Fast, Sample-Efficient, Affine-Invariant Private Mean and Covariance Estimation for Subgaussian Distributions

Gavin Brown[*] GRBROWN@BU.EDU
Samuel B. Hopkins[†] SAMHOP@MIT.EDU
Adam Smith[*] ADS22@BU.EDU

Abstract

We present a fast, differentially private algorithm for high-dimensional *covariance-aware* mean estimation with nearly optimal sample complexity. Only exponential-time estimators were previously known to achieve this guarantee. Given n samples from a (sub-)Gaussian distribution with unknown mean μ and covariance Σ, our (ε, δ)-differentially private estimator produces $\tilde{\mu}$ such that $\|\mu - \tilde{\mu}\|_\Sigma \leq \alpha$ as long as $n \gtrsim \frac{d}{\alpha^2} + \frac{d\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d \log 1/\delta}{\varepsilon}$. The Mahalanobis error metric $\|\mu - \hat{\mu}\|_\Sigma$ measures the distance between $\hat{\mu}$ and μ relative to Σ; it characterizes the error of the sample mean. Our algorithm runs in time $\tilde{O}(nd^{\omega-1} + nd/\varepsilon)$, where $\omega < 2.38$ is the matrix multiplication exponent.

We adapt an exponential-time approach of Brown, Gaboardi, Smith, Ullman, and Zakynthinou (2021), giving efficient variants of stable mean and covariance estimation subroutines that also improve the sample complexity to the nearly optimal bound above.

Our stable covariance estimator can be turned to private covariance estimation for unrestricted subgaussian distributions. With $n \gtrsim d^{3/2}$ samples, our estimate is accurate in spectral norm. This is the first such algorithm using $n = o(d^2)$ samples, answering an open question posed by Alabi et al. (2022). With $n \gtrsim d^2$ samples, our estimate is accurate in Frobenius norm. This leads to a fast, nearly optimal algorithm for private learning of unrestricted Gaussian distributions in TV distance.

Duchi, Haque, and Kuditipudi (2023) obtained similar results independently and concurrently.

Contents

[*] Department of Computer Science, Boston University. Supported by NSF Awards CNS-2120667 and CCF-1763786 and an Apple Faculty Award.

[†] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. Supported in part by funds from the MLA@CSAIL initiative and by NSF Award CCF-2238080.

1. Introduction

We consider the statistical task of estimating the mean of a high-dimensional subgaussian distribution from independent samples under the constraint of differential privacy. Differential privacy allows algorithm designers to control and reason about privacy loss in statistics and machine learning and is the standard approach for protecting the privacy of personal data, with adoption in academia, industry, and the United States government. We focus on the following *covariance-aware* version of the mean estimation problem.

Problem 1 (Covariance-Aware Mean Estimation) *Given n samples from a subgaussian distribution D on $\mathbb{R}^d$ with unknown mean μ and covariance Σ, output $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_\Sigma \leq \alpha$, where $\|\hat{\mu} - \mu\|_\Sigma$ denotes $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|_2$.*

The *Mahalanobis distance* $\|\hat{\mu} - \mu\|_\Sigma$ is the natural affine-invariant way to measure the statistical accuracy of an estimator for the mean. An estimator with small Mahalanobis error automatically adapts to directions of low uncertainty in the location of μ—in any particular direction, the estimator's error is scaled to the variance in that direction. The family of subgaussian distributions generalizes the Gaussian distribution; we defer definitions for now.

Absent privacy constraints, the empirical mean already achieves the best possible guarantees for Problem 1, requiring $n \gtrsim d/\alpha^2$ samples. (Here $\gtrsim$ hides constants.) However, it does not generally protect its input data—for example, an adversary with a rough idea of the distribution could easily tell from observing the empirical mean if the data set contained an extreme outlier in any particular direction. We thus aim for estimators that are differentially private:

Definition 2 (Differential Privacy (DP)) *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets. A (randomized) algorithm $A : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-DP if for every $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $x' = (x'_1, \ldots, x'_n) \in \mathcal{X}^n$ such that x, x' agree on all but one coordinate, $A(x)$ and $A(x')$ are (ε, δ)-indistinguishable (denoted $A(x) \approx_{(\varepsilon,\delta)} A(x')$); that is, for any event $Y \subseteq \mathcal{Y}$,*

$$\Pr[A(x) \in Y] \leq e^\varepsilon \Pr[A(x') \in Y] + \delta\,.$$

A differentially private estimator provides a strong guarantee: no matter what an outside adversary knows about the data set ahead of time, they will learn roughly the same things about any particular individual Alice *regardless of whether Alice's data is used in the computation* (as in, e.g., Kasiviswanathan and Smith (2014)). We call data sets that differ in one entry *adjacent*.

Until recently, covariance-aware mean estimation subject to differential privacy was assumed by many researchers to require at least as many samples as privately estimating the *covariance*—while the latter task also takes $O(d)$ samples nonprivately, $\Omega(d^{3/2})$ are required for private estimators (Kamath et al. (2022)).

Brown, Gaboardi, Smith, Ullman, and Zakynthinou (2021) (henceforth, "BGSUZ") disproved this assumption by giving a modification of the exponential mechanism that solves Problem 1 (for Gaussian data) as long as

$$n \gtrsim \underbrace{\frac{d}{\alpha^2}}_{\substack{\text{nonprivate}\\ \text{sample}\\ \text{complexity}}} + \frac{d}{\alpha\varepsilon} + \frac{\log 1/\delta}{\varepsilon} \tag{1}$$

This expression hides an absolute constant and an additive term of $\log(1/\varepsilon\alpha)/(\varepsilon\alpha)$. It is nearly tight, even when the covariance is known exactly: the first term in Equation (1) is necessary for nonprivate estimation, the second nearly matches the $\Omega(d/(\alpha\varepsilon\log(d))$ lower bound of Kamath et al. (2019), and the third is required even for estimating the mean of a one-dimensional Gaussian with unit variance (but unrestricted mean). Subsequent work by Liu, Kong, and Oh (2022) extended the exponential-mechanism-based approach to many statistical tasks; in particular, they solve Problem 1 for general subgaussian distributions with the same near-optimal sample complexity. Unfortunately, all these estimators appear to require at least exponential time to compute.

Our main result is a polynomial-time algorithm for Problem 1 with sample complexity also depending linearly on d:

Theorem 3 (Informal, see Theorem 8) *Algorithm 1 is (ε, δ)-differentially private. Given*

$$n \gtrsim \frac{d}{\alpha^2} + \frac{d\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d\log 1/\delta}{\varepsilon}$$

samples from a subgaussian distribution on $\mathbb{R}^d$ with mean μ and covariance Σ, with high probability it outputs $\tilde{\mu}$ such that $\|\tilde{\mu} - \mu\|_\Sigma \leq \alpha$. It runs in time $\tilde{O}\big(nd^{\omega-1} + nd/\varepsilon\big)$, where $\omega < 2.38$ is the matrix multiplication exponent.

When d is large relative to $1/\varepsilon$, the running time of our algorithm is dominated by the time to compute the covariance matrix, $nd^{\omega-1}$. The terms $d\sqrt{\log 1/\delta}/(\alpha\varepsilon)$ and $d\log(1/\delta)/\varepsilon$ in the sample complexity of our algorithm are slightly suboptimal; however, the nonprivate sample complexity continues to dominate for modest privacy parameters.

As we discuss below, our main algorithm relies on a new nonprivate covariance estimator `StableCovariance`. This estimator, combined with the "Gaussian sampling mechanism" of Alabi et al. (2022), yields a fast algorithm for private covariance estimation that provides strong error guarantees in both spectral and Frobenius norm. It is the first private algorithm achieving low spectral-norm error for unrestricted subgaussian distributions with $n = o(d^2)$ samples. It also recovers guarantees similar to those of Ashtiani and Liaw (2021) for estimating in Frobenius norm using $O(d^2)$ samples; such an error guarantee, in combination with our mean estimation guarantee, is known to suffice for learning unrestricted Gaussians distributions to low total variation distance.

Theorem 4 (Informal, see Theorem 42) *Algorithm 7 is (ε, δ)-differentially private and returns a vector $\tilde{\mu}$ and matrix $\tilde{\Sigma}$. Suppose it receives n samples drawn i.i.d. from a subgaussian distribution with mean μ and covariance Σ over $\mathbb{R}^d$.*

- *If $n \gtrsim \frac{d}{\alpha^2} + \frac{d^{3/2}\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d\log 1/\delta}{\varepsilon}$, then, with high probability, $\left\|\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} - \mathbb{I}\right\|_2 \leq \alpha$.*
- *If $n \gtrsim \frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d\log 1/\delta}{\varepsilon}$, then, with high probability, $\left\|\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} - \mathbb{I}\right\|_F \leq \alpha$; furthermore, if the distribution is Gaussian, then $\mathrm{TV}(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\tilde{\mu},\tilde{\Sigma})) = O(\alpha)$.*

It runs in time $\tilde{O}(nd^{\omega-1} + nd/\varepsilon)$, where $\omega < 2.38$ is the matrix multiplication exponent.

Throughout, our accuracy analyses assume the underlying distribution has a full-rank covariance matrix. There exist sample- and time-efficient preprocessing algorithms to deal with distributions where this assumption fails (Singhal and Steinke (2021); Kamath et al. (2021); Ashtiani and Liaw (2021)). Our privacy analysis makes no assumptions on the data.

In the remainder of this introduction we sketch our techniques and overview related work.

1.1. Our Techniques

Background: The Empirical Covariance Approach In addition to their modification of the exponential mechanism, BGSUZ give a second mean estimation algorithm, also requiring exponential time, which works for subgaussian distributions but has a worse dependence on the privacy parameters. This second algorithm is the basis for our efficient construction.

The basic idea of this algorithm is to first solve the problem for "good" data sets y, where goodness means a collection of conditions that are typical for subgaussian data. For example, a good data set y should have no significant outliers as measured by $\|(\cdot) - \mu_y\|_{\Sigma_y}$, where μ_y and Σ_y are the empirical mean and covariance matrix of y. BGSUZ show that, if we can restrict our inputs to good data sets, the mechanism that releases a single draw from $\mathcal{N}(\mu_y, \sigma^2\Sigma_y)$, for $\sigma \approx \frac{\sqrt{d}}{\varepsilon n}$, is in fact differentially private. The data-dependent choice of covariance Σ_y is crucial to achieving the Mahalanobis-distance guarantee, because it ensures that the noise added for privacy remains small in directions where y itself has small variance.

Given a data set x, BGSUZ's algorithm first projects x to the nearest "good" data set y, then checks that y is not too far from x and finally, if that test passes, releases $\tilde{\mu} \sim \mathcal{N}(\mu_y, \sigma^2\Sigma_y)$. The projection step ensures that privacy holds for all data sets.

Although the noise addition is computationally efficient, the BGSUZ algorithm requires brute-force search over data sets for the projection step. It also comes with an additional cost in sample complexity—it roughly requires $n \gtrsim \frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon^2}$. The extra factor of $\frac{1}{\varepsilon}$ compared to Equation (1) comes from the fact that the projection is not *stable*: it is possible for two neighboring data sets x and x' to be projected to "good" data sets y and y' that are far apart.

Overview of Our Approach In this work, we overcome the key computational bottleneck of the empirical covariance approach; along the way, we recover the optimal dependency on ε.

Our main technical tools are new efficient, *non*-private estimators $\texttt{StableCovariance}(x)$ and $\texttt{StableMean}(x)$. We also design an algorithm $\texttt{SCORE}(x)$ which approximates the Hamming distance from x to the nearest "good" dataset. We show that $|\texttt{SCORE}(x) - \texttt{SCORE}(x')| \leq 2$ for adjacent x, x' and that if $\texttt{SCORE}(x)$ is small—as it is with high probability for subgaussian x—then $\texttt{StableCovariance}(x) \approx \texttt{StableCovariance}(x')$ and $\texttt{StableMean}(x) \approx \texttt{StableMean}(x')$.

At a high level, on input x, our algorithm first privately tests if $\texttt{SCORE}(x)$ less than a threshold, roughly $1/\varepsilon$. If the test fails, we stop and return nothing (as in the "PTR" framework of Dwork and Lei (2009)). If the test passes, we output a sample from $\mathcal{N}(\texttt{StableMean}(x),\ \sigma^2 \cdot \texttt{StableCovariance}(x))$, where $\sigma \approx \frac{\sqrt{d}}{\varepsilon n}$ is the scale parameter as before.

$\texttt{StableMean}(x)$ and $\texttt{StableCovariance}(x)$ stand in for the empirical parameters μ_y and Σ_y of the projection y from BGSUZ's algorithm. The advantages are computational tractability and stability: `StableMean` and `StableCovariance` are more stable than their counterparts in BGSUZ, leading to our optimal dependence on ε. We establish the following key properties of `StableCovariance` (the ideas for `StableMean` are similar):

- *Accuracy:* When the x_i are drawn i.i.d. from a subgaussian distribution, with high probability we have $\texttt{StableCovariance}(x) = \Sigma_x$, where Σ_x is the empirical covariance of x.[1]

1. Actually, the "paired" empirical covariance $\frac{2}{n}\sum_{i \leq n/2}(x_i - x_{i+n/2})(x_i - x_{i+n/2})^T$.

- *Stability:* If x is a data set with $\texttt{SCORE}(x) \lesssim 1/\varepsilon$, then for any adjacent x', the matrices $\Sigma_1 = \texttt{StableCovariance}(x)$ and $\Sigma_2 = \texttt{StableCovariance}(x')$ are close. Specifically,

$$\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}_d\right\|_{\mathrm{tr}}, \left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \mathbb{I}_d\right\|_{\mathrm{tr}} \leq O(d/n),$$

  where $\|\cdot\|_{\mathrm{tr}}$ denotes the sum of absolute values of eigenvalues.

- *Efficiency:* `StableCovariance` runs in polynomial time.

Stable Covariance We now sketch `StableCovariance`. Assume for now we have a data set x that we believe to be drawn from $\mathcal{N}(0, \Sigma)$ for some unknown Σ. (We reduce to zero-mean case via a standard sample-pairing trick, see Equation (3).) Ultimately, $\texttt{StableCovariance}(x)$ produces a weight vector $w \in [0, 1/n]^n$ and outputs $\Sigma_w = \sum_{i\in[n]} w_i \cdot x_i x_i^T$. We will show the weights satisfy the following conditions.

(a) *Uniform on Subgaussian Inputs:* When the x_i are drawn i.i.d. from a subgaussian distribution, we have $w = (1/n, \ldots, 1/n)$ with high probability.

(b) λ-*Good Weighting:* For any $i \in [n]$ such that $w_i > 0$, we have $\left\|\Sigma_w^{-1/2} x_i\right\|_2^2 \leq \lambda$.

(c) *Stability:* If x is a data set with $\texttt{SCORE}(x) \lesssim 1/\varepsilon$, then for any adjacent x', the associated weight vectors w, w' have $\|w - w'\|_1 \leq O(1/n)$.

Here $\lambda > 0$ is a hyperparameter, which we will eventually set to roughly d. Achieving Property (a) implies our accuracy claim about `StableCovariance`, since setting all weights to $1/n$ yields exactly the second moment matrix of x. Since we have assumed x is zero-mean, this serves in place of the empirical covariance.

Properties (b) and (c) together imply the stability of `StableCovariance`'s parameter estimates. For intuition, consider the case where w and w' differ on exactly one value: suppose $w'_i = w_i + \eta_i$ for $\eta_i > 0$. Further suppose both weights are nonzero and that x and x' agree on index i. Thus we have $\Sigma_{w'} = \Sigma_w + \eta_i \cdot x_i x_i^T$.

$$\begin{aligned}\Sigma_w^{-1/2}\Sigma_{w'}\Sigma_w^{-1/2} - \mathbb{I} &= \Sigma_w^{-1/2}\big(\Sigma_w + \eta_i \cdot x_i x_i^T\big)\Sigma_w^{-1/2} - \mathbb{I} \\ &= \eta_i \cdot \Sigma_w^{-1/2} x_i x_i^T \Sigma_w^{-1/2}.\end{aligned}$$

This is a rank-one matrix whose non-zero eigenvalue is exactly $\eta_i \|x_i\|_{\Sigma_w}^2$, which is at most $\eta_i\lambda$ by Property (b).

The remainder of the stability proof is relatively straightforward: in general $\Sigma_w - \Sigma_{w'}$ is the sum of rank-one matrices whose coefficients η_i satisfy $\sum_i |\eta_i| = \|w - w'\|_1$. The trace norm is the sum of the absolute values of the eigenvalues, so we arrive at an upper bound of roughly $\lambda \|w - w'\|_1 \approx d/n$. The proof requires some more work to account for the one index i^* where $x_{i^*} \neq x'_{i^*}$, as well as accommodating indices where w_i is zero and w'_i is not (in this case $\|x_i\|_{\Sigma_w}^2$ is not bounded by assumption).

Good Subsets We now discuss how `StableCovariance` computes the weight vector w.

Definition 5 *A special case of a λ-good weighting is a λ-good* subset. *A subset $S \subseteq [n]$ is λ-good if its associated weight vector w_S is λ-good, where $(w_S)_i = 1/n$ for $i \in S$ and 0 otherwise.*

`StableCovariance` computes a nested sequence of good subsets $S_0 \subseteq S_1 \subseteq \ldots \subseteq S_{2k}$, where S_ℓ is λ_ℓ good for some sequence of outlier thresholds $\lambda_1 \leq \ldots \leq \lambda_{2k}$ and k is a parameter to set set later. Then it averages a subset of the associated weight vectors to produce $w = \frac{1}{k}\sum_{\ell=k+1}^{2k} w_{S_\ell}$. The subsets it chooses to average are described in the following lemma:

Lemma 6 (Informal) *Let x be a dataset and $\lambda > 0$ be an outlier threshold. There is a unique largest λ-good subset $S \subseteq [n]$ for x which contains all other λ-good subsets. Furthermore, this largest good subset can be found using a greedy algorithm,* `LargestGoodSubset`.

`StableCovariance` takes S_ℓ to be the `LargestGoodSubset` for outlier threshold λ_ℓ. By Lemma 6, this implies immediately that $S_0 \subseteq \ldots \subseteq S_{2k}$, because S_ℓ is $\lambda_{\ell+1}$-good for all ℓ.

For reasons we will see below, we choose the outlier thresholds according to $\lambda_{\ell+1} = e^\varepsilon \lambda_\ell$ and set $k \approx 1/\varepsilon$, so that $\lambda_{2k} \leq O(\lambda_0)$. We now have a nearly complete description of `StableCovariance`, and enough information to see why the weights it computes satisfy properties (a) and (b) above. We will ensure that if $[n]$ itself is a good subset, then $S_0 = \ldots = S_{2k} = [n]$, leading to Property (a) above. Because the subsets S_ℓ are all $O(\lambda_0)$-good, it will not be hard to show that w is itself $O(\lambda_0)$-good, leading to Property (b) above.

Now we turn to Property (c), which is contingent on x, x' having $\texttt{SCORE}(x), \texttt{SCORE}(x') \lesssim 1/\varepsilon$. To proceed, we need to define $\texttt{SCORE}(x)$.

Score Function We have computed S_ℓ, the largest λ_ℓ-good subsets of x, for $\ell = 0, 1, \ldots, 2k$. While $\texttt{StableCovariance}(x)$ averages over the items from $\ell = k+1, \ldots, 2k$, $\texttt{SCORE}(x)$ uses the first half of the sequence:

$$\texttt{SCORE}(x) \stackrel{\text{def}}{=} \min_{0 \leq \ell \leq k} \left(n - |S_\ell| + \ell\right).$$

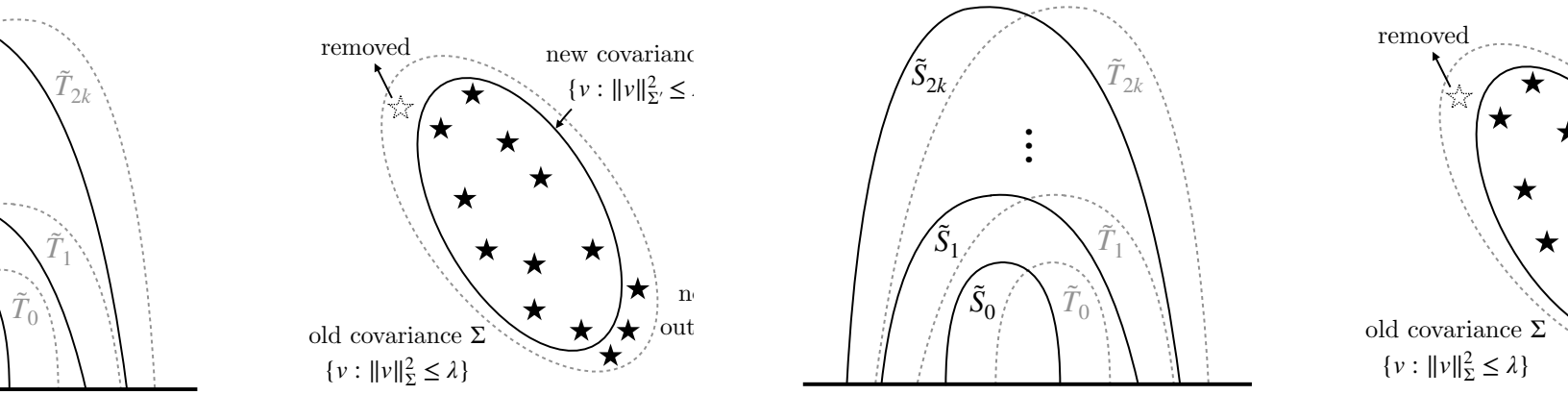
As we will discuss below, this function has sensitivity 2: adjacent data sets have good subsets with similar sizes and outlier thresholds. Our algorithm ultimately uses $\texttt{SCORE}'(x) = \min\{k, \texttt{SCORE}(x)\}$, which simplifies some statements.

For some intuition, note that if $[n]$ is good with respect to the outlier thresholds $\lambda_0, \ldots, \lambda_k$, then $\texttt{SCORE}(x) = 0$. Together with the stability of $\texttt{SCORE}(x)$, this means $\texttt{SCORE}(x)$ roughly tracks the Hamming distance between x and a dataset for which $[n]$ is a good subset.

Good Subsets on Adjacent Data Sets Towards establishing Property (c) for the algorithm we have just described, we first consider what happens to a λ-good subset S for a dataset x if a single sample is *removed* from x.

If we remove from x a sample x_i where $i \notin S$, then of course $S \setminus \{i\} = S$, which is still λ-good for $x \setminus \{x_i\}$. On the other hand, if we remove x_i for some $i \in S$, the resulting $S \setminus \{i\}$ may no longer be λ-good for $x \setminus \{x_i\}$; it may have many outliers! See Figure 1(*a*) for an illustration of this. However, for the parameter regimes we consider, we can show that $S \setminus \{i\}$ is still λ'-good for $x \setminus \{x_i\}$, where $\lambda' = e^\varepsilon \lambda$.

This allows us to reason about pairs of data sets under removal *and addition*, which is the notion we must consider to accommodate adjacent x, x'. If S is a λ-good subset for x and data set x' differs from it in index i^*, then $S \setminus \{i^*\}$ is an $e^\varepsilon\lambda$-good subset for x'.



(*a*) Removing a point from a good s

(*b*) Intertwined: $\tilde{S}_\ell \stackrel{\text{def}}{=} S_\ell \setminus \{i^*\}$

Figure 1: (*a*) A good subset: no points are outliers with respect to the empirical covariance. Removing one point may change the covariance (from Σ to Σ') and cause points to become (slight) outliers. (*b*) Let S_ℓ denote the largest λ_ℓ-good subset of x and T_ℓ the same for adjacent x'. Let $\tilde{S}_\ell$ denote $S_\ell \setminus \{i^*\}$, where x, x' differ in index i^* (and similarly $\tilde{T}_\ell$). These sets are intertwined: for all ℓ, we have $\tilde{S}_\ell \cup \tilde{T}_\ell \subseteq \tilde{S}_{\ell+1} \cap \tilde{T}_{\ell+1}$.

Families of Largest Good Subsets on Adjacent Datasets Now let us return to the family $S_0, \ldots, S_{2k}$ of largest good subsets computed with respect to outlier thresholds $\lambda_0, \ldots, \lambda_{2k}$ on input x and the analogous family $T_0, \ldots, T_{2k}$ for input x' adjacent to x. We prove a crucial "intertwining" relationship between $S_0, \ldots, S_{2k}$ and $T_0, \ldots, T_{2k}$; for a visual depiction, see Figure 1(*b*).

Lemma 7 (Intertwining, Informal) *Suppose x, x' differ on index i^*. For all ℓ, defining $\tilde{S}_\ell \stackrel{\text{def}}{=} S_\ell \setminus \{i^*\}$ and $\tilde{T}_\ell \stackrel{\text{def}}{=} T_\ell \setminus \{i^*\}$, we know that $\tilde{S}_\ell \cup \tilde{T}_\ell$ is a subset of $\tilde{S}_{\ell+1} \cap \tilde{T}_{\ell+1}$.*

Proof [*sketch*] From our discussion above, $S_\ell \setminus \{i^*\}$ is a $\lambda_{\ell+1}$-good subset of both x and x' (assuming $\lambda_{\ell+1} \leq e^\varepsilon \lambda_\ell$, which will be true by construction). Furthermore, $T_{\ell+1}$, the largest $\lambda_{\ell+1}$-good subset of x', contains all $\lambda_{\ell+1}$-good subsets of x', by Lemma 6, and similarly for $S_{\ell+1}$. So, $T_{\ell+1}, S_{\ell+1} \supseteq \tilde{S}_\ell$. And since $\tilde{S}_\ell$ by definition does not contain $\{i^*\}$, also $\tilde{T}_{\ell+1}, \tilde{S}_{\ell+1} \supseteq \tilde{S}_\ell$. A symmetric argument applies to $\tilde{T}_\ell$. ■

Now we sketch the main ideas to show that intertwining implies property (c), that $\|w - w'\|_1 \leq O(1/n)$, where w are the weights associated to input x and w' are those associated with an adjacent x'. Since $\texttt{SCORE}(x), \texttt{SCORE}(x') < k \approx 1/\varepsilon$, there exists $\ell < k$ such that $|S_\ell| \geq n - k$, and since $S_\ell \subseteq S_{k+1}$, also $|S_{k+1}| \geq n - k$. By the nesting property $S_\ell \subseteq S_{\ell+1}$ and similarly for the T_ℓ's, we have

$$\left| \bigcap_{\ell=k+1}^{2k} S_\ell \cap T_\ell \right| \geq n - O(k)$$

By definition,

$$|w_i - w_i'| = \frac{1}{kn} \left| \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in S_\ell\} - \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in T_\ell\} \right|. \tag{2}$$

For $i \in \bigcap_{\ell=k+1}^{2k} S_\ell \cap T_\ell$, the difference in Equation (2) is 0. This leaves just $O(k)$ indices i where $|w_i - w'_i|$ could be nonzero. We divide these into two cases.

If $i = i^*$, the sums in Equation (2) differ by at most k, so $|w_{i^*} - w'_{i^*}| \le 1/n$.

For $i \neq i^*$, by the intertwining property (Lemma 7) we know that if $w_i \neq w'_i$ then there exists only one ℓ such that $\mathbb{1}\{i \in S_\ell\} \neq \mathbb{1}\{i \in T_\ell\}$. So for such i we have $|w_i - w'_i| \le 1/(kn)$, and the total contribution of those terms is $O(1/n)$. This concludes our sketch of `StableCovariance`.

From `StableCovariance` to Mean Estimation Let's return to the task of estimating the mean, given the output $\hat{\Sigma}$ of `StableCovariance`(x). A natural strategy would be to draw n fresh samples x_i and set $\tilde{x}_i = \hat{\Sigma}^{-1/2} x_i$, then hand these samples to a *non*-covariance-aware private mean estimation algorithm. If $\hat{\Sigma}$ were actually *private*, composition would show that this is differentially private. However, our stability (rather than privacy) guarantees of `StableCovariance` do not translate into an overall privacy guarantee for this scheme.

Instead, the second phase of our algorithm computes a stable mean, to which we add noise. Our estimator uses the same idea from `StableCovariance` of averaging over large subsets of the data which contain no outliers; it has a different notion of "outlier" that depends on pairwise distances between points (relative to the estimated covariance $\hat{\Sigma}$). It is similar in flavor to recent approaches in private estimation that look for large sets of points (or "cores") which are all close to each other (Tsfadia et al., 2022; Ashtiani and Liaw, 2021).

Beyond Mean Estimation: Private Covariance Estimation and Learning of Gaussian Distributions A key piece of our privacy proof is the stability of `StableCovariance`: on adjacent inputs x and x' we either fail or produce covariance estimates $\Sigma_1 \leftarrow$ `StableCovariance`(x) and $\Sigma_2 \leftarrow$ `StableCovariance`(x') such that $\mathcal{N}(0, \Sigma_1) \approx_{(\varepsilon,\delta)} \mathcal{N}(0, \Sigma_2)$. Our main algorithm combines this with a stable mean estimation procedure, but by itself it gives rise to a differentially private algorithm: compute $\hat{\Sigma} \leftarrow$ `StableCovariance`(x) and, if this does not fail, return $z \sim \mathcal{N}(0, \hat{\Sigma})$. This algorithm has dubious utility: z contains no information about the mean of x and little information about the covariance. However, as Alabi et al. (2022) recently observed, *multiple* such samples are useful: given private samples $z_1, \ldots, z_N \overset{\text{iid}}{\sim} \mathcal{N}(0, \hat{\Sigma})$, one can post-process them to produce a private covariance estimate $\tilde{\Sigma} = \frac{1}{N}\sum_{i=1}^N z_i z_i^T$. Our main analysis says that we can draw $N = 1$ sample when $n \gtrsim d$; by advanced composition, we can draw N samples when $n \gtrsim d\sqrt{N}$. Alabi et al. (2022) call this the *Gaussian sampling mechanism*; they combine it with a stable estimator based on the sum-of-squares framework to privately and robustly learn the covariance of a Gaussian in polynomial time. Combined with `StableCovariance`, it yields a fast algorithm for private covariance estimation with strong error guarantees in both spectral and Frobenius norm, in particular achieving the optimal dimension-dependence for both.

1.2. Comparison with a Concurrent Result

Independently and concurrently, Duchi et al. (2023) (henceforth, "DHK") provided an algorithm for private mean estimation with similar running time and accuracy guarantees. Both papers fill in the basic framework of BGSUZ by devising stable mean and covariance estimators. DHK's stable covariance estimator has a similar flavor to ours: iteratively removing outlier and updating the empirical covariance. However, the analysis and algorithmic details are quite different. Both DHK's stable mean estimator and our own consider pairwise distances between points (after rescaling by the stable covariance). However, DHK's algorithm looks at distances within small, randomly generated

subsets of the data. Their algorithm and analysis, which make clever use of the randomness in the subset assignments, is substantially different from our own.

In the basic setting where the data are drawn i.i.d. from a full-rank subgaussian distribution, our accuracy guarantees are slightly stronger. Informally, the noise we add for privacy incurs error of magnitude $\frac{d\sqrt{\log 1/\delta}}{\varepsilon n}$, while the comparable term in their analysis is $\frac{d \log 1/\delta}{\varepsilon n}$. Additionally, their algorithm requires $n \gtrsim \frac{d \log^2 1/\delta}{\varepsilon^2}$ examples to ensure privacy, while ours requires asymptotically fewer: $n \gtrsim \frac{d \log 1/\delta}{\varepsilon}$.

DHK extend their algorithm and analysis to address distributions that are heavy-tailed or rank-deficient. Similar extensions would likely apply to our algorithm. We apply our stable covariance estimator to the task of learning the covariance of a subgaussian distributions; DHK's analogous estimator would fill the same role with similar guarantees.

1.3. Related Work

In this section, we overview the most relevant work on learning the mean and/or covariance of high-dimensional, unrestricted Gaussian and subgaussian distributions subject to differential privacy. These results deal with approximate differential privacy; algorithms satisfying stronger privacy notions (such as pure or concentrated DP) cannot be accurate without prior bounds on the parameters.

There is a great deal of research on differentially private statistics outside of these bounds. The primer of Kamath and Ullman (2020) contains an introduction to differentially private statistics along with a survey. One notable point is the work of Karwa and Vadhan (2017), which established approaches for learning univariate Gaussians with bounds that are independent of the data location and scale (using tools of Dwork and Lei (2009)). Another influential work is that of Kamath, Li, Singhal, and Ullman (2019), who gave algorithms for learning the mean and covariance of high-dimensional Gaussians that satisfy concentrated differential privacy. They require prior bounds on the mean and covariance, but their error bounds depend on these quantities only polylogarithmically. Both papers proved lower bounds that apply to our setting.

The way we use the stability of our nonprivate parameter estimates is somewhat nontraditional. Stability-like properties, such as bounds on Lipschitz constants or "global sensitivity," come up frequently in the design of differentially private algorithms. Most often, the idea is to argue that some function is stable so that we can add limited noise to the output of the function (as with the Laplace and usual Gaussian mechanisms). By contrast, we use the stability of `StableCovariance` to argue that the noise itself (added to $\texttt{StableMean}(x)$) is indistinguishable on adjacent inputs. This concept appears elsewhere recently in various forms (Kothari et al., 2021; Ashtiani and Liaw, 2021; Tsfadia et al., 2022; Alabi et al., 2022), though the details of our application are quite different.

Gaussian Mean Estimation The sample complexity of private covariance-aware mean estimation for unrestricted Gaussians is well-understood, with nearly matching upper (Brown et al., 2021) and lower bounds (Karwa and Vadhan, 2017; Kamath et al., 2019). Algorithms with nearly optimal sample complexity exist for subgaussian distributions as well (Liu et al., 2022), with slightly stronger lower bounds known for some related families of distributions (Cai et al., 2019). Previously, known approaches for this task required exponential time or $\Omega(d^2)$ samples, corresponding to the best-known sample complexity for privately learning a the covariance in spectral norm. One

can use such an approximation as a preconditioner: after rescaling, the distribution will be nearly isotropic and one can apply the techniques of Karwa and Vadhan (2017) dimension-by-dimension.

Gaussian Covariance Estimation Differentially private Gaussian covariance estimation has received much attention recently, especially as the key stepping stone to privately learning a Gaussian to small total variation distance. For this task the sample complexity is well-understood, with lower bounds (Vadhan, 2017; Kamath et al., 2022) and clean information-theoretic upper bounds (Aden-Ali et al., 2021) matching the non-private sample complexity for modest privacy parameters. A series of improving upper bounds has established polynomial-time algorithms nearly matching these results (Liu et al., 2022; Kothari et al., 2021; Kamath et al., 2021; Ashtiani and Liaw, 2021; Alabi et al., 2022; Hopkins et al., 2022). For more comprehensive overviews (which also discuss robustness and stronger notions of privacy) refer to Alabi et al. (2022) and Hopkins et al. (2022).

The task of privately producing a spectral-norm covariance approximation, often useful as a preconditioner, has seen less progress. Nonprivately, this task requires only $n \approx d$ examples; our algorithm requires $d^{3/2}$. This dependence on the dimension was recently proved be optimal in the regime of $\alpha = O(1/\sqrt{d})$ (Kamath et al., 2022). To the best of our knowledge, ours is the first differentially private polynomial-time algorithm achieving such a guarantee for unrestricted Gaussian distributions, closing an open question posed by Alabi et al. (2022). Under the assumption that $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, the standard Gaussian mechanism requires $\Omega(\mathrm{poly}(\kappa) \cdot d^{3/2})$ samples and the private preconditioner of Kamath et al. (2019) requires $\Omega(\mathrm{polylog}(\kappa) \cdot d^{3/2})$ samples (ignoring the dependence on other parameters).

2. Main Result and Analysis

Theorem 8 (Main Theorem) *Fix $\varepsilon, \beta \in (0,1)$, $\delta \in (0, \varepsilon/10]$, and $n, d \in \mathbb{N}$. Algorithm 1 takes a data set of n points, each in $\mathbb{R}^d$, privacy parameters ε, δ, and an outlier threshold λ_0.*

- *For any $\lambda_0 \geq 1$, Algorithm 1 is (ε, δ)-differentially private.*
- *For $\mu \in \mathbb{R}^d$ and positive definite $\Sigma \in \mathbb{R}^{d \times d}$, let $x = x_1, \ldots, x_n$ be drawn i.i.d. from a subgaussian distribution $\mathcal{D}$ with parameter $K_{\mathcal{D}}$, mean μ, and covariance Σ. There exist absolute constants K_1, K_2, and K_3 such that, if $\lambda_0 = K_1 K_{\mathcal{D}}^2(d + \log n/\beta)$ and*

$$n \geq K_2 K_{\mathcal{D}}^2 \cdot \frac{\log 1/\delta}{\varepsilon}\left(d + \log\left(\frac{K_{\mathcal{D}} \log 1/\delta}{\varepsilon\beta}\right)\right),$$

  then, with probability at least $1 - \beta$, Algorithm 1 returns $\hat{\mu}$ such that

$$\|\hat{\mu} - \mu\|_\Sigma \leq K_3 K_{\mathcal{D}}\left(\sqrt{\frac{d + \log 1/\beta}{n}} + \frac{d\sqrt{\log 1/\delta}}{\varepsilon n} + \frac{\log n/\beta \sqrt{\log 1/\delta}}{\varepsilon n}\right).$$

- *Algorithm 1 can be implemented to require: one product of the form $A^T A$ for $A \in \mathbb{R}^{n \times d}$, two products of the form AB for $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{d \times d}$, one inversion of a matrix in $\mathbb{R}^{d \times d}$ to logarithmic bit complexity, and further computational overhead of $\tilde{O}(nd/\varepsilon)$.*

We prove Theorem 8 by proving each subclaim separately: privacy in Lemma 15, accuracy in Lemma 19, and running time in Lemma 20.

Algorithm 1: Private Mean Estimation, $\mathcal{A}_{\text{main}}^{\varepsilon,\delta,\lambda_0}(x)$

input : data set $x \in \mathbb{R}^{n\times d}$; privacy parameters ε, δ; outlier threshold λ_0
require: $n \geq \frac{192e^2\lambda_0 \log 6/\delta}{\varepsilon} + 160e^2\lambda_0$

`/* Initialize */`
$k \leftarrow \left\lceil \frac{6\log 6/\delta}{\varepsilon} \right\rceil + 4; \quad M \leftarrow 6k + \lceil 18 \log 16n/\delta \rceil; \quad c^2 \leftarrow \frac{720e^2\lambda_0 \log 12/\delta}{\varepsilon^2 n^2};$
$R \sim \text{Uniform}(\{R' \subseteq [n] : |R'| = M\});$

`/* Compute nonprivate parameter estimates */`
$\hat{\Sigma}, \texttt{SCORE}_1 \leftarrow \texttt{StableCovariance}(x, \lambda_0, k);$ `/*` $\hat{\Sigma} \in \mathbb{R}^{d\times d}, \texttt{SCORE}_1 \in \mathbb{N}$ `*/`
$\hat{\mu}, \texttt{SCORE}_2 \leftarrow \texttt{StableMean}(x, \hat{\Sigma}, \lambda_0, k, R);$ `/*` $\hat{\mu} \in \mathbb{R}^d, \texttt{SCORE}_2 \in \mathbb{N}$ `*/`

`/* Test the scores and release */`
if $\mathcal{M}_{\text{PTR}}^{\varepsilon/3,\delta/6}(\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\}) = \texttt{PASS}$ **then**
  return $\tilde{\mu} \sim \mathcal{N}(\hat{\mu}, c^2\hat{\Sigma})$;
else
  return `FAIL`
end

2.1. Privacy Analysis

Let x and x' be adjacent data sets. Our main algorithm, Algorithm 1, either fails or computes nonprivate estimates $(\hat{\mu}, \hat{\Sigma})$ and outputs a sample $\tilde{\mu} \sim \mathcal{N}(\hat{\mu}, c^2\hat{\Sigma})$ for some real number c. By standard propose-test-release-style analysis, it suffices to establish that (i) the probabilities of failing under x and x' are (ε, δ)-indistinguishable and (ii) for any pairs $(\hat{\mu}, \hat{\Sigma})$ computed by Algorithm 1 on x and $(\hat{\mu}', \hat{\Sigma}')$ computed on x', if we do not fail then

$$\mathcal{N}(\hat{\mu}, c^2\hat{\Sigma}) \approx_{(\varepsilon,\delta)} \mathcal{N}(\hat{\mu}', c^2\hat{\Sigma}').$$

The indistinguishability of failure probabilities follows from the low sensitivity of our score function. As in BGSUZ, to establish the indistinguishability of Gaussians we "change the mean" and "change the covariance" separately and use composition to argue that the result is indistinguishable. We prove the overall privacy statement in Lemma 15 at the end of this subsection.

We call our propose-test-release function on the maximum of the two scores (one for the covariance and one for the mean). We begin by showing that this maximum has low sensitivity.

Lemma 9 *Fix outlier threshold $\lambda_0 > 0$, reference set $R \subseteq [n]$, and privacy parameters $\varepsilon > 0$ and $0 < \delta < 1$; set $k = \left\lceil \frac{6\log 6/\delta}{\varepsilon} \right\rceil + 4$ as in Algorithm 1. Assume $n \geq 32e^2\lambda_0 k$. Let x and x' be adjacent data sets of size n. Let*

$$\hat{\Sigma}, \texttt{SCORE}_1 \leftarrow \texttt{StableCovariance}(x, \lambda_0, k)$$
$$\hat{\Sigma}', \texttt{SCORE}_1' \leftarrow \texttt{StableCovariance}(x', \lambda_0, k)$$
$$\hat{\mu}, \texttt{SCORE}_2 \leftarrow \texttt{StableMean}(x, \hat{\Sigma}, \lambda_0, k, R)$$
$$\hat{\mu}', \texttt{SCORE}_2' \leftarrow \texttt{StableMean}(x', \hat{\Sigma}', \lambda_0, k, R).$$

Then $|\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\} - \max\{\texttt{SCORE}_1', \texttt{SCORE}_2'\}| \leq 2$.

Proof Lemma 31, proved in Section 3.4, says that $|\texttt{SCORE}_1 - \texttt{SCORE}'_1| \leq 2$, provided $k \leq \frac{n}{4e^2\lambda_0}$ (recall m, the size of the paired data set, is $n/2$). This is satisfied by assumption.

Lemma 41, proved in Section 4.4, says that $|\texttt{SCORE}_2 - \texttt{SCORE}'_2| \leq 2$ provided $(1-\gamma)\hat{\Sigma} \preceq \hat{\Sigma}' \preceq \frac{1}{1-\gamma}\hat{\Sigma}$ for $\gamma \leq \frac{1}{2}$ and $k \leq \frac{1}{2\gamma}$. By Lemma 11 (below), if $\texttt{SCORE}_1, \texttt{SCORE}'_1 < k$, these inequalities are satisfied for $\gamma = \frac{16e^2\lambda_0}{n}$ as long as $n \geq 32e^2\lambda_0 k$, as we have assumed. If one of $\texttt{SCORE}_1$ or $\texttt{SCORE}'_1$ is equal to k, then we are still low-sensitivity: assuming without loss of generality that $\texttt{SCORE}_1 = k$, we know that $\texttt{SCORE}'_1 \geq k - 2$, and thus

$$\max\left\{\texttt{SCORE}_1, \texttt{SCORE}_2\right\} = k$$
$$\max\left\{\texttt{SCORE}'_1, \texttt{SCORE}'_2\right\} \geq \texttt{SCORE}'_1 \geq k - 2.$$

We have used the fact that all four scores are at most k. ■

Algorithm 1 feeds this maximum of the two scores to $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}$, a private propose-test-release-style check that the score is low and it is safe to proceed. The textbook approach to this task uses Laplace noise; our mechanism is an analogue of this that always passes on zero inputs and always fails on large inputs. This modification cleans up our arguments, decoupling parameter stability from the (random) outcome of the private check. We provide the (elementary) proof of Claim 10 in Appendix B.

Claim 10 *Fix $0 < \varepsilon \leq 1$ and $0 < \delta \leq \frac{\varepsilon}{10}$. There is an algorithm $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}} : \mathbb{R} \to \{\texttt{PASS}, \texttt{FAIL}\}$ that satisfies the following conditions:*

(1) Let $\mathcal{U}$ be a set and $g : \mathcal{U}^n \to \mathbb{R}_{\geq 0}$ a function. If, for all $x, x' \in \mathcal{U}^n$ that differ in one entry, $|g(x) - g(x')| \leq 2$, then $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(g(\cdot))$ is (ε, δ)-DP.

(2) $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(0) = \texttt{PASS}$.

(3) For all $z \geq \frac{2\log 1/\delta}{\varepsilon} + 4$, $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(z) = \texttt{FAIL}$.

We now present the stability guarantees for our nonprivate parameter estimates. Recall that `StableCovariance` is deterministic and outputs an estimate $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ along with an integer `SCORE`. The value of `SCORE` is 2-sensitive (under adjacent inputs) and, when `SCORE` is not too large, the parameter estimate $\hat{\Sigma}$ is itself stable in the sense we need for privacy.

Lemma 11 *Fix data set size n, outlier threshold $\lambda_0 > 0$, and discretization parameter $k \in \mathbb{N}$. Assume $k \leq \frac{n}{4e^2\lambda_0}$ and let $\gamma = \frac{16e^2\lambda_0}{n}$. Let x, x' be adjacent data sets and*

$$\Sigma_1, \texttt{SCORE} \leftarrow \texttt{StableCovariance}(x, \lambda_0, k)$$
$$\Sigma_2, \texttt{SCORE}' \leftarrow \texttt{StableCovariance}(x', \lambda_0, k).$$

Assume $\texttt{SCORE}, \texttt{SCORE}' < k$. Then $\Sigma_1, \Sigma_2 \succ 0$ and $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$. If $\gamma \leq \frac{1}{2}$, then

$$\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}}, \left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}} \leq (1+2\gamma)\gamma.$$

The bulk of Section 3 is devoted to proving Lemma 11. Together with the following claim, it immediately implies indistinguishabilty of zero-mean, rescaled Gaussians.

Claim 12 *Fix $\varepsilon \in (0,1)$ and $\delta \in (0, 1/10]$ and let $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ be positive definite matrices. If*

$$\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}}, \left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}} \leq \frac{\varepsilon}{3\log 2/\delta},$$

then $\mathcal{N}(0,\Sigma_1) \approx_{(\varepsilon,\delta)} \mathcal{N}(0,\Sigma_2)$.

Claim 12 follows from a straightforward Gaussian concentration argument. See BGSUZ, Lemma 4.15, and note that their proof works as-is with a better constant (3 in place of 5). See also Alabi et al. (2022), whose similar Theorem 5.1 we use in Section 5.

The main result in Section 4 is similar: `StableMean` is deterministic and returns an estimate $\hat{\mu} \in \mathbb{R}^d$ and an integer `SCORE`. The latter is 2-sensitive and, when it is not too big, the mean estimates are stable. We have a few additional concerns: on inputs x and x', we might have previously calculated slightly different covariances Σ_1 and Σ_2. Thus we analyze the stability of `StableMean` when simultaneously moving from (x, Σ_1) to (x', Σ_2). Furthermore, for computational efficiency `StableMean` accepts a set $R \subseteq [n]$ of "reference points" on which to estimate whether points are outliers. For our stability conditions to hold, this set needs to be both sufficiently large and sufficiently representative (in a precise sense: see Definition 32).

Lemma 13 *Fix data set size n, dimension d, outlier threshold $\lambda_0 \geq 1$ and discretization parameter $k \in \mathbb{N}$. Use reference set $R \subseteq [n]$ with $|R| > 6k$. Let x and x' be adjacent d-dimensional data sets of size n. Let $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma = 16e^2\frac{\lambda_0}{n}$. Assume that $k \leq \frac{1}{2\gamma}$. Let*

$$\hat{\mu}, \texttt{SCORE} \leftarrow \texttt{StableMean}(x, \Sigma_1, \lambda_0, k, R)$$
$$\hat{\mu}', \texttt{SCORE}' \leftarrow \texttt{StableMean}(x', \Sigma_2, \lambda_0, k, R).$$

If $\texttt{SCORE}, \texttt{SCORE}' < k$ and R is degree-representative for both x and x' (see Definition 32), then $\|\hat{\mu} - \hat{\mu}'\|_{\Sigma_1}^2 \leq \frac{(1+2\gamma)38e^2\lambda_0}{n^2}$.

The nonprivate means are close in exactly the sense we need to establish indistinguishability of Gaussians (with the same covariance). Recall that $\|\hat{\mu} - \hat{\mu}'\|_{\Sigma_1}^2 = \|\Sigma_1^{-1/2}\hat{\mu} - \Sigma_1^{-1/2}\hat{\mu}'\|_2^2$.

Claim 14 (Standard Gaussian Mechanism) *Let vectors u and v satisfy $\|u - v\|_2^2 \leq \Delta^2$. For any $\varepsilon, \delta \in (0,1)$, let $c^2 = \Delta^2 \cdot \frac{2\log 2/\delta}{\varepsilon^2}$. Then $\mathcal{N}(u - v, c^2\mathbb{I}) \approx_{(\varepsilon,\delta)} \mathcal{N}(0, c^2\mathbb{I})$.*

We are ready to prove, via standard tools, the privacy guarantee for Algorithm 1.

Lemma 15 (Main Privacy Claim) *Fix $0 < \varepsilon < 1$ and $0 < \delta < \frac{\varepsilon}{10}$. For any $\lambda_0 \geq 1$, Algorithm 1 is (ε, δ)-differentially private.*

Proof Set $\varepsilon' = \varepsilon/3$ and $\delta' = \delta/6$. We will prove that Algorithm 1 is $(3\varepsilon', 6\delta')$-differentially private.

The algorithm requires $n \geq \frac{192e^2\lambda_0\log 6/\delta}{\varepsilon} + 160e^2\lambda_0$ to proceed. This is set so that $n \geq 32e^2\lambda_0 k$ for $k = \left\lceil \frac{6\log 6/\delta}{\varepsilon} \right\rceil + 4$, a fact we will require later. Assume this lower bound is true, since otherwise the algorithm immediately fails (and thus is private).

Fix adjacent data sets x and x'. Since $M = |R| > 18\log 16n/\delta = 18\log 4n/\delta'$, Claim 50 and a union bound over all data points in x and x' establish that R is degree-representative for x and x'

with probability at least $1-\delta'$. We will show that, conditioned on R being degree-representative for Y and Y', the mechanism is $(3\varepsilon', 5\delta')$-differentially private. This will prove that the entire algorithm is private with parameters $(3\varepsilon', 6\delta') = (\varepsilon, \delta)$.

We want $\mathcal{M}_{\text{PTR}}^{\varepsilon',\delta'}$ to be (ε', δ')-differentially private. For any fixed reference set $R = r \subseteq [n]$, Lemma 9 tells us this happens when $k = \left\lceil \frac{2\log 1/\delta'}{\varepsilon'} \right\rceil + 4 = \left\lceil \frac{6\log 6/\delta}{\varepsilon} \right\rceil + 4$ and $n \geq 32e^2\lambda_0 k$. Our assumed lower bound on n is set so that this is true.

When we do not fail, by the guarantees of $\mathcal{M}_{\text{PTR}}^{\varepsilon,\delta}$ we know that both scores are strictly less than k. Let (μ_1, Σ_1) be the nonprivate parameters computed on x and (μ_2, Σ_2) the nonprivate parameters computed on x'. Lemma 11 and Claim 12 together imply that $\mathcal{N}(0, \Sigma_1) \approx_{(\varepsilon',\delta')} \mathcal{N}(0, \Sigma_2)$. To apply Lemma 11, we require $n \geq 4e^2\lambda_0 k$ and $n \geq 32e^2\lambda_0$, which are satisfied by assumption. The lemma establishes a trace norm bound of $\Delta = (1+2\gamma)\gamma$ for $\gamma = \frac{16e^2\lambda_0}{n}$. To apply Claim 12, we require $\Delta \leq \frac{\varepsilon'}{3\log 2/\delta'}$. (These constraints, along with $\varepsilon \leq 1$ and $\delta \leq 1/10$, imply $\gamma \leq 0.025$, so $(1+2\gamma)\gamma \leq 1.05 \cdot \gamma$ and $\Delta \leq 17e^2\lambda_0/n$.) Rearranging, we see that (ε', δ')-indistinguishability of $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ requires $n \geq \frac{153e^2\lambda_0 \log 12/\delta}{\varepsilon}$, a weaker bound than we initially assumed. In particular, this implies $\mathcal{N}(\mu_2, c^2\Sigma_1) \approx_{(\varepsilon',\delta')} \mathcal{N}(\mu_2, c^2\Sigma_2)$, as adding a fixed vector or multiplying by a fixed value do not affect indistinguishability.

Lemma 13, with the above observation that $2\gamma \leq 0.05$ for the parameters we consider, states that $\|\mu_1 - \mu_2\|_{\Sigma_1}^2 \leq \frac{40e^2\lambda_0}{n^2}$. With Claim 14, this means that $\mathcal{N}(\mu_1, c^2\Sigma_1) \approx_{(\varepsilon',\delta')} \mathcal{N}(\mu_2, c^2\Sigma_1)$, where

$$c^2 = \left(\frac{40e^2\lambda_0}{n^2}\right)\left(\frac{2\log 2/\delta'}{(\varepsilon')^2}\right) = \frac{720e^2\lambda_0 \log 12/\delta}{\varepsilon^2 n^2},$$

as in Algorithm 1. To apply Lemma 13, we require a few conditions. First, $|R| > 6k$ is satisfied by construction. The condition that $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma = \frac{16e^2\lambda_0}{n}$ is satisfied; it is a consequence of Lemma 11. We also require $k \leq \frac{1}{2\gamma}$, which is equivalent to $n \geq 32e^2\lambda_0 k$; we already assumed this to be true. Finally, the scores are strictly less than k (since we did not fail) and R is degree-representative for x and x' (by assumption).

We now apply Fact 52 to combine the statements of $\mathcal{N}(\mu_1, c^2\Sigma_1) \approx_{(\varepsilon',\delta')} \mathcal{N}(\mu_2, c^2\Sigma_1)$ and $\mathcal{N}(\mu_2, c^2\Sigma_1) \approx_{(\varepsilon',\delta')} \mathcal{N}(\mu_2, c^2\Sigma_2)$. We get that $\mathcal{N}(\mu_1, c^2\Sigma_1) \approx_{(2\varepsilon',(1+e^{\varepsilon'})\delta')} \mathcal{N}(\mu_2, c^2\Sigma_2)$. Since $e^{\varepsilon'} \leq e \leq 3$, we have the same statement with values $(2\varepsilon', 4\delta')$. Basic composition (Fact 51) allows us to combine the guarantees for these Gaussians with those for $\mathcal{M}_{\text{PTR}}^{\varepsilon,\delta}$ to establish that the algorithm, conditioned on the fact that R is degree-representative for x and x', is $(3\varepsilon', 5\delta')$-differentially private. Since R fails to be degree-representative for x and x' with probability at most δ', we have finished the proof. ■

2.2. Accuracy Analysis

Let x be a d-dimensional data set of size n. We define the empirical mean and paired empirical covariance:

$$\mu_x \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \Sigma_x \stackrel{\text{def}}{=} \frac{1}{\sqrt{2}\,\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(x_i - x_{i+\lfloor n/2 \rfloor}\right)\left(x_i - x_{i+\lfloor n/2 \rfloor}\right)^T. \tag{3}$$

This "pairing" is a standard trick that centers the data at the cost of halving our sample size: given input $x = (x_1, \ldots, x_n)$ we construct $y = (y_1, \ldots, y_m)$ by setting $y_i = (1/\sqrt{2})(x_i - x_{i+m})$ where $m = \lfloor n/2 \rfloor$. If we have $x_i, x_{i+m} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$, then $y_i \sim \mathcal{N}(0, \Sigma)$. The transformation preserves adjacency: if x and x' are adjacent, then so are y and y'.

We now introduce a notion of good data sets, on which `StableCovariance` and `StableMean` "let the data through," always returning exactly the paired empirical covariance and empirical mean, respectively. Furthermore, Algorithm 1 never returns `FAIL` on such inputs. We formalize these points in Observation 17, which is simple but serves as the linchpin of our accuracy analysis.

Definition 16 (Well-Concentrated) *Fix $\lambda_0 > 0$ and let x be a d-dimensional data set of size n samples. Let Σ_x the paired empirical covariance. We say x is λ_0-*well-concentrated *if Σ_x is invertible and, for all $i, j \in [n]$, $\|x_i - x_j\|_{\Sigma_x}^2 \leq \lambda_0$.*

Observation 17 (Always `PASS` on Well-Concentrated Inputs) *Fix $\lambda_0 > 0$. If data set $x \in \mathbb{R}^{n \times d}$ is λ_0-well-concentrated and $n \geq \frac{192e^2\lambda_0 \log 6/\delta}{\varepsilon} + 160e^2\lambda_0$ (so we do not fail immediately), then Algorithm 1 computes $\texttt{SCORE}_1 = \texttt{SCORE}_2 = 0$, passes deterministically, and returns a sample from $\mathcal{N}(\mu_x, c^2\Sigma_x)$, where μ_x is the empirical mean, Σ_x is the paired empirical covariance, and c is the noise scale set in Algorithm 1.*

The essential fact we use about our subgaussian input distributions is that they concentrate in exactly this way. For definitions and concentration inequalities about subgaussian distribution, see Section A.2. The proof of Claim 18, which we omit, is straightforward from these statements. Recall that the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is subgaussian with parameter $K = O(1)$.

Claim 18 (Subgaussian Data are Well-Concentrated) *Let $x \in \mathbb{R}^{n \times d}$ be drawn i.i.d. from a subgaussian distribution $\mathcal{D}$ with parameter $K_{\mathcal{D}}$, arbitrary mean, and full-rank covariance. There exist constants K_1 and K_2 such that, when $\lambda_0 \geq K_1 K_{\mathcal{D}}^2(d + \log n/\beta)$ and $n \geq K_2 K_{\mathcal{D}}^2(d + \log 1/\beta)$, then x is λ_0-well-concentrated with probability at least $1 - \beta$.*

We now prove our main accuracy claim. The calculation is standard: with high probability the true mean is close to the empirical mean, which is close to our private estimate.

Lemma 19 (Main Accuracy Claim) *Fix $\varepsilon, \delta \in (0, 1)$, $\delta \in (0, \varepsilon/10]$, and $n, d \in \mathbb{N}$. For $\mu \in \mathbb{R}^d$ and positive definite $\Sigma \in \mathbb{R}^{d \times d}$, let $x = x_1, \ldots, x_n$ be drawn i.i.d. from a subgaussian distribution $\mathcal{D}$ with parameter $K_{\mathcal{D}}$, mean μ, and covariance Σ. There exist absolute constants K_1, K_2, and K_3 such that, if $\lambda_0 = K_1 K_{\mathcal{D}}^2(d + \log n/\beta)$ and*

$$n \geq K_2 K_{\mathcal{D}}^2 \cdot \frac{\log 1/\delta}{\varepsilon}\left(d + \log\left(\frac{K_{\mathcal{D}} \log 1/\delta}{\varepsilon\beta}\right)\right),$$

then, with probability at least $1 - \beta$, Algorithm 1 returns $\hat{\mu}$ such that

$$\|\hat{\mu} - \mu\|_\Sigma \leq K_3 K_{\mathcal{D}}\left(\sqrt{\frac{d + \log 1/\beta}{n}} + \frac{d\sqrt{\log 1/\delta}}{\varepsilon n} + \frac{\log n/\beta\sqrt{\log 1/\delta}}{\varepsilon n}\right).$$

Proof In order to not fail immediately, we require

$$n \geq \frac{192e^2\lambda_0 \log 6/\delta}{\varepsilon} + 160e^2\lambda_0 \gtrsim K_{\mathcal{D}}^2 \cdot \frac{(d + \log n/\beta)\log 1/\delta}{\varepsilon}.$$

This expression has n on both sides; using the fact that $\frac{n}{\log n} \gtrsim \eta$ implies $n \gtrsim \eta \log \eta$, we can rewrite this as

$$n \gtrsim K_{\mathcal{D}}^2 \cdot \frac{\log 1/\delta}{\varepsilon}\left(d + \log\left(\frac{K_{\mathcal{D}} \log 1/\delta}{\varepsilon\beta}\right)\right),$$

as we require. In particular, this implies $n \gtrsim K_{\mathcal{D}}^2(d + \log 1/\beta)$.

Beyond this, we are concerned about four bad events: (i) the data set is not λ_0-well-concentrated, (ii) the empirical mean is far from the true mean, (iii) the empirical covariance is unlike the true covariance, and (iv) the noise added for privacy is large. We will show each of these events happens with probability at most $\beta/4$, analyze the accuracy under the assumption that none of them happens, and use a union bound to finish the proof.

By Observation 17 and Claim 18, with probability at least $1 - \beta/4$ data set x is λ_0-well-concentrated and Algorithm 1 releases $\tilde{\mu} \sim \mathcal{N}(\mu_x, c^2\Sigma_x)$, where μ_x is the empirical mean, Σ_x is the paired empirical covariance, and

$$c^2 = \frac{72e^2\lambda_0 \log 2/\delta}{\varepsilon^2 n^2} \lesssim K_{\mathcal{D}}^2(d + \log n/\beta) \cdot \frac{\log 1/\delta}{\varepsilon^2 n^2}. \tag{4}$$

We add and subtract the empirical mean and apply the triangle inequality:

$$\begin{aligned}\|\mu - \tilde{\mu}\|_\Sigma &= \|\mu - \mu_x + \mu_x - \tilde{\mu}\|_\Sigma \\ &\leq \|\mu - \mu_x\|_\Sigma + \|\mu_x - \tilde{\mu}\|_\Sigma .\end{aligned}$$

μ_x is the empirical mean, so by Claim 48, with probability at least $1 - \beta/4$, we have

$$\|\mu - \mu_x\|_\Sigma \lesssim K_{\mathcal{D}}\sqrt{\frac{d + \log 1/\beta}{n}}. \tag{5}$$

Since $n \gtrsim K_{\mathcal{D}}^2(d + \log 1/\beta)$, Claim 49 implies that, with probability at least $1 - \beta/4$, we have $\|v\|_\Sigma \lesssim \|v\|_{\Sigma_x}$ for all vectors v. Thus $\|\mu_x - \tilde{\mu}\|_\Sigma \lesssim \|\mu_x - \tilde{\mu}\|_{\Sigma_x}$ This is the "correct" norm in which to control $\mu_x - \tilde{\mu}$, since $\tilde{\mu} \sim \mathcal{N}(\mu_x, c^2\Sigma_x)$. Abusing notation and conflating distributions with random variables, we have

$$\|\mu_x - \tilde{\mu}\|_{\Sigma_x} = \left\|\Sigma_x^{-1/2}\big(\mathcal{N}(0, c^2\Sigma_x)\big)\right\|_2 = c \cdot \|\mathcal{N}(0, \mathbb{I})\|_2 .$$

Again applying Claim 48, with probability at least $1-\beta/4$ we have $\|\mu_x - \tilde{\mu}\|_{\Sigma_x} \lesssim c\sqrt{d + \log 1/\beta}$. Plugging in the value of c from Equation (4) and using $d + \log n/\beta \geq d + \log 1/\beta$, we have

$$\begin{aligned}\|\mu_x - \tilde{\mu}\|_{\Sigma_x} &\lesssim K_{\mathcal{D}}\sqrt{d + \log n/\beta} \cdot \frac{\sqrt{\log 1/\delta}}{\varepsilon n} \cdot \sqrt{d + \log 1/\beta} \\ &\lesssim K_{\mathcal{D}} \cdot (d + \log n/\beta) \cdot \frac{\sqrt{\log 1/\delta}}{\varepsilon n}.\end{aligned}$$

Combining this with the upper bound on $\|\mu - \mu_x\|_\Sigma$ in Equation (5) finishes the proof. ■

2.3. Running Time

For modest privacy parameters, our running time is dominated by a few matrix operations related to computing the empirical covariance and rescaling the data.

Lemma 20 *Algorithm 1 can be implemented to require at most*

(1) One product of the form $A^T A$ for $A \in \mathbb{R}^{n \times d}$;

(2) Two products of the form AB for $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{d \times d}$;

(3) One inversion of a positive definite matrix in $\mathbb{R}^{d \times d}$ to polynomial precision (logarithmic bit complexity); and

(4) Further computational overhead of $\tilde{O}(nd/\varepsilon)$.

All of our algorithms can be implemented with finite-precision arithmetic in the word RAM model, setting the bit complexity to be logarithmic in the other parameters. Informally, matrix products of the form in (1) or (2) can be executed in time $\tilde{O}(nd^{\omega-1})$ and inversion can be accomplished in time $\tilde{O}(d^\omega)$, for an overall running time of

$$\tilde{O}\big(nd^{\omega-1} + nd/\varepsilon\big).$$

We have used $n = \Omega(kd) = \Omega(d)$, where $k = \Theta(\log(1/\delta)/\varepsilon)$ is our discretization parameter. These relationships are required by our privacy argument. In what follows we ignore the precision of our operations, as an exact understanding of the running time is not a focus of our work. Also note that the asymptotically fastest algorithms for matrix operations are not practical; "schoolbook" matrix operations can be performed with the number 3 in place of ω.

Covariance Estimation We compute the covariance $\hat{\Sigma} = \frac{1}{n} y^T y$ and its inverse $\hat{\Sigma}^{-1}$. For each (paired) point $y_i \in [m]$ we rescale it: $\tilde{y}_i \leftarrow \hat{\Sigma}^{-1} y_i$. This operation can be written as $\tilde{y} \leftarrow y\hat{\Sigma}^{-1}$, where $y \in \mathbb{R}^{n \times d}$ and $\hat{\Sigma}^{-1} \in \mathbb{R}^{d \times d}$. (Alternatively, one may solve the system $\hat{\Sigma}\tilde{y}^T = y^T$ directly.) Computing each squared Mahalanobis norm (and thus identifying any outliers) can be computed from y and $\tilde{y}$ (via entrywise multiplication and summation) in time $\tilde{O}(nd)$.

A naive implementation of `LargestGoodSet` would repeat the above procedure from scratch after finding an outlier. On top of this, `StableCovariance` calls `LargestGoodSet` $2k + 1$ times to generate $\{S_\ell\}_{\ell=0}^{2k}$, the family of sets where S_ℓ is the largest $e^{\ell/k}\lambda_0$-good subset for y. To improve upon this naive approach, we note three points.

(1) Once we have found an outlier, removing it requires a rank-one update to the Mahalanobis norms (in order to check for new outliers). There are well-known efficient techniques for updates of this type, taking $\tilde{O}(nd + d^2) = \tilde{O}(nd)$ time. We provide details below.

(2) For any $\ell \le \ell'$ we know from Lemma 27 that $S_\ell \subseteq S_{\ell'}$. Therefore, by proceeding through the outlier thresholds in *decreasing* order, each outlier we find can be ignored going forward.

(3) Once we have found an ℓ^* such that $|S_{\ell^*}| \le m - k$, we can immediately halt and either return `FAIL` or output the covariance estimate. Recall that we compute

$$\texttt{SCORE} = \min\left\{k, \min_{0 \le \ell \le k} \{m - |S_\ell| + \ell\}\right\}.$$

If $\ell^* \geq k+1$, then we know that $\texttt{SCORE} = k$, since for all $\ell \leq k$ we have $|S_\ell| \leq |S_{k+1}| \leq n-k$. This means the call to $\mathcal{M}^{\varepsilon,\delta}_{\text{PTR}}$ will fail with probability one. If $\ell^* < k+1$, we can still move to the next step: for all $\ell \leq \ell^*$ we have $m - |S_\ell| + \ell \geq k$, so the size of the smaller S_ℓ's will not affect our score computation.

Combined, these facts imply that we need to perform at most k outlier removals, updating all n Mahalanobis distances each time. Tracking Σ across removals, we can also compute the final weighted covariance.

Let us discuss the process of updating after an outlier removal in more detail. Suppose we want to recompute the norm of a point v after removing u from the covariance matrix. Let Σ_1 be the initial covariance matrix and $\Sigma_2 = \Sigma_1 - \frac{1}{m}uu^T$ the matrix after removing u. We want to compute $v^T(\Sigma_2)^{-1}v$ for each point in the data set. The Sherman-Morrison formula tells us

$$\Sigma_2^{-1} = \left(\Sigma_1 - \frac{1}{m}uu^T\right)^{-1} = \Sigma_1^{-1} + \frac{\Sigma_1^{-1}uu^T\Sigma_1^{-1}}{m - u^T\Sigma_1^{-1}u} \tag{6}$$

Therefore

$$\begin{aligned} v^T\Sigma_2^{-1}v &= v^T\Sigma_1^{-1}v + v^T\left(\frac{\Sigma_1^{-1}uu^T\Sigma_1^{-1}}{m - u^T\Sigma_1^{-1}u}\right)v \\ &= v^T\Sigma_1^{-1}v + \frac{\left(v^T\Sigma_1^{-1}u\right)^2}{m - u^T\Sigma_1^{-1}u}. \end{aligned}$$

We have already calculated $v^T\Sigma_1^{-1}v$ and $u^T\Sigma_1^{-1}u$. We can calculate $\Sigma_1^{-1}u$ once (in time $\tilde{O}(d^2)$), which allows us to compute $\left(v^T\Sigma_1^{-1}u\right)^2$ in time $\tilde{O}(d)$ for any vector v. Observe furthermore that Equation (6) allows us to compute Σ_2^{-1} directly in time $\tilde{O}(d^2)$.

We have established that each update after an outlier's removal can be accomplished in time $\tilde{O}(d^2 + nd) = \tilde{O}(nd)$. Since we perform at most k updates, this takes time $\tilde{O}(knd)$.

Note that we can simultaneously calculate the final weighted covariance with an additional cost of $\tilde{O}(kd^2) = \tilde{O}(knd)$ time. Furthermore, we can return the weighted inverse; it also requires at most k rank-one updates to the original $\left(\frac{1}{m}y^Ty\right)^{-1}$ we computed.

Thus we can execute `StableCovariance` with one computation of $\hat{\Sigma} = y^Ty$, one matrix inversion to obtain $\hat{\Sigma}^{-1}$, one product $y\hat{\Sigma}^{-1}$, and additional overhead of $\tilde{O}(knd) = \tilde{O}(nd/\varepsilon)$.

Mean Estimation As with covariance estimation, naively repeating `LargestCore` all $2k+1$ times would waste effort. Instead, suppose we had a matrix $D \in \mathbb{R}^{n \times M}$, where $M = |R| = \Theta(k + \log n/\delta)$, and D has entries $D_{i,j} = \|x_i - R_j\|^2_{\hat{\Sigma}}$. Observe that, with this matrix D in hand, one can compute the family of cores $\{S_\ell\}_{\ell=0}^{2k}$ and thus the score and weights in time $\tilde{O}(nM) = \tilde{O}(n/\varepsilon)$. Given the weights, computing the mean takes time $\tilde{O}(nd)$.

We must compute this matrix D of squared distances. As we pointed out, `StableCovariance` can return $\hat{\Sigma}^{-1}$, the inverse of the final weighted covariance. We rescale the data: $\tilde{x} \leftarrow x\hat{\Sigma}^{-1}$. With $\tilde{x}$ and x in hand, each entry in D can be computed in time $O(d)$.

Thus we can execute `StableMean` with one product $x\hat{\Sigma}^{-1}$ and $\tilde{O}(nd/\varepsilon)$ time.

Main Algorithm We can draw from $\mathcal{N}(\hat{\mu}, c^2\hat{\Sigma})$ in time $\tilde{O}(nd)$ by drawing $z \sim \mathcal{N}(0, \mathbb{I}_m)$ and returning $\hat{\mu} + y^T W^{1/2} z$, where $y \in \mathbb{R}^{m \times d}$ is the matrix of paired examples and W is the diagonal matrix whose entries are the weights computed by `StableCovariance`. (Digital computers do not sample exactly from Gaussian distributions, but we expect one can discretize appropriately, as in Canonne et al. (2020).)

To see that this is the correct distribution, note that $\hat{\Sigma} = y^T W y$ by definition and write out the expectation of $\left(y^T W^{1/2} z\right)\left(y^T W^{1/2} z\right)^T$.

3. Stable Covariance Estimation

In this section we present `StableCovariance`, our deterministic algorithm for nonprivate covariance estimation. Crucially, its parameter estimate is stable when the score it computes is small.

Lemma 21 (Lemma 11 Restated) *Fix data set size n, outlier threshold $\lambda_0 > 0$, and discretization parameter $k \in \mathbb{N}$. Assume $k \le \frac{n}{4e^2\lambda_0}$ and let $\gamma = \frac{16e^2\lambda_0}{n}$. Let x, x' be adjacent data sets and*

$$\Sigma_1, \texttt{SCORE} \leftarrow \texttt{StableCovariance}(x, \lambda_0, k)$$
$$\Sigma_2, \texttt{SCORE}' \leftarrow \texttt{StableCovariance}(x', \lambda_0, k).$$

Assume $\texttt{SCORE}, \texttt{SCORE}' < k$. Then $\Sigma_1, \Sigma_2 \succ 0$ and $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$. If $\gamma \le \frac{1}{2}$, then

$$\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}}, \left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}} \le (1+2\gamma)\gamma.$$

We can prove Lemma 21 quickly once we assemble the following definition and lemmas. The first step in Algorithm 2 is a standard trick to center the data: if $x_i, x_j \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$ then $\frac{1}{\sqrt{2}}(x_i - x_j) \sim \mathcal{N}(0, \Sigma)$. (A similar statement holds for subgaussian distributions.) With this in mind, the lemmas in this section will concern data sets y of size $m = \lfloor n/2 \rfloor$, where $y_i = \frac{1}{\sqrt{2}}\left(x_i - x_{i+\lfloor n/2 \rfloor}\right)$. We emphasize that this section is part of the *privacy* analysis and does not assume the data is in fact mean-zero. We analyze accuracy in Section 2.2

Definition 22 (Good Weighting/Subset) *Let $\lambda > 0$. For data set $y = y_1, \ldots, y_m$, a vector $w \in [0,1]^m$ is a λ-*good weighting of y *if, for $\Sigma_w = \sum_{j=1}^m w_j \cdot y_i y_i^T$,*

$$\forall i \in \mathrm{supp}(w), \quad \left\|\Sigma_w^{-1/2} y_i\right\|_2^2 \le \lambda.$$

*If the weights take the value $\frac{1}{m}$ over a set $S \subseteq [m]$ and are zero elsewhere, we will call S a λ-*good subset of y.

The weights can be thought of as a distribution over data points; our analysis is simpler if we do not require them to sum to one. Note that this definition requires invertibility of the covariance matrix.

Our proof of Lemma 21 requires the following lemma, which shows that if adjacent datasets y, y' have "similar" good weightings w, v respectively, then the corresponding weighted covariances are close in spectral and trace distances. Nearly identical calculations appeared in BGSUZ; for completeness we prove our version (restated as Lemma 53) in Appendix B.

Lemma 23 *Let $\lambda > 0$ and let y and y' be adjacent data sets of size m. Let w be a λ-good weighting for y and let v be a λ-good weighting for y' with $\|w - v\|_1 \leq \rho$ and $\|v\|_\infty, \|w\|_\infty \leq \eta$. Let Σ_w and Σ_v be the corresponding weighted covariances, as in Definition 22. Let $\gamma = \lambda(\rho + 2\eta)$.*

Then $\Sigma_w, \Sigma_v \succ 0$ and $(1-\gamma)\Sigma_w \preceq \Sigma_v \preceq \frac{1}{1-\gamma}\Sigma_w$. Furthermore, if $\gamma \leq \frac{1}{2}$, then

$$\left\|\Sigma_v^{-1/2}\Sigma_w\Sigma_v^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}}, \left\|\Sigma_w^{-1/2}\Sigma_v\Sigma_w^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}} \leq (1+2\gamma)\gamma.$$

This lemma, and its mean-estimation analog Lemma 36, are statements about "identifiability." Informally, identifiability is used in robust statistics via a standard blueprint: if y is a "good" data set (for some notion of "good") and y' is a corruption of y, then finding a "good" data set z that has high overlap with y' means z will also have high overlap with y, and "goodness" will ensure that empirical parameter estimates based on z therefore will be accurate for y.

Following BGSUZ, we use identifiability to establish privacy: on adjacent data sets we will produce "good" vectors of weights that are very similar. These will ultimately give rise to indistinguishable Gaussian distributions, from which we produce our final mean estimates.

To apply Lemma 23, we must argue that `StableCovariance` computes good weightings.

Lemma 24 *Fix data set $y \in \mathbb{R}^{m\times d}$, outlier threshold $\lambda_0 > 0$, and discretization parameter $k \in \mathbb{N}$. Assume $k \leq \frac{m}{2e^2\lambda_0}$. If* `StableCovariance` *computes* `SCORE` $< k$*, then the weights $\{w_i\}_{i\in[m]}$ it produces are a $2e^2\lambda_0$-good weighting of y.*

Finally, to prove Lemma 21 we must establish that on adjacent datasets, `StableCovariance` computes weightings which are close in ℓ_1 and not too large in ℓ_∞.

Lemma 25 *Fix data set size m, outlier threshold $\lambda_0 > 0$, and discretization parameter $k \in \mathbb{N}$. Assume $k \leq \frac{m}{2e^2\lambda_0}$. Let y and y' be adjacent data sets of size m. Let w be the weights computed by Algorithm 2 on y and let w' be the weights computed on y'. Assume in both cases we compute* `SCORE` $< k$*. Then $\|w - w'\|_1 \leq 2/m$. Furthermore, $\|w\|_\infty, \|w'\|_\infty \leq \frac{1}{m}$.*

Algorithm 2: `StableCovariance`(x, λ_0, k), for nonprivate covariance estimation

input: data set $x = (x_1, \ldots, x_n) \in \mathbb{R}^{n\times d}$, outlier threshold λ_0, discretization parameter $k \in \mathbb{N}$

$m \leftarrow \lfloor n/2 \rfloor$;
$\forall i \in [m], y_i \leftarrow \frac{1}{\sqrt{2}}(x_i - x_{i+m})$; `/* pair and rescale */`

for $\ell = 0, 1, \ldots, 2k$ **do**
  $S_\ell \leftarrow$ `LargestGoodSubset`$(y, e^{\ell/k}\lambda_0)$; `/* Algorithm 3 */`
end
`SCORE` $\leftarrow \min\{k, \min_{0\leq\ell\leq k}\{m - |S_\ell| + \ell\}\}$;
for $i = 1, \ldots, m$ **do**
  $w_i \leftarrow \frac{1}{km}\sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in S_\ell\}$;
end
$\hat{\Sigma} \leftarrow \sum_{i\in[m]} w_i \cdot y_i y_i^T$;
return $\hat{\Sigma}$, `SCORE`;

We are now ready to prove the main result of this section.

Proof [Proof of Lemma 21] On adjacent x and x', we pair and rescale to obtain y and y' (also adjacent). Lemma 25 says that we get vectors w and w' such that $\|w - w'\|_1 \leq 2/m$ and both $\|w\|_\infty$ and $\|w'\|_\infty$ are at most $\frac{1}{m}$. By Lemma 24, both w and w' are $2e^2\lambda_0$-good weightings of their respective data set. Setting $\rho = 2/m$ and $\lambda = 2e^2\lambda_0$, Lemma 23 tells us that, for

$$\gamma = \lambda(\rho + \|w\|_\infty + \|v\|_\infty) \leq \frac{8e^2\lambda_0}{m},$$

we have the desired inequalities. Finally, use $m = n/2$. ■

In the remainder of this section, we prove Lemma 23, Lemma 24, and Lemma 25. Finally, in Section 3.4, we prove that the score returned by `StableCovariance` is low-sensitivity.

Algorithm 3: `LargestGoodSubset`(y, λ), subroutine for covariance estimation

input: data set $y = (y_1, \ldots, y_m) \in \mathbb{R}^{m \times d}$, outlier threshold λ

`/* For vector` v `and singular matrix` A`, define` $\left\|A^{-1}v\right\|_2 = +\infty$ `*/`

$S \leftarrow [m]$;

repeat

  $\texttt{OUT} \leftarrow \left\{ i \in S : \left\| \left(\frac{1}{m} \sum_{j \in S} y_j y_j^T \right)^{-1/2} y_i \right\|_2^2 > \lambda \right\}$;

  $S \leftarrow S \setminus \texttt{OUT}$

until $\texttt{OUT} = \emptyset$;

return S;

3.1. Families of Largest Good Subsets

Before proceeding to the proofs of Lemmas 24 and 25, we establish a few facts about the operation of Algorithm 2.

Algorithm 2, `StableCovariance`, takes as input a discretization parameter $k \in \mathbb{N}$. For each $\ell = 0, 1, \ldots, 2k$, it calls `LargestGoodSubset` to find a set S_ℓ. Denote by $\{S_\ell\}_{\ell=0}^{2k}$ this family of subsets. In this subsection we prove that each S_ℓ is a unique largest good subset and discuss some properties of families of sets of this form.

Lemma 26 *For any data set y and $\lambda > 0$, there is a largest λ-good subset S^* that is unique and satisfies the following property: if S_0 is a λ-good subset, then $S_0 \subseteq S^*$. Furthermore, Algorithm 3 returns S^*.*

Proof Observe that Algorithm 3 either returns a λ-good subset or the empty set. We will show that, if y has a λ-good set S_0, then the set returned by Algorithm 3 contains S_0. This finishes the proof: for any two distinct λ-good sets, the output contains their union and thus must be strictly larger than both.

Suppose y has a λ-good set S_0. We prove that $S_0 \subseteq S$ is an algorithmic invariant, where S denotes the set operated upon in Algorithm 3. At initialization we have $S_0 \subseteq S = [m]$. Pick an

index $j \in S_0$ and consider a value of S at the start of the **repeat-until** block of Algorithm 3. We will show that j is not added to OUT. Let $\Sigma_S = \frac{1}{m}\sum_{i \in S} y_i y_i^T$ and similarly define Σ_{S_0}.

Removing points cannot increase the matrix in the positive semidefinite order, so

$$\Sigma_S = \frac{1}{m}\sum_{i \in S} y_i y_i^T \succeq \frac{1}{m}\sum_{i \in S_0} y_i y_i^T = \Sigma_{S_0}.$$

Since the Σ_{S_0} is invertible (by the goodness assumption), Σ_S is as well. So we also know that $\Sigma_S^{-1} \preceq \Sigma_{S_0}^{-1}$, and thus in turn

$$\|y_j\|_{\Sigma_S}^2 \leq \|y_j\|_{\Sigma_{S_0}}^2 .$$

The right term is at most λ by the λ-goodness assumption on S_0, so j is not removed as an outlier.

To finish, note that Algorithm 3 only terminates when it has found a λ-good subset. ■

The following lemma shows that families of good subsets computed on *adjacent* data sets are closely interwoven.

Lemma 27 *Let $\lambda_0 > 0$ and $k \in \mathbb{N}$. Let y and y' be adjacent data sets of size m that differ in index i^*. Let $\{S_\ell\}_{\ell=0}^{2k}$ be the family of sets where S_ℓ is the largest $e^{\ell/k}\lambda_0$-good subset for y. Let $\{T_\ell\}_{\ell=0}^{2k}$ be the corresponding family of largest $e^{\ell/k}\lambda_0$-good sets computed on y'. Assume $k \leq \frac{m}{2e^2\lambda_0}$. We have the following properties:*

(i) For any $0 \leq \ell \leq \ell' \leq 2k$, we have $S_\ell \subseteq S_{\ell'}$.

(ii) For any $0 \leq \ell < 2k$, we have $S_\ell \setminus \{i^*\} \subseteq T_{\ell+1}$.

Before we prove Lemma 27, we will prove a statement relating to the stability of good subsets (on a single data set y) under the removal of a index. If we remove an index j from a λ-good subset S, the covariance may shrink and the resulting subset may no longer be λ-good. However, we can show that the resulting set is λ'-good, where λ' is larger than λ by a small multiplicative factor that depends on λ and m.

Claim 28 *Let $\lambda > 0$ and $m \geq 2\lambda$. Let y be a data set of size m and let S be a λ-good subset for y. For any $j \in [m]$, $S \setminus \{j\}$ is a $e^{2\lambda/m}\lambda$-good subset for y.*

Proof Let $S' = S \setminus \{j\}$. Assume $j \in S$, otherwise we are done. By construction, we have $\Sigma_{S'} = \Sigma_S - \frac{1}{m} y_j y_j^T$. We claim that $\Sigma_{S'} \succeq \Sigma_S - (\lambda/m)\Sigma_S$. This is equivalent to $\frac{1}{m} y_j y_j^T \preceq (\lambda/m)\Sigma_S$, or

$$\Sigma_S^{-1/2} y_j y_j^T \Sigma_S^{-1/2} \preceq \lambda \mathbb{I}.$$

This holds because $\left\|\Sigma_S^{-1/2} y_j\right\|_2^2 \leq \lambda$ (since $j \in S$). By assumption $\frac{\lambda}{m}$ is at most $\frac{1}{2}$, so we have

$$\Sigma_{S'}^{-1} \preceq \frac{1}{1 - \lambda/m} \cdot \Sigma_S^{-1} \preceq e^{2\lambda/m} \cdot \Sigma_S^{-1}.$$

For any $i \in S' \subseteq S$ we had $\|y_i\|_{\Sigma_S}^2 \leq \lambda$, so we also have $\|y_i\|_{\Sigma_{S'}}^2 \leq e^{2\lambda/m}\lambda$. ■

Proof [Proof of Lemma 27] To prove statement (i), pick two indices $\ell \le \ell'$. We know that S_ℓ is an $e^{\ell'/k}\lambda_0$-good subset for y, because increasing the outlier threshold cannot cause any points in S_ℓ to become outliers. By construction, $S_{\ell'}$ is the largest $e^{\ell'/k}\lambda_0$-good subset for y. What's more, by Lemma 26, $S_{\ell'}$ contains all $e^{\ell'/k}\lambda_0$-good subsets. In particular, $S_\ell \subseteq S_{\ell'}$.

To prove statement (ii), consider $S_\ell \setminus \{i^*\}$, where data sets y and y' differ only in index i^*. We know that S_ℓ is an $e^{\ell/k}\lambda_0$-good subset for y, so by Claim 28 we know that $S_\ell \setminus \{i^*\}$ is λ'-good for

$$\begin{aligned}\lambda' &= \exp\left\{\frac{2}{m}\cdot e^{\ell/k}\lambda_0\right\}\cdot e^{\ell/k}\lambda_0 \\ &\le \exp\left\{\frac{2}{m}\cdot e^{2}\lambda_0\right\}\cdot e^{\ell/k}\lambda_0 \\ &\le e^{(\ell+1)/k}\lambda_0,\end{aligned}$$

where the first inequality uses $\ell \le 2k$ and the second uses $k \le \frac{m}{2e^2\lambda_0}$.

The next key fact is that, on the indices i in $S_\ell \setminus \{i^*\}$, we have $y_i = y'_i$. Therefore $S_\ell \setminus \{i^*\}$ is $e^{(\ell+1)/k}\lambda_0$-good for y'. Since, by Lemma 26, $T_{\ell+1}$ contains all such sets, $S_\ell \setminus \{i^*\} \subseteq T_{\ell+1}$. ■

3.2. The Weights Are Good (Proofs of Lemma 24)

Lemma 24 says that, when `StableCovariance` computes $\texttt{SCORE} < k$, the weights it produces are in fact a good weighting.

Proof [Proof of Lemma 24] For $\ell \in \{0,\dots,2k\}$, let S_ℓ denote the largest $e^{\ell/k}\lambda_0$-good subset for y. Let $\Sigma_\ell = \frac{1}{m}\sum_{i\in S_\ell} y_i y_i^T$. Algorithm 2 computes weights $w_i = \frac{1}{km}\sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in S_\ell\}$, meaning that we can rewrite the released $\hat{\Sigma}$ as

$$\hat{\Sigma} = \sum_{i=1}^{m} w_i \cdot y_i y_i^T = \frac{1}{k}\sum_{\ell=k+1}^{2k} \Sigma_\ell.$$

The larger sets contain the smaller ones, so we have a lower bound on $\hat{\Sigma}$ in terms of Σ_{k+1}:

$$\hat{\Sigma} \succeq \frac{1}{k}\sum_{\ell=0}^{2k} \Sigma_{k+1} = \Sigma_{k+1}.$$

Now, any two values ℓ and ℓ' such that $k+1 \le \ell \le \ell' \le 2k$ correspond to sets $S_\ell, S_{\ell'}$ that differ in at most k elements: S_ℓ is a subset of $S_{\ell'}$, $S_{\ell'}$ has size at most m, and S_ℓ has size at least $m-k+1 > m-k$. (To see this last fact, observe that $\texttt{SCORE} < k$ means there is an $\ell^* \le k$ such that $n - |S_{\ell^*}| + \ell^* < k$, which implies $|S_{\ell^*}| > n-k$. Then note that $S_{\ell^*} \subseteq S_\ell$, since $\ell^* \le k \le \ell$.) By the identifiability lemma for good sets, Lemma 23, we have $\Sigma_\ell \succeq (1-\gamma)\Sigma_{\ell'}$ for $\gamma = e^2\lambda_0 k/m$. (Both sets are $e^2\lambda_0$-good, since $\ell, \ell' \le 2k$.) This in turn implies an upper bound on the squared Mahalanobis distance for Σ_ℓ.

Pick any point $i \in \bigcup_{\ell=k+1}^{2k} S_\ell$. We know that $i \in S_{2k}$, so $\|y_i\|^2_{\Sigma_{2k}} \le e^2\lambda_0$. Using our above calculations, we have

$$\begin{aligned}\|y_i\|^2_{\hat{\Sigma}} &\le \|y_i\|^2_{\Sigma_{k+1}} \\ &\le (1 + 2e^2\lambda_0 k/m)\,\|y_i\|^2_{\Sigma_{2k}} \\ &\le (1 + 2e^2\lambda_0 k/m)\cdot e^2\lambda_0.\end{aligned}$$

This is at most $2e^2\lambda_0$, since we assumed $k \le \frac{m}{2e^2\lambda_0}$. ■

3.3. The Weights Are Stable (Proof of Lemma 25)

Lemma 25 says that, on adjacent inputs, `StableCovariance` produces nearly identical weight vectors (assuming the scores are not large).

Proof [Proof of Lemma 25] For $\ell \in \{0, \dots, 2k\}$, let S_ℓ denote the largest $e^{\ell/k}\lambda_0$-good subset for y and let T_ℓ denote the largest $e^{\ell/k}\lambda_0$-good subset for y'. Recall the notation from Algorithm 2: for each i we compute counts $c_i = \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in S_\ell\}$ and let $c_i + \Delta_i = \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in T_\ell\}$. Suppose y and y' differ in index i^*. Then $|\Delta_i^*| \le k$, since the counts are bounded between 0 and k. By Lemma 27, for all $i \neq i^*$, we know that $\Delta_i \in \{-1, 0, +1\}$.

We can bound the number of indices $i \neq i^*$ that have nonzero Δ_i. Because the score on y is less than k, the set S_k has size at least $m-k+1$. By Lemma 27, for every $\ell > k$ we have $S_k \setminus \{i^*\} \subseteq S_\ell$ and $S_k \setminus \{i^*\} \subseteq T_\ell$, so for all the indices $i \in S_k \setminus \{i^*\}$ we compute weight exactly $\frac{1}{m}$. So the number of indices $i \neq i^*$ with $\Delta_i \neq 0$ is at most k. Thus

$$\begin{aligned}
\left\|w - w'\right\|_1 &= \left|w_{i^*} - w'_{i^*}\right| + \sum_{i \in [m]\setminus\{i^*\}} \left|w_i - w'_i\right| \\
&= \left|\frac{c_{i^*}}{km} - \frac{c_{i^*} + \Delta_{i^*}}{km}\right| + \sum_{i \in [m]\setminus\{i^*\}} \left|\frac{c_i}{km} - \frac{c_i + \Delta_j}{km}\right| \\
&= \left|\frac{\Delta_{i^*}}{km}\right| + \sum_{i \in [m]\setminus\{i^*\}} \left|\frac{\Delta_i}{km}\right| \\
&\le \frac{1}{m} + k \cdot \frac{1}{km}.
\end{aligned}$$

Finally, note that all weights are at most $1/m$ by construction. ■

3.4. The Score Has Low Sensitivity

In this subsection we introduce the function $g(\cdot)$ that `StableCovariance` computes to determine the value of `SCORE`, which it returns. We prove that `SCORE` is low-sensitivity, which will allow us to use it as an input to $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}$, our propose-test-release subroutine.

Definition 29 *Fix data set y, outlier threshold $\lambda_0 > 0$, and discretization parameter $k \in \mathbb{N}$. For $\ell \in \{0, 1, \dots, k\}$, let S_ℓ denote the largest $e^{\ell/k}\lambda_0$-good subset for y, with $S_\ell = \emptyset$ if no good subset exists. Define*

- $f(y) \stackrel{def}{=} \min_{\ell \in \{0,\dots,k\}} \{m - |S_\ell| + \ell\}$ *and*
- $g(y) \stackrel{def}{=} \min\{f(y), k\}$.

The function $f(y)$ is small when there is λ not too much larger than λ_0 that yields a large set S_λ. In particular, when a data set y has $S = [m]$ as a λ_0-good subset, then $f(y) = g(y) = 0$.

For additional intuition, recall that a key subroutine in BGSUZ was computing the Hamming distance to the space of data sets with no outliers. Our score function approximates this quantity.

We make this precise in the following claim, which is stated to aid the reader and not directly used elsewhere.

Claim 30 *Let $\mathfrak{S}_\lambda$ be the space of size-m data sets where $[m]$ is a λ-good subset. Let $d(y, \mathfrak{S}_\lambda) = \min_{z \in \mathfrak{S}_\lambda} d(y, z)$ denote the Hamming distance to this space. If we compute $g(y) < k$, then we know that $d(y, \mathfrak{S}_{e\lambda_0}) \leq g(y) \leq 2 \cdot d(y, \mathfrak{S}_{\lambda_0})$.*

Proof For the upper bound, assume $d(y, \mathfrak{S}_{\lambda_0}) = r$, so there exists a z with $d(y, z) = r$ and $r \in \mathfrak{S}_{\lambda_0}$. Removing r points from z, we have an $e^{r/k}\lambda_0$-good subset of y with size $m - r$, so $g(y) \leq m - m + r + r = 2 \cdot d(y, \mathfrak{S}_{\lambda_0})$.

Now assume $g(y) = r < k$, so there exists some ℓ with $m - |S_\ell| + \ell = r$. We know that $S_\ell \subseteq S_k$, so augmenting S_ℓ with at most r copies of $\mathbf{0}$ yields a set $z \in \mathfrak{S}_{e\lambda_0}$ with $d(y, z) = r$. Thus $d(y, \mathfrak{S}_{e\lambda_0}) \leq r = g(y)$. ■

The main result of this section says that our score function has low sensitivity.

Lemma 31 *Fix data set size m, outlier threshold $\lambda_0 > 0$, and discretization parameter $k \in \mathbb{N}$. Assume $k \leq \frac{m}{2e^2\lambda_0}$. For all adjacent y and y' we have $|g(y) - g(y')| \leq 2$.*

We argued in Lemma 27 that, for certain parameters, if data set y has a large good set then adjacent y' must have a (slightly *smaller*) good set (for a slightly *larger* outlier threshold). Under these conditions, then, we have have $f(y) \approx f(y')$. These conditions might be violated when y has no good sets under modest λ, causing $f(y)$ to be large. To combat this, we introduce $g(y)$, which cannot be larger than k.

Proof [Proof of Lemma 31] Without loss of generality, assume $g(y) \leq g(y')$. We will show that $g(y') \leq g(y) + 2$ by analyzing two cases.

Case 1: Suppose $g(y) = k$. Then $g(y') \leq g(y) \leq g(y) + 2$ by construction, since $g(\cdot)$ is capped at k.

Case 2: Suppose $g(y) < k$. This can only happen when $f(y) < k$, so there exists an $\ell^* \in \{0, \ldots, k\}$ and subset S_{ℓ^*} that (i) is $e^{\ell^*/k}\lambda_0$-good for y and (ii) satisfies $m - |S_{\ell^*}| + \ell^* < k$. Furthermore, we know $\ell^* < k$, since $m - |S_{\ell^*}|$ is nonnegative.

We can now apply Lemma 27, since we have assumed $k \leq \frac{m}{2e^2\lambda_0}$. For $\ell \in \{0, \ldots, k\}$, let T_ℓ be the largest $e^{\ell/k}\lambda_0$-good subset for y'. Lemma 27 says that $S_{\ell^*} \setminus \{i^*\} \subseteq T_{\ell^*+1}$. This allows us to upper bound $g(y')$:

$$\begin{aligned} g(y') \leq f(y') &= \min_{\ell \in \{0,\ldots,k\}} m - |S_\ell| + \ell \\ &\leq m - |T_{\ell^*+1}| + \ell^* + 1 \\ &\leq m - (|S_{\ell^*}| - 1) + \ell^* + 1 \\ &= g(y) + 2. \end{aligned}$$

So we are done. ■

4. Stable Mean Estimation

In this section we present `StableMean`, our algorithm for nonprivate mean estimation, and prove its stability guarantee. It is based on a notion of "outlyingness" where inliers are sufficiently close to many other points in the data. More formally, one may think of a graph with n vertices corresponding to a data set x, where vertices i and j are adjacent if they are close. Under this interpretation, outliers are points with degree below some threshold.

`StableMean` takes as input a data set x, a matrix $\hat{\Sigma}$ that serves as a preconditioner, an outlier threshold λ_0, a "discretization" parameter k, and a set of "reference points" $R \subseteq [n]$.

We analyze this algorithm's stability with respect to simultaneously changing a single data point as well as slightly changing the preconditioner $\hat{\Sigma}$. The exact meaning of "slight change" we use is that of being multiplicatively close in the positive definite order, as in the consequence of Lemma 11.

The reference points $R \subseteq [n]$ are used to estimate which points are outliers: rather than check the distance to every one of the other $n-1$ points in the data set, we compute distances to a few randomly chosen data points. One can think of this as estimating the degree of vertices in the graph defined above: we will argue that the algorithm is private when R is sufficiently representative. A standard concentration argument says that a sufficiently large random R will be representative with probability at least $1-\delta$. (Such an argument appeared, with a similar application to private estimation, in the work of Tsfadia et al. (2022).)

Definition 32 (Degree-Representative) *Fix data set x, covariance Σ, outlier threshold λ_0, and reference set $R \subseteq [n]$. For all i, let*

$$N_i \stackrel{def}{=} \left\{ j \in [n] : \|x_i - x_j\|_\Sigma^2 \leq e^2 \lambda_0 \right\}$$
$$\tilde{N}_i \stackrel{def}{=} \left\{ j \in R : \|x_i - x_j\|_\Sigma^2 \leq e^2 \lambda_0 \right\}.$$

Let $z_i = |N_i|$ and $\tilde{z}_i = |\tilde{N}_i|$ be the sets' sizes. We say that R is degree-representative for x *if for every index i we have $\left| \frac{1}{|R|}\tilde{z}_i - \frac{1}{n} z_i \right| \leq \frac{1}{6}$.*

The main result in this section is the following stability guarantee for our nonprivate mean estimates on adjacent data sets.

Lemma 33 (Lemma 13 Restated) *Fix data set size n, dimension d, outlier threshold $\lambda_0 \geq 1$ and discretization parameter $k \in \mathbb{N}$. Use reference set $R \subseteq [n]$ with $|R| > 6k$. Let x and x' be adjacent d-dimensional data sets of size n. Let $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$ be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma = 16e^2 \frac{\lambda_0}{n}$. Assume that $k \leq \frac{1}{2\gamma}$. Let*

$$\hat{\mu}, \texttt{SCORE} \leftarrow \texttt{StableMean}(x, \Sigma_1, \lambda_0, k, R)$$
$$\hat{\mu}', \texttt{SCORE}' \leftarrow \texttt{StableMean}(x', \Sigma_2, \lambda_0, k, R).$$

If $\texttt{SCORE}, \texttt{SCORE}' < k$ and R is degree-representative for both x and x' (see Definition 32), then $\|\hat{\mu} - \hat{\mu}'\|_{\Sigma_1}^2 \leq \frac{(1+2\gamma)38e^2\lambda_0}{n^2}$.

As in Section 3, we assemble another definition and a few lemmas before proving Lemma 33.

In this section "outliers" are those points with insufficient degree. We formalize this notion (without explicit reference to graphs), below. We consider weighted subsets, as in Section 3.

Algorithm 4: $\texttt{StableMean}(x, \hat{\Sigma}, \lambda_0, k, R)$ for nonprivate mean estimation

input: data set $x = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$, covariance $\hat{\Sigma}$, outlier threshold λ_0, discretization parameter $k \in \mathbb{N}$, reference set $R \subseteq [n]$

for $\ell = 0, 1, \ldots, 2k$ **do**
  $S_\ell \leftarrow \texttt{LargestCore}(x, \hat{\Sigma}, e^{\ell/k}\lambda_0, R)$;
end
$\texttt{SCORE} \leftarrow \min\{k, \min_{0 \le \ell \le k}\{n - |S_\ell| + \ell\}\}$;
for $i = 1, \ldots, n$ **do**
  $c_i \leftarrow \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in S_\ell\}$;
end
$Z \leftarrow \sum_{i \in [n]} c_i$; /* normalizing constant */
$\forall i \in [n], w_i \leftarrow \frac{c_i}{Z}$; /* set $w_i \leftarrow 0$ if $Z = 0$ */
$\hat{\mu} \leftarrow \sum_{i \in [n]} w_i x_i$;
return $\hat{\mu}$, SCORE;

Algorithm 5: $\texttt{LargestCore}(x, \hat{\Sigma}, \lambda, \tau, R)$, subroutine for mean estimation

input: data set $x = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$, covariance $\hat{\Sigma}$, outlier threshold λ, degree threshold τ, reference set $R \subseteq [n]$

for $i \in [n]$ **do**
  $N_i = \left\{j \in R : \|x_i - x_j\|_\Sigma^2 \le \lambda\right\}$; /* Nearby points in R */
end
return $\{i \in [n] : |N_i| \ge \tau\}$;

Definition 34 (Weighted Core) *Let x be a data set of size n. Let $\lambda > 0$, $\tau \in \mathbb{N}$, and $R \subseteq [n]$. A distribution w over $[n]$ is a (τ, λ, Σ)*-weighted-core for x with respect to R *if, for each $i \in \text{supp}(w)$, there is a set $N_i \subseteq R$ of size at least τ such that $\forall j \in N_i$, $\|x_i - x_j\|_\Sigma^2 \leq \lambda$. If w is uniform over a subset S, we will refer to S as a (τ, λ, Σ)*-core for x with respect to R.

Remark 35 *For a given data set x, reference set R, and parameters τ, λ, and Σ, one can directly find $S^* \subseteq [n]$, the largest (τ, λ, Σ)-core for x with respect to R (if y_i has τ nearby points in R, put i in S^*, otherwise omit it). Contrast this with Section 3, where it was nontrivial that an efficient algorithm recovers the largest λ-good sets. In addition to such an algorithmic guarantee, Lemma 26 established that the largest λ-good set is unique and contains all other λ-good sets. The analogous observation is obvious for (τ, λ, Σ)-cores.*

The first statement we need in the proof of Lemma 33 tells us that, if two weighted cores (on adjacent data sets) are close in ℓ_1 distance, then their means are close. We give the proof of Corollary 36 (restated as Corollary 55) in Appendix B; nearly identical calculations appeared in BGSUZ.

Corollary 36 *Let x and x' be adjacent data sets of size n. Let $\lambda > 0$, $\tau \in \mathbb{N}$, and $R \subseteq [n]$. Let Σ_1 and Σ_2 be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Let w be a $(\tau, \lambda, \Sigma_1)$-weighted-core for x with respect to R and let v be a $(\tau, \lambda, \Sigma_2)$-weighted-core for x' with respect to R, with $\tau > \frac{n}{2} + 1$. If $\|w - v\|_1 \leq \rho$ then*

$$\|\mu_w - \mu_v\|_{\Sigma_1}^2 = \left\| \sum_{i \in [n]} w_i x_i - \sum_{i \in [n]} v_i x_i' \right\|_{\Sigma_1}^2 \leq (1 + 2\gamma)\lambda(\rho + \|w\|_\infty + \|v\|_\infty)^2.$$

To apply Corollary 36, we need to show that the weights produced by `StableMean` are in fact a weighted core for x.

Lemma 37 *Fix a set of inputs to* `StableMean` *(Algorithm 4): data set x, covariance $\hat{\Sigma}$, outlier threshold λ_0, discretization parameter $k \in \mathbb{N}$, and reference set $R \subseteq [n]$. Assume $|R| > 6k$. If* `StableMean` *computes* `SCORE` $< k$ *and R is degree-representative for x, then the weights $w \in [0,1]^n$ it produces are an $\left(\frac{n}{2} + 1, e^2\lambda_0, \hat{\Sigma}\right)$-weighted-core for x with respect to $[n]$.*

Additionally, the weights produced by `StableMean` on adjacent data sets must be close in ℓ_1 distance.

Lemma 38 *Fix outlier threshold $\lambda_0 \geq 1$, discretizaion parameter $k \in \mathbb{N}$, and reference set $R \subseteq [n]$. Let x and x' be adjacent data sets of size n. Let Σ_1 and Σ_2 be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Further assume $k \leq \frac{1}{2\gamma}$.*

Let w be the weights Algorithm 4 computes on inputs $(x, \Sigma_1, \lambda_0, k, R)$ and w' the weights it computes on inputs $(x', \Sigma_2, \lambda_0, k, R)$. Assume in both cases that it computes `SCORE` $< k$. *Then*

$$\|w - w'\|_1 \leq \frac{4(1 + 2k/n)^2}{n}.$$

Furthermore $\|w\|_\infty, \|w'\|_\infty \leq \frac{1+2k/n}{n}$.

We are now ready to prove the stability guarantee for the nonprivate mean estimate.

Proof [Proof of Lemma 33] Let w be the vector of weights computed by `StableMean` on input $(x, \Sigma_1, \lambda_0, k, R)$ and w' the weights computed on input $(x', \Sigma_2, \lambda_0, k, R)$. By Lemma 37 and the fact that R is degree-representative for x, w is an $(n/2+1, e^2\lambda_0, \Sigma_1)$-weighted-core for x. For the same reasons, w' is an $(n/2+1, e^2\lambda_0, \Sigma_2)$-weighted-core for x'. By Lemma 38, w and w' satisfy $\|w - w'\|_1 \leq \frac{4(1+2k/n)^2}{n}$ and $\|w\|_\infty, \|w'\|_\infty \leq \frac{(1+2k/n)}{n}$. Thus, setting $\rho = \frac{4(1+2k/n)^2}{n}$ and $\lambda = e^2\lambda_0$, Corollary 36 tells us that the means must be close:

$$\begin{aligned}\left\|\hat{\mu} - \hat{\mu}'\right\|_{\Sigma_1}^2 &\leq 2\lambda(\rho + \|w\|_\infty + \|v\|_\infty)^2 \\ &\leq (1+2\gamma)e^2\lambda_0 \cdot \left(\frac{6(1+2k/n)^2}{n}\right)^2.\end{aligned}$$

To clean this expression up, note that $\gamma = \frac{16e^2\lambda_0}{n}$ and $k \leq \frac{1}{2\gamma}$ imply

$$\frac{2k}{n} \leq \frac{1}{16e^2\lambda_0} \leq \frac{1}{16e^2}.$$

Thus $(1 + \frac{2k}{n})^4 \leq 1.04$ and we arrive at $\|\hat{\mu} - \hat{\mu}'\|_{\Sigma_1}^2 \leq \frac{(1+2\gamma)38e^2\lambda_0}{n^2}$. ■

4.1. Families of Largest Cores

Before presenting the proofs of Lemmas 37 and 38, we establish a few facts about the operation of `StableMean`, Algorithm 4.

Fix a reference set $R \subseteq [n]$ and a discretization parameter $k \in \mathbb{N}$. For each $\ell = 0, 1, \ldots, 2k$, `StableMean` calls `LargestCore` (Algorithm 5) to find the largest $\left(|R| - \ell, e^{\ell/k}\lambda_0, \hat{\Sigma}\right)$-core for x with respect to R, where $\hat{\Sigma}$ is the nonprivate estimate produced by `StableCovariance`. Denote by $\{S_\ell\}_{\ell=0}^{2k}$ this family of largest cores for x with respect to R. In this section we will discuss some properties of families of this form. As in Section 3 and Lemma 27, we show that families of cores on adjacent data sets are tightly interwoven.

Lemma 39 *Fix $R \subseteq [n]$, $\lambda_0 > 0$, and $k \in \mathbb{N}$. Let x and x' be adjacent data sets that differ in index i^*. Let Σ_1 and Σ_2 be positive definite matrices that satisfy $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Further assume $k \leq \frac{1}{2\gamma}$.*

Let $\{S_\ell\}_{\ell=0}^{2k}$ be the family of sets where S_ℓ is the largest $(|R| - \ell, e^{\ell/k}\lambda_0, \Sigma_1)$-core for x with respect to R. Let $\{T_\ell\}_{\ell=0}^{2k}$ the corresponding family of largest $(|R| - \ell, e^{\ell/k}\lambda_0, \Sigma_2)$-cores for x' with respect to R. The following properties hold:

(i) For any $0 \leq \ell \leq \ell' \leq 2k$, we have $S_\ell \subseteq S_{\ell'}$.

(ii) For any $0 \leq \ell < 2k$, we have $S_\ell \setminus \{i^*\} \subseteq T_{\ell+1}$.

Proof Statement (i) is direct from our definition of core: decreasing the neighbor threshold and increasing the outlier threshold results in a less restrictive condition, so S_ℓ is a $(|R| - \ell', e^{\ell'/k}\lambda_0, \Sigma_1)$-core for x with respect to R. Since $S_{\ell'}$ contains all such cores (recall Remark 35), $S_\ell \subseteq S_{\ell'}$.

When $\Sigma_1 \succeq (1-\gamma)\Sigma_2$, for any vector v we have $\|v\|_{\Sigma_1}^2 \leq \frac{1}{1-\gamma}\|v\|_{\Sigma_2}^2$. Now, $S_\ell \setminus \{i^*\}$ is a $(|R|-\ell, e^{\ell/k}\lambda_0, \Sigma_1)$-core for x with respect to R: for all $i \in S_\ell \setminus \{i^*\}$ there exists a set $N_i \subseteq R$ such that

$$|N_i| \geq |R| - \ell \quad \text{and} \quad \forall j \in N_i,\ \|x_i - x_j\|_{\Sigma_1}^2 \leq e^{\ell/k}\lambda_0.$$

(This condition holds for all $i \in S_\ell$, so it holds for all $i \in S_\ell \setminus \{i^*\}$.) Changing the covariance, we see that $S_\ell \setminus \{i^*\}$ is a $(|R|-\ell, \lambda', \Sigma_2)$-core for x with respect to R, where $\lambda' = \frac{1}{1-\gamma}e^{\ell/k}\lambda_0$. We assumed $\gamma \leq \frac{1}{2}$ and $k \leq \frac{1}{2\gamma}$, so $\frac{1}{1-\gamma} \leq 1 + 2\gamma \leq e^{2\gamma} \leq e^{1/k}$, which means $\lambda' \leq e^{(\ell+1)/k}\lambda_0$.

We now want to argue that $S_\ell \setminus \{i^*\}$ is a core for x'. This is not trivial, since i^* could still be a member of our reference set. We must decrease the degree threshold by one: for all $i \in S_\ell \setminus \{i^*\}$, there exists a set $N_i' = N_i \setminus \{i^*\} \subseteq R$ such that

$$\left|N_i'\right| \geq |R| - (\ell+1) \quad \text{and} \quad \forall j \in N_i',\ \left\|x_i' - x_j'\right\|_{\Sigma_1}^2 \leq e^{(\ell+1)/k}\lambda_0,$$

where we have used the fact that for all i in $S_\ell \setminus \{i^*\}$ or $N_i' = N_i \setminus \{i^*\}$ we have $x_i = x_i'$. Thus $S_\ell \setminus \{i^*\}$ is a $(|R| - (\ell+1), e^{(\ell+1)/k}\lambda_0, \Sigma_2)$-core for x' with respect to R, which implies $S_\ell \setminus \{i^*\} \subseteq T_{\ell+1}$ as we previously argued. ■

4.2. The Weights Are Good (Proof of Lemma 37)

Lemma 37 says that, when `StableMean` on input x computes `SCORE` $< k$ and the set of reference points is degree-representative, the weights produced are a core for x.

Proof [Proof of Lemma 37] For $\ell \in \{0, \ldots, 2k\}$, let S_ℓ denote the largest $(|R| - \ell, e^{\ell/k}\lambda_0, \hat{\Sigma})$-core for x with respect to R. We know that S_{2k} contains all such cores (recall Remark 35). In particular, if an index i is in the support of w, then $i \in S_{2k}$. This directly implies that w is an $(|R| - 2k, e^2\lambda_0, \hat{\Sigma})$-weighted-core for x with respect to R. (As a technical point, note that `SCORE` $< k$ implies that there exists an $\ell^* < k$ with $S_{\ell^*} \neq \emptyset$, so the weights returned are not the zero vector.)

Now suppose for contradiction that the weights w are not an $(n/2+1, e^2\lambda_0, \hat{\Sigma})$-weighted-core for x with respect to $[n]$. Then there is an index $i \in \mathrm{supp}(w)$ such that $z_i \leq n/2$ (here z_i and $\tilde{z}_i$ are as in Definition 32). Since R is degree-representative for x, this implies that $\frac{1}{|R|}\tilde{z}_i \leq \frac{2}{3}$. However, since i is in the weighted core, it has many neighbors in R: $\tilde{z}_i$ is at least $|R| - 2k > |R| - \frac{2|R|}{3} = \frac{2|R|}{3}$. Thus we have arrived at a contradiction. ■

4.3. The Weights Are Stable (Proof of Lemma 38)

Lemma 38 says that `StableMean` produces nearly identical weight vectors on adjacent inputs (assuming the scores are not too large).

Proof [Proof of Lemma 38] For $\ell \in \{0, \ldots, 2k\}$, let S_ℓ denote the largest $(|R| - \ell, e^{\ell/k}\lambda_0, \hat{\Sigma})$-core for x with respect to R. Similarly let T_ℓ denote the largest $(|R| - \ell, e^{\ell/k}\lambda_0, \Sigma_2)$-core for x' with respect to R. Recall notation from Algorithm 4: for each i we compute counts $c_i = \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in S_\ell\}$, normalizing constant $Z = \sum_{i \in [n]} c_i$, and weights $w_i = \frac{c_i}{Z}$. Let c_i', Z', and w_i' denote the same values computed on x'. As in the proof of Lemma 25, let $\Delta_i = c_i - c_i'$.

We first show that the normalizing constants computed under x and x' are similar. Suppose x and x' differ on index i^*. We know that $|\Delta_{i^*}| \leq k$, since the counts are bounded between 0 and k.

Lemma 39 tells us that, for all $i \neq i^*$, $\Delta_i \in \{-1, 0, +1\}$. We also have an upper bound on the number of indices $i \neq i^*$ where $\Delta_i \neq 0$. Since we computed $\texttt{SCORE} < k$ in both cases, the core S_k has size at least $n - k + 1$. Furthermore, for all $\ell > k$ we know that $S_k \setminus \{i^*\} \subseteq S_\ell$ and $S_k \setminus \{i^*\} \subseteq T_\ell$. Thus, for all but at most k indices $i \neq i^*$ (namely, those indices not in $S_k \setminus \{i^*\}$), we have $\Delta_i = 0$. So $|Z - Z'| = |\sum_i c_i - c_i - \Delta_i| \leq 2k$, since the i^* term may have magnitude k and at most k indices have magnitude one.

The same facts imply that Z' is large:

$$Z' = \sum_{i=1}^{n} \sum_{\ell=k+1}^{2k} \mathbb{1}\{i \in T_\ell\} = \sum_{\ell=k+1}^{2k} |T_\ell| \geq k \cdot |S_k \setminus \{i^*\}|.$$

This is true for Z, too, so we have $Z, Z' \geq k(n-k)$.

We can now break the ℓ_1 difference in weights across three cases: i^*, those $i \neq i^*$ where $\Delta_i = 0$, and those $i \neq i^*$ where $\Delta_i \neq 0$.

$$\begin{aligned} \left\|w - w'\right\|_1 &= \sum_{i \in [n]} \left| \frac{c_i}{Z} - \frac{c_i}{Z'} - \frac{\Delta_i}{Z'} \right| \\ &\leq \sum_{i \in [n]} c_i \left| \frac{1}{Z} - \frac{1}{Z'} \right| + \left| \frac{\Delta_i}{Z'} \right| \\ &\leq n \cdot \left| \frac{1}{Z} - \frac{1}{Z'} \right| \cdot \max_i \{c_i\} + \frac{2k}{Z'}, \end{aligned} \tag{7}$$

again arriving at $2k$ by combining the (at most k indices) where $|\Delta_i| = 1$ with the (one, namely i^*) index that has $|\Delta_{i^*}| \leq k$. Because Z' is large, the second term is at most $\frac{2}{n-k} \leq \frac{2(1+2k/n)}{n}$. The first term in Equation (7) is also small: we have

$$\begin{aligned} \left| \frac{1}{Z} - \frac{1}{Z'} \right| &= \left| \frac{Z'}{ZZ'} - \frac{Z}{ZZ'} \right| \\ &\leq \frac{2k}{k^2(n-k)^2}. \end{aligned}$$

Since $c_i \leq k$ we can bound the first term as

$$n \cdot \left| \frac{1}{Z} - \frac{1}{Z'} \right| \cdot \max_i \{c_i\} \leq \frac{2nk^2}{k^2(n-k)^2} \leq \frac{2(1+2k/n)^2}{n}.$$

We combine this with the second term in Equation (7) by using $(1+2k/n)^2 \geq (1+2k/n)$. Finally, note that all the weights are at most $\frac{1}{n-k} \leq \frac{1+2k/n}{n}$. ∎

4.4. The Score Has Low Sensitivity

In this subsection we introduce the function $g(\cdot)$ that `StableMean` computes to determine the value of `SCORE`, which it returns along with the mean estimate. We prove that `SCORE` has low sensitivity, which will allow us to use it as an input to $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}$, our propose-test-release subroutine.

Definition 40 *Fix x, Σ, λ_0, k, and $R \subseteq [n]$. For $\ell \in \{0, 1, \ldots, k\}$, let S_ℓ denote the largest $(|R| - \ell, e^{\ell/k}\lambda_0, \Sigma)$-core for x with respect to R, with $S_\ell = \emptyset$ if no such core exists. Define*

- $f(x, \Sigma) \stackrel{def}{=} \min_{\ell \in \{0,\ldots,k\}} (n - |S_\ell| + \ell)$ *and*
- $g(x, \Sigma) \stackrel{def}{=} \min\{f(x, \Sigma), k\}$.

Lemma 41 *Fix $\lambda_0 > 0$, $k \in \mathbb{N}$, and $R \subseteq [n]$. Let x and x' be adjacent data sets. Let Σ_1 and Σ_2 be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Further assume $k \leq \frac{1}{2\gamma}$. Then $|g(x, \Sigma_1) - g(x', \Sigma_2)| \leq 2$.*

This proof is nearly identical to that of Lemma 41, except we apply Lemma 39 instead of Lemma 27. For completeness, Appendix B contains a restatement (as Claim 56) and proof.

5. Private Covariance Estimation and Fast Learning of Gaussians

As discussed in Section 1.1, our privacy analysis argues that on adjacent inputs x and x' we either fail or produce covariance estimates Σ_1 and Σ_2 such that $\mathcal{N}(0, \Sigma_1) \approx_{(\varepsilon,\delta)} \mathcal{N}(0, \Sigma_2)$. As Alabi et al. (2022) point out, with a stronger stability guarantee one could offer indistinguishability guarantees for multiple independent draws from these distributions. With enough such samples, one could form an accurate private estimate of the covariance matrix.

We formalize this connection below. To simplify our presentation, we assume $d = \Omega(\log n)$ and focus on guarantees for Gaussian data that hold with high constant probability. We state the guarantees for spectral norm; one obtains the guarantees for Frobenius norm immediately via the inequality $\|\cdot\|_F \leq \sqrt{d}\,\|\cdot\|_2$.

Theorem 42 *Fix $\varepsilon \in (0, 1)$, $\delta \in (0, \varepsilon/10)$, and $n, d \in \mathbb{N}$. Assume $d = \Omega(\log n)$. Algorithm 6 takes a data set of n points, each in $\mathbb{R}^d$, privacy parameters ε, δ, and an outlier threshold λ_0.*

- *For any $\lambda_0 \geq 1$, Algorithm 6 is (ε, δ)-differentially private.*
- *Suppose x is drawn i.i.d. from $\mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite. There exists absolute constants K_1, K_2, and K_3 such that, if $\lambda_0 = K_1 d$ and $n \geq K_2 d \log(1/\delta)/\varepsilon$, then with high constant probability Algorithm 6 does not fail and instead returns a positive semidefinite matrix $\tilde{\Sigma} \in \mathbb{R}^{d\times d}$ such that*

$$\left\|\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} - \mathbb{I}\right\|_2 \leq K_3\left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}\sqrt{\log 1/\delta}}{\varepsilon n}\right).$$

- *Algorithm 6 can be implemented to require: one product of the form A^TA for $A \in \mathbb{R}^{n\times d}$, one product of the form AB for $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{d\times d}$, one product of the form A^TA for $A \in \mathbb{R}^{N\times d}$, one product of the form AB for $A \in \mathbb{R}^{N\times d}$ and $B \in \mathbb{R}^{d\times d}$, one inversion and one eigenvalue decomposition of a positive definite matrix in $\mathbb{R}^{d\times d}$ to logarithmic bit complexity, and further computational overhead of $\tilde{O}(nd/\varepsilon)$. Here N is the number of synthetic samples, set in Algorithm 6.*

Claim 43 connects Gaussian parameter estimation (the mean to low Mahalanobis error, the covariance to low Frobenius error) to learning the distribution itself in low total variation distance. With this fact, the utility guarantee for Algorithm 7 (which simply runs Algorithms 1 and 6) is immediate. The privacy guarantee follows basic composition (Fact 51).

Claim 43 (Diakonikolas et al. (2016)) *There exists a constant K such that, for any $\alpha \le \frac{1}{2}$, vectors $\mu_1, \mu_2 \in \mathbb{R}^d$, and positive definite $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$, if $\|\mu_1 - \mu_2\|_{\Sigma_1} \le \alpha$ and $\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}\right\|_F \le \alpha$ then $\mathrm{TV}(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)) \le K\alpha$.*

Corollary 44 *Fix $\varepsilon \in (0,1)$, $\delta \in (0, \varepsilon/10)$, and $n, d \in \mathbb{N}$. Assume $d = \Omega(\log n)$. Algorithm 7 takes a data set of n points, each in $\mathbb{R}^d$, privacy parameters ε, δ, and an outlier threshold λ_0.*

- *For any $\lambda_0 \ge 1$, Algorithm 7 is $(2\varepsilon, 2\delta)$-differentially private.*
- *Suppose $x \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\Sigma)$ where Σ is positive definite. There exists absolute constants K_1 and K_2 such that, if $\lambda_0 = K_1 d$ and*

$$n \ge K_2\left(\frac{d^2}{\alpha^2} + \frac{d^2\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d\log 1/\delta}{\varepsilon}\right),$$

then, with high constant probability, Algorithm 7 does not fail and instead returns a pair $(\tilde\mu, \tilde\Sigma)$ such that $\mathrm{TV}(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\tilde\mu,\tilde\Sigma)) \le \alpha$.

5.1. Discussion of Private Covariance Estimation

The topic of differentially private Gaussian covariance estimation has received much attention recently. Under the assumption that the true covariance satisfies $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$ for some $\kappa \ge 1$, Kamath, Li, Singhal, and Ullman (2019) provided efficient algorithms for learning in both the spectral and Frobenius norms. These algorithms are nearly optimal but require a priori information in the form of κ. Their error depends only polylogarithmically on κ (in contrast with prior approaches, with error $\Omega(\mathrm{poly}(\kappa))$), but such prior knowledge is not generally available). In what follows, we focus on learning *unrestricted* covariances, when we do not have such prior knowledge.

With $n \gtrsim d^{3/2}$ samples, Algorithm 6 returns a private covariance estimate that is accurate in spectral norm. To the best of our knowledge, this is the first differentially private polynomial-time algorithm achieving such a guarantee for unrestricted Gaussian distributions, closing an open question posed by Alabi et al. (2022). This dependence on the dimension is known to be optimal in the regime of $\alpha = O(1/\sqrt{d})$ (Kamath et al., 2022) and matches the standard Gaussian mechanism applied under the assumption that $\mathbb{I} \preceq \Sigma \preceq O(1)\mathbb{I}$.

In the past two years, many papers have established techniques for privately learning the covariance of unrestricted Gaussian distributions to α error in rescaled Frobenius norm, also called the Mahalanobis norm for matrices. Such an estimator, combined with a standard mean estimation procedure, allows one to learn the entire distribution to $O(\alpha)$ total variation distance (see Claim 43). The optimal sample complexity for this task is

$$n \gtrsim \frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{\log 1/\delta}{\varepsilon}.$$

The information-theoretic upper bound is due to Aden-Ali et al. (2021); the first term is required for nonprivate estimation, the third is required even for estimating the mean of a univariate Gaussian with known variance and unrestricted mean (Karwa and Vadhan, 2017), and the second was recently proved to be tight (Kamath et al., 2022). The upper bound of Aden-Ali et al. (2021) is nonconstructive; later work by Liu et al. (2022) gave an exponential-time algorithm with nearly the same guarantees. A group of concurrent and independent papers soon gave polynomial-time algorithms (Kamath et al., 2021; Ashtiani and Liaw, 2021; Kothari et al., 2021), with Ashtiani and Liaw (2021) achieving the optimal d^2 dimension-dependence. (Using tools from Ashtiani and Liaw (2021), the framework of Tsfadia et al. (2022) achieves a similar error bound.) Very recently, simultaneous work by Hopkins et al. (2022) and Alabi et al. (2022) gave polynomial-time and sample-optimal estimators that also satisfy robustness against adversarial perturbation to the input. The former's guarantees for this task are also the best-known in polynomial time, matching Equation 5.1 up to logarithmic factors in d, α, ε, and $\log 1/\delta$.

Our work (combining Algorithms 1 and 6) matches the optimal sample complexity up to logarithmic factors and runs in time $\tilde{O}(nd^{\omega-1} + nd/\varepsilon)$. To the best of our knowledge, the fastest known algorithm takes time $\tilde{O}(nd^{\omega-1} + d^{\omega+1}/\varepsilon)$ and arises by combining ideas of Ashtiani and Liaw (2021) and Tsfadia et al. (2022). For this task's typical parameter setting of $n \gtrsim d^2/\varepsilon$, both of these expressions are asymptotically dominated by their first terms, corresponding to the time needed to compute the covariance of the entire input data.

We remark that the exact computation of eigenvalue decompositions is not possible; suitable approximation algorithms are well-understood (see, e.g., Trefethen and Bau, 1997).

Algorithm 6: Private Covariance Estimation, $\mathcal{A}_{\text{cov}}^{\varepsilon,\delta,\lambda_0}(x)$

input : data set $x \in \mathbb{R}^{n\times d}$, privacy parameters ε, δ, outlier threshold λ_0
require: $n \geq \frac{272e^2\lambda_0 \log 2/\delta}{\varepsilon}$
$k \leftarrow \left\lceil \frac{4\log 2/\delta}{\varepsilon} \right\rceil + 4$;
$N \leftarrow \left\lfloor 10^{-6} \cdot \frac{n^2\varepsilon^2}{\lambda_0^2 \log 2/\delta} \right\rfloor$;
$\hat{\Sigma}, \texttt{SCORE} \leftarrow \texttt{StableCovariance}(x, \lambda_0, k)$; /* $\hat{\Sigma} \in \mathbb{R}^{d\times d}, \texttt{SCORE} \in \mathbb{N}$ */
if $\mathcal{M}_{\text{PTR}}^{\varepsilon/2,\delta/2}(\texttt{SCORE}) = \texttt{PASS}$ **then**
  Draw $Z_1, \ldots, Z_N \overset{iid}{\sim} \mathcal{N}(0, \hat{\Sigma})$;
  return $\frac{1}{N}\sum_{i=1}^N Z_i Z_i^T$;
else
  return FAIL
end

5.2. Proof of Theorem 42

We use the following lemma, which can be interpreted as an analog of Claim 12 (used in BGSUZ) that uses advanced composition, requiring roughly a factor of $\sqrt{N}$ more samples to repeat the process N times.

Claim 45 (Alabi et al. (2022), Theorem 5.1) *Fix $\varepsilon, \delta \in (0,1)$. Let Σ_1, Σ_2 be positive definite matrices satisfying*

$$\left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \mathbb{I}\right\|_F, \left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}\right\|_F = \Delta$$

for any Δ satisfying $\Delta \le \frac{\varepsilon}{8\log 1/\delta} < 1$. For any $N \le \frac{\varepsilon^2}{8\Delta^2 \log 1/\delta}$, we have $\mathcal{N}(0,\Sigma_1)^{\otimes N} \approx_{(\varepsilon,\delta)} \mathcal{N}(0,\Sigma_2)^{\otimes N}$, where $p^{\otimes N}$ denotes the N-fold product distribution, i.e., N independent copies of p.

Proof [Proof of Theorem 42] Set $\varepsilon' = \varepsilon/2$ and $\delta' = \delta/2$. Recall from Lemma 11 that, when the scores are below k, we have $\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - \mathbb{I}\right\|_{\mathrm{tr}} \le (1+2\gamma)\gamma$ for $\gamma = 16e^2\lambda_0/n$ (and the same inequality with Σ_1 and Σ_2 swapped). Thus, to apply Claim 45, we use the fact that $\|\cdot\|_F \le \|\cdot\|_{\mathrm{tr}}$ and require $(1+2\gamma)\gamma \le \frac{\varepsilon'}{8\log 1/\delta} = \frac{\varepsilon}{16\log 2/\delta}$. Our assumptions on ε and δ mean that $\gamma \le 0.03$, so $(1+2\gamma) \le 1.06$ and $(1+2\gamma)\gamma \le 17e^2\lambda_0/n$. Thus, to draw a single sample we require $n \ge \frac{272e^2\lambda_0 \log 2/\delta}{\varepsilon}$; otherwise we fail. This is a stronger assumption than $k \le \frac{n}{4e^2\lambda_0}$, which Lemma 11 also requires. To draw N samples, we require $\frac{n^2}{17^2e^4\lambda_0^2} \cdot \frac{\varepsilon^2}{32\log 2/\delta}$. (The constant in Algorithm 6 is 10^6, which is larger than $17^2 \times e^4 \times 32$.) The remainder of the privacy argument follows from the guarantees of propose-test-release (Lemma 9 for sensitivity of the score function and Claim 10 for privacy of $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon,\delta}$) and basic composition.

We analyzed the running time of `StableCovariance` in Section 2.3. To generate the private covariance estimate, we first we compute the eigenvalue decomposition of $\hat{\Sigma}$ and from it produce $\hat{\Sigma}^{1/2}$. We then draw N samples from $\mathcal{N}(0,\mathbb{I})$ (in time $\tilde{O}(Nd)$), rescale them by $\hat{\Sigma}^{1/2}$, and compute the empirical covariance of the rescaled points.

For accuracy, first note that we have assumed $n \gtrsim d\log(1/\delta)/\varepsilon$ so that the input requirement on n is satisfied and we do not fail immediately. Beyond that, the accuracy argument uses two straightforward applications of Claim 49. Let Σ_y be the empirical covariance of y and $\tilde{\Sigma}$ the private estimate we release. We add and subtract Σ_y and apply the triangle inequality:

$$\begin{aligned}
\left\|\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} - \mathbb{I}\right\|_2 &= \left\|\Sigma^{-1/2}\Big(\tilde{\Sigma} - \Sigma\Big)\Sigma^{-1/2}\right\|_2 \\
&= \left\|\Sigma^{-1/2}\Big(\tilde{\Sigma} - \Sigma_y + \Sigma_y - \Sigma\Big)\Sigma^{-1/2}\right\|_2 \\
&\le \left\|\Sigma^{-1/2}\Big(\tilde{\Sigma} - \Sigma_y\Big)\Sigma^{-1/2}\right\|_2 + \left\|\Sigma^{-1/2}(\Sigma_y - \Sigma)\Sigma^{-1/2}\right\|_2 \\
&= \left\|\Sigma^{-1/2}\Big(\tilde{\Sigma} - \Sigma_y\Big)\Sigma^{-1/2}\right\|_2 + \left\|\Sigma^{-1/2}\Sigma_y\Sigma^{-1/2} - \mathbb{I}\right\|_2.
\end{aligned}$$

Since $n \gtrsim d + \log 1/\beta$, with high probability Σ_y is invertible and $\left\|\Sigma^{1/2}\Sigma_y^{-1/2}\right\|_2 = 1 + o(1)$, so applying Cauchy-Schwarz we have

$$\begin{aligned}
\left\|\Sigma^{-1/2}\Big(\tilde{\Sigma} - \Sigma_y\Big)\Sigma^{-1/2}\right\|_2 &= \left\|\Sigma^{-1/2}\Big(\Sigma_y^{1/2}\Sigma_y^{-1/2}\Big)\Big(\tilde{\Sigma} - \Sigma_y\Big)\Big(\Sigma_y^{-1/2}\Sigma_y^{1/2}\Big)\Sigma^{-1/2}\right\|_2 \\
&\lesssim \left\|\Sigma_y^{-1/2}\Big(\tilde{\Sigma} - \Sigma_y\Big)\Sigma_y^{-1/2}\right\|_2 \\
&= \left\|\Sigma_y^{-1/2}\tilde{\Sigma}\Sigma_y^{-1/2} - \mathbb{I}\right\|_2.
\end{aligned}$$

Thus we have to control two terms:

$$\left\|\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} - \mathbb{I}\right\|_2 \lesssim \left\|\Sigma^{-1/2}\Sigma_y\Sigma^{-1/2} - \mathbb{I}\right\|_2 + \left\|\Sigma_y^{-1/2}\tilde{\Sigma}\Sigma_y^{-1/2} - \mathbb{I}\right\|_2. \tag{8}$$

Since Σ_y is an empirical estimate of Σ with n samples, with high probability the first term in Equation (8) will be at most $\sqrt{d/n}$.

Similarly, $\tilde{\Sigma}$ is an empirical estimate of Σ_y with N samples, so with high probability the second term in Equation 8 will be at most $\sqrt{d/N}$. We have $N \approx \frac{n^2\varepsilon^2}{d^2 \log 1/\delta}$, and thus rearranging get an upper bound of $\frac{d^{3/2}\sqrt{\log 1/\delta}}{\varepsilon n}$, as promised. ■

Algorithm 7: Private Learner for Unrestricted Gaussians

input : data set $x \in \mathbb{R}^{n \times d}$, privacy parameters ε, δ, outlier threshold λ_0

require: $n \geq \frac{272e^2\lambda_0 \log 2/\delta}{\varepsilon}$

$\tilde{\mu} \leftarrow \mathcal{A}^{\varepsilon,\delta,\lambda_0}_{\text{main}}(x)$; `/* Algorithm 1 */`

$\forall i \in [\lfloor n/2 \rfloor], y_i \leftarrow \frac{1}{\sqrt{2}}(x_i - x_{i+\lfloor n/2 \rfloor})$; `/* Pair and rescale */`

$\tilde{\Sigma} \leftarrow \mathcal{A}^{\varepsilon,\delta,\lambda_0}_{\text{cov}}(y)$; `/* Algorithm 6 */`

if $\tilde{\mu} =$ `FAIL` *or* $\tilde{\Sigma} =$ `FAIL` **then**

  return `FAIL`;

else

  return $\tilde{\mu}, \tilde{\Sigma}$;

end

References

Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Algorithmic Learning Theory*, pages 185–216. PMLR, 2021.

Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a gaussian: Efficient, robust and optimal. *arXiv preprint arXiv:2212.08018*, 2022.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. *arXiv preprint arXiv:2111.11320*, 2021.

Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34:7950–7964, 2021.

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.

Clément L. Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. *CoRR*, abs/2004.00010, 2020. URL `https://arxiv.org/abs/2004.00010`.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 2016.

John Duchi, Saminul Haque, and Rohith Kuditipudi. A fast algorithm for adaptive private mean estimation, 2023.

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.

Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. *arXiv preprint arXiv:2212.05015*, 2022.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. *arXiv preprint arXiv:2111.04609*, 2021.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *arXiv preprint arXiv:2205.08532*, 2022.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

Shiva Prasad Kasiviswanathan and Adam D. Smith. On the 'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 2014.

Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. *arXiv preprint arXiv:2112.03548*, 2021.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022.

Vikrant Singhal and Thomas Steinke. Privately learning subspaces. *Advances in Neural Information Processing Systems*, 34:1312–1324, 2021.

Matthew Skala. Hypergeometric tail inequalities: ending the insanity. *arXiv preprint arXiv:1311.5939*, 2013.

Lloyd N Trefethen and David Bau. *Numerical linear algebra*, volume 181. Siam, 1997.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *International Conference on Machine Learning*, pages 21828–21863. PMLR, 2022.

Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Appendix A. Preliminaries

A.1. Notation and Elementary Facts

Unless stated otherwise, "$\gtrsim$" and "$\lesssim$" hide absolute constants. We interpret data set $x = (x_1, \ldots, x_n)$ of points $x_i \in \mathbb{R}^d$ as both an ordered n-tuple and a matrix $x \in \mathbb{R}^{n \times d}$. Vectors are columns. Logarithms are base e. We use $[n]$ for the set $\{1, \ldots, n\}$ and $\mathbb{N} = \{1, 2, \ldots\}$ for the natural numbers. For a vector $v \in \mathbb{R}^d$, its support is $\operatorname{supp}(v) = \{i \in [d] : v_i \neq 0\}$. The indicator $\mathbb{1}\{P\} \in \{0, 1\}$ equals one when predicate P is true and zero otherwise. With zero-mean data, we will often use "covariance" to refer to the second moment matrix: $\Sigma_x = \frac{1}{n}\sum_{i=1}^n x_i x_i^T = \frac{1}{n} x^T x$.

Throughout, we use the facts that $\frac{1}{1-x} \leq 1 + 2x$ (for $x \in [0, 1/2]$) and $1 + x \leq e^x$ (for all x).

We say that a symmetric matrix $A = \mathbb{R}^{d \times d}$ is *positive semidefinite* if $v^T A v \geq 0$ for all $v \in \mathbb{R}^d$. If the same statement holds with a strict inequality, we say that A is *positive definite*. For positive semidefinite matrices A and B, we denote the positive semidefinite order (also called the Loewner order) in the standard way: $A \succeq B$ if and only if $A - B$ is positive semidefinite. This notation extends to $\preceq$, $\succ$, and $\prec$. Matrix inversion respects the positive semidefinite order (Horn and Johnson (2012), Corollary 7.7.4.a): if $A, B \succ 0$ then $A \succeq B$ if and only if $A^{-1} \preceq B^{-1}$.

For a matrix $A \in \mathbb{R}^{d \times d}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_d$, we use the *spectral norm* $\|A\|_2 = \sigma_1$, the *Frobenius norm* $\|A\|_F = \sqrt{\sum_{i=1}^d \sigma_i^2}$, and the *trace norm* $\|A\|_{\mathrm{tr}} = \sum_{i=1}^d \sigma_i$. We have $\|A\|_2 \leq \|A\|_F \leq \|A\|_{\mathrm{tr}}$.

A.2. Subgaussian Random Variables and Concentration Inequalities

For further discussion on subgaussian random variables, see the textbook by Vershynin (2018).

Definition 46 (Subgaussian Norm) *Let $y \in \mathbb{R}$ be a random variable. The* subgaussian norm of y, *denoted $\|y\|_{\psi_2}$, is $\|y\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp\left(y^2/t^2\right) \leq 2\right\}$.*

Definition 47 (Subgaussian Random Variable) *Let $y \in \mathbb{R}^d$ be a random variable with mean μ and covariance Σ. Call y* subgaussian with parameter K *if there exists $K \geq 1$ such that for all $v \in \mathbb{R}^d$ we have*

$$\|\langle y - \mu, v\rangle\|_{\psi_2} \leq K\sqrt{v^T \Sigma v}.$$

The Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is subgaussian with parameter $K = O(1)$.

Claim 48 (Concentration of Norm) *Let $y_1, \ldots, y_n$ be drawn i.i.d. from a d-dimensional subgaussian distribution with parameter K_y, mean μ, and (full-rank) covariance Σ. There exists a constant K_1 such that, with probability at least $1 - \beta$, we have both*

$$\|y_1 - \mu\|_\Sigma^2 \leq K_1 K_y^2 (d + \log 1/\beta) \quad \textit{and} \quad \left\|\frac{1}{n}\sum_{i=1}^n y_i - \mu\right\|_\Sigma^2 \leq K_1 K_y^2 \cdot \frac{d + \log 1/\beta}{n}.$$

Claim 49 (Concentration of Covariance) *Let $y_1, \ldots, y_n$ be drawn i.i.d. from a d-dimensional subgaussian distribution with parameter K_y, mean $\mu = 0$, and (full-rank) covariance Σ. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$ be the empirical covariance. There exists absolute constants K_1 and K_2 such that, for any $\beta \in (0,1)$, if $n \geq K_2(d + \log 1/\beta)$, then with probability at least $1-\beta$ we have*

$$\left\| \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - \mathbb{I} \right\|_2 \leq K_1 K_y^2 \sqrt{\frac{d + \log 1/\beta}{n}}.$$

Finally, we use the following tail bound for hypergeometric distributions to argue that the random set R of reference points (selected in Algorithm 1 and provided to `StableMean`) is degree-representative with high probability.

Claim 50 (See Skala (2013)) *Suppose an urn contains N balls and exactly k of them are black. Let random variable y be the number of black balls selected when drawing n balls uniformly from the urn without replacement. Then for all $t \geq 0$ we have $\Pr[|y - kn/N| \geq tn] \leq 2e^{-2t^2 n}$.*

A.3. Differential Privacy

For these facts and further background on differential privacy, see the monograph by Vadhan (2017).

Fact 51 (Basic Composition) *For all $1 \leq i \leq K$, suppose mechanism $\mathcal{M}_i$ is (ε, δ)-differentially private. Then $(\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K)$ is $(K\varepsilon, K\delta)$-differentially private. Moreover, this holds even when the mechanisms are chosen adaptively.*

Fact 52 *Suppose for some ε and δ that distributions p_1, p_2, and p_3 satisfy $p_1 \approx_{(\varepsilon,\delta)} p_2$ and $p_2 \approx_{(\varepsilon,\delta)} p_3$. Then $p_1 \approx_{(2\varepsilon,(1+e^{\varepsilon})\delta)} p_3$.*

Appendix B. Deferred Proofs: Identifiability and Sensitivity

In this appendix we prove three statements whose proofs appeared, with only minor changes from the versions we present, in BGSUZ.

We also prove Lemma 41 (restated as Lemma 56), whose proof is almost identical to that of Lemma 31, and Claim 10 (restated as Claim 57), the guarantees for our propose-test-release variant.

Lemma 53 (Restatement of Lemma 23) *Let $\lambda > 0$ and let y and y' be adjacent data sets of size m. Let w be a λ-good weighting for y and let v be a λ-good weighting for y' with $\|w - v\|_1 \leq \rho$ and $\|v\|_\infty, \|w\|_\infty \leq \eta$. Let Σ_w and Σ_v be the corresponding weighted covariances, as in Definition 22.*

Then Σ_w and Σ_v are positive definite and, for $\gamma = \lambda(\rho + 2\eta)$, satisfy

$$(1-\gamma)\Sigma_w \preceq \Sigma_v \preceq \frac{1}{1-\gamma}\Sigma_w. \tag{9}$$

Furthermore, if $\gamma \leq \frac{1}{2}$, then

$$\left\| \Sigma_v^{-1/2} \Sigma_w \Sigma_v^{-1/2} - \mathbb{I} \right\|_{\mathrm{tr}}, \left\| \Sigma_w^{-1/2} \Sigma_v \Sigma_w^{-1/2} - \mathbb{I} \right\|_{\mathrm{tr}} \leq (1+2\gamma)\gamma.$$

Proof Assume without loss of generality that y and y' differ on index 1 and let $z_1 \stackrel{\text{def}}{=} y'_1$. Then we have

$$\Sigma_v = \sum_{i\in[m]} v_i \cdot \left(y'_i\right)\left(y'_i\right)^T = \Sigma_w + v_1 \cdot z_1 z_1^T - w_1 \cdot y_1 y_1^T + \sum_{\substack{i\in[m]\\ i>1}} (v_i - w_i) \cdot y_i y_i^T \tag{10}$$

$$\succeq \Sigma_w + 0 - w_1 \cdot y_1 y_1^T + \sum_{\substack{i\in\text{supp}(w)\\ i>1}} (v_i - w_i) \cdot y_i y_i^T.$$

The last line follows from the fact that, by restricting to the support of w, we only drop positive semidefinite terms from the sum (when $w_i = 0$, we have $v_i - w_i \geq 0$). We want to lower bound this with $(1-\gamma)\Sigma_w$, so it remains to show that $-w_1 \cdot y_1 y_1^T + \sum_{i\in\text{supp}(w)}(v_i - w_i) y_i y_i^T \succeq -\gamma\Sigma_w$, or equivalently that $w_1 \cdot y_1 y_1^T + \sum_{i\in\text{supp}(w)}(w_i - v_i) y_i y_i^T \preceq \gamma\Sigma_w$. This is an upper bound on the operator norm; to prove it we conjugate by $\Sigma_w^{-1/2}$ and apply the triangle inequality plus the assumption that w is λ-good. (Also recall that $\left\|uu^T\right\|_2 = u^T u$ for any vector u.)

$$\begin{aligned}
&\left\| w_1 \cdot \Sigma_w^{-1/2} y_1 y_1^T \Sigma_w^{-1/2} + \sum_{i\in\text{supp}(w)} (w_i - v_i) \cdot \Sigma_w^{-1/2} y_i y_i^T \Sigma_w^{-1/2} \right\|_2 \\
&\qquad \leq |w_1| \cdot \left\| \Sigma_w^{-1/2} y_1 y_1^T \Sigma_w^{-1/2} \right\|_2 + \sum_{i\in\text{supp}(w)} |w_i - v_i| \cdot \left\| \Sigma_w^{-1/2} y_i y_i^T \Sigma_w^{-1/2} \right\|_2 \\
&\qquad = |w_1| \cdot \left\| \Sigma_w^{-1/2} y_1 \right\|_2^2 + \sum_{i\in\text{supp}(w)} |w_i - v_i| \cdot \left\| \Sigma_w^{-1/2} y_i \right\|_2^2 \\
&\qquad \leq |w_1| \cdot \lambda + \sum_{i\in\text{supp}(w)} |w_i - v_i| \cdot \lambda.
\end{aligned}$$

This is at most $\lambda(\|w\|_\infty + \|w - v\|_1)$, which is less than $\lambda(2\eta + \rho)$. A symmetrical argument shows that $(1-\gamma)\Sigma_v \preceq \Sigma_w$, which gives us our upper bound on Σ_v.

A straightforward argument establishes the trace norm inequalities. Continuing from Equation (10) and conjugating by $\Sigma_w^{-1/2}$, we have

$$\begin{aligned}
\Sigma_w^{-1/2}\Sigma_v\Sigma_w^{-1/2} - \mathbb{I} &= v_1 \cdot \Sigma_w^{-1/2} z_1 z_1^T \Sigma_w^{-1/2} - w_1 \cdot \Sigma_w^{-1/2} y_1 y_1^T \Sigma_w^{-1/2} \\
&\quad + \sum_{i=2}^m (v_i - w_i) \cdot \Sigma_w^{-1/2} y_i y_i^T \Sigma_w^{-1/2}.
\end{aligned}$$

As before, we apply the triangle inequality and $\left\|uu^T\right\|_{\text{tr}} = \|u\|_2^2$:

$$\begin{aligned}
\left\| \Sigma_w^{-1/2}\Sigma_v\Sigma_w^{-1/2} - \mathbb{I} \right\|_{\text{tr}} &\leq |v_1| \cdot \left\| \Sigma_w^{-1/2} z_1 \right\|_2^2 + |w_1| \cdot \left\| \Sigma_w^{-1/2} y_1 \right\|_2^2 \\
&\quad + \sum_{i=2}^m |v_i - w_i| \cdot \left\| \Sigma_w^{-1/2} y_i \right\|_2^2.
\end{aligned}$$

We cannot uniformly upper bound these squared norms by λ, since our assumption of w's goodness only applies to points in the support of w. However, by Equation (9), for all i in the support of v, we have $\left\| \Sigma_w^{-1/2} y'_i \right\|_2^2 \leq (1 + 2\gamma)\lambda$.

This argument, combined with $\|w - v\|_1 \leq \rho$, yields

$$\left\|\Sigma_w^{-1/2}\Sigma_v\Sigma_w^{-1/2} - \mathbb{I}\right\|_{\text{tr}} \leq (1 + 2\rho\lambda) \cdot (\|v\|_\infty + \|w\|_\infty + \rho) \cdot \lambda.$$

The second trace norm inequality is symmetrical. ■

Observe that, when $R = [n]$ and the degree threshold τ is at least $\frac{n}{2} + 1$, any two points in $\text{supp}(w)$ must have a "neighbor" in common, and thus cannot themselves be too far apart. We have the following claim, which says that weighted cores which are close in ℓ_1 norm (equivalently, total variation distance) have close means. Furthermore, this holds even when the cores are computed under covariances that differ slightly.

Lemma 54 *Let $x = x_1, \ldots, x_n$ be a data set. Let $\lambda > 0$, $\tau \in \mathbb{N}$, and $R \subseteq [n]$. Let Σ_1 and Σ_2 be positive definite matrices satisfying $(1 - \gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Let w be a $(\tau, \lambda, \Sigma_1)$-weighted-core for x with respect to R and let v be a $(\tau, \lambda, \Sigma_2)$-weighted-core for x with respect to R. Assume $\tau > \frac{n}{2}$. If $\|w - v\|_1 \leq \rho$, then*

$$\|\mu_w - \mu_v\|_{\Sigma_1}^2 = \left\|\sum_{i\in[n]} w_i x - \sum_{i\in[n]} v_i x_i\right\|_{\Sigma_1}^2 \leq (1 + 2\gamma)\rho^2\lambda.$$

Proof For any $i, j \in \text{supp}(w) \cup \text{supp}(v)$, by definition there exist sets of "nearby" data points $N_i, N_j \subseteq R$ of size at least τ. As $\tau > \frac{n}{2}$, we know that N_i and N_j have a nonempty intersection. Thus there is a point $\iota \in R$ such that $\|x_i - x_\iota\|_{\Sigma_1}^2 \leq \lambda$ and $\|x_\iota - x_j\|_{\Sigma_2}^2 \leq \lambda$. By the assumptions on Σ_1, Σ_2, and γ, the second inequality also implies $\|x_\iota - x_j\|_{\Sigma_1}^2 \leq (1 + 2\gamma)\lambda$. Now, working with the (unsquared) norm, we add and subtract x_ι and apply the triangle inequality. For any $i, j \in \text{supp}(w) \cup \text{supp}(v)$,

$$\begin{aligned}\|x_i - x_j\|_{\Sigma_1} &= \|x_i - x_\iota + x_\iota - x_j\|_{\Sigma_1} \\ &\leq \|x_i - x_\iota\|_{\Sigma_1} + \|x_\iota - x_j\|_{\Sigma_1} \\ &\leq 2\sqrt{(1 + 2\gamma)\lambda}.\end{aligned}$$

Let C be a multivariate random variable obeying distribution w, and let random variable D obey v. Thus $\|\mu_w - \mu_v\|_{\Sigma_1}$ is just $\|\mathbb{E}[C - D]\|_{\Sigma_1}$. The distributions w and v have total variation distance at most $\rho/2$ (since total variation distance is exactly half the ℓ_1 distance), so there exists a maximum coupling of w and v and random variables C', D' such that $\mathbb{E}[C - D] = \mathbb{E}[C' - D']$ and $\Pr[C' \neq D'] \leq \rho/2$. We apply Jensen's inequality to the norm and drop the part of the expectation corresponding to $C' = D'$, because it is zero:

$$\begin{aligned}\|\mu_w - \mu_v\|_{\Sigma_1} &= \left\|\mathbb{E}[C' - D']\right\|_{\Sigma_1} \\ &\leq \mathbb{E}\left[\left\|C' - D'\right\|_{\Sigma_1}\right] \\ &= \Pr[C' \neq D'] \cdot \mathbb{E}\left[\left\|C' - D'\right\|_{\Sigma_1} \mid C' \neq D'\right].\end{aligned}$$

With an upper bound on $\Pr[C' \neq D]$ and a uniform upper bound on the values inside the expectation, we arrive at

$$\|\mu_w - \mu_v\|_{\Sigma_1} \leq \frac{\rho}{2} \cdot 2\sqrt{(1+2\gamma)\lambda} = \rho\sqrt{(1+2\gamma)\lambda}.$$

Squaring gives an upper bound of $(1+2\gamma)\rho^2\lambda$. ■

Corollary 36 shows that the same argument applies, with little modification, to the setting where the cores are computed on adjacent data sets. Note that we make a slightly stronger assumption, asking that $\tau > \frac{n}{2} + 1$ instead of $\frac{n}{2}$.

Corollary 55 (Restatement of Corollary 36) *Let x and x' be adjacent data sets of size n. Let $\lambda > 0$, $\tau \in \mathbb{N}$, and $R \subseteq [n]$. Let Σ_1 and Σ_2 be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Let w be a $(\tau, \lambda, \Sigma_1)$-weighted-core for x with respect to R and let v be a $(\tau, \lambda, \Sigma_2)$-weighted-core for x' with respect to R, with $\tau > \frac{n}{2} + 1$. If $\|w - v\|_1 \leq \rho$ then*

$$\|\mu_w - \mu_v\|_{\Sigma_1}^2 = \left\| \sum_{i \in [n]} w_i x_i - \sum_{i \in [n]} v_i x_i' \right\|_{\Sigma_1}^2 \leq (1+2\gamma)\lambda(\rho + \|w\|_\infty + \|v\|_\infty)^2.$$

Proof [Proof of Corollary 36] From w and v we construct cores w^+ and v^+ over $x \cup x'$ and apply Lemma 54.

Let $i^* \in [n]$ be the index in which x and x' differ. Our data sets are ordered and not strictly sets, so interpret $x \cup x' = (x_1, x_2, \ldots, x_n, x'_{i^*})$ as a data set of size $n+1$. Let $R^+ \leftarrow R \cup \{n+1\}$ be the similarly extended reference set. Already, $w^+ = (w_1, \ldots, w_n, 0)$ is an (m, λ, Σ_1)-core for $x \cup x'$ with respect to R^+. Similarly,

$$v^+ = (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n, v_i)$$

is an $(\tau, \lambda, \Sigma_2)$-core for $x \cup x'$ with respect to R^+. Lemma 54 asks that $m > \frac{n+1}{2}$, which is satisfied: we have assumed that $m > \frac{n}{2} + 1$. So it remains to calculate $\|w^+ - v^+\|_1$, which is simple:

$$\begin{aligned}
\left\|w^+ - v^+\right\|_1 &= \sum_{j \in [n+1]} \left| w_j^+ - v_j^+ \right| \\
&= \left|w_i^+ - v_i^+\right| + \left|w_{n+1}^+ - v_{n+1}^+\right| + \sum_{j \in [n+1] \setminus \{i, n+1\}} \left| w_j^+ - v_j^+ \right| \\
&= \left|w_i^+ - 0\right| + \left|0 - v_{n+1}^+\right| + \sum_{j \in [n+1] \setminus \{i, n+1\}} \left| w_j^+ - v_j^+ \right| \\
&\leq \|w\|_\infty + \|v\|_\infty + \|w - v\|_1 .
\end{aligned}$$

■

Lemma 56 (Restatement of Lemma 41) *Fix $\lambda_0 > 0$, $k \in \mathbb{N}$, and $R \subseteq [n]$. Let x and x' be adjacent data sets. Let Σ_1 and Σ_2 be positive definite matrices satisfying $(1-\gamma)\Sigma_1 \preceq \Sigma_2 \preceq \frac{1}{1-\gamma}\Sigma_1$ for $\gamma \leq \frac{1}{2}$. Further assume $k \leq \frac{1}{2\gamma}$. Then $|g(x, \Sigma_1) - g(x', \Sigma_2)| \leq 2$.*

Proof Without loss of generality, assume $g(x, \Sigma_1) \leq g(x', \Sigma_2)$. We will show that $g(x', \Sigma_2) \leq g(x, \Sigma_1) + 2$ by analyzing two cases.

Case 1: Suppose $g(x, \Sigma_1) = k$. Then $g(x', \Sigma_2) \leq g(x, \Sigma_1) \leq g(x, \Sigma_1) + 2$ by construction, since $g(\cdot)$ is capped at k.

Case 2: Suppose $g(x, \Sigma_1) < k$. This can only happen when $g(x, \Sigma_1) < k$, so there exists an $\ell^* \in \{0, \ldots, k\}$ and subset S_{ℓ^*} that (i) is a $(|R| - \ell, e^{\ell/k}\lambda_0, \Sigma_1)$-core for x with respect to R and (ii) satisfies $n - |S_{\ell^*}| + \ell^* < k$. Furthermore, we know that $\ell^* \neq k$, since $n - |S_{\ell^*}|$ is nonnegative.

We can now apply Lemma 39. For $\ell \in \{0, \ldots, k\}$, let T_ℓ be the largest $(|R| - \ell, e^{\ell/k}\lambda_0, \Sigma_2)$-core for x' with respect to R. Lemma 39 says that $S_{\ell^*} \setminus i^* \subseteq T_{\ell^*+1}$. This allows us to upper bound $g(x', \Sigma_2)$:

$$\begin{aligned} g(x', \Sigma_2) \leq g(x', \Sigma_2) &= \min_{\ell \in \{0,\ldots,k\}} n - |T_\ell| + \ell \\ &\leq n - |T_{\ell^*+1}| + \ell^* + 1 \\ &\leq n - (|S_{\ell^*}| - 1) + \ell^* + 1 \\ &= g(x, \Sigma_1) + 2. \end{aligned}$$

So we are done. ■

Claim 57 (Restatement of Claim 10) *Fix $0 < \varepsilon \leq 1$ and $0 < \delta \leq \frac{\varepsilon}{10}$. There is an algorithm $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}} : \mathbb{R} \to \{\texttt{PASS}, \texttt{FAIL}\}$ that satisfies the following conditions:*

(1) Let $\mathcal{U}$ be a set and $g : \mathcal{U}^n \to \mathbb{R}_{\geq 0}$ a function. If, for all $x, x' \in \mathcal{U}^n$ that differ in one entry, $|g(x) - g(x')| \leq 2$, then $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(g(\cdot))$ is (ε, δ)-DP.

(2) $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(0) = \texttt{PASS}$.

(3) For all $z \geq \frac{2\log 1/\delta}{\varepsilon} + 4$, $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(z) = \texttt{FAIL}$.

Proof Let $\tau = \frac{2\log\frac{1-\delta}{\delta}}{\varepsilon} + 4$ and let $p(z)$ be the probability that $\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(z)$ passes, defined as follows:

$$p(z) = \Pr\left[\mathcal{M}^{\varepsilon,\delta}_{\mathrm{PTR}}(z) = \texttt{PASS}\right] = \begin{cases} 1 & \text{if } z = 0 \\ 0 & \text{if } z \geq \tau \\ 1 - e^{\frac{\varepsilon}{2}(z-2)}\delta & \text{otherwise} \end{cases}.$$

What properties must $p(z)$ have? It satisfies the second and third conditions by construction, so we only have to argue privacy, and consider $p(z)$ and $p(z')$ for $|z - z'| \leq 2$. Without loss of generality, assume $z \leq z'$. Since $p(\cdot)$ is monotone decreasing (and thus $p(z) \geq p(z')$), we must show that

$$p(z) \leq e^{\varepsilon} p(z') + \delta.$$

Since $z' \leq z + 2$, we know that $p(z') \geq p(z + 2)$ so it suffices to show that

$$p(z) \leq e^{\varepsilon} p(z + 2) + \delta.$$

We will prove this via cases, dealing with the boundaries (near 0 and τ) first. (Note that τ in the lemma statement is slightly larger than $\frac{2\log\frac{1-\delta}{\delta}}{\varepsilon}+4$ for simplicity; we prove the stronger statement.)

Case 1: Suppose $z = 0$. Then $p(z) = 1$ and $p(z+2) = 1 - e^{\frac{\varepsilon}{e}(2-2)}\delta = 1-\delta$. Thus $p(z) \le p(z+2) + \delta \le e^{\varepsilon}p(z+2)+\delta$.

Case 2: Suppose $z \ge \tau - 2$, so that $p(z+2) = 0$. We calculate

$$\begin{aligned} p(z) \le p(\tau-2) &= 1 - e^{\frac{\varepsilon}{2}(\tau-2-2)}\delta \\ &= 1 - e^{\frac{\varepsilon}{2}\left(\frac{2\log\frac{1-\delta}{\delta}}{\varepsilon}\right)}\delta \\ &= 1 - \frac{1-\delta}{\delta}\cdot\delta = \delta. \end{aligned}$$

Thus $p(z) \le p(z+2)+\delta \le e^{\varepsilon}p(z+2)+\delta$.

Case 3: Now suppose $0 < z < \tau - 2$ and set $\delta' \stackrel{\text{def}}{=} e^{\frac{\varepsilon}{2}(z-2)}\delta$ for brevity. Since $0 < z$, we have $\delta' \ge e^{-\varepsilon}\delta$. We calculate

$$p(z) = 1-\delta' \quad \text{and} \quad p(z+2) = 1 - e^{\frac{\varepsilon}{2}(z+2-2)}\delta = 1 - e^{\varepsilon}\delta'.$$

We now $e^{\varepsilon}p(z+2) - p(z)$ and show that it is positive:

$$\begin{aligned} e^{\varepsilon}p(z+2) - p(z) &= e^{\varepsilon}\big(1 - e^{\varepsilon}\delta'\big) - \big(1-\delta'\big) \\ &= e^{\varepsilon} - e^{2\varepsilon}\delta' - 1 + \delta' \\ &= (e^{\varepsilon}-1) - \delta'\big(e^{2\varepsilon}-1\big). \end{aligned}$$

This is positive when $\frac{e^{\varepsilon}-1}{e^{2\varepsilon}-1} \ge \delta' \ge e^{-\varepsilon}\delta$. Since $\varepsilon \le 1$, this is true. Observe:

$$\begin{aligned} \frac{e^{\varepsilon}-1}{e^{2\varepsilon}-1} &\ge \frac{\varepsilon}{e^{2\varepsilon}-1} \\ &\ge \frac{\varepsilon}{e^2-1}. \end{aligned}$$

This is greater than δ' when $\delta \le \frac{e^{\varepsilon}}{e^2-1}\cdot\varepsilon$. Using $\frac{e^{\varepsilon}}{e^2-1} \ge \frac{1}{10}$ finishes the proof, since we assumed $\delta \le \frac{\varepsilon}{10}$. ■