Fisher information under local differential privacy

Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür

Stanford University, Stanford, CA 94305

Email: {lpb, wnchen, aozgur}@stanford.edu

Abstract

We develop data processing inequalities that describe how Fisher information from statistical samples can scale with the privacy parameter ε under local differential privacy constraints. These bounds are valid under general conditions on the distribution of the score of the statistical model, and they elucidate under which conditions the dependence on ε is linear, quadratic, or exponential. We show how these inequalities imply order optimal lower bounds for private estimation for both the Gaussian location model and discrete distribution estimation for all levels of privacy $\varepsilon > 0$. We further apply these inequalities to sparse Bernoulli models and demonstrate privacy mechanisms and estimators with order-matching squared ℓ^2 error.

I. Introduction

In the model of local differential privacy [1], [2], [3], [4], sensitive data is released to an aggregator or centralized processor only after having been processed by a privatization mechanism. This privatization mechanism distorts the data in such a way that it is statistically guaranteed to not reveal too much about the underlying sensitive data. There is an inherent trade-off between the degree to which the data is distorted by the mechanism (and therefore the amount of privacy achieved), and the utility of the data for performing statistical inference and estimation tasks.

One measure of the information conveyed by a statistical sample for estimating a parameter is the so-called Fisher information, which describes how a family of probability distributions changes as one varies the parameter of interest. Under some mild regularity conditions, the Fisher information at a point θ in the space of possible parameters immediately gives a lower bound

on the squared ℓ^2 risk for estimating θ via the well-known Cramér-Rao bound for unbiased estimators [5], [6], [7], [8]. More generally, Fisher information describes the complexity of estimation problems in an asymptotic sense locally around θ [9], [10]; and a Bayesian version of the Cramér-Rao bound known as the van Trees inequality can be used to give lower bounds that hold for any estimator (including arbitrarily biased estimators) [11].

In this paper, we consider the problem of estimating a parameter $\theta \in \mathbb{R}^d$ from n independent statistical samples that have been processed by an ε-locally differentially private mechanism. See Figure 1. We characterize the Fisher information from the privatized samples, and provide strong data processing inequalities that describe how the Fisher information can scale with the privatization parameter ε. These data processing inequalities are valid under very general conditions on the tail of the score function random variable, and elucidate under which conditions the dependence on ε is linear, quadratic, or exponential. Using the van Trees inequality, we recover in a unified way order-wise optimal lower bounds on the minimax squared ℓ^2 risk for Gaussian mean estimation and discrete distribution estimation at all levels of privacy $\varepsilon > 0$, matching the lower bounds from [12], [13], [14] with simpler and more transparent proofs. Our results also apply to a sequential interaction model for local differential privacy via a straightforward consequence of the chain-rule for Fisher information; and they can even be applied to a fully interactive blackboard model by characterizing the Fisher information from the entire blackboard transcript, therefore extending, for example, earlier bounds in [13] to fully interactive models.

We further demonstrate the utility of this framework by developing lower bounds for other statistical models such as a sparse Bernoulli model with $X_i \sim \prod_{j=1}^{d} \mathsf{Bern}(\theta_j)$ and $\sum_{j=1}^{d} \theta_j \leq s$, and demonstrate a privatization mechanism with matching error. This model is interesting in that when $s = 1$, it provides an example where the dependence of the minimax squared ℓ^2 risk on ε is exponential even if the d components of each sample X_i are independent of each other. This is in contrast to mean estimation for the Gaussian and the dense Bernoulli model (when s is of order d), where the dependence on ε is linear [14]. When $s > 1$, the sample-size penalty due to privatization is of the order $\frac{s \log d}{\varepsilon}$ in the privacy regime $\log d \preceq \varepsilon \preceq s \log d$. This penalty scales linearly only with the sparsity s rather than the ambient dimension d, which opens up the possibility of private estimation with a more modest penalty provided that the data can be assumed to be sparse in a certain sense.

Our Fisher information approach to lower bounds under privacy constraints is motivated by recent results for statistical estimation under communication constraints such as [15], [16], [17]. In both the privacy constrained and communication constrained cases, the tail behavior of the score function plays a central role in determining how the risk can scale. Other works such as [18], [19], [14] have also noted the connection between communication and privacy constraints in statistical estimation. In contrast with [20], we analyze the Fisher information from the induced distribution of the privatized samples $Y_1, \ldots, Y_n$, rather than that from the original statistical model of the X_i's. In [20], Ruan and Duchi observe that the latter has limited applicability for capturing the local complexity of private estimation problems (see also [21]), while our paper shows that the former is a powerful measure for the same. Indeed, from the local asymptotic minimax point of view, it is natural to expect the Fisher information from the privatized samples to play a role in the complexity of the private estimation problem, but until now it remained unclear how to characterize or bound this Fisher information for any privatization mechanism satisfying the local differential privacy condition.

The main contributions of our paper can be summarized as follows:

- We introduce a framework for characterizing Fisher information from ε-differentially privatized samples. Under very general conditions on the statistical model, we provide upper bounds on the Fisher information that show that the dependence on ε is dictated by the tail of the score function random variable. These bounds continue to hold even when samples are released in an interactive fashion through a shared blackboard. Even though statistical estimation under privacy constraints has been of significant recent interest, to the best of our knowledge there are no known bounds on the Fisher information from privatized samples.
- We show that the bounds on Fisher information easily lend themselves to order optimal lower bounds on the minimax squared ℓ^2 risk of statistical estimation under privacy constraints. In particular, we recover in a unified way lower bounds developed separately for different statistical models in the literature, such as Gaussian mean estimation [14] and discrete distribution estimation [13], in the latter case extending the bounds to fully interactive models.
- To demonstrate the generality of our approach, we apply our bounds to a sparse Bernoulli model, for which we also develop optimal privacy mechanisms. We show that our framework can be flexibly applied to different parameter regimes of this model and the dependence of

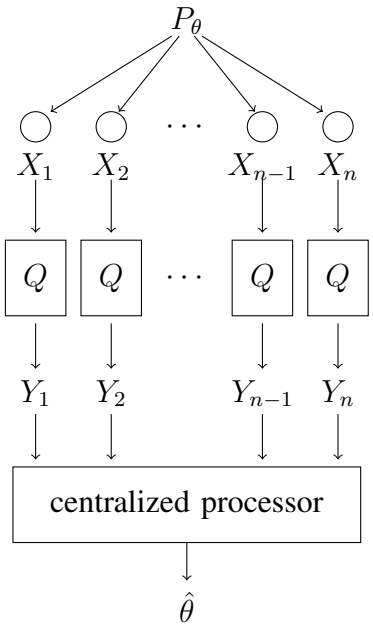

Fig. 1: An estimation system where sensitive data $X_1, \dots X_n$ is processed by the privatization mechanism $Q(y|x)$ before being released to the centralized processor that will use the data for statistical inference tasks such as estimating the parameter θ.

the minimax risk on ε can be exponential or linear depending on the parameter regime of interest.

A. *Preliminaries*

Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces and suppose that $\{P_\theta\}_{\theta \in \Theta}$ for $\Theta \subseteq \mathbb{R}^d$ is a family of probability measures on $(\mathcal{X}, \mathcal{A})$ that is dominated by some sigma-finite measure μ. Denote the density of P_θ with respect to μ by $f(x|\theta)$. Let

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$$

with θ being the parameter of interest that we are trying to estimate. We say that a regular conditional distribution

$$Q : \mathcal{B} \times \mathcal{X} \to [0, 1]$$

is an ε-differentially private mechanism if

$$\frac{Q(S|x)}{Q(S|x')} \leq e^{\varepsilon} \tag{1}$$

for any $x, x' \in \mathcal{X}$ and $S \in \mathcal{B}$.

Suppose that conditioned on $X_i = x_i$, we draw Y_i independently from $Q(\cdot|x_i)$ where Q satisfies the ε differentially priviate condition (1) with $\varepsilon > 0$. The privatization mechanisms can either be the same for each sample, or they can vary across different samples as long as each mechanism satisfies the condition (1). We will generally assume for simplicity that each mechanism is the same. In this case the Y_i have the marginal probability distribution

$$Q_\theta(S) = \int Q(S|x) f(x|\theta) d\mu(x) \, .$$

By (1), if $Q(S|x) = 0$ for some $x \in \mathcal{X}$ then $Q(S|x') = 0$ for all $x' \in \mathcal{X}$. We can therefore assume that both $\{Q(\cdot|x)\}_{x\in\mathcal{X}}$ and $\{Q_\theta\}_{\theta\in\Theta}$ form dominated families with $Q(\cdot|x) << \nu$ for all $x \in \mathcal{X}$ and $Q_\theta << \nu$ for all $\theta \in \Theta$ for some sigma-finite measure ν. Abusing notation slightly let $f(y|\theta)$ be the density of Q_θ with respect to ν, and let $Q(y|x)$ denote the density of $Q(\cdot|x)$ with respect to ν.

Instead of working with measures, we will find it more convenient to work with the corresponding densities of the probability distributions. The following proposition shows that the condition (1) implies a similar condition for the density $Q(y|x)$, and this is the form that will be most useful in the subsequent sections.

Proposition 1. *For any $x, x' \in \mathcal{X}$ and ν-almost any $y \in \mathcal{Y}$, $\frac{Q(y|x)}{Q(y|x')} \leq e^\varepsilon$.*

A proof of Proposition 1 is included in Appendix A.

II. Upper Bounds on Fisher Information

In this section, we introduce the relevant Fisher information quantities and then show how the Fisher information from the privatized samples Y_i can be upper bounded bounded in terms of the local differential privacy parameter ε. Recall that in the context of Fisher information, the score function associated with the statistical model P_θ is defined by

$$\begin{aligned} S_\theta(x) &= \nabla_\theta \log f(x|\theta) \\ &= \left(\frac{\partial}{\partial \theta_1} \log f(x|\theta), \ldots, \frac{\partial}{\partial \theta_d} \log f(x|\theta) \right)^T \, , \end{aligned}$$

and the Fisher information matrix for estimating θ from a sample Y is

$$I_Y(\theta) = \mathbb{E}\left[\left(\nabla_\theta \log f(Y|\theta)\right) \left(\nabla_\theta \log f(Y|\theta)\right)^T \right]$$

where the expectation is understood to be taken with respect to the "true" distribution with parameter θ. All logs are taken with respect to the natural base. In order to ensure that these quantities are well-defined, and that we can apply the van Trees inequality below, we require certain regularity conditions on the statistical model P_θ. In particular we assume that the square-root densities $\sqrt{f(x|\theta)}$ are continuously differentiable with respect to each θ_j and that the Fisher information from X exists and is finite. For more on these conditions and how they are used see Appendix D.

We are interested in this Fisher information quantity, in part, because it can provide bounds on the risk in estimating θ from the samples $Y_1, \ldots, Y_n$. Fisher information is by definition a local quantity that is defined at each $\theta \in \Theta$ and describes the local complexity of estimating that particular θ value asymptotically as the number of samples n increases. More concretely, if the statistical model Q_θ is *differentiable in quadratic mean* [1] [10], meaning that $\sqrt{f(y|\theta)}$ has a derivative in a certain L^2 sense, then the local asymptotic risk around θ is lower bounded as follows:

$$\sup_A \liminf_{n\to\infty} \sup_{h\in A} \mathbb{E}_{\theta+h/\sqrt{n}} \left\| \sqrt{n}\left(\hat{\theta}_n(Y_1,\ldots,Y_n) - \left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right\|_2^2 \geq \mathsf{Tr}(I_Y(\theta)^{-1})$$
$$\geq \frac{d^2}{\mathsf{Tr}(I_Y(\theta))}$$

where A is any finite subset of $\mathbb{R}^d$ (e.g. see Theorem 8.11 in [10]). Similarly, under these conditions, for each θ, there exists a sequence of estimators $\hat{\theta}_n(Y_1, \ldots, Y_n)$ such that

$$\sup_A \limsup_{n\to\infty} \sup_{h\in I} \mathbb{E}_{\theta+h/\sqrt{n}} \left\| \sqrt{n}\left(\hat{\theta}_n(Y_1,\ldots,Y_n) - \left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right\|_2^2 \leq \mathsf{Tr}(I_Y(\theta)^{-1})$$

(e.g. see Theorem 8.14 in [10]). In this way the Fisher information can determine both upper and lower bounds and is of fundamental importance in the parameter estimation problem. In the sequel, we develop upper bounds on $\mathsf{Tr}(I_Y(\theta))$, which immediately lead to lower bounds on local asymptotic risk. Additionally, by upper bounding Fisher information uniformly across all $\theta \in \Theta$ (or a subset of Θ), and using a Bayesian Cramér-Rao bound, we are able to get

[1] is implied by the assumptions we have made above without any additional assumptions on $Q(y|x)$ other than it being an ε differentially private mechanism

more global minimax lower bounds. For this we'll use a multivariate version of the van Trees inequality due to Gill and Levit [11], which bounds the average ℓ^2 risk by

$$\int_\Theta \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \lambda(\theta) d\theta \geq \frac{d^2}{\int_\Theta \mathsf{Tr}(I_{Y_1,\ldots,Y_n}(\theta))\lambda(\theta)d\theta + J(\lambda)} \tag{2}$$

where $\lambda(\theta) = \prod_{j=1}^d \lambda_j(\theta_j)$ is a prior for the parameter θ and $J(\lambda)$ is the Fisher information associated with the prior λ:

$$J(\lambda) = \sum_{j=1}^d \int \frac{\lambda_j'(\theta_j)^2}{\lambda_j(\theta_j)} d\theta_j \ .$$

Assuming that $\Theta = [-B, B]^d$, the prior λ can be chosen to minimize $J(\lambda)$ [8], [22]. This observation along with the independence of the Y_i, and upper bounding the average risk by the maximum risk, leads to

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\|\hat{\theta} - \theta\|_2^2 \geq \frac{d^2}{n \sup_{\theta\in\Theta} \mathsf{Tr}(I_Y(\theta)) + \frac{d\pi^2}{B^2}} \ . \tag{3}$$

We are therefore interested in upper bounding $\mathsf{Tr}(I_Y(\theta))$. To this end, we will need the following lemma:

Lemma 1 (Barnes et. al 2018 [15])**.** *The trace of the Fisher information matrix $I_Y(\theta)$ can be written as*

$$\mathsf{Tr}(I_Y(\theta)) = \mathbb{E}_Y\|\mathbb{E}_X[S_\theta(X)|Y]\|_2^2 \ .$$

For completeness a proof of Lemma 1 is included in Appendix B. Using the characterization of Fisher information from Lemma 1, the following Propositions 2-5 show how in the differentially private setting, $\mathsf{Tr}(I_Y(\theta))$ can be upper bounded under various assumptions on the tail of the score function random vector $S_\theta(X)$. We see that depending on the tail behavior, there can be a qualitatively different upper bound in terms of ε, i.e., it can be linear, quadratic, or exponential in ε. In Figure 2 we summarize these conditions and the corresponding upper bounds.

Proposition 2. *If $\mathbb{E}[\langle u, S_\theta(X)\rangle^2] \leq I_0$ for any unit vector $u \in \mathbb{R}^d$, then*

$$\mathsf{Tr}(I_Y(\theta)) \leq I_o(e^\varepsilon - 1)^2 \ .$$

Proof. Using Lemma 1,

$$\mathsf{Tr}(I_Y(\theta)) = \mathbb{E}_Y\|\mathbb{E}_X[S_\theta(X)|Y]\|_2^2 \ . \tag{4}$$

For a fixed y let

$$u = \frac{\mathbb{E}[S_\theta(X)|Y=y]}{\|\mathbb{E}[S_\theta(X)|Y=y]\|_2}$$

so that

$$\begin{aligned}\|\mathbb{E}[S_\theta(X)|Y=y]\|_2 &= \langle u, \mathbb{E}[S_\theta(X)|Y=y]\rangle \\ &= \mathbb{E}[\langle u, S_\theta(X)\rangle|Y=y] \\ &= \frac{1}{f(y|\theta)}\mathbb{E}[\langle u, S_\theta(X)\rangle Q(y|X)] \,. \end{aligned} \tag{5}$$

Let $c_{\min}(y) = \min_x Q(y|x)$ and $c_{\max}(y) = \max_x Q(y|x)$. We can assume that $c_{\min}(y) > 0$. Following (5),

$$\begin{aligned}\frac{1}{f(y|\theta)}\mathbb{E}[\langle u, S_\theta(X)\rangle Q(y|X)] &\le \frac{1}{c_{\min}(y)}\Bigg(\int_{\{x:\langle u,S_\theta(x)\rangle\ge 0\}}\langle u, S_\theta(x)\rangle Q(y|x)f(x|\theta)d\mu(x) \\ &\quad + \int_{\{x:\langle u,S_\theta(x)\rangle< 0\}}\langle u, S_\theta(x)\rangle Q(y|x)f(x|\theta)d\mu(x)\Bigg) \\ &\le \frac{1}{c_{\min}(y)}\Bigg(c_{\max}(y)\int_{\{x:\langle u,S_\theta(x)\rangle\ge 0\}}\langle u, S_\theta(x)\rangle f(x|\theta)d\mu(x) \\ &\quad + c_{\min}(y)\int_{\{x:\langle u,S_\theta(x)\rangle< 0\}}\langle u, S_\theta(x)\rangle f(x|\theta)d\mu(x)\Bigg) \,. \end{aligned} \tag{6}$$

Note that score functions are mean zero and thus

$$\int_{\{x:\langle u,S_\theta(x)\rangle\ge 0\}}\langle u, S_\theta(x)\rangle f(x|\theta)d\mu(x) + \int_{\{x:\langle u,S_\theta(x)\rangle< 0\}}\langle u, S_\theta(x)\rangle f(x|\theta)d\mu(x) = 0 \,. \tag{7}$$

Putting (6) together with (7),

$$\frac{1}{f(y|\theta)}\mathbb{E}[\langle u, S_\theta(X)\rangle Q(y|X)] \le (e^\varepsilon - 1)\int_{\{x:\langle u,S_\theta(x)\rangle\ge 0\}}\langle u, S_\theta(x)\rangle f(x|\theta)d\mu(x) \,,$$

and then squaring both sides and using Jensen's inequality yields

$$\left(\frac{1}{f(y|\theta)}\mathbb{E}[\langle u, S_\theta(X)\rangle Q(y|X)]\right)^2 \le (e^\varepsilon - 1)^2 I_0 \,. \tag{8}$$

The proposition is proved by combining (4), (5), and (8). □

Remark 1. *If* $0 < \varepsilon < 1$ *then* $(e^\varepsilon - 1)^2 \le (e-1)^2\varepsilon^2$ *and Proposition 2 implies*

$$\mathsf{Tr}(I_Y(\theta)) \le (e-1)^2 I_o \varepsilon^2 \,.$$

Proposition 3. *If* $\mathbb{E}[\langle u, S_\theta(X)\rangle^2] \le I_0$ *for any unit vector* $u \in \mathbb{R}^d$ *then*

$$\mathsf{Tr}(I_Y(\theta)) \le I_o e^\varepsilon \,.$$

Proof. Following (5) from the proof above, Jensen's inequality implies

$$\mathbb{E}\left[\langle u, S_\theta(X)\rangle \frac{Q(y|X)}{\mathbb{E}[Q(y|X)]}\right]^2 \leq \mathbb{E}\left[\langle u, S_\theta(X)\rangle^2 \frac{Q(y|X)}{\mathbb{E}[Q(y|X)]}\right] \leq I_0 e^{\varepsilon} .$$

□

Using the super-exponential definition of a sub-Gaussian and sub-exponential random variable [23], we say that a random vector $V \in \mathbb{R}^d$ is sub-Gaussian with parameter σ if $\mathbb{E}\left[e^{\left(\frac{\langle u,V\rangle}{\sigma}\right)^2}\right] \leq 2$ for any unit vector $u \in \mathbb{R}^d$. We say that a random vector $V \in \mathbb{R}^d$ is sub-exponential with parameter σ if $\mathbb{E}\left[e^{\left|\frac{\langle u,V\rangle}{\sigma}\right|}\right] \leq 2$ for any unit vector $u \in \mathbb{R}^d$.

Proposition 4. *If $S_\theta(X)$ is sub-Gaussian with parameter σ and $\varepsilon \geq 1$ then*

$$\mathsf{Tr}(I_Y(\theta)) \leq 2\sigma^2\varepsilon .$$

Proof. Using the convexity of $x \mapsto e^{x^2}$,

$$\begin{aligned}\exp\left(\mathbb{E}\left[\frac{1}{\sigma}\langle u, S_\theta(X)\rangle \frac{Q(y|X)}{\mathbb{E}[Q(y|X)]}\right]^2\right) &\leq \mathbb{E}\left[\exp\left(\left(\frac{\langle u, S_\theta(X)\rangle}{\sigma}\right)^2\right)\frac{Q(y|X)}{\mathbb{E}[Q(y|X)]}\right] \\ &\leq e^{\varepsilon}\mathbb{E}\left[\exp\left(\frac{\langle u, S_\theta(X)\rangle}{\sigma}\right)^2\right] \\ &\leq 2e^{\varepsilon} .\end{aligned}$$

Taking logs,

$$\mathbb{E}\left[\langle u, S_\theta(X)\rangle \frac{Q(y|X)}{\mathbb{E}[Q(y|X)]}\right]^2 \leq \sigma^2(\varepsilon + \log 2)$$

so that for $\varepsilon \geq 1$,

$$\mathbb{E}\left[\langle u, S_\theta(X)\rangle \frac{Q(y|X)}{\mathbb{E}[Q(y|X)]}\right]^2 \leq 2\sigma^2\varepsilon .$$

□

With a nearly identical proof we also have the following sub-exponential result.

Proposition 5. *If $S_\theta(X)$ is sub-exponential with parameter σ and $\varepsilon \geq 1$ then*

$$\mathsf{Tr}(I_Y(\theta)) \leq 2\sigma^2\varepsilon^2 .$$

condition on $S_\theta(X)$	upper bound on $\mathsf{Tr}(I_Y(\theta))$	lower bound for ℓ_2^2 risk (with n samples)
finite variance I_0	$I_0(e^\varepsilon - 1)^2$	$\frac{d^2}{nI_0(e^\varepsilon-1)^2}$
finite variance I_0	$I_0 e^\varepsilon$	$\frac{d^2}{nI_0 e^\varepsilon}$
σ^2-sub-Gaussian	$\sigma^2\varepsilon$	$\frac{d^2}{n\sigma^2\varepsilon}$
σ-sub-exponential	$\sigma^2\varepsilon^2$	$\frac{d^2}{n\sigma^2\varepsilon^2}$

Fig. 2: A summary of the (order-wise) upper bounds on $\mathsf{Tr}(I_Y(\theta))$ under different conditions on the score function random vector $S_\theta(X)$.

A. *Interactive Models*

So far we have assumed that there are no interactions between the different samples during privatization, so that the privatized samples $Y_1, \ldots, Y_n$ are independent. It is worth pointing out that our Fisher information bounds, and the corresponding lower bounds in the estimation error, can also be extended to interactive communication models where the privatized samples are no longer necessarily independent. We describe both a sequential interactive model [12] and a more general fully interactive blackboard model [14] below:

(i) **Sequential Interaction:** In this scenario, the samples $X_1, \ldots, X_n$ are ordered and the mechanism for sample i can depend on the previously privatized samples $Y_1, \ldots, Y_{i-1}$. Formally, there is a collection of ε differentially private mechanisms $Q_i(Y_i|x_i, y_1, \ldots, y_{i-1})$ such that

$$\frac{Q_i(S|x_i, y_1, \ldots, y_{i-1})}{Q_i(S|x_i', y_1, \ldots, y_{i-1})} \leq e^\varepsilon$$

for any $i = 1, \ldots, n$, event S, and $x_i, x_i', y_1, \ldots, y_{i-1}$. Using the chain rule for Fisher information,

$$\mathsf{Tr}(I_{Y_1,\ldots,Y_n}(\theta)) = \sum_{i=1}^{n} \mathbb{E}_{Y_1,\ldots,Y_{i-1}}\left[\mathsf{Tr}(I_{Y_i|Y_{i-1},\ldots,Y_1}(\theta))\right] \tag{9}$$

where $I_{Y_i|y_{i-1},\ldots,y_1}(\theta)$ denotes the Fisher information computed using the distribution for Y_i conditioned on $y_1, \ldots, y_{i-1}$. For a given θ, X_i is independent of $Y_1, \ldots, Y_{i-1}$, so conditioning on $y_1, \ldots, y_{i-1}$ just determines which mechanism is used, and because each mechanism satisfies the ε differentially private condition, each term inside the expectation in (9) can be

upper bounded just as in Propositions 2-5, and the total bound on the Fisher information from all n samples will remain the same.

(ii) **Fully Interactive Blackboard Model:** In this scenario, there are multiple rounds of sequential communication, and a public blackboard with all of the information released after each round is available to all nodes in future rounds of communication. Suppose there are T rounds of communication indexed by $t = 1, \dots, T$. Node i releases $Y_{i,t}$ on round t which is drawn from $Q_{i,t}(Y_{i,t}|b_1, \dots, b_{t-1}, y_{1,t}, \dots, y_{i-1,t}, x_i)$ where $B_t = (Y_{1,t}, \dots, Y_{n,t})$ is the blackboard that is visible to all nodes after round t. Call the total transcript after all rounds of communication $Z = (B_1, \dots, B_T)$. We assume that there is a total "privacy budget" of ε for each X_i in the sense that

$$\frac{\mathsf{Pr}(Z \in S|x_1, \dots, x_i, \dots, x_n)}{\mathsf{Pr}(Z \in S|x_1, \dots, x_i', \dots, x_n)} \leq e^{\varepsilon}$$

for any event S, i, $x_1, \dots, x_n$, x_i'. Note that by the chain rule, the conditional density of Z can be written as

$$\frac{Q(Z|x_1, \dots, x_i, \dots, x_n)}{Q(Z|x_1, \dots, x_i', \dots, x_n)} = \frac{\prod_t Q_{i,t}(Y_{i,t}|B_1, \dots, B_{t-1}, Y_{1,t}, \dots, Y_{i-1,t}, x_i)}{\prod_t Q_{i,t}(Y_{i,t}|B_1, \dots, B_{t-1}, Y_{1,t}, \dots, Y_{i-1,t}, x_i')}. \tag{10}$$

The total Fisher information from the transcript Z can be upper bounded by n times the bounds in Propositions 2-5 under the same conditions on the score $S_\theta(X)$, so that the same lower bounds in estimation error also apply for this more general interaction model. The details for this are found in Appendix E.

III. Applications

In this section we show how the upper bounds on $\mathsf{Tr}(I_Y(\theta))$ developed in the last section can be used to imply order-optimal lower bounds on the private estimation problem using (3). We are interested in characterizing the minimax risk

$$\inf_{(Q,\hat{\theta})} \max_{\theta \in \Theta} \mathbb{E}\|\hat{\theta}(Y_1, \dots, Y_n) - \theta\|_2^2$$

where the estimator $\hat{\theta}$ is a function of the privatized samples $Y_1, \dots, Y_n$, and the infimum is taken jointly over the estimator $\hat{\theta}$ and privatization mechanism Q. Upper bounds can therefore be found by jointly designing an estimator and privatization mechanism that achieve a certain worst-case error.

Corollary 1 (Gaussian location model)**.** *Suppose* $X_i \sim \mathcal{N}(\theta, \sigma_0^2 I_d)$ *and* $\Theta = [-B, B]^d$*. In this case* $S_\theta(X_i) \sim \mathcal{N}\left(0, \frac{1}{\sigma_0^2} I_d\right)$ *and the conditions for Proposition 2 and Proposition 4 are satisfied with* $I_0 = \frac{1}{\sigma_0^2}$ *and* $\sigma^2 = O\left(\frac{1}{\sigma_0^2}\right)$*, respectively. Using the van Trees inequality we have*

$$\max_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \geq c \frac{\sigma_0^2 d^2}{n \min\{\varepsilon^2, \varepsilon\}}$$

for an absolute constant c *if* $\frac{n \min\{\varepsilon^2, \varepsilon\}}{\sigma_0^2} \geq \frac{d}{B^2}$*.*

This lower bound for the Gaussian location model matches both the lower and upper bounds detailed in [14]. The condition on n is a mild technical condition that ensures the second term in the denominator of (3) will not dominate the order of the lower bound. We will make similar assumptions in the following examples.

Corollary 2 (discrete distribution estimation)**.** *Suppose* $\mathcal{X} = [1 : d+1]$ *and* $X_i \sim$ *Mult*$(1, \theta)$ *where* $\Theta = \{\theta \in \mathbb{R}^{d+1} \, : \, \sum_{i=1}^{d+1} \theta_i = 1\}$*. Then*

$$\max_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \geq c \frac{d}{n \min\{(e^\varepsilon - 1)^2, e^\varepsilon\}}$$

for an absolute constant c *if* $n \min\{(e^\varepsilon - 1)^2, e^\varepsilon\} \geq d^2$*.*

Corollary 2 follows by applying Propositions 2 and 3 with variance $I_0 \leq 6d$. The details are included in Appendix C. The lower bound for the discrete distribution example gives the same order as the upper and lower bounds from [12] when ε is close to zero, and it also matches the upper and lower bounds from [13] when $1 \preceq e^\varepsilon \preceq d$. Other mechanisms for this model are discussed in [24].

Corollary 3 (Sparse Bernoulli models)**.** *Suppose* $P_\theta = \prod_{j=1}^d \mathsf{Bern}(\theta_j)$*.*

(i) *Sparse Bernoulli: If* $\Theta = \left\{\theta \in [0,1]^d \, : \, \sum_{j=1}^d \theta_j \leq 1\right\}$ *and* $n \min\{(e^\varepsilon - 1)^2, e^\varepsilon\} \geq d^2$*, then*

$$\inf_{(Q, \hat{\theta})} \max_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \asymp \frac{d}{n \min\{(e^\varepsilon - 1)^2, e^\varepsilon, d\}} \, .$$

(ii) *s-Sparse Bernoulli: If* $\Theta = \left\{\theta \in [0,1]^d \, : \, \sum_{j=1}^d \theta_j \leq s\right\}$*,* $s \leq d^{1-\delta}$ *for some* $\delta > 0$*.*

- *High privacy regime: if* $\varepsilon \leq \log(d/s)$ *and*

$$\frac{n}{\log n} \geq \frac{20}{\min(\varepsilon^2, 1)} d^3 \log d$$

then

$$\inf_{(Q, \hat{\theta})} \sup_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \asymp \frac{sd}{n \min\left(e^\varepsilon, (e^\varepsilon - 1)^2, d\right)} \, . \tag{11}$$

- *Low privacy regime: if* $20 \log d \leq \varepsilon \leq s \log d$ *and* $n\varepsilon \geq d \log d$*, then*

$$\inf_{(Q,\hat{\theta})} \sup_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \asymp \frac{s^2 \log d}{n\varepsilon} \,. \tag{12}$$

We sketch the proof of Corollary 3 in the next section. The sparse Bernoulli example is illustrative of several interesting phenomena. In the sparse Bernoulli model (i), the minimax risk scales exactly the same as in the discrete distribution estimation problem. This shows that even in a statistical model with independent components, the dependence on ε can be exponential instead of linear. In this way, the scaling is dictated by the properties of the score function $S_\theta(X)$ rather than the independence of the model. In the s-sparse Bernoulli model (ii), we see that in the privacy range $\log d \preceq \varepsilon \preceq s \log d$, the minimax risk scales as $\frac{s^2 \log d}{n\varepsilon}$ instead of the centralized (i.e. without privacy constraints) rate $\frac{s}{n}$. This means that the sample size penalty is of the order $\frac{s \log d}{\varepsilon}$ for privacy. This is noteworthy in that the penalty scales linearly only with the sparsity s, rather than with the underlying dimension d, which is the case for, e.g, the sparse Gaussian location model.

Note that cases (i) and (ii) above focus on two different parameter regimes of the same model, and the domain Θ in (i) is a subset of that in (ii). As such, the lower bound in (i) can be regarded as a more local minimax risk bound, while the one in (ii) with s close to d can be regarded as a worst-case minimax bound. These two bounds together illustrate how having additional information that restricts the range of the parameter as in (i) can change the dependence of the risk on the privacy parameter ε.

IV. Proof of Corollary 3

A. Sparse Bernoulli (i)

The lower bound follows by applying Propositions 2 and 3 with a score function that satisfies $I_0 \leq 3d$, as we check below. We restrict our attention to $\Theta' = \left[\frac{1}{2d}, \frac{1}{d}\right]^d \subset \Theta$ using the fact that $\sup_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \geq \sup_{\theta \in \Theta'} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2$. The score function for each component is

$$S_{\theta_j}(x_j) = \frac{\partial}{\partial \theta_j} \log f(x_j|\theta_j) = \begin{cases} \frac{1}{\theta_i} & \text{, if } \; x_j = 1 \\ \frac{-1}{1-\theta_i} & \text{, if } \; x_j = 0 \,. \end{cases} \tag{13}$$

so that the variance of each component is

$$\mathbb{E}[S_{\theta_j}(x_j)^2] = \theta_j \frac{1}{\theta_j^2} + (1-\theta_j)\frac{1}{(1-\theta_j)^2} = \frac{1}{\theta_j} + \frac{1}{(1-\theta_j)}$$
$$\leq 3d\ .$$

By taking sums of independent variables we also have $\mathbb{E}[\langle u, S_\theta(x)\rangle^2] \leq 3d$ for any unit vector $u \in \mathbb{R}^d$ and $\theta \in \Theta'$, as desired.

For the upper bound, we demonstrate an estimator that works by reducing the problem to the discrete distribution estimation problem as follows. We perform the analysis for $\sum_i \theta_i = 1$, but the mechanism and the derivation also hold for $\sum_i \theta_i \leq 1$. Moreover, for any *constant* sparsity, say $\sum_i \theta_i \leq c$, we can always perform a randomized mapping to $\lceil c \rceil \cdot d$ symbols and reduce the problem to 1-Sparse Bernoulli problems (with $\lceil c \rceil$ repetition). Therefore the result holds for any constant sparsity case, i.e. $\sum_i \theta_i \leq c$. To convert the product Bernoulli model into the distribution estimation problem, define the mapping

$$f(X_k) = \begin{cases} i, & \text{if } \|X_k\|_1 = 1 \text{ and } X_k(i) = 1, \\ d+1, & \text{if } \|X_k\|_1 \neq 1. \end{cases}$$

Then $f(X_k)$ follows $(p_1, ..., p_{d+1})$, with

$$p_i = \theta_i \cdot \prod_{j \neq i}(1-\theta_j),\ \forall i \in [d],\ p_{d+1} = 1 - \sum_{j=1}^{d} p_j.$$

Also define $P_S \triangleq \prod_{j=1}^{d}(1-\theta_j)$, then we have

$$\theta_i = \frac{\theta_i \cdot \prod_{j\neq i}(1-\theta_j)}{\theta_i \cdot \prod_{j\neq i}(1-\theta_j) + \prod_{j=1}^{d}(1-\theta_j)} = \frac{p_i}{p_i + P_S}.$$

Therefore, our strategy is to estimate p_i and P_S separately, and the final estimator will be

$$\hat{\theta}_i \triangleq \frac{\hat{p}_i}{\hat{p}_i + \hat{P}_S}.$$

It remains to complete the descriptions of the estimators $\hat{p}_i$ and $\hat{P}_S$ and to analyze the error from this strategy.

For ease of analysis, we assume that $\theta_i \leq \frac{1}{2}$. for all $i \in \{1, ..., d\}$. Note that this assumption can be easily circumvented by using a randomized mapping $h : X_k \to \{0,1\}^{2d}$, such that if $X_k(i) = 1$ then set $h(X_k)_{2i}$ and $h(X_k)_{2i+1}$ to 1 with probability $\frac{1}{2}$, and otherwise set them to 0. Obviously $h(X_k)$ follows product Bernoulli with parameter $\frac{\theta_i}{2} \leq \frac{1}{2}$ in each dimension.

Discrete distribution estimator under LDP

First we review the distribution estimation problem with LDP constraint, where each node observes a sample $X_k \in \mathcal{X} = \{1, ..., d\}$ from a discrete distribution $p = (p_1, ..., p_d)$ and is allowed to transmit information under ε-local privacy constraint. The LDP mechanisms can be viewed as a pair of

- locally privatization mapping $Q_{\text{DE}}(y|i)$ that maps each observation X_k to $Y_k \in \mathcal{Y}$
- an estimator $\hat{p}\left(Y^n\right) = \left(\hat{p}_1\left(Y^n\right), ..., \hat{p}_k\left(Y^n\right)\right)$.

In particular, [13] propose the following privatization mapping that maps each X_k into $y \in \mathcal{Y}_{d,w} \triangleq \left\{y \in \{0,1\}^d : \sum_i y_i = w\right\}$ with the following transitional probability:

$$Q_{\text{DE}}(y|i) = \frac{e^{\varepsilon} y_i + (1 - y_i)}{e^{\varepsilon}\binom{d-1}{w-1} + \binom{d-1}{w}}.$$

The estimator is

$$\left(\frac{(d-1)e^{\varepsilon} + \frac{(d-1)(d-w)}{w}}{(d-w)(e^{\varepsilon}-1)}\right)\frac{T_i}{n} - \frac{(w-1)e^{\varepsilon} + d - w}{(d-w)e^{\varepsilon} - 1},$$

where $T_i \triangleq \sum_{k=1}^{n} Y_k(i)$.

Theorem 1 (Proposition III.1 [13])**.**

$$\mathbb{E}\|\hat{p}\left(Y^n\right) - p\|_2^2 = \frac{1}{n}\left(\frac{(w(d-2)+1)\,e^{2\varepsilon}}{(d-w)\left(e^{\varepsilon}-1\right)^2} + \frac{2(d-2)}{\left(e^{\varepsilon}-1\right)^2} + \frac{(d-2)(d-w)+1}{w\left(e^{\varepsilon}-1\right)^2} - \sum_i p_i^2\right). \quad (14)$$

For the low privacy regime $e^{\varepsilon} \succeq d$, we select $w = 1$ and the ℓ_2 estimation error is

$$\mathbb{E}\|\hat{p}\left(Y^n\right) - p\|_2^2 = O\left(\frac{e^{2\varepsilon}}{n\left(e^{\varepsilon}-1\right)^2}\right) = O\left(\frac{1}{n}\right). \quad (15)$$

For the regime $e^{\varepsilon} \prec d$, we select $w = \left[\frac{d}{e^{\varepsilon}+1}\right]$, (see [13, Proposition III.3]) and the ℓ_2 estimation error is given by

$$\Theta\left(\frac{d}{n\min\left\{\left(e^{\varepsilon}-1\right)^2, e^{\varepsilon}\right\}}\right).$$

The estimation error matches the lower bounds in both high and medium privacy regimes [12], [13] and thus is rate-optimal. For the low privacy regime, [13] coincides with the k-RR scheme from [24] and achieves optimal rate (i.e. $1/n$) too.

We will use the rate-optimal distribution estimators $\hat{p}$ to construct an estimator under sparse Bernoulli model with LDP constraint.

Estimating p_i: We use the first half of nodes to estimate p_i. For $k \in \{1, ..., n/2\}$, node k transmit Y_k according to $Q_{\text{DE}}(\cdot|f(X_k))$, and let $\hat{p}_i$ be the rate-optimal estimator of p_i as defined in previous subsection. The ℓ_2 estimation error of $\hat{p}$ is controlled by

$$\sum_{i=1}^{d} \mathbb{E}\left(\hat{p}_i - p_i\right)^2 \preceq \frac{d}{n \min\left\{e^{\varepsilon}, (e^{\varepsilon}-1)^2\right\}}. \tag{16}$$

We can truncate $\hat{p}_i$ and obtain a better estimator (i.e. with smaller ℓ_2 risk) since by definition p_i cannot take negative values:

$$\hat{p}_i^* \triangleq \max\left\{\hat{p}_i, 0\right\},$$

and thus $\hat{p}_i^* > 0$ almost surely.

Estimating P_S: The second half of nodes are used to estimate $P_S \triangleq \prod_{j=1}^{d}(1-\theta_j)$, which maps its observation X_k via

$$g(X_k) = \begin{cases} 1, & \text{if } \|X_k\|_1 = 0, \\ 0, & \text{else.} \end{cases}$$

Note that

- $g(X_k) \sim \text{Ber}(P_S)$
- P_S is lower bounded by some positive constant (see [18, Section 4.2]):

$$\begin{aligned} P_S = \prod_{i=1}^{d}(1-\theta_i) &\le \exp\left(-\sum_{i=1}^{d}\ln(1-\theta_i)\right) \\ &= \exp\left(-\sum_{i=1}^{d}\ln(1-\theta_i)\right) \\ &= \exp\left(-1-\sum_{t>1}\|\theta\|_t^t/t\right) \\ &\ge \exp\left(-1-\sum_{t>1}\|\theta\|_2^t/t\right) \\ &\ge \frac{1-\frac{1}{\sqrt{2}}}{\exp\left(1-\frac{1}{\sqrt{2}}\right)}. \end{aligned}$$

Therefore, for $k \in [n/2+1 : n]$, node k transmits Y_k according to $Q_{\text{DE}}(\cdot|g(X_k))$, and let $\hat{P}_S$ be the rate-optimal estimator of P_S. Estimating P_S is equivalent to distribution estimation problem with $d' = 2$ (which may falls into low privacy regime (15)), and by (14) the previous privatization scheme guarantees

$$\mathbb{E}\left(\hat{P}_S - P_S\right)^2 \preceq \frac{e^{2\varepsilon}}{n\left(e^{\varepsilon}-1\right)^2}. \tag{17}$$

Notice that if $e^{\varepsilon} \leq d$, then

$$\frac{e^{2\varepsilon}}{n\left(e^{\varepsilon}-1\right)^2} \leq \frac{de^{\varepsilon}}{n\left(e^{\varepsilon}-1\right)^2} \asymp \frac{d}{n\min\left\{e^{\varepsilon},(e^{\varepsilon}-1)^2\right\}}$$

(the last "$\asymp$" is derived by separating e^{ε} into $e^{\varepsilon} \succ 1$ and $e^{\varepsilon} \asymp 1$). Otherwise $e^{\varepsilon} \geq 2$ and (17) is $O(1/n)$. Since we already know that $P_S \geq \frac{1-\frac{1}{\sqrt{2}}}{\exp\left(1-\frac{1}{\sqrt{2}}\right)}$, the truncated estimator

$$\hat{P}_S^* \triangleq \max\left\{\hat{P}_S, \frac{1-\frac{1}{\sqrt{2}}}{\exp\left(1-\frac{1}{\sqrt{2}}\right)}\right\}$$

must have smaller ℓ_2 estimation error and is bounded

$$\frac{1}{\hat{P}_S^*} \leq \frac{\exp\left(1-\frac{1}{\sqrt{2}}\right)}{1-\frac{1}{\sqrt{2}}} \text{ almost surely.} \tag{18}$$

Analysis of ℓ_2 error of $\hat{\theta}$: Our final estimator is

$$\hat{\theta}_i\left(Y^n\right) \triangleq \frac{\hat{p}_i^*(Y_1,...,Y_{\frac{n}{2}})}{\hat{p}_i^*(Y_1,...,Y_{\frac{n}{2}}) + \hat{P}_S^*(Y_{\frac{n}{2}+1},...,Y_n)}.$$

The ℓ_2 error is

$$\begin{aligned}
\mathbb{E}\left(\hat{\theta}_i - \theta_i\right)^2 &= \mathbb{E}\left(\frac{\hat{p}_i^*}{\hat{p}_i^* + \hat{P}_S^*} - \frac{p_i}{p_i + P_S}\right)^2 \\
&= \mathbb{E}\left(\frac{\hat{p}_i^* - p_i}{\hat{p}_i^* + \hat{P}_S^*} + p_i\left(\frac{1}{\hat{p}_i^* + \hat{P}_S^*} - \frac{1}{p_i + P_S}\right)\right)^2 \\
&\leq 2\underbrace{\mathbb{E}\left(\frac{\hat{p}_i^* - p_i}{\hat{p}_i^* + \hat{P}_S^*}\right)^2}_{\text{(a)}} + 2p_i^2\underbrace{\mathbb{E}\left(\frac{1}{\hat{p}_i^* + \hat{P}_S^*} - \frac{1}{p_i + P_S}\right)^2}_{\text{(b)}}
\end{aligned}$$

By (16) and (18), (a) can be bounded by

$$\text{(a)} \leq 2\mathbb{E}\left[\left(\frac{1}{\hat{p}_i^* + \hat{P}_S^*}\right)^2\left(\hat{p}_i^* - p_i\right)^2\right] \leq 2\left(\frac{\exp\left(1-\frac{1}{\sqrt{2}}\right)}{1-\frac{1}{\sqrt{2}}}\right)^2 \mathbb{E}\left(\hat{p}_i^* - p_i\right)^2,$$

By (16), (17) and (18) (b) can be bounded by

$$(b) = 2p_i^2\mathbb{E}\left[\left(\frac{1}{\left(\hat{p}_i^* + \hat{P}_S^*\right)(p_i + P_S)}\right)^2 \left(\hat{p}_i^* - p_i + \hat{P}_S^* - P_S\right)^2\right]$$
$$\leq 4\left(\frac{\exp\left(1 - \frac{1}{\sqrt{2}}\right)}{1 - \frac{1}{\sqrt{2}}}\right)^4 p_i^2\left(\mathbb{E}\left(\hat{p}_i^* - p_i\right)^2 + \mathbb{E}\left(\hat{P}_S^* - P_S\right)^2\right).$$

Finally, summing over all $i \in [d]$, we have

$$\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq C_0 \sum_i \mathbb{E}\left(\hat{p}_i^* - p_i\right)^2 + C_1\mathbb{E}\left(\hat{P}_S^* - P_S\right)^2$$
$$\asymp \frac{d}{n\min\left\{e^\varepsilon, (e^\varepsilon - 1)^2\right\}},$$

where C_0 and C_1 are some universal constants.

B. s-Sparse Bernoulli (ii)

1) High privacy regime: The lower bound follows in the same way as that of (i) above, except focusing on $\frac{s}{2d} \leq \theta_j \leq \frac{s}{d}$ for each $j = 1, \ldots, d$. For the upper bound, we describe a scheme that achieves this error but requires at least sequential interaction between the nodes. The general idea is to group $\{\theta_1, ..., \theta_d\}$ into subgroups $\mathcal{G}_1, .., \mathcal{G}_s$, such that $\sum_{i\in\mathcal{G}_j}\theta_i = O(1)$. If we can do so, then for each sub-group we apply Part (i) of Corollary 3 with effective sample size $n_j' = n\left|\mathcal{G}_j\right|/d$, and the resulting ℓ_2 estimation error will be

$$\sum_{j=1}^{s} \frac{|\mathcal{G}_j|}{n_j' \min\{(e^\varepsilon - 1)^2, e^\varepsilon, d\}} = \frac{ds}{n\min\{(e^\varepsilon - 1)^2, e^\varepsilon, d\}}.$$

In order to grouping $\{\theta_1, ..., \theta_d\}$, in the first phase we estimate each of them up to precision $1/d$ with the first half of samples. This requires roughly $n \approx d^3/\min(\varepsilon^2, 1)$ samples. Once we obtain a coarse estimate of each θ_j, in the second phase we perform the mechanism for 1-Sparse Bernoulli model with the rest of the samples and refine the estimate.

Phase 1: grouping parameters: Let $n' = n/2d$, and by assumption

$$n' \geq 10\left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right)^2 d^2 \log d \log n.$$

We will use n' samples to estimate θ_j, $\forall j \in [d]$. For the j-th component of the i-sample $X_i(j) \sim \mathrm{Ber}(\theta_j)$, let $Y_i(j)$ be the ε-privatized version of it, so

$$Y_i(j) \sim \mathrm{Ber}\left(\left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right)\theta_j + \frac{1}{e^\varepsilon + 1}\right).$$

Therefore, by Hoeffding's inequality we have

$$\begin{aligned}\Pr\left\{\left|\hat{\theta}_j\left(Y^{n'}(j)\right)-\theta_j\right|\geq\frac{1}{d}\right\}&=\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n'}\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)\left(Y_i(j)-\mathbb{E}[Y_i(j)]\right)\right|\geq\frac{1}{d}\right\}\\&\leq 2\exp\left(-\frac{n'}{2d^2\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2}\right)\\&\leq\frac{1}{d^{10}n}.\end{aligned}$$

Let $\mathcal{E}=\left\{\exists j,\ \left|\hat{\theta}_j-\theta_j\right|\geq 1/d\right\}$ be the event of failure. By union bound,

$$\Pr\{\mathcal{E}\}\leq\frac{1}{nd^9}.$$

For the rest of analysis, we will condition on $\mathcal{E}^c$.

Since $\forall j,\ \hat{\theta}_j\leq\theta_j+1/d$, we must have $\sum_j\hat{\theta}_j\leq s+1$, and therefore we can find s groups $\mathcal{G}_1,..,\mathcal{G}_s$, such that

$$\forall k\in[s],\ \sum_{j\in\mathcal{G}_k}\hat{\theta}_j\leq 2.$$

On the other hand, $\forall j,\ \theta_j\leq\hat{\theta}_j+1/d$, so we also have

$$\forall k\in[s],\ \sum_{j\in\mathcal{G}_k}\theta_j\leq 3.$$

Phase 2: reducing to 1*-Bernoulli model:* Conditioning on $\mathcal{E}^c$ and applying Part (i) of Corollary 3 for each group $\mathcal{G}_j$, with effective sample size $n'_j=n\left|\mathcal{G}_j\right|/d$, the ℓ_2 estimation error is upper bounded by

$$\frac{|\mathcal{G}_j|}{n'_j\min\{(e^\varepsilon-1)^2,e^\varepsilon,d\}}=\frac{d}{n\min\{(e^\varepsilon-1)^2,e^\varepsilon,d\}}.$$

Summing over s groups yields

$$\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\middle|\mathcal{E}^c\right]\asymp\frac{ds}{n\min\{(e^\varepsilon-1)^2,e^\varepsilon,d\}}.\tag{19}$$

On the other hand, if phase 1 fails, we have a trivial upper bound:

$$\|\hat{\theta}-\theta\|_2^2\leq d.$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\right]&\leq\Pr\{\mathcal{E}\}\,\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\middle|\mathcal{E}\right]+\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\middle|\mathcal{E}^c\right]\\&\leq\frac{2}{nd^8}+\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\middle|\mathcal{E}^c\right]\\&\asymp\frac{ds}{n\min\{(e^\varepsilon-1)^2,e^\varepsilon,d\}},\end{aligned}$$

achieving the desired result.

2) Low privacy regime: Let us first derive the lower bound for estimation error. Restricting our attention to $\frac{s}{2d} \leq \theta_j \leq \frac{s}{d}$, the score for each component of the Bernoulli model (13) is sub-Gaussian with parameter

$$\sigma^2 \leq \frac{c_0 d^2}{s^2 \log\left(\frac{d}{s}\right)}.$$

This can be checked by letting

$$\sigma = \max\left\{ \frac{1}{\theta_j \sqrt{\log \frac{1}{\theta_j}}} \ , \ \frac{1}{(1-\theta_j)\sqrt{\log \frac{1}{(1-\theta_j)}}} \right\} \ ,$$

and then

$$\begin{aligned} \mathbb{E}\left[e^{\left(\frac{S_{\theta_j}(X_j)}{\sigma}\right)^2} \right] &= \theta_j e^{\left(\frac{1}{\theta_j \sigma}\right)^2} + (1-\theta_j) e^{\left(\frac{1}{(1-\theta_j)\sigma}\right)^2} \\ &\leq 2 \end{aligned}$$

and thus $S_\theta(X)$ is σ^2 sub-Gaussian. Then note that for $\theta_j \leq \frac{1}{2}$ the first term is the maximizer and

$$\sigma^2 = \frac{1}{\theta_j^2 \log \frac{1}{\theta_j}} \leq \frac{c_0 d^2}{s^2 \log\left(\frac{d}{s}\right)} \ .$$

Applying Propositon 4, we obtain the desired lower bound. In the rest of the section, we give an explicit construction of Q and $\hat{\theta}$ and characterize the error for this mechanism.

k-Randomized Response (k-RR) Scheme : If the support size of the input alphabet is k, k-RR scheme outputs the input symbol with probability $e^\varepsilon/(k-1+e^\varepsilon)$ and the rest of $k-1$ symbols with probability $1/(k-1+e^\varepsilon)$:

$$Q(y|x) = \begin{cases} \frac{e^\varepsilon}{(k-1)+e^\varepsilon}, & \text{if } y = x, \\ \frac{1}{(k-1)+e^\varepsilon}, & \text{if } y \neq x. \end{cases}$$

k-RR scheme works well for low privacy regime, i.e. when e^ε is large, and will be used later as our privatization mapping.

In general, as long as $e^\varepsilon \approx k$, the privatization error is with the same order of estimation error, so we can estimate the discrete distribution without increasing additional estimation error by too much.

LDP Scheme via Sub-sampling and k-RR: In our problem, for each node we aim to transmit its local observation X_k reliably to the fusion center, and with high probability there will be roughly s's 1 in X_k, so the expected number of possible X_k is roughly $\binom{d}{s}$. Unfortunately this means we need $\binom{d}{s} \approx \exp(s \log d)$ symbols to represent it, and notice that $\varepsilon \leq s \log d$, so we cannot send X_k reliably under ε-local privacy constraint.

To address this issue, we use *sub-sampling* trick to reduce the effective support size. First let k be the largest integer such that

$$\sum_{0 \leq i \leq k} \binom{d}{i} \leq \exp(\varepsilon - 10 \log d).$$

Notice that $k \asymp \frac{\varepsilon}{\log d}$ since

$$\sum_{0 \leq i \leq k} \binom{d}{i} \leq d \binom{d}{k} \leq \exp(k \log d + \log d),$$

so we must have

$$k \geq \frac{\varepsilon}{\log d} - 11 \succeq \frac{\varepsilon}{\log d}.$$

Next, for each local observation X_k, consider the sub-sampled version $\tilde{X}_k$ as follows:

$$\tilde{X}_k \triangleq \begin{cases} X_k, \text{ if } \|X\|_1 \leq k, \\ \text{randomly keep } k\text{'s 1 in } X, \text{ if } \|X_k\|_1 > k. \end{cases}$$

If we let R_k be the reciprocal of sampling rate $\frac{\max(\|X\|_1, k)}{k}$, then

$$\mathbb{E}\left[\mathbb{E}\left[R_k \cdot \tilde{X}_k(i) \middle| R_k\right]\right] = \mathbb{E}\left[\mathbb{E}\left[X_k(i) \middle| R_k\right]\right] = \theta_i.$$

Finally, each node transmits $\tilde{X}_k$ via k-RR scheme with privacy level ε:

$$Q\left(Y|\tilde{X}\right) = \begin{cases} \frac{e^\varepsilon}{(N-1)+e^\varepsilon}, \text{ if } Y = \tilde{X}, \\ \frac{1}{(N-1)+e^\varepsilon}, \text{ if } Y \neq \tilde{X}, \end{cases}$$

where N is the number of possible $\tilde{X}_k$:

$$N \triangleq \sum_{i \leq k} \binom{d}{i} \leq \exp(\varepsilon)/d^{10}.$$

We also use p_e to denote the probability of privatization error, i.e.

$$p_e \triangleq \Pr\left\{Q\left(Y|\tilde{X}\right) \neq \tilde{X}\right\} = \frac{N-1}{(N-1)+e^\varepsilon} \leq \frac{1}{d^{10}}.$$

Now we compute $\Pr\{Y_k(i)=1|R_k\}$ for some node k:

$$\begin{aligned}\Pr\{Y_k(i)=1|R_k\} =& \Pr\left\{\tilde{X}_k(i)=1\cap Y_k=\tilde{X}_k|R_k\right\}\\ &+\Pr\left\{Y_k(i)=1\cap\tilde{X}_k(i)=1\cap Y_k\neq\tilde{X}_k|R_k\right\}\\ &+\Pr\left\{Y_k(i)=1\cap\tilde{X}_k(i)\neq 1\cap Y_k\neq\tilde{X}_k|R_k\right\}\end{aligned}$$

It not hard to see that the first term is $(1-p_e)\cdot\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]$, and we further derive the second and the third terms:

$$\begin{aligned}&\Pr\left\{Y_k(i)=1\cap\tilde{X}_k(i)=1\cap Y_k\neq\tilde{X}_k|R_k\right\}\\ &=p_e\cdot\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]\cdot\Pr\left\{Y_k(i)=1|\tilde{X}_k(i)=1,Y_k\neq\tilde{X}_k,R_k\right\}\\ &=p_e\cdot\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]\frac{\sum_{1<i\leq k}\binom{d-1}{i-1}}{\sum_{0\leq i\leq k}\binom{d}{i}-1},\end{aligned}$$

and

$$\begin{aligned}&\Pr\left\{Y_k(i)=1\cap\tilde{X}_k(i)\neq 1\cap Y_k\neq\tilde{X}_k|R_k\right\}\\ &=p_e\cdot\left(1-\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]\right)\cdot\Pr\left\{Y_k(i)=1|\tilde{X}_k(i)\neq 1,Y_k\neq\tilde{X}_k,R_k\right\}\\ &=p_e\cdot\left(1-\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]\right)\frac{\sum_{1\leq i\leq k}\binom{d-1}{i}}{\sum_{0\leq i\leq k}\binom{d}{i}-1}.\end{aligned}$$

Summing the three terms together, we have

$$\Pr\{Y_k(i)=1|R_k\}=A\cdot\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]+B, \tag{20}$$

for some known constants A and B. Notice that $1\geq A\geq(1-2p_e)\geq(1-\frac{2}{d^{10}})$, and $B\leq p_e\leq\frac{1}{d^{10}}$, which implies $\mathbb{E}\left[Y_k(i)|R_k\right]\approx\mathbb{E}\left[\tilde{X}_k(i)|R_k\right]$.

If we know R_k, then $\hat{\theta}_i\triangleq R_k\cdot\frac{Y_k-B}{A}$ is an unbiased estimator of θ_i since

$$\begin{aligned}\mathbb{E}\hat{\theta}_i&=\mathbb{E}\left[R_k\cdot\frac{Y_k(i)-B}{A}\right]=\mathbb{E}\left[R_k\mathbb{E}\left[\frac{Y_k(i)-B}{A}\middle|R_k\right]\right]\\ &=\mathbb{E}\left[\mathbb{E}\left[X_k(i)\middle|R_k\right]\right]=\mathbb{E}X_k(i)=\theta_i.\end{aligned}$$

Unfortunately, we cannot directly obtain R_k since otherwise the LDP constraint will be violated. So instead, we replace R_k with an estimate of $\mathbb{E}[R_k]$, denoted as $\hat{\mu_R}$, in our estimator:

$$\hat{\theta}_i(Y_k) \triangleq \hat{\mu}_R \cdot \frac{Y_k(i) - B}{A}.$$

Notice that

$$\begin{aligned}\mathbb{E}[R_i] = \mathbb{E}\left[\frac{\max(\|X_i\|_1, k)}{k}\right] &= \mathbb{E}\left[\frac{\|X_i\|_1}{k}\right] + \mathbb{E}\left[\frac{\max(k - \|X_i\|_1, 0)}{k}\right] \\ &= \frac{s}{k} - \mathbb{E}\left[\frac{\max(k - \|X_i\|_1, 0)}{k}\right].\end{aligned}$$

Since $\mathbb{E}\left[\frac{\max(k-\|X_i\|_1,0)}{k}\right]$ is bounded by 1 and $\varepsilon = \Omega(1)$, by using first $\theta(s) = o(n)$ of samples, we can estimate $\mathbb{E}[R_i]$ to precision $1/s$ privately, i.e. $\mathbb{E}\left[(\hat{\mu}_R - \mathbb{E}[R_i])^2\right] \leq \frac{1}{s}$.

Our final estimator is the aggregation of $Y_1, ..., Y_n$:

$$\hat{\theta}(Y_1, ..., Y_n) \triangleq \frac{1}{n}\sum_{k=1}^{n} \hat{\mu}_R \frac{Y_k - B}{A},$$

where A, B are constants defined in (20) and are independent of θ. It remains to show that for this mechanism and estimator,

$$\mathbb{E}\|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \preceq \frac{s^2 \log d}{n\varepsilon}.$$

Analysis of ℓ_2 error: Now let us analyze the ℓ_2 error of $\hat{\theta}$. As stated in previous section, $\hat{\theta}_i$ is unbiased to θ_i, so

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} \hat{\mu}_R \frac{Y_k(i) - B}{A} - \theta_i\right)^2\right] \\ &= \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}\left[\left(\hat{\mu}_R \frac{Y_k(i) - B}{A} - \theta_i\right)^2\right] \\ &= \frac{1}{n}\mathbb{E}\left[\left(\hat{\mu}_R \frac{Y_1(i) - B}{A} - \theta_i\right)^2\right] \\ &\leq \frac{3}{n}\underbrace{\mathbb{E}\left[\left(\hat{\mu}_R \frac{Y_1(i) - B}{A} - \hat{\mu}_R \tilde{X}_1(i)\right)^2\right]}_{\text{(a)}} + \frac{3}{n}\underbrace{\mathbb{E}\left[\left(\hat{\mu}_R \tilde{X}_1(i) - R_1 \tilde{X}_1(i)\right)^2\right]}_{\text{(b)}} \\ &\qquad + \frac{3}{n}\underbrace{\mathbb{E}\left[\left(R_1 \tilde{X}_1(i) - \theta_i\right)^2\right]}_{\text{(c)}}.\end{aligned}$$

Note that (a) and (b) can be viewed as *privatization* errors due to the LDP constraint, and (c) is the estimation error. We bound (a), (b) and (c) separately.

To bound (a), we leverage the following facts

- $1 > A \geq (1 - \frac{2}{d^{10}})$,
- $0 < B < \frac{1}{d^{10}}$,
- $\Pr\left\{\tilde{X}_1(i) = Y_1(i)\right\} \geq 1 - p_e \geq 1 - \frac{1}{d^{10}}$.

$$
\begin{aligned}
&\mathbb{E}\left[\left(\hat{\mu}_R \frac{Y_1(i) - B}{A} - \hat{\mu}_R \tilde{X}_1(i)\right)^2\right] \\
&= \mathbb{E}\left[(\hat{\mu}_R)^2\right] \cdot \mathbb{E}\left[\mathbb{E}\left[\left(\frac{Y_1(i) - B}{A} - \tilde{X}_1(i)\right)^2 \Big| R_1\right]\right] \\
&\leq (\hat{\mu}_R)^2 \Pr\left\{\tilde{X}_1(i) = Y_1(i)\right\} \cdot \left(\left(\frac{1 - B - A}{A}\right)^2 + \left(\frac{B}{A}\right)^2\right) \\
&+ \mathbb{E}\left[(\hat{\mu}_R)^2\right] \Pr\left\{\tilde{X}_1(i) \neq Y_1(i)\right\} \cdot \left(\left(\frac{A - B}{A}\right)^2 + \left(\frac{A + B}{A}\right)^2\right) \\
&\leq \mathbb{E}\left[(\hat{\mu}_R)^2\right] \left(\left(\left(\frac{1 - B - A}{A}\right)^2 + \left(\frac{B}{A}\right)^2\right) + p_e \left(\left(\frac{A - B}{A}\right)^2 + \left(\frac{A + B}{A}\right)^2\right)\right) \\
&\leq \frac{C_0}{d^{10}} \mathbb{E}\left[(\hat{\mu}_R)^2\right].
\end{aligned}
$$

Note that, $\mathbb{E}\left[(\hat{\mu}_R)^2\right] \leq \mathbb{E}\left[R_1^2\right] + \mathsf{Var}\left(\hat{\mu}_R\right) \leq 2d^2$, so term (a) can eventually be bounded by

$$\text{(a)} \leq \frac{C_0}{d^8},$$

for some constant C_0. To bound (b), observe that

$$\begin{aligned}
\sum_{i=1}^{d}\mathbb{E}\left[\left(\hat{\mu}_R\tilde{X}_1(i)-R_1\tilde{X}_1(i)\right)^2\right] &= \sum_{i=1}^{d}\mathbb{E}\left[\tilde{X}_1(i)^2\cdot\left(\hat{\mu}_R-R_1\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{d}\tilde{X}_1(i)\right)\cdot\left(\hat{\mu}_R-R_1\right)^2\right] \\
&= k\cdot\mathbb{E}\left[\left(\hat{\mu}_R-R_1\right)^2\right] \\
&\leq 2k\cdot\left(\mathbb{E}\left[\left(\hat{\mu}_R-\mathbb{E}\left[R_1\right]\right)^2\right]+\frac{1}{k^2}\mathsf{Var}\left(\max\left(k,\left\|X_1\right\|_1\right)\right)\right) \\
&\overset{(1)}{\leq} \frac{2k}{s}+\frac{2}{k}\cdot\mathsf{Var}\left(\left\|X_1\right\|_1\right) \\
&\overset{(2)}{\leq} C_1 s/k,
\end{aligned}$$

where (1) is due to $\hat{\mu}_R$ is of precision $O(1/s)$ by using first $o(n)$ samples, and (2) is because $k \asymp \varepsilon/\log d \preceq s$. Finally we bound (c) as follow:

$$\begin{aligned}
\mathbb{E}\left[\left(R_1\tilde{X}_1(i)-\theta_i\right)^2\right] &= \mathbb{E}\left[R_1^2\tilde{X}_1^2(i)\right]-\theta_i^2 \\
&\leq \mathbb{E}\left[\mathbb{E}\left[R_1^2\tilde{X}_1(i)\middle|R_1\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[R_1X_1(i)\middle|R_1\right]\right] \\
&\leq \mathbb{E}\left[\left(\frac{\|X_1\|_1}{k}+1\right)X_1(i)\right] \\
&\leq \mathbb{E}\left[\frac{\|X_1\|_1}{k}X_1(i)\right]+\theta_i.
\end{aligned}$$

Combining (a), (b) and (c) and summing across all dimensions $i = 1, ..., d$, we obtain

$$\begin{aligned}
\mathbb{E}\|\hat{\theta}-\theta\|_2^2 &= \sum_{i=1}^{d}\mathbb{E}\left[\left(\hat{\theta}_i-\theta_i\right)^2\right] \\
&\leq \frac{1}{n}\left(C_0'\frac{1}{\left(d^7\right)^2}+C_1'\frac{s}{k}+C_2'\left(\frac{\mathbb{E}\left[\|X_1\|_1^2\right]}{k}+s\right)\right) \\
&\asymp \frac{s^2\log d}{n\varepsilon},
\end{aligned}$$

where in the last step we bound the second moment of Poisson binomial distribution by

$$\mathbb{E}\|X_1\|_1^2 = \left(\sum_i\theta_i\right)^2+\sum_i\theta_i(1-\theta_i) \leq s^2+s,$$

and observe that k is $\Omega\left(\frac{\varepsilon}{\log d}\right)$.

ACKNOWLEDGEMENTS

This work was supported in part by NSF award CCF-1704624 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

REFERENCES

[1] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, S. Halevi and T. Rabin, Eds. Springer, Berlin, Heidelberg, 2006.

[3] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proceedings of the 41st annual ACM symposium on Theory of computing*. ACM, 2009, pp. 371–380.

[4] C. Dwork, "Differential privacy," *Automata*, pp. 1–12, 2016.

[5] H. Cramér, *Mathematical Methods of Statistics*. Princeton Univ. Press, 1946.

[6] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, 1945.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[8] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.

[9] J. Hájek, "Local asymptotic minimax and admissibility in estimation," in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1972, pp. 175–194.

[10] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[11] R. D. Gill and B. Y. Levit, "Applications of the van trees inequality: a Bayesian cramér-rao bound," *Bernoulli*, vol. 1, no. 1/2, pp. 059–079, 1995.

[12] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[13] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Transactions on Information Theory*, vol. 64, no. 8, 2018.

[14] J. Duchi and R. Rogers, "Lower bounds for locally private estimation via communication complexity," in *Proceedings of the 32nd Conference On Learning Theory*, vol. 99. PMLR, 2019, pp. 1161–1191.

[15] L. P. Barnes, Y. Han, and A. Özgür, "A geometric characterization of fisher information from quantized samples with applications to distributed statistical estimation," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, 2018.

[16] ——, "Fisher information for distributed estimation under a blackboard communication protocol," *Proceedings of the IEEE International Symposium on Information Theory (ISIT), Paris, France*, 2019.

[17] ——, "Lower bounds for learning distributions under communication constraints via fisher information," arXiv preprint, arXiv*:1902.02890*, 2019.

[18] J. Acharya, C. Canonne, and H. Tyagi, "Distributed simulation and distributed inference," arXiv preprint, arXiv*:1804.06952*, 2018.

[19] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proceedings of the 32nd Conference on Learning Theory*, vol. 99. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 3–17.

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

[21] A. Rohde and L. Steinberger, "Geometrizing rates of convergence under local differential privacy constraints," arXiv preprint, arXiv*:1805.01422*, 2018.

[22] A. A. Borovkov, *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[23] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint* arXiv*:1011.3027*, 2010.

[24] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, New York, New York, USA, 20–22 Jun 2016, pp. 2436–2444.

APPENDIX

A. *Proof of Proposition 1*

Fix some $\delta > 0$. Suppose, for contradiction, that

$$\frac{Q(y|x)}{Q(y|x')} > e^{\varepsilon} + \delta$$

for all $y \in S$ with $\nu(S) > 0$. We have

$$\frac{Q(S|x)}{Q(S|x')} = \frac{\int_S Q(y|x) d\nu(y)}{\int_S Q(y|x') d\nu(y)} \geq \inf_{y \in S} \frac{Q(y|x)}{Q(y|x')} \geq e^{\varepsilon} + \delta \, .$$

This contradicts $Q(\cdot|\cdot)$ being an ε-differentially private mechanism, and thus we must have

$$\frac{Q(y|x)}{Q(y|x')} \leq e^{\varepsilon} + \delta$$

for ν-almost all y. Taking $\delta \to 0$ and using the measure's continuity from above completes the proof.

B. *Proof of Lemma 1*

$$\begin{aligned}
\mathsf{Tr}(I_Y(\theta)) &= \sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(Y|\theta)\right)^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\frac{\partial}{\partial \theta_i} f(Y|\theta)}{f(Y|\theta)}\right)^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\int Q(Y|x) \frac{\partial}{\partial \theta_i} f(x|\theta) d\mu(x)}{f(Y|\theta)}\right)^2\right] \qquad (21) \\
&= \sum_{i=1}^{d} \mathbb{E}\left[\left(\int \frac{Q(Y|x) f(x|\theta)}{f(Y|\theta)} \frac{\frac{\partial}{\partial \theta_i} f(x|\theta)}{f(x|\theta)} d\mu(x)\right)^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_Y \mathbb{E}_X\left[\frac{\partial}{\partial \theta_i} \log f(x|\theta) \middle| Y\right]^2 \\
&= \mathbb{E}_Y \|\mathbb{E}_X[S_\theta(X)|Y]\|_2^2 \, .
\end{aligned}$$

The key step (21) relies on interchanging integration over the sample space and differentiation with respect to the components θ_j which can be made precise via Lebesgue's Dominated Convergence Theorem as shown in Appendix D regarding regularity conditions.

C. *Proof of Corollary 2*

Without loss of generality we focus on a subset $\Theta' \subset \Theta$ defined by

$$\Theta' = \left\{ \theta \in \Theta \ : \ \frac{1}{4d} \leq \theta_i \leq \frac{1}{2d} \text{ for } i = 1, \ldots, d \right\} \ ,$$

and only consider the error from the first d components of θ_i. We can do this because

$$\max_{\theta \in \Theta} \mathbb{E} \|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|_2^2 \geq \max_{\theta \in \Theta'} \left[\mathbb{E} \sum_{i=1}^{d} \left(\hat{\theta}_i - \theta_i \right)^2 \right] \ .$$

It remains to show that for all $\theta \in \Theta'$ and unit vectors $u \in \mathbb{R}^d$,

$$\mathbb{E}[\langle u, S_\theta(X) \rangle^2] \leq 6d$$

where

$$\begin{aligned} S_\theta(X) &= (S_{\theta_1}(x), \ldots, S_{\theta_d}(x)) \\ &= \left(\frac{\partial}{\partial \theta_1} \log f(x|\theta), \ldots, \frac{\partial}{\partial \theta_d} \log f(x|\theta) \right) \end{aligned}$$

is the score from just the first d components. To see this note that

$$\theta_{d+1} = 1 - \sum_{i=1}^{d} \theta_i,$$

and

$$S_{\theta_i}(x) = \begin{cases} \frac{1}{\theta_i}, & x = i \\ -\frac{1}{\theta_{d+1}}, & x = d+1 \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, d$. Then for any unit vector $u = (u_1, \ldots, u_d)$,

$$\begin{aligned} \mathbb{E}[\langle u, S_\theta(X) \rangle^2] &= \sum_{x=1}^{d+1} \theta_x \left(\sum_{i=1}^{d} u_i S_{\theta_i}(x) \right)^2 \\ &= \theta_{d+1} \frac{1}{\theta_{d+1}^2} \left(\sum_{i=1}^{d} u_i \right)^2 + \sum_{x=1}^{d} \theta_x \left(\sum_{i=1}^{d} u_i S_{\theta_i}(x) \right)^2 \\ &\leq 2d + \sum_{x=1}^{d} \theta_x u_x^2 \frac{1}{\theta_x^2} \leq 6d \ . \end{aligned}$$

The corollary then follows by applying Propositions 2 and 3 with $I_0 = 6d$ to equation (3).

D. Regularity Conditions

We make the following assumptions on the statistical model P_θ:

(i) The density $f(x|\theta)$ is such that $\sqrt{f(x|\theta)}$ is continuously differentiable with respect to θ_j for $j = 1, \ldots, d$ and μ-almost all $x \in \mathcal{X}$. Note that this is the same as assuming that the density $f(x|\theta)$ itself is continuously differentiable if we assume that $f(x|\theta) > 0$, and this positivity assumption can always be made valid by considering all integrals to only be over the subset of $\mathcal{X}$ with $f(x|\theta) > 0$.

(ii) The Fisher information for each component $I_X(\theta_j) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_j} \log f(x|\theta)\right)^2\right]$ exists and is a continuous function of θ_j for each $j = 1, \ldots, d$.

It can easily be checked that for the Gaussian location model, discrete distribution estimation, and sparse Bernoulli models these conditions are met for an appropriate subset of the space of possible parameter values Θ.

These conditions are relatively standard sufficient conditions for a statistical model to be *differentiable in quadratic mean* [10]. Unfortunately the differentiable in quadratic mean condition itself is not appropriate for developing Cramér-Rao type lower bounds, and so it will not work for our purposes. One important aspect of these conditions is that we make assumptions on the statistical model P_θ, but not Q_θ, so that there are no implicit assumptions on the privacy mechanism.

Lemma 2. *Under the conditions above, $f(y|\theta)$ is continuously differentiable with respect to θ_j and*

$$\begin{aligned}\frac{\partial}{\partial \theta_j} f(y|\theta) &= \frac{\partial}{\partial \theta_j} \int Q(y|x) f(x|\theta) d\mu(x) \\ &= \int Q(y|x) \frac{\partial}{\partial \theta_j} f(x|\theta) d\mu(x)\end{aligned}$$

at ν-almost any y.

Proof. For simplicity we consider the scalar case with $d = 1$. The proof for each component of

the vector case is identical. Without yet knowing if the limit exists, formally we have

$$\begin{aligned}\frac{\partial}{\partial\theta}f(y|\theta) &= \lim_{h\to 0}\frac{1}{h}\left(\int Q(y|x)f(x|\theta+h)d\mu(x) - \int Q(y|x)f(x|\theta)d\mu(x)\right)\\ &= \lim_{h\to 0}\int Q(y|x)\int_0^1 f'(x|\theta+hu)du\,d\mu(x)\\ &= \lim_{h\to 0} 2\int\int_0^1 Q(y|x)\sqrt{f(x|\theta+hu)}\sqrt{f(x|\theta+hu)}'du\,d\mu(x)\ .\end{aligned}$$

For each h define the set

$$A_h = \left\{x\in\mathcal{X}\ :\ \sup_{v:|\theta-v|<h}\sqrt{f(x|v)} < 2\sqrt{f(x|\theta)}\quad,\quad \sup_{v:|\theta-v|<h}|\sqrt{f(x|v)}'| < 2|\sqrt{f(x|\theta)}'|\right\}$$

and split the integral into two terms, considering the set A_h and its complement A_h^C separately:

$$\int\int_0^1 Q(y|x)\sqrt{f(x|\theta+hu)}\sqrt{f(x|\theta+hu)}'du\,d\mu(x) \tag{22}$$

$$= \int_{A_h}\int_0^1 Q(y|x)\sqrt{f(x|\theta+hu)}\sqrt{f(x|\theta+hu)}'dud\mu(x) \tag{23}$$

$$+ \int_{A_h^C}\int_0^1 Q(y|x)\sqrt{f(x|\theta+hu)}\sqrt{f(x|\theta+hu)}'dud\mu(x). \tag{24}$$

To deal with term (23) we can use Lebesgue's dominated convergence theorem noting that

$$|1_{A_h}(x)Q(y|x)\sqrt{f(x|\theta+hu)}\sqrt{f(x|\theta+hu)}'| \le 4|Q(y|x)\sqrt{f(x|\theta)}\sqrt{f(x|\theta)}'|\ . \tag{25}$$

The right-hand side of display (25) is absolutely integrable by the Cauchy-Schwarz inequality:

$$\begin{aligned}\int|Q(y|x)\sqrt{f(x|\theta)}\sqrt{f(x|\theta)}'|d\mu(x) &\le \left(\int Q(y|x)^2 f(x|\theta)d\mu(x)\right)^{\frac{1}{2}}\left(\int\frac{f'(x|\theta)^2}{f(x|\theta)}d\mu(x)\right)^{\frac{1}{2}}\\ &\le f(y|\theta)\sqrt{e^{\varepsilon}I_X(\theta)}\ .\end{aligned}$$

This allows us to switch the limit inside the integral to get

$$\lim_{h\to 0}\int_{A_h^C}\int_0^1 Q(y|x)\sqrt{f(x|\theta+hu)}\sqrt{f(x|\theta+hu)}'dud\mu(x) = \int Q(y|x)\sqrt{f(x|\theta)}\sqrt{f(x|\theta)}'d\mu(x)$$

where we have used the continuity of $\sqrt{f(x|\theta)}$ and $\sqrt{f(x|\theta)}'$ to see that $1_{A_h}(x)\to 1$.

It remains to show that term (24) approaches zero as $h \to 0$. For this we again use the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\int_{A_h^C} \int_0^1 |Q(y|x)\sqrt{f(x|\theta+uh)}\sqrt{f(x|\theta+uh)}'|dud\mu(x) \\
&\qquad \leq \left(\int\int_0^1 Q(y|x)^2 f(x|\theta+uh)dud\mu(x)\right)^{\frac{1}{2}} \left(\int_{A_h^C}\int_0^1 \frac{f'(x|\theta+uh)^2}{f(x|\theta+uh)}dud\mu(x)\right)^{\frac{1}{2}} \\
&\qquad \leq e^{\varepsilon} f(y|\theta) \left(\int_{A_h^C}\int_0^1 \frac{f'(x|\theta+uh)^2}{f(x|\theta+uh)}dud\mu(x)\right)^{\frac{1}{2}}. \qquad (26)
\end{aligned}$$

The term inside the parentheses in (26) goes to zero as $h \to 0$ since

$$I_X(\theta) = \lim_{h\to 0}\left(\int_{A_h}\int_0^1 \frac{f'(x|\theta+uh)^2}{f(x|\theta+uh)}dud\mu(x) + \int_{A_h^C}\int_0^1 \frac{f'(x|\theta+uh)^2}{f(x|\theta+uh)}dud\mu(x)\right)$$

and

$$\int_{A_h}\int_0^1 \frac{f'(x|\theta+uh)^2}{f(x|\theta+uh)}dud\mu(x) \to I_X(\theta)$$

as $h \to 0$ using the dominated convergence theorem just as above. □

1) Applying the van Trees Inequality: In order to apply the van Trees inequality we will need

$$\int \left(\frac{\partial}{\partial\theta_j}\log f(y|\theta)\right) f(y|\theta)d\nu(y) = \int \frac{\partial}{\partial\theta_j} f(y|\theta)d\nu(y) = 0$$

for each $j = 1, \ldots, d$. In this subsection we check this condition under assumptions (i) and (ii) above regarding the distributions P_θ and their densities $f(x|\theta)$. We will make no further assumptions on $f(y|\theta)$ so that there are no implicit assumptions on the privacy mechanism Q other than it being a regular conditional distribution and an ε-differentially private mechanism.

Using Lemma 2 and the Fubini-Tonelli Theorem,

$$\begin{aligned}
\int \frac{\partial}{\partial \theta_j} f(y|\theta) d\nu(y) &= \int \frac{\partial}{\partial \theta_j} \int Q(y|x) f(x|\theta) d\mu(x) d\nu(y) \\
&= \int \int Q(y|x) \frac{\partial}{\partial \theta_j} f(x|\theta) d\mu(x) d\nu(y) \\
&= \int \int_{\left\{x: \frac{\partial}{\partial \theta_j} f(x|\theta) \geq 0\right\}} Q(y|x) \frac{\partial}{\partial \theta_j} f(x|\theta) d\mu(x) d\nu(y) \\
&\quad + \int \int_{\left\{x: \frac{\partial}{\partial \theta_j} f(x|\theta) < 0\right\}} Q(y|x) \frac{\partial}{\partial \theta_j} f(x|\theta) d\mu(x) d\nu(y) \\
&= \int_{\left\{x: \frac{\partial}{\partial \theta_j} f(x|\theta) \geq 0\right\}} \frac{\partial}{\partial \theta_j} f(x|\theta) d\mu(x) + \int_{\left\{x: \frac{\partial}{\partial \theta_j} f(x|\theta) < 0\right\}} \frac{\partial}{\partial \theta_j} f(x|\theta) d\mu(x) \\
&= \frac{\partial}{\partial \theta_j} \int f(x|\theta) d\mu(x) = 0 \ .
\end{aligned}$$

E. Blackboard Model

The density of the total transcript Z can be written as

$$\begin{aligned}
f(z|\theta) &= \mathbb{E}_{X_1,\ldots,X_n} \left[\prod_{i,t} Q_{i,t}(y_{i,t}|b_1, \ldots, b_{t-1}, y_{1,t}, \ldots, y_{i-1,t}, X_i) \right] \\
&= \prod_{i=1}^{n} \mathbb{E}_{X_1,\ldots,X_n} \left[\prod_{t} Q_{i,t}(y_{i,t}|b_1, \ldots, b_{t-1}, y_{1,t}, \ldots, y_{i-1,t}, X_i) \right] \\
&= \prod_{i=1}^{n} \mathbb{E}_{X_i} \left[p_{i,z}(X_i) \right]
\end{aligned}$$

where

$$p_{i,z}(x_i) = \prod_{t} Q_{i,t}(y_{i,t}|b_1, \ldots, b_{t-1}, y_{1,t}, \ldots, y_{i-1,t}, x_i) \ .$$

The score for this total transcript has components

$$\frac{\partial}{\partial \theta_j} \log f(z|\theta) = \sum_{i=1}^{n} \frac{\mathbb{E}_{X_i} \left[S_{\theta_j}(X_i) p_{i,z}(X_i) \right]}{\mathbb{E}_{X_i} \left[p_{i,z}(X_i) \right]} \ .$$

To get the above display we require interchanging differentiation and integration just like in the proof of Lemma 1. The trace of the Fisher information from the whole transcript is thus

$$\begin{aligned}
\mathsf{Tr}(I_Z(\theta)) &= \sum_{j=1}^{d} \mathbb{E}_Z \left[\left(\frac{\partial}{\partial \theta_j} \log f(Z|\theta) \right)^2 \right] \\
&= \mathbb{E}_Z \left[\sum_{i,j} \left(\frac{\mathbb{E}_{X_i} \left[S_{\theta_j}(X_i) p_{i,Z}(X_i) \right]}{\mathbb{E}_{X_i} \left[p_{i,Z}(X_i) \right]} \right)^2 \right] \qquad (27)
\end{aligned}$$

where (27) follows because the cross terms

$$\mathbb{E}_Z\left[\frac{\mathbb{E}_{X_i}\left[S_{\theta_j}(X_i)p_{i,Z}(X_i)\right]}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\frac{\mathbb{E}_{X_k}\left[S_{\theta_k}(X_k)p_{k,Z}(X_k)\right]}{\mathbb{E}_{X_k}\left[p_{k,Z}(X_k)\right]}\right]$$
$$=\mathbb{E}_{X_i,X_k}\left[S_{\theta_j}(X_i)S_{\theta_j}(X_j)\right]=0$$

for $i\neq k$.

1) Blackboard Proposition 2: Let

$$u_Z=\frac{\mathbb{E}_{X_i}\left[S_\theta(X_i)\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]}{\left\|\mathbb{E}_{X_i}\left[S_\theta(X_i)\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]\right\|_2}.$$

Following from (27),

$$\mathsf{Tr}(I_Z(\theta))=\sum_{i=1}^n\mathbb{E}_Z\left[\left\langle u_Z,\mathbb{E}_{X_i}\left[S_\theta(X_i)\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]\right\rangle^2\right]$$
$$=\sum_{i=1}^n\mathbb{E}_Z\left[\mathbb{E}_{X_i}\left[\left\langle u_Z,S_\theta(X_i)\right\rangle\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]^2\right].$$

We split up the expectation over X_i as follows:

$$\mathbb{E}_{X_i}\left[\left\langle u_Z,S_\theta(X_i)\right\rangle\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]$$
$$\leq\frac{1}{\min_x p_{i,Z}(x)}\mathbb{E}_{X_i}\left[\left\langle u_Z,S_\theta(X_i)\right\rangle p_{i,Z}(X_i)\right]$$
$$=\frac{1}{\min_x p_{i,Z}(x)}\int_{\{x:\langle u_Z,S_\theta(x)\rangle\geq 0\}}\left\langle u_Z,S_\theta(x)\right\rangle p_{i,Z}(x)f(x|\theta)d\mu(x)$$
$$+\frac{1}{\min_x p_{i,Z}(x)}\int_{\{x:\langle u_Z,S_\theta(x)\rangle<0\}}\left\langle u_Z,S_\theta(x)\right\rangle p_{i,Z}(x)f(x|\theta)d\mu(x)$$
$$\leq(e^\varepsilon-1)\int_{\{x:\langle u_Z,S_\theta(x)\rangle\geq 0\}}\left\langle u_Z,S_\theta(x)\right\rangle f(x|\theta)d\mu(x)$$

so that

$$\mathbb{E}_{X_i}\left[\left\langle u_Z,S_\theta(X_i)\right\rangle\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]^2\leq I_0(e^\varepsilon-1)^2$$

and

$$\mathsf{Tr}(I_Z(\theta))\leq nI_0(e^\varepsilon-1)^2$$

as desired.

2) Blackboard Proposition 3: Using Jensen's inequality,

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}_Z\left[\mathbb{E}_{X_i}\left[\left\langle u_Z, S_\theta(X_i)\right\rangle\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]^2\right]&\\ \leq\sum_{i=1}^{n}\mathbb{E}_Z\left[\mathbb{E}_{X_i}\left[\left\langle u_Z, S_\theta(X_i)\right\rangle^2\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]\right]&\\ \leq nI_0e^{\varepsilon}&\end{aligned}$$

where the last step uses (10) and the blackboard differential privacy condition.

3) Blackboard Proposition 4: By the convexity of $x\mapsto e^{x^2}$,

$$\begin{aligned}\exp\left(\left(\frac{\mathbb{E}_{X_i}\left[\left\langle u_Z, S_\theta(X_i)\right\rangle\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]}{\sigma}\right)^2\right)&\leq\mathbb{E}_{X_i}\left[\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\exp\left(\left(\frac{\left\langle u_Z, S_\theta(X_i)\right\rangle}{\sigma}\right)^2\right)\right]\\ &\leq 2e^{\varepsilon}\ .\end{aligned}$$

Taking logs,

$$\left(\mathbb{E}_{X_i}\left[\left\langle u_Z, S_\theta(X_i)\right\rangle\frac{p_{i,Z}(X_i)}{\mathbb{E}_{X_i}\left[p_{i,Z}(X_i)\right]}\right]\right)^2\leq\sigma^2(\varepsilon+\log 2)\ .$$

Blackboard Proposition 5 follows in the same way.