Proceedings of Machine Learning Research vol 99:1–31, 2019 32nd Annual Conference on Learning Theory

Lower Bounds for Locally Private Estimation via Communication Complexity

John Duchi JDUCHI@STANFORD.EDU
Departments of Statistics and Electrical Engineering, Stanford University, and Apple

Ryan Rogers RYAN_ROGERS@APPLE.COM
Apple

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We develop lower bounds for estimation under local privacy constraints—including differential privacy and its relaxations to approximate or Rényi differential privacy—by showing an equivalence between private estimation and communication-restricted estimation problems. Our results apply to arbitrarily interactive privacy mechanisms, and they also give sharp lower bounds for all levels of differential privacy protections, that is, privacy mechanisms with privacy levels $\varepsilon \in [0, \infty)$. As a particular consequence of our results, we show that the minimax mean-squared error for estimating the mean of a bounded or Gaussian random vector in d dimensions scales as $\frac{d}{n} \cdot \frac{d}{\min\{\varepsilon, \varepsilon^2\}}$.

1. Introduction

Estimation problems in which users keep their personal data private even from data collectors are of increasing interest in large-scale machine learning applications in both industrial (Erlingsson et al., 2014; Apple Differential Privacy Team, 2017; Bhowmick et al., 2018) and academic (e.g. Warner, 1965; Beimel et al., 2008; Kasiviswanathan et al., 2011; Duchi et al., 2018) settings. These notions of privacy are satisfying because a user or data provider can be confident that his or her data will remain private irrespective of what data collectors do, and they mitigate risks for data collectors, limiting challenges of hacking or other interference. Because of their importance, a parallel literature on optimality results in local privacy is developing. Yet this theory fails to address a number of important issues. Most saliently, many of these results only apply to non-adaptive privatization schemes, that is, the scheme remains static for all data contributors, except in 1-dimensional problems (Duchi et al., 2018; Ye and Barg, 2018; Gaboardi et al., 2018). A second issue is that these results provide meaningful bounds only for certain types of privacy. Typically, the results are sharp only for high levels of privacy (in the language of differential privacy, privacy parameters $\varepsilon \leq 1$), as in the papers of Duchi et al. (2018), Rohde and Steinberger (2018), and Duchi and Ruan (2018), or at most logarithmic in dimension (Ye and Barg, 2018); given the promise of privacy amplification in local settings (Erlingsson et al., 2019) and challenges of high-dimensional problems (Duchi et al., 2018; Duchi and Ruan, 2018), it is important to address limits in the case that $\varepsilon \gg 1$. With the exception of Duchi and Ruan, they also fail to apply to weakenings of differential privacy.

We remove many of these restrictions by framing the problem of estimation and learning under local privacy constraints as a problem in the communication complexity of statistical estimation. By doing so, we can build off of a line of sophisticated results due to Zhang et al. (2013), Garg et al. (2014), and Braverman et al. (2016), who develop minimax lower bounds on distributed estimation

problems. To set the stage for our results and give intuition for what follows, we recall the intuitive consequences of these results. Each applies in a setting in which n machines each receive a sample X_i from an underlying (unknown) population distribution P. These machines then interactively communicate with a central server, or a public blackboard, sending $n \cdot I$ bits of (Shannon) information in total, so that each sends an average of I bits. For d-dimensional mean estimation problems, where $X_i \in \mathbb{R}^d$ and the goal is to estimate $\mathbb{E}_P[X]$, the main consequences of these papers is that the mean-squared error for estimation must scale as $\frac{d}{n} \cdot \max\{\frac{d}{I}, 1\}$, where d/n is the optimal (communication unlimited) mean-squared error based on a sample of size n. Such scaling is intuitive, as to estimate a d-dimensional quantity, we expect each machine must send roughly d bits to achieve optimal complexity, and otherwise we receive information about only d/I coordinates. The strength of these results is that, in the most general case (Braverman et al., 2016), they allow essentially arbitrary interaction between the machines, so long as it is mitigated by the information constraints.

We leverage these results on information-limited estimation to establish lower bounds for locally private estimation. By providing bounds on the information released by locally private protocols—even when data release schemes are adaptive and arbitrarily interactive—we can nearly immediately provide minimax lower bounds on rates of convergence in estimation and learning problems under privacy. By using this information-based-complexity framework, we can simultaneously address each of the challenges we identify in previous work on estimation under privacy constraints, in that our results apply to differential privacy and its weakenings, including approximate, concentrated, and Rényi differential privacy (Dwork et al., 2006b,a; Dwork and Rothblum, 2016; Bun and Steinke, 2016; Mironov, 2017). They also apply to arbitrarily interactive data release scenarios. Roughly, what we show is that so long as we wish to estimate quantities for d-dimensional parameters that are "independent" of one another—which we define subsequently—the effective sample size available to a private procedure reduces from n to $n \cdot \min\{\varepsilon, \varepsilon^2, d\}/d$ for all ε-private procedures.

The use of information and communication complexity in determining fundamental limits in differential privacy is not uniquely ours. McGregor et al. (2010) show strong relationships between approximating functions by low-error differentially private protocols and low communication protocols. In their case, however, they study low error approximation of *sample* quantities, where one wishes to estimate $f(X_1, \ldots, X_n)$ for a function f. Here, as in most work in statistics and learning (Wainwright, 2019; Yu, 1997; Duchi et al., 2018), we provide limits on the estimation functions of the population from which the sample comes. In recent work, Joseph et al. (2018, Sec. 5) give communication-based bounds for locally-private estimation of a 1-dimensional Gaussian mean; their bound requires a single pass through the data and privacy parameter $\varepsilon = O(1)$.

As a consequence of our lower bounds, we identify several open questions. Work in information-limited estimation (Zhang et al., 2013; Garg et al., 2014; Braverman et al., 2016) typically strongly relies on independence among estimands, which allows decoupling them. Our results similarly suffer these restrictions, which is essential: when correlations exist among different coordinates of the sample vectors X, it is possible to achieve faster convergence. Thus, we argue that we should have renewed focus on *local* (non-minimax) notions of complexity (Le Cam and Yang, 2000; van der Vaart, 1998; Duchi and Ruan, 2018), which address the difficulty of the particular problem at hand.

Notation We index several quantities. We always indicate coordinates of a vector by j, and (independent) vectors we index by i. We consider private protocols communicating in rounds indexed by time t. We let $Z_{\le i} := (Z_1, \ldots, Z_i)$ and $Z_{<i} := (Z_1, \ldots, Z_{i-1})$, and similarly for superscripts. For distributions P and Q, $D_\alpha(P\|Q) := \frac{1}{\alpha-1} \log \int (dP/dQ)^\alpha dQ$ is the Rényi α-divergence.

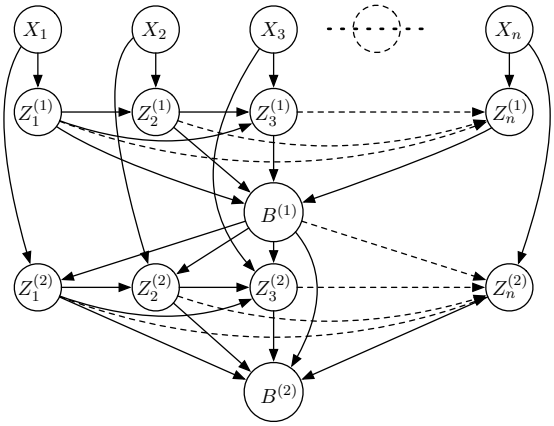

Figure 1. Two rounds of communication of variables, writing to public blackboards $B^{(1)}$ and $B^{(2)}$.

2. Problem setting and main results

We first describe our problem setting in detail, providing graphical representations of our privacy (or communication) settings. We present corollaries of our main lower bounds to highlight their application, then (in Section 4) give the main techniques, which extend Assouad's method.

2.1. Local privacy and interactivity

In our local privacy setting, we consider n individuals, each with private data X_i, $i = 1, \ldots, n$, and each individual i communicates privatized views Z_i of X_i. This private communication may depend on other data providers' private data. We consider communication of privatized data in rounds $t = 1, 2, \ldots, T$, where T may be infinite, and in round t, individual i communicates private datum $Z_i^{(t)}$, which may depend on all previous private communications. This is the standard blackboard communication model; at round t the $Z_i^{(t)}$ and previous blackboards $B^{(t-1)}$ join into $B^{(t)} = (Z_{\le n}^{(t)}, B^{(t-1)})$. Thus, at round t, individual i generates the private variable $Z_i^{(t)}$ according to the channel

$$Q_{i,t}\big(\cdot \mid X_i, Z_{<i}^{(t)}, B^{(t-1)}\big).$$

Figure 1 illustrates this communication scheme over two rounds of communication. We require that the channels be regular conditional probabilities (Billingsley, 1986).

Our main assumptions are that the channels satisfy quantitative privacy definitions.

Definition 1 *Let $\varepsilon \ge 0$. A random variable Z is (ε, δ)-*differentially private *(Dwork et al., 2006b,a) for $X \in \mathcal{X}$ if conditional on $X = x$, Z has distribution $Q(\cdot \mid x)$ and for all measurable S and x, x',*

$$Q(Z \in S \mid x) \le e^{\varepsilon} Q(Z \in S \mid x') + \delta.$$

*When $\delta = 0$, we say Q is ε-*differentially private. *For $\alpha \ge 1$, the channel is (ε, α)-*Rényi differentially private *(Mironov, 2017) if for all $x, x' \in \mathcal{X}$,*

$$D_\alpha\big(Q(\cdot \mid x) \| Q(\cdot \mid x')\big) \le \varepsilon.$$

By taking $\alpha = 1$ in Definition 1, we obtain *ε-KL-privacy*. If the channel Q is ε-differentially private, then for any $\alpha \ge 1$, it also satisfies (Mironov, 2017, Lemma 1)

$$D_\alpha\big(Q(\cdot \mid x) \| Q(\cdot \mid x')\big) \le \min\left\{2(\alpha - 1)\varepsilon^2 + \min\{2, (e^\varepsilon - 1)\}\varepsilon, \varepsilon\right\}. \tag{1}$$

Because Rényi-divergence is non-decreasing in α, any (ε, α)-Rényi differentially private channel is also (ε, α')-Rényi private for $\alpha' \le \alpha$, making KL-privacy the weakest Rényi privacy.

We consider channel and disclosure scenarios where users and data providers obtain a given amount of privacy, but multiple notions of privacy are possible. We separate these by allowing either full interactivity or requiring a type of compositionality of the private data releases; for more on this distinction and examples separating the classes, see Joseph et al. (2019).

2.1.1. FULLY INTERACTIVE PRIVACY MECHANISMS

The first and weakest assumptions on privacy we make are that the private $\boldsymbol{Z} := \{Z_i^{(t)}\}_{i,t}$, or the entire communication transcript, is private. To define this locally private setting, we require an appropriate definition of privacy, for which we use Feldman and Steinke (2018).

Definition 2 *Let $Q(\boldsymbol{Z} \in \cdot \mid x_{\le n})$ denote the distribution of the collection $\boldsymbol{Z}$ conditional on $X_{\le n} = x_{\le n}$, and for $i = 1, \ldots, n$, let the samples $x_{\le n}$ and $x_{\le n}^{(i)} \in \mathcal{X}^n$ differ in only example i, otherwise being arbitrary. The output $\boldsymbol{Z}$ is* $\varepsilon_{\mathrm{kl}}$-KL-locally private on average *if*

$$\frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{kl}} \left(Q(\boldsymbol{Z} \in \cdot \mid x_{\le n}) \| Q(\boldsymbol{Z} \in \cdot \mid x_{\le n}^{(i)}) \right) \le \varepsilon_{\mathrm{kl}}.$$

Definition 2 is weaker than most versions of local privacy. The most general standard notion of Rényi (ε, α)-privacy is that $D_\alpha\big(Q(\boldsymbol{Z} \in \cdot \mid x_{\le n}) \| Q(\boldsymbol{Z} \in \cdot \mid x'_{\le n}\big) \le \varepsilon$ for all samples $x_{\le n}$ and $x'_{\le n}$ differing in a single entry; this immediately implies Definition 2. We thus make the following

Assumption A1 (Fully interactive local differential privacy) *The entire output collection $\boldsymbol{Z}$ is $\varepsilon_{\mathrm{kl}}$-KL-locally private on average (Definition 2).*

Assumption A1 makes no assumptions on the local randomizers, requiring that the entire set of communicated private views $\boldsymbol{Z}$ is private for each individual i. There may be challenges in the implementation of general protocols satisfying Assumption A1 if the privacy of user i depends on the behavior of user i'—adversarial users—though for the purposes of lower bounds, this appears to be the weakest model of local privacy. Assumption A1 is also weaker than the assumption that the private variables $\boldsymbol{Z}$ are ε-differentially private: inequality (1) shows that ε-differential privacy implies $\varepsilon_{\mathrm{kl}} = \min\{\varepsilon, \varepsilon^2 / \log 2\}$-KL privacy. Thus, if each individual i is guaranteed (differential) privacy loss ε_i, the KL-privacy loss satisfies

$$\varepsilon_{\mathrm{kl}} \le \frac{1}{n} \sum_{i=1}^{n} \min \left\{ \varepsilon_i, \frac{\varepsilon_i^2}{\log 2} \right\}. \tag{2}$$

We can also consider (fully interactive) local approximate differential privacy, though in the case that $\delta > 0$, our lower bounds require a slight technical modification of Assumption A1, requiring that the domain $\mathcal{X}$ of the data X_i be finite and δ be appropriately small.

Assumption A1′ (Fully interactive local approximate differential privacy) *The output $\boldsymbol{Z}$ is (ε, δ)-differentially private: for each $S \subset \mathcal{Z}^{nT}$ and pair of samples $x_{\le n}, x'_{\le n} \in \mathcal{X}^n$ differing in at most a single element,*

$$Q\left(\boldsymbol{Z} \in S \mid X_{\le n} = x_{\le n}\right) \le e^{\varepsilon} Q\left(\boldsymbol{Z} \in S \mid X_{\le n} = x'_{\le n}\right) + \delta.$$

In addition, the parameters (ε, δ) *satisfy*

$$\delta \le \frac{\min\{\varepsilon, 1\}}{256}, \quad \delta \max\{\varepsilon^{-1}, 1\} \log \frac{1}{\delta \max\{\varepsilon^{-1}, 1\}} \le \varepsilon^2, \quad \delta \le \frac{\min\{\varepsilon, \varepsilon^2\}}{\log^2 |\mathcal{X}|}$$

and, if $\varepsilon \le \frac{1}{6}$, *also* $\delta \le \frac{\varepsilon^5}{64 \log^2 |\mathcal{X}|}$ *and* $\delta \log^2 \frac{\varepsilon}{\delta} \le \varepsilon^5/16$.

2.1.2. COMPOSITIONAL LOCAL PRIVACY MECHANISMS

A different modification of Assumption A1 is to require the individual randomizations be private while imposing a summability (compositionality) condition. This limits the interaction between private communications, and there are problems for which fully interactive mechanisms are more powerful than compositional ones (Joseph et al., 2019). To be concrete, let $Z_{\to i}^{(t)} := (Z_{<i}^{(t)}, B^{(t-1)})$ be the "messages" coming into the channel generating $Z_i^{(t)}$, so $Z_i^{(t)} \sim Q_{i,t}(\cdot \mid X_i, Z_{\to i}^{(t)})$ as in Fig. 1. The starting point is a *1-compositional* privacy definition (Joseph et al., 2019), where in the (weakest) KL-privacy case we assume that there exists a function $\varepsilon_{i,t}$ such that

$$D_{\rm kl}\left(Q_{i,t}(\cdot \mid X_i = x, Z_{\to i}^{(t)} = z_{\to i}^{(t)}) \| Q_{i,t}(\cdot \mid X_i = x', Z_{\to i}^{(t)} = z_{\to i}^{(t)})\right) \le \varepsilon_{i,t}(z_{\to i}^{(t)}) \tag{3}$$

and the $\varepsilon_{i,t}$ satisfy $n^{-1} \sum_{i=1}^n \sum_{t=1}^T \mathbb{E}[\varepsilon_{i,t}(Z_{\to i}^{(t)}) \mid x_{\le n}] \le \varepsilon_{\rm kl}$. Condition (3) implies Assumption A1 by the chain rule for KL-divergence; in the case that $T = 1$, this captures the familiar *sequentially interactive* local privacy mechanisms (Duchi et al., 2018).

In some cases, we can provide stronger results for compositional (ε, δ)-private channels than for the fully interactive case, leading us to consider the following assumption, which allows privacy levels chosen conditionally on the past so long as the expected privacy levels remain non-trivial.

Assumption A2 (Compositional differential privacy bounds) *For each* i *and* t *and all* $z_{\to i}^{(t)}$, *the channel mapping* X_i *to* $Z_i^{(t)}$ *is* $(\varepsilon_{i,t}(z_{\to i}^{(t)}), \delta_{i,t}(z_{\to i}^{(t)}))$*-approximately differentially private. There exist* $\delta_{\rm total} \le \frac{1}{2}$ *and* $\varepsilon_{\rm kl}$ *such that*

$$\sum_{i=1}^n \sum_{t=1}^T \mathbb{E}[\delta_{i,t}(Z_{\to i}^{(t)})] \le \delta_{\rm total} \quad \text{and} \quad \sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left[\min\left\{\frac{\varepsilon_{i,t}^2(Z_{\to i}^{(t)})}{\log 2}, \varepsilon_{i,t}(Z_{\to i}^{(t)})\right\}\right] \le n \cdot \varepsilon_{\rm kl},$$

where the expectations are taken over the randomness in the private variables $\boldsymbol{Z}$.

In Assumption A2, individual i compromises at most $(\sum_t \varepsilon_{i,t}, \sum_t \delta_{i,t})$-differential privacy.

2.2. Minimax lower bounds on private estimation

Given our definitions of (interactive) privacy and the interactive privacy bounds in Assumptions A1, A1′, and A2, we may now describe the minimax framework in which we work. Let $\mathcal{P}$ be a collection of distributions on a space $\mathcal{X}$, and let $\theta(P) \in \Theta$ be a parameter of interest. In the classical (non-information-limited) setting, we wish to estimate $\theta(P)$ given observations X_i drawn i.i.d. according to the distribution P. We focus on d-dimensional parameters θ, and the performance of an estimator $\widehat{\theta} : \mathcal{X}^n \to \mathbb{R}^d$ is its expected loss (or risk) for a loss $L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$,

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

We elaborate this classical setting by an additional privacy layer. For a sample $\{X_1, \ldots, X_n\}$, any (interactive) channel Q produces a set of private observations, each from some set $\mathcal{Z}$,

$$\boldsymbol{Z} := \left(Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_n^{(1)}, Z_1^{(2)}, \ldots, Z_n^{(2)}, \ldots, Z_n^{(T)}\right) \in \mathcal{Z}^{T \times n},$$

and we consider estimators $\widehat{\theta}$ that depend only on this private sample, which then suffer risk

$$\mathbb{E}_{P,Q}\left[L(\widehat{\theta}(\boldsymbol{Z}), \theta(P))\right],$$

where the expectation is taken over the n i.i.d. observations $X_i \sim P$ and the privatized views $\boldsymbol{Z}$. For the channel Q, we define the *channel minimax risk* for the family $\mathcal{P}$, parameter θ, and loss L by

$$\mathfrak{M}_n(\theta(\mathcal{P}), L, Q) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[L\left(\widehat{\theta}(\boldsymbol{Z}), \theta(P)\right)\right]. \tag{4}$$

We prove lower bounds on the quantity (4) for channels satisfying local privacy bounds.

Rather than stating and proving our main theorems, we present a number of corollaries of our main results, all of whose proofs we defer to Appendix C, to illustrate the power of the information-based framework we adopt. Our first corollary deals with estimating Bernoulli means.

Corollary 3 *Let $\mathcal{P}_d$ be the collection of Bernoulli distributions on $\{0,1\}^d$ and for a symmetric loss $\ell : \mathbb{R} \to \mathbb{R}_+$ minimized at 0, let $L(\theta, \theta') = \sum_{j=1}^d \ell(\theta_j - \theta'_j)$. There are numerical constants $c_1, c_2, c_3 > 0$ such that for any channel Q satisfying any of Assumptions A1, A1′ with $\varepsilon_{\rm kl} := \min\{\varepsilon, \varepsilon^2\}$, or Assumption A2 with privacy budget $\varepsilon_{\rm kl}$,*

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), L, Q) \geq c_1 \cdot d \cdot \ell\left(\sqrt{c_2 \frac{d}{n\varepsilon_{\rm kl}}} \wedge c_3\right).$$

In particular, if $\ell(t) = t^2$ and if the private data releases of each individual are ε-locally differentially private (under any model of interaction), then inequality (1) and the corollary imply that for a constant $c > 0$, for any estimator $\widehat{\theta}$ there exists a Bernoulli distribution P with mean θ such that

$$\mathbb{E}_{P,Q}\left[\left\|\widehat{\theta}(\boldsymbol{Z}) - \theta\right\|_2^2\right] \geq c\left(\frac{d^2}{n \min\{\varepsilon, \varepsilon^2\}} \vee \frac{d}{n}\right).$$

A counterpart to the lower bound of Corollary 3 is that ε-differentially-private channels achieve this risk when $1 \leq \varepsilon \leq d$, and they require no interactivity. To within numerical constant factors, weakenings of local differential privacy—down to KL-privacy—provide no rate of convergence improvement over differentially private mechanisms. Bhowmick et al. (2018, Sec. 4.1) exhibit a mechanism ($\texttt{PrivUnit}_2$), based on sampling from spherical caps, that given x satisfying $\|x\|_2 \leq r$ samples ε-differentially private $Z \in \mathbb{R}^d$ satisfying $\mathbb{E}[Z \mid x] = x$ and $\|Z\|_2 \leq Cr\sqrt{d/\min\{\varepsilon, \varepsilon^2\}}$ for a numerical constant C. Taking the radius $r = \sqrt{d}$ the estimator $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ satisfies

$$\mathbb{E}[\|\widehat{\theta}_n - \theta\|_2^2] \leq \frac{1}{n}\mathbb{E}[\|Z_1\|_2^2] \leq C\frac{d^2}{n \min\{\varepsilon, \varepsilon^2\}}.$$

For the simpler case of KL-privacy, Gaussian noise addition suffices. We have thus characterized the complexity of locally private d-dimensional estimation of bounded vectors.

By a reduction, the lower bound of Corollary 3 applies to logistic regression. In this case, we let $\mathcal{P}_d$ be the collection of logistic distributions on $(X, Y) \in \{-1, 1\}^d \times \{\pm 1\}$, where for $\theta \in \mathbb{R}^d$, $P(Y = y \mid X = x) = 1/(1 + \exp(-y\langle x, \theta\rangle))$. We take the loss L as the gap in prediction risk: for

$$\ell(\theta; (x, y)) = \log(1 + \exp(-y\langle x, \theta\rangle)), \quad \text{we set} \quad R_P(\theta) := \mathbb{E}_P[\ell(\theta; (X, Y))]$$

and $\theta(P) = \operatorname{argmin}_\theta R_P(\theta)$. We define the excess risk $L(\theta, \theta(P)) = R_P(\theta) - R_P(\theta(P))$.

Corollary 4 *Let $\mathcal{P}_d$ be the family of logistic distributions and L be the excess logistic risk as above. There exists a numerical constant $c > 0$ such that for any sequence Q_n of channels satisfying any of Assumption A1, or A1′ with $\varepsilon_{\text{kl}} = \min\{\varepsilon, \varepsilon^2\}$, or Assumption A2, for all suitably large n we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), L, Q_n) \geq c \cdot \frac{d}{n} \cdot \frac{d}{\varepsilon_{\text{kl}}}.$$

It is also of interest to consider continuous distributions. For concreteness, we consider estimation of general and sparse Gaussian means, showing results that follow as corollaries of our information bounds and Braverman et al. (2016). We prove the lower bounds for channels satisfying Assumption A1 or A2; proving them under Assumption A1′ remains a challenge.[1]

Corollary 5 *Let $\mathcal{P}$ be the collection of Gaussian distributions $\mathsf{N}(\theta, \sigma^2 I)$ where $\theta \in [-1, 1]^d$, $\sigma^2 > 0$ is known, and consider the squared ℓ_2 loss $L(\theta, \theta') = \|\theta - \theta'\|_2^2$. There exist numerical constants $c, c_0 > 0$ such that if the channel Q satisfies Assumption A1 or A2 with $\delta_{\text{total}} \leq c_0$,*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2, Q) \geq c \cdot \min\left\{d, \max\left\{\frac{d}{\varepsilon_{\text{kl}}} \cdot \frac{d\sigma^2}{n}, \frac{d\sigma^2}{n}\right\}\right\}.$$

We demonstrate how to achieve this risk in Section 3.1, showing (as is the case for our other results) that it is achievable by differentially private schemes.

We can also state lower bounds for the sparse case, using Braverman et al. (2016, Theorem 4.5). Let $\mathcal{N}_{k,\sigma^2}^d$ denote the collection of k-sparse Gaussian distributions $\mathsf{N}(\theta, \sigma^2 I)$, $\theta \in [-1, 1]^d$.

Corollary 6 *There exist numerical constants $c, c_0 > 0$ such that for any channel Q satisfying Assumptions A1 or A2 with $\delta_{\text{total}} \leq c_0$, and $d \geq 2k$,*

$$\mathfrak{M}_n(\theta(\mathcal{N}_{k,\sigma^2}^d), \|\cdot\|_2^2, Q) \geq c \min\left\{k, \max\left\{\frac{d}{\varepsilon_{\text{kl}}} \cdot \frac{k\sigma^2}{n}, \frac{k\sigma^2 \log \frac{d}{k}}{n}\right\}\right\}.$$

3. Achievability, information complexity, independence, and correlation

The lower bounds in our corollaries are achievable—we demonstrate each of these here—but we highlight a more subtle question regarding correlation. Each of our lower bounds relies on the independence structure of the data: roughly, all the communication-based bounds we discuss require the coordinates of X to follow a product distribution. The lower bounds in this case are intuitive:

1. We use mutual information-based bounds, and on the (negligible) δ_{total}-probability event of a privacy failure under Assumption A1′, it is possible to release infinite information. For compositional channels satisfying Assumption A2, we show (see Lemma 15 in Sec. 4.2.2) that each channel is within $\delta_{i,t}$-variation distance to a differentially private ($\delta_{i,t} = 0$) channel, so lower bounds based on testing apply. The argument fails in the fully interactive setting, because the interaction may break the independence structure of the communication upon which our results rely.

we must estimate d-dimensional quantities using (on average) ε bits, so we expect penalties scaling as d/ε because one coordinate carries no information about the others. In cases where there is correlation, however, we might hope for more efficient estimation; we view this as a major open question in privacy and, more broadly, information-constrained estimators. To that end, we briefly show (Section 3.1) that each of our lower bounds in Corollaries 3–5 is achievable. After this, we mention asymptotic results for sparse estimation (Sec. 3.2) and correlated data problems (Sec. 3.3).

3.1. Achievability by differentially-private estimators

We first demonstrate that the results in each of our corollaries are achievable by ε-differentially private channels with limited interactivity. We have already done so for Corollary 3. For Corollary 4, Corollary 3.2 of Bhowmick et al. (2018) gives the achievability result. We provide the Gaussian results for the sake of completeness. (For the one dimensional case, see also Joseph et al. (2018).)

We begin by demonstrating a one-dimensional Gaussian estimator. Let $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$, where σ^2 is known and $\theta \in [-1, 1]$. Consider ε-differentially private version of X_i defined by

$$B_i := \sign(X_i) \text{ and } Z_i = \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \cdot \begin{cases} B_i & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon + 1} \\ -B_i & \text{otherwise.} \end{cases} \tag{5}$$

Then $\mathbb{E}[Z_i \mid X_i] = \sign(X_i)$, and for $\Phi(t) = \mathbb{P}(\mathsf{N}(0, 1) \le t)$ the standard Gaussian CDF, we have $\mathbb{E}_\theta[Z_i] = 1 - 2\Phi(-\theta/\sigma)$. Letting $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ be the average of the Z_i, the estimator defined by solving $\overline{Z}_n = 1 - 2\Phi(\widehat{\theta}_n/\sigma)$ is nearly unbiased. Projecting this quantity onto $[-1, 1]$ gives the estimator

$$\widehat{\theta}_n := \text{Proj}_{[-1,1]}\left(\sigma\Phi^{-1}\left(\frac{1 - \overline{Z}_n}{2}\right)\right). \tag{6}$$

This estimator satisfies the following, which we prove in Appendix D.3 via a Taylor expansion.

Lemma 7 *Let $\widehat{\theta}_n$ be the estimator (6) for the $\mathsf{N}(\theta, \sigma^2)$ location family, where $\sigma^2 > 0$ is at least a constant. Assume $|Z_i| \le b$ and $\mathbb{E}[Z_i] = 1 - 2\Phi(-\theta/\sigma)$. For numerical constants $0 < c \le C < \infty$,*

$$|\widehat{\theta}_n - \theta| \le C\sqrt{\frac{b^2\sigma^2 t}{n}} \text{ w.p. } \ge 1 - e^{-t} \text{ and } \mathbb{E}[|\widehat{\theta}_n - \theta|^2] \le C\frac{b^2\sigma^2}{n} + Ce^{-cn/b^2}.$$

To achieve an upper bound matching Corollary 5, consider the following non-interactive estimator, which provides ε of differential privacy. We consider the cases $\varepsilon \le 1$ and $\varepsilon \ge 1$ separately.

i. In the case that $\varepsilon \ge 1$, choose $\lfloor \varepsilon \rfloor \wedge d$ coordinates $j \in [d]$ uniformly at random. On each chosen coordinate j, release $Z_{i,j}$ via mechanism (5) using privacy level $\varepsilon_0 = 1$, and use the estimator (6) applied to each coordinate; this mechanism is ε-differentially private, each coordinate (when sampled) takes values $|Z_{i,j}| \le \frac{e+1}{e-1}$, and so the resulting vector $\widehat{\theta}_n \in \mathbb{R}^d$ satisfies

$$\mathbb{E}[\|\widehat{\theta}_n - \theta\|_2^2] \le \frac{Cd\sigma^2}{n((\lfloor\varepsilon\rfloor \wedge d)/d)} \le C\min\left\{\frac{d^2}{n\varepsilon}, \frac{d}{n}\right\}.$$

ii. When $\varepsilon < 1$, we use the ℓ_∞-based mechanism of Duchi et al. (2018) applied to the vector $\sign(X_i) \in \{-1, 1\}^d$, which then releases a vector $Z_i \in C\sqrt{d/\varepsilon^2} \cdot \{-1, 1\}^d$ for a numerical constant C chosen to guarantee $\mathbb{E}[Z \mid \sign(X)] = \sign(X)$. Thus each coordinate of Z_i satisfies the conditions of Lemma 7, and applying the inversion (6) to each coordinate independently yields $\mathbb{E}[\|\widehat{\theta}_n - \theta\|_2^2] \le \frac{Cd^2}{n\varepsilon^2}$. In this setting, the value $\varepsilon_{\text{kl}} \le 2\varepsilon^2$ by inequality (1).

3.2. Sparse Estimation

We now turn to settings in which the coordinates exhibit dependence, assuming individuals have $\varepsilon \le 1$-differential privacy to make the discussion concrete. Consider the sparse Gaussian mean problem, $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(\theta, I_d)$ for $\|\theta\|_0 = k$. For simplicity, let us consider that $k = 1$ and is known; Corollary 5 gives the minimax lower bound $d/(n\varepsilon^2)$ under ε-differential privacy, which Duchi et al. (2018, Sec. 4.2.2) achieve to within a logarithmic factor; the non-private minimax risk (Johnstone, 2013) is the exponentially smaller $\frac{\log d}{n}$. In the case of a (very) large sample size n, however, we observe a different phenomenon: the non-private and private rates coincide.

Let us assume that $n \gg d$, and that $n \to \infty$ as d remains fixed. Let the sample be of size $2n$, which we split. On the first half, we further split the sample into d bins of size n/d; for each of these d bins, we construct a 1-dimensional estimator of the mean of coordinate j via (6), which gives us d preliminary estimates $\widehat{\theta}_1^{\text{pre}}, \ldots, \widehat{\theta}_d^{\text{pre}}$, each of which is ε-locally differentially private. Lemma 7 shows that we can identify the non-zero coordinate of θ by $\widehat{j} := \operatorname{argmax}_j |\widehat{\theta}_j^{\text{pre}}|$ with exponentially high probability. Then, on the second half of the sample, we apply the private estimator (6) to estimate the mean of coordinate $\widehat{j}$. In combination, this yields an estimator $\widehat{\theta}_{2n}$ that achieves $\mathbb{E}[\|\widehat{\theta}_{2n} - \theta\|_2^2] \le C/(n\varepsilon^2)$ for large n, while the non-private analogue in this case has risk $1/n$.

We have moved from an *exponential* gap in the dimension to one that scales only as $1/\varepsilon^2$, as soon as n is large enough. This example is certainly stylized and relies on a particular flavor of asymptotics ($n \to \infty$); we believe this transformation from "independent" structure, with risk scaling as d/n, to an identified structure with risk scaling as $1/n$, merits more investigation.

3.3. Correlated Data

We consider an additional stylized example of correlation. Let $b \in \{\pm 1\}^d$ be a *known* bit vector and assume the data $X_i = b \cdot B_i$ where $B_i \in \{\pm 1\}$, $P(B_i = 1) = p$ for an unknown p. Without privacy, $\widehat{p} = \frac{1+\overline{B}_n}{2}$ achieves minimax optimal ℓ_2^2 risk $\frac{d}{n}$; the error is d times that for the one-dimensional quantity. In the private case, as $b \in \{\pm 1\}^d$ is known, the private channel for user i may privatize only the bit B_i using randomized response, setting Z_i as in Eq. (5). Using the private estimate $\widehat{p}_\varepsilon = \frac{1+\overline{Z}_n}{2}$ yields $\mathbb{E}[(\widehat{p}_\varepsilon - p)^2] \le C/(n \min\{\varepsilon^2, 1\})$, so $\widehat{\theta}_n = b(2\widehat{p}_\varepsilon - 1)$ has mean square error

$$\mathbb{E}\left[\|\widehat{\theta}_n - b \cdot (2p-1)\|_2^2\right] \le Cd \cdot \mathbb{E}[(\widehat{p}_\varepsilon - p)^2] \le C\frac{d}{n \min\{\varepsilon^2, 1\}}.$$

In contrast to the case with independent coordinates in Corollary 3, here the locally private estimator achieves (to within a factor of ε^{-2}) the same risk as the non-private estimator. This example is again special, but it suggests that leveraging correlation structures may close some of the substantial gaps between private and non-private estimation that prevent wider adoption of private estimators.

4. Lower bounds via information complexity

We turn to stating and proving our main minimax lower bounds, which build out of work by Zhang et al. (2013), Garg et al. (2014), and Braverman et al. (2016) on communication limits in estimation.

We begin with an extension of Assouad's method (Assouad, 1983; Yu, 1997), which transforms a d-dimensional estimation problem into one of testing d binary hypotheses, to information-limited settings. We consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ indexed by the hypercube $\mathcal{V} = \{-1, 1\}^d$,

where nature chooses $V \in \mathcal{V}$ uniformly at random. Conditional on $V = v$, we draw $\{X_i\}_{i=1}^n \stackrel{\text{iid}}{\sim} P_v$, from which we obtain the observed (privatized) $\boldsymbol{Z}$. Letting $\theta_v = \theta(P_v)$, we follow Duchi et al. (2018) and say that $\mathcal{V}$ induces a δ-Hamming separation if there exists $\mathsf{v} : \Theta \to \{-1, 1\}^d$ such that

$$L(\theta, \theta_v) \geq \delta \sum_{j=1}^d 1\left\{\mathsf{v}_j(\theta) \neq v_j\right\}. \tag{7}$$

Example 1 (Location families) *Let $\mathcal{P}$ be a family of distributions, each specified by a mean $\theta(P)$, and for each $v \in \{-1, 1\}^d$ set $\theta(P_v) = \delta \cdot v$ for some $\delta > 0$. Then for any symmetric $\ell : \mathbb{R} \to \mathbb{R}_+$ and loss of the form $L(\theta, \theta') = \sum_{j=1}^d \ell(\theta_j - \theta'_j)$, we have $L(\theta, \theta_v) \geq \ell(\delta) \sum_{j=1}^d 1\left\{\operatorname{sgn}(\theta_j) \neq v_j\right\}$.*

As our proof of Corollary 4 demonstrates, similar separations hold for convex risk minimization.

Letting $\mathbb{P}_{+j}$ and $\mathbb{P}_{-j}$ be the marginal distributions of the privatized $\boldsymbol{Z}$ conditional on $V_j = 1$ and $V_j = -1$, respectively, we have Assouad's method (Duchi et al. (2018, Lemma 1) gives this form):

Lemma 8 (Assouad's method) *Let the conditions of the previous paragraph hold and let $\mathcal{V}$ induce a δ-separation in Hamming metric. Then*

$$\mathfrak{M}_n(\theta(\mathcal{P}), L, Q) \geq \delta \sum_{j=1}^d \inf_{\widehat{V}} \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \neq V_j) = \frac{\delta}{2} \sum_{j=1}^d \left(1 - \|\mathbb{P}_{+j} - \mathbb{P}_{-j}\|_{\text{TV}}\right).$$

Consequently, if we can show that the total variation distance $\|\mathbb{P}_{+j} - \mathbb{P}_{-j}\|_{\text{TV}}$ is small while the δ-separation (7) is large for our family, we have shown a strong lower bound.

4.1. Strong data processing and information contraction

To prove lower bounds via Lemma 8, we build off of ideas that originate from Zhang et al. (2013), which Braverman et al. (2016) develop elegantly. Braverman et al. show how *strong data processing* inequalities, which quantify the information loss in classical information processing inequalities (Cover and Thomas, 2006), extend from one observation to multiple observations. They use this to prove lower bounds on the information complexity of distributed estimators, and we show how their results imply strong lower bounds on private estimation. We first provide a definition.

Definition 9 *Let $U \to X \to Z$ be a Markov chain, where U takes values $\{-1, 1\}$, and conditional on $U = u$ we draw $X \sim P_u$, then draw Z conditional on X. The* strong data processing constant *$\beta(P_{-1}, P_1)$ is the smallest $\beta \leq 1$ such that for all distributions $X \to Z$,*

$$I(U; Z) \leq \beta I(X; Z).$$

Many distributions satisfy strong data processing inequalities; Gaussians do (Braverman et al., 2016), as do distributions with bounded likelihood ratio dP_1/dP_{-1} (see Lemma 24 in Appendix C).

We consider families of distributions where the coordinates of X are independent, dovetailing with Assouad's method. For $v \in \{-1, 1\}^d$, conditional on $V = v$ we assume that

$$X \sim P_v = P_{v_1} \otimes P_{v_2} \otimes \cdots \otimes P_{v_d}, \tag{8}$$

a d-dimensional product distribution. That is, conditional on $V_j = v_j$, the coordinates $X_{i,j}$ are i.i.d. and independent of $V_{\setminus j} = (V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_d)$. When we have the generation strategy (8), we can use Garg et al. and Braverman et al.'s results to prove the following lower bound.

Theorem 10 *Let $V \in \{-1,1\}^d$ and consider the Markov chain $V \to X_{\le n} \to \boldsymbol{Z}$, where conditional on $V = v$ the X_i are i.i.d., follow the product distribution (8), and $\boldsymbol{Z}$ follows the protocol of Fig. 1. Assume that for each coordinate j, the chain $V_j \to X_{i,j}$ satisfies a strong data processing inequality with $\beta(P_{-1}, P_1) = \beta$, and $|\log \frac{dP_1}{dP_{-1}}| \le b$ for some $b < \infty$. Then for any estimator $\widehat{V}$,*

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \neq V_j) \ge \frac{d}{2}\left(1 - \sqrt{\frac{7(e^b+1)}{d}\beta \cdot I(X_{\le n}; \boldsymbol{Z} \mid V)}\right).$$

We defer the proof of Theorem 10 to Appendix A. Lemma 24 to come shows that if $|\log \frac{dP_1}{dP_{-1}}| \le b$, then $\beta(P_{-1}, P_1) \le 2(e^b - 1)^2$, often allowing easier application of the theorem.

By combining Theorem 10 with Lemma 8, we can prove strong lower bounds on minimax rates of convergence if we can both (i) provide a strong data processing constant for P_{-1} and P_1 and (ii) bound the mutual information $I(X_{\le n}; \boldsymbol{Z} \mid V)$. We do both presently, but we note that Theorem 10 relies strongly on the repeated communication structure in Figure 1 (as does Corollary 16, Braverman et al.'s Theorem 3.1 in the sequel). Similar techniques appear challenging in centralized settings. Key to our applications of the theorem, which rely on i.i.d. sampling of the vector $X_{\le n}$ to provide bounds on mutual information via privacy, is that Braverman et al.'s results allow us to take the information *conditional* on V; without this our results fail.

4.2. Information bounds

To apply Theorem 10, the first step is to develop information bounds on private communication. We present our three main lemmas that accomplish this, based on Assumptions A1, A1′, and A2 here. As in the development of our assumptions, we divide our information bounds into two cases, depending on whether we work in the fully interactive or compositional privacy setting.

4.2.1. INFORMATION BOUNDS FOR FULLY INTERACTIVE MECHANISMS

In this section, we provide the two bounds on mutual information bounds that give our results. Before stating them, however, we give the corollary to Theorem 10 that they immediately imply.

Corollary 11 *Let the conditions of Theorem 10 hold and assume additionally that the channels Q satisfy Assumption A1 or A1′, setting $\varepsilon_{\text{kl}} = \min\{9\varepsilon, 75\varepsilon^2\}$ in this case. Then*

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \neq V_j) \ge \frac{d}{2}\left(1 - \sqrt{\frac{7(e^b+1)}{d}\beta n \varepsilon_{\text{kl}}}\right).$$

The corollary is immediate from Lemmas 12 and 13 to come. We begin with the former, which extends McGregor et al. (2010, Prop. 7 or 4.3) and simplifies Feldman and Steinke (2018, Prop. 3.4).

Lemma 12 *Let the channel Q and transcript satisfy Assumption A1. Then for any Markov chain $V \to X_{\le n} \to \boldsymbol{Z}$, where the X_i are independent conditional on V, we have*

$$I(\boldsymbol{Z}; X_{\le n} \mid V) \le n \cdot \varepsilon_{\text{kl}}.$$

See Section B.1 for the proof.

In the more complicated (ε, δ)-differential privacy cases, we require more care. Because of lack of space, we must defer the argument to Appendix B.2, stating only the final conclusion here. The lynchpin of our argument is based on the development of Rogers et al. (2016), who develop mutual information bounds for discrete random variables under (ε, δ)-differential privacy.

Lemma 13 *Let the private variables $\boldsymbol{Z}$ satisfy Assumption A1′. Then*

$$I(X_{\le n}; \boldsymbol{Z} \mid V) \le n \min\left\{9\varepsilon, 75\varepsilon^2\right\}.$$

4.2.2. INFORMATION BOUNDS FOR COMPOSITIONAL MECHANISMS

The main result of the section, which follows by combining Theorem 10 with the lemmas to come, gives the following corollary.

Corollary 14 *Let the conditions of Theorem 10 hold and assume additionally that the channels Q satisfy Assumption A2. Then*

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \neq V_j) \ge \frac{d}{2}\left(1 - \sqrt{\frac{7(e^b+1)}{d}\beta n \varepsilon_{\rm kl}} - \delta_{\rm total}\right).$$

The corollary follows from Lemma 12 once we subtract $\delta_{\rm total}$ and use the following approximation guarantee, which shows that (ε, δ) channels are nearly differentially private.

Lemma 15 *Let Assumption A2 hold on the channel Q. Let $\mathbb{P}_{-1}$ and $\mathbb{P}_1$ be the marginal distributions of $\boldsymbol{Z}$ under the communication model of Fig. 1 with channel Q and base distributions P_{-1} and P_1 on $X_{\le n}$, so that $\mathbb{P}_v(S) = \int Q(S \mid x_{\le n}) dP_v(x_{\le n})$. For each i, t there exist channels $\overline{Q}(Z_i^{(t)} \in \cdot \mid x_i, z_{\to i}^{(t)})$ from X_i to $Z_i^{(t)}$, conditional on $z_{\to i}^{(t)}$, where each channel is $\varepsilon_{i,t}(z_{\to i}^{(t)})$-differentially private. The induced marginal distributions $\overline{\mathbb{P}}_{-1,1}$ under the channels $\overline{Q}$ satisfy*

$$\|\mathbb{P}_{-1} - \mathbb{P}_1\|_{\rm TV} \le \left\|\overline{\mathbb{P}}_{-1} - \overline{\mathbb{P}}_1\right\|_{\rm TV} + \delta_{\rm total}.$$

The most challenging part of Lemma 15 is to establish the existence of regular conditional probabilities $\overline{Q}$ (i.e., veryifying measurability) that are close to Q; we do so in Appendix D.1.

5. Conclusion

By building off of the results in information-limited statistical estimation that Zhang et al. (2013), Garg et al. (2014), and Braverman et al. (2016) establish, we have developed fundamental limits for locally private estimation at all privacy levels and for all the acceptable and common models of privacy. We do not believe this paper closes any doors, however: there is a substantial gap between the worst-case minimax bounds and asymptotic results, highlighted by the challenges of correlated data. Identifying structures we can leverage for more efficient private or information-constrained estimation—an analogue of the geometric theory available in the case of classical statistics, where Fisher information and related ideas play an essential role—presents a challenging direction that, we hope, may allow more frequent practical use of private procedures.

Acknowledgments We thank Vitaly Feldman, Aleksandar Nikolov, Aaron Roth, Adam Smith, and Salil Vadhan for clarifying discussions and feedback on an earlier version of this draft, which (among other things) led us to the general Definition 2 of local privacy. We also thank the Simons Institute for hosting our visit as part of the *Data Privacy: Foundations and Applications* semester.

References

Apple Differential Privacy Team. Learning with privacy at scale, 2017. Available at https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus des Séances de l'Académie des Sciences, Série I*, 296(23):1021–1024, 1983.

A. Beimel, K. Nissim, and E. Omri. Distributed private data analysis: Simultaneously solving how and what. In *Advances in Cryptology*, volume 5157 of *Lecture Notes in Computer Science*, pages 451–468. Springer, 2008.

A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv:1812.00984 [stat.ML]*, 2018.

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

M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, 2016. URL https://arxiv.org/abs/1506.07216.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, pages 635–658, 2016.

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

J. C. Duchi and F. Ruan. The right complexity measure in locally private estimation: It is not the Fisher information. *arXiv:1806.05756 [stat.TH]*, 2018.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation (with discussion). *Journal of the American Statistical Association*, 113(521):182–215, 2018.

C. Dwork and G. Rothblum. Concentrated differential privacy. *arXiv:1603.01887 [cs.DS]*, 2016.

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006a.

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

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

U. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS)*, 2014.

U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

V. Feldman and T. Steinke. Calibrating noise to variance in adaptive data analysis. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, 2018. URL http://arxiv.org/abs/1712.07196.

M. Gaboardi, R. Rogers, and O. Sheffet. Locally private mean estimation: Z-test and tight confidence intervals. *arXiv:1810.08054 [cs.DS]*, 2018.

A. Garg, T. Ma, and H. L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 28*, 2014.

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

J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

I. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. 2013.

M. Joseph, J. Kulkarni, J. Mao, and Z. S. Wu. Locally private gaussian estimation. *arXiv:1811.08382 [cs.LG]*, 2018.

M. Joseph, J. Mao, S. Neel, and A. Roth. The role of interactivity in local differential privacy. *arXiv:1904.03564 [cs.LG]*, 2019.

S. P. Kasiviswanathan and A. Smith. On the 'semantics' of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014. doi: 10.29012/jpc.v6i1.634. URL http://arxiv.org/abs/0803.3946v3.

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.

L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. The limits of two-party differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, pages 81–90. IEEE, 2010.

I. Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

R. M. Rogers, A. Roth, A. D. Smith, and O. Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *57th Annual Symposium on Foundations of Computer Science*, pages 487–494, 2016.

A. Rohde and L. Steinberger. Geometrizing rates of convergence under differential privacy constraints. *arXiv:1805.01422 [stat.ML]*, 2018.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

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

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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

M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018.

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

Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed estimation with communication constraints. In *Advances in Neural Information Processing Systems 27*, 2013.

Appendix A. Proof of Theorem 10

Our proofs build essentially directly out of the work of Garg et al. (2014) and Braverman et al. (2016). The starting point for all of these results is a due to Braverman et al. (2016), where we have carefully controlled the constants.

Corollary 16 (Braverman et al. (2016), Theorem 3.1) *Consider a Markov chain $U \to Y_{\le n} \to Z$, where $U \in \{\pm 1\}$ is uniform, and $Y_i \stackrel{\text{iid}}{\sim} P_u$ conditional on $U = u$. Assume that $|\log \frac{dP_1}{dP_{-1}}| \le b$ and that the strong data processing inequality constant of P_1, P_{-1} is $\beta(P_{-1}, P_1)$. Let M_1 and M_{-1} denote the marginal distributions on Z conditional on $U = 1$ or -1, respectively. Then*

$$d_{\text{hel}}^2(M_{-1}, M_1) \le \frac{7}{2}(e^b + 1)\beta(P_{-1}, P_1) \sum_{i=1}^n \min\{I(Y_i; Z \mid U = -1), I(Y_i; Z \mid U = 1)\}.$$

The Y_i are i.i.d. conditional on U in the corollary, so as an immediate consequence, we have

$$d_{\text{hel}}^2(M_{-1}, M_1) \le \frac{7}{2}(e^b + 1)\beta(P_0, P_1) \min\{I(Y_{\le n}; Z \mid U = -1), I(Y_{\le n}; Z \mid U = 1)\} \tag{9}$$

for any $U \to Y_{\le n} \to Z$ when the Y_i are conditionally independent given U. To see this, note that

$$\begin{aligned} I(Y_{\le n}; Z \mid U = u) &= \sum_{i=1}^n H(Y_i \mid Y_{<i}, U = u) - H(Y_i \mid Y_{<i}, Z, U = u) \\ &\ge \sum_{i=1}^n H(Y_i \mid U = u) - H(Y_i \mid Z, U = u) = \sum_{i=1}^n I(Y_i; Z \mid U = u), \end{aligned}$$

where we use $H(Y_i \mid Y_{<i}, U = u) = H(Y_i \mid U = u)$ and that conditioning reduces entropy.

The key in Theorem 10, which uses the chain $V \to X_{\le n} \to \boldsymbol{Z}$, is (as in the case of Garg et al. (2014) and Braverman et al. (2016)) that each individual i draws coordinate j in $X_{i,j}$ conditional on only coordinate V_j of $V \in \{-1, 1\}^d$, that is, *independently* of $V_{\setminus j}$. Now, let $X_{\le n, j} = (X_{i,j})_{i=1}^n$ be the jth coordinate of the data, and let $X_{\le n, \setminus j}$ denote the remaining $d - 1$ coordinates across all $i = 1, \ldots, n$. By construction of our product sampling distribution (8), we thus have Markov structure

$$V_j \to X_{\le n, j} \to \boldsymbol{Z} \leftarrow X_{\le n, \setminus j} \leftarrow V_{\setminus j},$$

in turn implying (by marginalizing over $X_{\le n, \setminus j}$ and $V_{\setminus j}$) the Markov structure

$$V_j \to X_{\le n, j} \to \boldsymbol{Z}. \tag{10}$$

Now, define $M_{\pm j}$ to be the marginal distributions over the total communicated private variables $\boldsymbol{Z}$ conditional on $V_j = \pm 1$. Then Le Cam's inequalities and Cauchy-Schwarz imply that

$$\begin{aligned} 2\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \ne V_j) \ge \sum_{j=1}^d (1 - \|M_{-j} - M_{+j}\|_{\text{TV}}) &\ge \sum_{j=1}^d (1 - \sqrt{2} d_{\text{hel}}(M_{-j}, M_{+j})) \\ &\ge d\left(1 - \sqrt{\frac{2}{d}\sum_{j=1}^d d_{\text{hel}}^2(M_{-j}, M_{+j})}\right). \end{aligned} \tag{11}$$

It remains to bound the summed Hellinger distances. By inequality (9) and the particular Markov structure (10), we have

$$d_{\mathrm{hel}}^2(M_{-j}, M_{+j}) \le \frac{7}{2}(e^b + 1)\beta(P_0, P_1) I(X_{\le n,j}; \boldsymbol{Z} \mid V_j). \tag{12}$$

Using the fact that conditioning reduces entropy and that conditional on V_j, the values $X_{\le n,j}$ are i.i.d. and independent of $V_{\setminus j}$, we have

$$\begin{aligned} I(X_{\le n,j}; \boldsymbol{Z} \mid V_j) &= H(X_{\le n,j} \mid V_j) - H(X_{\le n,j} \mid V_j, \boldsymbol{Z}) \\ &\le H(X_{\le n,j} \mid V_j, V_{\setminus j}) - H(X_{\le n,j} \mid V_j, V_{\setminus j}, \boldsymbol{Z}) \\ &= I(X_{\le n,j}; \boldsymbol{Z} \mid V). \end{aligned}$$

The following lemma relates the individual informations to the global information $I(X_{\le n}; \boldsymbol{Z} \mid V)$.

Lemma 17 *Let $V, X_{\le n}, \boldsymbol{Z}$ be as in Theorem 10. Then*

$$\sum_{j=1}^d I(X_{\le n,j}; \boldsymbol{Z} \mid V) \le I(X_{\le n}; \boldsymbol{Z} \mid V).$$

Proof We have

$$\begin{aligned} \sum_{j=1}^d I(X_{\le n,j}; \boldsymbol{Z} \mid V) &= \sum_{j=1}^d [H(X_{\le n,j} \mid V) - H(X_{\le n,j} \mid \boldsymbol{Z}, V)] \\ &\overset{(i)}{=} H(X_{\le n} \mid V) - \sum_{j=1}^d H(X_{\le n,j} \mid \boldsymbol{Z}, V) \\ &\overset{(ii)}{\le} H(X_{\le n} \mid V) - \sum_{j=1}^d H(X_{\le n,j} \mid X_{\le n,<j}, \boldsymbol{Z}, V) = I(X_{\le n}; \boldsymbol{Z} \mid V), \end{aligned}$$

where the equality (i) follows because conditional on V, the coordinates $X_{\le n,j}$ are independent, and inequality (ii) because conditioning reduces entropy. ■

Substituting the bound of Lemma 17 via the consequence (12) of the strong data processing inequality (9) into inequality (11), we have

$$2\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \ne V_j) \ge d\left(1 - \sqrt{7(e^b + 1)\beta I(X_{\le n}; \boldsymbol{Z} \mid V)/d}\right).$$

This is the desired result.

Appendix B. Proofs of mutual information bounds

B.1. Proof of Lemma 12

We have

$$I(\boldsymbol{Z}; X_{\le n} \mid V) = \sum_{i=1}^{n} I(\boldsymbol{Z}; X_i \mid X_{<i}, V)$$
$$= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[D_{\rm kl}\left(Q(\boldsymbol{Z} \in \cdot \mid X_{\le i}, V) \| Q(\boldsymbol{Z} \in \cdot \mid X_{<i}, V)\right) \mid V\right]\right] \tag{13}$$

where the first equality is the chain rule and the second equality uses the equivalence of mutual information and expected KL-divergence, where $Q(\boldsymbol{Z} \in \cdot \mid X_{\le i}, V)$ denotes the conditional distribution of the full set of private variables $\boldsymbol{Z}$ given $X_{\le i}$. Now we note that

$$Q(\boldsymbol{Z} \in \cdot \mid x_{\le i}, v) = \int Q(\boldsymbol{Z} \in \cdot \mid x_{\le n}) dP_v(x_{i+1}) \cdots dP_v(x_n)$$

because the X_i are independent conditional on $V = v$, and similarly for $Q(\boldsymbol{Z} \in \cdot \mid x_{<i}, v)$. The joint convexity of the KL-divergence then implies

$$D_{\rm kl}\left(Q(\boldsymbol{Z} \in \cdot \mid x_{\le i}, v) \| Q(\boldsymbol{Z} \in \cdot \mid x_{<i}, v)\right)$$
$$\le \int_{\mathcal{X}^{n-i}} \int_{\mathcal{X}} \underbrace{D_{\rm kl}\left(Q(\boldsymbol{Z} \in \cdot \mid x_{\le n}) \| Q(\boldsymbol{Z} \in \cdot \mid x_{<i}, x_i', x_{>i})\right)}_{=:\varepsilon_i(x_{\le n}, x_i')} dP_v(x_{i+1}) \cdots dP_v(x_n) dP_v(x_i')$$

where we let ε_i be as above. Assumption A1 gives that $\sum_{i=1}^n \varepsilon_i(x_{\le n}, x_i') \le n\varepsilon_{\rm kl}$, and substituting in the chain rule (13) gives the result.

B.2. Proof of Lemma 13

The result actually follows from two more sophisticated lemmas, which we state here and proof subsequently (see Section B.3).

Lemma 18 *Let X_i be i.i.d. and Z be (ε, δ)-differentially private for $X_{\le n}$, where each X_i takes values in the finite set $\mathcal{X}$. Let $\eta > 0$ and define $p_\eta = 2(\frac{\delta}{\eta} + \eta\frac{e^{3\varepsilon}}{e^{3\varepsilon}-1} + \frac{\delta e^{\varepsilon}}{e^{\varepsilon}-1})$ and the binary entropy $h_2(p) = -p\log p - (1-p)\log(1-p)$. If $p_\eta \le 1$, then*

$$I(X_{\le n}; Z) \le n \cdot [6\varepsilon + p_\eta \log|\mathcal{X}| + h_2(p_\eta)]$$

Additionally, if $\eta > 0$ is small enough that $\eta(2e^{6\varepsilon}/(e^{3\varepsilon}-1)+1) \le \frac{1}{2}$, then

$$I(X_{\le n}; Z) \le n \cdot \left(6\varepsilon(e^{6\varepsilon}-1) + 3\eta\left[e^{3\varepsilon} + 3\eta\frac{e^{12\varepsilon}}{(e^{3\varepsilon}-1)^2}\right] + p_\eta \log|\mathcal{X}| + h_2(p_\eta)\right).$$

Extending this lemma for particular δ allows us to provide more intepretable results.

Lemma 19 *In addition to the conditions of Lemma 18, assume that* $16\sqrt{\delta \max\{\varepsilon^{-1}, 1\}} \le 1$, $\delta \max\{\varepsilon^{-1}, 1\} \log \frac{1}{\delta \max\{\varepsilon^{-1}, 1\}} \le \varepsilon^2$, *and* $\delta \max\{\varepsilon^{-1}, 1\} \log^2 |\mathcal{X}| \le \varepsilon^2$. *Then*

$$I(X_{\le n}; Z) \le 9n\varepsilon.$$

If $\varepsilon \le 1/6$ *and we additionally have* $\delta \le \frac{\varepsilon^5}{64 \log^2 |\mathcal{X}|}$ *and* $\delta \log^2 \frac{\varepsilon}{\delta} \le \varepsilon^5/16$, *then*

$$I(X_{\le n}; Z) \le 75n\varepsilon^2.$$

The proof is mostly algebraic manipulations; see Section B.4.

By recalling that in our packing of the hypercube, the Markov chain $V \to X_{\le n} \to \boldsymbol{Z}$ guarantees that the X_i are i.i.d. conditional on V, Lemma 19 implies Lemma 13 immediately.

B.3. Proof of Lemma 18

In this section, we provide the proof of Lemma 18. We require a number of different claims. First, we assume w.l.o.g. that all random variables of interest are discrete and finitely supported (as we note earlier, the mutual information $I(X;Y)$ is arbitrarily approximated by finite partitions of the ranges of X and Y Gray (1990)). We make a few definitions and give examples.

Definition 20 *Let* X, Y *be arbitrary random variables. They are* (ε, δ)-indistinguishable, *which we denote* $X \approx_{\varepsilon,\delta} Y$, *if the set* $E := \{x : |\log \frac{P(X=x)}{P(Y=x)}| \le \delta\}$ *satisfies* $P(Y \notin E) \ge 1 - \delta$ *and* $P(X \notin E) \le \delta$.

With this definition, we introduce a few notational shorthands for ease of use later. Let Z be the random variable distributed as $Q_Z(\cdot \mid X)$ (i.e. conditional on X), and we let $X|_{Z=z}$ be the random variable X conditional on $Z = z$, that is, the posterior on X given $Z = z$. With this, we can follow Rogers et al. (2016) and their development of mutual information bounds based on approximate differential privacy. The key is to bound the sequence of *privacy loss* random variables,

$$\ell_i^{\mathsf{pr}}(x_{\le i}, z) := \log \frac{P(X_i = x \mid Z = z, X_{<i} = x_{<i})}{P(X_i = x)},$$

as the mutual information between discrete variables $(X_{\le n}, Z)$ where the X_i are i.i.d. is

$$I(X_{\le n}; Z) = \sum_{i=1}^n I(X_i; Z \mid X_{<i}) = \sum_{i=1}^n \mathbb{E}\left[\ell_i^{\mathsf{pr}}(X_{\le i}, Z)\right]. \tag{14}$$

We begin with two of Rogers et al.'s claims, which in turn build off of Kasiviswanathan and Smith (2014). For $\delta > 0$ and $i \in [n]$, define the sets

$$\begin{aligned}
E_i(\delta) &:= \left\{(x_{<i}, z) \in \mathcal{X}^{i-1} \times \mathcal{Z} : X_i \approx_{3\varepsilon,\delta} X_i|_{Z=z, X_{<i}=x_{<i}}\right\} \\
F_i &:= \left\{(x_{\le i}, z) \in \mathcal{X} \times \mathcal{Z} : |\ell_i^{\mathsf{pr}}(x_{\le i}, z)| \le 6\varepsilon\right\} \\
G_i(\delta) &:= \left\{(x_{\le i}, z) \in \mathcal{X}^i \times \mathcal{Z} : (x_{<i}, z) \in E_i(\delta), (x_{\le i}, z) \in F_i\right\},
\end{aligned}$$

so that G is essentially the "good" set where the pair (X, Z) behaves as though Z is ε-differentially private.

We then have

Lemma 21 (Rogers et al. (2016), Claims 3.4–3.6) *Let the channel $Q_Z(\cdot \mid X)$ be (ε, δ)-differentially private. Then for any $\eta > 0$ and $z \in E(\eta)$,*

$$\mathbb{P}\left((X_{<i}, Z) \in E_i(\eta) \mid X_{<i} = x_{<i}\right) \geq 1 - \frac{2\delta}{\eta} - \frac{2\delta e^{\varepsilon}}{e^{\varepsilon} - 1} \tag{15a}$$

$$\mathbb{P}\left((X_{\leq i}, Z) \in F_i \mid Z = z, X_{<i} = x_{<i}\right) \geq 1 - \frac{2\eta e^{3\varepsilon}}{e^{3\varepsilon} - 1} \tag{15b}$$

$$\mathbb{P}\left((X_{\leq i}, Z) \in G_i(\eta)\right) \geq 1 - \frac{2\delta}{\eta} - \frac{2\delta e^{\varepsilon}}{e^{\varepsilon} - 1} - \frac{2\eta e^{3\varepsilon}}{e^{3\varepsilon} - 1}. \tag{15c}$$

With Lemma 21, we can bound the mutual information between X and Z. We begin by decomposing the mutual information into two sums, as for any $\eta > 0$,

$$\begin{aligned} &I(X_i; Z \mid X_{<i}) \\ &= \mathbb{E}[\ell_i^{\mathsf{pr}}(X_{\leq i}, Z) 1\left\{(X_{\leq i}, Z) \in G_i(\eta)\right\}] + \mathbb{E}[\ell_i^{\mathsf{pr}}(X_{\leq i}, Z) 1\left\{(X_{\leq i}, Z) \notin G_i(\eta)\right\}]. \end{aligned} \tag{16}$$

We control each of the terms in turn.

Lemma 22 *Let $\eta > 0$, and define the shorthands $P(G_\eta^c) = P((X_{\leq i}, Z) \notin G_i(\eta))$ and $h_2(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$. Then*

$$\mathbb{E}[\ell_i^{\mathsf{pr}}(X_{\leq i}, Z) 1\left\{(X_{\leq i}, Z) \notin G_i(\eta)\right\}] \leq P(G_\eta^c) \log |\mathcal{X}| + h_2(P(G_\eta^c)).$$

Proof For shorthand, let $G \equiv G_i(\eta)$. Let $X' = X_i|_{(X_{\leq i}, Z) \notin G}$ and $Z' = Z|_{(X_{\leq i}, Z) \notin G}$. Then as $\mathcal{X}$ is finite, we have

$$\begin{aligned} \log |\mathcal{X}| &\geq H(X_i' \mid X_{<i} = x_{<i}) \geq I(X_i'; Z_i' \mid X_{<i} = x_{<i}) \\ &= \sum_{x,z} \frac{P(X_i = x, Z = z, G^c \mid x_{<i})}{P(G^c \mid x_{<i})} \log \frac{P(X_i = x, Z = z \mid x_{<i}) P(G^c \mid x_{<i})}{P(X_i = x, G^c \mid x_{<i}) P(Z = z, G^c \mid x_{<i})} \\ &\geq \sum_{x,z} \frac{P(X_i = x, Z = z, G^c \mid x_{<i})}{P(G^c \mid x_{<i})} \log \frac{P(X_i = x, Z = z \mid x_{<i}) P(G^c \mid x_{<i})}{P(X_i = x \mid x_{<i}) P(Z = z \mid x_{<i})} \\ &= \sum_{x,z} \frac{P(X = x, Z = z, G^c \mid x_{<i})}{P(G^c \mid x_{<i})} \left[\ell_i^{\mathsf{pr}}(x_{\leq i}, z) + \log P(G^c \mid x_{<i})\right] \end{aligned}$$

Rearranging gives that

$$\begin{aligned} \mathbb{E}\left[\ell_i^{\mathsf{pr}}(X_{\leq i}, Z) 1\left\{(X_{\leq i}, Z) \notin G\right\} \mid X_{<i} = x_{<i}\right] &\leq P(G^c \mid x_{<i}) \left[\log |\mathcal{X}| + \log \frac{1}{P(G^c \mid x_{<i})}\right] \\ &\leq P(G^c \mid x_{<i}) \log |\mathcal{X}| + H(1\left\{G\right\} \mid X_{<i} = x_{<i}). \end{aligned}$$

Integrating over the marginal of $X_{<i}$ and noting that conditioning always reduces entropy, we obtain

$$\mathbb{E}\left[\ell_i^{\mathsf{pr}}(X_{\leq i}, Z) 1\left\{(X_{\leq i}, Z) \notin G\right\}\right] \leq P(G^c \mid x_{<i}) \log |\mathcal{X}| + H(1\left\{G\right\})$$

as desired. ∎

We now turn to the first term in the expansion (16). We always have $\ell_i^{\mathsf{pr}}(X, Z) \le 6\varepsilon$ on the event $G(\eta)$, so that Lemma 22, coupled with the chain rule (14) and probability bound (15c) gives

$$I(X_{\le n}; Z) \le \sum_{i=1}^{n} (6\varepsilon + p_\eta \log |\mathcal{X}| + h_2(p_\eta))$$

for $p_\eta = \frac{2\delta}{\eta} + \frac{2\delta e^\varepsilon}{e^\varepsilon - 1} + \frac{2\eta e^{3\varepsilon}}{e^{3\varepsilon}-1}$. This is evidently the first claim of Lemma 18.

To see the second claim requires a bit more work, though the next lemma suffices.

Lemma 23 *Let $\eta > 0$ be small enough that $\eta(2e^{6\varepsilon}/(e^{3\varepsilon} - 1) + 1) \le \frac{1}{2}$. Then*

$$\mathbb{E}[\ell_i^{\mathsf{pr}}(X_{\le i}, Z) 1\{(X, Z) \in G(\eta)\}] \le 6\varepsilon(e^{6\varepsilon} - 1) + 3\eta \left[e^{3\varepsilon} + 3\eta \frac{e^{12\varepsilon}}{(e^{3\varepsilon} - 1)^2}\right].$$

Proof Let $(x_{<i}, z) \in E(\eta)$. Then

$$\begin{aligned}
&\mathbb{E}[\ell_i^{\mathsf{pr}}(X_{\le i}, Z) 1\{(X, Z) \in G(\eta)\} \mid Z = z, X_{<i} = x_{<i}] \\
&= \sum_{x_i : (x_{\le i}, z) \in F_i} \ell_i^{\mathsf{pr}}(x_{\le i}, z) P(X_i = x_i \mid Z = z, x_{<i}) \\
&= \sum_{x_i : (x_{\le i}, z) \in F_i} \ell_i^{\mathsf{pr}}(x_{\le i}, z)[P(X_i = x_i \mid Z = z, x_{<i}) - P(X_i = x_i)] + \sum_{x_i : (x_{\le i}, z) \in F_i} \ell_i^{\mathsf{pr}}(x_{\le i}, z) P(X_i = x_i) \\
&\le 6\varepsilon(e^{6\varepsilon} - 1) + \sum_{x_i : (x_{\le i}, z) \in F_i} \ell_i^{\mathsf{pr}}(x_{\le i}, z) P(X_i = x_i) \qquad (17)
\end{aligned}$$

where we have used that

$$|P(X_i = x_i \mid Z = z, x_{<i}) - P(X_i = x_i)| \le e^{6\varepsilon} - 1$$

by definition of the set F_i and that $(x_{\le i}, z) \in F_i$, and that similarly $|\ell_i^{\mathsf{pr}}(x_{\le i}, z)| \le 6\varepsilon$.

To bound the second term in the sum (17), we note that

$$\begin{aligned}
&\sum_{x_i : (x_{\le i}, z) \in F_i} \ell_i^{\mathsf{pr}}(x_{\le i}, z) P(X_i = x_i) \\
&= P((X_{\le i}, z) \in F_i \mid x_{<i}) \sum_{x_i : (x_{\le i}, z) \in F_i} \ell_i^{\mathsf{pr}}(x_{\le i}, z) \frac{P(X = x)}{P((X_{\le i}, z) \in F_i \mid x_{<i})} \\
&\le P((X_{\le i}, z) \in F_i \mid x_{<i}) \log \frac{P((X_{\le i}, Z) \in F_i \mid x_{<i}, Z = z)}{P((X_{\le i}, z) \in F_i \mid x_{<i})} \\
&= P((X_{\le i}, z) \in F_i \mid x_{<i}) \log \frac{1 - P((X_{\le i}, Z) \notin F_i \mid x_{<i}, Z = z)}{1 - P((X_{\le i}, z) \notin F_i \mid x_{<i})}
\end{aligned}$$

by Jensen's inequality. Let us bound the logarithmic terms. As $(x_{<i}, z) \in E_i(\eta)$ by assumption, we have $P((X_{\le i}, z) \notin F_i \mid x_{<i}) \le e^{3\varepsilon} P((X_{\le i}, Z) \notin F_i \mid Z = z, x_{<i}) + \eta$. Letting $q = P((X_{\le i}, Z) \notin F_i \mid Z = z, x_{<i})$ for shorthand, Lemma 21 (Eq. (15b)) implies that $q \le \frac{2\eta e^{3\varepsilon}}{e^{3\varepsilon}-1}$, and thus

$$\log \frac{1 - P((X_{\le i}, Z) \notin F_i \mid Z = z, x_{<i})}{1 - P((X_{\le i}, z) \notin F_i \mid x_{<i})} \le \log \frac{1 - q}{1 - e^{3\varepsilon} q - \eta} \le (e^{3\varepsilon} - 1)q + \eta + (e^{3\varepsilon} q + \eta)^2,$$

where we have used that $-\log(1-t) \le t + t^2$ for $t \le \frac{1}{2}$ and the assumption that $e^{3\varepsilon}q + \eta < \frac{1}{2}$. Returning to our bounds on the sum (17), we see that

$$\mathbb{E}[\ell_i^{\mathsf{pr}}(X_{\le i}, Z)1\left\{(X,Z) \in G(\eta)\right\} \mid z, x_{<i}] \le 6\varepsilon(e^{6\varepsilon} - 1) + \eta\left[2e^{3\varepsilon} + 1 + \eta\left(\frac{2e^{6\varepsilon}}{e^{3\varepsilon} - 1} + 1\right)^2\right].$$

Noting that $e^{6\varepsilon}/(e^{3\varepsilon} - 1) > 5/4$ gives the result. ■

B.4. Proof of Lemma 19

We begin by addressing the exponential in ε terms, which will allow easier derivation. For all $\varepsilon \ge 0$, we have

$$\frac{e^{3\varepsilon}}{e^{3\varepsilon} - 1} \le \max\left\{\frac{1}{\varepsilon}, \frac{e}{e-1}\right\}, \quad \frac{e^{\varepsilon}}{e^{\varepsilon} - 1} \le \max\left\{\frac{2}{\varepsilon}, \frac{e}{e-1}\right\}, \tag{18a}$$

and for $\varepsilon \le \frac{1}{6}$,

$$\frac{e^{6\varepsilon}}{e^{3\varepsilon} - 1} \le \frac{3}{4\varepsilon}, \quad \frac{e^{12\varepsilon}}{(e^{3\varepsilon} - 1)^2} \le \frac{1}{2\varepsilon^2}, \quad 6\varepsilon(e^{6\varepsilon} - 1) \le 62\varepsilon^2. \tag{18b}$$

Using the bounds (18), we can provide our desired mutual information bounds. In the case that $\varepsilon \ge 0$ is arbitrary, we use the first bound of Lemma 18. In this case, to apply the bound it is sufficient that $p_\eta \le 2(\frac{\delta}{\eta} + \eta\max\{\varepsilon^{-1}, 2\} + \delta\max\{2\varepsilon^{-1}, 2\}) \le \frac{1}{2}$, and taking $\eta = \sqrt{\delta\min\{\varepsilon, 1/2\}}$ gives that

$$p_\eta \le 4\sqrt{\delta\max\{\varepsilon^{-1}, 2\}} + 2\delta\max\{\varepsilon^{-1}, 1\} \le 8\sqrt{\delta\max\{\varepsilon^{-1}, 1\}} \le \frac{1}{2}$$

whenever $\sqrt{\delta\max\{\varepsilon^{-1}, 1\}} \le 1/16$. Assuming additionally that $\sqrt{\delta\max\{\varepsilon^{-1}, 1\}}\log\frac{1}{\delta\max\{\varepsilon^{-1},1\}} \le \varepsilon$ and $\sqrt{\delta\max\{\varepsilon^{-1}, 1\}}\log|\mathcal{X}| \le \varepsilon$ gives the first claimed result as $h_2(p) \le -2p\log p$ for $p \le \frac{1}{2}$.

For the second result, under the additional condition that $\varepsilon \le 1/6$, our chosen $\eta = \sqrt{\delta\min\{\varepsilon, 1/2\}} = \sqrt{\delta\varepsilon}$ satisfies $\eta(\frac{3}{2\varepsilon} + 1) \le \frac{1}{2}$ (as $\delta \le 1/(64\varepsilon)$). When $\delta \le \varepsilon^3$, we have

$$3\eta\left[e^{3\varepsilon} + 3\eta\frac{e^{12\varepsilon}}{(e^{3\varepsilon} - 1)^2}\right] \le 3\sqrt{\delta\varepsilon}\left(2 + \frac{3}{2}\sqrt{\frac{\delta}{\varepsilon^3}}\right) \le 11\varepsilon^2$$

by inequalities (18). Finally, in this case we again have $p_\eta \le 8\sqrt{\delta/\varepsilon}$, and so if

$$\delta \le \frac{\varepsilon^5}{64\log^2|\mathcal{X}|} \text{ and } \delta\log^2\frac{\varepsilon}{\delta} \le \frac{\varepsilon^5}{16}$$

then $p_\eta\log|\mathcal{X}| + h_2(p_\eta) \le 2\varepsilon^2$. These bounds and Lemma 18 give the second result.

Appendix C. Proofs of Corollaries

Before proving the corollaries from Section 2.2, we present one lemma that will be useful throughout. It is similar to, but simpler than, a result of Zhang et al. (2013, Lemma 8).

Lemma 24 *Let $V \to X \to Z$, where $X \sim P_v$ conditional on $V = v$. If $|\log \frac{dP_v}{dP_{v'}}| \le \alpha$ for all v, v', then*

$$I(V;Z) \le 4(e^\alpha - 1)^2 \mathbb{E}_Z[\|P_X(\cdot \mid Z) - P_X\|_{\mathrm{TV}}^2] \le 2(e^\alpha - 1)^2 I(X;Z).$$

Proof By approximation, there is no loss of generality to assume that each random variable is discrete (Gray, 1990), so that our variables may have probability mass functions, which we denote by p. We first claim that

$$|p(v \mid z) - p(v)| \le 2(e^\alpha - 1)p(v) \left\|P_X(\cdot \mid z) - P_X(\cdot)\right\|_{\mathrm{TV}}. \tag{19}$$

Indeed, we have that $p(v \mid x) = p(x \mid v)p(v)/p(x) \in [e^{-\alpha}, e^\alpha]p(v)$ by assumption on $dP_v/dP_{v'}$. Thus, the Markov structure $V \to X \to Z$ implies

$$\begin{aligned}
|p(v \mid z) - p(v)| &= \left|\sum_x p(v \mid x)p(x \mid z) - p(v \mid x)p(x)\right| \\
&= \left|\sum_x (p(v \mid x) - p(v))(p(x \mid z) - p(x))\right| \\
&\le |e^\alpha - 1| p(v) \sum_x |p(x \mid z) - p(x)| = 2(e^\alpha - 1)p(v) \left\|P_X(\cdot \mid z) - P_X\right\|_{\mathrm{TV}}.
\end{aligned}$$

Then using the definition of mutual information and that χ^2-divergence upper bounds the KL-divergence (Tsybakov, 2009, Lemma 2.7),

$$\begin{aligned}
I(V;Z) &= \mathbb{E}_Z[D_{\mathrm{kl}}\left(P_V(\cdot \mid Z) \| P_V\right)] \\
&\le \mathbb{E}_Z\left[\sum_v \left(\frac{p(v \mid Z) - p(v)}{p(v)}\right)^2 p(v)\right] \le 4(e^\alpha - 1)^2 \mathbb{E}_Z\left[\sum_v p(v) \left\|P_X(\cdot \mid Z) - P_X\right\|_{\mathrm{TV}}^2\right],
\end{aligned}$$

where the second inequality used inequality (19). By Pinsker's inequality, we have the bound $\|P_X(\cdot \mid Z) - P_X\|_{\mathrm{TV}}^2 \le \frac{1}{2} D_{\mathrm{kl}}\left(P_X(\cdot \mid Z) \| P_X\right)$, and using that $I(Z;X) = \mathbb{E}_Z[D_{\mathrm{kl}}\left(P_X(\cdot \mid Z) \| P_X\right)]$ gives the lemma. ■

C.1. Proof of Corollary 3

By Corollaries 11 and 14, it will be sufficient to provide a good enough strong data processing inequality for Bernoulli random variables. We give the proof under Assumption A2 (which relies on Corollary 14), as the other cases are completely similar. Let $P_{-1} = \mathsf{Bernoulli}(\frac{1}{2})$ and, for some $\delta < 1$, let $P_1 = \mathsf{Bernoulli}(\frac{1+\delta}{2})$. Then $|\log dP_1/dP_{-1}| \le -\log(1-\delta)$, and consequently, for V uniform on $\{-1, 1\}$, we obtain

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

In particular, we have $\beta(P_{-1}, P_1) \le \frac{2\delta^2}{(1-\delta)^2}$, and in the notation of Theorem 10, we have $b = -\log(1-\delta)$ as well. Thus, for any $\delta < 1$, we have

$$\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(Z) \ne V_j) \ge \frac{d}{2}\left(1 - \sqrt{\frac{7(2-\delta)}{(1-\delta)} \frac{2\delta^2}{(1-\delta)^2} \frac{n\varepsilon_{\mathrm{kl}}}{d}} - \delta_{\mathrm{total}}\right).$$

Taking $\delta^2 = c \min\{1, d/(n\varepsilon_{\rm kl})\}$, using that $\delta_{\rm total} \le \frac{1}{2}$, and noting that the separation is at least $\delta/2$ in Assouad's Lemma 8 gives the corollary.

C.2. Proof of Corollary 4

We give a brief example before beginning the proof to show that similar ideas extend to other convex risk minimization problems.

Example 2 (Convex risk minimization) *Consider the problem of minimizing a convex risk functional $R_P(\theta) := \mathbb{E}_P[\ell(\theta; X)]$, where ℓ is convex in its first argument and the expectation is over $X \sim P$. Now, define $\theta(P) = \operatorname{argmin}_\theta \mathbb{E}[\ell(\theta; X)]$, and let $L(\theta, \theta(P)) = R_P(\theta) - R_P(\theta(P))$. If R_P is λ-strongly convex in a neighborhood of radius r of $\theta(P)$, then a straightforward convexity argument (Hiriart-Urruty and Lemaréchal, 1993) yields*

$$R_P(\theta) - R_P(\theta(P)) \ge \min\left\{\frac{\lambda}{2}\left\|\theta - \theta(P)\right\|_2^2, \lambda r \left\|\theta - \theta(P)\right\|_2\right\}.$$

Thus, if as in the previous example we can construct distributions P such that $\theta(P_v) = \delta \cdot v \in \{-\delta, \delta\}^d$, where $\delta \le r$, then $L(\theta, \theta(P))$ induces a $\lambda\delta^2/2$-separation in Hamming metric.

Our proof proceeds in two steps. First, we argue that the gap in the logistic risk is lower bounded by a quadratic (cf. Example 2); we then argue that this quadratic lower bound can be reduced to estimation in a model with independent Bernoulli coordinates. To avoid somewhat tedious constants, we perform the analysis in an asymptotic sense.

We first describe the precise problem setting. Let $\delta > 0$, to be chosen later, and let $v \in \mathcal{V} := \{\pm 1\}^d$ as is standard for our applications of Assouad's method, and for each $v \in \mathcal{V}$ let $\theta^v = \delta v$. Now, for any $\theta \in \{\pm\delta\}^d$, consider the class-conditional distributions with coordinates of $X \in \mathbb{R}^d$ independent and distributed (conditional on $Y \in \{\pm 1\}$) as

$$X_j \mid Y = \begin{cases} Y & \text{w.p. } \frac{e^{\theta_j/2}}{e^{\theta_j/2}+e^{-\theta_j/2}} = \frac{e^{\theta_j X_j Y/2}}{e^{\delta/2}+e^{-\delta/2}} \\ -Y & \text{w.p. } \frac{e^{-\theta_j/2}}{e^{\theta_j/2}+e^{-\theta_j/2}} = \frac{e^{-\theta_j X_j Y/2}}{e^{\delta/2}+e^{-\delta/2}}. \end{cases}$$

Let the prior probabilities $P(Y = y) = \frac{1}{2}$ for $y \in \{\pm 1\}$ Then conditional on $X = x \in \{\pm 1\}^d$, we have

$$P(Y = y \mid X = x) = \frac{\prod_{j=1}^d e^{\theta_j x_j y/2}}{\prod_{j=1}^d e^{\theta_j x_j y/2} + \prod_{j=1}^d e^{-\theta_j x_j y/2}} = \frac{e^{\theta^T x y}}{1 + e^{\theta^T x y}},$$

so that $Y \mid X$ follows the logistic model.

Quadratic lower bounds on risk: Fixing v, let $R_{\delta v}(\theta) = \mathbb{E}_{\delta v}[\ell(\theta; (X, Y))]$, where $\mathbb{E}_{\delta v}$ indicates expectation under the logistic model above with $\theta = \delta v$; note that $\theta^\star := \operatorname{argmin}_\theta R_{\delta v}(\theta) = \delta v$ here. We claim that for all $\epsilon > 0$ there exists a $\gamma > 0$ such that

$$\liminf_{\delta \downarrow 0} \inf_{\|\theta\|_2 \le \gamma} \lambda_{\min}(\nabla^2 R_{\delta v}(\theta)) \ge \frac{1 - \epsilon}{4}. \tag{20}$$

We return to prove inequality (20) at the end of the proof of the corollary, noting that by Example 2, it immediately implies that if $\delta > 0$ is small enough then

$$R_{\delta v}(\theta) - \inf_\theta R_{\delta v}(\theta) \ge \min\left\{\frac{1-\epsilon}{8}\left\|\theta - \delta v\right\|_2^2, \frac{1-\epsilon}{4}\gamma \left\|\theta - \delta v\right\|_2\right\}.$$

Projecting θ into the set $[-\delta, \delta]^d$ can only decrease the right hand side of the previous display, and thus (again for small enough $\delta > 0$ and using that $\gamma > 0$ is fixed relative to δ) we see that

$$R_{\delta v}(\theta) - \inf_\theta R_{\delta v}(\theta) \geq \delta^2 \frac{1-\epsilon}{8} \sum_{j=1}^d 1\left\{\mathrm{sgn}(\theta_j) \neq v_j\right\}. \tag{21}$$

This is exactly the separation condition (7) necessary for application of Assouad's method.

Reduction to Bernoulli estimation By construction, for each coordinate j, we have $YX_j \sim \mathsf{Bernoulli}(e^{\theta_j}/(1+e^{\theta_j}))$, independent of the others. As a consequence, we see for any estimator $\widehat{V}$ of the signs of the parameters of the logistic model, there exists an estimator $\widehat{V}^{\mathsf{bern}}$ and channel Q^{bern}, which is equally private to Q (and both are independent of the true $\theta = \delta v$), such that

$$\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \neq v_j) \geq \sum_{j=1}^d \mathbb{P}_{Q^{\mathsf{bern}}}(\widehat{V}_j^{\mathsf{bern}}(\boldsymbol{Z}) \neq v_j), \tag{22}$$

where the first expectation is taken over our logistic model with parameters θ and the second over the distribution on X with independent $\mathsf{Bernoulli}(e^{\theta_j}/(1+e^{\theta_j}))$ coordinates.

We now apply an argument completely parallel to that in the proof of Corollary 3, again focusing on Assumption A2 for simplicity—the parallel case under Assumptions A1 or A1′ is similarly immediate from Corollary 11. Let $P_{-1} = \mathsf{Bernoulli}(e^{-\delta}/(1+e^{-\delta}))$ and $P_1 = \mathsf{Bernoulli}(e^{\delta}/(1+e^{\delta}))$. Then $|\log dP_1/dP_{-1}| \leq 2\delta$, and Lemma 24 implies that the strong data processing constant $\beta(P_1, P_{-1}) \leq 2(e^{2\delta}-1)^2$. Randomizing over V uniform in $\mathcal{V}$, Lemma 15 and Theorem 10 (coupled with Corollary 14) yield the lower bound

$$\begin{aligned}\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\boldsymbol{Z}) \neq V_j) &\geq \frac{d}{2}\left(1 - \sqrt{\frac{14(e^{2\delta}+1)}{d}(e^{2\delta}-1)^2 I(X_{\leq n}; \boldsymbol{Z} \mid V)} - \delta_{\mathrm{total}}\right) \\ &\geq \frac{d}{2}\left(1 - \sqrt{\frac{14(e^{2\delta}+1)}{d}(e^{2\delta}-1)^2 n\varepsilon_{\mathrm{kl}}} - \delta_{\mathrm{total}}\right).\end{aligned}$$

Setting $\delta^2 = c\frac{d}{n\varepsilon_{\mathrm{kl}}}$ for small enough constant $c > 0$, inequality (21) coupled with inequality (22) immediately yields

$$\mathbb{E}\left[R_{\delta V}(\widehat{\theta}_n(\boldsymbol{Z})) - \inf_\theta R_{\delta V}(\theta)\right] \geq Cd\delta^2 = C'd\frac{d}{n\varepsilon_{\mathrm{kl}}}$$

as desired, where $C, C' > 0$ are numerical constants.

Proof of inequality (20): As $\theta \mapsto \nabla^2 R_{\delta v}(\theta)$ is $\mathcal{C}^\infty$ in θ, as is $\delta \mapsto R_{\delta v}(\theta)$ by the logistic model, we may swap the limit infimum and infimum over $\|\theta\|_2 \leq \gamma$. Now, fix any θ with $\|\theta\|_2 \leq \gamma$, where we will choose γ momentarily. Then Lebesgue's dominated convergence theorem and the continuity of the minimum eigenvalue $\lambda_{\min}$ gives

$$\liminf_{\delta \downarrow 0} \lambda_{\min}(\nabla^2 \mathbb{E}_{\delta v}[\ell(\theta; (X,Y))]) = \lambda_{\min}(\mathbb{E}[p_\theta(X)(1-p_\theta(X))XX^T])$$

where $X \sim \mathsf{Uni}(\{\pm 1\}^d)$ and $p_\theta(X) = 1/(1+e^{\theta^T X})$. As $\theta^T X$ is $\|\theta\|_2^2$-sub-Gaussian (Vershynin, 2012), meaning that $\mathbb{E}[e^{\theta^T X}] \leq \exp(\|\theta\|_2^2/2)$, standard sub-Gaussian concentration inequalities

and that $\|\theta\|_2 \le \gamma$ imply that with probability over at least $1-\alpha$ over X, we have $|\langle\theta, X\rangle| \le \sqrt{2\gamma^2\log(2/\alpha)}$. Setting $t = \sqrt{2\gamma^2\log(2/\alpha)}$, if $\gamma > 0$ is small enough that $e^t/(1+e^t)^2 \ge (1-\epsilon/2)/4$ we have

$$\lambda_{\min}(\mathbb{E}[p_\theta(X)(1-p_\theta(X))XX^T]) \ge \lambda_{\min}\left(\frac{1-\epsilon/2}{4}\mathbb{E}[XX^T]\right) - d\alpha = \frac{1-\epsilon/2}{4}\lambda_{\min}(I_{d\times d}) - d\alpha.$$

Choosing α and γ small enough, we have $\lambda_{\min}(\mathbb{E}[p_\theta(X)(1-p_\theta(X))]) \ge \frac{1-\epsilon}{4}$ as desired.

C.3. Proof of Corollary 5

We provide a slightly different proof, beginning with the reduction to Assouad's method. Let $\delta > 0$ to be chosen presently. We first observe that if $\theta \in [-\delta,\delta]^d$, then it is no loss of generality to assume the estimator $\widehat{\theta} \in [-\delta,\delta]^d$, as otherwise, we may simply project to $[-\delta,\delta]^d$. Then for any distributions P and $\overline{P}$ and any coordinate j, we have

$$\mathbb{E}_P[(\widehat{\theta}_j - \theta_j)^2] = \mathbb{E}_{\overline{P}}[(\widehat{\theta}_j - \theta_j)^2] + \int(\widehat{\theta}_j - \theta_j)^2(dP - d\overline{P}) \ge \mathbb{E}_{\overline{P}}[(\widehat{\theta}_j - \theta_j)^2] - 8\delta^2\left\|P - \overline{P}\right\|_{\mathrm{TV}}.$$

Now, let $\mathcal{P}_\delta$ be the collection of normal distributions with means in $[-\delta,\delta]$. Using Lemma 15, we then obtain that for any channel Q satisfying Assumption A2, there exist $\varepsilon_{i,t}$-differentially private channels $\overline{Q}$ satisfying $\sum_{i,t}\min\{\varepsilon_{i,t}, \varepsilon_{i,t}^2\} \le n\varepsilon_{\mathrm{kl}}$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2, Q) \ge \mathfrak{M}_n(\theta(\mathcal{P}_\delta), \|\cdot\|_2^2, Q) \ge \mathfrak{M}(\theta(\mathcal{P}_\delta), \|\cdot\|_2^2, \overline{Q}) - 8d\delta^2\delta_{\mathrm{total}}. \tag{23}$$

In the lower bound (23), choosing

$$\delta^2 = c\min\left\{\frac{d}{\varepsilon_{\mathrm{kl}}}\frac{d\sigma^2}{n}, 1\right\}$$

and using Theorem 4.5 of Braverman et al. (2016) (with the choice $k = d/2$ in their result, along with the specified separation δ), coupled with Lemma 12, we obtain the lower bound $c\min\{\frac{d}{\varepsilon_{\mathrm{kl}}}\frac{d\sigma^2}{n}, d\}$. The $d\sigma^2/n$ term is the standard minimax bound for estimation of a Gaussian mean.

C.4. Proof of Corollary 6

The proof is nearly identical to that of Corollary 5, except that in the lower bound (23), we may replace the $8d\delta^2\delta_{\mathrm{total}}$ term with $16k\delta^2\delta_{\mathrm{total}}$, which follows by assuming w.l.o.g. that $\widehat{\theta}$ is k-sparse, in which case we estimate at most $2k$ entries of θ incorrectly. Then the lower bound of $\frac{d}{\varepsilon_{\mathrm{kl}}}\frac{k\sigma^2}{n}$ follows by Theorem 4.5 of Braverman et al. (2016), coupled with Lemma 12. The minimum involving k follows because $\|\theta - \theta'\|_2^2 \le 4k$ for all $\theta, \theta' \in [-1,1]^d$ with $\|\theta\|_0 \le k$. The $k\sigma^2\log(\frac{d}{k})/n$ term is the standard minimax lower bound for sparse Gaussian sequence estimation (Johnstone, 2013).

Appendix D. Technical proofs

D.1. From approximate to pure differential privacy (proof of Lemma 15)

In this section, we prove Lemma 15. The idea in the lemma is simple (though measurability issues preclude trivial proof): we can construct alternative channels $\overline{Q}$ that are close in variation distance to Q, where $\overline{Q}$ satisfy pure differential privacy.

We use Lemma 25 along with the fact that it is no loss of generality to assume that, by approximations and continuity of f-divergences, the $\boldsymbol{Z}$ are discrete (Liese and Vajda, 2006, Thm. 15). Indeed, the variation distance $\|\cdot\|_{\mathrm{TV}}$ is an f-divergence and $\mathbb{P}_{\pm 1}$ are marginal distributions over $\boldsymbol{Z} = Z_{\le n}^{(\le T)} \in \mathcal{Z}^{nT}$. Thus, letting $\mathcal{A}$ denote a finite rectangular partition of $\mathcal{Z}^{nT}$, meaning that the sets in $A \in \mathcal{A}$ are of the form

$$A = \prod_{t=1}^{T} \left(A_{1,1} \otimes A_{2,1} \otimes \cdots \otimes A_{n,1}\right), \quad A_{i,t} \subset \mathcal{Z},$$

and recalling that rectangles generate the Borel σ-algebra on $\mathcal{Z}^{nT}$, we have the equality (cf. Liese and Vajda, 2006, Theorem 15)

$$\|\mathbb{P}_1 - \mathbb{P}_{-1}\|_{\mathrm{TV}} = \sup_{\mathcal{A}} \sum_{A \in \mathcal{A}} |\mathbb{P}_1(\boldsymbol{Z} \in A) - \mathbb{P}_{-1}(\boldsymbol{Z} \in A)|, \tag{24}$$

where the supremum is taken over all finite rectangular partitions of $\mathcal{Z}^{nT}$.

We use equality (24) to prove the result. Without loss of generality, we assume the supremum (24) is attained (otherwise, we simply approximate). As the partition $\mathcal{A}$ is finite and consists of rectangular sets, we can assume the communicated $Z_i^{(t)}$ are discrete. We then have the following lemma, whose proof we defer to Section D.2. This is an extension of the result Dwork et al. (2010) that (ε, δ)-private channels are close to $(\varepsilon, 0)$-private channels; naive application of earlier constructions can yield in non-measurable objects and non-regular conditional probabilities.

Lemma 25 *Assume that $\mathcal{Z}$ is countable and that for each $i, t \in \mathbb{N}$, the channel $Q(\cdot \mid x_i, z_{\to i}^{(t)})$ is a regular conditional probability and that it is (ε, δ)-differentially private. Then there exists a regular conditional probability $\overline{Q}(\cdot \mid x_i, z_{\to i}^{(t)})$ such that $\overline{Q}$ is ε-differentially private and*

$$\sup_{x_i \in \mathcal{X}} \left\| Q(\cdot \mid x_i, z_{\to i}^{(t)}) - \overline{Q}(\cdot \mid x_i, z_{\to i}^{(t)}) \right\|_{\mathrm{TV}} \le \frac{1}{2}\left[\frac{\delta}{1 + e^{\varepsilon}} + \frac{\delta}{1 + e^{\varepsilon} - \delta}\right].$$

Let $\overline{Q}$ be the channels Lemma 25 guarantees, and let $\overline{\mathbb{P}}_{\pm 1}$ be the induced marginal distributions on $\mathcal{Z}^{nT}$. Then

$$\|\mathbb{P}_1 - \mathbb{P}_{-1}\|_{\mathrm{TV}} \le \left\|\mathbb{P}_1 - \overline{\mathbb{P}}_1\right\|_{\mathrm{TV}} + \left\|\overline{\mathbb{P}}_1 - \overline{\mathbb{P}}_{-1}\right\|_{\mathrm{TV}} + \left\|\overline{\mathbb{P}}_{-1} - \mathbb{P}_{-1}\right\|_{\mathrm{TV}}$$

by the triangle inequality. Letting q denote the p.m.f. of Q, we bound $\left\|\mathbb{P}_v - \overline{\mathbb{P}}_v\right\|_{\mathrm{TV}}$ by expanding

$$\begin{aligned}
\left\|\mathbb{P}_v - \overline{\mathbb{P}}_v\right\|_{\mathrm{TV}} &= \frac{1}{2} \sum_{\boldsymbol{z} \in \mathcal{Z}^{nT}} \left| \int \left(q(\boldsymbol{z} \mid x_{\le n}) - \overline{q}(\boldsymbol{z} \mid x_{\le n})\right) dP_v(x_{\le n}) \right| \\
&= \frac{1}{2} \sum_{\boldsymbol{z} \in \mathcal{Z}^{nT}} \left| \int \left(\prod_{i,t} q(z_i^{(t)} \mid x_{\le n}, z_{\to i}^{(t)}) - \prod_{i,t} \overline{q}(z_i^{(t)} \mid x_{\le n}, z_{\to i}^{(t)}) \right) dP_v(x_{\le n}) \right| \\
&= \frac{1}{2} \sum_{\boldsymbol{z} \in \mathcal{Z}^{nT}} \left| \int \left(\prod_{i,t} q(z_i^{(t)} \mid x_i, z_{\to i}^{(t)}) - \prod_{i,t} \overline{q}(z_i^{(t)} \mid x_i, z_{\to i}^{(t)}) \right) \prod_{i \le n} dP_v(x_{\le n}) \right|
\end{aligned}$$

where we have used that $Z_i^{(t)}$ is conditionally independent of $X_{\setminus i}$ given X_i and $Z_{\to i}^{(t)}$. Now, let $(j,\tau) \prec (i,t)$ indicate the ordering that either $\tau < t$ or $j < i$ and $\tau = t$ (and similarly $(j,\tau) \succ (i,t)$ means that $\tau > t$ or $\tau = t$ and $j > i$), and define the shorthand

$$q_{\prec(i,t)}(\boldsymbol{z} \mid x_{\le n}) := \prod_{(j,\tau)\prec(i,t)} q(z_j^{(\tau)} \mid x_j, z_{\to j}^{(\tau)})$$

and similarly for $\overline{q}$ and $q_{\succ(i,t)}$. Using the telescoping identity that

$$\prod_i a_i - \prod_i b_i = \sum_i \Big(\prod_{j<i} a_j\Big)(a_i - b_i)\Big(\prod_{j>i} b_j\Big)$$

and the triangle inequality, we have

$$\begin{aligned} &2\left\|\mathbb{P}_v - \overline{\mathbb{P}}_v\right\|_{\mathrm{TV}} \\ &\le \sum_{\substack{v:v_j=1\\ i,t}} \int_{\mathcal{X}^n} \underbrace{\sum_{\boldsymbol{z}\in\mathcal{Z}^{nT}} q_{\prec(i,t)}(\boldsymbol{z} \mid x_{\le n}) \left|q(z_i^{(t)} \mid x_i, z_{\to i}^{(t)}) - \overline{q}(z_i^{(t)} \mid x_i, z_{\to i}^{(t)})\right| \overline{q}_{\succ(i,t)}(\boldsymbol{z} \mid x_{\le n})}_{=:T_{it}} \, dP_v(x_{\le n}). \end{aligned} \tag{25}$$

The term T_{it} satisfies

$$T_{it} = \sum_{\substack{(j,\tau)\prec(i,t),\\ z_j^{(\tau)}}} q_{\prec(i,t)}(\boldsymbol{z} \mid x_{\le n}) \sum_{z_i^{(t)}\in\mathcal{Z}} \left|q(z_i^{(t)} \mid x_i, z_{\to i}^{(t)}) - \overline{q}(z_i^{(t)} \mid x_i, z_{\to i}^{(t)})\right| \sum_{\substack{(j,\tau)\succ(i,t),\\ z_j^{(\tau)}}} \overline{q}_{\succ(i,t)}(\boldsymbol{z} \mid x_{\le n}),$$

where the variation distance guarantee of Lemma 25 (coupled with the privacy Assumption A2) guarantees that

$$\begin{aligned} \sum_{z_i^{(t)}\in\mathcal{Z}} \left|q(z_i^{(t)} \mid x_i, z_{\to i}^{(t)}) - \overline{q}(z_i^{(t)} \mid x_i, z_{\to i}^{(t)})\right| &\le \frac{1}{2}\left[\frac{\delta_{i,t}(z_{\to i}^{(t)})}{1 + e^{\varepsilon_{i,t}(z_{\to i}^{(t)})}} + \frac{\delta_{i,t}(z_{\to i}^{(t)})}{1 + e^{\varepsilon_{i,t}(z_{\to i}^{(t)})} - \delta_{i,t}(z_{\to i}^{(t)})}\right] \\ &\le \delta_{i,t}(z_{\to i}^{(t)}) \end{aligned}$$

as $\delta_{i,t} \in [0,1]$. We thus obtain

$$\begin{aligned} T_{it} &\le \sum_{\substack{(j,\tau)\prec(i,t),\\ z_j^{(\tau)}}} q_{\prec(i,t)}(\boldsymbol{z} \mid x_{\le n}) \delta_{i,t}(z_{\to i}^{(t)}) \max_{z_i^{(t)}\in\mathcal{Z}} \sum_{\substack{(j,\tau)\succ(i,t),\\ z_j^{(\tau)}}} \overline{q}_{\succ(i,t)}(\boldsymbol{z} \mid x_{\le n}) \\ &= \sum_{\substack{(j,\tau)\prec(i,t),\\ z_j^{(\tau)}}} q_{\prec(i,t)}(\boldsymbol{z} \mid x_{\le n}) \delta_{i,t}(z_{\to i}^{(t)}) = \mathbb{E}_Q\left[\delta_{i,t}(Z_{\to i}^{(t)}) \mid X_{\le n} = x_{\le n}\right], \end{aligned}$$

where the equality follows because p.m.f.s sum to 1. Substituting this into inequality (25) yields

$$\left\|\mathbb{P}_v - \overline{\mathbb{P}}_v\right\|_{\mathrm{TV}} \le \frac{1}{2}\sum_{i,t} \mathbb{E}_{P_v}[\delta_{i,t}(Z_{\to i}^{(t)})] = \frac{1}{2}\sum_{i,t} \mathbb{E}_{\mathbb{P}_v}\left[\delta_{i,t}(Z_{\to i}^{(t)})\right] \le \frac{\delta_{\mathrm{total}}}{2},$$

the final inequality following again by Assumption A2. This gives Lemma 15.

D.2. Proof of Lemma 25

If the space $\mathcal{X}$ is countable, then this result is essentially due to Dwork et al. (2010) (see Lemma 2.1 in the long version of their paper) once we apply the averaging technique in the end of this proof. When the space $\mathcal{X}$ is not countable, we must be more careful to maintain measurability, so that our construction actually yields a valid channel. Because $\mathcal{Z}$ is countable, however, it is possible to achieve our desired result. Without loss of generality, because $\mathcal{Z}$ is countable, we may assume that Q has a density (p.m.f.) q on $\mathcal{Z}$, as each $Q(\cdot \mid x_i, z_{\to i}^{(t)})$ is absolutely continuous w.r.t. the counting measure on $\mathcal{Z}$.

Let us take $x, x' \in \mathcal{X}$ otherwise arbitrary, and let $w = z_{\to i}^{(t)}$ for shorthand, so that we have densities $q(z \mid x, w)$ and $q(z \mid x', w)$, both of which are measurable in their (three) arguments. Then define the two sets

$$S_x := \{z \in \mathcal{Z} \mid q(z \mid x, w) > e^\varepsilon q(z \mid x', w)\} \text{ and } S_{x'} := \{z \in \mathcal{Z} \mid q(z \mid x', w) > e^\varepsilon q(z \mid x, w)\}$$

and the intermediate densities

$$\begin{aligned} q_1(z \mid x; x', w) &:= \left[q(z \mid x, w) + q(z \mid x', w)\right] \left(\frac{e^\varepsilon}{e^\varepsilon + 1} 1\left\{z \in S_x\right\} + \frac{1}{e^\varepsilon + 1} 1\left\{z \in S_{x'}\right\}\right) \\ &\quad + q(z \mid x, w) 1\left\{z \notin S_x \cup S_{x'}\right\}, \\ q_1(z \mid x'; x, w) &:= \left[q(z \mid x, w) + q(z \mid x', w)\right] \left(\frac{1}{e^\varepsilon + 1} 1\left\{z \in S_x\right\} + \frac{e^\varepsilon}{e^\varepsilon + 1} 1\left\{z \in S_{x'}\right\}\right) \\ &\quad + q(z \mid x, w) 1\left\{z \notin S_x \cup S_{x'}\right\}. \end{aligned}$$

Evidently these quantities satisfy

$$e^{-\varepsilon} \le \frac{q_1(z \mid x; x', w)}{q_1(z \mid x'; x, w)} \le e^\varepsilon$$

for all $z \in \mathcal{X}$, and moreover, by inspection they are (z, x, x', w)-measurable as they are the product of measurable functions. Let Q_1 denote the induced measure (not necessarily probabilities) on $\mathcal{Z}$ by the constructed q_1.

With this definition of Q_1, we may define the two quantities

$$\begin{aligned} \alpha_x := Q(S_x \mid x, w) - Q_1(S_x \mid x; x', w) &= Q(S_x \mid x, w) - \frac{e^\varepsilon}{1 + e^\varepsilon}(Q(S_x \mid x, w) + Q(S_x \mid x', w)) \\ &= \frac{Q(S_x \mid x, w) - e^\varepsilon Q(S_x \mid x', w)}{1 + e^\varepsilon} \in \left[0, \frac{\delta}{1 + e^\varepsilon}\right] \end{aligned}$$

and similarly

$$\alpha_{x'} := Q(S_{x'} \mid x', w) - Q_1(S_{x'} \mid x'; x, w) \in \left[0, \frac{\delta}{1 + e^\varepsilon}\right].$$

We also have $Q(S_x \mid x, w) - Q_1(S_x \mid x; x', w) = Q_1(S_x \mid x'; x, w) - Q(S_x \mid x', w)$ and $Q(S_{x'} \mid x', w) - Q_1(S_{x'} \mid x'; x, w) = Q_1(S_{x'} \mid x; x', w) - Q(S_{x'} \mid x, w)$ by construction. With these

definitions and equalities, we have the variation bound

$$\begin{aligned}
&\left\|Q(\cdot \mid x, w) - Q_1(\cdot \mid x'; x, w)\right\|_{\mathrm{TV}} \\
&= \frac{1}{2}\left(Q(S_x \mid x, w) - Q_1(S_x \mid x; x', w)\right) + \frac{1}{2}\left(Q_1(S_{x'} \mid x; x', w) - Q(S_{x'} \mid x, w)\right) \\
&= \frac{1}{2}\alpha_x + \frac{1}{2}\alpha_{x'} \le \frac{\delta}{1 + e^\varepsilon}.
\end{aligned}$$

The normalized densities

$$q_0(z \mid x; x', w) := \frac{q_1(z \mid x; x', w)}{\sum_z q_1(z \mid x; x', w)} \text{ and } q_0(z \mid x'; x, w) := \frac{q_1(z \mid x'; x, w)}{\sum_z q_1(z \mid x; x', w)}$$

are both (z, x, x', w)-measurable, and they satisfy the ratio guarantee $|\log \frac{q_0(z|x;x',w)}{q_0(z|x';x,w)}| \le \varepsilon$. Moreover, we have $Q_1(\mathcal{Z} \mid x; x', w) = 1 - \alpha_x + \alpha_{x'}$ and $Q_1(\mathcal{Z} \mid x'; x, w) = 1 - \alpha_{x'} + \alpha_x$. We then have

$$\begin{aligned}
&\left\|Q(\cdot \mid x, w) - Q_0(\cdot \mid x; x', w)\right\|_{\mathrm{TV}} \\
&\le \left\|Q(\cdot \mid x, w) - Q_1(\cdot \mid x; x', w)\right\|_{\mathrm{TV}} + \left\|Q_1(\cdot \mid x; x', w) - Q_0(\cdot \mid x; x', w)\right\|_{\mathrm{TV}} \\
&= \frac{\alpha_x + \alpha_{x'}}{2} + \frac{1}{2}\left|\frac{1}{Q_1(\mathcal{Z} \mid x; x', w)} - 1\right| = \frac{\alpha_x + \alpha_{x'}}{2} + \frac{|\alpha_x - \alpha_{x'}|/2}{1 - \alpha_x + \alpha_{x'}} \le \frac{1}{2}\left[\frac{\delta}{1 + e^\varepsilon} + \frac{\delta}{1 + e^\varepsilon - \delta}\right].
\end{aligned}$$

where we have taken $\alpha_x = \delta/(1 + e^\varepsilon)$ and $\alpha_{x'} = 0$ to maximize the sum above. An identical bound holds on $\|Q(\cdot \mid x', w) - Q_0(\cdot \mid x'; x, w)\|_{\mathrm{TV}}$.

It remains to construct our desired regular conditional distribution $\overline{Q}$. To that end, note that each of $q_0(z \mid x; x', w)$ and $q_0(z \mid x'; x, w)$ are measurable in (z, x, x', w) by our construction. Choosing an arbitrary probability measure λ on the space $\mathcal{X}$, we may then define

$$\overline{q}(z \mid x, w) := \int q_0(z \mid x; x', w) d\lambda(x')$$

for all z, x, w. Taking $\overline{Q}$ to be the associated probability measure, we evidently have that $\overline{Q}$ is a regular conditional probability, that $\|\overline{Q}(\cdot \mid x, w) - Q(\cdot \mid x, w)\|_{\mathrm{TV}} \le \frac{1}{2}(\frac{\delta}{1+e^\varepsilon} + \frac{\delta}{1+e^\varepsilon-\delta})$, and that $e^{-\varepsilon} \le \overline{q}(z \mid x, w)/\overline{q}(z \mid x', w) \le e^\varepsilon$ as desired.

D.3. Proof of Lemma 7

We allow c, C to be numerical constants whose value may change from line to line. We also assume $\sigma^2 > 0$ is at least a numerical constant. First, we have that $|Z_i| \le b$. Thus

$$\mathbb{P}(|\overline{Z}_n - \mathbb{E}[\overline{Z}_n]| \ge t) \le \exp\left(-\frac{nt^2}{2b^2}\right) \text{ for } t \ge 0$$

by Hoeffding's inequality. Note that $\mathbb{E}[\overline{Z}_n] \in [1 - 2\Phi(1/\sigma), 1 + 2\Phi(-1/\sigma)] \subset [e^{-c/\sigma^2}, 1 - e^{-c/\sigma^2}] = [e^{-C}, 1 - e^{-C}]$ by our assumption that σ is at least a constant. Now, let $\mathcal{E}$ denote the

event that $\overline{Z}_n \in [e^{-c/\sigma^2}/2, 1-e^{-c/\sigma^2}/2]$, which happens with probability at least $1-\exp(-cn/b^2)$. On this event, a Taylor expansion of Φ^{-1} gives

$$\begin{aligned}\sigma\Phi^{-1}\left(\frac{1-\overline{Z}_n}{2}\right) &= \sigma\Phi^{-1}\left(\frac{1-\mathbb{E}[\overline{Z}_n]}{2}\right) + \sigma\frac{\overline{Z}_n - \mathbb{E}[\overline{Z}_n]}{\phi(\theta)} \pm C\sigma(\overline{Z}_n - \mathbb{E}[\overline{Z}_n])^2 \\ &= \theta + \sigma\frac{\overline{Z}_n - \mathbb{E}[\overline{Z}_n]}{\phi(\theta)} \pm C\sigma(\overline{Z}_n - \mathbb{E}[\overline{Z}_n])^2.\end{aligned}$$

We have $|\overline{Z}_n - \mathbb{E}[\overline{Z}_n]| \le \sqrt{2b^2t/n}$ with probability at least $1-e^{-t}$ by Hoeffding's inequality, and we also have

$$\begin{aligned}\mathbb{E}_\theta[\|\widehat{\theta}_n - \theta\|_2^2] &\le \frac{2\sigma^2}{\phi(\theta)^2}\mathbb{E}[(\overline{Z}_n - \mathbb{E}[\overline{Z}_n])^2] + C^2\sigma^2\mathbb{E}[(\overline{Z}_n - \mathbb{E}[\overline{Z}_n])^4] + C\mathbb{P}(\mathcal{E}^c) \\ &\le C\frac{b^2\sigma^2}{n} + C\frac{b^4\sigma^2}{n^2} + Ce^{-cn/b^2},\end{aligned}$$

where the second inequality follows by the b-boundendess of the Z_i and standard moment bounds for sub-Gaussian random variables (Vershynin, 2012).