\gamma} + 2d - 2)^2} \cdot \frac{\gamma^2}{2} + \mathcal{O}\big(f^{(3)}(\gamma)\gamma^3\big). \end{aligned}$$

A few simpler Taylor expansions yield that $f^{(3)}(\gamma) = \mathcal{O}(\gamma/d)$, which means that all we have left to tackle is $f^{(2)}(\gamma)$. But noting that

$$2(e^{\gamma} + e^{-\gamma}) = 4\left(1 + \frac{\gamma^2}{2!} + \frac{\gamma^4}{4!} + \cdots\right) = 4 + \mathcal{O}(\gamma^2)$$

implies that $f^{(2)}(\gamma) = (4d + \mathcal{O}(d\gamma^2))/4d^2$, and hence $(\gamma^2/2)f^{(2)}(\gamma) = \gamma^2/2d + \mathcal{O}(\gamma^4/d)$, which yields the result. □

ACKNOWLEDGMENTS

The authors would like to thank Cynthia Dwork, Guy Rothblum, and Kunal Talwar for feedback on earlier versions of this work. We also thank the reviewers for several constructive and clarifying suggestions.

REFERENCES

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. 2012. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Trans. Inf. Theory* 58, 5, 3235–3249.

A. Beck and M. Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* 31, 167–175.

D. P. Bertsekas and J. N. Tsitsiklis. 1989. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc.

P. Billingsley. 1986. *Probability and Measure*, 2nd Ed. Wiley.

A. Blum, K. Ligett, and A. Roth. 2008. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th Annual ACM Symposium on the Theory of Computing*.

S. Boyd and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. 2011. Differentially private empirical risk minimization. *J. Machine Learn. Res.* 12, 1069–1109.

T. M. Cover and J. A. Thomas. 2006. *Elements of information theory,* 2nd Ed. Wiley.

L. H. Cox, A. F. Karr, and S. K. Kinney. 2011. Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act. *Int. Stat. Rev.* 79, 2, 160–199.

I. Csisz'ar and J. Körner. 1981. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press.

I. Dinur and K. Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the 22nd Symposium on Principles of Database Systems*. 202–210.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. Local privacy and statistical minimax rates. arXiv:1302.3203 [math.ST].

G. T. Duncan and D. Lambert. 1986. Disclosure-limited data dissemination. *J. Amer. Stat. Assoc.* 81, 393, 10–18.

G. T. Duncan and D. Lambert. 1989. The risk of disclosure for microdata. *J. Busin. Economic Statistics* 7, 2, 207–217.

C. Dwork. 2008. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation,* Lecture Notes in Computer Science Series, vol. 4978, Springer, 1–19.

C. Dwork and J. Lei. 2009. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*.

C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*. 265–284.

C. Dwork, G. N. Rothblum, and S. P. Vadhan. 2010. Boosting and differential privacy. In *Proceedings of the 51st Annual Symposium on Foundations of Computer Science*. 51–60.

A. V. Evfimievski, J. Gehrke, and R. Srikant. 2003. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22nd Symposium on Principles of Database Systems*. 211–222.

I. P. Fellegi. 1972. On the question of statistical confidentiality. *J. Amer. Stat. Assoc.* 67, 337, 7–18.

S. R. Ganta, S. Kasiviswanathan, and A. Smith. 2008. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD Conference on Knowledge and Data Discovery (KDD)*.

A. Ghosh, T. Roughgarden, and M. Sundararajan. 2009. Universally utility-maximizing privacy mechanisms. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*.

R. M. Gray. 1990. *Entropy and information theory*. Springer.

R. Hall, A. Rinaldo, and L. Wasserman. 2011. Random differential privacy. arXiv:1112.2680 [stat.ME].

M. Hardt and K. Talwar. 2010. On the geometry of differential privacy. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*. 705–714.

J. Hiriart-Urruty and C. Lemaréchal. 1996. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York.

O. Kallenberg. 1997. *Foundations of Modern Probability*. Springer.

A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. 2006. A framework for evaluating the utility of data altered to protect confidentiality. *Amer. Statistician* 60, 3, 224–232.

S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3, 793–826.

S. P. Kasiviswanathan, M. Rudelson, and A. Smith. 2013. The power of linear reconstruction attacks. In *Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*.

M. Kearns. 1998. Efficient noise-tolerant learning from statistical queries. *J. ACM* 45, 6, 983–1006.

L. Le Cam. 1956. On the asymptotic theory of estimation and hypothesis testing. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, 129–156.

L. Le Cam. 1973. Convergence of estimates under dimensionality restrictions. *Ann. Stat.* 1, 1, 38–53.

Y. Liang, H. V. Poor, and S. Shamai. 2008. Information theoretic security. *Found. Trends Commun. Inf. Theory* 5, 4, 355–580.

O. L. Mangasarian. 1979. Uniqueness of solution in linear programming. *Linear Algebra Appl.* 25, 151–162.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimiz.* 19, 4, 1574–1609.

A. Nemirovski and D. Yudin. 1983. *Problem Complexity and Method Efficiency in Optimization*. Wiley.

A. Nikolov, K. Talwar, and L. Zhang. 2013. The geometry of differential privacy: The sparse and approximate case. In *Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*.

R. R. Phelps. 2001. *Lectures on Choquet's Theorem,* 2nd Ed. Springer.

B. T. Polyak and A. B. Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM J. Cont. Optimiz.* 30, 4, 838–855.

J. P. Reiter. 2005. Estimating risks of identification disclosure in microdata. *J. Amer. Stat. Assoc.* 100, 1103–1113.

B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. 2012. Learning in a large function space: privacy-preserving mechanisms for SVM learning. *J. Priv. Confident.* 4, 1, 65–100.

L. Sankar, S. R. Rajagopalan, and H. V. Poor. 2010. An information-theoretic approach to privacy. In *Proceedings of the 48th Allerton Conference on Communication, Control, and Computing*. 1220–1227.

A. Smith. 2011. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*.

A. W. Van Der Vaart. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

A. Wald. 1939. Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Stat.* 10, 4, 299–326.

S. Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Stat. Assoc.* 60, 309, 63–69.

L. Wasserman and S. Zhou. 2010. A statistical framework for differential privacy. *J. Amer. Stat. Assoc.* 105, 489, 375–389.

Y. Yang and A. Barron. 1999. Information-theoretic determination of minimax rates of convergence. *Ann. Statics* 27, 5, 1564–1599.

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

S. Zhou, J. Lafferty, and L. Wasserman. 2009a. Compressed regression. *IEEE Trans. Inf. Theory* 55, 2, 846–866.

S. Zhou, K. Ligett, and L. Wasserman. 2009b. Differential privacy with compression. In *Proceedings of the IEEE International Symposium on Information Theory*.

M. Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*.

Received October 2012; revised October 2013 and June 2014; accepted June 2014