Privacy Aware Learning

John C. Duchi[1] **Michael I. Jordan**[1,2] **Martin J. Wainwright**[1,2]
[1]Department of Electrical Engineering and Computer Science, [2]Department of Statistics
University of California, Berkeley
Berkeley, CA USA 94720
{jduchi,jordan,wainwrig}@eecs.berkeley.edu

Abstract

We study statistical risk minimization problems under a version of privacy 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, measured by convergence rate, of any statistical estimator.

1 Introduction

There are natural tensions between learning and privacy that arise whenever a learner must aggregate data across multiple individuals. The learner wishes to make optimal use of each data point, but the providers of the data may wish to limit detailed exposure, either to the learner or to other individuals. It is of great interest to characterize such tensions in the form of quantitative tradeoffs that can be both part of the public discourse surrounding the design of systems that learn from data and can be employed as controllable degrees of freedom whenever such a system is deployed.

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 [28] in the 1930's, and which 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 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, providing (under certain conditions) minimax lower bounds that apply to any specific procedure. Finally, the use of loss functions, in particular convex loss functions, in the design of a learning system allows powerful tools of optimization theory to be brought to bear.

In more formal detail, our framework is as follows. 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

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

In the standard formulation of statistical risk minimization, a method $\mathcal{M}$ is given n samples $X_1, \ldots, X_n$, and outputs an estimate θ_n approximately minimizing $R(\theta)$. Instead of allowing $\mathcal{M}$ access to the samples X_i, however, we study the effect of giving only a perturbed view Z_i of each datum X_i, quantifying the rate of convergence of $R(\theta_n)$ to $\inf_{\theta \in \Theta} R(\theta)$ as a function of both the number of samples n and the amount of privacy Z_i provides for X_i.

There is a long history of research at the intersection of privacy and statistics, where there is a natural competition between maintaining the privacy of elements in a dataset $\{X_1, \ldots, X_n\}$ and the output of statistical procedures. Study of this issue goes back at least to the 1960s, when Warner [29] suggested privacy-preserving methods for survey sampling. Recently, there has been substantial work on privacy—focusing on a measure known as differential privacy [12]—in statistics, computer science, and other fields. We cannot hope to do justice to the large body of related work, referring the reader to the survey by Dwork [10] and the statistical framework studied by Wasserman and Zhou [30] for background and references.

In this paper, we study *local privacy* [13, 17], 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 access only disguised views of each datum X_i. Local algorithms are among the most classical approaches to privacy, tracing back to Warner's work on randomized response [29], and rely on communication only of some disguised view Z_i of each true sample X_i. Locally private algorithms are natural when the providers of the data—the population sampled to give $X_1, \ldots, X_n$—do not trust even the statistician or statistical method $\mathcal{M}$, but the providers are interested in the parameters θ^* minimizing $R(\theta)$. For example, in medical applications, a participant may be embarrassed about his use of drugs, but if the loss ℓ is able to measure the likelihood of developing cancer, the participant has high utility for access to the optimal parameters θ^*. In essence, we would like the statistical procedure $\mathcal{M}$ to learn *from* the data $X_1, \ldots, X_n$ but not *about* it.

Our goal is to understand the fundamental tradeoffs between maintaining privacy while still retaining the utility of the statistical inference method $\mathcal{M}$. Though intuitively there must be some tradeoff, quantifying it precisely has been difficult. In the machine learning literature, Chaudhuri et al. [7] develop differentially private empirical risk minimization algorithms, and Dwork and Lei [11] and Smith [26] analyze similar statistical procedures, but do not show that there must be negative effects of privacy. Rubinstein et al. [24] are able to show that it is impossible to obtain a useful parameter vector θ that is substantially differentially private; it is unclear whether their guarantees are improvable. Recent work by Hall et al. [15] gives sharp minimax rates of convergence for differentially private histogram estimation. Blum et al. [5] 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. [25] provide rate-distortion theorems for utility models involving information-theoretic quantities, which has some similarity to our risk-based framework, but it appears challenging to map their setting onto ours. The work most related to ours is probably that of Kasiviswanathan et al. [17], who show that that locally private algorithms coincide with concepts that can be learned with polynomial sample complexity in Kearns's statistical query (SQ) model. In contrast, our analysis addresses sharp rates of convergence, and applies to estimators for a broad class of convex risks (1).

2 Main results and approach

Our approach to local privacy is based on 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. Recalling that the method sees only the perturbed version Z_i of X_i, we adopt a uniform variant of the mutual information $I(Z_i; X_i)$ between the random variables X_i and Z_i as our measure for privacy. This use of mutual information is by no means original [13, 25], but because standard mutual information has deficiencies as a measure of privacy [e.g. 13], we say 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 constraints). This is similar to the worst-case information approach of Evfimievski et al. [13], which limits privacy breaches. (In the long version of this paper [9] we also consider differentially private algorithms.)

The central consequences of our main results are, under standard conditions on the loss functions ℓ, sharp upper and lower bounds on the possible convergence rates for estimation procedures when we wish to guarantee a level of privacy $I(X_i; Z_i) \le I^*$. We show there are problem dependent constants $a(\Theta, \ell)$ and $b(\Theta, \ell)$ such that the rates of convergence of *all possible procedures* are lower bounded by $a(\Theta, \ell)/\sqrt{nI^*}$ and that *there exist* procedures achieving convergence rates of $b(\Theta, \ell)/\sqrt{nI^*}$, where the ratio $b(\Theta, \ell)/a(\Theta, \ell)$ is upper bounded by a universal constant. Thus, we establish and quantify explicitly the tradeoff between statistical estimation and the amount of privacy.

We show that stochastic gradient descent is one procedure that achieves the optimal convergence rates, which means additionally that our upper bounds apply in streaming and online settings, requiring only a fixed-size memory footprint. Our subsequent analysis builds on this favorable property of gradient-based methods, whence we focus on statistical estimation procedures that access data through the subgradients of the loss functions $\partial\ell(X, \theta)$. This is a natural restriction. Gradients of the loss ℓ are asymptotically sufficient [18] (in an asymptotic sense, gradients contain *all* of the statistical information for risk minimization problems), stochastic gradient-based estimation procedures are (sample) minimax optimal and Bahadur efficient [23, 1, 27, Chapter 8], many estimation procedures are gradient-based [20, 6], and distributed optimization procedures that send gradient information across a network to a centralized procedure $\mathcal{M}$ are natural [e.g. 3]. Our mechanism gives $\mathcal{M}$ access to a vector Z_i that is a stochastic (sub)gradient of the loss evaluated on the sample X_i at a parameter θ of the method's choosing:

$$\mathbb{E}[Z_i \mid X_i, \theta] \in \partial\ell(X_i, \theta), \tag{2}$$

where $\partial\ell(X_i, \theta)$ denotes the subgradient set of the function $\theta \mapsto \ell(X_i, \theta)$. In a sense, the unbiasedness of the subgradient inclusion (2) is information-theoretically necessary [1].

To obtain upper and lower bound on the convergence rate of estimation procedures, we provide a two-part analysis. One part requires studying saddle points of the mutual information $I(X; Z)$ (as a function of the distributions P of X and $Q(\cdot \mid X)$ of Z) under natural constraints that allow inference of the optimal parameters θ^* for the risk R. We show that for certain classes of loss functions ℓ and constraints on the communicated version Z_i of the data X_i, there is a unique distribution $Q(\cdot \mid X_i)$ that attains the smallest possible mutual information $I(X; Z)$ for all distributions on X. Using this unique distribution, we can adapt information-theoretic techniques for obtaining lower bounds on estimation [31, 1] to derive our lower bounds. The uniqueness results for the conditional distribution Q show that no algorithm guaranteeing privacy between $\mathcal{M}$ and the samples X_i can do better. We can obtain matching upper bounds by application of known convergence rates for stochastic gradient and mirror descent algorithms [20, 21], which are computationally efficient.

3 Optimal learning rates and tradeoffs

Having outlined our general approach, we turn in this section to providing statements of our main results. Before doing so, we require some formalization of our notions of privacy and error measures, which we now provide.

3.1 Optimal Local Privacy

We begin by describing in slightly more detail the communication protocol by which information about the random variables X is communicated to the procedure $\mathcal{M}$. We assume throughout that there exist two d-dimensional compact sets C, D, where $C \subset \operatorname{int} D \subset \mathbb{R}^d$, and we have that $\partial\ell(x, \theta) \subset C$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. We wish to maximally "disguise" the random variable X with the random variable Z satisfying $Z \in D$. Such a setting is natural; indeed, many online optimization and stochastic approximation algorithms [34, 21, 1] assume that for any $x \in \mathcal{X}$ and $\theta \in \Theta$, if $g \in \partial\ell(x, \theta)$ then $\|g\| \le L$ for some norm $\|\cdot\|$. We may obtain privacy by allowing a perturbation to the subgradient g so long as the perturbation lives in a (larger) norm ball of radius $M \ge L$, so that $C = \{g \in \mathbb{R}^d : \|g\| \le L\} \subset D = \{g \in \mathbb{R}^d : \|g\| \le M\}$.

Now let X have distribution P, and for each $x \in \mathcal{X}$, let $Q(\cdot \mid x)$ denote the regular conditional probability measure of Z given that $X = x$. Let $Q(\cdot)$ denote the marginal probability defined by $Q(A) = \mathbb{E}_P[Q(A \mid X)]$. The mutual information between X and Z is the expected Kullback-Leibler (KL) divergence between $Q(\cdot \mid X)$ and $Q(\cdot)$:

$$I(P, Q) = I(X; Z) := \mathbb{E}_P \left[D_{\rm kl} \left(Q(\cdot \mid X) || Q(\cdot) \right) \right]. \tag{3}$$

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 $\sup_P I(P, Q) \le I^*$. (Evfimievski et al. [13, Definition 6] present a similar condition.) Thus we seek a saddle point P^*, Q^* such that

$$\sup_P I(P, Q^*) \le I(P^*, Q^*) \le \inf_Q I(P^*, Q), \tag{4}$$

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 $Z \sim Q(\cdot \mid X)$, 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 (4), then 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

$$\mathcal{P}(C) := \{\text{Distributions } P \text{ such that } \operatorname{supp} P \subset C\} \tag{5a}$$

and the set of regular conditional distributions (r.c.d.'s), or communicating distributions,

$$\mathcal{Q}(C, D) := \left\{\text{r.c.d.'s } Q \text{ s.t. } \operatorname{supp} Q(\cdot \mid c) \subset D \text{ and } \int_D z dQ(z \mid c) = c \text{ for } c \in C\right\}. \tag{5b}$$

The definitions (5a) and (5b) 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 if $\nabla \ell(x, \theta) \in C$ then $\mathbb{E}_Q[Z \mid X, \theta] := \int_D z dQ(z \mid \nabla \ell(x, \theta)) = \nabla \ell(x, \theta)$. We now make the following key definition:

Definition 1. *The conditional distribution Q^* satisfies* optimal local privacy *for the sets $C \subset D \subset \mathbb{R}^d$ at level I^* if*

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

where the supremum is taken over distributions $P \in \mathcal{P}(C)$ and the infimum is taken over regular conditional distributions $Q \in \mathcal{Q}(C, D)$.

If a distribution Q^* satisfies optimal local privacy, then it guarantees that even for the worst possible distribution on X, the information communicated about X is limited. In a sense, Definition 1 captures 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 this mechanism, if we can characterize a unique distribution Q^* attaining the infimum (4) for P^* (and by extension, for any P), then we may study the effects of privacy between the method $\mathcal{M}$ and X.

3.2 Minimax error and loss functions

Having defined our privacy metric, we now turn to our original goal: quantification of the effect privacy has on statistical estimation rates. Let $\mathcal{M}$ denote any statistical procedure or method (that uses n stochastic gradient samples) and let θ_n denote the output of $\mathcal{M}$ after receiving n such samples. Let P denote the distribution according to which samples X are drawn. We define the (random) error of the method $\mathcal{M}$ on the risk $R(\theta) = \mathbb{E}[\ell(X, \theta)]$ after receiving n sample gradients as

$$\epsilon_n(\mathcal{M}, \ell, \Theta, P) := R(\theta_n) - \inf_{\theta \in \Theta} R(\theta) = \mathbb{E}_P[\ell(X, \theta_n)] - \inf_{\theta \in \Theta} \mathbb{E}_P[\ell(X, \theta)]. \tag{6}$$

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. Let Q denote the regular conditional probability—the channel distribution—whose conditional part is defined on the range of the subgradient mapping $\partial \ell(X, \cdot)$. As the output θ_n of the statistical procedure $\mathcal{M}$ is a random function of both P and Q, we measure the expected sub-optimality of the risk according to both P and Q. Now, let $\mathfrak{L}$ be a collection of loss functions, where $\mathfrak{L}(P)$ denotes the losses $\ell : \operatorname{supp} P \times \Theta \to \mathbb{R}$ belonging to $\mathfrak{L}$. We define the minimax error

$$\epsilon_n^*(\mathfrak{L}, \Theta) := \inf_{\mathcal{M}} \sup_{\ell \in \mathfrak{L}(P), P} \mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)], \tag{7}$$

where the expectation is taken over the random samples $X \sim P$ and $Z \sim Q(\cdot \mid X)$. We characterize the minimax error (7) for several classes of loss functions $\mathfrak{L}(P)$, giving sharp results when the privacy distribution Q satisfies optimal local privacy.

We assume that our collection of loss functions obey certain natural smoothness conditions, which are often (as we see presently) satisfied. We define the class of losses as follows.

Definition 2. *Let $L > 0$ and $p \geq 1$. The set of* (L, p)-loss functions *are those measurable functions $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ such that $x \in \mathcal{X}$, the function $\theta \mapsto \ell(x, \theta)$ is convex and*

$$|\ell(x, \theta) - \ell(x, \theta')| \leq L \left\|\theta - \theta'\right\|_q \tag{8}$$

for any $\theta, \theta' \in \Theta$, where q is the conjugate of p: $1/p + 1/q = 1$.

A loss ℓ satisfies the condition (8) if and only if for all $\theta \in \Theta$, we have the inequality $\left\|g\right\|_p \leq L$ for any subgradient $g \in \partial \ell(x, \theta)$ (e.g. [16]). We give a few standard examples of such loss functions. First, we consider finding a multi-dimensional median, in which case the data $x \in \mathbb{R}^d$ and $\ell(x, \theta) = L \left\|\theta - x\right\|_1$. This loss is L-Lipschitz with respect to the ℓ_1 norm, so it belongs to the class of (L, ∞) losses. A second example includes classification problems, using either the hinge loss or logistic regression loss. In these cases, the data comes in pairs $x = (a, b)$, where $a \in \mathbb{R}^d$ is the set of regressors and $b \in \{-1, 1\}$ is the label; the losses are

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

By computing (sub)gradients, we may verify that each of these belong to the class of (L, p)-losses if and only if the data a satisfies $\left\|a\right\|_p \leq L$, which is a common assumption [7, 24].

The privacy-guaranteeing channel distributions Q^* we construct in Section 4 are motivated by our concern with the (L, p) families of loss functions. In our model of computation, the learning method $\mathcal{M}$ queries the loss $\ell(X_i, \cdot)$ at the point θ; the owner of the datum X_i then computes the subgradient $\partial \ell(X_i, \theta)$ and returns a masked version Z_i with the property that $\mathbb{E}[Z_i \mid X_i, \theta] \in \partial \ell(X_i, \theta)$. In the following two theorems, we give lower bounds on ϵ_n^* for the (L, ∞) and $(L, 1)$ families of loss functions under the constraint that the channel distribution Q *must* guarantee that a limited amount of information $I(X_i; Z_i)$ is communicated: the channel distribution Q satisfies our Definition 1 of optimal local privacy.

3.3 Main theorems

We now state our two main theorems, deferring proofs to Appendix B. Our first theorem applies to the class of (L, ∞) loss functions (recall Definition 2). We assume that the set to which the perturbed data Z must belong is $[-M_\infty, M_\infty]^d$, where $M_\infty \geq L$. We state two variants of the theorem, as one gives sharper results for an important special case.

Theorem 1. *Let $\mathfrak{L}$ be the collection of (L, ∞) loss functions and assume the conditions of the preceding paragraph. Let Q satisfy be optimally private for the collection $\mathfrak{L}$. Then*

(a) If Θ contains the ℓ_∞ ball of radius r,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \cdot \frac{M_\infty r d}{\sqrt{n}}.$$

(b) If $\Theta = \{\theta \in \mathbb{R}^d : \left\|\theta\right\|_1 \leq r\}$,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{r M_\infty \sqrt{\log(2d)}}{17\sqrt{n}}.$$

For our second theorem, 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, i.e., $Z \in \{z \in \mathbb{R}^d \mid \left\|z\right\|_1 \leq M_1\}$. Setting $M = M_1 / L$, we define (these constants relate to the optimal local privacy distribution for ℓ_1-balls)

$$\gamma := \log\left(\frac{2d - 2 + \sqrt{(2d-2)^2 + 4(M^2 - 1)}}{2(M-1)}\right), \quad \text{and} \quad \Delta(\gamma) := \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d-1)}. \tag{9}$$

Theorem 2. *Let $\mathfrak{L}$ be the collection of $(L, 1)$ loss functions and assume the conditions of the preceding paragraph. Let Q be optimally locally private for the collection $\mathfrak{L}$. Then*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \cdot \frac{r L \sqrt{d}}{\sqrt{n} \Delta(\gamma)}.$$

Remarks We make two main remarks about Theorems 1 and 2. First, we note that each result yields a minimax rate for stochastic optimization problems when there is no random distribution Q. Indeed, in Theorem 1, we may take $M_\infty = L$, in which case (focusing on the second statement of the theorem) we obtain the lower bound $rL\sqrt{\log(2d)}/17\sqrt{n}$ when $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$. Mirror descent algorithms [20, 21] attain a matching upper bound (see the long version of this paper [9, Sec. 3.3] for more substantial explanation). Moreover, our analysis is sharper than previous analyses [1, 20], as none (to our knowledge) recover the logarithmic dependence on the dimension d, which is evidently necessary. Theorem 2 provides a similar result when we take $M_1 \downarrow L$, though in this case stochastic gradient descent attains the matching upper bound.

Our second set of remarks are somewhat more striking. In these, we show that the lower bounds in Theorems 1 and 2 give sharp tradeoffs between the statistical rate of convergence for any statistical procedure and the desired privacy of a user. We present two corollaries establishing this tradeoff. In each corollary, we look ahead to Section 4 and use one of Propositions 1 or 2 to derive a bijection between the size M_∞ or M_1 of the perturbation set and the amount of privacy—as measured by the worst case mutual information I^*—provided. We then combine Theorems 1 and 2 with results on stochastic approximation to demonstrate the tradeoffs.

Corollary 1. *Let the conditions of Theorem 1(b) hold, and assume that $M_\infty \ge 2L$. Assume Q^* satisfies optimal local privacy at information level I^*. For universal constants $c \le C$,*

$$c \cdot \frac{rL\sqrt{d \log d}}{\sqrt{nI^*}} \le \epsilon_n^*(\mathfrak{L}, \Theta) \le C \cdot \frac{rL\sqrt{d \log d}}{\sqrt{nI^*}}.$$

Proof Since $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$, mirror descent [2, 21, 20, Chapter 5], using n unbiased stochastic gradient samples whose ℓ_∞ norms are bounded by M_∞, obtains convergence rate $\mathcal{O}(M_\infty r\sqrt{\log d}/\sqrt{n})$. This matches the second statement of Theorem 1. Now fix our desired amount of mutual information I^*. From the remarks following Proposition 1, if we must guarantee that $I^* \ge \sup_P I(P, Q)$ for any distribution P and loss function ℓ whose gradients are bounded in ℓ_∞-norm by L, we *must* (by the remarks following Proposition 1) have

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

Up to higher-order terms, to guarantee a level of privacy with mutual information I^*, we must allow gradient noise up to $M_\infty = L\sqrt{d/I^*}$. Using the bijection between M_∞ and the maximal allowed mutual information I^* under local privacy that we have shown, we substitute $M_\infty = L\sqrt{d}/\sqrt{I^*}$ into the upper and lower bounds that we have already attained. □

Similar upper and lower bounds can be obtained under the conditions of part (a) of Theorem 1, where we need not assume Θ is an ℓ_1-ball, but we lose a factor of $\sqrt{\log d}$ in the lower bound. Now we turn to a parallel result, but applying Theorem 2 and Proposition 2.

Corollary 2. *Let the conditions of Theorem 2 hold and assume that $M_1 \ge 2L$. Assume that Q^* satisfies optimal local privacy at information level I^*. For universal constants $c \le C$,*

$$c \cdot \frac{rLd}{\sqrt{nI^*}} \le \epsilon_n^*(\mathfrak{L}, \Theta) \le C \cdot \frac{rLd}{\sqrt{nI^*}}.$$

Proof By the conditions of optimal local privacy (Proposition 2 and Corollary 3), to have $I^* \ge \sup_P I(P, Q)$ for any loss ℓ whose gradients are bounded in ℓ_1-norm by L, we must have

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

using Corollary 3. 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. Indeed, stochastic gradient descent (SGD) enjoys the following convergence guarantees (e.g. [21]). Let $\Theta \subseteq \mathbb{R}^d$ be contained in the ℓ_∞ ball of radius r and the gradients of the loss ℓ belong to the ℓ_1-ball of radius M_1. Then SGD has $\epsilon_n^*(\mathfrak{L}, \Theta) \le CM_1 r\sqrt{d}/\sqrt{n}$. Now apply the lower bound provided by Theorem 2 and substitute for M_1. □

4 Saddle points, optimal privacy, and mutual information

In this section, we explore conditions for a distribution Q^* to satisfy optimal local privacy, as given by Definition 1. We give characterizations of necessary and sufficient conditions based on the compact sets $C \subset D$ for distributions P^* and Q^* to achieve the saddle point (4). Our results can be viewed as rate distortion theorems [14, 8] (with source P and channel Q) for certain compact alphabets, though as far as we know, they are all new. Thus we sometimes refer to the conditional distribution Q, which is designed to maintain the privacy of the data X by communication of Z, as the channel distribution. Since we wish to bound $I(X; Z)$ for arbitrary losses ℓ, we must address the case when $\ell(X, \theta) = \langle \theta, X \rangle$, in which case $\nabla \ell(X, \theta) = X$; by the data-processing inequality [14, Chapter 5] it is thus no loss of generality to assume that $X \in C$ and that $\mathbb{E}[Z \mid X] = X$.

We begin by defining the types of sets C and D that we use in our characterization of privacy. As we see in Section 3, such sets are reasonable for many applications. We focus on the case when the compact sets C and D are (suitably symmetric) norm balls:

Definition 3. *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* 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 rotationally invariant norm balls with finite numbers of extreme points by providing saddle point distributions P^* and Q^*. We provide the proof of Theorem 3 in Section A.1.

Theorem 3. *Let C be a compact, convex, polytope rotationally invariant through its extreme points $\{u_i\}_{i=1}^m$ and $D = (1 + \alpha)C$ for some $\alpha > 0$. Let Q^* be the conditional distribution on $Z \mid X$ 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 $(1 + \alpha)u_i$ for $i = 1, \ldots, m$. Then Q^* satisfies Definition 1, optimal local privacy, and Q^* is (up to measure zero sets) unique. Moreover, the distribution P^* uniform on $\{u_i\}_{i=1}^m$ uniquely attains the saddle point (4)*.*

Remarks: While in the theorem we assume that $Q^*(\cdot \mid X = x)$ maximizes the entropy for each $x \in C$, this is not in fact essential. In fact, we may introduce a random variable X' between X and Z: let X' be distributed among the extreme points $\{u_i\}_{i=1}^m$ of C in *any* way such that $\mathbb{E}[X' \mid X] = X$, then use the maximum entropy distribution $Q^*(\cdot \mid u_i)$ defined in the theorem when $X \in \{u_i\}_{i=1}^m$ to sample Z from X'. The information processing inequality [14, Chapter 5] guarantees the Markov chain $X \to X' \to Z$ satisfies the minimax bound $I(X; Z) \le \inf_Q \sup_P I(P, Q)$.

With Theorem 3 in place, we can explicitly characterize the distributions achieving optimal local privacy (recall Definition 1) for ℓ_1 and ℓ_∞ balls. We present the propositions in turn, providing some discussion here and deferring proofs to Appendices A.2 and A.3.

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

Proposition 1. *Let $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$ be random variables with $M \ge 1$ and $\mathbb{E}[Z \mid X] = X$ 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*

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

Then Q^* satisfies Definition 1, optimal local privacy, and moreover,

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

Before continuing, we give a more intuitive understanding of Proposition 1. 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 in particular

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

That is, we have for *any* distribution P on $X \in [-1,1]^d$ that (in natural logarithms)

$$I(P, Q^*) \le \frac{d}{M^2} \quad \text{and} \quad I(P, Q^*) = \frac{d}{M^2} + \mathcal{O}(M^{-3}).$$

We now consider the case when $X \in \left\{x \in \mathbb{R}^d \mid \|x\|_1 \le 1\right\}$ and $Z \in \left\{z \in \mathbb{R}^d \mid \|z\|_1 \le M\right\}$. Here the arguments are slightly more complicated, as the coordinates of the random variables are no longer independent, but Theorem 3 still allows us to explicitly characterize the saddle point of the mutual information.

Proposition 2. *Let $X \in \{x \in \mathbb{R}^d \mid \|x\|_1 \le 1\}$ and $Z \in \{z \in \mathbb{R}^d \mid \|z\|_1 \le M\}$ be random variables with $M > 1$. Define the parameter γ as in Eq.* (9)*, and let Q^* be the distribution on $Z \mid X$ such that Z is supported on $\{\pm M e_i\}_{i=1}^d$, and*

$$Q^*(Z = Me_i \mid X = e_i) = \frac{e^\gamma}{e^\gamma + e^{-\gamma} + (2d-2)}, \tag{10a}$$

$$Q^*(Z = -Me_i \mid X = e_i) = \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + (2d-2)}, \tag{10b}$$

$$Q^*(Z = \pm Me_j \mid X = e_i, j \ne i) = \frac{1}{e^\gamma + e^{-\gamma} + (2d-2)}. \tag{10c}$$

(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 conditioned on X' according to (10a)–(10c)*.) Then Q^* satisfies Definition 1, optimal local privacy, and*

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

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 3. Additionally, 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 1. We have the following corollary, proved in Appendix E.1.

Corollary 3. *Let Q^* denote the conditional distribution in Proposition 2. Then*

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

5 Discussion and open questions

This study leaves a number open issues and areas for future work. We study procedures that access each datum only once and through a perturbed view Z_i of the subgradient $\partial \ell(X_i, \theta)$, which allows us to use (essentially) any convex loss. A natural question is whether there are restrictions on the loss function so that a transformed version $(Z_1, \ldots, Z_n)$ of the data are sufficient for inference. Zhou et al. [33] study one such procedure, and nonparametric data releases, such as those Hall et al. [15] study, may also provide insights. Unfortunately, these (and other) current approaches require the data be aggregated by a trusted curator. 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 under other conditions. A better understanding of general privacy-preserving channels Q for alternative constraints to those we have proposed is also desirable.

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 herein prove fruitful, and we hope to investigate some of the above issues in future work.

Acknowledgments We thank Cynthia Dwork, Guy Rothblum, and Kunal Talwar for feedback on early versions of this work. 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. JCD was partially supported by an NDSEG fellowship and a Facebook fellowship.

References

[1] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Trans. on Information Theory*, 58(5):3235–3249, 2012.

[2] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

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

[4] P. Billingsley. *Probability and Measure*. Wiley, Second edition, 1986.

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

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[9] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. URL `http://arxiv.org/abs/1210.2085`, 2012.

[10] C. Dwork. Differential privacy: a survey of results. In *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, p. 1–19. Springer, 2008.

[11] C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the Fourty-First Annual ACM Symposium on the Theory of Computing*, 2009.

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

[13] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, p. 211–222, 2003.

[14] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

[15] R. Hall, A. Rinaldo, and L. Wasserman. Random differential privacy. URL `http://arxiv.org/abs/1112.2680`, 2011.

[16] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer, 1996.

[17] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[18] L. Le Cam. On the asymptotic theory of estimation and hypothesis testing. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, p. 129–156, 1956.

[19] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1(1):38–53, 1973.

[20] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[21] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[22] R. R. Phelps. *Lectures on Choquet's Theorem, Second Edition*. Springer, 2001.

[23] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[24] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

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

[26] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing*, 2011.

[27] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

[28] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

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

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

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

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

[33] S. Zhou, J. Lafferty, and L. Wasserman. Compressed regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009.

[34] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A Proofs of minimax mutual information results

In this section, we provide the proofs of the major results from Section 4. The proofs follow a broadly similar outline. We begin by providing a result, Lemma 2, that allows us 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 key result allows us to reduce computing maximal entropies and minimal mutual information values to finite dimensional convex optimization problems, whose optimality we can check using results from convex analysis and optimization.

A.1 Proof of Theorem 3

Before providing the proof of the theorem proper, we state two auxiliary lemmas that make our strategy cleaner. The first result is the data-processing inequality, which holds for essentially arbitrary random variables [14, Chapter 5].

Lemma 1 (Data processing)**.** *Let $X \to Z \to Y$ be a Markov chain. Then $I(X; Y) \le I(X; Z)$. Equality is attained if and only if X is conditionally independent of Y given Z.*

Coupled with Carathéodory and Minkowski's finite-dimensional version of the Krein-Milman theorem [e.g. 16], Lemma 1 nearly implies that regardless of P, any Q minimizing $I(P, Q)$ must supported on the extreme points of D (up to sets of measure 0). The next lemma, whose (technical) proof we defer to Appendix C, makes this precise (and addresses measurability issues involved in the choice of the extreme points).

Lemma 2. *Let $C \subset D \subset \mathbb{R}^d$ be compact convex sets. Let $Q(\cdot \mid x)$ be a regular conditional distribution defined for $x \in C$ with support contained in D such that for $Z \sim Q(\cdot \mid X = x)$, $\mathbb{E}[Z \mid X] = X$. Let P be a distribution supported on C. If there exists a set $A \subset C$ with $P(A) > 0$ and a set $B \subset D \setminus \mathrm{Ext}(D)$ with $Q(B \mid X = x) > 0$ for $x \in A$, there exists a regular conditional probability distribution Q' where $Q'(\cdot \mid x)$ has support contained in $\mathrm{Ext}(D)$ satisfying*

$$I(P, Q) > I(P, Q') \quad \text{and} \quad \mathbb{E}_{Q'}[Z \mid X] = X.$$

Paraphrasing Lemma 2 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 $\mathrm{Ext}(D)$. Now we proceed to the proof of the theorem.

Proof of Theorem 3 We first consider maximizing the entropy $H(Z \mid X = x)$ for a fixed $x = u_i$, where Z is supported on the extreme points $(1 + \alpha)u_i$ (which we know must be the case from Lemma 2). Now let $q(z \mid x)$ denote the probability mass function of $Z \mid X = x$. For a fixed x, consider the finite dimensional entropy maximization problem

$$\begin{aligned} \underset{q}{\text{minimize}} \ & \sum_z q(z \mid x) \log q(z \mid x) \\ \text{subject to} \ & \sum_z z q(z \mid x) = x, \quad \sum_z q(z \mid x) = 1, \ \ q(z \mid x) \ge 0 \text{ for all } z. \end{aligned} \tag{11}$$

We have the following lemma, which guarantees the form of the solution to the problem (11). See Appendix E.2 for the standard proof.

Lemma 3. *The p.m.f. $q(\cdot \mid x)$ solving problem* (11) *is given by*

$$q(z \mid x) = \frac{\exp(-\mu^\top z)}{\sum_{z'} \exp(-\mu^\top z')}, \tag{12}$$

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.

Having abstractly described the maximum entropy solution (12), we turn to the saddle point $I(P, Q)$. Setting Q^* as in the statement of the theorem, we begin by considering $\sup_P I(P, Q^*)$. Since the support of Q^* is finite (there are m extreme points of D), we have

$$\begin{aligned} I(P, Q^*) = I(X; Z) = H(Z) - H(Z \mid X) &\le \log(m) - H(Z \mid X) \\ &= \log(m) - \int H(Z \mid X = x) dP(x). \end{aligned}$$

For any distribution P on the set C and for any $x \in \operatorname{supp} P$, we can write $x = \sum_i \lambda_i(x) u_i$, where $\lambda_i(x) \geq 0$ and $\sum_i \lambda_i(x) = 1$ (using the Krein-Milman theorem). Define the individual probability mass functions q^i to be the maximum entropy p.m.f. (12) for each of the extreme points u_i of C. Then we can define the conditional probability mass function by

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

(Without loss of generality, we may assume that the λ_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 completely formal using the techniques in the proof of Lemma 2.) Denoting $H(q(\cdot \mid x)) := H(Z \mid X = x)$, we can use the convexity of the negative entropy to see that

$$I(P, Q^*) \leq \log(m) - \int \sum_i \lambda_i(x) H(q^i(\cdot)) dP(x). \tag{13}$$

By symmetry, the entropy $H(q^i(\cdot)) = H(Q^*(\cdot \mid X = u_i))$ is a constant determined by the maximum entropy distribution (12), and thus

$$I(P, Q^*) \leq \log(m) - H(Q^*(\cdot \mid X = u_i)). \tag{14}$$

Equality in the upper bound (14) is attained by taking P^* to be the uniform distribution on the extreme points $\{u_i\}$ of C.

What remains is to argue the identical lower bound for $I(P^*, Q)$ over all conditional distributions Q satisfying the constraints of the theorem statement. We know from Lemma 2 that Q must be supported on $(1+\alpha)u_i$ 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

$$\begin{aligned} \underset{q}{\text{minimize}} \;\; & \sum_z \left(\sum_x q(z \mid x) p(x) \right) \log \left(\sum_x q(z \mid x) p(x) \right) - \sum_x p(x) \sum_z p(z \mid x) \log p(z \mid x) \\ \text{subject to} \;\; & \sum_z p(z \mid x) = 1 \text{ for all } x, \;\; \sum_z z p(z \mid x) = x \text{ for all } x, \;\; p(z \mid x) \geq 0 \text{ for all } x, z. \end{aligned} \tag{15}$$

In the problem (15), 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 [8]. Moreover, an inspection of Cover and Thomas's proof of this fact shows that mutual information is strictly convex except when $q(z \mid x) = \sum_{x'} q(z \mid x') p(x')$ for all x, z; 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 (15). 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

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

If we can satisfy the Karush-Kuhn-Tucker (KKT) conditions (see, e.g., [6]) for optimality of the problem (15), we will be done. Taking derivatives with respect to $q(z \mid x)$, we see

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

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 (12) 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 complimentarity. 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 (12) of $q(z \mid x)$, and denoting the maximum entropy parameters μ there by $\mu^*(x)$, we have

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

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 (15).

Summarizing, the conditional distribution Q^* specified in the statement of the theorem as the maximum entropy distribution (12) 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

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

This is the desired saddle-point (4). □

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

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

Inspecting the proof of Theorem 3 (see the bound (13)) 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—even Lipschitz—in x (see Chapter X of Hiriart-Urruty and Lemaréchal [16]) so this distribution $q(\cdot \mid x)$ defines a measurable random variable as desired.

Additionally, though Theorem 3 assumes that the sets C and D satisfy $D = (1 + \alpha)C$ for some $\alpha > 0$, inspection of the proof yields a somewhat stronger result. Assume the distribution Q maximizing the entropy $H(Z \mid X = x)$ for satisfies $H(Q(\cdot \mid X = x)) = H(Q(\cdot \mid X = x'))$ for each extreme point x of C and additionally satisfies that for each extreme point z of D the sum $\sum_x Q(Z = z \mid X = x)$ is a constant (the sum is over extreme points x of C). Then the upper bound (14) is attained with equality, and a similar calculation yields that Q solves the mtual information problem (15). Thus, as long as C and D are suitably jointly symmetric, Z should be chosen to maximize the entropy $H(Z \mid X = x)$ for each $x \in C$.

A.2 Proof of Proposition 1

Using Theorem 3 (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

$$Z_i \mid X = \begin{cases} M & \text{w.p. } \frac{1}{2} + \frac{X_i}{2M} \\ -M & \text{w.p. } \frac{1}{2} - \frac{X_i}{2M}. \end{cases} \tag{16}$$

We verify that the distribution (16) maximizes the entropy $H(Z \mid X = x)$. Indeed, ignoring the conditioning we write the entropy maximization problem

$$\underset{q}{\text{minimize}} \;\; -H(q) \quad \text{subject to } \sum_z q(z) = 1, \; q(z) \geq 0, \; \sum_z z q(z) = x. \tag{17}$$

Here all sums are over $z \in \text{Ext}([-M, M]^d) = \{-M, M\}^d$. Writing the Lagrangian for the problem (17), we introduce Lagrange multipliers $\mu \in \mathbb{R}^d$, $\lambda(z) \geq 0$, and $\theta \in \mathbb{R}$ and have

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

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 (16) 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 3 says that without loss of generality we may assume that $x \in \{-1, 1\}^d$, the full probability mass function q can be written

$$q(z) = \left(\frac{1}{2} + \frac{1}{2M} \right)^{\frac{d}{2} + \frac{x^\top z}{2M}} \left(\frac{1}{2} - \frac{1}{2M} \right)^{\frac{d}{2} - \frac{x^\top z}{2M}}. \tag{18}$$

Plugging the conditional (18) results in

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

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 complimentarity conditions for optimality [6] imply that $\lambda(z) = 0$, and since the equality constraints in the problem (17) 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 our choice of Q^* is optimal.

A.3 Proof of Proposition 2

The proof outline is similar to that for the ℓ_∞ case: we compute the maximum entropy distribution of $H(Z)$ under the constraint that $\mathbb{E}[Z] = x$ for some $x \in \mathbb{R}^d$ with $\|x\|_1 \le 1$, and Z must be supported on the extreme points $\pm M e_i$ of the ℓ_1-ball of radius M (here e_i are the standard basis vectors). According to Theorem 3, to find the minimax mutual information, we need only consider the cases where $x = \pm e_i$ for some $i \in \{1, \ldots, d\}$.

Following this plan, we recall the entropy maximization problem (17), 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 1, we can write the Lagrangian and take its derivatives, finding that for $z = \pm M e_i$ 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 complimentarity [6] 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, $q(M e_i) \propto e^{-\mu_i}$ and $q(-M e_i) \propto e^{\mu_i}$. Without loss of generality, let $x = e_i$. Symmetry suggests we take (and we verify this to be true)

$$q(z) = \exp(-1 - \theta) \begin{cases} \exp(\mu_i) & \text{if } z = M e_i \\ \exp(-\mu_i) & \text{if } z = -M e_i \\ \exp(0) & \text{otherwise.} \end{cases} \tag{19}$$

Indeed, with the choice (19) of q, we have $q(M e_j) - q(-M e_j) = 0$ for $j \neq i$, while (setting $\gamma = \mu_i$ and normalizing appropriately)

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

Thus, if we can solve $M q(M e_i) - M q(-M e_i) = 1$, we will be nearly done. To that end, we write

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

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

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

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 $M\left[q(M e_i) - q(-M e_i)\right] = 1$ and $q(M e_j) - q(-M e_j) = 0$ for $j \neq i$. Since q belongs to the exponential family and satisfies the constraints, it maximizes the entropy $H(Z)$ as desired [8].

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

B Proofs of statistical rates

In this section, we prove Theorems 1 and 2. Our proofs build on classical information-theoretic techniques from statistical minimax theory [31, 32] while more closely paralleling recent techniques due to Agarwal et al. [1]. At a very high level, our approach is as follows. We begin with a finite set $\mathcal{V}$, and to each member $\alpha \in \mathcal{V}$ we assign a risk functional R_α. We construct the collection $\{R_\alpha\}_{\alpha \in \mathcal{V}}$ so that they "separate" points in the set $\mathcal{V}$ well in the sense that any θ nearly minimizing R_α cannot minimize R_β for $\beta \neq \alpha$. Then we can argue that statistical estimation implies the

existence of a testing procedure that distinguishes α for $\beta \neq \alpha$. By applying known lower bounds (Fano's inequality) for statistical hypothesis testing and careful argument of the mutual information between the random variable X_i and the vector Z_i communicated, we can then argue that the testing problem—and hence the estimation problem—are difficult.

Before continuing, we give a slightly more detailed outline of our arguments. The first step in our proof parallels Agarwal et al.'s construction of "difficult" risk functionals, which allow us (roughly) to show that minimization of the risk functional R is equivalent to being able to estimate the bias of d biased coins. Our starting point is the formulation of classes of loss functions ℓ that make optimization somewhat difficult. To begin, we assume we have a finite index set $\mathcal{V}$ that induces a set of risk functionals R_α for $\alpha \in \mathcal{V}$ (we specify the mapping and the sets $\mathcal{V}$ later). One key insight of Agarwal et al. [1] is to define a discrepancy measure between functionals based on how well minimizers of one risk behave on another from the collection. Thus we let $\theta_\alpha^* \in \operatorname{argmin}_{\theta \in \Theta} R_\alpha(\theta)$ and define the discrepancy measure between two risk functionals as

$$\rho(R_\alpha, R_\beta) := \inf_{\theta \in \Theta} \left[R_\alpha(\theta) + R_\beta(\theta) - R_\alpha(\theta_\alpha^*) - R_\beta(\theta_\beta^*) \right]. \tag{20}$$

We define the minimal discrepancy, which we call the ρ-separation of the set $\mathcal{V}$, to be

$$\rho^\star(\mathcal{V}) := \min \left\{ \rho(R_\alpha, R_\beta) : \alpha, \beta \in \mathcal{V}, \alpha \neq \beta \right\}. \tag{21}$$

When the set $\mathcal{V}$ is clear from context, we use $\rho^\star$ for shorthand. The key to the definition (21) 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. First, note that for any $\theta \in \Theta$, there is at most one $\alpha \in \mathcal{V}$ such that $R_\alpha(\theta) - R_\alpha(\theta_\alpha^*) < \rho^\star/2$. Indeed, if this inequality holds for both α and $\beta \neq \alpha$,

$$\rho^\star(\mathcal{V}) \leq R_\alpha(\theta) + R_\beta(\theta) - R_\alpha(\theta_\alpha^*) - R_\beta(\theta_\beta^*) < \rho^\star(\mathcal{V}),$$

a contradiction. We obtain the following lemma, a variant of Agarwal et al.'s Lemma 2.

Lemma 4 (Agarwal et al. [1])**.** *Let P be a joint distribution over $X \in \mathbb{R}^d$ and $A \in \mathcal{V}$ such that X are i.i.d. given A and*

$$\mathbb{E}_P \left[\ell(X, \theta) \mid A = \alpha \right] = R_\alpha(\theta).$$

Let Q be the conditional distribution of Z given the subgradients $\partial \ell(X, \cdot)$. For any minimization procedure $\mathcal{M}$, one may construct a hypothesis test $\widehat{\alpha}(\mathcal{M}) : (Z_1, \ldots, Z_n) \to \mathcal{V}$ such that

$$\mathbb{E}_{P,Q} \left[\epsilon_n(\mathcal{M}, \ell, \Theta, P) \right] \geq \frac{\rho^\star(\mathcal{V})}{2} \mathbb{P}_{P,Q} \left[\widehat{\alpha} \neq A \right].$$

In particular, if we can bound the probability of error of any hypothesis test for identifying A based on stochastic subgradient samples $Z_1, \ldots, Z_n$, then we have lower bounded the rate at which it is possible to minimize the risk R.

To the end of providing a lower bound on the error of a hypothesis testing problem, we apply Fano's inequality [8]. Let $A \in \mathcal{V}$ be chosen uniformly at random from $\mathcal{V}$. If a procedure observes random variables $Z_1, \ldots, Z_n$, Fano's inequality gives that for any estimate $\widehat{\alpha}$ of A—that is, any measurable function $\widehat{\alpha}$ of $Z_1, \ldots, Z_n$—we have

$$\mathbb{P}(\widehat{\alpha}(Z_1, \ldots, Z_n) \neq A) \geq 1 - \frac{I(Z_1, \ldots, Z_n; A) + \log 2}{\log |\mathcal{V}|}. \tag{22}$$

Using the lower bound provided by Lemma 4 and Fano's inequality (22), 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 2, computing the minimal separation (21) so that we may apply Lemma 4. (See Sections B.1.1–B.1.3.)
2. To be able to apply Fano's inequality (22), we provide an upper bound on the mutual information $I(Z_1, \ldots, Z_n; A)$ for our specific choice of loss from step 1. To do so, we use the fact that for each of Theorems 1 and 2, we used a distribution Q that satisfies our Definition 1 of optimal local privacy, necessitating some subtlety in providing the bound. (See Lemmas 8, 9, and 10 in Section B.2.)
3. The final step is to apply Fano's inequality (22) and Lemma 4 by using the results of steps 1 and 2, which will yield the theorems.

We provide the formal proofs of Theorems 1 and 2 in Sections B.3 and B.4, respectively, in the next two sections performing steps 1 and 2.

B.1 Collections of loss functions

In this section, we construct three example sets of functions, each yielding a different collection of risks, enumerating their separation properties to be able to apply Lemma 4.

B.1.1 Linear Losses

Our first collection of risk functionals is somewhat simpler than our second, and it is not quite so general. Nonetheless, it yields sharper lower bounds than the losses in the sequel for some choices of the set Θ. We assume that the random variables $X \in \mathbb{R}^d$, and we use the linear loss functions

$$\ell(X, \theta) := \langle \theta, X \rangle . \tag{23}$$

For this collection of loss functions, we let $\mathcal{V} = \{\pm e_i\}_{i=1}^d$, where the vectors e_i are the standard basis vectors in $\mathbb{R}^d$, whence $|\mathcal{V}| = 2d$. We also fix a $\delta \in (0, 1/4]$, which we specify later, and choose the distribution P on X so that the final risk is equal to

$$R_\alpha(\theta) = \mathbb{E}_P[\langle \theta, X \rangle] = \frac{c\delta}{d} \langle \alpha, \theta \rangle . \tag{24}$$

We choose the constant c so that the linear loss functions (24) belong to the appropriate loss class.

To construct a risk of the form (24), we draw the random vector $X \in \mathbb{R}^d$ conditional on the parameter α, choosing X from among the 2^d vectors in the scaled hypercube $\{-c, c\}^d$:

$$\text{choose } X \in \{-c, c\}^d \text{ with independent coordinates, where } X_j = \begin{cases} c/d & \text{w.p. } \frac{1+\delta\alpha_j}{2} \\ -c/d & \text{w.p. } \frac{1-\delta\alpha_j}{2}. \end{cases} \tag{25}$$

Under the sampling strategy (25), when $\alpha = \pm e_i$, the coordinate X_j is independent and uniformly chosen from $\{-c/d, c/d\}$ for $j \neq i$. Additionally, we have that $\mathbb{E}[\ell(X, \theta)] = R_\alpha(\theta)$, and we obtain

Lemma 5. *In the sampling scheme* (25)*, take $c = Ld$. Then*

(a) The loss (23) *is L-Lipschitz with respect to the $p = \infty$-norm.*

(b) Let $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$. 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. For the second, we compute minimizers. Indeed, by definition of the dual norm, we see that for $\alpha \in \mathcal{V}$,

$$\inf_{\|\theta\|_1 \leq r} R_\alpha(\theta) = \inf_{\|\theta\|_1 \leq r} \frac{c\delta}{d} \langle \alpha, \theta \rangle = -\frac{c\delta}{d} r \|\alpha\|_\infty = -L\delta r,$$

and the minimizer is uniquely attained at $\theta_\alpha^* = -r\alpha$. Then we have for any $\beta \neq \alpha$ that

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

since no identical coordinates of α and β have the same sign. Multiply the result by $Lr\delta$. □

B.1.2 Hinge (SVM) Losses

We use the set of losses we construct here to study convergence rates of procedures that receive stochastic subgradients bounded in ℓ_1-norm, though the construction is not so simple as that in the previous section. Let $\mathcal{V} \subset \{-1, 1\}^d$ be a subset of the hypercube $\{-1, 1\}^d$ chosen such that for any $\alpha, \alpha' \in \mathcal{V}$ with $\alpha \neq \alpha'$, we have $\|\alpha - \alpha'\|_1 \geq d/2$ (this is equivalent to $\|\alpha - \alpha'\|_0 \geq d/4$). From the Gilbert-Varshamov bound (e.g. [32, Lemma 4]), there are sets of this form with cardinality at least $\operatorname{card}(\mathcal{V}) \geq \exp(d/8)$. Let $\delta \in (0, 1/4]$, and let us assume that Θ contains the ℓ_∞ ball of radius r. Fix $c > 0$ and define the risk

$$R_\alpha(\theta) := \frac{c}{d} \sum_{j=1}^d \frac{1 + \delta\alpha_j}{2} [r - \langle e_j, \theta \rangle]_+ + \frac{c}{d} \sum_{j=1}^d \frac{1 - \delta\alpha_j}{2} [r + \langle e_j, \theta \rangle]_+ . \tag{26}$$

The (unique) minimizer of the risk is

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

To construct the risk (26) as the expectation of a loss, for $x \in \mathbb{R}^d$ we define the hinge loss

$$\ell(x,\theta) = c\left[r - \langle x, \theta\rangle\right]_+. \tag{27}$$

As our sampling process for the data, we choose X from among the $2d$ positive and negative standard basis vectors $\pm e_j$:

$$\text{choose index } j \in \{1,\ldots,d\} \text{ uniformly at random, set } X = \begin{cases} e_j & \text{w.p. } \frac{1+\delta\alpha_j}{2} \\ -e_j & \text{w.p. } \frac{1-\delta\alpha_j}{2}. \end{cases} \tag{28}$$

By inspection, the sampling strategy (28) yields $R_\alpha(\theta) = \mathbb{E}[\ell(X,\theta)]$. Moreover, we have

Lemma 6. *Assume that Θ contains $[-r,r]^d$ and let R_α be defined by the risk* (26)*. Then*

(a) For P with support $\operatorname{supp} P \subseteq \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$, the loss function $\ell(x,\theta) = c\left[1 - \langle\theta, x\rangle\right]_+$ is c-Lipschitz with respect to the ℓ_1-norm.

(b) If $\alpha, \beta \in \mathcal{V}$ with $\alpha \neq \beta$, the discrepancy $\rho(R_\alpha, R_\beta) \ge rc\delta/2$.

Proof The first statement of the lemma is immediate (e.g. [16]), since $\|\partial\ell(x,\theta)\|_1 \le c\|x\|_1 \le c$. For the second statement of the lemma, we see that the minimum of

$$\begin{aligned} &R_\alpha(\theta) + R_\beta(\theta) \\ &= \frac{c}{d}\sum_{j=1}^d \left(\left[r - \langle e_j,\theta\rangle\right]_+ + \left[r + \langle e_j,\theta\rangle\right]_+\right) + \frac{c\delta}{d}\sum_{j:\alpha_j=\beta_j}\left(\alpha_j\left[r - \langle e_j,\theta\rangle\right]_+ - \alpha_j\left[r + \langle e_j,\theta\rangle\right]_+\right) \end{aligned}$$

is attained by any $\theta \in \mathbb{R}^d$ with $\theta_j \in [-r,r]$ for j such that $\alpha_j \neq \beta_j$ and $\theta_j = r\alpha_j$ for j such that $\alpha_j = \beta_j$. Thus we have

$$\begin{aligned} \inf_{\theta\in\Theta}\left\{R_\alpha(\theta) + R_\beta(\theta)\right\} - R_\alpha(\theta^*_\alpha) - R_\beta(\theta^*_\beta) &= \frac{c}{d}\sum_{j=1}^d 2r - \frac{2c}{d}\sum_{j:\alpha_j=\beta_j} r\delta - cr(1-\delta) - cr(1-\delta) \\ &= 2cr - 2cr + 2cr\delta - \frac{2cr\delta}{d}\left(d - \|\alpha-\beta\|_0\right) = \frac{2cr\delta}{d}\|\alpha-\beta\|_0. \end{aligned}$$

Since $\|\alpha - \beta\|_0 \ge d/4$ by construction, we have $\rho(R_\alpha, R_\beta) \ge rc\delta/2$, as desired. □

B.1.3 Median-type Losses

The set of losses we now construct is more generally applicable than the linear losses of Sec. B.1.1, though the resulting lower bounds are somewhat weaker. As in Sec. B.1.2, let $\mathcal{V} \subset \{-1,1\}^d$ be a $d/4$-packing of the hypercube in ℓ_0-norm. Let $\delta \in (0, 1/4]$, and let us assume that Θ contains the ℓ_∞ ball of radius r. Define the risk

$$R_\alpha(\theta) := \frac{c}{d}\sum_{j=1}^d \frac{1+\delta\alpha_j}{2}|\theta - r| + \frac{1-\delta\alpha_j}{2}|\theta + r| = \frac{c}{d}\left(\frac{1+\delta}{2}\|\theta - r\alpha\|_1 + \frac{1-\delta}{2}\|\theta + r\alpha\|_1\right). \tag{29}$$

Notably, the (unique) minimizer of the risk is

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

Our sampling strategy to yield the risk (29) begins by sampling the 2^d vectors of $\{-r,r\}^d$:

$$\text{choose } X \in r\{-1,1\}^d \text{ with independent coordinates, where } X_j = \begin{cases} r & \text{w.p. } \frac{1+\delta\alpha_j}{2} \\ -r & \text{w.p. } \frac{1-\delta\alpha_j}{2}. \end{cases} \tag{30a}$$

In this sampling scheme, we use the loss

$$\ell(X, \theta) = \frac{c}{d} \left\| X - \theta \right\|_1 . \tag{30b}$$

By inspection, the loss (30b) yields the risk R_α through the expectation $R_\alpha(\theta) = \mathbb{E}[\ell(X, \theta)]$ when we use the sampling strategy (30a). The following lemma, due to Agarwal et al. [1], captures the separation properties of the collection $\{R_\alpha\}_{\alpha \in \mathcal{V}}$ of risk functionals:

Lemma 7. *Assume that Θ contains $[-r, r]^d$ and let R_α be defined by the risk* (29)*. If $\alpha, \beta \in \mathcal{V}$ with $\alpha \neq \beta$, the discrepancy $\rho(R_\alpha, R_\beta) \geq rc\delta/2$.*

As a final remark, for random variables $X \in \mathbb{R}^d$, then the loss function (30b) satisfies the Lipschitz continuity bound (8) (for appropriate choice of c) for *any* distribution P on X. Specifically, the subgradient set $\partial \ell(x, \theta) = (c/d) \operatorname{sign}(\theta - x)$, where the sign function $\operatorname{sign}(\cdot)$ is defined coordinate-wise. Thus, taking $c = Ld^{1/q}$, where $1/q + 1/p = 1$, yields a member of the collection of (L, p)-loss functions.

B.2 Mutual information bounds and hypothesis testing

Let $Z_1, \ldots, Z_n$ be unbiased subgradient estimates of the loss $\theta \mapsto \ell(X_i, \theta)$, where X_i are independent samples according to a distribution $P(\cdot \mid A)$. We assume that the samples Z_i are conditionally independent of A given X_i and the parameters θ, which is natural as 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 $A \in \mathcal{V}$.

From Propositions 1 and 2, we know that the channel distributions Q guaranteeing privacy are supported on a finite set: in the case of $p = 1$, on (a multiple) of the standard basis vectors $\{\pm e_i\}_{i=1}^d$, and for $p = \infty$, on (a multiple of) the corners of the hypercube $\{-1, 1\}^d$. Thus, we can decompose the mutual information using the chain rule for mutual information [8]

$$\begin{aligned} I(Z_1, \ldots, Z_n; A) &= \sum_{i=1}^n I(Z_i; A \mid Z_1, \ldots, Z_{i-1}) \\ &= \sum_{i=1}^n \left[H(Z_i \mid Z_1, \ldots, Z_{i-1}) - H(Z_i \mid A, Z_1, \ldots, Z_{i-1}) \right] . \end{aligned}$$

If we let θ_i denote the estimator that the procedure $\mathcal{M}$ uses to query the ith gradient, 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 A and θ_i and conditioning decreases entropy, we have

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

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

$$I(Z_1, \ldots, Z_n; A) \leq \sum_{i=1}^n \int_\Theta I(Z_i; A \mid \theta) dF_i(\theta) \leq \sum_{i=1}^n \sup_{\theta \in \Theta} I(Z_i; A \mid \theta). \tag{31}$$

We now state three lemmas, each bounding the mutual information between observed subgradients Z_i and the random variable A, for different choices of loss function ℓ and conditional distribution Q. The proof of each lemma begins by using the bound (31) to reduce the problem to estimating the mutual information $I(Z; A \mid \theta)$ for a single randomized gradient sample Z. Then, careful calculaion of the distribution of $Z \mid A$ yields the final inequalities. As the proofs are somewhat long and technical, we defer them to Appendix D.

Lemma 8. *Let A be drawn uniformly at random from $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Let X have the distribution* (25) *conditional on $A = \alpha$ and let the loss function $\ell(X, \theta) = \langle X, \theta \rangle$. Let Z be constructed according to the conditional distribution specified by Proposition 1 given a subgradient $\partial \ell(X_i; \theta)$ with $Z \in [-M_\infty, M_\infty]^d$, where $M_\infty \geq c/d$. Then*

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

See Appendix D.1 for a proof of Lemma 8.

Lemma 9. *Let A be drawn uniformly at random from a set $\mathcal{V} \subset \{-1,1\}^d$. Define the distribution $P(\cdot \mid A)$ on X to be such that the jth coordinate $X_j = rA_j$ with probability $(1+\delta)/2$ and $X_j = -rA_j$ with probability $(1-\delta)/2$, each coordinate independent of the others, where $r > 0$ is a constant. Let the loss function*

$$\ell(X, \theta) = \frac{c}{d} \left\|\theta - X\right\|_1.$$

Let Z be constructed according to the distribution specified by Proposition 1 conditional on a subgradient $\partial\ell(X_i; \theta)$, where $Z \in [-M_\infty, M_\infty]^d$ and $M_\infty \geq c/d$. Then

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

See Appendix D.2 for a proof of Lemma 9.

Lemma 10. *Let A 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* (28) *and use the loss* (27)*. Let Z be constructed according to the conditional distribution specified by Proposition 2, where $Z \in \{z \in \mathbb{R}^d : \left\|z\right\|_1 \leq M_1\}$. Define $M = M_1/c$ and the constants*

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

Then

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

We provide the proof of the lemma in Appendix D.3.

In the proof of Theorems 1 and 2, we require one additional result for the cases when the dimension d is small; we apply the result instead of Fano's inequality. Specifically, we use Le Cam's method [19, 32], which 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 $A \in \mathcal{V}$ 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 $A = 1$ and Q_{-1}^n if $A = -1$, then any estimate $\widehat{\alpha}$ of A satisfies the lower bound

$$\mathbb{P}\left(\widehat{\alpha}(Z_1, \ldots, Z_n) \neq A\right) \geq \frac{1}{2} - \frac{1}{2}\left\|Q_1^n - Q_{-1}^n\right\|_{\mathrm{TV}}. \tag{32}$$

See, for example, Yu [32, Lemma 1] and Le Cam [19, Section 2]. Moreover, we have the following lemma.

Lemma 11. *Let Q_1 and Q_{-1} be distributions on $\{-1, 1\}$, where*

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

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

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

We provide the proof of the lemma in Appendix D.4.

B.3 Proof of Theorem 1

We break the proof of Theorem 1 into three parts. In the first, we prove part (a) of the theorem assuming that the dimension $d \geq 9$. Next, we show part (a) for smaller values of the dimension, which requires Le Cam's bounding technique (32). Finally, we prove part (b). Roughly, our strategy is to apply Lemma 4 and one of Lemmas 8 or 9 to achieve a lower bound on the rate of convergence of any estimation procedure. We first recall the beginning of the previous section, stating the following application of Lemma 4 and Fano's inequality (22):

$$\frac{2}{\rho^\star(\mathcal{V})}\mathbb{E}_{P,Q}\left[\epsilon_n(\mathcal{M}, \ell, \Theta, P)\right] \geq \mathbb{P}_{P,Q}\left(\widehat{\alpha}(\mathcal{M}) \neq A\right) \geq 1 - \frac{I(Z_1, \ldots, Z_n; A) + \log 2}{\log|\mathcal{V}|}. \tag{33}$$

Now we give the proof of the first statement of the theorem in the case that $d \geq 9$. Applying Lemmas 7 and 9, we immediately have the following specialization of the inequality (33):

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

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 described in Sections B.1.2 and B.1.3, we see that

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

By the remarks following Lemma 7, we may take $L = c/d$. The numerical inequality $8\log 2 < 6$ coupled with the preceding bound implies

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

By our assumption that $d \geq 9$, if we choose $\delta = M_\infty/8L\sqrt{n}$, then 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

$$\mathbb{E}_{P,Q}\left[\epsilon_n(\mathcal{M}, \ell, \Theta, P)\right] > \frac{rdL\delta}{20} = \frac{1}{160} \cdot \frac{M_\infty rd}{\sqrt{n}}.$$

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, which has $\rho^\star(\mathcal{V}) = Lr\delta$. In this case, the marginal distribution $Q(\cdot \mid A)$ is given by

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

Now, let $Q^n(\cdot \mid A)$ denote the distribution of $Z_1, \ldots, Z_n$ conditional on A. Then applying Lemma 4 and Le Cam's lower bound (32), we obtain the inequality

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

By inspection, the distributions Q^n place us precisely in the conditions Lemma 11 specifies, so if $\delta \leq M/(3L)$, we have the bound

$$\frac{2}{rL\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{2} - \frac{\sqrt{3n}}{2\sqrt{2}} \cdot \frac{\delta L}{M}. \tag{34}$$

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

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{(3\sqrt{2} - \sqrt{3})rM}{36\sqrt{2n}} \geq \frac{rM}{20.3\sqrt{n}}.$$

In turn, for any $d \leq 8$, we immediately find that $1/20.3 \geq d/163$, which completes the proof of Theorem 1(a).

For the second statement of the theorem, we use the linear losses of Section B.1.1 and apply Lemmas 5 and 8 with the choice $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Note that the case $d = 1$ was proved above, so we may assume $d \geq 2$. Since we are in the (L, ∞)-Lipschitz class of loss functions, we take $c = Ld$ in the sampling scheme (25). In this case, the lower bound (33) and Lemma 5's separation guarantee imply that

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

By assumption that $d \geq 2$, we have $\log 2/\log(2d) \leq 1/2$, which, after an application of Lemma 8, yields

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

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

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

which is equivalent in this case to

$$\mathbb{E}_{P,Q}\left[\epsilon_n(\mathcal{M}, \ell, \Theta, P)\right] \geq \frac{rL\delta}{8} = \frac{1}{16} \cdot \frac{M_\infty r\sqrt{\log(2d)}}{\sqrt{n}}.$$

B.4 Proof of Theorem 2

The proof of Theorem 2 is quite similar to that of Theorem 1, except that we apply Lemma 10 in place of Lemmas 8 or 9. Indeed, following identical steps to those in the proof of Theorem 1, we see that with the packing $\mathcal{V}\{-1, 1\}^d$ of size $|\mathcal{V}| \geq \exp(d/8)$, we have

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

Consequently, if we choose $\delta = \sqrt{d}/(8\Delta(\gamma)\sqrt{n})$, 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

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{rL\delta}{20} = \frac{1}{160} \cdot \frac{rL\sqrt{d}}{\sqrt{n}\Delta(\gamma)},$$

which completes the proof (as the case $d \leq 8$ is identical to that in Theorem 1).

C Conditional Probabilities and Measurability

In this appendix, we present some basic lemmas on conditional independence and regular conditional probabilities that will be useful. We begin with a precise definition of a regular conditional probability.

Definition 4. *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*

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

Any Markov chain has a transition probability; conversely, any set of consistent transition probabilities define a Markov chain [4].

Proof of Lemma 2 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 [22]. We begin our proof by stating a measurable selection theorem [22, Theorem 11.4], though we restrict the theorem's statement to subsets of finite dimensional space.

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

In Proposition 3, 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 3 (we take $\nu(x, \cdot) = \mu_x$) with the additional properties that

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

Now suppose the conditions of the lemma: that Y is distributed according to Q (conditional on X) and there exists a set $A \subset C$ of positive P-measure and a set $B \subset D \setminus \mathrm{Ext}(D)$ for which $Q(B \mid X = x) > 0$ for all $x \in A$. For any $y \in D$, Proposition 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 [4] 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

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

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 1. □

D Calculation of the mutual information for sampling strategies

This section contains the proofs of Lemma 8, Lemma 9, and Lemma 10. The proofs of the latter two require a minor lemma, which we present here before giving the proofs proper.

Lemma 12. *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. □

D.1 Proof of Lemma 8

It is clear that the subgradient set $\partial \ell(X_i; \theta)$ is independent of θ, so we may use the inequality (31) to bound the mutual information of A and a single sample Z. Define $M = M_\infty d/c$. Since the sampling scheme (25) is independent per-coordinate, we see immediately that if Z_j denotes the jth coordinate of Z then

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

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

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

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

$$Q(Z_j = M_\infty \mid A_j = \alpha_j = 0) = \frac{1}{2}, \quad \text{and} \quad Q(Z_j = -M_\infty \mid A_j = \alpha_j = 0) = \frac{1}{2}.$$

On the other hand, by our definition of M and choice of sampling scheme, for the "on" index in A, we have

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

Consequently, we see that if we define $h(p) = p \log p + (1-p)\log(1-p)$, then

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

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

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

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

D.2 Proof of Lemma 9

By using the inequality (31), a bound on the mutual information $I(Z; A \mid \theta)$ 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 $M = M_\infty d/c$ 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

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

Now consider the distribution of Z_j given A_j and θ_j. Since the distribution has identical entropy (by symmetry) for any value of A_j, we can without loss of generality fix $A = \alpha$ and assume $\alpha_j = 1$. Then for $\theta_j \in (-1, 1)$, the jth component of the subgradient $\partial\ell(X; \theta)$ is $-X_j$, whence we see that

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

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

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

which increases the entropy $H(Z_j \mid A_j, \theta_j)$ by Lemma 12. Thus we see that $\theta_j \in (-1, 1)$, yielding the Bernoulli marginal $(\frac{1}{2} + \delta/2M, \frac{1}{2} - \delta/2M)$ on $Z_j \mid A_j$, has the smallest entropy $H(Z_j \mid A_j, \theta_j)$. Summarizing, we have

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

As in the proof of Lemma 8, we use the concavity of $\log$ to see that

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

Applying the bound (31) and replacing $M = M_\infty d/c$ completes the proof.

D.3 Proof of Lemma 10

Letting Z denote a single subgradient sample using the conditional distrubtion Q specified by Proposition 2, we prove that for any $\theta \in \mathbb{R}^d$, we have

$$I(Z; A \mid \theta) \le \delta^2 \Delta(\gamma)^2 \tag{35}$$

Recall our construction of the SVM risk (26) using the indivial hinge losses (27), where we see that if $X = e_i$, the loss is equal to $c\left[r - \theta_i\right]_+$. We have

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

For the remainder of this proof, we use the shorthand

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

for the denominator in many of our expressions. By the construction in Proposition 2, we have

$$Q(Z = M_1 e_i \mid X = e_i, \theta) = \begin{cases} \frac{e^{-\gamma}}{D} & \text{if } \theta_i \le r \\ \frac{1}{2d} & \text{if } \theta_i > r \end{cases} \tag{36}$$

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

$$Q(Z = M_1 e_j \mid X = e_i, \theta) = \begin{cases} \frac{1}{D} & \text{if } \theta_i \le r \\ \frac{1}{2d} & \text{if } \theta_i > r. \end{cases} \tag{37}$$

For $X = -e_i$, we have the conditional distribution parallel to (36):

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

For any given θ, we have that

$$I(Z; A \mid \theta) = H(Z \mid \theta) - H(Z \mid A, \theta) \le \log(2d) - \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} H(Z \mid \theta, A = \alpha) \tag{38}$$

since the choice of A is uniform and Z takes on at most $2d$ values. We thus use the conditional distributions (36) and (37) to compute the entropy $H(Z \mid \theta, A)$ (specifically, the minimal such entropy across all values of θ). To do this, we compute the marginal distribution $Q(z \mid \alpha)$, arguing that $H(Z \mid \theta, A)$ is minimal for $\theta \in \operatorname{int}[-r, r]^d$. When $\theta_j \in (-r, r)$ for all j, we have

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

When $\alpha_i = 1$, we thus have that

$$\begin{aligned} Q(Z = M_1 e_i \mid A = \alpha, \theta) &= \frac{1+\delta}{2d} \frac{e^{-\gamma}}{D} + \frac{1-\delta}{2d} \frac{e^{\gamma}}{D} + \sum_{j \neq i} \frac{1}{D} \left(\frac{1+\delta\alpha_j}{2d} + \frac{1-\delta\alpha_j}{2d} \right) \\ &= \frac{e^\gamma + e^{-\gamma} + \delta(e^{-\gamma} - e^\gamma)}{2dD} + \frac{d-1}{dD} = \frac{1}{2d} + \frac{\delta(e^{-\gamma} - e^\gamma)}{2dD}, \end{aligned} \tag{39a}$$

and under the same condition,

$$Q(Z = -M_1 e_i \mid A = \alpha, \theta) = \frac{e^\gamma + e^{-\gamma} + \delta(e^{\gamma} + e^{-\gamma})}{2dD} + \frac{d-1}{dD} = \frac{1}{2d} + \frac{\delta(e^{\gamma} - e^{-\gamma})}{2dD}. \tag{39b}$$

If for any (possibly multiple) indices j we have $\theta_j \notin (-r, r)$, then via a bit of algebra and the conditional distributions (36) and (37), we see that there exists an $\epsilon \in (0, 1)$ such that

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

Lemma 12 then implies that if $\theta \in \operatorname{int}[-r,r]^d$ while $\theta' \notin \operatorname{int}[-r,r]^d$, then

$$H(Z \mid \theta, A = \alpha) < H(Z \mid \theta', A = \alpha).$$

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

Now we compute the entropy $H(Z \mid \theta, \alpha)$ using the marginal conditional distributions (39a) and (39b), which describe $Z \mid A$ when $\theta \in \operatorname{int}[-r,r]^d$. Indeed, recall the definition in the statement of the lemma of the difference $\Delta(\gamma)$. If we define the relation $z \sim \alpha$ for $z \in \{\pm M_1 e_j\}_{j=1}^d$ to mean that if $z = M_1 e_i$, then $\alpha_i = 1$ and if $z = -M_1 e_i$ then $\alpha_i = -1$, then we see that the entropy is

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

As in the proofs of Lemmas 8 and 9, we use the concavity of $\log(\cdot)$ to see that

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

Invoking the earlier bound (38) and adding $\log(2d)$ to the above expression completes the proof of the claim (35).

D.4 Proof of Lemma 11

Recall that for any two probability distributions P, Q, Pinsker's inequality [8] asserts that the total variation norm is bounded as $\|P - Q\|_{\mathrm{TV}} \le \sqrt{D_{\mathrm{kl}}\left(P||Q\right)/2}$. Applying this inequality in our setting, we find that

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

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

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

which completes the proof.

E Technical Lemmas

E.1 Proof of Corollary 3

First, we claim that as $\gamma \to 0$, the following expansion holds:

$$\log(2d) - \log\left(e^\gamma + e^{-\gamma} + 2d - 2\right) + \gamma \frac{e^\gamma}{e^\gamma + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2} = \frac{\gamma^2}{2d} + \Theta\left(\frac{\gamma^4}{d}\right). \tag{40}$$

Before proving this, we use the expansion (40) to prove Corollary 3. 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 $\gamma = \frac{d}{M} + \Theta\left(\frac{d^2}{M^2}\right)$. Thus the mutual information in Proposition 2 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 (40). Indeed, define $f(\gamma) = \log(e^\gamma + e^{-\gamma} + 2d - 2)$. Taking several derivatives, we have

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

and

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

We see via a Taylor expansion that the difference

$$\log(2d) = \log(e^\gamma - e^{-\gamma} + 2d - 2) + (0-\gamma)f^{(1)}(\gamma) + \frac{(0-\gamma)^2}{2} f^{(2)}(\gamma) + \mathcal{O}\left(f^{(3)}(\gamma)\gamma^3\right).$$

Recalling our calculation of the first derivative $f^{(1)}(\gamma)$, we thus see that

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

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

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

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

E.2 Proof of Lemma 3

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

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

Since the problem (11) has convex cost, linear constraints, and non-empty domain, strong duality obtains 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. □