Four accuracy bounds and one estimator for frequency estimation under local differential privacy

Milan Lopuhaa-Zwakenberg*, Zitao Li†, Boris Škorić* and Ninghui Li†

*Department of Mathematics and Computer Science
Eindhoven University of Technology
m.a.lopuhaa@tue.nl, b.skoric@tue.nl
†Department of Computer Sciences
Purdue University
li2490@purdue.edu, ninghui@cs.purdue.edu

Abstract—We present four lower bounds on the mean squared error of both frequency and distribution estimation under Local Differential Privacy (LDP). Each of these bounds is stated in a different mathematical ‘language’, making them all interesting in their own right. Two are particularly important: First, for a given LDP protocol, we give formulas for the optimal frequency and distribution estimators for the aggregator. This theoretically ‘solves’ the issue of postprocessing, although it may be computationally infeasible in practice. We present methods to approximate the optimal estimator, yielding an estimator that outperforms state of the art methods. Second, we give a lower bound on MSE in terms of the LDP parameter $\varepsilon$, giving a quantitative statement of the privacy-utility tradeoff in LDP. In the high privacy domain, this improves on existing results by giving tighter constants even where earlier results consider a worst-case error metric, or giving constants where up to now, no constants were known.

Index Terms—Local Differential Privacy, frequency estimation, accuracy bound, privacy-utility tradeoff.

1. Introduction

In a context where a data aggregator collects potentially sensitive data, there is an inherent tension between the aggregator’s desire to obtain accurate population statistics and the individuals’ desire to protect their private data. One approach to protect privacy is offered by Local Differential Privacy (LDP) protocols [12]. Under this framework, each user randomises his private data before sending it to the aggregator. This hides the users’ true data, while for a large population size the randomness of the users cancels out, allowing the aggregator to obtain accurate estimates of the population statistics. Because of these properties, LDP-mechanisms are widely used in industry by companies such as Apple [1], Google [9], and Microsoft [6].

One of the main settings in which LDP protocols are used is that of frequency estimation [9], [16], [18]. In this setting, every user has a private data item from a (finite) set $\mathcal{A}$, and the aggregator’s goal is to determine the frequencies of the elements of $\mathcal{A}$ among the user population.

There is a tradeoff between user privacy and frequency estimation accuracy. Intuitively, the more ‘random’ the privacy protocol is, the better it will hide an individual’s private data, but the more noisy the aggregator’s estimations will be. Therefore, it is natural to ask the following question:

Question 1.1. Given an LDP protocol $Q$, what is the frequency estimation error for the aggregator?

The answer depends on a number of factors. The first of these is the metric used to assess estimation error; throughout this paper we consider the mean squared error (MSE), as this is an often-used metric in the literature [8], [16], [17]. Second, the answer will depend on the number of users: the more users, the more the random noise induced by $Q$ will ‘cancel out’, leading to better estimations. The total error over all data items naturally also depends on $a = \# \mathcal{A}$.

Furthermore, the estimation depends on the frequency estimator used by the aggregator. LDP protocols in the literature typically come with unbiased frequency estimators [9], [16], [18]. These estimators, however, often yield negative frequency estimations, or estimations that do not add up to 1. While one can improve the accuracy of the estimation by postprocessing these estimations, there are many ways to do so, and it is not clear what the optimal postprocessing method is for general $Q$ and input distribution [17]. Nevertheless, the most promising approaches depend crucially on prior expectations of properties of the input distribution. Input distributions typically follow a power law, and fitting a power law distribution to the data will ‘cancel out’, leading to better estimations. The total error over all data items naturally also depends on $a = \# \mathcal{A}$.

Setting the issue of estimator choice aside, the question remains what properties of $Q$ play a role in the estimation error. Of particular importance is the LDP privacy parameter $\varepsilon$, which is a positive real number measuring the worst-case leakage of the protocol. It is known that for a large number of users $n$, and small $\varepsilon$, the error is at best proportional to $\frac{\varepsilon}{\sqrt{n}}$ [3], [7], and protocols have been found that achieve this asymptotic rate [7], [16]. Nevertheless, depending on the precise method of estimation, the constant factor of the lower bound is either unknown or far from what has been achieved in practice.
1.1. Our contributions

In this paper, we consider two different, but related estimation problems; where the aggregator wants to estimate the probability distribution \( P \) from which the users’ private data is drawn, and where the aggregator wants to estimate the actual tallies \( T \) of the private data. The main contribution of this paper is to answer Question 1.1 by giving four lower bounds for the accuracy of estimators under LDP, for a given privacy protocol, for both of these problems. As we get down the list, we get lower bounds that are progressively less tight, but easier to compute. Each of these lower bounds is stated in a different mathematical ‘language’, making them all interesting in their own right. We have the following lower bounds:

**Optimal estimators:** We formalise the approach of incorporating prior information to improve estimator accuracy [10], [17]. This allows us to find, for a given protocol and a given prior distribution of \( P \), the distribution and tally estimation that minimises the average mean squared error (Theorem 3.1). Unfortunately, these can be computationally expensive in practice, but we show that they can be approximated by the MLE. We also assess the accuracy of the MLE estimator in practice, showing that it outperforms state of the art techniques [17].

**Information theory:** We give a lower bound of estimator accuracy in terms of \( \epsilon \) (Theorem 4.1). This relates information-theoretical bounds to what is currently achievable as average-case.

2. Preliminaries

2.1. Notation

For a finite set \( A = \{1, \ldots, a\} \), we write \( \mathcal{P}_A = \{p \in \mathbb{R}_{\geq 0} : \sum_{x} p_x = 1\} \) for the space of probability distributions on \( A \). This space is \((a-1)\)-dimensional, and a convenient way to parametrise it is to consider, for an \( x \in A \), the space

\[
\mathcal{P}^{(x)}_A = \left\{ p^{(x)}(x) \in [\mathbb{R}_{\geq 0}]^{|A|} : \sum_{x' \neq x} p_{x'} \leq 1 \right\},
\]

which is diffeomorphic to \( \mathcal{P}_A \) via the map \( p \mapsto p^{(x)} \). A random variable \( P \in \mathcal{P}_A \) is called continuous if there is a continuous function \( f(x) : \mathcal{P}^{(x)}_A \rightarrow \mathbb{R}_{\geq 0} \) such that for each measurable subset \( U \subset \mathcal{P}_A \) one has

\[
\mathbb{P}(P \in U) = \int_{p^{(x)} \in U(x)} f^{(x)}(p^{(x)}) dp^{(x)},
\]

where \( U(x) = \{p^{(x)} : p \in U\} \). If this is the case, then the differential entropy of \( P \) is defined as

\[
h(P) := \int_{p^{(x)} \in \mathcal{P}^{(x)}_A} -f^{(x)}(p^{(x)}) \log f^{(x)}(p^{(x)}) dp^{(x)}. \]

The definitions of continuity and of \( h(P) \) do not depend on the choice of \( x \in A \).

2.2. Privacy protocols

Let \( A \) be a finite set. A privacy protocol for \( A \) is a pair \( Q = (Q, B) \) where \( B \) is a finite set, and \( Q = (Q_{y|x})_{y \in B, x \in A} \in \mathbb{R}_{\geq 0}^{B \times A} \) is a collection of nonnegative reals such that \( \sum_{y} Q_{y|x} = 1 \) for all \( x \). We will often identify \( A = \{1, \ldots, a\} \) and \( B = \{1, \ldots, b\} \), and consider \( Q \) as a \((b \times a)\)-matrix. We consider \( Q \) as a random function: for \( x \in A \), let \( Q(x) \in B \) be the random variable with probability distribution \( Q_{y|x} \), i.e.,

\[
Q_{y|x} \overset{\text{d}}{=} \mathbb{P}(Q(x) = y).
\]

The standard way to measure the privacy of a privacy protocol is via \( \epsilon \)-Local Differential Privacy (LDP):

**Definition 2.1.** (\( \epsilon \)-LDP [12]) Let \( Q = (Q, B) \) be a privacy protocol for \( A \). We say that \( Q \) satisfies \( \epsilon \)-LDP if for all \( x, x' \in A \) and all \( y \in B \) one has

\[
Q_{y|x'} \leq e^{\epsilon} Q_{y|x}.
\]

Intuitively, this means that for small \( \epsilon \), given the output \( y \), it is difficult to decide whether the input was \( x \) or \( x' \). The smaller \( \epsilon \), the more privacy the protocol offers. Throughout, we make two assumptions on our privacy protocols:

1) We assume that \( \text{rk}(Q) = a \). We do this because privacy protocols of lower rank are unable to distinguish all possible input distributions, which makes them unsuitable for frequency estimation [13].

2) We assume that for all \( x \in A \), \( y \in B \) one has \( Q_{y|x} > 0 \). If there is an \( y \) such that \( Q_{y|x} = 0 \) for all \( x \), then we can remove \( y \) from \( B \) without changing the protocol. If there is an \( y \) such that \( Q_{y|x} = 0 \) but \( Q_{y|x'} > 0 \), then \( Q \) does not satisfy \( \epsilon \)-LDP for any \( \epsilon \). Since we are only interested in

| known WC bound | Distribution | Tally |
|----------------|-------------|-------|
| attainable MSE | \( \frac{\Omega(\frac{1}{\epsilon^2})}{n} \) | \( \Omega(\frac{1}{\epsilon^2}) \) [8] |
| MSE bound (us) | \( \frac{\Omega(\frac{1}{\epsilon^2})}{n} \) | \( \approx \frac{\Omega(\frac{1}{\epsilon^2})}{n} \) [3] |

**TABLE 1:** The average-case lower bounds obtained in this paper (for \( n \gg 0 \) and \( \epsilon \ll 1 \)) compared to previous worst-case lower bounds and to what is currently achievable as average-case.
privacy protocols that offer privacy, disregarding this case is harmless.

The main reason for these assumptions is that they simplify the expressions in Theorem 5.1 compared to [13], while still being general enough to describe all protocols that are used in practice.

2.3. Setting

We consider the setting of Local Differential Privacy. There are \( n \) users, and user \( i \) has a private data item \( X_i \in A = \{1, \ldots, a\} \). The aggregator publishes a privacy protocol \( Q = (Q, B) \), where we take \( B = \{1, \ldots, b\} \) for convenience. User \( i \) calculates \( Y_i := Q(X_i) \) and sends \( Y_i \) to the aggregator. We write \( \bar{Y} = (Y_i)_{i=1}^n \); this is the data the aggregator has at its disposal. Note that the LDP setting differs from DP in that the users do not give the aggregator their bare private data. We consider \( X_1, \ldots, X_n \) to be drawn independently from a probability distribution \( P \in \mathcal{P}_A \). The probability distribution \( P \) is unknown to the aggregator as well: hence we consider it to be a random variable. The distribution of \( P \) is given by a probability measure \( \mu \) on \( \mathcal{P}_A \); this distribution reflects the prior knowledge of the aggregator. Throughout this paper, we assume that \( P \) is continuous. If the aggregator does not have any prior knowledge, a good choice for \( \mu \) is the Jeffreys prior, i.e. the symmetric Dirichlet distribution with parameter \( \frac{1}{2} \) [13].

We consider two possible goals of the aggregator:

1) The aggregator wants to know \( P \) as accurately as possible. This occurs, for instance, when the aggregator is a scientist whose userbase is a sample of a greater population. The aggregator is not concerned with this specific userbase, but rather with the characteristics of the general population, which are modelled by \( P \).

2) For \( x \in A^n \) and \( \gamma \in A \), let \( t_\gamma(x) = \#\{i : x_i = \gamma\} \) and \( t(x) = (t_\gamma(x))_{\gamma \in A} \). Then the aggregator wants to know the tallies \( T := t(\bar{X}) \) as accurately as possible. This occurs when the users are customers of a service, and the service provider wants to know statistics about its customer base.

In this paper, we assess the ability of the aggregator to estimate \( P \) and \( T \) from \( \bar{Y} \). Specifically, the aggregator has an estimator function \( \Psi_n : B^n \to \mathbb{R}^A \) that estimates \( P \) from \( \bar{Y} \), and a \( \Phi_n : B^n \to \mathbb{R}^A \) that estimates \( T \) from \( \bar{Y} \).

We use the mean squared error to measure an estimator's performance:

**Definition 2.2.** We define the mean squared error of \((Q, \Psi_n)\) and \((Q, \Phi_n)\) to be

\[
\text{MSE}_{\mu}(Q, \Psi_n) := \mathbb{E}_{P, \bar{Y}} \left[ ||P - \Psi_n(\bar{Y})||^2 \right],
\]

\[
\text{MSE}_{\mu}(Q, \Phi_n) := \mathbb{E}_{T, \bar{Y}} \left[ ||T - \Phi_n(\bar{Y})||^2 \right].
\]

Note that the dependence on \( \mu \) is implicit, as \( \mu \) is the probability distribution of \( P \) (and hence affects the probability distribution of \( T \)). The metric means that the best estimator is the one that gives on average the lowest squared error, when averaging over all possible input distributions, and all possible outputs. As an alternative, one can consider the following worst-case metrics [3], [8]:

**Definition 2.3.** We define the worst-case MSE of \((Q, \Psi_n)\) and \((Q, \Phi_n)\) to be

\[
\text{WSE}_{\mu}(Q, \Psi_n) := \sup_P \mathbb{E}_P \left[ ||P - \Psi_n(\bar{Y})||^2 \right],
\]

\[
\text{WSE}_{\mu}(Q, \Phi_n) := \sup_T \mathbb{E}_T \left[ ||T - \Phi_n(\bar{Y})||^2 \right].
\]

The ‘WC bound’ column in Table 1 refers to these metrics. Note that

\[
\text{WSE}_{\mu}(Q, \Psi_n) = \max_{\mu} \text{WSE}_{\mu}(Q, \Psi_n),
\]

\[
\text{WSE}_{\mu}(Q, \Phi_n) \geq \max_{\mu} \text{WSE}_{\mu}(Q, \Phi_n).
\]

The notation employed in this paper is summarised in Table 2.

| \(A\) | \(a\) | \(n\) | \(\bar{X}\) | \(T\) |
|---|---|---|---|---|
| input space | # of users | number of users | vector of obfuscated data | tallies of private data |

| \(P_A\) | \(P\) | \(\mu\) | \(Q\) | \(B\) | \(\bar{Y}\) | \(S\) | \(\Psi_n\) | \(\Phi_n\) |
|---|---|---|---|---|---|---|---|---|
| space of prop. distr. on \(A\) | prob. distr. of private data | privacy protocol | matrix associated to \(Q\) | output space associated to \(Q\) | vector of obfuscated data | tallies of obfuscated data | estimator of \(P\) given \(\bar{Y}\) | estimator of \(T\) given \(\bar{Y}\) |

**Table 2: Notation employed in this paper.**

### 2.4. Multivariate de Moivre–Laplace

Let \( r, n \in \mathbb{Z}_{\geq 1} \), \( R := \{1, \ldots, r\} \), \( w \in \mathcal{P}_R \), and let \( L \) be drawn from a multinomial distribution of \( n \) samples with probability vector \( w \); then \( M := n^{-1}L \) is an element of \( \mathcal{P}(\{1, \ldots, r\}) \). The multivariate analog of the de Moivre–Laplace theorem tells us that for large \( n \), \( M \) can be approximated by a multivariate normal distribution. We introduce two bits of notation:

1) For \( j \leq r \), we write \( D(w)_{(j)} \in \mathbb{R}(R \setminus j)^2 \) for the matrix given by \( D(w)_{(j)} = \text{diag}(w_{(j)}) - w_{(j)} w_{(j)}^T \). Note that \( D(w)_{(j)} \) is positive definite and that \( \text{det}(D(w)_{(j)}) = \prod_{j' \neq j} w_{j'} \).

2) For \( m_{(j)} \in \mathbb{R}(R \setminus j)^2 \) and \( C_{(j)} \in \mathbb{R}(R \setminus j)^2 \) positive definite, we define \( N(m_{(j)}, C_{(j)}) \) for a multivariate normal distribution on \( \mathbb{R}(R \setminus j) \) with mean \( m_{(j)} \) and covariance matrix \( C_{(j)} \).

**Theorem 2.4.** (Multivariate de Moivre–Laplace [15]) For large \( n \) and for any \( j \), \( M_{(j)} \) converges in probability to the discretisation of \( N(m_{(j)}, n^{-1}D(w)_{(j)}) \) around the lattice \( (n^{-1/2}) \mathbb{Z}(R \setminus j) \).

In our setting, the de Moivre–Laplace theorem surfaces when describing the distribution of \( S \). Although \( S \) itself generally cannot be approximated by a normal
distribution, we can find such an approximation for $S$ given $T$.

For a $p \in \mathcal{P}_A$, we define

$$ G_p^{(y)} := \sum_{x} p_x D_p^{(y)} . \quad (11) $$

**Proposition 2.5.** Let $p \in \mathcal{P}_A$, and let $t \in \mathbb{Z}_{\geq 0}$ such that $\sum t_x = n$. Let $d = n^{-1} (s - \hat{Q} t) \in \mathbb{R}^b$. Then for large $n$, $\mathbb{P}(S = s, T = t)$ is approximately equal to

$$ \sqrt{\frac{(2\pi n)^{1-b}}{\det(G_p^{(y)})}} \exp \left( -\frac{n}{2} d^T G_p^{(y)-1} d \right) . \quad (12) $$

**Proof.** Write $M = n^{-1} S \in \mathcal{P}_B$, and for $x \in A$, define $E_x^{(y)} := D_p^{(y)}$. We can write $S = \sum_x S_x$, where for a given $T = t$, the random variable $S_x \in \mathbb{Z}_{\geq 0}$ follows a multinomial distribution with $t_x$ samples and probability vector $\hat{Q}_x^{n,t}$. By Theorem 2.4, we know that for large $t_x$ the random variable $M_x^{(y)} = (n^{-1} S_x^{(y)})$ approximates a discretisation of a normal distribution $N(\mu_x^{(y)}; \frac{t_x}{n} E_x^{(y)})$; hence $M^{(y)}$ approximates a discretisation of a normal distribution with mean $(n^{-1} \hat{Q} t)^{y}$ and covariance matrix $\sum_x \frac{t_x}{n} E_x^{(y)} = n^{-1} G_p^{(y)}$, around the lattice $(n^{-1} \mathbb{Z})^{b-1}$. Substituting this in the pdf of a multivariate normal distribution now proves the statement. \qed

### 3. Optimal estimators

We obtain our first upper bound for the MSE by giving formulas for the optimal estimators, given $Q, n, \mu$. Essentially, this is a well-known fact about minimum mean square error (MMSE) estimators.

**Theorem 3.1.** Let $Q, n, \mu$ be given. The $\hat{\Psi}_n$ and $\hat{\Phi}_n$ minimising $(5)$ and $(6)$, respectively, are given by

$$ \hat{\Psi}_n(y) = \mathbb{E}[P^{|T = y}] = \tilde{g}_l, \quad (13) $$

$$ \hat{\Phi}_n(y) = \mathbb{E}[T^{|\tilde{Y} = y}] = y'. \quad (14) $$

**Proof.** For any estimator $\Psi_n$ of $P$ one has

$$ \text{MSE}_{n,\mu}(Q, \Psi_n) = \text{E}_y \text{E}_\mathbb{P} \left[ |P - \Psi_n(y)|^2 \right] \frac{\tilde{y}}{\tilde{y}} \quad (15) $$

$$ \geq \text{E}_y \text{E}_\mathbb{P} \sum_{x=1}^a \text{Var}(P_x^{|\tilde{Y} = y}) \quad (17) $$

One has equality in $(17)$ if and only if $\Psi_n(y) = \text{E}_y \text{E}_P[P_x^{|\tilde{Y} = y}] = \tilde{g}_l$ for all $x$ and $\tilde{y}$. The proof for $\Phi_n$ is similar. \qed

Theorem 3.1 formalises the intuition in [10, 17] that one gets better estimators by incorporating prior knowledge of the distribution. While Theorem 3.1 gives us a direct formula for the optimal estimators for a given privacy protocol, the disadvantage is that these can be computationally difficult to evaluate. With regards to $\Psi_n$, for $x' \in A$ we find that for any $x$ we have

$$ \hat{\Psi}_n(y)_{x'} = \frac{1}{P(Y = y)} \int_{p \in \mathcal{P}} f(p(x')) p_x \prod_{y \in B} (\hat{Q}p)^{s_y} dp . \quad (18) $$

For large $a$, the integral over $\mathcal{P}_A$ can be computationally involved. One approach is to do a Monte Carlo approximation of the integral. Typically, $f$ will be a Dirichlet distribution, for which several efficient Monte Carlo methods exist [2, 4]. Furthermore, if the aggregator is interested in estimating $P$, we can give an expression for the optimal $\hat{\Psi}_n(y)$, as well as for its MSE, that does not involve any integration:

**Proposition 3.2.** Let $\hat{\Psi}_n$ be as in Theorem 3.1. Let $\mu$ be a Dirichlet distribution with parameter vector $\alpha \in \mathbb{R}^a$. For $\gamma \in A, \tilde{\gamma} \in A^n$ and $\tilde{y} \in B^n$, we define

$$ C_{g,\gamma} = \sum_{x \in A} \frac{B(\alpha + t(\tilde{y}))}{B(\alpha)} \hat{Q}_y[x_i], \quad (19) $$

$$ w_{\tilde{y}} g,\gamma = C_{g,\gamma}^{-1} \frac{B(\alpha + t(\tilde{y}))}{B(\alpha)} \hat{Q}_y[x_i], \quad (20) $$

$$ \mu_{x,\gamma} = \alpha_x + t_x(\tilde{y}), \quad (21) $$

$$ \mu_{\tilde{y},\gamma} = \sum_{\tilde{y} \in A} w_{\tilde{y}} g,\gamma, \quad (22) $$

$$ \sigma_{\tilde{y},\gamma}^2 = \mu_{x,\gamma}(1 - \mu_{x,\gamma}), \quad (23) $$

Then $\hat{\Psi}_n(y)_{\gamma} = \mu_{\tilde{y},\gamma}$, and $\text{MSE}_{n,\mu}(Q, \hat{\Psi}_n)$ is equal to

$$ \sum_{\tilde{y} \in B^n} \sum_{\gamma \in A} C_{g,\gamma} \left( \sum_{\tilde{y} \in A} w_{\tilde{y}} g,\gamma (\hat{\Psi}_n(y)_{\gamma} - \mu_{\tilde{y},\gamma}) - \frac{\mu_{\tilde{y},\gamma}}{2} \right). \quad (24) $$

**Proof.** By the proof of Theorem 9.1 of [13] we know that $P|\tilde{Y} = y$ is the mixture of Dirichlet distributions. These are parametrised by $x \in A^n$, have weight $w_{\tilde{y}} g,\gamma$, and parameter vector $\alpha + t(\tilde{y})$. The formulas for $\hat{\Psi}_n(y)_{\gamma}$ and $\text{MSE}_{n,\mu}(Q, \hat{\Psi}_n)$ now follow from standard formulas for the mean and variance of both Dirichlet distributions and mixture distributions. \qed

Although in the expressions of Proposition 3.2 we sum over domains whose size is exponential in $n$, following [13] we see that we only need to calculate a polynomial number of terms, $O(n^{b(a-1)})$ to be precise.

### 3.1. Approximation by MLE

For large $n$, the calculation in Proposition 3.2 will still be computationally involved. However, as $n \to \infty$, $\hat{\Psi}_n(y)$ and $\hat{\Phi}_n(y)$ converge to the maximum likelihood estimators of $P$ and $T$ given $Y = y$ [14]. For $p$ we can write this as follows. Since $P(Y = y) = p^T \prod_{y \in B} (\hat{Q}p)^{s_y}$, we find the MLE by solving the optimisation problem

$$ \text{maximise}_p \sum_{y \in B} s_y \log((\hat{Q}p)^y) \quad (25) $$

subject to $p \in \mathcal{P}_A$. 

---

1. There is also such a description for $S$ given $P$, but we will not need this.
Since the objective function is smooth and concave in \( p \), this can be solved quickly numerically. Using the MLE rather than the posterior also has the advantage that it is independent of the choice of \( \mu \), making it a good choice of estimator when the prior is unknown.

Unfortunately, the expression for \( \mathbb{P}(S = s|T = t) \) is more complicated. However, for large \( n \), we can approximate it via Proposition 2.5.2. As \( n \) grows larger, the \( t \)-dependence of \( \frac{(2\pi n)^{1/2}}{\det(G_{n, t}^{-1})} \) will become negligible compared to that of \( \exp \left( -\frac{1}{2}G(\hat{s} - \hat{q}t)^T G^{-1}(\hat{s} - \hat{q}t) \right) \). Therefore, we can approximate the MLE by finding the \( t \) that solves the following optimisation problem (after choosing a \( y \in B \)):

\[
\begin{align*}
\text{minimise} & \quad (s - \hat{q}t)^T G^{-1}(s - \hat{q}t) \\
\text{subject to} & \quad t \in \mathbb{Z}^b_{\geq 0}, \\
& \quad \sum_x t_x = n.
\end{align*}
\]

(26)

Any choice of \( y \) gives the same objective function. Unfortunately, this optimisation problem is quite a bit more complicated than the one for \( P \). However, the advantage of this approximation is that the \( n \)-dependence only appears in the discrete solution of \( n^{-1} t \in \mathcal{P}_A \). Furthermore, for large \( n \) we have \( nP \approx T \), so we can use the optimisation problem for \( P \) to get an approximate value for \( T \). Note that our approach to approximating the MLE is different from that of [17], as there the MLE is approximated based on a non-injective transformation of the obfuscated tallies \( S \), rather than on \( S \) itself. This transformation is protocol-specific, and hence this method does not easily extend to general protocols. The advantage of (25) is that it can be used to improve the estimation of any protocol.

3.2. MLE experiments

We perform synthetic experiments to see how well the MLE of (25) performs in practice. We apply the MLE to two well-established privacy protocols, GRR [18] and OUE [16]. Both of these are parametrised by the LDP parameter \( \varepsilon \). We take \( n = 10^4 \) and \( \varepsilon \in [0.2, 2] \). Since \( b = 2^a \) for OUE, the number of summands in (25) grows too large to handle for large \( a \). Therefore we take \( a = 10 \) for OUE. For GRR we do not have this problem, and we take the more general setting \( a = 1024 \). We draw \( P \) from the Jeffrey’s prior on \( \mathcal{P}_A \) 100 times and generate a dataset of \( n \) users from it. We then randomise the data via the LDP protocol, and perform MLE on the outcome to produce an estimate \( \hat{p} \) of \( P \); we furthermore use \( \ell = n\hat{p} \) as an estimator of \( T \). Finally, we measure \( ||P - \hat{p}||^2 \) and \( ||T - \ell||^2 \), and average this over all samples to determine the MSE. To calculate the MLE, we use projected gradient descent to solve (25) for OUE. For GRR, there is a direct method to find \( \hat{p} \), see Appendix A.

We compare the MLE estimator to two other estimators: First, the baseline Frequency Oracles (FO), which are affine transformations of \( S \) used to produce unbiased estimators of \( P \) and \( T \). Second, we look at Norm-Sub, which was found in [17] to be postprocessing method of the FO outcome that gives the best MSE, among a wide selection of considered postprocessing methods.

Figures 1 (GRR) and 2 (UE) show the experimental results. The simulation results show that both Norm-sub and MLE post-processing can elevate the accuracy of the results. For GRR, we see that Norm-Sub and MLE give the same accuracy. This is not unexpected, as the results in [17] show that for frequency estimation, Norm-Sub gives similar accuracy to MLE, and for GRR the MLEs defined here and in that paper are equivalent. For OUE, we see that MLE gives more accurate frequency estimations than Norm-sub when \( \varepsilon > 1.5 \).

4. Accuracy bounds from information theory

The disadvantage of the optimal estimators of the previous section is that their MSEs can be hard to quantify. In this section, we give lower bounds for these, based on the information-theoretical quantities \( h(P|\bar{Y}) \) and \( h(T|\bar{Y}) \). Intuitively, the lower the uncertainty about the value of \( P \) or \( T \), the lower the average error on the estimation should be. The following Theorem quantifies this intuition.

Theorem 4.1. For any estimators \( \Psi_n \) of \( P \) and \( \Phi_n \) of \( T \) one has

\[
\begin{align*}
\text{MSE}_{n,\mu}(Q, \Psi_n) & \geq \frac{a}{2\pi e} e^{\frac{1}{2\pi e}} h(\bar{P}|\bar{Y}), \\
\text{MSE}_{n,\mu}(Q, \Phi_n) & \geq \frac{a}{2\pi e} e^{\frac{1}{2\pi e}} h(\bar{T}|\bar{Y}).
\end{align*}
\]

Proof. Again, we prove this only for \( \Psi_n \), as the proof for \( \Phi_n \) is similar. By (17) we have

\[
\text{MSE}_{n,\mu}(Q, \Psi_n) \geq E_{\bar{Y}} \sum_{x=1}^n \text{Var}(P_x|\bar{Y} = \bar{y}) \geq \frac{1}{n-1} E_{\bar{Y}} \sum_{x \leq a, x' \neq x} \text{Var}(P_x|\bar{Y} = \bar{y}) = \frac{1}{n-1} \sum_{x=1}^n \text{Var}(P_x|\bar{Y} = \bar{y}) \geq \frac{1}{n-1} \prod_{x \neq y} \text{Var}(P_x|\bar{Y} = \bar{y}) \geq \text{det} \text{Cov}(P(x)|\bar{Y} = \bar{y}).
\]

The \{Var(P_x|\bar{Y} = \bar{y})\}_{x \neq y} are the diagonal coefficients of the positive definite matrix \( \text{Cov}(P(x)|\bar{Y} = \bar{y}) \). By Hadamard’s inequality we have

\[
\prod_{x \neq y} \text{Var}(P_x|\bar{Y} = \bar{y}) \geq \text{det} \text{Cov}(P(x)|\bar{Y} = \bar{y}).
\]

Furthermore, Theorem 9.6.5 of [5] shows that for all \( x \) we have

\[
\text{det} \text{Cov}(P(x)|\bar{Y} = \bar{y}) \geq (2\pi e)^{a-1} e^{2\pi e} h(\bar{P}|\bar{Y} = \bar{y}).
\]

The convexity of the exponential function tells us that \( E_{\bar{Y}} e^{\frac{1}{2\pi e}} h(\bar{P}|\bar{Y} = \bar{y}) \geq e^{\frac{1}{2\pi e}} h(\bar{P}|\bar{Y}) \). Combining this with (31), (32), (33) now proves the Theorem.

5. Accuracy bounds from linear algebra

In the previous section, we gave a lower bound for the MSE in terms of the information-theoretical concepts \( h(P|\bar{Y}) \) and \( h(T|\bar{Y}) \). We can obtain new lower bounds for the limit case by studying the behaviour of \( h(P|\bar{Y}) \) and \( h(T|\bar{Y}) \) as \( n \to \infty \); this expands on work in [13]. The resulting lower bound is weaker not in the sense that the
bound is lower, but rather that it only applies to the limit case $n \to \infty$, rather than all $n$. However, the advantage is that this limit case can be formulated purely in terms of linear algebra, and as it does not depend on $n$, it is computationally more feasible for large amounts of users, which is typical in the LDP setting.

Before we can state the result, we first need a bit more notation. Choose a $y \in \mathcal{B}$; for $p \in \mathcal{P}_A$ we define

$$
F_p := \tilde{Q}^T \text{diag}(\tilde{Q} p)^{-1} \tilde{Q} \in \mathbb{R}^{a \times a}.
$$

and constants $\gamma_p(Q), \delta_p(Q)$ by

$$
\gamma_p(Q) = -2^{-1} \log(2\pi e) - \frac{1}{2} \log \det F_p,
$$

$$
\delta_p(Q) = \gamma_p(Q) + \frac{1}{2} \log \frac{\det G_p^{(y)}}{\prod_{y' = 1}^b (Q P)'_{y'}}.\tag{36}
$$

While the matrix $G_p^{(y)}$ depends on the choice of the ‘distinguished’ element $y \in \mathcal{B}$, the resulting constant $\delta_p(Q)$ does not. The introduction of these constants allows us to formulate the following Theorem:

**Theorem 5.1.** One has

$$
\lim_{n \to \infty} h(P|\tilde{Y}) + \frac{a-1}{2} \log n = \gamma_p(Q),
$$

$$
\lim_{n \to \infty} H(T|\tilde{Y}) - \frac{a-1}{2} \log n = \delta_p(Q).\tag{38}
$$

**Proof.** Since $h(P|\tilde{Y}) = h(P) - I(\tilde{Y}; P)$, this follows directly from Theorem 6.7.1 of [13]. Similarly, we have $H(T|\tilde{Y}) = H(T) - I(\tilde{Y}; T) = I(S; T)$, where $S \in \mathcal{Z}_{\geq 0}$ is the tally vector of $\tilde{Y}$. Start by describing the limit behaviour of $I(S; T)$. The limit behaviour of $H(S)$ is given in [13, Lems. C.4 & C.7], so it suffices to study $H(S|T)$. Write $M = n^{-1} S \in \mathcal{P}_B$. By the proof of Proposition 2.5 we know that given $T = t$, $M^{(b)}$ is approximated by a discretisation of a multivariate normal distribution with mean $n^{-1}(Q(t))^{(b)}$ and covariance $n^{-1} G_n^{(b)}(t)$, around the lattice $(n^{-1} \mathbb{Z})^{b-1}$. Since $n^{-1} T$ approximates a discretisation of $P$ for large $n$, this implies

$$
H(S|T) = \mathbb{E}_t[H(S|T = t)]
$$

$$
\approx \frac{b-1}{2} \log(2\pi e n^2) + \frac{1}{2} \log \det(G_n^{(b)}(t))\tag{40}
$$

$$
\approx \frac{b-1}{2} \log(2\pi e n) + \frac{1}{2} \log \det(G_n^{(b)}).\tag{41}
$$
Combining this with Lemmas C.4 and C.7 of [13] now finishes the proof.

It follows from this Theorem that we have the following limit behaviour of our estimator accuracy metrics.

**Corollary 5.2.** Let \( \Psi_n \) and \( \Phi_n \) be as in Theorem 4.1. Then

\[
\lim_{n \to \infty} n \text{MSE}_{n, \mu}(Q, \Psi_n) \geq \frac{a}{2\pi e} e^{\frac{1}{2} - \sqrt{\frac{2}{\pi} e}} \gamma_{\mu}(Q),
\]

(42)

\[
\lim_{n \to \infty} n^{-1} \text{MSE}_{n, \mu}(Q, \Phi_n) \geq \frac{a}{2\pi e} e^{\frac{1}{2} - \sqrt{\frac{2}{\pi} e}} \delta_{\mu}(Q).
\]

(43)

We should interpret this as stating that at best we have MSE\(_{n, \mu}\) for \( Q, \Psi_n \) = \( \Omega(n^{-1}) \), MSE\(_{n, \mu}\) for \( Q, \Phi_n \) = \( \Omega(n) \) (for fixed \( Q \) and \( \mu \)), and we can bound the constants involved.

**6. Accuracy bounds from \( \varepsilon \)-LDP**

While the constants \( \gamma_{\mu}(Q) \) and \( \delta_{\mu}(Q) \) do not depend on \( n \), they still involve integration over \( \mathcal{P}_A \), and as such can be computationally difficult for large \( a \). However, it is possible to give lower bounds for these constants that are independent of \( \mu \), and whose \( Q \)-dependence only appears in the privacy parameter \( \varepsilon \).

**Theorem 6.1.** One has

\[
\gamma_{\mu}(Q) \geq (a - 1) \log \frac{\sqrt{2\pi e}}{e^{\frac{1}{2}} - 1},
\]

(44)

\[
\delta_{\mu}(Q) \geq (a - 1) \log \frac{\sqrt{2\pi e}}{e^{\frac{1}{2}} - 1} - b e.
\]

(45)

**Proof.** The first statement follows from Proposition 8.1 of [13]. As for the second statement, let \( E_y^{(y)} \) be as in the proof of Proposition 2.5. Note that for every \( x, y \) and \( p \) we have \( \det E_y^{(y)} = \prod_{i=1}^{y_1} Q_{y|x} \geq e^{-b e} \prod_{i=1}^{y_1} (Q_{y|x})_{y_i} \). Furthermore, since \( \log \det \) is concave on the space of positive symmetric matrices, we have \( \log \det G^{(b)} \geq \sum_x p_x \log \det E_x^{(y)} \) for every \( p \). This proves the Theorem.

As a corollary of this Theorem, we find a lower bound for the MSE in terms of \( \varepsilon \). We write \( f(\varepsilon) \geq g(\varepsilon) \) if \( \lim_{\varepsilon \to 0} \frac{g(\varepsilon)}{f(\varepsilon)} \geq 1 \).

**Corollary 6.2.** Let \( \Psi_n \) and \( \Phi_n \) be as in Theorem 4.1. Then for every \( \mu \) one has

\[
\lim_{n \to \infty} n \text{MSE}_{n, \mu}(Q, \Psi_n) \geq \frac{a}{(e^{\frac{1}{2}} - 1)^2},
\]

(46)

\[
\lim_{n \to \infty} n^{-1} \text{MSE}_{n, \mu}(Q, \Phi_n) \geq \frac{a}{e^{\frac{1}{2}} - 1}. \quad (47)
\]

For \( n \in \mathbb{Z}_{\geq 0} \) and \( \varepsilon \in \mathbb{R}_{>0} \), let \((Q_n, \varepsilon, \Psi_n, \varepsilon)\) be the pair of an \( \varepsilon \)-LDP privacy protocol and an estimator for \( P \) minimising 5, and let \((Q_n, \varepsilon, \Phi_n, \varepsilon)\) be the pair of an \( \varepsilon \)-LDP privacy protocol and an estimator for \( T \) minimising 6. Then as \( \varepsilon \to 0 \), we have

\[
\lim_{n \to \infty} n \text{MSE}_{n, \mu}(Q_n, \varepsilon, \Psi_n, \varepsilon) \geq \frac{a}{\varepsilon},
\]

(48)

\[
\lim_{n \to \infty} n^{-1} \text{MSE}_{n, \mu}(Q_n, \varepsilon, \Phi_n, \varepsilon) \geq \frac{a}{\varepsilon} \quad (49).
\]

Note that these results are strictly better than what is known in the literature, to the best of our knowledge: (46) improves the result in the proof of Proposition 6 of [8] in three ways: First, our result does not just give a bound for the WSE, but also for the MSE. Also, we improve the lower bound by a factor \( 64 \). Furthermore, the result of [8] only holds for \( \varepsilon \leq 1 \). The downside of our result, however, is that it only holds for the limit case \( n \to \infty \).

As for the results for \( \Phi_n \), it follows from results in [3] that the optimal \((Q, \Phi_n, \varepsilon)\) satisfies \( \text{WSE}_{n, \mu}(Q, \Phi_n, \varepsilon) = \Omega(\frac{1}{e^{\frac{1}{2}}} \varepsilon) \) for small \( \varepsilon \), but the authors do not make a statement about the constants involved. (47) improves upon this by giving a quantitative lower bound for the MSE, which itself is a lower bound for the WSE. Also, the OLH protocol from [16] performs as \( \approx \frac{a}{\varepsilon^2} \) for large \( n \) and small \( \varepsilon \), while our lower bound is of the form \( \frac{a}{\varepsilon^2} \). Therefore, our bound is quite near to what is possible in practice.

The bound in (47) seems to imply that when looking for optimal \( Q \), we should take \( b = a \) as large as possible. This seemingly contradicts results in [11], where it is found that taking \( b = a \) is always sufficient. There are two possible explanations for this. First, it is possible that our bounds are not sharp enough to accurately detect the dependence on \( b \); this is especially probable since (47) is only the latest in a chain of inequalities. However, the discrepancy between our results and those of [11] can also be caused by the fact that different utility metrics were being used. Whereas we focus on asymptotically many users, the utility metric in [11] looks at the KL-divergence between the probability distributions induced by the private datum of one user. It is a possibility that the optimal protocols for these different metrics do not coincide. Overall, the \( b \)-dependence of the estimation error is difficult to assess. On one hand, the "typical" LDP protocol with \( b = a \), GRR [18], performs poorly for large \( a \). The methods that are shown to perform better typically have \( b \) growing exponentially in terms of \( a \) [16], which suggests a large \( b \) is optimal. On the other hand, we can always increase \( b \) while maintaining the same privacy and utility [13], so a larger \( b \) does not automatically mean a better protocol.

**6.1. Tightness for \( a = 2 \)**

In this section, we show that if \( a = 2 \) and as \( \varepsilon \to 0 \), the bounds in Corollary 6.2 are tight. For this, we recall the Randomised Response protocol [18], which, for an \( \varepsilon > 0 \), is the LDP protocol \( R_{\varepsilon}: \{1, 2\} \to \{1, 2\} \) given by the matrix

\[
\begin{pmatrix}
\frac{\varepsilon}{\varepsilon + 1} & \frac{1}{\varepsilon + 1} \\
\frac{1}{\varepsilon + 1} & \frac{\varepsilon}{\varepsilon + 1}
\end{pmatrix}.
\]

(50)

Note that \( R_{\varepsilon} \) satisfies \( \varepsilon \)-LDP. As estimators for \( P \) and \( T \), respectively, we use the maps \( F_n, G_n: \{1, 2\}^n \to \mathbb{R}^2 \) given by

\[
F_n(y) = \left( \frac{(e^{\frac{1}{2}} + 1)s_1 - 1}{(e^{\frac{1}{2}} - 1)n}, \frac{(e^{\frac{1}{2}} + 1)s_2 - 1}{(e^{\frac{1}{2}} - 1)n} \right),
\]

(51)

\[
G_n(y) = n F_n(y).
\]

(52)

**Proposition 6.3.** For any prior \( \mu \) one has

\[
\text{MSE}_{n, \mu}(R_{\varepsilon}, F_n) = \frac{a}{n} \left( \frac{\varepsilon^{\frac{1}{2}} + 1}{e^{\frac{1}{2}} - 1} + E_P[P_1 P_2] \right),
\]

(53)

\[
\text{MSE}_{n, \mu}(R_{\varepsilon}, G_n) = \frac{2 \mu e}{(e^{\frac{1}{2}} - 1)^2}.
\]

(54)
Since $G_n$ is an unbiased estimator of $P$, we have
\[ \text{MSE}_{n,\mu}(\hat{F}_n, F_n) = \mathbb{E}_p \text{Var}(F_n(\tilde{Y})_1 | P = p) \]
+ \[ \mathbb{E}_p \text{Var}(F_n(\tilde{Y})_2 | P = p). \]
Since $F_n(\tilde{Y})_1 + F_n(\tilde{Y})_n = 1$, the two terms in the
expected value above are actually equal. For a given $p$, we
know that $S_1$ is binomially distributed with $n$ samples and
probability $\frac{e^\epsilon + 1}{e^\epsilon - 1} \mu$. Substituting this into (51) yields
\[ \text{Var}(F_n(\tilde{Y})_1 | P = p) = \frac{(e^\epsilon + 1)^2}{(e^\epsilon - 1)^2 n^2} \text{Var}(S_1 | P = p) \]
\[ = \frac{(1 + (e^\epsilon - 1)p_1)(e^\epsilon - (e^\epsilon - 1)p_1)}{(e^\epsilon - 1)^2 n}. \]
(57) \hspace{1cm} (58)

After some straightforward rewriting this yields (53).
Since $G_n$ is an unbiased estimator of $T$, we analogously
find that we only need to determine $\text{Var}_{T=1} G_n(\tilde{Y})_1$. Let
$S_{11} := \{ #i : x_i = y_i = 1 \}$ and $S_{12} := \{ #i : x_i = 2, y_i = 1 \}$. Then $S_1 = S_{11} + S_{12}$, and $S_{11}$ is binomially
distributed with $t$ samples and probability $\frac{e^\epsilon}{e^\epsilon + 1}$, while $S_{12}$ is binomially distributed with $n-t$ samples and probability
$\frac{1}{e^\epsilon + 1}$. Since $S_{11}$ and $S_{12}$ are independent given $T$, it follows from (52) that
\[ \text{Var}(G_n(\tilde{Y})_1 | T = t) = \frac{(e^\epsilon + 1)^2}{(e^\epsilon - 1)^2} (\text{Var}(S_{11} | T = t) + \text{Var}_{T=1}(S_{12} | T = t)) \]
\[ = t e^\epsilon + (n-t) e^\epsilon \]
\[ = \frac{e^\epsilon - 1}{e^\epsilon} \]
\[ = \frac{e^\epsilon - 1}{e^\epsilon - 1} \]
\[ = \frac{e^\epsilon}{(e^\epsilon - 1)^2}. \]
\hspace{1cm} (59) \hspace{1cm} (60) \hspace{1cm} (61)

(54) follows directly from this.

This Proposition yields the following Corollary. Below, we use
$f(\epsilon) \sim g(\epsilon)$ to denote $\lim_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} = 1$.

**Corollary 6.4.** Let $a = 2$, and let $\mu$ be given. Let
$(Q_{n,\epsilon}, \Psi_{n,\epsilon})$ and $(Q_{n,\epsilon}', \Psi_{n,\epsilon})$ be as in Corollary 6.2. Then as $\epsilon \to 0$ we have
\[ \lim_{n \to \infty} n \text{MSE}_{n,\mu}(Q_{n,\epsilon}, \Psi_{n,\epsilon}) \sim \frac{\epsilon^2}{2}, \]
\[ \lim_{n \to \infty} n^{-1} \text{MSE}_{n,\mu}(Q_{n,\epsilon}', \Psi_{n,\epsilon}) \sim \frac{\epsilon^2}{2}. \]
\hspace{1cm} (62) \hspace{1cm} (63)

**Proof.** A lower bound (of the behaviour in $\epsilon$ as $\epsilon \to 0$)
is provided by Corollary 6.2, while an upper bound is
provided by Proposition 6.3.

In other words, the bounds in Corollary 6.2 become
tight when $\epsilon \to 0$, provided that $a = 2$. This shows that
our bounds in Table 1 give the best possible coefficient
for $\frac{\epsilon^2}{2}$ that holds for all $a$.

**7. Further research**

Although we have shown that our bounds are tight for
$a = 2, n \gg 0$ and $\epsilon \ll 1$, it would be interesting to
know what frequency estimation accuracy is achievable
in more general settings. In particular, it is interesting to
know what happens for large $a$, and for large $\epsilon$ (i.e. in
the low privacy domain). For large $\epsilon$, the dependence on $b$
will probably also come into play, which might be able
to help point us towards optimal protocols.

Another useful approach would be to give computationally feasible approximations to $\Phi_n$ and $\Psi_n$. The formulas in Proposition 3.2 can become too complex for large $a, b, n$, and while the MLE approach reduces computational complexity, it is still too involved for large $a$. This is important, as for small $a$ the MLE is shown
to give accurate frequency estimations; it would be interesting
to find a way to extend this approach to larger $a$. Furthermore, both our Theorem 3.1, as well as empirical results [10], [17], show that prior knowledge about the
distribution can have a significant impact on frequency estimation.
The MLE ignores this prior knowledge, hence
would be good to find a way to enhance the MLE estimation
by taking prior knowledge into account.

**References**

[1] Apple Differential Privacy Team. Learning with Privacy at Scale. 2017.

[2] Michael Betancourt. “Cruising the simplex: Hamiltonian Monte Carlo and Dirichlet distribution”. In: AIP Conference Proceedings, Vol. 1443.1. AIP. 2012, pp. 157–164.

[3] Jaroslaw Błasiok et al. “Towards instance-optimal private query release”. In: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms; Society for Industrial and Applied Mathematics. 2019, pp. 2480–2497.

[4] Russell C.H. Cheng. “Random Variate Generation”. In: Handbook of Simulation: Principles, Methodology, Advences, Applications, and Practice 139 (1998), p. 172.

[5] Thomas M. Cover and Joy A. Thomas. Elements of Information theory. John Wiley & Sons, 2012.

[6] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. “Collecting telemetry data privately”. In: Advances in Neural Information Processing Systems. 2017, pp. 3571–3580.

[7] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. “Local privacy and statistical minimax rates”. In: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. IEEE. 2013, pp. 429–438.

[8] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local Privacy, Data Processing Inequalities, and Minimax Rates. On arXiv:1302.3201 (2013).

[9] ´Ulfar 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. ACM. 2014, pp. 1054–1067.

[10] Jinyuan Jia and Neil Zhengqiang Gong. “Calibrate: Frequency Estimation and Heavy Hitter Identification with Local Differential Privacy via Incorporating Prior Knowledge”. Preprint, arXiv:1812.02055 (2018).

[11] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. “Extremal mechanisms for local differential privacy”. In: Advances in neural information processing systems. 2014, pp. 2879–2887.

[12] Shiva Prasad Kasiviswanathan et al. “What can we learn privately?” In: SIAM Journal on Computing 40.3 (2011), pp. 793–826.

[13] Milan Lopuhai-Zwakkenberg, Boris Skoric, and Ninghui Li. Information-theoretic metrics for Local Differential Privacy protocols. On arXiv:1910.07826, 2019.

[14] Helmut Strasser. “The asymptotic equivalence of Bayes and maximum likelihood estimation”. In: Journal of Multivariate Analysis 5.2. 1975, pp. 206–226.

[15] Jerry Alan Veeh. “The multivariate Laplace-De Moivre theorem”. In: Journal of multivariate analysis 18.1 (1986), pp. 46–51.
Therefore, the problem of finding \( \hat{p} \) reduces to finding \( \mathcal{A}' \). We determine \( \mathcal{A}' \) by starting with \( \mathcal{A}' = \mathcal{A} \), and then repeatedly removing the \( x \) with the lowest value of \( s_x \), until the estimation (71) is unonnegative for all \( x \).

**Algorithm 1:** Exact MLE post-processing for GRR

**Input:** Total number of users \( n \); privacy budget \( \varepsilon \); tallies of obfuscated data \( s = [s_0, \ldots, s_{n-1}] \)

**Output:** MLE estimate distribution \( \hat{p} = [\hat{p}_1, \ldots, \hat{p}_n] \)

\[
\begin{align*}
[s'_0, \ldots, s'_a, \ldots, s'_{a-1}] & \leftarrow \text{Sort}(s), \text{ such that } \forall i < j, s'_i < s'_j ; \\
k & = 0 ; \\
\text{while } a - k + e^\varepsilon - 1 \geq \frac{\sum_{i=k}^{a-1} s'_i}{s'_k} < 0 \text{ do} \\
& | k \leftarrow k + 1 ; \\
\text{end} \\
\text{for } j = 0, \ldots, a - 1 \text{ do} \\
\ & | \theta_j = \frac{s'_j}{\sum_{i=k}^{a-1} s'_i} ; \\
\ & | \hat{p}_j = \max(0, \theta_j \left( \frac{a - k - 1}{e^\varepsilon - 1} \right) - \frac{1}{e^\varepsilon - 1}) \\
\text{end} \\
\end{align*}
\]

Appendix A.

**MLE Algorithm for GRR**

We describe an efficient algorithm solving (25) for GRR with \( O(a \log a) \) time complexity in Algorithm 1. This algorithm is justified as follows. Recall that GRR (with privacy parameter \( \varepsilon \)) is given by the \( a \times a \)-matrix \( Q \) satisfying

\[
\tilde{Q}_{y|x} = \frac{1 + (e^\varepsilon - 1)\delta_{x=y}}{e^\varepsilon + d - 1}.
\]

(64)

We rewrite (25) to

\[
\begin{align*}
\begin{align*}
\text{maximise}_{p} & \quad f(p) = \sum_{y \in B} s_y \log(\tilde{Q}(p)_y) \\
\text{subject to } & \quad p \geq 0, \sum_{x \in A} p_x = 1.
\end{align*}
\end{align*}
\]

(65)

From this, we obtain the Karush-Kuhn-Tucker (KKT) conditions (with extra variables \((u_x)_{x \in A}\) and \(v\)):

\[
\begin{align*}
\forall x \in \mathcal{A}, \frac{\partial f(p)}{\partial p_x} + u_x - v = 0, \quad \text{(stationarity)}
\end{align*}
\]

\[
\begin{align*}
u_x p_x = 0, \quad \text{(complementary slackness)}
\end{align*}
\]

\[
\begin{align*}
u_x \geq 0, \quad \text{(dual feasibility)}
\end{align*}
\]

\[
\begin{align*}
p_x \geq 0, \sum_{x \in A} p_x = 1, \quad \text{(primal feasibility)},
\end{align*}
\]

in which

\[
\frac{\partial f(p)}{\partial p_x} = \sum_{y \in B} \frac{s_y \tilde{Q}_{y|x}}{\sum_{k \in A} \tilde{Q}_{y|x} p_k} = \frac{(e^\varepsilon - 1)s_x}{(e^\varepsilon - 1)p_x + 1} + \sum_{y \in B} \frac{s_y}{(e^\varepsilon - 1)p_y + 1}.
\]

(66)

(67)

By stationarity and complementary slackness we find for all \( x \in \mathcal{A} \) that

\[
p_x \left( v - \frac{\partial f(p)}{\partial p_x} \right) = 0.
\]

(68)

By summing all \( x \in \mathcal{A} \), we find \( v = n \). Suppose we have found the optimal \( \hat{p} \), and define \( \mathcal{A}' = \{ x : \hat{p}_x > 0 \} \). For \( x, x' \in \mathcal{A}' \), it follows from (68) that we have

\[
\frac{s_{x'}}{s_x} = \frac{1}{(e^\varepsilon - 1)\hat{p}_{x'} + 1} ;
\]

(69)

\[
\text{hence } \frac{s_{x'}}{s_x}((e^\varepsilon - 1)\hat{p}_x + 1) = (e^\varepsilon - 1)\hat{p}_{x'} + 1.
\]

(70)

Summing over all \( x' \in \mathcal{A}' \), we can solve \( \hat{p}_x \) as

\[
\hat{p}_x = \frac{\# \mathcal{A}' - \sum_{x' \in \mathcal{A}' \setminus \mathcal{A}} s_{x'} + (e^\varepsilon - 1)}{(e^\varepsilon - 1) \left( \sum_{x' \in \mathcal{A}' \setminus \mathcal{A}} s_{x'} \right)}.
\]

(71)