Proceedings of Machine Learning Research vol 195:1–41, 2023 36th Annual Conference on Learning Theory

A Pretty Fast Algorithm for Adaptive Private Mean Estimation

John Duchi JDUCHI@STANFORD.EDU
Departments of Statistics and Electrical Engineering, Stanford University

Saminul Haque SAMINULH@STANFORD.EDU
Department of Computer Science, Stanford University

Rohith Kuditipudi ROHITHK@STANFORD.EDU
Department of Computer Science, Stanford University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We design an (ε, δ)-differentially private algorithm to estimate the mean of a d-variate distribution, with unknown covariance Σ, that is adaptive to Σ. To within polylogarithmic factors, the estimator achieves optimal rates of convergence for sub-Gaussian data with respect to the induced Mahalanobis norm $\|\cdot\|_\Sigma$, takes time $\widetilde{O}(nd^2)$ to compute, has near linear sample complexity for sub-Gaussian distributions, allows Σ to be degenerate or low rank, and adaptively extends beyond sub-Gaussianity. Prior to this work, other methods required exponential computation time or the superlinear scaling $n = \Omega(d^{3/2})$ to achieve non-trivial error with respect to the norm $\|\cdot\|_\Sigma$.

1. Introduction

We cannot consider the theory of differential privacy complete until we have—at least—a sample and computationally efficient estimator of the mean. To within logarithmic factors in the dimension d and sample size n, we achieve both.

To make this a bit more precise, let P be a distribution on $\mathbb{R}^d$ with unknown mean $\mu = \mathbb{E}_P[X]$ and unknown covariance $\Sigma = \mathbb{E}_P[(X-\mu)(X-\mu)^T]$, and let $X_i \stackrel{\text{iid}}{\sim} P$, $i \le n$. Define the extended-value Mahalanobis norm $\|\cdot\|_A$ for $A \succeq 0$ by

$$\|v\|_A^2 := \lim_{t \downarrow 0} v^T (A + tI)^{-1} v = \begin{cases} v^T A^\dagger v & v \in \mathsf{Col}(A) \\ +\infty & \text{otherwise,} \end{cases}$$

and for an estimator $\widehat{\mu}$ let $\mathrm{err}_\Sigma(\widehat{\mu}, \mu) := \|\widehat{\mu} - \mu\|_\Sigma^2$ be the covariance-normalized error. We give an (ε, δ)-differentially private estimator $\widetilde{\mu}$ of μ such that, assuming the vectors $\Sigma^{-1/2} X_i$ are sub-Gaussian and $n = \widetilde{\Omega}(\max\{\mathrm{rank}(\Sigma), 1\}/\varepsilon^2)$,

$$\mathrm{err}_\Sigma(\widetilde{\mu}, \mu) = \|\widehat{\mu} - \mu\|_\Sigma^2 \le \widetilde{O}(1) \left[\frac{\mathrm{rank}(\Sigma) + \log \frac{1}{\delta}}{n} + \frac{\mathrm{rank}(\Sigma)^2 \log^2 \frac{1}{\delta}}{n^2 \varepsilon^2} \right] \tag{1}$$

with probability at least $1 - \delta$, where the $\widetilde{O}(1)$ term hides dependence on the sub-Gaussian parameter of $\Sigma^{-1/2} X$ and logarithmic factors in n. Except for the dependence on $\log \frac{1}{\delta}$ and the hidden logarithmic factors in n, this is optimal, and the method extends naturally to distributions with heavier tails for which we can provide similar near-optimal guarantees.

By measuring error with respect to the covariance Σ of the data itself, we adopt the familiar efficiency goals of classical theoretical statistics: that an estimator should be adaptive to structure in covariates and should have (near)-optimal covariance. Mean estimation is, of course, one of the most basic problems in statistics, and we have known for seventy-odd years that the sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is efficient (Cramér, 1946; Le Cam, 1986), achieving the optimal error $\mathbb{E}[\|\widehat{\mu} - \mu\|_\Sigma^2] = \frac{\mathrm{rank}(\Sigma)}{n}$, with high-probability guarantees under appropriate moment assumptions (Wainwright, 2019). Perhaps stating the obvious, the sample mean is adaptive to the covariance of the distribution: no matter Σ, the sample mean is efficient.

When we require estimators to be private, however, the story is less clear. While differential privacy (Dwork et al., 2006b,a) has become the *de facto* choice for protecting sensitive data in the sixteen or so years since its release—with substantial theoretical advances and successful applications (Erlingsson et al., 2014; Apple Differential Privacy Team, 2017; Abadi et al., 2016; Lauger et al., 2017; Garfinkel et al., 2018; Dwork and Roth, 2014)—we know of no computationally efficient procedures that achieve order-optimal sample complexity with respect to the natural geometry the population P induces via its covariance. Brown et al. (2021) highlight this, developing sample efficient procedures that achieve small error in the Mahalanobis metric even when Σ is unknown. When the covariance Σ is known, estimators that truncate the data relative to Σ and add Gaussian noise to such a trimmed mean with covariance proportional to Σ suffice to privately estimate μ (under approximate differential privacy) with the essentially optimal rate (1), so that $n = \Omega(d)$ observations suffice to estimate μ in the full-rank case (see, e.g., Biswas et al., 2020; Brown et al., 2021). But in the more realistic setting that Σ is unknown, to the best of our knowledge all prior work either requires a sample of size $n = \Omega(d^{3/2})$; is intractable, taking time exponential in n or d to compute; or assumes P is isotropic. Many of these further assume P is Gaussian, a stringent assumption that never obtains in practice. See Section 1.1 for more discussion.

Our contribution is a polynomial-time private estimator (Algorithm 3, `PRIVMEAN`) whose error matches the error achievable when the covariance is known (equivalently, the data is isotropic) to polylogarithmic factors. In essence, our estimator privatizes a stable estimate of the empirical mean by adding Gaussian noise with covariance proportional to a stable estimate of the empirical covariance; it takes time $\tilde{O}(nd^2)$ to compute, has (nearly) linear sample complexity for sub-Gaussian distributions, allows Σ to be degenerate or low-rank, and naturally extends beyond sub-Gaussianity.

1.1. Related work

There are many connections between differential privacy and robust statistics (Dwork and Lei, 2009), in that the major focus of robust statistics is to develop estimators insensitive to outliers and corrupted data (Tukey, 1960; Huber, 1964; Huber and Ronchetti, 2009; Hampel et al., 1986), while differential privacy makes the output (distributions) of estimators similar even when individuals in the underlying data change (Dwork et al., 2006b,a; Dwork and Lei, 2009). While Tukey and Huber's initiation of robust statistics is more than sixty years old (Tukey, 1960; Huber, 1964), studying statistical limits of estimation and inference from corrupted data, computational tractability was elusive: only in the last decade have researchers developed computationally efficient methods for even robustly estimating a sample mean (Diakonikolas and Kane, 2022). Similarly, only recently has the community elucidated trade-offs between statistical and computational considerations in robust estimation (Diakonikolas and Kane, 2022).

It is natural to wonder whether such trade-offs also arise with privacy. For example, classical procedures in private query evaluation require exponential time in natural problem parameters (Hardt and Rothblum, 2010; Dwork et al., 2010). Likewise, in estimation, following the "propose, test, release" framework of Dwork and Lei (2009), a number of sample efficient private estimators (Liu et al., 2022; Brown et al., 2021; Hopkins et al., 2022b) require testing whether a given statistic is robust to the removal of groups of data points, which can be computationally intractable in high-dimensions. In a number of these settings, computationally efficient estimators achieving comparable sample efficiency have emerged only within the last year or so (e.g., Hopkins et al., 2022b; Kothari et al., 2022; Ashtiani and Liaw, 2022; Alabi et al., 2022). Our mean estimation setting is a striking example of a seemingly simple problem for which no known sub-exponential time and sample efficient algorithm exists. In particular, to the best of our knowledge, all previous work has either (i) exponential runtime (Brown et al., 2021; Liu et al., 2021, 2022); (ii) is sample inefficient (Kamath et al., 2018; Liu et al., 2021), requiring sample size at least $n = \Omega(d^{3/2})$; or (iii) otherwise essentially assumes the population covariance Σ is isotropic (Kamath et al., 2018; Biswas et al., 2020; Huang et al., 2021; Liu et al., 2021) (nominally, Huang et al. (2021) allow arbitrary covariance, but the squared error of its estimator scales at least linearly with the condition number of the population covariance Σ, which is effectively equivalent to assuming isotropic covariance (Biswas et al., 2020)). Here we highlight the most relevant (recent) examples; see the paper (Brown et al., 2021) for coverage of earlier work.

The work most closely related to ours is that of Brown et al. (2021), who also consider covariance-adaptive mean estimation and also achieve (nearly) linear sample complexity. They give a roadmap to adaptive private mean estimation that circumvents private covariance estimation, a task whose sample complexity is necessarily $\Omega(d^{3/2})$ (see for example Dwork et al. (2014, Corollary 32) and Kamath et al. (2022)), and are the first to achieve sample complexity $o(d^{3/2})$, let alone linear. However, their estimators take exponential time to compute; moreover, while their accuracy analysis is independent of the condition number of Σ, it assumes Σ is full rank. Finally, they only consider Gaussian and sub-Gaussian distributions.

Hopkins et al. (2022b) give a generic reduction from private to robust estimation, leveraging it to obtain private estimators with (near) optimal sample complexity. While their reduction is generic, the resulting estimators are efficient only in certain special cases, e.g., for Gaussian distributions whose algebraic moment relationships allow efficient formulation, and their results for mean estimation assume bounded covariance. This extends a line of work on obtaining efficient approximations of inefficient private mechanisms via sum-of-squares (SoS) relaxations (Kothari et al., 2022; Hopkins et al., 2022a), and while efficient in theory, SoS estimators typically incur large polynomial runtime and thus scale poorly to high-dimensional settings or large amounts of data. Unlike our estimator, however, they are robust to corruption of a constant fraction the data. Also independently of and concurrently to this paper, Brown, Hopkins, and Smith (2023) develop a private, polynomial time covariance-adaptive mean estimator, taking a similar approach as we do to give an efficient version of the sub-Gaussian estimator proposed by Brown et al. (2021).

1.2. Organization

We provide a brief outline of the paper to come. Section 2 introduces notation and covers the preliminary privacy definitions we require for our development. Our main estimator, `PRIVMEAN`, consists of two main parts: stably estimating the covariance of the data to reasonable accuracy and then estimating a truncated mean to which we add noise. We present our algorithms in Section 3, where Section 3.1 gives the covariance estimator, Section 3.2 the mean estimator, and Section 3.3 presents the full procedure; we analyze `PRIVMEAN`'s privacy in Section 4, deferring some of the requisite proofs to Sections C and D. We provide accuracy analysis in Section 5, where we also present `ADAMEAN` (Algorithm 5), which allows `PRIVMEAN` to adapt to the scale of the observed data.

2. Preliminary definitions, privacy properties, and mechanisms

To make our coming development smoother and easier, here we introduce notation and recapitulate the privacy definitions we use throughout. We also review a few standard privacy mechanisms, providing guarantees on their behavior; for those results that are not completely standard, we include proofs in the appendices for completeness.

2.1. Notation

Semidefinite matrices and norms For a positive semidefinite (PSD) matrix $A \in \mathbb{R}^{d\times d}$, we let $\mathsf{Col}(A)$ denote its columnspace and $A^\dagger$ its pseudoinverse, while the square-root of the pseudoinverse is $A^{\dagger/2}$. We let $\Pi_A := A^\dagger A = A^{\dagger/2}A^{1/2} \in \mathbb{R}^{d\times d}$ denote the orthogonal projector onto $\mathsf{Col}(A)$. Using the nuclear norm $\|A\|_* = \sum_{i=1}^n \sigma_i(A)$ (the sum of A's singular values), we define the distance-like quantity for PSD matrices A, B as

$$d_{\rm psd}(A,B) := \begin{cases} \max\left\{\left\|A^{\dagger/2}(B-A)A^{\dagger/2}\right\|_*, \left\|B^{\dagger/2}(A-B)B^{\dagger/2}\right\|_*\right\} & \text{if } \mathsf{Col}(A) = \mathsf{Col}(B) \\ \infty & \text{otherwise,} \end{cases}$$

setting $d_{\rm psd}(A,B) = \infty$ if A or B are not PSD. When A and B are invertible, $d_{\rm psd}(A,B) = \max\{\|A^{-1/2}BA^{-1/2} - I\|_*, \|B^{-1/2}AB^{-1/2} - I\|_*\}$, though we note in passing that it is not a distance. Recall the extended-value Mahalanobis norm $\|\cdot\|_A$ corresponding to $A \succeq 0$ is

$$\|v\|_A^2 := \lim_{t\downarrow 0} v^T(A+tI)^{-1}v = \begin{cases} v^T A^\dagger v & v \in \mathsf{Col}(A) \\ +\infty & \text{otherwise.} \end{cases}$$

When A is non-singular, this is the standard $\|v\|_A = \sqrt{v^T A^{-1} v}$, and the norm has the monotonicity property that if $A \preceq B$, then $\|v\|_A \geq \|v\|_B$ for all $v \in \mathbb{R}^d$.

Sets and Partitions For sets S, S', define the distance $d_{\text{sym}}(S, S') := \max\{|S \setminus S'|, |S' \setminus S|\}$. Given integers n and b, where we assume b divides n for simplicity, we let $\mathcal{P}_{n,b}$ be the set of all partitions of $[n]$ such that each subset constituting the partition has b elements. We represent a given partition in $\mathcal{P}_{n,b}$ as a tuple of subsets $\mathcal{S} = (S_1, \ldots, S_{n/b})$, where each $S_j \subset [n]$ has b elements and are pairwise disjoint.

Distributions We let $W \sim \mathsf{Lap}(\sigma)$ denote that W has Laplace distribution with scale σ, with density $p(w) = \frac{1}{2\sigma}\exp(-|w|/\sigma)$. $X \sim \mathsf{N}(\mu, \Sigma)$ indicates that X is normal with mean μ and covariance $\Sigma \succeq 0$, where if Σ is not full rank we mean that X has support $\mu + \mathsf{Col}(\Sigma)$.

2.2. Privacy definition and basic properties

It will be convenient for us to use closeness of distributions in our derivations (cf. Dwork and Roth, 2014, Ch. 3.5), so we frame differential privacy as a type of closeness in distribution.

Definition 1 ((ε, δ)-closeness) *Probability distributions P and Q are (ε, δ)-*close in distribution*, denoted $P \overset{d}{=}_{\varepsilon,\delta} Q$, if for all measurable sets $A \subset \mathcal{X}$,*

$$P(A) \le e^{\varepsilon} Q(A) + \delta \quad \text{and} \quad Q(A) \le e^{\varepsilon} P(A) + \delta.$$

Random variables $X \overset{d}{=}_{\varepsilon,\delta} Y$ if their induced distributions satisfy $\mathbb{P}(X \in \cdot) \overset{d}{=}_{\varepsilon,\delta} \mathbb{P}(Y \in \cdot)$

Differential privacy (Dwork et al., 2006b,a) is equivalent to this notion of closeness: a randomized function (or mechanism) M from an input space $\mathcal{X}^n$ to $\mathcal{Y}$ is (ε, δ)-differentially private if and only if for any vectors $x, x' \in \mathcal{X}^n$ differing in only a single element,

$$M(x) \overset{d}{=}_{\varepsilon,\delta} M(x').$$

The following results on closeness are standard (Dwork and Roth, 2014, Ch. 3).

Lemma 2 (Basic composition) *Let X, X', Y, Y' be independent random variables satisfying $X \overset{d}{=}_{\varepsilon_X,\delta_X} X'$, and $Y \overset{d}{=}_{\varepsilon_Y,\delta_Y} Y'$. Then $(X, Y) \overset{d}{=}_{\varepsilon_X+\varepsilon_Y,\delta_X+\delta_Y} (X', Y')$.*

Lemma 3 (Group composition) *Let $X_1, \ldots, X_k$ be random variables with $X_i \overset{d}{=}_{\varepsilon_i,\delta_i} X_{i+1}$ for each i. Let $\varepsilon_{>i} := \sum_{j=i+1}^{k-1} \varepsilon_j$, $\varepsilon = \sum_{i=1}^{k-1} \varepsilon_i$, and $\delta = \sum_{i=1}^{k} e^{\varepsilon_{>i}} \delta_i$. Then $X_1 \overset{d}{=}_{\varepsilon,\delta} X_k$.*

Lemma 4 (Post-Processing) *Let X, Y, W be random variables. Then for any function f, if $X \overset{d}{=}_{\varepsilon,\delta} Y$, then $f(X, W) \overset{d}{=}_{\varepsilon,\delta} f(Y, W)$.*

3. Algorithms

As our estimator and its full analysis are fairly involved, we provide a broad overview of our procedures here. We compute the estimator, whose full treatment we give in Algorithm 3 (`PRIVMEAN`) in section 3.3, in two phases, consisting of a stable covariance estimate and a stable mean estimate. Each carefully prunes outliers from the data, using plug-in quantities from the remaining observations as substitutes for the usual plug-in mean and covariance.

In the first phase (Algorithm 1, COVSAFE), we obtain a robust but non-private estimate $\widehat{\Sigma}$ of the covariance. Assuming for convenience n is even, we pair observations and initialize

$$\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n/2}(x_i - x_{n/2+i})(x_i - x_{n/2+i})^T.$$

As $x_i - x_{n/2+i}$ is symmetric, we can prune pairs of observations for which $\|x_i - x_{n/2+i}\|_{\widehat{\Sigma}}$ is large (regardless of the population mean μ), recompute $\widehat{\Sigma}$ on the remaining observations, and repeat until convergence. The key is that while this pruning provides no formal robustness guarantees, it is stable to changes of a single example x_i, ensuring $\widehat{\Sigma}$ itself is stable.

In the second phase (Algorithm 2, MEANSAFE), we first obtain a robust estimate $\widehat{\mu}$ of the empirical mean by trimming outliers with respect to $\|\cdot\|_{\widehat{\Sigma}}$. Using $\|x_i\|_{\widehat{\Sigma}}$ to determine whether x_i is influential for a mean estimate is unreliable, as the quantity may be arbitrarily large even for non-outliers if $\|\mu\|_{\widehat{\Sigma}}$ itself is large; unfortunately, paired observations (as in the stable covariance estimation phase) are similarly unhelpful, as $\|x_i - x_j\|_{\widehat{\Sigma}}$ could be small if both x_i, x_j are "outlying" in the same way. Instead, we randomly partition the n observations into groups S of size $O(\log \frac{n}{\delta})$ and prune *all* observations in a group S if *any* two observations in S are far with respect to $\|\cdot\|_{\widehat{\Sigma}}$, so that there is at least a pair of outlying observations in the group. Assuming the total number of pruned observations across both phases is not too large—and much of our analysis shows how to make the pruned observations stable across different samples x, x'—we let $\widehat{\mu}$ be the empirical mean of the un-pruned observations, then release $\widetilde{\mu} \sim \mathsf{N}(\widehat{\mu}, \sigma^2(\varepsilon, \delta)\widehat{\Sigma})$, where the privacy budget determines $\sigma^2(\varepsilon, \delta)$.

3.1. Stable covariance estimation

The first component of the private mean estimation algorithm is the covariance estimation procedure COVSAFE in Alg. 1, which removes "unusual" pairs of data points from the sample $x \in (\mathbb{R}^d)^n$, then uses the remaining pairs to actually construct the covariance. The procedure maintains an empirical covariance Σ_t of the remaining data at each iteration t, so that $\{\Sigma_t\}$ is a non-increasing (in the semidefinite order) sequence of matrices, and stores removed indices in an iteratively growing collection R_t for $t = 1, 2, \ldots$; the procedure thus necessarily terminates after at most $n/2$ rounds of index removal. For convenience of analysis, COVSAFE returns a transcript Γ of the removed indices and iteratively constructed covariances, returning $\perp$ if the data is so unstable that it removes too many indices.

The key is that the covariance estimates are appropriately stable (see Conditions (C.i) and (C.ii) to come in Section 4), and with high probability on any given input x, the algorithm guarantees that its output changes little when we remove index i or, if the data has too much variance relative to itself, that the procedure simply returns $\widehat{\Sigma} = \perp$. To allow cleaner description of the precise results we require in our main privacy result in Section 4, for a putative bound B on $\|x_i - x_j\|_{\Sigma}^2$, acceptable number of outliers m, and privacy random variables Z and W to be specified, let

$$(\widehat{\Sigma}, \Gamma) := \texttt{COVSAFE}_{B,m}(x; Z, W), \tag{2a}$$

Algorithm 1 Stable Covariance Estimation (COVSAFE)

Input : data $x_{1:n}$
Params: threshold B, threshold m
Noise : $z \in \mathbb{R}^{n/2+1}$, $w \in \mathbb{R}$

1 $\tilde{x} \leftarrow x_{1:n/2} - x_{n/2+1:n}$
2 $R_0 \leftarrow \emptyset$, $\Sigma_0 \leftarrow \frac{1}{n}\sum_{i=1}^{n/2} \tilde{x}_i \tilde{x}_i^T$
3 `converged` $\leftarrow$ **false**, $t \leftarrow 0$
4 **while not** `converged` **do**
5     $t \leftarrow t+1$, $R_t \leftarrow R_{t-1}$, `truncated` $\leftarrow 0$
6   **for** $i \in [n/2] \setminus R_{t-1}$ **do**
7   **if** $2\log \|\tilde{x}_i\|_{\Sigma_{t-1}} + z_i + z_{n/2+1} > \log(B)$ **then**
8             $R_t \leftarrow R_t \cup \{i\}$
9   `truncated` $\leftarrow$ `truncated` $+1$
10   **end**
11   **end**
12   **if** `truncated` $=0$ **then**
13   `converged` $\leftarrow$ **true**, $T \leftarrow t$
14   **end**
15     $\Sigma_t \leftarrow \frac{1}{n}\sum_{i \in [n/2]\setminus R_t} \tilde{x}_i \tilde{x}_i^T$
16 **end**
17 $\Gamma \leftarrow \left([R_i]_{t=0}^T, [\Sigma_t]_{t=0}^T, T\right)$
18 **if** $|R_T| > m + w$ **then**
19   **return** $\perp, \Gamma$
20 **end**
21 **return** Σ_T, Γ

where $\Gamma = ([\Sigma_t]_{t\le T}, [R_t]_{t \le T}, T)$ is the transcript of intermediate covariances and removed indices, and for $\tilde{x} = x_{1:n/2} - x_{n/2+1:n}$ (as in Line 1) define the leave-one-out covariance

$$\widehat{\Sigma}_{-i} := \begin{cases} \widehat{\Sigma} - \frac{1}{n} 1\{i \in R_T\}\, \tilde{x}_i \tilde{x}_i^T & \text{if } \widehat{\Sigma} \ne \perp \\ \perp & \text{otherwise,} \end{cases} \tag{2b}$$

which is $\widehat{\Sigma}$ whenever COVSAFE does not remove index pair $(i, n/2+i) \in [n]^2$.

3.2. Stable mean estimation

The second component of the private mean estimation algorithm is a sample mean estimator, adding noise commensurate with an estimated (positive semidefinite) noise covariance that we abstractly call $A \in \mathbb{R}^{d \times d}$. The procedure MEANSAFE removes elements x_i of the data x that are "too far" from the bulk of the data, measured by $\|x_i - x_{i'}\|_A$, using randomization to be sure that the removed indices are appropriately private. The algorithm uses TOPk to select groups of indices that contain too many outlying datapoints, then removes all data associated with these groups. By evaluating (random) groups of data, the procedure enforces privacy in that if the majority of the data are appropriately close to a center point as measured by covariance, then few groups have large diameter, and adding or removing a single datapoint x_i can only effect the removal of one group and the method may privately return a noisy empirical mean. When many datapoints are outliers, the method is likely to return $\perp$ regardless of the behavior of any individual datapoint.

Algorithm 2 Stable Mean Estimation (MEANSAFE)

Input : data $x_{1:n}$, PSD matrix $A \in \mathbb{R}^{d\times d}$
Params: threshold B, batchsize b, threshold k
Noise : $\mathcal{S} = (S_1, \ldots, S_{n/b}) \in \mathcal{P}_{n,b}$, $z, z' \in \mathbb{R}^{n/b}$, $w \in \mathbb{R}$, $z^{\mathsf{N}} \in \mathbb{R}^d$
Output: mean estimate $\widetilde{\mu}$

for $j \in [n/b]$ **do**
  $D_j \leftarrow \log(\operatorname{diam}_A(x_{S_j}))$
end
$\widetilde{D} \leftarrow$ TOPk$(D, k; z, z')$
$R \leftarrow \emptyset$, $t \leftarrow 0$ `/* initialize removed indices to empty */`
for $j \in [n/b]$ **do**
  if $\widetilde{D}_j \neq \perp$ ***and*** $\widetilde{D}_j > \log(\sqrt{B}/4)$ **then**
    $R \leftarrow R \cup S_j \quad t \leftarrow t + 1$
  end
end
$\widehat{\mu} \leftarrow \frac{1}{n-|R|} \sum_{i \notin R} x_i$
if $t > \frac{2k}{3} + w$ **then**
  $\widetilde{\mu} \leftarrow \perp$
else
  $\widetilde{\mu} \leftarrow \widehat{\mu} + A^{1/2} z^{\mathsf{N}}$
end
$\Gamma \leftarrow \big(D, \widetilde{D}, R, t, \widehat{\mu}\big)$
return $\widetilde{\mu}, \Gamma$

For use in Section 4, as with COVSAFE, we assign notation to the outputs of MEANSAFE. Let $x \in \mathbb{R}^{n\times d}$ be an arbitrary sample and A an arbitrary positive semidefinite matrix. For parameters defining the supposed bound B on $\|x_i - x_j\|_\Sigma^2$, group size b, acceptable outlier count k, and privacy random variables $(\mathcal{S}, Z, Z', W, Z^{\mathsf{N}})$, all to be specified later, define

$$(\widetilde{\mu}(x, A), \Gamma(x, A)) := \texttt{MEANSAFE}_{B,b,k}(x, A; \mathcal{S}, Z, Z', W, Z^{\mathsf{N}}). \tag{3}$$

3.3. The private mean estimation algorithm

Given COVSAFE and MEANSAFE, Algorithm 3 (PRIVMEAN) combines the two to perform private mean estimation. First, PRIVMEAN computes a stable covariance estimate via COVSAFE, and assuming the returned covariance estimate $\widehat{\Sigma} \neq \perp$, then computes a trimmed mean to which it adds Gaussian noise with covariance proportional to $\widehat{\Sigma}$ using MEANSAFE. Theorem 6 in Section 4 shows that the parameter choices guarantee privacy.

We remark briefly on the runtime of PRIVMEAN. Each iteration of the **while** loop (beginning in Line 4) of COVSAFE involves a $d \times d$ matrix inversion whose runtime is $O(d^3)$

Algorithm 3 Covariance Adaptive Private Mean Estimation (PRIVMEAN)

Input : data $x_{1:n}$
Params: threshold B, privacy budget (ε, δ)
Output: mean estimate $\widetilde{\mu}$

$m \leftarrow \frac{16}{\varepsilon} \log \frac{1}{\delta}$, $m_{\max} \leftarrow m + \frac{16}{\varepsilon} \log \frac{1+e^{\varepsilon/4}}{\delta}$
$\sigma_Z \leftarrow \frac{32\sqrt{e}B(m_{\max}+1)}{n\varepsilon}$, $\sigma_{W_{\text{cov}}} \leftarrow \frac{16}{\varepsilon}$
$Z_{\text{cov}} \sim \mathsf{Lap}(\sigma_Z)^{n/2+1}$, $W_{\text{cov}} \sim \mathsf{Lap}(\sigma_{W_{\text{cov}}})$
$\widehat{\Sigma}, \Gamma_{\text{cov}} \leftarrow \texttt{COVSAFE}_{B,m}(x; Z_{\text{cov}}, W_{\text{cov}})$
if $\widehat{\Sigma} = \perp$ **then**
  return $\perp$
end
$b \leftarrow 1 + \log_2 \frac{6n^2}{\delta}$, $k \leftarrow \frac{24}{\varepsilon} \log \frac{3}{\delta} - 3$
$\sigma_{\text{top}} \leftarrow \frac{8k}{n\varepsilon} \frac{B\sqrt{e}}{1-B\sqrt{e}/n}$, $\sigma_{\mathsf{N}} \leftarrow \frac{20b\sqrt{B}}{n\varepsilon} \exp(3\sigma_{\text{top}} \log \frac{12n}{b\delta})$, $\sigma_{W_{\text{mean}}} \leftarrow \frac{8}{\varepsilon}$
$\mathcal{S} \sim \mathsf{Uni}(\mathcal{P}_{n,b})$, $Z_{\text{top}}, Z'_{\text{top}} \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma_{\text{top}})^{n/b}$, $W \sim \mathsf{Lap}(\sigma_{W_{\text{mean}}})$, $Z^{\mathsf{N}} \sim \mathsf{N}(0, \sigma_{\mathsf{N}}^2 I_{d\times d})$
$\widetilde{\mu}, \Gamma_{\text{mean}} \leftarrow \texttt{MEANSAFE}_{B,b,k}(x, \widehat{\Sigma}; \mathcal{S}, Z_{\text{top}}, Z'_{\text{top}}, W, Z^{\mathsf{N}})$
return $\widetilde{\mu}$

followed by (at most) n matrix-vector products, each of whose runtime is $O(d^2)$. Because $n \geq d$, the total runtime of each iteration is thus $O(nd^2)$. We may modify COVSAFE without changing its behavior to terminate after $m+W_{\text{cov}}$ iterations, as rejecting more than $m+W_{\text{cov}}$ indices guarantees that COVSAFE (and hence PRIVMEAN) returns $\perp$ (see Line 18). With high probability, we have $m + W_{\text{cov}} = O(\frac{1}{\varepsilon} \log \frac{1}{\delta})$, giving runtime $O(nd^2 \min\{n, \frac{1}{\varepsilon} \log \frac{1}{\delta}\})$. COVSAFE's runtime dominates MEANSAFE's, and so the total (high probability) runtime of PRIVMEAN is $O(nd^2 \min\{n, \frac{1}{\varepsilon} \log \frac{1}{\delta}\})$. As an aside, we may convert this expected runtime into a worst-case runtime of the same order without affecting the privacy of PRIVMEAN by truncating W_{cov} to scale $\frac{1}{\varepsilon} \log \frac{1}{\delta}$.

4. Main privacy result

The analysis of PRIVMEAN is fairly involved, though there are four key building blocks. The first two conditions involve what we term *internal* and *external* leave-one-out stability of the covariance estimates (2a) and (2b) COVSAFE returns. These conditions require that the covariance estimates (2) are appropriately stable, both in terms of removing a single element contributing to the covariance estimate $\widehat{\Sigma}$ on input x and in terms of stability across two inputs x, x' whose transformations in Line 1 of COVSAFE, i.e., $\tilde{x} = x_{1:n/2} - x_{n/2+1:n}$ and $\tilde{x}' = x'_{1:n/2} - x'_{n/2+1:n}$, differ only in a single element. Letting $0 \leq a < \infty$ be a constant to be determined later and $\gamma \in (0, 1)$ be a probability, consider the conditions

(C.i) *Internal leave-one-out stability.* Let $\widehat{\Sigma}$ and $\widehat{\Sigma}_{-i}$ be the outputs (2) of COVSAFE on an arbitrary input x of size n. Then for any $i \in [n/2]$, with probability at least $1 - \gamma$,

$$d_{\text{psd}}(\widehat{\Sigma}, \widehat{\Sigma}_{-i}) \leq \frac{a}{n} \quad \text{or} \quad \widehat{\Sigma} = \perp .$$

(C.ii) *External leave-one-out stability.* Let $\widehat{\Sigma}$ and $\widehat{\Sigma}'$ be the outputs of COVSAFE on inputs x, x' of size n such that $\tilde{x}$ and $\tilde{x}'$ differ only in index $i \in [n/2]$, where $\widehat{\Sigma}_{-i}$ and $\widehat{\Sigma}'_{-i}$ are defined as in (2b). Then

$$\widehat{\Sigma}_{-i} \stackrel{d}{=}_{\varepsilon,\delta} \widehat{\Sigma}'_{-i}.$$

The second two conditions involve the noisy truncated mean estimate (3) MEANSAFE outputs. The first of these conditions (C.iii) essentially states MEANSAFE is stable over inputs x and x' differing in a single element, while the second states that MEANSAFE applied with identical input samples x, x' but different covariance estimates A, A' is stable so long as A, A' are close in the same sense as in Condition (C.i).

(C.iii) *Mean sample stability.* Let $\widetilde{\mu}(A, x)$ be the mean MEANSAFE outputs (3) on input covariance A and data x, and let x, x' differ only in one element. Then

$$\widetilde{\mu}(x, A) \stackrel{d}{=}_{\varepsilon,\delta} \widetilde{\mu}(x', A).$$

(C.iv) *Mean covariance stability.* If $d_{\rm psd}(A, A') \le \frac{a}{n}$, then $\widetilde{\mu}(x, A) \stackrel{d}{=}_{\varepsilon,\delta} \widetilde{\mu}(x, A')$.

Conditions (C.i)–(C.iv) form the basic privacy building blocks to show that PRIVMEAN is differentially private, and the following proposition—a warm-up for the full Theorem 6 to come—shows how we may relatively easily synthesize the conditions to achieve privacy.

Proposition 5 *Let samples x, x' differ in a single element, and let $\widehat{\Sigma}$ and $\widetilde{\mu}(x, \widehat{\Sigma})$ and $\widehat{\Sigma}'$ and $\widetilde{\mu}(x', \widehat{\Sigma}')$ be the covariance and mean estimates (2) and (3) for inputs x and x', respectively. Let Conditions (C.i)–(C.iv) hold. Then*

$$\widetilde{\mu}(x, \widehat{\Sigma}) \stackrel{d}{=}_{4\varepsilon,(e^{3\varepsilon}+e^{\varepsilon})\delta+(e^{2\varepsilon}+1)(\delta+\gamma)} \widetilde{\mu}(x', \widehat{\Sigma}').$$

Proof As x and x' are adjacent, there exists $i \in [n/2]$ such that $\tilde{x}_{-i} = \tilde{x}'_{-i}$. We have a string of approximate distributional equalities that, together with the transitivity of distributional closeness implied by group privacy (Lemma 3), make the proposition immediate. First, we show that conditions (C.i) and (C.iv) imply

$$\widetilde{\mu}(x, \widehat{\Sigma}) \stackrel{d}{=}_{\varepsilon,\delta+\gamma} \widetilde{\mu}(x, \widehat{\Sigma}_{-i}) \quad \text{and} \quad \widetilde{\mu}(x', \widehat{\Sigma}') \stackrel{d}{=}_{\varepsilon,\delta+\gamma} \widetilde{\mu}(x', \widehat{\Sigma}'_{-i}).$$

We prove the first equality as the argument for the second is identical. Treating x as fixed, let $\mathcal{E}$ be the event that $d_{\rm psd}(\widehat{\Sigma}, \widehat{\Sigma}_{-i}) \le \frac{a}{n}$ or $\widehat{\Sigma} = \perp$. Then for any measurable set O we have

$$\begin{aligned}
\mathbb{P}(\widetilde{\mu}(x, \widehat{\Sigma}) \in O) &= \mathbb{E}\left[\mathbb{P}(\widetilde{\mu}(x, \widehat{\Sigma}) \in O \mid \widehat{\Sigma}) 1\{\mathcal{E}\}\right] + \mathbb{E}\left[\mathbb{P}(\widetilde{\mu}(x, \widehat{\Sigma}) \in O \mid \widehat{\Sigma}) 1\{\mathcal{E}^c\}\right] \\
&\stackrel{(i)}{\le} \mathbb{E}\left[\left(e^{\varepsilon}\mathbb{P}(\widetilde{\mu}(x, \widehat{\Sigma}_{-i}) \in O \mid \widehat{\Sigma}_{-i}) + \delta\right) 1\{\mathcal{E}\}\right] + \mathbb{P}(\mathcal{E}^c) \\
&\le e^{\varepsilon}\mathbb{P}(\widetilde{\mu}(x, \widehat{\Sigma}_{-i}) \in O) + \delta + \gamma,
\end{aligned}$$

where inequality (i) is Condition (C.iv) and the final inequality follows from the γ probability bound in Condition (C.i). Second, we have the distributional approximations

$$\widetilde{\mu}(x, \widehat{\Sigma}_{-i}) \stackrel{d}{=}_{\varepsilon,\delta} \widetilde{\mu}(x, \widehat{\Sigma}'_{-i})$$

by Condition (C.ii), because post-processing preserves distributional closeness (Lemma 4). Finally, we observe from the mean sample stability condition (C.iii) that

$$\widetilde{\mu}(x, \widehat{\Sigma}') \stackrel{d}{=}_{\varepsilon,\delta} \widetilde{\mu}(x', \widehat{\Sigma}').$$

Combining each distributional equality, we have

$$\widetilde{\mu}(x, \widehat{\Sigma}) \stackrel{d}{=}_{\varepsilon,\delta+\gamma} \widetilde{\mu}(x, \widehat{\Sigma}_{-i}) \stackrel{d}{=}_{\varepsilon,\delta} \widetilde{\mu}(x, \widehat{\Sigma}'_{-i}) \stackrel{d}{=}_{\varepsilon,\delta+\gamma} \widetilde{\mu}(x, \widehat{\Sigma}') \stackrel{d}{=}_{\varepsilon,\delta} \widetilde{\mu}(x', \widehat{\Sigma}').$$

Apply Lemma 3. ■

Finally, then, we come to our main privacy theorem, which verifies that the procedures making up `PRIVMEAN` indeed satisfy Conditions (C.i)–(C.iv) with appropriate constants. We state the theorem here and defer the proof to Appendix B, where we make precise the constants that appear in the conditions.

Theorem 6 *Let $B < \infty$, $\delta \in (0,1)$, and let $x, x' \in (\mathbb{R}^d)^n$ be adjacent samples, and let $\varepsilon \le 8$. Define $\delta' = (e^{3\varepsilon/4} + e^{\varepsilon/4})\delta + 2(e^{\varepsilon/2}+1)\delta$ and let $m \in \mathbb{N}$ be as in line 1 of `PRIVMEAN`. Assume that $\delta \le \frac{1}{n}$ and n is large enough that*

$$n \ge \frac{128\sqrt{e}B \log \frac{n(1+e^{\varepsilon/4})}{\delta}}{\varepsilon}\left(m + 1 + \frac{16}{\varepsilon}\log\frac{1+e^{\varepsilon/4}}{\delta}\right) = O(1)\frac{B\log^2\frac{1}{\delta}}{\varepsilon^2}.$$

Then $\texttt{PRIVMEAN}_{B,(\varepsilon,\delta)}(x)$ is (ε, δ')-differentially private.

As a brief remark, the condition $\varepsilon \le 8$ is only for convenience; a minor modification of the proof allows arbitrary ε at the expense of a more convoluted theorem statement but in which n remains of the same order.

5. Accuracy analysis

The second important component of our analysis of `PRIVMEAN` is its accuracy. We provide two accuracy results: the first (Theorem 7) covers the case in which the data is sub-Gaussian, where we assume the method has some knowledge of the sub-Gaussian parameter of the sampling distribution. Of course, it is unreasonable to assume that a given distribution is sub-Gaussian or that we know its sub-Gaussian norms, and thus we extend `PRIVMEAN` via a procedure that adapts to the actual scale of the data in Alg. 5 below.

Throughout this section, we let P be a distribution on $\mathbb{R}^d$ with mean μ and covariance Σ, and we assume $X_i \stackrel{\text{iid}}{\sim} P$, $i = 1, \ldots, n$. The classical (non-private) sample mean and covariance are $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\overline{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n/2}(X_i - X_{n/2+i})(X_i - X_{n/2+i})^T$. We assume throughout that P enjoys certain concentration properties, though we emphasize that our methods will be adaptive to the parameters we specify here.

Assumption A1 (Sample concentration) *Let $c_1 \ge 64e$. There exists $\beta \in (0,1)$ and $M < \infty$ such that for $X_i \stackrel{\text{iid}}{\sim} P$ with $\mathbb{E}[X] = \mu$ and $\mathsf{Cov}(X) = \Sigma$, the event*

$$\mathcal{E}_{\text{samp}} := \left\{ \max_{i\in[n]} \|X_i - \mu\|_\Sigma^2 \le M^2/c_1 \text{ and } \tfrac{1}{2}\Sigma \preceq \overline{\Sigma}_n \preceq \tfrac{3}{2}\Sigma \right\}$$

occurs with probability at least $1 - \beta$.

We give our promised accuracy guarantee whenever Assumption A1 holds, letting c be a numerical constant. We defer the proof of Theorem 7 to Appendix E.

Theorem 7 *Let $\varepsilon > 0$ and $\delta \le \frac{1}{n}$ be privacy parameters and let Assumption A1 hold with β and M. Let $B \ge M^2$ and suppose $n \ge \frac{c}{\varepsilon^2} B \log^2 \frac{1}{\delta}$. Let $\widetilde{\mu} = \texttt{PRIVMEAN}_{B,\varepsilon,\delta}(X_{1:n})$. Then with probability at least $1 - (\beta + 5\delta)$, $\widetilde{\mu} \neq \perp$ and*

$$\left\|\widetilde{\mu} - \overline{X}_n\right\|_\Sigma \le \frac{c\sqrt{B \max\{\mathrm{rank}(\Sigma), \log\frac{1}{\delta}\}\log(\frac{1}{\delta})}}{n\varepsilon}.$$

Because $\mathbb{E}[\|X_i - \mu\|_\Sigma^2] = \mathrm{rank}(\Sigma)$, the constant M^2 typically scales at least as $\mathrm{rank}(\Sigma)$. In Appendix F.2, we demonstate that if the vectors $Z_i = \Sigma^{\dagger/2}(X_i - \mu)$ are τ^2-sub-Gaussian, meaning that $\mathbb{E}[e^{\langle Z,v\rangle}] \le e^{\|v\|_2^2\tau^2/2}$ for all v, then Assumption A1 holds with the parameter $M^2 = O(1)\tau^2(\mathrm{rank}(\Sigma) + \log\frac{n}{\beta})$, while if we have the moment bound $\mathbb{E}[\|X_i - \mu\|_\Sigma^p] < \infty$ for some $p \ge 4$, then $M = o(n^{1/p})$.

Given the necessarily slowed rates of convergence for private mean estimators when random variables have only p moments (Barber and Duchi, 2014), it is essential to be adaptive to the actual scale (and number of moments) of the problem. We therefore develop ADAMEAN, which automatically tunes the threshold parameter B by repeatedly calling PRIVMEAN and doubling B until $\widetilde{\mu} \neq \perp$. The key is that upon termination of ADAMEAN, the effective B is at most twice the realized scale $O(1)\max_{i\le n}\|Z_i\|^2$ of the random variables, while the total privacy loss is bounded.

Algorithm 4 Fully Adaptive Private Mean Estimation (ADAMEAN)

Input : data $x_{1:n}$
Params: privacy budget (ε, δ)
Output: mean estimate $\widetilde{\mu}$

for $t = 1, 2, \dots$ **do**
  $\widetilde{\mu}_t \leftarrow \texttt{PRIVMEAN}_{2^{t-1},(\varepsilon/t^2,\delta/t^2)}(x)$ **if** $\widetilde{\mu}_t \neq \perp$ **then**
    return $\widetilde{\mu}_t$
  end
end

Theorem 8 (Accuracy of ADAMEAN) *Let $\varepsilon > 0$ and $\delta \le \frac{1}{n}$ be privacy parameters and let event $\mathcal{E}_{\mathrm{samp}}$ hold. Let $s = \max\{1, \log_2 4M^2\}$ and assume $n \ge \frac{c\max\{\mathrm{rank}(\Sigma), M^2\}s^2}{\varepsilon^2}\log^2\frac{s^2}{\delta}$. Let $\widetilde{\mu} = \texttt{ADAMEAN}_{\varepsilon,\delta}(X_{1:n})$. Then with probability at least $1 - (5 + \pi^2/3)\delta$,*

$$\left\|\widetilde{\mu} - \overline{X}_n\right\|_\Sigma \le \frac{cs^2\log(\frac{s^2}{\delta})\max\{M\sqrt{\mathrm{rank}(\Sigma)}, M\log(\frac{s^2}{\delta}), 1\}}{n\varepsilon}. \tag{4}$$

See Appendix F.1 for a proof.

6. Discussion

The simplicity of mean estimation in classical statistics belies the sophistication necessary to adaptively and accurately estimate a mean under differential privacy constraints. While we have developed (nearly) minimax optimal procedures for mean estimation, a number of questions remain open, and we hope that we or others will tackle them. From a practical perspective, while our procedure is implementable, the numerical constant factors we have maintained to guarantee privacy—in addition to the logarithmic factors in n and $\log \frac{1}{\delta}$—may make effective use of the procedure challenging. From a theoretical perspective, it is still interesting to attempt to remove the logarithmic factors present in our bounds. Additionally, while we can adapt to weaker than sub-Gaussian moment bounds (via the method ADAMEAN), it may be possible to provide a sharper procedure or tighter analysis to achieve optimal dependence on dimension d and privacy level ε, as in the case that p moments are available, our results appear to be roughly a factor of $(\sqrt{d}/\varepsilon)^{1/p}$ loose (recall Examples 2). It will be interesting to see the extent of possibilities for differentially private estimation in these more general cases.

References

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016.

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

Apple Differential Privacy Team. Learning with privacy at scale, 2017. Available at https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In *Proceedings of the Thirty Fifth Annual Conference on Computational Learning Theory*, 2022.

Rina Foygel Barber and John C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv:1412.4451 [math.ST]*, 2014.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. CoinPress: Practical private mean and covariance estimation. In *Advances in Neural Information Processing Systems 33*, 2020.

Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *arXiv:2106.13329 [cs.LG]*, 2021.

Gavin Brown, Samuel B. Hopkins, and Adam Smith. Fast, sample-efficient, affine-invariant private mean and covariance estimation for subgaussian distributions. *arXiv:2301.12250 [cs.LG]*, 2023.

Richard Chen, Alex Gittens, and Joel A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, to appear, 2012.

Harald Cramér. *Mathematical Methods in Statistics.* Princeton University Press, 1946.

Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-dimensional Robust Statistics.* Cambridge University Press, 2022.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on the Theory of Computing*, pages 371–380, 2009.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3 & 4):211–407, 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006b.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on the Theory of Computing*. ACM, 2014.

Ulfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS)*, 2014.

Simson Garfinkel, John Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society (WPES)*, pages 133–137, 2018.

Frank Hampel, Elvezio Ronchetti, Peter Rousseeuw, and Werner Stahel. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, 1986.

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

Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the Fifty-Fourth Annual ACM Symposium on the Theory of Computing*, pages 1406–1417, 2022a.

Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. *arXiv:2212.05015 [cs.DS]*, 2022b.

Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. In *Advances in Neural Information Processing Systems 34*, volume 34, pages 25993–26004, 2021.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley and Sons, second edition, 2009.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan R. Ullman. Privately learning high-dimensional distributions. *arXiv:1805.00216 [cs.DS]*, 2018.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems 35*, 2022.

Pravesh K. Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Proceedings of the Thirty Fifth Annual Conference on Computational Learning Theory*, pages 1–55, 2022.

Aref N. Dajani Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Leclerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The modernization of statistical disclosure limitation at the U.S. Census Bureau. Available online at https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf, 2017.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302 – 1338, 2000.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

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

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of the Thirty Fifth Annual Conference on Computational Learning Theory*, pages 1167–1246, 2022.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

Gang Qiao, Weijie Su, and Li Zhang. Oneshot differentially private top-k selection. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2):3–22, 2017.

John W. Tukey. A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, volume 2 of *Stanford Studies in Mathematics and Statistics*. Stanford University Press, 1960.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

Roman Vershynin. *High Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2019.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Appendix A. Standard Private Mechanisms

We use several known mechanisms, and our procedures rely on their distributional closeness properties. The first is the `TOPk` mechanism, which (approximately) returns the largest k elements of a sample. In our analysis, it will be convenient to call the procedures we develop with noise as an argument to allow easier tracking of distributional closeness.

Algorithm 5 Top-k DP (`TOPk`)

Input : data $x \in \mathbb{R}^p$, threshold k
Noise : $\xi_1, \xi_2 \in \mathbb{R}^p$
Output: $R \subseteq [n]$ such that $|R| = k$, $\tilde{x} \in (\mathbb{R} \cup \{\perp\})^p$

```
y_1 ← x + ξ_1
y_2 ← x + ξ_2
R ← index set comprising the k largest y_{1,j}'s
for j ∈ [p] do
    if j ∈ R then
        x̃_j ← y_{2,j}
    else
        x̃_j ← ⊥
    end
end
return x̃
```

Lemma 9 (`TOPk` mechanism, Qiao et al. (2021), Theorem 2.1) *Let $\gamma, \varepsilon \in \mathbb{R}_+$. Let $x, x' \in \mathbb{R}^p$ be such that $\|x - x'\|_\infty \leq \gamma$. Then for $\xi_1, \xi_2 \sim \mathsf{Lap}(\frac{2k\gamma}{\varepsilon})^p$,*

$$\texttt{TOPk}(x, k; \xi_1, \xi_2) \stackrel{d}{=}_{\varepsilon,0} \texttt{TOPk}(x', k; \xi_1, \xi_2).$$

As our procedures rely on adding Gaussian noise, we require two distributional closeness results for normal distributions. We prove them here for completeness, as they are both tweaks of existing results (Dwork and Roth, 2014; Mironov, 2017).

Lemma 10 (Gaussians, distinct means) *Let $\mu_1, \mu_2 \in \mathbb{R}^d$ and let $\Sigma \in \mathbb{R}^{d \times d}$ be PSD. Suppose $\|\mu_1 - \mu_2\|_\Sigma \le \rho$ and define*

$$\tau = \begin{cases} \frac{\rho}{\varepsilon}\sqrt{2\log\frac{5}{4\delta}} & \text{if } 0 < \varepsilon \le 1 \\ \rho/\left(\sqrt{2\log\frac{1}{\delta} + 2\varepsilon} - \sqrt{2\log\frac{1}{\delta}}\right) & \text{otherwise.} \end{cases}$$

Then $\mathsf{N}(\mu_1, \tau^2\Sigma) \overset{d}{=}_{\varepsilon,\delta} \mathsf{N}(\mu_2, \tau^2\Sigma)$.

Proof The first case follows from Dwork and Roth (2014, Theorem 3.22). For the second, we use Mironov's Rényi-differential privacy (Mironov, 2017). The Rényi α-divergence between distributions P and Q is $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int(\frac{dP}{dQ})^\alpha dQ$, and (Mironov, 2017, Proposition 3) if $D_\alpha(P\|Q) \le c$, then for all measurable A and $\delta > 0$ we have $P(A) \le \exp(c + \frac{\log(1/\delta)}{\alpha-1})Q(A) + \delta$. The Rényi divergence for Gaussians has the explicit form

$$D_\alpha\left(\mathsf{N}(\mu_1, \tau^2\Sigma)\|\mathsf{N}(\mu_2, \tau^2\Sigma)\right) = \frac{\alpha}{2\tau^2}\|\mu_1 - \mu_2\|_\Sigma^2.$$

When $\rho \ge \|\mu_1 - \mu_2\|_\Sigma$, we set $\beta = \alpha - 1$ and see that for ε satisfying

$$\varepsilon = \frac{\rho^2}{2\tau^2} + \frac{\beta\rho^2}{2\tau^2} + \frac{\log(1/\delta)}{\beta}$$

we obtain $\mathsf{N}(\mu_1, \tau^2\Sigma) \overset{d}{=}_{\varepsilon,\delta} \mathsf{N}(\mu_2, \tau^2\Sigma)$. Choosing β to minimize the preceding ε we obtain $\varepsilon = \frac{\rho^2}{2\tau^2} + \frac{\rho}{\tau}\sqrt{2\log\frac{1}{\delta}}$, and solving for $\eta = \frac{1}{\tau}$ in $\frac{\rho^2}{2}\eta^2 + \sqrt{2\log\frac{1}{\delta}}\rho\eta - \varepsilon$ yields

$$\tau = \frac{1}{\eta} = \frac{\rho}{\sqrt{2\log\frac{1}{\delta} + 2\varepsilon} - \sqrt{2\log\frac{1}{\delta}}}$$

is always sufficient to guarantee $\mathsf{N}(\mu_1, \tau^2\Sigma) \overset{d}{=}_{\varepsilon,\delta} \mathsf{N}(\mu_2, \tau^2\Sigma)$. ∎

Brown et al. (2021, Lemma 4.15) essentially give the next result, but we allow low rank covariance matrices.

Lemma 11 (Gaussians, distinct covariances) *Let $\mu \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ be PSD and satisfy $d_{\mathrm{psd}}(\Sigma_1, \Sigma_2) \le \gamma < \infty$. Then $\mathsf{N}(\mu, \Sigma_1) \overset{d}{=}_{\varepsilon,\delta} \mathsf{N}(\mu, \Sigma_2)$ for $\varepsilon \ge 6\gamma\log(2/\delta)$.*

Proof Without loss of generality, we may assume $\mu = 0$. We first reduce to the case where Σ_1 and Σ_2 are full-rank. Because $d_{\mathrm{psd}}(\Sigma_1, \Sigma_2) < \infty$, we have immediately that there exists a vector space $V \subset \mathbb{R}^d$ with $V = \mathsf{Col}(\Sigma_1) = \mathsf{Col}(\Sigma_2)$. Letting $k = \dim(\mathsf{Col}(\Sigma_1))$, take $U \in \mathbb{R}^{d\times k}$ to be an orthonormal matrix such that $V = \mathsf{Col}(U)$. The random variables $X \sim \mathsf{N}(0, \Sigma_1)$ and $Y \sim \mathsf{N}(0, \Sigma_2)$ have support V and multiplication by U^T is an isomorphism between V and $\mathbb{R}^k$, so $X \overset{d}{=}_{\varepsilon,\delta} Y$ if and only if $U^T X \overset{d}{=}_{\varepsilon,\delta} U^T Y$. Of course, $U^T X \sim \mathsf{N}\left(0, U^T\Sigma_1 U\right)$ and $U^T Y \sim \mathsf{N}\left(0, U^T\Sigma_2 U\right)$ and both $U^T\Sigma_1 U$ and $U^T\Sigma_2 U$ are full rank. The orthogonality of U gives $d_{\mathrm{psd}}(U^T\Sigma_1 U, U^T\Sigma_2 U) = d_{\mathrm{psd}}(\Sigma_1, \Sigma_2) \le \gamma$. Hence, by showing the lemma for the full-rank matrices $U^T\Sigma_1 U$ and $U^T\Sigma_2 U$, we will have shown the claim for Σ_1 and Σ_2.

We proceed with the full-rank case with an argument similar to Brown et al. (2021, Lemma 4.15). Define $D_1 = \Sigma_1^{1/2}\Sigma_2^{-1}\Sigma_1^{1/2} - I$ and $D_2 = \Sigma_2^{1/2}\Sigma_1^{-1}\Sigma_2^{1/2} - I$. As D_1 has the same spectrum as $\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - I$, we have by assumption that $\|D_1\|_* \le \gamma$ and $\|D_2\|_* \le \gamma$.

Let f_1 be the density of $P_1 = \mathsf{N}(0, \Sigma_1)$ and f_2 that of $P_2 = \mathsf{N}(0, \Sigma_2)$. Then, to show (ε, δ)-closeness, it suffices to show $\ell(W) := \left|\log \frac{f_1(W)}{f_2(W)}\right| \le \varepsilon$ with probability at least $1-\delta$ when W is drawn from either P_1 or P_2. By symmetry, it suffices to only show this bound for the case when $W \sim P_1$. Expanding ℓ, we have

$$\ell(w) = \left|\frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \frac{1}{2}w^T(\Sigma_2^{-1} - \Sigma_1^{-1})w\right| \le \frac{1}{2}\left|w^T(\Sigma_2^{-1} - \Sigma_1^{-1})w\right| + \frac{1}{2}\left|\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right|. \tag{5}$$

The final term is independent of w and has the bound

$$\begin{aligned}\left|\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right| &= \max\left\{\log\det(\Sigma_2^{1/2}\Sigma_1^{-1}\Sigma_2^{1/2}), \log\det(\Sigma_1^{1/2}\Sigma_2^{-1}\Sigma_1^{1/2})\right\} \\ &\le \max\{\mathrm{tr}(D_2), \mathrm{tr}(D_1)\} \le \max\{\|D_2\|_*, \|D_1\|_*\} \le \gamma,\end{aligned}$$

where the first inequality holds because $\log\det(A) \le \mathrm{tr}(A - I)$ for any positive definite A.

Now we bound the first term on the right hand side of inequality (5) with high probability. Since $W \sim P_1$, the whitened random variable $Z = \Sigma_1^{-1/2}W \sim \mathsf{N}(0, I)$. We then have

$$\left|W^T(\Sigma_2^{-1} - \Sigma_1^{-1})W\right| = \left|Z^T D_1 Z\right|,$$

and so by the Hanson-Wright inequality (e.g. Vershynin, 2019, Thm. 6.2.1), we have with probability at least $1-\delta$ that

$$\left|Z^T D_1 Z\right| \le |\mathrm{tr}(D_1)| + 2\|D_1\|_{\mathrm{Fr}}\sqrt{\log\frac{2}{\delta}} + 2\|D_1\|_{\mathrm{op}}\log\frac{2}{\delta} \le 5\gamma\log\frac{2}{\delta},$$

where the inequality holds because $\|D_1\|_{\mathrm{op}} \le \|D_1\|_{\mathrm{Fr}} \le \|D_1\|_* \le \gamma$ and $\log\frac{2}{\delta} \ge 1$. Thus $\ell(W) \le 6\gamma\log\frac{2}{\delta} \le \varepsilon$ with probability at least $1-\delta$. ∎

We conclude with a standard guarantee for Laplacian random vectors (e.g. Dwork and Roth, 2014, Thm. 3.6).

Lemma 12 (Laplace mechanism) *Let $\alpha, \beta > 0$ and $Z \overset{\mathrm{iid}}{\sim} \mathsf{Lap}(\beta/\alpha)$. Then for any $A \subseteq \mathbb{R}^m$ and $\eta \in \mathbb{R}^m$ such that $\|\eta\|_1 \le \beta$,*

$$\mathbb{P}(Z \in A) \le \exp(\alpha)\mathbb{P}(Z \in A + \eta).$$

Appendix B. Proof of Theorem 6

By Proposition 5, it suffices to verify Conditions (C.i)–(C.iv), where we demonstrate each holding with parameters $(\varepsilon/4, \delta)$. Throughout the proof, the value $m \in \mathbb{N}$ (line 1 in `PRIVMEAN`) and parameter $B < \infty$ remain tacit, as the privacy guarantee holds regardless.

First, we consider Conditions (C.i) and (C.ii) on the covariance estimates. We prove the coming two lemmas in Section C, which begins with preliminaries that we require for their proofs before giving the proofs proper. Our first lemma provides sufficient conditions to verify Condition (C.i), internal stability. Let $z \in \mathbb{R}^{n/2+1}$ and $w \in \mathbb{R}$ be variables—these will be random to allow privacy presently, but we use them for the definition—and let

$$\widehat{\Sigma}(x, z, w), \left([R_t]_{t=0}^{T}, [\Sigma_t]_{t=0}^{T}, T\right) := \texttt{COVSAFE}_{B,m}(x; z, w), \tag{6a}$$

where we leave the dependence of the transcript $([R_t], [\Sigma_t], T)$ on (x, z, w) implicit, and redefine $\widehat{\Sigma}_{-i}$ as in the definitions (2):

$$\widehat{\Sigma}_{-i}(x, z, w) := \widehat{\Sigma}(x, z, w) - \frac{1}{n} 1\{i \in R_T\}\, \tilde{x}_i \tilde{x}_i^T \tag{6b}$$

whenever $\widehat{\Sigma}(x, z, w) \neq \perp$, and $\perp$ otherwise. Then we have

Lemma 13 (Internal stability) *Let $Z_j \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma)$ and $i \in [n/2]$. Then with probability at least $1 - \exp(-\frac{1}{4\sigma})$, either $\widehat{\Sigma}(x, Z, w) = \perp$ or*

$$d_{\text{psd}}(\widehat{\Sigma}(x, Z, w), \widehat{\Sigma}_{-i}(x, Z, w)) \leq \frac{1}{1 - B\sqrt{e}/n} \frac{B\sqrt{e}}{n}.$$

See Section C.2 for a proof of Lemma 13. Turning to the condition (C.ii) on external stability of COVSAFE, we compare the leave-one-out covariances $\widehat{\Sigma}_{-i}(x, z, w)$ and $\widehat{\Sigma}_{-i}(x', z, w)$ with input samples x and x', respectively, with identical (randomization) parameters z, w. Recalling $\tilde{x} = x_{1:n/2} - x_{n/2+1:n}$ and $\tilde{x}' = x'_{1:n/2} - x'_{n/2+1:n}$, we have the following guarantee:

Lemma 14 (External stability) *Let $\gamma \in (0, 1)$, $Z_j \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z)$, $W \sim \mathsf{Lap}(\sigma_{W_{\text{cov}}})$ and $k \in \mathbb{N}$. Define $m_{\max} = m + \sigma_{W_{\text{cov}}} \log \frac{1}{\gamma}$ and $\alpha = \frac{1}{\sigma_{W_{\text{cov}}}} + \frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}$ and $\beta = \frac{\gamma}{2} + \frac{n}{2}\exp(-\frac{1}{4\sigma_Z})$. For all $i \in [n/2]$, if $\tilde{x}_{-i} = \tilde{x}'_{-i}$ then*

$$\widehat{\Sigma}_{-i}(x, Z, W) \overset{d}{=}_{2\alpha,(1+e^{\alpha})\beta} \widehat{\Sigma}_{-i}(x', Z, W).$$

The lemma effectively shows that the sets of removed indices R and R' are stable, and as they determine $\widehat{\Sigma}_{-i}$ and $\widehat{\Sigma}'_{-i}$, this yields their closeness. See Section C.3 for a proof of Lemma 14.

We turn to the guarantees of MEANSAFE, realizing Conditions (C.iii) and (C.iv). Recall the definition (3) of $\widetilde{\mu}(x, A)$ as the output of MEANSAFE on input $x \in (\mathbb{R}^d)^n$ with positive semidefinite $A \in \mathbb{R}^{d\times d}$, with parameters bound B, batchsize b, and threshold $k \in \mathbb{N}$, and $\mathcal{S}$, Z, Z', and W as noise. We take $Z, Z' \in \mathbb{R}^{n/b}, W \in \mathbb{R}$ to be Laplacian random variables, $Z^{\mathsf{N}} \in \mathbb{R}^d$ to be Gaussian, and $\mathcal{S}$ to be a uniformly random partition of $[n]$ into blocks of size n/b; we track their scales in giving our distributional approximation guarantees.

To more cleanly state a general sample stability guarantee, which we may use to verify Condition (C.iii), we define a number of additional parameters whose values we can determine. Let the batchsize $b \in \mathbb{N}$ and threshold $k > 0$ satisfy $b \geq 4$ and $2b(k+1) \leq n$. Let $\beta_1, \gamma \in (0, 1)$, let $\alpha \geq 0$, and let $\sigma_{\text{top}} > 0$ and $\sigma_{W_{\text{mean}}} > 0$. Define the constants

$$\Delta := \frac{5b\sqrt{B}}{2n} \exp\left(3\sigma_{\text{top}} \log \frac{2n}{b\gamma}\right), \;\; \beta_2 := \frac{1}{2} e^{-(k/3-1)/\sigma_{W_{\text{mean}}}} + \gamma + n^2 2^{1-b}$$

and

$$\sigma_{\mathsf{N}} = \begin{cases} \frac{\Delta}{\alpha}\sqrt{2\log\frac{5}{4\beta_1}} & \text{if } 0 \le \alpha \le 1 \\ \frac{\Delta}{\sqrt{2\log\frac{1}{\beta_1}+2\alpha}-\sqrt{2\log\frac{1}{\beta_1}}} & \text{otherwise.} \end{cases}$$

With these, we have a mean-sample stability result from which Condition (C.iii) develops:

Lemma 15 *Let the conditions above hold and let* $Z_j, Z_j' \stackrel{\text{iid}}{\sim} \mathsf{Lap}(\sigma_{\text{top}}), W \sim \mathsf{Lap}(\sigma_{W_{\text{mean}}}), Z_j^{\mathsf{N}} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma_{\mathsf{N}}^2)$ *in* (3)*. If* x *and* x' *are adjacent, then*

$$\widetilde{\mu}(x, A) \stackrel{d}{=}_{\alpha+1/\sigma_{W_{\text{mean}}}, \beta_1+\beta_2} \widetilde{\mu}(x', A).$$

See Section D.1 for a proof.

The last building block in the argument is to demonstrate Condition (C.iv), that the estimates $\widetilde{\mu}(x, A)$ and $\widetilde{\mu}(x, A')$ are close when A, A' are close. For this, we give the following lemma with general noise parameters.

Lemma 16 *Let* $b, k \in \mathbb{N}$*,* $\beta \in (0, 1)$*, and* $a, \sigma_{\mathsf{N}}, \alpha_2 > 0$*. Define* $\alpha_1 = \frac{6a}{n}\log\frac{2}{\beta}$*, and define the noise scale* $\sigma_{\text{top}} = \frac{ka}{n\alpha_2}$*. Then for* $Z_j, Z_j' \stackrel{\text{iid}}{\sim} \mathsf{Lap}(\sigma_{\text{top}}), Z_j^{\mathsf{N}} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma_{\mathsf{N}}^2)$*, if*

$$d_{\text{psd}}(A, A') \le \frac{a}{n} \quad \textit{then} \quad \widetilde{\mu}(x, A) \stackrel{d}{=}_{\alpha_1+\alpha_2, \beta} \widetilde{\mu}(x, A').$$

See Section D.2 for a proof.

For the final step, we put all the pieces together to prove the theorem. We give each of the lemmas so the associated condition (of (C.i)–(C.iv)) holds with parameters $(\varepsilon/4, \delta)$, after which we can then apply Proposition 5 directly. We do this in a somewhat odd order because of the dependence on the noise scale between the different lemmas, beginning with

Condition (C.ii). For $Z_j \stackrel{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z)$ and $W \sim \mathsf{Lap}(\sigma_{W_{\text{cov}}})$, we use Lemma 14 to guarantee Condition (C.ii) that $\widehat{\Sigma}_{-i} \stackrel{d}{=}_{\varepsilon/4,\delta} \widehat{\Sigma}_{-i}'$. From the lemma statement, we have $\widehat{\Sigma}_{-i} \stackrel{d}{=}_{2\alpha,(1+e^{\alpha})\beta} \widehat{\Sigma}_{-i}'$, where $\alpha = \frac{1}{\sigma_{W_{\text{cov}}}} + \frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}$, $\beta = \frac{\gamma}{2} + \frac{n}{2}\exp(-\frac{1}{4\sigma_Z})$, and $m_{\max} = m + \sigma_{W_{\text{cov}}}\log\frac{1}{\gamma}$ for the m in line 1 of PRIVMEAN (though privacy does not depend on its value). We first achieve $2\alpha \le \frac{\varepsilon}{4}$. Setting $\sigma_{W_{\text{cov}}} = \frac{16}{\varepsilon}$, it is sufficient that σ_Z is large enough that $\frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z} \le \frac{\varepsilon}{16}$, i.e.,

$$\sigma_Z \ge \frac{32\sqrt{e}B(m_{\max}+1)}{n\varepsilon} = \frac{32\sqrt{e}B}{n\varepsilon}\left(m + 1 + \frac{16\log\frac{1}{\gamma}}{\varepsilon}\right).$$

For the δ privacy component, we wish to have $(1 + e^{\alpha})\beta \le \delta$. As we have guaranteed $\alpha \le \frac{\varepsilon}{4}$, taking $\gamma = \frac{\delta}{1+e^{\varepsilon/4}}$ and making sure σ_Z is small enough that $n\exp(-\frac{1}{4\sigma_Z}) \le \gamma = \frac{\delta}{1+e^{\varepsilon/4}}$ suffices. For this, it is evidently sufficient that $\frac{1}{\sigma_Z} \ge 4\log\frac{n(1+e^{\varepsilon/4})}{\delta}$, i.e., (substituting for σ_Z)

$$n \ge \frac{128\sqrt{e}B\log\frac{n(1+e^{\varepsilon/4})}{\delta}}{\varepsilon}\left(m + 1 + \frac{16}{\varepsilon}\log\frac{1+e^{\varepsilon/4}}{\delta}\right)$$

guarantees $\widehat{\Sigma}_{-i} \stackrel{d}{=}_{\varepsilon/4,\delta} \widehat{\Sigma}_{-i}'$.

Condition (C.i). In Lemma 13, if the scale of the noise σ_Z on $Z_j \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z)$ satisfies $\exp(-\frac{1}{4\sigma_Z}) \le \gamma$, Condition (C.i) holds. The choice of σ_Z to satisfy Condition (C.ii) above and the lower bound on n are evidently sufficient.

Condition (C.iii). Lemma 15 guarantees that if $Z_j^{\mathsf{N}} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_{\mathsf{N}}^2)$, $W \sim \mathsf{Lap}(\sigma_{W_{\text{mean}}})$, and $\mathcal{S} \sim \mathsf{Uni}(\mathcal{P}_{n,b})$, the call to MEANSAFE in line 11 of PRIVMEAN gives $\widetilde{\mu}(x, A) \overset{d}{=}_{\alpha+1/\sigma_{W_{\text{mean}}},\beta_1+\beta_2} \widetilde{\mu}(x', A)$, with Δ, β_2 and σ_{N} as defined in the lemma. To achieve $\alpha + \frac{1}{\sigma_{W_{\text{mean}}}} = \frac{\varepsilon}{4}$, take $\sigma_{W_{\text{mean}}} = \frac{8}{\varepsilon}$ and choose $\alpha = \frac{\varepsilon}{8}$. To achieve $\beta_1 + \beta_2 \le \delta$, choose $\beta_1 = \frac{\delta}{2}$ and then recognize that $\beta_2 \le \frac{\delta}{2}$ as long as $\gamma \le \frac{\delta}{6}$, $n^2 2^{1-b} \le \frac{\delta}{6}$ (or $b \ge \log_2 \frac{6n^2}{\delta} + 1$) and $\frac{1}{2}\exp(-\frac{k/3+1}{\sigma_{W_{\text{mean}}}}) \le \frac{\delta}{6}$ (or $k \ge \frac{24}{\varepsilon} \log \frac{3}{\delta} - 3$). Thus, we arrive at

$$\sigma_{\mathsf{N}} = \frac{8\Delta}{\varepsilon}\sqrt{\log \frac{5}{2\delta}} = \frac{20\sqrt{B}b}{n\varepsilon} \exp\left(3\sigma_{\text{top}} \log \frac{12n}{b\delta}\right)$$

for (any) $b \ge 2\log \frac{6n}{\delta} + 1$ so long as $\frac{\varepsilon}{8} \le 1$. (Otherwise we may use the alternative value for σ_{N} preceding Lemma 15, which the (ε, δ)-differential privacy guarantee of Lemma 10 justifies.)

Condition (C.iv). The last condition to verify is that $\widetilde{\mu}(x, A) \overset{d}{=}_{\varepsilon/4,\delta} \widetilde{\mu}(x, A')$ for close enough A, A'. For this, we use Lemma 16, which guarantees that $\widetilde{\mu}(x, A) \overset{d}{=}_{\alpha_1+\alpha_2,\delta} \widetilde{\mu}(x, A')$ for $\alpha_1 = \frac{6a}{n} \log \frac{2}{\delta}$, where we take $a = \frac{B\sqrt{e}}{1-B\sqrt{e/n}}$ via Lemma 13, and arbitrary $\alpha_2 > 0$. Set $\alpha_2 = \frac{\varepsilon}{8}$ and obtain $\sigma_{\text{top}} = \frac{8ka}{n\varepsilon}$. When $n \ge \frac{48a}{\varepsilon} \log \frac{2}{\delta}$, we have $\alpha_1 \le \frac{\varepsilon}{8}$, and so the desired privacy holds.

Making appropriate substitutions gives that each of conditions (C.i)–(C.iv) holds with parameters $(\varepsilon/4, \delta)$. Proposition 5 gives the theorem.

Appendix C. Proofs for stable covariance estimation

In this section, we provide the proofs of Lemmas 13 and 14, though we begin with a collection of preliminary results that allow us to actually prove the main two lemmas. In the proofs, we refer to each execution of the while loop beginning in Line 4 of COVSAFE as an *iteration* of COVSAFE and use the transcript Γ as a convenient means for tracking the full execution of COVSAFE through all iterations.

C.1. Properties of COVSAFE

We first formalize deterministic properties about the execution of COVSAFE, giving conditions under which outputs of COVSAFE are quite stable. In the sequel, we use these to give sets to which the noise variables Z and W belong with high probability, guaranteeing stability. Recall the notation (6) that $\widehat{\Sigma}(x, z, w)$ is the output of COVSAFE on input sample x and noise $z \in \mathbb{R}^{n/2+1}, w \in \mathbb{R}$, with transcript $\Gamma = ([R_t]_{t \le T}, [\Sigma_t]_{t \le T}, T)$ depending implicitly on (x, z, w), and $\widehat{\Sigma}_{-i}(x, z, w)$ is the corresponding leave-one-out covariance. We shorthand $\widehat{\Sigma} = \widehat{\Sigma}(x, z, w)$ and $\widehat{\Sigma}_{-i} = \widehat{\Sigma}_{-i}(x, z, w)$ and take $\tilde{x} = x_{1:n/2} - x_{n/2+1:n}$ as in Line 1 of COVSAFE.

Lemma 17 gives necessary and sufficient conditions for pruning $\tilde{x}_i$ in iteration $t+1$ of COVSAFE, i.e., $i \in R_{t+1}$, and Lemma 18 gives similar conditions for ever pruning $\tilde{x}_i$ (that is, whether $i \in R_T$).

Lemma 17 *Index $i \in R_{t+1}$ if and only if $\log\left(\|\tilde{x}_i\|_{\Sigma_t}^2\right) + z_i + z_{n/2+1} > \log(B)$.*

Proof The "if" direction is immediate from the condition for adding an element to R_{t+1} (see line 7 of COVSAFE). For the other direction, if $i \in R_{t+1}$ then (again from the same condition) we must have for some $s \le t$ that

$$\log\left(\|\tilde{x}_i\|_{\Sigma_s}^2\right) + z_i + z_{n/2+1} > \log(B).$$

Because $s \le t$, we have $R_s \subset R_t$ and therefore $\Sigma_s \succeq \Sigma_t$, this in turn implies $\log\left(\|\tilde{x}_i\|_{\Sigma_t}^2\right) + z_i + z_{n/2+1} > \log(B)$. ■

Lemma 18 *Index $i \notin R_T$ if and only if $\log(\|\tilde{x}_i\|_{\Sigma_T}^2) + z_i + z_{n/2+1} \le \log(B)$.*

Proof Observe that $\Sigma_{T-1} = \Sigma_T$ because the inner while loop of COVSAFE terminates only if the algorithm prunes no observations in the previous iteration (see line 12 of COVSAFE). Then the claim follows by applying Lemma 17 with $t = T-1$. ■

Finally, we may completely characterize $\widehat{\Sigma}_{-i}$ via the removed indices $R_{T,-i}$ and the threshold $m+w$, as prescribed by the lemma below.

Lemma 19 *$\widehat{\Sigma}_{-i} = \frac{1}{n}\sum_{j \notin R_{T,-i} \cup \{i\}} \tilde{x}_j \tilde{x}_j^T$ if and only if $|R_T| \le m + w$.*

Proof The claim follows immediately from the return condition in Line 18 of COVSAFE, as $\widehat{\Sigma} \neq \perp$ implies $\widehat{\Sigma}_{-i} = \widehat{\Sigma} - \frac{1}{n}1\{i \notin R_T\}\, \tilde{x}_i \tilde{x}_i^T$, where $\widehat{\Sigma} = \Sigma_T = \frac{1}{n}\sum_{j \notin R_T} \tilde{x}_j \tilde{x}_j^T$ by definition. ■

C.2. Proof of Lemma 13

We shorthand $\widehat{\Sigma} = \widehat{\Sigma}(x, Z, w)$ and $\widehat{\Sigma}_{-i} = \widehat{\Sigma}_{-i}(x, Z, w)$ throughout the proof. Assume that $\widehat{\Sigma} \neq \perp$, as otherwise the result is trivial, and recall $Z_j \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma)$. Observe if $i \in R_T$ we have $\widehat{\Sigma} = \widehat{\Sigma}_{-i}$ by definition; thus, we need only consider $i \notin R_T$. Proceeding, Lemma 18 gives

$$\log\left(\|\tilde{x}_i\|_{\widehat{\Sigma}}^2\right) + Z_i + Z_{n/2+1} \le \log(B)$$

for all $i \notin R_T$, from which it follows that $\|\tilde{x}_i\|_{\widehat{\Sigma}}^2 \le B\sqrt{e}$ whenever $Z_i + Z_{n/2+1} \ge -1/2$. We now use that $\|\tilde{x}_i\|_{\widehat{\Sigma}}^2 \le B\sqrt{e}$ implies that $d_{\rm psd}(\widehat{\Sigma}, \widehat{\Sigma}_{-i})$ is small, which follows from the following linear algebraic lemma.

Lemma 20 *Let $A \in \mathbb{R}^{d \times d}$ be positive semi-definite and $a \in \mathbb{R}^d$ satisfy $\|a\|_A^2 < 1$. Then*

$$d_{\rm psd}(A, A - aa^T) \le \frac{1}{1 - \|a\|_A^2} \|a\|_A^2 .$$

Proof Define $C = A - aa^T$ for shorthand. We first establish that $\mathsf{Col}(C) = \mathsf{Col}(A)$. Because $\|a\|_A^2$ is finite, it follows that $a \in \mathsf{Col}(A)$ and so $\mathsf{Col}(C) \subset \mathsf{Col}(A)$. On the other hand, by expanding C we have

$$C = A^{1/2}(I - A^{\dagger/2}aa^TA^{\dagger/2})A^{1/2} \succeq (1 - \|a\|_A^2)A, \tag{7}$$

thus implying that $\mathsf{Col}(C) = \mathsf{Col}(A)$.

We also have from (7) that $C^\dagger \preceq \frac{1}{1-\|a\|_A^2}A^\dagger$, and so

$$\left\|C^{\dagger/2}(A-C)C^{\dagger/2}\right\|_* = \left\|C^{\dagger/2}aa^TC^{\dagger/2}\right\|_* = \|a\|_C^2 \le \frac{\|a\|_A^2}{1-\|a\|_A^2}.$$

A parallel calculation yields $\|A^{\dagger/2}(C-A)A^{\dagger/2}\|_* = \|a\|_A^2$, proving the claim. ■

Lemma 20 immediately shows that $\widehat{\Sigma}_{-i} = \widehat{\Sigma} - \frac{1}{n}\tilde{x}_i\tilde{x}_i^T 1\{i \in R_T\}$ satisfies

$$d_{\mathrm{psd}}(\widehat{\Sigma}, \widehat{\Sigma}_{-i}) \le \frac{1}{1 - B\sqrt{e}/n}\frac{B\sqrt{e}}{n}$$

whenever $Z_i + Z_{n/2+1} \ge -\frac{1}{2}$. To show that this occurs with high probability, we use the following result, which follows from the observation that if $c \ge 0$, then for any independent variables X, Y we have $\mathbb{P}(X + Y > c) \le \mathbb{P}(X > c/2) + \mathbb{P}(Y > c/2)$ by a union bound:

Observation C.1 *Let $X, Y \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma)$ and $c \ge 0$. Then $\mathbb{P}(X + Y > c) \le \exp(-\frac{c}{2\sigma})$.*

We see that $\mathbb{P}(Z_i + Z_{n/2+1} < -\frac{1}{2}) \le \exp(-\frac{1}{4\sigma})$ as claimed.

C.3. Proof of Lemma 14

The proof of Lemma 14 comes in four steps. The crux of the proof is a coupling argument where, via the running assumption $\tilde{x}_{-i} = \tilde{x}'_{-i}$, we equate the execution of `COVSAFE` on x to that on x' by perturbing Z in a careful way that changes the distribution of Z little. Step one in this approach, which we provide in Lemma 21, is a deterministic lemma relating the collections R and R' of indices `COVSAFE` removes on adjacent inputs x and x' via the noise input values z. In the second and third steps, which consist of Lemmas 22 and 23 respectively, we construct a map $\pi : \mathbb{R}^{n/2+1} \to \mathbb{R}^{n/2+1}$ such that Z and $\pi(Z)$ have similar distributions and for which $\widehat{\Sigma}_{-i}(x, z, W)$ and $\widehat{\Sigma}_{-i}(x', \pi(z), W)$ (recall the definition (6b)) likewise have similar distributions for all z, where we use the randomness in W for the latter distributional approximation. Lemma 22 relates the distributions of the removed indices $R_{T,-i}$, while Lemma 23 relates the probabilities that `COVSAFE` aborts and returns $\perp$. In Sec. C.3.1, we finally synthesize the intermediate lemmas to give the proof of Lemma 14.

Our first step is the deterministic lemma relating the collections of removed indices.

Lemma 21 *Let $z, z' \in \mathbb{R}^{n/2+1}$ and $w \in \mathbb{R}$, and let*

$$\widehat{\Sigma}, \left([R_t]_{t=0}^T, [\Sigma_t]_{t=0}^T, T\right) := \texttt{COVSAFE}_{B,m}(x; z, w),$$
$$\widehat{\Sigma}', \left([R'_t]_{t=0}^{T'}, [\Sigma'_t]_{t=0}^{T'}, T'\right) := \texttt{COVSAFE}_{B,m}(x'; z', w).$$

Assume $\tilde{x}'_i = 0$. The following hold.

(a) *If* $z'_j + z'_{n/2+1} \geq z_j + z_{n/2+1}$ *for all* $j \in R_{T,-i}$*, then* $R_{T,-i} \subset R'_{T',-i}$.

(b) *If* $z'_j + z'_{n/2+1} \geq z_j + z_{n/2+1}$ *for all* $j \notin R'_{T',-i}$*, then* $R_{T,-i} \subset R'_{T',-i}$.

Additionally assume that $n \geq 2B\sqrt{e}$ *and* $z_i + z_{n/2+1} \geq -1/2$*. Then the following hold.*

(c) *If* $z'_j + z'_{n/2+1} \leq z_j + z_{n/2+1} - 2B\sqrt{e}/n$ *for all* $j \in R'_{T',-i}$*, then* $R'_{T',-i} \subset R_{T,-i}$.

(d) *If* $z'_j + z'_{n/2+1} \leq z_j + z_{n/2+1} - 2B\sqrt{e}/n$ *for all* $j \notin R_{T,-i}$*, then* $R'_{T',-i} \subset R_{T,-i}$.

Proof We prove each claim by induction over $t \in \mathbb{N}$.

Proceeding with the first claim (a), observe trivially that $R_{0,-i} = \emptyset \subset R'_{T',-i}$. Now suppose for the sake of induction that $R_{t,-i} \subset R'_{T',-i}$. If $t = T$ then there is nothing to show, so we take $t < T$. Our assumption that $\tilde{x}_{-i} = \tilde{x}'_{-i}$ and $\tilde{x}'_i = 0$ implies $\Sigma_t \succeq \Sigma'_{T'}$. Thus, for all $j \in R_{t,-i} \subset R_{T,-i}$ we have

$$\begin{aligned}\log(B) &\overset{(i)}{<} \log\left(\|\tilde{x}_j\|^2_{\Sigma_t}\right) + z_j + z_{n/2+1} \overset{(ii)}{\leq} \log\left(\|\tilde{x}_j\|^2_{\Sigma'_{T'}}\right) + z_j + z_{n/2+1} \\ &\overset{(iii)}{\leq} \log\left(\|\tilde{x}_j\|^2_{\Sigma'_{T'}}\right) + z'_j + z'_{n/2+1},\end{aligned}$$

where inequality (i) follows from Lemma 17 (applied with noise z), inequality (ii) because $\Sigma_t \succeq \Sigma'_{T'}$, and inequality (iii) is by assumption in case (a). Lemma 18 (with data x' and noise z') then gives that $j \in R'_{T',-i}$ if and only if $\log B < \log(\|\tilde{x}_j\|^2_{\Sigma'_{T'}}) + z'_j + z'_{n/2+1}$, and as $j \in R_{t,-i}$ was arbitrary we have $R_{t+1,-i} \subset R'_{T',-i}$. This completes the inductive step and so $R_{t,-i} \subset R'_{T',-i}$ for all $t \leq T$ and the first claim holds.

The proof of claim (b) relies on a similar inductive argument as that for the first claim (a). Equivalent to the inclusion $R_{T,-i} \subset R'_{T',-i}$ is that if $j \notin R'_{T',-i}$, then $j \notin R_{T,-i}$. Consider $j \notin R'_{T',-i}$, and begin with the inductive assumption that $R_{t,-i} \subset R'_{T',-i}$; it suffices to show that $j \notin R_{t+1,-i}$. Because $\Sigma_t \succeq \Sigma'_{T'}$ by construction of Σ_t, we obtain

$$\begin{aligned}\log\left(\|\tilde{x}_j\|^2_{\Sigma_t}\right) + z_j + z_{n/2+1} &\leq \log\left(\|\tilde{x}_j\|^2_{\Sigma'_{T'}}\right) + z_j + z_{n/2+1} \\ &\overset{(i)}{\leq} \log\left(\|\tilde{x}_j\|^2_{\Sigma'_{T'}}\right) + z'_j + z'_{n/2+1} \leq \log(B),\end{aligned}$$

where step (i) is by the assumption that $z_j + z_{n/2+1} \leq z'_j + z'_{n/2+1}$ in part (b) and the final inequality is Lemma 18. Applying Lemma 17 with the inequality $\log\left(\|\tilde{x}_j\|^2_{\Sigma_t}\right) + z_j + z_{n/2+1} \leq \log B$ then guarantees that $j \notin R_{t+1,-i}$ as desired, completing the proof of claim (b).

For the proof of claim (c), we induct on $R'_{t,-i}$ for $t \leq T'$ and must account for the possibility that $\Sigma_T \not\preceq \Sigma'_t$ even if $R'_{t,-i} \subset R_{T,-i}$, because Σ_T may include the term $\tilde{x}_i\tilde{x}_i^T$ (i.e., $i \notin R_T$). The base case for $t = 0$ is trivial, so assume that $R'_{t,-i} \subset R_{T,-i}$. If $i \notin R_T$, then Lemma 18 and the standing assumption that $z_i + z_{n/2+1} \geq -\frac{1}{2}$ guarantee that

$$\log\left(\|\tilde{x}_j\|^2_{\Sigma_T}\right) \leq \log B - z_i - z_{n/2+1} \leq \log B + \frac{1}{2} = \log(B\sqrt{e}),$$

i.e., $\|\tilde{x}_j\|^2_{\Sigma_T} \leq B\sqrt{e}$. We require the following technical observation about positive definite matrices, whose proof we temporarily defer.

Observation C.2 *Let $\Sigma \in \mathbb{R}^{d\times d}$ be positive semi-definite, $\alpha \ge 0$, and $u \in \mathbb{R}^d$. Define $\Sigma' := \Sigma - \alpha uu^T$. If $\|u\|_\Sigma^2 \le \frac{1}{2\alpha}$, then $\Sigma' \succeq \frac{1}{2}\Sigma$ and for any $v \in \mathsf{Col}(\Sigma)$,*

$$\left|\log\left(\|v\|_\Sigma^2\right) - \log\left(\|v\|_{\Sigma'}^2\right)\right| \le 2\alpha \|u\|_\Sigma^2 .$$

As $\|\tilde{x}_i\|_{\Sigma_T}^2 \le B\sqrt{e}$, Observation C.2 applies with $u = \tilde{x}_i$ and $\alpha = \frac{1}{n}$ when $n \ge 2B\sqrt{e}$, and thus

$$\log\left(\|v\|^2_{\Sigma_T - \frac{1}{n}1\{i \notin R_T\}\tilde{x}_i\tilde{x}_i^T}\right) \le \log\left(\|v\|_{\Sigma_T}^2\right) + \frac{2B\sqrt{e}}{n} \tag{8}$$

for all v, and in particular, for $v = \tilde{x}_j$ for each $j \in [n/2]$. On the other hand, regardless of whether $i \in R_T$, the inductive assumption that $R'_{t,-i} \subset R_{T,-i}$ guarantees that

$$\Sigma'_t \succeq \Sigma_T - \frac{1}{n}1\{i \notin R_T\}\,\tilde{x}_i\tilde{x}_i^T. \tag{9}$$

Considering $j \in R'_{t+1,-i}$, then, Lemma 17 implies

$$\begin{aligned}
\log B &< \log(\|\tilde{x}_j\|^2_{\Sigma'_t}) + z'_j + z'_{n/2+1} \\
&\overset{(i)}{\le} \log\left(\|\tilde{x}_j\|^2_{\Sigma_T - \frac{1}{n}1\{i\notin R_T\}\tilde{x}_i\tilde{x}_i^T}\right) + z'_j + z'_{n/2+1} \\
&\overset{(ii)}{\le} \log\left(\|\tilde{x}_j\|^2_{\Sigma_T}\right) + \frac{2B\sqrt{e}}{n} + z'_j + z'_{n/2+1} \\
&\overset{(iii)}{\le} \log\left(\|\tilde{x}_j\|^2_{\Sigma_T}\right) + z_j + z_{n/2+1}.
\end{aligned}$$

Here inequality (i) follows from the ordering relation (9); inequality (ii) holds because if $i \in R_T$, then $\Sigma_T = \Sigma_T - \frac{1}{n}1\{i \notin R_T\}\,\tilde{x}_i\tilde{x}_i^T$ and if $i \notin R_T$ then inequality (8) holds; the final inequality (iii) follows by assumption under claim (c). This gives the induction that $R'_{t+1,-i} \subset R_{T,-i}$, as Lemma 17 shows that $j \in R_{T,-i}$.

Claim (d) follows from an essentially identical induction argument, *mutatis mutandis*, as that for claim (c).

Proof of Observation C.2. We finally return to prove the claimed observation. That $\Sigma' \succeq \frac{1}{2}\Sigma$ follows by observing that $u \in \mathsf{Col}(\Sigma)$ and hence

$$\Sigma - \alpha uu^T = \Sigma^{1/2}\underbrace{(I - \alpha\Sigma^{\dagger/2}uu^T\Sigma^{\dagger/2})}_{\succeq(1/2)I}\Sigma^{1/2} \succeq \frac{1}{2}\Sigma.$$

This also implies that $\mathsf{Col}(\Sigma') = \mathsf{Col}(\Sigma)$.

To prove the remainder of the lemma, it suffices to show for $v \in \mathsf{Col}(\Sigma)$ that $\log(\|v\|_{\Sigma'}^2) \le \log(\|v\|_\Sigma^2) + 2\alpha\|u\|_\Sigma^2$, since the other direction is immediate from $\Sigma \succeq \Sigma'$. Observe that

$$(\Sigma - \alpha uu^T)^\dagger = \Sigma^{\dagger/2}(I - \alpha\Sigma^{\dagger/2}uu^T\Sigma^{\dagger/2})^{-1}\Sigma^{\dagger/2}.$$

By the inequality $I - \alpha\Sigma^{\dagger/2}uu^T\Sigma^{\dagger/2} \succeq (1 - \alpha\|u\|_\Sigma^2)I$ we have

$$(I - \alpha\Sigma^{\dagger/2}uu^T\Sigma^{\dagger/2})^{-1} \preceq (1 - \alpha\|u\|_\Sigma^2)^{-1}I \preceq (1 + 2\alpha\|u\|_\Sigma^2)I,$$

where the final inequality follows from the assumption that $\|u\|_\Sigma^2 \le \frac{1}{2\alpha}$. Combining this with the preceding display implies that

$$\Sigma'^{\dagger} \preceq (1 + 2\alpha \|u\|_\Sigma^2)\Sigma^{\dagger}$$

and so

$$\log\left(\|v\|_{\Sigma'}^2\right) \le \log\left((1 + 2\alpha\|u\|_\Sigma^2)\|v\|_\Sigma^2\right) \le \log\left(\|v\|_\Sigma^2\right) + 2\alpha\|u\|_\Sigma^2$$

as desired. ■

We move to the second step we outline at the beginning of this section, which relates the distributions of removed indices R_T in the execution of COVSAFE on adjacent inputs x and x'. The key idea is to construct a deterministic map π so that the execution of COVSAFE on input x with noise z and that on x' with noise $\pi(z)$ is similar—leveraging Lemma 21—and to show that the distributions of $\pi(Z)$ and Z are similar. Lemma 19 shows that the set of outlier indices $R_{T,-i}$ completely determines $\widehat{\Sigma}_{-i}$ except in the case that $\widehat{\Sigma}_{-i} = \perp$, which occurs with high probability if $|R_{T,-i}|$ is large, so the next lemma controls the distribution of the sets of removed indices. To state the lemma, we require a few events whose probabilities we can control. Recalling that $Z_j \stackrel{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z)$, define

$$\mathcal{E}_{\text{prune}} := \{z \in \mathbb{R}^{n/2+1} \mid z_j + z_{n/2+1} \ge -1/2 \text{ for all } j \in [n/2]\}. \tag{10}$$

To set notation for the remainder of the proof, we shorthand the definition (6a) as

$$\begin{aligned} \widehat{\Sigma}, ([R_t]_{t=0}^T, [\Sigma_t]_{t=0}^T, T) &:= \texttt{COVSAFE}_{B,m}(x; Z, W) \\ \widehat{\Sigma}', ([R'_t]_{t=0}^{T'}, [\Sigma'_t]_{t=0}^{T'}, T') &:= \texttt{COVSAFE}_{B,m}(x'; Z, W), \end{aligned} \tag{11}$$

and the definition (6b) as

$$\widehat{\Sigma}_{-i} = \widehat{\Sigma} - \frac{1}{n}1\{i \notin R_T\}\,\tilde{x}_i\tilde{x}_i^T \quad \text{and} \quad \widehat{\Sigma}'_{-i} = \widehat{\Sigma}' - \frac{1}{n}1\{i \notin R'_T\}\,\tilde{x}'_i\tilde{x}_i'^T,$$

where $\perp + v = \perp$ for any vector v.

We have the following distributional guarantee on the removed indices regardless of W.

Lemma 22 *Let $S \subset [n/2] \setminus \{i\}$ and define $\alpha = \frac{2\sqrt{e}B(|S|+1)}{n\sigma_Z}$. If $\tilde{x}'_i = 0$, then*

(a) $\mathbb{P}(R_{T,-i} = S, Z \in \mathcal{E}_{\text{prune}}) \le \exp(\alpha)\,\mathbb{P}(R'_{T',-i} = S)$.

(b) $\mathbb{P}(R'_{T',-i} = S, Z \in \mathcal{E}_{\text{prune}}) \le \exp(\alpha)\,\mathbb{P}(R_{T,-i} = S)$.

Proof The input noise Z completely determines R_T and $R'_{T'}$ in COVSAFE (see the **while** loop constructing R_t in lines 4–15). Consequently, we may define sets of input noise Z yielding a given set of removed indices, letting

$$\begin{aligned} \mathcal{Z}(S) &:= \{z \in \mathbb{R}^{n/2+1} \mid R_{T,-i} = S \text{ for } Z = z\} \\ \mathcal{Z}'(S) &:= \{z \in \mathbb{R}^{n/2+1} \mid R'_{T',-i} = S \text{ for } Z = z\}, \end{aligned}$$

so $Z \in \mathcal{Z}(S)$ is equivalent to $R_{T,-i} = S$. It suffices to show that

$$\begin{aligned} \mathbb{P}\left(Z \in \mathcal{Z}(S) \cap \mathcal{E}_{\text{prune}}\right) &\le e^{\alpha} \mathbb{P}\left(Z \in \mathcal{Z}'(S)\right) \quad \text{and} \\ \mathbb{P}\left(Z \in \mathcal{Z}'(S) \cap \mathcal{E}_{\text{prune}}\right) &\le e^{\alpha} \mathbb{P}\left(Z \in \mathcal{Z}(S)\right), \end{aligned} \tag{12}$$

as claim (a) follows via the first bound and claim (b) the second.

Proceeding with the first bound, define $\eta \in \mathbb{R}^{n/2+1}$ and $\pi : \mathbb{R}^{n/2+1} \to \mathbb{R}^{n/2+1}$ by

$$\eta_j := \begin{cases} 2B\sqrt{e}/n & j \in S \\ -2B\sqrt{e}/n & j = n/2+1 \\ 0 & \text{otherwise,} \end{cases} \qquad \pi(z) := z + \eta.$$

The deterministic removals Lemma 21 shows that on $Z \in \mathcal{E}_{\text{prune}}$, if we let $z = Z$ and $z' = \pi(z)$ so that $z_j + z_{n/2+1} \ge z'_j + z'_{n/2+1}$ for $j \in S$, then parts (a) and (d) of Lemma 21 give

$$\pi(\mathcal{Z}(S) \cap \mathcal{E}_{\text{prune}}) \subset \mathcal{Z}'(S).$$

The first bound in (12) then follows by the standard Laplacian ratio bounds in Lemma 12. Indeed, we have $Z_j \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z) = \mathsf{Lap}(\|\eta\|_1 \frac{\sigma_Z}{\|\eta\|_1})$ and $\|\eta\|_1 = \frac{2\sqrt{e}B(|S|+1)}{n}$. Then setting $\beta = \|\eta\|_1$ yields $\beta/\sigma_Z \le \alpha$, so we can apply Lemma 12 to obtain the claimed bound (12) via

$$\mathbb{P}(Z \in \mathcal{Z}(S) \cap \mathcal{E}_{\text{prune}}) \le e^{\alpha} \mathbb{P}(Z \in \pi(\mathcal{Z}(S) \cap \mathcal{E}_{\text{prune}})) \le e^{\alpha} \mathbb{P}(Z \in \mathcal{Z}'(S)).$$

The proof of the second bound (12) is essentially the same, only this time we let

$$\eta_j := \begin{cases} 2B\sqrt{e}/n & j \in S \\ 0 & \text{otherwise,} \end{cases} \qquad \pi(z) = z + \eta.$$

Then the event $Z \in \mathcal{E}_{\text{prune}}$ implies $\pi(Z) \in \mathcal{E}_{\text{prune}}$, as $\pi(Z)_j + \pi(Z)_{n/2+1} \ge Z_j + Z_{n/2+1}$ for all $j \in [n/2]$. We may thus appeal to cases (b) and (c) of Lemma 21 with the settings $z = \pi(Z)$, $z' = Z$ and $R'_{T',-n} = S$ and proceed with the same argument as above. ■

The third step we outline at the beginning of the proof of Lemma 14 is to relate the probabilities that `COVSAFE` aborts on neighboring inputs x and x'. Recall $\widehat{\Sigma}$ and $\widehat{\Sigma}'$ are the covariances `COVSAFE` outputs on inputs x and x', respectively, as in definition (11).

Lemma 23 *Let $\sigma_Z, \sigma_W > 0$, $Z_j \overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z)$, and $W \sim \mathsf{Lap}(\sigma_W)$ in definition (11). If $\tilde{x}_{-i} = \tilde{x}'_{-i}$ and $\tilde{x}'_i = 0$, then*

(a) $\mathbb{P}(\widehat{\Sigma} = \perp) \le \exp(\frac{1}{\sigma_W}) \mathbb{P}(\widehat{\Sigma}' = \perp)$.

(b) $\mathbb{P}(\widehat{\Sigma}' = \perp, Z \in \mathcal{E}_{\text{prune}}) \le \exp(\frac{2\sqrt{e}B}{n\sigma_Z}) \mathbb{P}(\widehat{\Sigma} = \perp)$.

Proof Let f denote the density of W, so that $f(w) = \frac{1}{2\sigma_{W_{\text{cov}}}} \exp(-|w|/\sigma_{W_{\text{cov}}})$, and thus $|\log \frac{f(w)}{f(w-1)}| \le \frac{1}{\sigma_{W_{\text{cov}}}}$ for all w, and recall the threshold $m \in \mathbb{N}$ in line 18 of `COVSAFE`. Proceeding with the first claim of the lemma, we have the following sequence of inequalities:

$$\mathbb{P}\left(\widehat{\Sigma} = \perp\right) \overset{(i)}{\le} \int \mathbb{P}\left(|R_T| > m + w\right) f(w) dw \le \int \mathbb{P}\left(|R_{T,-i}| > m + w - 1\right) f(w) dw$$

$$\overset{(ii)}{\leq} \int \mathbb{P}\left(\left|R'_{T',-i}\right| > m+w-1\right) f(w)dw$$
$$\leq \exp(1/\sigma_{W_{\text{cov}}}) \int \mathbb{P}\left(\left|R'_{T',-i}\right| > m+w-1\right) f(w-1)dw$$
$$\overset{(iii)}{=} \exp(1/\sigma_{W_{\text{cov}}})\mathbb{P}(\widehat{\Sigma}' = \perp).$$

Here, step (i) follows from Lemma 19 that $\widehat{\Sigma}' = \perp$ if and only if $|R'_T| > m + w$. Step (ii) follows from the coupling argument in Lemma 21 part (a) : because the noise Z is identical in both executions of COVSAFE$(x; Z, W)$ and COVSAFE$(x'; Z, W)$, we have $R_{T,-i} \subset R'_{T',-i}$. Step (iii) applies because of Lemma 19 again, as the assumption $\widetilde{x}'_i = 0$ guarantees $R'_{T',-i} = R'_{T'}$ (recall the rejection threshold in Line 7). Claim (a) follows.

For the claim (b), again applying Lemma 19 we have

$$\mathbb{P}\left(\widehat{\Sigma}' = \perp, Z \in \mathcal{E}_{\text{prune}}\right) = \int \mathbb{P}\left(\left|R'_{T',-i}\right| > m+w, Z \in \mathcal{E}_{\text{prune}}\right) f(w)dw$$

and

$$\int \mathbb{P}\left(\left|R_{T,-i}\right| > m+w\right) f(w)dw \leq \mathbb{P}\left(\widehat{\Sigma} = \perp\right).$$

Combining these displays, it thus suffices to show for all w that

$$\mathbb{P}\left(\left|R'_{T',-i}\right| > m+w, Z \in \mathcal{E}_{\text{prune}}\right) \leq \exp\left(\frac{2\sqrt{e}B}{n\sigma_Z}\right)\mathbb{P}\left(\left|R_{T,-i}\right| > m+w\right). \tag{13}$$

To this end, we adopt a similar tack as in the proof of Lemma 22, defining

$$\mathcal{Z}(w) := \{z \in \mathbb{R}^{n/2+1} \mid |R_{T,-i}| > m + w \text{ for } Z = z\}$$
$$\mathcal{Z}'(w) := \{z \in \mathbb{R}^{n/2+1} \mid \left|R'_{T',-i}\right| > m + w \text{ for } Z = z\}$$

and the single coordinate perturbation $\pi(z) := z + \eta$ for $\eta \in \mathbb{R}^{n/2+1}$ the vector with all zeros except that $\eta_{n/2+1} = \frac{2\sqrt{e}B}{n}$. Similar to our proof of Lemma 22, the mapping π guarantees that $z' = \pi(z)$ satisfies $z'_j + z'_{n/2+1} \leq z_j + z_{n/2+1} - \frac{2B\sqrt{e}}{n}$ for all j, which is precisely the condition for case (c) of Lemma 21, and so $R'_{T',-i} \subset R_{T,-i}$ irrespective of $R'_{T',-i}$ and $R_{T,-i}$. It thus holds simultaneously for all $w \in \mathbb{R}$ that

$$\pi\left(\mathcal{Z}'(w) \cap \mathcal{E}_{\text{prune}}\right) \subset \mathcal{Z}(w).$$

Noting that $\|\eta\|_1 = \frac{2\sqrt{e}B}{n}$, Lemma 12 on likelihood ratios for Laplace random variables then guarantees

$$\mathbb{P}(Z \in \mathcal{Z}'(w) \cap \mathcal{E}_{\text{prune}}) \leq \exp\left(\frac{2\sqrt{e}B}{n\sigma_Z}\right)\mathbb{P}(Z \in \pi(\mathcal{Z}'(w) \cap \mathcal{E}_{\text{prune}})) \leq \exp\left(\frac{2\sqrt{e}B}{n\sigma_Z}\right)\mathbb{P}(Z \in \mathcal{Z}(w)),$$

which is equivalent to inequality (13). ■

C.3.1. Finalizing proof of Lemma 14

By combining Lemmas 22 and 23, we can prove the stability of `COVSAFE`. Recall the set $\mathcal{E}_{\text{prune}}$ in (10) and that $W \sim \mathsf{Lap}(\sigma_{W_{\text{cov}}})$, and additionally define

$$\mathcal{E}_{\text{thr}} := \left(-\infty, \sigma_{W_{\text{cov}}} \log \frac{1}{\gamma}\right].$$

The key part of our argument is to show that when x and x' are adjacent but $\tilde{x}'_i = 0$, if the noise variables Z, W satisfy $Z \in \mathcal{E}_{\text{prune}}$ and $W \in \mathcal{E}_{\text{thr}}$, then for $\widehat{\Sigma}$ and $\widehat{\Sigma}'$ as in the call (11) the leave-one-out covariances $\widehat{\Sigma}_{-i}$ and $\widehat{\Sigma}'_{-i}$ are similar. We then bound the probabilities of the individual events and use a group composition argument to give the lemma for arbitrary $\tilde{x}'_i$.

With this in mind, let $A \in \mathbb{R}^{d\times d}$ and note that for any fixed sample x, $\widehat{\Sigma}$ and $\widehat{\Sigma}_{-i}$ can take only finitely many values. For

$$\alpha = \frac{1}{\sigma_{W_{\text{cov}}}} + \frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z},$$

we show for any $A \in \mathbb{R}^{d\times d} \cup \{\perp\}$ that

$$\mathbb{P}\left(\widehat{\Sigma}_{-i} = A, Z \in \mathcal{E}_{\text{prune}}, W \in \mathcal{E}_{\text{thr}}\right) \le \exp(\alpha)\mathbb{P}\left(\widehat{\Sigma}'_{-i} = A\right) \quad \text{and} \tag{14}$$

$$\mathbb{P}\left(\widehat{\Sigma}'_{-i} = A, Z \in \mathcal{E}_{\text{prune}}, W \in \mathcal{E}_{\text{thr}}\right) \le \exp(\alpha)\mathbb{P}\left(\widehat{\Sigma}_{-i} = A\right). \tag{15}$$

Lemma 23 already implies both inequalities (14) and (15) hold for $A = \perp$, so all that remains is to show the same for $A \in \mathbb{R}^{d\times d}$.

Proceeding first with inequality (14), let $f(w) = \frac{1}{2\sigma_{W_{\text{cov}}}} \exp(-|w|/\sigma_{W_{\text{cov}}})$ be the density of W and $S(A) := \{S \subset [n/2] \setminus \{i\} \mid A = \frac{1}{n}\sum_{j \notin S\cup\{i\}} \tilde{x}_j \tilde{x}_j^T\}$. Marginalizing over W gives

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\Sigma}_{-i} = A, Z \in \mathcal{E}_{\text{prune}}, W \in \mathcal{E}_{\text{thr}}\right) \\
&\overset{(i)}{=} \int_{\mathcal{E}_{\text{thr}}} \mathbb{P}\left(R_{T,-i} \in S(A), |R_T| \le m + w, Z \in \mathcal{E}_{\text{prune}}\right) f(w)dw \\
&\le \int \mathbb{P}\left(R_{T,-i} \in S(A), |R_{T,-i}| \le \min\{m+w, m_{\max}\}, Z \in \mathcal{E}_{\text{prune}}\right) f(w)dw,
\end{aligned}$$

where step (i) follows from the condition that $|R_T| \le m + w$ if and only if $\widehat{\Sigma}_{-i} \ne \perp$ from Lemma 19, and the final inequality follows because $\mathcal{E}_{\text{thr}} = \{w \mid w \le m_{\max}\}$. Continuing, note for each $S \in S(A)$, we can have $S = R_{T,-i}$ with $|R_{T,-i}| \le \min\{m+w, m_{\max}\}$ only if $|S| \le \min\{m+w, m_{\max}\} \le m_{\max}$, so that by case (a) of Lemma 22

$$\begin{aligned}
&\mathbb{P}(R_{T,-i} \in S(A), |R_{T,-i}| \le \min\{m+w, m_{\max}\}, Z \in \mathcal{E}_{\text{prune}}) \\
&\le \exp\left(\frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}\right) \mathbb{P}\left(R'_{T',-i} \in S(A), \left|R'_{T',-i}\right| \le m + w\right).
\end{aligned}$$

Returning to the integral above, we obtain inequality (14) by integrating and applying Lemma 19:

$$\int \mathbb{P}\left(R'_{T',-i} \in S(A), \left|R'_{T',-i}\right| \le m + w\right) f(w)dw = \mathbb{P}(\widehat{\Sigma}'_{-i} = A)$$

as $R'_{T',-i} = R'_{T'}$ because $\tilde{x}'_i = 0$ by assumption.

The proof of inequality (15) is essentially the same, only now we must take additional care to account for the possibility that $i \in R_T$. As in the preceding integral inequalities, Lemma 19 gives

$$\begin{aligned}&\mathbb{P}\left(\widehat{\Sigma}'_{-i} = A, Z \in \mathcal{E}_{\text{prune}}, W \in \mathcal{E}_{\text{thr}}\right)\\ &\le \int_{\mathcal{E}_{\text{thr}}} \mathbb{P}\left(R'_{T',-i} \in S(A), \left|R'_{T',-i}\right| \le \min\{m + w, m_{\max}\}, Z \in \mathcal{E}_{\text{prune}}\right) f(w)dw.\end{aligned}$$

In this case, with reasoning identical to that above, we apply case (b) of Lemma 22 to achieve

$$\begin{aligned}&\mathbb{P}\left(R'_{T',-i} \in S(A), \left|R'_{T',-i}\right| \le \min\{m + w, m_{\max}\}, Z \in \mathcal{E}_{\text{prune}}\right)\\ &\le \exp\left(\frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}\right)\mathbb{P}\left(R_{T,-i} \in S(A), |R_{T,-i}| \le m + w\right),\end{aligned}$$

so

$$\begin{aligned}&\mathbb{P}\left(\widehat{\Sigma}'_{-i} = A, Z \in \mathcal{E}_{\text{prune}}, W \in \mathcal{E}_{\text{thr}}\right)\\ &\le \exp\left(\frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}\right)\int \mathbb{P}\left(R_{T,-i} \in S(A), |R_{T,-i}| \le m + w\right) f(w)dw.\end{aligned}$$

We upper bound the final integral by noting that

$$\begin{aligned}\mathbb{P}\left(\widehat{\Sigma}_{-i} = A\right) &\stackrel{(\star)}{=} \int \mathbb{P}\left(R_{T,-i} \in S(A), |R_T| \le m + w\right) f(w)dw\\ &\ge \int \mathbb{P}\left(R_{T,-i} \in S(A), |R_{T,-i}| \le m + w - 1\right) f(w)dw,\end{aligned}$$

where $(\star)$ follows from Lemma 19, and then using $|\log \frac{f(w)}{f(w+1)}| \le \frac{1}{\sigma_{W_{\text{cov}}}}$ for all w to see that

$$\int \mathbb{P}\left(R_{T,-i} \in S(A), |R_{T,-i}| \le m + w\right) f(w)dw \le \exp\left(\frac{1}{\sigma_{W_{\text{cov}}}}\right)\mathbb{P}\left(\widehat{\Sigma}_{-i} = A\right),$$

which gives inequality (15) once we substitute $\alpha = \frac{1}{\sigma_{W_{\text{cov}}}} + \frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}$.

We combine inequalities (14) and (15) to get Lemma 14. For any set $C \subset \mathbb{R}^{d\times d} \cup \{\perp\}$,

$$\begin{aligned}\mathbb{P}(\widehat{\Sigma}_{-i} \in C) &\le \mathbb{P}(\widehat{\Sigma}_{-i} \in C, Z \in \mathcal{E}_{\text{prune}}, W \in \mathcal{E}_{\text{thr}}) + \mathbb{P}(Z \notin \mathcal{E}_{\text{prune}}) + \mathbb{P}(W \notin \mathcal{E}_{\text{thr}})\\ &\le e^{\alpha}\mathbb{P}(\widehat{\Sigma}'_{-i} \in C) + \mathbb{P}(Z \notin \mathcal{E}_{\text{prune}}) + \mathbb{P}(W \notin \mathcal{E}_{\text{thr}})\end{aligned}$$

by inequality (14). We then have the immediate bounds $\mathbb{P}(W \notin \mathcal{E}_{\text{thr}}) = \mathbb{P}(W > \sigma_{W_{\text{cov}}} \log \frac{1}{\gamma}) = \frac{1}{2}\exp(-\log\frac{1}{\gamma}) = \frac{\gamma}{2}$. Similarly, $\mathbb{P}(Z \notin \mathcal{E}_{\text{prune}}) \le \frac{n}{2}\exp(-\frac{1}{4\sigma_Z})$ by Observation C.1. The upper bound on $\mathbb{P}(\widehat{\Sigma}'_{-i} \in C)$ is similar but uses inequality (15).

To this point, we have shown that if x and x'' are adjacent samples differing only in that the difference $\tilde{x}_i = x_i - x_{n/2+i}$ may be non-zero while $\tilde{x}''_i = x''_i - x''_{n/2+i} = 0$, then returning to the notation (6) and identifying $\widehat{\Sigma}_{-i} = \widehat{\Sigma}_{-i}(x, Z, W)$ and $\widehat{\Sigma}'_{-i} = \widehat{\Sigma}_{-i}(x'', Z, W)$,

$$\widehat{\Sigma}_{-i}(x, Z, W) \stackrel{d}{=}_{\alpha,\beta} \widehat{\Sigma}_{-i}(x'', Z, W)$$

for $\alpha = \frac{1}{\sigma_{W_{\text{cov}}}} + \frac{2\sqrt{e}B(m_{\max}+1)}{n\sigma_Z}$ and $\beta = \frac{\gamma}{2} + \frac{n}{2}\exp(-\frac{1}{4\sigma_Z})$. Thus we obtain that if x' is any sample satisfying $x'_{-i} = x_{-i}$,

$$\widehat{\Sigma}_{-i}(x, Z, W) \stackrel{d}{=}_{\alpha,\beta} \widehat{\Sigma}_{-i}(x'', Z, W) \stackrel{d}{=}_{\alpha,\beta} \widehat{\Sigma}_{-i}(x', Z, W).$$

Using group composition (Lemma 3), we obtain

$$\widehat{\Sigma}_{-i}(x, Z, W) \stackrel{d}{=}_{2\alpha,\beta+e^{\alpha}\beta} \widehat{\Sigma}_{-i}(x', Z, W),$$

which is the desired Lemma 14.

Appendix D. Proofs for mean estimation

In this section, we provide the proofs of Lemmas 15 and 16. Throughout, we differentiate outputs of `MEANSAFE` on inputs x versus x' (or A versus A') via tick marks, so that (for example) $\widehat{\mu}$ corresponds to the mean in Line 11 of `MEANSAFE` on input sample x, or D'_j corresponds to the log-diameter in Line 2 of `MEANSAFE` on input sample x'. We will make this precise using the function $\Gamma(x, A)$ from (3), which is the transcript `MEANSAFE` outputs on input x, A.

D.1. Proof of Lemma 15

We shorthand $\widetilde{\mu}(x, A)$ and $\widetilde{\mu}(x', A)$ as $\widetilde{\mu}$ and $\widetilde{\mu}'$ respectively, and unpack the corresponding execution transcripts:

$$(D, \widetilde{D}, R, t, \widehat{\mu}) := \Gamma(x, A) \quad \text{and} \quad (D', \widetilde{D}', R', t', \widehat{\mu}') := \Gamma(x', A).$$

Throughout our arguments, $i \in [n]$ denotes the index at which the samples x, x' differ, that is, $x_{-i} = x'_{-i}$ while we may have $x_i \neq x'_i$.

The main idea in the proof of Lemma 15 is to first bound the sensitivity of the mean, showing that (with high probability) $\|\widehat{\mu} - \widehat{\mu}'\|_A$ is small, unless there are too many outlying entries x_j. We do this in Lemma 26 by showing that for appropriate subgroup sizes b (recall the random partition $\mathcal{S}$ of $[n]$ into blocks of size n/b in `MEANSAFE`), the `MEANSAFE` algorithm correctly identifies all outliers without pruning many inlying datapoints. In the second step, we finalize the proof (section D.1.1) by combining the sensitivity bound with more or less standard distributional stability guarantees for Gaussian distributions, which we list in the preliminary section 2.

We begin by formalizing two properties that will be helpful to proving the sensitivity bound in Lemma 26. We recall the notation t (respectively t') for denoting the number of pruned groups in Lines 6–10 of `MEANSAFE` on inputs x and x', while R and R' denote the sets of all pruned indices. Of the next two lemmas, Lemma 24 bounds differences between R and R' and t and t', while Lemma 25 is a generic lemma that bounds the difference of empirical means with nested index sets. These two lemmas are combined in Lemma 26 to bound the difference between the estimated mean $\widehat{\mu} = \frac{1}{n-|R|}\sum_{j \notin R} x_j$ and $\widehat{\mu}'$.

Before stating Lemma 24, recall for sets S, S' that $d_{\text{sym}}(S, S') = \max\{|S \setminus S'|, |S' \setminus S|\}$.

Lemma 24 (Stability of rejected indices) *Let t, t' and R, R' be as above. Then $|t - t'| \leq 1$ and $d_{\text{sym}}(R, R') \leq b$.*

Proof Let the set $J := \{j \mid \widetilde{D}_j \neq \perp, \widetilde{D}_j \geq \log(\sqrt{B}/4)\}$ index the subgroups pruned by the execution of MEANSAFE on the sample x', and similarly define J' relative to $\widetilde{D}'$ for the sample x'. Then $t = |J|$ and $R = \cup_{j \in J} S_j$, and also $t' = |J'|$ and $R' = \cup_{j \in J'} S'_j$. We show $d_{\text{sym}}(J, J') \leq 1$, from which the claim $|t - t'| \leq 1$ follows immediately and the claim $d_{\text{sym}}(R, R') \leq b$ follows from the fact that $|S_j| = b$ for all $j \in [n/b]$.

Recalling x and x' differ only at index i, suppose that $i \in S_\ell$ for $\ell \in [n/b]$. Then $x_{S_j} = x'_{S_j}$ for all $j \neq \ell$; in particular, $\text{diam}_A(x_{S_j}) = \text{diam}_A(x'_{S_j})$ and so $D_j = D'_j$ for $j \neq \ell$. Thus, the indices of the k largest elements of $D + Z$ and $D' + Z$, i.e., those subgroups identified by TOPk as having the largest diameters, which we denote by $K = \{j \mid \widetilde{D}_j \neq \perp\}$ and $K' = \{j \mid \widetilde{D}'_j \neq \perp\}$ respectively, differ by at most one index: $d_{\text{sym}}(K, K') \leq 1$ with equality obtaining only if ℓ is in exactly one of K or K'. If ℓ is in neither K nor K', then $J = J'$ and the claim $d_{\text{sym}}(J, J') \leq 1$ follows. Otherwise, supposing $\ell \in K$, the bound $d_{\text{sym}}(K, K') \leq 1$ implies $K \setminus \{\ell\} \subset K'$ and thus $\widetilde{D}_{K \setminus \{\ell\}} = \widetilde{D}'_{K \setminus \{\ell\}}$, or vice versa if $\ell \in K'$; $d_{\text{sym}}(J, J') \leq 1$ then follows from $J \subset K$ and $J' \subset K'$. ■

Lemma 25 *Let $\{y_1, \ldots, y_n\}$ be an arbitrary collection of vectors and $S \subset S' \subset [n]$. Define $\mu_S := \frac{1}{|S|} \sum_{i \in S} y_i$ and $\mu_{S'} := \frac{1}{|S'|} \sum_{i \in S'} y_i$. Then*

$$\|\mu_S - \mu_{S'}\| \leq \frac{|S' \setminus S| \, \text{diam}_{\|\cdot\|}(y_{S'})}{|S'|}.$$

Proof Observe

$$\mu_S - \mu_{S'} = \mu_S - \left(\frac{|S|}{|S'|} \mu_S + \frac{1}{|S'|} \sum_{i \in S' \setminus S} y_i \right) = \frac{1}{|S'|} \sum_{i \in S' \setminus S} (\mu_S - y_i),$$

where from the assumption that $S \subset S'$ we have

$$\max_{i \in S' \setminus S} \|y_i - \mu_S\| \leq \max_{j \in S, i \in S'} \|y_i - y_j\| \leq \text{diam}_{\|\cdot\|}(S').$$

The claim then follows as $\|\mu_S - \mu_{S'}\| \leq \frac{1}{|S'|} \sum_{i \in S' \setminus S} \text{diam}_{\|\cdot\|}(y_{S'}) = \frac{|S' \setminus S| \text{diam}_{\|\cdot\|}(x_{S'})}{|S'|}$. ■

We now turn to the first step we outline, providing an explicit bound on $\|\widehat{\mu} - \widehat{\mu}'\|_A$ except on the event that $\max\{t, t'\} = k$. Recall the definition $\Delta = \frac{5b\sqrt{B}}{2n} \exp(3\sigma_{\text{top}} \log \frac{2n}{b\gamma})$ in the statement of Lemma 15.

Lemma 26 *With probability at least $1 - \gamma - n^2 2^{1-b}$, $\max\{t, t'\} = k$ or $\|\widehat{\mu} - \widehat{\mu}'\|_A \leq \Delta$.*

Proof We first show that with probability at least $1 - n^2 2^{-b}$ over the random partition $\mathcal{S} \sim \mathsf{Uni}(\mathcal{P}_{n,b})$, $\mathcal{S} = (S_1, \ldots, S_{n/b})$,

$$\text{diam}_A(x_{R^c}) \leq \frac{1}{2} \exp(2 \|Z\|_\infty + \|Z'\|_\infty) \sqrt{B}, \tag{16}$$

with the same bound holding for x' by symmetry. To this end, observe that for the index set

$$J := \{j \in [n/b] \mid \widetilde{D}_j \neq \perp, \widetilde{D}_j \geq \log(\sqrt{B}/4)\},$$

MEANSAFE constructs the removed indices R in Lines 6–10 via the union $R = \cup_{j \in J} S_j$. The first step in the bound (16) is to bound the diameter of the set x_{R^c} by the diameters of the constituent sets within R, which the following generic lemma allows (see Section D.1.2 for a proof).

Claim 1 *Let $\{y_1, \ldots, y_n\}$ be an arbitrary collection of vectors and $\mathcal{S} \sim \mathsf{Uni}(\mathcal{P}_{n,b})$. With probability at least $1 - n^2 2^{-b}$, for all index sets $J \subset [n/b]$, the set $S_J := \cup_{j \in J} S_j$ satisfies* $\mathrm{diam}_{\|\cdot\|}(y_{S_J}) \leq 2 \max_{j \in J} \mathrm{diam}_{\|\cdot\|}(y_{S_j})$.

In light of Claim 1, inequality (16) follows by showing

$$\mathrm{diam}_A(x_{S_j}) \leq \exp(2\|Z\|_\infty + \|Z'\|_\infty)\sqrt{B}/4 \tag{17}$$

for all $j \notin J$ on the event $t < k$. When $t < k$, there exists an index $\ell \in [n/b]$ such that $\widetilde{D}_\ell \neq \perp$ and $\widetilde{D}_\ell \leq \log(\sqrt{B}/4)$, i.e., ℓ indexes one of the k largest elements of $D + Z$ but $\ell \notin J$. Thus, for $j \notin J$ such that $\widetilde{D}_j = \perp$, i.e., $\log(\mathrm{diam}_A(x_{S_j})) + Z_j$ is not among the k largest elements of $D + Z$ (by the construction in TOPk), we have

$$\log(\mathrm{diam}_A(x_{S_j})) \leq \log(\mathrm{diam}_A(x_{S_\ell})) + 2\|Z\|_\infty.$$

Meanwhile, for all $j \notin J$ such that $\widetilde{D}_j \neq \perp$, including $j = \ell$, from the definition of J we immediately have

$$\log(\mathrm{diam}_A(x_{S_j})) + Z'_j \leq \log(\sqrt{B}/4).$$

The claim (17), and hence claim (16), thus follows from the preceding two displays. Moreover, via a union bound over the two executions of MEANSAFE, Claim 1 gives

$$\max\{\mathrm{diam}_A(x_{R^c}), \mathrm{diam}_A(x'_{R'^c})\} \leq \exp(2\|Z\|_\infty + \|Z'\|_\infty)\frac{\sqrt{B}}{2} \quad \text{or} \quad \max\{t, t'\} = k \tag{18}$$

with probability at least $1 - n^2 2^{1-b}$.

We can now bound $\|\widehat{\mu} - \widehat{\mu}'\|_A$ for $\widehat{\mu} = \frac{1}{n-|R|}\sum_{j \notin R} x_j$ and $\widehat{\mu} = \frac{1}{n-|R'|}\sum_{j \notin R'} x'_j$ via the following claim (essentially, a number of applications of the triangle inequality), whose proof we also defer (see Section D.1.3).

Claim 2 $\|\widehat{\mu} - \widehat{\mu}'\|_A \leq \frac{4(b+1)}{n} \max\{\mathrm{diam}_A(x_{R^c}), \mathrm{diam}_A(x'_{R'^c})\}$.

Using Claim 2, the main Lemma 26 follows relatively quickly. By combining the display (18) with the fact that, by elementary calculation,

$$\mathbb{P}(\max\{\|Z\|_\infty, \|Z'\|_\infty\} > \sigma_{\mathrm{top}} \log(2n/b\gamma)) \leq \gamma,$$

we obtain that with probability at least $1 - \gamma - n^2 2^{1-b}$, $\max\{t, t'\} = k$ or

$$\|\widehat{\mu}' - \widehat{\mu}'\|_A \leq \frac{2(b+1)\sqrt{B}}{n} \exp\left(2\|Z\|_\infty + \|Z'\|_\infty\right) \leq \frac{2(b+1)\sqrt{B}}{n} \exp\left(3\sigma_{\mathrm{top}} \log\frac{2n}{b\gamma}\right).$$

Recalling the assumption that the batchsize $b \geq 4$ (so $2(b+1) \leq \frac{5}{2}b$), we obtain the lemma. ■

D.1.1. Finalizing proof of Lemma 15

We prove for any (measurable) event $O \subset \mathbb{R}^d \cup \{\perp\}$ that

$$\mathbb{P}(\widetilde{\mu} \in O) \le e^{\alpha + 1/\sigma_{W_{\text{mean}}}} \mathbb{P}(\widetilde{\mu}' \in O) + \beta_1 + \beta_2, \tag{19}$$

where $\alpha > 0$ and $\beta_1 \in (0,1)$ determine the Gaussian noise scale for $Z^{\mathsf{N}} \sim \mathsf{N}(0, \sigma_{\mathsf{N}}^2 I)$ via

$$\sigma_{\mathsf{N}} = \begin{cases} \frac{\Delta}{\alpha}\sqrt{1.25 \log \frac{1}{\beta_1}} & \text{if } \alpha \le 1 \\ \frac{\Delta}{\sqrt{2\log\frac{1}{\beta_1} + 2\alpha} - \sqrt{2\log\frac{1}{\beta_1}}} & \text{otherwise,} \end{cases} \quad \text{and} \quad \beta_2 = \frac{1}{2} e^{-(k/3-1)/\sigma_{W_{\text{mean}}}} + \gamma + n^2 2^{1-b}.$$

The other direction follows by symmetry. We treat $O \subset \mathbb{R}^d$ and $O = \perp$ separately, merging the two cases at the end to show the claim (19). Supposing first $O \subset \mathbb{R}^d$, the following observation delineates necessary and sufficient conditions for $\widetilde{\mu} \in O$.

Observation D.1 *Let $O \subset \mathbb{R}^d$. Then $\widetilde{\mu} \in O$ if and only if $t \le 2k/3 + W$ and $\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O$.*

Proof From the condition for returning $\perp$ in Line 13 of `MEANSAFE`, we immediately have $\widetilde{\mu} = \perp \notin \mathbb{R}^d$ if and only if $t > 2k/3 + W$; thus, the condition $t \le 2k/3 + W$ is necessary and sufficient for $\widetilde{\mu} \in \mathbb{R}^d$. As either $\widetilde{\mu} = \perp$ or $\widetilde{\mu} = \widehat{\mu} + A^{1/2} Z^{\mathsf{N}}$ by definition, it then follows trivially that $t \le 2k/3 + W$ and $\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O$ together suffice to obtain $\widetilde{\mu} \in O$. ■

Marginalizing over the number of sets of rejected indices t and $\widehat{\mu}$ we have the following sequence of inequalities:

$$\begin{aligned}
&\mathbb{P}(\widetilde{\mu} \in O) \\
&= \mathbb{E}\left[\mathbb{P}(\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O \mid \widehat{\mu}) \mathbb{P}(t \le 2k/3 + W \mid t)\right] \\
&\overset{(i)}{\le} \mathbb{E}\left[\mathbb{P}(\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O \mid \widehat{\mu}) \mathbb{P}(t \le 2k/3 + W \mid t) 1\left\{\left\|\widehat{\mu}' - \widehat{\mu}'\right\|_A \le \Delta\right\}\right] \\
&\qquad + \mathbb{E}\left[\mathbb{P}(t \le 2k/3 + W \mid t) 1\left\{\max\{t, t'\} = k\right\}\right] + \gamma + n^2 2^{1-b} \\
&\overset{(ii)}{\le} \mathbb{E}\left[\mathbb{P}(\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O \mid \widehat{\mu}) \mathbb{P}(t \le 2k/3 + W \mid t) 1\left\{\left\|\widehat{\mu}' - \widehat{\mu}'\right\|_A \le \Delta\right\}\right] \\
&\qquad + \mathbb{P}(W \ge k/3 - 1) + \gamma + n^2 2^{1-b} \\
&= \mathbb{E}\left[\mathbb{P}(\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O \mid \widehat{\mu}) \mathbb{P}(t \le 2k/3 + W \mid t) 1\left\{\left\|\widehat{\mu}' - \widehat{\mu}'\right\|_A \le \Delta\right\}\right] + \beta_2
\end{aligned} \tag{20}$$

Here, step (i) follows because $\|\widehat{\mu}' - \widehat{\mu}'\|_A \le \Delta$ or $\max\{t, t'\} = k$ occurs with probability at least $1 - \gamma - n^2 2^{1-b}$ by Lemma 26; step (ii) because $|t - t'| \le 1$ by Lemma 24 and so $\max\{t, t'\} = k$ implies $t \ge k-1$; the final equality follows from the identity $\mathbb{P}\left(W \ge k/3 - 1\right) = \frac{1}{2} e^{-(k/3-1)/\sigma_{W_{\text{mean}}}}$ and definition of β_2.

Continuing, we can bound the last expectation in the preceding display by

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}(\widehat{\mu} + A^{1/2} Z^{\mathsf{N}} \in O \mid \widehat{\mu}) \mathbb{P}(t \le 2k/3 + W \mid t) 1\left\{\left\|\widehat{\mu}' - \widehat{\mu}'\right\|_A \le \Delta\right\}\right] \\
&\overset{(i)}{\le} \exp(\alpha) \mathbb{E}\left[\mathbb{P}(\widehat{\mu}' + A^{1/2} Z^{\mathsf{N}} \in O \mid \widehat{\mu}') \mathbb{P}(t \le 2k/3 + W \mid t)\right] + \beta_1
\end{aligned}$$

$$\overset{(ii)}{\leq} \exp(\alpha + 1/\sigma_{W_{\mathrm{mean}}})\mathbb{E}\left[\mathbb{P}(\widehat{\mu}' + A^{1/2}Z^{\mathsf{N}} \in O \mid \widehat{\mu}')\mathbb{P}(t' \leq 2k/3 + W \mid t')\right] + \beta_1$$
$$= \exp(\alpha + 1/\sigma_{W_{\mathrm{mean}}})\mathbb{P}(\widetilde{\mu}' \in O) + \beta_1, \tag{21}$$

with step (i) following from the privacy of the Gaussian mechanism with noise σ_N and sensitivity bound $\|\widehat{\mu} - \widehat{\mu}'\| \leq \Delta$ (Lemma 10); step (ii) from $|t - t'| \leq 1$ by Lemma 24 and that $W \sim \mathsf{Lap}(\sigma_{W_{\mathrm{mean}}})$; and the final equality follows directly from Observation D.1, applied here to the execution of `MEANSAFE` on data x'. Combining inequalities (20) and (21) yields the claim (19) when $O \subset \mathbb{R}^d$.

For the case that $O = \{\perp\}$, we have

$$\mathbb{P}(\widetilde{\mu} = \perp) = \mathbb{E}\left[\mathbb{P}(t > 2k/3 + W \mid t)\right]$$
$$\leq e^{\frac{1}{\sigma_{W_{\mathrm{mean}}}}}\mathbb{E}\left[\mathbb{P}(t' > 2k/3 + W \mid t')\right] = \exp^{\frac{1}{\sigma_{W_{\mathrm{mean}}}}}\mathbb{P}(\widetilde{\mu} = \perp).$$

Here, the two equalities follow from the condition for returning $\perp$ in Line 13 of `MEANSAFE`, while the inequality follows because $|t - t'| \leq 1$ by Lemma 24 and that $W \sim \mathsf{Lap}(\sigma_{W_{\mathrm{mean}}})$. The claim (19) for arbitrary O is immediate.

D.1.2. Proof of Claim 1

Consider the event $\mathcal{E}$ that for all indices $i_1, i_2 \in [n]$, with $i_1 \in S_{j_1}$ and $i_2 \in S_{j_2}$, we have $\|y_{i_1} - y_{i_2}\| \leq 2\max\{\mathrm{diam}(y_{S_{j_1}}), \mathrm{diam}(y_{S_{j_2}})\}$. The claim holds on $\mathcal{E}$: for any $J \subset [n/b]$ and $S_J := \cup_{j \in J} S_j$, there exist $j_1, j_2 \in J$ with $i_1 \in S_{j_1}$ and $i_2 \in S_{j_2}$ attaining $\mathrm{diam}(y_{S_J}) = \|y_{i_1} - y_{i_2}\|$, and so

$$\|y_{i_1} - y_{i_2}\| \leq 2\max\{\mathrm{diam}(y_{S_{j_1}}), \mathrm{diam}(y_{S_{j_2}})\} \leq 2\max_{j \in J}\mathrm{diam}(y_{S_j}).$$

It remains to show that $\mathcal{E}$ occurs with probability at least $1 - n^2 2^{-b}$. As there are $\binom{n}{2} \leq \frac{1}{2}n^2$ unordered pairs of distinct indices $i_1, i_2 \in [n]$, the result obtains from a union bound if we show that $\|y_{i_1} - y_{i_2}\| > 2\max\{\mathrm{diam}(y_{S_{j_1}}), \mathrm{diam}(y_{S_{j_2}})\}$ occurs with probability at most 2^{1-b}.

Proceeding, let $i_1, i_2 \in [n]$ and $i_1 \in S_{j_1}, i_2 \in S_{j_2}$ and let $c = \frac{1}{2}\|y_{i_1} - y_{i_2}\|$. If $i_1 = i_2$ or $j_1 = j_2$, there is nothing to show, so assume $i_1 \neq i_2$ and $j_1 \neq j_2$. Let $C_1 = \{i \in [n] \setminus \{i_1, i_2\} \mid \|y_{i_1} - y_i\| < c\}$ and $C_2 = \{i \in [n] \setminus \{i_1, i_2\} \mid \|y_{i_2} - y_i\| < c\}$ be those indices i for which y_i is close to y_{i_1} or y_{i_2}, respectively. By the triangle inequality, C_1 is disjoint from C_2, and so without loss of generality, we suppose that $|C_1| \leq (n-2)/2$.

We wish to show that $\mathrm{diam}(y_{S_{j_1}}) \geq c$, for which it is sufficient that there exists an index in $S_{j_1} \setminus \{i_1\}$ not in C_1. So by showing $S_{j_1} \setminus \{i_1\} \subset C_1$ occurs with probability at most 2^{1-b}, we will be done. As $\mathcal{S} \sim \mathsf{Uni}(\mathcal{P}_{n,b})$, the set $S_{j_1} \setminus \{i_1\}$ is a uniformly distributed subset of $[n] \setminus \{i_1, i_2\}$ of size $b-1$. Consequently, there are $\binom{n-2}{b-1}$ distinct values it can take and $\binom{|C_1|}{b-1}$ values such that $S_{j_1} \setminus \{i_1\} \subset C_1$. Therefore, the probability that $S_{j_1} \setminus \{i_1\} \subset C_1$ is

$$\mathbb{P}(S_{j_1} \setminus \{i_1\} \subset C_1) = \frac{\binom{|C_1|}{b-1}}{\binom{n-2}{b-1}} = \prod_{i=0}^{b-2}\frac{(|C_1| - i)_+}{n - 2 - i} \leq \left(\frac{|C_1|}{n-2}\right)^{b-1} \leq 2^{1-b},$$

where the last inequality follows because $|C_1| \leq (n-2)/2$.

D.1.3. Proof of Claim 2

Recall that

$$\widehat{\mu} = \frac{1}{n - |R|} \sum_{j \notin R} x_j \quad \text{and} \quad \widehat{\mu}' = \frac{1}{n - |R'|} \sum_{j \notin R'} x'_j,$$

and define

$$R_{\mathrm{all}} := R \cup R', \quad \widehat{\mu}_{\mathrm{all}} := \frac{1}{n - |R_{\mathrm{all}}|} \sum_{j \notin R_{\mathrm{all}}} x_j, \quad \widehat{\mu}'_{\mathrm{all}} := \frac{1}{n - |R_{\mathrm{all}}|} \sum_{j \notin R_{\mathrm{all}}} x'_j.$$

Lemma 24 gives $|R^c \setminus R^c_{\mathrm{all}}| = |R_{\mathrm{all}} \setminus R| \le b$, and by assumption on the batchsize b and rejection threshold k we also have $|R_{\mathrm{all}}| \le b + |R| \le b + kb \le \frac{n}{2}$.

Applying Lemma 25 with $S = R^c_{\mathrm{all}}$, and $S' = R^c$, we get

$$\|\widehat{\mu} - \widehat{\mu}_{\mathrm{all}}\|_A \le \frac{|R^c \setminus R^c_{\mathrm{all}}| \operatorname{diam}_A(x_{R^c})}{|R^c|} \le \frac{2b \operatorname{diam}_A(x_{R^c})}{n}$$

as $|R| \le |R_{\mathrm{all}}| \le \frac{n}{2}$. Applying Lemma 25 again, this time with dataset x', $S = R'^c_{\mathrm{all}}$ and $S' = R'^c$, we get $\|\widehat{\mu}' - \widehat{\mu}'_{\mathrm{all}}\|_A \le \frac{2b}{n} \operatorname{diam}_A(x'_{R'^c})$.

Now we bound $\widehat{\mu}_{\mathrm{all}} - \widehat{\mu}'_{\mathrm{all}}$, where recalling that index i is the sole (potentially) differing index in x, x' (that is, $x_{-i} = x'_{-i}$), we can write as

$$\widehat{\mu}_{\mathrm{all}} - \widehat{\mu}'_{\mathrm{all}} = \frac{1}{n - |R_{\mathrm{all}}|} \sum_{j \notin R_{\mathrm{all}}} (x_j - x'_j) = \frac{\mathbf{1}\{i \notin R_{\mathrm{all}}\}}{n - |R_{\mathrm{all}}|} (x_i - x'_i).$$

If $i \in R_{\mathrm{all}}$, this difference is 0. Otherwise, $i \notin R$ and $i \notin R'$. As $|R_{\mathrm{all}}| \le \frac{n}{2}$, we may pick some $j' \notin R_{\mathrm{all}} \cup \{i\}$. Because $x_{j'} = x'_{j'}$, we have both $\left\|x_i - x_{j'}\right\|_A \le \operatorname{diam}_A(x_{R^c})$ and both $\left\|x'_i - x_{j'}\right\|_A \le \operatorname{diam}_A(x'_{R'^c})$. The triangle inequality then gives $\|x_i - x'_i\|_A \le 2 \max\{\operatorname{diam}_A(x_{R^c}), \operatorname{diam}_A(x'_{R'^c})\}$, and so $\|\widehat{\mu}_{\mathrm{all}} - \widehat{\mu}'_{\mathrm{all}}\|_A \le \frac{4}{n} \max\{\operatorname{diam}_A(x_{R^c}), \operatorname{diam}_A(x'_{R'^c})\}$. Combining the above, the claim follows immediately from

$$\begin{aligned}
\left\|\widehat{\mu} - \widehat{\mu}'\right\|_A &\le \|\widehat{\mu} - \widehat{\mu}_{\mathrm{all}}\|_A + \left\|\widehat{\mu}_{\mathrm{all}} - \widehat{\mu}'_{\mathrm{all}}\right\|_A + \left\|\widehat{\mu}'_{\mathrm{all}} - \widehat{\mu}'\right\|_A \\
&\le \frac{2b \operatorname{diam}_A(x_{R^c})}{n} + \frac{4 \max\{\operatorname{diam}_A(x_{R^c}), \operatorname{diam}_A(x'_{R'^c})\}}{n} + \frac{2b \operatorname{diam}_A(x'_{R'^c})}{n}.
\end{aligned}$$

D.2. Proof of Lemma 16

Unpacking the execution transcripts $\Gamma(x, A)$ and $\Gamma(x, A')$ from (3) as

$$(D, \widetilde{D}, R, t, \widehat{\mu}) := \Gamma(x, A) \quad \text{and} \quad (D', \widetilde{D}', R', t', \widehat{\mu}') := \Gamma(x, A'),$$

observe that given the pair $(\widetilde{D}, A^{1/2} Z^{\mathsf{N}})$, $\widetilde{\mu}(x, A)$ is independent of A (see the execution of MEANSAFE), and analogously, $\widetilde{\mu}(x, A')$ is independent of A' given $(\widetilde{D}', A'^{1/2} Z^{\mathsf{N}})$. Therefore, by showing $A^{1/2} Z^{\mathsf{N}} \stackrel{d}{=}_{\alpha_1, \beta} A'^{1/2} Z^{\mathsf{N}}$ and $\widetilde{D} \stackrel{d}{=}_{\alpha_2, 0} \widetilde{D}'$, basic composition (Lemma 2) and the post-processing property (Lemma 4) will imply the claimed result that $\widetilde{\mu} \stackrel{d}{=}_{\alpha_1 + \alpha_2, \beta} \widetilde{\mu}'$.

Recalling $Z^{\mathsf{N}} \sim \mathsf{N}(0, \sigma_{\mathsf{N}}^2 I)$, we have $A^{1/2} Z^{\mathsf{N}} \sim \mathsf{N}(0, \sigma_{\mathsf{N}}^2 A)$ and $A'^{1/2} Z^{\mathsf{N}} \sim \mathsf{N}(0, \sigma_{\mathsf{N}}^2 A')$, and so $A^{1/2} Z^{\mathsf{N}} \stackrel{d}{=}_{\alpha_1, \beta} A'^{1/2} Z^{\mathsf{N}}$ follows immediately from the assumption $d_{\mathrm{psd}}(A, A') \le \frac{a}{n}$ and the closeness of Gaussian distributions with differing covariances (Lemma 11).

To show $\widetilde{D} \overset{d}{=}_{\alpha_2,0} \widetilde{D}'$, we make the following observation to bound the sensitivity of the log-Mahalanobis norm for A and A'.

Observation D.2 *Suppose $A, A' \in \mathbb{R}^{d\times d}$ and $d_{\text{psd}}(A, A') \le \gamma < \infty$. Then for any $v \in \mathsf{Col}(A)$, $|\log \|v\|_A - \log \|v\|_{A'}| \le \gamma/2$. For any $v \notin \mathsf{Col}(A)$, $\log \|v\|_A = \log \|v\|_{A'} = \infty$.*

Proof Observe $d_{\text{psd}}(A, A') < \infty$ trivially implies A and A' are PSD and their columnspaces coincide, from which the second claim immediately follows. For $v \in \mathsf{Col}(A)$, we only show $\log \|v\|_{A'} \le \frac{1}{2}\gamma + \log \|v\|_A$, as the reverse inequality holds by symmetry. By assumption,

$$\left\| A'^{\dagger/2}(A - A')A'^{\dagger/2} \right\|_{\text{op}} \le d_{\text{psd}}(A, A') \le \gamma$$

and hence $A'^{\dagger/2}(A - A')A'^{\dagger/2} \preceq \gamma I$. Conjugating by $A'^{1/2}$ and rearranging terms, we have $\Pi_{A'} A \Pi_{A'} \preceq (1+\gamma)A'$. Because $\Pi_{A'} = \Pi_A$, we have $\Pi_{A'} A \Pi_{A'} = A$, which yields $A \preceq (1+\gamma)A'$, or equivalently $A'^{\dagger} \preceq (1+\gamma)A^{\dagger}$. Therefore $\|v\|_{A'}^2 \le (1+\gamma)\|v\|_A^2$. Taking square roots and logarithms on both sides proves the claim as $\log(\sqrt{1+\gamma}) \le \frac{\gamma}{2}$. ∎

This observation, coupled with our construction that both $\widetilde{\mu}(x, A)$ and $\widetilde{\mu}(x, A')$ use the same (random) partition $\mathcal{S} = (S_1, \ldots, S_{n/b})$, implies $|D_j - D'_j| \le \frac{a}{2n}$ for all $j \in [n/b]$; hence $\|D - D'\|_\infty \le \frac{a}{2n}$ (the indices where the entries are infinite coincide). The closeness properties of TOPk (Lemma 9) and our choice $\sigma_{\text{top}} = \frac{ka}{n\alpha_2}$ then give $\widetilde{D} \overset{d}{=}_{\alpha_2,0} \widetilde{D}'$.

Appendix E. Proof of Theorem 7

We first show under the event $\mathcal{E}_{\text{samp}}$ that with probability at least $1-4\delta$ over the randomness in PRIVMEAN, both COVSAFE and MEANSAFE prune no observations, meaning the sets of removed indices $R = \emptyset$ in both procedures (so that Line 7 in COVSAFE and Line 7 in MEANSAFE never fail), and thus $\widetilde{\mu} = \overline{X}_n + \overline{\Sigma}_n^{1/2} Z^{\mathsf{N}}$. Let Π_Σ denote the projection matrix onto the columnspace of Σ. As $\overline{\Sigma}_n \preceq \frac{3}{2}\Sigma$ on $\mathcal{E}_{\text{samp}}$, we have $\|\overline{\Sigma}_n^{1/2} Z^{\mathsf{N}}\|_\Sigma^2 \le \frac{3}{2}\|\Pi_\Sigma Z^{\mathsf{N}}\|_2^2$. The result then follows follows once we show that $\|\Pi_\Sigma Z^{\mathsf{N}}\|_2 \le \frac{c\sqrt{B \max\{\text{rank}(\Sigma), \log\frac{1}{\delta}\}\log\frac{1}{\delta}}}{n\varepsilon}$ with probability at least $1-\delta$ and take a union bound over these events and $\mathcal{E}_{\text{samp}}$.

Rearranging the condition in Line 7 of COVSAFE, the element $X_i - X_{n/2+i}$ is pruned in the first iteration only if

$$\begin{aligned}(Z_{\text{cov}})_i + (Z_{\text{cov}})_{n/2+1} &> \log(B) - \log(\left\|X_i - X_{n/2+i}\right\|_{\overline{\Sigma}_n}^2) \\ &\overset{(\star)}{\ge} \log(c_1 B/8M^2) \ge \log(c_1/8),\end{aligned}$$

where $(\star)$ holds for all $i \in [n/2]$ on event $\mathcal{E}_{\text{samp}}$ because

$$\left\|X_i - X_{n/2+i}\right\|_{\overline{\Sigma}_n}^2 \le 2\left\|X_i - X_{n/2+i}\right\|_\Sigma^2 \le 4\|X_i - \mu\|_\Sigma^2 + 4\left\|X_{n/2+i} - \mu\right\|_\Sigma^2 \le 8M^2/c_1. \quad (22)$$

As $c_1 \ge 64e$ by Assumption A1, if $\|Z_{\text{cov}}\|_\infty \le 1/2$ then COVSAFE in line 7 prunes no entries, instead simply passing $\overline{\Sigma}_n$ to MEANSAFE so long as $W_{\text{cov}} + m > 0$ (see line 18). Recall that

$(Z_{\text{cov}})_j \stackrel{\text{iid}}{\sim} \mathsf{Lap}(\sigma_Z)$ for $j = 1, \ldots, n/2+1$ and $\sigma_Z = \frac{32\sqrt{e}B(m_{\max}+1)}{n\varepsilon} \le \frac{cB\log\frac{1}{\delta}}{n\varepsilon^2}$, so by taking a union bound over the entries, we have with probability at least $1-\delta$ that

$$\|Z_{\text{cov}}\|_\infty \le \frac{cB\log\frac{1}{\delta}}{n\varepsilon^2}\log\left(\frac{n/2+1}{\delta}\right) \le \frac{1}{2},$$

where the last inequality is by the assumptions that $n \ge \frac{cB\log^2(\frac{1}{\delta})}{\varepsilon^2}$ and $\delta \le \frac{1}{n}$. Also recall that $W_{\text{cov}} \sim \mathsf{Lap}(\frac{16}{\varepsilon})$ and $m = \frac{16}{\varepsilon}\log\frac{1}{\delta}$, so $W_{\text{cov}} + m > 0$ with probability at least $1 - \frac{\delta}{2}$.

Continuing to the next phase of PRIVMEAN, MEANSAFE with input $A = \overline{\Sigma}_n$ prunes the indices S_j only if

$$\widetilde{D}_j = D_j + (Z'_{\text{top}})_j > \log(\sqrt{B}/4).$$

By the same argument we used to obtain inequality (22), on $\mathcal{E}_{\text{samp}}$ we have for all $j \in [n/b]$ that

$$D_j = \log(\text{diam}_{\overline{\Sigma}_n}(X_{S_j})) \le \log(\sqrt{8M^2/c_1}),$$

and so if $\left\|Z'_{\text{top}}\right\|_\infty \le 1/2$ then

$$\widetilde{D}_j \le \log(\sqrt{8M^2/c_1}) + \frac{1}{2} \le \log(\sqrt{B}/4),$$

where the last inequality follows from $c_1 \ge 64e$ and $B \ge M^2$. Thus, MEANSAFE prunes no entries, and $\widetilde{\mu} = \overline{X}_n + \overline{\Sigma}_n^{1/2} Z^{\mathsf{N}}$ so long as $W_{\text{mean}} + \frac{2k}{3} > 0$ (see Line 12). Recall that $(Z'_{\text{top}})_j \stackrel{\text{iid}}{\sim} \mathsf{Lap}(\sigma_{\text{top}})$ for $j = 1, \ldots, n/b$ and $\sigma_{\text{top}} = \frac{8k}{n\varepsilon}\frac{B\sqrt{e}}{1-B\sqrt{e}/n} \le \frac{cB\log(\frac{1}{\delta})}{n\varepsilon^2}$. Another union bound gives that with probability at least $1-\delta$,

$$\|Z_{\text{cov}}\|_\infty \le \frac{cB\log(\frac{1}{\delta})}{n\varepsilon^2}\log\frac{n}{b\delta} \le \frac{1}{2},$$

where the last inequality follows from the assumption $n \ge \frac{cB\log^2\frac{1}{\delta}}{\varepsilon^2}$ and $\delta \le \frac{1}{n}$. Also, $W_{\text{mean}} \sim \mathsf{Lap}(\frac{8}{\varepsilon})$ and $\frac{2k}{3} = \frac{16}{\varepsilon}\log\frac{3}{\delta} - 2$, so $W_{\text{mean}} - k > \frac{8}{\varepsilon}\log\frac{3}{\delta} - 2 \ge 0$ with probability at least $1-\delta$.

Therefore, PRIVMEAN returns $\overline{X}_n + \overline{\Sigma}_n^{1/2} Z^{\mathsf{N}}$ with probability at least $1-4\delta$ on the event $\mathcal{E}_{\text{samp}}$. Recall that $Z^{\mathsf{N}} \sim \mathsf{N}(0, \sigma_{\mathsf{N}}^2 I)$ with $\sigma_{\mathsf{N}} = \frac{20b\sqrt{B}}{n\varepsilon}\exp(3\sigma_{\text{top}}\log\frac{12n}{b\delta})$, and because $\sigma_{\text{top}} \le \frac{cB\log(\frac{1}{\delta})}{n\varepsilon^2}$, $\delta \le \frac{1}{n}$, and $n \ge \frac{cB\log^2(\frac{1}{\delta})}{\varepsilon^2}$, we have that $\sigma_{\mathsf{N}} \le \frac{c\sqrt{B}\log(\frac{1}{\delta})}{n\varepsilon}$. Classical tail bounds on the χ^2-distribution (Laurent and Massart, 2000, Lemma 1) give with probability at least $1-\delta$ that

$$\left\|\Pi_\Sigma Z^{\mathsf{N}}\right\|_2^2 \le \sigma_{\mathsf{N}}^2\left[\text{rank}(\Sigma) + 2\sqrt{\text{rank}(\Sigma)\log\frac{1}{\delta}} + 2\log\frac{1}{\delta}\right] \le \frac{cB\max\{\text{rank}(\Sigma), \log\frac{1}{\delta}\}\log^2\frac{1}{\delta}}{n^2\varepsilon^2},$$

where the last inequality follows from the bound on σ_{N}.

Appendix F. Adapting to heavy-tailed data

F.1. Proof of Theorem 8

Let $t^\star$ be the smallest positive integer such that $M^2 \le 2^{t^\star - 1}$ and let t_{stop} be the iteration when ADAMEAN terminates (which may be infinite). Note that $t^\star \le s$. The proof comes in two parts: on the event $\mathcal{E}_{\text{samp}}$, we first show that either $t_{\text{stop}} > t^\star$ or $\widetilde{\mu}$ satisfies the claim (4) with probability at least $1 - \pi^2\delta/6$; secondly, we show ADAMEAN terminates with $t_{\text{stop}} \le t^\star$ with probability at least $1 - 5\delta$. The result then follows via a union bound.

We carry out the first part with the help of the following lemma.

Lemma 27 *Let $\varepsilon > 0$ and $\delta \le \frac{1}{n}$ be privacy parameters and let $B \ge 0$. Suppose the event $\mathcal{E}_{\text{samp}}$ holds and let $\widetilde{\mu} = \texttt{PRIVMEAN}_{B,(\varepsilon,\delta)}(X_{1:n})$. Then with probability at least $1 - 2\delta$ over the randomness of PRIVMEAN, $\widetilde{\mu} = \perp$ or*

$$\left\|\widetilde{\mu} - \overline{X}_n\right\|_\Sigma \le \frac{c \log\frac{1}{\delta} \max\left\{M \log\frac{1}{\delta}, \sqrt{B \max\{\operatorname{rank}(\Sigma), \log\frac{1}{\delta}\}}\right\}}{n\varepsilon}.$$

Proof Suppose $\widetilde{\mu} \ne \perp$ as otherwise the claim is trivial. Let $\widehat{\Sigma}$ denote the covariance estimate of COVSAFE (that is, Σ_T at the final iteration of COVSAFE), and let $\widehat{\mu}$ denote the empirical mean of the observations not pruned by MEANSAFE so that $\widetilde{\mu} = \widehat{\mu} + \widehat{\Sigma}^{1/2} Z^{\mathsf{N}}$. Then by the condition for returning $\perp$ in Line 13 of MEANSAFE, MEANSAFE prunes at most $b(\frac{2k}{3} + W_{\text{mean}})$ points and so

$$\begin{aligned}
\left\|\widehat{\mu} - \overline{X}_n\right\|_\Sigma &\le \frac{b(\frac{2k}{3} + W_{\text{mean}}) \max_i \|X_i - \widehat{\mu}\|_\Sigma}{n} \\
&\le \frac{b(\frac{2k}{3} + W_{\text{mean}}) \left(\max_i \|X_i - \mu\|_\Sigma + \|\mu - \widehat{\mu}\|_\Sigma\right)}{n} \\
&\le \frac{2b(\frac{2k}{3} + W_{\text{mean}}) \max_i \|X_i - \mu\|_\Sigma}{n} \overset{(\star)}{\le} \frac{2b(\frac{2k}{3} + W_{\text{mean}})M}{\sqrt{c_1} n},
\end{aligned}$$

with $(\star)$ following directly from $\mathcal{E}_{\text{samp}}$. Recalling that $k = \frac{24}{\varepsilon}\log\frac{3}{\delta} - 3$ and $W \sim \mathsf{Lap}(\frac{8}{\varepsilon})$, it follows that $\frac{2k}{3} + W < \frac{24}{\varepsilon}\log\frac{3}{\delta}$ with probability at least $1 - \frac{\delta}{6}$. Recalling also that $b = 1 + \log_2\frac{6n^2}{\delta}$, it follows that on $\mathcal{E}_{\text{samp}}$,

$$\left\|\widehat{\mu} - \overline{X}_n\right\|_\Sigma \le \frac{2(1 + \log_2\frac{6n^2}{\delta})(\frac{24}{\varepsilon}\log\frac{3}{\delta})M}{\sqrt{c_1} n} \le \frac{cM\log^2\frac{1}{\delta}}{n\varepsilon}$$

with probability at least $1 - \frac{\delta}{6}$.

Meanwhile, observe $\mathcal{E}_{\text{samp}}$ implies $\widehat{\Sigma} \preceq \overline{\Sigma}_n \preceq \frac{3}{2}\Sigma$ as pruning entries (line 7) in COVSAFE only shrinks its covariance estimate. Thus, from essentially the same argument as in the proof of Theorem 7, we have

$$\|\widehat{\mu} - \widetilde{\mu}\|_\Sigma = \left\|\widehat{\Sigma}^{1/2} Z^{\mathsf{N}}\right\|_\Sigma \le \frac{c\sqrt{B \max\{\operatorname{rank}(\Sigma), \log\frac{1}{\delta}\}}\log\frac{1}{\delta}}{n\varepsilon}$$

with probability at least $1 - \delta$.

The preceding two displays together imply Lemma 27 after taking a union bound. ■

Applying Lemma 27 with the mappings $B \leftarrow 2^{t-1}$, $\varepsilon \leftarrow \varepsilon/t^2$ and $\delta \leftarrow \delta/t^2$, we have for any $1 \le t \le t^\star$ that on the event $\mathcal{E}_{\text{samp}}$, with probability at least $1 - 2\delta/t^2$, either $\widetilde{\mu}_t = \perp$ or

$$\left\|\widetilde{\mu}_t - \overline{X}_n\right\|_\Sigma \le \frac{ct^2 \log(t^2/\delta) \max\left\{M \log(t^2/\delta), 2^{(t-1)/2}\sqrt{\max\{\operatorname{rank}(\Sigma), \log \frac{t^2}{\delta}\}}\right\}}{n\varepsilon},$$

where the latter case $\widetilde{\mu}_t$ satisfies Eq. (4) as $t \le t^\star \le s$. Then via a union bound this same event holds simultaneously for all $1 \le t \le t^\star$ with probability at least $1 - \pi^2\delta/3$, and thus either $t_{\text{stop}} > t^\star$ or `ADAMEAN` terminates and $\widetilde{\mu}$ satisfies the claim (4).

Proceeding to the second part of the proof, recall we have $2^{t^\star - 1} \ge M^2$ and so applying Theorem 7 with $B = 2^{t^\star-1}$, it follows under $\mathcal{E}_{\text{samp}}$ that $\widetilde{\mu}_{t^\star} \neq \perp$, and thus `ADAMEAN` terminates after $t_{\text{stop}} \le t^\star$ iterations, with probability at least $1 - 5\delta/(t^\star)^2 \ge 1 - 5\delta$. The claim (4) follows.

F.2. Example calculations for bounding maximal deviations from the mean

We give detailed examples, assuming throughout for simplicity Σ is full rank. In each example, we let $Z_i = \Sigma^{-1/2}(X_i - \mu)$ be the whitened data, defining the sample covariance $\overline{\Sigma}_Z = \frac{1}{n}\sum_{i=1}^{n/2}(Z_i - Z_{n/2+i})(Z_i - Z_{n/2+i})^T$. As $\|X_i - \mu\|_\Sigma = \|Z_i\|_2$ and $\overline{\Sigma}_n = \Sigma^{1/2}\overline{\Sigma}_Z\Sigma^{1/2}$, we have the equivalence

$$\mathcal{E}_{\text{samp}} = \left\{\max_{i\in[n]} \|Z_i\|_2^2 \le M^2/c_1 \quad \text{and} \quad \left\|\overline{\Sigma}_Z - I\right\|_{\text{op}} \le \frac{1}{2}\right\}.$$

Example 1 (Sub-Gaussian random vectors) *If for all v satisfying $\|v\|_2 \le 1$ the scalar random variable $\langle Z, v\rangle$ is τ^2-sub-Gaussian,*

$$M^2 \le O(1)\tau^2\left[d + \log\frac{n}{\beta}\right].$$

Indeed, a standard covering argument (see, e.g., Wainwright (2019, Ch. 5) or Vershynin (2019, Ch. 4)) gives that for all $t \ge 0$, $\mathbb{P}(\|Z\|_2 \ge t) \le 4^d \exp(-ct^2/\tau^2)$, where $c > 0$ is a numerical constant. Replacing t^2 with $O(1)(d\tau^2 + \tau^2 t^2)$ gives that $\mathbb{P}(\|Z\|_2 \ge C\tau\sqrt{d + t^2}) \le \exp(-t^2)$, and for any $\gamma > 0$, setting $t^2 = \log\frac{n}{\gamma}$ yields that with probability at least $1-\gamma$,

$$\max_{i\le n} \|Z_i\|_2^2 \le O(1)\tau^2\left[d + \tau^2 \log\frac{n}{\gamma}\right].$$

To control the covariance, we use Vershynin (2012, Theorem 5.39), which gives that with probability at least $1 - 2e^{-ct^2}$, $\left\|\overline{\Sigma}_Z - I\right\|_{\text{op}} \le O(1)\tau^2 \max\{\sqrt{d/n} + t/\sqrt{n}, d/n + t^2/n\}$, so that (igoring the sub-Gaussian constant) for $n \gtrsim d$, setting $t^2 = O(1)\log\frac{1}{\gamma}$ gives $\left\|\overline{\Sigma}_Z - I\right\|_{\text{op}} \le \frac{1}{2}$ with probability at least $1 - \gamma$. Set $\gamma = \beta/2$.

In this case, $\Sigma^{-1/2}X_i$ is τ^2-sub-Gaussian, so $M^2 \lesssim \tau^2(d + \log\frac{n}{\beta})$ in Assumption A1. Thus, the sample mean concentrates as $\left\|\overline{X}_n - \mu\right\|_\Sigma \lesssim \tau\sqrt{(d + \log(1/\delta))/n}$ with probability *at least* $1-\delta$*, and assuming* $\delta \geq e^{-d}$*, Theorem 8 then implies with probability at least* $1-O(\delta)$ *over* $\widetilde{\mu} = \texttt{ADAMEAN}_{\varepsilon,\delta}(X_{1:n})$ *that (ignoring polylogarithmic factors in* n*)*

$$\|\widetilde{\mu}-\mu\|_\Sigma = \widetilde{O}\left(\tau\sqrt{\frac{d}{n}} + \frac{\tau d \log\frac{1}{\delta}}{n\varepsilon}\right).$$

This rate is, up to a factor of $\log\frac{1}{\delta}$ *and polylogarithmic factors in* n*, minimax-optimal for the sub-Gaussian setting (see Steinke and Ullman (2017) or Kamath et al. (2018, Lemma 6.7) for a lower bound on Gaussian mean estimation with known covariance matrix).*

Example 2 (General moment bounds) *Suppose for some* $p \geq 4$ *we have* $\mathbb{E}[\|X_i-\mu\|_\Sigma^p] = \mathbb{E}[\|Z_i\|_2^p] \leq \tau^p d^{p/2}$*, where necessarily* $\tau \geq 1$*. Then we can give two results: the first being that asymptotically* $M = o(n^{1/p})$ *and the second more quantitative. For the first, we claim that* $\max_{i\leq n}\|Z_i\|_2/n^{1/p} \overset{a.s.}{\to} 0$*. To see this, note that for any* $\varepsilon > 0$*,*

$$\infty > \frac{1}{\varepsilon^p}\mathbb{E}[\|Z_1\|_2^p] = \int_0^\infty \mathbb{P}(\|Z_1\|_2^p \geq \varepsilon^p t)\,dt \geq \sum_{i=1}^\infty \mathbb{P}(\|Z_i\|_2^p \geq \varepsilon^p i).$$

By the Borel-Cantelli lemma, the event $\|Z_n\|_2 \geq \varepsilon n^{1/p}$ *occurs only finitely often, and so the claim follows. Meanwhile, the strong law of large numbers guarantees that* $\overline{\Sigma}_Z \overset{a.s.}{\to} I$*.*

For more quantitative parameters, we first get by Markov's inequality that

$$\mathbb{P}(\max_{i\in[n]}\|Z_i\|_2 > t) \leq \frac{n\mathbb{E}[\|Z_1\|_2^p]}{t^p} \leq \frac{n\tau^p d^{p/2}}{t^p},$$

so setting $M \asymp \tau\sqrt{d}n^{1/p}/\beta^{1/p}$*, we have* $\max_i\|Z_i\|_2 \leq M/c_1$ *with probability at least* $1-\beta$*. To show concentration of the covariance matrix, we apply Chen et al. (2012, Theorem A.1 Part 2), treating p as a constant, obtaining*

$$\begin{aligned} n\mathbb{E}\left[\left\|\overline{\Sigma}_Z - I\right\|_{\text{op}}^{p/2}\right]^{2/p} &\lesssim \sqrt{n\log d}\sqrt{\mathbb{E}[\|Z\|_2^4]} + (n^{2/p}\log d)\mathbb{E}[\|Z_1\|_2^p]^{2/p} \\ &\lesssim \max\{\sqrt{n\log d}, n^{2/p}\log d\}\mathbb{E}[\|Z_1\|_2^p]^{2/p}, \end{aligned}$$

and so by Markov's inequality

$$\mathbb{P}(\left\|\overline{\Sigma}_Z - I\right\|_{\text{op}} > \tfrac{1}{2}) \lesssim \max\{n^{-p/4}\log^{p/4} d, n^{1-p/2}\log^{p/2} d\}\mathbb{E}[\|Z_1\|_2^p] \lesssim \frac{\tau^p(d\log d)^{p/2}}{n^{p/2-1}},$$

which has bound β *when* $n \gtrsim (\tau^2 d\log d)^{p/(p-2)}\beta^{-2/(p-2)}$*.*

Recall here that $\mathbb{E}[\|X_i-\mu\|_\Sigma^p] \leq \tau^p d^{p/2}$ *for* $p \geq 4$ *and* $\tau \geq 1$*. By Theorem 8, with probability at least* $1-5\delta$ *over* $\widetilde{\mu} = \texttt{ADAMEAN}_{\varepsilon,\delta}(X_{1:n})$*, we have*

$$\|\widetilde{\mu}-\mu\|_\Sigma \leq \left\|\overline{X}_n - \mu\right\|_\Sigma + \widetilde{O}\left(\frac{\max_i\|X_i-\mu\|_\Sigma\sqrt{d}\log\frac{1}{\delta}}{n\varepsilon}\right)$$

so long as the empirical covariance satisfies $\frac{1}{2}\Sigma \preceq \overline{\Sigma}_n \preceq \frac{3}{2}\Sigma$*. As this occurs with constant probability and* $\|\overline{X}_n-\mu\|_\Sigma \lesssim \sqrt{d/n}$ *with constant probability, we substitute the bounds on* $\max_i\|X_i-\mu\|_\Sigma$ *from Example 2 to obtain that with (any) constant probability,*

$$\|\widetilde{\mu}-\mu\|_\Sigma = \widetilde{O}\left(\sqrt{\frac{d}{n}} + \frac{\tau d\log\frac{1}{\delta}}{n^{1-1/p}\varepsilon}\right).$$