Query-Efficient Locally Private Hypothesis Selection via the Scheffe Graph*

Gautam Kamath
University of Waterloo and Vector Institute
g@csail.mit.edu

Alireza F. Pour
University of Waterloo
alireza.fathollahpour@uwaterloo.ca

Matthew Regehr
University of Waterloo
matt.regehr@uwaterloo.ca

David P. Woodruff
Carnegie Mellon University
dwoodruf@cs.cmu.edu

Abstract

We propose an algorithm with improved query-complexity for the problem of hypothesis selection under local differential privacy constraints. Given a set of k probability distributions Q, we describe an algorithm that satisfies local differential privacy, performs $\tilde{O}(k^{3/2})$ non-adaptive queries to individuals who each have samples from a probability distribution p, and outputs a probability distribution from the set Q which is nearly the closest to p. Previous algorithms required either $\Omega(k^2)$ queries or many rounds of interactive queries. Technically, we introduce a new object we dub the Scheffé graph, which captures structure of the differences between distributions in Q, and may be of more broad interest for hypothesis selection tasks.

1 Introduction

Hypothesis selection refers to the following statistical question: given n samples from a distribution p, and descriptions of k distributions Q, output a distribution $\hat{q}$ which is as close to p as possible. More precisely, for an $\alpha > 0$, the goal is to output a distribution $\hat{q}$ such that

$$\|\hat{q} - p\|_1 \leq O(1) \cdot \min_{q \in Q} \|q - p\|_1 + \alpha.$$

In other words, the ℓ_1-distance between p and the output distribution $\hat{q}$ is at most a constant factor larger than that of the closest distribution $q^* \in Q$, up to some additional additive error α. How many samples are needed for this task, and what algorithms do we use to do it? This fundamental primitive serves as an important building block for many other statistical estimation tasks. Furthermore, it generalizes one of the most classic problems in statistics, simple hypothesis testing, wherein the distribution p is promised to be exactly equal to one of the distributions $q \in Q$.

Many classical works (e.g., [Yat85, DL96, DL97, DL01]) address and resolve these questions, showing that $n = O(\log k)$ samples suffice. That is, we require only *logarithmically*-many samples in order to identify the (near-)best distribution. Subsequently, many other works have studied hypothesis selection subject to other constraints and desiderata, including computational efficiency, robustness, weaker access to hypotheses, and more (see, e.g., [MS08, DDS12, DK14, SOAJ14, AJOS14, DKK+16, AFJ+18, BKM19, BKSW19, GKK+20]).

We focus on the constraint of differential privacy (DP) [DMNS06], a rigorous notion of data privacy that guarantees that a procedure will not leak too much information about individual points in the

*Authors are listed in alphabetical order.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

dataset. DP has been adopted in practice by numerous organizations, including Google [XZA+23], Apple [Dif17], and the US Census Bureau [AACM+22]. Under the *central* notion of DP, wherein there exists a trusted curator who may observe the sensitive dataset directly, Bun, Kamath, Steinke, and Wu showed that $O(\log k)$ samples still suffice to perform hypothesis selection [BKSW19, BKSW21].

However, central DP requires a trusted curator, a strong assumption when operating on sensitive data. Instead, one can consider *local* DP (LDP) [War65, EGS03, KLN+11]: in this setting, every dataholder makes their own outputs DP before sharing them with anyone else. This offers much stronger privacy semantics than central DP, but also requires more data for most tasks.

Work by Gopi, Kamath, Kulkarni, Nikolov, Wu, and Zhang [GKK+20] initiated the study of hypothesis section under local DP [GKK+20]. In this case, each user holds an independent sample from the unknown distribution p. Unfortunately, lower bounds of Duchi and Rogers for sparse mean estimation imply that $\Omega(k)$ samples are necessary for this problem [DR19], exponentially more than the $O(\log k)$ samples which suffice for the central DP setting.

With this barrier in mind, [GKK+20] proved two main results. First, they provided an $\tilde{O}(k)$-sample algorithm for locally private hypothesis selection. This matches the lower bound of Duchi and Rogers up to logarithmic factors, and subsequent work by Pour, Ashtiani, and Asoodeh improves the upper bound to $O(k)$ [PAA24], matching the lower bound up to constant factors. The major caveat of both these algorithms is that they require *interactivity*. This is because the specific queries asked to each dataholder depend on the outputs provided by earlier dataholders (i.e., the queries are selected adaptively). This style of interactivity can be a non-starter for real-world deployments of local DP. If one desires a *non-interactive* algorithm, a straightforward privatization of the celebrated Scheffé tournament requires $O(k^2)$ samples. [GKK+20] improve upon this with their second main result, a non-interactive $\tilde{O}(k)$-sample algorithm, but for the simpler problem of k-wise simple hypothesis testing, where the distribution p is promised to be *equal* to one of the distributions $q \in Q$.[2]

To summarize, we highlight three existing results under LDP:

- An interactive $O(k)$-sample algorithm for hypothesis selection;
- A non-interactive $O(k^2)$-sample algorithm for hypothesis selection; and
- A non-interactive $\tilde{O}(k)$-sample algorithm for simple hypothesis testing.

1.1 Results and Techniques

Our main result improves upon all of these, providing a non-interactive $\tilde{O}(k^{3/2})$-sample algorithm for LDP hypothesis selection, where $\tilde{O}(f) = f \cdot \text{polylog}(f)$. Definitions are given in Section 2.

Theorem 1. *Given a set of k distributions Q and $\tilde{O}(k^{5/2})$ expected preprocessing time[3], there exists a non-interactive ε-locally differentially private algorithm with the following guarantees. For any $\alpha, \beta > 0$, there is*

$$n_0 \leq O\left(\frac{k^{3/2}\sqrt{\log k}\log(k/\beta)}{\alpha^2\varepsilon^2}\right)$$

such that given $n \geq n_0$ samples from a distribution p, then with probability at least $1-\beta$ the algorithm outputs a distribution $\hat{q} \in Q$ satisfying

$$\|\hat{q} - p\|_1 \leq 13 \cdot \min_{q \in Q} \|q - p\|_1 + \alpha.$$

To prove this result, we first introduce in Section 3 a generalization of the classical minimum distance estimator that accepts any collection of queries that is rich enough to facilitate comparisons between any pair of distributions. Next, we show in Section 4 how standard tools from differential privacy can

[2]We simplify for the sake of presentation: they actually show a slightly stronger result. Let $OPT = \min_{q \in Q} \|q - p\|$. Roughly speaking, their Lemma 4.1 provides a non-interactive ε-LDP algorithm such that, given $n = \tilde{O}(k/\alpha^4\varepsilon^2)$ samples, it outputs a distribution $\hat{q}$ such that $\|\hat{q} - p\| \leq O(\sqrt{\log k}) \cdot \sqrt{OPT} + O(\alpha)$. Note that, compared to our desired $O(1) \cdot OPT$ guarantees, their result degrades quadratically in the value of OPT, and weakens as the number of hypotheses k becomes large.

[3]Preprocessing involves computing many probabilities $q(E)$ for $q \in Q$, which we treat as constant-time.

be used to non-interactively estimate the queries under local privacy. Our final and most crucial step is to introduce in Section 5 a new combinatorial object, the Scheffé graph. This is a directed graph whose vertices correspond to possible queries that a locally private hypothesis selection algorithm may ask and whose directed edges indicate when one query gives sufficient information to answer another query. It is natural to ask for a minimal set of queries that yield sufficient information to answer all queries and we show that there indeed exists such a small such set of queries. Theorem 1 then follows immediately by combining Theorem 11 in Section 4 with Theorem 13 in Section 5 below.

The natural question is whether our bound can be strengthened to achieve an $\tilde{O}(k)$ sample complexity for non-interactive locally private hypothesis selection. A core part of our analysis involves showing an $\tilde{O}(k^{3/2})$ bound on the domination number of any Scheffé graph – if this could be improved to $\tilde{O}(k)$, then it would produce the desired result. However, in Section 6.1, we provide a nearly-matching lower bound on the domination number, showing that additional structure must be employed to go beyond this barrier.

Another approach to designing an $\tilde{O}(k)$ sample algorithm is based on a suggestion of [GKK+20]. A key technical component of their work is a so-called *flattening lemma* – they point out that a specific strengthening would lead to an $\tilde{O}(k)$ sample algorithm. In Section 6.2, we provide a concrete counterexample to such a strengthening, showing that it is not achievable.

1.2 Related Work

Hypothesis selection is a classical statistical task. This style of approach was introduced by Yatracos [Yat85], and further developed in subsequent work by Devroye and Lugosi [DL96, DL97, DL01]. The most relevant line of work to ours studies hypothesis selection under differential privacy constraints [BKSW19, AAK21, GLW21, GKK+23, PAA24]. However, all of these works either study a weaker notion of privacy, require interactivity, require more data, or apply to a weaker problem than our work. Another line of work focuses on algorithms for (non-private) hypothesis selection that minimize the number of comparisons or the amount of computation [MS08, DDS12, DK14, SOAJ14, AJOS14, AFJ+18, ABS24]. While many of these algorithm require only a near-linear number of comparisons between hypotheses, they are unsuitable for our purposes as they perform adaptive queries, which would result in an interactive protocol in our setting. There are a number of other works on hypothesis selection, focusing on desiderata including robustness [DKK+16, BBKL23], approximation factor [BKM19, BBK+22], memory constraints [ABS23], and more [QCR20, AAC+23, AAC+24]. There has also been significant work into hypothesis testing under local DP [DJW13, DJW17, GR18, She18, ACFT19, ACT19, JMNR19, AZ24, PAJL24, PJL24], though this often focuses on the non-agnostic case (i.e., when the distribution is exactly equal to one of the given distributions) and $k = 2$. For more coverage of private statistics, see [KU20].

2 Preliminaries

We recall the classic definitions of differential privacy (DP) and local differential privacy (LDP):

Definition 2 ([DMNS06])**.** *An algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is ε-differentially private if, for all $X, X' \in \mathcal{X}^n$ that differ in exactly one entry and $S \subseteq \mathcal{Y}$, we have that*

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S].$$

Definition 3 ([War65, EGS03, KLN+11])**.** *Suppose there are n individuals, where the i-th individual has datapoint X_i. A protocol is non-interactive and ε-local differentially private if, for every $i \in [n]$, individual i computes and outputs a (randomized) message $m_i(X_i)$ (where each $m_i : \mathcal{X} \to \mathcal{Y}$ is independently randomized), and m_i is ε-differentially private. That is, for all $i \in [n]$, any $X_i, X_i' \in \mathcal{X}$, and any event $E \subseteq \mathcal{Y}$, we have that*

$$\Pr[m_i(X_i) \in E] \leq e^{\varepsilon} \Pr[m_i(X_i') \in E].$$

We also recall the notion of a dominating set in a directed graph.

Definition 4. *Let $G = (V, E)$ be a digraph. A dominating set for G is a subset $D \subseteq V$ of vertices such that that, for every vertex $w \in V$, either $w \in D$, or there is $v \in D$ such that $(v, w) \in E$, i.e. there is an edge $v \to w$. We call the size of a minimal dominating set the domination number of G, which we write as dom(G).*

We will sometimes say a vertex v dominates a set of vertices W, which means that for each $w \in W$, either $w = v$ or there is an edge $v \to w$. In the same vein, we say that a set of vertices U dominates a set W if each $w \in W$ is dominated by some $v \in U$.

Finally, we recall the classical Scheffé test. Note that we frequently conflate a distribution q with its mass function (density in the continuous setting) to make expressions such as $q(x)$ and $\langle q, T\rangle$ legible.

Definition 5. *For a pair of distributions $q, q' \in \Delta(\mathcal{X})$ over a domain $\mathcal{X}$, we denote by $\delta(x) := q(x) - q'(x)$ the difference functional from q to q' and we denote by*

$$S(x) := \mathit{sgn}(\delta(x)) = \begin{cases} +1 & \text{if } q(x) \geq q'(x) \\ -1 & \text{if } q(x) < q'(x) \end{cases}$$

the signed Scheffé set from q to q'.

In some sense, "querying" the signed Scheffé set S is the best possible measure of the ℓ_1 distance between q and q', formalized in the following lemma.

Lemma 6. *For any distributions q and q' with signed Scheffé set S,*

$$\|q - q'\|_1 = \langle \delta, S\rangle = \sup_{T \in \{-1,1\}^{\mathcal{X}}} |\langle q - q', T\rangle|.$$

The classical Scheffé test between q and q' involves sampling data from some unknown distribution p, calculating an estimate $\hat{p}_S$ of $\langle p, S\rangle$, then returning q if $\langle q, S\rangle$ is closer to $\hat{p}_S$ than $\langle q', S\rangle$ and returning q' otherwise. This estimator can be shown [DL01] to obtain ℓ_1-error at most

$$3\min\{\|q - p\|_1, \|q' - p\|_1\} + 2|\langle p, S\rangle - \hat{p}_S|.$$

3 The Relaxed Minimum Distance Estimator

In this section, we develop an estimator for k distributions with a similar guarantee to that of the classical Scheffé test. We assume that we are given access to some estimates $\hat{p}_T$ of $\langle p, T\rangle$ where p is an unknown distribution and where T belongs to a family of queries $\mathcal{T}$. Moreover, under the LDP constraints, each query $\hat{p}_T$ requires fresh data, so we would like some estimator that only makes a small number of distinct queries to p.

Definition 7 (Relaxed Minimum Distance Estimator (RMDE))**.** *Let $Q \subseteq \Delta(\mathcal{X})$ be a finite set of distributions and suppose we have collection $\mathcal{T}$ of functionals $T \in \{-1, 1\}^{\mathcal{X}}$ as well as a sequence of query results $\hat{p}_{\mathcal{T}} = (\hat{p}_T)_{T \in \mathcal{T}}$. The relaxed minimum distance estimate given the query results is*

$$\hat{q}(\hat{p}_{\mathcal{T}}) := \arg\min_{q \in Q} \sup_{T \in \mathcal{T}} |\langle q, T\rangle - \hat{p}_T|.$$

The following theorem provides theoretical guarantees for the RMDE – similar to the Scheffé test, it can be decomposed into the error from the optimal hypothesis plus approximation error over the set of functionals $\mathcal{T}$.

Theorem 8. *Let Q be a finite set of distributions over $\mathcal{X}$ and let $\mathcal{T} \subseteq \{-1, 1\}^{\mathcal{X}}$ be a set of functionals with the property that, for each $q, q' \in Q$, there is some $T \in \mathcal{T}$ satisfying*

$$|\langle q - q', T\rangle| \geq \phi \|q - q'\|_1. \tag{$\star$}$$

Then, for any distribution p and query results $\hat{p}_{\mathcal{T}} = (\hat{p}_T)_{T \in \mathcal{T}}$,

$$\|\hat{q}(\hat{p}_{\mathcal{T}}) - p\|_1 \leq (1 + 2\phi^{-1})\|q^* - p\|_1 + 2\phi^{-1} \sup_{T \in \mathcal{T}} |\langle p, T\rangle - \hat{p}_T|$$

where $q^* := \arg\min_{q \in Q} \|q - p\|_1$ denotes the closest distribution to p.

One could take the query set $\mathcal{T}$ to be all $\binom{k}{2}$ signed Scheffé sets between pairs $q, q' \in Q$. This recovers the classical minimum distance estimator [DL01], and would satisfy the theorem condition with $\phi = 1$. Our goal will be to obtain a smaller query set $\mathcal{T}$ (which will translate into fewer queries and samples), at the cost of a smaller value of ϕ.

Proof. Write $\hat{q} := \hat{q}(\hat{p}_{\mathcal{T}})$ for short. Clearly, $\|\hat{q} - p\|_1 \le \|q^* - p\|_1 + \|\hat{q} - q^*\|_1$, so we will just focus on bounding $\|\hat{q} - q^*\|_1$.

Now, by $(\star)$, there must be some $\hat{T} \in \mathcal{T}$ for which

$$\begin{aligned}
\|\hat{q} - q^*\|_1 &\le \phi^{-1} |\langle \hat{q} - q^*, \hat{T} \rangle| \\
&\le \phi^{-1} \sup_{T \in \mathcal{T}} |\langle \hat{q} - q^*, T \rangle| \\
&\le \phi^{-1} \left(\sup_{T \in \mathcal{T}} |\langle \hat{q}, T \rangle - \hat{p}_T| + \sup_{T \in \mathcal{T}} |\langle q^*, T \rangle - \hat{p}_T| \right) \\
&\le 2\phi^{-1} \sup_{T \in \mathcal{T}} |\langle q^*, T \rangle - \hat{p}_T| \\
&\le 2\phi^{-1} \left(\sup_{T \in \mathcal{T}} |\langle q^* - p, T \rangle| + \sup_{T \in \mathcal{T}} |\langle p, T \rangle - \hat{p}_T| \right) \\
&\le 2\phi^{-1} \|q^* - p\|_1 + 2\phi^{-1} \sup_{T \in \mathcal{T}} |\langle p, T \rangle - \hat{p}_T|
\end{aligned}$$

where the last inequality follows from Lemma 6. □

4 Non-Interactive Locally Differentially Private RMDE

In this section, we explain how to get accurate estimates of $\hat{p}_T$ of $\langle p, T \rangle$ under ε-LDP. Consequently, we will achieve non-interactive LDP hypothesis selection by calculating all of these estimates in parallel and then supplying them to the relaxed minimum distance estimator.

We first recall randomized response, which is a classical mechanism that ensures local privacy by flipping the response bit with small probability. This introduces a (correctable) bias.

Lemma 9 ([War65, EGS03, KLN+11])**.** *Randomized response is the randomized function RR_ε that receives $x \in \{-1, 1\}$ and outputs x with probability $\frac{e^\varepsilon}{e^\varepsilon + 1}$ and $-x$ with probability $\frac{1}{e^\varepsilon + 1}$. Randomized response satisfies ε-LDP.*

Assuming each user holds an independent datapoint $x \sim p$, we can estimate our workload of queries $\langle p, T \rangle$ under LDP by applying randomized response to $T(x)$ for each user, averaging, and correcting the bias introduced by RR_ε.

Proposition 10. *Let $\mathcal{T}$ be a collection of functionals $T \in \{-1, 1\}^{\mathcal{X}}$. Then there is an ε-LDP mechanism which requires $m = O\left(\frac{|\mathcal{T}| \log(|\mathcal{T}|/\beta)}{\alpha^2 \varepsilon^2}\right)$ samples and computes estimates $\hat{p}_{\mathcal{T}} = (\hat{p}_T)_{T \in \mathcal{T}}$ such that with probability at least $1 - \beta$, we have $|\langle p, T \rangle - \hat{p}_T| \le \alpha$ for all $T \in \mathcal{T}$.*

Proof. Assume we have a sample S from p distributed locally among users. The curator divides the sample into $|\mathcal{T}|$ disjoint subsets $S_1, \ldots, S_{|\mathcal{T}|}$ each of size $\ell = |S|/|\mathcal{T}| = O\left(\frac{\log(|\mathcal{T}|/\beta)}{\alpha^2 \varepsilon^2}\right)$. Fix an enumeration π on the functionals in $\mathcal{T}$. For each $T \in \mathcal{T}$, every user in $S_{\pi(T)}$ with sample x sends $m(x) := \mathsf{RR}_\varepsilon(T(x))$ to the curator, who computes $\hat{p}_T = \frac{1}{\ell} \cdot \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \left(\sum_{x \in S_{\pi(T)}} m(x) \right)$. This protocol satisfies ε-LDP by Lemma 9. Now, we claim that $\frac{e^\varepsilon + 1}{e^\varepsilon - 1} m(x)$ is an unbiased estimate of $\langle p, T \rangle$. Indeed, $\mathbb{E}_{\mathsf{RR}_\varepsilon, x}\left[\frac{e^\varepsilon + 1}{e^\varepsilon - 1} m(x)\right] = \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \left(\frac{e^\varepsilon}{e^\varepsilon + 1} \mathbb{E}_x[T(x)] - \frac{1}{e^\varepsilon + 1} \mathbb{E}_x[T(x)] \right) = \mathbb{E}_x[T(x)] = \langle p, T \rangle$. Moreover, $\frac{e^\varepsilon + 1}{e^\varepsilon - 1} m(x)$ for $x \in S_{\pi(T)}$ are ℓ i.i.d. random variables with values in $\left[-\frac{e^\varepsilon + 1}{e^\varepsilon - 1}, \frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right]$. We can therefore apply Hoeffding's inequality to conclude that

$$\begin{aligned}
\mathbb{P}\left[|\hat{p}_T - \langle p, T \rangle| \ge \alpha\right] &= \mathbb{P}\left[\left| \frac{1}{\ell} \cdot \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \left(\sum_{x \in S_{\pi(T)}} m(x) \right) - \langle p, T \rangle \right| \ge \alpha \right] \\
&\le \exp\left(- \frac{\ell \alpha^2}{2 \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1} \right)^2} \right) \le \beta / |\mathcal{T}|,
\end{aligned}$$

where the last line follows from the fact that for $\varepsilon \in (0,1)$, we have $(\frac{e^\varepsilon+1}{e^\varepsilon-1})^2 = \Theta(1/\varepsilon^2)$. The union bound yields $|\hat{p}_T - \langle p, T\rangle| \le \alpha$ for all $T \in \mathcal{T}$ with probability at least $1-\beta$, as desired. □

Combining Theorem 8 and Proposition 10, we have the following.

Theorem 11. *Let Q be a set of k distributions over $\mathcal{X}$ and let $\mathcal{T}$ be a set of functionals $T \in \{-1,1\}^{\mathcal{X}}$ with the property that, for each $q, q' \in Q$, there is some $T \in \mathcal{T}$ satisfying*

$$|\langle q - q', T\rangle| \ge \phi \|q - q'\|_1. \tag{$\star$}$$

Then, RMDE is an ε-LDP algorithm that requires $m = O\left(\frac{|\mathcal{T}|\log(|\mathcal{T}|/\beta)}{\phi^2\alpha^2\varepsilon^2}\right)$ samples with the following property. For any distribution p, it outputs a distribution $\hat{q}$ such that with probability at least $1-\beta$

$$\|\hat{q} - p\|_1 \le (1 + 2\phi^{-1})\|q^* - p\|_1 + \alpha,$$

where $q^* := \arg\min_{q\in Q}\|q-p\|_1$ denotes the closest distribution to p.

Our remaining task to prove Theorem 1 is to find a small set $\mathcal{T}$ of queries that satisfies the property ($\star$) with $\phi \ge \Omega(1)$.

5 The Scheffé Graph

In order to outfit RMDE with an appropriate test set $\mathcal{T}$, we will begin with the Scheffé sets and pare them down by exploiting their shared information structure.

Definition 12. *Given distributions $q_1, \dots, q_k \in \Delta(\mathcal{X})$, the induced ϕ-Scheffé graph is the digraph with vertices*[4] $\binom{[k]}{2} = \{\{j,j'\} : 1 \le j < j' \le k\}$ *and an edge $\{i,i'\} \to \{j,j'\}$ whenever*

$$|\langle \delta_{jj'}, S_{ii'}\rangle| \ge \phi\|\delta_{jj'}\|_1$$

where $\delta_{jj'} := q_j - q_{j'}$ and $S_{ii'}$ is the signed Scheffé set from q_i to $q_{i'}$.

Now recall that a dominating set in a digraph is a subset D of its vertices V such that every vertex $v \in V$ either belongs to D or $v \in N_{\text{out}}(D)$, namely v is an out-neighbour of some vertex in D.

Since $|\langle \delta_{jj'}, S_{jj'}\rangle| = \|\delta_{jj'}\|_1$ by Lemma 6, then clearly for any dominating set D in the ϕ-Scheffé graph, $\mathcal{T} := \{S_{jj'} : \{j,j'\} \in D\}$ will satisfy condition ($\star$) of Theorem 8, so our main goal in this section is to demonstrate the existence of a small dominating set. In particular, we show that the $1/6$-Scheffé graph for any set of k distributions has domination number $\tilde{O}(k^{3/2})$.

Theorem 13. *For $\phi = 1/6$ and any distributions $q_1, \dots, q_k$, the induced ϕ-Scheffé graph has domination number at most $4k^{3/2}\sqrt{\log k}$. Moreover, there exists a randomized preprocessing algorithm that finds a dominating set of this size in $O(k^{5/2}\sqrt{\log k})$ expected time.*

Note that this bound may be loose. We ran simulations for randomly selected Q and observed weak empirical evidence that the domination number behaves as $\tilde{O}(k)$. This is because, for small values of k (namely < 20), the Scheffé graph appears to be much denser than the following analysis suggests.

In any case, the first step to proving the bound is to examine the structure of a fixed *triangle* $\{\{j,j'\},\{j',j''\},\{j,j''\}\}$. We will argue that, if there are no vertices whose corresponding δ has small ℓ_1-length (relative to the other vertices), then each vertex sends an edge to at least one other vertex in the triangle. On the other hand, if a vertex has very small ℓ_1-length, then we can show that the remaining two vertices must share a bidirectional edge.

Proposition 14 (Triangular Substructure)**.** *For $\phi = 1/6$, the induced ϕ-Scheffé graph on any distributions $q_1, \dots, q_k$ has the following triangular structure. For every $\{j,j',j''\} \in \binom{[k]}{3}$, the graph has at least one of the following edge structures:*

(i) $\{j,j''\} \leftrightarrow \{j',j''\}$

(ii) $\{j,j'\} \to \{j,j''\}$

[4] More generally, we write $\binom{X}{t} := \{A \subseteq X : |A| = t\}$ for shorthand.

(iii) $\{j, j'\} \to \{j', j''\}$

The argument relies on the following geometric property of a triangle in a metric space.

Lemma 15. *For any three points x, y, z in a metric space with metric d, let $a := d(x, y)$, $b := d(x, z)$, and $c := d(y, z)$ denote the lengths of each leg of the triangle xyz. Then either*

$$a \leq \frac{1}{2}b \quad \textit{and} \quad a \leq \frac{1}{2}c \qquad \textit{or} \qquad a > \frac{1}{3}b \quad \textit{and} \quad a > \frac{1}{3}c.$$

To prove the lemma, assume the first condition fails, namely $a > \frac{1}{2}b$ or $a > \frac{1}{2}c$. If $a > \frac{1}{2}b$, then certainly $a > \frac{1}{3}b$ and, by the triangle inequality,

$$a > \frac{1}{2}b \geq \frac{1}{2}(c - a) \implies \frac{3}{2}a > \frac{1}{2}c \implies a > \frac{1}{3}c.$$

The case $a > \frac{1}{2}c$ is analogous.

Proof of Proposition 14. For simplicity, let $j = 1$, $j' = 2$, and $j'' = 3$. With an eye toward the geometric lemma, assume first that δ_{12} is "short", i.e.

$$\|\delta_{12}\|_1 \leq \frac{1}{2}\|\delta_{13}\|_1, \frac{1}{2}\|\delta_{23}\|_1.$$

In this case, since $\delta_{23} = \delta_{13} - \delta_{12}$, the triangle inequality yields

$$\|\delta_{23}\|_1 = |\langle \delta_{23}, S_{23}\rangle| \leq |\langle \delta_{13}, S_{23}\rangle| + |\langle \delta_{12}, S_{23}\rangle| \leq |\langle \delta_{13}, S_{23}\rangle| + \|\delta_{12}\|,$$

so

$$|\langle \delta_{13}, S_{23}\rangle| \geq \|\delta_{23}\|_1 - \|\delta_{12}\|_1 \geq \frac{1}{2}\|\delta_{23}\|_1 \geq \frac{1}{2}(\|\delta_{13}\|_1 - \|\delta_{12}\|_1) \geq \frac{1}{4}\|\delta_{13}\|_1$$

and thus we have an edge $\{2, 3\} \to \{1, 3\}$. By symmetry we also have an edge $\{1, 3\} \to \{2, 3\}$.

On the other hand, by Lemma 15, the remaining case to consider is that

$$\|\delta_{12}\|_1 > \frac{1}{3}\|\delta_{13}\|_1, \frac{1}{3}\|\delta_{23}\|_1.$$

Then, since $\delta_{12} = \delta_{13} - \delta_{23}$, we have

$$\|\delta_{12}\|_1 = |\langle \delta_{12}, S_{12}\rangle| \leq |\langle \delta_{13}, S_{12}\rangle| + |\langle \delta_{23}, S_{12}\rangle|.$$

By averaging, either $|\langle \delta_{13}, T_{12}\rangle| \geq \frac{1}{2}\|\delta_{12}\|_1 > \frac{1}{6}\|\delta_{13}\|_1$, so we have an edge $\{1, 2\} \to \{1, 3\}$, or $|\langle \delta_{23}, S_{12}\rangle| \geq \frac{1}{2}\|\delta_{12}\|_1 > \frac{1}{6}\|\delta_{23}\|_1$, in which case we get an edge $\{1, 2\} \to \{2, 3\}$. □

The consequence of this triangular substructure is that the graph must have relatively dense edges and thus only relatively few vertices can be supported by a small number of other signed Scheffé sets.

Proposition 16. *For any $r \geq 1$, the $1/6$-Scheffé graph on any k distributions has at most $3kr$ vertices with in-degree less than r.*

Proof. Suppose not. By averaging over the following covering of the vertex set

$$V = V_1 \cup \cdots \cup V_k$$

where $V_j := \{v \in V : j \in v\}$, there must be some j for which V_j contains at least $3r$ vertices with in-degree less than r. Call these vertices B_j. Now, for any pair of vertices $\{j, j'\} \neq \{j, j''\}$ in B_j, either $\{j, j'\} \leftrightarrow \{j, j''\}$, $\{j', j''\} \to \{j, j'\}$, or $\{j', j''\} \to \{j, j''\}$ by Proposition 14. That is, for each pair of vertices $v \neq v'$ in B_j, there is at least one corresponding edge that lands in B_j, so

$$\sum_{v \in B_j} d_{\text{in}}(v) \geq \binom{|B_j|}{2} = \frac{1}{2}|B_j|(|B_j| - 1).$$

By averaging again, there must be some $v \in B_j$ for which

$$d_{\text{in}}(v) \geq \frac{1}{2}(|B_j| - 1) \geq \frac{1}{2}(3r - 1) \geq r,$$

which is a contradiction. □

Proof of Theorem 13. We proceed by dominating vertices with small in-degree separately from vertices with large in-degree.

To that end, set $r := \sqrt{k \log k}$ and let B be those vertices with in-degree less than r. By the previous proposition, $|B| \leq 3k^{3/2}\sqrt{\log k}$.

Now, draw uniformly at random a subset R of size $\ell := k^{3/2}\sqrt{\log k}$ from the whole vertex set V. For a fixed $v \in V \setminus B$,

$$\begin{aligned}\mathbb{P}(v \notin N_{\text{out}}(R)) = \frac{\binom{|V|-d_{\text{in}}(v)}{\ell}}{\binom{|V|}{\ell}} &= \left(\frac{|V| - d_{\text{in}}(v)}{|V|}\right)\left(\frac{|V| - d_{\text{in}}(v) - 1}{|V| - 1}\right)\cdots\left(\frac{|V| - d_{\text{in}}(v) - \ell + 1}{|V| - \ell + 1}\right) \\ &\leq \left(\frac{|V| - d_{\text{in}}(v)}{|V|}\right)^{\ell} \leq 2^{-d_{\text{in}}(v)\ell/|V|} \leq 2^{-2\log k} = \frac{1}{k^2}.\end{aligned}$$

By the union bound, $\mathbb{P}(V \setminus B \nsubseteq N_{\text{out}}(R)) \leq \sum_{v \in V\setminus B} \mathbb{P}(v \notin N_{\text{out}}(R))) \leq |V|/k^2 \leq 1/2$, so, by the probabilistic method, there must be some $R \in \binom{V}{\ell}$ that dominates $V \setminus B$, in which case $D = B \cup R$ is a dominating set of size at most $|B| + \ell \leq 4k^{3/2}\sqrt{\log k}$.

As for finding such a dominating set algorithmically, pick $R \in \binom{V}{k^{3/2}\sqrt{\log k}}$ uniformly at random. Iterate over $v = \{a, b\} \in R$, add v and its triangular out-neighbours to a hashtable in $O(k)$ time[5] by checking $\{a, i\}$ and $\{b, i\}$ for each $i \in [k] \setminus v$. By the preceding calculation, R together with all uncovered vertices in the hashtable is a dominating set of size at most $4k^{3/2}\sqrt{\log k}$ with probability greater than $1/2$, so repeating until this is the case achieves the desired expected runtime. □

6 Barriers to a Near-Linear Algorithm

The ideal algorithm for non-interactive locally private hypothesis selection would require only $\tilde{O}(k)$ samples, matching known lower bounds for the problem [DR19, GKK+20]. In this section, we rule out two different approaches one could employ to design such an algorithm.

6.1 An $\tilde{\Omega}(k^{3/2})$ Lower Bound under Triangular Substructure Assumption

One could conceive of a strengthening of Theorem 13, which argues that the domination number of any Scheffé graph is $\tilde{O}(k)$. Unfortunately, we argue that the triangular substructure described in Proposition 14 is insufficient to yield a better bound than $\tilde{O}(k^{3/2})$. Therefore, to go beyond this bound, one must employ additional structure of the Scheffé graph.

Theorem 17. *For all sufficiently large k, there is a digraph G_k on vertices $\binom{[k]}{2}$ satisfying the triangular substructure condition of Proposition 14 for which*

$$\textit{dom}(G_k) \geq \frac{k^{3/2}}{8\sqrt{\log k}} = \tilde{\Omega}(k^{3/2}).$$

Proof. Draw uniformly at random a set R of size $\ell := \frac{1}{4}k^{3/2}\sqrt{\log k}$ from the vertex set $V = \binom{[k]}{2}$. For a fixed vertex $v = \{a, b\} \in V$, consider the set of indices

$$I_v^R := \{i \in [k] \setminus v : \{a, i\}, \{b, i\} \in R\}$$

[5]We treat the time to check whether $|\langle q_j - q_{j'}, S_{ii'}\rangle| \geq \phi\|q_j - q_{j'}\|_1$ as a single unit of computation. In practice, this requires computing a sum or integral and depends on how the distributions are stored in memory.

that form a triangle with v in which both vertices other than v land in R. We aim to show that this set is small with high probability. Indeed, setting $t := 2\log k$, the union bound yields

$$\mathbb{P}(|I_v^R| \geq t) \leq \sum_{J \in \binom{[k]\setminus v}{t}} \mathbb{P}(\{\{a,i\} : i \in J\} \cup \{\{b,i\} : i \in J\} \subseteq R) = \binom{k-2}{t} \cdot \frac{\binom{|V|-2t}{\ell-2t}}{\binom{|V|}{\ell}}$$

$$\leq \left(\frac{e(k-2)}{t}\right)^t \left(\frac{\ell}{|V|}\right)^{2t} = \left(\frac{e(k-2)}{2\log k} \cdot \frac{\frac{1}{16}k^3 \log k}{\frac{1}{4}k^2(k-1)^2}\right)^t = \left(\underbrace{\frac{e}{8}}_{\leq 1/2} \cdot \underbrace{\frac{k(k-2)}{(k-1)^2}}_{\leq 1}\right)^t$$

$$\leq 2^{-2\log k} = 1/k^2.$$

By another union bound, we can show that all such sets are likely to be small simultaneously, i.e. $\mathbb{P}(\exists v \in V, |I_v^R| \geq t) \leq |V|/k^2 \leq 1/2$, so there must be some $R \subseteq V$ of size $\ell = \frac{1}{4}k^{3/2}\sqrt{\log k}$ for which all I_v^R have size less than $t = 2\log k$.

We can now form the bad digraph G_k on the vertex set V by adding edges as follows. For every $v = \{a,b\} \in V$ and every $i \in [k] \setminus v$,

$$\begin{aligned} \{a,i\} \in R, \{b,i\} \notin R &\implies \{a,b\} \to \{b,i\} \\ \{a,i\} \notin R, \{b,i\} \in R &\implies \{a,b\} \to \{a,i\} \\ \text{otherwise} &\implies \{a,b\} \to \{a,i\} \text{ or } \{a,b\} \to \{b,i\} \text{ arbitrarily.} \end{aligned}$$

Clearly, every triangle of G_k satisfies either condition (ii) or (iii) of Proposition 14. Moreover, for a vertex $v = \{a,b\}$, it can only dominate an element $\{a,i\}$ or $\{b,i\}$ of R if both $\{a,i\}$ and $\{b,i\}$ belong to R, namely $i \in I_v^R$. Therefore, including itself, v dominates at most $|I_v^R| + 1 \leq 2\log k$ elements of R, so any dominating set must have size at least $\frac{|R|}{2\log k} = \frac{k^{3/2}}{8\sqrt{\log k}}$. □

6.2 A Counterexample to Flattening

Another possible technique for non-interactive LDP hypothesis selection is that of *flattening*, which is discussed in [GKK+20]. The following conjecture (Question 4.4 in that work), states that any collection of distributions over a finite domain can be mapped to distributions close to uniform while still preserving their pairwise ℓ_1-distances. Note that the original conjecture contained minor mistakes as it was written—including a missing factor of two—which we have corrected.

Conjecture 18 (Flattening)**.** *Let $q_1, \ldots, q_k \in \Delta([n])$ be distributions that are separated in ℓ_1-distance by at most 2α. Then there exists a randomized map $\phi : [n] \to [m]$ satisfying*

1. *For all $1 \leq j \leq k$ and $y \in [m]$, $\frac{1-\alpha}{m} \leq \phi q_j(y) \leq \frac{1+\alpha}{m}$ and*
2. *For all $1 \leq j < j' \leq k$, $\|\phi q_j - \phi q_{j'}\|_1 \in \Theta(\|q_j - q_{j'}\|_1)$*

where ϕq means the distribution of $\phi(x), x \sim q$.

If the conjecture is true, then one can in effect compare any two distributions q_j and $q_{j'}$ by applying the mapping ϕ to each and then comparing them separately to the uniform distribution $\mathcal{U}([m])$. In this case, only k comparisons to the intermediary uniform distribution are required to gain information about all $\binom{k}{2}$ pairwise comparisons. These comparisons can be carried out in parallel with a small number of samples each, leading to non-interactive LDP hypothesis selection. For more details see the proof of Lemma 4.1 in [GKK+20]. Unfortunately, we show that this conjecture is false.

Counterexample to Flattening Conjecture. For simplicity, we identify a distribution with its mass function as a column vector in $\mathbb{R}^n$ or $\mathbb{R}^m$ and a sequence of distributions $(q_1, \ldots, q_k)$ with the matrix Q whose columns are $q_1, \ldots, q_k$. Similarly, we identify a stochastic map ϕ with a left stochastic matrix (LSM) in $\mathbb{R}^{m \times n}$ so that $(\phi q_1, \ldots, \phi q_k)$ is just a matrix multiplication $\phi(q_1, \ldots, q_k)$

Consider the $n \times n$ identity matrix $E := I_n$. We construct n additional distributions that, for the purposes of flattening, will conflict with the columns of E. To that end, let H denote the $n \times n$

Hadamard matrix, i.e., $H_{ij} := (-1)^{\langle i,j\rangle \bmod 2}$ where i and j are viewed as binary strings of length $\log n$. Key properties of this matrix are that $H/\sqrt{n}$ is orthonormal, every pair of columns differs in exactly $n/2$ entries, and every column sums to 0, except for the first column which is all ones. Now, let F be an $n \times n$ matrix whose first column f_1 is the uniform distribution $\mathbf{1}/n$ and whose j^{th} column f_j is the j^{th} column of H with -1 replaced by 0 and $+1$ replaced by $2/n$. In this case, $Q = (E, F) \in \mathbb{R}^{n\times k}$ ($k := 2n$) consists of distributions separated in ℓ_1-distance by at most 2.

Now, assume we have an LSM $\phi \in \mathbb{R}^{m\times n}$ satisfying the first condition $\phi Q \in [0, 2/m]^{m\times k}$. In particular, all entries of $\phi = \phi E$ must fall in $[0, 2/m]$. On the other hand, by construction we have $H = n(f_1, f_2 - f_1, \ldots, f_n - f_1)$, so, since $\frac{1}{\sqrt{n}}H$ is orthonormal, we have

$$\|\phi(f_1, f_2 - f_1, \ldots, f_n - f_1)\|_F^2 = \|\phi H/\sqrt{n}\|_F^2/n = \|\phi\|_F^2/n \le mn(2/m)^2/n = 4/m.$$

By averaging, there must be some $v \in \{f_1, f_2 - f_1, \ldots, f_n - f_1\}$ for which

$$\|\phi v\|_1^2 \le m\|\phi v\|_2^2 \le m\|\phi(f_1, f_2 - f_1, \ldots, f_n - f_1)\|_F^2/n \le 4/n.$$

Provided that $n > 4$, this is impossible for $v = f_1$ because $\|\phi f_1\|_1 = 1$, so there must be $1 < i \le n$ such that $\|\phi f_i - \phi f_1\|_1 \le 2/\sqrt{n} = o(1) = o(\|f_i - f_1\|_1)$. □

7 Conclusion

In this work we introduce two new techniques for hypothesis selection.

The first is a relaxation of the classical minimum distance estimator in which the Scheffé sets are replaced by any collection of queries that is diverse enough for ℓ_1-comparisons between any pair of candidate distributions.

The second is a new object called the Scheffé graph that contains structural information about the relationship between queries a hypothesis selection algorithm might ask. Our analysis of the Scheffé graph reveals a dense triangular substructure that can be exploited to yield a non-trivial reduction in query complexity. We show that our analysis of query complexity arising from the triangular substructure is nearly tight, so any further reduction in query complexity via the Scheffé graph will require the discovery of additional graph substructure.

Combining these two techniques yields an algorithm for non-interactive hypothesis selection under LDP constraints with state-of-the-art sample complexity, though we stress that our techniques are relevant to hypothesis selection problems more broadly.

Acknowledgments

GK is supported by a Canada CIFAR AI Chair, an NSERC Discovery Grant, and an Ontario Early Researcher Award. MR is supported by an NSERC CGS-D scholarship. DW is supported by a Simons Investigator Award and Office of Naval Research award number N000142112647.

References

[AAC+23] Anders Aamand, Alexandr Andoni, Justin Y. Chen, Piotr Indyk, Shyam Narayanan, and Sandeep Silwal. Data structures for density estimation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 1–18. JMLR, Inc., 2023.

[AAC+24] Anders Aamand, Alexandr Andoni, Justin Y. Chen, Piotr Indyk, Shyam Narayanan, Sandeep Silwal, and Haike Xu. Statistical-computational trade-offs for density estimation. In *Advances in Neural Information Processing Systems 37*, NeurIPS '24, pages 97907–97927. Curran Associates, Inc., 2024.

[AACM+22] John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*, Special Issue 2, 2022.

[AAK21] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 185–216. JMLR, Inc., 2021.

[ABS23] Maryam Aliakbarpour, Mark Bun, and Adam Smith. Hypothesis selection with memory constraints. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23, pages 50453–50481. Curran Associates, Inc., 2023.

[ABS24] Maryam Aliakbarpour, Mark Bun, and Adam Smith. Optimal hypothesis selection in (almost) linear time. In *Advances in Neural Information Processing Systems 37*, NeurIPS '24, pages 141490–141527. Curran Associates, Inc., 2024.

[ACFT19] Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS '19, pages 2067–2076. JMLR, Inc., 2019.

[ACT19] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1–15, 2019.

[AFJ+18] Jayadev Acharya, Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Maximum selection and sorting with adversarial comparators. *Journal of Machine Learning Research*, 19(1):2427–2457, 2018.

[AJOS14] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sorting with adversarial comparators and application to density estimation. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, ISIT '14, pages 1682–1686. IEEE Computer Society, 2014.

[AZ24] Shahab Asoodeh and Huanyu Zhang. Contraction of locally differentially private mechanisms. *IEEE Journal on Selected Areas in Information Theory*, 5:385–395, 2024.

[BBK+22] Olivier Bousquet, Mark Braverman, Gillat Kol, Klim Efremenko, and Shay Moran. Statistically near-optimal hypothesis selection. In *Proceedings of the 62nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '21, pages 909–919. IEEE Computer Society, 2022.

[BBKL23] Shai Ben-David, Alex Bie, Gautam Kamath, and Tosca Lechner. Distribution learnability and robustness. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23, pages 52732–52758. Curran Associates, Inc., 2023.

[BKM19] Olivier Bousquet, Daniel M. Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 318–341, 2019.

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 156–167. Curran Associates, Inc., 2019.

[BKSW21] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *IEEE Transactions on Information Theory*, 67(3):1981–2000, 2021.

[DDS12] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning Poisson binomial distributions. In *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, STOC '12, pages 709–728. ACM, 2012.

[Dif17] Differential Privacy Team, Apple. Learning with privacy at scale. `https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf`, December 2017.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '13, pages 429–438. IEEE Computer Society, 2013.

[DJW17] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2017.

[DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory*, COLT '14, pages 1183–1213, 2014.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 655–664. IEEE Computer Society, 2016.

[DL96] Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimation. *The Annals of Statistics*, 24(6):2499–2512, 1996.

[DL97] Luc Devroye and Gábor Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *The Annals of Statistics*, 25(6):2626–2637, 1997.

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DR19] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1161–1191, 2019.

[EGS03] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 211–222. ACM, 2003.

[GKK+20] Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, pages 1785–1816, 2020.

[GKK+23] Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Raghu Meka, and Chiyuan Zhang. User-level differential privacy with few examples per user. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23, pages 19263–19290. Curran Associates, Inc., 2023.

[GLW21] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 11631–11642. Curran Associates, Inc., 2021.

[GR18] Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 1626–1635. JMLR, Inc., 2018.

[JMNR19] Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *Proceedings of the 60th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '19, pages 94–105. IEEE Computer Society, 2019.

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[KU20] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

[MS08] Satyaki Mahalanabis and Daniel Stefankovic. Density estimation in linear time. In *Proceedings of the 21st Annual Conference on Learning Theory*, COLT '08, pages 503–512, 2008.

[PAA24] Alireza F Pour, Hassan Ashtiani, and Shahab Asoodeh. Sample-optimal locally private hypothesis selection and the provable benefits of interactivity. In *Proceedings of the 37th Annual Conference on Learning Theory*, COLT '24, pages 4240–4275, 2024.

[PAJL24] Ankit Pensia, Amir Reza Asadi, Varun Jog, and Po-Ling Loh. Simple binary hypothesis testing under local differential privacy and communication constraints. *IEEE Transactions on Information Theory*, 71(1):592–617, 2024.

[PJL24] Ankit Pensia, Varun Jog, and Po-Ling Loh. In *Proceedings of the 37th Annual Conference on Learning Theory*, COLT '24, pages 4205–4206, 2024.

[QCR20] Yihui Quek, Clément L. Canonne, and Patrick Rebentrost. Robust quantum minimum finding with an application to hypothesis selection. *arXiv preprint arXiv:2003.11777*, 2020.

[She18] Or Sheffet. Locally private hypothesis testing. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 4605–4614. JMLR, Inc., 2018.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 1395–1403. Curran Associates, Inc., 2014.

[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.

[XZA+23] Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher A Choquette-Choo, Peter Kairouz, H Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of Gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, ACL '23, pages 629–639. Association for Computational Linguistics, 2023.

[Yat85] Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The title and abstract skip our negative results but include our more important positive results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our improved query complexity bound is $\sqrt{k}$ larger than the known lower bound for the problem. We give a hard construction that shows our bound to be tight for our technique and suggests how the bound could be improved by taking our technique further.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Only one lemma (Lemma 6) has no proof but it is a standard easy result for the area.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper conforms to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our paper is a theoretical work with indirect societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.