Discrete Distribution Estimation under User-level Local Differential Privacy

Jayadev Acharya
Cornell University

Yuhan Liu
Cornell University

Ziteng Sun
Google Research

Abstract

We study discrete distribution estimation under user-level local differential privacy (LDP). In user-level ε-LDP, each user has $m \geq 1$ samples and the privacy of all m samples must be preserved simultaneously. We resolve the following dilemma: While on the one hand having more samples per user should provide more information about the underlying distribution, on the other hand, guaranteeing privacy of all m samples should make estimation task more difficult. We obtain tight bounds for this problem under almost all parameter regimes. Perhaps surprisingly, we show that in suitable parameter regimes, having m samples per user is equivalent to having m times more users, each with only one sample. Our results demonstrate interesting phase transitions for m and the privacy parameter ε in the estimation risk. Finally, connecting with recent results on shuffled DP, we show that combined with random shuffling, our algorithm leads to optimal error guarantees (up to logarithmic factors) under the central model of user-level DP in certain parameter regimes. We provide several simulations to verify our theoretical findings.

1 Introduction

Modern distributed machine learning systems such as federated learning (Kairouz et al., 2021) collects data from users to provide better service. Without proper design, a learning algorithm can reveal sensitive information about the users. Differential privacy (DP) (Dwork et al., 2006), which requires the algorithm's output to be "similar" when a single contribution changes, has become the gold standard for privacy protection in many machine learning and database applications.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

In the distributed setting, the more stringent version, local differential privacy (LDP) (Warner, 1965; Kasiviswanathan et al., 2011) requires users to privatize their data before sending it to the data collector (see Definition 1.2). In other words, the true data never leaves the user. However, LDP comes at a significant drop in utility compared to central DP where a trusted central data collector performs differentially private analysis on user data. To circumvent this, a sequence of recent works (Erlingsson et al., 2019; Cheu et al., 2019; Balle et al., 2019; Girgis et al., 2021; Feldman et al., 2022), has shown that combined with random shuffling, locally randomized data can lead to an amplified DP guarantee in the central model. This setting is often referred to as the *shuffle model of DP* and motivates more study of local randomizers with large privacy parameters.

For the task of discrete distribution under LDP constraints, efficient algorithm and fundamental limits have been established in Duchi et al. (2013); Erlingsson et al. (2014); Kairouz et al. (2016); Ye and Barg (2018); Acharya et al. (2018); Chen et al. (2020). However, these works consider the setting where each user contributes a single data point. The setting where multiple samples per user are allowed, which is common in practice, is largely unexplored.

We study discrete distribution estimation when each user has multiple data samples and must privatize all their samples under LDP. Notice that without privacy constraints, more samples per user means an increase in overall number of samples thus leading to a reduction in the estimation error. When each user has multiple samples, one can choose to ignore all but one sample from each user and obtain the same performance as the item-level LDP where user has one data sample, which will lead to the case performance as the one-sample case. When the users have multiple samples, we have hope of using information from these samples to obtain better estimators. However, the noise addition mechanism also becomes stringent because now changing the value of a data point means changing all the samples of a user. We ask the following question.

Can multiple samples per user help with estimation while maintaining
the same local privacy budget at each user?

We settle this question affirmatively and obtain a nearly

tight characterization of the estimation error for all values of m (the number of samples per user) and ε (the privacy parameter). We show that in certain regimes, *having m samples per user is equivalent to having mn users each with one sample and each with a privacy budget of ε*. Our results also demonstrate interesting phase transitions of the estimation risk in terms of privacy budget ε and number of samples per user m.

Moreover, we show that combined with random shuffling, our results lead to optimal (up to logarithmic factors) estimation error in the central model of user-level privacy (Liu et al., 2020; Narayanan et al., 2022) in certain regimes of ε while maintaining the local privacy guarantee that the server only has access to a properly randomized version of user data. This also establishes the tight estimation error in the shuffle model of DP.

Organization. We define the problem in Section 1.1 and state our results in Section 2. We introduce our algorithms for the high privacy regime ($\varepsilon < 1$) in Section 3. Algorithms for the low privacy regime are discussed in Section 4. Finally we discuss lower bounds in Section 5. Missing proofs are presented in the supplementary material.

1.1 Problem setup and preliminaries

Let $\Delta_k := \{\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_k) \in \mathbb{R}^k : \mathbf{p} \geq 0, \|\mathbf{p}\|_1 = 1\}$ be the $(k-1)$-dimensional probability simplex, which is the set of all k-ary distributions. In this paper, we consider the homogeneous case where there are a total of n users, each observing m i.i.d. samples from the same (unknown) distribution $\mathbf{p} \in \Delta_k$. We write $X_i := (X_{i,1}, X_{i,2}, \ldots, X_{i,m})$ for the samples at user i, and $X^n = (X_1, X_2, \ldots, X_n) \in [k]^{nm}$ for all nm samples.

Remark on heterogeneity. In practice, the data generation process can be heterogeneous. Our results can be extended to the case with limited heterogeneity on user distribution, *e.g.*, $\forall i \in [n], X_i \sim \mathbf{p}_i$ and $d_{\mathrm{TV}}(\mathbf{p}_i, \mathbf{p}) \leq \gamma$. Using the coupling argument in (Levy et al., 2021, Appendix B), the same results can be obtained when γ is small (polynomial in $1/m$ and $1/n$). We leave the study of the more general heterogeneous case as an interesting future work.

To preserve privacy of users, (central) differential privacy requires an algorithm $\mathcal{A}$ has "similar" outputs on neighbouring datasets, formally defined below.

Definition 1.1. An algorithm $\mathcal{A} : [k]^{nm} \to \mathcal{Y}$ is said to be (ε, δ)-DP at user level if for any X^n and X'^n which differ at one user's contribution, we have for any $S \subset \mathcal{Y}$,

$$\Pr\left(\mathcal{A}(X^n) \in S\right) \leq e^{\varepsilon} \Pr\left(\mathcal{A}(X'^n) \in S\right) + \delta.$$

The case of $\delta > 0$ is called approximate DP and $\delta = 0$ is pure DP, denoted as ε-DP. When $m = 1$, this is the same as item-level DP.

In the local model of DP, user i sends a message $Y_i \in \mathcal{Y}$ to the central server through a channel W_i, which describes the randomized mapping from $[k]^m$ to $\mathcal{Y}$. We require each W_i to satisfy LDP constraints:

Definition 1.2. A randomized scheme $W_i : [k]^m \to \mathcal{Y}$ satisfies (ε, δ)-LDP at user-level if $\forall \mathbf{x}, \mathbf{x}' \in [k]^m$ and $S \subset \mathcal{Y}$,

$$W_i(y \in S \mid \mathbf{x}) \leq e^{\varepsilon} \cdot W_i(y \in S \mid \mathbf{x}') + \delta. \tag{1}$$

For LDP, we will focus on the case when $\delta = 0$, denoted as ε-LDP. All messaging schemes satisfying (1) with $\delta = 0$ are denoted $\mathcal{W}_\varepsilon$.

Upon receiving $Y^n := (Y_1, Y_2, \ldots, Y_n)$, the server outputs an estimator $\widehat{\mathbf{p}} \colon \mathcal{Y}^n \to \Delta_k$ for the underlying distribution $\mathbf{p}$. The performance of the estimator is measured by the expected total variation (TV) distance between $\widehat{\mathbf{p}}$ and $\mathbf{p}$, where for $\mathbf{p}, \mathbf{q} \in \Delta_k$, $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) := (1/2)\sum_{x \in [k]} |\mathbf{p}(x) - \mathbf{q}(x)|$. In this work, we are interested in the *minimax risk* of the estimation problem, defined as

$$\mathcal{R}(\varepsilon, k, n, m) := \min_{W^n} \min_{\hat{\mathbf{p}}} \max_{\mathbf{p} \in \Delta_k} \mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}(Y^n), \mathbf{p})] \tag{2}$$

where the minimum over W^n is taken over all ε-LDP messaging schemes.

In general, the choice of W_i may depend on the previous messages $Y^{i-1} := (Y_1, Y_2, \ldots, Y_{i-1})$ and a common random seed U (independent of the observations) available to all users. A protocol is called **noninteractive** if all the channels W_is are chosen independently of each other conditioned on the shared random seed. In distributed systems, noninteractive schemes are easier to implement and lead to lower latency.

Next we introduce composition property of differential privacy and privacy amplification by shuffling, which we will use in later sections.

Theorem 1.3 (Advanced composition (Dwork et al., 2010; Dwork and Roth, 2014))**.** *If messaging schemes $W_1, W_2, \ldots, W_t$ satisfy ε-LDP, then their composition $W^t = (W_1, W_2, \ldots, W_t)$ is ε'-LDP with $\varepsilon' = t\varepsilon$ and (ε'', δ)-LDP with $\varepsilon'' = \varepsilon\sqrt{2t\log(1/\delta)} + t\varepsilon(e^{\varepsilon} - 1)$. Moreover, the choice of W_i is allowed to depend on the outputs of $W_1, W_2, \ldots, W_{i-1}$.*

Theorem 1.4 (Amplification by shuffling (Feldman et al., 2022))**.** *Suppose messaging schemes $W_1, W_2, \ldots, W_n$ satisfy ε-LDP. Let $\mathcal{A}$ be the algorithm that applies $(W_1, W_2, \ldots, W_n)$ on $X_\pi^n = (X_{\pi(1)}, \ldots, X_{\pi(n)})$ where π is a uniform premutation of $[n]$, then we have for $\delta \in (0, 1)$ satisfying $\varepsilon \leq \log(\frac{n}{16\log(1/\delta)})$, $\mathcal{A}$ is (ε', δ) - central DP for*

$$\varepsilon' \leq \log\left(1 + \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}\left(\frac{8\sqrt{e^{\varepsilon}\log(4/\delta)}}{\sqrt{n}} + \frac{8e^{\varepsilon}}{n}\right)\right).$$

When $\varepsilon > 1$, we have $\varepsilon' = O\left(\sqrt{\frac{e^\varepsilon \log(1/\delta)}{n}}\right)$ and when $\varepsilon \le 1$, $\varepsilon' = O\left(\varepsilon\sqrt{\frac{\log(1/\delta)}{n}}\right)$. In the distributed setting, random shuffling is often performed by a secure multi-party communication protocol. Hence besides central DP guarantee, the model also guarantees that the server does not have access to the true user data. This model is often referred to as *shuffle model* (Erlingsson et al., 2019; Cheu et al., 2019; Balle et al., 2019; Girgis et al., 2021; Feldman et al., 2022).

2 Prior work and our results

Distribution estimation under local privacy when each user has one sample ($m = 1$) has been well-studied and it has been established that Erlingsson et al. (2014); Duchi et al. (2013); Kairouz et al. (2016); Ye and Barg (2018); Acharya et al. (2018); Acharya and Sun (2019),[1]

$$\mathcal{R}(\varepsilon, k, n, m = 1) = \Theta\left(\sqrt{\frac{k}{n}} \vee \sqrt{\frac{k^2}{n((e^\varepsilon - 1)^2 \wedge e^\varepsilon)}}\right). \tag{3}$$

The first term is the centralized minimax risk without privacy constraints and the second term is the additional loss due to privacy. In our setup when each player has m samples, without privacy constraints when $\varepsilon = \infty$, the server has unconstrained access to all nm samples giving a risk of

$$\mathcal{R}(\varepsilon = \infty, k, n, m) = \Theta\left(\sqrt{\frac{k}{nm}}\right). \tag{4}$$

Therefore the first term of minimax risk reduces by a factor of $1/\sqrt{m}$ compared to the case when each user has one sample. The conundrum we try to resolve is about the second term. Can one take advantage of the multiple samples per user or does the requirement of guaranteeing privacy to all samples overwhelm the minimax risk?

Consider the case when $\varepsilon = O(1)$. If we only use one sample from each user, we recover the rate $O\left(\sqrt{k^2/(n\varepsilon^2)}\right)$ for the case of $m = 1$ under ε-LDP. Another approach is to use a naive element-level LDP algorithm and apply composition of LDP (Theorem 1.3) to get user-level privacy guarantee. This leads to a rate of $O\left(\sqrt{mk^2/(n\varepsilon^2)}\right)$ under pure LDP or $\tilde{O}(\sqrt{k^2/(n\varepsilon^2)})$ if we relax to approximate LDP and use advanced composition. Either case, the risk does not decrease with m. The question of whether increasing m brings an advantage is still unclear.

Another important question for the general $m > 1$ case is the dependence on the privacy parameter ε. From (3), when $m = 1$, the error rate decreases exponentially with respect to ε when $\varepsilon \in (1, \ln k)$. With $m > 1$, can we still enjoy this exponential rate, and if so, for what ranges of ε?

In this work, we answer these questions, showing that increasing m can *indeed help* in certain regimes and the rate can be as steep as $O(1/\sqrt{m})$ as in the centralized case. Moreover, we characterize the precise dependency on ε, which has more sophisticated phase transitions compared to the case where $m = 1$. Our results are summarized in Table 1.

For sufficiently large n, our rates are tight up to constant factors in all regimes except in $k/e^\varepsilon \le m < k$ where it is tight up to log factors. Somewhat surprisingly, for $\varepsilon < 1$ or $m < k/e^\varepsilon$, the error rate is the same as having m times more users in the one sample case, but the sum of privacy budgets of all users is m times smaller. Next we look at m and ε separately and discuss their rates in different regimes.

Dependence on m. When $\varepsilon < 1$, the error rate always decays as $\Theta(1/\sqrt{m})$. For $\varepsilon \ge 1$, the error rate with respect to m differs for small m ($m < k/e^\varepsilon$), medium m ($k/e^\varepsilon < m < k$), and large m ($m > k$). For small m and large m, the error decays as $\sqrt{m}$, but the dependence on ε is different. For medium m, however, the error barely improves with m by at most a logarithmic factor. It is an interesting future direction to study whether this logarithmic factor is tight.

Dependence on ε. In the high privacy regime ($\varepsilon < 1$), the error decays at a rate of $\Theta(1/\varepsilon)$. The situation in the low privacy regime ($\varepsilon > 1$) is more complicated. When $m < k$, we observe a phase transition at $\varepsilon = \ln(k/m)$. Below this threshold, there is an exponential decay with respect to ε. Beyond $\ln(k/m)$, the rate of decay becomes $\Theta(\sqrt{\varepsilon})$. If $m > k$, then the exponential phase does not exist. When $\varepsilon \ge k$, the error matches that of $\varepsilon = \infty$ and cannot be improved further by increasing ε.

2.1 Connection to central and shuffled DP at user level

Our results imply almost tight rates in the central and shuffle model of DP under certain parameter regimes through amplification by shuffling. In particular, we get the following result.

Theorem 2.1. *For $m < k$ and ε and δ satisfying $\varepsilon < \sqrt{\frac{k\log(1/\delta)^2}{mn}}$ and $\delta \in (0, 1/n)$, using algorithms in Theorem 3.1 and Theorem 4.3 combined with random shuffling, the estimation risk under (ε, δ) user-level DP in the shuffle model is*

$$O\left(\frac{k\log(1/\delta)}{n\sqrt{m}\varepsilon}\right).$$

[1] We use $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

<table>
<tr><td rowspan="2">Regime</td><td rowspan="2">$\varepsilon < 1$</td><td colspan="3">$\varepsilon > 1$</td></tr>
<tr><td>$m < k/e^{\varepsilon}$</td><td>$k/e^{\varepsilon} \le m < k$</td><td>$m \ge k$</td></tr>
<tr><td>Upper bound</td><td rowspan="2">$\sqrt{\frac{k^2}{mn\varepsilon^2}}$ (Thm 3.1)</td><td rowspan="2">$\sqrt{\frac{k^2}{mne^{\varepsilon}}}$ (Thm 4.3)</td><td>$\sqrt{\frac{k}{mn}}+\sqrt{\frac{k\log(\frac{k}{m}+1)}{n\varepsilon}}$ (Thm 4.5)</td><td rowspan="2">$\sqrt{\frac{k}{mn}} + \sqrt{\frac{k^2}{mn\varepsilon}}$† (Cor 4.1, Thm 4.2)</td></tr>
<tr><td>Lower bound</td><td>$\sqrt{\frac{k}{mn}} + \sqrt{\frac{k}{n\varepsilon}}$† (Thm 4.2)</td></tr>
</table>

Table 1: Estimation risks for different parameter regimes of m and ε (omitting constants). The upper bounds hold with mild regularity (see theorem statements for details). For risks marked with †, the lower bounds hold only when $n > (k/\varepsilon)^2$.

Up to logarithmic factors, the bound matches the tight user-level central DP risk established in Liu et al. (2020); Narayanan et al. (2022). which scales as $\tilde{O}(\sqrt{k/(nm)} + k/(n\sqrt{m}\varepsilon))$. Hence it is also tight up to logarithmic factors under shuffle DP. An interesting observation is that the privacy term for central/shuffle DP and local DP have different dependence on n.

We obtain the bound by applying amplification by shuffling (Theorem 1.4) to the LDP algorithm for $\varepsilon < 1$ and $1 \le \varepsilon \le \log(k/m)$. The above regime of ε covers both the $1/\varepsilon$ decay rate when $\varepsilon < 1$ and the $1/e^{\varepsilon/2}$ decay rate when $\varepsilon \ge 1$ and $m < k/e^{\varepsilon}$ in the local setting, showing the benefit of obtaining tight rates for large ε in LDP. Whether this can be achieved for a wider range of ε is an interesting future question. We present the details in the supplementary.

2.2 Our approach

How to utilize the increased sample size at each user while preserving the same level of privacy is the central question to be resolved to design algorithms for $m > 1$. A natural observation is that with m samples, each user can obtain a rough estimate of the entire distribution $\mathbf{p}$ with its local samples.

Observation 2.2. For $x \in [k]$, let $Z_i(x)$ be the counts of x in user i's samples. Then the empirical frequency $Z_i(x)/m$ is concentrated around $\mathbf{p}(x)$ with a standard deviation of $O(\sqrt{\mathbf{p}(x)(1-\mathbf{p}(x))/m})$.

Our algorithm for $\varepsilon \le 1$ relies on this observation. We provide a motivation for the special case of $k = 2$, where we just need to estimate $p := \mathbf{p}(1)$. Define following binomial tail function for a threshold $t \in [0, 1]$

$$P_{m,t}(q) := \Pr_{Z \sim \mathrm{Bin}(m,q)}[\, Z/m > t \,].$$

If p and t are known to be in an interval I of length $O(\sqrt{p(1-p)/m})$, then the derivative of is large is in I. See Figure 1 for an illustration.

To achieve the centralized rate, it suffices to send the indicator $\mathbb{1}\{Z_i(1)/m > t\}$ where $t \in I$. The server then obtains an empirical estimate of $\Pr[\, Z_i(1)/m > t \,]$, and evaluate the inverse function $P^{-1}_{m,t}$ at the empirical estimate to obtain $\hat{p}$. To ensure privacy, the bits of users can be privatized using Randomized Response (Warner, 1965). One remaining ingredient is how to obtain the interval I. For this part, we will rely on Observation 2.2 and apply a private selection-type algorithm, which we will elaborate in Section 3.

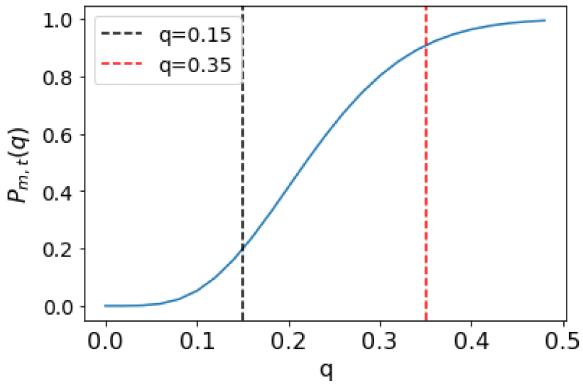


Figure 1: Plot of $P_{m,t}(q)$ with $m = 21$ and $t = 0.25$. Let $s = 0.1$ which is the standard deviation of $\mathrm{Bin}(m, p)/m$ with $p = 0.3$. The vertical lines mark the lines $q = t - s$ and $q = t + s$. We can see that $p \in [t-s, t+s]$ and $P_{m,t}(q)$ increases rapidly in this interval.

For $\varepsilon > 1$, the situation becomes more complicated since we also want to enjoy the benefit of increased privacy budget, especially for $m < k/e^{\varepsilon}$ where an exponential decay in ε is expected. We observe another benefit of having more local samples.

Observation 2.3. For any subset $S \subseteq [k]$, The probability that a user observes at least one sample in S is $P_m(S) = 1 - (1 - \mathbf{p}(S))^m$.

The idea is to divide the domain $[k]$ into m subsets of equal size, denoted by $B_1, \ldots, B_m$. The users are also divided into m groups, each responsible for estimating the probability of symbols in just one subset. A user can only send

useful information about a subset B_j if it observes at least one sample in B_j. If $m = 1$, this happens with probability $\mathbf{p}(B_j)$. However, with m samples, the probability increases to $P_m(S)$. At least 90% of the blocks satisfy $\mathbf{p}(B_j) \le 10/m$, in which case $P_m(B_j) = \Theta(m\mathbf{p}(B_j))$. Hence, the number of effective messages sent by the users roughly increases by a factor of m.

Connection to Acharya et al. (2021a). Acharya et al. (2021a) studied a similar problem under communication constraints where each user sends a message of at most ℓ bits. They show that more samples per user decreases the error by $O(1/\sqrt{m})$ in certain parameter regimes. While our algorithms are inspired by their algorithms, nontrivial extensions and novel ideas are needed to obtain tight rates in the LDP case. We highlight the important differences in terms of algorithm design and proof technique below.

Localization stage. In the localization stage, the analysis for the Gray code scheme in Acharya et al. (2021a) fails since the bits are not private. This issue cannot be resolved by flipping the bits sent by the Gray code scheme using Randomized Response since it requires the error probability for most of the bits in the Gray code to decrease exponentially.

In this work, we view the localization localization stage as a private selection problem and resolve it based on private sparse distribution estimation in Acharya et al. (2021b). In addition to circumventing the failure issue mentioned above in the LDP case, this new idea can also be used in the communication-constrained case considered in Acharya et al. (2021a), which leads to a simpler analysis and better regularity condition. For example, Theorem 2.1 of Acharya et al. (2021a) requires $n/\log n = \Omega(k \log m)$ for 1-bit algorithms, while using communication-limited sparse distribution estimation algorithm in Acharya et al. (2021b) only requires $n = \Omega(k \log m)$.

A unified algorithm for $m \le k/e^\varepsilon$ and $k/e^\varepsilon \le m \le k$. For the algorithms with $m \le k/e^\varepsilon$, we divide the domain $[k]$ into m bins instead of k/e^ε as suggested by Acharya et al. (2021a). Intuitively, this modification ensures that for a uniform $\mathbf{p}$, for any block B_j, $P_m(B_j)$ is some constant away from 0 and 1, which ensures that privatization does not lose too much information. Moreover, the algorithms for $m \le k/e^\varepsilon$ and $k/e^\varepsilon \le m \le k$ are now unified. We can make the same modification to the algorithms in Acharya et al. (2021a) for $m \le k/2^\ell$ and $k/2^\ell \le m \le k$.

Lower bound proof. Acharya et al. (2021a) uses a Poissonization trick, where each user needs to send one bit to indicate whether they get enough samples under Poisson sampling, which might violate privacy constraints. We resolve this issue differently in different regimes. See Section 5 for a detailed discussion.

3 Algorithms for high privacy regime ($\varepsilon \le 1$)

We focus on the high privacy regime ($\varepsilon \le 1$) and show that having more samples per user indeed brings an advantage and the rate decreases as $\Theta(1/\sqrt{m})$.

Theorem 3.1. *When $\varepsilon < 1$ and $n \ge Ck\log(m)/\varepsilon^2$ for some constant C,*

$$\mathcal{R}(\varepsilon, k, n, m) = \Theta\left(\sqrt{\frac{k^2}{mn\varepsilon^2}}\right).$$

Moreover, the bound is achieved by a non-interactive *protocol.*

We describe the upper bound part in this section and discuss the lower bound idea in Section 5. For simplicity, we describe the *interactive* algorithm in this section, which carries most of the algorithmic ideas. We discuss how to modify the algorithm to a *non-interactive* version in Appendix A.1.

Inspired by Acharya et al. (2021a), we start with the special case of $k = 2$ and then generalize to $k > 2$.

3.1 Coin estimation ($k = 2$)

We first consider a simple coin estimation problem, which corresponds to the special case of $k = 2$: There are n users, each has m i.i.d. samples from $\mathrm{Bern}(p)$. The goal is to estimate p under ε-LDP. Our solution to this simple problem will become a crucial building block for algorithms in the general case. The formal guarantee is stated below.

Theorem 3.2. *For $\varepsilon < 1$, there exists an algorithm with an estimate $\hat{p}$ such that if $n \ge C\log(m)/\varepsilon^2$ for some constant C,*

$$\mathbb{E}\left[(\hat{p} - p)^2\right] = O\left(1/\left(mn\varepsilon^2\right)\right).$$

Let $Z_u \sim \mathrm{Bin}(m, p)$ be the number of 1's in user u. Our algorithm is inspired by Acharya et al. (2021a, Section 2.1) and consists of two stages. In the first stage (**localization**), we estimate p up to accuracy $O(\sqrt{p(1-p)/m})$, the standard deviation of the local empirical estimate Z_i/m. Then in the second stage (**refinement**), we try to obtain a more accurate estimate by inverting a binomial density function.

Similar to Acharya et al. (2021a), we divide the $[0, 1]$ interval into $\Theta(\sqrt{m})$ sub-intervals. At a high level, the intervals are designed such that if $p \in I_i$, there exists c, such that

$$(p - c\sqrt{\frac{p(1-p)}{m}}, p + c\sqrt{\frac{p(1-p)}{m}}) \subset I_{i-1} \cup I_i \cup I_{i+1}.$$

This is important for the localization stage since by Observation 2.2, we know that the empirical estimate of p will lie in an interval close to p. Let C_I be a constant

and $r := \lfloor \sqrt{\frac{m}{2C_I}} \rfloor$. We define a partition $\{I_i\}_{i\in[2r]}$. Let $I_i := [l_{i-1}, l_i]$ for $1 \le i \le r$, where

$$l_i := \min\left\{\frac{C_I i^2}{m}, \frac{1}{2}\right\}, \quad 0 \le i \le r.$$

Furthermore $I_{2r+1-i} := [1 - l_i, 1 - l_{i-1}]$.

Next we describe the algorithm, we divide users into two groups S_1, S_2 with equal size, which will be used for the localization stage and refinement stage respectively.

Localization stage. In this stage, the server obtains a crude estimation of p based on messages from S_1.

1. *Privatization scheme.* For $u \in S_1$, let V_u be a $2r$-dimensional binary vector with $\forall i \in [2r], V_u(i) = \mathbb{1}\{Z_u \in I_i\}$, which is a one-hot vector indicating the index of the interval that Z_u falls in. Let Y_u be obtained by flipping each coordinate of V_u with probability $\beta := 1/(e^{\varepsilon/2}+1)$, *i.e.*, $\forall i \in [2r]$,

$$Y_u(i) = \begin{cases} V_u(i) & \text{with prob } 1-\beta, \\ 1 - V_u(i) & \text{with prob } \beta. \end{cases}$$

2. *Estimation scheme.* Here we obtain a confidence interval of p using Y_u's, whose index is given by

$$\hat{i} = \arg\max_{i\in[2r]} \sum_{u\in S_1} Y_u(i).$$

Refinement stage. In this stage, users in S_2 send messages based on $\hat{i}$ and the server obtains a refined estimate of p.

1. *Privatization scheme.* We choose t as follows: if $2 < \hat{i} < 2r-2$, then let t be the mid point of $I_{\hat{i}}$; if $\hat{i} \le 2$, let $t = 1/m$; else let $t = 1 - 1/m$. Users in S_2 send a privatized version of $\mathbb{1}\{Z_u/m > t\}$, *i.e.*,

$$Y_u = \begin{cases} \mathbb{1}\{Z_u/m > t\} & \text{with prob } 1-\beta, \\ 1 - \mathbb{1}\{Z_u/m > t\} & \text{with prob } \beta. \end{cases}$$

2. *Estimation scheme.* Let $P_{m,t}(p) := \Pr\left(Z_u/m > t \mid Z_u \sim \mathrm{Bin}(n/2, p)\right)$ and

$$\hat{P} := \frac{e^\varepsilon+1}{e^\varepsilon-1}\left(\frac{2}{n}\sum_{u\in S_2} Y_u - \frac{1}{e^\varepsilon+1}\right),$$

which is the empirical estimate of $P_{m,t}(p)$. Return $\hat{p} = P_{m,t}^{-1}(\hat{P})$.

The interactive scheme achieves the error rate in Theorem 3.2. We provide a proof sketch here and defer the detailed proof to Appendix A.1. In the localization stage, we show that $p \in \hat{I} := I_{\hat{i}} \cup I_{\hat{i}-1} \cup I_{\hat{i}+1}$ with high probability. More precisely, the failure probability is at most $O(1/(mn\varepsilon^2))$ when $n = \Omega(\log m/\varepsilon^2)$.

In the refinement stage, we condition on the event that $p \in \hat{I}$ (i.e. all expectations are conditioned on $p \in \hat{I}$). Using the analysis for Randomized Response, we have

$$\mathbb{E}\left[(\hat{P} - P_{m,t}(p))^2\right] = O\left(\frac{1}{n\varepsilon^2}\right).$$

Furthermore, since $|\hat{I}| = O(\sqrt{p(1-p)/m})$, it is shown in Acharya et al. (2021a) that for $p \in \hat{I}$,

$$\frac{d}{dp}P_{m,t}(p) = \Omega(\sqrt{m}).$$

Hence, evaluating $\hat{p} = P_{m,t}^{-1}(\hat{P})$ yields the desired accuracy,

$$\begin{aligned}\mathbb{E}\left[(\hat{p}-p)^2\right] &\le \max_{p\in\hat{I}}\left(\frac{dP_{m,t}(p)}{dp}\right)^{-2}\mathbb{E}\left[(\hat{P}-P_{m,t}(p))^2\right] \\ &= O\left(\frac{1}{mn\varepsilon^2}\right).\end{aligned}$$

Combining with the failure probability in the localization stage proves Theorem 3.2 in the interactive setting.

3.2 General case $k > 2$

Using the algorithm for coin estimation, we can design an algorithm for $k > 2$ using ideas from the 1-bit Hadamard Response algorithm (Acharya and Sun, 2019).

Without loss of generality assume k is a power of 2. Let H_k be the Hadamard matrix defined as

$$H_1 = 1, \; H_{2^l} = \begin{bmatrix} H_{2^{l-1}} & H_{2^{l-1}} \\ H_{2^{l-1}} & -H_{2^{l-1}} \end{bmatrix}, \forall l \ge 1.$$

Let $T_i = \{j \in [k] : H_k(i,j) = 1\}$ be the locations of 1's in the ith row of H_k. Users are divided into k groups of size n/k, each responsible for estimating one of $\mathbf{p}(T_i)$. By Theorem 3.2, we can obtain $\widehat{\mathbf{p}}_T(i)$ such that

$$\mathbb{E}\left[(\mathbf{p}(T_i) - \widehat{\mathbf{p}}_T(i))^2\right] = O\left(\frac{k}{mn\varepsilon^2}\right).$$

Let $\widehat{\mathbf{p}}_T = (\widehat{\mathbf{p}}_T(1), \ldots, \widehat{\mathbf{p}}_T(k))$. We obtain $\hat{\mathbf{p}}$ with inverse Hadamard transform $\widehat{\mathbf{p}} = H_k^{-1}(2\widehat{\mathbf{p}}_T - \mathbf{1}_k)$. Let $\mathbf{p}_T = (\mathbf{p}(T_1), \ldots, \mathbf{p}(T_k))$. Since $H_k^\top H_k = kI$, we have

$$\mathbb{E}\left[\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2\right] = \frac{1}{k}\mathbb{E}\left[\|\widehat{\mathbf{p}}_T - \mathbf{p}_T\|_2^2\right] = O\left(\frac{k}{mn\varepsilon^2}\right).$$

Applying Cauchy-Schwarz inequality, we can obtain the desired accuracy in Theorem 3.1.

4 Algorithms for low privacy regime ($\varepsilon > 1$)

In this regime, the main challenge is to design algorithms that takes full advantage of both the increasing sample size

m and extra privacy budget ε. One may easily propose a simple extension of the algorithm for $\varepsilon < 1$: each user split the privacy budget into $\lfloor \varepsilon \rfloor$ parts using the composition property of LDP (Theorem 1.3), each with a budget of 1 (the excess budget is omitted). Now each user can send information about $\lfloor \varepsilon \rfloor \wedge k$ different rows in H_k. Hence the effective sample size increases by a factor of $\varepsilon \wedge k$. Using Theorem 3.1, the guarantee of this algorithm is given by Corollary 4.1

Corollary 4.1. *For $\varepsilon > 1$, if $n > Ck\log(m)/\varepsilon$ for some constant C, the simple extension outputs an estimate $\widehat{\mathbf{p}}$ with*

$$\mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p})] = O\left(\sqrt{\frac{k}{mn}} \vee \sqrt{\frac{k^2}{mn\varepsilon}}\right).$$

Hence we can easily achieve a risk with $1/\sqrt{\varepsilon}$ decay. Can we acheive better rates? It turns out that when $n > (k/\varepsilon)^2$, for large m ($m > k$) the simple extension achieves the following optimal risk.

Theorem 4.2. *For $n > (k/\varepsilon)^2$, $m \geq k$, and $\varepsilon > 1$, the minimax error rate satisfies*

$$\mathcal{R}(\varepsilon, k, n, m) = \Omega\left(\sqrt{\frac{k}{mn}} \vee \sqrt{\frac{k^2}{mn\varepsilon}}\right).$$

For small m and medium m, we can design better algorithms, which we will elaborate in this section.

4.1 Small m ($m \leq k/e^{\varepsilon}$)

For small m, we are able to obtain the same $\Theta(1/\sqrt{m})$ decrease in the rate as in the high privacy case. Moreover, the error rate decays exponentially with ε, similar to the error rate for $m = 1$.

Theorem 4.3. *When $\varepsilon > 1$ and $m < k/e^{\varepsilon}$, if $n > Cm\log(m)$, we have*

$$\mathcal{R}(\varepsilon, k, n, m) = \Theta\left(\sqrt{\frac{k^2}{mne^{\varepsilon}}}\right).$$

The bound is achieved by a non-interactive protocol.

We focus on the upper bound part in this seciton and discuss the lower bound proof in Section 5. At first glance, it may seem overly ambitious to achieve both exponential decay in ε and $1/\sqrt{m}$ improvement in m. Nevertheless, we accomplish this goal by taking advantage of both Observation 2.2 and 2.3, and using the algorithm for $m = 1$ which enjoys exponential dependence on ε as a subroutine. Details of the algorithm are described as follows,

1. Let $\varepsilon_0 = 0.5$[2]. Divide the domain $[k]$ into m blocks $B_1, \ldots, B_m$, each with size k/m.
2. Each user uses $\varepsilon_0 = 0.5$ to estimate the block distribution $\mathbf{p}_B := [\mathbf{p}(B_1), \ldots, \mathbf{p}(B_m)]$ with the algorithm for $\varepsilon \leq 1$ in Section 3. Denote the estimate as $\widehat{\mathbf{p}}_B = [\widehat{\mathbf{p}}_B(1), \ldots, \widehat{\mathbf{p}}_B(m)]$.
3. Divide all users into m groups. The jth group tries to estimate $\bar{\mathbf{p}}_j := \mathbf{p}(\cdot|B_j)$, the distribution conditioned on a sample is in B_j (treated as uniform if $\mathbf{p}(B_j) = 0$). Note that for $x \in B_j$, $\bar{\mathbf{p}}_j(x) = \frac{\mathbf{p}(x)}{\mathbf{p}(B_j)}$.

   To do this, each user in the jth group considers the distribution $\tilde{\mathbf{p}}_j$ over $B_j \cup \{\perp\}$ where

   $$\tilde{\mathbf{p}}_j(\perp) := P_{X^m \sim \mathbf{p}}(\forall x \in B_j, x \notin X^m) = (1 - \mathbf{p}(B_j))^m,$$

   and for $x \in B_j$, $\tilde{\mathbf{p}}_j(x)$ is the probability that x is the first symbol in B_j that appears in X^m . It can be obtained that

   $$\tilde{\mathbf{p}}_j(x) = \bar{\mathbf{p}}_j(x)(1 - \tilde{\mathbf{p}}(\perp)).$$

   A user can simulate a sample from $\tilde{\mathbf{p}}_j$ by getting $\perp$ if $B_j \cap X^m = \varnothing$ and getting the first sample in $X^m \cap B_j$ if it is not empty. Each user then sends a message using Hadamard Response (Acharya et al., 2018) for $(\varepsilon - \varepsilon_0)$-LDP.

   The server can then get an estimate $\hat{\tilde{\mathbf{p}}}_j$ for $\tilde{\mathbf{p}}_j$ using the messages above. Using $\hat{\tilde{\mathbf{p}}}_j$, an estimate $\widehat{\mathbf{p}}_j$ for $\bar{\mathbf{p}}_j$ can be obtained by $\forall x \in B_j$

   $$\widehat{\mathbf{p}}_j(x) = \frac{\hat{\tilde{\mathbf{p}}}_j(x)}{1 - \hat{\tilde{\mathbf{p}}}_j(\perp)},$$

   or m/k if $1 - \hat{\tilde{\mathbf{p}}}_j(\perp) = 0$.
4. To obtain an estimate $\widehat{\mathbf{p}}$ for the underlying distribution, for each $x \in B_j$,

   $$\hat{\mathbf{p}}(x) = \hat{\mathbf{p}}_B(j) \cdot \widehat{\mathbf{p}}_j(x).$$

To derive the guarantee for the algorithm, we need to relate the estimation errors for $\mathbf{p}$, $\tilde{\mathbf{p}}_j$, and $\bar{\mathbf{p}}_j$.

Lemma 4.4. *The estimation errors can be decomposed as*

$$\mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p})] \leq \sum_{j \in [m]} \frac{\mathbf{p}(B_j)}{(m\mathbf{p}(B_j)) \wedge 1} \mathbb{E}\left[d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right)\right] + \mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}_B, \mathbf{p}_B)] \tag{5}$$

From Theorem 3.1, when $n/m > C\log(m)$,

$$\mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}_B, \mathbf{p}_B)] = O\left(\sqrt{\frac{m^2}{mn}}\right) = O\left(\sqrt{\frac{k^2}{mne^{\varepsilon}}}\right). \tag{6}$$

The second inequality is due to $m \leq k^2/(me^{\varepsilon})$ whenever $m \leq k/e^{\varepsilon/2}$. By the guarantee of the Hadamard Response algorithm (Acharya et al., 2018, Corollary 8),

$$\mathbb{E}\left[d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right)\right] = O\left(\sqrt{\frac{(k/m)^2}{(n/m)e^{\varepsilon}}}\right) = O\left(\sqrt{\frac{k^2}{mne^{\varepsilon}}}\right).$$

Plugging in (5) yields the desired bound. Detailed proofs of Lemma 4.4 and Theorem 4.3 are in Appendix D.2.

[2] We choose $\varepsilon_0 = 0.5$ for simplicity. Any constant $\varepsilon_0 < 0.5$ will work without changing the bounds up to constant.

4.2 Medium m ($k/e^\varepsilon < m < k$)

In this regime, we discover that increasing m barely helps with improving the error rates in certain parameter regimes.

Theorem 4.5. *For $\varepsilon > 1$ and $k/e^\varepsilon < m < k$, if $n > Cm\log(m)/\varepsilon$ for some constant C, we have*

$$\mathcal{R}(\varepsilon, k, n, m) = O\left(\sqrt{\frac{k}{mn}} \vee \sqrt{\frac{k\ln(k/m+1)}{n\varepsilon}}\right).$$

The bound is achieved by a non-interactive protocol.

Note that $\mathcal{R}(\varepsilon, k, n, m)$ is non-increasing with m. Setting $m = k$ in Theorem 4.2 yields a lower bound of $\Omega\left(\sqrt{k/mn} \vee \sqrt{k/n\varepsilon}\right)$ for $k/e^\varepsilon < m < k$ when $n > (k/\varepsilon)^2$. Thus Theorem 4.5 is tight up to logarithmic factors.

When $m \le k/e^{\varepsilon/2}$, we use the same algorithm as $m \le k/e^\varepsilon$, and the guarantee is proved similarly (see Appendix B.2 for details). When $m > k/e^{\varepsilon/2}$, we make the following changes,

1. To learn $\mathbf{p}_B = [\mathbf{p}(B_1), \dots, \mathbf{p}(B_m)]$, we use $\varepsilon/2$ privacy budget with the algorithm for $m \ge k$. Hence, the estimation error for $\mathbf{p}_B$ satisfies

$$\mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}_B, \mathbf{p}_B)] = O\left(\sqrt{\frac{m^2}{mn\varepsilon}}\right) = O\left(\sqrt{\frac{k}{n\varepsilon}}\right).$$

   The final equality is due to $m < k$.

2. To estimate $\mathbf{p}(\cdot|B_j)$, we divide the remaining budget of $\varepsilon/2$ into $t' := \lfloor \frac{\varepsilon}{2\ln(k/m)} \rfloor$ parts. Note that with $\ln(k/m)$ privacy budget and n/m samples, we can learn $\tilde{\mathbf{p}}_j$ with accuracy $O(\sqrt{k/n})$. Since $k/m < e^{\varepsilon/2}$, we can assign $m \wedge t'$ blocks to each user. The effective sample size increases by a factor of $m \wedge t'$. Thus

$$\begin{aligned}\mathbb{E}\left[d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right)\right] &= O\left(\sqrt{\frac{k}{n(m \wedge t')}}\right) \\ &= O\left(\sqrt{\frac{k}{mn}} \vee \sqrt{\frac{k\ln(k/m+1)}{n\varepsilon}}\right).\end{aligned}$$

Applying Lemma 4.4 yields the desired upper bound.

5 Lower bound

In this section, we discuss the proof of lower bounds in Theorem 3.1, Theorem 4.3, and Theorem 4.2. We use the information contraction framework in Acharya et al. (2020) and the lower bound construction in Acharya et al. (2021a). Our hard instances are from the "Paninski" family (Paninski, 2008). Let $\gamma \in (0, 1/2)$ be a parameter related to the expected error. We consider a family of distributions defined as follows: for each vector $z \in \mathcal{Z} := \{-1, 1\}^{k/2}$, define a discrete distribution p_z as

$$p_z(2i-1) = \frac{1+\gamma z_i}{k}, \quad p_z(2i) = \frac{1-\gamma z_i}{k}.$$

The m samples observed by each user can be viewed as a k-dimensional vector indicating the histogram from a multinomial distribution. While the proof builds on Acharya et al. (2021a, 2020) for the communication-constrained case, their techniques cannot directly translate to ε-LDP. Their proof relies on the "Poissonization" trick to make each coordinate independent. However, for the trick to work, each user needs to send one bit to indicate whether they get enough samples, which might violate privacy constraints. Our solution is as follows,

- For $m \le k/e^\varepsilon$ and $\varepsilon < 1$, we compute the information contraction bound for multinomial distributions directly, which leads to tight lower bounds without "Poissonization".
- For $m > k/e^\varepsilon$ and $\varepsilon > 1$, "Poissonization" is still used. We show that even if we allow each user to send an extra clean bit, which we term "ε-LDP + 1-bit" channels, the desired lower bound still holds.

We defer the detailed proof to Appendix D.

6 Experiments

The main goal of the section is to demonstrate the effectiveness of our algorithmic ideas and verify our theoretical findings[3]. The experiments are based on prototype algorithms without extensive tuning on constants. We mainly focus on the *interactive* versions since they give better constants than the non-interactive ones numerically. We compare our algorithms to Hadamard Response (Acharya et al., 2018) with 1 sample per user on either n users (referred as **1-sample HR**) or mn users (**all-sample HR**) in various parameter regimes. They serve as baseline upper and lower bounds on the achievable rates under user-level LDP [4]. Average TV error and standard deviation over 5 independent runs are reported. Additional results for both interactive and non-interactive algorithms are provided in Appendix E.

High privacy $\varepsilon \le 1$. Figure 2 shows the result for the high privacy regime for $k = 2$ and $k = 32$, with $m = [32, 64, 128, 256, 512]$. In both cases, the performance of the 1-sample HR remains nearly the same, while the performance of our algorithm is always within a constant (2.5) factor to that of all-sample HR, as Theorem 3.1 suggests.

[3] Code at https://github.com/Azulgrana1/user_level_LDP

[4] RAPPOR Erlingsson et al. (2014) outperforms HR numerically by a small margin (e.g., Acharya et al. (2018)). We compare with HR here since our algorithms use ideas from HR and both algorithms are optimal up to constants.

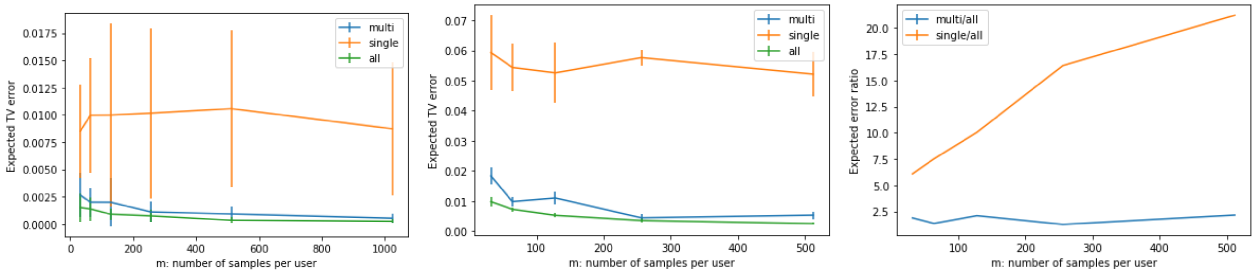


Figure 2: **Left and Middle**: expected TV error with respect to m of our algorithm (blue), 1-sample HR (orange), and all-sample HR (green) in the high privacy regime with $\varepsilon = 0.9$. **Left**: $k = 2, n = 9000, \mathbf{p} = (0.6, 0.4)$. **Middle**: $k = 32, n = 9000k$, $\mathbf{p}$ uniform. **Right**: orange/green and blue/green ratio in the middle plot.

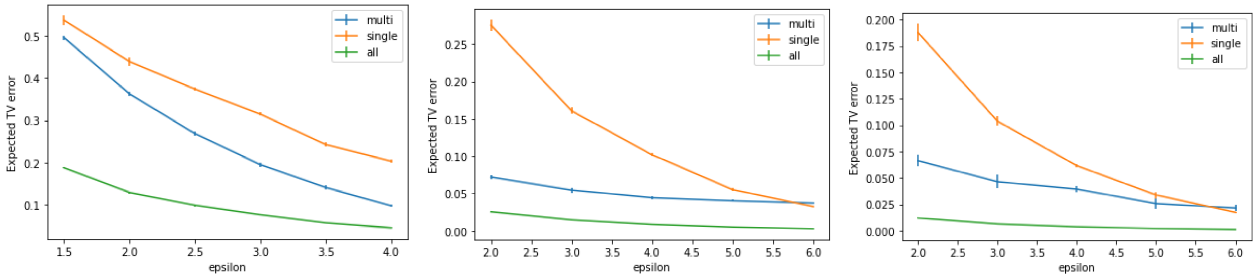


Figure 3: Expected TV error with respect to ε of our algorithm (blue), 1-sample HR (orange), and all-sample HR (green) in the $\varepsilon > 1$ regime with $\mathbf{p}$ uniform. **Left:** $m < k/e^{\varepsilon}$, $k = 1000, m = 20, n = 600k$. **Middle:** $m \in [k/e^{\varepsilon}, k)$. $k = 500, m = 128, n = 1200k$. **Right:** $m \geq k$, $k = 200, m = 256, n = 1200k$.

Low privacy $\varepsilon > 1$. In this regime, we mainly focus on the dependence on ε. Figure 3 shows the expected TV error with respect to ε. When $m < k/e^{\varepsilon}$, our algorithm approaches all-sample HR as ε increases. The rate of decay is much faster than 1-sample HR, indicating an exponential decay with ε as suggested by Theorem 4.3.

When $m \geq k/e^{\varepsilon}$, as suggested by Theorem 4.5 and Corollary 4.1, our algorithms no longer improve exponentially with ε and gradually approaches 1-sample HR (near $\varepsilon = 6$). This is expected, as their rates differ by at most a factor of $\Theta(\sqrt{\ln k})$ when $m = \Theta(k)$ and $\varepsilon \simeq \ln k$.

7 Conclusion

We prove tight min-max rates for discrete distribution learning under user-level LDP, verified by experimental results. The limitations are that we assume users receive the same number of i.i.d. samples from the same unknown distribution, and our algorithms may be inefficient in practice. We leave these directions as future work.

Acknowledgements

Jayadev Acharya and Yuhan Liu are supported in part by the grant NSF-CCF-1846300 (CAREER), and a Fellowship from Google.

References

J. Acharya and Z. Sun. Communication complexity in locally private distribution estimation and heavy hitters. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 51–60, Long Beach, California, USA, June 2019. PMLR. 2, 3.2, A.2

J. Acharya, Z. Sun, and H. Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS'19)*, volume abs/1802.04705, 2018. 1, 2, 3, 4.1, 6, 4, A.2

J. Acharya, C. L. Canonne, Z. Sun, and H. Tyagi. Unified lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562*, 2020. 5, D, D.1, D.1, D.2

J. Acharya, C. Canonne, Y. Liu, Z. Sun, and H. Tyagi. Distributed estimation with multiple samples per user: Sharp rates and phase transition. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18920–18931. Curran Associates, Inc., 2021a. URL https://proceedings.

neurips.cc/paper/2021/file/9d740bd0f36aaa312c8d504e28c42163-Paper.pdf. 2.2, 3, 3.1, 3.1, 5, A.1.2, 1, A.1.3, A.1.3, A.1.3, B.1, D, D.3, D.3, D.3

J. Acharya, P. Kairouz, Y. Liu, and Z. Sun. Estimating sparse discrete distributions under privacy and communication constraints. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 79–98. PMLR, 16–19 Mar 2021b. URL https://proceedings.mlr.press/v132/acharya21b.html. 2.2

B. Balle, J. Bell, A. Gascón, and K. Nissim. The privacy blanket of the shuffle model. In A. Boldyreva and D. Micciancio, editors, *Advances in Cryptology – CRYPTO 2019*, pages 638–667, Cham, 2019. Springer International Publishing. ISBN 978-3-030-26951-7. 1, 1.1

C. L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. doi: 10.4086/toc.gs.2020.009. URL http://www.theoryofcomputing.org/library.html. D

W.-N. Chen, P. Kairouz, and A. Ozgur. Breaking the communication-privacy-accuracy trilemma. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3312–3324. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/222afbe0d68c61de60374b96f1d86715-Paper.pdf. 1

A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev. Distributed differential privacy via shuffling. In Y. Ishai and V. Rijmen, editors, *Advances in Cryptology – EUROCRYPT 2019*, pages 375–403, Cham, 2019. Springer International Publishing. ISBN 978-3-030-17653-2. 1, 1.1

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013*, pages 429–438. IEEE Computer Society, 2013. 1, 2

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, Aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042. 1.3

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, Berlin, 2006. 1

C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010. doi: 10.1109/FOCS.2010.12. 1.3

Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014. 1, 2, 4

Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019. 1, 1.1

V. Feldman, A. McMillan, and K. Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964, 2022. doi: 10.1109/FOCS52979.2021.00096. 1, 1.4, 1.1

A. M. Girgis, D. Data, S. Diggavi, A. T. Suresh, and P. Kairouz. On the renyi differential privacy of the shuffle model. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2321–2341, 2021. 1, 1.1

P. Kairouz, K. Bonawitz, and D. Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2436–2444. JMLR.org, 2016. 1, 2

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. A. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14 (1-2):1–210, 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083. 1

S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011. ISSN 0097-5397. 1

D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh. Learning with user-level privacy. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12466–12479. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/67e235e7f2fa8800d8375409b566e6b6-Paper.pdf. 1.1

Y. Liu, A. T. Suresh, F. X. X. Yu, S. Kumar, and M. Riley. Learning discrete distributions: user vs item-level privacy. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20965–20976. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f06edc8ab534b2c7ecbd4c2051d9cb1e-Paper.pdf. 1, 2.1

S. Narayanan, V. Mirrokni, and H. Esfandiari. Tight and robust private mean estimation with few users. In *International Conference on Machine Learning*, pages 16383–16412. PMLR, 2022. 1, 2.1

L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 5, D

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. ISSN 01621459. URL http://www.jstor.org/stable/2283137. 1, 2.2

M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018. ISSN 0018-9448. 1, 2

A Detailed algorithm and proof for $\varepsilon < 1$

A.1 $\varepsilon < 1, k = 2$

In this section, we provide the detailed proof of Theorem 3.2. Recall the coin estimation problem: there are n users, each has m i.i.d. samples from $\mathrm{Bern}(p)$. We want to estimate p under LDP.

A.1.1 Localization stage

We first prove the guarantee of the localization stage. Let $C_I = 10$ and $r := \lfloor \sqrt{\frac{m}{2C_I}} \rfloor$. We now recall the definition of the intervals $\{I_i\}_{i\in[2r]}$. Let $I_i := [l_{i-1}, l_i]$ for $1 \le i \le r$, where

$$l_i := \min\left\{\frac{C_I i^2}{m}, \frac{1}{2}\right\}, \quad 0 \le i \le r.$$

Furthermore $I_{2r+1-i} := [1 - l_i, 1 - l_{i-1}]$.

For user u, let Z_u be the number of 1s. Z_u induces a random variable $V_u := \arg\max_{i\in[2r]} \mathbb{1}\{Z_u \in I_i\}$, which follows a discrete distribution $\mathbf{q}$, with $\mathbf{q}_i = \Pr[Z_u \in I_i],\ i \in [2r]$.

Recall that $\beta = 1/(e^\varepsilon + 1)$ and define $\gamma = 1 - 2\beta = \frac{e^\varepsilon - 1}{e^\varepsilon + 1}$. We start with the following observation about the partition $\{I_i\}_{i\in[2r]}$.

Lemma A.1. *Suppose that $p \in I_i$ and $p \le 1/2$. Then*

$$\max\left\{\frac{C_I}{m}, \frac{5}{3}\sqrt{C_I}\sqrt{\frac{p}{m}}\right\} \le |I_i| \le \max\left\{\frac{C_I}{m}, 2.5\sqrt{C_I}\sqrt{\frac{p}{m}}\right\}$$

Proof. If $i = 1$, then $|I_i| = C_I/m$. If $i \ge 2$, then since $p \ge C_I i^2/m$,

$$|I_i| = \frac{C_I(2i+1)}{m} \le \frac{C_I(2.5i)}{m} \le 2.5\sqrt{C_I}\sqrt{\frac{p}{m}}.$$

Since $p \le C_I(i+1)^2/m$ and $1 \le (i+1)/3$,

$$|I_i| = \frac{C_I(2i+1)}{m} \ge \frac{C_I(2-1/3)(i+1)}{m} = \frac{5}{3}\sqrt{C_I}\sqrt{\frac{p}{m}}.$$

□

We can obtain a result similar to Lemma A.1 for $p > 1/2$ by replacing p with $1 - p$.

Theorem A.2. *Recall that $\hat{I} = \cup_{|i-\hat{i}|\le 1} I_i$. There exists a constant C such that if $n_1 \ge C\log(m)/\varepsilon^2$, we have*

$$\mathbb{E}\left[(\hat{p} - p)^2 \mathbb{1}\left\{p \notin \hat{I}\right\}\right] = O\left(\frac{1}{mn_1\varepsilon^2}\right)$$

Proof. Let i_p be such that $p \in I_{i_p}$. Due to Lemma A.1, by concentration inequalities for binomials

$$\sum_{|i-i_p|\le 1} \mathbf{q}_i \ge 0.96 =: 1 - \alpha.$$

For $i \in [2r]$, let $M_i = \sum_{u\in S_1} Y_{u,i}$. By Chernoff bound, for i such that $|i - i_p| > 1$, with probability at least $1 - \delta$,

$$M_i \le n_1(\beta + \alpha\gamma) + \sqrt{3n_1(\beta + \alpha\gamma)\log\left(\frac{1}{\delta}\right)} =: M^*.$$

By union bound, with probability at least $1 - 2r\delta = 1 - \Theta(\sqrt{m}\delta)$, $M_i \le M^*$ for all $|i - i_p| > 1$.

Let $i^* = \arg\max_i \mathbf{q}_i$. It is clear that $|i^* - i_p| \le 1$, and $\mathbf{q}_{i^*} \ge (1-\alpha)/3 = 0.32$. Next we argue that with high probability, $M_{i^*} > M^*$, and hence the maximum of M_i's must be achieved at some $i \in [i_p - 1, i_p + 1]$.

First, there exists a constant C_1 such that when $n_1 \ge C_1 \log(1/\delta)/\varepsilon^2$,

$$\mathbb{E}[M_{i^*} - M^*] = \frac{1-4\alpha}{3} n_1 \gamma - \sqrt{3n_1(\beta + \alpha\gamma)\log\frac{1}{\delta}} \ge \frac{1-4\alpha}{6} n_1 \gamma.$$

Therefore, by Chernoff bound,

$$\begin{aligned}\Pr[\, M_{i^*} \le M^* \,] &= \Pr[\, M_{i^*} \le \mathbb{E}[M_{i^*}] - \mathbb{E}[M_{i^*} - M^*] \,] \\ &\le \exp\left(-\frac{\gamma^2 0.28^2}{\beta + 0.32\gamma} n_1\right),\end{aligned}$$

which is at most δ as long as $n_1 \ge C_2 \log(1/\delta)/\varepsilon^2$ for some constant C_2.

Set $\delta = \frac{1}{m^2 n_1 \varepsilon^2}$ and $C' = \max\{C_1, C_2\}$. Then

$$\mathbb{E}\left[(\hat{p} - p)^2 \mathbb{1}\left\{p \notin \hat{I}\right\}\right] \le \Pr\left[p \notin \hat{I}\right] \le (\sqrt{m} + 1)\delta = O\left(\frac{1}{m n_1 \varepsilon^2}\right),$$

as long as

$$n_1 \varepsilon^2 \ge C' \log(m^2 n_1 \varepsilon^2) = 2C' \log m + C' \log(n_1 \varepsilon^2).$$

In addition, there exists a constant C_3 such that as long as $n_1\varepsilon^2 \ge C_3$, we can guarantee $n_1\varepsilon^2/2 \ge C' \log(n_1\varepsilon^2)$. Hence, let $C = \max\{C', C_3\}$, we can guarantee the desired error as long as $n_1 \ge C \log(m)/\varepsilon^2$ □

A.1.2 Interactive scheme

With Theorem A.2, we are now ready to prove the estimation error of the interactive scheme described in Section 3.1. Recall that $P_{m,t}(p) = \Pr_{X \sim \mathrm{Bin}(m,p)}[X/m \ge t]$. We first prove a lower bound on its derivative, similar to (Acharya et al., 2021a, Claim A.10).

Lemma A.3. *Let $S_m(p) \sim \mathrm{Bin}(m, p)$ be a binomial random variable, then*

$$\Pr[\, S_m(p) = s \,] \ge \frac{\sqrt{2\pi}}{e^2\sqrt{s}} e^{-m\frac{(s/m-p)^2}{p(1-p)}}.$$

Proof. Using Stirling's approximation,

$$\sqrt{2\pi} n^{n+1/2} e^{-n} \le n! \le e n^{n+1/2} e^{-n},$$

we have

$$\begin{aligned}\Pr[\, S_m(p) = s \,] &= \binom{m}{s} p^s (1-p)^{m-t} \\ &\ge \frac{\sqrt{2\pi}}{e^2\sqrt{m}} \frac{1}{\sqrt{s/m}\sqrt{1-s/m}} \frac{p^s(1-p)^{m-s}}{(s/m)^s(1-s/m)^{m-s}} \\ &= \frac{\sqrt{2\pi}}{e^2\sqrt{s}} e^{-m D_{KL}(s/m||p)} \\ &\ge \frac{\sqrt{2\pi}}{e^2\sqrt{s}} e^{-m\frac{(s/m-p)^2}{p(1-p)}}.\end{aligned}$$

□

Lemma A.4. *Let $C/m \le p \le 1/2$ for some constant C and $t - 1/m \in [p - C\sqrt{\frac{p}{m-1}}, p + C\sqrt{\frac{p}{m-1}}]$. Then,*

$$\frac{dP_{m,t}(p)}{dp} \ge \frac{\sqrt{\pi}}{e^2} e^{-2C^2} \sqrt{\frac{m}{p}}.$$

For $1 - C/m \ge p \ge 1/2$, the same inequality hods with p replaced by $1 - p$.

Proof. Without loss of generality assume that $s = mt$ is an integer since binomial random variables are integer-valued, and thus we only need to consider integer thresholds. Let $S_m(p) \sim \mathrm{Bin}(m, p)$ be a binomial random variable. From the binomial-beta relation,

$$\Pr[\, S_m(p) \geq s \,] = m \int_0^p \Pr[\, S_{m-1}(u) = s - 1 \,] du.$$

Therefore,

$$\frac{dP_{m,t}(p)}{dp} = m \Pr[\, S_{m-1}(p) = s - 1 \,]. \tag{7}$$

Thus using Lemma A.3, for $s/m \in [p - C\sqrt{pm}, p + C\sqrt{p/m}]$,

$$\Pr[\, S_{m-1}(p) = s - 1 \,] \geq \frac{\sqrt{2\pi}}{e^2\sqrt{s-1}} e^{-C^2/(1-p)} \geq \frac{\sqrt{2\pi}}{e^2\sqrt{s-1}} e^{-2C^2} \geq \frac{\sqrt{2\pi}}{e^2\sqrt{2mp}} e^{-2C^2}$$

The final inequality is due to $mp \geq C^2$, and thus $C\sqrt{mp} \leq mp$ and $s - 1 \leq mp + C\sqrt{mp} \leq 2mp$ □

We also need to bound the derivative of $P_{m,t}(p)$ when $p \leq C/m$,

Lemma A.5. *Let $p \leq C/m$ and $t = 1/m$ for some integer constants $C, C' > 0$. Then,*

$$\frac{dP_{m,t}(p)}{dp} = \Omega(m).$$

For $p \geq 1 - C/m$ and $t = 1 - 1/m$, the same holds with p replaced by $1 - p$.

Proof. When $t = 1/m$, then $P_{m,t}(p) = 1 - (1-p)^m$ and

$$\frac{dP_{m,t}(p)}{dp} = m(1-p)^{m-1}.$$

Since $p \leq C/m$, there exists a constant C' that depends on C such that $(1-p)^{m-1} \geq C'$, hence proving the lemma. □

Now we can prove the guarantee of the interactive algorithm

Theorem A.6. *Let $\varepsilon \leq 1$ and let $\hat{p}$ be the output of the interactive scheme. Then*

$$\mathbb{E}\left[(\hat{p} - p)^2\right] = O\left(\frac{1}{mn\varepsilon^2}\right).$$

Proof. Recall that $\hat{I}$ is a confidence interval obtained from the localization stage. It suffices to prove that

$$\mathbb{E}\left[(\hat{p} - p)^2 \mathbb{1}\left\{p \in \hat{I}\right\}\right] = O\left(\frac{1}{mn\varepsilon^2}\right) \tag{8}$$

since Theorem A.2 establishes

$$\mathbb{E}\left[(\hat{p} - p)^2 \mathbb{1}\left\{p \notin \hat{I}\right\}\right] = O\left(\frac{1}{mn\varepsilon^2}\right),$$

Thus combining the two parts proves the theorem. We now proceed to prove (8). To achieve this, we can safely condition on the event that $p \in \hat{I}$.

Using the analysis for Randomized Response, for all $t \in (0, 1)$,

$$\mathbb{E}\left[(\hat{P} - P_{m,t}(p))^2\right] \leq \frac{2}{n}\left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} + \frac{e^\varepsilon + 1}{(e^\varepsilon - 1)^2}\right) = O\left(\frac{1}{n\varepsilon^2}\right)$$

First, we consider $p \leq 1/2$. If $\hat{i} \leq 2$, recall that in this case $t = 1/m$. Then we must have $p \in \hat{I} \subseteq I_1 \cup I_2 \cup I_3$ and hence $p \leq C/m$ for some constant C. Thus, we can apply Lemma A.5

$$\mathbb{E}\left[(\hat{p} - p)^2 | p \in \hat{I}\right] \leq \max_{p \in \hat{I}} \left(\frac{dP_{m,t}(p)}{dp}\right)^{-2} \mathbb{E}\left[(\hat{P} - P_{m,t}(p))^2 | p \in \hat{I}\right] = O\left(\frac{1}{n\varepsilon^2} \cdot \frac{1}{m^2}\right) = O\left(\frac{1}{mn\varepsilon^2}\right)$$

If $\hat{i} > 2$, then $p \geq C/m$ for some constant C. We use a similar argument and apply Lemma A.4

$$\mathbb{E}\Big[(\hat{p}-p)^2 | p \in \hat{I}\Big] \leq \max_{p \in \hat{I}} \left(\frac{dP_{m,t}(p)}{dp}\right)^{-2} \mathbb{E}\Big[(\hat{P} - P_{m,t}(p))^2 | p \in \hat{I}\Big] = O\left(\frac{1}{n\varepsilon^2} \cdot \frac{1}{m}\right) = O\left(\frac{1}{mn\varepsilon^2}\right)$$

For $p \geq 1/2$ similar holds by replacing p with $1-p$ when applying Lemma A.5 and A.4. Combining all the parts proves the theorem.

□

A.1.3 Non-interactive scheme

Let $C_R = 100{C_I}^2$ and $r' = \lfloor \sqrt{\frac{m}{2C_R}} \rfloor$. For $1 \leq i \leq r'$, define $L_i = [l'_{i-1} - l'_i]$ similarly as $\{I_i\}_{i\in[2r]}$ with C_I replaced by C_R. Let $j_i = (l'_{i-1} + l_i)/2$ and $\{J_i\}_{i\in[2r+1]}$ be the partition defined by j_is. The detailed protocol is described in Algorithm 1.

In Algorithm 1 we define the functions R_2, R_3, R_4 as

$$R_2(p) := \Pr\left(\frac{Z}{m} \in \bigcup_i L_{2i}\right), \quad R_3(p) := \Pr\left(\frac{Z}{m} \in \bigcup_i J_{2i}\right), \quad R_4(p) = \Pr\left(Z \geq 1\right), \tag{9}$$

where $Z \sim \mathrm{Bin}(m, p)$.

Algorithm 1 Non-interactive binomial Estimation Protocol.

Divide users into 4 groups $S_1, \ldots, S_4$. $|S_1| = \frac{n}{2} =: n_1$, $|S_2| = |S_3| = |S_4| = \frac{n}{6} =: N$.

Localization stage. In this stage, the goal is to obtain an interval I, which corresponds to a crude estimate of p.

- **Users**: $u \in S_1$ computes the one-hot encoding of V_u and flips each coordinate with probability $\beta := 1/(e^\varepsilon + 1)$. Denote the flipped vector as $Y_u := (Y_{u,1}, \ldots, Y_{u,2r})$.
- **The server**: Let
$$\hat{i} = \arg\max_{i\in[2r]} \sum_{u\in S_1} Y_{u,i}.$$
Set the confidence interval $\hat{I} = \cup_{i:|i-\hat{i}|\leq 1} I_i$.

Refinement stage. In this stage, we improve the accuracy to $\Theta(1/mn)$.

- **Users**:
  1. $u \in S_2$ flip $\mathbb{1}\{Z_u/m \in \cup_i L_{2i}\}$ with probability β.
  2. $u \in S_3$ flip $\mathbb{1}\{Z_u/m \in \cup_i J_{2i}\}$ with probability β.
  3. $u \in S_4$ flips $\mathbb{1}\{Z_u \geq 1\}$ with probability β.

  Denote Y_u as the flipped bit and send Y_u
- **The server**: According to (Acharya et al., 2021a, Lemma A.8), one of the 3 cases must hold.

  If $\hat{I} \subseteq [0, 65C_R/m]$, let $\bar{Y}_4 = \big(\frac{1}{N}\sum_{u\in S_4} Y_u - \beta\big)/\gamma$
  $$\hat{p} = R_4^{-1}\left(\bar{Y}_4\right) := \{\, p \in [0,1] :\ R_4(p)\ \} = \bar{Y}_4\}.$$
  Else if there exists $i \in [2r]$ such that $\hat{I} \subseteq I'_i := \left[l'_i - \frac{0.55C_R i}{m}, l'_i + \frac{0.55C_R i}{m}\right]$, let $\bar{Y}_2 = \big(\frac{1}{N}\sum_{u\in S_2} Y_u - \beta\big)/\gamma$
  $$\hat{p} = R_{2,I'_i}^{-1}\left(\bar{Y}_2\right) := \left\{\, p \in I'_i :\ R_2(p) = \bar{Y}_2\ \right\}.$$
  Else if there exists $i \in [2r+1]$ such that $\hat{I} \subseteq J'_i := \left[j_i - \frac{0.55C_R i}{m}, j_i + \frac{0.55C_R i}{m}\right]$, let $\bar{Y}_3 = \big(\frac{1}{N}\sum_{u\in S_3} Y_u - \beta\big)/\gamma$
  $$\hat{p} = R_{3,J'_i}^{-1} := \left\{\, p \in J'_i :\ R_3(p) = \bar{Y}_3\ \right\}.$$

Using a similar argument in the proof of (Acharya et al., 2021a, Lemma A.8), the interval $\hat{I}$ has the following property.

Lemma A.7. *Conditioned on $p \in \hat{I}$, at least one of the following must hold,*

1. *There exists $i \in [2r]$, such that $\hat{I} \subseteq I_i' = \left[\frac{C_R i^2}{m} - \frac{0.55 C_R i}{m}, \frac{C_R i^2}{m} + \frac{0.55 C_R i}{m}\right]$*
2. *There exists $i \in [2r+1]$ such that $\hat{I} \subseteq J_i' = \left[j_i - \frac{0.55 C_R i}{m}, j_i + \frac{0.55 C_R i}{m}\right]$*
3. *$\hat{I} \subseteq [0, 65 C_R/m]$*

The proof is identical to (Acharya et al., 2021a, Lemma A.8). Furthermore, in the respective intervals stated in Lemma A.7, there is at least one of $R_2(p), R_3(p), R_4(p)$ with large derivatives.

Lemma A.8. *There exists some absolute constant $C > 0$ such that the following holds.*

1. *For all $i \in [2r]$, $R_2(p)$ is monotonic in $I_i' := \left[l_i' - \frac{0.55 C_R i}{m}, l_i' + \frac{0.55 C_R i}{m}\right]$, and for $p \in I_i'$,*

$$\left|\frac{dR_2(p)}{dp}\right| \geq C\sqrt{\frac{m}{p}}.$$

2. *For all $i \in [2r+1]$, $R_3(p)$ is monotonic in $J_i' := \left[j_i - \frac{2 C_R i}{m}, j_i + \frac{0.55 C_R i}{m}\right]$, and for $p \in J_i'$,*

$$\left|\frac{dR_3(p)}{dp}\right| \geq C\sqrt{\frac{m}{p}}.$$

3. *$R_4(p)$ is monotonic in $[0, 65 C_R/m]$, and for $p \in [0, 65 C_R/m]$,*

$$\frac{dR_4(p)}{dp} \geq Cm.$$

The proof is identical to (Acharya et al., 2021a, Lemma A.9)

Proof of Theorem 3.2. We note that for $h \in \{2, 3, 4\}$, using the analysis for Randomized response,

$$\mathbb{E}\left[(\bar{Y}_h - R_h(p))^2\right] \leq \frac{1}{N} \cdot \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} + \frac{e^\varepsilon + 1}{(e^\varepsilon - 1)^2}\right) = O\left(\frac{1}{N\varepsilon^2}\right)$$

If Case 1 holds in Lemma A.7, we have

$$\mathbb{E}\left[(\hat{p} - p)^2 \,|p \in \hat{I}\right] \leq \max_{p \in \hat{I}} \left(\frac{dR_2(p)}{dp}\right)^{-2} \mathbb{E}\left[(\bar{Y}_2 - R_2(p))^2 | p \in \hat{I}\right] \leq \frac{1}{N\varepsilon^2} \left(\frac{1}{C}\sqrt{\frac{1}{m}}\right)^2 = O\left(\frac{1}{mN\varepsilon^2}\right),$$

where we use Lemma A.8.

When Case 2 holds, we can prove it similarly by inverting $R_3(p)$. When Case 3 holds, we have

$$\mathbb{E}\left[(\hat{p} - p)^2 \,|p \in \hat{I}\right] \leq \max_{p \in \hat{I}} \left(\frac{dR_4(p)}{dp}\right)^{-2} \mathbb{E}\left[(\bar{Y}_4 - R_4(p))^2 | p \in \hat{I}\right] \leq \frac{1}{N\varepsilon^2} \cdot \frac{1}{C^2 m^2} \leq O\left(\frac{1}{mN\varepsilon^2}\right).$$

Together this implies $\mathbb{E}\left[(\hat{p} - p)^2 \mathbb{1}\left\{p \in \hat{I}\right\}\right] = O\left(\frac{1}{mn\varepsilon^2}\right)$. Combining with Theorem A.2, this concludes the proof of Theorem 3.2. □

A.2 $\varepsilon < 1, k > 2$

Theorem A.9. *There exists a constant C and an ε-LDP algorithm such that when $n \geq Ck\log(m)/\varepsilon^2$,*

$$\mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p})] = O\left(\sqrt{\frac{k^2}{mn\varepsilon^2}}\right).$$

Proof. We use an idea considered in Acharya et al. (2018) and estimate the probabilities of subsets of $[k]$ defined below.

Let $K = 2^{\lceil \log_2(k+1) \rceil}$ be the smallest power of 2 larger than k and H_K be the $K \times K$ Hadamard matrix. Define $T_i = \{j \in [k] : H_K(i,j) = 1\}$, i.e., the locations of 1's in the ith row of H_K. Let $\mathbf{p}_T = (\mathbf{p}(T_1), \mathbf{p}(T_2), \ldots, \mathbf{p}(T_K))$. The following two claims are shown in Acharya and Sun (2019).

Claim A.10. *For any distribution* $\mathbf{p}$*, we have*

$$\mathbf{p}_T = \frac{H_K \cdot \mathbf{p} + \mathbf{1}_K}{2},$$

where we append 0's to $\mathbf{p}$ *to make it of dimension* K*.*

Claim A.11. *For all* $\mathbf{p}, \hat{\mathbf{p}}$*, we have*

$$\|\mathbf{p}_T - \hat{\mathbf{p}}_T\|_2^2 = \frac{K}{4}\|\mathbf{p} - \hat{\mathbf{p}}\|_2^2.$$

The above two claims show that for any estimate for the set probabilities $\hat{\mathbf{p}}_T$, we can obtain an estimate for $\mathbf{p}$ by inverting the formula in Claim A.10. Moreover, Claim A.11 establishes the relation between the errors for the two estimates.

Now we described the protocol.

1. Divide users into K subsets, each with size n/K.
2. Users in the ith set count the number of samples in T_i and apply the ε-LDP protocol in Theorem 3.2 to estimate $\mathbf{p}(T_i)$.
3. After obtaining the estimates $\hat{\mathbf{p}}(T_i)$, the server returns the first k coordinates of $\hat{\mathbf{p}}$ where

$$\hat{\mathbf{p}} = H_K^{-1}(2\hat{\mathbf{p}}_T - \mathbf{1}_K),$$

where $\widehat{\mathbf{p}}_T = (\widehat{\mathbf{p}}(T_1), \widehat{\mathbf{p}}(T_2), \ldots, \widehat{\mathbf{p}}(T_K))$.

By Theorem 3.2, for $n/K \geq C \log m/\varepsilon^2$ where C is the constant in Theorem 3.2,

$$\mathbb{E}\left[\|\widehat{\mathbf{p}}_T - \mathbf{p}_T\|_2^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[(\widehat{\mathbf{p}}(T_i) - \mathbf{p}(T_i))^2\right] = O\left(\frac{K}{m(n/K)\varepsilon^2}\right) = O\left(\frac{K^2}{mn\varepsilon^2}\right).$$

Combining with Claim A.11, we get

$$\mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p})] \leq \frac{1}{2}\sqrt{K\mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2\right]} = \frac{1}{2}\sqrt{K\mathbb{E}\left[\frac{4}{K}\|\hat{\mathbf{p}}_T - \mathbf{p}_T\|_2^2\right]} = O\left(\sqrt{\frac{K^2}{mn\varepsilon^2}}\right).$$

Then the upper bound of Theorem A.9 follows by $K \leq 2k$.

□

B Missing proofs for $\varepsilon > 1$

B.1 $m \leq k/e^{\varepsilon}$

To prove Lemma 4.4, we use (Acharya et al., 2021a, Lemma 3.2), which states

$$\mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{p})] \leq \mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}_B, \mathbf{p}_B)] + \sum_j \mathbf{p}(B_j) d_{\mathrm{TV}}(\hat{\mathbf{p}}_j, \bar{\mathbf{p}}_j).$$

The only missing part is the following claim.

Claim B.1. *For all* $j \in [t]$*,*

$$d_{\mathrm{TV}}(\hat{\mathbf{p}}_j, \bar{\mathbf{p}}_j) \leq \frac{d_{\mathrm{TV}}(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j)}{1 - \tilde{\mathbf{p}}_j(\perp)}.$$

Proof.

$$d_{\mathrm{TV}}(\hat{\mathbf{p}}_j, \bar{\mathbf{p}}_j) = \sum_{x \in B_j} |\hat{\mathbf{p}}_j(x) - \bar{\mathbf{p}}_j(x)| \tag{10}$$

$$= \sum_{x \in B_j} \left| \frac{\hat{\tilde{\mathbf{p}}}_j(x)}{1 - \hat{\tilde{\mathbf{p}}}_j(\perp)} - \frac{\tilde{\mathbf{p}}_j(x)}{1 - \tilde{\mathbf{p}}_j(\perp)} \right| \tag{11}$$

$$\leq \sum_{x \in B_j} \left(\left| \frac{\hat{\tilde{\mathbf{p}}}_j(x)}{1 - \hat{\tilde{\mathbf{p}}}_j(\perp)} - \frac{\hat{\tilde{\mathbf{p}}}_j(x)}{1 - \tilde{\mathbf{p}}_j(\perp)} \right| + \left| \frac{\hat{\tilde{\mathbf{p}}}_j(x)}{1 - \tilde{\mathbf{p}}_j(\perp)} - \frac{\tilde{\mathbf{p}}_j(x)}{1 - \tilde{\mathbf{p}}_j(\perp)} \right| \right) \tag{12}$$

$$= \sum_{x \in B_j} \frac{\hat{\tilde{\mathbf{p}}}_j(x) \left| \hat{\tilde{\mathbf{p}}}_j(\perp) - \tilde{\mathbf{p}}_j(\perp) \right|}{(1 - \hat{\tilde{\mathbf{p}}}_j(\perp))(1 - \tilde{\mathbf{p}}_j(\perp))} + \frac{\sum_{x \in B_j} \left| \hat{\tilde{\mathbf{p}}}_j(x) - \tilde{\mathbf{p}}_j(x) \right|}{1 - \tilde{\mathbf{p}}_j(\perp)} \tag{13}$$

$$= \frac{\left| \hat{\tilde{\mathbf{p}}}_j(\perp) - \tilde{\mathbf{p}}_j(\perp) \right| + \sum_{x \in B_j} \left| \hat{\tilde{\mathbf{p}}}_j(x) - \tilde{\mathbf{p}}_j(x) \right|}{1 - \tilde{\mathbf{p}}_j(\perp)} \tag{14}$$

$$= \frac{d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right)}{1 - \tilde{\mathbf{p}}_j(\perp)}. \tag{15}$$

□

Noting that $1 - \tilde{\mathbf{p}}_j(\perp) = \Theta(m\mathbf{p}(B_j) \wedge 1)$ completes the proof of Lemma 4.4.

Finally to prove Theorem 4.3, recall that

$$d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right) = O\left(\sqrt{\frac{(k/m)^2}{(n/m)e^\varepsilon}}\right) = O\left(\sqrt{\frac{k^2}{mne^\varepsilon}}\right).$$

Therefore,

$$\sum_{j \in [t]} \mathbf{p}(B_j) \mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}_j, \bar{\mathbf{p}}_j)] \leq O\left(\sum_{j \in [t]} \frac{\mathbf{p}(B_j)}{m\mathbf{p}(B_j) \wedge 1} \mathbb{E}\left[d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right)\right] \right)$$

$$= O\left(\sqrt{\frac{k^2}{mne^\varepsilon}}\right) \cdot \sum_{j \in [m]} \left(\mathbf{p}(B_j) + \frac{1}{m} \right) = O\left(\sqrt{\frac{k^2}{mne^\varepsilon}}\right).$$

Combining with (6) completes the proof of Theorem 4.3.

B.2 $k/e^\varepsilon < m < k$

We provide the detailed proof for $m \leq k/e^{\varepsilon/2}$. Recall that in this regime we use the algorithm for $m \leq k/e^\varepsilon$. Since $m \leq k/e^{\varepsilon/2}$, by Theorem 3.1

$$\mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}_B, \mathbf{p}_B)] = O\left(\sqrt{\frac{m}{n}}\right) = O\left(\sqrt{\frac{k}{ne^{\varepsilon/2}}}\right)$$

However, since each block only has $k/m \leq e^\varepsilon$ elements, the error for estimating $\tilde{\mathbf{p}}_j$ satisfies

$$\mathbb{E}\left[d_{\mathrm{TV}}\left(\hat{\tilde{\mathbf{p}}}_j, \tilde{\mathbf{p}}_j\right)\right] = O\left(\sqrt{\frac{k/m}{(n/t)}}\right) = O\left(\sqrt{\frac{k}{n}}\right).$$

Using the same argument as $m \leq k/e^\varepsilon$, we have

$$\mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathbf{p})] = O\left(\sqrt{\frac{k}{n}}\right) = O\left(\sqrt{\frac{k \ln(k/m + 1)}{n\varepsilon}}\right).$$

The final equality is due to $\varepsilon/2 \leq \ln(k/m)$.

C Connection to central DP and the shuffle model.

In this section, we provide the proof of Theorem 2.1. The bound can be obtained by a combination of amplification by shuffling (Theorem 1.4) and the upper bound results in Theorem 3.1 and Theorem 4.3. We assume without shuffling, each user sends an ε_0-LDP message.

Small ε_0 : $\varepsilon_0 \le 1$. Note that in this case, in the shuffle model, $\varepsilon = O\left(\varepsilon_0\sqrt{\frac{\log(1/\delta)}{n}}\right)$. More specifically, for $\varepsilon < \sqrt{\frac{9e\log(4/\delta)}{n}}$, there exists $\varepsilon_0 = \varepsilon\cdot\sqrt{\frac{n}{9e\log(4/\delta)}} < 1$ such that the ε_0-LDP algorithm is (ε,δ)-DP in the shuffle model. Plugging this into Theorem 3.1, we get the desired bound in Theorem 2.1.

Large ε_0 : $1 \le \varepsilon_0 \le \log(k/m)$. In this case, in the shuffling model, $\varepsilon = O\left(\sqrt{\frac{e^{\varepsilon_0}\log(1/\delta)}{n}}\right)$. More specifically, when $\varepsilon < \sqrt{\frac{k\log(1/\delta)^2}{mn}}$, there exists $\varepsilon_0 = \frac{1}{2}\log\frac{n\varepsilon^2}{\log(1/\delta)} < \log(k/m)$ such that the ε_0-LDP algorithm is (ε,δ)-DP in the shuffle model. Plugging this into Theorem 4.3, we get the desired bound in Theorem 2.1.

D Missing proofs for the lower bounds

In this section, we present complete proofs for lower bound part of Theorem 3.1, Theorem 4.3, and Theorem 4.2. We use the information contraction framework in Acharya et al. (2020) and the lower bound construction in Acharya et al. (2021a). Our hard instances are from the "Paninski" family Paninski (2008). Let $\gamma \in (0, 1/2)$ be a parameter related to the expected error. We consider a family of distributions defined as follows: for each vector $z \in \mathcal{Z} := \{-1,1\}^{k/2}$, define a discrete distribution $\mathbf{p}_z$ as

$$\mathbf{p}_z(2i-1) = \frac{1+\gamma z_i}{k}, \quad \mathbf{p}_z(2i) = \frac{1-\gamma z_i}{k}.$$

The m samples observed by each user can be viewed as a k-dimensional vector indicating the histogram from a multinomial distribution. We denote this distribution as $\mathbf{p}_z^{\mathrm{mul}} = \mathrm{Multinomial}(m, \mathbf{p}_z)$. In this section, we use $\mathbf{m} = (\mathbf{m}(1), \mathbf{m}(2), \ldots, \mathbf{m}(k))$ to denote the histogram observed from $\mathbf{p}_z^{\mathrm{mul}}$ where $\mathbf{m}(x)$ denotes the number of times x appears in the observed m samples.

When m is large ($m \ge k$), we will consider the "Poissonization" of the multinomial distribution, which we denote as $\mathbf{p}_z^{\mathrm{poi}} = \mathrm{Poisson}(m, \mathbf{p}_z)$. To generate a sample from $\mathbf{p}_z^{\mathrm{poi}}$, first a random integer M is generated from $\mathrm{Poi}(m)$ and the final observed samples are generated from $\mathrm{Multinomial}(M, \mathbf{p}_z)$. It is a folklore (*e.g.*, Canonne (2020)) for $\mathbf{m} = (\mathbf{m}(1), \mathbf{m}(2), \ldots, \mathbf{m}(k)) \sim \mathbf{p}_z^{\mathrm{poi}}$, we have: (1) All $\mathbf{m}(x)$'s are mutually independent; (2) $\forall x \in [k]$, $\mathbf{m}(x)$ follows a Poisson distribution with mean $m\mathbf{p}_z(x)$.

As discussed in Section 5, we provide our proof in two separate regimes. In Appendix D.2, we prove the lower bound part of Theorem 3.1 and Theorem 4.3 by directly analyzing the multinomial setting. In Appendix D.3, we prove Theorem 4.2 using the Poissonization trick introduced above. We first introduce the information contraction framework in Acharya et al. (2020) and necessary results.

D.1 Information contraction bounds

Let $\mathcal{Z} := \{-1,+1\}^k$ and $\{\mathbf{q}_z\}_{z\in\mathcal{Z}}$ be a collection of distributions over $\mathcal{X}$, indexed by $z \in \mathcal{Z}$. For $z \in \mathcal{Z}$, denote by $z^{\oplus i} \in \mathcal{Z}$ the vector obtained by flipping the sign of the ith coordinate of z. The following two assumptions on the density functions are needed.

Assumption 1. For every $z \in \mathcal{Z}$ and $i \in [k]$ it holds that $\mathbf{q}_{z^{\oplus i}} \ll \mathbf{q}_z$, and there exist measurable functions $\phi_{z,i}\colon \mathcal{X} \to \mathbb{R}$ such that

$$\frac{d\mathbf{q}_{z^{\oplus i}}}{d\mathbf{q}_z} = 1 + \alpha_{z,i}\phi_{z,i},$$

where $|\alpha_{z,i}| \le \alpha$ for some constant $\alpha \in \mathbb{R}$ independent of z, i. Moreover, for all $z \in \mathcal{Z}$ and $i, j \in [k]$, $\mathbb{E}_{\mathbf{q}_z}[\phi_{z,i}\phi_{z,j}] = \mathbb{1}\{i = j\}$. (In particular, $\mathbb{E}_{\mathbf{p}_z}[\phi_{z,i}^2] = 1$.)

Assumption 2. There exists some $\sigma \ge 0$ such that, for all $z \in \mathcal{Z}$, the random vector $\phi_z(X) := (\phi_{z,i}(X))_{i\in[k]} \in \mathbb{R}^k$ is σ^2-subgaussian for $X \sim \mathbf{q}_z$, with independent coordinates.

Consider the following generating process. We first pick Z uniformly at random from $\mathcal{Z}$. Then each user observes a sample from $\mathbf{q}_Z$. The users follow the protocol Π where each user uses a messaging scheme from a constrained set $\mathcal{W}$ (*e.g.*, $\mathcal{W}_\varepsilon$ denotes all ε-LDP schemes) to send a message Y_i about there sample. The server observes all the messages Y^n and estimate the distribution as $\hat{\mathbf{p}}$.

We denote the distribution of Y^n when the samples are from $\mathbf{q}_Z$ as $\mathbf{q}_Z^{Y^n}$. We also denote the mixture of message distributions conditioned on a fixed Z_i as the following $\mathbf{q}_{+i}^{Y^n} := \mathbb{E}\left[\mathbf{q}_Z^{Y^n} \,\middle|\, Z_i = 1\right]$, $\mathbf{q}_{-i}^{Y^n} := \mathbb{E}\left[\mathbf{q}_Z^{Y^n} \,\middle|\, Z_i = 1\right]$. Note that $d_{\mathrm{TV}}\left(\mathbf{q}_{+i}^{Y^n}, \mathbf{q}_{-i}^{Y^n}\right)$ can be viewed as an information measure that describes how much information Y^n carries about Z_i. The following theorem provides an upper bound on this information measure.

Theorem D.1 (Main theorem of Acharya et al. (2020))**.** *Let Π be a sequentially interactive protocol using messaging schemes from $\mathcal{W}$ and (Y^n, U) be the transcript of Π when the input $X_1, \ldots, X_n$ is i.i.d. with common distribution $\mathbf{q}_Z$. Then, under Assumption 1, we have*

$$\left(\frac{1}{k}\sum_{i=1}^{k} d_{\mathrm{TV}}\left(\mathbf{q}_{+i}^{Y^n}, \mathbf{q}_{-i}^{Y^n}\right)\right)^2 \le \frac{7}{k} n\alpha^2 \max_{z\in\mathcal{Z}} \max_{W\in\mathcal{W}} \sum_{y\in\mathcal{Y}} \frac{\mathrm{Var}_{\mathbf{q}_z}[W(y \mid X)]}{\mathbb{E}_{\mathbf{q}_z}[W(y \mid X)]}, \tag{16}$$

Finally, if Assumption 2 holds as well, we have

$$\left(\frac{1}{k}\sum_{i=1}^{k} d_{\mathrm{TV}}\left(\mathbf{q}_{+i}^{Y^n}, \mathbf{q}_{-i}^{Y^n}\right)\right)^2 \le \frac{14\ln 2}{k} n\alpha^2\sigma^2 \max_{z\in\mathcal{Z}} \max_{W\in\mathcal{W}} I(\mathbf{q}_z; W), \tag{17}$$

where $I(\mathbf{q}_z; W)$ denotes the mutual information $I(X;Y)$ between the input $X \sim \mathbf{q}_z$ and the output Y of the channel W with X as input.

In particular, it is proved in Acharya et al. (2020) that when $\mathcal{W}_\varepsilon$ is the set of all ε-LDP channels, we have for any $\mathbf{q}_z$,

$$\max_{z\in\mathcal{Z}} \max_{W\in\mathcal{W}_\varepsilon} \sum_{y\in\mathcal{Y}} \frac{\mathrm{Var}_{\mathbf{q}_z}[W(y \mid X)]}{\mathbb{E}_{\mathbf{q}_z}[W(y \mid X)]} \le \min\left\{4\varepsilon^2, e^\varepsilon\right\}. \tag{18}$$

D.2 $m \le k/e^\varepsilon$ or $\varepsilon < 1$

We prove the minimax lower bound presented below.

Theorem D.2. *The minimax error rate satsifies*

$$\mathcal{R}(\varepsilon, k, n, m) = \Omega\left(\sqrt{\frac{k}{mn}} \vee \sqrt{\frac{k^2}{mn(\varepsilon^2 \wedge e^\varepsilon)}}\right)$$

Note that when $\varepsilon < 1$, ε^2 is the dominating term, leading to the tight lower bound in the high privacy regime (Theorem 3.1). When $\varepsilon \ge 1$, e^ε is the dominating term, which yields the desired lower bound for $m < k/e^\varepsilon$ (Theorem 4.2).

Proof. The first term is the lower bound in the centralized setting. We will mainly focus on the second term. Consider the same generating process described in Appendix D.1 with $\mathbf{q}_z = \mathbf{p}_z^{\mathrm{mul}}$. The following lemma shows that if $\Pi, \hat{\mathbf{p}}$ is a good estimator for $\mathbf{p}_Z$, we must be able to extract enough information about Z from Y^n. The result follows from (Acharya et al., 2020, Lemma).

Lemma D.3. *If $\Pi, \hat{\mathbf{p}}$ satisfies*

$$\mathbb{E}[d_{\mathrm{TV}}(\hat{\mathbf{p}}(Y^n), \mathbf{p})] \le \frac{\gamma}{4},$$

we must have

$$\sum_{i=1}^{k/2} d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{\mathrm{mul},Y^n}, \mathbf{p}_{-i}^{\mathrm{mul},Y^n}\right) = \Omega(k). \tag{19}$$

Next we upper bound the left hand side of (19) using Theorem D.1. In particular, we will prove $\mathbf{p}_z^{\mathrm{mul}}$ satisfies Assumption 1 with appropriate parameters..

Lemma D.4. *$\{\mathbf{p}_z^{\mathrm{mul}}\}_{z\in\mathcal{Z}}$ satisfies Assumption 1 with $\alpha = O(\sqrt{m\gamma^2/k})$ for $\gamma < \min\{1/2, \sqrt{k/(8m+k)}\}$.*

Proof. For a vector $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_k) \in \mathbb{N}^k$, the probability mass is

$$\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m}) = m! \prod_{i=1}^{k} \frac{\mathbf{p}_z(i)^{\mathbf{m}_i}}{\mathbf{m}_i!}.$$

Therefore,

$$\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})} = \left(\frac{1-\gamma z_i}{1+\gamma z_i}\right)^{\mathbf{m}_{2i-1}} \left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^{\mathbf{m}_{2i}} = \left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^{\mathbf{m}_{2i}-\mathbf{m}_{2i-1}}.$$

We want to compute

$$\mathbb{E}_{\mathbf{p}_z^{\mathrm{mul}}}\left[\left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})} - 1\right)^2\right] = \mathbb{E}_{\mathbf{p}_z^{\mathrm{mul}}}\left[\left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})}\right)^2\right] - 1.$$

First let $N = m_{2i-1} + m_{2i}$. For fixed N, $\mathbf{m}_{2i}$ follows $\mathrm{Bin}(N, p)$ where $p = (1-\gamma z_i)/2$. Hence we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})}\right)^2 \,\middle|\, N\right] &= \mathbb{E}_{\mathbf{m}_{2i}\sim\mathrm{Bin}(N,p)}\left[\left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^{4\mathbf{m}_{2i}-2N}\right] \\
&= \left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^{-2N}\left(p\left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^4 + 1 - p\right)^N \\
&= \left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^{-2N}\left(\frac{1+\gamma z_i}{2}\left(\left(\frac{1+\gamma z_i}{1-\gamma z_i}\right)^3 + 1\right)\right)^N \\
&= \left(\frac{1}{2}\left(\frac{(1+\gamma z_i)^2}{1-\gamma z_i} + \frac{(1-\gamma z_i)^2}{1+\gamma z_i}\right)\right)^N \\
&= \left(\frac{1+3\gamma^2}{1-\gamma^2}\right)^N.
\end{aligned}$$

The second equality follows by the generating function of binomial distribution. Notice that $N \sim \mathrm{Bin}(m, 2/k)$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})} - 1\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})}\right)^2 \,\middle|\, N\right]\right] - 1 \\
&= \mathbb{E}_{N\sim\mathrm{Bin}(m,2/k)}\left[\left(\frac{1+3\gamma^2}{1-\gamma^2}\right)^N\right] - 1 \\
&= \left(\frac{2}{k}\frac{1+3\gamma^2}{1-\gamma^2} + 1 - \frac{2}{k}\right)^m - 1 \\
&= \left(1 + \frac{8\gamma^2}{k(1-\gamma^2)}\right)^m - 1 =: \alpha = O(m\gamma^2/k).
\end{aligned}$$

Setting $\alpha_{z,i} = \sqrt{\mathbb{E}_{\mathbf{p}_z^{\mathrm{mul}}}\left[\left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})} - 1\right)^2\right]}$ and $\phi_{z,i} = \left(\frac{\mathbf{p}_{z^{\oplus i}}^{\mathrm{mul}}(\mathbf{m})}{\mathbf{p}_z^{\mathrm{mul}}(\mathbf{m})} - 1\right)/\alpha_{z_i}$ yields the desired result. It is obvious that $\mathbb{E}[\phi_{z,i}\phi_{z,j}] = \mathbb{1}\{i=j\}$. □

Combining Lemma D.4, Theorem D.1, and Eq. (18), we get:

$$\gamma = \Omega\left(\sqrt{\frac{k^2}{mn\min\{\varepsilon^2, e^\varepsilon\}}}\right),$$

completing the proof. □

D.3 Large m: $m > k$.

We prove Theorem 4.2, restated below.

Theorem D.5. *For $n > (k/\varepsilon)^2$, $m \geq k$, and $\varepsilon > 1$, the minimax error rate satisfies*

$$\mathcal{R}(\varepsilon, k, n, m) = \Omega\left(\sqrt{\frac{k}{mn}} \vee \sqrt{\frac{k^2}{mn\varepsilon}}\right).$$

For $m > k$, we prove the lower bound via Poissonization. Formally, define the following problems.

MULTINOMIAL$(\mathcal{W}, n, m)$: each of the n users obtains m samples from $\mathbf{p}$, and chooses a channel from $\mathcal{W}$. The mn samples are i.i.d.

POISSONIZED$(\mathcal{W}, n, m)$: For $1 \leq t \leq n$, user t observes M_t samples from $\mathbf{p}$, where $(M_t)_{1\leq t\leq n}$ are independent $\mathrm{Poi}(m)$, and chooses a channel from $\mathcal{W}$. The $\sum_{t=1}^{n} M_t$ samples are i.i.d.

We do not reduce MULTINOMIAL$(\mathcal{W}_\varepsilon, n, m)$ to POISSONIZED$(\mathcal{W}_\varepsilon, n, m)$ as (Acharya et al., 2021a, Lemma C.1) suggests. Instead, we consider the following channel.

Definition D.6. We define the family of channels 'ε-LDP+1bit', denoted as $\mathcal{W}_{\varepsilon,1}$. A channel $W = W_1 \otimes W_2 \in \mathcal{W}_{\varepsilon,1}$ consists of two independent channels such that satisfies the following property given X, each user can send two messages Y_1, Y_2 through two independent channels W_1 and W_2: $Y_1 \in \{0, 1\}$, and Y_2 satisfies LDP constraints.

We have the following lemma:

Lemma D.7. *If there exists a protocol that solves* MULTINOMIAL$(\mathcal{W}_\varepsilon, n, m)$ *with accuracy γ, then there also exists a protocol that solves* POISSONIZED$(\mathcal{W}_{\varepsilon,1}, 20n, 2m)$ *with accuracy $\gamma + e^{-2n/3}$. Moreover, the latter one is non-interactive if the former one is.*

Proof. To design an algorithm that solves POISSONIZED$(\mathcal{W}_{\varepsilon,1}, 20n, 2m)$ with an algorithm for MULTINOMIAL$(\mathcal{W}_\varepsilon, n, m)$, user u first sends a bit $Y_{u,1}$ indicating whether it receives more than m samples. Then, if the user has more than m samples, then it keeps only m samples and sends a message $Y_{u,2}$ according to the ε-LDP protocol for MULTINOMIAL$(\mathcal{W}_\varepsilon, n, m)$. Otherwise, duplicate the existing samples so that the user has m samples, and also send $Y_{u,2}$ according to the ε-LDP protocol. $Y_{u,2}$ obviously satisfies ε-LDP constraints. Hence $Y_u = (Y_{u,1}, Y_{u,2})$ is a valid message from a channel in $\mathcal{W}_{\varepsilon,1}$.

The server keeps the messages such that $Y_{u,1} = 1$, and use the corresponding $Y_{u,2}$ to estimate the underlying distribution.

To bound the accuracy of the above protocol, first note that for $M \sim \mathrm{Poi}(2m)$, we have

$$\Pr[\, M < \mathbb{E}[M]/2 = m \,] \leq e^{-m/6} \leq e^{-1/6}.$$

Therefore, each user receives at least m samples with probability at least $1 - e^{-1/6} > 3/20$. Using Chernoff bound, with probability at least $1 - \delta := 1 - e^{2n/3}$, at least n users has at least m samples. Hence the expected error is at most

$$\gamma(1 - \delta) + \delta \leq \gamma + e^{-2n/3}.$$

□

Next we focus on the Poisonized setting. Similar to Lemma D.8, we can obtain the following lemma.

Lemma D.8. *Under the Poissonized sampling model, if $\Pi, \hat{\mathbf{p}}$ satisfies*

$$\mathbb{E}[d_{\mathrm{TV}}(\widehat{\mathbf{p}}(Y^n), \mathbf{p})] \leq \frac{\gamma}{4},$$

we must have

$$\sum_{i=1}^{k/2} d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{\mathrm{poi},Y^n}, \mathbf{p}_{-i}^{\mathrm{poi},Y^n}\right) = \Omega(k). \tag{20}$$

Following the proof of (Acharya et al., 2021a, Theorem C.7, C.10), we can obtain the following upper bound on the obtained information for the Poissonized problem under $\mathcal{W}_{\varepsilon,1}$.

Lemma D.9. *For any interactive protocol with channels from $\mathcal{W}_{\varepsilon,1}$, when $m > k \log k$, we have there exists a constant C such that*

$$\sum_{i=1}^{k/2} d_{\mathrm{TV}}\left(\mathbf{p}_{+i}^{\mathrm{poi},Y^n}, \mathbf{p}_{-i}^{\mathrm{poi},Y^n}\right) \le C \cdot n\frac{\gamma^2 m}{k} \cdot \left(m\gamma^2 + \max_{z\in\mathcal{Z}} \max_{W\in\mathcal{W}_{\varepsilon,1}} I(\mathbf{p}_z^{\mathrm{poi}}; W)\right).$$

The final ingredient is to prove a mutual information bound for $\mathcal{W}_{\varepsilon,1}$ to apply (Acharya et al., 2021a, Theorem 2)

Lemma D.10. *The mutual information* $\max_{z\in\mathcal{Z}} \max_{W\in\mathcal{W}_{\varepsilon,1}} I(\mathbf{p}_z^{\mathrm{poi}}; W) \le \varepsilon \log_2 e + 1$.

Proof. Let $X \sim \mathbf{p}_z^{\mathrm{poi}}$ and $Y = (Y_1, Y_2)$ be a message sent through a channel in $\mathcal{W}_{\varepsilon,1}$.

$$\begin{aligned}
I(Y_1, Y_2; X) &= \mathbb{E}_X\left[\mathrm{KL}\big(p_{Y|X} \,||\, p_Y\big)\right] \\
&= \mathbb{E}_X\left[\sum_y p_{Y|X}(y) \log \frac{p_{Y|X}(y)}{p_Y(y)}\right] \\
&= \mathbb{E}_X\left[\sum_{y_1} W_1(y_1|X) \sum_{y_2} W_2(y_2|X) \left(\log \frac{W(y_2|X)}{p_Y(y_2|y_1)} + \log \frac{W(y_1|X)}{p_Y(y_1)}\right)\right] \\
&= \mathbb{E}_X\left[\sum_{y_1} W_1(y_1|X) \sum_{y_2} W_2(y_2|X) \log \frac{W(y_2|X)}{p_Y(y_2|y_1)} + \mathrm{KL}\big(p_{Y_1|X} \,||\, p_{Y_1}\big)\right] \\
&\le \varepsilon \log e + I(Y_1; X) \\
&\le \varepsilon \log e + 1
\end{aligned}$$

The second to last inequality is due to LDP constraint on Y_2. The final inequality is due to $I(Y_1; X) \le H(p^{W_1})$ where $p^{W_1} = \mathbb{E}_p[\, W_1(Y_1|X)\,]$. Since $Y_1 \in \{0, 1\}$, the entropy must be at most 1. □

Combining Lemma D.8, Lemma D.9, and Lemma D.10, we have

$$n\frac{m\gamma^2}{k}(\varepsilon + 1 + m\gamma^2)) = \Omega(k),$$

which implies

$$\gamma = \Omega\left(\min\left\{\sqrt{\frac{k^2}{mn\varepsilon}}, \sqrt{\frac{k}{m\sqrt{n}}}\right\}\right) = \Omega\left(\sqrt{\frac{k^2}{mn\varepsilon}}\right)$$

The final equality is due to $n > (k/\varepsilon)^2$. By Lemma D.7 the same bound holds for $\textsc{Multinomial}(\mathcal{W}_\varepsilon, n, m)$ up to constant factors.

E Additional experiment results

E.1 Interactive algorithm

In this section, we present additional experiment results for our interactive algorithms.

High privacy regime $\varepsilon \le 1$ We show an additional result with larger alphabet size ($k = 100$). We can see that our algorithm outperforms 1-sample HR by a large margin, and the error is always within a constant factor of all-sample HR.

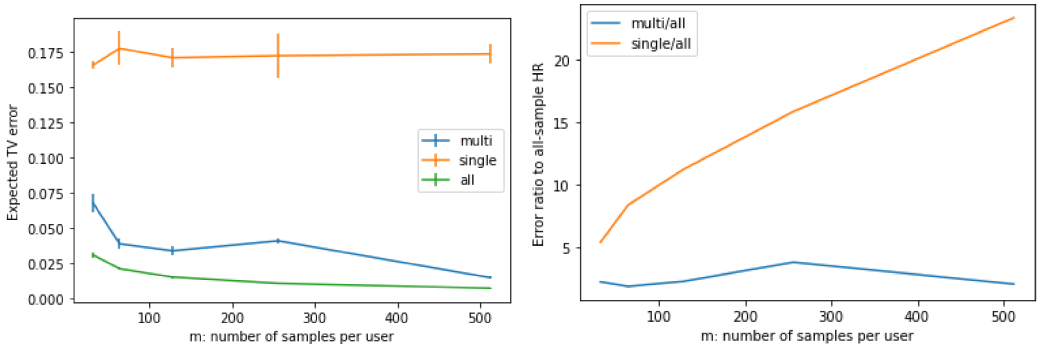


Figure 4: Performance in the high privacy regime with $\varepsilon = 0.9$, $k = 100$, $\mathbf{p}$ uniform. **Left**: expected error of our algorithm (blue), 1-sample HR (orange), and all-sample HR (green). **Right**: orange/green and blue/green ratio in the left plot.

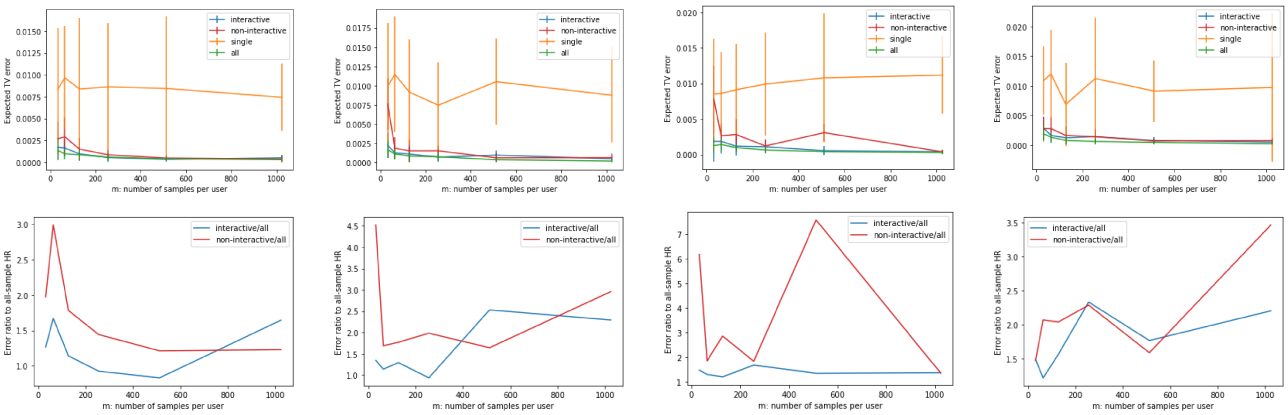
E.2 Non-interactive algorithm

In this section we present experiment results for the non-interactive algorithm. We mainly focus on the case when $k = 2$ (binomial estimation) and the high privacy regime ($\varepsilon = O(1)$) as this is the only part where interactivity is needed in the interactive version of the algorithm for all other regimes. Thus, it is sufficient to demonstrate the difference between the two versions when $k = 2, \varepsilon = O(1)$ since we can substitute this part in other regimes to make them non-interactive as well.

We make some minor changes in our implementation,

1. We choose $C_I = 0.6$ and $C_R = 2.1$, much smaller than the constants used in our proofs.
2. We divide users into 3 groups instead of 4 with $|S_1| = |S_2| = |S_3| = n/3$, dropping the users that send $\mathbb{1}\{Z_u \geq 1\}$. Users in S_1 are used for the localization stage. Users S_2 and S_3 are used in the refinement stage to obtain empirical estimates of R_2 and R_3.

We compare the non-interactive version with the interactive algorithm and the baselines (1-sample HR and all-sample HR). The results are shown in Fig. 5.



$p = 0.1$ $p = 0.2$ $p = 0.3$ $p = 0.4$

Figure 5: Performance of the non-interactive algorithm in the high privacy regime with $k = 2$, $\varepsilon = 0.9, n = 9000$. $m = [32, 64, 128, 256, 512, 1024]$. **Top row**: the expected TV error of non-interactive (orange), interactive (red), 1-sample HR (orange) and all-sample HR (green) with respect to m. Mean and std are reported over 20 independent runs. **Bottom row**: the ratio to all-sample HR of non-interactive (red) and interactive (blue) algorithms w.r.t. m.

Fig. 5 shows that the non-interactive algorithm significantly outperforms the 1-sample HR baseline, and the performance is reasonably close to the interactive version and all-sample HR. This demonstrates the possibility of implementing a non-interactive algorithm that improves with increasing m and matches our theoretical bounds. However, we do observe that

the non-interactive algorithm is less stable and usually performs worse than the interactive one. We view our work mainly as a theoretical investigation of the role of multiple samples in user-level LDP, and the experiments are mainly used to demonstrate algorithmic ideas. We leave optimizing the constants and implementation details to make the algorithm more stable as future work.