Locally differentially private estimation of nonlinear functionals of discrete distributions

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of estimating non-linear functionals of discrete distributions in the context of local differential privacy. The initial data $x_1, \ldots, x_n \in [K]$ are supposed i.i.d. and distributed according to an unknown discrete distribution $p = (p_1, \ldots, p_K)$. Only α-locally differentially private (LDP) samples $z_1, ..., z_n$ are publicly available, where the term 'local' means that each z_i is produced using one individual attribute x_i. We exhibit privacy mechanisms (PM) that are interactive, i.e. they are allowed to use already published confidential data, or non-interactive. We describe the behavior of the squared error risk for estimating the power sum functional $F_\gamma = \sum_{k=1}^{K} p_k^\gamma$, $\gamma > 0$ as a function of K, n and α. In the non-interactive case, we study several plug-in type estimators of F_γ, for all $\gamma > 0$, that are similar to the MLE which has been analyzed by Jiao *et al.* [14] in the multinomial model. However, due to the privacy constraint the rates we attain are slower and similar to those obtained in the Gaussian model by Collier *et al.* [6]. In the interactive case, we introduce for all $\gamma > 1$ a two-step procedure which attains the faster parametric rates $(n\alpha^2)^{-1/2}$ when $\gamma \geq 2$. We give lower bounds results over all $\alpha-$LDP privacy mechanisms and over all estimators using the private samples.

1 Introduction

Information theoretic measures have become of utmost importance and extensively used in image processing, physics, genetics and more recently in machine learning and statistics. Indeed, such functionals of probability distributions are useful to design estimators, to choose most informative features in different algorithms, to test identity or closeness of distributions. Most popular such measures are the power-sum, the entropy and more general Rényi entropies of discrete distributions.

In this work, we study the effect that keeping the original data private has on the estimators of such information theoretic measures and, in particular, on the power sum functional of a discrete distribution. Differential privacy [9] has prevailed in the recent literature as the most convenient approach to randomize samples and still have a control of the amount $\alpha > 0$ of privacy that has been introduced. The randomized samples, also known as private samples, are provided to the users who want to take out of them the largest amount of information on the original underlying distribution of the now confidential sample. Local differential privacy (LDP) is a particular setup where no trusted curator or third party is needed. Indeed, each private sample has been produced using a privacy mechanism that has access to only one individual of the original sensitive data. However we distinguish local differential privacy mechanisms that are interactive, i.e. they are allowed to use publicly available data (already privatized), or not-interactive.

It is quite popular now, that privacy, in particular local differential privacy, comes at the cost of slower rates for learning and testing methods in many cases. There is a rush nowadays to better understand

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

when the loss is unavoidable and to show that such loss is optimal over all privacy mechanisms and procedures at hand. More concretely, estimation of the probability density is known to be achieved with slower rates in the minimax sense under $\alpha-$LDP constraints, see [7], [19], [16] and [3]. However, estimating a power sum functional with power $\gamma = 2$ (quadratic functional) is possible with parametric rates in some cases [4]. Moreover, this paper first showed that interactive methods attain faster rates than non-interactive ones. This has also been proved for identity testing of discrete distributions in [2].

In order to describe phase transitions for the estimation problem of nonlinear functionals under LDP constraints, we consider the power sum functionals F_γ of discrete distributions $p = (p_1, ..., p_K)$ with power $\gamma > 0$: $F_\gamma = \sum_{k=1}^K p_k^\gamma$. We compare to the phase transitions known for non private data and, within the private setup, distinguish non-interactive and interactive privacy mechanisms.

1.1 Related literature

In the non-private setting we distinguish different behaviours for estimating F_γ according to the observations scheme, namely faster rates are attained in the multinomial setup above, than in the Gaussian vector/sequence model below.

Multinomial setting: When the r.v. $x_1, ..., x_n$ are observed, [14] show that the worst case squared error risk of the maximum likelihood estimator (MLE) in estimating F_γ, $\gamma > 0$, is

$$\frac{K^2}{n^{2\gamma}}\mathbf{1}_{\gamma\in(0,1)} + \frac{K^{2(1-\gamma)}}{n}\mathbf{1}_{\frac{1}{2}<\gamma<1} + n^{-2(\gamma-1)}\mathbf{1}_{1<\gamma<\frac{3}{2}} + \frac{1}{n}\mathbf{1}_{\gamma\geq\frac{3}{2}}.$$

Recall that in this model, the MLE of p_k is the average of $\mathbf{1}_{x_i=k}$ over i from 1 to n. These results entail that the MLE achieves the minimax rate n^{-1} when $\gamma \geq 3/2$. However, when the regularity γ decreases below $3/2$, the rates of the MLE get slower. In particular, the MLE does not achieve the minimax rates of estimation of F_γ when $0 < \gamma < 3/2$. For $0 < \gamma < 1$, the difficulty of estimation increases, as the alphabet size K may become problematic and prevent from a consistent estimation of F_γ, in the sense that the risk of the MLE converges to zero. The MLE consistently estimates F_γ, $0 < \gamma < 1$, if and only if, the sample size is large enough to satisfy $n \gg K^{1/\gamma}$.

For the values of $\gamma < 3/2$, the minimax rates roughly correspond to the performance of the MLE with n replaced by $n \log n$ [13]. These faster rates have been attained through a different estimator that uses best polynomial approximation of the functional p_k^γ for small values of p_k, which allows to gain a log factor in the bias. This method has been largely employed in the literature since the pioneer work by [15], in the Gaussian white noise model. The best polynomial approximation technique has been used for estimating the $\mathbb{L}_1$ norm in a Gaussian setting by [5], the $\mathbb{L}_r$ norm in the Gaussian white noise model by [11], the entropy of discrete distributions by [20], the entropy of lipschitz continuous densities by [12]. In the multimnomial setting, [1] studied the estimation of the Rényi entropy of order $a > 0$ and [10] extended these results to more general additive functionals of the form $\sum_{k=1}^K \varphi(p_k)$, for 4-times differentiable functions φ.

In the *Gaussian vector setting* the observations are $y_i = \theta_i + \epsilon \cdot \xi_i$, $i = 1, \ldots, K$ where $(\theta_1, \ldots, \theta_K)$ are unknown parameters, $\epsilon > 0$ is known, and ξ_i are i.i.d. standard Gaussian random variables. [6] show that there exists an estimator based on best polynomial approximation with the following error bound (see (14) in the proof of their Theorem 1)

$$\mathbb{E}\left[\left(\hat{F}_\gamma - F_\gamma\right)^2\right] \lesssim \frac{\epsilon^{2\gamma}K^2}{\log^\gamma(K)} + \frac{\epsilon^2}{\log(K)}\left(\sum_{i=1}^K \theta_i^{\gamma-1}\right)^2 \mathbb{1}_{\gamma>1} + \epsilon^2 \sum_{i=1}^K \theta_i^{2\gamma-2}\mathbb{1}_{\gamma>1} \tag{1}$$

for any $\gamma > 0$. When $\gamma \in (0, 1)$, they show that their bound $\epsilon^{2\gamma}K^2/\log^\gamma(K)$ is minimax optimal. Besides, [6] show that better rates can be achieved if γ is an integer. Namely, there exists an estimator with the following error bound, see Theorem 2 in [6],

$$\mathbb{E}\left[\left(\hat{F}_\gamma - F_\gamma\right)^2\right] \lesssim \epsilon^{2\gamma}K + \epsilon^2\|\theta\|_{2\gamma-2}^{2\gamma-2} \tag{2}$$

for any integer $\gamma \geq 1$.

1.2 Our contributions

To the best of our knowledge, in the context of privacy only the power sum functional with $\gamma = 2$ has been solved, for continuous distributions in [4] and can be easily inferred for discrete distributions, see [2].

We study here for the first time the estimation of nonlinear and nonquadratic power-sum functionals F_γ, that are also nonsmooth for γ in (0,1). We privatize the initial sample via the non-interactive Laplace mechanism and use the resulting data to build two estimators. First, the analogue of the MLE in the multinomial model is described and studied. Averages of the binned and privatized samples form an unbiased estimator of the vector p. However, the resulting estimator of F_γ is not a MLE in this setup and is therefore called a plug-in estimator. Even though our plug-in presents similarities with the MLE in the multinomial setting [14], we note that the rates attained by our plug-in are slower than the rates obtained by the MLE [14] and similar, slightly faster in some cases, to those obtained in the Gaussian sequence model by [6]. Our rates benefit from the constraint that p is a probability vector (belongs to the simplex).

Next, we notice that small probabilities p_k, below the fluctuation level of their unbiased estimator, need not be estimated. Indeed, the bias introduced by skipping those values is smaller than the squared error risk of the plug-in estimator in estimating these small probabilities. This leads to the thresholded procedure, where half of the sample is used to detect significant probabilities, and the other half is used to estimate the functional associated to the significant values. This thresholded procedure outperforms the plug-in typically when the alphabet size K is large.

In the interactive case for any $\gamma > 1$, we mimick the two-step optimal procedure of [4]. We attain faster rates in this case than those attained by the thresholded estimator in the non-interactive case. The improvement is at the level of the log factors induced by the thresholding in the non-interactive case. This two-step procedure uses half of the non-interactive private sample to produce the unbiased estimator of p. It then encodes this information in the private versions of the second half of the sample in such a way that they produce an estimator of $\sum_{k=1}^{K} p_k \cdot p_k^{\gamma-1}$.

We sum up in Table 1 these upper bounds. We also provide lower bounds results that are optimal for $\gamma \geq 2$ and nearly optimal by a factor K^γ for $\gamma \in (0, 2)$.

These procedures can be easily extended to the estimation of other information theoretic measures such as the entropy and the Rényi entropy, in the context of privacy.

2 Problem and notation

We assume that the sensitive random variables $x_1, ..., x_n$ are i.i.d. distributed according to the discrete density model

$$\mathbb{P}(x_i = k) = p_k, \qquad k = 1, \ldots, K, \tag{3}$$

where $p = (p_1, \ldots, p_K)$ is an unknown vector of parameters. The $x_1, \ldots, x_n$ are not publicly available and instead they are used as inputs into a privacy mechanism (PM), in order to produce available sanitized observations $Z_1, ..., Z_n$. Using the sanitized observations, we estimate the functional

$$F_\gamma = \sum_{k=1}^{K} p_k^\gamma \ ,$$

where γ is a real number in $(0, \infty)$ and where p belongs to

$$\mathcal{P}_K = \left\{ p \in [0,1]^K \ : \sum_{k=1}^{K} p_k = 1 \right\} \ .$$

The PM also known as channel or randomization, is submitted to the constraint that it is $\alpha-$locally differentially private (LDP) for some $\alpha > 0$, that is it generates the private samples Z_i using the conditional distribution $Q_i(\cdot|x_i, Z_1, ..., Z_{i-1})$ such that

$$\sup_A \sup_{Z_1,\ldots,Z_{i-1},x_i,x_i'} \frac{Q_i(A|x_i, Z_1, ..., Z_{i-1})}{Q_i(A|x_i', Z_1, ..., Z_{i-1})} \leq e^\alpha, \quad \text{for all } i = 1, ..., n.$$

In particular, a non-interactive PM generates each Z_i via the conditional distribution $Q_i(\cdot|x_i)$, without knowledge of previously published samples, as opposed to the more general ones above that are called interactive.

3 Non-interactive private samples and estimators

We first present a plug-in estimator based on a non-interactive mechanism. With the objective of reducing the dependence of the risk on the alphabet size K, we then introduce a thresholded estimation procedure that is built on the previous plug-in estimator.

3.1 Non-interactive PM

Given the original data x_i, individual i generates a random vector $Z_i = (z_{i1}, \dots, z_{iK})$ using the Laplace non-interactive privacy mechanism $Q^{(NI)}$ defined by

$$Q^{(NI)}: \qquad z_{ik} = \mathbf{1}_{\{x_i=k\}} + \frac{\sigma}{\alpha} \cdot w_{ik}, \qquad k = 1, \dots, K, \tag{4}$$

where the w_{ik} are i.i.d. Laplace distributed with density $f^w(x) = \frac{1}{2}\exp(-|x|)$. Note that w_{ik} are all centered, with variance 2, and we put $\sigma = 2$. One can readily check that the above channel $Q^{(NI)}$ is an α-LDP non-interactive mechanism, see also [2]. We denote by $\hat{z}_k = \frac{1}{n}\sum_{i=1}^n z_{ik}$, and similarly $\hat{w}_k = \frac{1}{n}\sum_{i=1}^n w_{ik}$.

3.2 Plug-in estimator

Denoting by $F_\gamma(k) = p_k^\gamma$ the k-th term in the sum F_γ, we have $F_\gamma = \sum_{k=1}^K F_\gamma(k)$. Let us study the plug-in estimator of F_γ

$$\hat{F}_\gamma = \sum_{k=1}^K \hat{F}_\gamma(k), \text{ where } \hat{F}_\gamma(k) = \left(T_{[0,2]}\left[\hat{z}_k\right]\right)^\gamma, \tag{5}$$

and using the clipping operation $T_{[0,2]}[\cdot]$ defined by $T_{[0,2]}[y] = (y \vee 0) \wedge 2$.

Note that $\hat{z}_k$ is an unbiased estimator of p_k with fluctuation of order

$$\sqrt{\frac{p_k(1-p_k)}{n}} + \frac{\sigma}{\sqrt{\alpha^2 n}}.$$

We denote by $\tau = c/\sqrt{(\alpha^2 \wedge 1)n}$ for some numerical constant $c > 0$ which can be seen as the noise level in this data. We write $p^{\geq\tau}$ (respectively $p^{<\tau}$) the vector containing thresholded values of p: $p_k \cdot \mathbf{1}(p_k > \tau)$ (respectively $p_k \cdot \mathbf{1}(p_k \leq \tau)$) for all k in $[K]$.

Theorem 3.1. *Let $p \in \mathcal{P}_K$, and $\alpha, \gamma \in (0, \infty)$. Then, the estimator $\hat{F}_\gamma$ in (5) has*

$$\begin{aligned} |\,\mathbb{E}_p\,\hat{F}_\gamma - F_\gamma| &\lesssim_\gamma \frac{K}{((\alpha^2 \wedge 1)n)^{\gamma/2}} + \mathbf{1}_{\{\gamma\geq 2\}}\frac{\|p^{\geq\tau}\|_{\gamma-2}^{\gamma-2}}{(\alpha^2 \wedge 1)n} \\ \mathrm{Var}(\hat{F}_\gamma) &\lesssim_\gamma \frac{K}{((\alpha^2 \wedge 1)n)^{\gamma}} + \mathbf{1}_{\{\gamma\geq 1\}}\frac{\|p^{\geq\tau}\|_{2\gamma-2}^{2\gamma-2}}{(\alpha^2 \wedge 1)n} \end{aligned}$$

and the quadratic risk is bounded by

$$\mathbb{E}_p\left[(\hat{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \frac{K^2}{((\alpha^2 \wedge 1)n)^{\gamma}} + \mathbf{1}_{\{\gamma\geq 2\}}\frac{\|p^{\geq\tau}\|_{\gamma-2}^{2(\gamma-2)}}{((\alpha^2 \wedge 1)n)^{2}} + \mathbf{1}_{\{\gamma\geq 1\}}\frac{\|p^{\geq\tau}\|_{2\gamma-2}^{2\gamma-2}}{(\alpha^2 \wedge 1)n} \tag{6}$$

where $\lesssim_\gamma$ means the inequality holds up to constants depending only on γ.

The proof of the theorem is in Appendix.

Remark Theorem 3.1 implies a tighter upper bound, up to log factors, than (1) from [6] in the Gaussian setting with noise $\epsilon = 1/\sqrt{(\alpha^2 \wedge 1)n}$. Indeed, $\|p^{\geq \tau}\|_{\gamma-2}^{\gamma-2}/(\alpha^2 \wedge 1)n$ is bounded from above by

$$\frac{\sum_{k=1}^K p_k^{\gamma-2} \mathbb{1}_{\{p_k > \tau\}}}{(\alpha^2 \wedge 1)n} \leq \sum_{k=1}^K \frac{p_k^{\gamma-1}}{\sqrt{(\alpha^2 \wedge 1)n}} \frac{p_k^{-1} \mathbb{1}_{\{p_k > \tau\}}}{\sqrt{(\alpha^2 \wedge 1)n}} \leq \sum_{k=1}^K \frac{p_k^{\gamma-1} c^{-1}}{\sqrt{(\alpha^2 \wedge 1)n}} = \frac{\|p\|_{\gamma-1}^{\gamma-1} c^{-1}}{\sqrt{(\alpha^2 \wedge 1)n}}$$

which corresponds to the squared root of the second term of (1).

Although the Laplace privacy mechanism (4) produces homoscedastic random variables having similar characteristics to those of the Gaussian setting [6], the rate (1) from [6] is not directly comparable to our rate since we benefit from the extra constraint that p is a probability vector, i.e. $p \in \mathcal{P}_K$ (whereas the vector of parameters in [6] lives in $\mathbb{R}^K$).

Note that for $\gamma \geq 2$ we have $p_k^{\gamma-1} \leq p_k$ so that we can further bound $\|p\|_{\gamma-1}^{\gamma-1} \leq \sum_{k=1}^K p_k = 1$ using the fact that $(p_k)_k$ are probabilities. Therefore the second term in the upper bound of the bias is smaller than $1/\sqrt{(\alpha^2 \wedge 1)n}$.

The number of the significant probabilities $p_k \geq \tau$ is smaller than $\tau^{-1} \asymp \sqrt{(\alpha^2 \wedge 1)n}$ since p is a vector of probability. With the notation $K_{\wedge \tau^{-1}} := K \wedge \sqrt{(\alpha^2 \wedge 1)n}$, we take the supremum over all $p \in \mathcal{P}_K$ in (6) to get the following corollary.

Corollary 3.2. *Let $p \in \mathcal{P}_K$, and $\alpha, \gamma \in (0, \infty)$. Then, the quadratic risk is uniformly bounded over all $p \in \mathcal{P}_K$ by*

$$\sup_{p \in \mathcal{P}_K} \mathbb{E}\left[(\hat{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \frac{K^2}{((\alpha^2 \wedge 1)n)^\gamma} + \mathbf{1}_{\{\gamma \geq 1\}} \frac{1 \vee K_{\wedge \tau^{-1}}^{3-2\gamma}}{(\alpha^2 \wedge 1)n} \quad .$$

Reciprocally, we show a lower bound on the squared error risk of the plug-in estimator in the next theorem. The proof can be found in Appendix.

Theorem 3.3. *Let $p \in \mathcal{P}_K$, and $\alpha, \gamma \in (0, \infty)$. Then, the quadratic risk of $\hat{F}_\gamma$ is asymptotically lower bounded by*

$$\mathbb{E}\left[(\hat{F}_\gamma - F_\gamma)^2\right] \gtrsim_\gamma \frac{K^2}{((\alpha^2 \wedge 1)n)^\gamma} \qquad \text{as } n \to \infty \ .$$

Therefore, it is necessary and sufficient to have $n \gg K^{2/\gamma}/(\alpha^2 \wedge 1)$ initial observations for the plug-in estimator to be consistent (in the sense that its risk converges to zero). In contrast, the analogous MLE in the multinomial model [14] is consistent regardless of the alphabet size K, when $\gamma > 1$. When $\gamma \in (0, 1)$, the MLE is consistent if, and only if $n \gg K^{1/\gamma}$. These two plug-in type estimators have therefore two distinct behaviors. This can be explained by the fact that the former employs nearly homogeneous random variables resulting from the Laplace privacy mechanism (4), in the sense that the variance of $\hat{z}_k$ is of the order of $\tau^2 \asymp 1/((\alpha^2 \wedge 1)n)$, whereas the latter has non-homogeneous fluctuations, as the variance of $\hat{x}_k := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{x_i = k\}}$ is equal to $p_k(1 - p_k)/n$.

It should be possible to show that this lower bound actually holds up to log factors for any estimator, not just the plug-in $\hat{F}_\gamma$, when $\gamma \in (0, 1)$ and $K \lesssim_{\sqrt{\log(K)}} \sqrt{(\alpha^2 \wedge 1)n}$ (where $\lesssim_{\sqrt{\log(K)}}$ means the inequality holds up to some $\sqrt{\log(K)}$ factor). The rationale for this conjecture is the strong connection between the estimation of F_γ in the Laplace private mechanism (4) used here, and the Gaussian vector problem [6] discussed in the introduction where such lower bound is proved for all estimators and $\gamma \in (0, 1)$. For a potential adaptation of their lower bound to our set-up, we could start by restricting their noise level ϵ to $1/\sqrt{(\alpha^2 \wedge 1)n}$, and their parameters to a vector of probability in $\mathcal{P}_K$, which seems to require the condition $K \lesssim_{\sqrt{\log(K)}} \sqrt{(\alpha^2 \wedge 1)n}$ in their proof. When $K \gtrsim_{\sqrt{\log(K)}} \sqrt{(\alpha^2 \wedge 1)n}$, the lower bound produced by their proof would be, up to a $\log$ factor, equal to $\big((\alpha^2 \wedge 1)n\big)^{1-\gamma}$ regardless of the alphabet size K.

Due to its quadratic dependence in K, the squared error risk of $\hat{F}_\gamma$ becomes very large when K is big. A natural way of reducing this dependence is to not estimate all K terms p_k^γ in the functional F_γ,

but only a subset of them, so that the quadratic risk will not scale with K^2 but with the square of the cardinality of the chosen subset. In fact, the lemmas in appendix show that the small probabilities that are below the noise τ should not be considered in the plug-in estimator, since the risk of $\hat{F}_\gamma(k)$ is larger than the target p_k^γ itself, for those p_k satisfying $p_k < \tau$. This suggests not incorporating the $p_k < \tau$ in the plug-in estimator, so that we only pay the quadratic risk of $\hat{F}_\gamma$ on the large probabilities $p_k \geq \tau$, as well as the squared approximation error $\|p^{<\tau}\|_\gamma^{2\gamma}$. The former error is bounded by Corollary 3.2, replacing K by the number of significant probabilities $p_k \geq \tau$, which is smaller than $K_{\wedge\tau^{-1}} = K \wedge \tau^{-1}$. As a consequence, this threshold type estimator, denoted by $\overline{F}_\gamma$, would satisfy

$$\mathbb{E}\left[(\overline{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \|p^{<\tau}\|_\gamma^{2\gamma} + \frac{K_{\wedge\tau^{-1}}^2}{((\alpha^2 \wedge 1)n)^\gamma} + \mathbf{1}_{\{\gamma \geq 1\}} \frac{1 \vee K_{\wedge\tau^{-1}}^{3-2\gamma}}{(\alpha^2 \wedge 1)n} \tag{7a}$$

$$\lesssim_\gamma \|p^{<\tau}\|_\gamma^{2\gamma} + \frac{1}{((\alpha^2 \wedge 1)n)^{(\gamma-1)\wedge 1}} \quad . \tag{7b}$$

However, at this point $\overline{F}_\gamma$ is purely theoretical as we do not known which p_k are small and which p_k are large; so we need to design a data-driven threshold procedure that achieves the rates (7a-7b) up to log factors, which is the object of the sequel.

3.3 Thresholded plug-in estimator

Our objective in this section is to define an estimator whose squared error risk is bounded by (7b), up to $\log$ factors. Besides, this estimation procedure uses the same private sample produced by the non-interactive Laplace privacy mechanism.

The case $\gamma \in (0,1)$ is actually trivial. We bound the approximation error as follows $\|p^{<\tau}\|_\gamma^\gamma \leq F_\gamma \leq K(\sum_{k=1}^K p_k/K)^\gamma = K^{1-\gamma}$ using the concavity of the power function p^γ. Note that this bound is tight in the sense that there exist distributions (e.g. uniform or nearly uniform distributions) in $\mathcal{P}_K$ fulfilling $\|p^{<\tau}\|_\gamma^\gamma \asymp K^{1-\gamma}$. When $K \gtrsim \tau^{-1}$, the squared of the latter bound, $K^{2(1-\gamma)}$, is larger than the second error term of (7b), thus even the trivial estimator zero achieves the rate (7b). When $K \lesssim \tau^{-1}$, the performance of the plug-in estimator $\hat{F}_\gamma$ (written in Corollary 3.2) matches the second term of (7b), even better the second term of (7a). Thus, defining $\overline{F}_\gamma$ for $\gamma \in (0,1)$ as the combination of the trivial estimator (when $K \geq \tau^{-1}$) and the plug-in estimator (when $K \leq \tau^{-1}$), we get

$$\sup_{p \in \mathcal{P}_K} \mathbb{E}\left[(\overline{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \mathbb{1}_{K \geq \tau^{-1}} K^{2(1-\gamma)} + \frac{K_{\wedge\tau^{-1}}^2}{((\alpha^2 \wedge 1)n)^\gamma}$$

where the second term can be further bounded by $((\alpha^2 \wedge 1)n)^{-[(\gamma-1)\wedge 1]}$.

For $\gamma > 1$, observe that the approximation error is bounded by $\|p^{<\tau}\|_\gamma^\gamma \leq \tau^{\gamma-1}$, so that we can hope to estimate the functional F_γ without a risk that suffers from a dependence on K.

Let us now design a thresholded estimation procedure. We assume that the sample size is $2n$ for convenience and we split the data in two samples $X^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})$ and $X^{(2)} = (x_1^{(2)}, \ldots, x_n^{(2)})$. The individuals owning the data $X^{(s)}$, $s \in \{1,2\}$, use the non-interactive mechanism (4), which generates $Z_i^{(s)} = (z_{i1}^{(s)}, \ldots, z_{iK}^{(s)})$ for $i = 1, \ldots, n$. Define the sample means of the kth bin:

$$\hat{z}_k^{(s)} = \frac{1}{n} \sum_{i=1}^n z_{ik}^{(s)}, \quad k \in [K].$$

One of the means will help detect large values of the underlying probabilities and the other one is used to estimate the functional. They are independent as they were built using independent samples. Introduce the threshold

$$\hat{\tau} = 192\sigma \sqrt{\frac{\log(Kn^{1/3})}{(\alpha^2 \wedge 1)n}}.$$

We define $\hat{\tau}$ larger than τ by a logarithmic factor, only for $\hat{z}_k^{(s)}$ to concentrate around its means p_k with high probability. For each $k \in [K]$, we proceed as follows:

- if $\hat{z}_k^{(1)} < \hat{\tau}$, then do not estimate $F_\gamma(k)$.
- otherwise, when $\hat{z}_k^{(1)} \geq \hat{\tau}$, estimate $F_\gamma(k)$ using the same plug-in estimator as in (5).

That gives

$$\overline{F}_\gamma(k) =: \left(T_{[0,2]}\left[\hat{z}_k^{(2)} \cdot \mathbf{1}_{\hat{z}_k^{(1)} \geq \hat{\tau}}\right]\right)^\gamma \quad . \tag{8}$$

We bound the quadratic risk of this threshold estimator in the next theorem. Its proof is written in appendix.

Theorem 3.4. *For any $p \in \mathcal{P}$, any $\alpha \in (0, \infty)$ and any $\gamma \in (1, \infty)$, the quadratic risk of $\overline{F}_\gamma = \sum_{k=0}^K \overline{F}_\gamma(k)$ is uniformly bounded over all $p \in \mathcal{P}$ by*

$$\sup_{p \in \mathcal{P}_K} \mathbb{E}\left[(\overline{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \left(\frac{\log(Kn^{1/3})}{((\alpha^2 \wedge 1)n)}\right)^{\gamma-1} + \frac{1}{(\alpha^2 \wedge 1)n} \quad .$$

This matches the desired bound (7b), up to a $\log$ factor, as we used the threshold $\hat{\tau}$ instead of the theoretical one τ, and we bounded the approximation error by $\|p^{<\hat{\tau}}\|_\gamma^\gamma \leq \hat{\tau}^{\gamma-1}$.

We remove the log-factor in the next section, using an interactive LDP-mechanism instead of the current non-interactive mechanism (4).

4 Interactive case

In this section we study an interactive procedure that proceeds in two steps. First, using half of the sample we build via the Laplace mechanism private samples and a preliminary estimator of the probability distribution. Then we encode this partial knowledge in the private samples of the second half of the sample and build our final estimator from these data. This reasoning is based on the idea on re-writing the functional as

$$F_\gamma = \sum_{k=1}^K p_k \cdot F_{\gamma-1} \quad .$$

Therefore, the procedure is only applicable for $\gamma > 1$. Note that it has been first introduced by [2] for $\gamma = 2$ and applied to identity testing.

Interactive PM We consider the following interactive channel $Q^{(I)}$. Assuming that the sample size is $2n$ for convenience, we split the data in two groups $X^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})$ and $X^{(2)} = (x_1^{(2)}, \ldots, x_n^{(2)})$. The individuals owning the data $X^{(1)}$ use the non-interactive mechanism seen in (4), i.e.

$$Q^{(NI)}: \qquad z_{ik}^{(1)} = \mathbf{1}_{\{x_i^{(1)}=k\}} + \frac{\sigma}{\alpha} \cdot w_{ik}, \qquad k = 1, \ldots, K, \tag{9}$$

which generates $Z_i^{(1)} = (z_{i1}^{(1)}, \ldots, z_{iK}^{(1)})$ for $i = 1, \ldots, n$.

These sanitized data allow us to estimate $F_{\gamma-1}(k) = (p_k)^{\gamma-1}$ using the plug-in estimator seen in (5), that is :

$$\hat{F}_{\gamma-1}^{(1)}(k) = \left(T_{[0,2]}\left[\frac{1}{n}\sum_{i=1}^n z_{ik}^{(1)}\right]\right)^{\gamma-1} =: \left(T_{[0,2]}\left[\hat{z}_k^{(1)}\right]\right)^{\gamma-1} \quad . \tag{10}$$

Then, in order to estimate the functional of interest F_γ, we are going to estimate the following proxy functional based on (10)

$$\sum_{k=1}^K p_k \hat{F}_{\gamma-1}^{(1)}(k) \quad .$$

This is possible with the following interactive mechanism that encodes prior information in the released random variables:

$$Q^{(I)}: \qquad z_i^{(2)} = \pm z_\alpha, \text{with probability } \frac{1}{2}\left(1 \pm \frac{1}{z_\alpha}\hat{F}_{\gamma-1}^{(1)}(x_i^{(2)})\right) \tag{11}$$

where $z_\alpha = 2^{\gamma-1}\frac{e^\alpha+1}{e^\alpha-1}$. Let us denote by $Z^{(2)} = (z_1^{(2)}, \ldots, z_n^{(2)})$. Then, $(Z^{(1)}, Z^{(2)})$ is $\alpha-$LDP. This privatized sample is interactive in the sense that the second half also uses the first half of private samples in the process of randomizing locally each individual data x_i. The final estimator of F_γ is

$$\widetilde{F}_\gamma = \frac{1}{n}\sum_{i=1}^{n} z_i^{(2)} \quad . \tag{12}$$

Next Theorem gives upper bounds on the bias, the variance and the mean square error of this estimator, as a function of K, n which can be large, and of α which can be small.

Theorem 4.1. *Let $p \in \mathcal{P}_K$ and $\alpha \in (0,\infty)$, $\gamma > 1$. Then, the estimator $\widetilde{F}_\gamma$ in (12) has*

$$\begin{aligned} |\,\mathbb{E}(\widetilde{F}_\gamma) - F_\gamma| \quad &\lesssim_\gamma \quad \frac{1}{((\alpha^2\wedge 1)n)^{(\gamma-1)/2}} + \mathbf{1}_{\{\gamma\geq 3\}}\frac{\|p^{\geq\tau}\|_{\gamma-2}^{\gamma-2}}{(\alpha^2\wedge 1)n} \\ \mathrm{Var}(\widetilde{F}_\gamma) \quad &\lesssim_\gamma \quad \frac{1}{(\alpha^2\wedge 1)n} + \frac{\|p\|_2^2}{((\alpha^2\wedge 1)n)^{\gamma-1}} + \mathbf{1}_{\{\gamma\geq 2\}}\frac{\|p^{\geq\tau}\|_{2\gamma-2}^{2\gamma-2}}{(\alpha^2\wedge 1)n} \end{aligned}$$

and the quadratic risk

$$\mathbb{E}\left[(\widetilde{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \frac{1}{((\alpha^2\wedge 1)n)^{(\gamma-1)\wedge 1}} + \mathbf{1}_{\{\gamma\geq 3\}}\frac{\|p^{\geq\tau}\|_{\gamma-2}^{2(\gamma-2)}}{((\alpha^2\wedge 1)n)^2} + \mathbf{1}_{\{\gamma\geq 2\}}\frac{\|p^{\geq\tau}\|_{2\gamma-2}^{2\gamma-2}}{(\alpha^2\wedge 1)n}.$$

The proof of the Theorem is written in Appendix. We can take the supremum over all $p \in \mathcal{P}_K$ to get the following Corollary.

Corollary 4.2. *Let $p \in \mathcal{P}_K$, and $\alpha \in (0,\infty)$, $\gamma > 1$. Then, the quadratic risk is uniformly bounded over all $p \in \mathcal{P}_K$ by*

$$\sup_{p\in\mathcal{P}_K} \mathbb{E}\left[(\widetilde{F}_\gamma - F_\gamma)^2\right] \lesssim_\gamma \frac{1}{((\alpha^2\wedge 1)n)^{(\gamma-1)\wedge 1}}.$$

This matches the desired bound (7b), as the approximation error satisfies the tight bound $\|p^{<\tau}\|_\gamma^\gamma \leq \tau^{\gamma-1}$ in this case.

Corollary 4.2 generalizes the result in [2] to the case $\gamma \geq 2$, and it also shows a slower rate for γ in (1,2). These rates do not depend on K – the support size of the probability distribution p.

5 Optimality of the results

We give the lower bounds over all privacy mechanisms and over all estimators.

Theorem 5.1. *Let $\gamma > 0$, $\gamma \neq 1$ and $\alpha \in (0,1]$ satisfying $\alpha^2 n \geq 1$. Then, for any positive even integer K,*

$$\inf_Q \inf_{\hat{F}} \sup_{p\in\mathcal{P}(2)} \mathbb{E}\left[(\hat{F} - F_\gamma)^2\right] \gtrsim_\gamma \frac{1}{\alpha^2 n} + \frac{1}{(\alpha^2 n)^\gamma}$$

where $\gtrsim$ denotes inequality up to constants possibly depending on γ, and where the infimum is taken over all α-locally differentially private mechanisms Q and over all estimators $\hat{F}$.

When $\gamma \geq 2$, Theorem 5.1 involves minimax optimality of the interactive estimator presented in section 4. Notably, the minimax rate for $\gamma \geq 2$ is independent of the alphabet size K. In contrast, our lower bound for $\gamma \in (0,2)$ depends on K – see Theorem 5.2 below.

Theorem 5.2. *Let $\gamma \in (0,2)$, $\gamma \neq 1$ and $\alpha > 0$. Then, for any positive even integer K,*

$$\inf_Q \inf_{\hat{F}} \sup_{p\in\mathcal{P}(K)} \mathbb{E}\left[(\hat{F} - F_\gamma)^2\right] \gtrsim_\gamma K^{2(1-\gamma)} \wedge \frac{K^{2-\gamma}}{(n(e^{2\alpha}-e^{-2\alpha})^2)^\gamma} \tag{13}$$

where the infimum is taken over all α-locally differentially private mechanisms Q as well as over all estimators $\hat{F}$.

Notice that $(e^{2\alpha}-e^{-2\alpha})^2 \asymp \alpha^2$ when α is small. The lower bound (13) is of the order of $K^{2(1-\gamma)}$ if $K \gtrsim n(e^{2\alpha}-e^{-2\alpha})^2$, and it is of the order of $K^{2-\gamma}/(n(e^{2\alpha}-e^{-2\alpha})^2)^\gamma$ otherwise. Although the dependence on K in the lower bounds may seem in contradiction with our upper bounds that are independent of K, one can readily check that the lower bounds are smaller than the upper bounds.

The proofs of the lower bounds can be found in the Appendix.

Table 1: Upper bounds on the MSE for estimating F_γ by different estimation procedures, $\alpha \in (0,1)$

	$\gamma \in (0,1)$	$\gamma > 1$
Plug-in non-interactive	$\frac{K^2}{(\alpha^2 n)^\gamma}$	$\frac{K^2}{(\alpha^2 n)^\gamma} + \frac{(K \wedge \sqrt{\alpha^2 n})^{3-2\gamma}}{\alpha^2 n} + \frac{1}{\alpha^2 n}$
Thresholded non-interactive	$\frac{(K \wedge \sqrt{\alpha^2 n})^2}{(\alpha^2 n)^\gamma} + \mathbb{1}_{K \geq \sqrt{\alpha^2 n}} K^{2(1-\gamma)}$	$\frac{\log^{\gamma-1}(Kn)}{(\alpha^2 n)^{\gamma-1}} + \frac{1}{\alpha^2 n}$
Interactive		$\frac{1}{(\alpha^2 n)^{\gamma-1}} + \frac{1}{\alpha^2 n}$

6 Discussion

The estimation of power sum functionals in the context of local differential privacy proves to be a very rich topic potentially difficult to solve sharply in all possible cases. We discuss here the numerous directions that we plan to follow or that we could have included here, should the space be allowed.

In the case of smooth continuous distributions, interactive privacy mechanisms improved significantly over the non-interactive ones for estimating F_2 ($\gamma = 2$), see [4]. Similar interactive procedures were designed here for the case $\gamma > 1$. An open question is to know whether there is a gap between the minimax rates in non-interactive and interactive private mechanisms. Our upper bounds are faster in the interactive setting by a $\log$ factor. However, there is still a missing piece in the lower bounds to understand the nature of this gap. In particular, it is not yet obvious whether a different interactive procedure could improve on the rates in the case γ in (0,1).

It is important to highlight that all results could be transferred to the estimation of the Shannon entropy $H(p) = \sum_{k=1}^K p_k \log(1/p_k)$. Due to the non-smoothness of the function $x \log(1/x)$ the rates of the plug-in estimator would be similar to those obtained for the case $\gamma < 1$.

If we restrict the class of possible probability vectors p, our lower bounds become sharp. Indeed, the lower bounds of Theorem 5.2 are in fact obtained from the construction of simple vectors p whose set of coordinates $\{p_1, \ldots, p_K\}$ take only a constant number of values (which is smaller than 3). Conversely, for the analysis of upper bounds, we could restrict our attention to a sub-class of $\mathcal{P}_K$ where the vectors p only take a constant number of different values, which are sufficiently spaced from each other. Note that this constraint does not exclude K to grow with n, neither p_k to take any value in $[0,1]$. On this sub-class, we think that a refinement of our plug-in estimator would achieve the lower bound rate of Theorem 5.2, thus implying the minimax rates for this sub-class of parameters.

We have discussed in Section 3 that the proof of the lower bound (Theorem 8 in [6]) could be investigated to get a lower bound in the non-interactive privacy mechanism (4) that holds for any estimator, and matches our upper bounds. However, an open question is whether there exist lower bounds with respect to any estimator and *any privacy mechanism*, which matches our upper bounds. In [6], the lower bounds of Theorem 8 are obtained using a standard tool − the two fuzzy hypothesis theorem introduced by [15], see also [17]− which is widely used for the proof of lower bounds on such functionals, e.g. [13], [20] and [6]. It is our most important challenge to provide a turnkey tool in LDP settings to prove lower bounds with respect to all privacy mechanisms.

References

[1] Acharya, J., Orlitsky, A., Suresh, A.T., and Tyagi, H. (2015) The complexity of estimating Rényi entropy. *Proceedings of the Twenty-Sixth Annual ACMSIAM Symposium on Discrete Algorithms, SODA 2015*, pages 1855–1869.

[2] Berrett, T. and Butucea, C. (2020) Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *NeurIPS*, 33.

[3] Butucea, C., Dubois, A., Kroll, M. and Saumard, A. (2020) Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids, *Bernoulli*, **26**(3):1727–1764.

[4] Butucea, C. and Rohde, A. and Steinberger, L. (2020) Interactive versus non-interactive locally, differentially private estimation: Two elbows for the quadratic functional. *arXiv:2003.04773*

[5] Cai, T. T. and Low, M. G. (2011) Testing composite hypothesis, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. STatist.*, **39**: 1012–1041.

[6] Collier, O., Comminges, L. and Tsybakov, A. B. (2020) On estimation of nonsmooth functionals of sparse normal means. *Bernoulli*, **26**(3): 1989–2020.

[7] Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2018) Local privacy and minimax bounds: sharp rates for probability estimation. *NeurIPS*, **26**.

[8] Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2018) Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, **113**(521):182–201.

[9] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 265-–284.

[10] Fukuchi, K. and Sakuma, J. (2017) Minimax optimal estimators for additive scalar functionals of discrete distributions,*IEEE International Symposium on Information Theory (ISIT)*, 2017, 2103–2107.

[11] Han, Y., Jiao, J., Mukherjee, R. and Weissman, T. (2017) On estimation of $\mathbb{L}_r$ norm in Gaussian white noise models. *arxiv: 1710.03863*

[12] Han, Y., Jiao, J., Weissman, T. and Wu, Y. (2017) Optimal rates of entropy estimation over Lipschitz balls. *arxiv: 1711.02141*

[13] Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015) Minimax estimation of functionals of discrete distributions.*IEEE Transactions on Information Theory*, **61**(5): 2835–2885.

[14] Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2017) Maximum Likelihood Estimation of Functionals of Discrete Distributions. *Institute of Electrical and Electronics Engineers (IEEE)*, **63**(10):6774-–6798.

[15] Lepski, O., Nemirovski, A. and Spokoiny, V. (1999) On estimation of the $\mathbb{L}_r$ norm of a regression function. *Probab. Theory Related Fields*, **113**: 221-253

[16] Rohde, A. and Steinberger, L. (2020) Geometrizing rates of convergence under local differential privacy constraints, *Annals of Statistics*, **48**(5):2646–2670.

[17] Tsybakov, A. B. (2009) *Introduction to Nonparametric Estimation*, Springer series in statistics.

[18] Wainwright, M. J. (2019) *High-dimensional statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, **48**, Cambridge University Press, Cambridge

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

[20] Wu, Y. and Yang, P. (2016) Minimax rates of entropy estimation on large alphabets via best polynomial approximation. IEEE Transactions on Information Theory, 62(6): 3702–3720.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section .
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We don't see any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Most of them in the Supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We don't include any
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] Not concerned
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Not concerned
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Not concerned
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] Not concerned
   (b) Did you mention the license of the assets? [N/A] Not concerned
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] Not concerned
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Not concerned
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Not concerned
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable