Privacy Induces Robustness: Information-Computation Gaps and Sparse Mean Estimation

Kristian Georgiev
MIT EECS
Cambridge, MA 02139
krisgrg@mit.edu

Samuel B. Hopkins
MIT EECS
Cambridge, MA 02139
samhop@mit.edu

Abstract

We establish a simple connection between robust and differentially-private algorithms: private mechanisms *which perform well with very high probability* are automatically robust in the sense that they retain accuracy even if a constant fraction of the samples they receive are adversarially corrupted. Since optimal mechanisms typically achieve these high success probabilities, our results imply that optimal private mechanisms for many basic statistics problems are robust.

We investigate the consequences of this observation for both algorithms and computational complexity across different statistical problems. Assuming the Brennan-Bresler secret-leakage planted clique conjecture, we demonstrate a fundamental tradeoff between computational efficiency, privacy leakage, and success probability for sparse mean estimation. Private algorithms which match this tradeoff are not yet known – we achieve that (up to polylogarithmic factors) in a polynomially-large range of parameters via the Sum-of-Squares method.

To establish an information-computation gap for private sparse mean estimation, we also design new (exponential-time) mechanisms using fewer samples than efficient algorithms must use. Finally, we give evidence for privacy-induced information-computation gaps for several other statistics and learning problems, including PAC learning parity functions and estimation of the mean of a multivariate Gaussian.

1 Introduction

Avoiding leakage of sensitive data and robustness to data corruption or model misspecification are often key goals for designers of statistical estimators. Both these properties admit mathematical formalizations, and a great deal of recent work in (algorithmic) statistics has gone into designing and analyzing algorithms that satisfy them.

Differential privacy (DP) is the gold-standard formal definition of privacy for algorithms processing sensitive data [DMNS06]. DP requires that the distribution of outputs of an algorithm (or "mechanism") M is insensitive to exchanging a small number of individuals in its input dataset. It offers such strong guarantees against privacy leakage that in addition to its ongoing adoption in industry, the US Census Bureau employs DP to satisfy its legal mandate to protect privacy [AACM+22].

The η-*contamination model* is a stringent formalization of robustness against model misspecification. It generalizes the classical model of i.i.d. samples: a dataset is η-corrupted for some $\eta > 0$ if it is first drawn i.i.d, but then an η-fraction of samples have been arbitrarily corrupted by a malicious adversary (who may look at the whole dataset) [Hub65, Tuk75, Hub11]. An η-*robust* algorithm is one which maintains guarantees of accuracy when given η-corrupted samples.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Both robustness and privacy demand that the output of some statistical method not "change too much" when one or a few input samples are modified arbitrarily. This conceptual similarity has not gone unnoticed: [DL09] observe that "robust statistical estimators present an excellent starting point for differentially private estimators," and recent works have even made good on this idea in high-dimensional settings, taking inspiration from robust statistics to design private mechanisms [LKKO21, HBK22]. However, a general account of which private algorithms can be made robust, or vice versa, remains an open problem. Here we take a small step by tackling:

Question 1: When are private algorithms (also) robust?

While common wisdom holds that privacy and robustness are not formally comparable, we give a meta-theorem Theorem 2.1 which quantifies the degree of robustness that private mechanisms exhibit: at a high level, mechanisms satisfying quantitatively-strong versions of DP are automatically robust.

This connection between robustness and privacy gives us the tools to investigate a second basic question, concerning tradeoffs among computational resources, privacy, and statistical accuracy. *Accuracy*-privacy tradeoffs appear in even the most elementary statistical settings. For instance, estimating the mean of a d-dimensional Gaussian to ℓ_2 error α requires $\Theta(d/\alpha^2)$ samples non-privately, but subject to ε-DP* requires (roughly) $\Theta(d/\alpha^2 + d/(\alpha\varepsilon))$ samples [BKSW19]. While important in their own right, these two-way tradeoffs are not the whole story. In numerous cases, even including the Gaussian mean estimation problem, *computationally efficient algorithms which achieve the optimal privacy-accuracy tradeoffs are not known*. This brings up the question:

Question 2: Does requiring differential privacy introduce computational barriers in statistics?

While computational barriers to efficient private algorithms are known in some settings [GHRU13, Ull16, Bun20], these apply only to algorithms with *worst-case* accuracy guarantees – in statistical settings we are asking only for *average-case* accuracy guarantees (although privacy should still hold with respect to all possible datasets).

An archetypal problem for which computational barriers arise when estimators are required to satisfy criteria beyond accuracy is that of *sparse mean estimation*. The goal there is to estimate a k-sparse vector $\mu \in \mathbb{R}^d$ using independent samples from a distribution with mean μ. With no requirements on privacy or robustness, this can be accomplished with $O(k \log d)$ samples in polynomial time via simple thresholding-based estimators. In exponential time, it is possible to retain $O(k \log d)$ sample complexity and satisfy privacy and robustness (as we show in this paper), but in polynomial time $\Omega(k^2)$ samples are required (under a variant of the planted clique conjecture) just to satisfy robustness [BB20]. Given the outlined connection between robustness and privacy, the existence of an information-computation gap for *robust* sparse mean estimation naturally leads to:

Question 3: How many samples do poly-time private algorithms for sparse mean estimation require?

1.1 Our Contributions

We make three main contributions: (a) a meta-theorem characterizing robustness of private mechanisms; (b) a case study of sparse mean estimation, including a computational lower bound arising from robustness, and a new Sum-of-Squares-based algorithm whose sample complexity matches that lower bound in parameter regimes where no such algorithms were previously known; and (c) a collection of computational and information-theoretic lower bounds for private mechanisms, inherited from lower bounds for robust algorithms.

Robustness of Optimal Private Mechanisms Our first contribution is a simple but useful observation: *mechanisms with strong **group privacy** guarantees are automatically robust!* By "strong group privacy," we mean guarantees strong enough to retain privacy and accuracy when a constant fraction of individuals in the dataset are exchanged with others. We capture this in Theorem 2.1.

In spite of its simplicity, robustness of strongly-group-private mechanisms has significant consequences for simultaneously robust and private mechanisms, a topic of much recent interest [KMV21, LKO21, LKKO21, UKRK22, CS22], because mechanisms with optimal privacy-accuracy tradeoffs often automatically satisfy strong group privacy.

*See Section 2 for formal definitions.

Table 1: **Algorithms for sparse mean estimation.** In the input column, $\mathcal{N}$ signifies that the algorithm takes i.i.d. samples from $\mathcal{N}(\mu, I)$, and $\Sigma \preceq I$ — samples from a distribution with bounded covariance. The sample complexity column hides polylogarithmic factors in the ambient dimension d and a priori mean bound (in ℓ_2) R. In the auto-robust column we indicate whether our meta-theorem (Theorem 2.1) implies that the algorithm is robust to corruptions of an $\eta = 1/\mathrm{poly}\log(d, R)$ fraction of the samples.

Algorithm	Input	Runtime	Sample Complexity	Auto-robust
`Hypothesis Sel.` [BKSW19]	$\mathcal{N}$	exp	$\frac{k+\log(1/\beta)}{\alpha\varepsilon} + \frac{k+\log(1/\beta)}{\alpha^2}$	✓
`Subset Sel.` [Theorem C.7]	$\Sigma \preceq I$	exp	$\frac{k+\log(1/\beta)}{\alpha^2\varepsilon}$	✓
`SoS` [Theorem 1.2]	$\mathcal{N}$	poly	$\frac{k^2+\log(1/\beta)}{\alpha^2\varepsilon}$	✓
`Threshold` [Theorem 4.1]	$\mathcal{N}$	poly	$\frac{k^2\log(1/\beta)}{\alpha^2\varepsilon}$	✗
`Peeling` [CWZ21] [1]	$\mathcal{N}$	poly	$R\left(\frac{k^{1.5}\log(1/\beta)}{\alpha\varepsilon} + \frac{k^{1.5}\log(1/\beta)}{\alpha^2}\right)$	✗

[1] `Peeling` is stated only for (ε, δ)-DP in [CWZ21]. We use a (mildly) modified ε-DP version, which we formally state in Appendix G.

The privacy guarantees needed for automatic robustness rely on quantitatively-strong forms of differential privacy: either so-called *pure* DP, or *approximate* DP, but with the additive error parameter δ taken exponentially small, and high success probability. (By contrast, typical convention in the privacy literature is to take δ only polynomially small in other parameters.) For this reason, we focus primarily on mechanisms satisfying pure DP and achieving high success probability.

Sparse Mean Estimation Recall that in sparse mean estimation the goal is to estimate a k-sparse vector $\mu \in \mathbb{R}^d$ to ℓ_2-error α, succeeding with probability $1-\beta$, using i.i.d. samples from a distribution with mean μ. We contribute (a) new exponential-time ε-DP mechanisms using $O(k \log d)$ samples, (b) evidence that poly-time DP algorithms with high success probabilities require $\tilde{\Omega}(k^2)$ samples, and (c) new poly-time algorithms using $\tilde{O}(k^2)$ samples (in certain parameter regimes). One of these algorithms, using the SoS exponential mechanism of [HKM21], is our main technical contribution.

Information-theoretic bounds: First, as a baseline, we study sparse mean estimation without worrying about running time. We show that even subject to both privacy and robustness, $O(k \log d)$ samples suffice, with sample complexities differing in their dependence on α and ε on between the cases that the underlying distribution is assumed to be Gaussian versus allowing for heavy-tailed distributions (assuming only bounded covariance), as in the case of non-sparse mean estimation. See the estimators `Hypothesis Sel` and `Subset Sel` in Table 1, and Appendix C for formal statements.

Computational lower bound: Next, we give evidence for a privacy-samples-success probability tradeoff for efficient private algorithms: a private algorithm for sparse mean estimation with high success probability will satisfy strong group privacy, and hence robustness, but efficient and robust algorithms for sparse mean estimation require $\Omega(k^2)$ samples (assuming the planted clique conjecture).

Corollary 1.1 (of Theorem 2.1 and [BB20], Theorem 3.1)**.** *Assume the* secret-leakage planted clique *conjecture [BB20]. For polynomially-related n, k, and d, with $k = o(\sqrt{d})$, $\beta \in (0,1)$, and $\alpha, \varepsilon > 0$, assume $\beta \le 2^{-\varepsilon\sqrt{n}}$. Let $m(\beta, n, \varepsilon)$ be the greatest value less than $\log(1/\beta)/(\varepsilon n)$ in the set $\{n^{-o(1)}\} \cup \{n^{-1/(2t)} : t \in \mathbb{N}, t \ge 1\}$. Every polynomial-time, ε-DP algorithm which, for any k-sparse $\mu \in \mathbb{R}^d$ with $\|\mu\| \le poly(d)$ can take n samples from $\mathcal{N}(\mu, I)$ and return $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le \alpha$ with probability $1-\beta$ requires $n \ge \frac{k^2 m^2}{\alpha^4 \cdot poly\log(d, \frac{1}{\varepsilon}, \frac{1}{\alpha})}$.*

Pretending $m = \log(1/\beta)/(\varepsilon n)$ and ignoring logarithmic factors in $d, \frac{1}{\varepsilon}, \frac{1}{\alpha}$, the lower bound says $n \gtrsim \left(\frac{k^2}{\alpha^2} \cdot \left(\frac{\log(1/\beta)}{\alpha\varepsilon}\right)^2\right)^{1/3}$ samples are required by efficient ε-DP algorithms for sparse mean estimation. This is the geometric mean of three terms: $\frac{k^2}{\alpha^2}$ and $\frac{\log(1/\beta)}{\alpha\varepsilon}$ (twice). We conjecture that the max of these is actually a lower bound: $n \gtrsim \frac{k^2}{\alpha^2} + \frac{\log(1/\beta)}{\alpha\varepsilon}$.

Even if this stronger lower bound were true, *existing efficient algorithms for sparse mean estimation would not match it*. The state-of-the-art for sparse ε-DP mean estimation using techniques in the literature is a modification (to achieve pure DP) of an algorithm by [CWZ21], whose sample complexity scales with $k^{1.5}\log(1/\beta)$ (see Appendix G).

Efficient algorithms – SoS: Are there polynomial-time algorithms which match the tradeoff from Corollary 1.1, or the stronger conjectured one above? Our main algorithmic contribution is a new algorithm for sparse mean estimation which matches the tradeoff of Corollary 1.1 (up to polylog factors) under the conditions: (1) $\alpha, \varepsilon \geq 1/\mathrm{poly}\log(d)$, (2) $k^2 \approx \log(1/\beta)$, and (3) $k \geq d^{0.4}$. Of these conditions, (2) could be removed if the stronger conjectural lower bound above were true, while we believe that (1) and (3) are shortcomings of our algorithm, and might be removable. We are not aware of any previous efficient private algorithm which matches the above conjectured tradeoff for any simultaneously super-constant k and $\log(1/\beta)$.

Theorem 1.2. *There exists $C > 0$ such that for every $\varepsilon, R > 0$, $, \alpha, \beta \in (0,1)$, and large-enough $d, k \in \mathbb{N}$ such that $k \geq d^{0.4}$, there is a polynomial-time ε-DP algorithm* `SoS` *with the following guarantees. For every k-sparse $\mu \in \mathbb{R}^d$ with $\|\mu\| \leq R$, given η-corrupted samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, I)$, with probability at least $1-\beta$, the algorithm outputs $\hat{\mu} \in \mathbb{R}^d$ such that $\|\mu - \hat{\mu}\| \leq \alpha + O\left(\sqrt{(\log(Rd))^C \eta}\right)$, so long as $n \gg (\log(Rd))^C \cdot \frac{k^2 + \log(1/\beta) + \log\log R}{\alpha^2 \varepsilon}$.*

Our algorithm employs the Sum-of-Squares exponential mechanism invented by [HKM21] for a private mean estimation algorithm, but adapting this approach to the sparse setting requires overcoming several technical roadblocks (see Section 3 and Appendix D). Our approach hits a technical obstacle related to the volume of the d-dimensional ℓ_1 ball when $k \ll d^{0.4}$, and we leave as an open problem to match or approach the tradeoff in Corollary 1.1 for a wider range of parameters.

Linear-time coordinate selection: Finally, what can be accomplished with a "truly efficient" algorithm – one which does not require solving large semidefinite programs as in the SoS exponential mechanism? The state-of-the-art private algorithm for sparse mean estimation, of [CWZ21], is a simple iterative coordinate-selection procedure. With a minor modification to achieve a pure DP guarantee, that algorithm uses $O_R(k^{1.5}\log d)$ to estimate k-sparse mean vectors $\mu \in \mathbb{R}^d$ with $\|\mu\| \leq R$, with probability $1-\beta$. However, the $O_R(\cdot)$ hides a *linear*, rather than logarithmic, dependence on R, which is very costly even for moderately-large values of R!

While it is now well understood how to obtain logarithmic-in-R sample complexity for non-sparse mean estimation, standard approaches introduce a linear dependence on ambient dimension d. We improve over the algorithm of [CWZ21] while maintaining linear running time by designing a simple thresholding procedure for estimating the support of μ requiring a number of samples which is *independent of* R. Once the support is known, the ambient dimension of the problem can be reduced from d to k, and a off-the-shelf private non-sparse mean estimation algorithm can be run. A formal statement is given in Section 4. We demonstrate with experiments on synthetic data (Section 3) that the coordinate-selection procedure `Threshold`, is substantially more accurate than that used by prior state of the art [CWZ21].

By contrast to `SoS`, the linear-time algorithm `Threshold` has sample complexity that scales with $k^2\log(1/\beta)$ rather than $k^2 + \log(1/\beta)$. While this difference might appear minor at first, it has significant consequences: the linear-time algorithm cannot match the computational lower bound in Corollary 1.1, and it does not have strong-enough privacy guarantees to be robust via Theorem 2.1. Note that our results do not preclude other tradeoffs between k and $\log(1/\beta)$. For instance, we leave it as an exciting open problem to design an efficient algorithm with sample complexity scaling with $k\log(1/\beta)$.

Information-Computation Gaps in Private Statistics A problem in private statistics has an *information-computation* gap if the accuracy-privacy tradeoffs achieved by optimal (exponential-time) mechanisms for that problem cannot be achieved by polynomial-time algorithms. Our connection between privacy and robustness can be used to give evidence for several information-computation gaps in private statistics beyond sparse mean estimation.

Gaussian mean estimation (Corollary B.2): Gaussian mean estimation, where the goal is to estimate $\mu \in \mathbb{R}^d$ given i.i.d. samples from $\mathcal{N}(\mu, I)$, is arguably even simpler than sparse mean estimation. We give evidence for an information-computation gap: polynomial-time ε-DP algorithms obtaining

accuracy α with probability at least $1-\beta$ require $n \geq \log(1/\beta) \cdot (\log(1/\alpha))^{1/2-o(1)}/(\alpha\varepsilon))$ samples, unless there exist robust polynomial-time algorithms for Gaussian mean estimation which would contradict known statistical query lower bounds [DKS17] – this is a $\log(1/\alpha)^{1/2-o(1)}$ factor gap.

Learning parities (Corollary B.3): We also consider one of the most fundamental *supervised* learning problems: privately PAC learning parity functions from labeled examples. For each $S \subseteq [n]$, we can define a parity function $f_S : \{\pm 1\}^n \to \{\pm 1\}$ by $f_S(x) = \prod_{i \in S} x_i$. The goal is to take labeled examples (x, y) drawn from some distribution D and find a parity function f_S such that $\Pr_{(x,y)\sim D}(f_S(x) = y) = 1$, assuming one exists. While polynomial-time private algorithms for learning parities are known [KLN+11], we show, via Theorem 2.1 that the failure probabilities of any such algorithms must be larger than what can be achieved in exponential time, or else $RP = NP$.

Information-Theoretic Lower Bounds in Private Statistics Finally, we show that the connection to robustness can provide information-theoretic lower bounds for private mechanisms. As an example, we study private covariance testing (Corollary B.4), where the goal is to take samples from $\mathcal{N}(0, \Sigma)$ and detect whether $\Sigma = I \in \mathbb{R}^{d\times d}$ or if $\|\Sigma - I\|_F \geq \gamma$. Appealing to the lower bound of [DK21] for robust covariance testing, we give a lower bound for private covariance testing, showing that $\Omega(d^2)$ samples are required by private algorithms with high success probabilities, while $O(d)$ suffice non-privately.

Formal statements for the results on Gaussian mean estimation, learning parities, and covariance testing can be found in Appendix B.

1.2 Related Work

Privacy and robustness. As mentioned in Section 1, there is a rich history of connections between DP and robustness, starting from the propose-test-release (PTR) framework of Dwork and Lei [DL09]. Building on top of PTR, a number of recent works tackle high-dimensional statistics problems by leveraging robust primitives [BGS+21, LKO21], themselves inspired by a recent revolution in high-dimensional robust statistics [DK19]. On the flip side, private algorithms for certain problems have been shown to "automatically" exhibit a small amount of robustness [TS13, HKM21].

Sparse mean estimation. Without privacy or robustness requirements, it is a folklore result that the truncated empirical mean achieves the information-theoretically optimal rate. In the approximate DP case, [TS13] show that the stability of LASSO can be leveraged for private support selection and private sparse regression via the sample-and-aggregate framework [NRS07]. Cai, Wang, and Zhang [CWZ21] show information-theoretic lower bounds for approximate DP (based on tracing attacks [HSR+08]) and computationally efficient algorithms that match those bounds with constant probability under additional assumptions on the ℓ_∞ norm of the mean μ. In the presence of η-corruptions of the samples, [BDLS17] gives an $\tilde{O}\left(k^2 \log(d)/\eta^2\right)$-sample algorithm matching SQ lower bounds from [DKS17].

Computational Roadblocks to Privacy. Several prior works investigate computational roadblocks to privacy arising from cryptographic considerations, e.g. [Ull16, UV11]. The hard problem instances constructed in such works have a worst-case flavor, while we are interested in computational hardness for typical datasets/those drawn i.i.d. from an underlying probability distribution.

Sum-of-Squares Method. The SoS method for algorithm design in high-dimensional statistics has led to a number recent of algorithmic advances – see the survey [RSS18]. [HKM21], which introduces the SoS exponential mechanism, is most closely related, and provides the foundations for Theorem 1.2.

Lower bounds for private mechanisms There are multiple works leveraging group privacy to derive lower bounds for private algorithms, both for pure and approximate DP [HT10, De12, SU15, BS16]. They are all information-theoretic in nature, while we also provide computational hardness results. Additionally, to the best of our knowledge, there are no existing work relating lower bounds for privacy to ones for robustness.

2 Automatic Robustness Meta-Theorem and Private Robust Mechanisms

We formally define DP in Appendix A. With this in hand, With this in hand, we are ready to state and prove our meta-theorem on automatic robustness of private algorithms.

Theorem 2.1 (Automatic Robustness Meta-Theorem)**.** *Let $M : \mathcal{X}^* \to \mathcal{O}$ be an (ε, δ)-private map from datasets $\mathcal{X}^*$ to outputs $\mathcal{O}$. For every dataset $X_1, \ldots, X_n$, let $G_{X_1,\ldots,X_n} \subseteq \mathcal{O}$ be a set of* good *outputs. Suppose that $M(X_1, \ldots, X_n) \in G_{X_1,\ldots,X_n}$ with probability at least $1 - \beta$ for some $\beta = \beta(n)$. Then, for every $n \in \mathbb{N}$, on n-element datasets M is* robust *to adversarial corruption of any $\eta(n)$-fraction of inputs, where*

$$\eta(n) = O\left(\min\left(\frac{\log 1/\beta}{\varepsilon n}, \frac{\log 1/\delta}{\varepsilon n \log n}\right)\right),$$

meaning that for every $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$ differing on only ηn elements, $M(X'_1, \ldots, X'_n) \in G_{X_1,\ldots,X_n}$ with probability at least $1 - \beta^{\Omega(1)}$.

Proof of Theorem 2.1. Consider ηn intermediate datasets $(X_1, \ldots, X_n) = \mathbf{X}_0, \ldots, \mathbf{X}_{\eta n} = (X'_1, \ldots, X'_n)$, where a single coordinate X_i is modified in passing from $\mathbf{X}_j$ to $\mathbf{X}_{j+1}$. Let $p_j = \Pr(M(\mathbf{X}_j) \in G_{X_1,\ldots,X_n})$. Then we have the following recurrence for $(1 - p_j)$:

$$(1 - p_j) \le e^{\varepsilon}(1 - p_{j-1}) + \delta \text{ for } j \ge 1, \text{ and } (1 - p_0) \le \beta\,,$$

from which we obtain $(1 - p_{\eta n}) \le e^{\varepsilon \eta n}(\beta + \eta n \delta)$. The conclusion follows. □

An analogous statement for concentrated DP [BS16] is presented in Appendix H.

Theorem 2.1 can be applied broadly to show that optimal private mechanisms are automatically robust, frequently even with optimal dependence of the lost accuracy on the rate of corruption. This breadth is possible because, for many statistical problems, *private mechanisms obtaining information-theoretically optimal privacy-accuracy tradeoffs automatically have the strong group privacy guarantees needed to apply Theorem 2.1*. This is because strong group privacy for a mechanism M is implied by two other desirable properties of private mechanisms: (1) M satisfies *pure* DP (or, (ε, δ)-DP for small choices of δ), and (2) M produces accurate results *with high probability* over the randomness used internally by the mechanism. We give two examples of this phenomenon below.

On the success probabilities of private algorithms: Before turning to examples, we observe that Theorem 2.1 only gives robustness to a constant fraction of corrupted samples for private algorithms which have very high probability of succeeding – to obtain $\eta \ge \Omega(1)$ requires $\beta \le 2^{-\Omega(n)}$. In most work on randomized algorithm design, the difference between succeeding with probability $2/3$ versus $1 - \beta$ for small β can be treated as an afterthought, because algorithms can be repeated to amplify success probability. *But this kind of naive repetition causes privacy leakage!*

In spite of this, optimal private mechanisms in statistics often do succeed with high probability, using more sophisticated approaches than naive repetition: in fact, high success probability is generally implied by M's outputs having (asymptotically) *optimal confidence intervals*. Our work points to a need for algorithm designers to focus on the confidence intervals/success probabilities of private algorithms: the payoff is robustness for free.

On Black-Box Robustification of Optimal Private Mechanisms: We observe that Theorem 2.1 can be used to automatically obtain robust and private mechanisms from private ones with high success probabilities. For instance, *(1) Hypothesis selection:* The private hypothesis selection procedure of [BKSW19], already proved robust against non-adaptive adversaries, is additionally robust to corruptions made by adaptive adversaries. *(2) Affine-invariant mean estimation:* [BGS+21, LKO21] study mean-estimation mechanisms which provide error guarantees in the *Mahalanobis distance* $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$ given samples from $\mathcal{N}(\mu, \Sigma)$. [BGS+21] give a private mechanism for this problem with high success probability, and [LKO21] give a simultaneously private and robust mechanism. In either case, because both mechanisms provide strong-enough privacy guarantees to apply Theorem 2.1, robustness can be obtained in a black-box fashion knowing only the privacy guarantees.

From Theorem 2.1 to Lower Bounds for Private Statistics: As we discussed in Section 1.1, we use Theorem 2.1 to prove both computational and information-theoretic sample-complexity lower bounds for private algorithms for sparse mean estimation (Corollary 1.1), non-sparse mean estimation, learning parities, and covariance testing. We prove Corollary 1.1 in the next section, and defer the remaining statements and proofs of lower bounds to the supplement.

3 Sparse Private Mean Estimation: Techniques

In this section, we overview ideas which go into our results on sparse mean estimation, starting from the information-theoretic results, and moving on to computational barriers and polynomial-time algorithms.

Information-theoretic bounds In the absence of computational considerations, the landscape for *Gaussian* sparse mean estimation can be understood via standard tools: a packing-based lower bound, and a matching (exponential-time) mechanism can be constructed as a direct corollary of existing results in the literature on private hypothesis selection [BKSW19]. For completeness, we carry out those in Appendix C.

In the *heavy-tailed* case, assuming only that the samples $X_1, \ldots, X_n$ are drawn from a distribution with k-sparse mean μ and with bounded covariance, one can no longer construct a small cover of the set of possible distributions; this precludes an approach as general as hypothesis selection from directly applying. Instead, we design a mechanism which first selects a subset of k coordinates, then hands off to a non-sparse mean estimation mechanism run just on those coordinates.

To select the coordinates, we use the exponential mechanism. To define a score function, we take inspiration from recent ideas in high-dimensional statistics [LM19] using empirical quantiles of univariate projections of the samples. The key idea for coordinate selection is to restrict attention to projection in sparse directions. For samples $X_1, \ldots, X_n$, we define the following score function on subsets of coordinates $T \subseteq [d]$:

$$S(\{X_i\}_{i \le n}, T; L) = \max_{v \in \mathbb{R}^k, \|v\|_2 = 1} \sum_{i=1}^{n} \mathbb{1}\{v^\top X_{iT} \ge L\}, \tag{1}$$

parametrized by a scalar threshold L[†]. Then we sample T from the distribution $\Pr(T) \propto \exp(\varepsilon S(\{X_i\}_{i \le n}, T; L)$. Since S has sensitivity $\Delta(S) = 1$,[‡] the resulting mechanism satisfies ε-DP. We show using standard concentration tools that if $n \gg O(k \log(d) + \log(1/\beta)/(\alpha^2 \varepsilon))$ then this mechanism identifies a subset T containing all but α of the ℓ_2-mass of μ, with probability at least $1 - \beta$. To do so, we choose the threshold L such that with high probability all "bad" subsets containing a small portion of the ℓ_2 mass of μ have "low" score compared to "good" ones for which $\|\mu_T\| \approx \|\mu\|$; finally, we use bucketing to control the variance of samples such that we can set L as close to zero as possible without allowing "bad" subsets to achieve high score.

We then delegate the mean estimation on the candidate set of coordinates to any information-theoretically optimal (non-sparse) mean estimation mechanism. For the matching lower bound, standard packing-based arguments suffice (see Proposition C.13 for a formal argument).

Computational barrier Now we turn to the landscape when we require polynomial-time. Starting with barriers, we first show how to get Corollary 1.1 from our meta-theorem.

Proof of Corollary 1.1. First of all, since $n \ge \log(1/\beta)^2/\varepsilon^2$, there exists a value m in the given set. Now suppose a polynomial-time private algorithm exists tolerating $n \le k^2 m^2/(\alpha^4 \mathrm{poly}\log(d, 1/\varepsilon, 1/\alpha))$. By adjusting parameters β, ε to artificially weaken the guarantees as necessary, we may assume that $m(\beta, n, \varepsilon) = \log(1/\beta)/(\varepsilon n)$. Now, Theorem 3.1 of [BB20] shows that no algorithm for sparse mean estimation under these parameters can tolerate an η-fraction of adversarial corruptions with $k^2\eta^2/\alpha^4 \gtrsim n$, where $\gtrsim$ hides polylogarithmic factors. But from Theorem 2.1 our hypothesized private algorithm tolerates an $\eta = \log(1/\beta)/(\varepsilon n)$ fraction of adversarial corruptions, which is a contradiction. □

Overview of SoS Algorithm for Sparse Mean Estimation (Theorem 1.2) We remain informal in this section and defer mathematical rigor to the supplement. For now, let $\alpha = \Theta(1)$. A standard trick reduces from $\alpha \ll 1$ to this case.

As with prior work on both private and robust mean estimation (e.g. [CFB19, HKM21]) our algorithm produces a series of iterates $x_0, x_1, \ldots, x_T$ for $T = O(\log d)$, where x_0 is the origin, with the invariant

[†]For a sample $X \in \mathbb{R}^d$, let X_T to be the projection onto the basis vectors with indices in the set $T \subseteq [d]$.

[‡]That is, it changes by at most 1 when any sample is exchanged for another.

that $\|x_t - \mu\| \le 0.9\|x_{t-1} - \mu\|$. To accomplish this, given x_{t-1} with $\|\mu - x_{t-1}\| \gg 1$, and samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, I)$, we find a unit vector v such that $\langle v, \mu - x_{t-1}\rangle \ge 0.9\|\mu - x_{t-1}\|$; then we could take $x_t = x_{t-1} + \Omega(\|\mu - x_{t-1}\|)v$.

Unlike prior works, for reasons we will see shortly, we also need the invariant that x_t is k-sparse. So, we actually take x_t to be $x_{t-1} + \Omega(\|\mu - x_{t-1}\|)v$ with all but the largest-magnitude k coordinates set to 0; we show that this thresholding step cannot increase the distance to μ by too much (Lemma D.4).

Picking a gradient in exponential time: To find such a gradient vector v, we use the SoS exponential mechanism [HKM21]. Let us first see how we would pick a gradient vector v using the (non-SoS) exponential mechanism, but allowing exponential running time. For a given dataset $\mathcal{X} = X_1, \ldots, X_n$, the goal is to find a *score function* $s_{\mathcal{X}}(v) \in \mathbb{R}$ which assigns each $(2k)$-*sparse* unit vector a score, such that for neighboring datasets $\mathcal{X}, \mathcal{X}'$ we always have $|s_{\mathcal{X}}(v) - s_{\mathcal{X}'}(v)| \le 1$ ("bounded differences"). Here we choose $2k$-sparse because $\mu - x_{t-1}$ is itself $2k$ sparse. Then outputting a random v, where each is chosen with probability $\propto \exp(\varepsilon s_{\mathcal{X}}(v))$, gives an ε-DP mechanism. The goal is that vs with high scores are closer to $\mu - x_{t-1}$.

A good choice of score function turns out to be $s_{\mathcal{X}}(v) = \sum_{i \le n} \mathbf{1}\{\langle v, X_i - x_{t-1}\rangle \ge \|x_{t-1} - \mu\| - O(1)\}$ – using standard concentration of measure one can show that as long as $n \gg k \log d$, for v s.t. $s_{\mathcal{X}}(v) \ge 0.9n$ we will have $\langle v, \mu - x_{t-1}\rangle \ge 0.9\|\mu - x_{t-1}\|$ ("utility"). Furthermore, if one were to sample a uniformly random k-sparse unit vector, it would satisfy $\langle v, \mu - x_{t-1}\rangle \ge 0.9\|\mu - x_{t-1}\|$ with probability $d^{-O(k)}$ (since $\mu - x_{t-1}$ is $2k$-sparse). So, the distribution given by $\Pr(v) \propto \exp(\varepsilon s_{\mathcal{X}}(v))$ puts $1 - \beta$ probability on v with $s_{\mathcal{X}}(v) \ge 0.9n$, so long as $n \gg (k \log d + \log(1/\beta))/\varepsilon$, since it "boosts" the probabilities of these high scoring vectors by a factor of $\exp(\Omega(d \log k)) \ge d^{\Omega(k)}$.

Here we have crucially used the fact that x_{t-1}, and hence $\mu - x_{t-1}$, is sparse: otherwise, the gradient v we need to select would not be sparse, and we would need to use exponential mechanism to sample v from a bigger set. This, in turn, would require us to draw more samples n, to ensure that the score function is well-behaved for a bigger set of vectors, because the probability of $\langle v, \mu - x_{t-1}\rangle \ge 0.9\|\mu - x_{t-1}\|$ for uniformly-random v would be $\ll d^{-k}$.

Of course, the major drawback of the above is that it is not clear how to sample from the necessary distribution of vs efficiently – in fact, doing so would violate the lower bound of Corollary 1.1. There is also a second drawback: to evaluate the score function given above, we would need to know $\|\mu - x_{t-1}\|$; however, we are able to adapt the strategy of [HKM21] for this task to the sparse setting, re-using several of the ideas below.

From exponential to polynomial time with SoS exponential mechanism: The SoS exponential mechanism allows potentially exponential-time instances of the exponential mechanism to be converted into polynomial time algorithms, so long as (a) the bounded-differences and utility properties of the score function can be proved in a certain restricted proof system (the *SoS proof system*), and (b) the set over which the exponential mechanism is run is convex.

Convexity: The $2k$-sparse unit vectors – used by the above exponential-time algorithm – do not form a convex set. A natural idea is to relax from the $2k$-sparse vectors to the (scaled) ℓ_1 ball. This creates a substantial difficulty: the set $\left\{v : \|v\|_1 \le O\left(\sqrt{k}\right)\right\}$ has much more volume near the origin than the set of $2k$-sparse unit vectors. In particular, it is no longer true that $\langle v, \mu - x_{t-1}\rangle \ge 0.9\|\mu - x_{t-1}\|$ with probability $d^{-O(k)}$ for uniformly-random v; this probability will be exponentially small in d.

To fix this, we "fatten" the ℓ_1 ball: we use the exponential mechanism over the set $C = \left\{v : \exists w \text{ s.t. } \|v - w\| \le 0.01, \|w\| \le 1, \|w\|_1 \le O\left(\sqrt{k}\right)\right\}$. While perhaps counterintuitive that *adding* vectors is helpful here, it is possible (Lemma E.5) to show roughly the following statement: over randomly-chosen v in this set, $\Pr(\langle v, \mu - x_{t-1}\rangle \ge 0.9\|\mu - x_{t-1}\|) \ge e^{-O(d/\sqrt{k})} \cdot d^{-O(k^2)} = d^{-O(k^2)}$ for $k \ge d^{0.4}$. To see this, note that all w with $\|w - \frac{\mu - x_{t-1}}{\mu - x_{t-1}}\| \le 0.01$ are in C and have $\langle w, \mu - x_{t-1}\rangle \ge 0.0\|\mu - x_{t-1}\|$, and, using Sudakov minoration, C can be covered by $d^{O(k^2)}$ ℓ_2-balls of radius $1/\sqrt{k} + 0.01$. So at least a $d^{-O(k^2)}\left(\frac{0.01}{0.01 + 1/\sqrt{k}}\right)^d = d^{-O(k^2)} \exp(-O(d/\sqrt{k}))$-fraction of the volume of C lies on such "good" ws. So, as long as $n \gg (k^2 \log d + \log(1/\beta))/\varepsilon$ and $k \ge d^{0.4}$, the arguments we used for the exponential-time method will still work.

Utility and Bounded-Differences in SoS: Finally, we turn to (a), the need to capture the proofs of the bounded-differences and utility in the SoS proof system. Here we take "intuitively simple" as a proxy for "expressible in SoS," deferring technical definitions to the supplement. The bounded-differences property for sum-of-threshold based score functions like $s_{\mathcal{X}}$ above has a standard SoS proof [Hop20], so we won't address it further here.

The concentration of measure arguments we referred to above to establish utility are not captured by the SoS proof system! So we actually need to devise a new proof of utility for the score function $s_{\mathcal{X}}$. The key step is a proof that if $n \gg (\log d)^{O(1)}(k^2+\log(1/\beta))$ then with probability $1-\beta$, for every unit vector w with $\|w\|_1 \le \sqrt{k}$, we have $\sum_{i\le n} \mathbf{1}(\langle X_i-\mu, w\rangle \gg 1) \le 0.1n$. To see that this implies utility, note that if v has $s_{\mathcal{X}}(v) \ge 0.9n$, then there exists some i such that $\langle v, X_i - x_{t-1}\rangle \ge \|\mu - x_{t-1}\| - O(1)$ but $\langle v, X_i-\mu\rangle \le O(1)$. Then $\langle v, \mu - x_{t-1}\rangle = \langle v, X_i - x_{t-1}\rangle + \langle v, \mu - X_i\rangle \ge \|\mu - x_{t-1}\| - O(1)$.

By a standard bounded-differences argument, it suffices to prove that $\mathbb{E}\max_w \sum_{i\le n} \mathbf{1}(\langle X_i-\mu, w\rangle \gg 1) \le 0.01n$. Letting $M = \sum_{i\le n}(X_i-\mu)(X_i-\mu)^\top$, by Cauchy-Schwarz, the quantity on the left-hand side is at most $0.01\sqrt{n}\cdot(\mathbb{E}\max_w \langle M, ww^\top\rangle)^{1/2}$. Splitting $M = M_{\text{diag}} + M_{\text{off-diag}}$, and applying Holder's inequality, $\langle M, ww^\top\rangle \le \|M_{\text{diag}}\|_\infty \|w\|_2^2 + \|M_{\text{off-diag}}\|_\infty \|w\|_1^2$, where $\|\cdot\|_\infty$ is the entry-wise ℓ_∞ norm. The important term here turns out to be the second one; it is possible to bound $\mathbb{E}\|M_{\text{off-diag}}\|_\infty \le (\log d)^{O(1)}\sqrt{n}$, for an overall bound of $0.01n^{3/4}k$ (up to log factors), which gives the desired bound if $n \gg k^2$. See Lemma E.3 and E.4 for the formal version of this argument.

4 Fast Algorithms and Experiments

In this section, we describe a practical algorithm — Algorithm 1 (`Threshold` in Table 1). We use standard DP tools to guarantee privacy and utility — the exponential mechanism [MT07] with a sensitivity-1 score function in a coordinate-wise fashion, followed by a black-box application of a univariate mean estimator.

Algorithm 1 The subroutine `exp-mech` refers to the exponential mechanism [MT07], and `KV-1D` — to the univariate sparse mean estimator of [KV17].

Input: $\{x_i\}_{i\le n}$, T, b, σ^2, n; $\hat{\mu} \leftarrow \vec{0}$, selected $\leftarrow$ []

1: $m_j \leftarrow \frac{1}{b}\sum_{j=i\cdot b}^{(i+1)b} x_i$ for $j \in [\lceil n/b \rceil]$ ▷ *compute bucketed means*
2: $z_i \leftarrow \sum_{j=1}^{\lceil n/b\rceil} \mathbb{1}\{(m_j)_i \ge T\}$ for $i \in [d]$. ▷ *coordinate-wise threshold*
3: **for** $j \le k$ **do**
4: $\quad t \leftarrow$ `exp-mech`$\big(\{z_i\}_{i\in[d]\setminus\text{selected}}\big)$
5: $\quad$ selected.insert(t)
6: $\quad \hat{\mu}_t \leftarrow$ `KV-1D`$\big(\{(m_i)_t\}_{i=1}^d\big)$ ▷ *univariate estimation*

Output: mean estimate $\hat{\mu}$

Theorem 4.1. *Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Algorithm 1 is ε-DP outputs an estimate $\hat{\mu}$ s.t. $\|\hat{\mu} - \mu\|_2 \le \alpha$ with probability at least $1-\beta$ as long as*

$$n = \Omega\left(\underbrace{\frac{k^2\sigma^2(\log d + \log(1/\beta))}{\alpha^2\varepsilon}}_{\textit{support estimation}} + \underbrace{\frac{\sigma^2 k\log(2k/\beta)}{\alpha^2} + \frac{\sigma k^{1.5}\log(2k/\beta)}{\alpha\varepsilon} + \frac{k\log(R)}{\varepsilon}}_{\textit{dense estimation in } k \textit{ dimensions}}\right) \tag{2}$$
$$= \Omega\left(\sigma^2\frac{k^2(\log d + \log(k/\beta))}{\alpha^2\varepsilon} + \frac{k\log(R)}{\varepsilon}\right).$$

The proof is given in Appendix F. While the statement and proof are given for Gaussians, we only use the concentration properties of the Gaussian distribution; hence the proof automatically applies for sub-Gaussian distributions as well.

Instead of the exponential mechanism, we could also use the peeling [DSZ18] algorithm, as is done for the linear-in-R algorithm in [CWZ21].

4.1 Experimental details

Now we turn to empirically validating the performance of Algorithm 1. To isolate the effects of each subroutine (support estimation and dense mean estimation), we plot (1) the performance of the corresponding support estimation steps alone (2) ℓ_2 error for both.

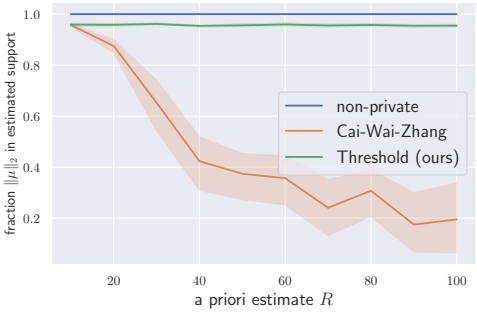

Figure 1: Empirical evaluation of 0.5-DP algorithms (and a non-private baseline) for support estimation for 1500 samples from $\mathcal{N}(\mu, I)$ in ambient dimension $d = 1000$ with $\|\mu\|_0 = 20$ (non-zero coordinates sampled uniformly from $[-10, 10]$) as a function of R, the a priori estimate of $\|\mu\|_2$.

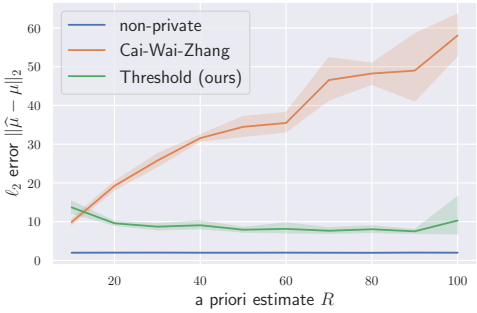

Figure 2: Empirical evaluation of `Threshold` and the sparse mean estimation algorithm of [CWZ21] under ε-DP with $\varepsilon = 0.5$; results are shown for Gaussian data $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 4 \cdot I)$ for a k-sparse μ in $\mathbb{R}^d$ for $k = 20, d = 1000, n = 1000$; non-sparse coordinates of μ are sampled uniformly from the interval $[-10, 10]$. The ℓ_2 error of the estimates is plotted against the a priori mean estimate R. A non-private baseline is also presented to highlight the cost of privacy.

We address (1) in Fig. 1 — we use the fraction (ℓ_2) mass of μ on the k coordinates the algorithms select as a metric of success, since it is well-suited for mean estimation. Fig. 1 shows that (a) we significantly outperform the previous state of the art as soon as we introduce very mild uncertainty in the a priori estimate of $\|\mu\|$; (b) our method does not introduce additional constant factors "hiding" in the asymptotics.

In Fig. 1 we presented results for support estimation, in order to highlight the improvement coming from coordinate selection alone. In Fig. 2 we evaluate Algorithm 1 directly in terms of our metric of interest – ℓ_2 error.

As can be seen from Fig. 2, the performance of the sparse mean estimation algorithm of [CWZ21] degrades rapidly even for very mild levels of uncertainty in the range of the mean - e.g. if we only know beforehand that the mean lies in the range $[-20, 20]$, instead of the tight range $[-10, 10]$, the ℓ_2 error (expectedly) doubles when we use the [CWZ21] algorithm; in contrast, the performance of Algorithm 1 is effectively unchanged.

Finally, we highlight a weakness of Algorithm 1 – we gain the mild dependence on R at the cost of losing the $\left(\frac{1}{\alpha\varepsilon} + \frac{1}{\alpha^2}\right)$-like sub-Gaussian rate; instead we have heavy-tailed-style $\left(\frac{1}{\alpha^2\varepsilon}\right)$-like rate. This comes with a practical cost in scenarios where the standard deviation σ is much larger than the a priori mean estimate R.

For all figures, we average results over 10 random seeds and report average results, together with 95% bootstrap confidence intervals. Code necessary to reproduce all plots is available[§]. For all experiments we use commodity hardware (CPU: Intel® Core™ i7-9750H CPU @ 2.60GHz).

5 Acknowledgements

Work supported in part by the NSF grants CCF-1553428 and CNS-1815221. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract

[§] https://github.com/kristian-georgiev/privacy-induces-robustness

No. HR001120C0015.

References

[AACM+22] John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Daniel Kifer, Philip Leclerc, Jeffrey Ocker, Michael Ratcliffe, and Pavel Zhuravlev. Geographic spines in the 2020 census disclosure avoidance system topdown algorithm. *arXiv preprint arXiv:2203.16654*, 2022. 1

[ADK+19] Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Differentially private covariance estimation. *Advances in Neural Information Processing Systems*, 32, 2019. 17

[AV19] Mehmet Eren Ahsen and Mathukumalli Vidyasagar. An approach to one-bit compressed sensing based on probably approximately correct learning theory. 2019. 19

[BB20] Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020. 2, 3, 7

[BDLS17] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212. PMLR, 2017. 5

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000. 17

[BGS+21] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34, 2021. 5, 6

[BIPW10] Khanh Do Ba, Piotr Indyk, Eric Price, and David P Woodruff. Lower bounds for sparse recovery. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1190–1197. SIAM, 2010. 24

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Steven Z Wu. Private hypothesis selection. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 6, 7, 16, 18, 19

[BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014. 15

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016. 5, 6, 34, 35

[Bun20] Mark Bun. A computational separation between private learning and online learning. *Advances in Neural Information Processing Systems*, 33:20732–20743, 2020. 2

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR, 2019. 7

[CM13] T Tony Cai and Zongming Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013. 17

[CS22] Julien Chhor and Flore Sentenac. Robust estimation of discrete distributions under local differential privacy. *arXiv preprint arXiv:2202.06825*, 2022. 2

[CWZ21] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021. 3, 4, 5, 9, 10, 19, 34

[De12] Anindya De. Lower bounds in differential privacy. In *Theory of cryptography conference*, pages 321–338. Springer, 2012. 5

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019. 5

[DK21] Ilias Diakonikolas and Daniel M Kane. The sample complexity of robust covariance testing. In *Conference on Learning Theory*, pages 1511–1521. PMLR, 2021. 5, 17

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017. 5, 16

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009. 2, 5

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006. 1, 15

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018. 19

[DSZ18] Cynthia Dwork, Weijie J Su, and Li Zhang. Differentially private false discovery rate control. *arXiv preprint arXiv:1807.04209*, 2018. 9

[GHRU13] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. *SIAM Journal on Computing*, 42(4):1494–1520, 2013. 2

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM (JACM)*, 48(4):798–859, 2001. 16, 17

[HBK22] Jamie Hayes, Borja Balle, and M Pawan Kumar. Learning to be adversarially robust and differentially private. *arXiv preprint arXiv:2201.02265*, 2022. 2

[HKM21] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. *arXiv preprint arXiv:2111.12981*, 2021. 3, 4, 5, 7, 8, 15, 24, 27, 28, 29, 30, 31

[HLZ20] Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912, 2020. 23

[Hop20] Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020. 9

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008. 5

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714, 2010. 5

[Hub65] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965. 1

[Hub11] Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011. 1

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011. 5, 16

[KMV21] Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. *arXiv preprint arXiv:2112.03548*, 2021. 2

[KSS94] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994. 16

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pages 2204–2235. PMLR, 2020. 24

[KV17] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017. 9, 34

[LKKO21] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[LKO21] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. *arXiv preprint arXiv:2111.06578*, 2021. 2, 5, 6

[LM19] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019. 7, 22

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991. 33

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007. 9, 15, 20

[MV21] Oren Mangoubi and Nisheeth K Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees and applications to differentially private optimization. *arXiv preprint arXiv:2111.04089*, 2021. 31

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007. 5

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3389–3423. World Scientific, 2018. 5

[SU15] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *arXiv preprint arXiv:1501.06095*, 2015. 5

[TS13] Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013. 5

[Tuk75] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975. 1

[UKRK22] Dmitrii Usynin, Helena Klause, Daniel Rueckert, and Georgios Kaissis. Can collaborative learning be private, robust and scalable?, 2022. 2

[Ull16] Jonathan Ullman. Answering n^2+o(1) counting queries with differential privacy is hard. *SIAM Journal on Computing*, 45(2):473–496, 2016. 2, 5

[UV11] Jonathan Ullman and Salil Vadhan. Pcps and the hardness of generating private synthetic data. In *Theory of Cryptography Conference*, pages 400–416. Springer, 2011. 5

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not envision any direct negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Any omitted proofs can be found in the appendicies.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplement.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bars are included for all figures in Section 3 and the supplement
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] A URL to codebase, together with implementation details needed to reproduce our results, are given in the supplement
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] All experiments are done on synthetic data without relation to any individual's private data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Preliminaries

We first introduce central definitions and results from differential privacy that we use throughout the paper. Next, we give background on the notion of robustness we work with. Finally, we briefly overview the central ideas in the Sum-of-Squares (SoS) tools we use.

Differential Privacy We start by formally defining the notion of differential privacy.

Definition A.1 ((Approximate) Differential Privacy)**.** *Let $\mathcal{X}$ be a set and $\mathcal{X}^* = \{(X_1, \ldots, X_n) : n \in \mathbb{N}, X_i \in \mathcal{X}\}$ be all possible datasets over $\mathcal{X}$. For $\varepsilon, \delta > 0$, a (randomized) map $M : \mathcal{X}^* \to \mathcal{O}$ (where $\mathcal{O}$ is an* output *set) is (ε, δ)-DP if for every $(X_1, \ldots, X_n), (X'_1, \ldots, X'_n) \in \mathcal{X}^*$ such that $X_i = X'_i$ except for a single index i and for every subset $S \subseteq \mathcal{O}$, $\Pr(M(X_1, \ldots, X_n) \in S) \leq e^{\varepsilon} \Pr(M(X'_1, \ldots, X'_n) \in S) + \delta$.*

The special case of $\delta = 0$ is referred to as *pure* DP and will be the main focus of our work. Throughout the paper, we refer to datasets that differ in one entry as *neighboring*.

We freely use basic primitives in private algorithm design like composition of private mechanisms, which we state for completeness.

Lemma A.2 ((Basic) Composition [DMNS06])**.** *Suppose that we have private mechanisms $M_1, \ldots, M_n$ where M_i is ε_i-DP. Then an adaptive composition M of $\{M_1, \ldots, M_n\}$ is $\sum_{i \leq n} \varepsilon_i$-DP.*

We also freely use basic DP mechanism like the Laplace mechanism defined below.

Lemma A.3 (Laplace mechanism [DMNS06])**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ such that $\max_{X, X' \text{ neighbors}} \|f(X) - f(X')\|_1 = \Delta$. Then the mechanism $M : \mathcal{X}^n \to \mathbb{R}^d$ defined as $M(X) = f(X) + L$, where $L \sim Lap\left(0, \frac{\varepsilon}{\Delta}\right)$ is sampled from the Laplace distribution, is ε-DP.*

The quantity Δ in Lemma A.3 is referred to as the ℓ_1 sensitivity of f. A central part of our design of private mean estimation algorithms revolves around using functions with low sensitivity as primitives for our private estimators.

We make extensive use of the exponential mechanism [MT07]. It is a technique to privately select the (approximately) "best" object in a universe $\mathcal{H}$ according to a *score function* S, which measures the "goodness" of a given object. An important property of the score function is its *sensitivity* Δ_S, defined as $\max_{h \in \mathcal{H}, \text{neighboring } D, D'} |S(D, h) - S(D', h)|$. Given a score function S with sensitivity Δ_S and a privacy parameter ε, the exponential mechanism samples an object h with probability proportional to $\exp(\varepsilon/(2\Delta_S) \cdot S(h))$. The exponential mechanism comes with the following privacy and utility guarantees.

Theorem A.4 ([MT07])**.** *For a dataset X and a score function $S : \mathcal{X}^n \times \mathcal{H} \to \mathbb{R}$, the exponential mechanism M on the score function S is ε-DP and with probability at least $1 - \beta$ outputs an object such that*

$$S(M(X)) \geq OPT(X) - \frac{2\Delta_S}{\varepsilon}\left(\log\left(\frac{|\mathcal{H}|}{|\mathcal{H}^*|} + \log(1/\beta)\right)\right),$$

where $\mathcal{H}^*$ is the set of objects achieving score OPT.

Robustness There are a variety of models for the adversary in the agnostic setting. We work with the case of *adaptive* adversaries:

Definition A.5 (η-contamination model)**.** *In the η-contamination model, given a "clean" distribution D, to draw n (η-contaminated) samples from D, first draw $X'_1, \ldots, X'_n \sim D$, and then output any $\{X_1, \ldots, X_n\}$ such that $X_i = X'_i$ for at least $(1 - \eta)n$ choices of i.*

Another popular model in the literature is that of an *oblivious* adversary who provides a distribution that is close in total variation distance to the original one. The name comes from the fact that, unlike in Definition A.5, the adversary is not allowed to inspect the samples and adaptively decide what samples to add and remove.

Sum of Squares In Appendix D we use the SoS exponential mechanism in a black-box fashion. For a derivation of it, as well as examples, see [HKM21]. Informally, we use SoS as a proofs-to-algorithms paradigm which automatically "simple" (expressible within the SoS proof system) proofs into polynomial time algorithms. For an overview of SoS, see e.g. [BS14].

B Lower Bounds in Private Statistics

In this section, we state and prove theorems giving information-computation gaps for private algorithms for Gaussian mean estimation and learning parities, and we give an information-theoretic lower bound for private covariance testing, again via Theorem 2.1.

B.1 Private Gaussian Mean Estimation

Consider the task of privately estimating the mean of class of spherical Gaussian distributions, and the following conjecture, supported by statistical query lower bounds [DKS17].

Conjecture B.1. *There exists $c > 0$, no poly$(n, d, 1/\eta)$-time algorithm η-robustly estimates the mean of an unknown d-dimensional spherical Gaussian $\mathcal{N}(\mu, I)$ with, $\|\mu\| \leq$ poly(d), from independent samples to ℓ_2 error $\eta(\log 1/\eta)^{1/2-\Omega(1)}$, so long as $\eta \geq 2^{-d^c}$.*

(By contrast, $O(\eta)$ error is achievable in exponential time.) We prove the following corollary:

Corollary B.2. *Assume Conjecture B.1. Every poly$(n, d, 1/\varepsilon)$-time ε-DP algorithm which takes i.i.d. samples from $\mathcal{N}(\mu, I)$ for $\mu \in \mathbb{R}^d$ and outputs a vector $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \leq \alpha$ with probability $1 - \beta$ for all $\|\mu\| \leq$ poly(d) requires $n \geq \min\{\frac{\log(1/\beta)}{\varepsilon\alpha} \cdot (\log(1/\alpha))^{1/2-o(1)}, \frac{2^{d^c}\log(1/\beta)}{\varepsilon}\}$, where $c > 0$ is a universal constant.*

To see that Corollary B.2 captures an information-computation gap, recall that in exponential time it is possible to perform ε-DP Gaussian mean estimation for $\|\mu\| \leq \mathrm{poly}(d)$ with

$$n = \Theta\left(\frac{d + \log(1/\beta)}{\alpha^2} + \frac{d + \log(1/\beta)}{\alpha\varepsilon} + \frac{d\log d + \log(1/\beta)}{\varepsilon}\right)$$

samples [BKSW19]. Unless the SQ lower bounds of [DKS17] are broken by some polynomial-time algorithm, this error rate can be matched, at best, up to logarithmic factors by polynomial time algorithms.

Proof of Corollary B.2. A polynomial-time private algorithm with the hypothesized guarantees would be $\eta = \log(1/\beta)/(\varepsilon n)$-robust by Theorem 2.1. So long as $\log(1/\beta)/\varepsilon n = \eta \geq 2^{-d^c}$ this satisfies the hypotheses of Conjecture B.1, so it can't estimate μ to error $\eta(\log(1/\eta))^{1/2-o(1)}$; the conclusion follows by substituting $\log(1/\beta)/(\varepsilon n)$ for η. □

B.2 Privately Learning Parities

We now turn to a fundamental supervised problem - PAC learning parities. In the absence of noise, Gaussian elimination provides a polynomial-time sample-efficient algorithm for learning parities.

Learning under adversarial label noise is known as agnostic learning [KSS94]. Learning parities in the agnostic case, unlike in the noiseless case, is notoriously difficult; in the proper case, the problem has been shown to be NP-hard [Hås01], even for getting accuracy of $1/2 + \varepsilon$ for any $\varepsilon > 0$. With this in mind, our goal is to probe where does private parity learning lie computationally.

More formally, let `PARITY` be the class of functions $f_S : \{\pm 1\}^d \to \{\pm 1\}$ defined by $f_S(x) = \prod_{i \in S} x_i$ for each $S \subseteq [n]$. An algorithm PAC learns `PARITY` if, for every S, given n samples $(x, f_S(x))$ where x comes from some distribution D on $\{\pm 1\}^d$, it finds T such that $\Pr_{x \sim D}(f_S(x) \neq f_T(x)) \leq \alpha$. [KLN+11] show that

$$n = O\left(\frac{d\log(1/\beta)}{\varepsilon\alpha}\right) \tag{3}$$

suffice to privately PAC learn `PARITY` in polynomial time [KLN+11, Theorem 4.4], succeeding with probability $1 - \beta$. In contrast, allowing exponential time, only

$$n = O\left(\frac{d + \log(1/\beta)}{\varepsilon\alpha}\right) \tag{4}$$

suffice. From Theorem 2.1 we obtain the result that this gap cannot be closed with polynomial-time private algorithms, unless $RP = NP$:

Corollary B.3. *Suppose $RP \neq NP$. Then every polynomial-time ε-DP algorithm which for any $\beta > 0$ can PAC-learn d-variable `PARITY`s to accuracy α, succeeding with probability $1 - \beta$, requires $n \geq \omega\left(\frac{\log(1/\beta)}{\varepsilon\alpha}\right)$ samples.*

Proof. Suppose otherwise, that there exists a private PAC learner using $n \leq O(\log(1/\beta)/(\alpha\varepsilon))$ samples. By Theorem 2.1, such an algorithm is also $\eta = \frac{\log(1/\beta)}{\varepsilon n}$-robust, with $\eta \geq \Omega(\alpha)$. We claim that it can be used to distinguish $1 - \Omega(\alpha)$-satisfiable instances of XOR-SAT from $1/2 + O(\alpha)$-satisfiable ones; by [Hås01] this is NP-hard for every constant $\alpha > 0$.

Given a d-variable instance ϕ of XOR-SAT, let D be the following distribution on the hypercube. First, draw a clause $\prod_{i \in C} y_i = b_C$ from ϕ uniformly at random. Then let $c \in \{\pm 1\}^d$ be given by $c_i = -1$ if $i \in C$ and otherwise $c_i = -1$.

Suppose ϕ is $1 - \alpha'$-satisfiable, by some $x \in \{\pm 1\}^n$. Define $f_x(c) = \prod_{i : x_i = -1} c_i$. We claim that $\Pr_{C \sim \phi}(f_x(c) = b_C) \geq 1 - \alpha'$. This holds because $f_x(c) = \prod_{i : x_i = -1} c_i = \prod_{i : c_i = -1} x_i = b_C$ if C is a clause satisfied by x, and by hypothesis a $(1 - \alpha')$-fraction of the clauses in ϕ are satisfied by x.

Now, given samples $(c, f_x(c))$, our hypothesized private PAC learner returns, with probability $1 - \beta$, some parity function f_S such that $\Pr_{C \sim \phi}(f_x(c) = f_S(c)) \geq 1 - \alpha'$. This means that $\Pr_{C \sim \phi}(f_S(c) = b_C) \geq 1 - 2\alpha'$.

By η-robustness with $\eta \geq \Omega(\alpha)$ for $\alpha' \ll \alpha$, even given samples (c, b_C), the learning algorithm returns such an f_S with probability at least 0.99, taking β sufficiently small. But given f_S it is easy to extract a $1 - O(\alpha')$-satisfying assignment to ϕ, which is NP hard. □

B.3 Private Covariance Testing

A fundamental question in high-dimensional statistics is that of distributional property testing [BFR+00]. Given independent samples, the goal is to design efficient algorithms to test whether their distribution satisfies a given property.

In this work, we are interested in a simple problem in that field - Gaussian covariance testing. In particular, given samples $X_1, \ldots, X_n$ in $\mathbb{R}^d$ from $\mathcal{N}(0, \Sigma)$ with unknown Σ, we want to *privately* determine whether $\Sigma = I$ or $\|\Sigma - I\|_F \geq \gamma$ with high probability using as few samples as possible.

In the non-private case, [CM13] show that Gaussian covariance testing can be achieved with $O\left(\frac{d}{\gamma^2}\right)$ samples, significantly less than the $O\left(\frac{d^2}{\gamma^2}\right)$ samples required to learn the distribution. Curiously, Diakonikolas and Kane [DK21] show that under η-corruptions for any constant $\eta > 0$ one needs $\Omega(d^2)$ samples for testing.

[ADK+19] show that one needs $O\left(\frac{d^2}{\gamma^2 \varepsilon}\right)$ samples to privately learn the covariance of a Gaussian in d dimensions. Thus, this is a natural upper bound on the sample complexity of *testing* the covariance.

We leverage Theorem 2.1 to show that any private algorithm with strong group privacy guarantees requires $\Omega(d^2)$ samples for covariance testing.

Corollary B.4. *For every $C > 0$ there is $C' > 0$ such that every ε-DP mechanism which takes independent samples $X_1, \ldots, X_n$ from $\mathcal{N}(0, \Sigma)$ for an unknown Σ and distinguishes with probability at least $1 - \beta$ whether $\Sigma = I$ or $\|\Sigma - I\|_F \geq 1/2$ takes $n \geq \Omega(\min(C \log(1/\beta)/\varepsilon, C' d^2))$ samples.*

Proof. [DK21] shows that for every $c > 0$ there is $c' > 0$ such that any c-robust algorithm for distinguishing $\Sigma = I$ from $\|\Sigma - I\| > 1/2$ requires at least $c' d^2$ samples. Any ε-private algorithm which accomplishes this task with probability at least $1 - \beta$ will be $\log(1/\beta)/(\varepsilon n)$-robust, by Theorem 2.1. If $n \leq C \log(1/\beta)/\varepsilon$, then this algorithm is $\Omega(1/C)$-robust. Hence, there is C' such that it must use $n \geq C' d^2$ samples.

□

C Information-theoretic Results for Private Sparse Mean Estimation

In this section we focus on algorithms and lower bounds for private sparse mean estimation in the absence of computational considerations. First, we consider the Gaussian case; then we develop algorithms for the more general heavy-tailed case where we only assume that the data has bounded covariance.

C.1 Gaussian Private Sparse Mean Estimation

Proposition C.1. *For every $\alpha, \varepsilon, \beta, R > 0$ and small-enough $\eta > 0$ there is an ε-DP, mechanism taking $n \gg \frac{k \log(d)}{\alpha^2} + \frac{\log(1/\beta) + k \log(Rk/\alpha)}{\alpha \varepsilon}$ η-corrupted samples from $\mathcal{N}(\mu, I)$ with $\mu \in \mathbb{R}^d$, $\|\mu\|_0 \leq k$, and $\|\mu\| \leq R$ and produces $\hat{\mu}$ such that $\|\mu - \hat{\mu}\| \leq \alpha + O(\eta)$, with probability $1 - \beta$.*

Proposition C.1 follows quickly from combining the private hypothesis selection mechanism of [BKSW19] with Theorem 2.1 to establish robustness. For convenience, we restate the result of [BKSW19]. Before that, we introduce the notion of a Scheffé set.

Definition C.2 (Scheffé set)**.** *Let $\mathcal{H}$ be a set of distributions on $\mathcal{X}$. The Scheffé set for $H, H' \in \mathcal{H}$ is defined as $w(H, H') = \{x \in \mathcal{X} \mid H(x) > H'(x)\}$. The set $\mathcal{W}$ of Scheffé sets for $\mathcal{H}$ is given by $\mathcal{W}(\mathcal{H}) = \{w(H, H') \mid H, H' \in \mathcal{H}\}$.*

Theorem C.3 ([BKSW19, Theorem 4.1])**.** *Suppose $\mathcal{H}$ is a set of distributions on $\mathcal{X}$. Let d be the VC dimension of the set of indicators of the Scheffé sets of $\mathcal{H}$. Then there exists an (ε, δ)-differentially private mechanism with the following guarantees. Suppose $D = \{X_1, \ldots, X_n\}$ is a set of private samples independently drawn from an unknown distribution P and suppose there exists some H^* such that $d_{\mathrm{TV}}(P, H^*) \leq \alpha$. If $n = \Omega\left(\frac{d + \log(1/\beta)}{\alpha^2} + \frac{\log(|\mathcal{H}|/\beta) + \min\{\log(|\mathcal{H}|), \log(1/\delta)\}}{\alpha \varepsilon}\right)$, then the output $\hat{H}$ of the algorithm is such that $d_{\mathrm{TV}}(P, \hat{H}) \leq 7\alpha$ with probability at least $1 - \beta$.*

Intuitively, Theorem C.3 tells us that if we have a hypothesis class of candidate distributions with a small cover, an example of which are Gaussians with bounded mean and known covariance, then we can privately select a distribution close to an unknown target distribution in total variation distance with only a mild additional cost of privacy. To make this more precise, we need the following definition.

Definition C.4. *A set $\mathcal{C}_\alpha$ of distributions is an α-cover for a set of distributions $\mathcal{H}$ if for every $H \in \mathcal{H}$, there is some $C \in \mathcal{C}_\alpha$ such that $d_{\mathrm{TV}}(H, C) \leq \alpha$.*

In order to apply Theorem C.3 to our use case, we first show a cover for Gaussian distribution with k-sparse bounded means.

Lemma C.5. *Let $\mathcal{S}$ be the set of Gaussian distributions $\mathcal{N}(\mu, I)$ in d dimensions with k-sparse μ such that $\|\mu\|_2 \leq R$. Then $\mathcal{S}$ admits an α-cover of size*

$$O\left(\binom{d}{k}\left(\frac{Rk}{\alpha}\right)^k\right).$$

Proof. First, observe that it suffices to obtain an α-cover for each of the $\binom{d}{k}$ possible choices of the support of μ, and return the union of the covers. For any particular subset, the problem becomes equivalent to finding an α-cover of the set of non-sparse Gaussian distributions with bounded mean in k dimensions. By Lemma 6.8 in [BKSW19], there exists an α-cover of size $O\left(\frac{Rk}{\alpha}\right)$. For completeness, we note that the proof of Lemma 6.8 in [BKSW19] constructs the α-cover by taking the Cartesian product of (α/k)-covers in each standard basis direction, which by the triangle inequality gives an α-cover. □

Covering numbers are tightly related to VC dimension. The final piece we need for Proposition C.1 is a bound on the VC dimension of the Scheffé sets of the set of k-sparse d-dimensional Gaussians with identity covariance.

Lemma C.6. *Let $\mathcal{H}_k^d$ be the set of Gaussian distributions with covariance matrix I_d and mean $\mu \in \mathbb{R}^d$ such that $\|\mu\|_0 \leq k$. Let $\mathcal{W}\left(\mathcal{H}_k^d\right)$ be the set of indicators of Scheffé sets of $\mathcal{H}_k^d$. Then $\mathcal{W}$ has VC dimension at most $4k \log(de) = O(k \log(d))$.*

Proof. We want to bound the VC dimension of the set of functions $f_{H,H'} : \mathcal{X} \to \{0, 1\}$ defined by $f_{H,H'}(x) \mapsto \mathbb{1}\{H(x) > H'(x)\}$ for $H, H' \in \mathcal{H}_k^d$.

First, observe that for any two Gaussians H, H', $f_{H,H'}$ correspond to a linear threshold function $L(x) = \left\{a \in \mathbb{R}^d \mid \langle a, x\rangle \geq 0\right\}$. Restricting H, H' to belong to $\mathcal{H}_k^d$, i.e. have k-sparse means, we now have that each $f_{H,H'}$ corresponds to a linear threshold function $L(x)$ generated by $2k$-sparse vector x. We now leverage a result of Ahsen and Vidyasagar [AV19, Theorem 6] which states that the set of linear threshold functions generated by k-sparse vectors in d dimensions has VC dimension at most $2k\log(de)$; this concludes the argument. □

We are now ready to prove Proposition C.1.

Proof of Proposition C.1. Consider the case of uncorrupted data ($\eta = 0$). We construct an estimator that gives a guarantee in ℓ_2 distance, which we show is equivalent to total variation distance for spherical Gaussians.

From Lemma C.5 and a bound on the binomial coefficients we obtain a cover of size

$$O\left((de/k)^k \left(\frac{Rk}{\alpha}\right)^k\right). \tag{5}$$

From Lemma C.6, we know that the VC dimension of the set of indicators of the Scheffé sets of k-sparse Gaussians with identity covariance in d dimensions is $O(k\log(d))$.

With the size of the α-cover and the bound on the VC dimension established, we are ready to invoke Theorem C.3. This gives us the guarantee that for the output distribution $\hat{H}$ we have $d_{\mathrm{TV}}(P, \hat{H}) \leq 7\alpha$ with probability at least $1-\beta$. Suppose the selected distribution $\hat{H}$ has mean $\hat{\mu}$. Since both P and $\hat{H}$ are spherical Gaussians, we have that

$$\|\mu - \hat{\mu}\|_2 \leq d_{\mathrm{TV}}(P, \hat{H}) \tag{6}$$

following from [DMR18, Theorem 1.2]. Therefore, we can obtain an estimate with the desired properties by outputting the mean of the selected distribution $\hat{H}$.

Finally, robustness follows directly from Theorem 2.1. □

As for an information-theoretic lower bound, [CWZ21] give a tracing-based lower bound (which also holds for approximate DP) which states that any ε-DP algorithm must use at least $n = \Omega\left(\frac{k\log(d)}{\alpha^2} + \frac{k\log(d)}{\alpha\varepsilon}\right)$ samples to estimate the mean of a sub-Gaussian distribution up to ℓ_2 error α, assuming $\|\mu\|_\infty \leq 1$. Alternatively, we can leverage hypothesis-selection based arguments again, to get a $\Omega\left(\frac{k\log(d)+\log(Rk/\alpha)}{\varepsilon}\right)$ bound, assuming $\|\mu\|_2 \leq R$.

C.2 Heavy-Tailed Private Sparse Mean Estimation

Now we turn to a more general set of distributions – ones with bounded covariance. We give an (exponential-time) algorithm – Algorithm 2, which has strong group privacy properties allowing us to invoke Theorem 2.1.

Theorem C.7. *For every $\alpha, \varepsilon, \beta, R > 0$ and small-enough $\eta > 0$ Algorithm 2 is an ε-DP mechanism with the following properties. Given n η-corrupted samples from any distribution D with mean $\mu \in \mathbb{R}^d$ having $\|\mu\|_0 \leq k, \|\mu\| \leq R$, and covariance $\Sigma \preceq I$, it produces an estimate $\hat{\mu}$ such that $\|\mu - \hat{\mu}\| \leq \alpha + O(\sqrt{\eta})$, with probability $1-\beta$, as long as*

$$n = \Omega\left(\underbrace{\frac{k\log(d) + \log(1/\beta)}{\alpha^2\varepsilon}}_{\text{support estimation}} + \underbrace{\frac{k\log(k) + \log(1/\beta)}{\alpha^2\varepsilon} + \frac{k\log(R)}{\varepsilon}}_{\text{dense estimation in } k \text{ dimensions}}\right).$$

To prove Theorem C.7 we can no longer rely on the hypothesis selection mechanism of [BKSW19], as the hypothesis class of distributions with bounded second moments doesn't have a finite cover in

Algorithm 2 Exponential-time algorithm for private sparse heavy-tailed mean estimation

Input: bucket size b, coordinate-wise variance σ^2, number of samples n

`// Support estimation`

$m_j \leftarrow \frac{1}{b}\sum_{j=i\cdot b}^{(i+1)b} x_i$ for all $j \in [\lceil n/b \rceil]$ ▷ *compute bucketed means*

$s_T \leftarrow S(\{m_i\}_{i\le\lceil n/b\rceil}, T, L)$ for all $T \subseteq [d]$ with $|T| = k$ ▷ *compute scores*

Run the exponential mechanism [MT07] on $\{s_T\}_{T:T\subseteq[d],|T|=k}$ to get candidate support T

`// Coarse dense mean estimation (in ambient dimension` k`)`

$S_{coarse}(\{x_{iT}\}_{i\le n}, x, L) = \sum_{i\le n} \mathbb{1}\{\|x_{iT} - x\| \ge L\}$

$s_x \leftarrow S_{coarse}(\{x_{iT}\}_{i\le n}, x, L)$ for all x in a $\sqrt{k}$-packing of the ℓ_2 ball of radius R

Run exponential mechanism on $\{s_x\}$ to get $\hat{\mu}_{coarse}$

`// Fine dense mean estimation (in ambient dimension` k`)`

$S_{fine}(\{x_{iT}\}_{i\le n}, x, L; \hat{\mu}_{coarse}) = \max_{v:\|v\|=1} \sum_{i\le n} \mathbb{1}\{\langle x_{iT} - x, v\rangle \ge L\}$

$s_x \leftarrow S_{fine}(\{m_{iT}\}_{i\le\lceil n/b\rceil}, x, L)$ for all x in a 1-packing of the ℓ_2 ball of radius $\sqrt{k}$ centered at $\hat{\mu}_{coarse}$

Run exponential mechanism on $\{s_x\}$ to get $\hat{\mu}$

$\hat{\mu}_i \leftarrow \begin{cases} (\hat{\mu}_T)_i & \text{if } i \in T \\ 0 & \text{o.w.} \end{cases}$

Output: mean estimate $\hat{\mu}$

total variation distance. Instead, we design a new exponential-mechanism-based approach to identify the nonzero coordinates of μ, inspired by quantile-based score functions used in robust statistics.

We show that the exponential mechanism is going to select a subset $\hat{T}$ of coordinates such that $\|\mu_{\hat{T}}\|$ is close to $\|\mu\|$ with high probability. We are guaranteed that the exponential mechanism will select a candidate with score

$$\mathrm{OPT} - 2\frac{\Delta^{(S)}}{\varepsilon}\log\left(|\mathcal{H}|/\beta\right)) \tag{7}$$

with probability at least β. We first show that the sensitivity of the score function S is 1, which directly affects the guarantee in Eq. (7).

For convenience, we restate the definition of S below:

$$S(\{x_i\}_{i\le n}, T; L) = \max_{v\in\mathbb{R}^k, \|v\|_2=1} \sum_{i=1}^{n} \mathbb{1}\{v^\top x_{iT} \ge L\}. \tag{8}$$

To ease notation, we will omit the $\{x_i\}_{i\le n}$ argument whenever it is clear from context.

Lemma C.8. *The score function $S(\cdot,\cdot;L)$ has sensitivity $\Delta^{(S)} = 1$ for all values of L.*

Proof. Changing a single element x_i in the dataset affects at most one of the indicators in the sum, thus changing S by at most 1. □

Next, we show that the ground-truth subset has a high score.

Lemma C.9. *Let $x_1, \ldots, x_n \sim P$ be independent samples and suppose that P has a k-sparse mean μ such that $\|\mu\|_2 \le R$, and $\|\Sigma\|_{op} \le \lambda$. Then for the ground-truth subset T^* and some $c > 0$ we have $S(\{x_i\}_{i\le n}, T^*; \|\mu\| - c) \ge n - \frac{n\lambda}{c^2} - \sqrt{n\log(3/\beta)/2}$ with probability at least $1 - \beta/3$.*

Proof. Since T^* is the optimal subset, we have that $\|\mu_{T^*}\| = \|\mu\|$. Using $v = \frac{\mu_{T^*}}{\|\mu_{T^*}\|}$ as a representative, we get

$$\begin{aligned} \mathbb{E}S(T^*; L) &\ge \mathbb{E}\sum_{i=1}^{n} \mathbb{1}\{\mu_{T^*}^\top x_{iT^*} \ge \|\mu_{T^*}\|L\} \\ &= \sum_{i=1}^{n} \mathbb{P}\left(\mu_{T^*}^\top x_{iT^*} \ge \|\mu\|L\right). \end{aligned} \tag{9}$$

Now we bound $\mathrm{Var}(\mu_{T^*}^\top x_{iT^*})$. Let Σ_{T^*} be the covariance matrix restricted to the coordinates in T^*. By Cauchy's interlacing theorem, we have that $\|\Sigma_{T^*}\|_{op} \leq \lambda$. Thus,

$$\mathrm{Var}(\mu_{T^*}^\top x_{iT^*}) \leq \|\mu_{T^*}\|^2 \lambda = \|\mu\|^2 \lambda. \tag{10}$$

This allows us to use Chebyshev's inequality on each term of the sum to get

$$\begin{aligned}
\mathbb{P}\left(\mu_{T^*}^\top x_{iT^*} \geq \|\mu\| L\right) &= 1 - \mathbb{P}\left(\mu_{T^*}^\top x_{iT^*} - \|\mu\|^2 < \|\mu\| L - \|\mu\|^2\right) \\
&\geq 1 - \frac{\|\mu\|^2 \lambda}{\left(\|\mu\| L - \|\mu\|^2\right)^2} \\
&= 1 - \frac{\lambda}{(L - \|\mu\|)^2} = 1 - \frac{\lambda}{c^2}
\end{aligned} \tag{11}$$

as long as $L < \|\mu\|$, which is guaranteed from the condition that $c > 0$. Thus, in total, we have

$$\mathbb{E} S(T^*; \|\mu\| - c) \geq n - \frac{n\lambda}{c^2}. \tag{12}$$

Finally, from McDiarmid's inequality, we have that

$$\mathbb{P}\left(S(T; L) - \mathbb{E} S(T; L) \geq t\right) \leq \exp\left(-\frac{2t^2}{n}\right), \tag{13}$$

which concludes the proof. □

In addition to the ground-truth subset of coordinates having a large score, we want "bad" subsets to have a low score.

Lemma C.10. *Let $x_1, \ldots, x_n \sim P$ be independent samples and suppose that P has a k-sparse mean μ such that $\|\mu\|_2 \leq R$, and $\|\Sigma\|_{op} \leq \lambda$. Let $T \subseteq [d]$. Then for any L such that $\|\mu_T\| < L$ we have that $S(T; L) \leq n \frac{\sqrt{\lambda}}{L - \|\mu_T\|} + \frac{2}{L}\sqrt{nk\lambda} + \sqrt{n \log(3/\beta)}$ with probability at least $1 - \beta/3$.*

Proof. First, note that S has bounded differences with respect to the sequence of random vectors $X = \{X_1, \ldots, X_n\}$. In particular, from Lemma C.8 we have that $|S(X; L) - S(X'; L)| \leq 1$ for any X, X' that differ in one entry. Thus, we can apply McDiarmid's inequality to obtain

$$\mathbb{P}\left(S(T; L) - \mathbb{E} S(T; L) \geq t\right) \leq \exp\left(-\frac{2t^2}{n}\right) \tag{14}$$

for all L, T, and t. The rest of the proof is devoted to bounding $\mathbb{E} S(T; L)$. To ease notation, for $X = \{x_1, \ldots, x_n\}$ define

$$f_{T,L,v}(X) = \sum_{i=1}^{n} \mathbb{1}\{v^\top x_{iT} \leq L\}. \tag{15}$$

We start with

$$\begin{aligned}
&\mathbb{E} S(T; L) = \mathbb{E}_X \max_{v:\, \|v\|=1} f_{T,L,v}(X) \\
&\leq \mathbb{E}_X \left[\max_{v:\, \|v\|=1} \left(f_{T,L,v}(X) - \mathbb{E} f_{T,L,v}(X)\right) + \max_{v:\, \|v\|=1} \mathbb{E} f_{T,L,v}(X)\right] \\
&= \mathbb{E}_X \max_{v:\, \|v\|=1} \left(f_{T,L,v}(X) - \mathbb{E} f_{T,L,v}(X)\right) + \max_{v:\, \|v\|=1} \mathbb{E} f_{T,L,v}(X) \\
&= \mathbb{E}_X \max_{v:\, \|v\|=1} \left(f_{T,L,v}(X) - \mathbb{E} f_{T,L,v}(X)\right) + n \max_{v:\, \|v\|=1} \mathbb{P}_X\left(v^\top x_T \leq L\right).
\end{aligned} \tag{16}$$

We follow with a symmetrization argument for the first term. Let $\{\sigma_i\}_{i=1}^n$ be independent Rademacher random variables. Then

$$\begin{aligned}
&\mathbb{E}_X \max_{v:\,\|v\|=1} (f_{T,L,v}(X) - \mathbb{E} f_{T,L,v}(X)) \\
&\leq \mathbb{E}_{X,X',\sigma_i} \max_{v:\,\|v\|=1} (\sigma_i (f_{T,L,v}(X) - f_{T,L,v}(X'))) \\
&\leq \mathbb{E}_{X,X',\sigma_i} \max_{v:\,\|v\|=1} \left(\sum_{i\leq n} \sigma_i \left(\mathbb{1}\{v^\top x_{iT} \leq L\} - \mathbb{1}\{v^\top x'_{iT} \leq L\} \right) \right) \\
&\leq 2\mathbb{E}_{X,\sigma_i} \max_{v:\,\|v\|=1} \left(\sum_{i\leq n} \sigma_i \mathbb{1}\{v^\top x_{iT} \leq L\} \right),
\end{aligned} \tag{17}$$

via the triangle inequality. Now we can use the inequality $\mathbb{1}\{\eta^\top v \leq L\} \leq \frac{|\eta^\top v|}{L}$ and Ledoux-Talagrand contraction [LM19] to get

$$\begin{aligned}
&2\mathbb{E}_{X,\sigma_i} \max_{v:\,\|v\|=1} \left(\sum_{i\leq n} \sigma_i \mathbb{1}\{v^\top x_{iT} \leq L\} \right) \\
&\leq 2\mathbb{E}_{X,\sigma_i} \max_{v:\,\|v\|=1} \sum_{i\leq n} \sigma_i \frac{|v^\top x_{iT}|}{L} \\
&\leq \frac{2}{L} \mathbb{E}_{X,\sigma_i} \max_{v:\,\|v\|=1} \sum_{i\leq n} \sigma_i \cdot v^\top x_{iT} \\
&= \frac{2}{L} \mathbb{E}_{X,\sigma_i} \max_{v:\,\|v\|=1} v^\top \left(\sum_{i\leq n} \sigma_i \cdot x_{iT} \right) \\
&= \frac{2}{L} \mathbb{E}_{X,\sigma_i} \left\| \sum_{i\leq n} \sigma_i \cdot x_{iT} \right\| \\
&\leq \frac{2}{L} \sqrt{n \mathrm{Tr}(\Sigma_T)},
\end{aligned} \tag{18}$$

where Σ_T is the restriction of the covariance matrix Σ to the coordinates of the subset $T \subseteq [d]$. By the Cauchy interlacing theorem we have that $\|\Sigma_T\|_{op} \leq \|\Sigma_T\|$ and thus overall we have

$$\mathbb{E}_X \max_{v:\,\|v\|=1} (f_{T,L,v}(X) - \mathbb{E} f_{T,L,v}(X)) \leq \frac{2}{L} \sqrt{nk\lambda}. \tag{19}$$

For the second term we have

$$\begin{aligned}
&n \max_{v:\,\|v\|=1} \mathbb{P}_X(v^\top x_T \leq L) \\
&= n \max_{v:\,\|v\|=1} \mathbb{P}_X(v^\top x_T - \|\mu_T\| \leq L - \|\mu_T\|) \\
&\leq n \max_{v:\,\|v\|=1} \mathbb{P}_X(v^\top x_T - v^\top \mu_T \leq L - \|\mu_T\|) \\
&= n \max_{v:\,\|v\|=1} \mathbb{E}_X \left[\mathbb{1}\left\{ (x_T - \mu_T)^\top v \leq L - \|\mu_T\| \right\} \right] \\
&\leq n \max_{v:\,\|v\|=1} \mathbb{E}_X \left[\frac{|(x_T - \mu_T)^\top v|}{L - \|\mu_T\|} \right] \leq \frac{n\|\Sigma\|^{1/2}}{L - \|\mu_T\|}.
\end{aligned} \tag{20}$$

□

Finally, we show that the dense (fine and coarse) private mean estimation algorithms satisfy the properties desired by Algorithm 2 in Lemmas C.11 and C.12.

Lemma C.11 (Coarse dense estimation)**.** *Let $x_1, \ldots, x_n \sim P$ be independent samples and suppose P has covariance $\Sigma \preceq I$ and mean $\mu \in \mathbb{R}^k$ such that $\|\mu\|_2 \leq R$. Then running exponential*

mechanism on $\{x_i\}_{i\le n}$ *with the score function* $S_{coarse}(\{x_i\}_{i\le n}, x, L) = \sum_{i\le n} \mathbb{1}\{\|x_i - x\| \ge L\}$ *gives us an estimate* $\hat{\mu}$ *such that* $\|\hat{\mu} - \mu\|_2 \le \sqrt{k}$ *with probability at least* $1-\beta$ *so long as* $n = \Omega\left(\frac{k\log R + \log(1/\beta)}{\varepsilon}\right)$.

Lemma C.12 (Fine dense estimation)**.** *Let* $x_1, \ldots, x_n \sim P$ *be independent samples and suppose* P *has covariance* $\Sigma \preceq I$ *and mean* $\mu \in \mathbb{R}^k$ *such that* $\|\mu\|_2 \le 1$ *and* $\|\mu\|_0 \le k$*. Then running exponential mechanism on* $\{x_i\}_{i\le n}$ *with the score function* $S_{fine}(\{x_i\}_{i\le n}, x, L) = \max_{v:\|v\|=1} \sum_{i\le n} \mathbb{1}\{\langle x_i - x, v\rangle \ge L\}$ *gives us an estimate* $\hat{\mu}$ *such that* $\|\hat{\mu} - \mu\|_2 \le \alpha$ *with probability at least* $1-\beta$ *so long as* $n = \Omega\left(\frac{k\log k + \log(1/\beta)}{\alpha^2\varepsilon}\right)$.

For both the coarse and fine estimation, our proof strategy will be as follows: first, argue that the sizes $|P_{coarse}|$ and $|P_{fine}|$ of the packings are small enough such that $\max\{\log(|P_{coarse}|), \log(|P_{fine}|)\} \le k\log k/\alpha^2 + k\log(R)$. After that, we will show that low score (few "outliers") is achieved if and only if we are at a good mean candidate (α-close to μ).

Proof of Lemma C.11. The size of $\sqrt{k}$-packing of the ℓ_2 ball of radius R in k dimensions is $O\left(R^k\right)$. Thus, $\log(|P_{coarse}|) \le k\log R$, as desired.

The coarse estimation score function S_{coarse} has sensitivity 1, by an argument analogous to the one in Lemma C.8.

Now we turn to showing that S_{coarse} is low if and only if we are at a mean candidate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le \sqrt{k}$ We can directly leverage standard concentration arguments to argue that any ball of poly(k) radius that contains at least $0.9n$ of the samples must be at a distance at most $O\left(\sqrt{k}\right)$ to μ [HLZ20] with probability at least $1-\beta$. This is enough to show both directions of the desired implication.

Thus, in total, the coarse exponential mechanism gives us a coarse estimate with good utility with probability at least $1-\beta$ as long as $n = \Omega\left(\frac{\Delta}{\varepsilon}\left(k\log R + \log(1/\beta)\right)\right) = \Omega\left(\frac{k\log R + \log(1/\beta)}{\varepsilon}\right)$. □

Proof of Lemma C.12. As in the proof of Lemma C.11, the size of ε-packing (with ℓ_2 balls) of the unit ℓ_2 ball in k dimensions is $O\left(\left(\frac{1}{\varepsilon}\right)^k\right)$. Thus, $\log(|P_{fine}|) \le k\log k/\alpha^2$, as desired.

For the fine estimation part, without loss of generality assume that $\alpha = 1$; bucketed means as employed in the support estimation part get us at the cost of an α^2 factor in the sample complexity. We now show that a low score for a point x implies that $\|\mu - x\| \le 1$. More formally, let "low" mean $\le 0.1n$. A low score implies that

$$\max_{v:\|v\|=1} \langle X_i - x, v\rangle = \max_{v:\|v\|=1} \left(\langle X_i - \mu, v\rangle + \langle \mu - x, v\rangle\right) \le L \tag{21}$$

for at least $0.9n$ indices i. Taking the maximum for each term, this becomes $\max_{v:\|v\|=1}\langle X_i - x, v\rangle \le \max_{v:\|v\|=1}\langle X_i - \mu, v\rangle + \|\mu - x\|$. From the bounded covariance assumption, we get that for at least $0.9n$ of the samples, we must have $\|\mu - x\| \le L - 1$. Setting $L = 2$ shows this direction. Suppose now we have a good mean candidate x. Then

$$\max_{v:\|v\|=1} \langle X_i - x, v\rangle = \max_{v:\|v\|=1} \left(\langle X_i - \mu, v\rangle + \langle \mu - x, v\rangle\right) \le \max_{v:\|v\|=1} \langle X_i - \mu, v\rangle + 1. \tag{22}$$

Now from standard concentration we have that $\max_{v:\|v\|=1}\langle X_i - \mu, v\rangle \le 1$ for at least $0.9n$ of the samples, and we are done.

□

With this, we are ready to prove Theorem C.7.

Proof of Theorem C.7. First, we note that bucketing with bucket size b is done in order to reduce the variance of the samples. In particular, the bucketed means m_i have covariance matrix Σ_m s.t. $\Sigma_m \preceq \lambda/bI$, where $\lambda := \lambda_{\max}(\Sigma)$.

From Eq. (7), together with Lemma C.9 and Lemma C.10, we get that to avoid choosing a subset T for which $\|\mu_T - \mu\| \geq \alpha$ with probability at most β, we need

$$n - \frac{n\lambda/b}{\alpha^2} - \sqrt{\frac{n\log(3/\beta)}{2}} - \frac{2(k\log(d) + \log(3/\beta))}{\varepsilon} \geq n\frac{\sqrt{\lambda}}{2\sqrt{b}\alpha} + \frac{2}{L}\sqrt{\frac{nk\lambda}{b}} + \sqrt{\frac{n\log(3/\beta)}{2}}, \quad (23)$$

or equivalently

$$\sqrt{n}\left(1 - \frac{\lambda}{b\alpha^2} - \frac{1}{2}\sqrt{\frac{\lambda}{b\alpha^2}}\right) \geq \frac{2(k\log(d) + \log(3/\beta))}{\sqrt{n}\,\varepsilon} + \frac{2}{L}\sqrt{\frac{k\lambda}{b}} + 2\sqrt{\frac{\log(3/\beta)}{2}}. \quad (24)$$

This holds true when $b \geq 25\frac{\lambda}{\alpha^2}$ and
$n \geq \max\left\{10(k\log(d) + \log(3/\beta))/\varepsilon, 10k\lambda/\left(bL^2\right), 10\log(3/\beta)\right\}$. This gets us the support estimation part of the sample complexity in the statement of Theorem C.7.

Combining the support estimation step with the dense estimation steps from Lemma C.11 and Lemma C.12 gives the desired result in the absence of corruptions. The $\alpha + \sqrt{\eta}$ rate for η-corrupted inputs follows directly from Theorem 2.1. □

It is worth noting that if we directly leverage existing heavy-tailed estimators for the dense estimation subroutine [KSU20, HKM21], we would get suboptimal robustness guarantees – either $\alpha + \sqrt{d\eta}$ or $\alpha + \sqrt{\log(R)\eta}$ in the η-contamination model.

Now we turn to an information-theoretic lower bound for heavy-tailed sparse private mean estimation.

Proposition C.13. *Suppose M is an ε-DP algorithm such that for every distribution D with k-sparse mean μ with $\|\mu\| \leq R$ and covariance $\Sigma \preceq I$, M produces an estimate $\hat{\mu}$ from n samples such that $\|\mu - \hat{\mu}\| \leq \alpha$, with probability $1 - \beta$. Then*

$$n = \Omega\left(\frac{k\log d}{\alpha^2\varepsilon}\right).$$

Proof. Our proof is a straightforward adaptation of the methods used to prove [KSU20, Theorem 6.1]. In particular, we show the statement is true for a class of product distributions that satisfy the mean and covariance requirements.

Formally, let

$$Q_0 = \begin{cases} -\frac{\sqrt{k}}{\alpha} & w.p.\ \frac{\alpha^2}{k} \\ 0 & w.p.\ 1 - 2\frac{\alpha^2}{k} \\ \frac{\sqrt{k}}{\alpha} & w.p.\ \frac{\alpha^2}{k}. \end{cases} \quad (25)$$

and

$$Q_1 = \begin{cases} -\frac{\sqrt{k}}{\alpha} & w.p.\ \frac{1}{2}\frac{\alpha^2}{k} \\ 0 & w.p.\ 1 - 2\frac{\alpha^2}{k} \\ \frac{\sqrt{k}}{\alpha} & w.p.\ \frac{3}{2}\frac{\alpha^2}{k}. \end{cases} \quad (26)$$

For $c \in \{0,1\}^d$, define $Q_c = \bigotimes_{i=1}^d Q_{c_i}$. We will work only with k-sparse distributions, so we only need to consider c such that $\|c\|_1 = k$ Let C_k denote the set of such k-sparse vectors in $\{0,1\}^d$. For any $c, c' \in C_k$ we have that the maximum distance in total variation between Q_c and $Q_{c'}$ is at most α^2. Now let $\mathcal{H}$ be a linear code with Hamming distance $k/4$. This implies that for every $c, c' \in \mathcal{H}$, we have that c and c' differ on at least $k/4$ coordinates. From coding theory (Gilbert-Varshamov bound, [BIPW10, Lemma 3.1] for a more direct argument for k-sparse codes) we know that there exists a code of size $|\mathcal{H}| = \Omega\left(d^k\right)$. The result follows from standard concentration and packing arguments. □

D Polynomial-Time Sparse Mean Estimation with $\tilde{O}(k^2)$ Samples

In this section we describe a polynomial-time ε-DP algorithm for sparse mean estimation with strong group privacy guarantees – it maintains privacy of groups up to size $n/\mathrm{poly}\log(d)$. (Hence, the algorithm is automatically robust to corruption of a $1/\mathrm{poly}\log(d)$ fraction of inputs, per Theorem 2.1.)

Algorithm 3 Polynomial-time algorithm for sparse sub-Gaussian mean estimation

Input: iterations N, bucket size b, number of samples n, initial estimate μ_0, step size η

$m_j \leftarrow \frac{1}{b}\sum_{j=i\cdot b}^{(i+1)b} x_i$ for all $j \in [\lceil n/b \rceil]$ ▷ *compute bucketed means*

for $1 \le i \le N$ **do**

  Check `Halt-Estimation`$\left(\{m_j\}_{j=1}^n, \mu_{i-1}\right)$, return μ_{i-1} if `halt`

  $d_i \leftarrow$`Distance-Estimation`$\left(\{m_j\}_{j=1}^n, \mu_{i-1}\right)$

  $g_i \leftarrow$`Gradient-Estimation`$\left(\{m_j\}_{j=1}^n, \mu_{i-1}, d_i\right)$

  $\bar{\mu}_i \leftarrow \mu_{i-1} + \eta d_i g_i$

  $\mu_i \leftarrow$ `Sparsify`$(\bar{\mu}_i, k)$

Output: mean estimate μ_T

Theorem D.1. *There exist universal constants $C > 0, C' > 0$ such that for every $\varepsilon, R > 0$, $\beta \in (0,1)$, and $d, k \in \mathbb{N}$ such that $k \ge d^{0.4}$, there is a polynomial-time ε-DP algorithm with the following guarantees. For every k-sparse $\mu \in \mathbb{R}^d$ with $\|\mu\| \le R$, given independent samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, I)$, with probability at least $1-\beta$ the algorithm outputs $x \in \mathbb{R}^d$ such that $\|\mu - x\| \le \alpha$, so long as*

$$n \ge C(\log d)^{C'} \log R \cdot \frac{k^2 + \log(1/\beta) + \log\log R}{\alpha^2 \varepsilon}.$$

To prove Theorem D.1 we assemble our main lemmas here, and prove them in subsequent sections. The first key lemma captures a private gradient-finding procedure, finding a direction correlated with $x - \mu$ for k-sparse x.

Lemma D.2 (Finding Gradients)**.** *For every $c > 0$ there exists $C > 0$ such that for every $\beta, R, \varepsilon > 0$ there is a polynomial-time ε-DP algorithm,* `Gradient-Estimation`*, with the following guarantees. For every $d \in \mathbb{N}$ and $k \in \mathbb{N}$ with $k \ge d^{0.4}$ and every $\mu \in \mathbb{R}^d$ with $\|\mu\| \le R$, given*

- *n independent samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, I)$,*
- *a vector $x \in \mathbb{R}^d$ such that*
  - *$\|x\|_2 \le R$,*
  - *$\|x\|_0 \le k$, and*
  - *$\|x - \mu\| \ge C$*
- *a number $r \in [(1-2/C)\|x-\mu\|, (1-1/C)\|x-\mu\|]$*

and if

$$n \ge (\log d)^C \cdot (k^2 + \log(1/\beta)) \cdot \frac{1}{\varepsilon},$$

then with probability at least $1-\beta$, the algorithm produces a vector v such that $\left\| v - \frac{\mu - x}{\|\mu - x\|} \right\| \le c$. Furthermore, the algorithm runs in time poly$(n, d, \log(1/\varepsilon), \log R)$.

The second lemma shows that there is a private procedure to estimate the distance from the current iterate x to μ, as long as x is k-sparse.

Lemma D.3 (Estimating distance)**.** *There exists a polynomial-time ε-DP algorithm* `Distance-Estimation` *taking as input a vector $x \in \mathbb{R}^d$ with $\|x\| \le R$ and independent samples $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, I)$, where $\|\mu\| \le R$, with the following properties. For every $c > 0$ there is $C > 0$ such that given $n \ge \max\left\{\log(d)^C(k^2 + \log(1/\beta))/\varepsilon, 100\log(R)\log(\log(R)/\beta)/\varepsilon\right\}$ samples, and if $\|x - \mu\| \ge C$ the algorithm outputs a distance estimate $\hat{r}$ such that $(1-c)\|x-\mu\| \le \hat{r} \le (1+c)\|x-\mu\|$ with probability at least $1-\beta$.*

The next lemma shows that we can preserve the sparsity of our iterates by a simple thresholding procedure.

Lemma D.4 (Sparsifying iterates)**.** *Let* `Sparsify`(x, k) *denote the procedure of projecting x onto its k largest (in absolute value) coordinates, breaking ties at random. Let $x, y \in \mathbb{R}^d$ with $\|y\|_0 \le k$, and let $x' =$* `Sparisfy`(x, k)*. Then $\|x' - y\| \le 4\|x - y\|$.*

The last lemma shows that the algorithm can detect when the current iterate x is close enough to μ.

Lemma D.5 (Deciding to halt)**.** *There exists a universal constant C such that for all $C' > C$ there exists a polynomial-time ε-DP algorithm* `Halt-Estimation` *with the following guarantees. For every k-sparse x and μ such that $\|x\|, \|\mu\| \le R$, given x and $n \ge (\log d)^C(k^2 + \log(1/\beta))/\varepsilon$ samples from $\mathcal{N}(\mu, I)$,* `Halt-Estimation` *outputs "halt" or "do not halt". If $\|x - \mu\| \le C'$, then with probability at least $1-\beta$,* `Halt-Estimation` *outputs "halt", and if $\|x - \mu\| \ge 200C'$ then with probability at least $1-\beta$* `Halt-Estimation` *outputs "do not halt". Furthermore,* `Halt-Estimation` *runs in time poly$(n, d, \log(1/\varepsilon), \log R)$.*

Now we can put together the pieces to prove Theorem D.1.

Proof of Theorem D.1. First of all, by standard arguments, using buckets of size $b = 1/\alpha^2$ yields samples from $\mathcal{N}(\mu, \alpha^2 I)$. So by rescaling, it suffices therefore to prove the theorem in the case $\alpha = \Theta(1)$, which is what we will do.

For some β' we choose momentarily, let us condition on all of the $1-\beta'$-probability events specified in Lemmas D.2,D.3, and D.5, for all $O(\log R)$ calls to `Halt-Estimation`, `Distance-Estimation`, and `Gradient-Estimation`. Choosing $\beta' = \Omega(\beta/\log R)$, so long as $n \ge (\log d)^{O(1)} \log R(k^2 + \log(1/\beta) + \log\log R)/\varepsilon$, by a union bound, with probability $1-\beta$ all these events occur.

Having conditioned on all these events, from Lemma D.5, it suffices to show that there exists $C > 0$ such that within $O(\log R)$ iterations of the main loop of Algorithm 3 some iterate μ_i has $\|\mu_i - \mu\| \le C$. Then, we can choose C' in Lemma D.5 to have $C' \ge C$ and we obtain the guarantee that (a) `Halt-Estimation` outputs "halt" in some iteration and (b) when it does so, the output μ_i satisfies $\|\mu_i - \mu\| \le O(1)$.

It suffices therefore to show that there exists universal constant $C' > 0$ such that each iteration of gradient descent (lines 4-7 in Algorithm 3) shrinks $\|\mu_t - \mu\|$ by a factor of at least 0.9, i.e. if $\|\mu_{t-1} - \mu\| \ge C'$ then $\|\mu_t - \mu\| \le 0.9\|\mu_{t-1} - \mu\|$. Now we show that this follows from the guarantees of Lemma D.2 and Lemma D.3. In particular, it is enough to show that $\left\|g_t - \frac{\mu - \mu_{t-1}}{\|\mu - \mu_{t-1}\|}\right\| \le 0.01$. Given that $(1-c)\|\mu_{t-1} - \mu\| \le d_t \le (1+c)\|\mu_{t-1} - \mu\|$ for a constant c we will choose below (Lemma D.3), before calling `Sparsify` we have that

$$\begin{aligned}
\|\bar{\mu}_t - \mu\| &= \left\|\left(\mu_{t-1} - \mu + \eta\frac{\mu - \mu_{t-1}}{\|\mu - \mu_{t-1}\|}d_t\right) + \eta d_t g_t - \eta d_t \frac{\mu - \mu_{t-1}}{\|\mu - \mu_{t-1}\|}\right\| \\
&\le \left\|(\mu_{t-1} - \mu) - \eta\frac{\mu_{t-1} - \mu}{\|\mu - \mu_{t-1}\|}d_t\right\| + 0.01\eta d_t \\
&\le \left\|1 - \frac{\eta d_t}{\|\mu_{t-1} - \mu\|}\right\| \|\mu_{t-1} - \mu\| + 0.01(1+c)\eta\|\mu_{t-1} - \mu\| \\
&\le (\max\{|1 - \eta(1+c)|, |1 - \eta(1-c)|\} + 0.01(1+c)\eta)\,\|\mu_{t-1} - \mu\|.
\end{aligned} \tag{27}$$

Setting $\eta = 1$, we get that

$$\|\bar{\mu}_t - \mu\| \le (1.01c + 0.01)\|\mu_{t-1} - \mu\|. \tag{28}$$

Choosing $c \le 0.1$, and applying the guarantees of Lemma D.4 (`Sparsify` loses at most a factor of 4) finishes the argument. □

E Omitted Proofs from Appendix D

To set up, we define two key systems of polynomial inequalities.

Definition E.1 (The polynomial systems $\mathcal{Q}$ and $\mathcal{P}$)**.** *Let $\mathcal{Q}$ be the following polynomial system in variables $v_1, \dots, v_d, c_1, \dots, c_d$:*

$$\|v\|^2 = 1,\ \|c\|^2 = 1,\ v_i v_j \le c_i c_j \text{ and } -v_i v_j \le c_i c_j \text{ for all } i, j \in [d],\ \left(\sum_{i\in[d]} c_i\right)^2 \le s\,.$$

For $X_1, \dots, X_n \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$ and $\alpha, s > 0$, we define the following system of inequalities $\mathcal{P}(x, \delta, \alpha, s, X_1, \dots, X_n)$ in variables $w_1, \dots, w_d, v_1, \dots, v_d, b_1, \dots, b_n, c_1, \dots, c_d$ to be:

$$\mathcal{Q} \cup \{b_i^2 = b_i, b_i\langle X_i - x, v\rangle \ge b_i\alpha\}_{i\in[n]} \cup \{\|w - v\|^2 \le \delta\}\,.$$

E.1 Estimating Distance

Here we prove Lemma D.3.

Proof. Let $X = \{X_i\}_{i\le n}$ and let $f(r, X; \delta, s) = \max \tilde{\mathbb{E}} \sum_{i\le n} b_i$ s.t. $\tilde{\mathbb{E}}$ satisfies $\mathcal{P}(x, \delta, r, s, X)$; in words, $f(r, X)$ (roughly) counts the number of samples considered as "inliers" when we use radius r. The key idea is that this number is "large" when $r \gg \|x - \mu\|$ and "small" when $r \ll \|x - \mu\|$.

More formally, we invoke the following private binary search primitive [HKM21, Theorem 6.15]: Given a function $f\colon [0, R] \times \mathcal{X}^n \to [0, n]$, which is decreasing in its first argument, and has sensitivity 1 in its second argument, $S \stackrel{\text{def}}{=} \log(R/a)$ rounds of binary search suffice to output an estimate $\hat{r} \in [r_e - a, r_s + a]$, where $r_e = \min r$ s.t. $f(r, X) \le ek + \Delta$, and $r_s = \max r$ s.t. $f(r, X) \ge sk - \Delta$ for $\Delta = S\log(S/\beta)/\varepsilon$. We choose $s = 1 - 2\delta, e = 1 - \delta, \Delta = \frac{\delta}{10}n$ for some $\delta > 0$.

First, monotonicity follows from the fact that the solution for a given $r = r' > r''$ is a feasible solution for $r = r''$, and bounded sensitivity follows from the fact that the value of the (non-relaxed) quadratic program has sensitivity 1 since we can change at most one of the indicators upon changing a single sample; furthermore, there is an SoS proof of that [HKM21].

Now we turn to quantifying r_e and r_s; we show the following two facts:

1. For all $c_e > 0, \delta > 0$, there exists $C_e > 0$ such that if $\|x - \mu\| \ge C_e$, then $r_e \ge (1 - c_e)\|x - \mu\|$.
2. For all $c_s > 0, \delta > 0$, there exists $C_s > 0$ such that if $\|x - \mu\| \ge C_s$, then $r_s \le (1 + c_s)\|x - \mu\|$.

For Fact 1, it suffices to show that $f((1 - c_e)\|x - \mu\|, X) \ge (1 - \delta)n$. Now, it is enough to exhibit a feasible direction. In particular, set $v = \frac{\mu - x}{\|\mu - x\|}$. Let $S \subseteq [n]$ be a subset of samples such that $|S| \ge (1 - \delta)n$ and $\mathcal{Q} \vdash_{O(1)} \sum_{i\in S}\langle X_i - \mu, v\rangle^2 \le 2n$. We know such a subset exists from Lemma E.3 as long as $n \ge \log(d)^C \left(k^2 + \log(1/\beta)\right)/\varepsilon$. Then we have that there exists $S' \subseteq S$ such that $|S'| \ge (1 - 2\delta)n$ for which we have $|\langle X_i - \mu, v\rangle| \le 1/(2\delta)$ for all $i \in S'$. From this, we know that

$$\sum_{i\in S'}\langle X_i - x, v\rangle = \sum_{i\in S'}\langle X_i - \mu, v\rangle + \sum_{i\in S'}\langle \mu - x, v\rangle \ge -\frac{1}{2\delta} + \|x - \mu\|. \tag{29}$$

Now we need $\|x - \mu\| \ge \frac{1}{2c_e\delta}$ for Fact 1 to hold. Setting C_e to the RHS finishes the argument.

For Fact 2, it is enough to show that $f((1 + c_s)\|x - \mu\|, X) \le (1 - 2\delta)n$. Fix δ' and L to be chosen later. Let $S \subseteq [n]$ be a subset of samples such that $|S| \ge (1 - \delta')n$ and $\mathcal{Q} \vdash_{O(1)} \sum_{i\in S}\langle X_i - \mu, v\rangle^2 \le 2n$ (such exists from Lemma E.3).

Note that $(a + b)^2 \le (1 + L^2)a^2 + (1 + 1/L^2)b^2$ is true for any $L \in \mathbb{R}$ and has an SoS proof. Using this and the SoS Cauchy-Schwarz inequality, we have the following sequence of inequalities, all with

degree-2 SoS proofs.

$$
\begin{aligned}
\left(\sum_{i\le n} b_i\right)^2 &\le \left((n-|S|)+\sum_{i\in S} b_i\right)^2 \le \left((n-|S|)+\sum_{i\in S} b_i \frac{\langle X_i - x, v\rangle}{r}\right)^2 \\
&\le \left(1+L^2\right)(n-|S|)^2 + \left(1+\frac{1}{L^2}\right)\left(\sum_{i\in S} b_i \frac{\langle X_i - x, v\rangle}{r}\right)^2 \\
&\le \left(1+L^2\right)(\delta' n)^2 + \left(1+\frac{1}{L^2}\right)\left(\sum_{i\in S} b_i \frac{\langle X_i - \mu, v\rangle + \langle \mu - x, v\rangle}{r}\right)^2 \\
&\le \left(1+L^2\right)(\delta' n)^2 + \\
&\quad \left(1+\frac{1}{L^2}\right)\left(\left(1+L^2\right)\left(\sum_{i\in S} b_i \frac{\langle X_i - \mu, v\rangle}{r}\right)^2 + \left(1+\frac{1}{L^2}\right)\left(\sum_{i\in S} b_i \frac{\langle \mu - x, v\rangle}{r}\right)^2\right) \qquad (30) \\
&\le \left(1+L^2\right)(\delta' n)^2 + \left(1+\frac{1}{L^2}\right)\left(1+L^2\right)\left(\sum_{i\in S} b_i^2\right)\frac{2n}{r^2} + \\
&\quad \left(1+\frac{1}{L^2}\right)^2\left(\sum_{i\in S} b_i^2\right)\left(\sum_{i\in S}\frac{\|\mu - x\|^2}{r^2}\right) \\
&\le n^2 \cdot \left(\left(1+L^2\right)(\delta')^2 + \left(L^2+2+\frac{1}{L^2}\right)\frac{|S|}{n}\frac{2}{r^2} + \left(1+\frac{1}{L^2}\right)^2 \frac{|S|^2}{n^2}\frac{\|\mu - x\|^2}{r^2}\right) \\
&\le n^2 \cdot \left(\left(1+L^2\right)(\delta')^2 + \left(2L^2+4+\frac{2}{L^2} + \left(1+\frac{1}{L^2}\right)^2 \|\mu - x\|^2\right)\frac{1}{r^2}\right)
\end{aligned}
$$

We can choose $\delta' \le \frac{\sqrt{1-2\delta}}{100\sqrt{1+L^2}}$ such that the first term is sufficiently small. Next, we choose L so that the overall sum is at most $(1-2\delta)n$. In particular, we want that

$$
\frac{2L^2+4+\frac{2}{L^2}}{(1+c_s)^2\|\mu - x\|^2} + \left(\frac{1+1/L^2}{1+c_s}\right)^2 \le \frac{2L^2+4+\frac{2}{L^2}}{(1+c_s)^2 C_s^2} + \left(\frac{1+1/L^2}{1+c_s}\right)^2 \le (1-2\delta)^2. \qquad (31)
$$

Thus, we can choose L such that $L \ge \frac{1}{\sqrt{c_s - 2\delta(1+c_s)}}$ and $2L^2+4+2/L^2 \le ((1-2\delta)(1+c_s)C_s)^2$ (solving self-consistently).

Combining the conclusion of [HKM21, Theorem 6.15] with the conclusions of Facts 1 and 2, we receive as output a distance estimate $\hat{r} \in [(1-c_e)\|x-\mu\| - a, (1+c_s)\|x-\mu\| + a]$. Noting that $\|x-\mu\| \ge \max\{C_e, C_s\}$, we can set $c_s = c_e = c/2$ and $a = \max\{C_e, C_s\}\cdot c/2$ to obtain the desired result.

□

E.2 Sparsifying

Lemma E.2 (Lemma D.4 restated)**.** *Let* $\texttt{Sparsify}(x,k)$ *denote the procedure of projecting* x *onto its* k *largest (in absolute value) coordinates, breaking ties at random. Let* $x, y \in \mathbb{R}^d$ *with* $\|y\|_0 \le k$*, and let* $x' = \texttt{Sparisfy}(x,k)$*. Then* $\|x' - y\| \le 4\|x-y\|$*.*

Proof. Let $S = \mathrm{supp}(y)$ and $T = \mathrm{supp}(x')$. Now note that $|S \setminus T| = |T \setminus S| \le k$. Additionally, from the assumption that we select the largest k components, we have that $\|x_{T\setminus S}\| \ge \|x_{S\setminus T}\|$.

Using the above observation, together with the triangle inequality, we have that

$$\begin{aligned}\|x' - y\| &\le \|x'_{S\setminus T} - y_{S\setminus T}\| + \|x'_{T\setminus S} - y_{T\setminus S}\| + \|x - y\| \\ &\le \|x'_{S\setminus T} - x_{S\setminus T}\| + \|x_{S\setminus T} - y_{S\setminus T}\| + \|x_{T\setminus S} - y_{T\setminus S}\| + \|x - y\| \\ &\le \|x_{S\setminus T}\| + 3\|x - y\| \\ &\le \|x_{T\setminus S}\| + 3\|x - y\| \\ &\le 4\|x - y\|,\end{aligned} \tag{32}$$

as desired.

□

E.3 Deciding to Halt

In this section we prove Lemma D.5.

Proof of Lemma D.5. The algorithm is as follows. Compute $Z \stackrel{\text{def}}{=} \max \tilde{\mathbb{E}} \sum_{i=1}^n b_i$ over all $\tilde{\mathbb{E}}$ satisfying $\mathcal{P}(x, 1, 100C', X_1, \dots, X_n)$ and add noise $Lap(1/\varepsilon)$. If the result is $\le 0.1n$, output "halt", otherwise output "do not halt".

Sensitivity of Z was already proved in Lemma D.3. Privacy follows directly from the guarantees of the Laplace mechanism.

Next we establish correctness. This also follows directly from the arguments in Lemma D.3 since we can view halt estimation as a one-step binary search. In particular, let $S \subseteq [n]$ be a subset such that $|S| \ge (1-\delta)n$ and $\mathcal{Q} \vdash \sum_{i\in S} \langle X_i - x, v\rangle^2 \le 2n$. Then we claim that the algorithm above will output "halt" with probability at least $1-\beta$, using that C' is at least some universal constant.

At the same time, suppose that the algorithm above outputs "halt". Conditioning again on an event of probability $1-\beta$, this means that $\max \tilde{\mathbb{E}} \sum_{i=1}^n b_i$ over all $\tilde{\mathbb{E}}$ satisfying $\mathcal{P}(x, 1, 100C', X_1, \dots, X_n)$ is at most $0.2n$. In particular, for every $2k$-sparse unit vector v, at least $0.8n$ choices of i have $\langle X_i - x, v\rangle \le 100C'$, and the same is true for $\langle X_i - \mu, v\rangle$. Choosing some i such that this holds simultaneously for $v = (\mu - x)/\|\mu - x\|$ in the first case and $v = (x-\mu)/\|x-\mu\|$ in the second, we obtain by adding the two,

$$200C' \ge \langle X_i - x, (\mu - x)/\|\mu - x\|\rangle - \langle X_i - \mu, (\mu - x)/\|\mu - x\|\rangle = \|\mu - x\| .$$

□

E.4 Finding Gradients

In this section we prove Lemma D.2. The promised algorithm will use the SoS exponential mechanism of [HKM21].

Our first lemma says that the quadratic form of the covariance matrix of $X_1, \dots, X_n$, after throwing out a few samples, has bounded quadratic form in sparse directions, and that furthermore this has an SoS proof.

Lemma E.3. *For every $c > 0$ there exists a universal constant $C > 0$ such that for all large-enough $d \in \mathbb{N}$, all $s \le d$, and all $\beta \in (0,1)$, for all $n \ge (\log d)^C (s^2 + \log(1/\beta))$, with probability at least $1-\beta$ over i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(0, I)$ there is a subset $S \subseteq [n]$ with $|S| \ge (1-c)n$ and such that*

$$\mathcal{Q} \vdash_{O(1)} \sum_{i\in S} \langle X_i, v\rangle^2 \le 2n ,$$

The next lemma gives the SoS proof of utility we need for SoS exponential mechanism.

Lemma E.4. *For every $c > 0$ there is $C > 0$ such that the following holds. Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, I)$ for $\mu \in \mathbb{R}^d$ with $\|\mu\|_0 \le k$, with $d \in \mathbb{N}$ sufficiently large. For all $x \in \mathbb{R}^d$ with $\|x - \mu\| > C$, all $\beta \in (0,1)$, and $r \in [(1-1/C)\|x-\mu\|, (1+1/C)\|x-\mu\|]$, and*

$n \geq (\log d)^C(k^2 + \log(1/\beta))$, with probability at least $1-\beta$ over $X_1, \ldots, X_n$, there is a degree-$O(1)$ SoS proof

$$\mathcal{P}(x, 1/C, r, k/c, X_1, \ldots, X_n), \sum_{i \leq n} b_i \geq (1-1/C)n \vdash_{O(1)} \left\| w - \frac{\mu - x}{\|\mu - x\|} \right\|^2 \leq c\,,$$

and, furthermore, this proof has degree-1 in the constraint $\sum_{i \leq n} b_i \geq (1-1/C)n$.

The last lemma proves key properties of the convex set over which SoS exponential mechanism will run a log-concave sampling algorithm.

Lemma E.5. *Let $d, k \in \mathbb{N}$ with $k \leq d$, and let $\delta > 0$. Then the set*

$$C(k, \delta) = \{x \in \mathbb{R}^d \,:\, \exists y \in \mathbb{R}^d \text{ s.t. } \|y\|_2 \leq 1, \|y\|_1 \leq \sqrt{k}, \|x - y\|_2 \leq \delta\}$$

has the following properties:

- *C is compact and convex.*
- *C has diameter $O(1+\delta)$.*
- *C admits polynomial-time projection and membership oracles.*
- *For every $v \in \mathbb{R}^d$ with $\|v\| \leq 1$ and $\|v\|_1 \leq \sqrt{k}$, there exists a set $S_v \subseteq C$ such that for all $w \in S_v$, $\|w - v\| \leq \delta$, and $|S_v| \geq \exp(-O(d/(\delta\sqrt{k}) + k^2 \log d)) \cdot |C|$, where $|\cdot|$ denotes Lebesgue measure.*

Now we can apply Theorem 4.5 of [HKM21] to prove Lemma D.2.

Proof of Lemma D.2. We verify that the conditions of Theorem 4.5 of [HKM21] apply to our polynomial system $\mathcal{P}$, with $p(b) = \sum_{i \in [n]} b_i - (1-1/C)n$ for a large-enough constant C, and the convex set $C(k, 1/C)$ of Lemma E.5.

- Compactness and convexity of C, projection and membership oracle, diameter at most $\mathrm{poly}(n, d)$: guaranteed by Lemma E.5.
- $\mathcal{P}$ is Archimedian: $\mathcal{P}$ contains constraints upper-bounding $\|v\|^2, \|c\|^2, \|b\|^2$ by $\mathrm{poly}(d, n)$, by inspection.
- Robust satisfiability: let $\eta = 1/\mathrm{poly}(n, d)$. We claim that for all $X_1, \ldots, X_n, \alpha$, the system $\mathcal{P}(x, 2/C, \alpha, s, X_1, \ldots, X_n)$ is η-robustly satisfiable with respect to C and $p(b) = \sum_{i \leq n} b_i - (1-1/C)n$.

  To see this, consider any $y \in C$ and any y' such that $\|y' - y\| \leq 1/\mathrm{poly}(d, n)$. Then, setting $w = y'$, we can satisfy $\mathcal{P}$ by choosing v to be the element of C satisfying $\|v - y\| \leq 1/C$ and having $\|v\|_2 \leq 1, \|v\|_1 \leq \sqrt{k/c}$; then we will have $\|w - v\| \leq 1/C + 1/\mathrm{poly}(d, n) \leq 2/C$. Choosing $c_i = |v_i|$, and choosing $b_i = 0$ for all i satisfies the rest of the constraints in $\mathcal{P}$.
- SoS proof of bounded sensitivity: same as Example 4.2 in [HKM21].
- SoS proof of utility: existence with probability at least $1-\beta$ over choice of samples follows directly from Lemma D.2. This proof if expressible in at most $\mathrm{poly}(d, n, \log R)$ bits by inspection of the proof of Lemma D.2.
- Volume ratio: We claim that for $n \geq (\log d)^C(k^2 + \log(1/\beta))$ and $k \geq d^{0.4}$, there exists a set $S \subseteq C$ with $|S|/|C| \geq d^{-O(k^2)}$ (where $|\cdot|$ denotes Lebesgue measure) such that for all $z \in S$ there is a solution v, w, b, c to $\mathcal{P}$ with $p(b) \geq \Omega(n)$ and $w = z$.

  To see this, we take S to be the set of z such that $\|z - (\mu - x)/\|\mu - x\|\| \leq 1/C$. First, $\|\mu - x\|_0 \leq 2k$, so $\|\mu - x\|_1 \leq \sqrt{2k}\|\mu - x\|_2$, hence $S \subseteq C$. For any $z \in S$ we can set $w = z$ and $v = (\mu - x)/\|\mu - x\|$, then $\|w - v\|^2 \leq 2/C$. Take $c_i = |v_i|$; since v is

$2k$-sparse this satisfies the constraints of $\mathcal{Q}$. Take $b_i = 1$ if $\langle X_i - x, \frac{\mu - x}{\|\mu - x\|}\rangle = \langle X_i - \mu, \frac{\mu - x}{\|\mu - x\|}\rangle + \|\mu - x\| \geq r$, again satisfying $\mathcal{P}$.

It remains to show that $\sum_{i\in[n]} b_i - (1 - 1/C)n \geq \Omega(n)$. Applying Lemma E.3 and Markov's inequality, for large-enough $n \gg (\log d)^{C'}(k^2 + \log(1/\beta))$ for some other constant C', there is a set $T \subseteq [n]$ with $|T_i| \geq (1 - 1/(10C))n$ and $|\langle X_i - \mu, (\mu - x)/\|\mu - x\|\rangle| \leq O(1)$. So, as long as $r \leq \|\mu - x\| - O(1)$, $\sum b_i$ is large enough.

It follows that for fixed x and μ, with probability at least $1 - \beta$ over $n \geq (\log d)^C(k^2 + \log(1/\beta))/\varepsilon$ samples from $\mathcal{N}(\mu, I)$, SoS exponential mechanism is ε-DP and outputs a vector v such that $\|v - (\mu - x)/\|\mu - x\| \leq c$.

For the running time dependence on $\log(1/\varepsilon)$, note that the dependence of $1/\varepsilon$ claimed in [HKM21] for SoS exponential mechanism can be improved by appeal to the main result of [MV21]. □

E.4.1 Proof of Lemma E.4, Lemma E.3

Proof of Lemma E.3. Consider the random variable $B = \max\{|S| : \mathcal{Q} \vdash_{O(1)} \sum_{i\in S}\langle X_i, v\rangle^2 \leq 2n\}$. By McDiarmid's bounded-differences inequality, $\Pr(|B - \mathbb{E}B| > t) \leq 2\exp(-\Omega(t^2/n))$, so as long as $n \gg \log(1/\beta)$, we have $|B - \mathbb{E}B| \leq cn/10$ with probability at least $1 - \beta$. It will suffice therefore to show that $\mathbb{E}B \geq (1 - c/10)n$. For this in turn it suffices to show that with probability at least $1 - c/20$ we have $B \geq (1 - c/20)n$.

Let $S \subseteq [n]$ be the set of X_i such that $\|X_i\|_\infty \leq c'\sqrt{\log d}$, with $c' > 0$ chosen so that $|S| \geq (1 - c/20)n$ with probability at least $(1 - c/100)$. We claim that if $n \gg \operatorname{poly}\log(d) \cdot s^2$ then $\mathcal{Q} \vdash_{O(1)} \sum_{i\in S}\langle X_i, v\rangle^2 \leq 2n$ with probability at least $1 - c/100$; then a union bound finishes the proof.

Let $M = \sum_{i\in S} X_i X_i^\top$. By Bernstein's inequality, with probability at least $1 - \delta$ the following both hold:

$$\max_{a\neq b\in[d]} |M_{ab}| \leq O(\sqrt{n\log(1/\delta)} + \log d \cdot \log(1/\delta))$$

$$\max_{a\in[d]} |M_{aa}| \leq n + O(\sqrt{n\log(1/\delta)} + \log d \cdot \log(1/\delta)) .$$

For all $a, b \in [d]$, $\mathcal{Q} \vdash_{O(1)} v_a v_b M_{ab} \leq |M_{ab}| c_a c_b$, so

$$\begin{aligned}
\mathcal{Q} \vdash \sum_{i\leq n}\langle X_i, v\rangle^2 &\leq \max_{a\in[d]} |M_{aa}| \sum_{a\in[d]} c_a^2 + \max_{a\neq b\in[d]} |M_{ab}| \cdot \left(\sum_{a\in[d]} c_a\right)^2 \\
&\leq \max_{a\in[d]} |M_{aa}| + s \cdot \max_{a\neq b\in[d]} |M_{ab}| \\
&\leq s \cdot O(\sqrt{n\log(1/\delta)} + \log d \cdot \log(1/\delta)) + n .
\end{aligned}$$

Choosing δ a small enough, and then C large enough, completes the proof. □

Lemma E.6. *For an indeterminate X and $C > 0$,*

$$X^2 \leq C^2 \vdash_2 X \leq 2C .$$

Proof. Note that $X = \frac{1}{2C}((X + C)^2 - X^2 - C^2)$. So $X^2 \leq C^2 \vdash X \leq \frac{1}{2C}(X + C)^2 \leq \frac{1}{2C} \cdot 2(X^2 + C^2)$ via SoS triangle inequality. Using the axiom $X^2 \leq C^2$ completes the proof. □

Now we can prove Lemma E.4.

Proof of Lemma E.4. First, by SoS triangle inequality and the constraint $\|w - v\|^2 \leq 1/C$, for large-enough C, it suffices to show that

$$\mathcal{P}(x, 1/C, r, k, X_1, \ldots, X_n), \sum_{i\leq n} b_i \geq (1 - 1/C)n \vdash_{O(1)} \left\|v - \frac{\mu - x}{\|\mu - x\|}\right\|^2 \leq c/4 .$$

And, for this, it is enough to show $\mathcal{P}, \sum_{i\le n} b_i \ge (1-1/C)n \vdash_{O(1)} \langle \mu - x, v\rangle \ge (1-c/8)\|\mu - x\|$. (All these proofs must be degree-1 in $\sum_{i\le n} b_i \ge (1-1/C)n$, which will be true by construction.)

Let $S \subseteq [n]$ be the set of indices guaranteed to exist by Lemma E.3, of size $(1-1/C)n$, for the vectors $X_1 - \mu, \dots, X_n - \mu$. We first claim that

$$\mathcal{P} \vdash_{O(1)} \sum_{i\in S} b_i \langle X_i - \mu, v\rangle \le 4n\,.$$

For this, using Lemma E.6, we can show instead

$$\mathcal{P} \vdash_{O(1)} \left(\sum_{i\in S} b_i \langle X_i - \mu, v\rangle\right)^2 \le 2n^2\,,$$

but this follows immediately from SoS Cauchy-Schwarz and Lemma E.3.

Now,

$$\begin{aligned}\mathcal{P}, \sum_{i\in n} b_i \ge (1-1/C)n \vdash_{O(1)} r\cdot(1-2/C)n &\le r\cdot \sum_{i\in S} b_i \\ &\le \sum_{i\in S} b_i\langle X_i - x, v\rangle = \sum_{i\in S} b_i\langle X_i - \mu, v\rangle + b_i\langle \mu - x, v\rangle\,.\end{aligned}$$

Putting this together with the preceding, we have

$$\mathcal{P}, \sum_{i\in[n]} b_i \ge (1-1/C)n \vdash_{O(1)} (1-2/C)rn - 4n \le \langle \mu - x, v\rangle \cdot \sum_{i\in S} b_i = n\cdot\langle \mu - x, v\rangle + \left(\sum_{i\in S} b_i - n\right)\cdot\langle \mu - x, v\rangle\,.$$

Next we claim that
$\mathcal{P}, \sum_{i\in[n]} b_i \ge (1-1/C)n \vdash_{O(1)} (\sum_{i\in S} b_i - n)\cdot\langle \mu - x, v\rangle \le O(1/\sqrt{C})n\|\mu - x\|$.
We have

$$\begin{aligned}\|v\|^2 \le 1 &\vdash_{O(1)} \left(\sum_{i\in S} b_i - n\right)\cdot\langle \mu - x, v\rangle \\ &\le \frac{\sqrt{C}\|\mu - x\|}{n}\cdot\left(\sum_{i\in S} b_i - n\right)^2 + \frac{n}{\sqrt{C}\|\mu - x\|}\langle \mu - x, v\rangle^2 \\ &\le \frac{\sqrt{C}\|\mu - x\|}{n}\cdot\left(\sum_{i\in S} b_i - n\right)^2 + (1/\sqrt{C})n\|\mu - x\|\,.\end{aligned}$$

Now, $\{b_i^2 = b_i\}_{i\in[n]} \vdash_{O(1)} (\sum_{i\in S} b_i - n)^2 \le 2n\,(\sum_{i\in S} b_i - n)$, so via a proof which is degree-1 in $\sum_{i\in[n]} b_i \ge (1-1/C)n$ we have

$$\mathcal{P}, \sum_{i\in[n]} b_i \ge (1-1/C)n \vdash_{O(1)} \left(\sum_{i\in S} b_i - n\right)^2 \le 2\cdot 1/C\cdot n^2\,.$$

Putting everything together and using the assumption $r \ge (1-1/C)\|\mu - x\|$, we get

$$\mathcal{P}, \sum_{i\le n} b_i \ge (1-1/C)n \vdash_{O(1)} (1 - O(1/\sqrt{C}))\|\mu - x\| - 4 \le \langle \mu - x, v\rangle$$

which gives the conclusion by taking $C = C(c)$ large enough. □

E.4.2 Proof of Lemma E.5

Proof of Lemma E.5. Let $\mathcal{C} = \{x \in \mathbb{R}^d : \exists v \in \mathbb{R}^d \text{ s.t. } \|x - v\| \le \delta \text{ and } \|v\|_1 \le \sqrt{k}, \|v\|_2 \le 1\}$ be a "fattening" of the scaled ℓ_1 ball. $\mathcal{C}$ is convex by inspection.

To compute membership in C, given x, compute the projection x' to $\{v : \|v\|_1 \le \sqrt{k} \text{ and } \|v\|_2 \le 1\}$. If $\|x - x'\| \le \delta$ then $x \in C$ and otherwise $x \notin C$. Furthermore, if $\|x - x'\| > \delta$, then for some δ', a hyperplane through $x + \delta'(x' - x)$ separates x from C. With this separation oracle for C, projections can be computed by minimizing $\|x - y\|$ over $y \in C$.

Now we move on to the volume arguments. By Sudakov minoration (see e.g. [LT91]), $\{v \ \|v\|_1 \le \sqrt{k}\}$ can be covered by $d^{O(k^2)}$ ℓ_2 balls of radius $1/\sqrt{k}$. So, C can be covered by $d^{O(k^2)}$ ℓ_2 balls of radius $1/\sqrt{k} + \delta$. For any v with $\|v\|_1 \le \sqrt{k}$ and $\|v\|_2 \le 1$, we know $S_v = \{w : \|w - v\| \le \delta\} \subseteq C$. Furthermore,

$$\frac{|S_v|}{|C|} \ge d^{-O(k^2)} \cdot \left(\frac{\delta}{\delta + \frac{1}{\sqrt{k}}}\right)^d = d^{-O(k^2)} 2^{-O(d/(\sqrt{k}\delta))}.$$

□

F Proof of Theorem 4.1

Proof. The key idea in our analysis is to show that there exists a parameter regime (for the threshold T and bucket size b) for which there is a separation between the minimum score z_i of coordinates for which $|\mu_i| > T$ and the maximum expected score of coordinates for which $\mu_i = 0$.

Assume for simplicity that $n/b = \lceil n/b \rceil$. Note that $m_i \sim \mathcal{N}(\mu, \sigma^2/bI)$. Define $p_i := \mathbb{P}(|(m_1)_i| > T)$ as the probability of a (bucketed mean of a) sample exceeding the given threshold T. Then $z_i \sim \mathrm{Bin}(n/b, p_i)$. Note that from Chernoff we have for all coordinates i outside the support of μ that $p_i \le \exp\left(-bT^2/(2\sigma^2)\right) =: p_0$. Let $z_0 \sim \mathrm{Bin}(n/b, p_0)$. Since all p_is are bounded below by a constant, we can use the normal approximation to z_i and sub-Gaussian concentration. On one hand,

$$\begin{aligned} \mathbb{E} \max_{i:\mu_i=0} z_i &\le \frac{n}{b} p_0 + \sqrt{2(\log d + \log(4k/\beta)) \cdot \mathrm{Var}(z_0)} \\ &\le \frac{n}{b} \exp\left(\frac{-bT^2}{2\sigma^2}\right)\left(1 + \sqrt{\frac{2b(\log d + \log(4k/\beta))}{np_0}}\right) \end{aligned} \tag{33}$$

with probability at least $1 - \beta/(4k)$.

On the other, for large coordinates i where $\mu_i \ge T$, we have that $p_i \ge \frac{1}{2}$ and thus

$$\begin{aligned} \mathbb{E} \min_{i:|\mu_i|\ge T} z_i &\ge \frac{n}{b} \min_{i:|\mu_i|\ge T} p_i - \sqrt{2(\log d + \log(4k/\beta)) \cdot (n/b) \cdot \min_{i:|\mu_i|\ge T} p_i(1 - p_i)} \\ &\ge \frac{n}{b}\left(\frac{1}{2} - \sqrt{\frac{2b(\log d + \log(4k/\beta))}{n}}\right), \end{aligned} \tag{34}$$

again with probability at least $1 - \beta/(4k)$.

We can ensure that $\mathbb{E} \min_{i:|\mu_i|\ge T} z_i - \frac{2k(\log(d)+\log(4k/\beta))}{\varepsilon} > \mathbb{E} \max_{i:\mu_i=0} z_i$, by setting $T = 3.5\sigma/\sqrt{b}$ and $n \ge \max\{1, k/\varepsilon\} \cdot 20b(\log(d) + \log(4k/\beta))$. For all practical purposes, we can safely assume $k/\varepsilon \ge 1$ and thus get $n \ge 20kb(\log(d) + \log(4k/\beta))/\varepsilon$. For all

In k rounds of the exponential mechanism, each with privacy budget ε/k, we will pick each coordinate i such that $|\mu_i| \ge T$ with probability at least $1 - \beta/(4k)$. Taking a union bound over the success probabilities of the bounds, and the exponential mechanism rounds, we get that we will select k coordinates above the threshold T with probability $1 - \beta/2$ (if there are at least that many). Assuming we "give up" on potentially non-zero coordinates with mean magnitude lower than T, we get the following bound on the estimation error α.

$$\sum_{i:|\mu_i|\le T} \mu_i^2 \le kT^2. \tag{35}$$

To get to an estimation error of at most α, we need

$$\alpha^2 \geq kT^2 \geq 15k\frac{\sigma^2}{b} \geq 300k\sigma^2 \cdot \frac{k(\log d + \log(4k/\beta))}{n\varepsilon}. \tag{36}$$

This implies that for the support estimation part we will need

$$n = \Omega\left(\sigma^2 \frac{k^2(\log d + \log(k/\beta))}{\alpha^2\varepsilon}\right) \tag{37}$$

samples. It remains to privately estimate the mean on the selected coordinates. Since we have already used up a super-linear (in k) number of samples, we can afford to use a naive estimator for the mean estimation part: we can invoke the histogram-based univariate estimator of Karwa and Vadhan [KV17] for each coordinate, and require accuracy of $\alpha/\sqrt{k}$ on each coordinate. For non-selected coordinates, we return $\hat{\mu}_i = 0$. In total, that would give us an ℓ_2 guarantee of α, as desired.

From [KV17, Theorem 1.1] we know that we need

$$n = \Omega\left(\frac{\sigma^2 k\log(2k/\beta)}{\alpha^2} + \frac{\sigma k^{1.5}\log(2k/\beta)}{\alpha\varepsilon} + \frac{k\log(R)}{\varepsilon}\right) \tag{38}$$

samples to estimate each coordinate up to under ε/k DP with probability at least $1-\beta/(2k)$. Using coordinate, we will need $\tilde{\Omega}(k^{1.5}/(\alpha^2\varepsilon))$ samples in total for the dense mean estimation part.

Combining the support estimation and the dense mean estimation components, we get the desired result.

□

G Re-statement of sparse mean algorithm of [CWZ21]

In Section 3 and Section 4 we compare Algorithm 1 with [CWZ21, Algorithm 3.3] (referred to as CWZ algorithm hereafter). We focus solely on pure DP, and [CWZ21] state only an approximate DP version of the CWZ algorithm. To have a fair comparison, we make a (minor) modification of the CWZ algorithm to handle the ε-DP case. In short, we replace the advanced composition step in their analysis with a basic composition needed for the more stringent pure DP requirements.

More concretely, we keep [CWZ21, Algorithm 3.3] intact and only modify the peeling subroutine [CWZ21, Algorithm 3.2]. Given an a priori boun R_∞ satisfying $\|\mu\|_\infty \leq R_\infty$, we replace the scale $R_\infty \cdot \frac{\sqrt{k\log(1/\delta)}}{\varepsilon}$ of the Laplace noise added on lines 3 and 7 of Algorithm 3.2 with $R_\infty \cdot \frac{k}{\varepsilon}$. The new privacy analysis differs from [CWZ21, Lemma 3.3] only in the composition steps.

H Meta-theorem for concentrated DP

An alternative formulation of differential privacy if that of zero-concentrated DP (zCDP) based on Rényi divergence. More formally, we have the following definitions.

Definition H.1 (Rényi divergence)**.** *Given distributions P and Q on a common sample space Ω, the α-Rényi divergence between P and Q is given by*

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\left(\mathop{\mathbb{E}}_{x\sim P}\left[\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}\right]\right). \tag{39}$$

Definition H.2 (zCDP [BS16])**.** *Let $\mathcal{X}$ be a set and $\mathcal{X}^* = \{(X_1,\ldots,X_n) : n \in \mathbb{N}, X_i \in \mathcal{X}\}$ be all possible datasets over $\mathcal{X}$. For $\rho > 0$, a (randomized) map $M : \mathcal{X}^* \to \mathcal{O}$ (where $\mathcal{O}$ is an* output *set) is ρ-zCDP if for every $(X_1,\ldots,X_n),(X'_1,\ldots,X'_n) \in \mathcal{X}^*$ such that $X_i = X'_i$ except for a single index i and for every $\alpha \in (1,\infty)$ it holds that $D_\alpha(M(X_1,\ldots,X_n)\|M(X'_1,\ldots,X'_n)) \leq \rho\alpha$.*

With this, we are ready to state the zCDP version of Theorem 2.1.

Corollary H.3 (zCDP version of Theorem 2.1)**.** *Let $M : \mathcal{X}^* \to \mathcal{O}$ be an ρ-zCDP map from datasets $\mathcal{X}^*$ to outputs $\mathcal{O}$. For every dataset $X_1, \ldots, X_n$, let $G_{X_1,\ldots,X_n} \subseteq \mathcal{O}$ be a set of* good *outputs. Suppose that $M(X_1, \ldots, X_n) \in G_{X_1,\ldots,X_n}$ with probability at least $1 - \beta$ for some $\beta = \beta(n)$. Then, for every $n \in \mathbb{N}$ and every $\delta > 0$, on n-element datasets M is* robust *to adversarial corruption of any $\eta(n)$-fraction of inputs, where*

$$\eta(n) = \Omega\left(\min\left(\frac{\log 1/\beta}{\varepsilon(\rho,\delta)\cdot n}, \frac{\log 1/\delta}{\varepsilon(\rho,\delta)\cdot n \log n}\right)\right),$$

and $\varepsilon(\rho,\delta) = \rho + 2\sqrt{\rho \log(1/\delta)}$, meaning that for every $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$ differing on only ηn elements, $M(X'_1, \ldots, X'_n) \in G_{X_1,\ldots,X_n}$ with probability at least $1 - \beta^{\Omega(1)}$.

Proof. The result follows directly from the connection between zCDP and approximate differential privacy [BS16, Proposition 1.3]. □