The Communication Cost of Security and Privacy in Federated Frequency Estimation

Wei-Ning Chen
Stanford University

Ayfer Özgür
Stanford University

Graham Cormode
Meta AI

Akash Bharadwaj
Meta AI

Abstract

We consider the federated frequency estimation problem, where each user holds a private item X_i from a size-d domain and a server aims to estimate the empirical frequency (i.e., histogram) of n items with $n \ll d$. Without any security and privacy considerations, each user can communicate their item to the server by using $\log d$ bits. A naive application of secure aggregation protocols would, however, require $d \log n$ bits per user. Can we reduce the communication needed for secure aggregation, and does security come with a fundamental cost in communication?

In this paper, we develop an information-theoretic model for secure aggregation that allows us to characterize the fundamental cost of security and privacy in terms of communication. We show that with security (and without privacy) $\Omega\left(n \log d\right)$ bits per user are necessary and sufficient to allow the server to compute the frequency distribution. This is significantly smaller than the $d \log n$ bits per user needed by the naive scheme, but significantly higher than the $\log d$ bits per user needed without security. To achieve differential privacy, we construct a linear scheme based on a noisy sketch that locally perturbs the data and does not require a trusted server (a.k.a. distributed differential privacy). We analyze this scheme under ℓ_2 and ℓ_∞ loss. By using our information-theoretic framework, we show that the scheme achieves the optimal accuracy-privacy trade-off with optimal communication cost, while matching the performance in the centralized case where data is stored in the central server.

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

1 INTRODUCTION

Modern data is increasingly born at the edge and can carry sensitive user information. To make efficient use of this data while protecting individual information from being revealed to the public or service providers, in recent years there has been a strong desire for data science methods that allow servers to collect population-level information from a set of users without knowing each individual value. Consider, for instance, frequency estimation which serves as a fundamental building block for many analytics tasks. Each user holds an item X_i from a size-d domain $\mathcal{X}$, and the server aims to learn the empirical frequency (i.e., the histogram) of all items. Can the server learn the empirical frequency distribution of the items without learning each individual's item?

Recently, distributed protocols based on multi-party computation (MPC) such as secure aggregation (SecAgg) (Bonawitz et al., 2016) have emerged as a powerful tool to securely aggregate population-level information from a set of users. In particular, SecAgg allows a single server to compute the population sum (and hence also the average) of local variables (often vectors), while also ensuring no additional information, other than the sum, is released to the server or other participating entities. This can be achieved, for example, by having users apply additive masks on their local vectors which cancel out upon addition at the server. SecAgg is widely used within protocols for secure federated learning and secure statistics gathering, which both rely on vector summation.

A straightforward way to use SecAgg for the empirical frequency estimation problem above is to have each user represent their item X_i as a d-dimensional one-hot vector (i.e., a vector with a single 1 in the X_i-th coordinate and zero otherwise), so that the sum of all one-hot vectors (which is revealed to the server by SecAgg) gives the desired histogram. However, this requires $d \log n$ bits of communication per user since each user has to communicate a masked vector of dimension d (with each entry taking values in a finite field of size n). This is a drastic increase from the $\log d$ bits per user needed to communicate each item in the absence of any security considerations.

Can we reduce the communication cost of secure aggregation, and does security come with a fundamental cost in communication? This is the main question we investigate in this paper. We show:

- The communication cost for secure frequency estimation can be reduced from $O(d \log n)$ to $O(n \log d)$ when $n \ll d$. This is the relevant regime in many real-world applications (e.g., location tracking (Bagdasaryan et al., 2021), language modeling, web-browsing, etc.) where d can be very large compared to the number of participating users n (even in cross-device FL/FA settings) and computational constraints limit the number of users that can participate in each SecAgg round.
- Complementarily, any aggregation protocol that is information-theoretically secure needs $\Omega(n \log d)$ bits per user to perfectly recover the histogram. To show this we develop an information-theoretic model for secure aggregation and prove a lower bound on the communication cost of any secure aggregation protocol.

This reveals that while the communication cost of secure frequency estimation can be reduced with more carefully designed schemes (e.g., we show that one can formulate it as an ℓ_1 constrained integer linear inverse problem), there is a fundamental price to computing the histogram *securely*: in the absence of any security considerations, each user needs $\log d$ bits, and hence the total communication cost for *all* users is $n \log d$; with security, each user *individually* incurs the $n \log d$ bits communication cost (i.e., $\Theta(n^2 \log d)$ in total).

Secure aggregation alone does not provide any provable differential privacy (DP) guarantees (Dwork et al., 2006b). Sensitive information may still be revealed from the aggregated population statistics, causing potential privacy leakage. To address this issue, a common approach is to perturb the aggregated information by adding noise before passing it to downstream analytic tasks. With a privacy requirement, the empirical frequency can be estimated only approximately, with an amount of distortion that depends on the privacy level, number of participating users, and the loss function. This distortion due to DP also allows for some 'slack' in the secure aggregation framework – as long as secure aggregation returns an approximate sum with a distortion small enough compared to the distortion due to DP, we can achieve order-wise the same performance as with only the DP constraint. This observation leads us to study the communication cost of secure aggregation for computing an *approximate* sum rather than an exact sum of the user values. We show that computing an approximate sum requires less communication, and the optimal communication cost can be characterized by a rate-distortion function that depends on the error and the loss function. While security drastically increases the communication cost, we show that privacy helps us reduce it.

Our end goal is to arrive at secure and private frequency estimation protocols that provide differential privacy guarantees without putting trust in the service provider, while at the same time achieving the optimal privacy-accuracy-communication trade-off. To this end, we develop a *user-level* DP protocol for frequency estimation, where users compute a summary of their local data, perturb these slightly, and employ SecAgg to simulate some of the benefits of a trusted central party. The untrusted server has access only to the aggregated reports with the aggregated perturbations. We show that the end-to-end privacy-accuracy trade-off achieved by this scheme is optimal and matches the trade-off achievable with a trusted server, i.e., in a centralized setting where the server receives the data as it is and perturbs it after aggregation. Furthermore, by using our aforementioned information-theoretic framework for securely computing an approximate sum, we show that the communication cost of this scheme is also optimal.

Our contributions. The main contributions of our paper can be summarized as follows:

- We provide an information-theoretic view on secure aggregation and analyze the amount of communication needed for securely computing the sum either exactly or approximately. In the case of exact recovery, we show that the per-user communication cost is lower bounded by the entropy of the sum; for approximate recovery under a general loss function $\ell(\cdot)$, we specify the communication-distortion trade-offs.
- We specialize these information-theoretic lower bounds to frequency estimation with and without differential privacy constraints. We show that without privacy $\Omega(n \log d)$ bits per user are needed to allow the server to learn the exact histogram. We also characterize the minimal communication cost when differential privacy is required.
- We introduce schemes that match the above information-theoretic communication lower bounds. In particular, we show that to perfectly recover the exact histogram (without privacy), one can achieve the optimal $O(n \log d)$ bits per-user communication by applying SecAgg and solving a linear inverse problem. To achieve differential privacy, we construct a linear scheme based on noisy sketch (with proper modifications tailored to the specific loss function) which locally perturbs the data and does not require a trusted server (a.k.a. user-level DP). We show that this scheme achieves the (nearly) optimal accuracy-privacy trade-off with optimal communication cost, while matching the performance in the centralized case where data is stored in the central server.

Organization. The rest of the paper is organized as follows. We discuss the related works in Section 2. In Section 3, we introduce a general framework for SecAgg and the corresponding information-theoretic security it provides and proves general communication lower bounds on computing the exact or approximate sum. In Section 4, we apply SecAgg to the frequency estimation problem and specify the optimal communication cost. Finally, in Section 5, we incorporate the differential privacy constraint and characterize the optimal privacy-communication-accuracy trade-offs.

Notation. Throughout this paper, we use $[m]$ to denote the set of $\{1, ..., m\}$ for any $m \in \mathbb{N}$. Random variables (vectors) $(X_1, ..., X_m)$ are denoted as $X_{[m]}$ or X^m. We also make use of Bachmann-Landau asymptotic notation, i.e., $O, o, \Omega, \omega,$ and Θ. We use $H(X)$ (or $H(P_X)$) to denote the Shannon entropy of X with base 2. Finally, for random variables X, Y, $I(X;Y)$ denotes the mutual information, i.e., $I(X;Y) \triangleq \mathbb{E}_X\left[D_{\mathsf{KL}}\left(P_{Y|X} \| P_Y\right)\right]$.

2 RELATED WORK

Secure aggregation. Single-server SecAgg is a cryptographic secure multi-party computation (MPC) that enables users to submit vector inputs, such that the server learns just the sum of the users' vectors. This is usually achieved via additive masking over a finite group (Bonawitz et al., 2016; Bell et al., 2020). The single-server setup makes SecAgg particularly suitable for federated learning (Kairouz et al., 2021; Agarwal et al., 2021) or federated analytics (Choi et al., 2020b), and a recent line of works (Jahani-Nezhad et al., 2022; So et al., 2021; Choi et al., 2020a; Kadhe et al., 2020; Yang et al., 2021) aim to scale it up by improving the communication or computation overhead. However, all of the above works focus on a general setting where the local vectors can be arbitrary; meanwhile, in the frequency estimation problem with a large domain size, local vectors are one-hot and the histogram is typically sparse. Without secure aggregation such sparsity can be leveraged to reduce the communication cost (Acharya et al., 2019b; Han et al., 2018; Barnes et al., 2019; Acharya et al., 2019a, 2020, 2021b; Chen et al., 2021a,b). However, with secure aggregation, it is not clear if and how sparsity can be leveraged to reduce communication, which is the main focus of our work.

Differential Privacy. To achieve provable differential privacy guarantees SecAgg is insufficient[1] as even the sum of local model updates may still leak sensitive information (Melis et al., 2019; Song and Shmatikov, 2019; Carlini et al., 2019; Shokri et al., 2017) and so differential privacy (DP) (Dwork et al., 2006a) can be adopted. By having the noise added locally and letting the server aggregate local information via SecAgg, the DP guarantees do not rely on users' trust in the server. This *user-level DP* (also referred to as distributed DP in the literature) framework has recently been adopted in private federated learning (Agarwal et al., 2018; Kairouz et al., 2021; Agarwal et al., 2021; Chen et al., 2022a). In this work, we use the Poisson-binomial mechanism as a primitive (Chen et al., 2022b) to achieve user-level DP.

We also distinguish our setup from the local DP setting (Kasiviswanathan et al., 2011; Evfimievski et al., 2004; Warner, 1965), where the data is perturbed on the user-side before it is collected by the server. Local DP, which allows for a possibly malicious server, is stronger than distributed DP, which assumes an honest-but-curious server. Consequently, local DP suffers from worse privacy-utility trade-offs (Duchi et al., 2013; Ye and Barg, 2017; Barnes et al., 2020; Acharya et al., 2021a).

SecAgg can be viewed as a privacy amplification technique that amplifies weak local DP to much stronger central DP guarantees. Other amplification techniques are based on different cryptographic techniques such as secure shuffling (Erlingsson et al., 2019; Balle et al., 2019, 2020; Balcer and Cheu, 2019) or distributed point functions (Gilboa and Ishai, 2014). While the fundamental communication cost for SecAgg that we show in our paper can be potentially circumvented by these methods, these techniques either require the existence of a trusted shuffler or assume multiple servers that do not collude.

Private frequency estimation and heavy hitters. Private frequency estimation, a.k.a. histogram estimation, is a canonical task that has been heavily studied in the DP literature (Dwork et al., 2006b). When subject to ℓ_∞ loss, it is the same as the heavy hitter problem. Under the centralized setting, typical techniques for releasing a private histogram include the addition of noise (and thresholding the counts) (Dwork et al., 2006b; Ghosh et al., 2012; Korolova et al., 2009; Bun and Steinke, 2016; Balcer and Vadhan, 2017) or sampling-and-thresholding (Zhu et al., 2020; Cormode and Bharadwaj, 2022). The private heavy hitter problem has also been heavily studied under the local or multiparty DP model (Bassily and Smith, 2015; Bassily et al., 2017; Bun et al., 2018, 2019; Huang et al., 2022) (which can also be used to obtain a central DP guarantee when combined with a secure shuffling (Erlingsson et al., 2019; Cheu et al., 2019; Ghazi et al., 2021; Niu et al., 2011; Girgis et al., 2021; Luo et al., 2022)). Our work, however, is under the user-level DP model, under which most previous techniques cannot be directly applied. Our privatization technique makes use of noisy count-sketch, which is close to the work of Choi et al. (2020b), in which a general distributed noisy sketch framework is analyzed. In this work, we use a similar technique to characterize the exact communication cost and show that a noisy sketch can achieve the communication lower bound.

[1]Arguably, SecAgg may be private under different notions of privacy (e.g., Elkordy et al. (2022)). However, we only focus on DP in this work as it is typically considered the gold standard in both theory and practice.

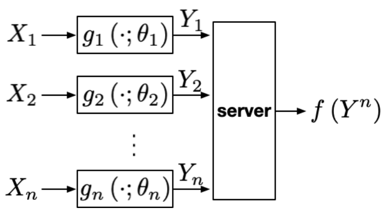

Figure 1: A framework for SecAgg (illustrated for the case without dropouts).

The private frequency estimation task can also be related to private set union (Frikken, 2007) or set operations (Kissner and Song, 2005), where the goal is to discover the union of the supports of the users' vectors. Though one can try to first estimate the support of the frequency vector and then its distribution over the support via a two-step approach, we note that the resulting accuracy will be highly sub-optimal.

3 SECURE AGGREGATION

In this section, we formulate a general framework for secure aggregation with a single server and n users. Assume each user $i \in [n]$ holds local information (as a vector) $X_i \in \mathcal{X}$, and the server aims to compute the sum $\mu(X_1, X_2, ..., X_n) \triangleq \sum_{i\in[n]} X_i$. During the aggregation, up to D clients may drop out, and the secure aggregation protocol should still be able to recover the sum of the remaining clients. In general, an aggregation protocol consists of encoding functions g_i, $i \in [n]$ at the users and an aggregation function f at the server such that:

1. Each user encodes their local information X_i as $Y_i = g_i(X_i;\theta_i)$, where θ_i is randomness available at the i'th user which is independent of X_i but may depend on other $\theta_{i'}$ for $i' \in [n] \setminus i$.
2. The server observes Y_i for $i \in [n] \setminus \mathcal{D}$, i.e., the messages of the available users. If $\mathcal{D} = \emptyset$, i.e. there are no dropouts, it estimates the sum $\mu(X^n)$ by a (deterministic) function $f(Y^n)$.
3. If $\mathcal{D} \neq \emptyset$, the server invokes a second round of communication with the surviving clients to recover the masks of the dropout users. In this round, the server collects $h\left(\theta_{[n]}, \mathcal{D}\right)$, where $h(\cdot)$ is a general function of local secrets $\theta_{[n]}$. Using the information it collects over the two rounds, $Y_{[n]\setminus\mathcal{D}}$ and $h\left(\theta_{[n]}, \mathcal{D}\right)$, the server estimates the sum of the surving clients $\mu\left(X_{[n]\setminus\mathcal{D}}\right)$.

Security constraints: We call the aggregation protocol that can tolerate D drop-outs *secure*, if it satisfies the following two conditions on mutual information for any distribution P_{X^n} imposed on the user data:

$$\forall \mathcal{D} \subseteq [n], I\left(Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right); X_{[n]\setminus\mathcal{D}} \middle| \mu\left(X_{[n]\setminus\mathcal{D}}\right)\right) = 0, \tag{S1}$$

$$\forall |\mathcal{D}| > D,\ I\left(Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right); X_{[n]\setminus\mathcal{D}}\right) = 0. \tag{S2}$$

(S1) implies that $X_{[n]\setminus\mathcal{D}} - \mu\left(X_{[n]\setminus\mathcal{D}}\right) - (Y_{[n]}, h)$ forms a Markov chain, and hence given $\mu\left(X_{[n]\setminus\mathcal{D}}\right)$ the server cannot deduce any further information about $X_{[n]\setminus\mathcal{D}}$ from the information it gathers over the two stages of the scheme, $Y_{[n]\setminus\mathcal{D}}$ and $h\left(\theta_{[n]}, \mathcal{D}\right)$; (S2) states that without a sufficient number of users participating (e.g., when $|\mathcal{D}| \geq D$), the server cannot learn any information about the user data.

Note that the same framework can be used to include colluding users by allowing $h(\cdot)$ to contain information about both the masks θ_D and the information $X_{\mathcal{D}}$ of the users in $\mathcal{D}$, i.e. $h\left(\theta_{[n]}, X_{\mathcal{D}}, \mathcal{D}\right)$. Security for the remaining users is ensured with the same constraints (S1) and (S2).

The two security requirements above are satisfied by most practical secure aggregation protocols such as Bonawitz et al. (2016) and Bell et al. (2020). In the next section, we show that these security requirements come at a fundamental and significant communication cost.

Correctness constraints: In the absence of any privacy considerations, we impose the following correctness requirement on the protocol, which ensures that it always outputs the correct sum:

$$\forall |\mathcal{D}| \leq D,\ \mathbb{P}\left\{f\left(Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right)\right) = \mu\left(X_{[n]\setminus\mathcal{D}}\right)\right\} = 1. \tag{C1}$$

We are also interested in the case where the server recovers the sum approximately under a certain loss function (this is the relevant setting under differential privacy constraints). Let $\ell(\cdot,\cdot)$ be a loss function defined on the domain of $\mu\left(X^n\right) = \sum_{i=1}^n X_i$. We refer to the following approximate recovery criterion as the β-distortion criterion:

$$\forall |\mathcal{D}| \leq D,\ \mathbb{E}\left[\ell\left(f\left(Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right)\right), \mu\left(X_{[n]\setminus\mathcal{D}}\right)\right)\right] \leq \beta. \tag{C1$'$}$$

Note that under this criterion the server recovers the sum with distortion β under the loss function ℓ.

Communication cost: The communication cost of an aggregation protocol for user i is given by $\max_{P_{X^n}} H(Y_i)$ (i.e., the worst-case entropy for any possible joint distributions over the local data). This is the (maximum over the choice of X^n) number of bits node i needs on average to communicate Y_i using an optimal compression scheme.

3.1 Communication Lower Bounds

Next, we present general communication lower bounds on estimating the sum of n random variables $X_{[n]}$ (where we do not make any assumptions on the domain of X_i) under the security constraints (S1) and (S2).

Lemma 3.1 (Lower bound for perfect recovery) *Let $\mathcal{D} \subset [n]$ be the set of dropout clients, such that $|\mathcal{D}| \leq D$ for some $D \leq \frac{n}{2}$. Under the correctness constraint* (C1) *and security constraints* (S1) *and* (S2) *on the protocol, it holds that for all $i \in [n] \setminus \mathcal{D}$, $H(Y_i) \geq H\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i\right)$,* *where $H(\cdot)$ is the Shannon entropy.*

Note that $H\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i\right)$ quantifies the information the server is able to learn about the user data. The lemma states that in a secure protocol, the entropy of each individual message should be at least as large the total information communicated to the server. In the following lemma, we characterize how this lower bound is modified when the server needs to recover the sum only approximately.

Lemma 3.2 *Let $n' \triangleq n - |\mathcal{D}|$. Let $\ell(\cdot,\cdot)$ be a loss function defined on the domain of $\mu\left(X^{n'}\right) = \sum_{i\in[n]\setminus\mathcal{D}} X_i$. Under the β-approximate recovery criterion* (C1′) *and the security constraints* (S1) *and* (S2)*, it holds that*

$$H(Y_i) \geq R(\beta), \text{ for all } i \in [n],$$

where $R(\beta)$ is the solution of the following rate-distortion problem:

$$R(\beta) \triangleq \begin{pmatrix} \min & I\left(Y^{n'}; \mu\left(X^{n'}\right)\right) \\ \textit{s.t.} & \min_{\hat{\mu}} \mathbb{E}\left[\ell\left(\hat{\mu}\left(Y^{n'}\right), \mu\left(X^{n'}\right)\right)\right] \leq \beta \end{pmatrix} \tag{1}$$

where the first minimization is taken over all conditional probability $P_{Y^{n'}|\mu(X^{n'})}$.

Lemma 3.2 suggests that under the β-approximate recovery criterion, the communication load of a secure aggregation protocol is lower-bounded by $R(\beta)$ per user. In Section 5, we explicitly characterize $R(\beta)$ for the frequency estimation problem.

4 SECURE FREQUENCY ESTIMATION

In this section, we formally define the frequency estimation problem with security constraints (S1) and (S2) and study the optimal communication cost. Assume each user i holds an item X_i in a size d domain $\mathcal{X}$ and the server aims to estimate the histogram of the n items. Let $X_i \in \mathcal{X} \triangleq \{e_1, ..., e_d\} \in \{0,1\}^d$, i.e., each item is expressed as a one-hot vector. Note that this is without loss of generality since the encoding functions g_i at the users can be arbitrary. Then, the histogram of the n items can be expressed as $\mu\left(X^n\right) \triangleq \sum_{i\in[n]} X_i \in [n]^d$. We mainly focus on the high-dimensional regime where $d \gg n$, and our goal is to characterize the communication needed to securely compute $\mu(X^n)$.

To this end, we first apply the (general) lower bound derived in Section 3.1 with $X_i \in \{e_1, e_2, ..., e_d\}$. For simplicity, we present our results without dropouts (i.e., $\mathcal{D} = \emptyset$), but extending to the $|\mathcal{D}| > 0$ case is immediate. Our lower bound is obtained by imposing a worst-case prior distribution on X^n we arrive at the following corollary:

Corollary 4.1 *Let $X_i \in \{e_1, ..., e_d\}$ for $i \in [n]$. Under the same set of constraints as in Lemma 3.1, there exists a worst-case prior distribution π_{X^n} such that*

$$H(Y_i) \geq H\left(\textstyle\sum_{i=1}^n X_i\right) = \Omega\left(n \log d\right), \tag{2}$$

where the entropy $H\left(\sum_{i=1}^n X_i\right)$ is computed with respect to $X^n \sim \pi_{X^n}$.

In the rest of this section, we outline a communication-efficient secure frequency estimation scheme based on solving a linear inverse problem, and the resulting per-user communication cost matches the lower bound in the corollary. We state this result, together with the lower bound in Corollary 4.1, as our main theorem:

Theorem 4.1 *To securely (i.e., under* (S1) *and* (S2)*) and correctly (i.e., under* (C1)*) compute the histogram from n users, it is both sufficient and necessary for each user to send $\Theta\left(n \log d\right)$ bits to the server.*

4.1 Reducing Communication via Sparse Recovery

In this section, we propose a scheme that shows that the communication cost can be reduced to the information-theoretic $\Omega\left(n \log d\right)$ bits lower bound. Our scheme depends on two main ingredients: (1) a specific construction of a secure aggregation protocol, often called SecAgg, due to Bonawitz et al. (2016), and (2) a linear binary compression scheme based on random coding. For simplicity, we describe our schemes for the case of no dropouts, but our schemes can be readily extended to handle dropouts or colluding users since they are based on SecAgg (which is designed to tolerate dropouts/colluding users).

In a nutshell, the encoding steps of SecAgg (Bonawitz et al., 2016) consist of (i) mapping X_i into an element of a finite group (where, without loss of generality, we assume the group is $\mathbb{Z}_M^m$ for some $m, M \in \mathbb{N}$), and then (ii) adding a random mask $\theta_i \in \mathbb{Z}_M^m$ so that $Y_i = \mathcal{A}_{\mathsf{enc}}(X_i) + \theta_i$. The mask θ_i has uniform marginal density, is independent of X^n, and satisfies $\sum_{i\in[n]} \theta_i = 0$. Upon receipt of Y^n, the server computes the sum of Y^n and decodes it via $\mathcal{A}_{\mathsf{dec}}\left(\sum_i \mathcal{A}_{\mathsf{enc}}(X_i)\right)$. The goal is to design mappings $(\mathcal{A}_{\mathsf{enc}}, \mathcal{A}_{\mathsf{dec}})$, so that

- the outcome correctly recovers $\mu\left(X^n\right)$, i.e., $\mathcal{A}_{\mathsf{dec}}\left(\sum_i \mathcal{A}_{\mathsf{enc}}(X_i)\right) = \sum_{i=1}^n X_i$;
- the per-user communication cost $m \log M$ is minimized.

Due to the linearity of SecAgg (i.e., the server obtains the sum of $\mathcal{A}_{\mathsf{enc}}$), $\mathcal{A}_{\mathsf{enc}}$ is usually constructed via a linear mapping, so that $\mathcal{A}_{\mathsf{enc}}(X_i) \triangleq S \cdot X_i$ for some $S \in \left(\mathbb{Z}_M\right)^{m\times d}$. In this case, the sum of the encodings is the same as the encoding of the sum, i.e.,

$$\textstyle\sum_i \mathcal{A}_{\mathsf{enc}}\left(X_i\right) = \mathcal{A}_{\mathsf{enc}}\left(\sum_i X_i\right) = S\mu\left(X^n\right). \tag{3}$$

To recover μ from $S\mu$, the server solves a linear inverse problem, which has a unique solution only if S is "invertible" for all possible μ's. For example, a naive choice of S can be the identity mapping I_d, which encodes each X_i as a one-hot vector. In this case, the size of the finite group $\mathbb{Z}_M^m$ is $(M, m) = (n, d)$, and the communication complexity is $d \log n$ bits. This is far from the lower bound $\Omega\left(n \log d\right)$ when $n \ll d$.

Can we design a better embedding matrix S with smaller range (i.e., with smaller (M, m)) than the naive choice I_d so that $y = S\mu$ is solvable? Specifically, define $\mathcal{H}_n$ to be the collection of all possible n-histogram, i.e., $\mathcal{H}_n \triangleq \left\{\mu \in \mathbb{Z}_+^d \middle| \|\mu\|_1 = n\right\}$. Our goal is to show that there exists an $S \in \{0,1\}^{m \times d}$ with $m = O\left(n \log d / \log n\right)$, such that $y = S\mu$ is solvable for all $\mu \in \mathcal{H}_n$. Using such S as our local embedding, the resulting communication cost becomes $O(n \log d)$ and hence matches the lower bound. We summarize this in the following theorem

Theorem 4.2 *Let $\mathcal{H}_n$ be the collection of all valid n-histograms formally defined as above. Then there exists an embedding matrix $S \in \{0,1\}^{m \times d}$ with $m = O\left(\frac{n \log d}{\log n}\right)$, such that*

$$\forall \mu_1, \mu_2 \in \mathcal{H}_n,\, \mu_1 \neq \mu_2 \Longrightarrow S\mu_1 \neq S\mu_2. \qquad (4)$$

Theorem 4.2 can be viewed as a generalization of classical (non-adaptive) Quantative Group Testing (QGT) (Bshouty, 2009; Wang et al., 2016; Scarlett and Cevher, 2017; Gebhard et al., 2019), in which the linear inverse problem is defined over the ℓ_1 constrained *binary* vectors $\mathcal{G}_n \triangleq \left\{\nu \in \{0,1\}^d \,\middle|\, \|\nu\|_0 = n\right\}$. To prove the existence of such S, we follow the idea of Wang et al. (2016) by constructing S in a probabilistic way, i.e., generating each element of S as an independent $\mathsf{Bern}(1/2)$ random variable. We then show that as long as $m = \Omega\left(n \log d / \log n\right)$, (4) holds with high probability, hence concluding the existence of S. One key step that generalizes the result from classical QGT is an application of Sperner's theorem (Sperner, 1928; Lubell, 1966), which may be of independent interest. The proof of Theorem 5.1 can be found in Appendix D.1.

Comparison to compressed sensing. Note that as the set of n-histograms is a subset of n-sparse vectors in $\mathbb{R}^d$, it may be tempting to use standard sparse recovery techniques such as compressed sensing (Donoho, 2006a,b) (e.g., with the classical Rademacher ensemble, see (Wainwright, 2019, Chapter 7)). This can allow us to reduce the dimensionality from d to $m = O(n \log d)$. However, each coordinate of the embedded vector can range from $-n$ to n (using the Rademacher ensemble), and requires $O(\log n)$ bits to represent it and the total communication cost is $O(n \log d \cdot \log n)$ leading to an extra $\log n$ factor. Theorem 4.2 is necessary in order to obtain a information-theoretically optimal solution.

On the other hand, the scheme proposed in the proof of Theorem 4.2, though optimal in terms of communication efficiency, is computationally infeasible. It requires exhaustively scanning over $\mathcal{H}_n$ to find the unique consistent histogram μ^*, and hence the computation cost is $\Omega\left(d^n\right)$. It remains open if one can design computationally efficient schemes (e.g., a scheme with computational cost $\mathsf{poly}(n, d)$ or ideally $\mathsf{poly}(n, \log d)$) that achieves the best $O(n \log d)$ communication cost.

5 SECURE AND PRIVATE FREQUENCY ESTIMATION

Secure aggregation alone does not provide any differential privacy guarantees. In this section, we study the private frequency estimation problem, in which, apart from security constraints (S1) and (S2), we also impose a privacy constraint on our protocol. Our goal is to characterize the communication required for the optimal accuracy-privacy tradeoff. We first state the definition of differential privacy (Dwork et al., 2006b).

Definition 5.1 (Differential Privacy (DP)) *For $\varepsilon, \delta \geq 0$, a randomized mechanism M satisfies (ε, δ)-DP if for all neighboring datasets D, D' and all $\mathcal{S}$ in the range of M, we have that*

$$\mathbb{P}\left(M(D) \in \mathcal{S}\right) \leq e^{\varepsilon} \mathbb{P}\left(M(D') \in \mathcal{S}\right) + \delta,$$

where $D = (X_1, ..., X_n)$ and $D' = (X_1, ..., X_i', ..., X_n)$ are neighboring pairs that can be obtained from each other by adding or removing all the records that belong to a particular user.

In our frequency estimation setting (see Figure 1), DP can be achieved in two different ways:

- Central-level DP criterion: $f(Y^n)$ is (ε, δ)-DP.
- User-level DP criterion: $(Y_1, \ldots, Y_n)$ is (ε, δ)-DP.

The central DP criterion requires the server to apply a DP mechanism to its computation to obtain a privatized estimate $f(Y^n)$, and hence puts trust in the service provider. The user-level DP criterion removes the need for a trusted server as noise is added to each message before it is sent to the server. By the data processing property of DP, the latter is a stronger notion and implies the former.

In this section, we provide a secure and private frequency estimation scheme that satisfies the user-level DP criterion above in addition to (S1) and (S2). The scheme consists of local perturbations, where local data is privatized by local randomizers, and secure aggregation, where the server aggregates the noisy sum via SecAgg. The next lemma states that we can achieve user-level DP if the (locally privatized) noisy sum is DP and the aggregation protocol satisfies the security condition (S1). This fact has been implicitly used in previous distributed DP works, dating all the way back to Dwork et al. (2006a). We provide an explicit statement in the following lemma.

Lemma 5.1 *Let $\mathcal{M}_i$ be the local randomizer at user $i \in [n]$ such that $\hat{\mu} = \sum_i \mathcal{M}_i(X_i)$ is (ε, δ)-DP. Let Y_i be the message sent by client i in the secure aggregation protocol with input $\mathcal{M}_i(X_i)$ (i.e., $Y_i = g_i(\mathcal{M}_i; \theta_i)$ in Figure 1). Then as long as the security constraint* (S1) *holds, $(Y_1, ..., Y_n)$ satisfies (ε, δ)-DP.*

Lemma 5.1 implies that with secure aggregation, we only need to ensure that the *sum* of locally randomized messages $\mathcal{M}_i$'s are DP (as opposed to requiring $(\mathcal{M}_1, ..., \mathcal{M}_n)$ to be jointly DP). This not only simplifies the construction of $\mathcal{M}_i$'s, but also significantly reduces the amount of perturbation needed.

In the rest of this section, we characterize the accuracy-privacy trade-off achieved by this scheme as well as its per-user communication cost in Theorem 5.1. Since this scheme satisfies (ε, δ)-user-level DP, it also satisfies the weaker (ε, δ)-central DP criterion. Moreover, the accuracy-privacy trade-off achieved by this scheme is (nearly) optimal in the sense that it (nearly) matches the best trade-off achievable by any scheme satisfying the central DP criterion (Balcer and Vadhan, 2017). Since our scheme is designed to satisfy the stronger user-level DP criterion, this means that we can achieve the optimal privacy-accuracy trade-off while removing the need for a trusted server. We show that the communication cost is also optimal by proving a lower bound on the communication cost of any scheme that achieves the optimal privacy-accuracy trade-off while satisfying (S1) and (S2). In other words, any secure frequency estimation scheme requires at least as many bits to achieve the optimal privacy-accuracy trade-off. This establishes the optimality of our scheme in terms of communication cost. Finally, we remark that although we present bounds in terms of standard DP (Definition 5.1), our scheme also satisfies Rényi differential privacy (Mironov, 2017) (RDP), which allows for tighter privacy accounting when applying a private mechanism iteratively. We defer the details to Appendix A.

We next state the main results of this section starting with our achievability result.

Theorem 5.1 (Private frequency estimation) *The scheme presented in Section 5.1 (see also Algorithm 1) satisfies* (S1), (S2) *and an (ε, δ)-user-level DP criterion (and also $\left(\left(\alpha, \varepsilon/\log\left(\frac{1}{\delta}\right)\right)\right)$-RDP), while achieving*

- *ℓ_∞ error $\mathbb{E}\left[\|\hat{\mu} - \mu\left(X^n\right)\|_\infty\right] = O\left(\frac{\sqrt{\log d \log(1/\delta)}}{\varepsilon}\right)$;*
- *ℓ_1 error $O\left(\frac{n\sqrt{\log(d)\log(1/\delta)}}{\varepsilon}\right)$;*
- *ℓ_2 error $\mathbb{E}\left[\|\hat{\mu} - \mu\left(X^n\right)\|_2^2\right] = O\left(\frac{n \log d \log(1/\delta)}{\varepsilon^2}\right)$;*

and uses $\tilde{O}\left(n \min\left(\varepsilon\sqrt{\log d/\log(1/\delta)}, \log d\right)\right)$ bits (where in $\tilde{O}$ we hide dependency on $\log n$ and $\log\log d$ terms).

The formal statement of Theorem 5.1 and the proof are provided in Appendix A (see Theorem A.3). Note that the (ε, δ)-user-level DP guarantee in the theorem implies a (ε, δ)-central DP guarantee. We contrast this with the optimal accuracy-privacy tradeoff achievable in the centralized case, i.e., when the only requirement imposed on the scheme is an (ε, δ)-central DP criterion. For the ℓ_2 and ℓ_∞ loss (i.e., setting the loss function in (C1′) to be $\|\cdot\|_2$ or $\|\cdot\|_\infty$ respectively), the minimax error is well-known (see, for instance, (Hardt and Talwar, 2010; Balcer and Vadhan, 2017)) as we state in the following lemma:

Lemma 5.2 (Minimax error under central DP) *Under a (ε, δ)-central DP, the minimax error for frequency estimation, defined as*

$$\min_{M(\cdot) \text{ satisfies } (\varepsilon,\,\delta)\text{-central DP}} \max_{X^n} \mathbb{E}\left[\ell\left(M\left(X^n\right), \mu\left(X^n\right)\right)\right],$$

is equal to

- *$\Theta\left(\frac{\min(\log d, \log(1/\delta))}{\varepsilon}\right)$ under the ℓ_∞ loss;*
- *$O\left(\frac{n \log d \log(1/\delta)}{\varepsilon^2}\right)$ under the ℓ_2^2 loss;*

We note that the the ℓ_∞ accuracy results in Theorem 5.1 matches that in Lemma 5.2 up to a $\max\left(\sqrt{\frac{\log d}{\log(1/\delta)}}, \sqrt{\frac{\log(1/\delta)}{\log d}}\right)$ factor, while the scheme in Theorem 5.1 satisfies the additional (S1), (S2) and the stronger user-level-DP constraints. This establishes the optimality of our scheme from an accuracy-privacy trade-off perspective. We also observe that the communication cost in Theorem 5.1 *decreases* with ε when $\varepsilon \leq \log d$, meaning that we can compress more aggressively with more stringent privacy constraint. This behavior aligns with the conclusions of Chen et al. (2022a) (under a federated learning setting) and Chen et al. (2020) (under a local DP model). We next show that the communication cost in Theorem 5.1 is optimal under the ℓ_∞ loss (up to a $\mathsf{poly}\left(\log n, \log\log d\right)$ factor).

Corollary 5.1 *Any (ε, δ)-central DP scheme that satisfies* (S1) *and* (S2) *such that:*

- *$\mathbb{E}\left[\|\hat{\mu} - \mu\left(X^n\right)\|_\infty\right] = O\left(\frac{\sqrt{\log d \log(1/\delta)}}{\varepsilon}\right)$ requires $\Omega\left(n \min\left(\varepsilon\sqrt{\log d/\log(1/\delta)}, \log d\right)\right)$ per-user communication;*
- *$\mathbb{E}\left[\|\hat{\mu} - \mu\left(X^n\right)\|_2^2\right] = O\left(\frac{n \log d \log(1/\delta)}{\varepsilon^2}\right)$ requires $\Omega\left(n \min\left(\frac{\varepsilon}{\log(1/\delta)}, \log d\right)\right)$ per-user communication.*

Recall from the previous section that we need $n \log d$ bits to securely compute the exact histogram. Corollary 5.1 characterizes the reduction in communication cost when the histogram is computed approximately due to the privacy constraint and $\varepsilon = O(\log d)$.

We establish this result as a corollary of the following lemma which specifies the (worst-case) asymptotic behavior of $R(\beta)$ (defined in (1)) under ℓ_2 and ℓ_∞ loss for secure frequency estimation. Corollary 5.1 is obtained by plugging in the corresponding errors for β in the lemma.

Lemma 5.3 *Let $R(\beta)$ be defined as in* (1). *When $X_i \in \{e_1, ..., e_d\}$ for $i \in [n]$, there is a worst-case prior distribution π_{X^n} (possibly correlated for X_i's), s.t.*

- *under the ℓ_∞ loss , $R(\beta) = O(n \log d/\beta)$;*
- *under the ℓ_2 loss, $R(\beta) = O\left(n^2 \log d/\beta\right)$.*

Lemma 5.3 is a special case of Lemma 3.2. However, to obtain the asymptotic scaling, we make use of Fano's inequality, with carefully constructed prior distributions via ℓ_∞ and ℓ_2 packing over the space of all histograms. The proof can be found in Appendix D.5

5.1 Frequency Estimation via Noisy Sketch

Next, we present a (nearly) optimal secure and private frequency estimation scheme in Algorithm 1 that uses the optimal communication in Corollary 5.1. We use the following ingredients in our scheme: (1) the specific SecAgg implementation of Bonawitz et al. (2016) (see Section 4.1 for a brief introduction), (2) count-sketch (Charikar et al., 2002) together with Hadamard transform, and (3) the Poisson-binomial mechanism (Chen et al., 2022b). Following the idea in Section 4.1, we use the SecAgg protocol introduced by Bonawitz et al. (2016) as a primitive and focus on designing $(\mathcal{A}_{\mathsf{enc}}, \mathcal{A}_{\mathsf{dec}})$. Since (ε, δ)-DP inevitably incurs $O\left(\frac{\log d}{\varepsilon}\right)$ error on the estimated frequency, it suffices to have SecAgg output an approximate sum (i.e., histogram) with distortion less than the DP error. This slack allows us to reduce the communication below the $\Omega(n \log d)$ lower bound per user for computing the exact histogram.

Count-sketch. We use count-sketch to achieve this goal. Count-sketch is a linear compression scheme (and hence can be represented in a matrix form $S = [S_1^\intercal, S_2^\intercal, ..., S_t^\intercal] \in \{-1,0,1\}^{wt \times d}$ for some $w, t \in \mathbb{N}$, where each $S_j \in \{-1,0,1\}^{w \times d}$ is generated according to an independent hash function) that allows for trading off the estimation error for communication cost. A count-sketch is determined by two parameters $w, t \in \mathbb{N}$; w is the bucket size that controls the magnitude of ℓ_∞ error, and t, the number of hash functions, determines the failure probability. To apply count-sketch in the frequency estimation problem, each user computes a local sketch of its data, i.e., SX_i, and sends it to the server. Upon receiving local sketches, the server can unsketch and obtain an estimate on $\mu(X^n)$. By setting $t = \Theta(\log(d/\gamma))$, count-sketch estimates μ with $O(\|\mu\|_1/w)$ error with failure probability at most γ^2.

Algorithm 1: Secure and private frequency estimation

Input: users' data (one-hot) $X_1, ..., X_n$, sketch parameters w, t, failure probability γ, PBM parameters L, θ.
Output: frequency estimate $\hat{\mu}$
Server broadcasts t i.i.d. generated sketch matrices $S_1, ..., S_t \in \{-1,0,1\}^{w \times d}$;
for *user* $i \in [n]$ **do**
 Compute t sketches $S_1X_i, ..., S_tX_i$;
 Perform Hadamard transform on each sketch;
 Apply PBM on each transformed sketch;
end
Server aggregates local noisy sketches via SecAgg, decodes PBM, and applies inverse Hadmard transform to obtain a noisy estimate $\widehat{S\mu}$;
Server unsketches $\widehat{S\mu}$ and obtains $\hat{\mu}$;
return $\hat{\mu}$

Hadmard transform. After computing the local sketch, each user performs the Hadamard transform to flatten each S_jX_i for $j \in [t]$ and $i \in [n]$, i.e., computes $H_wS_jX_i$, where H_w is the (normalized) Walsh-Hadamard matrix (assuming w is a power of 2) satisfying the following relation:

$$H_{2^n} = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{2^{n-1}}, & H_{2^{n-1}} \\ H_{2^{n-1}}, & -H_{2^{n-1}} \end{bmatrix}, \text{ and } H_0 = \left[1\right].$$

The flattening step reduces the dynamic range of S_iX_j in the sense that $\|H_wS_iX_j\|_\infty = \frac{1}{\sqrt{w}}\|S_iX_j\|_\infty$. This controls the ℓ_∞-sensitivity, which facilitates the following privatization steps.

Poisson-binomial mechanism. Last, to introduce DP, we make use of the Poisson-binomial mechanism (PBM) (Chen et al., 2022b). In PBM, users encode their locally flattened sketches H_wSX_i into parameters of binomial random variables (and hence the sum of n users' noisy reports follow a Poisson-binomial distribution). The main advantages of using PBM include: (1) the binomial distribution is closed under addition, and hence it is compatible with SecAgg; (2) it asymptotically converges to a Gaussian distribution and gives Rényi DP guarantees (which supports tight privacy accounting); (3) it does not require modular clipping and hence results in an unbiased estimate of μ (as opposed to other user-level discrete DP mechanisms, such as those of Kairouz et al. (2021); Agarwal et al. (2021)).

[2]Here we apply an ℓ_1 point-query bound due to the ℓ_1 geometry of $\mu(X^n)$.

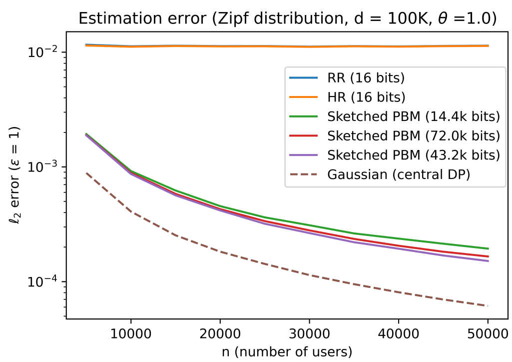

Figure 2: ℓ_2 loss with $\varepsilon = 1$. The error is computed with a normalization (the goal is to estimate $\frac{\mu(X^n)}{n}$).

By putting these pieces together, we arrive at Algorithm 1 with privacy guarantees, estimation error, and communication cost as stated in Theorem 5.1. A more detailed version is given in Algorithm 2 in Appendix A. In addition, in Appendix C, we show that we can improve the accuracy when additional knowledge on the sparsity of $\mu(X^n)$ is available.

Comparing the communication cost in Theorem 5.1 and the lower bounds in Corollary 5.1, we see that under ℓ_∞ loss, Algorithm 1 matches the lower bound up to a $\log n$ and $\sqrt{\log d}$ factor, where the small sub-optimality gap is due to the modular arithmetic used by SecAgg. Closing this gap is left to future work.

6 EXPERIMENTS

In this section, we provide empirical results for Algorithm 1, which we label as 'Sketched PBM' .

We compare sketched PBM with other decentralized (local) DP mechanisms, including randomized response (RR) (Warner, 1965; Kairouz et al., 2016) and the Hadamard response (HR) (Acharya et al., 2019b) (which is order-wise optimal for all $\varepsilon = O\left(\log d\right)$)

We set $d = 10^5$ and $n \in [10K, 50K]$, i.e., in a regime where $d \gg n$. Under this regime, it is well-known that local DP suffers from poor-utility (Duchi et al., 2013). We demonstrate that our proposed sketched PBM achieves a much better convergence rate (though admittedly at the cost of higher communication as predicted by our theoretical results). We also remark that the communication cost per user of the sketched PBM is fixed in this set of experiments, and thus the (normalized) estimation error does not strictly decrease with n (recall that our theory suggests in order to achieve the best performance, the communication cost has to be *increasing with* n). More detailed empirical results can be found in Appendix B.

7 ACKNOWLEDGEMENTS

This work was supported in part by NSF Award # CCF-2213223, a Meta Faculty research award, and a National Semiconductor Corporation Stanford Graduate Fellowship. The authors would like to thank Peter Kairouz for the helpful suggestions in the earlier version of this work.

References

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In *Conference on Learning Theory*, pages 3–17. PMLR, 2019a.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1120–1129, 2019b.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

Jayadev Acharya, Clément L Canonne, Cody Freitag, Ziteng Sun, and Himanshu Tyagi. Inference under information constraints iii: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory*, 2(1): 253–267, 2021a.

Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under privacy and communication constraints. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pages 79–98, 2021b.

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: communication-efficient and differentially-private distributed SGD. In *Advances in Neural Information Processing Systems*, pages 7564–7575, 2018.

Naman Agarwal, Peter Kairouz, and Ziyu Liu. The Skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Eugene Bagdasaryan, Peter Kairouz, Stefan Mellem, Adrià Gascón, Kallista A. Bonawitz, Deborah Estrin, and Marco Gruteser. Towards sparse federated analytics: Location heatmaps under distributed differential privacy with secure aggregation. *CoRR*, abs/2111.02356, 2021. URL `https://arxiv.org/abs/2111.02356`.

Victor Balcer and Albert Cheu. Separating local & shuffled differential privacy via histograms. *arXiv preprint arXiv:1911.06879*, 2019.

Victor Balcer and Salil Vadhan. Differential privacy on finite computers. *arXiv preprint arXiv:1709.05396*, 2017.

Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, pages 638–667. Springer, 2019.

Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 657–676, 2020.

Leighton Pate Barnes, Yanjun Han, and Ayfer Ozgur. Lower bounds for learning distributions under communication constraints via fisher information, 2019.

Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3): 645–659, 2020.

Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *ACM Symposium on Theory of Computing*, page 127–135, 2015. doi: 10.1145/2746539.2746632.

Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. Practical locally private heavy hitters. In *International Conference on Neural Information Processing Systems*, page 2285–2293, 2017.

James Henry Bell, Kallista A Bonawitz, Adrià Gascón, Tancrède Lepoint, and Mariana Raykova. Secure single-server aggregation with (poly) logarithmic overhead. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1253–1269, 2020.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

Nader H Bshouty. Optimal algorithms for the coin weighing problem with a spring scale. In *COLT*, volume 2009, page 82, 2009.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, page 435–447, 2018. doi: 10.1145/3196959.3196981.

Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. *ACM Transactions on Algorithms (TALG)*, 15(4):1–40, 2019.

Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete Gaussian for differential privacy. *arXiv preprint arXiv:2004.00010*, 2020.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, pages 267–284, 2019.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Interna-*

tional Colloquium on Automata, Languages, and Programming, pages 693–703. Springer, 2002.

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33, 2020.

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the dimension dependence in sparse distribution estimation under communication constraints. In *Conference on Learning Theory*, pages 1028–1059. PMLR, 2021a.

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Pointwise bounds for distribution estimation under communication constraints. *Advances in Neural Information Processing Systems*, 34:24593–24603, 2021b.

Wei-Ning Chen, Christopher A Choquette Choo, Peter Kairouz, and Ananda Theertha Suresh. The fundamental price of secure aggregation in differentially private federated learning. In *International Conference on Machine Learning*, pages 3056–3089. PMLR, 2022a.

Wei-Ning Chen, Ayfer Ozgur, and Peter Kairouz. The Poisson Binomial mechanism for unbiased federated learning with secure aggregation. In *International Conference on Machine Learning*, pages 3490–3506. PMLR, 2022b.

Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology–EUROCRYPT*, pages 375–403. Springer, 2019.

Beongjun Choi, Jy-yong Sohn, Dong-Jun Han, and Jaekyun Moon. Communication-computation efficient secure aggregation for federated learning. *arXiv preprint arXiv:2012.05433*, 2020a.

Seung Geol Choi, Dana Dachman-Soled, Mukul Kulkarni, and Arkady Yerukhimovich. Differentially-private multi-party sketching for large-scale statistics. *Cryptology ePrint Archive*, 2020b.

Graham Cormode and Akash Bharadwaj. Sample-and-threshold differential privacy: Histograms and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 1420–1431. PMLR, 2022.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006a.

David L Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(6):797–829, 2006b.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006b.

Ahmed Roushdy Elkordy, Jiang Zhang, Yahya H Ezzeldin, Konstantinos Psounis, and Salman Avestimehr. How much privacy does federated learning with secure aggregation guarantee? *arXiv preprint arXiv:2208.02304*, 2022.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep 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.

Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. *Information Systems*, 29(4): 343–364, 2004.

Keith Frikken. Privacy-preserving set union. In *Applied Cryptography and Network Security*, pages 237–252. Springer, 2007.

Oliver Gebhard, Max Hahn-Klimroth, Dominik Kaaser, and Philipp Loick. Quantitative group testing in the sublinear regime: Information-theoretic and algorithmic bounds. *arXiv preprint arXiv:1905.01458*, 2019.

Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh, and Amer Sinha. Differentially private aggregation in the shuffle model: Almost central accuracy in almost a single message. In *International Conference on Machine Learning*, pages 3692–3701. PMLR, 2021.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

Niv Gilboa and Yuval Ishai. Distributed point functions and their applications. In *Advances in Cryptology - EUROCRYPT*, pages 640–658, 2014. doi: 10.1007/978-3-642-55220-5_35.

Antonious M Girgis, Deepesh Data, Suhas Diggavi, Ananda Theertha Suresh, and Peter Kairouz. On the rényi differential privacy of the shuffle model. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2321–2341, 2021.

Yanjun Han, Pritam Mukherjee, Ayfer Ozgur, and Tsachy Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018*

IEEE International Symposium on Information Theory (ISIT), pages 506–510, 2018.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714, 2010.

Ziyue Huang, Yuan Qiu, Ke Yi, and Graham Cormode. Frequency estimation under multiparty differential privacy: One-shot and streaming. *Proc. VLDB Endow.*, 15(10): 2058–2070, 2022. doi: 10.14778/3547305.3547312.

Tayyebeh Jahani-Nezhad, Mohammad Ali Maddah-Ali, Songze Li, and Giuseppe Caire. Swiftagg+: Achieving asymptotically optimal communication load in secure aggregation for federated learning. *arXiv preprint arXiv:2203.13060*, 2022.

Swanand Kadhe, Nived Rajaraman, O Ozan Koyluoglu, and Kannan Ramchandran. Fastsecagg: Scalable secure aggregation for privacy-preserving federated learning. *arXiv preprint arXiv:2009.11248*, 2020.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, volume 48, pages 2436–2444, 2016.

Peter Kairouz, Ziyu Liu, and Thomas Steinke. The distributed discrete Gaussian mechanism for federated learning with secure aggregation. *arXiv preprint arXiv:2102.06387*, 2021.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3): 793–826, 2011.

Lea Kissner and Dawn Song. Privacy-preserving set operations. In *Advances in Cryptology–CRYPTO*, pages 241–257. Springer, 2005.

Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *International conference on World wide web*, pages 171–180, 2009.

David Lubell. A short proof of sperner's lemma. *Journal of Combinatorial Theory*, 1(2):299, 1966.

Qiyao Luo, Yilei Wang, and Ke Yi. Frequency estimation in the shuffle model with almost a single message. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2219–2232, 2022.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. HOGWILD! a lock-free approach to parallelizing stochastic gradient descent. In *Neural Information Processing Systems*, page 693–701, 2011.

Jonathan Scarlett and Volkan Cevher. Phase transitions in the pooled data problem. *Advances in Neural Information Processing Systems*, 30, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

Jinhyun So, Başak Güler, and A Salman Avestimehr. Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning. *IEEE Journal on Selected Areas in Information Theory*, 2(1):479–489, 2021.

Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206, 2019.

Emanuel Sperner. Ein satz über untermengen einer endlichen menge. *Mathematische Zeitschrift*, 27(1): 544–548, 1928.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

I-Hsiang Wang, Shao-Lun Huang, and Kuan-Yun Lee. Extracting sparse data via histogram queries. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 39–45. IEEE, 2016.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Chien-Sheng Yang, Jinhyun So, Chaoyang He, Songze Li, Qian Yu, and Salman Avestimehr. Lightsecagg: Rethinking secure aggregation in federated learning. *arXiv preprint arXiv:2109.14236*, 2021.

M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under local differential privacy. In *IEEE International Symposium on Information Theory (ISIT)*, pages 759–763, 2017. doi: 10.1109/ISIT.2017.8006630.

Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, and Wei Li. Federated heavy hitters discovery with differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 3837–3847. PMLR, 2020.

A Additional Details of Section 5

In this section, we provide additional details and empirical results of our noisy sketch scheme Algorithm 1 in Section 5 and give formal proofs on its privacy and utility guarantees. As mentioned in Section 5 our goal is to design a scheme that satisfies a stronger version of (distributed) DP, i.e., Rényi differential privacy, as it allows for tight privacy accounting. Therefore, in this section we first provide an RDP guarantee for our scheme, and then convert the RDP guarantee to (ε, δ)-DP using well-known conversion results such as Mironov (2017). To this end, we start by giving a brief introduction to Rényi DP.

A.1 Rényi Differential Privacy (RDP)

A useful variant of DP is the Rényi differential privacy (RDP), which allows for tight privacy accounting when a mechanism M is applied iteratively.

Definition A.1 (Rényi Differential Privacy (RDP)) *A randomized mechanism M satisfies (α, ε)-RDP if for any two neighboring datasets D, D', we have that $D_\alpha\left(P_{M(D)}, P_{M(D')}\right) \leq \varepsilon$ where $D_\alpha\left(P, Q\right)$ is the Rényi divergence between P and Q and is given by*

$$D_\alpha\left(P, Q\right) \triangleq \frac{1}{\alpha} \log\left(\mathbb{E}_Q\left[\left(\frac{P(X)}{Q(X)}\right)^\alpha\right]\right).$$

Note that one can convert an RDP guarantee to an (approximate) DP guarantee (for instance, see Mironov (2017)) but not the other way around in general. Although we presented our bounds in Section 5 in terms of approximate DP, our proposed schemes satisfy the RDP definition as we show next.

A.2 Details of Algorithm 1

We start by briefly recalling the details of count-sketch Charikar et al. (2002), which serves as our main compression tool for reducing communication costs. Count-sketch can be constructed via two sets of (pairwise independent) hash functions $h_i : [d] \to [w]$ and $\sigma_i : [d] \to \{-1, +1\}$ for $i \in [t]$. The functions can be organized in matrix form $S \in \{-1, 0, 1\}^{wt \times d}$, which can be viewed as a vertical stack of $S_1, ..., S_t \in \{-1, 0, 1\}^{w \times d}$, where for $i \in [t]$, $\left(S_i\right)_{j,k} = \sigma_i(j) \cdot \mathbb{1}_{\{h_i(j)=k\}}$. Note that $m \triangleq w \cdot t$ is the embedded dimension.

In Algorithm 2, we give a more detailed description of Algorithm 1, our private frequency estimation scheme from Section 5. We analyze the performance of Algorithm 2 in the next section.

A.3 Performance Analysis for Algorithm 2

We start by proving that Algorithm 2 satisfies the following RDP guarantee.

Theorem A.1 (RDP guarantee) *As long as $\theta \leq \frac{1}{4}$, Algorithm 2 satisfies $(\alpha, \tau(\alpha))$-RDP for all $\alpha > 1$ and $\tau(\alpha)$ such that*

$$\tau(\alpha) \geq C_0 \frac{\theta^2 L \alpha}{n} \cdot wt,$$

for some $C_0 > 0$

Proof. The proof follows from (Chen et al., 2022b, Corollary 3.2). ■

Once we obtain an RDP guarantee, we cast it into an (ε, δ)-DP guarantee by using results due to Canonne et al. (2020).

Theorem A.2 ((ε, δ)-DP guarantee) *Assume $\delta \leq \exp\left(-\frac{m\theta^2 wt}{n}\right)$. Then Algorithm 2 satisfies an (ε, δ) distributed DP guarantee for all ε and δ satisfying*

$$\varepsilon = \Omega\left(\sqrt{\frac{L\theta^2 wt \log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Proof. We apply (Canonne et al., 2020) to convert the RDP guarantee in Theorem A.1.

Algorithm 2: Secure and private frequency estimation with noisy sketch (detailed)

Input: users' data $X_1, ..., X_n \subseteq \{e_1, ..., e_d\}$, failure probability γ, sketch parameter w, t, PBM parameter L, θ
Output: frequency estimate $\hat{\mu}$
Server generates $(S_1, ..., S_t)$ (with $t = \Theta\left(\log\left(\frac{d}{\gamma}\right)\right)$ and w being a power of two and satisfying $w = \Theta\left(\min\left(n, \frac{n\varepsilon}{t}\right)\right)$);
The server broadcasts $S_1, ..., S_t$ to all users;
for $i \in [n]$ **do**
  Set $\theta =$ and $L =$;
  for $j \in [t]$ **do**
    user i computes $p_{ij} = \theta\left(H_w \cdot S_j \cdot X_i\right) + 1/2$, where $H_w \in \{-1/\sqrt{w}, 1/\sqrt{w}\}^{w\times w}$ is the Hadamard matrix;
    user i generates $Y_{ij} \triangleq \mathsf{Binom}\left(L, p_{ij}\right)$ coordinate-wisely (so $Y_{ij} \in [L]^w$);
  end
end
The server aggregates (via SecAgg Bonawitz et al. (2016)) noisy reports $\{Y_{ij}\}$ and computes the median

$$\left(\widehat{S_1\mu}, ..., \widehat{S_t\mu}\right) \triangleq \left(\frac{1}{\theta\sqrt{w}}\sum_{i=1}^n H_w \cdot \left(\frac{Y_{i1}}{L} - \frac{1}{2}\right), ..., \frac{1}{\theta\sqrt{w}}\sum_{i=1}^n H_w \cdot \left(\frac{Y_{it}}{L} - \frac{1}{2}\right)\right).$$

Server unsketches by computing the median:

$$\hat{\mu} = \mathsf{median}\left(S_1^\intercal \widehat{S_1\mu}, ..., S_t^\intercal \widehat{S_t\mu}\right).$$

return $\hat{\mu}$

Lemma A.1 (Renyi DP to approximate DP) *For any $\alpha \in (1, \infty)$, if*

$$D_\alpha\left(\mathcal{M}\left(\boldsymbol{x}\right) \| \mathcal{M}\left(\boldsymbol{x}'\right)\right) \leq \tau,$$

then $\mathcal{M}(\cdot)$ satisfies (ε, δ)-DP for

$$\varepsilon \geq \varepsilon^* \triangleq \tau + \frac{\log\left(\frac{1}{\delta}\right) + (\alpha-1)\log\left(1-\frac{1}{\alpha}\right) - \log\left(\alpha\right)}{\alpha - 1}.$$

Applying Theorem A.1 and Lemma A.1 above and plugging in $\tau = C_0\frac{\theta^2 L\alpha}{n}\cdot wt$, we see that $\hat{\mu}$ is (ε, δ)-DP for

$$\begin{aligned}
\varepsilon^* &= C_0\frac{\theta^2 L\alpha}{n}\cdot wt + \frac{\log\left(\frac{1}{\delta}\right) + (\alpha-1)\log\left(1-\frac{1}{\alpha}\right) - \log\left(\alpha\right)}{\alpha-1} \\
&\leq C_0\frac{\theta^2 L}{n}\cdot wt + C_0\frac{\theta^2 L(\alpha-1)}{n}\cdot wt + \frac{\log\left(\frac{1}{\delta}\right)}{\alpha-1} \\
&\overset{\text{(a)}}{=} C_0\frac{\theta^2 L}{n}\cdot wt + 2\sqrt{C_0\frac{L\theta^2 wt\log\left(\frac{1}{\delta}\right)}{n}} \\
&\overset{\text{(b)}}{=} O\left(\sqrt{\frac{L\theta^2 wt\log\left(\frac{1}{\delta}\right)}{n}}\right),
\end{aligned}$$

where (a) holds if we pick $\alpha - 1 = \sqrt{\frac{n\log(1/\delta)}{\theta^2 Lwt}}$ (i.e., such that AM-GM inequality holds with equality), and (b) holds if

$$\log\left(\frac{1}{\delta}\right) \geq \frac{L\theta^2 wt}{n} \iff \delta \leq \exp\left(-\frac{L\theta^2 wt}{n}\right).$$

■

Finally, in the following theorem, we compute the communication cost and control the ℓ_∞ and ℓ_2 estimation error of our algorithm.

Theorem A.3 (Privacy and Utility of Algorithm 2) *Let*

$$t = \log\left(\frac{d}{\gamma}\right),$$
$$w = \min\left(n, \left(\frac{n\varepsilon}{\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}}\right)\right),$$
$$L = \left\lceil \frac{n\varepsilon^2}{wt\log(1/\delta)} \right\rceil + 1,$$
$$\theta = O\left(\min\left(\frac{1}{4}, \sqrt{\frac{n\varepsilon^2}{wt\log(1/\delta)}}\right)\right).$$

Let $\hat{\mu}$ be the output of the Algorithm 2. Then:

- *$\hat{\mu}$ satisfies $(O(\varepsilon), \delta)$-DP and $\left(\alpha, O\left(\frac{\varepsilon^2\alpha}{\log(1/\delta)}\right)\right)$-Rényi DP.*
- *The communication complexity is $\tilde{O}(\min\left(n\varepsilon\log\left(\frac{1}{\gamma}\right), n\right))$ bits per user.*
- *With probability at least $1-\gamma$,*

$$\|\hat{\mu}-\mu\|_\infty = \max_{j\in[d]} |\mu_j - \hat{\mu}_j| \leq \frac{4n}{w} + O\left(\frac{\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right) = O\left(\frac{\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right).$$

- *By setting $\hat{\mu}_j = 0$ for all $j \in [d]$ such that $\hat{\mu}_j = O\left(\frac{\log\left(\frac{d}{\gamma}\right)\log\frac{1}{\delta}}{\varepsilon}\right)$, the ℓ_2^2 estimation error is bounded by*

$$O\left(\frac{n\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}\right),$$

  and the ℓ_1 error is bounded by

$$O\left(\frac{n\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right).$$

Proof.

Privacy guarantee. By plugging $L = \left\lceil \frac{n\varepsilon^2}{wt\log(1/\delta)} \right\rceil + 1$ and $\theta = O\left(\min\left(\frac{1}{4}, \sqrt{\frac{n\varepsilon^2}{wt\log(1/\delta)}}\right)\right)$ into Theorem A.1 and Theorem A.2, we immediately obtain the desired privacy guarantee.

Analysis of the communication cost. Let $\mathbb{Z}_M^m$ be the finite group that SecAgg operates on. In Algorithm 2, client i needs to communicate $\{Y_{ij}|i=1,...,t\}$ to the server. Notice that each $Y_{ij} \in [L]^w$, but for all $j \in [t]$, each coordinate of $\sum_i Y_{ij}$ can be as large as nL. Therefore, we will set $M = nL$. Now, if $w = \left(\frac{n\varepsilon}{\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}}\right) \leq n$, then the communication cost for each client becomes

$$\begin{aligned} &m\log(M+1) = m\log(nL+1) = wt\log(nL+1) \\ &= \frac{n\varepsilon\sqrt{\log(d/\gamma)}}{\sqrt{\log(1/\delta)}}\log\left(n\left(\left\lceil\frac{n\varepsilon^2}{wt\log(1/\delta)}\right\rceil+1\right)+1\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{n\varepsilon\sqrt{\log(d/\gamma)}}{\sqrt{\log(1/\delta)}} \log\left(n\left(\left\lceil \frac{\varepsilon}{\sqrt{\log(d/\gamma)\log(1/\delta)}} \right\rceil + 1\right) + 1\right) \\
&= \tilde{O}\left(\frac{n\varepsilon\sqrt{\log(d/\gamma)}}{\sqrt{\log(1/\delta)}}\right),
\end{aligned}
$$

where in the last equation we hide the $\log(n\lceil\varepsilon\rceil)$ term into $\tilde{O}(\cdot)$ for simplicity. On the other hand, if $w = n$, then

$$
\begin{aligned}
&m\log(M+1) = wt\log(nL+1) \\
&= n\log\left(\frac{d}{\gamma}\right)\log\left(n\left(\left\lceil \frac{n\varepsilon^2}{wt\log(1/\delta)} \right\rceil + 1\right) + 1\right) \\
&= n\log\left(\frac{d}{\gamma}\right)\log\left(n\left(\left\lceil \frac{\varepsilon^2}{\log\left(\frac{d}{\gamma}\right)\log(1/\delta)} \right\rceil + 1\right) + 1\right) \\
&= \tilde{O}\left(n\log(d/\gamma)\right).
\end{aligned}
$$

Bounding the ℓ_∞ error. We apply a similar analysis of error bounds using the count-sketch. Let

$$
\hat{\mu}^{(j)} = S_j^\intercal \widehat{S_j\mu} = S_j^\intercal\left(\frac{1}{\theta\sqrt{w}}\sum_{i=1}^n H_w\cdot\left(\frac{Y_{ij}}{L} - \frac{1}{2}\right)\right),
$$

for $j \in [t]$. Define $N^{(j)} \in \mathbb{R}^w$ be the estimation error of the j-th sketch, i.e.,

$$
N^{(j)} \triangleq S_j\mu - \frac{1}{\theta\sqrt{w}}\sum_{i=1}^n H_w\cdot\left(\frac{Y_{ij}}{m} - \frac{1}{2}\right).
$$

Then, for any $i \in [d]$, we can write the absolute error of the j-th sketch as

$$
\hat{\mu}_i^{(k)} - \mu_i = \sum_{j\neq i}\sigma_k(j)\sigma_k(i)\mathbb{1}_{\{h(j)=h(i)\}}\mu_j + N^{(j)}_{h_k(i)}.
$$

Therefore, we must have

$$
\begin{aligned}
\mathbb{E}\left[\left|\hat{\mu}_i^{(k)} - \mu_i\right|\right] &\leq \mathbb{E}\left[\left|\sum_{j\neq i}\sigma_k(j)\sigma_k(i)\mathbb{1}_{\{h_k(j)=h_k(i)\}}\mu_j\right| + \left|N^{(j)}_{h_k(i)}\right|\right] \\
&\overset{\text{(a)}}{\leq} \mathbb{E}\left[\sum_{j\neq i}\mathbb{1}_{\{h_k(j)=h_k(i)\}}\mu_j\right] + \sqrt{\mathbb{E}\left[\left(N^{(j)}_{h_k(i)}\right)^2\right]} \\
&\overset{\text{(b)}}{\leq} \frac{n}{w} + \sqrt{\mathbb{E}\left[\left(N^{(j)}_{h_k(i)}\right)^2\right]},
\end{aligned}
$$

where (a) follows due to Jensen's inequality and the fact that $\sigma_k(\cdot) \in \{-1,+1\}$, (b) holds since $h_k(i)$ and $h_k(j)$ are pairwise independent.

Next, we upper bound $\mathbb{E}\left[\left(N^{(j)}_{h_k(i)}\right)^2\right]$. For notational simplicity, assume $h_k(i) = h \in [w]$. Observe that

$$
\begin{aligned}
N^{(j)} &= S_j\mu - \frac{1}{\theta\sqrt{w}}\sum_{i=1}^n H_w\cdot\left(\frac{Y_{ij}}{L} - \frac{1}{2}\right) \\
&= H_w\left(H_wS_j\mu - \frac{1}{\theta\sqrt{w}}\sum_{i=1}^n\left(\frac{Y_{ij}}{L} - \frac{1}{2}\right)\right),
\end{aligned}
$$

where the second equality is due to the fact that $H_w \cdot H_w = I_w$.

Denote

$$\Delta_j \triangleq H_w S_j \mu - \frac{1}{\theta\sqrt{w}} \sum_{i=1}^{n} \left(\frac{Y_{ij}}{L} - \frac{1}{2} \right) \in \mathbb{R}^w.$$

Note that $H_w S_j \mu$ is the input to the PBM and $\frac{1}{\theta\sqrt{w}} \sum_{i=1}^{n} \left(\frac{Y_{ij}}{L} - \frac{1}{2} \right)$ is the estimate of PBM, so Δ_j satisfies the following properties (see (Chen et al., 2022b) for more details):

- $\Delta_j(h)$ is independent of $\Delta_j(h')$ for all $h \neq h'$ (where $\Delta_j(h)$ is the h-th coordinate of Δ_j).
- $\mathbb{E}\left[\Delta_j\right] = 0$.
- For any $h \in [w]$, $\mathbb{E}\left[\Delta_j^2(h)\right] = \frac{1}{wL\theta^2} \sum_{i=1}^{n} \mathsf{Var}\left(Y_{ij}\right) \leq \frac{n}{4wL\theta^2}$.

Let $H_w(h)$ be the h-th row of H_w. Then

$$\mathbb{E}\left[\left(N_h^{(j)} \right)^2 \right] = \mathbb{E}\left[\langle H_w(h), \Delta_j \rangle^2 \right] \overset{\text{(a)}}{=} \mathbb{E}\left[\frac{1}{w} \left\| \Delta_j \right\|^2 \right] \overset{\text{(b)}}{\leq} \frac{n}{4wL\theta^2} \overset{\text{(c)}}{=} \frac{n}{4w} O\left(\frac{wt\log(1/\delta)}{n\varepsilon^2} \right) = O\left(\frac{t\log(1/\delta)}{\varepsilon} \right),$$

where (a) holds since each coordinate of Δ_j is independent and each coordinate of $H_w(h)$ is either $\frac{1}{\sqrt{w}}$ or $-\frac{1}{\sqrt{w}}$, (b) holds since $\mathbb{E}\left[\Delta_j^2(h)\right] \leq \frac{n}{4wL\theta^2}$ for all $h \in [w]$, and (c) is because of our choice of L and θ.

Therefore, by Markov's inequality, we have

$$\mathbb{P}\left\{ \left| \hat{\mu}_i^{(k)} - \mu_i \right| \geq \frac{4n}{w} + O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right) \right\} \leq \frac{1}{4}.$$

Taking the median for $\left(\hat{\mu}_i^{(1)}, ..., \hat{\mu}_i^{(t)} \right)$ to apply the Chernoff bound, we obtain

$$\begin{aligned}
\mathbb{P}\left\{ \left| \hat{\mu}_i - \mu_i \right| \geq \frac{4n}{w} + O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right) \right\} &\leq \mathbb{P}\left\{ \sum_{k=1}^{t} \mathbb{1}_{\left\{ \left| \hat{\mu}_i^{(k)} - \mu_i \right| \geq \frac{4n}{w} + O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right) \right\}} \geq \frac{t}{2} \right\} \\
&\leq \mathbb{P}\left\{ \mathsf{Binom}\left(t, \frac{1}{4} \right) \geq \frac{t}{2} \right\} \\
&\leq \frac{\gamma}{d},
\end{aligned}$$

if we take $t = O\left(\log\left(\frac{d}{\gamma} \right) \right)$, where the last inequality is due to the Chernoff bound.

Taking the union bound over $i \in [d]$, we conclude that

$$\mathbb{P}\left\{ \max_{i \in [d]} \left| \hat{\mu}_i - \mu_i \right| \geq \frac{4n}{w} + O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right) \right\} \leq \gamma.$$

Setting $w = O\left(\frac{n\varepsilon}{\sqrt{t\log\left(\frac{1}{\delta}\right)}} \right) = O\left(\frac{n\varepsilon}{\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}} \right)$, we arrive at the desired result.

Bounding the ℓ_2 and ℓ_1 error. Since

$$\mathbb{P}\left\{ \max_{j \in [d]} \left| \hat{\mu}_i - \mu_i \right| = O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right) \right\} \leq \gamma,$$

we condition on the event

$$\mathcal{E} \triangleq \left\{ \max_{j\in[d]} |\hat{\mu}_i - \mu_i| = O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right) \right\}.$$

Under $\mathcal{E}$, when thresholding out every coordinate i such that $\hat{\mu}_i \leq O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right)$ (denoted as $\check{\mu}_i$), we must have

$$\begin{cases} |\check{\mu}_i - \mu_i| \leq O\left(\frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right), \text{ if } \mu_i \neq 0 \\ |\check{\mu}_i - \mu_i| = 0, \text{ if } \mu_i = 0. \end{cases}$$

Since there can be at most n coordinates such that $\mu_i \neq 0$, the ℓ_2^2 error can be at most

$$\sum_{i=1}^d (\check{\mu}_i - \mu_i)^2 \leq n \cdot \frac{t\log\left(\frac{1}{\delta}\right)}{\varepsilon^2} + (d-n)\cdot 0 = O\left(\frac{n\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \right).$$

Similarly, for the ℓ_1 error, we have

$$\sum_{i=1}^d |\check{\mu}_i - \mu_i| \leq n \cdot \frac{\sqrt{t\log\left(\frac{1}{\delta}\right)}}{\varepsilon} = O\left(\frac{n\sqrt{\log\left(\frac{d}{\gamma}\right)\log\left(\frac{1}{\delta}\right)}}{\varepsilon} \right).$$

This completes the proof of Theorem A.3.

Finally, setting $\gamma = \frac{1}{\mathsf{poly}(n,d)}$, we can cast the high-probability bound in Theorem A.3 into expected bounds shown in Theorem 5.1.

B Additional Experiments

In this section, we provide additional empirical results for Algorithm 1, which we label as 'sketched PBM'. As in Section 6, in the first set of experiments, we compare sketched PBM with other decentralized (local) DP mechanisms, including randomized response (RR) (Warner, 1965; Kairouz et al., 2016) and the Hadamard response (HR) (Acharya et al., 2019b) (which is order-wise optimal for all $\varepsilon = O(\log d)$)[3]. The data is generated under a (truncated) Geometric distribution (with $\theta = 0.8$) in Figure 3 and under a (truncated) Zipf distribution (with $\theta = 1.0$) in Figure 4. For the (centralized) Gaussian and the (distributed) sketched PBM mechanisms, δ is set to be 10^{-5}. For sketched PBM, we set the parameter $L = 10$.

We set $d = 10^5$ and $n \in [10k, 50k]$, i.e., in a regime where $d \gg n$ and compare the above schemes for $\varepsilon \in \{1, 5, 10\}$. Under this regime, it is well-known that local DP suffers from poor utility (Duchi et al., 2013). We demonstrate that our proposed sketched PBM mechanism achieves a much better convergence rate (though admittedly at the cost of higher communication) both for the Gemoetric and Zipf distributions. We also remark that the per user communication cost of the sketched PBM mechanism is fixed in this set of experiments, and thus the (normalized) estimation error does not strictly decrease with n (recall that our theory suggests in order to achieve the best performance, the per user communication cost has to be *increasing with* n). We note that in the low privacy regimes (e.g., when $\varepsilon = 10$), the communication budget has a greater impact on the accuracy of sketched PBM. This suggests that in this regime the performance of the scheme is limited by the compression error. Equivalently, the number of bits used by the scheme are below the threshold characterized by our theory to achieve the central DP performance.

[3] For the local DP mechanisms, we partly use the implementation from `https://github.com/zitengsun/hadamard_response`.

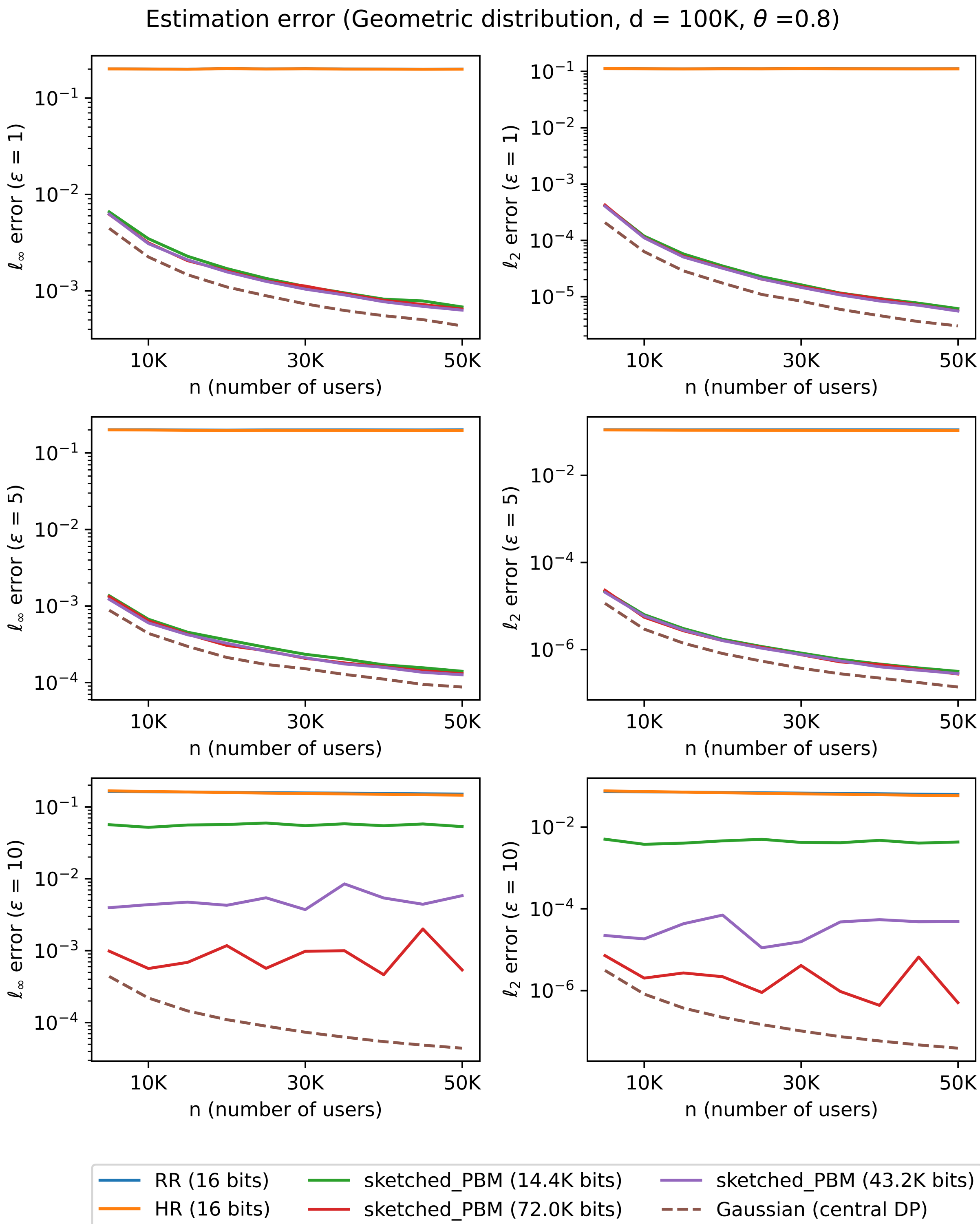

Figure 3: ℓ_∞ and ℓ_2 loss with $\varepsilon = \{1, 5, 10\}$. The error is computed with a normalization (the histogram is normalized by a factor of n, i.e., $\frac{\mu(X^n)}{n}$). The y-axis is under a log-scale. In addition, when computing the ℓ_2 error, we project all the estimated histograms into the probability simplex to further reduce the estimation error (also been adopted by Acharya et al. (2019b)).

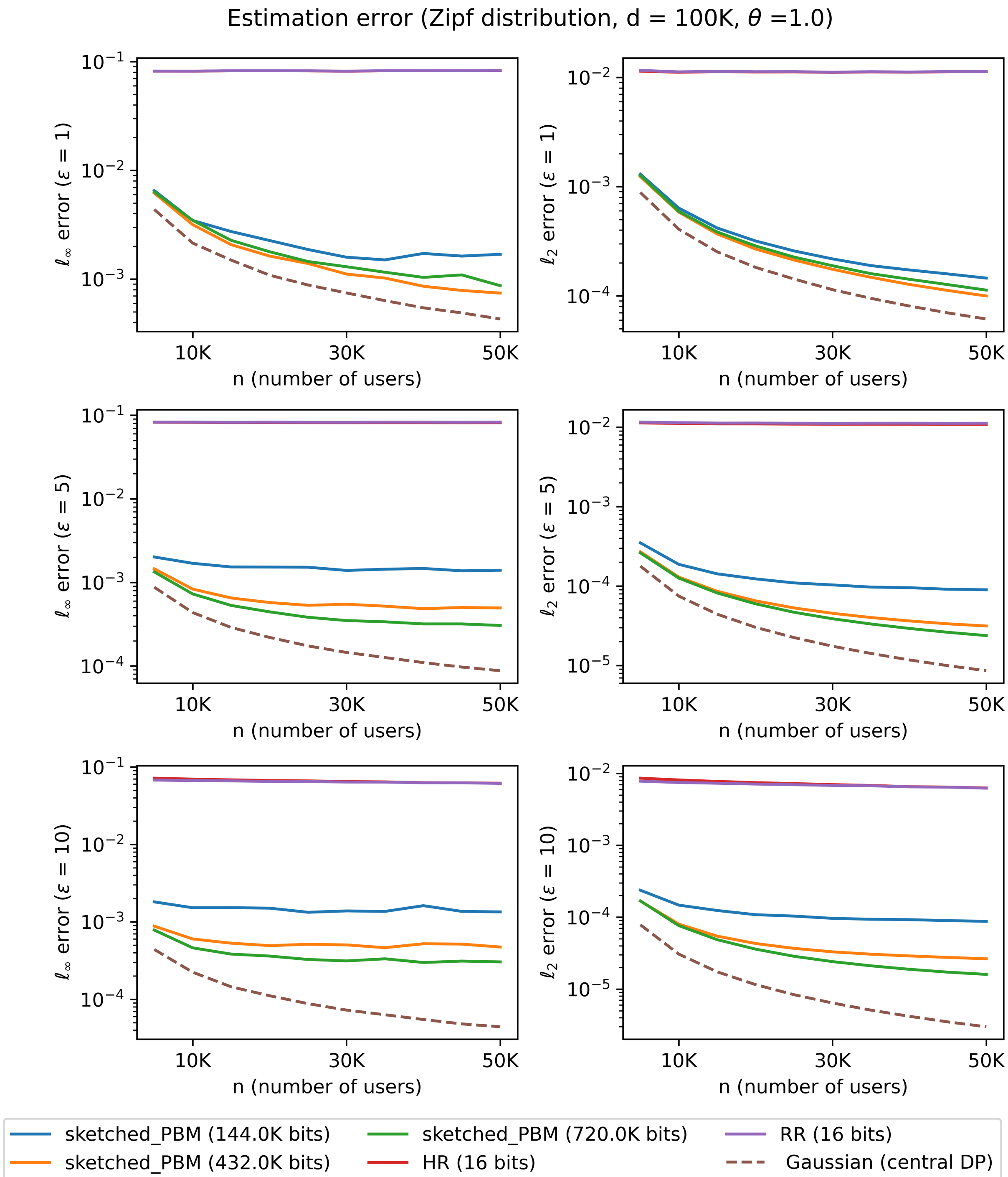

Figure 4: ℓ_∞ and ℓ_2 loss with $\varepsilon = \{1, 5, 10\}$.

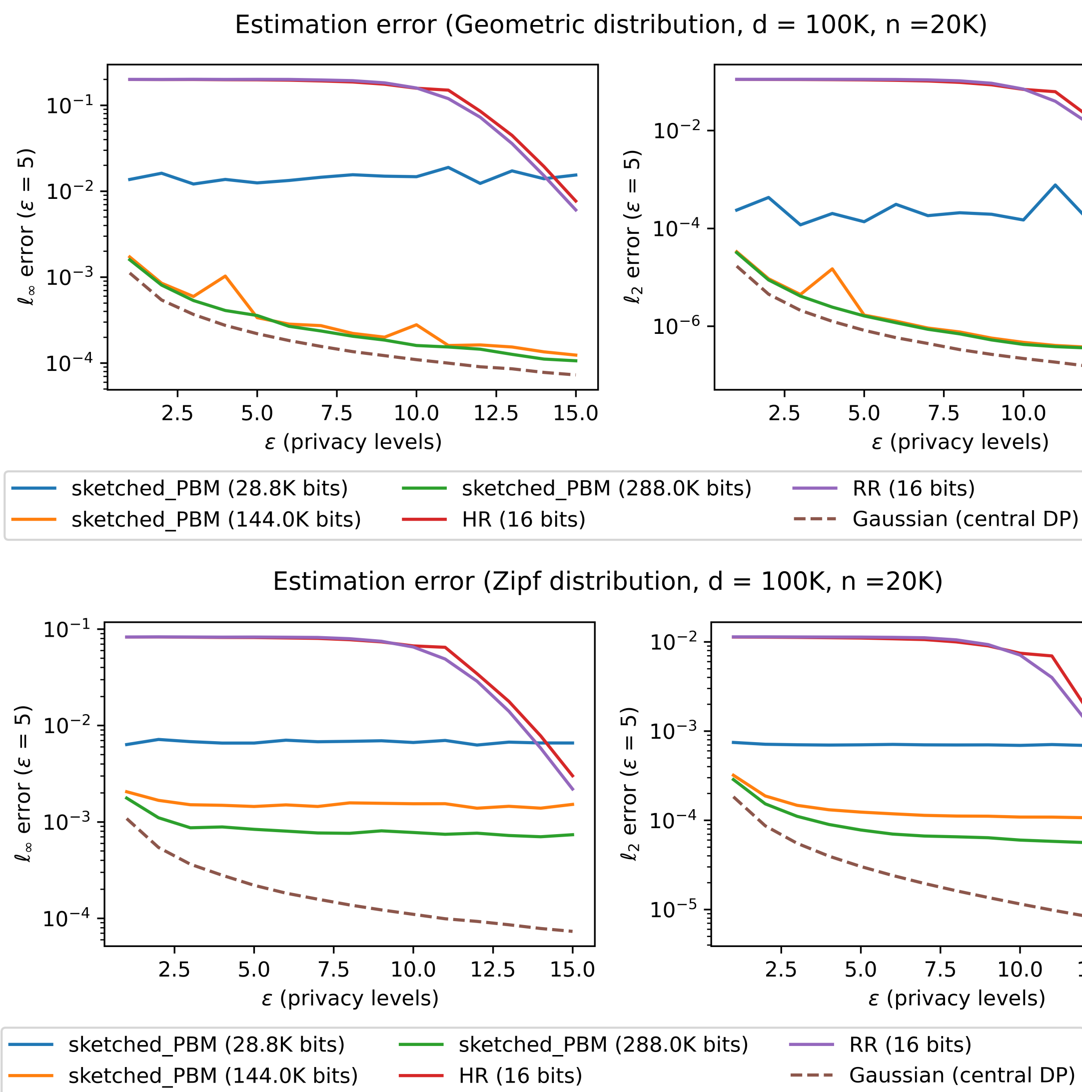

Figure 5: ℓ_∞ and ℓ_2 loss with $\varepsilon = [1, 15]$.

In the next set of experiments (Figure 5), we fix $d = 10^5$ and $n = 2 \cdot 10^4$ and vary $\varepsilon \in [1, 15]$. We compare the ℓ_2 and ℓ_∞ error from different mechanisms under the Geometric distribution and Zipf distribution. We see that the sketched PBM mechanism significantly outperforms local DP mechanisms in high-privacy regime.

C Sparse Private Frequency Estimation

Finally, we briefly discuss the sparse frequency estimation setting, where the true histogram is assumed to be s-sparse $\|\mu(X^n)\|_0 \leq s$ for some $s \ll n \ll d$ (i.e., X_i belongs to a size-s subset of $[d]$). When s is known ahead of time, the server can generate a sketch matrix S according to s instead of n, and all the analysis carries through with n replaced by s. This improves both the communication cost and the ℓ_2 estimation error.

On the other hand, if s is unknown but we are allowed to run a protocol with multiple rounds (this may or may not be possible in federated analytic settings where users may frequently drop out), we can first estimate s (subject to privacy and security constraints) via a private F_0 sketch (using, for example, (Choi et al., 2020b)). In the second round, we can set the size of the count-sketch in Algorithm 1 according to $\hat{s}$. The communication cost of estimating s is negligible compared to that of estimating $\hat{\mu}$, and hence we can still replace the dependency on n with s in our results.

Finally, we note that if interaction (multiple-rounds) is not allowed and s is unknown, we cannot reduce the communication from linear in n to s. However, the thresholding trick used in the proof of Theorem 5.1 can still be applied (which does not require knowledge of s) and hence the ℓ_2 error can be reduced to $O\left(\frac{s\log^2 d}{\varepsilon^2}\right)$.

D Omitted Proofs

D.1 Proof of Theorem 4.2

Recall that $\mathcal{H}_n \triangleq \left\{\mu \in \mathbb{Z}_+^d \middle| \|\mu\|_1 = n\right\}$ is the collection of all n-histograms. Then (4) is the same as

$$\forall \Delta\mu \in \Delta\mathcal{H}_n,\, S \cdot \Delta\mu \neq 0, \tag{5}$$

where $\Delta\mathcal{H}_n = \mathcal{H}_n - \mathcal{H}_n \triangleq \{\mu_1 - \mu_2 | \mu_1, \mu_2 \in \mathcal{H}_n, \mu_1 \neq \mu_2\}$. Note that for any $\Delta\mu \in \Delta\mathcal{H}_n$, we must have (1) $\Delta\mu_j \in \mathbb{Z}^d$; (2) $\sum_j \Delta\mu_j = 0$; and (3) $\|\Delta\mu_j\|_1 \leq 2n$.

To show that (5) holds when m is large enough, we construct S in the following probabilistic way:

$$\forall i \in [m], j \in [d],\, S_{ij} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(1/2).$$

We denote the resulting probability distribution over all possible S as Q. In addition, let $s_i \in \mathbb{R}^d$ be the i-th row of S, i.e., $S = [s_1, s_2, ..., s_m]^\intercal$. Then, to prove (5) holds for some S, it suffices to show

$$\mathbb{P}_Q\left\{\forall \Delta\mu \in \Delta\mathcal{H}_n,\, S\cdot\Delta\mu = 0\right\} < 1,$$

as long as $m = O\left(n\log d/\log n\right)$, where the probability is taken with respect to the randomization over S.

To this end, observe that

$$\mathbb{P}_Q\left\{\forall \Delta\mu \in \Delta\mathcal{H}_n,\, S\cdot\Delta\mu = 0\right\} \overset{\text{(a)}}{\leq} \sum_{\Delta\mu\in\Delta\mathcal{H}_n} \mathbb{P}_Q\left\{S\cdot\Delta\mu = 0\right\} \overset{\text{(b)}}{=} \sum_{\Delta\mu\in\Delta\mathcal{H}_n} \left(\mathbb{P}_Q\left\{s_1\cdot\Delta\mu = 0\right\}\right)^m, \tag{6}$$

where (a) is due to the union bound, and (b) holds since each row of S is generated i.i.d.

Additional notation. Before we proceed to upper bound $\mathbb{P}_Q\left\{s_1\cdot\Delta\mu = 0\right\}$, we introduce some necessary notations. Let $\Delta\mu^+$ be the positive part of $\Delta\mu$, i.e., $\Delta\mu_j^+ \triangleq \min(\Delta\mu_j, 0)$ for $j \in [d]$. Similarly, $\Delta\mu_j^- \triangleq \min(-\Delta\mu_j, 0)$ (so we must have $\Delta\mu = \Delta\mu^+ - \Delta\mu^-$).

For a vector $\nu \in \mathbb{Z}^d$, let $\iota(\nu)$ be the multi-set containing all the *non-zero* values of ν. Let $|\iota(\nu)|$ be the (multi-set) cardinality of $\iota(\nu)$. For instance, if $\nu = [0, 1, 3, 3, 2]$, then $\iota(\nu) = \{1, 2, 3, 3\}$ and $|\iota(\nu)| = |\{1, 2, 3, 3\}| = 4$.

Finally, let $\mathsf{sum}(\Delta\mu^+)$ be the set of all possible partial sums of $\iota(\Delta\mu^+)$, i.e., $\mathsf{sum}(\Delta\mu^+) = \left\{v\cdot\Delta\mu^+ \middle| v \in \{0,1\}^d\right\}$. Similarly, $\mathsf{sum}(\Delta\mu^-) = \left\{v\cdot\Delta\mu^- \middle| v \in \{0,1\}^d\right\}$.

Claim D.1 *For any* $\Delta\mu \in \Delta\mathcal{H}_n$, $\mathbb{P}_Q\left\{s_1\cdot\Delta\mu = 0\right\} \leq \sqrt{\frac{\pi}{2}\left\lceil\frac{|\iota(\Delta\mu)|}{2}\right\rceil}^{-1}$.

Proof of claim. Observe that

$$\begin{aligned}\mathbb{P}_Q\left\{s_1\cdot\Delta\mu=0\right\} &\overset{\text{(a)}}{=} \mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=s_1\cdot\Delta\mu^-\right\} && (7)\\ &= \sum_{\ell\in\mathsf{sum}(\Delta\mu^-)\cup\mathsf{sum}(\Delta\mu^+)} \mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\cap s_1\cdot\Delta\mu^-=\ell\right\}\\ &\overset{\text{(b)}}{=} \sum_{\ell\in\mathsf{sum}(\Delta\mu^-)\wedge\mathsf{sum}(\Delta\mu^+)} \mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\right\}\cdot\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^-=\ell\right\}\\ &\leq \max_{\ell\in\mathsf{sum}(\Delta\mu^+)}\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\right\}, && (8)\end{aligned}$$

where (a) holds since $\Delta\mu=\Delta\mu^+-\Delta\mu^-$, (b) holds since $\Delta\mu^+$ and $\Delta\mu^-$ have disjoint supports and that each coordinate of s_1 is generated independently. Similarly, by symmetry, we have $\mathbb{P}_Q\left\{s_1\cdot\Delta\mu=0\right\}\leq\max_{\ell\in\mathsf{sum}(\Delta\mu^-)}\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^-=\ell\right\}$, so

$$\mathbb{P}_Q\left\{s_1\cdot\Delta\mu=0\right\}\leq\min\left(\max_{\ell\in\mathsf{sum}(\Delta\mu^+)}\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\right\},\max_{\ell\in\mathsf{sum}(\Delta\mu^-)}\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^-=\ell\right\}\right). \tag{9}$$

Therefore, it remains to upper bound $\max_{\ell\in\mathsf{sum}(\Delta\mu^+)}\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\right\}$. To this end, observe that since each coordinate of s_1 is i.i.d. $\mathsf{Bern}(1/2)$,

$$\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\right\}=\left|\left\{v\middle|v\in\{0,1\}^d,v\cdot\Delta\mu^+=\ell\right\}\right|\cdot2^{-d}=\left|\left\{A\middle|A\in2^{\iota(\Delta\mu^+)},\sum_{a\in A}a=\ell\right\}\right|\cdot2^{-|\iota(\Delta\mu^+)|},$$

where $2^{\iota(\Delta\mu^+)}$ denotes the power set of the multi-set $\iota(\Delta\mu^+)$. Notice that for the multi-set $\iota(\Delta\mu^+)$, we treat each element as a different one even some of them may possess the same value, so the cardinality of $2^{\iota(\Delta\mu^+)}$ is $2^{|\iota(\Delta\mu^+)|}$.

Now, observe that $\mathcal{F}_\ell\triangleq\left\{A\middle|A\in2^{\iota(\Delta\mu^+)},\sum_{a\in A}a=\ell\right\}$ must form a Sperner family (Sperner, 1928; Lubell, 1966), that is, for any $A_1,A_2\in\mathcal{F}_\ell$, neither $A_1\subset A_2$ nor $A_2\subset A_1$ holds. This is because otherwise, if $A_1\subset A_2$, we must have $\sum_{A_2}a>\sum_{A_1}a$, and thus at least one of them must be not equal to ℓ. Therefore, applying Sperner's theorem (Sperner, 1928; Lubell, 1966), we must have

$$\left|\left\{A\middle|A\in2^{\iota(\Delta\mu^+)},\sum_{a\in A}a=\ell\right\}\right|\leq\binom{|\iota(\Delta\mu^+)|}{\left\lceil\frac{|\iota(\Delta\mu^+)|}{2}\right\rceil},$$

which implies

$$\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^+=\ell\right\}\leq\binom{|\iota(\Delta\mu^+)|}{\left\lceil\frac{|\iota(\Delta\mu^+)|}{2}\right\rceil}\cdot2^{-|\iota(\Delta\mu^+)|}\leq\sqrt{\frac{\pi}{2}\left|\iota(\Delta\mu^+)\right|}^{-1},$$

where the last inequality is due to basic combinatorial fact (Cover, 1999, Chapter 17). Similarly, by symmetry, we also have

$$\mathbb{P}_Q\left\{s_1\cdot\Delta\mu^-=\ell\right\}\leq\sqrt{\frac{\pi}{2}\left|\iota(\Delta\mu^-)\right|}^{-1},$$

and hence plugging in (9) we obtain

$$\mathbb{P}_Q\left\{s_1\cdot\Delta\mu=0\right\}\leq\min\left(\sqrt{\frac{\pi}{2}\left|\iota(\Delta\mu^+)\right|}^{-1},\sqrt{\frac{\pi}{2}\left|\iota(\Delta\mu^-)\right|}^{-1}\right)\leq\sqrt{\frac{\pi}{2}\left\lceil\frac{|\iota(\Delta\mu)|}{2}\right\rceil}^{-1},$$

where the last inequality holds since

$$\max\left(\left|\iota(\Delta\mu^-)\right|,\left|\iota(\Delta\mu^+)\right|\right)\geq\left\lceil\frac{|\iota(\Delta\mu^-)|+|\iota(\Delta\mu^+)|}{2}\right\rceil=\left\lceil\frac{|\iota(\Delta\mu)|}{2}\right\rceil.$$

□

Now, with Claim D.1, we proceed to bound (6) as follows:

$$\begin{aligned}
\mathbb{P}_Q\left\{\forall \Delta\mu \in \Delta\mathcal{H}_n,\, S\cdot\Delta\mu = 0\right\} &\leq \sum_{\Delta\mu\in\Delta\mathcal{H}_n} \left(\mathbb{P}_Q\left\{s_1\cdot\Delta\mu = 0\right\}\right)^m \\
&\leq \sum_{\Delta\mu\in\Delta\mathcal{H}_n} \sqrt{\frac{\pi}{2}\left\lceil\frac{|\iota(\Delta\mu)|}{2}\right\rceil}^{-m} \\
&= \sum_{\ell=1}^{2n} \sum_{\Delta\mu:|\iota(\Delta\mu)|=\ell} \sqrt{\frac{\pi}{2}\left\lceil\frac{\ell}{2}\right\rceil}^{-m} \\
&\leq \sum_{\ell=1}^{2n} \binom{d}{\ell}(2n+1)^\ell \sqrt{\frac{\pi}{2}\left\lceil\frac{\ell}{2}\right\rceil}^{-m} \\
&= \sum_{\ell=1}^{n^*} \binom{d}{\ell}(2n+1)^\ell \left(\frac{\pi}{2}\left\lceil\frac{\ell}{2}\right\rceil\right)^{-m/2} + \sum_{\ell=n^*}^{2n} \binom{d}{\ell}(2n+1)^\ell \left(\frac{\pi}{2}\left\lceil\frac{\ell}{2}\right\rceil\right)^{-m/2}, \quad (10)
\end{aligned}$$

where $n^* \in [n]$ is a tuning parameter that will be specified later. Now, we bound the last two terms separately. For the first term, we have

$$\begin{aligned}
\sum_{\ell=1}^{n^*} \binom{d}{\ell}(n+1)^\ell \left(\frac{\pi}{2}\left\lceil\frac{\ell}{2}\right\rceil\right)^{-m/2} &\leq (2n+1)^{n^*}\left(\frac{\pi}{2}\right)^{-m/2} \sum_{\ell=1}^{n^*}\binom{d}{\ell} \\
&\leq (2n+1)^{n^*}\left(\frac{\pi}{2}\right)^{-m/2}(d+1)^{n^*+1} \\
&\leq \exp\left((n^*+1)\log(d+1) + n^*\log(2n+1) - \frac{m}{2}\log\left(\pi/2\right)\right) \to 0,
\end{aligned}$$

as long as $m = \Omega\left(n^*\log(d+1) + n^*\log(2n+1)\right) = \Omega\left(n^*\log d\right)$ (since $n \ll d$). For the second term, observe that

$$\begin{aligned}
\sum_{\ell=n^*}^{2n} \binom{d}{\ell}(n+1)^\ell \left(\frac{\pi}{2}\left\lceil\frac{\ell}{2}\right\rceil\right)^{-m/2} &\leq (2n+1)^{2n}\left(\frac{\pi n^*}{4}\right)^{-m/2}\left(\sum_{\ell=n^*}^{2n}\binom{d}{\ell}\right) \\
&\leq (2n+1)^{2n}\left(\frac{\pi n^*}{4}\right)^{-m/2}\left(\sum_{\ell=0}^{2n}\binom{d}{\ell}\right) \\
&= \exp\left(2n\log(2n+1) + 2n\log(d+1) - \frac{m}{2}\left(\log n^* + \log\left(\pi/4\right)\right)\right).
\end{aligned}$$

Therefore, as long as $m = \Omega\left(\frac{2n\log(2n+1)+n\log(d+1)}{\log n^*+\log(\pi/4)}\right) = \Omega\left(\frac{2n\log d}{\log n^*+\log(\pi/4)}\right)$.

Putting both upper bounds on m together, and select $n^* = \lceil n/\log n + 3\rceil$, we conclude that as long as

$$m = \Omega\left(\max\left(\frac{n\log d}{\log n} + 3\log d, \frac{n\log d}{\log n - \log\log n + 3 - \log\left(\pi/4\right)}\right)\right) = \Omega\left(n\log d/\log n\right),$$

then $\mathbb{P}_Q\left\{\forall\Delta\mu \in \Delta\mathcal{H}_n,\, S\cdot\Delta\mu = 0\right\} \to 0$, which implies that there must exists a feasible S that distinguish all elements in $\Delta\mathcal{H}_n$.

■

D.2 Proof of Lemma 3.1

First of all, observe that for any $\mathcal{D} \subset [n]$ such that $\mathcal{D} \leq d$

$$\begin{aligned}
&I\left(X_{[n]}; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right)\right) \\
&= I\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right)\right) + I\left(Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]}, \mathcal{D}\right); X_{[n]} \middle| \sum_{i\in[n]\setminus\mathcal{D}} X_i\right)
\end{aligned}$$

$$
\begin{aligned}
&= I\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right)\\
&= H\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i\right) - H\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i \middle| Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right)\\
&= H\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i\right),
\end{aligned}
$$

where the second equality holds due to (S1) and the third equality holds since (C1) implies

$$
H\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i \middle| Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right) = 0.
$$

On the other hand, let $\mathcal{D}' \subset [n]$ be such that $|\mathcal{D}'| = d$ and let $j \in \mathcal{D}\setminus j$ we also have

$$
\begin{aligned}
&I\left(X_{[n]}; Y_{[n]\setminus\mathcal{D}'}, \theta_{\mathcal{D}'}\right)\\
&= I\left(Y_{[n]\setminus\{\{j\}\vee\mathcal{D}'\}}, \theta_{\mathcal{D}'\vee\{j\}}; X_{[n]}\right) + I\left(Y_j; X_{[n]}\middle|Y_{[n]\setminus\{\{j\}\vee\mathcal{D}'\}}, \theta_{\mathcal{D}'\vee\{j\}}\right)\\
&= I\left(Y_j; X_{[n]}\middle|Y_{[n]\setminus\{\{j\}\vee\mathcal{D}'\}}, \theta_{\mathcal{D}'\vee\{j\}}\right)\\
&= H\left(Y_j\middle|Y_{[n]\setminus\{\{j\}\vee\mathcal{D}'\}}, \theta_{\mathcal{D}'\vee\{j\}}\right) - H\left(Y_j\middle|Y_{[n]\setminus\{\{j\}\vee\mathcal{D}'\}}, \theta_{\mathcal{D}'\vee\{j\}}, X_{[n]}\right)\\
&\leq H\left(Y_j\middle|Y_{[n]\setminus\{\{j\}\vee\mathcal{D}'\}}, \theta_{\mathcal{D}'\vee\{j\}}\right)\\
&\leq H\left(Y_j\right), \qquad (11)
\end{aligned}
$$

where the second equality is due to (S2). ■

D.3 Proof of Lemma 3.2

Notice that by (11), we have $H(Y_i) \geq I\left(X_{[n]}; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right)$. Therefore, it suffices to lower bound $I\left(X_{[n]}; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right)$ subject to (C1′), (S1), and (S2). Using (S1), we have

$$
I\left(X_{[n]}; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right) = I\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i; Y_{[n]\setminus\mathcal{D}}, h\left(\theta_{[n]},\mathcal{D}\right)\right) \geq I\left(\sum_{i\in[n]\setminus\mathcal{D}} X_i; Y_{[n]\setminus\mathcal{D}}\right).
$$

Constrained on (C1′), this quantity is lower bounded by $R(\beta)$. ■

D.4 Proof of Corollary 4.1

Let $\mathcal{H}_n \triangleq \left\{(n_1, n_2, ..., n_d)\middle|\sum_{j=1}^d n_j = n, n_j \in \mathbb{Z}_+\right\}$ be the collection of all n-histograms (over a size-d domain). To construct a worst-case prior π_{X^n} over $\mathcal{X}^n$ such that $H\left(\sum_{i=1}^n X_i\right) = H\left(\mu\left(X^n\right)\right)$ is maximized, it suffices to find a π_μ over $\mathcal{H}_n$ that has large entropy. This is because one can generate π_{X^n} according to the following compound procedure such that $\sum_i X_i$ has marginal distribution π_μ: first select $\mu \sim \pi_\mu$ and then draw X_i from histogram μ without replacement.

To this end, we simply set $\pi_\mu = \mathsf{uniform}\left(\mathcal{H}_n\right)$. The entropy is thus given by

$$
H\left(\mu\left(X^n\right)\right) = \log\left|\mathcal{H}_n\right| = \log\left(\binom{d+n-1}{n-1}\right) = \Omega\left(n\log\left(\frac{d+n-1}{n-1}\right)\right) = \Omega\left(n\log d\right),
$$

where the last equality holds when $d \gg n$.

D.5 Proof of Lemma 5.3

Note that characterizing the rate function $R(\beta)$ (i.e., solving (1)) is equivalent to solving the following dual form:

$$
\mathsf{err}(b) \triangleq \begin{pmatrix} \min_{P_{Y^n|\mu(X^n)}} & \min_{\hat{\mu}} \mathbb{E}\left[\ell\left(\hat{\mu}\left(Y^n\right), \mu\left(X^n\right)\right)\right] \\ \text{subject to} & I\left(Y^n; \mu\left(X^n\right)\right) \leq b. \end{pmatrix} \qquad (12)
$$

The dual form can be interpreted as the minimum distortion (under loss function $\ell(\cdot)$) subject to a b-bit communication constraint. Moreover, since $\hat{\mu}(\cdot)$ can be any arbitrary (measurable) function of Y^n, we suppress its dependency on Y^n and simplify (12) to

$$\mathsf{err}(b) \triangleq \begin{pmatrix} \min_{P_{\hat{\mu}|\mu(X^n)}} & \min_{\hat{\mu}} \mathbb{E}\left[\ell\left(\hat{\mu}, \mu\left(X^n\right)\right)\right] \\ \text{subject to} & I\left(\hat{\mu}; \mu\left(X^n\right)\right) \leq b. \end{pmatrix} \tag{13}$$

To obtain the lower bound on $\mathsf{err}(b)$, our strategy is to construct a hard prior distribution π_{X^n}. Following the same argument as in Corollary 4.1, it suffices to construct a prior π_μ over $\mathcal{H}_n$, such that when $\mu_1, \mu_2 \overset{\text{i.i.d.}}{\sim} \pi_\mu$, with high-probability $\ell(\mu_1, \mu_2)$ will be large. Once obtaining a hard π_μ, we make use of the following Fano's inequality to obtain a lower bound on the smallest distortion $\mathbb{E}_{\mu\sim\pi_\mu}\left[\ell\left(\hat{\mu}, \mu\right)\right]$ one can possibly hope for.

Lemma D.1 (Fano's inequality) *Let $V \sim \mathsf{uniform}(\mathcal{V})$ for some finite set $\mathcal{V}$ and $V - U - \hat{V}$ form a Markov chain. Then*

$$\mathbb{P}\left\{\hat{V}(U) \neq V\right\} \geq 1 - \frac{I(U;V)+1}{\log|\mathcal{V}|}.$$

Bounding the ℓ_∞ distortion. Recall that our goal is to find a prior π_μ over $\mathcal{H}_n$, such that when $\mu_1, \mu_2 \overset{\text{i.i.d.}}{\sim} \pi_\mu$, $\|\mu_1 - \mu_2\|_\infty$ is large. We proceed by finding a (large) subset of $\Pi_R \subseteq \mathcal{H}_n$, such that

- $|\Pi_R| \geq 2^{2b}$ (where R is a tuning parameter);
- for any $\mu_1, \mu_2 \in \Pi_R$ such that $\mu_1 \neq \mu_2$, $\|\mu_1 - \mu_2\|_\infty \geq \Theta\left(\frac{n \log d}{b}\right)$.

If we can find such Π_R, then by setting $\pi_\mu = \mathsf{uniform}(\Pi_R)$ and together with Fano's inequality (Lemma D.1), we obtain

$$\begin{aligned}
\min_{\hat{\mu}} \mathbb{E}_\mu\left[\|\hat{\mu} - \mu\|_\infty\right] &\geq \min_{\hat{\mu}} \mathbb{E}_{\mu\sim\pi_\mu}\mathbb{E}_\mu\left[\|\hat{\mu} - \pi\|_\infty\right] && (14)\\
&\geq \min_{\hat{\mu}} \mathbb{P}_{\mu\sim\pi_\mu}\left\{\hat{\mu} \neq \mu\right\} \cdot \min_{\mu_1\neq\mu_2, \mu_1,\mu_2\in\Pi_R} \|\mu_1 - \mu_2\|_\infty && (15)\\
&\geq \Theta\left(\frac{n\log d}{b}\right) \min_{\hat{\mu}} \mathbb{P}_{\mu\sim\pi_\mu}\left\{\hat{\mu} \neq \mu\right\} && (16)\\
&\overset{\text{(a)}}{\geq} \Theta\left(\frac{n\log d}{b}\right)\left(1 - \frac{I(\hat{\mu};\mu)+1}{\log|\Pi_R|}\right) && (17)\\
&\geq \Theta\left(\frac{n\log d}{b}\right)\left(1 - \frac{b+1}{2b}\right) && (18)\\
&= \Theta\left(\frac{n\log d}{b}\right), && (19)
\end{aligned}$$

where (a) follows from Lemma D.1.

Therefore, it suffices to find a Π_R that satisfies the above two criteria. To this end, consider the following construction of Π_R:

$$\Pi_R \triangleq \left\{\left(\frac{n}{R}n_1, \frac{n}{R}n_2, ..., \frac{n}{R}n_d\right) \middle| \sum_i n_i = R, n_j \in \mathbb{Z}_+\right\}.$$

For a given b, we will pick $R = \Theta\left(\frac{b}{\log d}\right)$. It is then straightforward to see that

$$|\Pi_R| = \binom{d+R-1}{R-1} \geq \left(\frac{d+R-1}{R-1}\right)^{R-1} \geq 2^{2b}.$$

In addition, for any distinct $\mu_1, \mu_2 \in \Pi_R$, $\|\mu_1 - \mu_2\|_\infty \geq \frac{n}{R} = \Theta\left(\frac{n\log d}{b}\right)$.

Bounding the ℓ_2 distortion. We follow the same steps of analysis as in the ℓ_∞ case (with $\|\cdot\|_\infty$ being replaced by $\|\cdot\|_2^2$), except for requiring the set Π_R to satisfy

- $|\Pi_R| \geq 2^{\theta(b)}$;
- for any $\mu_1, \mu_2 \in \Pi_R$ such that $\mu_1 \neq \mu_2$, $\|\mu_1 - \mu_2\|_2^2 \geq \Theta\left(\frac{n^2 \log d}{b}\right)$.

The construction of Π_R under ℓ_2 loss is slightly more involved than that in the ℓ_∞ case, but the central idea is to obtain a set Π_R that matches a packing lower bound, similar to the proof of the GV bound.

We begin with a few notations: Let $\mathcal{H}_R$ be the Hamming surface with radius R (over a d-dimensional cube), i.e., $\mathcal{H}_R \triangleq \left\{(n_1, ..., n_d) \middle| \sum_{i=1}^d n_i = R, n_i \in \{0,1\}\right\}$. Now, we construct a $\tilde{\Pi}_R \subset \mathcal{H}_R$, such that for any distinct $\pi_1, \pi_2 \in \tilde{\Pi}_R$, $d_H(\pi_1, \pi_2) \geq \frac{R}{8}$ (where $d_H(\cdot,\cdot)$ is the Hamming distance between π_1 and π_2, i.e. $\sum_{j=1}^d \mathbb{1}_{\{\pi_1(i) \neq \pi_2(i)\}}$).

We claim that there exists such $\tilde{\Pi}_R$ with $\left|\tilde{\Pi}_R\right| = 2^{\theta(R \log d)}$, when $R = o(d)$. To see this, let $\tilde{\Pi}_R$ be the largest subset that satisfies the requirement. Then this would imply that for any $\pi \in \mathcal{H}_R$ there exists a $\tilde{\pi} \in \tilde{\Pi}_R$, such that $d_H(\pi, \tilde{\pi}) \leq R/4$ (otherwise, one can add π into $\tilde{\Pi}_R$ while still satisfying the requirement). This would imply the following covering bound:

$$|\mathcal{H}_R| \leq \left|\tilde{\Pi}_R\right| \cdot \left|\{\pi \in \mathcal{H}_R : d_H(\pi, \tilde{\pi}) \leq R/4\}\right|. \tag{20}$$

Now, notice that $|\mathcal{H}_R| = \binom{d}{R}$, and the volume of the Hamming ball can be upper bounded by

$$\begin{aligned}
\left|\{\pi \in \mathcal{H}_R : d_H(\pi, \tilde{\pi}) \leq R/4\}\right| &= \sum_{i=1}^{R/8} \binom{d-R}{i}\binom{R}{i} \\
&\leq \binom{d-R}{R/8} \sum_{i=0}^{R/8} \binom{R}{i} \\
&\leq d^{R/8} \cdot 2^{R h_{\mathsf{b}}(1/8)},
\end{aligned}$$

where in the last inequality we use upper bound on binomial partial sum: $\sum_{i=1}^k \binom{R}{k} \leq 2^{R h_{\mathsf{b}}\left(\frac{k}{R}\right)}$ where $h_{\mathsf{b}}(\cdot)$ is the binary entropy function.

Plugging the upper bound into (20), we obtain

$$\left|\tilde{\Pi}_R\right| \geq \frac{\binom{d}{R}}{d^{R/8} \cdot 2^{R h_{\mathsf{b}}(1/8)}} = 2^{\left(R \log\left(\frac{d}{R}\right) - R\left(\frac{1}{8} + h_{\mathsf{b}}\left(\frac{1}{8}\right)\right)\right)} = 2^{\Theta(R \log d)} = 2^{\Theta(b)},$$

when $d \gg R$ and $R = \Theta\left(\frac{b}{\log d}\right)$.

Finally, we rescale $\tilde{\Pi}_R$ to obtain Π_R: $\Pi_R \triangleq \left\{\frac{n}{R}\pi : \pi \in \tilde{\Pi}_R\right\}$. Obviously, we have $|\Pi_R| = \left|\tilde{\Pi}_R\right| \geq 2^{\Theta(b)}$. Moreover, for any distinct $\mu_1, \mu_2 \in \Pi_R$, $\|\mu_1 - \mu_2\|_2^2 \geq d_H(\mu_1, \mu_2) \cdot \frac{n^2}{R^2} = \Theta\left(\frac{n^2 \log d}{b}\right)$. This completes the lower bound on $\mathsf{err}(b)$ under the ℓ_2 loss.

D.6 Proof of Lemma 5.1

Let $\mathcal{M}'_j \triangleq \mathcal{M}_j(X'_j)$. By security constraint (S1), we know that $I\left(\mathcal{M}_1, ..., \mathcal{M}_n; Y_1, ..., Y_n \middle| \sum_i \mathcal{M}_i\right) = 0$. Therefore, we must have

$$\begin{aligned}
\mathbb{P}\left\{(Y_1, ..., Y_n)\right\} &= \mathbb{P}\left\{(Y_1, ..., Y_n) \middle| \sum_i \mathcal{M}_i\right\} \cdot \mathbb{P}\left\{\sum_i \mathcal{M}_i\right\} \\
&\leq \mathbb{P}\left\{(Y_1, ..., Y_n) \middle| \sum_i \mathcal{M}_i\right\} \cdot \left(e^{\varepsilon} \mathbb{P}\left\{\sum_{i \neq j} \mathcal{M}_i + \mathcal{M}'_j\right\} + \delta\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left\{(Y_1,...,Y_n)\middle|\sum_{i\neq j}\mathcal{M}_i+\mathcal{M}_j'\right\}\cdot\left(e^{\varepsilon}\mathbb{P}\left\{\sum_{i\neq j}\mathcal{M}_i+\mathcal{M}_j'\right\}+\delta\right)\\
&\leq e^{\varepsilon}\mathbb{P}\left\{\sum_{i\neq j}\mathcal{M}_i+\mathcal{M}_j'\right\}\mathbb{P}\left\{(Y_1,...,Y_n)\middle|\sum_{i\neq j}\mathcal{M}_i+\mathcal{M}_j'\right\}+\delta\\
&= \mathbb{P}\left\{\left(Y_1,...,Y_j',...,Y_n\right)\right\},
\end{aligned}
$$

where the second inequality is due to the DP assumption of $\sum_i \mathcal{M}_i$, the third equality is due to the fact of $I\left(\mathcal{M}_1,...,\mathcal{M}_n;Y_1,...,Y_n|\sum_i\mathcal{M}_i\right)=0$.