Breaking the Communication-Privacy-Accuracy Trilemma

Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür

***Abstract*— Two major challenges in distributed learning and estimation are 1) preserving the privacy of the local samples; and 2) communicating them efficiently to a central server, while achieving high accuracy for the end-to-end task. While there has been significant interest in addressing each of these challenges separately in the recent literature, treatments that simultaneously address both challenges are still largely missing. In this paper, we develop novel encoding and decoding mechanisms that simultaneously achieve optimal privacy and communication efficiency in various canonical settings. In particular, we consider the problems of mean estimation and frequency estimation under ε-local differential privacy and b-bit communication constraints. For mean estimation, we propose the SQKR mechanism, a scheme based on Kashin's representation and random sampling, with order-optimal estimation error under both constraints. We further apply SQKR to distributed SGD and obtain a communication efficient and (locally) differentially private distributed SGD protocol. For frequency estimation, we present the RHR mechanism, a scheme that leverages the recursive structure of Walsh-Hadamard matrices and achieves order-optimal estimation error for *all* privacy levels and communication budgets. As a by-product, we also construct a distribution estimation mechanism that is rate-optimal for all privacy regimes and communication constraints, extending recent work that is limited to $b = 1$ and $\varepsilon = O(1)$. Our results demonstrate that intelligent encoding under joint privacy and communication constraints can yield a performance that matches the optimal accuracy achievable under either constraint alone. In other words, the optimal performance is determined by the more stringent of the two constraints, and the less stringent constraint can be satisfied for free.**

***Index Terms*— Differential privacy, distributed estimation, communication, Kashin's representation, stochastic gradient descend.**

I. Introduction

THE rapid growth of large-scale datasets has been stimulating interest in and demands for distributed learning and estimation, where datasets are often too large and too sensitive to be stored on a centralized machine. When data is distributed across multiple devices, communication cost often becomes a bottleneck of modern machine learning tasks [1]. This is even more so in federated learning type settings, where communication occurs over bandwidth-limited wireless links [2]. Moreover, as more personal data is entrusted to data aggregators, in many applications it carries sensitive individual information, and hence finding ways to protect individual privacy is of crucial importance. In particular, local differential privacy (LDP) [3], [4], [5], [6] is a widely adopted privacy paradigm, which guarantees that the outcome from a privatization mechanism will not release too much individual information statistically. In this paper, we study the relationship between utility (often in forms of accuracy for certain statistical tasks), privacy, and communication *jointly*.

At first glance, privacy and communication may seem to be in conflict with each other: achieving privacy requires the addition of noise, therefore increasing the entropy of the data and making it less compressible. For instance, consider the mean estimation problem, which appears as a fundamental subroutine in many distributed optimization tasks, e.g. distributed stochastic gradient descent (SGD). Here, the goal is to estimate the empirical mean of a collection of d-dimensional vectors. If we first privatize each vector via $\mathsf{privUnit}$ in [7] (which is optimal under LDP constraints) and then quantize via the $\mathsf{RandomSampling}$ quantizer in [8] (which is optimal under communication constraints), a tedious but straightforward calculation shows that the resulting ℓ_2 estimation error grows with d^2. However, this is far from matching the error rate under each constraint separately, which has a linear dependence on d. A similar phenomenon happens in the distribution estimation problem, where each client's data is drawn independently from a discrete distribution $\boldsymbol{p}$ with domain size d. One can satisfy both constraints by first perturbing the data via the Subset Selection (SS) mechanism [9] (which is optimal under LDP constraints) and then quantizing the noised data to b bits. Again, it can be shown that under such strategy, the ℓ_2 estimation error of $\boldsymbol{p}$ has a quadratic dependence on d. This leaves a huge gap to the lower bounds under each constraint separately, which have a linear dependence on d. See Section A of the appendix for a detailed discussion.

While there has been significant recent progress on understanding how to achieve optimal accuracy under separate privacy [9], [10] and communication [11], [12] constraints, as illustrated above a simple concatenated application of these optimal schemes can yield a highly suboptimal performance. Recent works that attempt to break this communication-privacy-accuracy trilemma have been either limited to specific

Manuscript received 2 March 2022; revised 15 August 2022; accepted 17 October 2022. Date of publication 9 November 2022; date of current version 20 January 2023. This work was supported in part by NSF under Award NeTS-1817205 and Award CCF-2213223, in part by the Stanford Graduate Fellowship, and in part by the Google Research Award. *(Corresponding author: Wei-Ning Chen.)*

Wei-Ning Chen and Ayfer Özgür are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: wnchen@stanford.edu; aozgur@stanford.edu).

Peter Kairouz is with Google Research, Seattle, WA 98101 USA (e-mail: kairouz@google.com).

Communicated by M. Skoglund, Associate Editor At Large for Security and Privacy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3218772.

Digital Object Identifier 10.1109/TIT.2022.3218772

regimes or, as we show, are far from optimal. For example, [13] provides a 1-bit ε-LDP scheme for distribution estimation which is order-optimal only in the low communication regime ($b = O(1)$) and high privacy regime ($\varepsilon = O(1)$), while [8] tries to address both constraints in the mean estimation setting, but the error rate achieved under their mechanism is quadratic in d and therefore does not improve on the above baseline. We note that the general privacy regime (i.e. $\varepsilon = \Omega(1)$) is also of both theoretical and practical interest. For instance, when $n = \Omega(d)$, one can combine LDP with amplification techniques [14], [15], [16] to ensure stronger central differential privacy.

This paper closes the above gaps for any given privacy level ε and communication budget b. Indeed, our results show that the fundamental trade-offs are determined by the more stringent of the two constraints, and with careful encoding we can satisfy the less stringent constraint *for free*, thus breaking the privacy-communication-accuracy trilemma. For the same privacy level ε, this allows us to achieve the accuracy of existing mechanisms in the literature with a drastically smaller communication budget, or equivalently, for the same communication budget achieve higher privacy. It also explains, for example, why 1-bit communication budget is sufficient under the high privacy regime [13], [17]. We will demonstrate this phenomenon in various canonical tasks and answer the following question: "given arbitrary privacy budget ε and communication budget b, what are the fundamental limits for estimation accuracy?" We next formally define the settings and the problem formulations we consider in this paper.

A. Problem Formulation

The general distributed statistical tasks we consider in this paper can be formulated as follows: each one of the n clients has local data $X_i \in \mathcal{X}$ and sends a message $Y_i \in \mathcal{Y}$ to the server, who upon receiving Y^n aims to estimate some pre-specified quantity of X^n. Note that X^n are not necessarily drawn from some distribution. At client i, the message Y_i is generated via some mechanism (a randomized mapping that possibly uses shared randomness across participating clients and the server) denoted by a conditional probability $Q_i(y|X_i)$ satisfying the following constraints.

Local differential privacy: Let $(\mathcal{Y}, \mathcal{B})$ be a measurable space, and $Q(\cdot|x)$ be probability measures for all $x \in \mathcal{X}$, with $\{Q(\cdot|x)|x \in \mathcal{X}\}$ dominated by some σ-finite measure μ so that the density $Q(y|x)$ exists. A mechanism Q is ε-LDP if

$$\forall x, x' \in \mathcal{X},\, y \in \mathcal{Y},\, \frac{Q(y|x)}{Q(y|x')} \leq e^{\varepsilon}.$$

b-bit communication constraint: $\mathcal{Y}$ satisfies b-bit communication constraint if each of its elements can be described by b bits, i.e. $|\mathcal{Y}| \leq 2^b$.

The goal is to jointly design a mechanism (on the clients' sides) and an estimator (on the server side) so that the accuracy of estimating some target function $\sum_{i=1}^n f(X_i)$ is maximized. In this paper, we are mainly interested in the *distribution-free* framework; that is, we do not assume any underlying distribution on X_i, but we also demonstrate that our results can be extended to probabilistic settings. To this end, we will focus on the following four canonical tasks.

Mean estimation: For real-valued data, we consider the d-dimensional unit euclidean ball $\mathcal{X} = \mathcal{B}_d(\mathbf{0}, 1)$ and are interested in estimating the *empirical mean* $\bar{X} \triangleq \frac{1}{n}\sum_i X_i$. The goal is to minimize the worst-case ℓ_2 estimation error defined as

$$r_{\mathtt{ME}}\left(\ell_2, \varepsilon, b\right) \triangleq \min_{\left(\hat{X}, Q^n\right)} \max_{X^n \in \mathcal{X}^n} \mathbb{E}\left[\left\|\hat{X} - \bar{X}\right\|_2^2\right], \qquad (1)$$

where Q^n satisfies ε-LDP and b-bit communication constraints. When the context is clear, we may omit ε and b in $r_{\mathtt{ME}}\left(\ell, \varepsilon, b\right)$.

Statistical mean estimation: In the probabilistic version of the mean estimation problem, we assume that X_i's are drawn from some common but unknown distribution P supported on $\mathcal{B}_d(\mathbf{0}, 1)$, the goal is to estimate the *statistical mean* $\theta\left(P\right) = \mathbb{E}_P\left[X_1\right]$ and to minimize the ℓ_2 estimation error:

$$r_{\mathtt{SME}}\left(\ell_2, \varepsilon, b\right) \triangleq \min_{\left(\hat{\theta}, Q^n\right)} \max_{X^n \in \mathcal{X}^n} \mathbb{E}\left[\left\|\hat{\theta}\left(X^n\right) - \theta\left(P\right)\right\|_2^2\right].$$

Frequency estimation: When $\mathcal{X}$ consists of categorical data, i.e. $\mathcal{X} = [d] = \{1, \ldots, d\}$, we are interested in estimating $D_{X^n}(x) \triangleq \frac{1}{n}\sum_i \mathbb{1}_{\{X_i = x\}}$ for $x \in [d]$. With a slight abuse of notation, D_{X^n} is viewed as a vector $(D_{X^n}(1), \ldots, D_{X^n}(d))$ lying in the d-dimensional probability simplex. The worst-case estimation error is defined by

$$r_{\mathtt{FE}}\left(\ell, \varepsilon, b\right) \triangleq \min_{\left(\hat{D}, Q^n\right)} \max_{X^n \in \mathcal{X}^n} \mathbb{E}\left[\ell\left(\hat{D}, D_{X^n}\right)\right],$$

where $\ell = \|\cdot\|_\infty, \|\cdot\|_1$, or $\|\cdot\|_2^2$ and again Q^n satisfies ε-LDP and b-bit communication constraints.

Distribution estimation: A closely related setting is that of discrete distribution estimation, where we assume that the X_i's are drawn independently from a discrete distribution $\boldsymbol{p}$ on the alphabet $\mathcal{X} = [d]$, and the goal is to estimate $\boldsymbol{p}$. In this case, the worst-case error is given by

$$r_{\mathtt{DE}}\left(\ell, \varepsilon, b\right) \triangleq \inf_{(Q^n, \hat{\boldsymbol{p}})} \sup_{\boldsymbol{p} \in \mathcal{P}_d} \mathbb{E}\left[\ell(\hat{\boldsymbol{p}}, \boldsymbol{p})\right],$$

where $\mathcal{P}_d$ is the d-dimensional probability simplex.

We note that these canonical tasks serve as fundamental subroutines in many distributed optimization and learning problems. For instance, the convergence rate of distributed SGD is determined by the ℓ_2 error of estimating the mean of the local gradient vectors (see [18] for more on this connection). Lloyd's algorithm [19] for k-means clustering or the power-iteration method for PCA can also be reduced to the mean estimation task.

Remark 1.1: In this work, we generally assume the availability of shared randomness across the participating clients and the server. In this case, the encoding functions at each node can be explicitly denoted as $Q_i(y|X_i, U)$ where U is a shared random variable that is independent of data, referred to as a public coin. U is also available at the server and the estimator implicitly depends on U. In our notation, we suppress this dependence on U for simplicity. The entropy of U is referred as the amount of shared randomness needed by a scheme.

TABLE I

COMPARISON BETWEEN OUR MEAN ESTIMATION SCHEME AND VQSGD [8]. OUR SCHEME APPLIES TO GENERAL COMMUNICATION AND PRIVACY REGIMES, AND ACHIEVES OPTIMAL ESTIMATION ERROR FOR ALL SCENARIOS

	Privacy	Comm.	ℓ_2 error
SQKR (Thm. 3.1)	$\forall\, \varepsilon > 0$	$\forall\, b > 0$	$O\left(\frac{d}{n \min\left(\varepsilon^2, \varepsilon, b\right)}\right)$
Cross-polytope [8]	$\varepsilon = \Omega\left(1\right)$	$b = \Omega\left(\log d\right)$	$O\left(\frac{d^2}{n}\right)$
Simplex [8]	$\varepsilon = \Omega\left(\log d\right)$	$b = \Omega\left(\log d\right)$	$O\left(\frac{d}{n}\right)$

In Section V, we discuss the amount of shared randomness required by our schemes in order to achieve the optimal estimation error. We point out that in the statistical settings (i.e. statistical mean estimation and distribution estimation), the optimal estimation error can be achieved without shared randomness.

Notation: Throughout this paper, we use $[d]$ to denote the set of $\{1, \ldots, m\}$ for any $d \in \mathbb{N}$. For two sets $\mathcal{S}_1$ and $\mathcal{S}_2$, let $\mathcal{S}_1 \setminus \mathcal{S}_2 \triangleq \{j | j \in \mathcal{S}_1 \text{ and } j \notin \mathcal{S}_2\}$.

We also make use of Bachmann-Landau asymptotic notation: for two positive sequences a_n and b_n, if $\lim_{n\to\infty} a_n/b_n \leq C$ for some $C > 0$, we denote $a_n = O(b_n)$ or $a_n \preceq b_n$; on the other hand, $\lim_{n\to\infty} a_n/b_n = 0$ will be denoted as $a_n = o(b_n)$ (or $a_n \prec b_n$). Similarly, when $\lim_{n\to\infty} a_n/b_n \geq C$, we write $a_n = \Omega(b_n)$ or $a_n \succeq b_n$, and when $\lim_{n\to\infty} a_n/b_n = \infty$, we write $a_n = \omega(b_n)$ or $a_n \succ b_n$. Finally, we use $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ hold.

II. Prior Work

Previous works in the mean estimation problem [8], [12], [20], [21], [22], [23] mainly focus on reducing communication cost, for instance, by random rotation [12] and sparsification [20], [21], [24], [25]. Among them, [8] considers LDP simultaneously. It proposes vector quantization and takes privacy into account, developing a scheme for $\varepsilon = \Theta(1)$ and $b = \Theta(\log d)$ with estimation error $O(d^2/n)$. In contrast, the scheme we develop in Theorem 3.1 achieves an estimation error $O(d/n)$ when $\varepsilon = \Theta(1)$ and $b = \Theta(\log d)$. Moreover, our scheme is applicable for any ε and b and achieves the optimal estimation error, which we show by proving a matching information-theoretic lower bound. See Table I for a comparison of our results with [8]. A key step in our scheme is to pre-process the local data via Kashin's representation [26]. While various compression schemes, based on quantization, sparsification, and dithering have been proposed in the recent literature and Kashin's representation for communication efficiency [27], [28], [29], [30] or for LDP [31] has also been explored in a few works, it is particularly powerful in the case of joint communication and privacy constraints as it helps spread the information in a vector evenly in every dimension. This helps mitigate the error due to subsequent noise introduced by privatization and compression.

The recent works of [32] and [33] also consider estimating empirical mean under ε-LDP. They show that if the data is from a d-dimensional unit ℓ_∞ ball, i.e. $X_i \in [-1, 1]^d$, then directly quantizing, sampling and perturbing each entry can achieve optimal ℓ_∞ estimation error that matches the LDP lower bound in [34], where their privatization steps are based on techniques developed in [34] and [7]. Nevertheless, their approach does not yield good ℓ_2 error in general. Indeed, as in the case of separation schemes discussed in Section A, the ℓ_2 error of their scheme can grow with d^2. We emphasize that in many applications the ℓ_2 estimation error (i.e. MSE) is a more appropriate measure than ℓ_∞. For instance, [18] shows a direct connection between the MSE in mean estimation and the convergence rate of distributed SGD.

Frequency estimation under local differential privacy has been studied in [35], where they propose schemes for estimating the frequency of an individual symbol and minimizing the variance of the estimator. Some of their schemes, while matching the information-theoretic lower bound on ℓ_2 estimation error under privacy constraints, require large communication. For instance, the scheme Optimal Unary Encoding (OUE), which can be viewed as an asymmetric version of RAPPOR [36], achieves optimal ℓ_2 estimation error, but the communication required is $O(d)$ bits, which, as we show in this work, can be reduced to $O(\min(\lceil\varepsilon\rceil, \log d))$ bits. We do this by developing a new scheme for frequency estimation under joint privacy and communication constraints. We establish the optimality of our proposed schemes by deriving matching information-theoretic lower bounds on $r_{\mathrm{FE}}\left(\ell_2, \varepsilon, b\right)$.

Frequency estimation is also closely related to heavy hitter estimation [10], [13], [17], [36], [37], [38], [39], where the goal is to discover symbols that appear frequently in a given data set and estimate their frequencies. This can be done if the error of estimating the frequency of each individual symbol can be controlled uniformly (i.e. by a common bound), and thus is equivalent to minimizing the ℓ_∞ error of estimated frequencies, i.e. $r_{\mathrm{FE}}\left(\ell_\infty, \varepsilon, b\right)$. It is shown in [10] that in the high privacy regime $\varepsilon = O(1)$, $r_{\mathrm{FE}}\left(\ell_\infty, \varepsilon, b\right) = \Theta\left(\sqrt{\frac{\log d}{n\varepsilon^2}}\right)$, and this rate can be achieved via a 1-bit public-coin scheme that has a runtime almost linear in n [17]. An extension, which we describe in Section IV-B, generalizes the achievability in [10] to arbitrary ε and b, achieving $r_{\mathrm{FE}}\left(\ell_\infty, \varepsilon, b\right) = O\left(\sqrt{\frac{\log d}{n \min\left(\varepsilon^2, \varepsilon, b\right)}}\right)$.

After the initial conference version of this paper [40], [41] shows that the optimal private mean estimation scheme privUnit [7] and private frequency scheme asymmetric RAPPOR [35], [36] can be efficiently and losslessly compressed by having clients communicate seeds

TABLE II

COMPARISON OF DIFFERENT FREQUENCY ESTIMATION SCHEMES

	Loss	Estimation error	Communication
Asymmetric RAPPOR	ℓ_2	$\Theta\left(\frac{d}{n\min\left((e^\varepsilon-1)^2, e^\varepsilon\right)}\right)$	d bits
RHR (this work, Thm 4.1)	ℓ_2	$\Theta\left(\frac{d}{n\min\left((e^\varepsilon-1)^2, e^\varepsilon\right)}\right)$	$\min\left(\lceil\varepsilon\rceil, \log d\right)$ bits
Heavy hitter	ℓ_∞	$\Theta\left(\sqrt{\frac{\log d}{n\min\left(\varepsilon, \varepsilon^2\right)}}\right)$	$\lceil\varepsilon\rceil$ bits

TABLE III

COMPARISON BETWEEN LDP DISTRIBUTION ESTIMATION SCHEMES. UNDER THE SAME PRIVACY GUARANTEE, OUR SCHEME IS MORE COMMUNICATION EFFICIENT WHILE ACHIEVING THE SAME ACCURACY

Privacy regime	$\varepsilon \in (0,1)$	$\varepsilon \in (1, \log d)$
SS	d bits	$\max\left(\frac{d}{e^\varepsilon}, \log d\right)$
HR	$\log d$ bits	$\log d$ bits
1bit-HR	1 bit	-
RHR (this work, Thm. 4.1)	1 bit	$\min\left(\lceil\varepsilon\rceil, \log d\right)$

(or indices of seeds if shared randomness is available) of a cryptographically-secure pseudo number generator (PRG) to the server. With a careful sampling of random seeds (e.g., via rejection sampling), this method also achieves the optimal privacy-communication-accuracy trade-off shown in our results. However, their results rely on the assumption of the existence of an *exponentially strong* PRG. In addition, the decoding time complexity of their methods is more expensive than ours. For example, for frequency estimation, the server runtime of their scheme is $\tilde{O}\left(nd\right)$ as opposed to ours $O\left(n + d\log d\right)$ (see Remark 4.2).

We compare our scheme and existing results in Table II.

If we further assume X^n are drawn from some discrete distribution $\boldsymbol{p}$, then the problem falls into distribution estimation under local differential privacy [9], [13], [34], [36], [42], [43], [44], [45], [46] and limited communication [11], [25], [45], [46], [47], [48], [49], [50]. Tight lower bounds are given separately: for instance [9], [44] shows $r_{\mathrm{DE}}\left(\ell_1, \varepsilon, \log d\right) = \Omega\left(\sqrt{\frac{d^2}{n\min((e^\varepsilon-1)^2, e^\varepsilon)}}\right)$ and [49] shows $r_{\mathrm{DE}}\left(\ell_1, \infty, b\right) = \Omega\left(\sqrt{\frac{d^2}{n2^b}}\right)$.

We show that these lower bounds can be achieved simultaneously (Theorem 4.1). Our result recovers the result of [13] when $b = 1$ and $\varepsilon = O(1)$ as a special case. See Table III for a comparison.

III. MEAN ESTIMATION

In the mean estimation problem, each client has a d-dimensional vector X_i from the Euclidean unit ball, and the goal is to estimate the empirical mean $\bar{X} = \frac{1}{n}\sum_i X_i$ under ε-LDP and b bits communication constraints. This problem has applications in private and communication-efficient distributed SGD. The following theorem characterizes the optimal ℓ_2 estimation error for this setting.

Theorem 3.1: For mean estimation under ε-LDP and b-bit communication constraints, we can achieve

$$r_{\mathrm{ME}}\left(\ell_2, \varepsilon, b\right) = O\left(\frac{d}{n\min\left(\varepsilon^2, \varepsilon, b\right)}\right). \tag{2}$$

Moreover, if $\min(\varepsilon^2, \varepsilon, b) = o(d)$ and $n \cdot \min(\varepsilon^2, \varepsilon, b) > d$, the above error is optimal.

Note that by taking $\varepsilon \to \infty$ for a fixed b, or by taking $b \to \infty$ for a fixed ε in part (i), Theorem 3.1 provides the optimal error when we have the corresponding constraint alone. Furthermore, for finite ε and b we see that the optimal error is dictated by the error due to one of these constraints, the one that leads to a larger error, and hence the less stringent constraint is satisfied for free. This also implies that to achieve the optimal accuracy under ε-LDP constraints, we do not need more than $\lceil\varepsilon\rceil$ bits.

The lower bounds are obtained by connecting the problem to a specific parametric estimation problem with a distribution supported on the unit ball. To match this lower bound, we propose a public-coin scheme, Subsampled and Quantized Kashin's Response (SQKR), based on Kashin's representation [26] and random sampling.

Remark 3.1: We note that the two conditions for optimality in the theorem are standard and are needed to restrict the problem to the interesting parameter regime. To see this, observe that if the first lower bound condition $\min(\varepsilon^2, \varepsilon, b) = o(d)$ is not met, then it would imply the privatization error of our proposed scheme is $o(1/n)$, which is dominated by the sampling error $O(1/n)$, and thus asymptotically the privacy and communication constraints cause no effect to the accuracy. In addition, $\varepsilon = \Omega(d)$ is typically not a meaningful regime, as it preserves very weak privacy in practice.

On the other hand, if the second condition is not satisfied, it would imply $\frac{d}{n\min(\varepsilon^2, \varepsilon, b)} \geq O(1)$. In this case, the server could always output $\mathbf{0} \in \mathbb{R}^d$ as the final estimator regardless of local samples – which requires no communication from the server, and still obtain an $O(1)$ ℓ_2 error. Indeed, the lower bounds from [50] and [51] imply that under this regime, $\Omega(1)$ error is inevitable and that the trivial achievability scheme (i.e., having the server always output $\mathbf{0}$) achieves it.

A. Subsampled and Quantized Kashin's Response (Achievability of Theorem 3.1)

For each observation X_i, we aim to construct an unbiased estimator $\hat{X}_i$ which is ε-LDP, can be described in b bits, and has a small variance. Towards this goal, our general strategy is to quantize, subsample, and privatize the data X_i. However, before this, it is crucial to pre-process each X_i by a carefully designed mechanism to increase the robustness of the signal to noise introduced by sampling and privatization.

1) Kashin's Representation and Randomized Rounding: We first introduce the idea of a tight frame in Kashin's representation. We begin by introducing tight frames and Kashin's representation [26].

Definition 3.1 (Tight Frame): A tight frame is a set of vectors $[u_1, \ldots, u_N] \in \mathbb{R}^{d \times N}$ that obeys Parseval's identity

$$\|x\|_2^2 = \sum_{j=1}^{N} \langle u_j, x \rangle^2, \text{ for all } x \in \mathbb{R}^d.$$

A frame can be viewed as a generalization of an orthogonal basis in $\mathbb{R}^d$, which can improve the encoding stability by adding redundancy to the representation system when $N > d$. To increase robustness, we wish the information to spread evenly in each coefficient.

Definition 3.2 (Kashin's Representation): For a set of vectors $[u_1, \ldots, u_N]$, we say the expansion

$$x = \sum_{j=1}^{N} a_j u_j, \text{ with } \max_j |a_j| \leq \frac{K}{\sqrt{N}} \|x\|_2$$

is a Kashin's representation of vector x at level K.

Therefore, if we can obtain unbiased estimators $(\hat{a}_1, \ldots, \hat{a}_N) \in \mathbb{R}^N$ of the Kashin's representation of X with respect to a tight frame $[u_1, \ldots, u_N]$, then the MSE can be controlled by

$$\begin{aligned}\mathbb{E}\left[\left(\hat{X} - X\right)^2\right] &= \mathbb{E}\left[\left\|\sum_{j=1}^{N} (\hat{a}_j - a_j) u_j\right\|_2^2\right] \\ &\overset{\text{(a)}}{\leq} \mathbb{E}\left[\sum_{j=1}^{N} (\hat{a}_j - a_j)^2\right] \\ &= \sum_{j=1}^{N} \mathsf{Var}(\hat{a}_j), \qquad (3)\end{aligned}$$

where (a) is due to the Cauchy-Schwarz inequality and the definition of a tight frame. Recall that X is deterministic, so here the expectation is taken with respect to the randomness on $\hat{a}_j$. Notice that the cardinality N of the frame determines the compression (i.e. quantization) rate, and Kashin's level K affects the variance. Hence we are interested in constructing tight frames with small N and K.

[26] shows that if $N > (1 + \mu) d$ for some $\mu > 0$, then there exists a tight frame $[u_1, \ldots, u_N]$ such that for any $x \in \mathbb{R}^d$, one can find a Kashin's representation at level $K = \Theta(1)$:

Lemma 3.1 (Uncertainty Principle and Kashin's Representation): For any $\mu > 0$ and $N > (1 + \mu)d$, there exists a tight frame $[u_1, \ldots, u_N]$ with Kashin's level $K = O\left(\frac{1}{\mu^3} \log \frac{1}{\mu}\right)$. Moreover, for each X, finding Kashin's coefficient requires $O(dN \log N)$ computation.

For our purpose, we choose μ to be a constant, i.e. $\mu = \Theta(1)$, so $N = \Theta(d), K = \Theta(1)$, and we can obtain representation of $X = \sum_{j=1}^{N} a_j u_j$, with $|a_j| \leq \frac{K}{\sqrt{N}} = \frac{c}{\sqrt{d}}$ for some constant c. Therefore, we quantize each a_j as follows:

$$q_j \triangleq \begin{cases} -\frac{c}{\sqrt{d}}, \text{ with probability } \frac{c/\sqrt{d} - a_j}{2c/\sqrt{d}} \\ \frac{c}{\sqrt{d}}, \text{ with probability } \frac{a_j + c/\sqrt{d}}{2c/\sqrt{d}}. \end{cases} \qquad (4)$$

$\boldsymbol{q} \triangleq (q_1, \ldots, q_N)$ yields an unbiased estimator of $\boldsymbol{a} \triangleq (a_1, \ldots, a_N)$ and can be described by $N = \Theta(d)$ bits.

2) Sampling: To further reduce the communication cost to $k = \min(\lceil \epsilon \rceil, b)$ bits, we sample k bits uniformly at random from $\boldsymbol{q}$ using public randomness. Let $s_1, \ldots, s_k \overset{\text{i.i.d.}}{\sim}$ uniform$[N]$ be the indices of the sampled elements, and define the sampled message as

$$Q\left(\boldsymbol{q}, (s_1, \ldots, s_k)\right) \triangleq (q_{s_1}, \ldots, q_{s_k}) \in \left\{-c/\sqrt{d}, c/\sqrt{d}\right\}^k. \qquad (5)$$

Then Q can be described in k bits, and each of q_{s_m} yields an independent and unbiased estimator of $\boldsymbol{a}$: for all $j \in [N]$

$$\begin{aligned}\mathbb{E}\left[N q_{s_m} \cdot \mathbb{1}_{\{j = s_m\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[N q_{s_m} \cdot \mathbb{1}_{\{j = s_m\}} \middle| q_1, \ldots, q_N\right]\right] \\ &= \mathbb{E}\left[q_j\right] = a_j. \qquad (6)\end{aligned}$$

3) Privatization: Each client then perturbs Q via 2^k-RR mechanism (as a k-bit string):

$$\tilde{Q} = \begin{cases} Q, \text{ with probability } \frac{e^\varepsilon}{e^\varepsilon + 2^k - 1} \\ Q' \in \left\{-\frac{c}{\sqrt{d}}, \frac{c}{\sqrt{d}}\right\}^k \setminus \{Q\}, \text{ with probability } \frac{1}{e^\varepsilon + 2^k - 1}, \end{cases} \qquad (7)$$

where recall that $\left\{-c/\sqrt{d}, c/\sqrt{d}\right\}^k \setminus \{Q\}$ denotes the difference between two sets, i.e., removing Q from $\left\{-\frac{c}{\sqrt{d}}, \frac{c}{\sqrt{d}}\right\}^k$. Since

$$\sum_{Q' \in \left\{-c/\sqrt{d}, c/\sqrt{d}\right\}^k / \{Q\}} Q' = -Q,$$

$\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right) \tilde{Q}$ yields an unbiased estimator of Q. Indeed, if we write $\tilde{Q} = (\tilde{q}_1, \ldots, \tilde{q}_k)$, then

$$\mathbb{E}\left[\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right) \cdot \tilde{q}_m \middle| q_1, \ldots, q_N, s_1, \ldots, s_k\right] = q_{s_m}, \qquad (8)$$

or equivalently

$$\mathbb{E}\left[\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right) \tilde{Q} \middle| Q\right] = Q.$$

4) Analysis of the ℓ_2 Error: Given $\tilde{Q} = (\tilde{q}_1, \ldots, \tilde{q}_k)$, define

$$\hat{a}_j = \frac{N}{k} \cdot \left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right) \sum_{m=1}^{k} \tilde{q}_m \cdot \mathbb{1}_{\{j = s_m\}}.$$

By (6) and (8), $\mathbb{E}\left[\hat{a}_j\right] = a_j$, and hence $\hat{X}\left(\tilde{Q}, (s_1, \ldots, s_k)\right) \triangleq \sum_{j=1}^{N} \hat{a}_j u_j$ gives an unbiased estimator of X.

Claim 3.1: Let $C > 0$ be some universal positive constant. The MSE of $\hat{X}$ can be bounded by

$$\mathbb{E}\left[\left\|\hat{X} - X\right\|_2^2\right] \leq C \left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right)^2 \frac{d}{k}.$$

Finally, each client encodes its data X_i independently, and the server computes $\frac{1}{n}\sum_i \hat{X}_i$. Since $\hat{X}_i$ is unbiased and by Claim 3.1, we get

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\hat{X}_i-\bar{X}\right\|_2^2\right]=\frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\hat{X}_i-X_i\right\|_2^2\right]$$

$$\leq C\left(\frac{e^{\varepsilon}+2^k-1}{e^{\varepsilon}-1}\right)^2\frac{d}{nk}.$$

Finally, picking $k=\min\left(\lceil\log_2 e\rceil\varepsilon,b\right)$ gives us the desired upper bound. □

Remark 3.2: In order to achieve optimal communication efficiency, SQKR uses public randomness at the sampling step. That being said, we can still turn SQKR into a private scheme by using additional communication. See Section V for more details.

At a high level, SQKR resembles vqSGD [8] as both schemes seek a suitably designed representation for X_i before quantizing it. vqSGD represents X_i by a basis $B=\{b_1,\dots,b_K\}\subset\mathbb{R}^d$ where B is chosen in such a way that its convex hull contains the unit ℓ_2 ball. Therefore we can write $X_i=\sum_{j=1}^{N}a_jb_j$ with $\sum_j a_j=1$. Equivalently, the pre-processing step of vqSGD corresponds to a linear transformation that embeds the d-dim ℓ_2 unit ball into a N-dim ℓ_1 ball. In contrast, Kashin's representation above embeds the d-dim ℓ_2 unit ball into an N-dim ℓ_∞ ball. Therefore, while both schemes have a pre-processing step of a similar flavor, what is achieved by these steps is quite different. The representation of vqSGD is most efficient when it concentrates the information in a few coefficients, while Kashin's representation spreads the information evenly across different coefficients. The first representation serves us well when we only seek to quantize the signal. However, the quantized signal becomes very sensitive to privatization noise. Therefore vqSGD ends up with $O(d^2)$ error in the case of both privacy and communication constraints, while we can achieve $O(d)$ error.

B. Converse of Theorem 3.1

The lower bound of Theorem 3.1 can be obtained by constructing a prior distribution on X_i and analyzing the statistical mean estimation problem. Therefore, we will impose a prior distribution P on $X_1,\dots,X_n$ and lower bound the ℓ_2 error of estimating the mean $\theta(P)$, where P is a distribution supported on the d-dimension unit ball.

For any $\hat{X}$, observe that

$$\begin{aligned}
&\mathbb{E}_{\hat{X},X^n\overset{\text{i.i.d.}}{\sim}P}\left[\left\|\hat{X}-\bar{X}\right\|_2^2\right]\\
&\overset{(a)}{\geq}\mathbb{E}\left[\left(\left\|\hat{X}-\theta\left(P\right)\right\|_2-\left\|\bar{X}-\theta\left(P\right)\right\|_2\right)^2\right]\\
&\geq\mathbb{E}\left[\left\|\hat{X}-\theta\left(P\right)\right\|_2^2\right]-2\mathbb{E}\left[\left\|\hat{X}-\theta\left(P\right)\right\|_2\left\|\bar{X}-\theta\left(P\right)\right\|_2\right]\\
&\overset{(b)}{\geq}\mathbb{E}\left[\left\|\hat{X}-\theta\left(P\right)\right\|_2^2\right]\\
&\qquad-2\sqrt{\mathbb{E}\left[\left\|\hat{X}-\theta\left(P\right)\right\|_2^2\right]\mathbb{E}\left[\left\|\bar{X}-\theta\left(P\right)\right\|_2^2\right]},
\end{aligned}\tag{9}$$

where (a) and (b) follow from the triangular inequality and the Cauchy-Schwartz inequality respectively. Since X_i and $\theta(P)$ are supported on the unit ball, $\mathbb{E}\left[\left\|\bar{X}-\theta\left(P\right)\right\|_2^2\right]\asymp 1/n$, so it remains to find a distribution P^* such that

$$\min_{\hat{X}}\mathbb{E}\left[\left\|\hat{X}-\theta\left(P^*\right)\right\|_2^2\right]\succeq\frac{d}{n\min\left(\varepsilon^2,\varepsilon,b\right)}.$$

Consider the product Bernoulli model $Y\sim\prod_{j=1}^{d}\mathrm{Ber}(\theta_j)$. If we set $\Theta=[1/2-\varepsilon,1/2+\varepsilon]^d$ for some $\frac{1}{2}>\varepsilon>0$, then it can be shown that both variance and sub-Gaussian norm of the score function of this model is $\Theta(1)$ [50, Corollary 4]. Therefore, applying [50, Corollary 8] and [51, Proposition 2, Proposition 4] yields

$$\min_{\hat{\theta}}\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|_2^2\right]\succeq\frac{d^2}{n\min\left(\varepsilon^2,\varepsilon,b\right)}.$$

Finally, if we set $X_i=Y_i/\sqrt{d}$, then each X_i is supported on the unit ball and $\mathbb{E}\left[X_i\right]=\theta/\sqrt{d}$. Therefore

$$\min_{\hat{X}}\mathbb{E}\left[\left\|\hat{X}-\frac{\theta}{\sqrt{d}}\right\|_2^2\right]\succeq\frac{d}{n\min\left(\varepsilon^2,\varepsilon,b\right)}.$$

Plugging into (9), as long as $\min(\varepsilon^2,\varepsilon,k)=o(d)$, the first term dominates and we get the desired lower bound. □

C. Application to Statistical Mean Estimation

Finally, we point out that SQKR easily extends to an optimal scheme for statistical mean estimation, where each local data is drawn from an unknown distribution P supported on $\mathcal{B}_d(\mathbf{0},1)$, and the goal is to estimate the statistical mean. Under the statistical setting, however, SQKR requires no shared randomness, as one can replace the random sampling step with a deterministic grouping and sampling of coordinates across all the clients (see the proof of Corollary 3.1 in Section B-A of the appendix for details). This allows bypassing the use of shared randomness and gives the following result:

Corollary 3.1: For statistical mean estimation under ε-LDP and b bits communication constraint, we can achieve

$$r_{\mathrm{SME}}\left(\ell_2,\varepsilon,b\right)=O\left(\frac{d}{n\min\left(\varepsilon^2,\varepsilon,b,d\right)}\right),\tag{10}$$

without shared randomness. Moreover, if $\min(\varepsilon^2,\varepsilon,b)=o(d)$, the above error is optimal (even in the presence of shared randomness).

IV. Frequency Estimation

Recall that in the frequency estimation problem, given $X_1,\dots X_n\in[d]$, we want to estimate the empirical frequency $D_{X^n}(x)$ under ε-LDP and b bits communication budgets on each X_i. The following theorem characterizes the optimal estimation error achievable in this setting.

Theorem 4.1: For frequency estimation under ε-LDP and b bits communication constraint, we can achieve

(i) $r_{\mathtt{FE}}\left(\ell_2\right)=O\left(\frac{d}{n\min\left\{e^{\varepsilon},(e^{\varepsilon}-1)^2,2^b,d\right\}}\right)$, and

$$r_{\mathtt{FE}}\left(\ell_1\right)=O\left(\frac{d}{\sqrt{n\min\left\{e^{\varepsilon},(e^{\varepsilon}-1)^2,2^b,d\right\}}}\right);$$

(ii) $r_{\mathrm{FE}}\left(\ell_\infty\right) = O\left(\sqrt{\frac{\log d}{n \min\left\{\varepsilon^2,\varepsilon,b\right\}}}\right)$.

Moreover, if $\min\left(e^{\varepsilon}, \left(e^{\varepsilon}-1\right)^2, 2^b\right) = o(d)$ and $n \min\left(e^{\varepsilon}, \left(e^{\varepsilon}-1\right)^2, 2^b\right) \geq d^2$, the errors in (i) are order-optimal.

Note that, similar to Theorem 3.1, Theorem 4.1 shows that for finite ε and b, the error is determined by the error due to one of these constraints, and hence the other less stringent constraint is satisfied for free. It also implies that to achieve the optimal accuracy under ε-LDP constraints, we do not need more than $\min\left(\lceil \log_2 e \cdot \varepsilon \rceil, \log d\right)$ bits. In the rest of the section, we overview the scheme we develop to achieve the optimal error in (2).

We next overview the scheme that achieves the error in (i) of Theorem 4.1. We call this scheme Recursive Hadamard Response (RHR) as it builds on the recursive structure of the Hadamard matrix. The complete proof of Theorem 4.1 can be found in Section IV-B (the achievability part) and Section IV-C (the converse part).

A. Recursive Hadamard Response: An Overview of the Scheme

For notational convenience, we will view D_{X^n} as a d-dimensional vector $\left(D_{X^n}(1), \ldots, D_{X^n}(d)\right)$ and assume X_i is one-hot encoded, i.e. $X_i = \boldsymbol{e}_j$ for some $j \in [d]$, so $D_{X^n} = \frac{1}{n}\sum_i X_i$. We further assume, without of loss of generality, that $d = 2^m$ for some $m \in \mathbb{N}$. Recall that a Hadamard matrix $H_d \in \{-1,+1\}^{d \times d}$ can be constructed in a recursive fashion as

$$H_m = \begin{bmatrix} H_{m/2} & H_{m/2} \\ H_{m/2} & -H_{m/2} \end{bmatrix},$$

where $H_1 = [1]$. It can be easily shown that $H_d^{-1} = H_d/d$.

Instead of directly estimating D_{X^n}, our strategy is to first estimate $H_d \cdot D_{X^n}$ and then perform the inverse transform H_d^{-1} to get an estimate for D_{X^n}. So each client will transmit information about $Y_i \triangleq H_d \cdot X_i \in \{-1,1\}^d$ rather than its original data X_i.

1) The 1-bit Case: In this case, each client transmits a uniformly at random chosen entry of Y_i via any 1-bit LDP channel (for instance, using the 2-randomized response (RR) scheme [3], [43], [52]). Once receiving all the bits of the clients, the server can construct an unbiased estimator of Y_i (since the randomness is public the server knows which entry is chosen for communication by each client). It turns out that this simple 1-bit scheme achieves optimal ℓ_1 (and ℓ_2) error $\Theta(\sqrt{d^2/n\varepsilon^2})$ in the high privacy regime $\varepsilon < 1$. This idea is not new and has been used in heavy hitter estimation [17] and distribution estimation [13]. However, a key question remains: how do we minimize the error given an arbitrary communication budget b and privacy level ε?

2) Moving Beyond the 1-bit Case: A natural way to extend the 1-bit scheme above to the case when each client can transmit b-bits is to have each client communicate b randomly chosen entries of its transformed data Y_i instead of a single entry. This will boost the sample size by a factor of b, equivalently decrease the ℓ_2 error by a factor of b ($\sqrt{b}$ for ℓ_1). Instead, we argue next that we can exploit the recursive structure of the Hadamard matrix to boost the sample size by a factor of 2^b, equivalently decreasing the error by an exponential factor.

Consider $b \leq \lfloor \log d \rfloor$ and let $B = d/2^{b-1}$. Note that $H_d = H_{2^{b-1}} \otimes H_B$, where $\otimes$ denotes the Kronecker product. To visualize, for $b = 3$, H_d has the following structure:

$$Y_i = H_d X_i = \begin{bmatrix} H_B & H_B & H_B & H_B \\ H_B & -H_B & H_B & -H_B \\ H_B & H_B & -H_B & -H_B \\ H_B & -H_B & -H_B & H_B \end{bmatrix} \begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \\ X_i^{(3)} \\ X_i^{(4)} \end{bmatrix},$$

where for $l = 1, \ldots, 2^{b-1}$, $X_i^{(l)}$ denotes the l'th block of X_i of length $B = d/2^{b-1}$. Therefore, in order to communicate Y_i, we can equivalently communicate $H_B X_i^{(l)}$ for $l = 1, \ldots, 2^{b-1}$. Since $H_{2^{b-1}}$ is known, this is sufficient to reconstruct Y_i. We next observe that while communicating Y_i requires $d = B \times 2^{b-1}$ bits, communicating $\{H_B X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$ requires $B + (b-1)$ bits. This is because X_i is one-hot encoded and all but one of the 2^{b-1} vectors $\{H_B X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$ are equal to zero. It suffices to communicate the index l of the non-zero vector, by using $(b-1)$ bits, and its B entries by using additional B bits. This is the key observation that RHR builds on.

When each client has only b bits, they cannot communicate sufficient information for fully reconstructing Y_i, i.e. all $\{H_B X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$. Instead, each client chooses a random index $r_i \in [B]$ and communicates the r_i'th row of $\{H_B\, X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$, equivalently $\{(H_B)_{r_i} X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$ where $(H_B)_{r_i}$ denotes the r_i'th row of H_B. Note that as before, only one of the 2^{b-1} numbers $\{(H_B)_{r_i}\, X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$ is non-zero and therefore these numbers can be communicated by using b bits, $b-1$ bits to represent the index of the non-zero number and a single bit to communicate its value. When there is a privacy constraint, client i perturbs their b bits by a 2^b-RR mechanism with privacy level ε, and this yields the privatized report of b bits.

Upon receiving the reports from clients, the server constructs an unbiased estimator for Y_i. To do this, it first constructs an unbiased estimator for $\{H_B\, X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$ and then employs the structure $H_d = H_{2^{b-1}} \otimes H_B$. Note that since the randomness is shared the server knows the index r chosen by each client, and since the clients choose their indices independently and uniformly at random, roughly speaking, they communicate information about different rows of $\{H_B\, X_i^{(l)}, l = 1, \ldots, 2^{b-1}\}$. Finally, an unbiased estimator $\hat{Y}_i$ for Y_i yields an unbiased estimator for X_i through the transformation $\hat{X}_i = \frac{1}{d} H_d \cdot \hat{Y}_i$, and due to the orthogonality of H_d, it can be shown that the variance of $\hat{X}_i$ is the same as the variance of $\hat{Y}_i$ divided by d.

A subtle issue is that if $e^{\varepsilon} \ll 2^b$, the noise due to 2^b-RR mechanism may be too large, so instead of using all b bits, we perform the above encoding and decoding procedure with $b' \triangleq \min\left(\lceil \log_2 e \cdot \varepsilon \rceil\right)$.

Note that this careful construction based on the recursive structure of the Hadamard matrix is only required in the case when there are joint privacy and communication constraints. When only one constraint is present, the optimal error can be achieved in a much simpler fashion. When there is only a b bit constraint, [49] shows that the optimal error can be achieved by simply having each client communicate a subset of the entries of its data vector X_i (without requiring Hadamard transform). When there is only a privacy constraint ε, the optimal error can be achieved by a number of schemes, such as subset selection (2^b-SS) [9] and Hadamard response (HR) [44]. We summarize our proposed scheme RHR in Algorithm 1 and Algorithm 2.

Remark 4.1: As in mean estimation, RHR requires public randomness to achieve optimal communication efficiency. Indeed, we can show that RHR uses the minimum amount of shared randomness. See Section V for more details.

Remark 4.2: The encoding mechanism above involves two operations: 1) sampling a random index r_i from $[B]$ at each client with the help of a public coin, and 2) computing $(H_d)_{r_i} \cdot X_i$. Since X_i is one-hot, the encoding complexity is $O(\log d)$. On the other hand, in order to efficiently decode, the server first computes the joint histogram of client i's report and r_i in $O(n)$ time, which in turn allows us to calculate $\frac{1}{n}\sum_i \hat{Y}_i$, and then apply the Fast Walsh-Hadamard transform (FWHT) to obtain the estimator of empirical frequency in $O(d \log d)$ time. Hence the overall decoding complexity is $O\left(n + d\log d\right)$.

B. Achievability of Theorem 4.1

Next, we show that Recursive Hadamard Response (RHR) achieves optimal ℓ_1 and ℓ_2 estimation error.

1) Decomposition of Hadamard Matrix: Let us set $B = d/2^{k-1}$. Since $H_d = H_{2^{k-1}} \otimes H_B$, for any $j \in [B]$ and $m \in [2^{k-1}]$, if $j' = (m-1)B + j$ (and thus $j \equiv j' \pmod{B}$), we must have $(H_d)_{j'} = \left(H_{2^{k-1}}\right)_m \otimes \left(H_b\right)_j$, where $\otimes$ is the Kronecker product. This allows us to decompose the j'-th component of $H_d \cdot X_i$ into

$$\begin{aligned}(H_d)_{j'} \cdot X_i &= \left(\left(H_{2^{k-1}}\right)_m \otimes (H_B)_j\right) \cdot X_i \\ &= \sum_{l=1}^{2^{k-1}} \left(H_{2^{k-1}}\right)_{m,l} (H_B)_j \cdot X_i^{(l)}, \qquad (11)\end{aligned}$$

where X_i^l is the l-th block of X_i, i.e. $X_i^{(l)} \triangleq X_i[(l-1)B + 1 : lB]$. Therefore, as long as we know $(H_B)_j \cdot X_i^{(l)}$ for $l = 1, \dots, 2^{k-1}$, we can reconstruct $(H_d)_{j'} \cdot X_i$, for all $j' \equiv j \pmod{B}$.

2) Encoding Mechanism: Let $r_i \sim \text{Uniform}(B)$ be generated from the shared randomness, and consider the following quantizer

$$Q(X_i, r_i) = \left((H_B)_{r_i} \cdot X_i^{(l)}\right)_{l=1,\dots,2^{k-1}} \in \{-1, 0, 1\}^{2^{k-1}}.$$

Since X_i is one-hot encoded, there is exactly one non-zero $X_i^{(l)}$, so $Q(X_i, r_i)$ can be described by a k-bit string (with $k-1$ bits indicating the location of the non-zero entry and 1 bit indicating its sign).

Given $Q(X_i, r_i)$, by (11) we can recover 2^{k-1} coordinates of $Y_i = H_d \cdot X_i$:

$$\begin{aligned}Y_i(r') = (H_d)_{r'} \cdot X_i &= \sum_{l=1}^{2^{k-1}} \left(H_{2^{k-1}}\right)_{m,l} (H_B)_{r_i} \cdot X_i^{(l)} \\ &= \left(H_{2^{k-1}}\right)_m \cdot Q(X_i, r_i), \qquad (12)\end{aligned}$$

for any $r' = (m-1)B + r_i$. Therefore, if we define

$$\hat{Y}_i(Q(X_i, r_i), r_i) \triangleq \begin{cases} \frac{1}{2^{k-1}} Y_i(r'), & \text{if } r' \equiv r_i \\ 0, \text{ else}, \end{cases} \qquad (13)$$

then $\mathbb{E}\left[\hat{Y}_i\right] = \frac{1}{d} H_d \cdot X_i$, where the expectation is taken with respect to r_i.

To protect privacy, client i then perturbs $Q(X_i, r_i)$ via 2^k-RR scheme, since Q takes values on an alphabet of size 2^k, denoted by $\mathcal{Q} = \{\pm e_1, \dots, \pm e_{2^{k-1}}\}$,

$$\tilde{Q}_i = \begin{cases} Q(X_i, r_i), \text{ w.p. } \frac{e^\varepsilon}{e^\varepsilon + 2^k - 1} \\ Q' \in \mathcal{Q} \setminus \{Q(X_i, r_i)\}, \text{ w.p. } \frac{1}{e^\varepsilon + 2^k - 1}, \end{cases}$$

where $\boldsymbol{e}_l$ denotes the l-th coordinate vector in $\mathbb{R}^{2^{k-1}}$.

Client i then sends the k-bit report $\tilde{Q}_i$ to the server, and with $\tilde{Q}_i$, the server can compute an estimate of Q_i since $\mathbb{E}\left[\tilde{Q}_i \middle| Q(X_i, r_i)\right] = \frac{e^\varepsilon - 1}{e^\varepsilon + 2^k - 1} Q(X_i, r_i)$.

3) Constructing Estimator for $\hat{D}$: For a given $\tilde{Q}_i$, we estimate Y_i by $\hat{Y}_i\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\tilde{Q}_i, r_i\right)$, where $\hat{Y}_i$ is given by (12) and (13), with $Q(X_i, r_i)$ in (12) replaced by $\tilde{Q}_i$.

Claim 4.1: $\hat{Y}_i$ is an unbiased estimator of Y_i.

The final estimator of $D_{X^n} = \frac{1}{n}\sum X_i$ is given by

$$\hat{D}\left(\left(\tilde{Q}_i, r_i\right)_{i=1,\dots,n}\right) \triangleq \frac{1}{n}\sum_{i=1}^{n} H_d \hat{Y}_i \left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\tilde{Q}_i, r_i\right). \qquad (14)$$

Note that by Claim 4.1, $\hat{D}$ is an unbiased estimator for D_{X^n}. Finally picking $k = \min\left(b, \lceil \varepsilon \log_2 e \rceil, \lfloor \log d \rfloor\right)$ yields the following bounds.

Claim 4.2: The estimator $\hat{D}$ in (14) achieves the optimal ℓ_1 and ℓ_2 errors:

$$\mathbb{E}\left[\left\|\hat{D} - D_{X^n}\right\|_2^2\right] \preceq \frac{d}{n\left(\min\left(e^\varepsilon, \left(e^\varepsilon - 1\right)^2, 2^b, d\right)\right)} \quad \text{and}$$

$$\mathbb{E}\left[\left\|\hat{D} - D_{X^n}\right\|_1\right] \preceq \frac{d}{\sqrt{n\left(\min\left(e^\varepsilon, \left(e^\varepsilon - 1\right)^2, 2^b, d\right)\right)}}.$$

This establishes part (i) of Theorem 4.1. □

To obtain an upper bound on ℓ_∞ error, we extend the `TreeHist` protocol in [17], a 1-bit LDP heavy hitter estimation mechanism, to communicate b bits and satisfy the desired privacy level ε. A simpler version of `TreeHist` protocol, which is not optimized for computational complexity, is as follows: we first perform Hadamard transform on X_i, and sample one random coordinate with public randomness r_i. The 1-bit message is then passed through a binary ε-LDP mechanism. We can show that from the perturbed outcomes,

Algorithm 1 Encoding Mechanism $\tilde{Q}_i$ (at Each Client)

```
Input: client index $i$, observation $X_i$, privacy level $\varepsilon$, alphabet size $d$
Result: Encoded message $\left(\tilde{\texttt{sign}}, \tilde{\texttt{loc}}\right)$
Set $D = 2^{\lceil \log d \rceil}$, $k = \min\left(b, \lceil \varepsilon \log_2 e \rceil\right)$, $B = D/2^{k-1}$;
Draw $r_i$ from uniform($B$) using public-coin ;
begin
    $\texttt{loc} \leftarrow \lceil \frac{X_i}{B} \rceil$;
    $\texttt{sign} \leftarrow (H_d)_{r_i, X_i}$;
    $\left(\tilde{\texttt{sign}}, \tilde{\texttt{loc}}\right) \leftarrow 2^k - \mathrm{RR}_\varepsilon\left((\texttt{sign}, \texttt{loc})\right)$        /* $(\texttt{sign}, \texttt{loc})$ as a $k$-bit string */;
end
```

Algorithm 2 Estimator of D_{X^n} (at the Server)

```
Input: $(\tilde{\texttt{sign}}[1:n], \tilde{\texttt{loc}}[1:n])$, privacy level $\varepsilon$, alphabet size $d$
Result: $\hat{D}$
Set $D = 2^{\lceil \log d \rceil}$, $k = \min\left(b, \lceil \varepsilon \log_2 e \rceil\right)$, $B = D/2^{k-1}$;
Partition messages into groups $\mathcal{G}_1, \ldots, \mathcal{G}_B$, with message $i$ in $\mathcal{G}_{r_i}$;
forall the $j = 1, \ldots, B$ do
    $\mathcal{G}_j^+ \leftarrow \left\{ \tilde{\texttt{loc}}(i) \,|\, i \in \mathcal{G}_j, \tilde{\texttt{sign}}(i) = +1 \right\}$;
    $\mathcal{G}_j^- \leftarrow \left\{ \tilde{\texttt{loc}}(i) \,|\, i \in \mathcal{G}_j, \tilde{\texttt{sign}}(i) = -1 \right\}$;
    $\mathsf{Emp}_j \leftarrow \left(\text{histogram}(\mathcal{G}_j^+) - \text{histogram}(\mathcal{G}_j^-)\right) \cdot \frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}$;
    forall the $l = 0, \ldots, 2^{k-1} - 1$ do
        $\hat{E}[l \cdot B + j] \leftarrow \mathsf{FWHT}(\mathsf{Emp}_j)[l]$        /* fast Walsh-Hadamard transform */
    end
end
$\hat{D} \leftarrow \frac{1}{d} \cdot \mathsf{FWHT}\left(\hat{E}\right)$;
```

the server can construct an unbiased estimator of X_i with a bounded sub-Gaussian norm, and the ℓ_∞ error will be $O(\sqrt{\log d / n\varepsilon^2})$.

To extend this scheme to an arbitrary privacy regime and an arbitrary communication budget of b bits, we independently and uniformly sample the Hadamard transform of X_i for $k = \min\left(b, \lceil \varepsilon \rceil\right)$ times. Each 1-bit sample is then perturbed via a ε'-LDP mechanism with $\varepsilon' \triangleq \varepsilon / k$.

Note that under the distribution-free setting, the randomness comes only from the sampling and the privatization steps, so we could view each re-sampled and perturbed message as generated from a fresh new copy of X_i since X_i is not random. Equivalently, this boils down to a frequency estimation problem with $n' = nk$ clients and under $\varepsilon' = \varepsilon / k$ and gives us the ℓ_∞ error

$$O\left(\sqrt{\frac{\log d}{n' \left(\varepsilon'\right)^2}}\right) = O\left(\sqrt{\frac{\log d}{n \min\left(\varepsilon^2, \varepsilon, b\right)}}\right).$$

Below we describe the details.

4) Encoding: Set $k = \min\left(b, \lceil \varepsilon \rceil\right)$. For each X_i, we randomly sample $(H_d)_{X_i}$ (i.e. the X_i-th column of H_d) k times, identically and independently by using the shared randomness. Let $r_i^{(1)}, \ldots, r_i^{(k)}$ be the sampled coordinates, which are known to both the server and node i, and $(H_d)_{X_i, r_i^{(\ell)}}$ be the sampling outcomes. Then due to the orthogonality of H_d, for all $j \in [d], \ell \in [k]$,

$$\mathbb{E}\left[(H_d)_{j, r_i^{(\ell)}} \cdot (H_d)_{X_i, r_i^{(\ell)}} \right] = \begin{cases} 1, & \text{if } j = X_i \\ 0, & \text{if } j \neq X_i, \end{cases} \tag{15}$$

where the expectation is taken over $r_i^{(\ell)}$.

We then pass $\left\{ (H_d)_{X_i, r_i^{(\ell)}} \middle| \ell = 1, \ldots, k \right\}$ through k binary ε'-LDP channels sequentially, with $\varepsilon' \triangleq \varepsilon / k$. By the composition theorem [5] of differential privacy, the privatized outcomes, denoted as $\left\{ (\tilde{H_d})_{X_i, r_i^{(\ell)}} \right\}$, satisfy ε-LDP.

5) Estimation: Observe that

$$\mathbb{E}\left[\left(\frac{e^{\varepsilon'} + 1}{e^{\varepsilon'} - 1} \right) (\tilde{H_d})_{X_i, r_i^{(\ell)}} \middle| (H_d)_{X_i, r_i^{(\ell)}} \right] = (H_d)_{X_i, r_i^{(\ell)}},$$

where the expectation is taken with respect to the randomness from the privatization step. Therefore

$$\hat{X}_i^{(\ell)}(j) \triangleq \left(\frac{e^{\varepsilon'} + 1}{e^{\varepsilon'} - 1} \right) (H_d)_{j, X_i} (\tilde{H_d})_{X_i, r_i^{(\ell)}}$$

defines an unbiased estimator of $X_i(j)$. Moreover,

$$\left| \hat{X}_i^{(\ell)}(j) - X_i(j) \right| \leq \left(\frac{e^{\varepsilon'} + 1}{e^{\varepsilon'} - 1} + 1 \right) \text{ a.s.},$$

so $\hat{X}_i^{(\ell)}(j)$ has sub-Gaussian norm bounded by

$$\sigma \leq 2 \frac{e^{\varepsilon'} + 1}{e^{\varepsilon'} - 1}. \tag{16}$$

Finally, we estimate $D_{X^n}(j)$ by

$$\hat{D}(j) = \frac{1}{nk}\sum_{i=1}^{n}\sum_{\ell=1}^{k}\hat{X}_i^{(\ell)}(j).$$

Observe that

$$\hat{D}(j) - D_{X^n}(j) = \frac{1}{nk}\sum_{i=1}^{n}\sum_{\ell=1}^{k}\left(\hat{X}_i^{(\ell)}(j) - X_i(j)\right) \qquad (17)$$

has sub-Gaussian norm bounded by $\sigma/\sqrt{nk}$, where σ is given by (16).

To bound the ℓ_∞ norm, we apply the maximum bound (see, for instance, [53, Chapter 2]) for sub-Gaussian random variables (note that for j, j', $\hat{D}(j)$ and $\hat{D}(j')$ are not independent):

$$\begin{aligned}\mathbb{E}\left[\max_{j\in[d]}\left|\hat{D}(j) - D_{X^n}(j)\right|\right] &\le 2\sqrt{\sigma^2\log d}\\ = 4\sqrt{\left(\frac{e^{\varepsilon'}+1}{e^{\varepsilon'}-1}\right)^2\frac{\log d}{nk}} &\overset{\text{(a)}}{\asymp} \sqrt{\frac{\log d}{n\min(\varepsilon,\varepsilon^2,k)}},\end{aligned} \qquad (18)$$

where (a) holds since if $\varepsilon = o(1)$, then $k = 1$ and hence

$$\left(\frac{e^{\varepsilon'}+1}{e^{\varepsilon'}-1}\right)^2 \asymp \frac{1}{\varepsilon^2};$$

otherwise $\varepsilon = \Omega(1)$ and $\varepsilon' = \Omega(1)$, so

$$\left(\frac{e^{\varepsilon'}+1}{e^{\varepsilon'}-1}\right)^2 \asymp 1.$$

Both cases are upper bounded by (18), so the result follows. This establishes part (ii) of Theorem 4.1. □

C. Converse of Theorem 4.1

We bound the error by imposing a prior distribution $\boldsymbol{p}$ on $X_1, \dots, X_n$ and applying the lower bounds for *distributional setting* from [9], [51] (under an LDP constraint) and [49], [50] (under a communication constraint).

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \boldsymbol{p}$. Then for any $\hat{D}(X^n)$, we must have

$$\begin{aligned}&\max_{X^n\sim\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D} - D_{X^n}\right\|_2^2\right]\\ &\overset{\text{(a)}}{\ge} \max_{\boldsymbol{p}}\mathbb{E}\left[\left(\left\|\hat{D}-\boldsymbol{p}\right\|_2 - \left\|D_{X^n}-\boldsymbol{p}\right\|_2\right)^2\right]\\ &\ge \max_{\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}\right\|_2^2\right] - 2\mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}\right\|_2\left\|D_{X^n}-\boldsymbol{p}\right\|_2\right]\\ &\overset{\text{(b)}}{\ge} \max_{\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}\right\|_2^2\right]\\ &\quad - 2\sqrt{\mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}\right\|_2^2\right]\mathbb{E}\left[\left\|D_{X^n}-\boldsymbol{p}\right\|_2^2\right]},\end{aligned} \qquad (19)$$

where (a) and (b) follow from the triangular inequality and the Cauchy-Schwarz inequality respectively. From [9] and [49], there exists a worst-case $\boldsymbol{p}^*$ such that

$$\begin{aligned}c\frac{d}{n}\left(\frac{1}{\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b\right)}\right) &\le \mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}^*\right\|_2^2\right]\\ &\le C\frac{d}{n}\left(\frac{1}{\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b\right)}\right),\end{aligned} \qquad (20)$$

for some positive constants c and C.

On the other hand, the ℓ_2 convergence of $D(X^n)$ to $\boldsymbol{p}$ is $O(1/n)$ for any $\boldsymbol{p}$, which gives us

$$\mathbb{E}\left[\left\|D_{X^n}-\boldsymbol{p}^*\right\|_2^2\right] \le c'\frac{1}{n}. \qquad (21)$$

Plugging (20) and (21) back into (19) yields

$$\begin{aligned}\max_{X^n\sim\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] &\ge C_1\frac{d}{n}\left(\frac{1}{\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b\right)}\right)\\ &\quad - C_2\frac{1}{n}\sqrt{\frac{d}{\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b\right)}}.\end{aligned}$$

Note that the first term of the above equation dominates as long as $\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b\right) = o(d)$ (see [54] for example), and hence the desired ℓ_2 lower bound follows.

For ℓ_1 error, similarly, we have

$$\begin{aligned}\max_{X^n\sim\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_1\right] &\ge \max_{\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}\right\|_1\right]\\ &\quad - \mathbb{E}\left[\left\|D_{X^n}-\boldsymbol{p}\right\|_1\right].\end{aligned} \qquad (22)$$

Observe that it holds that $\mathbb{E}\left[\left\|D_{X^n}-\boldsymbol{p}\right\|_1\right] \le \sqrt{d/n}$ (for instance, see [54]), and again from the lower bounds of [9] and [49],

$$\max_{\boldsymbol{p}}\mathbb{E}\left[\left\|\hat{D}-\boldsymbol{p}\right\|_1\right] \ge \sqrt{\frac{d^2}{n\min\left\{e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b\right\}}}.$$

Plugging this into (22) yields the desired ℓ_1 lower bound. □

D. Application to Distribution Estimation

For frequency estimation, RHR requires shared randomness so that the server can construct an unbiased estimator. However, for distribution estimation where $X_i \sim \boldsymbol{p}$, we can replace the random sampling with a deterministic partitioning of coordinates among the different clients and circumvent the need for shared randomness. This gives us the following theorem:

Corollary 4.1: For distribution estimation under ε-LDP and b-bit communication constraints,

$$r_{\text{DE}}\left(\ell_2\right) \asymp \frac{d}{n\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b,d\right)}, \text{ and}$$

$$r_{\text{DE}}\left(\ell_1\right) \asymp \frac{d}{\sqrt{n\min\left(e^\varepsilon,\left(e^\varepsilon-1\right)^2,2^b,d\right)}},$$

without shared randomness. Moreover, if

$$n \cdot \min\left(e^{\varepsilon}, \left(e^{\varepsilon}-1\right)^2, 2^b, d\right) \geq d^2,$$

the above errors are optimal even in the presence of shared randomness.

The lower bounds follow directly from the results of [9] (under LDP constraint) and [49], [50] (under communication constraint). We leave the formal proof of the achievability to Section B-B of the appendix.

V. Role of Shared Randomness

A. The Amount of Shared Randomness

In the achievability part of Theorem 3.1, our proposed scheme SQKR randomly and independently samples $b^*_{\mathrm{ME}} \triangleq \min\left(\lceil\varepsilon\rceil, b\right)$ bits from the quantized d-dimensional binary vector at each client. These bits are then privatized and communicated to the server. In addition to the values of these bits, the server needs to know the indices of the sampled bits, which corresponds to an additional $b^*_{\mathrm{ME}} \log d$ bits of information that needs to be shared between each client and the server. This information can be shared in two different ways: 1) sampling can be done by using a public coin shared a priori between the client and the server, or 2) sampling can be done by using a private coin on the client side, which is then communicated to the server. We can also combine both 1) and 2) when $b > b^*_{\mathrm{ME}}$: given b bits communication budget, SQKR compresses the data to b^*_{ME} bits, so the client can use the remaining $b - b^*_{\mathrm{ME}}$ bits to communicate the locally generated randomness required at the sampling step. Thus the amount of shared randomness is reduced to $b^*_{\mathrm{ME}} \log d - (b - b^*_{\mathrm{Me}})$ bits. Moreover, by extending [13, Th. 4], we also obtain a lower bound on the amount of shared randomness required, which we summarize in the following corollary:

Corollary 5.1: Under ε-LDP and b-bit communication constraints, SQKR uses $\min\left(b^*_{\mathrm{ME}} \log d, d\right) - (b - b^*_{\mathrm{ME}})$ bits of shared randomness to achieve $r_{\mathrm{ME}}\left(\ell_2, b, \varepsilon\right)$, where $b^*_{\mathrm{ME}} \triangleq \min\left(\lceil\varepsilon\rceil, b\right)$. Moreover, if $b < \log d - 2$, any b-bit consistent mean estimation scheme[1] requires at least $\log d - b - 2$ bits.

We contrast this with the amount of shared randomness needed in the generic scheme of [10] which provides ε-LDP by using 1 bit per client in the high privacy regime $\varepsilon = O(1)$. The shared randomness required by this scheme is d bits per client. In contrast, when $\varepsilon = O(1)$ and $b = 1$, SQKR requires $\log d$ bits of shared randomness.

Similarly, for frequency estimation, it can be seen that RHR requires $\log d - b^*_{\mathrm{FE}}$ bits of shared randomness in the random sampling step, where $b^*_{\mathrm{FE}} \triangleq \min\left(\lceil\varepsilon \log_2 e\rceil, b\right)$. Again, if the communication budget b is greater than the privacy budget $\lceil\varepsilon \log_2 e\rceil$, the clients can privately generate $b - \lceil\varepsilon \log_2 e\rceil$ random bits and send it to the server, which reduces the required public randomness to $\log d - b$ bits. Furthermore, as in mean estimation, we can show that at least $\log d - b - 2$ bits are needed to get a consistent scheme, so RHR is also optimal in the amount of public randomness it uses. We summarize it in the following corollary:

TABLE IV

The Amounts of Required Shared Randomness

	Communication	Shared randomness
SQKR (Thm. 3.1)	$\lceil\varepsilon\rceil$ bits	$\min\left(\lceil\varepsilon\rceil \log d, d\right)$ bits
RHR (Thm. 4.1)	$\lceil\log_2 e \cdot \varepsilon\rceil$ bits	$\log d - \lfloor\log_2 e \cdot \varepsilon\rfloor$ bits

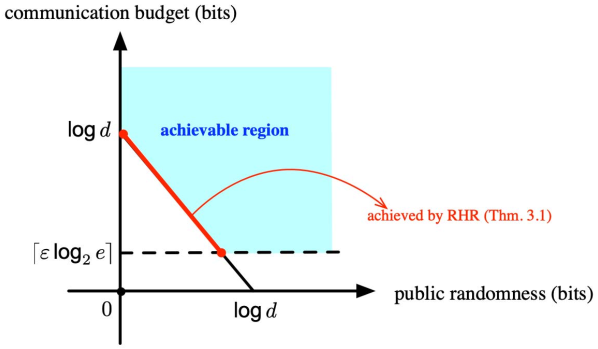

Fig. 1. Achievable region for frequency estimation with public randomness.

Corollary 5.2: Under ε-LDP and b-bit communication constraints, RHR uses $\log d - b$ bits of shared randomness to achieve $r_{\mathrm{FE}}\left(\ell_2, b, \varepsilon\right)$. Moreover, if $b < \log d - 2$, any b-bit consistent frequency estimation scheme requires at least $\log d - b - 2$ bits of shared randomness. Thus RHR is optimal in the amount of shared randomness it uses for frequency estimation, up to an additive constant.

The achievability parts of Corollary 5.1 and Corollary 5.2 follow directly from the analysis of SQKR and RHR, and we defer the proof of the converse part to Section B-C of the appendix. Given a ε-LDP constraint, we summarize the minimum amounts of communication and shared randomness required to achieve the optimal error $r_{\mathrm{ME}}\left(\ell_2, \varepsilon, \infty\right)$ and $r_{\mathrm{FE}}\left(\ell_2, \varepsilon, \infty\right)$ in Table IV.

In Figure 1, we plot the achievable region for the minimax frequency estimation error under ε-LDP constraint (i.e. $r_{\mathrm{FE}}\left(\ell_2, \varepsilon, \infty\right)$). Note that the red line in Figure 1 can be achieved by RHR.

Remark 5.1: Note that shared randomness is only needed for distribution-free settings; for distribution estimation and statistical mean estimation, one can achieve the same estimation error with only private randomness as noted in Theorems 3.1 and 4.1.

B. Converting Public-Coin Schemes to Private-Coin Schemes

As discussed above, we can always replace shared randomness with additional communication by first generating the random bits at the client side and then sending them to the server. Therefore, by Corollary 5.1 and Corollary 5.2, we automatically obtain private-coin SQKR and private-coin RHR by using additional communication. We next state these observations for completeness.

Corollary 5.3 (Private-Coin SQKR): Under ε-LDP and b-bit communication constraints with $b > \log d$ and

[1] A scheme is *consistent* if it has vanishing estimation error as $n \to \infty$.

$0 < \varepsilon \leq d$, the ℓ_2 minimax error for private-coin mean estimation, denoted as $\tilde{r}_{\mathtt{ME}}(\ell_2, \varepsilon, b)^2$ (to distinguish it from the minimax error $r_{\mathtt{ME}}(\ell_2, \varepsilon, b)$ achieved by public-coin schemes), is characterized as follows:

(i) if $\log d < b < d$, then

$$\tilde{r}_{\mathtt{ME}}(\ell_2, \varepsilon, b) = O\left(\frac{d}{n \min\left(\varepsilon^2, \varepsilon, b/\log d, d\right)}\right);$$

(ii) if $b \geq d$, then

$$\tilde{r}_{\mathtt{ME}}(\ell_2, \varepsilon, b) = O\left(\frac{d}{n \min\left(\varepsilon^2, \varepsilon, d\right)}\right),$$

and the above errors can be achieved by private-coin SQKR. Therefore private-coin SQKR requires $O\left(\min\left(\lceil\varepsilon\rceil \log d, d\right)\right)$ bits of communication to achieve $\tilde{r}_{\mathtt{ME}}\left(\ell_2, \varepsilon, \infty\right)$.
Similarly, the estimation error of private-coin RHR is characterized below:

Corollary 5.4 (Private-Coin RHR): Under ε-LDP and b-bit communication constraints with $b > \log d$ and $0 < \varepsilon \leq \log d$, the ℓ_2 minimax error for private-coin frequency estimation, denoted as $\tilde{r}_{\mathtt{FE}}(\ell_2, \varepsilon, b)$, is

$$\tilde{r}_{\mathtt{FE}}(\ell_2, \varepsilon, b) = O\left(\frac{d}{n \min\left(\left(e^{\varepsilon} - 1\right)^2, e^{\varepsilon}, d\right)}\right),$$

which can be achieved by private-coin RHR. In words, for any ε, private-coin RHR always uses $\log d$ bits of communication to achieve $\tilde{r}_{\mathtt{FE}}(\ell_2, \varepsilon, \infty)$.

Moreover, the following lemma, an extension of [13, Th. 4], establishes a lower bound on the communication required for consistent private-coin schemes:

Lemma 5.1: Any consistent private-coin scheme for both mean estimation and frequency estimation uses at least $b > \log d - 2$ bits of communication.

This shows that the $\log d$ lower bounds on b in both corollaries are fundamental (within 2 bits). The proof of the lemma is given in Section B-D of the appendix.

VI. Application to Private Stochastic Gradient Descent and Federated Learning

In this section, we apply our SQKR mean estimation scheme to (differentially private) stochastic gradient descent (SGD), which yields a distributed local DP-SGD. In each round, the server samples n out of N clients uniformly at random, each (sampled) client computes a local gradient from its data, and the server aggregates the mean of the local gradients via the SQKR. Since the SQKR ensures local DP, we call the resulting scheme local DP-SGD.

We summarize local DP-SGD in Algorithm 3, in which we use $\mathsf{SQKR}_{\mathsf{enc}}$ to denote the clients' procedure and $\mathsf{SQKR}_{\mathsf{dec}}$ to denote the server's procedure.

Algorithm 3 Local DP-SGD

```
Input: Clients local dataset $D_1, \dots, D_N \in \mathcal{D}$, SQKR
       parameters $\varepsilon > 0$, $b \in \mathbb{N}$, loss function
       $\ell(\cdot,\cdot) : \mathcal{W} \times \mathcal{D} \to \mathbb{R}_+$, learning rate $\gamma > 0$
Result: Compute $w_T \approx \arg\min_w \sum_{i=1}^{N} \ell\left(D_i, w\right)$
Server generates initial model weights $w_0 \in \mathbb{W}$;
forall the iteration $t = 1, \dots, T$ do
    Server samples a subset of $n$ clients $\mathcal{C}_t \subset [N]$ and
    broadcasts $w_{t-1}$ to them;
    forall the each client $i \in \mathcal{C}_t$ do
        Computes $g_i^t = \mathsf{Clip}_{\ell_2, c}\left(\nabla \ell\left(d_i, w_{t-1}\right)\right)$;
        Computes $Z_i^t = \mathsf{SQKR}_{\mathsf{enc}}\left(g_i^t\right)$;
        Send $Z_i^t$ to the server;
    end
    (Server) decodes $\hat{g}_i^t = \mathsf{SQKR}_{\mathsf{dec}}\left(Z_i\right)$, $\forall i \in \mathcal{C}_t$;
    (Server) updates the model by $w_t = w_{t-1} + \frac{\gamma}{n} \sum_i \hat{g}_i$;
end
Return: $w_T$;
```

A. Privacy of Local DP-SGD

Since the SQKR satisfies ε-LDP, for each round t the (local) privacy loss of each client is at most ε. Applying the composition theorem [5] for T rounds, we conclude that the local privacy guarantee for each client is no worse than $T\varepsilon$.

Nevertheless, $T\varepsilon$-LDP is the *worst-case* guarantee, as it considers the worst-case event in which a client is sampled for all T rounds. However, this event happens with an exponentially small probability. To mitigate these worst-case scenarios, one can consider *without-replacement* SGD (SGDo) [55], [56], [57], [58], which ensures each client being sampled exactly $\frac{T}{n}$ times, and hence the total privacy loss is reduced to $\frac{nT\varepsilon}{N}$. We provide a local DP-SGDo in Algorithm 4.[3]

We remark that although it is observed empirically that SGDo can potentially converge at a faster rate [56] than standard SGD, the theoretical convergence is less known and existing analysis only focuses on convex and smooth loss functions.

B. Convergence Analysis

To analyze the convergence rate of Algorithm 3, the next lemma(which originates from [59] but we use a version adapted from [18]) builds the connection between distributed SGD and mean estimation.

Lemma 6.1 ([18]): Assume $F(w) \triangleq \frac{1}{N}\sum_{i=1}^{N} \ell(w; d_i)$, where $\ell(\cdot, d)$ is an L-smooth and c-Lipschitz function for all $d \in \mathcal{D}$. Let w_0 satisfies $F\left(w_0\right) - F\left(w^*\right) \leq D_F$. Let μ_g^t be an unbiased estimate of $\nabla F(w_t)$ and let $\hat{\mu}_g^t$ be the noisy (privatized) version of μ_g^t. Let the learning rate $\gamma \triangleq \min\left\{L^{-1}, \sqrt{2D_F}\left(\sigma\sqrt{LT}\right)^{-1}\right\}$. Then after T rounds,

$$\mathbb{E}_{t \sim \mathsf{unif}(T)}\left[\left\|\nabla F(w_t)\right\|_2^2\right] \leq \frac{2 D_F L}{T} + \frac{2\sqrt{2}\sigma\sqrt{L D_F}}{\sqrt{T}} + cB,$$

[2] The definition of $\tilde{r}_{\mathtt{ME}}(\cdot)$ is the same as that of $r_{\mathtt{ME}}(\cdot)$ in (1), except that now the minimum is taken over all private-coin schemes.

[3] In Algorithm 4, we abuse notation of a random shuffling σ and let $\sigma\left(\left(a_1, \dots, a_N\right)\right) \triangleq \left(a_{\sigma(1)}, \dots, a_{\sigma(N)}\right)$.

Algorithm 4 Local DP-SGDo

```
Input: Clients local dataset $d_1, \dots, d_N \in \mathcal{D}$, SQKR
       parameters $\varepsilon > 0$, $b \in \mathbb{N}$, loss function
       $\ell(\cdot,\cdot) : \mathcal{W} \times \mathcal{D} \to \mathbb{R}_+$, learning rate $\gamma > 0$
Result: Compute $w_T \approx \arg\min_w \sum_{i=1}^{N} \ell(D_i, w)$
Server generates initial model weights $w_0 \in \mathbb{W}$;
forall the epochs $k = 0, \dots, K-1$ do
    Sever generates a random shuffling $\sigma_k \in \mathcal{S}_N$;
    forall the iteration $t = 1, \dots, N/n$ do
        Server broadcasts $w_{k,t-1}$to cohort
        $\mathcal{C}_t \triangleq \sigma_k([n(t-1) : nt])$;
        forall the each client $i \in \mathcal{C}_t$ do
            Computes $g_i^t = \mathsf{Clip}_{\ell_2,c}(\nabla \ell(d_i, w_{t-1}))$;
            Computes $Z_i^t = \mathsf{SQKR}_{\mathrm{enc}}(g_i^t)$;
            Send $Z_i^t$ to the server;
        end
        (Server) decodes $\hat{g}_i^t = \mathsf{SQKR}_{\mathrm{dec}}(Z_i)$, $\forall i \in \mathcal{C}_t$;
        (Server) updates the model by
        $w_{k,t} = w_{k,t-1} + \frac{\gamma}{n} \sum_i \hat{g}_i$;
    end
    Server updates $w_{k+1,0} \leftarrow w_{k,N/n}$;
end
Return: $w_T$;
```

where

$$\sigma^2 = 2\Big(\max_{1 \le t \le T} \mathbb{E}\left[\left\| \mu_g^t - \nabla F(w_t) \right\|_2^2 \right] + \max_{1 \le t \le T} \mathbb{E}_Q\left[\left\| \mu_g^t - \tilde{\mu}_g^t \right\|_2^2 \right] \Big),$$

and $B = \max_{1 \le t \le T} \left\| \mathbb{E}_Q\left[\mu_g^t - \hat{\mu}_g^t \right] \right\|_2$.

To apply Lemma 6.1 to Algorithm 3, observe that (1) μ_g^t (the true mean of gradients of $\mathcal{C}_t$) is an unbiased estimator of $\nabla F(w_t)$ (because clients are sampled uniformly at random), and (2) $\hat{\mu}_g^t$ is an unbiased estimator of μ_g^t since the SQKR is unbiased.[4] This implies $B = 0$ and $\sigma^2 = \max_t \mathsf{Var}\left(\mu_g^t\right) + \mathsf{Var}\left(\hat{\mu}_g^t \middle| \mu_g^t\right)$.

Note that the first term $\mathsf{Var}\left(\mu_g^t\right)$ is bounded by c^2. Applying Theorem 3.1, we can bound the second $\mathsf{Var}\left(\hat{\mu}_g^t \middle| \mu_g^t\right)$ by $\frac{c^2 d}{n \min(\varepsilon, b, d)}$. Thus we arrive at the following conclusion:

Corollary 6.1 (Convergence of Local DP-SGD): Under the same assumptions of Lemma 6.1, after $\tau \sim \mathsf{uniform}(T)$ iterations, the output of Algorithm 3 satisfies

$$\mathbb{E}_\tau\left[\|\nabla F(w_\tau)\|_2^2 \right] \le \frac{L D_F}{T} + C_0 \frac{\sqrt{8c^2 L D_F}}{\sqrt{T}} \sqrt{1 + \frac{d}{n \min(\varepsilon, \varepsilon^2, b, d)}},$$

for some universal constant $C_0 > 0$.

Remark 6.1: Since the convergence guarantees of Lemma 6.1 is derived for *the average* of all intermediate steps, i.e., w_t for $t\ in[T]$, in Corollary 6.1 we apply Algorithm 3 with a random stopping time τ.

Finally, we remark that one can also obtain convergence guarantees for SGDo (i.e., Algorithm 4) by following similar analysis in [55] (for generalized linear models) or in [56] (for convex and smooth loss functions). See [60, Corollary 2] for example.

1) Discussion on the Convergence Rates: Most existing results in private empirical risk minimization (ERM) problems, such as the exponential mechanism [61] and DP-SGD [62], focus on central DP instead of local DP. An important exception is [63], in which stochastic risk minimization under ε-*local* DP is studied (with an assumption that $\varepsilon = O(1)$). Under the stochastic setting, each local sample is assumed to be i.i.d. from an unknown distribution P, and the goal is to minimize the population risk $R(w) \triangleq \mathbb{E}_P\left[\ell(X, w)\right]$. Under an ε-LDP constraint, [63] gives a lower bound on the excess error for generalized linear or convex models: $R(w_T) - R(w^*) = \Omega\left(\frac{\sqrt{d}}{\varepsilon\sqrt{T}}\right)$. Moreover, for the generalized linear model, this lower bound is achievable via a private SGD [63, Theorem 5]. By replacing the private local randomizer in [34] with SQKR, we can obtain the same convergence rate $\left(\frac{\sqrt{d}}{\varepsilon\sqrt{T}}\right)$ but with much less communication, i.e., $\Theta\left(\lceil \varepsilon \rceil\right)$ bits per client. In addition, the resulting ERM algorithm is essentially the same as Algorithm 4 (note that under the stochastic setting, the convergence analysis for with-replacement SGD and without-replacement SGD remains the same), implying that the convergence rate in Corollary 6.1 is optimal for convex loss functions and $\varepsilon = O(1)$. For general loss or low privacy regime $\varepsilon = \Omega(1)$, however, the optimal rate remains open.

We remark that there have been extensive works studying ERM under central DP, e.g., [64], [65], [66], [67]. As opposed to the local DP setting, under central DP, it is shown that the optimal convergence rate for stochastic convex optimization (DP-SCO) becomes $O\left(\frac{\sqrt{d}}{n\varepsilon}\right)$ [64], [67] when $\varepsilon = O(1)$, and hence we see a $\sqrt{n}$ factor as the price for ensuring *local* DP.

VII. Experiments

In this section, we implement our mean estimation and frequency estimation schemes and present our experimental results.[5]

A. Mean Estimation

We implement our mean estimation scheme Subsampled and Quantized Kashin's Response (SQKR) as in Section III under *private-coin setting* and compare it with a baseline, a concatenation of `DJW` [7], [34] (which is order-optimal under ε-LDP for $\varepsilon = O(1)$) and the quantizer based on Kashin's representation [26] (which is optimal up to a logarithmic factor, under b-bit communication constraint).

`DJW` (Lemma 1 in [34]) samples a vector from the unit sphere with proper probability density (which depends on X_i), and scales it by a factor of $O(\sqrt{d})$ in order to make it unbiased. Although under public-coin setting, one can sample the vector with the help of public randomness and reduce the

[4] Notice that the clipping step in Algorithm 3 does not cause any bias since the Lipschitz condition implies $\|\nabla \ell\|_2 \le c$.

[5] The code can be found in https://github.com/WeiNingChen/Kashin-mean-estimation (for the SQKR scheme) and https://github.com/WeiNingChen/RHR (for the RHR scheme).

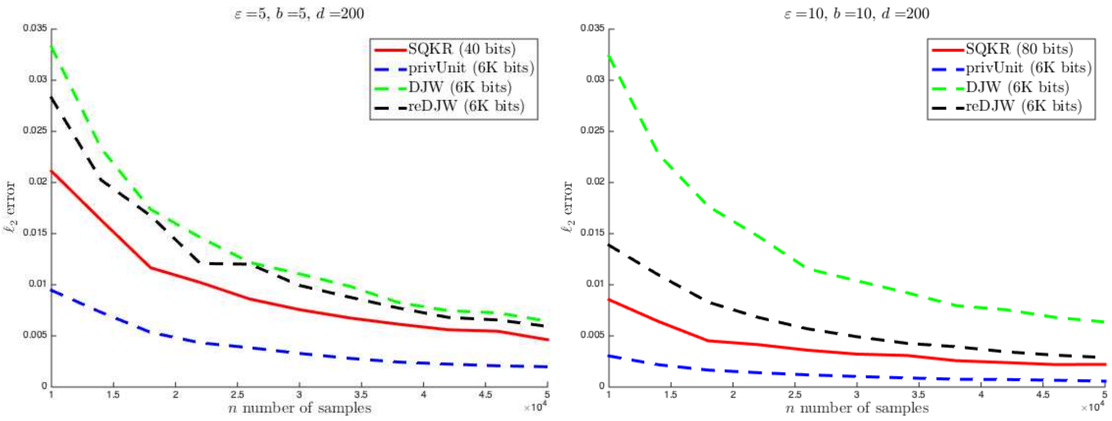

Fig. 2. ℓ_2 error of `privUnit`, `DJW`, `reDJW` and SQKR with different dimensions $d = 200$.

communication to $\lceil \varepsilon \rceil$ bits [17], for private-coin model each client has to send a d-dimensional vector to the server and hence requires to communicate $\Theta(d)$ bits.[6] To compare with SQKR under private-coin setting, we use an (order-optimal) quantizer based on Kashin's representation to further compress the communication to $b\lceil \log d \rceil$ bits. It can be shown that such direct concatenation will result in $\tilde{O}(d^2)$ error rate (see Section A in the appendix for more details).

1) Generating the Data: In order to capture the distribution-free setting, we generate data independently but non-identically; in particular, we set $Z_1, \ldots, Z_{n/2} \overset{\text{i.i.d.}}{\sim} N(1,1)^{\otimes d}$ and $Z_{n/2+1}, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} N(10,1)^{\otimes d}$ (this also makes the data non-central, i.e. $\mathbb{E}\left[\sum Z_i\right] \neq 0$). Since each sample has bounded ℓ_2 norm, we normalize each Z_i by setting $X_i = Z_i / \left\|Z_i\right\|_2$.

2) Generating the Tight Frame: We construct the tight frame by using the random partial Fourier matrices in [26]. Specifically, we set $N = 2^{\lceil \log_2 d \rceil + 1} = \Theta(d)$, and choose the basis $U = \left\{1/\sqrt{N}, -1/\sqrt{N}\right\}^{N \times d}$ by selecting the first d rows of $H_N \cdot D$, where H_N is a $N \times N$ Hadamard matrix and D is a random diagonal matrix with each diagonal entry generated from uniform $\{+1, -1\}$. It can be shown that the tight frame based on U has Kashin's level $K = \tilde{O}(1)$.

In Figure 4, we fix the sample size to $n = 10^5$ and ε, b, and increase the dimension d. From the result, we see that SQKR has a linear dependence on d, whereas the baseline (labeled as "Separation" since it is based on the idea of separately coding for privacy and communication efficiency) has super-linear dependence. Therefore the performance differs drastically when d increases.

a) Compare to optimal ε-LDP schemes [7]: We first compare our scheme SQKR, under private-coin setting, with 1) `privUnit` [7], which is order-optimal for all ε and 2) `DJW` [34], which is order-optimal for $\varepsilon = O(1)$. Note that although `DJW` is originally designed for high-privacy regime $\varepsilon = O(1)$, one can independently and repeatedly apply it with $\varepsilon' = 1$ for $\lfloor \varepsilon \rfloor$ times and return the mean of the $\lfloor \varepsilon \rfloor$ vectors. By the composition theorem [5] for DP, the output satisfies $\lfloor \varepsilon \rfloor$-LDP, and the MSE is reduced by a factor of $\lfloor \varepsilon \rfloor$. The repeated version of `DJW` (denoted as `reDJW`) is hence asymptotically optimal, and we also compare it with our scheme.

Note that the outcomes of `privUnit`, `DJW` and `reDJW` are d-dimensional vectors lying in a radius $O(\sqrt{d})$ sphere, so in general we need $32d$ bits to represent it (where we assume each float requires 32 bits). Figure 2 shows that SQKR achieves similar performance with significantly less communication budgets. For instance, under the private-coin model, when $\varepsilon = 5$ and $d = 200$, the communication cost of `privUnit` is roughly $32 \times 200 \approx 6K$ bits, while according to Corollary 5.3, SQKR uses only $5 \times \lceil \log_2 200 \rceil = 40$ bits.

b) Compare with the baseline scheme: Next, we compare SQKR with a combination of privUnit and an optimal quantizer.

i) Baseline: a direct concatenation of privUnit*, Kashin's quantizer and sampling:* For each X_i in unit ℓ_2 ball, privUnit maps it to a vector $\tilde{X}_i$ with length $\left\|\tilde{X}_i\right\|_2 = \Theta\left(\sqrt{d / \min\left(\varepsilon, \varepsilon^2\right)}\right)$. If we quantize $\tilde{X}_i$ according to its Kashin's representation and then subsample b bits from it as in Section III, then the ℓ_2 error (i.e. variance) will be

$$\tilde{O}\left(\frac{d}{b}\left\|\tilde{X}_i\right\|^2\right) = \tilde{O}\left(\frac{d^2}{b \min\left(\varepsilon, \varepsilon^2\right)}\right).$$

Therefore, averaging over n clients, the ℓ_2 error of estimating the empirical mean is

$$\tilde{O}\left(\frac{d^2}{n \cdot b \min\left(\varepsilon, \varepsilon^2\right)}\right).$$

However, in Theorem 3.1, we see that with a more sophisticated design, we can achieve smaller ℓ_2 error

$$O\left(\frac{d}{n \cdot \min\left(\varepsilon, \varepsilon^2, b\right)}\right).$$

In the experiment, we mainly focus on the *high-privacy low-communication* setting where $\varepsilon = b = 1$, and the *low-privacy high-communication* setting where $\varepsilon = b = 5$. We consider different dimensions d and plot the (log-scale) ℓ_2 estimation error (i.e. mean square error) with sample size n. For each

[6] We remark that after our paper being published, a recent work [41] shows that `DJW` and its improved version `privUnit` [7] can be compressed in a more efficient way. We refer the reader to [41] for more details.

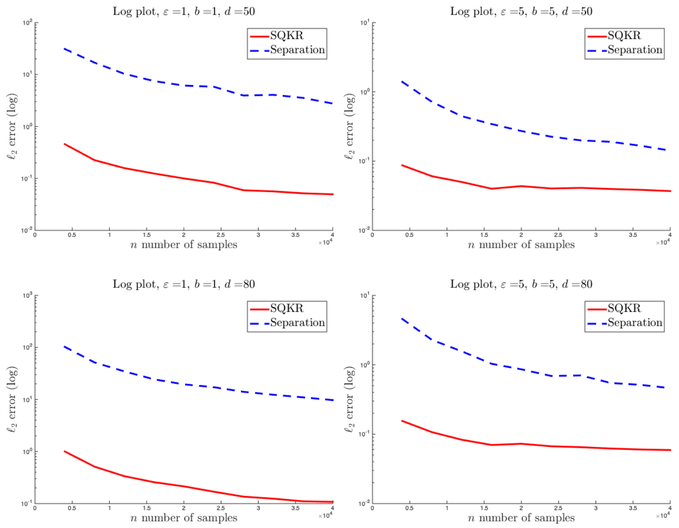

Fig. 3. Log-scale ℓ_2 error with different dimensions $d = 20, 50, 80$ and different privacy and communication budgets.

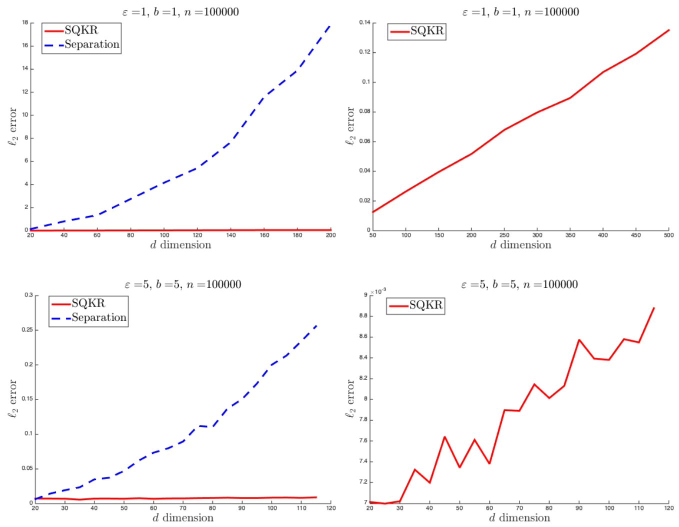

Fig. 4. ℓ_2 error with $n = 10^5$ and different dimensions d. In order to better emphasize the dependence to d, on the right-hand side we only plot the ℓ_2 error of SQKR.

point, i.e. each combination of parameters ε, b, d, n, we repeat the simulation for 8 iterations and compute the average. In Figure 3, we see that SQKR drastically outperforms the baseline (labeled as "Separation" since it is based on the idea of separately coding for privacy and communication efficiency). The gain increases in higher dimensions or with more stringent privacy/communication constraints.

In order to study the dependence on d, we fix the sample size to $n = 10^5$ and ε, b, and increase the dimension d. In Figure 4, We see that SQKR has linear dependence on d, and Separation has super-linear dependence. Therefore the performance differs drastically when d increases.

B. Frequency Estimation

For frequency estimation problem, we experimentally compare our scheme, Recursive Hadamard Response (RHR),

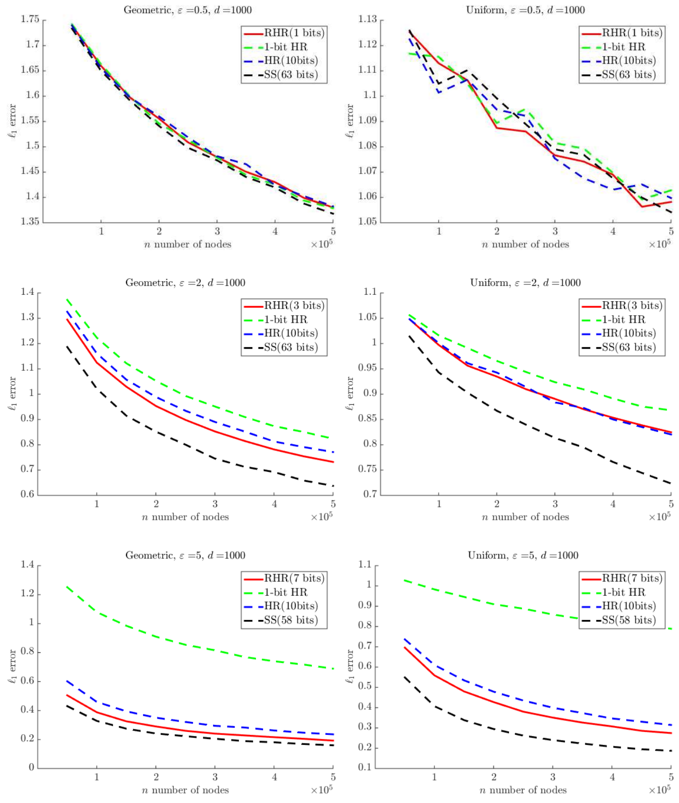

Fig. 5. ℓ_1 error with $d = 1000$. Left are $Geo(0.8)$ and right are *Uniform*.

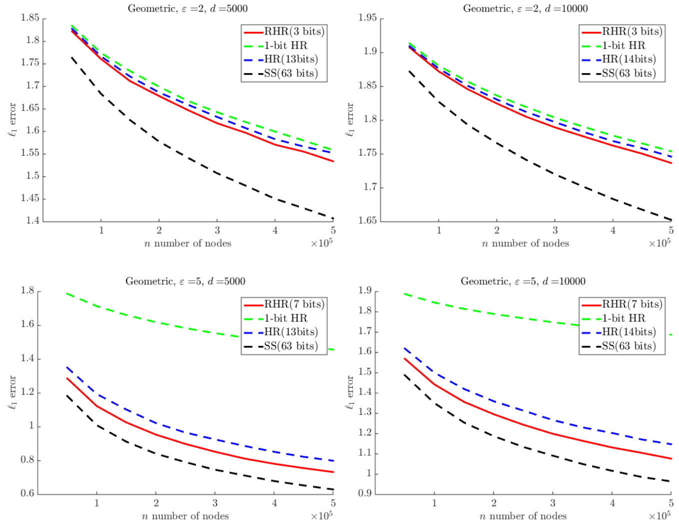

Fig. 6. ℓ_1 error with $d = 5000$ and $d = 10000$, under (truncated) $Geo(0.8)$ and different ε.

with SS [9], HR [44] and 1-bit HR [13].[7] We set $d = \{1000, 5000, 10000\}$, $\varepsilon \in \{0.5, 2, 5\}$ and $n = \{50000, 100000, \ldots, 500000\}$, and evaluate the ℓ_1 estimation errors on uniform distribution and truncated and normalized geometric distribution with $\lambda = 0.8$. For each point (i.e., for each parameter n, ε, d), we repeat the simulation 30 times and average the ℓ_2 errors. Figure 5 and Figure 6 show that RHR

[7] For HR, we use the codes from [44] (https://github.com/zitengsun/hadamard_response)

can achieve the same performance as HR but is significantly more communication efficient. For instance, in Figure 6 with $d = 10000, \varepsilon = 5$, RHR uses only half of the communication budget for HR and achieves better performance. In all settings, k-SS has the best statistical performance, but this comes with drastically higher communication and computation cost.

VIII. Conclusion

We have investigated mean estimation and frequency estimation under ε-LDP and b-bit communication constraints. A significant advantage of the approaches we presented is that they achieve the privacy and communication constraints simultaneously at the cost of the harsher one. We also study the role of shared randomness in distributed estimation and how it benefits communication costs and accuracy. Finally, we apply our mean estimation scheme SQKR to local DP-SGD and analyze its convergence rate.

Appendix A
Separate Quantization and Privatization Is Strictly Sub-Optimal

Distribution estimation: First, let us recap the subset selection (SS) scheme proposed by [9]. Assume $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \boldsymbol{p} = (p_1, \dots, p_d)$. Client i maps the local data X_i into $y \in \mathcal{Y}_{d,w} \triangleq \left\{ y \in \{0,1\}^d : \sum_j y_j = w \right\}$ with the transitional probability

$$Q_{\text{SS}}(y|X=j) = \frac{e^\varepsilon y_j + (1-y_j)}{e^\varepsilon \binom{d-1}{w-1} + \binom{d-1}{w}}.$$

The estimator for p_j is defined by

$$\hat{p}_j \triangleq \left(\frac{(d-1)e^\varepsilon + \frac{(d-1)(d-w)}{w}}{(d-w)(e^\varepsilon - 1)} \right) \frac{T_j}{n} - \frac{(w-1)e^\varepsilon + d - w}{(d-w)e^\varepsilon - 1}, \tag{23}$$

where $T_j \triangleq \sum_{i=1}^n Y_i(j)$. Note that by picking $w = \lceil \frac{d}{e^\varepsilon + 1} \rceil$, SS is order-optimal for all privacy regimes.

To demonstrate that separating privatization and quantization is strictly sub-optimal, we analyze the estimation error of directly concatenating the 2^b-SS mechanism with the grouping-based quantization in [49]. Note that both schemes are known to be optimal under the corresponding constraints, privacy and communication respectively. However, their direct combination yields an ℓ_2 error of order $O\left(d^2\right)$, which is far from the optimal accuracy established in Theorem 4.1.

We first group $[d]$ into $s = d/2^b$ equal-sized groups $\mathcal{G}_1, \dots, \mathcal{G}_s$, and each client is only responsible for sending information about one particular group. That is, let Y_i be the outcome of the 2^b-SS mechanism, i.e. $Y_i \sim Q_{\text{SS}}\left(\cdot|X_i\right)$, and client i only transmits $\{Y_i(j) | j \in \mathcal{G}_{s'}\}$, for some $s' \in [s]$. Since the server estimates each component of $\boldsymbol{p}$ separately as in (23), this grouping strategy reduces the effective sample size from n to $n' = n2^b/d$. Plugging n' into the ℓ_2 error (see [9, Proposition III.1], we conclude that the error grows as

$$O\left(\frac{d^2}{n 2^b \min\left(e^\epsilon, (e^\epsilon - 1)^2\right)} \right).$$

Note that since each Y_i contains exactly w ones, the required communication budget to describe $\{Y_i(j), j \in \mathcal{G}_l\}$ may be larger than b bits. But this is fine since it implies that even given more than b bits, the estimation error still grows with d^2. In Theorem 4.1, on the other hand, we show that the optimal ℓ_2 error is linear in d, so this demonstrates that separate quantization and privatization is sub-optimal.

Mean estimation: For the mean estimation problem, a straightforward combination is using the privUnit mechanism [7, Algorithm 1] to perturb the local data $X_i \in \mathcal{B}_d(\mathbf{0}, 1)$, and then using RandomSampling quantization in [8, Th. 6] to compress the perturbed data. Both schemes are known to be optimal under the corresponding constraints, privacy, and communication respectively. (Note that in the implementation, we replaced the RandomSampling quantization with a Kashin's quantizer, since implementing the theoretically optimal RandomSampling quantization is computationally infeasible.)

By [7, Proposition 4], the output of privUnit, denoted as

$$Z_i = \mathsf{privUnit}\left(X_i, \varepsilon\right),$$

has ℓ_2 norm of order $\Theta\left(\sqrt{\frac{d}{\min(\varepsilon, \varepsilon^2)}}\right)$. However, if we further apply RandomSampling to b bits, by Theorem 6 in [8], the ℓ_2 estimation error grows as

$$\Theta\left(\|Z_i\| \frac{d}{n \cdot b} \right) = \Theta\left(\frac{d^2}{nb \min\left(\varepsilon, \varepsilon^2\right)} \right),$$

showing a quadratic dependence in d. By Theorem 3.1, nevertheless, we can construct a better scheme with $O(d/n \min\left(\varepsilon, \varepsilon^2, b\right))$ dependence under both constraints.

Appendix B
Proof of Claims, Lemmas, and Corollaries

A. Proof of Corollary 3.1

The lower bounds follow directly from [7] (under ε-LDP constraint) and [12] (under b-bit communication constraint). For the achievability part, we apply SQKR except for replacing the random sampling step with deterministic grouping.

Let $X_i \overset{\text{i.i.d.}}{\sim} P$ with P supported on $\mathcal{B}(\mathbf{0}, 1)$. First, as in the proof of Theorem 4.1, by Lemma 3.1 we can write $X_i = \sum_{j=1}^N A_{ij} u_j$ with $N = c_0 d$ and $|A_{ij}| \leq K/\sqrt{d}, K = \Theta(1)$. Since $X_i \overset{\text{i.i.d.}}{\sim} P$, if we denote $A_i = [A_{i1}, \dots, A_{iN}]$, then $A_i \overset{\text{i.i.d.}}{\sim} Q$ for some $\tilde{P}$ supported on $\left[-\frac{K}{\sqrt{d}}, \frac{K}{\sqrt{d}} \right]^N$.

Now we group n clients into $m \triangleq N/b^*$ groups $\mathcal{G}_1, \dots, \mathcal{G}_m$, each with nb^*/N clients, where $b^* \triangleq \min\left(\lceil \varepsilon \log_2 e \rceil, b\right)$. Also, we divide all of N coordinates (of A_i) into m groups $\mathcal{I}_1, \dots, \mathcal{I}_m$, and each group of clients is responsible for estimating the corresponding group of coordinates of $\theta\left(\tilde{P}\right) \in \left[-\frac{K}{\sqrt{d}}, \frac{K}{\sqrt{d}} \right]^N$, where $\theta\left(\tilde{P}\right) = \mathbb{E}_{\tilde{P}}[A]$ is the population mean of $\tilde{P}$.

Quantization: If client i belongs to $\mathcal{G}_l$, then it quantizes A_{ij} to Q_{ij} according to

$$Q_{ij} \triangleq \begin{cases} -\frac{K}{\sqrt{d}}, \text{ with probability } \frac{K/\sqrt{d}-A_{ij}}{2K/\sqrt{d}}, \text{ if } j \in \mathcal{I}_l, \\ \frac{K}{\sqrt{d}}, \text{ with probability } \frac{A_{ij}+K/\sqrt{d}}{2K/\sqrt{d}}, \text{ if } j \in \mathcal{I}_l, \\ 0, \text{ else.} \end{cases} \tag{24}$$

Conditioned on A_i, $\{Q_{ij} \mid j \in \mathcal{I}_l\}$ yields an unbiased estimator of $\{A_{ij} \mid j \in \mathcal{I}_l\}$ and can be described by $|\mathcal{I}_l| = b^*$ bits.

Privatization: Client i then perturbs the b^*-bit message $\{Q_{ij} \mid j \in \mathcal{I}_l\}$ into $\left\{\hat{Q}_{ij} \mid j \in \mathcal{I}_l\right\}$ via 2^{b^*}-RR, as described in (7). Similarly,

$$\left\{\left(\frac{e^\varepsilon + 2^{b^*} - 1}{e^\varepsilon - 1}\right)\hat{Q}_{ij} \mid j \in \mathcal{I}_l\right\}$$

yields an unbiased estimator on $\{A_{ij} \mid j \in \mathcal{I}_l\}$.

Analysis of the ℓ_2 error: For all $j \in \mathcal{I}_l$, $\hat{A}_{ij} \triangleq \left(\frac{e^\varepsilon+2^{b^*}-1}{e^\varepsilon-1}\right)\hat{Q}_{ij}$ yields an unbiased estimator on $\mathbb{E}_{\tilde{P}}\left[A_{ij}\right]$, and note that $\hat{Q}_{ij} \in \left[-\frac{K}{\sqrt{d}}, \frac{K}{\sqrt{d}}\right]$, so the variance of $\hat{A}_{ij}$ is controlled by

$$\begin{aligned}\mathbb{E}_{\tilde{P}}\left[\left(\hat{A}_{ij} - \theta\left(Q\right)(j)\right)\right] &\leq \left(\frac{e^\varepsilon + 2^{b^*} - 1}{e^\varepsilon - 1}\right)^2\left(\frac{2K}{\sqrt{d}}\right)^2 \\ &= O\left(\frac{1}{d\min\left(1, \varepsilon^2\right)}\right).\end{aligned}$$

Since for each coordinate $j \in \mathcal{I}_l$, there are $|\mathcal{G}_l|$ clients (samples) that output independent and unbiased estimators $\hat{A}_{ij}$, the estimator $\hat{A}_j \triangleq \frac{1}{|\mathcal{G}_l|}\sum_{i\in\mathcal{G}_l}\hat{A}_{ij}$ has variance

$$O\left(\frac{1}{d\left|\mathcal{G}_l\right|}\right) = O\left(\frac{1}{n\min\left(b^*, \varepsilon^2\right)}\right).$$

Therefore, we arrive at

$$\mathbb{E}\left[\sum_{j=1}^N\left(\hat{A}_j - \mathbb{E}_{\tilde{P}}\left[A_j\right]\right)^2\right] = O\left(\frac{d}{n\min\left(b^*, \varepsilon^2\right)}\right).$$

Write $\hat{\theta} = \sum_{j=1}^N \hat{A}_j u_j$ and note that $\theta\left(P\right) = \sum_{j=1}^N \mathbb{E}_{\tilde{P}}\left[\hat{A}_j\right] u_j$, so by (3) we conclude that

$$\begin{aligned}\mathbb{E}_P\left[\|\hat{\theta} - \theta(P)\|_2^2\right] &= O\left(\frac{d}{n\min\left(b^*, \varepsilon^2\right)}\right) \\ &= O\left(\frac{d}{n\min\left(\varepsilon, \varepsilon^2, b\right)}\right).\end{aligned}$$

B. Proof of Corollary 4.1

The construction of the distribution estimation scheme mainly follows Section IV, except we replace the random sampling step by a deterministic grouping idea. We will use the same notation as in Section IV.

Encoding mechanism: We group n samples into B equal-sized groups, each with $n' = n/B$ samples. For sample $X_i \in \mathcal{G}_j$, we quantize it to a 2^{k-1}-dimensional $\{1, 0, -1\}$ vector:

$$Q_j(X_i) = \begin{bmatrix} (H_B)_j \cdot X_i^{(1)} \\ (H_B)_j \cdot X_i^{(2)} \\ \vdots \\ (H_B)_j \cdot X_i^{(2^{k-1})} \end{bmatrix} \in \{-1, 0, 1\}^{2^{k-1}}.$$

Since X_i is one-hot encoded, there is only one $l \in \{1, \dots, 2^{k-1}\}$ such that $(H_B)_j \cdot X_i^{(l)} \neq 0$, so $Q_j(X_i)$ can be described by k bits (1 bit for the sign and $(k-1)$ bits for the location of the non-zero element). Also, notice that

$$\mathbb{E}\left[Q_j(X_i)\right] = \begin{bmatrix} (H_B)_j \cdot \boldsymbol{p}^{(1)} \\ (H_B)_j \cdot \boldsymbol{p}^{(2)} \\ \vdots \\ (H_B)_j \cdot \boldsymbol{p}^{(2^{k-1})} \end{bmatrix},$$

where $\boldsymbol{p}^{(l)} \triangleq \boldsymbol{p}[(l-1)B + 1 : lB]$. By (11), the estimator $\hat{q}_{j'} = \langle (H_{2^{k-1}})_m, Q_j(X_i)\rangle$ is unbiased for $q_{j'}$ (where $j' = (m-1)B + j$).

We further perturb Q_j via 2^k-RR scheme, since Q takes values on an alphabet of size 2^k, denoted by $\mathcal{Q} = \{\pm e_1, \dots, \pm e_{2^{k-1}}\}$,

$$\tilde{Q}_j = \begin{cases} Q_j, \text{ w.p. } \frac{e^\varepsilon}{e^\varepsilon + 2^k - 1} \\ Q' \in \mathcal{Q} \setminus \{Q_j\}, \text{ w.p. } \frac{1}{e^\varepsilon + 2^k - 1}, \end{cases}$$

where $\boldsymbol{e}_l$ denotes the l-th coordinate vector in $\mathbb{R}^{2^{k-1}}$. This gives us

$$\mathbb{E}\left[\tilde{Q}_j\right] = \frac{e^\varepsilon - 1}{e^\varepsilon + 2^k - 1}\mathbb{E}\left[Q_j\right].$$

Therefore $\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\tilde{Q}_j$ yields an unbiased estimator of

$$\begin{bmatrix} (H_B)_j \cdot \boldsymbol{p}^{(1)} \\ (H_B)_j \cdot \boldsymbol{p}^{(2)} \\ \vdots \\ (H_B)_j \cdot \boldsymbol{p}^{(2^{k-1})} \end{bmatrix}.$$

Constructing the estimator for $\mathbf{p}$: For each $j' \equiv j \pmod{B}$, we estimate $(H_{2^{k-1}})_m \cdot Q_j(X_i), i \in \mathcal{G}_j$ (recall that $j' = j + (m-1)B$). Define the estimator

$$\begin{aligned}&\hat{q}_{j'}\left(\{X_i, i \in \mathcal{G}_j\}\right) \\ &= \frac{1}{|\mathcal{G}_j|}\sum_{i\in\mathcal{G}_j}(H_{2^{k-1}})_m \cdot \left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right)\tilde{Q}_j(X_i) \\ &= \frac{B}{n}\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right)\sum_{i\in\mathcal{G}_j}(H_{2^{k-1}})_m\tilde{Q}_j(X_i).\end{aligned}$$

The MSE of $\hat{q}_{i'}$ can be obtained by

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{q}_{j'} - q_{j'}\right)^2\right] \\ &\overset{\text{(a)}}{=} \mathsf{Var}\left(\hat{q}_{i'}\right) \\ &\overset{\text{(b)}}{=} \frac{d}{n2^{k-1}}\left(\frac{e^\varepsilon + 2^k - 1}{e^\varepsilon - 1}\right)^2 \mathsf{Var}\left(\left(H_{2^{k-1}}\right)_m \cdot \tilde{Q}_j(X_i)\right)\end{aligned}$$

$$\overset{\text{(c)}}{\le} \frac{d}{n2^{k-1}}\left(\frac{e^{\varepsilon}+2^{k}-1}{e^{\varepsilon}-1}\right)^{2}, \tag{25}$$

where (a) is due to the unbiasedness of $\hat{q}_{j'}$, (b) is due to the independence across X_i, and (c) is because $\langle (H_{2^{k-1}})_m, \tilde{Q}_j\rangle$ only takes value in $\{-1,1\}$.

Finally, let $\hat{\boldsymbol{p}}$ be the inverse Hadamard transform of $\hat{\boldsymbol{q}}$, the MSE is

$$\begin{aligned}
\mathbb{E}\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_2^2 &= \mathbb{E}\left[\langle \hat{\boldsymbol{p}}-\boldsymbol{p}, \hat{\boldsymbol{p}}-\boldsymbol{p}\rangle\right] \\
&= \mathbb{E}\left[\left(\hat{\boldsymbol{q}}-\boldsymbol{q}\right)^{\intercal}\left(H_d^{-1}\right)^{\intercal} H_d^{-1}\left(\hat{\boldsymbol{q}}-\boldsymbol{q}\right)\right] \\
&= \frac{1}{d}\mathbb{E}\left\|\hat{\boldsymbol{q}}-\boldsymbol{q}\right\|_2^2 \\
&\le \frac{d}{n2^{k}}\left(\frac{e^{\varepsilon}+2^{k}-1}{e^{\varepsilon}-1}\right)^{2} \\
&= O\left(\frac{d}{n2^{k}}\left(\frac{e^{\varepsilon}+2^{k}}{e^{\varepsilon}-1}\right)^{2}\right),
\end{aligned}$$

where the last inequality holds due to (25).

Picking $k=\min\left(b, \lceil \varepsilon\log_2 e\rceil, \lfloor \log d\rfloor\right)$ yields

$$\mathbb{E}\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_2^2 = O\left(\frac{d}{n\min\left(2^b, e^{\varepsilon}, d\right)}\left(\frac{e^{\varepsilon}}{e^{\varepsilon}-1}\right)^{2}\right).$$

Observe that if $e^{\varepsilon}=O(2^b)$, then $e^{\varepsilon}\preceq 2^b$, so $\mathbb{E}\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_2^2 = O\left(\frac{de^{\varepsilon}}{n(e^{\varepsilon}-1)^2}\right)$. On the other hand, if $e^{\varepsilon}=\Omega(2^b)$, then $\frac{e^{\varepsilon}}{e^{\varepsilon}-1}=\theta(1)$, and $\mathbb{E}\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_2^2 = O\left(\frac{d}{n\min(2^b,d)}\right)$.

Therefore we conclude that

$$\begin{aligned}
\mathbb{E}\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_2^2 &\preceq \max\left(\frac{d}{n\min\left(2^b,d\right)}, \frac{de^{\varepsilon}}{n\left(e^{\varepsilon}-1\right)^2}\right) \\
&\asymp \frac{d}{n}\left(\frac{1}{\min\left\{e^{\varepsilon}, \left(e^{\varepsilon}-1\right)^2, 2^b, d\right\}}\right).
\end{aligned}$$

Finally, by Jensen's inequality and Cauchy-Schwarz inequality, we also have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_1\right] &\le \left(\mathbb{E}\left[\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_1^2\right]\right)^{\frac{1}{2}} \le \left(d\cdot\mathbb{E}\left\|\hat{\boldsymbol{p}}-\boldsymbol{p}\right\|_2^2\right)^{\frac{1}{2}} \\
&\preceq \frac{d}{\sqrt{n\left(\min\left\{e^{\varepsilon}, \left(e^{\varepsilon}-1\right)^2, 2^b, d\right\}\right)}},
\end{aligned}$$

establishing the achievability part of Theorem 4.1.

C. Proof of Corollary 5.1 and Corollary 5.2

Notice that since one can always "simulate" the public coin by uplink communication (i.e. each client generates its private random bits and send them to the server), any b bits public-coin scheme can be cast into a private coin scheme with additional b bits communication. This implies the above impossibility results (Lemma 5.1) also serve a valid lower bound for the amount of public randomness: for any public-coin scheme with $b < \log d - 2$ bits communication budgets, we need at least $\log d - b - 2$ bits of shared randomness in order to obtain a consistent estimate of the empirical mean or empirical frequency.

D. Proof of Lemma 5.1

Without access to the public randomness, [13] shows that at least $\Theta(d)$ bits of communication is required for heavy hitter estimation in order to obtain a consistent estimator.[8] We state their result here:

Lemma 2.1 ([13] Theorem 4): Let $b \le \log d - 2$. For all private-coin schemes $\left(Q^n, \hat{D}\right)$ with only private randomness and b bits communication budgets, there exists a data sets $X_1, \dots, X_n$ with $n > 12(2^b+1)^2$, such that

$$\mathbb{E}\left[\left\|\hat{D}\left(Q^n\right) - D_{X^n}\right\|_{\infty}\right] \ge \frac{1}{2^{b+2}+4}.$$

Based on this, we claim that without a public coin, each client needs to transmit at least $\Theta(\log d)$ bits in order to construct consistent schemes for frequency estimation or mean estimation.

Frequency estimation: We lower bound ℓ_1 and ℓ_2 error by ℓ_∞ and apply Lemma 2.1.

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}\left(Q^n\right) - D_{X^n}\right\|_{1}\right] &\ge \mathbb{E}\left[\left\|\hat{D}\left(Q^n\right) - D_{X^n}\right\|_{\infty}\right] \\
&\ge \frac{1}{2^{b+2}+4}, \text{ and}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}\left(Q^n\right) - D_{X^n}\right\|_{2}^{2}\right] &\ge \mathbb{E}\left[\left\|\hat{D}\left(Q^n\right) - D_{X^n}\right\|_{\infty}^{2}\right] \\
&\ge \left(\mathbb{E}\left[\left\|\hat{D}\left(Q^n\right) - D_{X^n}\right\|_{\infty}\right]\right)^{2} \\
&\ge \left(\frac{1}{2^{b+2}+4}\right)^{2}.
\end{aligned} \tag{26}$$

This implies that it is impossible to construct consistent schemes with less than $\log d - 2$ bits per client in the absence of a public randomness. On the other hand, given $\log d$ bits, one can readily achieve the optimal estimation accuracy without any public randomness, for instance, by using Hadamard response [44] (see also the discussion in [13]). Therefore, the problem of frequency estimation is somewhat trivialized in the absence of public randomness.

Mean estimation: Let $X_i \in [d]$ be one-hot encoded, so $X_i \in \mathcal{B}_d(\mathbf{0}, 1)$. Then (26) implies the ℓ_2 error of mean estimation is at least $1/\left(2^{b+2}+4\right)^2$. Thus with less than $\log d - 2$ bits of communication budget, it is also impossible to construct a consistent scheme for mean estimation. □

E. Proof of Claim 3.1

According to (3), it suffices to control $\mathsf{Var}\left(\hat{a}_j\right)$. To bound the variance, consider

$$\begin{aligned}
&\mathsf{Var}\left(\hat{a}_j\right) \\
&= \frac{N^2}{k^2}\cdot\left(\frac{e^{\varepsilon}+2^{k}-1}{e^{\varepsilon}-1}\right)^{2}\mathsf{Var}\left(\sum_{m=1}^{k}\tilde{q}_m\cdot\mathbb{1}_{\{j=s_m\}}\right) \\
&\le \frac{N^2}{k^2}\cdot\left(\frac{e^{\varepsilon}+2^{k}-1}{e^{\varepsilon}-1}\right)^{2}\mathbb{E}\left[\left(\sum_{m=1}^{k}\tilde{q}_m\cdot\mathbb{1}_{\{j=s_m\}}\right)^{2}\right]
\end{aligned}$$

[8] Recall that an estimator is consistent if it has vanishing estimation error as n tends to infinity.

$$\overset{\text{(a)}}{\le} \frac{N^2}{k^2}\cdot\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\right)^2\left(\frac{c}{\sqrt{d}}\right)^2\mathbb{E}\left[\left(\sum_{m=1}^{k}\mathbb{1}_{\{j=s_m\}}\right)^2\right]$$

$$\overset{\text{(b)}}{\le} C\frac{N}{k^2}\cdot\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\right)^2\left(\frac{k^2}{N^2}+\frac{k}{N}\right)$$

$$= C\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\right)^2\left(\frac{1}{N}+\frac{1}{k}\right),$$

where (a) is due to $|\tilde{q}_m| = \frac{c}{\sqrt{d}}$, and (b) is due to the second moment bound on $\mathsf{Binomial}(k, 1/N)$ and the fact $N = \Theta(d)$. Therefore by (3),

$$\mathbb{E}\left[\left\|\hat{X}-X\right\|_2^2\right] \le C_0\sum_{i=1}^{N}\mathsf{Var}\left(\hat{a}_i\right) \le C_1\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\right)^2\frac{d}{k},$$

establishing the claim. □

F. Proof of Claim 4.1

$\hat{Y}_i$ yields an unbiased estimator since

$$\begin{aligned}
&\mathbb{E}\left[\hat{Y}_i\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\tilde{Q}_i, r_i\right)\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\hat{Y}_i\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\tilde{Q}_i, r_i\right)\middle| r_i\right]\right]\\
&\overset{\text{(a)}}{=} \mathbb{E}\left[\hat{Y}_i\left(\mathbb{E}\left[\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\tilde{Q}_i\middle| r_i\right], r_i\right)\right]\\
&= \mathbb{E}\left[\hat{Y}_i\left(Q(X_i, r_i), r_i\right)\right]\\
&= \frac{1}{d}H_d X_i, \qquad (27)
\end{aligned}$$

where (a) holds since conditioning on r_i, $\hat{Y}_i(Q, r_i)$ is a linear function of Q. □

G. Proof of Claim 4.2

The ℓ_2 error is

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|H_d\hat{Y}_i - H_d\mathbb{E}\left[\hat{Y}_i\right]\right\|_2^2\right]\\
&= \frac{d}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\hat{Y}_i - \mathbb{E}\left[\hat{Y}_i\right]\right\|_2^2\right]. \qquad (28)
\end{aligned}$$

It remains to bound $\mathbb{E}\left[\left\|\hat{Y}_i - \mathbb{E}\left[Y_i\right]\right\|_2^2\right]$. Observe that

$$\left|\mathbb{E}[\hat{Y}_i]\right| = \left|\frac{H_d\cdot X_i}{d}\right| = [1/d, \ldots, 1/d]^\intercal,$$

and from expression (13), given r_i, there are only 2^{k-1} non-zero coordinates, each with value bounded by $\left(\frac{e^\varepsilon+2^k-1}{e^\varepsilon-1}\right)/2^{k-1}$. Therefore we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{Y}_i-\mathbb{E}\left[\hat{Y}_i\right]\right\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{Y}_i-\mathbb{E}\left[\hat{Y}_i\right]\right\|_2^2\middle| r_i\right]\right]\\
&\le 2\left(d\left(\frac{1}{d}\right)^2+2^{k-1}\left(\frac{e^\varepsilon+2^k-1}{2^{k-1}\left(e^\varepsilon-1\right)}\right)^2\right).
\end{aligned}$$

Plugging this in to (28), we arrive at

$$\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] \preceq \frac{d}{n2^{k-1}}\left(\frac{e^\varepsilon+2^k-1}{(e^\varepsilon-1)}\right)^2.$$

Picking $k = \min\left(b, \lceil\varepsilon\log_2 e\rceil, \lfloor\log d\rfloor\right)$ yields

$$\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] = O\left(\frac{d}{n\min\left(2^b, e^\varepsilon, d\right)}\left(\frac{e^\varepsilon}{e^\varepsilon-1}\right)^2\right).$$

Observe that

(i) if $e^\varepsilon = O(2^b)$, then $e^\varepsilon \preceq 2^b$, so $\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] = O\left(\frac{de^\varepsilon}{n(e^\varepsilon-1)^2}\right)$.

(ii) If $e^\varepsilon = \Omega(2^b)$, then $\frac{e^\varepsilon}{e^\varepsilon-1} = \theta(1)$, and $\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] = O\left(\frac{d}{n\min(2^b,d)}\right)$.

Therefore we conclude that

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_2^2\right] &\preceq \max\left(\frac{d}{n\min\left(2^b,d\right)}, \frac{de^\varepsilon}{n\left(e^\varepsilon-1\right)^2}\right)\\
&\asymp \frac{d}{n}\left(\frac{1}{\min\left\{e^\varepsilon, \left(e^\varepsilon-1\right)^2, 2^b, d\right\}}\right).
\end{aligned}$$

By Jensen's inequality and Cauchy-Schwarz inequality, we also have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_1\right] &\le \left(\mathbb{E}\left[\left\|\hat{D}-D_{X^n}\right\|_1^2\right]\right)^{\frac{1}{2}}\\
&\le \left(d\cdot\mathbb{E}\left\|\hat{D}-D_{X^n}\right\|_2^2\right)^{\frac{1}{2}}\\
&\preceq \frac{d}{\sqrt{n\left(\min\left\{e^\varepsilon, \left(e^\varepsilon-1\right)^2, 2^b, d\right\}\right)}}.
\end{aligned}$$

□

Acknowledgment

The authors would like to thank Jakub Konečný for bringing Kashin's representation to their attention. This was helpful in achieving orderwise optimality for mean estimation. They would also like to thank Vitaly Feldman and Kunal Talwar for pointing out a mistake in the experiments of mean estimation as well as the connection between SQKR and [31].

References

[1] F. Niu, B. Recht, C. Re, and S. J. Wright, "Hogwild! A lock-free approach to parallelizing stochastic gradient descent," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2011, pp. 693–701.

[2] P. Kairouz et al., "Advances and open problems in federated learning," 2019, *arXiv:1912.04977.*

[3] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[4] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2003, pp. 211–222.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.* Berlin, Germany: Springer, Mar. 2006, pp. 265–284.

[6] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, Jun. 2011.

[7] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018, *arXiv:1812.00984*.

[8] V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar, "vqSGD: Vector quantized stochastic gradient descent," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, 2019, pp. 2197–2205.

[9] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under local differential privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 759–763.

[10] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proc. 47th Annu. ACM Symp. Theory Comput. (STOC)*. New York, NY, USA: Association for Computing Machinery, Jun. 2015, pp. 127–135, doi: 10.1145/2746539.2746632.

[11] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2328–2336.

[12] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 3329–3337.

[13] J. Acharya and Z. Sun, "Communication complexity in locally private distribution estimation and heavy hitters," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 51–60.

[14] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, 2019, pp. 2468–2479.

[15] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Proc. Annu. Int. Cryptol. Conf.* Cham, Switzerland: Springer, Aug. 2019, pp. 638–667.

[16] Ú. Erlingsson et al., "Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation," 2020, *arXiv:2001.03618*.

[17] R. Bassily, K. Nissim, U. Stemmer, and A. Thakurta, "Practical locally private heavy hitters," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 2285–2293.

[18] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7564–7575.

[19] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[20] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon et al., Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1709–1720. [Online]. Available: http://papers.nips.cc/paper/6768-qsgd-communication-efficient-sgd-via-gradient-quantization-and-encoding.pdf

[21] W. Wen et al., "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1509–1519.

[22] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, "ATOMO: Communication-efficient learning via atomic sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9850–9861.

[23] L. P. Barnes, H. A. Inan, B. Isik, and A. Ozgur, "rTOP-k: A statistical estimation approach to distributed SGD," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 897–907, Nov. 2020.

[24] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1299–1309.

[25] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Proc. 48th Annu. ACM Symp. Theory Comput.*, Jun. 2016, pp. 1011–1020.

[26] Y. Lyubarskii and R. Vershynin, "Uncertainty principles and vector quantization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3491–3501, Jul. 2010.

[27] J.-J. Fuchs, "Spread representations," in *Proc. Conf. Rec. 55th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 814–817.

[28] C. Studer, W. Yin, and R. G. Baraniuk, "Signal representations with minimum ℓ_∞-norm," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 1270–1277.

[29] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," 2018, *arXiv:1812.07210*.

[30] M. Safaryan, E. Shulgin, and P. Richtárik, "Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor," 2020, *arXiv:2002.08958*.

[31] V. Feldman, C. Guzmán, and S. Vempala, "Statistical query algorithms for mean vector estimation and stochastic convex optimization," in *Proc. 28th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2017, pp. 1265–1277.

[32] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin, "Collecting and analyzing data from smart device users with local differential privacy," 2016, *arXiv:1606.05053*.

[33] T. Wang, J. Zhao, X. Yang, and X. Ren, "Locally differentially private data collection and analysis," 2019, *arXiv:1906.01777*.

[34] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 429–438.

[35] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proc. 26th USENIX Secur. Symp. (USENIX Security)*, 2017, pp. 729–745.

[36] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. 21st ACM SIGSAC Conf. Comput. Commun. Secur.*, Scottsdale, AZ, USA, Nov. 2014, pp. 1054–1067.

[37] J. Hsu, S. Khanna, and A. Roth, "Distributed private heavy hitters," in *Proc. 39th Int. Colloq. Conf. Automata, Lang., Program. (ICALP)*. Berlin, Germany: Springer-Verlag, 2012, pp. 461–472, doi: 10.1007/978-3-642-31594-7_39.

[38] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 192–203, doi: 10.1145/2976749.2978409.

[39] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," in *Proc. 37th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.* New York, NY, USA: Association for Computing Machinery, May 2018, pp. 435–447, doi: 10.1145/3196959.3196981.

[40] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the communication-privacy-accuracy trilemma," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–13.

[41] V. Feldman and K. Talwar, "Lossless compression of efficient private local randomizers," 2021, *arXiv:2102.12099*.

[42] S. Wang et al., "Mutual information optimally local private discrete distribution estimation," 2016, *arXiv:1607.08025*.

[43] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, New York, NY, USA, Jun. 2016, pp. 2436–2444.

[44] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1120–1129.

[45] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proc. Conf. Learn. Theory*, 2019, pp. 3–17.

[46] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints II: Communication constraints and shared randomness," 2019, *arXiv:1905.08302*.

[47] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," 2018, *arXiv:1802.08417*.

[48] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2726–2734.

[49] Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman, "Distributed statistical estimation of high-dimensional and nonparametric distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 506–510.

[50] L. P. Barnes, Y. Han, and A. Özgür, "Lower bounds for learning distributions under communication constraints via Fisher information," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 9583–9612, 2019.

[51] L. P. Barnes, W.-N. Chen, and A. Özgür, "Fisher information under local differential privacy," 2020, *arXiv:2005.10783*.

[52] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 492–542, Jan. 2016.

[53] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[54] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2291–2295.

[55] O. Shamir, "Without-replacement sampling for stochastic gradient methods," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 46–54.

[56] D. Nagaraj, P. Jain, and P. Netrapalli, "SGD without replacement: Sharper rates for general smooth convex functions," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4703–4711.

[57] S. Rajput, A. Gupta, and D. Papailiopoulos, "Closing the convergence gap of SGD without replacement," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7964–7973.

[58] K. Ahn, C. Yun, and S. Sra, "SGD with shuffling: Optimal rates without component convexity and large epoch requirements," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17526–17535.

[59] S. Ghadimi and G. H. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[60] S. Asoodeh, W.-N. Chen, F. P. Calmon, and A. Özgür, "Differentially private federated learning: An information-theoretic perspective," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 344–349.

[61] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.

[62] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[63] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *J. ACM*, vol. 61, no. 6, pp. 1–57, Dec. 2014.

[64] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proc. IEEE 55th Annu. Symp. Found. Comput. Sci.*, Oct. 2014, pp. 464–473.

[65] R. Bassily, V. Feldman, K. Talwar, and A. Guha Thakurta, "Private stochastic convex optimization with optimal rates," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[66] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar, "Stability of stochastic gradient descent on nonsmooth convex losses," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4381–4391.

[67] V. Feldman, T. Koren, and K. Talwar, "Private stochastic convex optimization: Optimal rates in linear time," in *Proc. 52nd Annu. ACM SIGACT Symp. Theory Comput.*, Jun. 2020, pp. 439–449.

Wei-Ning Chen received the B.Sc. degree in electrical engineering and mathematics and the M.S. degree in communication engineering from National Taiwan University, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Stanford University. He was a recipient of the Stanford Graduate Fellowship (SGF). His research interests include information theory, statistics, and theoretical machine learning, with applications in differential privacy and federated learning.

Peter Kairouz is a Research Scientist with Google, where he leads research efforts focused on federated learning and privacy-preserving technologies. He was a recipient or co-recipient of the 2012 Roberto Padovani Scholarship from Qualcomm's Research Center, the 2015 ACM SIGMETRICS Best Paper Award, the 2015 Qualcomm Innovation Fellowship Finalist Award, the 2016 Harold L. Olesen Award for Excellence in Undergraduate Teaching from UIUC, and the 2021 ACM CCS Best Paper Award.

Ayfer Özgür is an Associate Professor with the Electrical Engineering Department, Stanford University, where she is the Chambers Faculty Scholar with the School of Engineering. Her research interests include information theory, wireless communication, statistics, and machine learning. She received the EPFL Best Ph.D. Thesis Award in 2010, the NSF CAREER award in 2013, the Okawa Foundation Research Grant, the Faculty Research Awards from Google and Facebook, and the IEEE Communication Theory Technical Committee (CTTC) Early Achievement Award in 2018. She was selected as the Inaugural Goldsmith Lecturer of the IEEE ITSoc in 2020.