Locally Private Mean Estimation: Z-test and Tight Confidence Intervals

Marco Gaboardi
SUNY at Buffalo

Ryan Rogers

Or Sheffet
University of Alberta

Abstract

This work provides tight upper- and lower-bounds for the problem of mean estimation under differential privacy in the local-model, when the input is composed of n i.i.d. drawn samples from a Gaussian. Our algorithms result in a $(1-\beta)$-confidence interval for the underlying distribution's mean μ of length $O\left(\sigma\sqrt{\log(n/\beta)\log(1/\beta)}/\epsilon\sqrt{n}\right)$. In addition, our algorithms leverage on binary search using local differential privacy for quantile estimation, a result which may be of separate interest. Moreover, our algorithms have a matching lower-bound, where we prove that any one-shot (each individual is presented with a single query) local differentially private algorithm must return an interval of length $\Omega\left(\sigma\sqrt{\log(1/\beta)}/\epsilon\sqrt{n}\right)$.

1 Introduction

This work focuses on the task of *mean estimation* in the local-model. The problem is composed of n samples drawn from a Gaussian $X_1, ..., X_n \sim_{\text{i.i.d}} \mathcal{N}(\mu, \sigma^2)$ such that $\mu \in [-R, R]$ for some known bound R, and σ is either provided as an input (known variance case) or left unspecified (unknown variance case). We point out that the privacy analysis in our algorithms holds even if the assumption of normal data is not satisfied, whereas our utility analysis relies on this assumption. The goal of our algorithms is to provide an estimation of μ, which may be represented in multiple forms. The classical approach in statistical inference is to represent the likelihood of each point on the real line to be μ as a probability distribution — where in the case of known variance (Z-test) the output is a Gaussian distribution, and in the case of unknown variance (T-test) the output is a t-distribution. This likelihood allows an analyst to estimate a *confidence interval* I s.t. $\mathbb{P}[\mu \in I] \geq 1-\beta$, where non-privately it holds that $|I| = O(\sigma/\sqrt{n})$ (assuming β is a constant). Based on confidence intervals, one is able to reject (or fail-to-reject) certain hypotheses about μ, such as the hypothesis that $\mu = 0$ or that the means of two separate collections of samples ($X_1, ..., X_n$ and $Y_1, ..., Y_m$) are identical.

Our Contribution. The goal of this work is to provide upper- and lower-bounds for the problem of mean-estimation under (ϵ, δ)-local differentially private (LDP) assuming the data is drawn from an unknown Gaussian. On the upper-bound side, in the case of known variance we design a (ϵ, δ)-LDP algorithm, which yields a confidence interval of length $O(\sigma \cdot \sqrt{\log(n)}/\epsilon\sqrt{n})$ provided that $n = \Omega(\frac{\log(R/\sigma)}{\epsilon^2})$; and in the case of unknown variance we give an algorithm that returns a confidence interval of similar length assuming we have a lower-bound on the value of the unknown σ. In the known variance case, our algorithm results in a private Z-test, which we also assess empirically. On the lower-bound side, we prove that any ϵ-LDP algorithm must return an interval whose length is $\Omega(\sigma/\epsilon\sqrt{n})$, proving the optimality of our technique up to a $\sqrt{\log(n)}$-factor.

1.1 Our Techniques: Overview

Basic Tools. In our algorithms, we use two basic LDP building blocks. These are the canonical *Randomized Response* (Warner, 1965; Kasiviswanathan et al., 2008) and *Bit Flipping* (in its various versions) (Erlingsson et al., 2014; Bassily and Smith, 2015; Bassily et al., 2017). The mechanisms are known, and, for completion, in Section 2 we provide utility bounds for these building blocks under randomly drawn input.

The Known Variance Case. In the known variance case, our approach is a direct LDP implementation of the ideas behind the algorithm of Karwa and Vadhan (2018) who provide a private confidence interval in the centralized model. We equipartition the interval where

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

μ is assumed to be between $[-R, R]$ into $d = \lceil \frac{2R}{\sigma} \rceil$ sub-intervals of length σ, and use the above-mentioned Bit Flipping mechanism to find the most likely interval. The most common interval must be within distance $\leq 2\sigma$ from the mean (with high probability) of the underlying Gaussian distribution. This allows us to narrow in on an interval I of length $O(\sigma\sqrt{\log(n/\beta)})$ which should hold n new points from the same distribution.

Once we have found this interval, we project each datapoint onto I and add Gaussian noise of $\mathcal{N}(0, \frac{2|I|^2\log(2/\delta)}{\epsilon^2})$ to the projection, and then average the outcomes. This implies we have n i.i.d sample points for a Gaussian of mean μ and variance $O(\frac{\sigma^2\log(n/\beta)\log(1/\delta)}{\epsilon^2})$.[1] Thus, $\widetilde{\mu}$, the average of these n noisy datapoints, is also sampled from a Gaussian, whose variance is $\widetilde{\sigma}^2 = O(\frac{\sigma^2\log(n/\beta)\log(1/\delta)}{\epsilon^2 n})$. We can thus represent the likelihood of each point on $\mathbb{R}$ to be the mean using a Gaussian $\mathcal{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$ which is our analog to the Z-test. Moreover, the interval of length $2\widetilde{\sigma}\sqrt{\ln(4/\beta)}$ centered at $\widetilde{\mu}$ is a $(1-\beta)$-confidence interval. Details appear in Section 3, where in Section 3.1 we present some empirical assessment of our Z-test.

The Unknown Small Variance Case. We then consider the case of unknown variance, where instead of knowing σ we are provided bounds on the smallest and largest (resp.) values of the variance: $\sigma_{\min}, \sigma_{\max}$. First, we illustrate our approach in the case where we know $\sigma_{\max} \leq 2R$. (This is of course the more natural case, as we think of R as large and σ as reasonable.) Later, we discuss how to deal with the case of general unknown variance.

In this case, the approach of Karwa and Vadhan (2018) is to estimate the variance using the pairwise differences of the datapoints. That is due to the property of Gaussians where the difference between two i.i.d samples is also a Gaussian of 0-mean and variance $2\sigma^2$. This however is an approach that only works in the centralized model, where one is able to observe two datapoints without noise. In the local model, we are forced to use a different approach.

The approach we follow is to do binary search for different quantiles of the Gaussian, a folklore approach which has appeared before in certain testers, and in particular in the work of Feldman (2017). Given a quantile $p \in (0, 1)$, a continuous and smooth distribution $\mathcal{P}$, our goal is to find the threshold point t such that $\mathbb{P}_{X\sim\mathcal{P}}[X < t] = p \pm \lambda$ for a given tolerance parameter $\lambda > 0$. In each iteration j, we hold an interval $I^{(j)}$ which is guaranteed to hold t, and we use the middle point of this interval as our current guess. Denoting $t^{(j)}$ as the current interval's mid-point, we use enough datapoint to estimate $\mathbb{P}_{X\sim\mathcal{P}}[X < t^{(j)}]$ up to error of λ, and then either halt (if the estimated probability is approximately p) or recurse on either the left- or right-half of the interval. Since our initial interval is $[-(R+\sigma_{\max}), R+\sigma_{\max}]$ (of length $< 6R$) and we must halt when we reach an interval of length $\Omega(\sigma_{\min})$ (we treat λ as a constant), then the number of iterations overall is $T = O(\log(R/\sigma_{\min}))$.

And so, we first run binary search till we find a point t_1 for which we estimate that $\mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}[X < t_1] \approx 50\%$. We then find a point t_2 for which we estimate that $\mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}[X < t_2] \approx 81.4\%$. Due to the properties of a Gaussian, $t_1 \approx \mu$ and $t_2 \approx \mu+\sigma$. Of course, we do not have access to the actual quantiles, but rather just an estimation of them, but we are still able to show that w.h.p. it holds that $0.5\sigma < t_2 - t_1 < 2\sigma$. (These bounds explain why taking λ as a constant suffices for our needs.) We can thus run the algorithm for known variance case with this estimation of the variance on the remainder of the datapoints. The full details of our algorithm appear in Section 4.

The General Unknown Variance Case. In the general case, where $\sigma_{\max}$ isn't known, we begin by testing to see if the variance is $> R$ or $< 2R$ by estimating the probability that a new datapoint falls inside the interval $[-2R, 2R]$. If this probability is large then we have that $\sigma < 2R$ and we can use the previous algorithm for unknown bounded variance; whereas if this probability is smaller it must be that $\sigma > R$, and we run a very different algorithm. Instead of binary search, we merely estimate $q_1 \stackrel{\text{def}}{=} \mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}[X < -R]$ using the first half of the points, and then estimate $q_2 \stackrel{\text{def}}{=} \mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}[X < R]$ using the latter half of the points. We then use the two resulting quantiles to plot a suitable curve of the Gaussian distribution based on comparing these thresholds ($-R$ and R) to the thresholds on the real line obtaining q_1 and q_2 over a standard normal $\mathcal{N}(0, 1)$. The key point is that both $-R$ and R are within distance $< 2\sigma$ of the true mean μ; so by known properties of the Gaussian distribution, estimating q_1 and q_2 up to an error of $O(1/\epsilon\sqrt{n})$ implies a similar error guarantee in estimating μ. Due to space considerations, this approach has been deferred to the full version (Gaboardi et al., 2018).

Lower Bounds. Lastly, we give bounds on any ϵ-LDP algorithm that approximates the mean of a Gaussian distribution. Formally, we say an algorithm (β, τ)-*solves the mean-estimation problem* if its input is a sample of n points drawn i.i.d from a Gaussian dis-

[1] Actually, this is an approximation of the distribution, since we clip the original Gaussian. However, since the probability mass we remove is $< {}^{\beta}/_{n}$, the TV-dist to this distribution is $< 1/n$.

tribution $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in [-R, R]$ for some given parameter R, and its output is an interval I such that $\mu \in I$ w.p. $\geq 1 - \beta$ and furthermore $\mathbb{E}\left[|I|\right] \leq \tau$. Note that the probability is taken over *both* the sample draws and the coin-tosses of the algorithm. We prove that any one-shot, where each datapoint is queried only once, ϵ-locally differentially private algorithm $\mathcal{M}$ that (β, τ)-solves that mean estimation problem must have that $\tau = \Omega\left(\sigma\sqrt{\log(1/\beta)}/\epsilon\sqrt{n}\right)$ and also $n = \Omega\left(\ln(\frac{R}{\beta \cdot \tau})/\epsilon^2\right)$. In addition, we also provide lower bounds for any one-shot ϵ-LDP algorithm that approximates the quantile of a given distribution $\mathcal{P}$ using i.i.d samples from $\mathcal{P}$. We comment that the recent result of Bun et al. (2018) shows that these bounds carry from ϵ-LDP mechanisms to (ϵ, δ)-LDP mechanisms.

1.2 Related Work

Several works have studied the intersection of differential privacy and statistics (Dwork and Lei, 2009; Smith, 2011; Chaudhuri and Hsu, 2012; Duchi et al., 2013a,b; Dwork et al., 2015) mostly focusing on robust statistics; but only a handful of works study rigorously the significance and power of hypotheses testing under differential privacy (Vu and Slavkovic, 2009; Uhler et al., 2013; Wang et al., 2015; Gaboardi et al., 2016; Kifer and Rogers, 2017; Cai et al., 2017; Sheffet, 2017; Karwa and Vadhan, 2018). Vu and Slavkovic (2009) looked at the sample size for privately testing the bias of a coin. Johnson and Shmatikov (2013), Uhler et al. (2013) and Yu et al. (2014) focused on the Pearson χ^2-test, showing that the noise added by differential privacy vanishes asymptotically as the number of datapoints goes to infinity, and propose a private χ^2-based test which they study empirically. Wang et al. (2015), Gaboardi et al. (2016), and Kifer and Rogers (2017) then revised the statistical tests themselves to incorporate the additional noise due to privacy as well as the randomness in the data sample. Cai et al. (2017) give a private identity tester based on noisy χ^2-test over large bins, Sheffet (2017) studies private Ordinary Least Squares using the JL transform, and Aliakbarpour et al. (2018) study identity and equivalence testing. All of these works however deal with the centralized-model of differential privacy.

Few additional works are highly related to this work. Karwa and Vadhan (2018) give matching upper- and lower-bounds on the confidence intervals for the mean of a population, also in the centralized model. Duchi et al. (2013a,b) give matching upper- and lower-bound on robust estimators in the local model, and in particular discuss mean estimation. However, their bounds are related to minmax bounds rather than mean estimation or Z-tests. Gaboardi and Rogers (2018) and Sheffet (2018) study the asymptotic power and the sample complexity (respectively) of a variety of chi-squared based hypothesis testing in the local model. Finally, we mention the related work of Feldman (2017) who also discusses mean estimation using a version of a statistical query oracle which is thus related to LDP. Similar to our approach, Feldman (2017) also uses the folklore approach of binary search in the case the input variance is significantly smaller than the given bounding interval.

2 Preliminaries

We will write the dataset $\boldsymbol{X} \stackrel{i.i.d.}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ where $\boldsymbol{X} = (X_1, \cdots, X_n)$. Our goals is to develop confidence intervals for the mean μ subject to local differential privacy in two settings: (1) known variance, (2) unknown variance. We assume that the mean μ is in some finite interval $\mu \in [-R, R]$ and similarly for the standard deviation $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, if it is not known a priori. We first present the definition of differential privacy in the *curator* model, where the algorithm takes a single element from universe $\mathcal{X}$ as input.

Definition 1 (Dwork et al. (2006b,a))**.** *An algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-differentially private (DP) if for all $x, x' \in \mathcal{X}$ and for all outcomes $S \subseteq \mathcal{Y}$, we have $\mathbb{P}\left[\mathcal{M}(x) \in S\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{M}(x') \in S\right] + \delta$.*

We then define *local* differential privacy, formalized by Kasiviswanathan et al. (2008), where each data entry is perturbed on its own.

Definition 2 (LR Oracle)**.** *Given a dataset $\boldsymbol{x}$, a* local randomizer oracle *$\mathtt{LR}_{\boldsymbol{x}}(\cdot, \cdot)$ takes as input an index $i \in [n]$ and an (ϵ, δ)-DP algorithm R, and outputs $y \in \mathcal{Y}$ chosen according to the distribution of $R(x_i)$, i.e. $\mathtt{LR}_{\boldsymbol{x}}(i, R) = R(x_i)$.*

Definition 3 (Kasiviswanathan et al. (2008))**.** *An algorithm $\mathcal{M} : \mathcal{X}^N \to \mathcal{Y}$ is (ϵ, δ)-*local differentially private *(LDP) if it accesses the input database $\boldsymbol{x} \in \mathcal{X}^n$ via the LR oracle $\mathtt{LR}_{\boldsymbol{x}}$ with the following restriction: if $\mathtt{LR}(i, R_j)$ for $j \in [k]$ are the $\mathcal{M}$'s invocations of $\mathtt{LR}_{\boldsymbol{x}}$ on index i, then each R_j for $j \in [k]$ is (ϵ_j, δ_j)- DP and $\sum_{j=1}^k \epsilon_j \leq \epsilon$, $\sum_{j=1}^k \delta_j \leq \delta$.*

In this work we present and prove bounds regarding *one-shot* mechanisms, where a user may be queried only once without any further rounds of interaction.

Definition 4. *We say a randomized mechanism $\mathcal{M}$ is a* one-shot local differentially private *if for any dataset input D, $\mathcal{M}$ interacts with datum x_i by first choosing a differentially private mechanism $\mathcal{M}_i$, applying $\mathcal{M}_i(x_i)$ and then only post-processes the resulting output without any further interaction with x_i. In other words, $\mathcal{M}$ has one-round of interaction with any datapoint.*

Note, the definition of a one-shot mechanism does not rule out choosing the separate mechanisms *adaptively.* It only rules out the possibility that $\mathcal{M}$ may re-visit the details of individual based on her prior responses.

We next define our utility goal, which is to find confidence intervals that contain the mean parameter μ with high probability. Our goal is to design an algorithm that is (ϵ, δ)-LDP and also produces a valid $(1-\beta)$-confidence interval.

Definition 5 (Confidence Interval)**.** *An algorithm $\mathcal{M}$ produces a $(1-\beta)$-confidence interval for the mean of the underlying Gaussian distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ if*

$$\underset{\boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}(\mu,\sigma^2), \mathcal{M}(\boldsymbol{X})}{\mathbb{P}} [\mu \in \mathcal{M}(\boldsymbol{X})] \geq 1-\beta$$

Useful Bounds. Throughout this paper, Φ is the cumulative distribution function of a standard normal $\mathcal{N}(0,1)$. We use several concentration bounds, especially for Gaussians, where it is known that for any $\beta \in (0, 1/2)$ we have $\mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[|X-\mu| > \sigma\sqrt{2\ln(2/\beta)}\right] \leq \beta$.

A useful tool in our analysis is the following well-known variation of McDiarmid's inequality.

Fact 6. *Let $X_1, ... X_n$ be n independent random variables. Denote $B_1, ..., B_n$ and $\mu_1, ..., \mu_n$ such that $\forall i, |X_i| \leq B_i$ and $\mathbb{E}[X_i] = \mu_i$. Then for any $t > 0$ we have $\mathbb{P}\left[|\sum_i X_i - \sum_i \mu_i| > t\right] \leq 2\exp\left(-2t^2/\sum_i B_i^2\right)$.*

Existing Locally Private Mechanisms. A basic approach to preserve differential privacy is to use additive random noise. Suppose each datum is sampled from an interval I of length ℓ. Then adding random noise taken from $\mathcal{N}\left(0, 2\ell^2\ln(2/\delta)/\epsilon^2\right)$ to each datum (independently) guarantees (ϵ,δ)-differential privacy (Dwork et al., 2006a).

Two other canonical ϵ-local differentially private algorithm are the *randomized response* algorithm (Warner, 1965) and the *bit flipping* mechanism (Erlingsson et al., 2014; Bassily and Smith, 2015). In the randomized response mechanism, each datum is a bit $b \in \{0,1\}$ and each datum is independently flipped w.p. $1/1+e^\epsilon$. The bit flipping mechanism is similar, but rather than associating with each datum one of two possible values, we associate it with one of d possible values by mapping it to one of the d vectors of the standard basis. Thus the bit flipping mechanism outputs a vector $\boldsymbol{V}_i \in \{0,1\}^d$ per datum, with each coordinate of $\boldsymbol{V}_i$ slightly skewed towards 0 or 1 in a fashion similar to randomized response. Building on these two mechanisms, there exists an estimator $\widehat{\theta_{\mathrm{RR}}}$ that leverages on the output of randomized response on n-bit input to estimate the number of 1s in the input; and an estimator $\widehat{\theta_{\mathrm{BF}}}$ that leverages on the n d-dimensional vectors outputted by bit flipping to estimate the histogram of the inputs on the d possible types. For brevity, we defer to the full version (Gaboardi et al., 2018) the formal description of both mechanisms and both estimators. However the following claim, which summarizes the utility of either mechanism under *randomly drawn* input, will be useful in the sequel for our results.

Claim 7. *Let $\mathcal{X}$ be a domain and let $\mathcal{D}$ be a distribution over this domain. Given a predicate $\phi : \mathcal{X} \to \{0,1\}$ we denote $p = \mathbb{E}_{X\sim\mathcal{D}}[\phi(X)]$; and given a partition $\psi : \mathcal{X} \to \{1,2,..,d\}$ and denote $\boldsymbol{q}$ as the vector $(\mathbb{E}_{X\sim\mathcal{D}}[\psi(X)=j])_{j=1}^d$. Given n i.i.d draws from $\mathcal{D}$, $x_1, ..., x_n$, denote by $\widehat{\theta_{\mathrm{RR}}}(n,\phi)$ the randomized response estimator applied to the n-bit input $\phi(x_1), \phi(x_2), ..., \phi(x_n)$; and denote by $\widehat{\theta_{\mathrm{BF}}}(n,\psi)$ the bit-flipping estimator over the n d-dimensional unit vectors $\boldsymbol{e}_{\psi(x_1)}, \boldsymbol{e}_{\psi(x_2)}, ..., \boldsymbol{e}_{\psi(x_n)}$. Fix any $\alpha, \beta \in (0, \frac{1}{2})$. Then if $n \geq \frac{2}{\alpha^2}\left(\frac{e^\epsilon+1}{e^\epsilon-1}\right)^2 \ln(\frac{4}{\beta})$ we have that $\mathbb{P}\left[|\frac{1}{n}\cdot\widehat{\theta_{\mathrm{RR}}}(n,\phi) - p| \leq \alpha\right] \geq 1-\beta$; and if $n \geq \frac{2}{\alpha^2}\left(\frac{e^{\epsilon/2}+1}{e^{\epsilon/2}-1}\right)^2 \ln(4d/\beta)$ we have that $\mathbb{P}\left[\|\frac{1}{n}\cdot\widehat{\theta_{\mathrm{BF}}}(n,\psi) - \boldsymbol{q}\|_\infty \leq \alpha\right] \geq 1-\beta$.*

3 Confidence Intervals for the Mean with Known Variance

In this section we assume that σ is known and we want to estimate a confidence interval for μ based on a sample of n users, subject to local differential privacy. As in Karwa and Vadhan (2018), we will break the algorithm into two parts. First, we discretize the interval $[-R-\sigma/2, R+\sigma/2]$ into bins of width σ, so that we have a collection of $d \overset{\mathrm{def}}{=} \lceil 2R/\sigma \rceil$ disjoint intervals

$$\mathcal{S}(\sigma) = \mathcal{S}_{-d}(\sigma) \cup \mathcal{S}_{-d+1}(\sigma) \cup \cdots \cup \mathcal{S}_d(\sigma) \qquad (1)$$

where $\mathcal{S}_i(\sigma) = [(i-1/2)\cdot\sigma, (i+1/2)\cdot\sigma]$. Denote $\phi : \mathbb{R} \to \{\boldsymbol{0}, \boldsymbol{e}_1, \boldsymbol{e}_2, .., \boldsymbol{e}_d\}$ as the function that maps each x to the indicating vector of the bin it resides in, and assigns any point outside the $[-R-\sigma/2, R+\sigma/2]$ interval the all-0 vector, we can now apply the Bit Flipping mechanism to estimate the histogram over the $d = \lceil 2R/\sigma \rceil$ bins. Next, we find the bin with the largest count, denoted j^*, and argue this bin is close up to two standard deviations to the true population mean μ. We then move to the second part of the algorithm, where we place an interval I of length $|I| = \widetilde{O}(\sigma)$ around the j^*-th bin which is likely to hold all remaining points (a point outside this interval is projected onto the nearest point in I). Adding Gaussian noise to each point suffices to make the noisy result (ϵ,δ)-differentially private, and yet we can still sum over all points and obtain an estimation of the population mean which is close up to $\widetilde{O}(\sigma/\sqrt{n})$. Details are given in Algorithm `KnownVar`.

Algorithm 1 Known Variance Case: `KnownVar`

Input: Data $\{x_1, \cdots, x_n\}$; σ, β, ϵ, R.

1: Set $n_1 = 800\left(\frac{e^{\epsilon/2}+1}{e^{\epsilon/2}-1}\right)^2 \ln\left(\frac{16R}{\sigma\cdot\beta}\right)$ and $n_2 = n - n_1$.

2: Partition the input into $\mathcal{U}_1 = \{1, \cdots, n_1\}$ and $\mathcal{U}_2 = \{n_1+1, \cdots, n\}$.

3: Denote ϕ as the partition of the real-line into the $d = \lceil 2R/\sigma \rceil$ bins as in Equation (1).

4: Apply Bit Flipping on $\mathcal{U}_1$: $\widetilde{\boldsymbol{p}} \leftarrow \frac{1}{n_1}\widehat{\theta_{\mathrm{BF}}}(n_1, \phi)$ and let j^* be the largest coordinate of $\widetilde{\boldsymbol{p}}$.

5: Set $\Delta = 2\sigma + \sigma\sqrt{2\log\left({}^{8n}/\beta\right)}$. Denote the interval

$$[s_1, s_2] = [j^*\sigma - \Delta, \quad j^*\sigma + \Delta] \tag{2}$$

and denote $\pi_{[s_1,s_2]}(x) = \min\{s_2, \max\{s_1, x\}\}$, namely the projection of x onto $[s_1, s_2]$.

6: Set $\widehat{\sigma}^2 = 8\Delta^2 \ln(2/\delta)/\epsilon^2$.

7: **foreach** $i \in \mathcal{U}_2$

set $\widetilde{x}_i = \pi_{[s_1,s_2]}(x_i) + N_i$ where $N_i \sim \mathcal{N}\left(0, \widehat{\sigma}^2\right)$.

8: Set $\widetilde{\mu} = \frac{1}{n_2}\sum_{i\in\mathcal{U}_2} \widetilde{x}_i$, and $\tau = \sqrt{\frac{\sigma^2+\widehat{\sigma}^2}{n_2}} \cdot \Phi^{-1}(1-\beta/8)$

Output: $I = [\widetilde{\mu} - \tau, \widetilde{\mu} + \tau] \cap [-R, R]$

The following two theorems prove that Algorithm `KnownVar` satisfies the required (proofs are deferred to the full version (Gaboardi et al., 2018)).

Theorem 8. *`KnownVar` is (ϵ, δ)-LDP.*

Theorem 9. *Let $\boldsymbol{X} \overset{i.i.d.}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and $I = $ `KnownVar`$(\boldsymbol{X}; \sigma, \beta, \epsilon, n, R)$. Set $d = \lceil {}^{2R}/\sigma \rceil$. If $n \geq 1600\left(\frac{e^{\epsilon/2}+1}{e^{\epsilon/2}-1}\right)^2 \log\left(\frac{8d}{\beta}\right)$, then $\mathbb{P}_{\boldsymbol{X},\texttt{KnownVar}}[\mu \in I] \geq 1 - \beta$. Furthermore,*

$$|I| = O\left(\sigma \cdot \frac{\sqrt{\log(n/\beta) \cdot \log(1/\beta) \cdot \log(1/\delta)}}{\epsilon\sqrt{n}}\right)$$

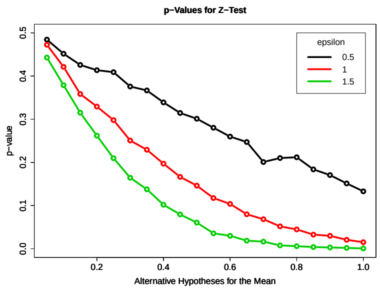
3.1 Experiment: Z-Test

As in Algorithm `KnownVar`, we denote $n_1 \stackrel{\text{def}}{=} 800 \cdot \left(\frac{e^{\epsilon/2}+1}{e^{\epsilon/2}-1}\right)^2 \cdot \log\left(\frac{8d}{\beta}\right)$ and $n_2 \stackrel{\text{def}}{=} n - n_1$. Following the proof of Theorem 9, we have that — under the assumption that no datapoint is clipped — all n_2 datapoints we use in the latter part of Algorithm 1 are sampled from $\mathcal{N}\left(\mu, \sigma^2 + \widehat{\sigma}^2\right)$. This allows us to infer that (w.p. $\geq 1-\beta$) the average of the n_2 datapoints in $\mathcal{U}_2$ is sampled from $\mathcal{N}\left(\mu, \frac{\sigma^2+\widehat{\sigma}^2}{n_2}\right)$. Just as in Algorithm 1, denoting $\widetilde{\mu}$ as the average of the noisy datapoints, we now can define an approximation of the likelihood: $\mathcal{P} = \mathcal{N}\left(\widetilde{\mu}, \frac{\sigma^2+\widehat{\sigma}^2}{n_2}\right)$. As a result, for any interval on the reals I we can associate a likelihood of $p_I \stackrel{\text{def}}{=} \mathbb{P}_{X\sim\mathcal{P}}[X \in I]$, and we know that w.p. $p_I \pm \beta$ it indeed holds that $\mu \in I$. This mimics the power of a Z-test (Hogg et al., 2005) — in particular we can now compare two intervals as to which one is more likely to hold μ, compare populations, etc.

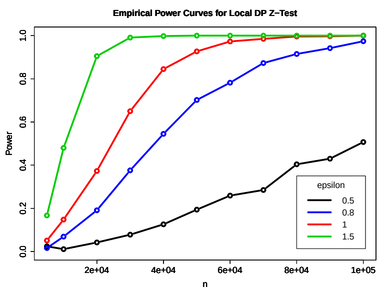
Note however that, as opposed to standard Z-test, the result of Algorithm 1 only gives confidence bounds up to an error of β. So for example, given two intervals I and I' we can safely argue that it is more likely that $\mu \in I$ than $\mu \in I'$ only when $p_I > p_{I'} + 2\beta$. Similarly, if we wish to draw an interval whose likelihood to contain μ is $1-\nu$ for some $\nu > 0$, we must pick a corresponding $(1-\nu+\beta)$-confidence interval from $\mathcal{P}$. Naturally, this limits us to the setting where $\beta < \nu$, or conversely: we can never allow for more certainty than the $1-\beta$ parameter specified as an input for Algorithm 1.

Subject to this caveat, Algorithm 1 allows us to perform Z-test in a similar fashion to the standard Z-test, after we omit the first n_1 datapoints from our sample. One of the more common uses of Z-test is to test whether a given sample behaves in a similar fashion to the general population. For example, suppose that the SAT scores of the entire population are distributed like a Gaussian of mean μ and variance σ^2. Taking a sample of SAT scores from one specific city, we can apply the Z-test to see if we can reject the null hypothesis that the score distribution in this city are distributed just as they are distributed in the general population. Should we have n samples of SAT scores which happen to be distributed from $\mathcal{N}\left(\mu', \sigma^2\right)$ for some $\mu' \neq \mu$, then sufficiently large n (with dependency on $|\mu' - \mu|$) should allow us to reject this null hypothesis with confidence $1-\nu$. We set to discover precisely this notion of utility, using our locally-private Z-test.

The Experiment: We tested our LDP Z-test on n i.i.d samples from a Gaussian. We set the null-hypothesis to be $H_0 : \mathcal{N}(0, 1)$, whereas the n samples were drawn from the alternative hypothesis $H_1 : \mathcal{N}(\mu', 1)$ with $\mu' > 0$. We run our experiments in the known variance $\sigma^2 = 1$ case with a fixed bound $R = 200$ and $\beta = 0.01$. In each set of experiments we vary ϵ while keeping $\delta = 10^{-9}$. In Figure 1a, we plot the average p-value over 1,000 trails for our Z-test when the data is actually generated with sample size $n = 200{,}000$ and mean μ' that varies. In Figure 1b, we plot the empirical power of our test over 1000 trails where we fix $\mu' = 3$ and vary the sample size n. Our figures show the tradeoffs between the privacy parameter, the alternate we are comparing the null to, and the sample size. The results themselves match the theory pretty well and emphasize the magnitude of the needed sample size. For $\epsilon = 1.5$ we need 10,000 sample points to reject the null hypothesis w.h.p. When $\epsilon = 0.5$, even 100,000 sample points do not suffice to reject the null hypothesis w.h.p despite the fact that the difference between the means of the null and the alternative is



(a) Average p-values with $n = 200,000$.



(b) Empirical power with alternate $\mu' = 3$.

Figure 1: Z-test experiments showing the empirical p-values and power averaged over 100 trials for various privacy parameters.

3 times greater than the variance. This is a setting where non-privately we can reject the null hypothesis with a sample size < 100. This illustrates (yet again) how LDP relies on the abundance of data.

4 Mean Estimation with Unknown (Bounded) Variance

In this section we discuss the problem of locally private mean estimation in the case the variance of the underlying population is unknown. For the ease of exposition, we separate this case into two sub-cases. The first sub-case is the one where we know that the variance is bounded by some $\sigma_{\max} \leq 2R$ and it is the sole focus of this section as it the more likely of the two. The second one is the case of very-large variance ($\sigma > R$), a case which Karwa and Vadhan (2018) do not analyze, and it is deferred to the full version (Gaboardi et al., 2018). As our lower bounds show, our algorithm must be provided bounds $\sigma_{\min} > 0$ and $\sigma_{\max} \leq 2R$ such that $\sigma \in [\sigma_{\min}, \sigma_{\max}]$. As we show, our parameters dependency on these upper- and lower-bounds on the variance is logarithmic (so, for example, $\sigma_{\min} > {}^1/_{R^2}$ is a useful bound for us).

Our overall approach in this section mimics the same approach from Algorithm 1. Our goal in to find a suitably large yet sufficiently tight interval $[s_1, s_2]$ that is likely to holds the latter part of the input. However, finding this $[s_1, s_2]$-interval cannot be done using the off-the-shelf Bit Flipping mechanism as that required we know the granularity of each bin in advance. Indeed, if we discretize the interval $[-R, R]$ with an upper-bound on the variance, each bin might be far too large and result in an interval $[s_1, s_2]$ which is far larger than the variance of the underlying population; and if we were to discretize $[-R, R]$ with a lower-bound on the variance we cannot guarantee substantial differences between the bins that are close to μ. And so, we abandon the idea of finding a histogram on the data. Instead, we propose finding a good approximation for σ using a quantile estimation based on a binary search. This result is likely to be of independent interest. Once we establish formal guarantees on our locally private binary search algorithm (privacy and utility bounds), we plug those into our confidence interval estimation algorithm in Subsection 4.2.

4.1 Locally Private Binary Search and Quantile Estimation

We now show how to estimate quantiles of a probability distribution using randomized response and binary search. We assume our domain $\mathcal{X}$ is contained in the real line and that there exists some distribution $\mathcal{P}$ over this domain. This defines the quantile of a threshold t as $p(t) = \mathbb{P}_{\mathcal{P}}[X < t]$. Given a target probability p^*, let t^* be the quantile we want to estimate, namely $p(t^*) = p^*$. Since our algorithm is randomized and therefore uses only estimations, we must allow for some error λ, and find some t such that $|p(t) - p^*| \leq \lambda$.

Our binary search begins with some bounded interval guaranteed to contain t^*, i.e. $t^* \in [Q_{\min}, Q_{\max}]$. Initially, we set $t^{(0)} = \frac{Q_{\max}+Q_{\min}}{2}$, and draw a subsample of size m, where m is chosen so that w.h.p. we can estimate $\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}\{X < t^{(0)}\}]$ using randomized response up to an error of λ. Denoting the randomized response estimator as $\widehat{\theta_{\rm RR}}^{(0)}$ one of the following three must holds. Either (i) $|\widehat{\theta_{\rm RR}}^{(0)} - p^*| \leq \lambda$, in which case we have found a good enough approximation for t^* and we may halt; or (ii) $\widehat{\theta_{\rm RR}}^{(0)} > p^* + \lambda$ in which case $t^{(0)}$ is too large, and so $t^* \in [Q_{\min}, t^{(0)}]$ and we recurse of the LHS half of the original interval; or (iii) $\widehat{\theta_{\rm RR}}^{(0)} < p^* - \lambda$ in which case $t^{(0)}$ is too small, and so $t^* \in [t^{(0)}, Q_{\max}]$ and we recurse of the RHS half of the original interval.

When does our binary search algorithm halts? If $\mathcal{P}$ is a pathological distribution, it may put 2λ probability mass on an infinitesimally small intervals to the left and right of t^*, forcing our binary search algorithm to continue for arbitrarily many rounds. To avoid such a case, we require an a-priori bound $\alpha_{\rm dist}$ on the length of an interval that can hold λ-probability mass; or alternatively, allow our algorithm to output any t such that $|t - t^*| \le \alpha_{\rm dist}$. The formal definition follows.

Definition 10. *An algorithm $\mathcal{M}$ is said to* $(\alpha_{\rm dist}, \alpha_{\rm quant}, \beta)$-approximate the p^*-quantile *over $\mathcal{P}$ under the guarantee that t^*,the bound such that $p^* = \mathbb{P}_{x\sim\mathcal{P}}[x \le t^*]$, is bounded $t^* \in [Q_{\min}, Q_{\max}]$, if it takes as input n i.i.d draws from $\mathcal{P}$ and returns $t \in [Q_{\min}, Q_{\max}]$ such that w.p. $\ge 1-\beta$ we have that either $|p^* - \mathbb{P}_{x\sim\mathcal{P}}[x \le t]| \le \alpha_{\rm quant}$ or that $|t - t^*| \le \alpha_{\rm dist}$.*

Provided with such a bound $\alpha_{\rm dist}$ we can bound the number of iterations in our binary search by T such that ${Q_{\max} - Q_{\min}}/{2^T} < \alpha_{\rm dist}$. A description of our binary search given such an iteration bound T is detailed in Algorithm `BinQuant`.

Algorithm 2 Quantile Estimation: `BinQuant`

Input: Data $\{x_1, \cdots, x_N\}$, target quantile p^*; ϵ, $[Q_{\min}, Q_{\max}]$, λ, T.
  Initialize $n = N/T$, $s_1 = Q_{\min}, s_2 = Q_{\max}$.
  for $j = 0, \cdots, T$ **do**
    Select users $\mathcal{U}^{(j)} = \{j\cdot n+1, j\cdot n+2, \cdots, (j+1)\cdot n\}$
    Set $t^{(j)} \leftarrow \frac{s_1+s_2}{2}$
    Denote $\phi^{(j)}(x) = \not\Vdash\{x < t^{(j)}\}$.
    Run randomized response on $\mathcal{U}^{(j)}$ and obtain $Z^{(j)} = \frac{1}{n}\widehat{\theta_{\rm RR}}(n, \phi^{(j)})$.
    if $(Z^{(j)} > p^* + \frac{\lambda}{2})$ **then**
      $s_2 \leftarrow t^{(j)}$
    else if $(Z^{(j)} < p^* - \frac{\lambda}{2})$ **then**
      $s_1 \leftarrow t^{(j)}$
    else
      break
Output: $t^{(j)}$

Two theorems summarize Algorithm 2's properties. Their proofs are deferred to the full version (Gaboardi et al., 2018).

Theorem 11. *`BinQuant` is ϵ-LDP.*

Theorem 12. *Let $\mathcal{P}$ be any distribution on the real line. For any $p^* \in (0,1)$ and any $Q_{\min}, Q_{\max}$ such that $q^* \in [Q_{\min}, Q_{max}]$, for any $\epsilon > 0$ and for any $\lambda, \tau, \beta \in (0, 1/2)$, Algorithm `BinQuant` indeed (τ, λ, β)-approximates the p^*-quantile if $T = \lceil \log_2(\frac{Q_{\max} - Q_{\min}}{\tau}) \rceil$ and its input is N i.i.d draws from $\mathcal{P}$, provided that $N \ge \frac{8T}{\lambda^2}\left(\frac{e^\epsilon+1}{e^\epsilon-1}\right)^2 \ln(4T/\beta)$.*

4.2 Locally Private Mean Estimation Using Quantile Estimation

We return to discuss the case where the underlying distribution of the data is Gaussian with unknown variance. Recall, our plan is to use quantile estimation to find an interval $[s_1, s_2]$ which is likely to contain most datapoints. This requires that we assess μ up to an error of about $\pm\sigma$ and also have an estimation of σ which is also fairly close to the true σ. I.e. denoting $\widetilde{\sigma}$ as our estimation, we would like to have $\frac{\sigma}{2} \le \widetilde{\sigma} \le 2\sigma$.

Our approach for obtaining such estimations of μ and σ is to apply the quantile estimation technique *twice*: once for $p^* = \frac{1}{2}$ where $t^* = \mu$, and once for the value of $p^* = \Phi(1) \approx 0.8413$ for which the corresponding threshold is $t^* = \mu + \sigma$. We argue next is that, since both thresholds are sufficiently close to the mean of the underlying distribution, we can set λ as a reasonable constant and guarantee that our estimations of the two thresholds are close up to a factor of $\sigma/4$ to the true thresholds. This required some calculations on the PDF of Gaussians which we defer to the full version (Gaboardi et al., 2018), but the end result is that it suffices to have error of $\lambda = 0.098$ in the first estimation, and an error of $\lambda = 0.052$ in the latter estimation. Note that in both cases we can set $\alpha_{\rm dist} = {\sigma_{\min}}/4$. Our LDP confidence interval estimator in the unknown variance case is given in Algorithm 3.

Algorithm 3 Unknown Variance Case: `UnkVar`

Input: Data $\{x_1, \cdots, x_N\}$;λ, R, $\sigma_{\min}$,$\sigma_{\max}$, ϵ, β
  Set $T^{\rm med} = \lceil \log_2(\frac{8R}{\sigma_{\min}}) \rceil$,$T^{\rm sd} = \lceil \log_2(\frac{8R + 4\sigma_{\max}}{\sigma_{\min}}) \rceil$.
  Set $n_1 = \frac{T^{\rm med}}{(0.098)^2} \cdot \left(\frac{e^\epsilon+1}{e^\epsilon-1}\right)^2 \cdot \ln(16T^{\rm med}/\beta)$, and $n_2 = \frac{T^{\rm sd}}{(0.052)^2} \cdot \left(\frac{e^\epsilon+1}{e^\epsilon-1}\right)^2 \cdot \ln(16T^{\rm sd}/\beta)$, and $n_3 = n - n_1 - n_2$.
  Init $\mathcal{U}_1 = \{1, \cdots, n_1\}$, $\mathcal{U}_2 = \{n_1+1, \cdots, n_1+n_2\}$, and $\mathcal{U}_3 = \{n_1+n_2+1, \cdots, n\}$.
  $\widehat{t_\mu} \leftarrow$ `BinQuant`$(\{x_i : i \in \mathcal{U}_1\}, 1/2; \ \ \epsilon, n, [-R, R], 0.098, T^{\rm med})$
  $\widehat{t_\sigma} \leftarrow$ `BinQuant`$(\{x_i : i \in \mathcal{U}_2\}, \Phi(1); \epsilon, n, [-R, R+\sigma_{\max}], 0.052, T^{\rm sd})$
  Set $\Delta = (\widehat{t_\sigma} - \widehat{t_\mu}) \cdot (\frac{1}{2} + 2\sqrt{2\ln(8n/\beta)})$
  Denote the interval $[s_1, s_2] = [\widehat{t_\mu} - \Delta, \widehat{t_\mu} + \Delta]$.
  Run steps 6-9 of Algorithm `KnownVar` over $\mathcal{U}_3$.

Theorem 13. *Let $\boldsymbol{X} \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Fix parameters ϵ, $\beta \in (0, 1/2)$. Given that $\mu \in [-R, R]$ and that $\sigma_{\min} \le \sigma \le \sigma_{\max} \le 2R$, if*

$$n \ge 1500\log_2(\tfrac{16R}{\sigma_{\min}}) \cdot \left(\tfrac{e^\epsilon+1}{e^\epsilon-1}\right)^2 \cdot \ln(\tfrac{16\log_2(16R/\sigma_{\min})}{\beta})$$

then the interval $\widehat{I}$ returned by Algorithm `UnkVar` satisfies that $\mathbb{P}_{\boldsymbol{X},\ \texttt{UnkVar}}\left[\widehat{I} \ni \mu\right] \ge 1-\beta$, and moreover

$$\widehat{I} = O\left(\sigma \cdot \frac{\sqrt{\log(n/\beta)\log(1/\beta)\log(1/\delta)}}{\epsilon\sqrt{n}}\right)$$

The proof of Theorem 13 is deferred to the full version (Gaboardi et al., 2018). It is interesting to compare the bounds of Theorems 9 and 13. "Replacing" the known quantity σ in Theorem 9 with the provided lower bound $\sigma_{\min}$ in Theorem 13, the sample complexity bound only increases by a $\log\log(R/\sigma_{\min})$-factor. Note in both algorithms we conclude in a similar fashion (averaging Gaussian noise), so, if we are to denote by m the number of points either algorithms use in their last parts, then both algorithms output intervals of length $\tilde{O}(\sigma/\epsilon\sqrt{m})$.

5 Lower Bounds

We begin our discussion on the bounds on the utility of any ϵ-locally private mechanism which is a one-shot mechanism, by presenting the following lemma. This lemma is a combination of two separate results. The one, Karwa and Vadhan's coupling argument that suggest that the "effective group privacy" between two n-size samples from either a distribution $\mathcal{P}$ or a distribution $\mathcal{Q}$ is roughly $n \cdot d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q})$. The second is a lemma, which originally appeared in Beimel et al. (2008) and then also appeared in a more formal way in Bun et al. (2018), that states that group privacy of altering k datums in the local scales proportional to $O(\epsilon\sqrt{k})$ rather $O(\epsilon k)$ in the centralized model. We combine the two into a single lemma, dealing with ϵ-LDP mechanisms over input drawn i.i.d from some distribution. This lemma is the main building block in all of our lower-bounds. Its proof, as well as all proofs in this section, are deferred to the full version (Gaboardi et al., 2018).

Lemma 14. *Let $\mathcal{M}$ be a one-shot local ϵ-differentially private mechanism. Let $\mathcal{P}$ and $\mathcal{Q}$ be two distributions, with $\Delta \stackrel{\text{def}}{=} d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q})$. Fix any $0 < \delta < e^{-1}$ and set $\epsilon^* = 8\epsilon\Delta\sqrt{n}\left(\sqrt{\frac{1}{2}\ln(2/\delta)} + 16\epsilon\Delta\sqrt{n}\right)$. Then, for any set of possible outputs S we have that*

$$\mathbb{P}_{\boldsymbol{X} \overset{\text{i.i.d}}{\sim} \mathcal{P}}[\mathcal{M}(\boldsymbol{X}) \in S] \le e^{\epsilon^*}\mathbb{P}_{\boldsymbol{X} \overset{\text{i.i.d}}{\sim} \mathcal{Q}}[\mathcal{M}(\boldsymbol{X}) \in S] + \delta$$

where the probability is taken over both the n i.i.d samples and over the coin-tosses of $\mathcal{M}$.

5.1 Lower Bounds for One-Shot ϵ-Locally Private Mechanisms

Leveraging on our main lemma, we can now prove lower bounds on the interval length and sample complexity of any one-shot ϵ-LDP algorithm that outputs a meaningful confidence interval. We focus on the case of a known variance, and our lower-bound shows the optimality of Algorithm `KnownVar` up to a $O(\sqrt{\log(n/\beta)})$-factor.

Theorem 15. *We say an algorithm (β, τ)*-solves the mean-estimation problem *(under known variance σ^2 and bound R) if its input is a sample of n points and its output is an interval I such that, if all n datapoints are i.i.d draws from $\mathcal{N}(\mu, \sigma^2)$ for some $\mu \in [-R, R]$ then w.p. $\ge 1 - \beta$ it holds that $\mu \in I$ and furthermore, $\mathbb{E}[|I|] \le \tau$. (The probability is taken over* both *the sample draws and the coin-tosses of the algorithm.)*

Fix any $\beta < 1/3$. Then any one-shot ϵ-locally differentially private algorithm $\mathcal{M}$ that (β, τ)-solves that mean estimation problem must have that $\tau = \Omega\left(\frac{\sigma\sqrt{\log(1/\beta)}}{\epsilon\sqrt{n}}\right)$ and also that $n = \Omega\left(\frac{1}{\epsilon^2}\ln(\frac{R}{\beta \cdot \tau})\right)$.

It is worth-while to discuss the implications of Theorem 15. Aside from showing the near optimality of our technique, it also shows that our dependency on R is of the essence. This is in sharp contrast to the *centralized*-model, when the results of Karwa and Vadhan (2018) show that there exists a (ϵ, δ)-differentially private algorithm whose sample complexity is independent of R. Our lower bounds, which, as shown by Bun et al. (2018) are carried from the ϵ-LDP setting to the (ϵ, δ)-LDP, show that some dependency on $\log(R)$ is required. This illustrates a sharp contrast between the centralized and the local model.

In addition, we prove a similar bound on the optimality of the `BinQuant`-Algorithm.

Theorem 16. *Let $\mathcal{M}$ be a ϵ-LDP mechanism which is $(\alpha_{\text{dist}}, \alpha_{\text{quant}}, \beta)$-accurate for the p-quantile problem over $\mathcal{P}$, given that the true p-quantile lies in the interval $[-R, R]$. Then, for any $\beta < \frac{1}{6}$ it must hold that $n \ge \Omega(\frac{1}{\alpha_{\text{quant}}^2\epsilon^2} \cdot \ln(\frac{R}{\alpha_{\text{dist}}\beta}))$.*

It is important to note that our lower bound shows how *all* three parameters are necessary for devising a suitable ϵ-LDP algorithm for the problem. For example, we must have both stopping conditions (α_{quant} and α_{dist}). If we didn't specify α_{dist} as well, then we could devise a collection of infinitely many distributions — for any point $z \in [-R, R]$ we would construct a similar $\mathcal{P}_z$ similar to $\mathcal{P}_i$ — resulting in infinite sample complexity. Then for any m we could create a m-size collection of distributions by repeating the same collection with R set to be any number $> m/\alpha_{\text{dist}}$, thus we could get a sample complexity as arbitrary large as we want. Lastly, if α_{quant} was unspecified, we could derive an arbitrarily large sample complexity even without privacy as finding the exact quantile of a distribution requires infinitely many samples.

Acknowledgments

We gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting O.S. with grant #2017–06701; O.S. is also an unpaid collaborator on NSF grant #1565387.

References

Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In *ICML*, pages 169–178, 2018.

Apple Press Info. Apple previews ios 10, the biggest ios release ever, 2016. URL https://www.apple.com/pr/library/2016/06/13Apple-Previews-iOS-10-The-Biggest-iOS-Release-Ever.html.

Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 127–135, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746632. URL http://doi.acm.org/10.1145/2746539.2746632.

Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. In *NIPS*, pages 2285–2293, 2017.

Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *CRYPTO*, pages 451–468, 2008.

Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *PODS*, pages 435–447, 2018.

Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'IT: Private and sample efficient identity testing. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 635–644, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/cai17a.html.

K. Chaudhuri and D. Hsu. Convergence rates for differentially private statistical estimation. In *ICML*, 2012.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 429–438, 2013a. doi: 10.1109/FOCS.2013.53. URL https://doi.org/10.1109/FOCS.2013.53.

John C. Duchi, Martin J. Wainwright, and Michael I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1529–1537, 2013b. URL http://papers.nips.cc/paper/5013-local-privacy-and-minimax-bounds-sharp-rates-for-probability-estimation.

C. Dwork and J. Lei. Differential privacy and robust statistics. In *STOC*, 2009.

C. Dwork, W. Su, and L. Zhang. Private false discovery rate control. *CoRR*, abs/1511.03803, 2015.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, pages 486–503, 2006a. doi: 10.1007/11761679_29.

Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284. Springer, 2006b.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2957-6. doi: 10.1145/2660267.2660348. URL http://doi.acm.org/10.1145/2660267.2660348.

Vitaly Feldman. Dealing with range anxiety in mean estimation via statistical queries. In *ALT*, pages 629–640, 2017.

Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 1612–1621. JMLR.org, 2018.

Marco Gaboardi, Hyun Woo Lim, Ryan Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2111–2120. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045613.

Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: Z-test and tight confidence intervals. *CoRR*, abs/1810.08054, 2018. URL http://arxiv.org/abs/1810.08054.

R.V. Hogg, J.W. McKean, and A.T. Craig. *Introduction to Mathematical Statistics*. Pearson education international. 2005.

Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1079–1087, New York, NY, USA, 2013. ACM.

Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 44:1–44:9, 2018. doi: 10.4230/LIPIcs.ITCS.2018.44. URL `https://doi.org/10.4230/LIPIcs.ITCS.2018.44`.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 531–540, 2008. doi: 10.1109/FOCS.2008.27. URL `https://doi.org/10.1109/FOCS.2008.27`.

Daniel Kifer and Ryan Rogers. A New Class of Private Chi-Square Hypothesis Tests. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 991–1000, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL `http://proceedings.mlr.press/v54/rogers17a.html`.

Or Sheffet. Differentially private ordinary least squares. In *ICML*, 2017.

Or Sheffet. Locally private hypothesis testing. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 4612–4621. JMLR.org, 2018.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

Caroline Uhler, Aleksandra Slavkovic, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *ICDM*, pages 138–143, 2009.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. *arXiv preprint arXiv:1511.03376*, 2015.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60:63–69, 1965.

Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavkovic, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50: 133–141, 2014.