Linear Queries Estimation with Local Differential Privacy

Raef Bassily
Department of Computer Science and Engineering
The Ohio State University
bassily.1@osu.edu

Abstract

We study the problem of estimating a set of d linear queries over some unknown distribution based on a sensitive data set under the constraint of local differential privacy (LDP). Let $\mathcal{J}$ be a data domain of size J. A linear query is uniquely identified by a vector $\mathbf{q} \in \mathbb{R}^J$, and is defined as the linear function $\langle \mathbf{q}, \cdot \rangle : \mathsf{Simplex}(J) \rightarrow \mathbb{R}$, where $\mathsf{Simplex}(J)$ is the probability simplex in $\mathbb{R}^J$. Given a set $D = \{v_i \in \mathcal{J} : \ i \in [n]\}$ of private data items of n individuals drawn i.i.d. from some unknown distribution $\mathbf{p} \in \mathsf{Simplex}(J)$, we wish to estimate the values of a set of d linear queries $\mathbf{q}_1, \ldots, \mathbf{q}_d$ over $\mathbf{p}$ under LDP. This problem subsumes a wide range of estimation tasks including distribution estimation and d-dimensional mean estimation. We provide new algorithms for both the offline (non-adaptive) and the adaptive versions of this problem.

In the offline setting, the set of queries are determined and fixed at the beginning of the algorithm. In the regime where $n \lesssim d^2/\log(J)$, our algorithms have L_2 estimation error (with respect to the distribution $\mathbf{p}$) that is independent of d, and is tight up to a factor of $\tilde{O}\left(\log^{1/4}(J)\right)$. Our algorithms combine different ideas such as L_2 projection on convex polytopes and rejection sampling. For the special case of distribution estimation, we show that projecting the output estimate of an algorithm due to [ASZ18] on the probability simplex yields an L_2 error that depends only sub-logarithmically on J in the regime where $n \lesssim J^2/\log(J)$. These results show the possibility of accurate estimation of linear queries in the high-dimensional settings under the L_2 error criterion.

In the adaptive setting, the queries are generated over d rounds; one query at a time. At the start of each round $k \in [d]$, a query $\mathbf{q}_k$ can be chosen *adaptively* based on all the history of previous queries and answers. We give an algorithm for this problem with optimal L_∞ estimation error (worst error in the estimated values for the queries w.r.t. the data distribution). Our bound matches a lower bound on the L_∞ error for the *offline* version of this problem [DJW13b].

1 Introduction

Differential privacy [DMNS06] is a rigorous mathematical definition that has emerged as one of the most successful notions of privacy in statistical data analysis. Differential privacy provides a rich and powerful algorithmic framework for private data analysis, which can help organizations mitigate users' privacy concerns. There are two main models for private data analysis that are studied in the literature of differential privacy: the centralized model and the local model. The centralized model assumes a trusted centralized curator that collects all the personal information and then analyzes it. In contrast, the *local model*, which dates back to [War65], does not involve a central repository. Instead, each individual holding a piece of private data randomizes her data herself via a local randomizer before it is collected for analysis. This local randomizer is designed to satisfy differential privacy, providing a strong privacy protection for each individual. The local model is attractive in many practical and industrial domains since it relieves organizations and companies from the liability of holding and securing their users private data. Indeed, in the last few years there have been many successful deployments of local differentially private algorithms in the industrial domain, most notably by Google and Apple [EPK14, TVV+17].

In this paper, we study the problem of linear queries estimation under local differential privacy (LDP). Let $\mathcal{J} = [J]$ be a data domain of size J. A linear query with respect to $\mathcal{J}$ is uniquely identified by a vector $\mathbf{q} \in \mathbb{R}^J$ that describes a linear function $\langle \mathbf{q}, \cdot \rangle : \mathsf{Simplex}(J) \to \mathbb{R}$, where $\mathsf{Simplex}(J)$ denotes the probability simplex in $\mathbb{R}^J$. In this problem, we have a set of n individuals (users), where each user $i \in [n]$ holds a private value $v_i \in \mathcal{J}$ drawn independently from some *unknown* distribution $\mathbf{p} \in \mathsf{Simplex}(J)$. An entity (server) generates a sequence of linear queries $\mathbf{q}_1, \ldots, \mathbf{q}_d$ and wishes to estimate, within a small error, the values of these queries over the unknown distribution $\mathbf{p}$, i.e., $\langle \mathbf{q}_1, \mathbf{p} \rangle, \ldots, \langle \mathbf{q}_d, \mathbf{p} \rangle$. To do this, the server collects signals from the users about their inputs and use them to generate these estimates. Due to privacy concerns, the signal sent by each user is generated via a local randomizer that outputs a randomized (privatized) version of the user's true input in a way that satisfies LDP. The goal is to design a protocol that enables the server to derive accurate estimates for its queries under the LDP constraint. This problem subsumes a wide class of estimation tasks under LDP, including distribution estimation studied in [DJW13b, BS15, DHS15, KBR16, BNST17, YB18, ASZ18] and mean estimation in d dimensions [DJW13a, DJW13b].

Non-adaptive versus Adaptive Queries: In this work, we consider two versions for the above problem. In the non-adaptive (*offline*) version, the set of d queries $\mathbf{q}_1, \ldots, \mathbf{q}_d$ are decided by the server before the protocol starts (i.e., before users send their signals). In this case, the set of d queries can be represented as the rows of a matrix $\mathbf{A} \in \mathbb{R}^{d \times J}$ that is published before the protocol starts. In the *adaptive* version of this problem, the d queries are submitted and answered over d rounds: one query in each round. Before the start of each round $k \in [d]$, the server can *adaptively* choose the query $\mathbf{q}_k$ based on all the history it sees, i.e., based on all the previous queries and signals from users in the past $k-1$ rounds. This setting is clearly harder than the offline setting. Both distribution estimation and mean estimation over a finite (arbitrary large) domain can be viewed as special cases of the offline queries model above. In particular, for distribution estimation, the queries matrix $\mathbf{A}$ is set to $\mathbb{I}_J$, the identity matrix of size J (in such case, the dimensionality $d = J$). For d-dimensional mean estimation, the columns of $\mathbf{A}$ are viewed as the set of all realizations of a d-dimensional random variable.

One of the main challenges in the local model is dealing with high-dimensional settings (i.e., when $d \gtrsim n$). Previous constructions for distribution estimation [DJW13b, KBR16, YB18, ASZ18] and mean estimation [DJW13b] suffer from an explicit polynomial dependence on the dimensions in the resulting L_2 estimation error.

In this work, we address this challenge and give new constructions for large, natural families of offline linear queries that subsumes the above estimation problems. The resulting L_2 estimation error[1] has no dependence on d in the high-dimensional setting and depends only sub-logarithmically on J. We also consider the adaptive version of the general linear queries problem, and give a new protocol with optimal L_∞ error (which is a more natural error criterion in the adaptive setting). We discuss these results below.

[1] In this work, we consider the true population risk not the empirical risk. We refer to it as the estimation error and sometimes as the *true* error.

1.1 Results and comparison to previous works

The accuracy guarantees of our ϵ-LDP protocols are summarized in Table 1.

General offline linear queries: We assume that the L_2 norm of any column of the queries matrix $\mathbf{A} \in \mathbb{R}^{d\times J}$ is bounded from above by some arbitrary constant $r > 0$. We note that this is weaker assumption than assuming that the spectral norm of $\mathbf{A}$ (largest singular value) is bounded by r. For any $r > 0$, let $\mathcal{C}_2(r)$ denote the collection of all matrices in $\mathbb{R}^{d\times J}$ satisfying this condition. We design ϵ-LDP protocol that given any queries matrix $\mathbf{A}$ from this family, it outputs an estimate for $\mathbf{Ap}$ with nearly optimal L_2 estimation error (see Section 2.2.1 for the definition of the L_2 estimation error). As noted earlier, the resulting L_2 estimation error does not depend on d in the high-dimensional setting: in particular, in the case where $n \lesssim d^2/\log(J)$ (which subsumes the high-dimensional setting when $\log(J) \lesssim d$). This improves over the upper bound in [DJW13b, Proposition 3] achieved by the ball sampling mechanism proposed therein. The near optimality of our protocol follows from the lower bound in the same reference (see Table 1). To construct our protocol, we start with an (ϵ, δ)-LDP protocol that employs the Gaussian mechanism together with the projection technique similar to the one used in [NTZ13] in the *centralized* model of differential privacy. We show the applicability of this technique in the local model. Next, we transform our (ϵ, δ)-LDP construction into a pure ϵ-LDP construction while maintaining the same accuracy (and the same computational cost). To do this, we give a technique based on rejection sampling ideas from [BS15, BNS18]. In particular, our technique can be viewed as a simpler, more direct version of the generic transformation of [BNS18] tuned to the linear queries problem. For this general setting, we focus on improving the estimation error. We do not consider the problem of optimizing communication or computational efficiency. We think that providing a succinct description of the queries matrix (possibly under more assumptions on its structure) is an interesting problem, which we leave to future work.

Problem/Error metric	Upper bound (This work)	Upper bound (Previous work)	Lower bound
General offline queries (L_2 error)	$r \cdot \min\left(\left(\frac{\log(J)\log(n)}{n\epsilon^2}\right)^{1/4}, \sqrt{\frac{d}{n\epsilon^2}}\right)$	$r \cdot \sqrt{\frac{d}{n\epsilon^2}}$ [DJW13b, Prop. 3]	$r \cdot \min\left(\left(\frac{1}{n\epsilon^2}\right)^{1/4}, \sqrt{\frac{d}{n\epsilon^2}}\right)$ ([DJW13b, Prop. 3])
Distribution estimation (L_2 error)	$\min\left(\left(\frac{\log(J)}{n\epsilon^2}\right)^{1/4}, \sqrt{\frac{J}{n\epsilon^2}}\right)$	$\sqrt{\frac{J}{n\epsilon^2}}$ [ASZ18, Thm. 3]	$\min\left(\left(\frac{1}{n\epsilon^2}\right)^{1/4}, \sqrt{\frac{J}{n\epsilon^2}}\right)$ ([DJW13b, YB18])
General adaptive queries (L_∞ error)	$r\sqrt{\frac{c_\epsilon^2 d\log(d)}{n}}$	–	$r\sqrt{\frac{c_\epsilon^2 d\log(d)}{n}}$ ([DJW13b, Prop. 4] for offline queries)

Table 1: Error bounds for the proposed ϵ-LDP protocols with comparison to previous results. Since the error in each case cannot exceed the trivial error r, each upper bound should be understood as the min of the stated bound and r.

Distribution estimation: For this special case, we extend the Hadamard-Response protocol of [ASZ18] to the high-dimensional setting. This protocol enjoys several computational advantages, particularly, $O(\log(J))$ communication and running time for each user. We show that this protocol when combined with a projection step onto the probability simplex gives L_2 estimation error that depends only sub-logarithmically on J for all $n \lesssim J^2/\log(J)$. The resulting error is also tight up to a sub-logarithmic factor in J. We note that the L_2 error bound in [ASZ18] is applicable only in the case where $n \gtrsim J/\epsilon^2$. Our result thus shows the possibility of accurate distribution estimation under the L_2 error criterion in the high-dimensional setting. Our bound also improves over the bound of [ASZ18] for all $n \lesssim \frac{J^2}{\epsilon^2 \log(J)}$. To the best of our knowledge, existing results do not imply L_2 error bound better than the trivial $O(1)$ error in the regime where $n \lesssim \frac{J}{\epsilon^2}$. It is worthy to point out that the L_2 error bound of [ASZ18] is optimal only when $n \gtrsim J^2/\epsilon^2$. Although this condition is not explicitly mentioned in [ASZ18], however, as stated in the same paper, their claim of optimality follows from the lower bound in [YB18]; specifically, [YB18, Theorem IV]. From this theorem, it is clear that the lower bound is only valid when $n \geq \text{const.} \frac{J^2}{\epsilon^2}$. Hence, our bound does not contradict with the results of these previous works. We also note that the idea of projecting the estimated distribution onto the probability simplex was proposed in [KBR16] (along with a different protocol than that of [ASZ18]). Although [KBR16] show empirically that the projection technique yield improvements in accuracy, no formal analysis or guarantees were provided for the resulting error in this case.

Note that the L_2 estimation error bounds in the previous works were derived for the expected L_2*-squared* error, and hence the expressions here are the square-root of the bounds appearing in these references. Moreover, we note that our bounds are obtained by first deriving bounds on the L_2*-squared* estimation error, which then imply our stated bounds on the L_2 error. Hence, squaring our bounds give valid bounds on the L_2-squared error.

Adaptive linear queries: We assume the following constraint on any sequence of adaptively chosen queries $\langle \mathbf{q}_1, \cdot \rangle, \ldots, \langle \mathbf{q}_d, \cdot \rangle$: for each $k \in [d]$, $\|\mathbf{q}_k\|_\infty \leq r$ for some $r > 0$. That is, each vector $\mathbf{q}$ defining a query has a bounded L_∞ norm. Unlike the offline setting, since the sequence of the queries is not fixed beforehand (i.e., the queries matrix $\mathbf{A}$ is not known a priori), the above L_∞ constraint is more natural than constraining a quantity related to the norm of the queries matrix as we did in the offline setting. For any $r > 0$, we let $\mathcal{Q}_\infty(r) = \{\langle \mathbf{q}, \cdot \rangle : \ \|\mathbf{q}\|_\infty \leq r\}$, i.e., $\mathcal{Q}_\infty(r)$ denote the family of all linear queries satisfying the above constraint. In this setting, we measure accuracy in terms of the true L_∞ error; that is, the maximum true error $\max\limits_{k \in [d]} |y_k - \langle \mathbf{q}_k, \ \mathbf{p} \rangle|$ in any of the estimates $\{y_k : \ k \in [d]\}$ for the d queries. (See Section 2.2.2 for a precise definition).

We give a construction of ϵ-LDP protocol that answers any sequence of d adaptively chosen queries from $\mathcal{Q}_\infty(r)$. Our protocol attains the optimal L_∞ estimation error. The optimality follows from the fact that our upper bound matches a lower bound on the same error in the *non-adaptive* setting given in [DJW13b, Proposition 4]. In our protocol, each user sends only a constant number of bits to the server, namely, $O(\log(r))$ bits/user. In our protocol, the set of users are partitioned into d disjoint subsets, and each subset is used to answer one query. Roughly speaking, this partitioning technique can be viewed as some version of sample splitting. In contrast, this technique is known to be suboptimal (w.r.t. the L_∞ estimation error) in the *centralized* model of differential privacy [BNS+16]. Moreover, given the offline lower bound in [DJW13b], our result shows that adaptivity does not pose any extra penalty in the *true* L_∞ estimation error for linear queries in the local model. In contrast, it is still not clear whether the same statement can be made in the *centralized* model of differential privacy. For instance, assuming $\epsilon = \Theta(1)$ and $n \gtrsim d^{3/2}$, then in the *centralized* model, the best known upper bound on the *true* L_∞ estimation error for this problem in the *adaptive* setting is $\approx d^{1/4}/\sqrt{n}$ [BNS+16, Corollary 6.1] (which combines [DMNS06] with the generalization guarantees of differential privacy). Whereas in the offline setting, the *true* L_∞ error is upper-bounded by $\approx \sqrt{\frac{\log(d)}{n}}$ (combining [DMNS06] with the standard generalization bound for the offline setting). There is also a gap to be tightened in the other regime of n and d as well. For example, this can be seen by comparing [BNS+16, Corollary 6.3] with the bound attained by the private multiplicative weights algorithm [HR10] in the offline setting.

2 Preliminaries and Definitions

2.1 (ϵ, δ)-Local Differential Privacy

In the local model, an algorithm $\mathcal{A}$ can access any entry in a private data set $D = (v_1, \ldots, v_n) \in \mathcal{J}^n$ only via a randomized algorithm (local randomizer) $\mathcal{R} : \mathcal{J} \rightarrow \mathcal{W}$that, given an index $i \in [n]$, runs on the input v_i and returns a randomized output $\mathcal{R}(v_i)$ to $\mathcal{A}$. Such algorithm $\mathcal{A}$ satisfies (ϵ, δ)-local differential privacy ((ϵ, δ)-LDP) if the local randomizer $\mathcal{R}$ satisfies (ϵ, δ)-LDP defined as follows.

Definition 2.1 ((ϵ, δ)-LDP)**.** A randomized algorithm $\mathcal{R} : \mathcal{J} \rightarrow \mathcal{W}$ is (ϵ, δ)-LDP if for any pair $v, v' \in \mathcal{J}$ and any measurable subset $\mathcal{O} \subseteq \mathcal{W}$, we have

$$\underset{\mathcal{R}}{\mathbb{P}}[\mathcal{R}(v) \in \mathcal{O}] \leq e^{\epsilon} \underset{\mathcal{R}}{\mathbb{P}}[\mathcal{R}(v) \in \mathcal{O}] + \delta,$$

where the probability is taken over the random coins of $\mathcal{R}$. The case of $\delta = 0$ is called pure ϵ-LDP.

2.2 Accuracy Definitions

2.2.1 Offline queries

For the non-adaptive (offline) setting, we measure accuracy in terms of the worst-case expected L_2-error in the responses to d queries. Let $\mathbf{p}$ be any (unknown) distribution over a data domain $\mathcal{J} = [J]$. To simplify presentation, we will overload notation and use $\mathbf{p} \in \mathsf{Simplex}(J)$ to also denote the probability mass function (p.m.f.) of the same distribution, where $\mathsf{Simplex}(J)$ refers to the probability simplex in $\mathbb{R}^J$ defined as $\mathsf{Simplex}(J) = \left\{(w_1, \ldots, w_J) \in \mathbb{R}^J : \; w_j \geq 0 \; \forall j \in [J], \; \sum_{j=1}^{J} w_j = 1\right\}$.

Let D denote the set of users' inputs $\{v_i : i \in [n]\}$ that are drawn i.i.d. from $\mathbf{p}$(this will be usually denoted as $D \sim \mathbf{p}^n$). For any $r > 0$, let $\mathcal{C}_2(r) = \left\{\mathbf{A} = [\mathbf{a}_1 \; \ldots \; \mathbf{a}_J] \in \mathbb{R}^{d \times J} : \; \|\mathbf{a}\|_2 \leq r\right\}$; that is, $\mathcal{C}_2(r)$ denote the family of all matrices in $\mathbb{R}^{d \times J}$ whose columns lie in $B_2^d(r)$ (the d-dim L_2 ball of radius r). Let $\mathbf{A} \in \mathcal{C}_2(r)$ be a queries matrix whose rows determine d offline linear queries. An (ϵ, δ)-LDP protocol Prot describes a set of procedures executed at each user and the server that eventually produce an estimate $\hat{\mathbf{y}} \in \mathbb{R}^d$ for the true answer vector $\mathbf{Ap} \in \mathbb{R}^d$ subject to (ϵ, δ)-LDP. Let $\mathsf{Prot}(\mathbf{A}, D)$ denote the final estimate vector $\hat{\mathbf{y}}$ generated by the protocol Prot for a data set D and queries matrix $\mathbf{A}$. The true expected L_2 error in the estimate $\mathsf{Prot}(\mathbf{A}, D)$ when $D \sim \mathbf{p}^n$is defined as

$$\mathrm{err}_{\mathsf{Prot}, L_2}(\mathbf{A}; \mathbf{p}^n) \triangleq \underset{\mathsf{Prot},\, D \sim \mathbf{p}^n}{\mathbb{E}}\left[\|\mathsf{Prot}(\mathbf{A}, D) - \mathbf{Ap}\|_2\right],$$

where the expectation is taken over the randomness in D and the random coins of the protocol.

True error: The worst-case expected L_2-error (with respect to *worst-case distribution* and *worst case queries matrix* in $\mathcal{C}_2(r)$) is defined as

$$\mathrm{err}_{\mathsf{Prot},L_2}(\mathcal{C}_2(r),n) \triangleq \sup_{\mathbf{A}\in\mathcal{C}_2(r)} \sup_{\mathbf{p}\in\mathsf{Simplex}(J)} \mathop{\mathbb{E}}_{\mathsf{Prot},\ D\sim\mathbf{p}^n}[\|\mathsf{Prot}(\mathbf{A};D)-\mathbf{A}\mathbf{p}\|_2] \tag{1}$$

Empirical error: Sometimes, we will consider the worst-case empirical L_2 error of an LDP protocol. Given any data set $D\in[J]^n$, let $\widehat{\mathbf{p}}(D)\in\mathsf{Simplex}(J)$ denote the histogram (i.e., the empirical distribution) of D. The worst-case empirical L_2 error of an LDP protocol Prot is defined as

$$\widehat{\mathrm{err}}_{\mathsf{Prot},L_2}(\mathcal{C}_2(r),n) \triangleq \sup_{\mathbf{A}\in\mathcal{C}_2(r)} \sup_{D\in[J]^n} \mathop{\mathbb{E}}_{\mathsf{Prot}}[\|\mathsf{Prot}(\mathbf{A};D)-\mathbf{A}\widehat{\mathbf{p}}(D)\|_2] \tag{2}$$

Note the expectation in this case is taken only over the random coins of Prot.

Optimal non-private estimators for offline linear queries The following is a simple observation that follows well-known facts in statistical estimation.

$$\sup_{\mathbf{A}\in\mathcal{C}_2(r)} \sup_{\mathbf{p}\in\mathsf{Simplex}(J)} \mathop{\mathbb{E}}_{D\sim\mathbf{p}^n}[\|\mathbf{A}\widehat{\mathbf{p}}(D)-\mathbf{A}\mathbf{p}\|_2] \leq \frac{r}{\sqrt{n}} \tag{3}$$

Note that $\mathbf{A}\widehat{\mathbf{p}}(D)$ is an unbiased estimator of $\mathbf{A}\mathbf{p}$. The above bound follows from a simple analysis of the variance of $\mathbf{A}\widehat{\mathbf{p}}(D)$.

Note: Given (3), if we have an LDP protocol Prot that has worst-case *empirical* L_2 error α, then such a protocol has worst-case true L_2 error $\mathrm{err}_{\mathsf{Prot},L_2}(\mathcal{C}_2(r),n) \leq \alpha + \frac{r}{\sqrt{n}}$.

2.2.2 Adaptive queries

For any $r>0$, we let $\mathcal{Q}_\infty(r)=\{\langle\mathbf{q},\cdot\rangle: \ \|\mathbf{q}\|_\infty\leq r\}$, i.e., $\mathcal{Q}_\infty(r)$ denote the family of all linear queries described by vectors in $\mathbb{R}^J$ of L_∞ norm bounded by r. In the adaptive setting, we consider the worst-case expected L_∞ error in the vector of estimates generated by LDP protocol for any sequence of d adaptively chosen queries $\mathbf{q}_1,\ldots,\mathbf{q}_d\in\mathcal{Q}_\infty$. Let $D\sim\mathbf{p}^n$ be a data set of users' inputs. Let Prot be LDP protocol for answering any such sequence. We define the worst-case L_∞ error as

$$\mathrm{err}_{\mathsf{Prot},L_\infty}(\mathcal{Q}_\infty(r),d,n) \triangleq \sup_{\mathbf{p}\in\mathsf{Simplex}(J)} \sup_{\substack{\text{adaptive strategy}\\ \text{choosing } \mathbf{q}_1,\ldots,\mathbf{q}_d}} \mathop{\mathbb{E}}_{\mathsf{Prot},\ D\sim\mathbf{p}^n}\left[\max_{k\in[d]} \mid \mathsf{Prot}^{(k)}(D)-\langle\mathbf{q}_k,\ \mathbf{p}\rangle \mid\right], \tag{4}$$

where $\mathsf{Prot}^{(k)}(D)$ denotes the estimate generated by the protocol in the k-th round of the protocol.

2.3 Geometry facts

For a convex body $K\subseteq\mathbb{R}^d$, the polar body K_o is defined as $\{\mathbf{y}:|\langle\mathbf{y},\mathbf{x}\rangle|\leq 1\ \forall\mathbf{x}\in K\}$. A convex body K is symmetric if $K=-K$. The Minkowski norm $\|\mathbf{x}\|_K$ induced by a symmetric convex body K is defined as $\|\mathbf{x}\|_K=\inf\{r\in\mathbb{R}:\mathbf{x}\in rK\}$. The Minkowski norm induced by the polar body K_o of K is the dual norm of $\|\mathbf{x}\|_K$, and has the form $\|\mathbf{y}\|_{K_o}=\sup_{\mathbf{x}\in K}|\langle\mathbf{x},\mathbf{y}\rangle|$. By Holder's inequality, we have $\langle\mathbf{x},\mathbf{y}\rangle\leq\|\mathbf{x}\|_K\|\mathbf{y}\|_{K_o}$.

Let $\mathbb{B}_1^J$ denote the unit L_1 ball in $\mathbb{R}^J$. A symmetric convex polytope $L \subset \mathbb{R}^d$ of J vertices that are represented as the columns of a matrix $\mathbf{A} \in \mathbb{R}^{d\times J}$ is defined as $L \triangleq \mathbf{A}\mathbb{B}_1^J = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^J$ with $\|\mathbf{x}\|_1 \leq 1\}$. The *dual* Minkowski norm induced by the convex symmetric polytope L is given by $\|\mathbf{x}\|_{L_o} = \max_{\mathbf{y}\in L}|\langle \mathbf{x}, \mathbf{y}\rangle| = \max_{j\in[J]}|\langle \mathbf{a}_j, \mathbf{x}\rangle|$, where the last equality is due to the fact that any linear function over a polytope attains its maximum at one of the vertices of the polytope.

The following is a useful lemma based on standard analysis that bounds the least squared estimation error over convex bodies. We restate here the version that appeared in [NTZ13].

Lemma 2.2 (Lemma 1 in [NTZ13])**.** *Let $L \subseteq \mathbb{R}^d$ be a symmetric convex body, and let $\mathbf{y} \in L$ and $\bar{\mathbf{y}} = \mathbf{y} + \mathbf{z}$ for some $\mathbf{z} \in \mathbb{R}^d$. Let $\hat{\mathbf{y}} = \arg\min_{\mathbf{w}\in L}\|\mathbf{w} - \bar{\mathbf{y}}\|_2^2$. Then, we must have*

$$\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 \leq 4\min\{\|\mathbf{z}\|_2^2,\ \|\mathbf{z}\|_{L_o}\}.$$

As a direct consequence of the above lemma and the preceding facts, we have the following corollary.

Corollary 2.3. *Let $L \subset \mathbb{R}^d$ be a symmetric convex polytope of J vertices $\{\mathbf{a}_j\}_{j=1}^J$, and let $\mathbf{y} \in L$ and $\bar{\mathbf{y}} = \mathbf{y} + \mathbf{z}$ for some $\mathbf{z} \in \mathbb{R}^d$. Let $\hat{\mathbf{y}} = \arg\min_{\mathbf{w}\in L}\|\mathbf{w} - \bar{\mathbf{y}}\|_2^2$. Then, we must have*

$$\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 \leq 4\max_{j\in[J]}|\langle \mathbf{z},\ \mathbf{a}_j\rangle|.$$

2.4 SubGaussian random variables

Definition 2.4 (σ^2-subGaussian random variable)**.** A zero mean random variable X is called σ-subgaussian if for all $\lambda \geq 0$, $\mathbb{P}[|X| \geq \lambda] \leq 2\,e^{-\frac{\lambda^2}{2\sigma^2}}$.

Another equivalent version of the definition is as follows: A zero-mean random variable X is σ-subgaussian if for all $t \in \mathbb{R}$, $\mathbb{E}\left[e^{tX}\right] \leq e^{\frac{1}{2}t^2\sigma^2}$. It is worth noting that these two versions of the definition are equivalent up to a small constant in σ (see, e.g., [Bul]).

3 LDP Protocols for Offline Linear Queries

In this section, we consider the problem of estimating d offline linear queries under ϵ-LDP. For any given $r > 0$, as discussed in Section 2.2.1, we consider a queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$; that is, the columns of $\mathbf{A}$ are assumed to lie in the L_2 ball $\mathbb{B}_2^d(r)$ of radius r.

As a warm-up, in Section 3.1, we first describe and analyze an (ϵ, δ)-LDP protocol. Our protocol is simple and is based on (i) perturbing the columns of $\mathbf{A}$ corresponding to users' inputs via Gaussian noise and (ii) applying a projection step, when appropriate, to the noisy aggregate similar to the technique of [NTZ13] in the centralized model. This projection step reduces the error significantly in the regime where $n \lesssim d^2/\log(J)$ (which subsumes the high-dimensional setting $d \gtrsim n$ when $\log(J) \lesssim d$). In particular, in such regime, our protocol yields an L_2 error $\approx r\left(\frac{\log(J)}{n}\right)^{1/4}$, which does not depend on d and depends only sub-logarithmically on J. Moreover, this error is within a factor of $\log^{1/4}(J)$ from the optimal error in this regime. Hence, this result establishes the possibility of accurate estimation of linear queries with respect to the L_2 error in high-dimensional settings. Adoption of all previously known algorithms (particularly, the ball sampling mechanism of [DJW13b]) do not provide any guarantees better than the trivial error for that problem in the regime where $n \lesssim d$.

In Section 3.2, we give a construction that transforms our (ϵ, δ) algorithm into a pure ϵ-LDP algorithm with essentially the same error guarantees. Our transformation is inspired by ideas from [BS15, BNS18]. In particular, [BNS18] gives a generic technique for transforming an (ϵ, δ)-LDP protocol to an $O(\epsilon)$-LDP protocol. Our construction can be viewed as a simpler, more direct version of this transformation for the case of linear queries.

3.1 (ϵ, δ) LDP Protocol for Offline Linear Queries

We first describe the local randomization procedure $\mathcal{R}_i^{\mathsf{Gauss}}$ carried out by each user $i \in [n]$. The local randomization is based on perturbation via Gaussian noise ; that is, it can be viewed as LDP version of the standard Gaussian mechanism [DKM+06].

Algorithm 1 $\mathcal{R}_i^{\mathsf{Gauss}}$: (ϵ, δ)-Local Randomization of user $i \in [n]$

```
Require: Queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$, User $i$ input $v_i \in [J]$, privacy parameters $\epsilon, \delta$.
  return $\tilde{\mathbf{y}}_i = \mathbf{a}_{v_i} + \mathbf{z}_i$ where $z_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_d)$ where $\mathbf{a}_{v_i}$ is the $v_i$-th column of $\mathbf{A}$, $\sigma^2 = 2r^2 \frac{\log(2/\delta)}{\epsilon^2}$,
  and $\mathbb{I}_d$ denotes the identity matrix of size $d$.
```

The desciption of our (ϵ, δ) protocol for linear queries is given in Algorithm 2.

Algorithm 2 $\mathsf{Prot}_{\mathsf{Gauss}}$: (ϵ, δ)-LDP protocol for answering offline linear queries from $\mathcal{C}_2(r)$

```
Require: Queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$, Users' inputs $\{v_i \in [J] : i \in [n]\}$, privacy parameters $\epsilon, \delta$.
for Users $i = 1$ to $n$ do
    User $i$ computes $\tilde{\mathbf{y}}_i = \mathcal{R}_i^{\mathsf{Gauss}}(v_i)$ and sends it to the server.
end for
Server computes $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{y}}_i$.
if $n < \frac{d^2 \log(2/\delta)}{8\epsilon^2 \log(J)}$: then
    $\hat{\mathbf{y}} = \arg\min_{\mathbf{w} \in \mathbf{A}\mathbb{B}_1^J} \|\mathbf{w} - \bar{\mathbf{y}}\|_2^2$ where $\mathbb{B}_1^J$ is the unit $L_1$ ball in $\mathbb{R}^J$.
else
    $\hat{\mathbf{y}} = \bar{\mathbf{y}}$
end if
return $\hat{\mathbf{y}}$.
```

We now state and prove the privacy and accuracy guarantee of our protocol. Note in the local model of differential privacy, the privacy of the entire protocol rests only on differential privacy of the local randomizers, which we prove now.

Theorem 3.1. *[Privacy Guarantee] Algorithm 1 is (ϵ, δ)-LDP.*

Proof. The proof follows directly from standard analysis of the Gaussian mechanism [DKM+06, NTZ13] applied in the context of (ϵ, δ)- LDP. □

Theorem 3.2 (Accuracy of Algorithm 2)**.** *Protocol* $\mathsf{Prot}_{\mathsf{Gauss}}$ *given by Algorithm 2 satisfies the following accuracy guarantee:*

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) \leq \; r \cdot \min\left(\left(\frac{32\log(J)\log(2/\delta)}{n\epsilon^2}\right)^{1/4}, \sqrt{\frac{2d\log(2/\delta)}{n\epsilon^2}}\right)$$

where $\mathrm{err}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n)$ *is as defined in (1).*

Proof. Fix any queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$. Let $\mathbf{y} = \mathbf{A}\widehat{\mathbf{p}}$ where $\widehat{\mathbf{p}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_{v_i}$ is the actual histogram of the users' data set (here, $\mathbf{e}_t \in \mathbb{R}^J$ denotes the vector with 1 in the t-th coordinate and zeros elsewhere). First, consider the case where $n \geq \frac{d^2\log(2/\delta)}{8\epsilon^2\log(J)}$. Note that $\hat{\mathbf{y}} = \bar{\mathbf{y}}$, and hence $\hat{\mathbf{y}} - \mathbf{y}$ is Gaussian random vector with zero mean and covariance matrix $\frac{\sigma^2}{n}\mathbb{I}_d$. Hence, in this case, it directly follows that $\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) = \sqrt{\frac{\sigma^2 d}{n}} = r\sqrt{\frac{2d\log(2/\delta)}{n\epsilon^2}}$, where $\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n)$ is the worst-case empirical error as defined in (2).

Next, consider the case where $n < \frac{d^2\log(2/\delta)}{8\epsilon^2\log(J)}$. Since $\hat{\mathbf{y}}$ is the projection of $\bar{\mathbf{y}}$ on the symmetric convex polytope $\mathbf{A}\mathbb{B}_1^J$, then by Corollary 2.3, it follows that

$$\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 \leq 4\max_{j\in[J]}|\langle \bar{\mathbf{y}} - \mathbf{y},\, \mathbf{a}_j\rangle|.$$

Hence, we have

$$\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) \leq 2\sqrt{\mathbb{E}\left[\max_{j\in[J]}|\langle \bar{\mathbf{y}} - \mathbf{y},\, \mathbf{a}_j\rangle|\right]}.$$

As before, note that $\bar{\mathbf{y}} - \mathbf{y} \sim \mathcal{N}\left(\mathbf{0},\, \frac{\sigma^2}{n}\mathbb{I}_d\right)$. Note also that $\|\mathbf{a}_j\| \leq r \;\; \forall j \in [J]$. Hence, for each $j \in [J]$, $\langle \bar{\mathbf{y}} - \mathbf{y},\, \mathbf{a}_j\rangle$ is Gaussian with zero mean and variance $\leq r^2\sigma^2/n$. By standard bounds on the maximum of Gaussian r.v.s (e.g., see [Rig15]), we have

$$\mathbb{E}\left[\max_{j\in[J]}|\langle \bar{\mathbf{y}} - \mathbf{y},\, \mathbf{a}_j\rangle|\right] \leq \sqrt{\frac{\sigma^2}{n}r^2\log(J)} \leq r^2\sqrt{2\frac{\log(J)\log(2/\delta)}{n\epsilon^2}}.$$

Hence, in this case, we have $\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) \leq \left(\frac{32\log(J)\log(2/\delta)}{n\epsilon^2}\right)^{1/4}$.

Putting the two cases above together, we get that $\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n)$ is upper-bounded by the expression in the theorem statement.

From (3) in Section 2.2.1 (and the succeeding note), we have

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) \leq \widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) + r/\sqrt{n}.$$

Note that the $r/\sqrt{n}$ term above is swamped by the bound on $\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n)$. This completes the proof.

□

3.2 $(\epsilon, 0)$ LDP Protocol for Offline Linear Queries

In this section, we give a pure LDP construction that achieves essentially the same accuracy (up to a constant factor of at most 2) as our approximate LDP algorithm above. Our construction is based on a direct transformation of the above approximate LDP protocol into a pure LDP one. Our construction is inspired by the idea of rejection sampling in [BS15, BNS18], and can be viewed as a simpler, more direct version of the generic technique in [BNS18] in the case of linear queries.

In our construction, we assume that $\epsilon \leq 1^2$. For any $\mathbf{a} \in \mathbb{R}^d$, let $f_{\mathbf{a}}$ denote the probability density function of the Gaussian distribution $\mathcal{N}(\mathbf{a}, \sigma^2\,\mathbb{I}_d)$ where $\sigma^2 = 4r^2\frac{\log(n)}{\epsilon^2}$. (Note that the setting of σ^2 is the same setting for the Gaussian noise used in Algorithm 2 with $\delta \approx 1/n^2$.)

In Algorithm 3, we describe the local randomization procedure $\mathcal{R}_i^{\mathsf{RejSamp}}$ executed independently by every user $i \in [n]$. Then, we describe our ϵ-LDP protocol for offline linear queries in Algorithm 4.

Algorithm 3 $\mathcal{R}_i^{\mathsf{RejSamp}}$: ϵ-Local Randomization of user $i \in [n]$ based on rejection sampling

```
Require: Queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$, User $i$ input $v_i \in [J]$, privacy parameter $\epsilon$.
1: Get $\mathbf{a}_{v_i}$: the $v_i$-th column of $\mathbf{A}$.
2: Sample a Gaussian vector $\tilde{\mathbf{y}}_i \sim \mathcal{N}\left(\mathbf{0},\ \sigma^2\,\mathbb{I}_d\right)$, where $\sigma^2 := 2\,r^2\,\frac{\log(2/\delta)}{\epsilon^2}$ and $\delta := \frac{2}{n^2}$.
3: Compute (scaled) ratio of the two Gaussian densities $f_{\mathbf{a}_{v_i}}$ and $f_{\mathbf{0}}$ at $\tilde{\mathbf{y}}_i$ : $\eta_i := \frac{1}{2}\,\frac{f_{\mathbf{a}_{v_i}}(\tilde{\mathbf{y}}_i)}{f_{\mathbf{0}}(\tilde{\mathbf{y}}_i)}$.
4: if $\eta_i \in [\frac{e^{-\epsilon/4}}{2},\ \frac{e^{\epsilon/4}}{2}]$ then
5:     Sample a bit $B_i \sim \mathsf{Ber}(\eta_i)$
6: else
7:     Let $B_i = 0$
8: end if
9: if $B_i = 1$ then
10:     return $\tilde{\mathbf{y}}_i$
11: else
12:     return $\perp$ {The output in this case indicates that user $i$ is dropped out of the protocol.}
13: end if
```

Algorithm 4 $\mathsf{Prot}_{\mathsf{RejSamp}}$: ϵ-LDP protocol for offline linear queries from $\mathcal{C}_2(r)$

```
Require: Queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$, Users' inputs $\{v_i \in [J] : i \in [n]\}$, privacy parameter $\epsilon$.
for All users $i \in [n]$ such that $\mathcal{R}_i^{\mathsf{RejSamp}}(v_i) \neq \perp$ do
    Let $\tilde{\mathbf{y}}_i = \mathcal{R}_i^{\mathsf{RejSamp}}(v_i)$ and send $\tilde{\mathbf{y}}_i$ to the server.
end for
Server receives the set of responses $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\hat{n}}$, where $\hat{n}$ is the number users whose response $\neq \perp$.
Server computes $\bar{\mathbf{y}} = \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} \tilde{\mathbf{y}}_i$.
if $\hat{n} < \frac{d^2\log(n)}{4\,\epsilon^2\,\log(J)}$: then
    $\hat{\mathbf{y}} = \arg\min_{\mathbf{w}\in\mathbf{A}\mathbb{B}_1^J} \|\mathbf{w} - \bar{\mathbf{y}}\|_2^2$ where $\mathbb{B}_1^J$ is the unit $L_1$ ball in $\mathbb{R}^J$.
else
    $\hat{\mathbf{y}} = \bar{\mathbf{y}}$
end if
return $\hat{\mathbf{y}}$.
```

We now state and prove the privacy and accuracy guarantees of our protocol.

Theorem 3.3. *[Privacy Guarantee] Algorithm 3 is ϵ-LDP.*

Proof. Consider any user $i \in [n]$. Let $v \in [J]$ be any input of user i. Define

$$\mathsf{Good}_i(v) \triangleq \left\{\mathbf{y} \in \mathbb{R}^d :\ \eta_i(v, \mathbf{y}) \in [\frac{e^{-\epsilon/4}}{2},\ \frac{e^{\epsilon/4}}{2}]\right\},$$

where $\eta_i(v,\ \mathbf{y}) = \frac{1}{2}\,\frac{f_{\mathbf{a}_v}(\mathbf{y})}{f_{\mathbf{0}}(\mathbf{y})}$. Note that by the standard analysis of the Gaussian mechanism, we have $\underset{\tilde{\mathbf{y}}_i \sim \mathcal{N}(\mathbf{0},\ \sigma^2)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \notin \mathsf{Good}_i(v)] \leq \delta$, where σ^2 and δ are set as in Step 2 of Algorithm 3). Now, we note that the output of Algorithm 3 is a function of only the bit B_i. Since differential privacy

[2] This is not a loss of generality in most practical scenarios where we aim at a reasonably strong privacy guarantee.

is resilient to post-processing, it suffices to show that for any $v, v' \in [J]$, any $b \in \{0,1\}$, we have $\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = b] \leq e^{\epsilon} \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i = b]$. First, observe that

$$\begin{aligned}\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = 0] &\leq \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = 0 |\, \tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v)] + \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \notin \mathsf{Good}_i(v)] \\ &\leq 1 - \frac{e^{-\epsilon/4}}{2} + \delta.\end{aligned}$$

We also have

$$\begin{aligned}\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i = 0] &\geq \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i = 0 |\, \tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v')] \cdot \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v')] \\ &\geq \left(1 - \frac{e^{\epsilon/4}}{2}\right)(1-\delta).\end{aligned}$$

Thus, $\frac{\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i=0]}{\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i=0]} \leq \frac{1-\frac{e^{-\epsilon/4}}{2}+\delta}{\left(1-\frac{e^{\epsilon/4}}{2}\right)(1-\delta)}$. Note that for any $t \in \mathbb{R}$, $1 + t \leq e^t$. Also, note that since $\epsilon \leq 1$, we have $1 + \epsilon/4 \leq e^{\epsilon/4} \leq 1 + \frac{5}{16}\epsilon$. Hence, this ratio can be upper bounded as

$$\frac{\frac{1}{2}(1+\epsilon/4)+\delta}{\frac{1}{2}(1-\frac{5}{16}\epsilon)(1-\delta)} = \frac{1+\epsilon/4}{1-\frac{5}{16}\epsilon} \cdot \frac{1+\frac{2\delta}{(1+4\epsilon)}}{1-\delta} \leq e^{\frac{7}{8}\epsilon} e^{4\delta} \leq e^{\epsilon}.$$

In the last step, we use the fact that $\delta = (1/n^2)$ and hence, $\delta \ll \epsilon/32$.

Now, we consider the event that $B_i = 1$. Note that $\forall v \in [J]$, $\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = 1 \mid \tilde{\mathbf{y}}_i \notin \mathsf{Good}_i(v)] = 0$. Hence, we have

$$\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = 1] \leq \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = 1 \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v)] \leq \frac{e^{\epsilon/4}}{2}.$$

We also have

$$\begin{aligned}\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i = 1] &= \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i = 1 \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v')] \cdot \underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v')] \\ &\geq \frac{e^{-\epsilon/4}}{2}(1-\delta).\end{aligned}$$

Hence,

$$\frac{\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}{\mathbb{P}}[B_i = 1]}{\underset{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}{\mathbb{P}}[B_i = 1]} \leq e^{\epsilon/2 + 2\delta} < e^{\epsilon}.$$

□

Theorem 3.4 (Accuracy of Algorithm 4)**.** *Suppose $n \geq 120$. Then, Protocol $\mathsf{Prot}_{\mathsf{RejSamp}}$ (Algorithm 4) satisfies the following accuracy guarantee:*

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{RejSamp}},\, L_2}(\mathcal{C}_2(r), n) \leq\, r \cdot \min\left(\left(\frac{280 \log(J) \log(n)}{n\epsilon^2}\right)^{1/4}, \sqrt{\frac{10\, d \log(n)}{n\epsilon^2}}\right)$$

where $\mathrm{err}_{\mathsf{Prot}_{\mathsf{RejSamp}},\, L_2}(\mathcal{C}_2(r), n)$ is as defined in (1).

The high-level idea of the proof can be described as follows. We first show that the number of users who end up sending a signal to the server (i.e., those users with $B_i = 1$) is at least a constant fraction of the total number of users ($\gtrsim n/4$). Hence, the effective reduction in the sample size will not have a pronounced effect on the true error (it can only increase the true expected L_2 error by at most a factor ≤ 2). Next, we show that *conditioned on* $B_i = 1$, the distribution of the user's signal $\tilde{\mathbf{y}}_i$ in Algorithm 4 is identical to the distribution of the user's signal in the (ϵ, δ) protocol of the previous section (Algorithm 2). That is, *conditioned on a high probability event*, the signals generated by the active users via the pure ϵ local randomizers $\mathcal{R}^{\mathsf{RejSamp}}$ (Algorithm 3) are statistically indistinguishable from the signals that could have been generated if those users have used the Gaussian local randomizers $\mathcal{R}_i^{\mathsf{Gauss}}$ (Algorithm 1). This allows us to show that the L_2 error resulting from Algorithm 4 is essentially the same as the one resulting from Algorithm 2.

We now give the formal proof. In the sequel, we call user $i \in [n]$ *active* if $B_i = 1$; that is, if $\mathcal{R}_i^{\mathsf{RejSamp}}(v_i) \neq \perp$ and hence, user i ends up sending a signal $\tilde{\mathbf{y}}_i$ to the server. As in the proof of Theorem 3.3, we define

$$\mathsf{Good}_i = \left\{\mathbf{y} \in \mathbb{R}^d : \ \eta_i(\mathbf{y}) \in \left[\frac{e^{-\epsilon/4}}{2}, \ \frac{e^{\epsilon/4}}{2}\right]\right\},$$

where $\eta_i(\mathbf{y}) = \frac{1}{2}\frac{f_{\mathbf{a}_{v_i}}(\mathbf{y})}{f_{\mathbf{0}}(\mathbf{y})}$.

We start by the following useful lemmas.

Lemma 3.5. *Suppose $n \geq 120$. With probability $\geq 1 - e^{-n/34}$, the number of active users $\hat{n}$ in Step 4 of Algorithm 4 satisfies $\hat{n} > n/4$.*

Proof. Given Algorithm 3, for any user $i \in [n]$, observe that

$$\mathbb{P}\left[\mathcal{R}^{\mathsf{RejSamp}}(v_i) = \perp\right] = \mathbb{P}[B_i = 0] \leq \mathop{\mathbb{P}}_{\tilde{\mathbf{y}}_i \sim \mathcal{N}(\mathbf{0},\, \sigma^2)}[\tilde{\mathbf{y}}_i \notin \mathsf{Good}_i] + \mathbb{P}[B_i = 0 \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i]$$
$$\leq \delta + (1 - \frac{e^{-\epsilon/4}}{2}) \leq \frac{2}{n^2} + 5/8.$$

where the last inequality follows from the fact that $\epsilon \leq 1$. Thus, we have $\mathbb{P}[B_i = 1] \geq 3/8 - 2/n^2$. Note that $\hat{n} = \sum_{i=1}^n B_i$. Since $n \geq 120$, then by Chernoff's bound, we have

$$\mathbb{P}[\hat{n} < n/4] < e^{-n/34}.$$

□

Lemma 3.6. *For any user $i \in [n]$, any input $v_i \in [J]$, and any measurable set $\mathcal{O} \subseteq \mathbb{R}^d$, we have*

$$\mathop{\mathbb{P}}_{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i)}[\tilde{\mathbf{y}}_i \in \mathcal{O} \mid B_i = 1] = \mathop{\mathbb{P}}_{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i)}[\tilde{\mathbf{y}}_i \in \mathcal{O} \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i]$$

Proof. Let $\tilde{\mathbf{y}}_i$ be the Gaussian r.v. generated in Step 2 of Algorithm 3. Note $(\tilde{\mathbf{y}}_i, B_i)$ has mixed probability distribution. For every realization $\mathbf{y}$ of $\tilde{\mathbf{y}}_i$ and every $b \in \{0,1\}$, the joint (mixed) density function of $(\tilde{\mathbf{y}}_i, B_i)$ can be expressed as $h_{\tilde{\mathbf{y}}_i|B_i}(\mathbf{y}|b)\mathbb{P}[B_i = b] = p_{B_i|\tilde{\mathbf{y}}_i}(b|\mathbf{y})f_{\mathbf{0}}(\mathbf{y})$, where $h_{\tilde{\mathbf{y}}_i|B_i}$ is the conditional density of $\tilde{\mathbf{y}}_i$ given B_i and $p_{B_i|\tilde{\mathbf{y}}_i}$ is the conditional density function of B_i given $\tilde{\mathbf{y}}_i$. Note that we have

$$p_{B_i|\tilde{\mathbf{y}}_i}(1 \mid \mathbf{y}) = \begin{cases} \frac{1}{2}\frac{f_{\mathbf{a}_{v_i}}(\mathbf{y})}{f_{\mathbf{0}}(\mathbf{y})} & \mathbf{y} \in \mathsf{Good}_i \\ \\ 0 & \mathbf{y} \notin \mathsf{Good}_i \end{cases} \tag{5}$$

Now, observe that for any measurable set $\mathcal{O} \in \mathbb{R}^d$,

$$\begin{aligned}
\underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathcal{O} \mid B_i = 1] &= \frac{\underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathcal{O},\ B_i = 1]}{\mathbb{P}[B_i = 1]} \\
&= \frac{\int_{\mathbf{y} \in \mathcal{O}} p_{B_i \mid \tilde{\mathbf{y}}_i}(1 \mid \mathbf{y}) f_{\mathbf{0}}(\mathbf{y})\, d\mathbf{y}}{\mathbb{P}[B_i = 1]} \\
&= \frac{\frac{1}{2} \int_{\mathbf{y} \in \mathcal{O} \cap \mathsf{Good}_i} f_{\mathbf{a}_{v_i}}(\mathbf{y})\, d\mathbf{y}}{\mathbb{P}[B_i = 1]} \qquad (6) \\
&= \frac{1}{2} \cdot \frac{\underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathcal{O},\ \tilde{\mathbf{y}}_i \in \mathsf{Good}_i]}{\mathbb{P}[B_i = 1]} \qquad (7)
\end{aligned}$$

where (6) follows from (5), and (7) follows from observing that the distribution of $\mathcal{R}^{\mathsf{Gauss}}(v_i)$ is $\mathcal{N}\left(\mathbf{a}_{v_i},\ \sigma^2 \mathbb{I}_d\right)$ (whose density is denoted as $f_{\mathbf{a}_{v_i}}$ as defined early in this section).

Next, we consider $\mathbb{P}[B_i = 1]$. Note that

$$\begin{aligned}
\mathbb{P}[B_i = 1] &= \underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i)}{\mathbb{P}}[B_i = 1, \tilde{\mathbf{y}}_i \in \mathsf{Good}_i] \\
&= \int_{\mathbf{y} \in \mathsf{Good}_i} p_{B_i \mid \tilde{\mathbf{y}}_i}(1 \mid \mathbf{y}) f_{\mathbf{0}}(\mathbf{y})\, d\mathbf{y} \\
&= \frac{1}{2} \cdot \underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathsf{Good}_i]
\end{aligned}$$

Plugging this in (7), then (7) reduces to $\underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathcal{O} \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i]$, which proves the lemma.

□

Proof of Theorem 3.4: Putting Lemmas 3.5 and 3.6 together with Theorem 3.2 leads us easily to the stated result. Fix any queries matrix $\mathbf{A} \in \mathcal{C}_2(r)$. First, note that $\{(\tilde{\mathbf{y}}_i, B_i) :\ i \in [n]\}$ are independent (independence across users). Hence, for any fixed subset $\tilde{\mathcal{S}} \subseteq [n]$, any sequence of users' inputs $\{v_i :\ i \in \tilde{\mathcal{S}}\}$, and any sequence of measurable sets $\{\mathcal{O}_i \subseteq \mathbb{R}^d :\ i \in \tilde{\mathcal{S}}\}$, we have

$$\begin{aligned}
\underset{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i):\ i \in \tilde{\mathcal{S}}\right\}}{\mathbb{P}}\left[\tilde{\mathbf{y}}_i \in \mathcal{O}_i\ \ \forall i \in \tilde{\mathcal{S}} \mid B_i = 1\ \ \forall i \in \tilde{\mathcal{S}}\right] &= \prod_{i \in \tilde{\mathcal{S}}} \underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathcal{O}_i \mid B_i = 1] \\
&= \prod_{i \in \tilde{\mathcal{S}}} \underset{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i)}{\mathbb{P}}[\tilde{\mathbf{y}}_i \in \mathcal{O}_i \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i] \\
&= \underset{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i):\ i \in \tilde{\mathcal{S}}\right\}}{\mathbb{P}}\left[\tilde{\mathbf{y}}_i \in \mathcal{O}_i\ \ \forall i \in \tilde{\mathcal{S}} \mid \tilde{\mathbf{y}}_i \in \mathsf{Good}_i\ \ \forall i \in \tilde{\mathcal{S}}\right] \\
&\leq \frac{\underset{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i):\ i \in \tilde{\mathcal{S}}\right\}}{\mathbb{P}}\left[\tilde{\mathbf{y}}_i \in \mathcal{O}_i\ \ \forall i \in \tilde{\mathcal{S}}\right]}{1 - |\tilde{\mathcal{S}}| \cdot \delta} \\
&\leq \left(\frac{n}{n-2}\right) \underset{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i):\ i \in \tilde{\mathcal{S}}\right\}}{\mathbb{P}}\left[\tilde{\mathbf{y}}_i \in \mathcal{O}_i\ \ \forall i \in \tilde{\mathcal{S}}\right] \\
&\leq 1.02 \cdot \underset{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i):\ i \in \tilde{\mathcal{S}}\right\}}{\mathbb{P}}\left[\tilde{\mathbf{y}}_i \in \mathcal{O}_i\ \ \forall i \in \tilde{\mathcal{S}}\right] \qquad (8)
\end{aligned}$$

where the second equality follows from Lemma 3.6, and the last two inequalities follow from the fact that $\delta = 2/n^2$, $|\tilde{\mathcal{S}}| \le n$, and the fact that $n \ge 120$.

Let $\mathcal{S} \subseteq [n]$ denote the subset of active users; that is $\mathcal{S} = \{i \in [n] : B_i = 1\}$. Fix any realization $\tilde{\mathcal{S}}$ of $\mathcal{S}$. Let $\tilde{D} = \left\{v_i :\ i \in \tilde{\mathcal{S}}\right\}$; that is, $\tilde{D}$ is the subset of the data set D of users' inputs congruent with $\tilde{\mathcal{S}}$. Now, *conditioned on* the event in Lemma 3.5 (i.e., conditioned on $\hat{n} \ge n/4$) and *conditioned on* any fixed realization $\tilde{\mathcal{S}}$ of $\mathcal{S}$, then by (8), we have

$$\mathop{\mathbb{E}}_{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{RejSamp}}(v_i):\ i\in\tilde{\mathcal{S}}\right\},\ \tilde{D}} \left[\|\hat{\mathbf{y}} - \mathbf{A}\mathbf{p}\|_2\right] \le 1.02 \cdot \mathop{\mathbb{E}}_{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i):\ i\in\tilde{\mathcal{S}}\right\},\ \tilde{D}} \left[\|\hat{\mathbf{y}} - \mathbf{A}\mathbf{p}\|_2\right] \tag{9}$$

Note that by Theorem 3.2, the expectation on the right-hand side is bounded as

$$\mathop{\mathbb{E}}_{\left\{\tilde{\mathbf{y}}_i \leftarrow \mathcal{R}_i^{\mathsf{Gauss}}(v_i):\ i\in\tilde{\mathcal{S}}\right\},\ \tilde{D}} \left[\|\hat{\mathbf{y}} - \mathbf{A}\mathbf{p}\|_2\right] \le r \cdot \min\left(\left(\frac{32\log(J)\log(n^2)}{(n/4)\,\epsilon^2}\right)^{1/4},\ \sqrt{\frac{2\,d\log(n^2)}{(n/4)\,\epsilon^2}}\right) \tag{10}$$

By Lemma 3.5, the event $\hat{n} \ge n/4$ occurs with probability at least $1 - e^{-n/34}$. Thus, if we remove conditioning on such event from the expectation on the left-hand side of (9), the unconditional version of such expectation can only increase by an additive term of at most $r\,e^{-n/34}$ (since the L_2 error cannot exceed r, and the probability that $\hat{n} < n/4$ is at most $e^{-n/34}$). This term is dominated by (10).

Putting these together, we finally arrive at

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{RejSamp}},\ L_2}(\mathcal{C}_2(r), n) \le\ r \cdot \min\left(\left(\frac{280\log(J)\log(n)}{n\epsilon^2}\right)^{1/4},\ \sqrt{\frac{10\,d\log(n)}{n\epsilon^2}}\right).$$

3.3 On Tightness of the Bound

The above bound (Theorem 3.4) is tight up to a factor $O\left((\log(J)\log(n))^{1/4}\right)$.. In particular, one can show a lower bound of $\Omega\left(\min\left(\frac{1}{n^{1/4}\sqrt{\epsilon}},\ \sqrt{\frac{d}{n\epsilon^2}}\right)\right)$ on the L_2 error. We note that it would suffice to show a tight lower bound on the minimax L_2 error in estimating the mean of a d-dimensional random variable with a finite support in $\mathbb{B}_2^d(r)$. Such lower bound follows from the lower bound in [DJW13b, Proposition 3]. We note that the packing constructed in the proof of [DJW13b, Proposition 3] is for a d-dimensional random variable with finite support. Hence, this lower bound is applicable to our case where the data universe is of finite size J. Tightening the remaining gap between the upper and lower bounds is left as an open problem. We conjecture that the $\log^{1/4}(J)$ factor in the upper bound is necessary.

4 $(\epsilon, 0)$-LDP Distribution Estimation

In this section, we revisit the problem of LDP distribution estimation under L_2 error criterion. First, we note that this problem is a special case of the linear queries problem, where the queries matrix $\mathbf{A} = \mathbb{I}_J$, i.e., the identity matrix of size J. Therefore, our results in Section 3 immediately give an upper bound on the L_2 error in this case. Namely, our results imply the existence of pure ϵ LDP protocol for distribution estimation whose L_2 error is bounded by $\min\left(\left(\frac{280\log(J)\log(n)}{n\epsilon^2}\right)^{1/4},\ \sqrt{\frac{10\,J\log(n)}{n\epsilon^2}}\right)$. However, in our protocol $\mathsf{Prot}_{\mathsf{RejSamp}}$, both communication complexity and running time per user would be $\Omega(J)$ in this case, which is prohibitive when J is large since the users are usually computationally limited (compared to the server). Our goal in this section is to have a construction with similar error guarantees but with better communication and running time at each user.

In the low-dimensional setting ($n \gtrsim J/\epsilon^2$), [ASZ18] give a nice construction (the Hadamard-Response protocol) whose L_2 error is $O(\sqrt{\frac{J}{\epsilon^2 n}})$. In this protocol, both communication and running time at each user are $O(\log(J))$. Also, the running time at the server is $\tilde{O}(n+J)$ (which is significantly better than the naive $O(nJ)$ running time). We show that this protocol can be extended to the *high-dimensional* setting. In particular, we show that, when $n \lesssim \frac{J^2}{\epsilon^2 \log(J)}$, projecting the output of the Hadamard-Response protocol onto the probability simplex yields L_2 error $\lesssim \left(\frac{\log(J)}{\epsilon^2 n}\right)^{1/4}$, which is tight up to a factor of $(\log(J))^{1/4}$ given the lower bounds in [DJW13b, YB18]. This improves the bound of [ASZ18] for all $n \lesssim \frac{J^2}{\epsilon^2 \log(J)}$. Moreover, to the best of our knowledge, existing results do not imply L_2 error bound better than the trivial $O(1)$ error in the regime where $n \lesssim \frac{J}{\epsilon^2}$.

The idea of projecting the estimated distribution onto the probability simplex was also proposed in [KBR16] and was empirically shown to yield improvements in accuracy, however, no formal analysis was provided for the error resulting from this technique.

We want to point out that the L_2 error of [ASZ18] is optimal only when $n \geq \Omega(J^2/\epsilon^2)$. Although this condition is not explicitly mentioned in [ASZ18], however, as stated in the same paper, their claim of optimality follows from the lower bound in [YB18]; specifically, [YB18, Theorem IV]. From this theorem, it is clear that the lower bound is only valid when $n \geq \text{const.} \frac{J^2}{\epsilon^2}$. Hence, our bound on the L_2 error does not contradict the results of these previous works.

Outline of Hadamard-Response Protocol of [ASZ18]: We will refer to this protocol as $\mathsf{Prot}_{\mathsf{HR}}$. We will use such a protocol as a black-box, so, we will not give a detailed description for it. The details can be found in [ASZ18, Section 4]. Let $\tilde{J} = 2^{\lceil \log_2(J+1) \rceil}$. Note that $J+1 \leq \tilde{J} \leq 2J+1$ Let $H_{\tilde{J}}$ denote the Hadamard matrix of size $\tilde{J}$. As before, the data set $D = \{v_i \in [J] : i \in [n]\}$ of users' inputs is assumed to be drawn i.i.d. from unknown distribution $\mathbf{p} = (p(1), \ldots, p(J)) \in \mathsf{Simplex}(J)$.
User procedure: In $\mathsf{Prot}_{\mathsf{HR}}$, each user $i \in [n]$ encode his input v_i as follows: first, select the (v_i+1)-th row of $H_{\tilde{J}}$, then, encode v_i as the subset $C_{v_i} \subset [\tilde{J}]$ of indices of that row that are incident with $+1$. Given C_{v_i}, user i invokes a generalized version of the basic randomized response technique to output a randomized index $z_i \in [\tilde{J}]$ as its ϵ-LDP report (See [ASZ18, Section 3]). Hence, the communication requirement per user is $\leq \log_2(2J+1)$ bits. Moreover, by the properties of the Hadamard matrix and the generalized randomized response, all the operations at any user can be executed in time $O(\log(J))$.
Server procedure: For every element $v \in [J]$ in the domain, the server generates an estimate $\bar{p}(v)$ for the true probability mass $p(v)$ as follows:

$$\bar{p}(v) = \left(\frac{e^\epsilon + 1}{e^\epsilon - 1}\right) \cdot \sum_{w \in [\tilde{J}]} H_{\tilde{J}}(v+1, w) \cdot q(w), \tag{11}$$

where $H_{\tilde{J}}(v+1, w)$ denotes the entry of $H_{\tilde{J}}$ at the $(v+1)$-th row and the w-th column, and $q(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(v_i = w)$ is the fraction of users whose reports are equal to w (See [ASZ18, Section 4]). Finally, the server outputs the vector of estimates: $\bar{\mathbf{p}} = (\bar{p}(1), \ldots, \bar{p}(J))$. As shown in [ASZ18], the total operations can be done in $\tilde{O}(n+J)$.

As noted in the same reference, equation (11) reduces to

$$\bar{p}(v) = 2\left(\frac{e^{\epsilon}+1}{e^{\epsilon}-1}\right)\cdot\left(\widehat{q(C_v)} - \frac{1}{2}\right), \tag{12}$$

where $\widehat{q(C_v)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(z_i \in C_v)$, which, by the properties of the local randomization, is the average of n Bernoulli r.v.s. Moreover, as shown in [ASZ18, Section 3], $\underset{\mathsf{Prot}_{\mathsf{HR}}}{\mathbb{E}}[\bar{p}(v)] = p(v) \;\; \forall v \in [J]$. Putting these observations together with Chernoff's inequality leads to the following fact.

Fact 4.1. *Let $\sigma^2 = 4\left(\frac{e^{\epsilon}+1}{e^{\epsilon}-1}\right)^2$. Each of the J components of $\bar{\mathbf{p}} - \mathbf{p}$ is $\frac{\sigma^2}{n}$-subGaussian random variable.*

We now give a construction $\mathsf{Prot}_{\mathsf{PHR}}$ (Projected Hadamard-Response) where the output estimate of $\mathsf{Prot}_{\mathsf{HR}}$ is projected onto the probability simplex whenever $n \lesssim \frac{J^2}{\epsilon^2 \log(J)}$. Clearly, $\mathsf{Prot}_{\mathsf{PHR}}$ has the same computational advantages of $\mathsf{Prot}_{\mathsf{HR}}$.

Algorithm 5 $\mathsf{Prot}_{\mathsf{PHR}}$: ϵ-LDP protocol for distribution estimation

```
Require: Data set of users' inputs $D = \{v_i \in [J] : i \in [n]\}$, privacy parameter $\epsilon$.
$\bar{\mathbf{p}} \leftarrow \mathsf{Prot}_{\mathsf{HR}}(D, \epsilon)$
$\hat{\mathbf{p}} = \arg\min_{\mathbf{w} \in \mathsf{Simplex}(J)} \|\mathbf{w} - \bar{\mathbf{p}}\|_2^2$.
return $\hat{\mathbf{p}}$.
```

First, note that since differential privacy is resilient to post-processing, ϵ-LDP of $\mathsf{Prot}_{\mathsf{PHR}}$ immediately follows from ϵ-LDP of $\mathsf{Prot}_{\mathsf{HR}}$ (shown in [ASZ18]).

Theorem 4.2 (Accuracy of Algorithm 5). *Let $c_\epsilon \triangleq \frac{e^{\epsilon}+1}{e^{\epsilon}-1}$. (Note that $c_\epsilon = O(1/\epsilon)$ when $\epsilon = O(1)$). Protocol $\mathsf{Prot}_{\mathsf{PHR}}$ satisfies the following accuracy guarantee:*

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{PHR}},\, L_2}(n) \triangleq \sup_{\mathbf{p} \in \mathsf{Simplex}(J)} \underset{\mathsf{Prot}_{\mathsf{PHR}},\, D \sim \mathbf{p}^n}{\mathbb{E}}\left[\|\mathsf{Prot}_{\mathsf{PHR}}(D) - \mathbf{p}\|_2\right] \leq \min\left(\left(\frac{256\, c_\epsilon^2 \log(J)}{n}\right)^{1/4}, \sqrt{\frac{4\, c_\epsilon^2 J}{n}}\right).$$

Proof. Fix any $\mathbf{p} \in \mathsf{Simplex}(J)$ as the true distribution. First, consider the case where $n \geq \left(\frac{e^{\epsilon}+1}{e^{\epsilon}-1}\right)^2 \frac{J^2}{16 \log(J)}$. Note that in this case the bound follows from [ASZ18] (since the projection step cannot increase the L_2 error).

Next, we consider the case where $n < \left(\frac{e^{\epsilon}+1}{e^{\epsilon}-1}\right)^2 \frac{J^2}{16 \log(J)}$. Note that the symmetric version of the polytope $\mathsf{Simplex}(J)$ is the L_1 Ball $\mathbb{B}_1^J$. Let $\mathbf{p}^* = \arg\min_{\mathbf{w} \in \mathbb{B}_1^J} \|\mathbf{w} - \bar{\mathbf{p}}\|_2^2$. Corollary 2.3 tells us that

$$\|\mathbf{p}^* - \mathbf{p}\|_2^2 \leq 4 \max_{j \in [J]} |\langle \bar{\mathbf{p}} - \mathbf{p},\, \mathbf{e}_j \rangle|,$$

where $\mathbf{e}_j \in \mathbb{R}^J$ denotes the vector with 1 in the j-th coordinate and zeros elsewhere. Now, as defined in $\mathsf{Prot}_{\mathsf{PHR}}$, let $\hat{\mathbf{p}} = \arg\min_{\mathbf{w} \in \mathsf{Simplex}(J)} \|\mathbf{w} - \bar{\mathbf{p}}\|_2^2$. Since $\mathbf{p} \in \mathsf{Simplex}(J)$, then for the special case where the symmetric polytope is $\mathbb{B}_1^J$, we always have $\|\hat{\mathbf{p}} - \mathbf{p}\|_2 \leq \|\mathbf{p}^* - \mathbf{p}\|_2$. This is because $\mathbf{p}^* - \mathbf{p}$ in this case can be written as the sum of two orthogonal components : $(\mathbf{p}^* - \hat{\mathbf{p}}) + (\hat{\mathbf{p}} - \mathbf{p})$. Hence, Corollary 2.3 implies that

$$\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 \leq 4 \max_{j \in [J]} |\langle \bar{\mathbf{p}} - \mathbf{p},\, \mathbf{e}_j \rangle|.$$

By Fact 4.1, for every $j \in [J]$, $\langle \bar{\mathbf{p}} - \mathbf{p},\ \mathbf{e}_j \rangle$ is $\frac{\sigma^2}{n}$-subGaussian where $\sigma^2 = 4\left(\frac{e^\epsilon+1}{e^\epsilon-1}\right)^2$. Now, by using the standard bounds on the maximum of subGaussian r.v.s (see [Rig15, Theorem 1.16]), we have

$$\mathop{\mathbb{E}}_{\mathsf{Prot}_{\mathsf{PHR}},\ D\sim\mathbf{p}^n}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2\right] \le 4\frac{\sigma}{\sqrt{n}}\sqrt{2\log(2J)} \le \sqrt{\frac{256\, c_\epsilon^2 \log(J)}{n}}.$$

Thus, we have $\mathrm{err}_{\mathsf{Prot}_{\mathsf{PHR}},\ L_2}(n) \le \left(\frac{256\, c_\epsilon^2 \log(J)}{n}\right)^{1/4}$.

□

5 ϵ-LDP Protocol for Adaptive Linear Queries

In this section, we consider the problem of estimating any sequence of d adaptively chosen linear queries $\mathbf{q}_1, \ldots, \mathbf{q}_d$ from $\mathcal{Q}_\infty(r)$ over some unknown distribution $\mathbf{p} \in \mathsf{Simplex}(J)$. As defined in Section 2.2.2, for any fixed $r > 0$, each query from $\mathcal{Q}_\infty(r)$ is a linear function $\langle \mathbf{q},\ \cdot \rangle$, which is uniquely identified by a vector $\mathbf{q} \in \mathbb{R}^J$ where $\|\mathbf{q}\|_\infty \le r$.

We measure accuracy in terms of the L_∞ estimation error in the d queries as defined in (4) in Section 2.2.2.

We give a construction of ϵ-LDP protocol that yields the optimal L_∞ error. The optimality follows from the fact that our upper bound matches a lower bound on the same error in the *weaker non-adaptive* setting, which follows from the lower bound in [DJW13b, Proposition 4]. Moreover, in our protocol each user sends only $O(\log(r))$ bits to the server[3]. In our protocol, the set of users are *randomly* partitioned into d disjoint subsets before the protocol starts, and each subset is used to answer one query. Assignment of the subsets to the queries is *fixed before the protocol starts*. Roughly speaking, this partitioning technique can be viewed as sample splitting. This avoids the trap of overfitting a query to the data samples it is evaluated on. In the centralized model, sample-splitting is generally sub-optimal. Our result shows that for adaptive linear queries in the local model, this technique is optimal.

The description of the protocol is given in Algorithm 6.

Theorem 5.1 (Privacy Guarantee)**.** *Protocol* $\mathsf{Prot}_{\mathsf{AdSamp}}$ *given by Algorithm 6 is* $(\epsilon, 0)$*-LDP.*

Proof. Fix any user i and any choice of $j_i = k \in [d]$. Observe that user i responds only to a single query: the k-th query. We show that the procedure described by Step 4 is ϵ-differentially private with respect to any such user's input item. First, note that $c_\epsilon \ge 1$ for all $\epsilon > 0$ and $|\mathbf{q}_k| \le r$. Hence, it is easy to verify that $\frac{1}{2}\left(1 + \frac{\mathbf{q}_k(v_i)}{c_\epsilon r}\right)$ and $\frac{1}{2}\left(1 - \frac{\mathbf{q}_k(v_i)}{c_\epsilon r}\right)$ in Step 4 are always in $(0,1)$ and they sum to 1, so they are legitimate probabilities. Observe that the ratio of the probabilities of the responses of user i when its data item is v_i and v_i' is given by

$$\frac{\mathbb{P}\left[\tilde{y}_{k,i} = c_\epsilon\, r \middle|\ v_i\right]}{\mathbb{P}\left[\tilde{y}_{k,i} = c_\epsilon\, r \middle|\ v_i'\right]} = \frac{c_\epsilon + \frac{\mathbf{q}_k(v_i)}{r}}{c_\epsilon + \frac{\mathbf{q}_k(v_i')}{r}} \le \frac{c_\epsilon + 1}{c_\epsilon - 1} = e^\epsilon$$

where the second inequality follows from the fact that $|\mathbf{q}_k(v)| \le r$ for all $k \in [d]$ and all $v \in [J]$. Similarly,

$$\frac{\mathbb{P}\left[\tilde{y}_{k,i} = -c_\epsilon\, r \middle|\ v_i\right]}{\mathbb{P}\left[\tilde{y}_{k,i} = -c_\epsilon\, r \middle|\ v_i'\right]} = \frac{c_\epsilon - \frac{\mathbf{q}_k(v_i)}{r}}{c_\epsilon - \frac{\mathbf{q}_k(v_i')}{r}} \le \frac{c_\epsilon + 1}{c_\epsilon - 1} = e^\epsilon$$

[3] Assuming fixed-precision representation of real numbers in $[-1,\ 1]$

Algorithm 6 $\mathsf{Prot}_{\mathsf{AdSamp}}$: ϵ-LDP protocol for adaptive linear queries from $\mathcal{Q}_\infty(r)$

```
Require: Data set of users' inputs $D = \{v_i \in [J] : i \in [n]\}$, privacy parameter $\epsilon$, sequence of $d$ adaptive linear queries $\mathbf{q}_1, \ldots, \mathbf{q}_d \in \mathcal{Q}_\infty(r)$ .
Each user $i \in [n]$ gets independently assigned (by itslef or via the server) a random uniform index $j_i \leftarrow [d]$. {This creates a random partition of the users, which is independent of the users' data, and is fixed before the protocol starts.}
for $k = 1, \ldots, d$ do
    for all users $i$ such that $j_i = k$ do
        User $i$ receives query $\mathbf{q}_k$ responds with $\tilde{y}_{k,i}$ generated as follows:
            $\tilde{y}_{k,i} = \begin{cases} c_\epsilon r & \text{w.p. } \frac{1}{2}\left(1 + \frac{\mathbf{q}_k(v_i)}{c_\epsilon r}\right) \\ -c_\epsilon r & \text{w.p. } \frac{1}{2}\left(1 - \frac{\mathbf{q}_k(v_i)}{c_\epsilon r}\right) \end{cases}$
        where $c_\epsilon = \frac{e^\epsilon + 1}{e^\epsilon - 1}$.
    end for
    Server computes an estimate $\bar{y}_k = \frac{1}{\hat{n}_k} \sum_{i: j_i = k} \tilde{y}_{k,i}$ based on the reports of the active users in round $k$ : $\{\tilde{y}_{k,i} : j_i = k\}$, where $\hat{n}_k = |\{i \in [n] : j_i = k\}|$ is the number of active users in round $k$.
    Server chooses a new query $\mathbf{q}_{k+1} \in \mathcal{Q}_\infty(r)$ (possibly based on all observations it received from the users until round $k$).
end for
return Estimated vector $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_d)$.
```

□

Theorem 5.2 (Accuracy). *Suppose $n \geq 8\, d \log(n)$. Then, Protocol $\mathsf{Prot}_{\mathsf{AdSamp}}$ given by Algorithm 6 satisfies the following accuracy guarantee for any sequence $\mathbf{q}_1, \ldots, \mathbf{q}_d \in \mathcal{Q}_\infty(r)$ of adaptive linear queries*

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{AdSamp}},\, L_\infty}(\mathcal{Q}_\infty(r), d, n) \leq 4\, r\, \sqrt{\frac{c_\epsilon^2 d \log(d)}{n}},$$

where $\mathrm{err}_{\mathsf{Prot}_{\mathsf{AdSamp}},\, L_\infty}(\mathcal{Q}_\infty(r), d, n)$ is as defined in (4). Moreover, this bound is optimal.

Our upper bound is optimal since it matches a lower bound on the L_∞ error in the *weaker non-adaptive* version of the same problem, which follows from [DJW13b, Proposition 4] .

This result shows that adaptivity does not pose any extra penalty in the *true* L_∞ estimation error for linear queries in the local model. In contrast, it is still not clear whether the same statement can be made about linear queries in the *centralized* model of differential privacy. For instance, assuming $\epsilon = \Theta(1)$ and $n \gtrsim d^{3/2}$, then in the *centralized* model, the best known upper bound on the *true* L_∞ estimation error in the *adaptive* setting is $\approx d^{1/4}/\sqrt{n}$ [BNS+16, Corollary 6.1] (which combines [DMNS06] with the generalization guarantees of differential privacy). Whereas in the offline setting, the *true* L_∞ error is upper-bounded by $\approx \sqrt{\frac{\log(d)}{n}}$ (combining [DMNS06] with the standard generalization bound for the offline setting). There is also a gap to be tightened in the other regime of n and d as well. This, for example, can be seen by comparing the bound for the adaptive setting [BNS+16, Corollary 6.3] with the bound attained by [HR10] in the offline setting.

Proof. Let $\mathbf{p}$ denote the true distribution over the data domain $[J]$. Let $D = \{v_i : i \in [n]\} \sim \mathbf{p}^n$. For every $i \in [n]$, $k \in [d]$, define $B_{k,i} = \mathbf{1}(j_i = k)$, where $j_i \leftarrow [d]$ is the uniform index generated for user i

in Step 1 of $\mathsf{Prot}_{\mathsf{AdSamp}}$. Note that for any fixed $k \in [d]$, $\{B_{k,i} :\ i \in [n]\}$ are i.i.d., and $\mathbb{P}[B_{k,i} = 1] = 1/d$ for every $i \in [n]$. For every $k \in [d]$, define $\mathcal{I}_k = \{i \in [n] :\ B_{k,i} = 1\}$; that is, $\mathcal{I}_k$ is the set of active users in round k. Hence, $\hat{n}_k = |\mathcal{I}_k|$, where $\hat{n}_k$ is the number of active users (as in Step 6). Let $D_k \subseteq D$ be defined as $D_k = \{v_i :\ i \in \mathcal{I}_k\}$; that is, D_k is the subset of data set D that contains the inputs of the active users in round k. For every round $k \in [d]$, as in Step 6, $\bar{y}_k$ is given by

$$\bar{y}_k = \frac{1}{\hat{n}_k} \sum_{i \in \mathcal{I}_k} \tilde{y}_{k,i}. \tag{13}$$

Suppose we *condition on* any fixed realization of the partition of users $\{\mathcal{I}_k :\ k \in [d]\}$. Conditioned on any such partition, since the users' inputs in D are i.i.d., then $D_1, \ldots, D_k$ are mutually independent. Hence, conditioned on any such partition, for every round $k \in [d]$, *the choice of the query* $\mathbf{q}_k$ *is independent of the subsample* D_k involved in the computation in round k, and thus, for every round $k \in [d]$, we have

$$\mathbb{E}\Big[\tilde{y}_{k,i}\Big|\ (\mathcal{I}_1, \ldots, \mathcal{I}_d)\Big] = \mathbb{E}\Big[\tilde{y}_{k,i}\ \Big|\ \mathcal{I}_k\Big] = \langle \mathbf{q}_k, \mathbf{p} \rangle \quad \forall i \in \mathcal{I}_k$$

where the expectation is taken w.r.t. $v_i \sim \mathbf{p}$, the randomization Step 4 in $\mathsf{Prot}_{\mathsf{AdSamp}}$, and any possible randomness in the choice of $\mathbf{q}_k$. Hence, *conditioned on any fixed partition of the users*, from (13) we get

$$\underset{\mathsf{Prot}_{\mathsf{AdSamp}},\ D \sim \mathbf{p}^n}{\mathbb{E}}\Big[\bar{y}_k\ \Big|\ (\mathcal{I}_1, \ldots, \mathcal{I}_d)\Big] = \frac{1}{\hat{n}_k} \sum_{i \in \mathcal{I}_k} \underset{\mathsf{Prot}_{\mathsf{AdSamp}},\ D_k \sim \mathbf{p}^{\hat{n}_k}}{\mathbb{E}}\Big[\tilde{y}_{k,i}\Big|\ \mathcal{I}_k\Big] = \langle \mathbf{q}_k, \mathbf{p} \rangle \quad \forall k \in [d] \tag{14}$$

Now, we define the set Good that contains "good" realizations for the partition $(\mathcal{I}_1, \ldots, \mathcal{I}_d)$:

$$\mathsf{Good} = \left\{(\mathcal{I}_1, \ldots, \mathcal{I}_d) :\ |\mathcal{I}_k| \geq \frac{n}{2\,d}\ \ \forall\ k \in [d]\right\}$$

Since for every $k \in [d]$, $\mathcal{I}_k$ is a $\mathsf{Bin}(n, 1/d)$ r.v., then by the multiplicative Chernoff's bound and the union bound, we have

$$\mathbb{P}[(\mathcal{I}_1, \ldots, \mathcal{I}_d) \notin \mathsf{Good}] \leq d\, e^{-\frac{n}{8d}} \tag{15}$$

Now, conditioned on any fixed realization of a partition $(\mathcal{I}_1, \ldots, \mathcal{I}_d) \in \mathsf{Good}$, it is easy to see that for every $k \in [d]$, $\bar{y}_k$ is the average of $\hat{n}_k \geq \frac{n}{2d}$ independent r.v.s, each taking a value in $\{-c_\epsilon\, r,\ c_\epsilon\, r\}$ w.p. 1. From this observation and using (14), it follows that for every $k \in [d]$, $\bar{y}_k - \langle \mathbf{q}_k, \mathbf{p} \rangle$ is σ^2-subGaussian, where $\sigma^2 = \frac{4\,d\,c_\epsilon^2\,r^2}{n}$. Hence, by a standard fact concerning the expectation of the maximum of subGaussians (see, e.g., [Rig15]), we have

$$\mathbb{E}\left[\max_{k \in [d]} |\ \bar{y}_k - \langle \mathbf{q}_k, \mathbf{p} \rangle\ |\ \Big|\ (\mathcal{I}_1, \ldots, \mathcal{I}_d) \in \mathsf{Good}\right] \leq 2\, c_\epsilon\, r \sqrt{\frac{2\,d\, \log(2d)}{n}} \tag{16}$$

Putting (16) and (15) together, and noting that the error is always bounded by r, we get

$$\mathbb{E}\left[\max_{k \in [d]} |\ \bar{y}_k - \langle \mathbf{q}_k, \mathbf{p} \rangle\ |\right] \leq 2\, c_\epsilon\, r \sqrt{\frac{2\,d\, \log(2d)}{n}} + r\, d\, e^{-\frac{n}{8d}} \tag{17}$$

By the assumption that $n \geq 8\,d\, \log(n)$, the second term on the right-hand side is bounded by $r\,\frac{d}{n}$, and hence, dominated by the first term. This gives the desired bound on $\mathrm{err}_{\mathsf{Prot}_{\mathsf{AdSamp}},\ L_\infty}(\mathcal{Q}_\infty(r), d, n)$.
□

References

[ASZ18] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Communication efficient, sample optimal, linear time locally private discrete distribution estimation. *arXiv preprint arXiv:1802.04705*, 2018.

[BNS$^+$16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, 2016.

[BNS18] Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 435–447. ACM, 2018.

[BNST17] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. Practical locally private heavy-hitters. *NIPS*, 2017.

[BS15] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pages 127–135. ACM, 2015.

[Bul] Valeriĭ Buldygin. *Metric characterization of random variables and random processes.*

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2566–2574. Curran Associates, Inc., 2015.

[DJW13a] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

[DJW13b] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

[DKM$^+$06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, 2014.

[HR10] Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 61–70. IEEE, 2010.

[KBR16] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387*, 2016.

[NTZ13] Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 351–360. ACM, 2013.

[Rig15] Philippe Rigollet. *Lecture Notes. 18.S997: High Dimensional Statistics*. MIT Courses/Mathematics, 2015. https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015.

[TVV$^+$17] A.G. Thakurta, A.H. Vyrros, U.S. Vaishampayan, G. Kapoor, J. Freudiger, V.R. Sridhar, and D. Davidson. Learning new words, 2017. US Patent 9,594,741.

[War65] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[YB18] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 2018.