Linear Queries Estimation with Local Differential Privacy

Raef Bassily
The Ohio State University

Abstract

We study the problem of estimating a set of d linear queries with respect to some unknown distribution $\mathbf{p}$ over a domain $\mathcal{J} = [J]$ based on a sensitive data set of n individuals under the constraint of *local differential privacy*. This problem subsumes a wide range of estimation tasks, e.g., distribution estimation and d-dimensional mean estimation. We provide new algorithms for both the offline (non-adaptive) and adaptive versions of this problem.

In the offline setting, the set of queries are fixed before the algorithm starts. In the regime where $n \lesssim d^2/\log(J)$, our algorithms attain L_2 estimation error that is independent of d. 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 $n \lesssim J^2/\log(J)$. Our bounds are within a factor of at most $(\log(J))^{1/4}$ from the optimal L_2 error when $n \lesssim d^2/\log(J)$. 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. In each round, a query 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].

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

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.

1.1 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.

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

[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.

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.

bound in [YB18]; specifically, [YB18, Theorem IV]. From this theorem, it is clear that the lower bound is only valid when $n \geq$ 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 $\max_{k\in[d]} |y_k - \langle \mathbf{q}_k, \ \mathbf{p}\rangle|$ 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 sub-optimal (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

A more detailed version of this section is provided in the attached full version.

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

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

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 $\mathbf{p} \in \mathsf{Simplex}(J)$ be any (unknown) distribution over a data domain $\mathcal{J} = [J]$, where $\mathsf{Simplex}(J)$ is 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 $\mathbf{p}$. 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 is defined as $\mathrm{err}_{\mathsf{Prot},L_2}(\mathbf{A}; \mathbf{p}^n) \triangleq \mathop{\mathbb{E}}_{\mathsf{Prot},\ D\sim\mathbf{p}^n}[\|\mathsf{Prot}(\mathbf{A}, D) - \mathbf{Ap}\|_2]$,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 $\mathrm{err}_{\mathsf{Prot},L_2}(\mathcal{C}_2(r), n) \triangleq$

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

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.

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

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 a LDP protocol for any sequence of d adaptively chosen queries $\mathbf{q}_1, \ldots, \mathbf{q}_d \in \mathcal{Q}_\infty$. We define the worst-case L_∞ error of a protocol Prot as $\mathrm{err}_{\mathsf{Prot},L_\infty}(\mathcal{Q}_\infty(r), d, n) \triangleq$

$$\sup_{\mathbf{p}} \sup_{\substack{\text{adaptive strategy}\\ \text{for } \mathbf{q}_1,\ldots,\mathbf{q}_d}} \mathop{\mathbb{E}}_{\mathsf{Prot},\ D\sim\mathbf{p}^n}\left[\max_{k\in[d]} \left|\, \mathsf{Prot}^{(k)}(D) - \langle \mathbf{q}_k,\ \mathbf{p}\rangle \,\right|\right], \tag{4}$$

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

2.3 Geometry facts

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 $L \triangleq \mathbf{A}\mathbb{B}_1^J = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y} = \mathbf{Ax} \text{ for some } \mathbf{x} \in \mathbb{R}^J \text{ with } \|\mathbf{x}\|_1 \leq 1\}$.

The following fact is a direct consequence of standard results in convex geometry (see the attached full version for details).

Fact 2.2. *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*

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

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 $\mathbf{A} \in \mathcal{C}_2(r)$.

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 $n \lesssim d^2/\log(J)$. In particular, in such regime, our protocol yields an L_2 error $\approx r \left(\frac{\log(J)}{n}\right)^{1/4}$, which is within a factor of $\log^{1/4}(J)$ from the optimal error. 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 .

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 ϵ, δ.

1: **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 $\mathbf{A}$, $\sigma^2 = 2\, r^2\, \frac{\log(2/\delta)}{\epsilon^2}$, 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.

We now give the privacy and accuracy guarantees of our protocol.

Theorem 3.1. *Algorithm 1 is (ϵ, δ)-LDP.*

The proof follows directly from standard analysis of the Gaussian mechanism [DKM+06, NTZ13].

Theorem 3.2. *The worst-case L_2 error of 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 ϵ, δ.

1: **for** Users $i = 1$ to n **do**
2: User i computes $\tilde{\mathbf{y}}_i = \mathcal{R}_i^{\mathsf{Gauss}}(v_i)$ and sends it to the server.
3: **end for**
4: Server computes $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{y}}_i$.
5: **if** $n < \frac{d^2 \log(2/\delta)}{8\,\epsilon^2\,\log(J)}$: **then**
6: $\hat{\mathbf{y}} = \arg\min_{\mathbf{w} \in \mathbf{A}\mathbb{B}_1^J} \|\mathbf{w} - \bar{\mathbf{y}}\|_2^2$ where $\mathbb{B}_1^J$ is the unit L_1 ball in $\mathbb{R}^J$.
7: **else**
8: $\hat{\mathbf{y}} = \bar{\mathbf{y}}$
9: **end if**
10: **return** $\hat{\mathbf{y}}$.

$\mathrm{err}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n)$ *is upper-bounded by*

$$r \cdot \min\left(\left(\frac{32\,\log(J)\log(2/\delta)}{n\epsilon^2}\right)^{1/4},\ \sqrt{\frac{2\,d\log(2/\delta)}{n\epsilon^2}}\right),$$

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 $\widehat{\mathrm{err}}_{\mathsf{Prot}_{\mathsf{Gauss}},\, L_2}(\mathcal{C}_2(r), n) = \sqrt{\frac{\sigma^2\, d}{n}} = r\sqrt{\frac{2\,d\log(2/\delta)}{n\epsilon^2}}$, 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 Fact 2.2, 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

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

Hence, in this case, we have

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

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

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

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

□

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 protocol into a pure LDP one. Our construction is inspired by the idea of rejection sampling in [BS15, BNS18].

In our construction, we assume that $\epsilon \le 1$[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 $\sigma^2 = 4\,r^2\,\frac{\log(n)}{\epsilon^2}$. (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.

The privacy and accuracy guarantees are given by the following theorems.

Theorem 3.3. *Algorithm 3 is ϵ-LDP.*

A detailed proof is provided in the attached full version. We give here a high-level idea of the proof technique. To show that $\mathcal{R}_i$ is ϵ-LDP, it suffices to show that for any $v, v' \in [J]$, any $b \in \{0,1\}$, we have $\mathbb{P}_{\mathcal{R}_i^{\mathsf{RejSamp}}(v)}[B_i = b] \le e^{\epsilon}\,\mathbb{P}_{\mathcal{R}_i^{\mathsf{RejSamp}}(v')}[B_i = b]$. We start by defining "good" event $\mathsf{Good}_i(v) \triangleq \left\{\mathbf{y} \in \mathbb{R}^d : \frac{1}{2}\,\frac{f_{\mathbf{a}_v}(\mathbf{y})}{f_{\mathbf{0}}(\mathbf{y})} \in [\frac{e^{-\epsilon/4}}{2}, \frac{e^{\epsilon/4}}{2}]\right\}$. Note that by the standard analysis of the Gaussian mechanism, we have $\mathbb{P}[\tilde{\mathbf{y}}_i \notin \mathsf{Good}_i(v)] \lesssim 1/n^2$, when $\tilde{\mathbf{y}}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2)$. We then proceed to obtain upper and lower bounds on $\mathbb{P}[B_i = b]$ in terms of $\mathbb{P}[B_i = b |\, \tilde{\mathbf{y}}_i \in \mathsf{Good}_i(v)]$ and $\mathbb{P}[\tilde{\mathbf{y}}_i \notin \mathsf{Good}_i(v)]$. Since the former always lies in $[\frac{e^{-\epsilon/4}}{2}, \frac{e^{\epsilon/4}}{2}]$ and the latter is bounded by $1/n^2$, we can bound the aforementioned ratio by e^{ϵ}.

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

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 ϵ.

Get $\mathbf{a}_{v_i}$: the v_i-th column of $\mathbf{A}$.
Sample a Gaussian vector $\tilde{\mathbf{y}}_i \sim \mathcal{N}(\mathbf{0},\, \sigma^2\,\mathbb{I}_d)$, where $\sigma^2 := 4\,r^2\,\frac{\log(n)}{\epsilon^2}$.
Compute (scaled) ratio of the two Gaussian densities $f_{\mathbf{a}_{v_i}}$ and $f_{\mathbf{0}}$ at $\tilde{\mathbf{y}}_i$: $\eta_i := \frac{1}{2}\,\frac{f_{\mathbf{a}_{v_i}}(\tilde{\mathbf{y}}_i)}{f_{\mathbf{0}}(\tilde{\mathbf{y}}_i)}$.
if $\eta_i \in [\frac{e^{-\epsilon/4}}{2}, \frac{e^{\epsilon/4}}{2}]$ **then**
  Sample a bit $B_i \sim \mathsf{Ber}(\eta_i)$
else
  Let $B_i = 0$
end if
if $B_i = 1$ **then**
  return $\tilde{\mathbf{y}}_i$
else
  return $\perp$ {The output in this case indicates that user i is dropped out of the protocol.}
end if

Theorem 3.4. *Suppose $n \ge 120$. Then, Protocol $\mathsf{Prot}_{\mathsf{RejSamp}}$ (Algorithm 4) has L_2 error $\mathrm{err}_{\mathsf{Prot}_{\mathsf{RejSamp}},\,L_2}(\mathcal{C}_2(r),n)$ that is upper bounded as*

$$r \cdot \min\left(\left(\frac{280\,\log(J)\log(n)}{n\epsilon^2}\right)^{1/4},\ \sqrt{\frac{10\,d\log(n)}{n\epsilon^2}}\right)$$

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

The detailed proof is provided in the attached full version. 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 signal generated by an active user via the pure ϵ local randomizer $\mathcal{R}_i^{\mathsf{RejSamp}}$ is identically distributed to (hence, statistically indistinguishable from) the signal that could have been generated if this users has 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.

3.3 On Tightness of the Bound

The above bound (Theorem 3.4) is tight up to a factor $\approx (\log(J)\log(n))^{1/4}$. In particular, one can show a lower bound of $\min\left(\frac{1}{n^{1/4}\sqrt{\epsilon}},\ \sqrt{\frac{d}{n\epsilon^2}}\right)$ on the L_2 error. We note

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 ϵ.

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 $\hat{\mathbf{y}}$.

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]. 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. 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 $O(n\,J)$ 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 $n \lesssim \frac{J}{\epsilon^2}$.

Hadamard-Response Protocol of [ASZ18]: In the attached full version, we give a brief description of this protocol for completeness. Due to limited space, we only give here the relevant facts about this protocol, which we denote as $\mathsf{Prot}_{\mathsf{HR}}$: (i) $\mathsf{Prot}_{\mathsf{HR}}$ is ϵ-LDP. (ii) $\mathsf{Prot}_{\mathsf{HR}}$ requires $O(\log(J))$ bits of communication per user, and the running time at each user is also $O(\log(J))$. (iii) The running time at the server is $\tilde{O}(n + J)$. (iv) When $n \gtrsim J/\epsilon^2$ (low-dimensional or large sample setting), then for any true distribution $\mathbf{p} \in \mathsf{Simplex}(J)$, $\mathsf{Prot}_{\mathsf{HR}}$ outputs an estimate $\bar{\mathbf{p}}$ for $\mathbf{p}$ satisfying: $\mathbb{E}_{\mathsf{Prot}_{\mathsf{HR}}, D \sim \mathbf{p}^n} [\|\bar{\mathbf{p}} - \mathbf{p}\|_2] = O\left(\sqrt{\frac{J}{\epsilon^2 n}}\right)$. Following the steps of this protocol, one can show the following fact (see the attached full version for details).

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.*

(See the Preliminaries section of the full version for a formal definition of subGaussian random variables).

We now describe $\mathsf{Prot}_{\mathsf{PHR}}$ (Projected Hadamard-Response) $\mathsf{Prot}_{\mathsf{PHR}}$ 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 ϵ.

$\bar{\mathbf{p}} \leftarrow \mathsf{Prot}_{\mathsf{HR}}(D,\ \epsilon)$
$\hat{\mathbf{p}} = \arg\min_{\mathbf{w} \in \mathsf{Simplex}(J)} \|\mathbf{w} - \bar{\mathbf{p}}\|_2^2$.
return $\hat{\mathbf{p}}$.

Note that $\mathsf{Prot}_{\mathsf{PHR}}$ is ϵ-LDP since $\mathsf{Prot}_{\mathsf{HR}}$ is ϵ-LDP .

Theorem 4.2. *Let* $c_\epsilon \triangleq \frac{e^\epsilon+1}{e^\epsilon-1}$. $\mathsf{Prot}_{\mathsf{PHR}}$ *has* L_2 *error*

$$\mathrm{err}_{\mathsf{Prot}_{\mathsf{PHR}},\ L_2}(n) \triangleq \sup_{\mathbf{p}} \mathbb{E}_{\mathsf{Prot}_{\mathsf{PHR}},\ D \sim \mathbf{p}^n} [\|\mathsf{Prot}_{\mathsf{PHR}}(D) - \mathbf{p}\|_2]$$

that is upper bounded by

$$\min\left(\left(\frac{256\, c_\epsilon^2\, \log(J)}{n}\right)^{1/4},\ \sqrt{\frac{4\, c_\epsilon^2\, J}{n}}\right).$$

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 $\mathbf{p}^* = \arg\min_{\mathbf{w}\in\mathbb{B}_1^J} \|\mathbf{w} - \bar{\mathbf{p}}\|_2^2$. Fact 2.2 tells us that $\|\mathbf{p}^* - \mathbf{p}\|_2^2 \le 4 \max_{j\in[J]} |\langle \bar{\mathbf{p}} - \mathbf{p},\ \mathbf{e}_j\rangle|$, 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 $\|\hat{\mathbf{p}} - \mathbf{p}\|_2 \le \|\mathbf{p}^* - \mathbf{p}\|_2$. 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, Fact 2.2 implies that $\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 \le 4 \max_{j\in[J]} |\langle \bar{\mathbf{p}} - \mathbf{p},\ \mathbf{e}_j\rangle|$. 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. Thus, by using the standard bounds on the maximum of subGaussian r.v.s (see [Rig15, Theorem 1.16]), we have $\mathbb{E}_{\mathsf{Prot}_{\mathsf{PHR}},\ D\sim\mathbf{p}^n} \left[\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2\right] \le \sqrt{\frac{256\, c_\epsilon^2\, \log(J)}{n}}$.

□

5 ϵ-LDP for Adaptive Linear Queries

We now consider the problem of estimating a sequence of d adaptively chosen linear queries $\mathbf{q}_1, \ldots, \mathbf{q}_d$ from $\mathcal{Q}_\infty(r)$ over unknown distribution $\mathbf{p} \in \mathsf{Simplex}(J)$. 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. 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. *Algorithm 6 is ϵ-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. First, note that $c_\epsilon \ge 1$ and $|\mathbf{q}_k| \le r$. Hence, $\frac{1}{2}\left(1 + \frac{\mathbf{q}_k(v_i)}{c_\epsilon r}\right)$ and $\frac{1}{2}\left(1 - \frac{\mathbf{q}_k(v_i)}{c_\epsilon r}\right)$ in Step 4 are legitimate probabilities. Observe that for any pair $v_i, v_i' \in [J]$,

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

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

Require: Data set $D = \{v_i \in [J] : i \in [n]\}$, privacy parameter ϵ, d adaptive queries from $\mathcal{Q}_\infty(r)$.

1: Each user $i \in [n]$ gets independently assigned (by itslef or via the server) a random uniform index $j_i \leftarrow [d]$.
2: **for** $k = 1, \ldots, d$ **do**
3: **for** all users i **such that** $j_i = k$ **do**
4: 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 $c_\epsilon = \frac{e^\epsilon + 1}{e^\epsilon - 1}$.
5: **end for**
6: Server computes an estimate $\bar{y}_k = \frac{1}{\hat{n}_k} \sum_{i:j_i=k} \tilde{y}_{k,i}$ where $\hat{n}_k = |\{i \in [n] :\ j_i = k\}|$ is the number of active users in round k.
7: Server chooses a new query $\mathbf{q}_{k+1} \in \mathcal{Q}_\infty(r)$
8: **end for**
9: **return** Estimated vector $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_d)$.

We can bound the ratio of the probabilities for the case of $\tilde{y}_{k,i} = -c_\epsilon\, r$ in a similar fashion. □

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

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

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.

The proof of the above theorem is provided in the attached full version. Given the *offline* lower bound in [DJW13b], this result shows that adaptivity does not pose any extra penalty in the *true* L_∞ 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. 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 [BNS+16, Corollary 6.3] with the bound attained by [HR10] in the offline setting.

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.