Covariance-Aware Private Mean Estimation Without Private Covariance Estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present two sample-efficient differentially private mean estimators for d-dimensional (sub)Gaussian distributions with unknown covariance. Informally, given $n \gtrsim d/\alpha^2$ samples from such a distribution with mean μ and covariance Σ, our estimators output $\tilde{\mu}$ such that $\|\tilde{\mu} - \mu\|_\Sigma \leq \alpha$, where $\|\cdot\|_\Sigma$ is the *Mahalanobis distance*. All previous estimators with the same guarantee either require strong a priori bounds on the covariance matrix or require $\Omega(d^{3/2})$ samples.

Each of our estimators is based on a simple, general approach to designing differentially private mechanisms, but with novel technical steps to make the estimator private and sample-efficient. Our first estimator samples a point with approximately maximum Tukey depth using the exponential mechanism, but restricted to the set of points of large Tukey depth. Proving that this mechanism is private requires a novel analysis. Our second estimator perturbs the empirical mean of the data set with noise calibrated to the empirical covariance. Only the mean is released, however; the covariance is only used internally. Its sample complexity guarantees hold more generally for subgaussian distributions, albeit with a slightly worse dependence on the privacy parameter. For both estimators, careful preprocessing of the data is required to satisfy differential privacy.

1 Introduction

Although the goal of statistics and machine learning is to infer properties of a population, there is a growing awareness that many statistical estimators and trained models reveal a concerning amount of information about their data set, which leads to significant concerns about the *privacy* of the individuals who have contributed sensitive information to that data set. These privacy violations have been demonstrated repeatedly via *reconstruction attacks* [21, 30, 27, 44], *membership-inference attacks* [37, 50, 10, 32, 51, 57], and instances of unwanted *memorization of training data* [13, 14, 34, 7]. In order to realize the benefits of analyzing sensitive data sets, it is crucial to develop statistical estimators and machine learning algorithms that make accurate inferences about the population but also protect the privacy of the individuals who contribute data.

In this work we study statistical estimators that satisfy a condition called *differential privacy* [29], which has become the standard criterion for individual privacy in statistics and machine learning. Informally, a differentially private algorithm guarantees that no attacker, regardless of their background knowledge or resources, can infer much more about any individual than they could have learned had that individual never contributed to the data set. A long body of work shows that differential privacy is compatible with a wide range of tasks in statistics and machine learning, and it is now seeing deployment at companies like Google [33, 6, 55], Apple [3], Facebook [53] and LinkedIn [49], as well as statistical agencies like the U.S. Census Bureau [1, 36].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Background: Differentially Private Mean Estimation. We revisit differentially private estimators for one of the most fundamental tasks in all of statistics and machine learning—given $x_1, \ldots, x_n \in \mathbb{R}^d$ sampled i.i.d. from a distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$, estimate the mean μ. Mean estimation is not only an essential summary statistic in its own right, but also a building block for more sophisticated tasks like regression and stochastic optimization.

Without privacy constraints, the natural solution is to output the empirical mean $\mu_x = \frac{1}{n}\sum_i x_i$. The natural way to state the sample-complexity guarantee of the empirical mean is

$$n \gtrsim \frac{d}{\alpha^2} \implies \|\mu_x - \mu\|_\Sigma \le \alpha.$$

Importantly, $\|\mu_x - \mu\|_\Sigma = \|\Sigma^{-1/2}(\mu_x - \mu)\|_2$ is the error in *Mahalanobis distance*. Intuitively, bounding the error in Mahalanobis distance requires that in every direction v, the squared error is proportional to the variance $v^T \Sigma v$ in that direction. The Mahalanobis distance is the natural way to measure the error for mean estimation, since it tightly captures the uncertainty about the true mean and is preserved under affine transformations.

Unfortunately, when the dimension is large, releasing the empirical mean leads to serious privacy breaches [32, 40]. A natural question is thus: can we design a differentially private mean estimator that performs nearly as well as the empirical mean?

Without making additional assumptions, the answer turns out to be no: every differentially private estimator incurs a large overhead in sample complexity compared to the empirical mean [41]. However, it is known that if we further assume that the distribution satisfies some additional concentration properties, we can do much better [52, 43, 40, 41]. In this work, we focus on one class of well-concentrated distributions, those that are *Gaussian* (or, in some of our results, *subgaussian*). Although assuming Gaussian data is restrictive, it is a natural starting point for understanding the complexity of estimation when the distribution is not pathological. Moreover, even in the Gaussian case, one cannot obtain comparable error to the empirical mean unless $n \gtrsim d$ [10, 32, 40], so we will also focus on the case where the sample size is at least as large as the dimension.

For Gaussian data, if the analyst has prior information about the covariance matrix in the form of a matrix A and bound $\kappa \ge 1$ such that $A \preceq \Sigma \preceq \kappa A$,[1] then there is a private estimator $\mathcal{A}(x)$, based on a line of work initiated by Karwa and Vadhan [43, 40, 5, 2], such that

$$n \gtrsim \frac{d}{\alpha^2} + \frac{\kappa d}{\alpha \varepsilon} \implies \|\mathcal{A}(x) - \mu\|_\Sigma \le \alpha, \tag{1}$$

where ε is the *privacy parameter* controlling the level of privacy, with stronger privacy as $\varepsilon \to 0$.[2] We can interpret this result as showing that the additional cost of privacy is small provided that the user has a strong a priori bound on the covariance so that κ is small (e.g. κ is a constant), and also that the privacy guarantee is not too strong (e.g. $\varepsilon \ge \alpha$). In particular, setting $\kappa = 1$ corresponds to the *known-covariance* setting, where the guarantee in (1) is known to be minimax optimal up to polylogarithmic factors [32, 40] among all differentially private estimators.

However, the sample complexity in (1) degrades linearly with κ, a large price to pay for the user's uncertainty. Intuitively, this degradation arises because the algorithm perturbs the empirical mean μ_x with noise from a *spherical* Gaussian distribution, whose magnitude must be proportional to the largest variance in any direction, so the noise is unnecessarily large in the directions with small variance. In contrast, when the user is very uncertain about the covariance, there are estimators with a weaker dependence on κ but a *superlinear* dependence on the dimension. In particular, there is an estimator [40] with an error guarantee of the form

$$n \gtrsim \frac{d}{\alpha^2} + \frac{d}{\alpha \varepsilon} + \frac{d^{3/2} \log^{1/2} \kappa}{\varepsilon} \implies \|\mathcal{A}(x) - \mu\|_\Sigma \le \alpha. \tag{2}$$

Without any prior information about the covariance, the best known approach is to estimate the mean by learning the entire distribution—both mean and covariance—which is the more difficult task

[1] Given two covariance matrices $A, B \in \mathbb{R}^{d \times d}$, the notation $A \preceq B$ indicates that in every direction $v \in \mathbb{R}^d$ the variance $v^T B v$ is at least as large as $v^T A v$. More generally, $A \preceq B$ iff $B - A$ is positive semidefinite.

[2] To simplify the discussion, we focus only on the ε parameter, although our results and many of those we discuss require relaxations of differential privacy such as approximate [28] or concentrated [26, 8] differential privacy, which have different parameterizations.

considered in [2]. Doing so incurs an even worse dependence on the dimension:

$$n \gtrsim \frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} \implies \|\mathcal{A}(x) - \mu\|_\Sigma \le \alpha. \tag{3}$$

The Covariance-Estimation Bottleneck. The bottleneck in the algorithms above is privately obtaining a good spectral approximation to the covariance, i.e. a matrix A such that $A \preceq \Sigma \preceq 2A$. With such an estimate, one can apply the known-covariance approach in (1). Without privacy constraints, the empirical covariance will have this spectral-approximation property when the sample size is $n \gtrsim d$. However, all known private covariance estimators require $n = \Omega(d^{3/2})$ samples, and there is evidence that this is an inherent limitation, as $\Omega(d^{3/2})$ samples are necessary for solving this task for a worst-case data distribution [31].

Our Work: Sample-Efficient Private Mean Estimation. We circumvent this apparent difficulty of covariance estimation by designing an algorithm that adapts the noise it adds to the distribution's covariance without actually providing an explicit covariance estimate, nearly matching the optimal sample complexity (1) for the known-covariance setting.

Theorem 1.1 (Informal)**.** *For $\alpha \le 1$, there is an (ε, δ)-differentially private estimator $\mathcal{A}(\cdot)$ such that if $x = (x_1, \ldots, x_n)$ are sampled from $\mathcal{N}(\mu, \Sigma)$ for unknown μ and Σ,*

$$n \gtrsim \frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\log(1/\delta)}{\varepsilon} \implies \|\mathcal{A}(x) - \mu\|_\Sigma \le \alpha.$$

The above guarantee holds with high probability over the sample x and the randomness of $\mathcal{A}$.

For the formal statement, see Theorem 3.2. Our estimator is based on privately sampling a point of large *Tukey depth*. Tukey depth generalizes the notion of quantiles to multiple dimensions; it is known to be a good robust estimator of the Gaussian mean. The natural way to sample such a point privately is to use the *exponential mechanism* [48], but sampling from a distribution over the entire domain $\mathbb{R}^d$ will not have finite sample complexity. We improve this approach by sampling from a *data-dependent* domain consisting only of points of large Tukey depth, which necessitates careful preprocessing and privacy analysis.

We emphasise that the sample-complexity of this estimator is optimal up to polylogarithmic factors. However, the estimator is not computationally efficient. An interesting open problem is to design an estimator matching the guarantee of Theorem 1.1 with running time polynomial in the dimension.

Beyond Gaussian Distributions. A natural question is how much the assumption of Gaussian data can be relaxed without blowing up the sample complexity. Our second result is an alternative estimator, based on a completely different technique, that will give a similar guarantee for any distribution with *subgaussian tails*. For our purposes, we say that P with mean μ and covariance Σ is subgaussian if, for every direction $u \in \mathbb{R}^d$, the tails of the distribution decay as fast as a univariate normal distribution with mean $u^T\mu$ and variance $Cu^T\Sigma u$ for some constant C. That is for every λ, $\mathbb{E}[e^{\lambda u^T(P-\mu)}] \le e^{C\lambda^2(u^T\Sigma u)/2}$. More generally, our estimator works for any distribution such that the empirical covariance matrix converges rapidly to the population covariance matrix and typical samples are close to the mean in Mahalanobis distance.

Theorem 1.2 (Informal)**.** *For $\alpha \le 1$, there is an (ε, δ)-differentially private estimator $\mathcal{A}(\cdot)$ such that if $x = (x_1, \ldots, x_n)$ are sampled from any subgaussian distribution with unknown mean μ and unknown covariance Σ,*

$$n \gtrsim \frac{d}{\alpha^2} + \frac{d\log(1/\delta)}{\alpha\varepsilon^2} \implies \|\mathcal{A}(x) - \mu\|_\Sigma \le \alpha.$$

The above guarantee holds with high probability over the sample x and the randomness of $\mathcal{A}$.

For the formal statement, see Theorem 4.4. This estimator is based on another simple approach—we perturb the empirical mean μ_x with noise scaled to Σ_x, where Σ_x is the exact (not private) empirical covariance. We show that this approach satisfies differential privacy if the data set satisfies certain concentration properties, which we enforce using a careful preprocessing step.

Both of our estimators generalize beyond Gaussian distributions in different directions, not fully captured by our theorems. Although the Tukey-depth estimator will only return an approximation

to the mean when the distribution is symmetric and will not generalize to arbitrary subgaussian distributions, it returns an approximate *median* for distributions satisfying some natural regularity conditions. In contrast, the empirically rescaled estimator generalizes to distributions that are well concentrated in the sense that typical samples from the distribution are close to the mean in Mahalanobis distance with respect to the empirical covariance, which captures much more than just subgaussian distributions. Exploring the extent to which each estimator can be generalized is an interesting direction for future work.

Limitations. Our estimators are not computationally efficient. We provide finite implementations which construct and search over a sufficiently fine grid, resulting in running time exponential in the dimension d and sample size n. We also stress that the error guarantees of our estimators only hold for families of distributions which satisfy certain regularity conditions, as described above. We leave these two interesting directions of improvement to future work.

Societal Impact. Since our algorithms are inefficient, they will not see practical use without further research. More broadly, our work serves the goal of increasing the array of differentially private algorithms, and thereby enable safer access to privacy-sensitive data. Although such tools might support corporations and governments in harmful data-collection efforts, we believe that the existence of these tools is beneficial on balance.

1.1 Techniques

Tukey-depth Mechanism. Our first algorithm adapts a well-known approach to estimating the location of a distribution with differential privacy. Briefly, we sample from the distribution defined by the exponential mechanism based on the Tukey depth, but restricted to a data-dependent set of possibe outputs—those points with Tukey depth at least $1/4$. To ensure differential privacy, we add a private check that the data set is "safe," which we perform before running the main mechanism.

In more detail, our starting point is the *exponential mechanism* [48]. In this context, the exponential mechanism samples a point $y \in \mathbb{R}^d$ from the distribution with probability density roughly proportional to $\exp(-\varepsilon q(x; y))$, where $q(x; y)$ is a score function that indicates how good a match y is for the data set at hand. To instantiate the mechanism, one must choose (i) a score function that rewards values y that are close to mean μ in the unknown Mahalanobis metric, and (ii) a set of candidate values y from which to sample. For (i), we choose $q(x; y) = nT_x(y)$ where T_x is the *Tukey depth* of a point, defined as

$$T_x(y) = \min_{v \in \mathbb{R}^d} \frac{1}{n} \cdot \left| \left\{ x_i \in x : \langle x_i, v \rangle \geq \langle y, v \rangle \right\} \right|. \tag{4}$$

For normally distributed data, the Tukey depth ranges from 0 (outside the convex hull of the data points) to about $1/2$, near the mean μ. The point of maximal Tukey depth, called the *Tukey median*, is well-known as a robust estimator of the mean of a Gaussian distribution.

Using the exponential mechanism with the Tukey depth as the score function is a well-established idea. In one dimension, it is now the standard algorithm for approximating the median (e.g. [52]), and its high-dimensional variant was studied in previous [42] and concurrent [45] work. However, on its own, it is not sufficient for our needs. The challenge is in specifying the set of potential outputs y from which we sample (step *(ii)* above). In order to reliably output a value y such that $\|y - \mu\|_\Sigma$ is small, we must sample from a set of outputs with Σ-norm that is not too large. For that, however, it would seem one needs to get an initial rough approximation to Σ, which is exactly what we want to avoid.

We circumvent the barrier by sampling from a data-defined set *without releasing a description of that set*. Specifically, consider the algorithm which samples from the exponential mechanism restricted to points with Tukey depth at least $1/4$. A standard concentration argument shows that this set is roughly the ellipsoid $\{y : \|y - \mu\|_\Sigma \leq c\}$ for a modest constant c. Running the exponential mechanism on this set returns a good approximation to the mean (with Σ-norm $o(1)$) when $n = \omega(d/\varepsilon)$.

This gives us an accurate algorithm, dubbed $\mathcal{M}$. The remaining challenge is that $\mathcal{M}$ is not, on its own, differentially private. Specifically, there are data sets x for which the volume of the set of Tukey-depth-$1/4$ points changes drastically when a small number of records in x are changed. To address this, we identify a set of *safe* data sets x—these are data sets such that $\mathcal{M}$ behaves similarly on all data sets x' that are neighbors of x. We show that normally distributed data sets are typically

safe and, furthermore, require many insertions or deletions of records to be made unsafe. This allows us to apply the *propose-test-release* (PTR) framework of [25] to obtain an algorithm that is accurate for nicely-distributed data and differentially private in the worst case.

Our modification of the exponential mechanism is quite general. It is similar in flavor to the GAP-MAX variant [11, 56, 15–17] as well as the top-k-of-k' approach of [24]. However, we do not know how to obtain our results using those variants since they are specific to the discrete setting and appear to require knowledge of the volume of the level sets of the score function. Such knowledge is not obviously available in our setting.

Empirically Rescaled Gaussian Mechanism. The well-known *Gaussian mechanism* perturbs the empirical mean μ_x with noise drawn from $\mathcal{N}(0, C^2\|\Sigma\|_2^2\mathbb{I})$ for a scale parameter $C > 0$. Since the noise is proportional to the direction with the largest variance and the same in every direction, this mechanism cannot in general achieve a good estimate in Mahalanobis distance.

Our approach relies on the following simple idea: If the data set x is drawn i.i.d. from $\mathcal{N}(\mu, \Sigma)$, and the number of samples is a bit larger than the dimension d, then the empirical covariance Σ_x is a good approximation to the true covariance in spectral norm. When this holds, perturbing μ_x with noise drawn from $\mathcal{N}(0, C^2\Sigma_x)$ will be a good estimate of the mean in Mahalanobis distance. Thus, we want to understand when perturbing μ_x in this way can be made differentially private.

Adding noise from $\mathcal{N}(0, C^2\Sigma_x)$ will not be private for worst-case data sets. For example, consider a pair of adjacent data sets, one of which lies in a proper subspace of dimension $d-1$ and the other of which has full rank. For one of these data sets, our mean estimate will always lie in the proper subspace, while for the other it will lie outside of this subspace with probability 1, making the two cases easy to distinguish.

Our main observation is that such pathological examples should not arise when the data sets are sampled from a distribution that satisfies strong *concentration properties*, such as a Gaussian. For example, if x and x' are adjacent data sets of i.i.d. samples from the same Gaussian, then μ_x and $\mu_{x'}$ will be similar, as will Σ_x and $\Sigma_{x'}$. To take advantage of these nice distributions, we define the family of *good data sets* that capture certain properties of typical samples from a Gaussian. Roughly, a data set x is *good* if Σ_x is invertible and, for every x_i, $\|x_i - \mu_x\|_{\Sigma_x} \lesssim \sqrt{d \log n}$. Our main technical contribution is to show that if x and x' differ on a small number of samples, and both data sets are good, then their empirical means and empirical covariances are close. Thus, $\mu_x + \mathcal{N}(0, C^2\Sigma_x)$ and $\mu_{x'} + \mathcal{N}(0, C^2\Sigma_{x'})$ will be indistinguishable in the sense required for (ε, δ)-differential privacy.

However, we need to define our estimator on data sets that are *not* good in such a way that the estimator will be differentially private in the worst case. To do so, we can privately test whether the input data set lies close to the good set and, if needed, we can project the data into the family of good data sets. This preprocessing step will have no effect when the data is Gaussian, and any pair (Gaussian or not) of adjacent data sets x and x' will be mapped to a pair of good data sets $\tilde{x}$ and $\tilde{x}'$ that differ on a small number of examples.

This projection step is stated abstractly in Algorithm 2 as finding the minimizer over an infinite family of data sets. In the supplementary material, we present a concrete, exponential-time algorithm that searches over a discrete grid of candidate datasets. We leave the task of identifying more efficient algorithms as an interesting problem for future work.

1.2 Additional Related Work

Differentially Private Mean and Covariance Estimation. The line of work most relevant to ours was initiated by Karwa and Vadhan [43], who established optimal private mean and variance estimators of univariate Gaussians with sample complexity $\tilde{O}(1/\alpha^2 + 1/\alpha\varepsilon)$, without requiring a priori bounds on the parameters. Previously, Smith [52] gave estimators for asymptotically normal statistics (which include the mean of a Gaussian) with optimal convergence rates for a certain range of privacy parameters. In the multivariate setting, a series of works [40, 11, 2] gives algorithms for Gaussian mean estimation with known covariance that have a near-optimal sample complexity of $\tilde{O}(d/\alpha^2 + d/\alpha\varepsilon)$. We note that [2] obtain the best bound among these works, but the guarantees of the estimator from [40] extend naturally to subgaussian distributions as well. [12] also studied mean and covariance estimation of subgaussian distributions, but in their setting strong a priori bounds

on the parameters are required. Beyond (sub)Gaussian distributions, [4, 9, 41] study differentially private mean estimation under weaker moment assumptions.

In addition to the work we discuss in the introduction [40, 2], recent work focuses on practical private mean and covariance estimation in univariate [23] and multivariate [5] settings, although these approaches still require explicit private covariance estimation. Dwork et al. [31] gave a lower bound of $\Omega(d^{3/2})$ for privately estimating the empirical covariance, albeit via a non-Gaussian distribution. Together, these results serve as evidence that $\Omega(d^{3/2})$ samples are necessary for private covariance estimation for Gaussian data as well.

Robust Statistical Estimation. Robust statistical estimation [39, 47], which dates to at least 1960 [54] and remains an active area of research [18–20, 22, 38], studies the problem of estimating distribution parameters when an α-fraction of the data may be adversarially corrupted. As noted by Dwork and Lei [25], robust statistics and differential privacy have similar goals, and private estimators are often inspired by robust estimators, but the models are formally incomparable.

More recent work aims to give algorithms which satisfy both constraints simultaneously [45, 35]. Specifically, independently from our work, Liu et al. [45] propose a simple mechanism for Gaussian mean estimation with known covariance, given α-corrupted data sets and an a priori bound on the range of the mean $\|\mu\|_\infty \le \rho$. This estimator has sample complexity $\tilde{O}(d/\alpha^2 + d\log\rho/\alpha\varepsilon)$. The algorithm runs the exponential mechanism [48] in the given range, using the Tukey depth of a point as its score. We observe that the same mechanism gives a solution for the problem we study—mean estimation in Mahalanobis distance with unknown covariance and no corruptions—but this solution requires a priori bounds on the mean and covariance, which are not required by our algorithms.

2 Preliminaries

We write $x \sim P$ to denote that x is drawn from distribution P and $x \sim P^{\otimes n}$ if x consists of n i.i.d. draws from P. We write $[n] = \{1, \ldots, n\}$. We define the Hamming distance between two data sets x, y of size n by $D_H(x, y) = |\{i \in [n] : x_i \neq y_i\}|$. For data set x and set $S \subseteq \mathbb{R}^{n\times d}$, we write $D_H(x, S) = \min_{z\in S} D_H(x, z)$.

Let $x, x' \in \mathcal{X}^n$ be two data sets of size n. We say that x, x' are *neighboring data sets* if $D_H(x, x') \le 1$, and denote this by $x \sim x'$. Differentially private algorithms have *indistinguishable* output distributions on neighboring data sets.

Definition 2.1 ((ε, δ)-indistinguishability)**.** *Two distributions P, Q over domain $\mathcal{W}$ are (ε, δ)-indistinguishable, denoted by $P \approx_{\varepsilon,\delta} Q$, if for any measurable subset $W \subseteq \mathcal{W}$,*

$$\Pr_{w\sim P}[w \in W] \le e^{\varepsilon} \Pr_{w\sim Q}[w \in W] + \delta \quad \textit{and} \quad \Pr_{w\sim Q}[w \in W] \le e^{\varepsilon} \Pr_{w\sim P}[w \in W] + \delta.$$

Definition 2.2 (Differential Privacy [29])**.** *A randomized algorithm $\mathcal{A}\colon \mathcal{X}^n \to \mathcal{W}$ is (ε, δ)-*differentially private *if for all neighboring datasets x, x' we have $\mathcal{A}(x) \approx_{\varepsilon,\delta} \mathcal{A}(x')$.*

We next describe well-known mechanisms which serve as building blocks for our algorithms.

Lemma 2.3 (Laplace Mechanism [29])**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$, data set $x \in \mathcal{X}^n$, and privacy parameter ε. The* Laplace Mechanism *returns $\tilde{f}(x) = f(x) + \mathrm{Lap}(\Delta_f/\varepsilon)$, where $\Delta_f = \max_{x\sim x'} |f(x) - f(x')|$ is the* global sensitivity *of f. The Laplace Mechanism is $(\varepsilon, 0)$-differentially private.*

Lemma 2.4 (Gaussian Mechanism [29])**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$, data set $x \in \mathcal{X}^n$, and privacy parameters ε, δ. The* Gaussian Mechanism *returns $\tilde{f}(x) = f(x) + \mathcal{N}(0, \sigma^2\mathbb{I})$, where $\sigma = \Delta_f\sqrt{2\log(5/4\delta)}/\varepsilon$ and $\Delta_f = \max_{x\sim x'} \|f(x) - f(x')\|_2$ is the* global ℓ_2-sensitivity *of f. The Gaussian Mechanism is (ε, δ)-differentially private.*

3 Tukey-Depth Mechanism

We start with a score function $q(x; y)$ and wish to sample from the exponential mechanism, proportional to $\exp\{\varepsilon q(x; y)/2\}$, but restricting the sampling to the set of points with score at least t. Denote this set $\mathcal{Y}_{t,x} = \{y \in \mathcal{Y} : q(x; y) \ge t\}$ and call the resulting distribution $\mathcal{M}_{\varepsilon,t}(x)$. Unfortunately, sampling directly from $\mathcal{M}_{\varepsilon,t}(x)$ may not be private. To address this, we try to sample only from data

sets that are "safe" with respect to privacy, i.e., have distributions that are indistinguishable from those of their neighbors.

Definition 3.1 (Safety)**.** *Data set x is (ε, δ, t)-safe if, for all $x' \sim x$, we have $\mathcal{M}_{\varepsilon,t}(x) \approx_{\varepsilon,\delta} \mathcal{M}_{\varepsilon,t}(x')$. Let $\texttt{SAFE}_{(\varepsilon,\delta,t)} \subseteq \mathcal{X}^n$ be the set of safe data sets, and let $\texttt{UNSAFE}_{(\varepsilon,\delta,t)} = \mathcal{X}^n \setminus \texttt{SAFE}_{(\varepsilon,\delta,t)}$ be its complement.*

Following the propose-test-release framework of [25], we check if the input is *far* from $\texttt{UNSAFE}_{(\varepsilon,\delta,t)}$. Since the distance check itself is private and indistinguishability *is* the definition of safety, the proof of privacy becomes straightforward. Such an abstract definition, however, does not yield much insight into what safe data sets look like. Below, we show that one can establish safety via a simple condition on the volumes of sets of the form $\mathcal{Y}_{t\pm\eta,x}$ (for certain values of η), which allows us to show that Gaussian data are far from $\texttt{UNSAFE}_{(\varepsilon,\delta,t)}$ with high probability.

Algorithm 1 Restricted Exponential Mechanism $\mathcal{A}^E_{\varepsilon,\delta,t}(x)$

Require: Data space $\mathcal{X}$. Output space $\mathcal{Y} \cup \{\texttt{FAIL}\}$. Data set $x \in \mathcal{X}^n$. Score function $q : \mathcal{X}^n \times \mathcal{Y} \to \mathbb{R}$, with global sensitivity 1 in the first argument. Privacy parameters $\varepsilon, \delta > 0$. Minimum score t.

1: $h \leftarrow D_H(x, \texttt{UNSAFE}_{(\varepsilon,\delta,t)})$

2: **if** $h + z < \frac{\log(1/2\delta)}{\varepsilon}$ for $z \sim \mathrm{Lap}(1/\varepsilon)$ **then return** $\texttt{FAIL}$.

3: **return** $\hat{y} \sim \mathcal{M}_{\varepsilon,t}(x)$, where $\mathcal{M}_{\varepsilon,t}(x) \propto \begin{cases} \exp\left\{\frac{\varepsilon q(x;y)}{2}\right\} & \text{if } y \in \mathcal{Y}_{t,x} \\ 0 & \text{otherwise} \end{cases}$

3.1 Main Algorithm

We estimate the mean by instantiating Algorithm 1 with $t = n/4$ and score function $q(x; y) = nT_x(y)$, where $T_x(y)$ is the (empirical) Tukey depth, defined in Equation (4). Observe that $nT_x(y)$ has sensitivity 1, since for any halfspace the fraction of points it contains can change by at most $\frac{1}{n}$ when we change one data point.

Theorem 3.2 (Privacy and Accuracy of the Tukey-Depth Mechanism)**.** *For any $\varepsilon, \delta > 0$, Algorithm 1 is $(2\varepsilon, e^{\varepsilon}\delta)$-differentially private. There exists an absolute constant C such that, for any $0 < \alpha, \beta, \varepsilon < 1$, $0 < \delta \leq \frac{1}{2}$, mean μ, and positive definite Σ, if $x \sim \mathcal{N}(\mu, \Sigma)^{\otimes n}$ and*

$$n \geq C\left(\frac{d + \log(1/\beta)}{\alpha^2} + \frac{d\log(1/\alpha) + \log(1/\beta)}{\alpha\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right), \tag{5}$$

then with probability at least $1 - 3\beta$, Algorithm 1 with $q(x; y) = nT_x(y)$ returns $\mathcal{A}^E_{\varepsilon,\delta,n/4}(x) = \hat{\mu}$ such that $\|\hat{\mu} - \mu\|_\Sigma \leq \alpha$.

In this submission, we focus on the accuracy analysis. Privacy follows from standard propose-test-release calculations [25], which we include in the supplementary material.

3.2 Accuracy Analysis

The proof of accuracy proceeds in four stages. Using standard analysis, we first relate the empirical Tukey depth of a point y to its Mahalanobis distance $\|y - \mu\|_\Sigma$ via the *expectation* of Tukey depth under $\mathcal{N}(\mu, \Sigma)$. Since Tukey depth is defined as a minimum over halfspaces, which have Vapnik-Chervonekis dimension $d + 1$, one can show via uniform convergence that the empirical measure concentrates around its expectation. This portion ends with a standard lemma relating the sets $\mathcal{Y}_{np,x}$, where $p \in (0, 1/2)$, to ellipsoids defined by Mahalanobis distance.

The remaining three steps, which are new to this work, begin with a characterization of the set $\texttt{SAFE}$ defined above, which provides conditions under which a data set is far from $\texttt{UNSAFE}$. We discuss this in more detail below. The third stage uses that characterization and the tools we developed to show that Gaussian data is typically far from $\texttt{UNSAFE}$, establishing that Algorithm 1 has a small probability of returning $\texttt{FAIL}$. Finally, conditioned on Algorithm 1 not returning $\texttt{FAIL}$, a similar analysis shows that with high probability the restricted sampler $\mathcal{M}_{\varepsilon,t}(x)$ returns a point with high empirical Tukey depth. Together, this yields a bound on the Mahalanobis distance to the true mean.

A Volume Condition for Safety We consider sets of the form $\mathcal{Y}_{t+\eta,x}$ for moderate positive and negative values of η. Recall that as η becomes smaller, the set $\mathcal{Y}_{t+\eta,x}$ grows. We show that, if the volume of $\mathcal{Y}_{t+\eta,x}$ does not increase too quickly as η shrinks, then x is far from every data set in $\texttt{UNSAFE}_{(\varepsilon,\delta,t)}$. In particular, this implies that x itself is in $\texttt{SAFE}_{(\varepsilon,\delta,t)}$. These lemmas do not rely on specific features of Gaussian data or Tukey depth, which enter in only in the last two stages as described above, when we argue about typical data sets.

Before arguing about volumes directly, we sketch the proof of a lemma about the weight assigned to sets by the exponential mechanism. For any set $S \subseteq \mathcal{Y}$, we denote this by $w_x(S) = \int_S \exp\left\{\frac{\varepsilon q(x;y)}{2}\right\} \mathrm{d}y$.

Lemma 3.3. *Assume $\delta < \frac{1}{2}$. If $w_x(\mathcal{Y}_{t+1,x}) \geq (1-\delta) w_x(\mathcal{Y}_{t-1,x})$, then $x \in \texttt{SAFE}_{(\varepsilon,\delta',t)}$ for $\delta' = 4e^{\varepsilon}\delta$.*

Proof Sketch. For an arbitrary adjacent x', we must establish that the distributions $\mathcal{M}_{\varepsilon,t}(x)$ and $\mathcal{M}_{\varepsilon,t}(x')$ are (ε, δ')-indistinguishable. The proof is elementary but non-trivial: our hypothesis only concerns x, not its neighbors. To illustrate, we show that $\Pr[\mathcal{M}_{\varepsilon,t}(x) \notin \mathrm{supp}(\mathcal{M}_{\varepsilon,t}(x'))] \leq \delta'$. By construction, $\mathrm{supp}(\mathcal{M}_{\varepsilon,t}(x')) = \mathcal{Y}_{t,x'}$. Now we move from x' to x: the score function has sensitivity 1, so every point with $q(x;y) \geq t+1$ satisfies $q(x';y) \geq t$. Thus $\mathcal{Y}_{t+1,x} \subseteq \mathcal{Y}_{t,x'}$, and we have

$$\Pr[\mathcal{M}_{\varepsilon,t}(x) \notin \mathrm{supp}(\mathcal{M}_{\varepsilon,t}(x'))] \leq \Pr[\mathcal{M}_{\varepsilon,t}(x) \notin \mathcal{Y}_{t+1,x}] = 1 - \frac{w_x\left(\mathcal{Y}_{t+1,x}\right)}{w_x\left(\mathcal{Y}_{t,x}\right)}.$$

Since $\mathcal{Y}_{t,x} \subseteq \mathcal{Y}_{t-1,x}$, $w_x(\mathcal{Y}_{t,x}) \leq w_x(\mathcal{Y}_{t-1,x})$ and our hypothesis implies $\Pr[\mathcal{M}_{\varepsilon,t}(x) \notin \mathrm{supp}(\mathcal{M}_{\varepsilon,t}(x'))] \leq \delta$, which is less than $\delta' = 4e^{\varepsilon}\delta$.

The rest of the proof requires similar (but slightly longer) calculations. □

We now use this lemma to establish when a data set is far from the set of unsafe data sets. Note that setting $k = 0$ below implies for all $z \in \texttt{UNSAFE}$ that we have $D_H(x,z) > 0$, i.e. $x \neq z$ and $x \in \texttt{SAFE}$.

Lemma 3.4. *For any $k \geq 0$, if there exists a $g > 0$ such that $\frac{\mathrm{Vol}(\mathcal{Y}_{t-k-1,x})}{\mathrm{Vol}(\mathcal{Y}_{t+k+g+1,x})} \cdot e^{-\varepsilon g/2} \leq \delta$, then for all $z \in \texttt{UNSAFE}_{(\varepsilon,\delta',t)}$, with $\delta' = 4e^{\varepsilon}\delta$, we have $D_H(x,z) > k$*

Proof. Take some z at distance at most k from x (if $k = 0$, set $z \leftarrow x$). We show $z \in \texttt{SAFE}_{(\varepsilon,\delta',t)}$. We have, from Lemma 3.3, that if $\frac{w_z(\mathcal{Y}_{t+1,z})}{w_z(\mathcal{Y}_{t-1,z})} \geq 1-\delta$, then z is safe. This assumption is equivalent to $\frac{w_z(\mathcal{Y}_{t-1,z} \setminus \mathcal{Y}_{t+1,z})}{w_z(\mathcal{Y}_{t-1,z})} \leq \delta$, which is the form we use.

First we lower bound the denominator:

$$w_z(\mathcal{Y}_{t-1,z}) \geq w_z(\mathcal{Y}_{t+g+1,z}) \geq \mathrm{Vol}(\mathcal{Y}_{t+g+1,z}) e^{\varepsilon(t+g+1)/2} \geq \mathrm{Vol}(\mathcal{Y}_{t+g+k+1,x}) e^{\varepsilon(t+g+1)/2},$$

where (crucially) the last inequality switches to the volume under x, and we have used the sensitivity of q. We use the same idea on the numerator, switching to x in the first inequality:

$$w_z(\mathcal{Y}_{t-1,z} \setminus \mathcal{Y}_{t+1,z}) \leq w_z(\mathcal{Y}_{t-k-1,x} \setminus \mathcal{Y}_{t+1,z}) \leq \mathrm{Vol}(\mathcal{Y}_{t-k-1,x}) e^{\varepsilon(t+1)/2}.$$

With an upper bound on the numerator, a lower bound on the denominator, and the fact that $\frac{e^{\varepsilon(t+1)/2}}{e^{\varepsilon(t+g+1)/2}} = e^{-\varepsilon g/2}$, we have

$$\frac{w_z(\mathcal{Y}_{t-1,z} \setminus \mathcal{Y}_{t+1,z})}{w_z(\mathcal{Y}_{t-1,z})} \leq \frac{\mathrm{Vol}(\mathcal{Y}_{t-k-1,x})}{\mathrm{Vol}(\mathcal{Y}_{t+k+g+1,x})} \cdot e^{-\varepsilon g/2} \leq \delta,$$

so $z \in \texttt{SAFE}_{(\varepsilon,\delta',t)}$. □

With these tools in hand, in the supplementary material we go on to show that, if $n = \omega(d/\varepsilon)$, typical Gaussian data satisfy the hypothesis of Lemma 3.4 for $g = n/8$ and thus are far from unsafe (implying that we are unlikely to return `FAIL`) and that with high probability the restricted sampler $\mathcal{M}_{\varepsilon,t}(x)$ returns a point with high Tukey depth.

4 Empirically Rescaled Gaussian Mechanism

In this section, we describe our second estimator. At a high level, the estimator first privately checks whether the data set x is $\frac{1}{\varepsilon}$-close in Hamming distance to a good set of "roughly Gaussian" data sets $\mathcal{G}(\lambda)$. If so, it finds the closest data set to x that belongs in $\mathcal{G}(\lambda)$, which we call $\tilde{x}$. Then it returns a sample $\hat{\mu}$ drawn from $\mathcal{N}(\mu_{\tilde{x}}, C^2\Sigma_{\tilde{x}})$, where $\mu_{\tilde{x}}$ and $\Sigma_{\tilde{x}}$ are the empirical mean and covariance of $\tilde{x}$ and C is a scale parameter appropriately set to ensure privacy. Specifically, we use the definitions:

Definition 4.1 (Empirical Mean and Covariance)**.** *For data set $x \in \mathbb{R}^{3n\times d}$, the empirical mean and covariance of x are respectively defined by $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_{i+2n}$ and $\Sigma_x = \frac{1}{2n}\sum_{i=1}^{n}(x_i - x_{i+n})(x_i - x_{i+n})^T$.*

In comparison with the standard empirical estimators, ours enable a simpler privacy analysis, since replacing one datapoint in x only affects one of the sums. For convenience, we choose the number of samples to be $3n$ so that we can pair the first two thirds to construct Σ_x and use the last third for μ_x. With these definitions at hand, the good set $\mathcal{G}(\lambda)$ is defined as follows:

Definition 4.2 (λ-goodness)**.** *For any $\lambda > 0$, define $\mathcal{G}(\lambda) \subseteq \mathbb{R}^{3n\times d}$ as*

$$\mathcal{G}(\lambda) \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^{3n\times d} : \Sigma_x \text{ is invertible and } \forall i \in [3n] \;\; \|x_i - \mu_x\|_{\Sigma_x}^2 \le \lambda\right\}.$$

We set $\lambda \approx d\log n$, since for this value (sub)Gaussian data will belong in $\mathcal{G}(\lambda)$ with high probability.

Finally, we note that the algorithm immediately aborts if the number of samples is less than $\frac{k\lambda\log(1/\delta)}{\varepsilon} \approx \frac{d\log(1/\delta)}{\varepsilon^2}$, a condition necessary to ensure privacy.

Algorithm 2 Empirically Rescaled Gaussian Mechanism $\mathcal{A}^G_{\varepsilon,\delta,\beta}(x)$

Require: Data set $x = (x_1, \ldots, x_{3n})^T \in \mathbb{R}^{3n\times d}$. Privacy parameters $\varepsilon, \delta > 0$. Failure probability β.

1: Initialize: $\lambda \leftarrow O\left(d\log\frac{n}{\beta}\right)$, $t \leftarrow \frac{1}{\varepsilon}\log\frac{1}{\beta}$, $k \leftarrow \frac{2}{\varepsilon}\log\frac{1}{\delta\beta} + 1$, $C^2 \leftarrow \frac{32k^2}{\varepsilon^2 n^2}\cdot\frac{\lambda}{1-2k\lambda/n}\cdot\log\frac{1.25}{\delta}$.

2: **if** $n = o\left(\frac{k\lambda}{\varepsilon}\log\frac{1}{\delta}\right)$ **then return** `FAIL`.

3: $\bar{x} \leftarrow \sigma(x)$ ▷ random permutation $\sigma : (\mathbb{R}^d)^{3n} \to (\mathbb{R}^d)^{3n}$

4: $h \leftarrow D_H(\bar{x}, \mathcal{G}(\lambda))$ ▷ distance between x and λ-goodness

5: **if** $h + r > t$ for $r \sim \mathrm{Lap}\,(1/\varepsilon)$ **then return** `FAIL`.

6: $\tilde{x} \leftarrow \arg\min_{z\in\mathcal{G}(\lambda)} D_H(\bar{x}, z)$ ▷ projection to λ-goodness

7: **return** $\hat{\mu} \sim \mathcal{N}(\mu_{\tilde{x}}, C^2\Sigma_{\tilde{x}})$

We remark that the sample size check in line 2 and the setting of λ are not well-defined as they are stated with asymptotic notation. Although it is possible to compute the constants for these steps, we exclude them in favor of a cleaner analysis.

We prove accuracy for the following family of distributions:

Definition 4.3 (Subgaussian distribution)**.** *Let $P_{\mu,\Sigma}$ be a distribution over $\mathbb{R}^d$ with mean μ and covariance $\Sigma \succ 0$. For a constant $c > 0$, we say that $P_{\mu,\Sigma}$ is subgaussian with parameter $c\Sigma$, if for all $u \in \mathbb{R}^d$ such that $\|u\|_2 = 1$, $\mathbb{E}_{v\sim P_{\mu,\Sigma}}[e^{\lambda u^T(v-\mu)}] \le e^{c\lambda^2(u^T\Sigma u)/2}$ for all $\lambda \in \mathbb{R}$.*

We write $P_{\mu,\Sigma} \in \mathrm{subG}(c\Sigma)$. The definition appears in relevant literature [46, 22] and intuitively asks that the distribution concentrates "at least as well" as a Gaussian along every univariate projection.

We now state the guarantees of Algorithm 2 and defer their proof to the supplementary material.

Theorem 4.4 (Privacy and Accuracy of the Empirically Rescaled Gaussian Mechanism)**.** *For any $\varepsilon > 0$, $0 < \delta < 1$, Algorithm 2 is $(3\varepsilon, (3 + e^{\varepsilon})\delta)$-differentially private. There exists an absolute constant C such that, for any $0 < \alpha, \beta, \varepsilon, \delta < 1$, mean μ, and positive definite Σ, if $x \sim P^{\otimes n}_{\mu,\Sigma}$, where $P_{\mu,\Sigma} \in \mathrm{subG}(c\Sigma)$ for some constant $c > 0$, and*

$$n \ge C\left(\frac{d}{\alpha^2}\log\frac{1}{\beta} + \frac{d}{\alpha\varepsilon^2}\log^3\frac{1}{\delta\beta}\cdot\log\frac{d\log(1/\delta\beta)}{\alpha\varepsilon}\right), \tag{6}$$

then with probability at least $1 - 3\beta$, Algorithm 2 returns $\mathcal{A}^G_{\varepsilon,\delta,\beta}(x) = \hat{\mu}$ such that $\|\hat{\mu} - \mu\|_\Sigma \le \alpha$.

References

[1] John M Abowd. The US Census Bureau adopts differential privacy. In *ACM International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2867–2867, 2018.

[2] 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 2021)*, ALT '21. JMLR, Inc., March 2021. URL `https://arxiv.org/abs/2010.09929`.

[3] Apple Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8), 2017. `https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf`.

[4] Rina Barber and John Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arxiv preprint arXiv:1412.4451*, 12 2014.

[5] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. COINPRESS: Practical Private Mean and Covariance Estimation. In *Advances in Neural Information Processing Systems*, NeurIPS '20, 2020. `https://arxiv.org/abs/2006.06618`.

[6] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. PROCHLO: Strong privacy for analytics in the crowd. In *ACM Symposium on Operating Systems Principles*, SOSP '17, pages 441–459, Shanghai, China, 2017. `https://arxiv.org/abs/1710.00901`.

[7] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? *arXiv*, 2012.06421, 2021.

[8] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, TCC '16, pages 635–658, Beijing, China, 2016. `https://arxiv.org/abs/1605.02065`.

[9] Mark Bun and Thomas Steinke. Smooth sensitivity, revisited. Manuscript, 2019.

[10] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014. `https://arxiv.org/abs/1311.3158`.

[11] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems*, NeurIPS '19, pages 156–167, Vancouver, Canada, 2019. `https://arxiv.org/abs/1905.13229`.

[12] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

[13] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In {*USENIX*} *Security Symposium*, ({USENIX} '19), pages 267–284, 2019. `https://arxiv.org/abs/1802.08232`.

[14] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

[15] Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimates. *The Annals of Statistics*, 24(6):2499 – 2512, 1996. doi: 10.1214/aos/1032181164. URL `https://doi.org/10.1214/aos/1032181164`.

[16] Luc Devroye and Gábor Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *The Annals of Statistics*, 25(6):2626 – 2637, 1997. doi: 10.1214/aos/1030741088. URL `https://doi.org/10.1214/aos/1030741088`.

[17] Luc Devroye and Gábor Lugosi. Combinatorial methods in density estimation. In *Springer series in Statistics*, 2001.

[18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE Annual Symposium on Foundations of Computer Science*, FOCS '16, pages 655–664. IEEE, 2016. https://arxiv.org/abs/1604.06443.

[19] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, ICML '17, pages 999–1008, 2017. https://arxiv.org/abs/1703.00893.

[20] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2018)*, SODA '18, Philadelphia, PA, USA, January 2018. SIAM. URL https://arxiv.org/abs/1704.03866.

[21] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems*, PODS '03, pages 202–210. ACM, 2003.

[22] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/a4d92e2cd541fca87e4620aba658316d-Paper.pdf.

[23] Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. Differentially private confidence intervals. *arxiv preprint arXiv:2001.02285*, 2020.

[24] David Durfee and Ryan M Rogers. Practical differentially private top-k selection with pay-what-you-get composition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b139e104214a08ae3f2ebcce149cdf6e-Paper.pdf.

[25] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st ACM Symposium on Theory of Computing*, STOC '09, pages 371–380. ACM, 2009.

[26] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016. https://arxiv.org/abs/1603.01887.

[27] Cynthia Dwork and Sergey Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In *Annual International Cryptology Conference*, pages 469–480. Springer, 2008.

[28] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *International Conference on the Theory and Applications of Cryptographic Techniques*, EUROCRYPT '06, pages 486–503, St. Petersburg, Russia, 2006.

[29] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Conference on Theory of Cryptography*, TCC '06, pages 265–284, New York, NY, USA, 2006.

[30] Cynthia Dwork, Frank McSherry, and Kunal Talwar. The price of privacy and the limits of LP decoding. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 85–94. ACM, 2007.

[31] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 11–20, New York, NY, 2014. ACM.

[32] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *IEEE Symposium on Foundations of Computer Science*, FOCS '15, 2015.

[33] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *ACM Conference on Computer and Communications Security*, CCS '14, 2014.

[34] Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing*, STOC '20, pages 954–959, 2020. `https://arxiv.org/abs/1906.05271`.

[35] Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thao Nguyen. Robust and private learning of halfspaces. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1603–1611. PMLR, 13–15 Apr 2021.

[36] Samuel Haney, Ashwin Machanavajjhala, John M Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. Utility cost of formal privacy for releasing national employer-employee statistics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 1339–1354, Chicago, IL, 2017. ACM.

[37] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[38] Samuel B Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *arxiv preprint arXiv:2007.15839*, 2020.

[39] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[40] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high dimensional distributions. In *Annual Conference on Learning Theory*, COLT '19. JMLR.org, 2019.

[41] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235. PMLR, 09–12 Jul 2020. URL `http://proceedings.mlr.press/v125/kamath20a.html`.

[42] Haim Kaplan, Micha Sharir, and Uri Stemmer. How to find a point in the convex hull privately. In Sergio Cabello and Danny Z. Chen, editors, *36th International Symposium on Computational Geometry, SoCG 2020, June 23-26, 2020, Zürich, Switzerland*, volume 164 of *LIPIcs*, pages 52:1–52:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.SoCG.2020.52. URL `https://doi.org/10.4230/LIPIcs.SoCG.2020.52`.

[43] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[44] Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam Smith, and Jonathan Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, STOC '10, pages 775–784. ACM, 2010.

[45] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *arXiv preprint arXiv:2102.09159*, 2021.

[46] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794, 2019. doi: 10.1214/17-AOS1639. URL `https://doi.org/10.1214/17-AOS1639`.

[47] Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[48] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Las Vegas, NV, USA, 2007.

[49] Ryan Rogers, Subbu Subramaniam, Sean Peng, David Durfee, Seunghyun Lee, Santosh Kumar Kancha, Shraddha Sahay, and Parvez Ahammad. Linkedin's audience engagements api: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839*, 2020.

[50] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[51] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P), Oakland*, 2017.

[52] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

[53] David Tastuggine and Ilya Mironov. Introducing Opacus: A high-speed library for training PyTorch models with differential privacy. Facebook AI Blog, 2020. `https://ai.facebook.com/blog/introducing-opacus-a-high-speed-library-for-training-pytorch-models-with-differential-privacy/`.

[54] John D. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 448–485, 1960.

[55] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution. *Proceedings on Privacy Enhancing Technologies*, 2020(2):230–250, 2020. `https://arxiv.org/abs/1909.01917`.

[56] Yannis G. Yatracos. Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy. *The Annals of Statistics*, 13(2):768 – 774, 1985. doi: 10.1214/aos/1176349553. URL `https://doi.org/10.1214/aos/1176349553`.

[57] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, CSF '18, pages 268–282, 2018. `https://arxiv.org/abs/1709.01604`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In addition to the discussion of the assumptions in our theorems, see the "Limitations" paragraph in the introduction.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see the "Societal Impact" paragraph in the introduction.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Our formal theorem statements, Theorems 3.2 and 4.4 explicitly specify our assumptions.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are contained in either the submission or supplementary materials.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]