Private Statistical Estimation of Many Quantiles

Anonymous Authors[1]

Abstract

This work studies the estimation of many statistical quantiles under differential privacy. More precisely, given a distribution and access to i.i.d. samples from it, we study the estimation of the inverse of its cumulative distribution function (the quantile function) at specific points. For instance, this task is of key importance in private data generation. We present two different approaches. The first one consists in privately estimating the empirical quantiles of the samples and using this result as an estimator of the quantiles of the distribution. In particular, we study the statistical properties of the recently published algorithm introduced by Kaplan et al. 2022 that privately estimates the quantiles recursively. The second approach is to use techniques of density estimation in order to uniformly estimate the quantile function on an interval. In particular, we show that there is a tradeoff between the two methods. When we want to estimate many quantiles, it is better to estimate the density rather than estimating the quantile function at specific points.

1. Introduction

Computing statistics from real users' data leads to new challenges, notably privacy concerns. Indeed, it is now well documented that the release of statistics computed on them can, without further caution, have disastrous repercussions (Narayanan & Shmatikov, 2006; Backstrom et al., 2007; Fredrikson et al., 2015; Dinur & Nissim, 2003; Homer et al., 2008; Loukides et al., 2010; Narayanan & Shmatikov, 2008; Sweeney, 2000; Wagner & Eckhoff, 2018; Sweeney, 2002). In order to solve this problem, differential privacy (DP) (Dwork et al., 2006b) has become the gold standard in privacy protection. It adds a layer of randomness in the estimator (i.e. the estimator does not only build on $X_1, \ldots, X_n$ but also on another source of randomness) in order to hide each user's data influence. It is notably used by the US Census Bureau (Abowd, 2018), Google (Erlingsson et al., 2014), Apple (Thakurta et al., 2017) and Microsoft (Ding et al., 2017) among others. This notion is properly defined in Section 2, but for now it is only important to view it as a constraint on the estimators that ensures that the observation of the estimator only leaks little information on the individual samples on which it is built on.

Any probability distribution $\mathbb{P}$ on $[0, 1]$ is fully characterized by its cumulative distribution function (CDF) defined by

$$F_{\mathbb{P}}(t) := \mathbb{P}\big((-\infty, t]\big), \quad \forall t \in \mathbb{R}\,.$$

The central topic of this article is the quantile function (QF), $F_{\mathbb{P}}^{-1}$, defined as the generalized inverse of $F_{\mathbb{P}}$:

$$F_{\mathbb{P}}^{-1}(p) = \inf\left\{t \in \mathbb{R} \mid p \leq F_{\mathbb{P}}(t)\right\}, \quad \forall p \in [0, 1]\ ,$$

with the convention $\inf \emptyset = +\infty$. When $\mathbb{P}$ is absolutely continuous w.r.t. Lebesgue's measure with a density that is bounded away from 0, $F_{\mathbb{P}}$ and $F_{\mathbb{P}}^{-1}$ are bijective and are inverse from one another. A well-known result is that, under mild hypotheses on $\mathbb{P}$, if $U \sim \mathcal{U}([0, 1])$ (U follows a uniform distribution on $[0, 1]$), then $F_{\mathbb{P}}^{-1}(U) \sim \mathbb{P}$ (Devroye, 1986). In other words, knowing $F_{\mathbb{P}}^{-1}$ allows to generate data with distribution $\mathbb{P}$. It makes the estimation of $F_{\mathbb{P}}^{-1}$ a key component in data generation.

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, this article studies the *private* estimation of $F_{\mathbb{P}}^{-1}(p_j)$ from these samples at prescribed values $\{p_1, \ldots, p_m\} \subset (0, 1)$. Without privacy and under mild hypotheses on the distribution, it is well-known (Van der Vaart, 2000) that for each $p \in (0, 1)$, the quantity $X_{(E(np))}$ is a good estimator of $F_{\mathbb{P}}^{-1}(p)$, where $X_{(1)}, \ldots, X_{(n)}$ are the order statistic of $X_1, \ldots, X_n$ (i.e. a permutation of the observations such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$) and $E(x)$ denotes the largest integer smaller or equal to x. The quantity $X_{(E(np))}$ is called the empirical (as opposed to statistical) quantile of the dataset $(X_1, \ldots, X_n)$ (as opposed to the distribution $\mathbb{P}$) of order p.

While the computation of private *empirical* quantiles has led to a rich literature, much less is known on the statistical properties of the resulting algorithms seen as estimators of the *statistical* quantiles of an underlying distribution, compared to more traditional ways of estimating a distribution.

1.1. Related work

Early approaches for solving the private empirical quantile computation used the Laplace mechanism (Dwork et al., 2006a;b) but the high sensitivity of the quantile query made it of poor utility (see Section 2 for a quick introduction to differential privacy, including the Laplace mechanism and the notion of sensitity). Smoothed sensitivity-based approaches followed (Nissim et al., 2007) and managed to achieve greatly improved utility.

The current state of the art for the computation of *a single empirical private quantile* (Smith, 2011) is an instantiation of the so-called exponential mechanism (McSherry & Talwar, 2007) with a specific utility function (see Section 2) that we will denote QExp (for exponential quantile) in the rest of this article. It is implemented in many DP software libraries (Allen; IBM).

For the computation of *multiple empirical private quantiles*, the problem gets more complicated. Indeed, with differential privacy, every access to the dataset has to be accounted for in the overall privacy budget. Luckily, and part of the reasons why differential privacy became so popular in the first place, composition theorems (Dwork et al., 2006b; Kairouz et al., 2015; Dong et al., 2019; 2020; Abadi et al., 2016) give general rules for characterizing the privacy budget of an algorithm depending on the privacy budgets of its subroutines. It is hence possible to estimate multiple empirical quantiles privately by separately estimating each empirical quantile privately (using the techniques presented above) and by updating the overall privacy budget with composition theorems. The algorithm IndExp (see Section 2) builds on this framework. However, recent research has shown that such approaches are suboptimal. For instance, Gillenwater et al. 2021 presented an algorithm (JointExp) based on the exponential mechanism again, with a utility function tailored for the joint computation of multiple private empirical quantiles directly. JointExp became the state of the art for about a year. It can be seen as a generalization of QExp, and the associated clever sampling algorithm is interesting in itself. Yet, more recently, Kaplan et al. 2022 demonstrated that an ingenious use of a composition theorem (as opposed to a more straightforward direct independent application) yields a simple recursive computation using QExp that achieves the best empirical performance to date. We will refer to their algorithm as RecExp (for recursive exponential). Furthermore, contrary to JointExp, RecExp is endowed with strong utility guarantees (Kaplan et al., 2022) in terms of the quality of estimation of the *empirical* quantiles.

In terms of *statistical* utility of the above-mentioned algorithms (i.e. when using the computed private empirical quantiles as statistical estimators of the statistical quantiles of the underlying distribution), under mild hypotheses, QExp is asymptotically normal (Smith, 2011; Asi & Duchi, 2020) and JointExp is consistent (Lalanne et al., 2022b).

1.2. Contributions

The main contribution of this paper is to obtain concentration properties for RecExp as a private estimator of multiple statistical quantiles (see Theorem 3.5) of a distribution. In order to do so, we adopt a proof framework that controls both the order statistic of $X_1, \dots, X_n$ relatively to the statistical quantiles (see Lemma 3.1), and the minimum gap in the order statistic, which is defined as $\min_i X_{(i+1)} - X_{(i)}$, and with the convention $X_{(0)} = 0$ and $X_{(n+1)} = 1$ (see Lemma 3.2). Indeed, this last quantity is of key interest in order to leverage the empirical utility provided by Kaplan et al. 2022. This framework also gives us concentration results for QExp when used to estimate multiple statistical quantiles (see Corollary 3.4). In particular, our results show that when m (the number of statistical quantiles to estimate) is large, RecExp has a much better statistical utility (both in term of proved statistical upper bounds and of experimental behavior) for a given privacy budget than the simple composition of QExp.

We then compare the statistical utility of RecExp to the one of a quantile function built on a simple histogram estimator of the density of $\mathbb{P}$. Since this estimator is a functional estimator that estimates all the quantiles in an interval, its statistical utility (see Theorem 4.4) obviously has no dependence on m, whereas the utility of RecExp has one. We show that for high values of m the histogram estimator has a better utility than RecExp for a given privacy budget. This theoretical result is confirmed numerically (see Section 5). For reasonable values of m however, our work consolidates the fact that RecExp is a powerful private estimator, both to estimate *empirical* quantiles of a dataset (Kaplan et al., 2022) and to estimate the *statistical* quantiles of a distribution (this work). Furthermore, a simple comparison of the upper bounds (Theorem 3.5 and Theorem 4.4) can serve as a guideline to decide whether to choose RecExp or an histogram estimator.

2. Background

This section presents technical details about differential privacy and private empirical quantiles computation.

2.1. Differential Privacy

A randomized algorithm A that takes as input a dataset $(X_1, \dots, X_n)$ (where each X_i lives in some data space, and the size n can be variable) is ϵ-differentially private (ϵ-DP) (Dwork et al., 2006a;b; 2014), where $\epsilon > 0$ is a privacy budget, if for any measurable S in the output space of A and any neighboring datasets $(X_1, \dots, X_n) \sim (X'_1, \dots, X'_{n'})$ (given some neighboring relation $\sim$) we have

$$\mathbb{P}\big(A(X_1, \dots, X_n) \in S\big) \leq e^{\epsilon} \times \mathbb{P}\big(A(X'_1, \dots, X'_{n'}) \in S\big)$$

where the randomness is taken w.r.t. A.

Differential privacy ensures that it is hard to distinguish between two neighboring datasets when observing the output of A. The neighboring relation has an impact on the concrete consequences of such a privacy guarantee. A usual goal is to make it hard to tell if a specific user contributed to the dataset. This is typically associated with an ”addition/removal” neighboring relation: $(X_1, \dots, X_n) \sim (X'_1, \dots, X'_{n'})$ if $(X'_1, \dots, X'_{n'})$ can be obtained from $(X_1, \dots, X_n)$ by adding/removing a single element, up to a permutation. Another choice is the ”replacement” neighboring relation: $(X_1, \dots, X_n) \sim (X'_1, \dots, X'_{n'})$ if $(X'_1, \dots, X'_{n'})$ can be obtained from $(X_1, \dots, X_n)$ up to a permutation by replacing a single entry.

There are multiple standard ways to design an algorithm that is differentially private. We focus on the ones that will be useful for this article.

Given a deterministic function f mapping a dataset to a quantity in $\mathbb{R}^d$, the sensitivity of f is

$$\Delta f := \sup_{(X_1,\dots,X_n)\sim(X'_1,\dots,X'_{n'})} \big\| f(X_1, \dots, X_n) - f(X'_1, \dots, X'_{n'}) \big\|_1 .$$

Given a dataset $(X_1, \dots, X_n)$, the *Laplace mechanism* returns $f(X_1, \dots, X_n) + \frac{\Delta f}{\epsilon}\mathrm{Lap}(I_d)$ where $\mathrm{Lap}(I_d)$ refers to a random vector of dimension d with independent components that follow a centered Laplace distribution of parameter 1. This mechanism is ϵ-DP (Dwork et al., 2014).

If the private mechanism has to output in a general space O equipped with a reference σ-finite measure μ, one can exploit the *exponential mechanism* (McSherry & Talwar, 2007) to design it. Given a utility function u that takes as input a dataset $(X_1, \dots, X_n)$ and a candidate output $o \in O$ and returns $u\big((X_1, \dots, X_n), o\big) \in \mathbb{R}$, which is supposed to measure how well o fits the result of a certain operation that we want to do on $(X_1, \dots, X_n)$ (with the convention that the higher the better), the sensitivity of u is

$$\Delta u := \sup_{o\in O,(X_1,\dots,X_n)\sim(X'_1,\dots,X'_{n'})} \Big| u\big((X_1, \dots, X_n), o\big) - \big((X'_1, \dots, X'_{n'}), o\big) \Big| .$$

Given a dataset $(X_1, \dots, X_n)$, the exponential mechanism returns a sample o on O of which the distribution of probability has a density w.r.t. μ that is proportional to $e^{\frac{\epsilon}{2\Delta u} u\left((X_1,\dots,X_n),o\right)}$. It is ϵ-DP (McSherry & Talwar, 2007).

Finally, a simple composition property (Dwork et al., 2006b) states that if $A_1, \dots, A_k$ are ϵ-DP, $(A_1, \dots, A_k)$ is $k\epsilon$-DP.

2.2. Private empirical quantile estimation

This subsection details the algorithms evoked in Section 1.1 that will be of interest for this article.

QExp. Given n points $X_1, \dots, X_n \in [0, 1]$ and $p \in (0, 1)$, the QExp mechanism, introduced by Smith et al. 2011, is an instantiation of the exponential mechanism w.r.t. μ the Lebesgue's measure on $[0, 1]$, with utility function u_{QExp} such that, for any $q \in [0, 1]$,

$$u_{\text{QExp}}\big((X_1, \dots, X_n), q\big) := -\big| |\{i | X_i < q\}| - E(np) \big| ,$$

where for a set, $|\cdot|$ represents its cardinality. The sensitivity of u_{QExp} is 1 for both of the above-mentioned neighboring relations. As the density of QExp is constant on all the intervals of the form $(X_{(i)}, X_{(i+1)})$, a sampling algorithm for QExp is to first sample an interval (which can be done by sampling a point in a finite space) and then to uniformly sample a point in this interval. This algorithm has complexity $O(n)$ if the points are sorted and $O(n \log n)$ otherwise. Its utility (as measured by a so-called ”empirical error”) is controlled, cf Kaplan et al. 2022 Lemma A.1. This is summarized as follows

Fact 2.1 (Empirical Error of QExp)**.** *Consider fixed real numbers* $X_1, \dots, X_n \in [0, 1]$ *that satisfy* $\min_i X_{(i+1)} - X_{(i)} \geq \Delta > 0$ *with the convention* $X_{(0)} = 0$ *and* $X_{(n+1)} = 1$. *Denote q the (random) output of QExp on this dataset, for the estimation of a single empirical quantile of order p, and*

$$\mathfrak{E} := \Big| \big|\{i | X_i < q\}\big| - E(np) \Big| ,$$

the empirical error *of QExp. For any* $\beta \in (0, 1)$, *we have*

$$\mathbb{P}\left(\mathfrak{E} \geq 2\frac{\ln\left(\frac{1}{\Delta}\right) + \ln\left(\frac{1}{\beta}\right)}{\epsilon}\right) \leq \beta .$$

Let us mention that in this article, we use the term *Fact* to refer to results that are directly borrowed from the existing literature in order to clearly identify them. In particular, it is not correlated with the technicality of the result.

IndExp. Given $p_1, \dots, p_m \in (0, 1)$, IndExp privately estimates the empirical quantiles of order $p_1, \dots, p_m$ by evaluating each quantile independently using QExp and the simple composition property. Each quantile is estimated with a privacy budget of $\frac{\epsilon}{m}$. The complexity is $O(mn)$ if the points are sorted, $O(mn + n \log n)$ otherwise.

RecExp. Introduced by Kaplan et al. 2022, RecExp is based on the following idea : Suppose that we already have a private estimate, q_i, of the empirical quantile of order p_i for a given i. Estimating the empirical quantiles of orders $p_j > p_i$ should be possible by only looking at the data points that are bigger than q_i, and similarly for the empirical quantiles of orders $p_j < p_i$. Representing this process as a tree, the addition or removal of an element in the dataset only affects at most one child of each node and at most one node per level of depth in the tree. The "per-level" composition of mechanisms comes for free in terms of privacy budget, hence only the tree depth matters for composition. By choosing a certain order on the quantiles to estimate, this depth can be bounded by $\log_2 m + 1$. More details can be found in the original article (Kaplan et al., 2022).

When using QExp with privacy budget $\frac{\epsilon}{\log_2 m+1}$ for estimating the individual empirical quantiles, RecExp is ϵ-DP with the addition/removal neighborhing relation. This remains valid with the replacement relation if we replace ϵ by $\epsilon/2$, as the replacement relation can be seen as a two-steps addition/removal relation. RecExp has a complexity of $O(n \log m)$ if the points are sorted and $O(n\log(nm))$ otherwise. The following control of its empirical error is adapted from Kaplan et al. 2022 Theorem 3.3.

Fact 2.2 (Empirical Error of RecExp)**.** *Consider fixed real numbers* $X_1, \dots, X_n \in [0,1]$ *that satisfy* $\min_i X_{(i+1)} - X_{(i)} \geq \Delta > 0$ *with the convention* $X_{(0)} = 0$ *and* $X_{(n+1)} = 1$*. Denote* $(q_1, \dots, q_m)$ *the (random) return of RecExp on this dataset, for the estimation of* m *empirical quantiles of orders* $(p_1, \dots, p_m)$*, and*

$$\mathfrak{E} := \max_j \left| \left| \{i | X_i < q_j\} \right| - E(np_j) \right| ,$$

the empirical error *of RecExp. For any* $\beta \in (0,1)$*, we have*

$$\mathbb{P}\left(\mathfrak{E} \geq 2(\log_2 m + 1)^2 \frac{\ln\left(\frac{1}{\Delta}\right) + \ln(m) + \ln\left(\frac{1}{\beta}\right)}{\epsilon} \right) \leq \beta \, .$$

3. Statistical utility of ⋆Exp

Fact 2.1 and Fact 2.2 control how well QExp, IndExp and RecExp privately estimates *empirical* quantiles of a given dataset. However, they do not tell how well those algorithms behave when the dataset is drawn from some probability distribution and the algorithm output is used to estimate the *statistical* quantiles of this distribution. This is precisely the objective of this section, where we notably highlight the fact that the utility of RecExp scales much better with m (the number of quantiles to estimate) than previous algorithms for this task.

3.1. How to leverage Fact 2.1 and Fact 2.2

Two difficulties arise when trying to control the *statistical* utility of QExp and IndExp based on Fact 2.1 and Fact 2.2.

First, the measure of performance (i.e. show mall the empirical error is) controls the deviation w.r.t. the empirical quantiles in terms of *order* :

$$\max_j \left| \left| \{i | X_i < q_j\} \right| - E(np_j) \right| .$$

In fact, $E(np_j) \approx np_j$ has no link with $F_{\mathbb{P}}$ a priori. In contrast, from a statistical point of view, the quantity of interest in the deviation w.r.t. the statistical quantiles $(F_{\mathbb{P}}^{-1}(p_1), \dots, F_{\mathbb{P}}^{-1}(p_m))$. We circumvent that difficulty with the following general purpose lemma :

Lemma 3.1 (Concentration of empirical quantiles)**.** *If* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* $\pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely, then for any* $p \in (0,1)$ *and* $\gamma > 0$ *such that* $\gamma < \min\left(F_X^{-1}(p), 1 - F_X^{-1}(p)\right)$*, we have*

$$\mathbb{P}\Big(\sup_{k \in J} |X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma \Big) \leq 2e^{-\frac{\gamma^2 \underline{\pi}^2}{8p} n} + 2e^{-\frac{\gamma^2 \underline{\pi}^2}{8(1-p)} n} \, ,$$

where

$$J := \left\{ \max\left(-E(np) + 1, -E\left(\frac{1}{2} n \gamma \underline{\pi}\right) + 1 \right), \dots, \min\left(n - E(np), E\left(\frac{1}{2} n \gamma \underline{\pi}\right) - 1 \right) \right\} .$$

The proof is postponed to Appendix A. The integer set J may be viewed as an error buffer : As long as an algorithm returns a point with an *order* error falling into J (compared to $E(np)$), the error on the *statistical* estimation will be small.

The second difficulty is the need to control the lower bound on the gaps Δ. For many distributions, this quantity can be as small as we want, and the guarantees on the empirical error of QExp, IndExp and RecExp can be made as poor as we want (Lalanne et al., 2022b). However, by imposing a simple condition on the density, the following lemma tells that the minimum gap in the order statistic is "not too small".

Lemma 3.2 (Concentration of the gaps)**.** *Consider* $n \geq 1$ *and* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely. Denote* $\Delta_i = X_{(i)} - X_{(i-1)}$, $1 \leq i \leq n+1$, *with the convention* $X_{(0)} = 0$ *and* $X_{(n+1)} = 1$*. For any* $\gamma > 0$ *such that* $\gamma < \frac{1}{4\bar{\pi}}$*, we have*

$$\mathbb{P}\left(\min_{i=1}^{n+1} \Delta_i > \frac{\gamma}{n^2} \right) \geq e^{-4\bar{\pi}\gamma} \, .$$

The proof is postponed to Appendix B.

3.2. Statistical utility of QExp and IndExp

As a first step towards the analysis of RecExp, and in order to offer a point of comparison, we first build on the previous results to analyze statistical properties of QExp and IndExp.

Theorem 3.3 (Statistical utility of QExp)**.** *Consider* $n \geq 1$ *and* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely. Denote* q *the (random) result of QExp on* $(X_1, \dots, X_n)$ *for the estimation of the quantile of order* p*, where* $\min(p, 1-p) > 2/n$*. For any* $\gamma \in (0, \frac{2\min(p,1-p)}{\underline{\pi}})$

$$\mathbb{P}\big(|q - F_\pi^{-1}(p)| > \gamma\big) \leq 4n\sqrt{2e\bar{\pi}}e^{-\frac{\epsilon n \gamma \underline{\pi}}{32}} + 4e^{-\frac{\gamma^2 \underline{\pi}^2}{8}n} .$$

Sketch of proof. We fix a buffer size K and define QC (for quantile concentration) the event "Any error of at most K points in the order statistic compared to $X_{(E(np))}$ induces an error of at most γ on the statistical estimation of $F_\pi^{-1}(p)$". The probability $\mathbb{P}(QC^c)$ is controlled by Lemma 3.1.
We fix a gap size $\Delta > 0$ and define the event G (for gaps) $\min_i \Delta_i \geq \Delta$, so that $\mathbb{P}(G^c)$ is controlled by Lemma 3.2.
Then, we notice that

$$\begin{aligned}
&\mathbb{P}\big(|q - F_\pi^{-1}(p)| > \gamma\big) \\
&\quad \leq \mathbb{P}\big(|q - F_\pi^{-1}(p)| > \gamma \big| QC, G\big) + \mathbb{P}(QC^c) + \mathbb{P}(G^c) \\
&\quad \leq \mathbb{P}\big(\mathfrak{E} \geq K+1 \big| QC, G\big) + \mathbb{P}(QC^c) + \mathbb{P}(G^c) \ ,
\end{aligned}$$

where $\mathfrak{E}$ refers to the empirical error of QExp. Using Fact 2.1 for a suited β controls $\mathbb{P}\Big(\mathfrak{E} \geq K+1 \Big| QC, G\Big)$. Tuning the values of K, Δ and β concludes the proof. □

The full proof can be found in Appendix C.

Applying this result to IndExp (ϵ becomes $\frac{\epsilon}{m}$) together with a union bound gives the following result :

Corollary 3.4 (Statistical utility of IndExp)**.** *Consider* $n \geq 1$ *and* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely. Denote* $\mathbf{q} := (q_1, \dots, q_m)$ *the (random) result of IndExp on* $(X_1, \dots, X_n)$ *for the estimation of the quantiles of orders* $\mathbf{p} := (p_1, \dots, p_m)$*, where* $\min_i[\min(p_i, 1-p_i)] > 2/n$*. For each* $\gamma \in \left(0, \frac{2\min_i[\min(p_i,1-p_i)]}{\underline{\pi}}\right)$ *we have*

$$\begin{aligned}
\mathbb{P}\big(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\big) &\leq 4nm\sqrt{2e\bar{\pi}}e^{-\frac{\epsilon n \gamma \underline{\pi}}{32m}} \\
&\quad + 4me^{-\frac{\gamma^2 \underline{\pi}^2}{8}n} ,
\end{aligned}$$

where $F_\pi^{-1}(\mathbf{p}) = (F_\pi^{-1}(p_1), \dots, F_\pi^{-1}(p_m))$*.*

The proof is postponed to Appendix D.

So, there exist a polynomial expression P and two positive constants C_1 and C_2 depending only on the distribution such that, under mild hypotheses,

$$\begin{aligned}
&\mathbb{P}\big(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\big) \\
&\qquad \leq P(n,m) \max\left(e^{-C_1 \frac{\epsilon n \gamma}{m}}, e^{-C_2 \gamma^2 n}\right) .
\end{aligned}$$

We factorized the polynomial expression since it plays a minor role compared to the values in the exponential.

Statistical complexity. The term $P(n,m)e^{-C_2\gamma^2 n}$ simply comes from the concentration of the empirical quantiles around the statistical ones. It is independent of the private nature of the estimation. It is the price that one usually expects to pay without the privacy constraint.

Privacy overhead. The term $P(n,m)e^{-C_1 \frac{\epsilon n \gamma}{m}}$ can be called the privacy overhead. It is the price paid for using this specific private algorithm for the estimation. For IndExp, if we want it to be constant, ϵn has to roughly scale as m times a polynomial expression in $\log_2 m$. As we will see later in Theorem 3.5, RecExp behaves much better, with $n\epsilon$ having to scale only as a polynomial expression in $\log_2 m$.

A privacy overhead of this type is not only an artifact due to a given algorithm (even if a suboptimal algorithm can make it worse), but in fact a constituent part of the private estimation problem, associated to a necessary price to pay, as captured by several works on generic lower bounds valid for *all* private estimators (Duchi et al., 2013a;b; Acharya et al., 2021e; 2018; 2021c;b;d;a; Barnes et al., 2020a;b; 2019; Kamath et al., 2022; Butucea et al., 2020; Lalanne et al., 2022a; Berrett & Butucea, 2019; Kroll, 2021).

3.3. Statistical properties of RecExp

With a similar proof technique as in the one of Theorem 3.3, the following result gives the statistical utility of RecExp :

Theorem 3.5 (Statistical utility of RecExp)**.** *Consider* $n \geq 1$ *and* $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ *where* π *is a density on* $[0,1]$ *w.r.t. Lebesgue's measure such that* $\bar{\pi} \in \mathbb{R} \geq \pi \geq \underline{\pi} \in \mathbb{R} > 0$ *almost surely. Denote* $\mathbf{q} := (q_1, \dots, q_m)$ *the result of RecExp on* $(X_1, \dots, X_n)$ *for the quantiles of orders* $\mathbf{p} := (p_1, \dots, p_m)$*, where* $\min_i[\min(p_i, 1-p_i)] > 2/n$*. For any* $\gamma \in (0, \frac{2\min_i[\min(p_i,1-p_i)]}{\underline{\pi}})$ *we have*

$$\begin{aligned}
\mathbb{P}\big(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\big) &\leq 4n\sqrt{2e\bar{\pi}}me^{-\frac{\epsilon n \gamma \underline{\pi}}{32\log_2(2m)^2}} \\
&\quad + 4me^{-\frac{\gamma^2 \underline{\pi}^2}{8}n} .
\end{aligned}$$

The proof is postponed to Appendix E.

As with Corollary 3.4, we can simplify this expression as

$$\mathbb{P}\left(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq P(n,m)\max\left(e^{-C_1\frac{\epsilon n\gamma}{\log_2 m}}, e^{-C_2\gamma^2 n}\right),$$

where P is a polynomial expression and C_1 and C_2 are constants, all depending only on the distribution.

Statistical complexity. On the one hand the statistical term of this expression, which is independent of ϵ, is the same as with IndExp. This is natural since the considered statistical estimation problem is unchanged, only the privacy mechanism employed to solve it under a DP constraint was changed.

Privacy overhead. On the other hand the privacy overhead $P(n,m)e^{-C_1\frac{\epsilon n\gamma}{\log_2 m}}$ is much smaller than the one of IndExp. The scaling of ϵn to reach a prescribed probability went from approximately linear in m to roughly a polynomial expression in $\log_2 m$.

In particular and to the best of our knowledge, this scaling in m places RecExp much ahead of its competitors (the algorithms that compute multiple private empirical quantiles) for the task of statistical estimation.

4. Uniform estimation of the quantile function

Private quantile estimators often focus on estimating the quantile function at specific points $p_1, \dots, p_m$, which is probably motivated by a combination of practical considerations (algorithms to estimate and representing finitely many numbers are easier to design and manipulate than algorithms to estimate a function) and of intuitions about privacy (estimating the whole quantile function could increase privacy risks compared to estimating it on specific points). It is however well-documented in the (non-private) statistical literature that, under regularity assumptions on the quantile function, it can also be approximated accurately from functional estimators, see e.g. (Györfi et al., 2002; Tsybakov, 2004).

Building on this, this section considers a simple private histogram estimator of the density (Wasserman & Zhou, 2010) in order to estimate the quantile function in functional infinite norm. This allows of course to estimate the quantile function at $(p_1, \dots, p_m)$ for arbitrary m. As a natural consequence, we show that when m is very high, for a given privacy level RecExp has suboptimal utility guarantees and is beaten by the guarantees of the histogram estimator. Theorem 4.4 and Theorem 3.5 give a decision criterion (by comparing the upper bounds) to decide whether to use RecExp or a histogram estimator for the estimation problem.

4.1. Motivation: lower bounds for IndExp and RecExp

Lower-bounding the density of the exponential mechanism for u_{QExp} gives a general lower-bound on its utility:

Lemma 4.1 (Utility of QExp; Lower Bound)**.** *Let $X_1, \dots, X_n \in [0,1]$. Denoting by q the result of QExp on $(X_1, \dots, X_n)$ for the quantile of order p, we have for any $t \in [0,1]$ and any positive $\gamma \in (0, \frac{1}{4}]$,*

$$\mathbb{P}\left(|q-t| > \gamma\right) \geq \frac{1}{2}e^{-\frac{n\epsilon}{2}} .$$

Note that this holds without any constraint relating p,n, or γ. The proof is postponed to Appendix F. As a consequence, if the points $X_1, \dots, X_n$ are randomized, the probability that QExp makes an error bigger than γ on the estimation of a quantile of the distribution is at least $\frac{1}{2}e^{-\frac{n\epsilon}{2}}$. A direct consequence is that for any $\gamma \in (0, \frac{1}{4}]$, the statistical utility of IndExp has a is lower-bounded:

$$\mathbb{P}\left(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \geq \frac{1}{2}e^{-\frac{n\epsilon}{2m}} ,$$

and the statistical utility of RecExp is also lower-bounded:

$$\mathbb{P}\left(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \geq \frac{1}{2}e^{-\frac{n\epsilon}{2(\log_2 m+1)}} .$$

These are consequences of lower-bounds on the estimation error of the first statistical quantile estimated by each algorithm in its respective computation graph (with privacy level ϵ/m for IndExp; $\epsilon/(\log_2 m+1)$ for RecExp).

In particular, for both algorithms, utility becomes arbitrarily bad when m increases. This is not a behavior that would be expected from any optimal algorithm. The rest of this section studies a better estimator for high values of m.

4.2. Histogram density estimator

The histogram density estimator is a well-known estimator of the density of a distribution of probability. Despite its simplicity, a correct choice of the bin size can even make it minimax optimal for the class of Lipschitz densities.

Under differential privacy, this estimator was first adapted and studied by Wasserman and Zhou 2010. It is studied both in terms of integrated squared errorand in Kolmogorov-Smirnov distance. In the sequel, we need a control in infinite norm. We hence determine the histogram concentration properties for this metric.

Given a a bin size $h > 0$ that satisfies $\frac{1}{h} \in \mathbb{N}$, we partition $[0,1]$ in $\frac{1}{h}$ intervals of length h. The intervals of this partition are called the bins of the histogram. Given $\frac{1}{h}$ i.i.d. centered Laplace distributions of parametter 1, $(\mathcal{L}_b)_{b\in\text{bins}}$, we define $\hat{\pi}^{\text{hist}}$, an estimator of the supposed density π of the distribution as: for each $t \in [0,1]$,

$$\hat{\pi}^{\text{hist}}(t) := \sum_{b \in \text{bins}} \mathbb{1}_b(t) \frac{1}{nh} \left(\sum_{i=1}^{n} \mathbb{1}_b(X_i) + \frac{2}{\epsilon} \mathcal{L}_b \right) .$$

The function that, given the bins of a histogram, counts the number of points that fall in each bin of the histogram has a sensitivity of 2 for the replacement neighboring relation. Indeed, replacing a point by another changes the counts of at most two (consecutive) bins by one. Hence, the construction of the Laplace mechanism ensures that $\hat{\pi}^{\text{hist}}$ is ϵ-DP.

Note that, as a common practice, we divided by n freely in terms of privacy budget in the construction of $\hat{\pi}^{\text{hist}}$. This is possible because we work with the replacement neighboring relation. The size n of the datasets is fixed and is a constant of the problem.

The deviation between π and $\hat{\pi}^{\text{hist}}$ can be controlled.

Lemma 4.2 (Utility of $\hat{\pi}^{\text{hist}}$; Density estimation)**.** *Consider $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that π is L-Lipschitz for some positive constant L, and the private histogram density estimator $\hat{\pi}^{hist}$ with bin size h. For any $\gamma > Lh$, we have*

$$\mathbb{P}\left(\|\hat{\pi}^{hist} - \pi\|_\infty > \gamma \right) \leq \frac{1}{h} e^{-\frac{\gamma h n \epsilon}{4}} + \frac{2}{h} e^{-\frac{h^2(\gamma - Lh)^2}{4} n} .$$

The proof is postponed to Appendix G.

4.3. Application to quantile function estimation

In order to use $\hat{\pi}^{\text{hist}}$ as an estimator of the quantile function, we need to properly define a quantile function estimator associated with it. Indeed, even if $\hat{\pi}^{\text{hist}}$ estimates a density of probability, it does not necessary integrate to 1 and can even be negative at some locations. Given any integrable function $\hat{\pi}$ on $[0,1]$, we define its generalized quantile function

$$F_{\hat{\pi}}^{-1}(p) = \inf \left\{ q \in [0,1] | \int_0^q \hat{\pi} \geq p \right\}, \forall p \in [0,1] \ ,$$

with the convention $\inf \emptyset = 1$. Even if this quantity has no reason to behave as a quantile function, the following lemma tells that $F_{\hat{\pi}}^{-1}$ is close to an existing quantile function when $\hat{\pi}$ is close to its corresponding density.

Lemma 4.3 (Inversion of density estimators)**.** *Consider a density π on $[0,1]$ w.r.t. Lebesgue's measure such that $\pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely. If $\hat{\pi}$ is an integrable function that satisfies $\|\hat{\pi} - \pi\|_\infty \leq \alpha$, and if $p \in [0,1]$ is such that $\left[F_\pi^{-1}(p) - \frac{2\alpha}{\underline{\pi}}, F_\pi^{-1}(p) + \frac{\alpha}{\underline{\pi}} \right] \subset (0,1)$, then*

$$\left| F_\pi^{-1}(p) - F_{\hat{\pi}}^{-1}(p) \right| \leq \frac{2\alpha}{\underline{\pi}} .$$

The proof is in Appendix H.

A direct consequence of Lemma 4.2 and Lemma 4.3 is Theorem 4.4. It controls the deviation of the generalized quantile function based on $\hat{\pi}^{\text{hist}}$ to the true quantile function.

Theorem 4.4 (Utility of $F_{\hat{\pi}^{\text{hist}}}^{-1}$; Quantile function estimation)**.** *Consider $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that π is L-Lipschitz for some positive constant L and that $\pi \geq \underline{\pi} \in \mathbb{R} > 0$ almost surely, and $h < \underline{\pi}/(4L)$ such that $\frac{1}{h} \in \mathbb{N}$. Let $F_{\hat{\pi}^{hist}}^{-1}$ be the quantile function estimator associated with the private histogram density estimator $\hat{\pi}^{hist}$ with bin size h. Consider $\gamma_0 \in (2Lh/\underline{\pi}, 1/2)$, $I := F_\pi\big((\gamma_0, 1-\gamma_0)\big)$, and $\|\cdot\|_{\infty,I}$ the sup-norm of functions on the interval I. We have*

$$\mathbb{P}\bigg(\|F_{\hat{\pi}^{hist}}^{-1} - F_\pi^{-1}\|_{\infty,I} > \gamma \bigg) \leq \frac{1}{h} e^{-\frac{\gamma \underline{\pi} h n \epsilon}{8}} + \frac{2}{h} e^{-\frac{h^2}{4}\left(\frac{\gamma \underline{\pi}}{2} - Lh\right)^2 n} ; ,$$

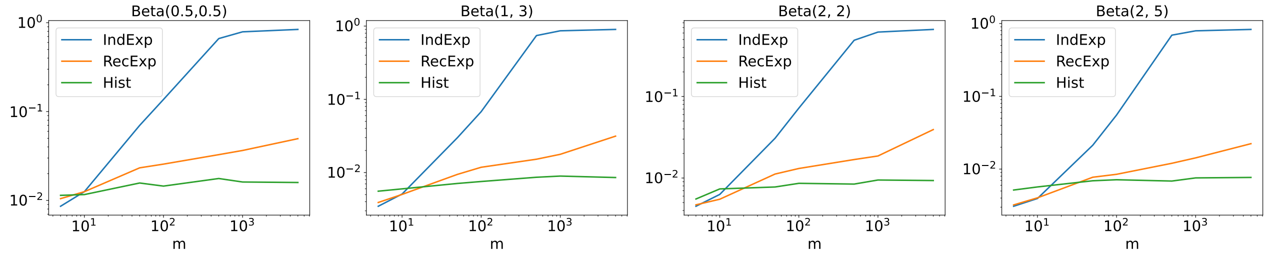
whenever $\gamma \in (2Lh/\underline{\pi}, \gamma_0)$.

The proof is postponed to Appendix I.

Analysis of Theorem 4.4. As with Theorem 3.3 and Theorem 3.5, the upper-bound provided by Theorem 4.4 can be split in two terms : The error that one usually expects without privacy constraint, $\frac{2}{h} \exp(-\frac{h^2}{4}(\frac{\gamma \underline{\pi}}{2} - Lh)^2 n)$, and the one that come from the private algorithm, $\frac{1}{h} \exp(-\frac{\gamma \underline{\pi} h n \epsilon}{8})$. The assumption $\frac{\gamma \underline{\pi}}{2} > Lh$ ensures that the bin size h and the desired level of precision γ are compatible.

Computational aspects. $\hat{\pi}^{\text{hist}}$ is constant on each bin. Hence, it can be stored in a single array of size $\frac{1}{h}$. If the data points are sorted, this array can be filled with a single pass over all data points and over the array. Then, given $p_1, \dots, p_m \in (0,1)$ sorted, estimating $F_{\hat{\pi}^{\text{hist}}}^{-1}(p_1), \dots, F_{\hat{\pi}^{\text{hist}}}^{-1}(p_m)$ can be done with a single pass over $p_1, \dots, p_m$ and over the array that stores $\hat{\pi}^{\text{hist}}$. Indeed, it is done by "integration" of the array until the thresholds of the p_i's are reached. The overall complexity of this procedure is $O\left(n + m + \frac{1}{h}\right)$ to which must be added $O(n \log n)$ if the data is not sorted and $O(m \log m)$ if the targeted quantiles p_i are not sorted.

Comparison with RecExp. Comparing this histogram-based algorithm to RecExp is more difficult than comparing RecExp to IndExp. First of all, the results are qualitatively different. Indeed, RecExp estimates the quantile function on a finite number of points and the histogram estimator estimates it on an interval. The second result is stronger in the sense that when the estimation is done on an interval,



The vertical axis reads the error $\mathbb{E}\left(\|\hat{\mathbf{q}} - F^{-1}(\mathbf{p})\|_\infty\right)$ where $\mathbf{p} = \left(\frac{1}{4} + \frac{1}{2(m+1)}, \dots, \frac{1}{4} + \frac{m}{2(m+1)}\right)$ for different values of m, $n = 10000$, $\epsilon = 0.1$, $\hat{\mathbf{q}}$ is the private estimator, and $\mathbb{E}$ is estimated by Monte-Carlo averaging over 50 runs. The histogram is computed on 200 bins.

Figure 1. Numerical performance of the different private estimators

it is done for any finite number of points in that interval. However, the error of RecExp for that finite number of points may be smaller than the one given by the histogram on the interval. Then, the histogram depends on a meta parameter h. With a priori information on the distribution, it can be tuned using Theorem 4.4. Aditionally, the hypothesis required are different : Theorem 3.5 does not require the density to be Lipschits contrary to Theorem 4.4. Finaly, we can observe that the histogram estimator is not affected by the lower bounds described in Section 4.1. Hence, when all the hypotheses are met, there will obviously always be a number m of targeted quantiles above which it is better to use histograms. The two algorithms are numerically compared in Section 5.

5. Numerical results

For the experiments, we benchmarked the different estimators on beta distributions, as they allow to easily tune the Lipschitz constants of the densities, which is important for characterizing the utility of the histogram estimator.

Figure 1 represents the performance of the estimator as a function of m. We estimate the quantiles of orders $\mathbf{p} = \left(\frac{1}{4} + \frac{1}{2(m+1)}, \dots, \frac{1}{4} + \frac{m}{2(m+1)}\right)$ since it allows us to stay in the regions where the density is not too small.

IndExp *vs* RecExp *vs* Histograms. Figure 1, confirms our claims about the scaling in m of IndExp and RecExp. Indeed, even if IndExp quickly becomes unusable, RecExp stays at a low error until really high values of m. The conclusions of Section 4.1 also seem to be verified : Even if RecExp performs well for small to intermediate values of m, there is always a certain value of m for which it becomes worse than the histogram estimator. This shift of regime occurs between $m \approx 10$ for the distribution Beta(0.5, 0.5) and $m \approx 40$ for the distribution Beta(2, 5).

Error of the histogram-based approach. The shape of the error for the histogram estimator is almost flat. Again, it is compatible with Theorem 4.4 : The control in infinite norm is well suited for the histograms.

Role of the Lipschitz constant. By crossing the shape of the beta distributions (see Appendix J) and Figure 1, a pattern becomes clear : The distributions on which the histogram estimator performs best (i.e. the distributions on which it becomes the best estimator for the lowest possible value of m) are the distributions with the smallest Lipschitz constant. This was expected since the guarantees of utility of Theorem 4.4 get poorer the higher this quantity is.

6. Conclusion

Privately estimating the (statistical) quantile function of a distribution has some interesting properties. For low to mid values of m, this article demonstrated that there is a real incentive in estimating it on a finite sample of m points. This was done by using algorithms recently introduced in order to estimate the *empirical* quantiles of a dataset. However, when the number m becomes too high, the previously-mentioned algorithms become suboptimal. It is then more effective to estimate the density with a histogram. Furthermore, the utility results are qualitatively stronger : The estimation is uniform over an interval, as opposed to pointwise on a finite set. Theorem 3.5 and Theorem 4.4 can be used to decide what method to choose.

An interesting question would be to know if it is possible to modify RecExp in such regimes in order to bridge the gap with histograms. Possibly by adapting the privacy budget to the depth in the computation tree.

Another interesting question would be to investigate the possible (minimax) optimality of the techniques of this article on restricted classes of distributions or regimes of m.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Abowd, J. M. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, 2018.

Acharya, J., Sun, Z., and Zhang, H. Differentially private testing of identity and closeness of discrete distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

Acharya, J., Canonne, C. L., Freitag, C., Sun, Z., and Tyagi, H. Inference under information constraints III: local privacy constraints. *IEEE J. Sel. Areas Inf. Theory*, 2(1): 253–267, 2021a.

Acharya, J., Canonne, C. L., Mayekar, P., and Tyagi, H. Information-constrained optimization: can adaptive processing of gradients help? *CoRR*, abs/2104.00979, 2021b.

Acharya, J., Canonne, C. L., Singh, A. V., and Tyagi, H. Optimal rates for nonparametric density estimation under communication constraints. *CoRR*, abs/2107.10078, 2021c.

Acharya, J., Canonne, C. L., Sun, Z., and Tyagi, H. Unified lower bounds for interactive high-dimensional estimation under information constraints. *CoRR*, abs/2010.06562, 2021d.

Acharya, J., Sun, Z., and Zhang, H. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pp. 48–78. PMLR, 2021e.

Allen, J. e. a. Smartnoise core differential privacy library. https://github.com/opendp/smartnoise-core.

Asi, H. and Duchi, J. C. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020.

Backstrom, L., Dwork, C., and Kleinberg, J. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web*, pp. 181–190, 2007.

Barnes, L. P., Han, Y., and Özgür, A. Fisher information for distributed estimation under a blackboard communication protocol. In *ISIT*, pp. 2704–2708. IEEE, 2019.

Barnes, L. P., Chen, W., and Özgür, A. Fisher information under local differential privacy. *IEEE J. Sel. Areas Inf. Theory*, 1(3):645–659, 2020a.

Barnes, L. P., Han, Y., and Özgür, A. Lower bounds for learning distributions under communication constraints via fisher information. *J. Mach. Learn. Res.*, 21:Paper No. 236, 30, 2020b. ISSN 1532-4435.

Berrett, T. and Butucea, C. Classification under local differential privacy. *arXiv preprint arXiv:1912.04629*, 2019.

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

Devroye, L. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, USA, 1986.

Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. *arXiv preprint arXiv:1712.01524*, 2017.

Dinur, I. and Nissim, K. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 202–210, 2003.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

Dong, J., Durfee, D., and Rogers, R. Optimal differential privacy composition for exponential mechanisms. In *International Conference on Machine Learning*, pp. 2597–2606. PMLR, 2020.

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

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy, data processing inequalities, and minimax rates. *arXiv preprint arXiv:1302.3203*, 2013b.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

Gillenwater, J., Joseph, M., and Kulesza, A. Differentially private quantiles. *In International conference on machine learning. PMLR*, 2021.

Györfi, L., Köhler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4 (8):e1000167, 2008.

IBM. Smartnoise core differential privacy library. https://github.com/IBM/differential-privacy-library.

Kairouz, P., Oh, S., and Viswanath, P. The composition theorem for differential privacy. In *International conference on machine learning*, pp. 1376–1385. PMLR, 2015.

Kamath, G., Liu, X., and Zhang, H. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10633–10660. PMLR, 2022.

Kaplan, H., Schnapp, S., and Stemmer, U. Differentially private approximate quantiles. In *International Conference on Machine Learning*, pp. 10751–10761. PMLR, 2022.

Kroll, M. On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics*, 15 (1):1783–1813, 2021.

Lalanne, C., Garivier, A., and Gribonval, R. On the statistical complexity of estimation and testing under privacy constraints. *arXiv preprint arXiv:2210.02215*, 2022a.

Lalanne, C., Gastaud, C., Grislain, N., Garivier, A., and Gribonval, R. Private quantiles estimation in the presence of atoms. *arXiv preprint arXiv:2202.08969*, 2022b.

Loukides, G., Denny, J. C., and Malin, B. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327, 2010.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Narayanan, A. and Shmatikov, V. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822, 2011.

Sweeney, L. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Thakurta, A. G., Vyrros, A. H., Vaishampayan, U. S., Kapoor, G., Freudiger, J., Sridhar, V. R., and Davidson, D. Learning new words. *Granted US Patents*, 9594741, 2017.

Tsybakov, A. B. Introduction to nonparametric estimation, 2009. *URL https://doi. org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wagner, I. and Eckhoff, D. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51 (3):1–38, 2018.

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

A. Proof of Lemma 3.1

We define

$$\bar{N} := \sum_{i=1}^{n} \mathbb{1}_{(F_X^{-1}(p)+\gamma,+\infty)}(X_i) \ .$$

Let $k \in \{-E(np)+1, \dots, n-E(np)\}$. We have the following event inclusion:

$$\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) \subset \left(\bar{N} \geq n-(E(np)+k)\right) \subset \left(\bar{N} \geq n(1-p)-k-1\right) \ .$$

$\bar{N}$ being a sum of independent Bernoulli random variables, we introduce $\eta := 1-p-\gamma\underline{\pi}$, a natural upper bound on the probability of success of each of these Bernoulli random variables. Hence, by multiplicative Chernoff bounds, whenever $\frac{\gamma\underline{\pi}}{\eta} - \frac{k+1}{n\eta} \geq 0$, which is equivalent to $k \leq n\gamma\underline{\pi} - 1$,

$$\begin{aligned}\mathbb{P}\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) &\leq \mathbb{P}\left(\bar{N} \geq n\eta\left(1+\frac{\gamma\underline{\pi}}{\eta}-\frac{k+1}{n\eta}\right)\right)\\ &\leq e^{-n\eta\left(\frac{\gamma\underline{\pi}}{\eta}-\frac{k+1}{n\eta}\right)^2/\left(2+\frac{\gamma\underline{\pi}}{\eta}-\frac{k+1}{n\eta}\right)} \ .\end{aligned}$$

By going further and imposing that $k+1 \leq \frac{1}{2}n\gamma\underline{\pi}$, we get

$$\mathbb{P}\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) \leq e^{-\frac{n\eta}{4}\left(\frac{\gamma\underline{\pi}}{\eta}\right)^2/\left(2+\frac{\gamma\underline{\pi}}{2\eta}\right)} \ .$$

Finally, by noticing that $\eta\left(\frac{\gamma\underline{\pi}}{\eta}\right)^2/\left(2+\frac{\gamma\underline{\pi}}{2\eta}\right) = \frac{\gamma^2\underline{\pi}^2}{2(1-p)-\frac{3}{2}\gamma\underline{\pi}} \geq \frac{\gamma^2\underline{\pi}^2}{2(1-p)}$,

$$\mathbb{P}\left(X_{(E(np)+k)} > F_X^{-1}(p)+\gamma\right) \leq e^{-\frac{\gamma^2\underline{\pi}^2}{8(1-p)}n} \ .$$

Now, looking at the other inequality, we define

$$\underline{N} := \sum_{i=1}^{n} \mathbb{1}_{(-\infty,F_X^{-1}(p)-\gamma)}(X_i) \ .$$

Like previously,

$$\left(X_{(E(np)+k)} < F_X^{-1}(p)-\gamma\right) \subset \left(\underline{N} \geq E(np)+k\right) \subset \left(\underline{N} \geq np+k-1\right) \ .$$

With the exact same techniques as previously, imposing the condition $k-1 \geq -\frac{1}{2}n\gamma\underline{\pi}$ gives

$$\mathbb{P}\left(X_{(E(np)+k)} < F_X^{-1}(p)-\gamma\right) \leq e^{-\frac{\gamma^2\underline{\pi}^2}{8p}n} \ .$$

Thus, under the various conditions specified for k, by union bound,

$$\mathbb{P}\left(|X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\right) \leq e^{-\frac{\gamma^2\underline{\pi}^2}{8p}n} + e^{-\frac{\gamma^2\underline{\pi}^2}{8(1-p)}n} \ .$$

Now define $I := \{k \in \{-E(np), \dots, n-E(np)\}||X_{(E(np)+k)} - F_X^{-1}(p)| \leq \gamma\}$. Notice that since $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, I is an integer interval. Which means that if $a \in I \leq b \in I$, then $[a,b] \cap \mathbb{Z} \subset I$. As a consequence, if $|X_{(E(np)+k)} - F_X^{-1}(p)| \leq \gamma$ for two integers k_1 and k_2, it is also the case for all the integers between them. By union bound, we get

$$\mathbb{P}\left(\sup_{k\in J}|X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\right) \leq 2e^{-\frac{\gamma^2\underline{\pi}^2}{8p}n} + 2e^{-\frac{\gamma^2\underline{\pi}^2}{8(1-p)}n} \ ,$$

where

$$J := \left\{\max\left(-E(np)+1, -E\left(\frac{1}{2}n\gamma\underline{\pi}\right)+1\right), \dots, \min\left(n-E(np), E\left(\frac{1}{2}n\gamma\underline{\pi}\right)-1\right)\right\} \ .$$

B. Proof of Lemma 3.2

The following fact is a direct consequence of Lemma 2.1 in Chapter 5 of Luc Devroye's book 1986.

Fact B.1 (Concentration of the gaps for uniform samples)**.** *Let* $X_1, \dots, X_n \overset{i.i.d.}{\sim} U([0,1])$*, the uniform distribution on* $[0,1]$*. Denoting* $\Delta_1 := X_{(1)}, \Delta_2 := X_{(2)} - X_{(1)}, \dots, \Delta_n := X_{(n)} - X_{(n-1)}$*, and* $\Delta_{n+1} := 1 - X_{(n)}$*, for any* $\gamma > 0$ *such that* $\gamma < \frac{1}{n+1}$*,*

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) = \left(1 - (n+1)\gamma\right)^n \ .$$

We give a proof here for completeness. The first step consists in characterizing the distribution of $(\Delta_1, \dots, \Delta_n)$. Let $h : \mathbb{R}^n \to \mathbb{R}$ be a positive Borelian function. By the transfer theorem,

$$\int h(\Delta_1, \dots, \Delta_n) d\mathbb{P}(\Delta_1, \dots, \Delta_n) = \int h(X_{(1)}, X_{(2)} - X_{(1)}, \dots, X_{(n)} - X_{(n-1)}) d\mathbb{P}(X_{(1)}, \dots, X_{(n)}) \ .$$

Furthermore, $(X_{(1)}, \dots, X_{(n)})$ follows a uniform distribution on the set of n ordered points in $[0,1]$. Hence,

$$\int h(\Delta_1, \dots, \Delta_n) d\mathbb{P}(\Delta_1, \dots, \Delta_n) = \frac{1}{n!} \int h(X_1, X_2 - X_1, \dots, X_n - X_{n-1}) \mathbb{1}_{0 \le X_1 \le \dots \le X_n \le 1} dX_1 \dots dX_n \ .$$

Finally, the variable swap $\delta_1 = X_1, \delta_2 = X_2 - X_1, \dots, \delta_n = X_n - X_{n_1}$ that has a jacobian of 1, same as its inverse (both transformations are triangular matrices with only 1's on the diagonal), gives that

$$\int h(\Delta_1, \dots, \Delta_n) d\mathbb{P}(\Delta_1, \dots, \Delta_n) = \frac{1}{n!} \int h(\delta_1, \dots, \delta_n) \mathbb{1}_{0 \le \delta_1, \dots, 0 \le \delta_n, \sum_{i=1}^n \delta_i \le 1} d\delta_1 \dots d\delta_n \ .$$

The last equation means that $(\Delta_1, \dots, \Delta_n)$ follows a uniform distribution on the simplex $\left\{0 \le \Delta_1, \dots, 0 \le \Delta_n, \sum_{i=1}^n \Delta_i \le 1\right\}$. The probability $\mathbb{P}\left(\min_i \Delta_i > \gamma\right)$ may now be computed as

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) = \frac{1}{n!} \int \mathbb{1}_{\gamma < \delta_1, \dots, \gamma < \delta_n, \sum_{i=1}^n \delta_i < 1 - \gamma} \mathbb{1}_{0 \le \delta_1, \dots, 0 \le \delta_n, \sum_{i=1}^n \delta_i \le 1} d\delta_1 \dots d\delta_n,$$

and by considering the variable swap $\delta_i' := \frac{\delta_i - \gamma}{1 - (n+1)\gamma}$ (which is separable) of which the jacobian of the inverse is $(1 - (n+1)\gamma)^n$,

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) = \frac{(1 - (n+1)\gamma)^n}{n!} \int \mathbb{1}_{0 < \delta_1', \dots, 0 < \delta_n', \sum_{i=1}^n \delta_i' < 1} d\delta_1' \dots d\delta_n' = (1 - (n+1)\gamma)^n \ .$$

This concludes the proof of Fact B.1. Now, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\pi$ where π is a density on $[0,1]$ w.r.t. Lebesgue's measure such that $\bar{\pi} \in \mathbb{R} \ge \pi \ge \underline{\pi} \in \mathbb{R} > 0$ almost surely. In particular, the data is not necessary uniform. By a coupling argument, if $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} U([0,1])$, $\left(F_\pi^{-1}(U_1), \dots, F_\pi^{-1}(U_n)\right)$ has the same distribution as $(X_1, \dots, X_n)$. We can furthermore notice that

$$\forall p, q \in (0,1), \epsilon > 0, \quad , |p - q| > \epsilon \implies \left|F_\pi^{-1}(p) - F_\pi^{-1}(q)\right| > \frac{\epsilon}{\bar{\pi}} \ .$$

Indeed, the lower bound $\pi \ge \underline{\pi}$ ensures that F_π is a bijection and that so does its inverse. The upper bound $\bar{\pi} \ge \pi$ ensures that F_π cannot grow too fast, and thus that its inverse is not too flat. Formally,

$$\forall a, b, \quad |F_\pi(b) - F_\pi(a)| = \left|\int_a^b \pi\right| \le \bar{\pi}|b - a|.$$

In particular, it holds for $b = F_\pi^{-1}(p)$ and $a = F_\pi^{-1}(q)$.

Consequently, if $\Delta_1' := U_{(1)}, \Delta_2' := U_{(2)} - U_{(1)}, \dots, \Delta_n' := U_{(n)} - U_{(n-1)}$, and $\Delta_{n+1}' := 1 - U_{(n)}$,

$$\mathbb{P}\left(\min_i \Delta_i > \gamma\right) \ge \mathbb{P}\left(\min_i \Delta_i' > \bar{\pi}\gamma\right) = \left(1 - (n+1)\bar{\pi}\gamma\right)^n \ .$$

Finally, let us simplify this expression to a easy-to-handle one. If $\gamma < \frac{n}{2\bar{\pi}}$,

$$\mathbb{P}\left(\min_i \Delta_i > \frac{\gamma}{n^2}\right) = \left(1 - \frac{n+1}{n}\frac{\bar{\pi}\gamma}{n}\right)^n \geq \left(1 - \frac{2n}{n}\frac{\bar{\pi}\gamma}{n}\right)^n = \left(1 - \frac{2\bar{\pi}\gamma}{n}\right)^n \ .$$

Furthermore, for any $x \in (0, 1/2)$ and $n \geq 1$, by the Taylor-Lagrange formula, there exist $c \in \left(-\frac{x}{n}, 0\right)$

$$\left(1 - \frac{x}{n}\right)^n = e^{n\ln\left(1-\frac{x}{n}\right)} = e^{n\left(-\frac{x}{n} - \frac{1}{2}\frac{1}{(1+c)^2}\frac{x^2}{n^2}\right)}$$

And so, when $n \geq 1$,

$$\left(1 - \frac{x}{n}\right)^n \geq e^{-2x}$$

In definitive, when $n \geq 1$ and $\gamma < \frac{1}{4\bar{\pi}}$

$$\mathbb{P}\left(\min_i \Delta_i > \frac{\gamma}{n^2}\right) \geq e^{-4\bar{\pi}\gamma} \ .$$

C. Proof of Theorem 3.3

For simplicity, let us assume that $E\left(\frac{1}{2}n\gamma\underline{\pi}\right) - 1 \leq \min\left(E(np) - 1, n - E(np)\right)$, which is for instance the case when $\gamma < \frac{2\min(p,1-p)}{\underline{\pi}}$, which we suppose. Furthermore, suppose that $\frac{1}{2}n\gamma\underline{\pi} \geq 2$, which is for instance the case when $n > 2/\min(p, 1-p)$ thank to the hypothesis on γ. By noting $K := E\left(\frac{1}{4}n\gamma\underline{\pi}\right)$, Lemma 3.1 says that,

$$\mathbb{P}\left(\sup_{k\in\{-K,\ldots,K\}} |X_{(E(np)+k)} - F_X^{-1}(p)| > \gamma\right) \leq 4e^{-\frac{\gamma^2\underline{\pi}^2}{8\max(p,(1-p))}n} \ ,$$

We call QC (for *quantile concentration*) this last event. Let $\delta > 0$ that satisfies $\delta < \frac{1}{4\bar{\pi}}$. We define the event $G := \left(\min_i \Delta_i > \frac{\delta}{n^2}\right)$ (for *gaps*). Lemma 3.2 ensures that

$$\mathbb{P}\left(G^c\right) \leq 1 - e^{-4\bar{\pi}\delta} \ .$$

Conditionally to QC, denoting by q the output of QExp, $|q - F_\pi^{-1}(p)| > \gamma \implies \mathfrak{E} \geq K - 1 \geq K/2$ whenever $n \geq 4/(\gamma\underline{\pi})$. By also working conditionally to G, and in order to apply Fact 2.1, we look for a $\beta > 0$ such that

$$K/2 = 2\frac{\ln(n^2) + \ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{1}{\beta}\right)}{\epsilon} \ ,$$

which gives

$$\beta = \frac{n^2}{\delta}e^{-\frac{\epsilon E\left(\frac{1}{4}n\gamma\underline{\pi}\right)}{4}} \ .$$

Note that even if Fact 2.1 is stated for $\beta \in (0, 1)$, its conclusion remains obviously true for $\beta \geq 1$.

Finally,

$$\begin{aligned}\mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma\right) &\leq \mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma, QC, G\right) + \mathbb{P}\left(QC^c\right) + \mathbb{P}\left(G^c\right) \\ &\leq \frac{en^2}{\delta}e^{-\frac{\epsilon n\gamma\underline{\pi}}{16}} + 1 - e^{-4\bar{\pi}\delta} + 4e^{-\frac{\gamma^2\underline{\pi}^2}{8\max(p,1-p)}n},\end{aligned}$$

and by fixing $\delta := \frac{n\sqrt{e}}{2\sqrt{2\bar{\pi}}}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32}}$, because $1 - e^{-4\bar{\pi}\delta} \leq 8\bar{\pi}\delta$ for any $\delta > 0$,

$$\mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma\right) \leq 4n\sqrt{2e\bar{\pi}}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32}} + 4e^{-\frac{\gamma^2\underline{\pi}^2}{8\max(p,(1-p))}n} \ .$$

D. Proof of Corollary 3.4

IndExp is the application of m independent QExp procedures but with privacy parameter $\frac{\epsilon}{m}$ in each. A union bound on the events that check if each quantile is off by at least γ gives the result by Theorem 3.3.

E. Proof of Theorem 3.5

For simplicity, let us assume that $E\left(\frac{1}{2}n\gamma\underline{\pi}\right) - 1 \leq \min\left(E(np_1) - 1, n - E(np_m)\right)$, which is for instance the case when $\gamma < \frac{2\min_i \min(p_i, 1-p_i)}{\underline{\pi}}$, which we suppose. Furthermore, suppose that $\frac{1}{2}n\gamma\underline{\pi} \geq 2$, which is for instance the case when $n > 2/\min_i \min(p_i, 1-p_i)$ thank to the hypothesis on γ. . By noting $K := E\left(\frac{1}{4}n\gamma\underline{\pi}\right)$, Lemma 3.1 says that for any $i \in \{1, \dots, m\}$,

$$\mathbb{P}\left(\sup_{k \in \{-K,\dots,K\}} |X_{(E(np_i)+k)} - F_X^{-1}(p_i)| > \gamma\right) \leq 4e^{-\frac{\gamma^2\underline{\pi}^2}{8C_{p_1,\dots,p_m}}n} ,$$

where $C_{p_1,\dots,p_m} := \max_i \left(\max\left(p_i, (1-p_i)\right)\right)$. We define the event QC (for *quantile concentration*),

$$QC := \bigcap_{i=1}^{m} \left(\sup_{k \in \{-K,\dots,K\}} |X_{(E(np_i)+k)} - F_X^{-1}(p_i)| \leq \gamma\right) .$$

By union bounds,

$$\mathbb{P}\left(QC^c\right) \leq 4me^{-\frac{\gamma^2\underline{\pi}^2}{8C_{p_1,\dots,p_m}}n} .$$

Let $\delta > 0$ that satisfies $\delta < \frac{1}{4\bar{\pi}}$. We define the event $G := \left(\min_i \Delta_i > \frac{\delta}{n^2}\right)$ (for *gaps*). Lemma 3.2 ensures that

$$\mathbb{P}\left(G^c\right) \leq 1 - e^{-4\bar{\pi}\delta} .$$

Conditionally to QC, denoting by $\mathbf{q}$ the output of RecExp, $\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma \implies \mathfrak{E} \geq K - 1 \geq K/2$ whenever $n \geq 4/(\gamma\underline{\pi})$. By also working conditionally to G, and in order to apply Fact 2.2, we look for a $\beta > 0$ such that

$$K/2 = 2(\log_2 m + 1)^2 \frac{\ln(n^2) + \ln\left(\frac{1}{\delta}\right) + \ln m + \ln\left(\frac{1}{\beta}\right)}{\epsilon} ,$$

which gives

$$\beta = \frac{n^2 m}{\delta} e^{-\frac{\epsilon E\left(\frac{1}{4}n\gamma\underline{\pi}\right)}{4(\log_2 m+1)^2}} .$$

Note that even if Fact 2.2 is stated for $\beta \in (0, 1)$, its conclusion remains obviously true for $\beta \geq 1$.

Finally,

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) &\leq \mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma, QC, G\right) + \mathbb{P}\left(QC^c\right) + \mathbb{P}\left(G^c\right) \\ &\leq \frac{en^2m}{\delta} e^{-\frac{\epsilon n\gamma\underline{\pi}}{32(\log_2 m+1)^2}} + 1 - e^{-4\bar{\pi}\delta} + 4me^{-\frac{\gamma^2\underline{\pi}^2}{8C_{p_1,\dots,p_m}}n},\end{aligned}$$

and by fixing $\delta := \frac{n\sqrt{em}}{2\sqrt{2\bar{\pi}}} e^{-\frac{\epsilon n\gamma\underline{\pi}}{32(\log_2 m+1)^2}}$, we get that,

$$\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq 4n\sqrt{2e\bar{\pi}m}e^{-\frac{\epsilon n\gamma\underline{\pi}}{32(\log_2 m+1)^2}} + 4me^{-\frac{\gamma^2\underline{\pi}^2}{8C_{p_1,\dots,p_m}}n} .$$

F. Proof of Lemma 4.1

By definition of u_{QExp} we have $-n \leq u_{\text{QExp}}\big((X_1, \dots, X_n), q\big) \leq 0$ for any input, hence using that $0 \leq \gamma \leq 1/4$ we get

$$\begin{aligned}\mathbb{P}\Big(|q - t| > \gamma\Big) &= \frac{\int_{[0,1]\setminus[t-\gamma,t+\gamma]} e^{\frac{\epsilon}{2}u_{\text{QExp}}\big((X_1,\dots,X_n),q\big)}dq}{\int_{[0,1]} e^{\frac{\epsilon}{2}u_{\text{QExp}}\big((X_1,\dots,X_n),q\big)}dq} \\ &\geq \frac{\int_{[0,1]\setminus[t-\gamma,t+\gamma]} e^{-\frac{\epsilon}{2}n}dq}{\int_{[0,1]} e^0 dq} \\ &\geq (1 - 2\gamma)e^{-\frac{\epsilon}{2}n} \\ &\geq \frac{1}{2}e^{-\frac{\epsilon}{2}n} .\end{aligned}$$

G. Proof of Lemma 4.2

Let us consider a specific bin of the histogram b. Let $\gamma > 0$. Denoting by $\|\cdot\|_{\infty,b}$ the infinite norm restrained to the support of b, which is a semi-norm, we have

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty,b} > \gamma\right) = \mathbb{P}\left(\left\|\frac{1}{nh}\left(\sum_{i=1}^{n} \mathbb{1}_b(X_i) + \frac{2}{\epsilon}\mathcal{L}\right) - \pi\right\|_{\infty,b} > \gamma\right)$$

where $\mathcal{L} \sim \text{Lap}(1)$, a centered Laplace distribution of parameter 1. So,

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty,b} > \gamma\right) &= \mathbb{P}\left(\left\|\left(\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_b(X_i) - \pi\right) + \frac{2}{nh\epsilon}\mathcal{L}\right\|_{\infty,b} > \gamma\right) \\
&\overset{\text{triangular inequality}}{\leq} \mathbb{P}\left(\left\|\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_b(X_i) - \pi\right\|_{\infty,b} > \gamma/2\right) + \mathbb{P}\left(\left|\frac{2}{nh\epsilon}\mathcal{L}\right| > \gamma/2\right)
\end{aligned}$$

Let us first control the first term. Since π is L Lipschitz, $\forall x \in b, \left|\pi(x) - \frac{1}{h}\int_b \pi\right| \leq \frac{Lh}{2}$. So, when $\gamma > Lh$,

$$\left(\left\|\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_b(X_i) - \pi\right\|_{\infty,b} > \gamma/2\right) \subset \left(\left|\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_b(X_i) - \frac{1}{h}\int_b \pi\right| > \gamma/2 - Lh/2\right) .$$

Finally, notice that the family $\left(\mathbb{1}_b(X_i)\right)_i$ is a family of i.i.d. Bernoulli random variables of probability of success $\int_b \pi$. By Hoeffding's inequality,

$$\mathbb{P}\left(\left\|\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_b(X_i) - \pi\right\|_{\infty,b} > \gamma/2\right) \leq 2e^{-\frac{h^2(\gamma-Lh)^2}{4}n} .$$

The second term is controlled via a tail bound on the Laplace distribution as

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{2}{nh\epsilon}\mathcal{L}\right| > \gamma/2\right) &= \mathbb{P}\left(|\mathcal{L}| > \frac{\gamma nh\epsilon}{4}\right) \\
&= \int_{\frac{\gamma nh\epsilon}{4}}^{\infty} e^{-t}dt \\
&= e^{-\frac{\gamma hn\epsilon}{4}} .
\end{aligned}$$

So, if $\gamma > Lh$,

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty,b} > \gamma\right) \leq 2e^{-\frac{h^2(\gamma-Lh)^2}{4}n} + e^{-\frac{\gamma hn\epsilon}{4}} .$$

Finally, a union bound on all the bins gives that if $\gamma > Lh$,

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_{\infty} > \gamma\right) \leq \frac{2}{h}e^{-\frac{h^2(\gamma-Lh)^2}{4}n} + \frac{1}{h}e^{-\frac{\gamma hn\epsilon}{4}} .$$

H. Proof of Lemma 4.3

We have,

$$\begin{aligned}
F_{\hat{\pi}}\left(F_{\pi}^{-1}(p) + \frac{\alpha}{\underline{\pi}}\right) &\overset{\|\hat{\pi}-\pi\|_{\infty}\leq\alpha}{\geq} F_{\pi}\left(F_{\pi}^{-1}(p) + \frac{\alpha}{\underline{\pi}}\right) - \alpha \\
&\overset{\pi\geq\underline{\pi}}{\geq} F_{\pi}\left(F_{\pi}^{-1}(p)\right) + \frac{\alpha}{\underline{\pi}}\underline{\pi} - \alpha \\
&= F_{\pi}\left(F_{\pi}^{-1}(p)\right) = p .
\end{aligned}$$

So,

$$F_{\hat{\pi}}^{-1}(p) \leq F_{\pi}^{-1}(p) + \frac{\alpha}{\underline{\pi}} \ .$$

Furthermore, for any $t \in \left[\frac{2\alpha}{\underline{\pi}}, F_{\pi}^{-1}(p)\right]$,

$$\begin{aligned} F_{\hat{\pi}}\left(F_{\pi}^{-1}(p) - t\right) &\overset{\|\hat{\pi}-\pi\|_\infty \leq \alpha}{\leq} F_{\pi}\left(F_{\pi}^{-1}(p) - t\right) + \alpha \\ &\overset{\pi \geq \underline{\pi}}{\leq} F_{\pi}\left(F_{\pi}^{-1}(p)\right) - t\underline{\pi} + \alpha \\ &< F_{\pi}\left(F_{\pi}^{-1}(p)\right) - \frac{2\alpha}{\underline{\pi}}\underline{\pi} + \alpha \\ &= F_{\pi}\left(F_{\pi}^{-1}(p)\right) - \alpha < p \ . \end{aligned}$$

So, for any $t \in \left(\frac{2\alpha}{\underline{\pi}}, F_{\pi}^{-1}(p)\right)$;

$$F_{\hat{\pi}}^{-1}(p) \geq F_{\pi}^{-1}(p) - t \ ,$$

and finally,

$$F_{\hat{\pi}}^{-1}(p) \geq F_{\pi}^{-1}(p) - \frac{2\alpha}{\underline{\pi}} \ .$$

I. Proof of Theorem 4.4

Given $\gamma \in \left(\frac{2Lh}{\underline{\pi}}, \gamma_0\right)$, $\frac{\gamma\underline{\pi}}{2} \geq \frac{2\underline{\pi}Lh}{2\underline{\pi}} = Lh$. So, Lemma 4.2 applies and gives that

$$\mathbb{P}\left(\|\hat{\pi}^{\text{hist}} - \pi\|_\infty > \frac{\gamma\underline{\pi}}{2}\right) \leq \frac{1}{h}e^{-\frac{\gamma\underline{\pi}hn\epsilon}{8}} + \frac{2}{h}e^{-\frac{h^2}{4}\left(\frac{\gamma\underline{\pi}}{2} - Lh\right)^2 n} \ .$$

Furthermore, $I = F_\pi\left((\gamma_0, 1-\gamma_0\right)$. So,

$$\forall p \in I, \quad \gamma_0 < F_{\pi}^{-1}(p) < 1 - \gamma_0 \ .$$

In particular, when $\hat{\pi}^{\text{hist}}$ satisfies $\|\hat{\pi}^{\text{hist}} - \pi\| \leq \frac{\gamma\underline{\pi}}{2}$, Lemma 4.2 applies and gives

$$\forall p \in I, \quad |F_{\hat{\pi}^{\text{hist}}}^{-1}(p) - F_{\pi}^{-1}(p)| \leq \gamma \ .$$

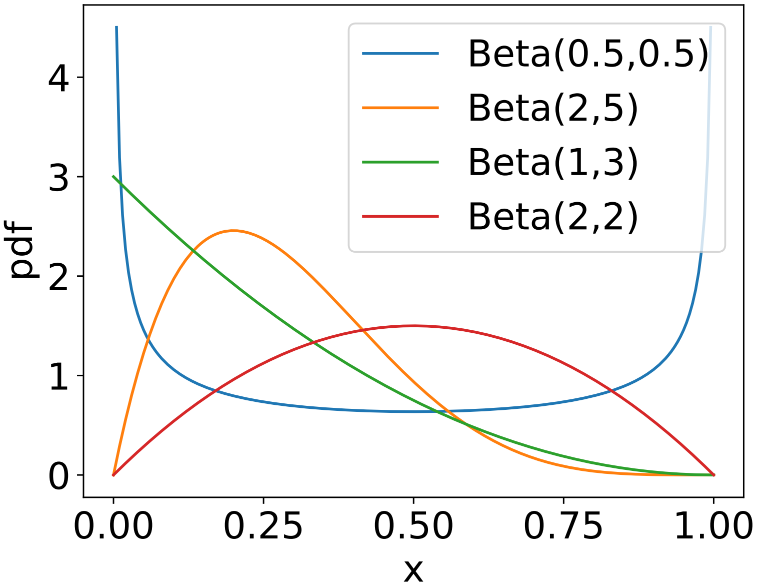
This is equivalent to

$$\forall p \in I, \quad \|F_{\hat{\pi}^{\text{hist}}}^{-1}(p) - F_{\pi}^{-1}(p)\|_{\infty, I} \leq \gamma \ .$$

Finally,

$$\mathbb{P}\left(\|F_{\hat{\pi}^{\text{hist}}}^{-1} - F_{\pi}^{-1}\|_{\infty, I} > \gamma\right) \leq \frac{1}{h}e^{-\frac{\gamma\underline{\pi}hn\epsilon}{8}} + \frac{2}{h}e^{-\frac{h^2}{4}\left(\frac{\gamma\underline{\pi}}{2} - Lh\right)^2 n} ;.$$

J. Distributions for the experiments



pdf : ”probability distribution function”, is the density w.r.t. Lebesgue’s measure.

Figure 2. Distributions used for the experiments