From Robustness to Privacy and Back

Hilal Asi [1] **Jonathan Ullman** [2] **Lydia Zakynthinou** [2]

Abstract

We study the relationship between two desiderata of algorithms in statistical inference and machine learning—differential privacy and robustness to adversarial data corruptions. Their conceptual similarity was first observed by Dwork and Lei (STOC 2009), who observed that private algorithms satisfy robustness, and gave a general method for converting robust algorithms to private ones. However, all general methods for transforming robust algorithms into private ones lead to suboptimal error rates. Our work gives the first black-box transformation that converts any adversarially robust algorithm into one that satisfies pure differential privacy. Moreover, we show that for any low-dimensional estimation task, applying our transformation to an optimal robust estimator results in an optimal private estimator. Thus, we conclude that for any low-dimensional task, the optimal error rate for ε-differentially private estimators is essentially the same as the optimal error rate for estimators that are robust to adversarially corrupting $1/\varepsilon$ training samples. We apply our transformation to obtain new optimal private estimators for several high-dimensional tasks, including Gaussian (sparse) linear regression and PCA. Finally, we present an extension of our transformation that leads to approximate differentially private algorithms whose error does not depend on the range of the output space, which is impossible under pure differential privacy.

1. Introduction

Both *differential privacy* and *robustness* are desirable properties for machine learning or statistical algorithms, and there are extensive, mostly separate bodies of research on each of these properties.

Differential privacy (DP) was proposed by Dwork, McSherry, Nissim, and Smith (Dwork et al., 2006) as a rigorous formalization of what it means for an algorithm to guarantee individual privacy, and has been widely deployed in both industry (Erlingsson et al., 2014; Bittau et al., 2017; Apple Differential Privacy Team, 2017; Tastuggine and Mironov, 2020; Wilson et al., 2020; Rogers et al., 2020) and government (Haney et al., 2017; Abowd, 2018; Abowd et al., 2022) applications. Informally, a DP algorithm ensures that no adversary who observes the algorithm's output can learn much more about an individual in the dataset than they could if that individual's data had been excluded. Formally, a randomized algorithm $\mathcal{A}$ satisfies ε-DP if for every dataset $\mathcal{S}$, and every dataset $\mathcal{S}'$ that differs on one, or a small number of entries, the distributions $\mathcal{A}(\mathcal{S})$ and $\mathcal{A}(\mathcal{S}')$ are ε-close in a precise sense (see Definition 2.1), where the privacy guarantee becomes stronger as ε becomes smaller.

Robustness, which was first systematically studied by Tukey and Huber in the 1960s (Tukey, 1960; Huber, 1965), formalizes algorithms that perform well under data corruptions or model misspecifications. Specifically, we consider a dataset $\mathcal{S}$ that is drawn iid from some well behaved distribution P, and allow an adversary to produce a dataset $\mathcal{S}'$ that differs from $\mathcal{S}$ in a τ fraction of its entries. An algorithm $\mathcal{A}$ is τ-robust if the distance $\|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}')\|$ is typically small in some particular error norm, where the robustness guarantee becomes stronger as τ becomes larger.

Although these two conditions have entirely different motivations, they are both notions of what it means for an algorithm to be insensitive to small changes in its input, which was first observed in the influential work of Dwork and Lei (2009). But even once we recognize their similarity, there are substantial technical differences. While differentially private algorithms are robust, DP is a more stringent requirement in a few ways: First, DP is *worst case*, meaning the algorithm $\mathcal{A}$ must be insensitive in a neighborhood around *every* dataset $\mathcal{S}$, whereas a robust algorithm only needs to be insensitive in the *average case* around datasets

Authors ordered alphabetically.

[1]Apple [2]Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA. Correspondence to: Hilal Asi <hilal.asi94@gmail.com>, Lydia Zakynthinou <zakynthinou.l@northeastern.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

$\mathcal{S}$ drawn from well behaved distributions P. Second, DP requires that $\mathcal{A}(\mathcal{S})$ an $\mathcal{A}(\mathcal{S}')$ be close *as probability distributions* in a strong sense, whereas robustness only requires the *distance between outputs* $\mathcal{A}(\mathcal{S})$ and $\mathcal{A}(\mathcal{S}')$ to be small. On the other hand, since DP is harder to satisfy, DP algorithms are typically insensitive to changes in a *small number* of inputs, whereas robust algorithms can often be insensitive to changes in a *small constant fraction* of inputs.

Although, differential privacy is stronger than robustness, Dwork and Lei (Dwork and Lei, 2009) designed a method, called *propose-test-release (PTR)*, which can be used to turn any accurate robust algorithm for a statistical estimation task into an accurate differentially private algorithm for the same task. However, this generic transformation typically does not lead to optimal algorithms for most specific tasks of interest. Nonetheless, there has been a recent flurry of works in differentially private statistics that use robust estimators as *inspiration* for differentially private estimators (see Related Work for a summary of this line of work). These works use a variety of methods for upgrading robust estimators to differentially private ones, and each of these methods is tailored to a specific task or set of tasks.

In this paper we demonstrate that there is, in fact, a *black-box* way to transform robust estimators into private estimators that provably gives optimal error rates for low-dimensional tasks, and often leads to optimal error rates for many high-dimensional tasks.

1.1. Our Contributions

In this section we give a high-level overview of our main contributions—black-box transformation from robust to private estimators, optimality results for our transformations for low-dimensional estimation tasks, and applications of our transformations to high-dimensional estimation tasks.

From robustness to privacy via inverse-sensitivity. Our first main contribution is a black-box transformation that takes a robust algorithm for any statistical task and converts it into an ε-differentially private algorithm for the same task with comparable accuracy. Intuitively, since robust estimators are insensitive to changing $n\tau$ inputs on a dataset of size n, and private estimators are insensitive to changing a $1/\varepsilon$ total inputs, the accuracy of the private estimator will be related to the accuracy of the robust estimator when a $\tau \approx 1/n\varepsilon$ fraction of inputs can be corrupted.

Our transformation applies in the standard setting of statistical estimation: assume that there exists a distribution P over domain $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\mathcal{S} = (S_1, \ldots, S_n)$ is a dataset consisting of n examples drawn independently from P, that is, $\forall i \in [n]$, $S_i \overset{\text{iid}}{\sim} P$. We wish to estimate a parameter $\mu(P)$ (e.g. the mean $\mu = \mathbb{E}_{s\sim P}[s]$) of distribution P. The error of an algorithm $\mathcal{A}$ for this task is measured with respect to a norm $\|\cdot\|$, i.e., $\|\mathcal{A}(S) - \mu\|$.

We use the following short-hand to denote the accuracy guarantees of τ-robust and ε-private estimators of a statistic μ for a distribution P.

Definition 1.1 ((τ, β, α)-robust estimator)**.** Let $\mathcal{A}$ be a (possibly randomized) algorithm for the estimation of statistic μ. We say that $\mathcal{A}$ is a (τ, β, α)-robust estimator for distribution P, if with probability $1 - \beta$ over dataset $\mathcal{S} \overset{\text{iid}}{\sim} P^n$ (and possibly the randomness of the algorithm), for all $\mathcal{S}'$ that differ in at most $n\tau$ points from $\mathcal{S}$, we have that

$$\|\mathcal{A}(\mathcal{S}') - \mu(P)\| \leq \alpha.$$

Definition 1.2 ($(\varepsilon, \beta, \alpha)$-private estimator)**.** Let $\mathcal{A}$ be a (possibly randomized) algorithm for the estimation of statistic μ. We say that $\mathcal{A}$ is an $(\varepsilon, \beta, \alpha)$-private estimator for distribution P, if $\mathcal{A}$ is ε-DP (Definition 2.1) and with probability $1 - \beta$ over $\mathcal{S} \overset{\text{iid}}{\sim} P^n$ (and possibly the randomness of the algorithm), we have that

$$\|\mathcal{A}(\mathcal{S}) - \mu(P)\| \leq \alpha.$$

We may refer to such algorithms as (τ, β, α)-robust and $(\varepsilon, \beta, \alpha)$-private for brevity.

Our main theorem shows that any robust algorithm can be transformed into an ε-DP algorithm with the same accuracy guarantees up to constants, as long as the fraction of corruptions $\tau \approx \frac{d\log(R)}{n\varepsilon}$, where R is the diameter of the range of the robust algorithm.

Theorem 1.3 (Informal, Theorem 3.1)**.** *Let $n \geq 1, \varepsilon \in (0, 1)$. Let P be a distribution over $\mathcal{Z} \subseteq \mathbb{R}^d$. Let $\mathcal{A}_{\text{rob}} : \mathcal{Z}^n \to \{t \in \mathbb{R}^d : \|t\| \leq R\}$ be any (τ, β, α)-robust algorithm for the statistic $\mu(P)$. Let $\alpha_0 \leq \alpha$. If*

$$\tau \gtrsim \frac{d\log(R/\alpha_0) + \log(1/\beta)}{n\varepsilon},$$

then there exists an $(\varepsilon, O(\beta), O(\alpha))$-private algorithm $\mathcal{A}_{\text{priv}}$ for $\mu(P)$. The notation $\gtrsim$ above hides constants.

The main idea behind our transformation is to use the *inverse-sensitivity mechanism* (Asi and Duchi, 2020a). At a high level, for a deterministic algorithm $\mathcal{A}_{\text{rob}}$, the inverse-sensitivity mechanism outputs t with probability (or density) proportional to

$$\Pr[M_{\text{Inv}}(\mathcal{S}; \mathcal{A}_{\text{rob}}) = t] \propto e^{-\mathsf{len}_{\mathcal{A}_{\text{rob}}}(\mathcal{S}; t)\cdot\varepsilon/2}$$

where $\mathsf{len}_{\mathcal{A}_{\text{rob}}}(\mathcal{S}; t)$ is the minimum number of entries of $\mathcal{S}$ that would have to be corrupted to obtain a dataset $\mathcal{S}'$ with $\mathcal{A}_{\text{rob}}(\mathcal{S}') = t$. This mechanism is an instance of the exponential mechanism (McSherry and Talwar, 2007), and to the best of our knowledge the idea to use len as a score function first appeared in (Johnson and Shmatikov, 2013) for

applications in genomic data, and its general accuracy guarantees were first studied systematically in (Asi and Duchi, 2020a;b). A standard analysis shows that this estimator will output t for which $\mathsf{len}_{\mathcal{A}_{\rm rob}}(\mathcal{S}; t)$ is small, and we can relate the accuracy of such a t to the robustness of the estimator $\mathcal{A}_{\rm rob}$ on corruptions of the sample $\mathcal{S}$.

We note that our transformation only preserves the accuracy of the robust mechanism, but not its computational efficiency, and an interesting open question is whether one can get a fully black-box, efficiency-preserving transformation from robustness to privacy (see the concurrent and independent work of (Hopkins et al., 2022b) for some progress on this question).

We also define and analyze an extension of this transformation, which is based on the restricted exponential mechanism of Brown et al. (Brown et al., 2021), that avoids the dependence on R that appears in Theorem 1.3 above, by relaxing the privacy definition to approximate DP.

An equivalence between private and robust estimation. We prove that, up to the factor of d in Theorem 1.3, our transformation is optimal, and in particular is optimal for low-dimensional tasks when d is a constant. That is, for any low-dimensional task, if $\mathcal{A}_{\rm rob}$ is an optimal robust estimator, then our transformation of $\mathcal{A}_{\rm rob}$ is an optimal private estimator. This is the first result to show a general conversion from robust estimators to *optimal* private estimators.

A consequence of this result is an equivalence between robust and private estimation in low dimensions, which shows that the optimal minimax error rates for ε-DP estimation and for τ-robust estimation for $\tau \approx 1/n\varepsilon$ are essentially the same. Specifically, for a given statistic μ and a family of distributions $\mathcal{P}$ over domain $\mathcal{Z}$, we define the minimax error of estimating μ under $\mathcal{P}$ for private and robust algorithms as follows. Having fixed β, τ, and ε, the (τ, β)-robust minimax error under $\mathcal{P}$ is

$$\alpha_{\rm rob}(\tau, \beta) = \inf_{\alpha}\{\alpha : \exists(\tau, \beta, \alpha)\text{-robust algorithm } \forall P \in \mathcal{P}\},$$

and the (ε, β)-private minimax error under $\mathcal{P}$ is

$$\alpha_{\rm priv}(\varepsilon, \beta) = \inf_{\alpha}\{\alpha : \exists(\varepsilon, \beta, \alpha)\text{-private algorithm } \forall P \in \mathcal{P}\}.$$

Theorem 1.4 (Informal, Corollary 4.1)**.** *Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}$ and let μ be a 1-dimensional statistic where $|\mu(P)| \le 1$ for all $P \in \mathcal{P}$. Suppose $\alpha_{\rm rob}(\tau, \beta)$ is a continuous function of β for all $\tau \le 1/2$. Let $n > 1$, $\varepsilon = \omega(\log(n)/n)$ and $\tau = \Theta(\log(n)/n\varepsilon)$. Suppose there exists constant c such that the non-private error $\alpha_{\rm rob}(0, \beta) \ge \frac{1}{n^c}$ for any $\beta \le 1/4$. Then there are constants $c_1 \ge c_2 > 0$ such that for $\beta_p = 1/n^{c_1}$ and $\beta_r = 1/n^{c_2}$,*

$$\alpha_{\rm priv}(\varepsilon, \beta_p) = \Theta\left(\alpha_{\rm rob}(\tau, \beta_r)\right).$$

The above theorem extends to any d-dimensional problem with a weaker conclusion roughly $\alpha_{\rm priv}(\varepsilon, \beta_p) = \Theta\left(d \cdot \alpha_{\rm rob}(\tau, \beta_r)\right)$, so in particular we obtain the same equivalence for any problem in constant dimension.

The theorem follows from the folklore observation that differentially private estimators are also robust (with $\tau \approx 1/\varepsilon n$). Thus, if we have an optimal differentially private algorithm $\mathcal{A}_{\rm priv}$ we can use $\mathcal{A}_{\rm priv}$ itself as the robust estimator. Thus, we can instantiate the inverse-sensitivity mechanism using $\mathcal{A}_{\rm priv}$ as the robust estimator and apply the inverse-sensitivity mechanism to $\mathcal{A}_{\rm priv}$ to obtain a new private mechanism with similar accuracy. Although this transformation would be a circular way to produce a private estimator, the argument shows that one can always obtain an optimal private estimator by instantiating our transformation with an optimal robust estimator.

Applications to high-dimensional private estimation. Although our general optimality result only applies to low-dimensional problems, we show that our transformation often yields optimal error bounds for several high-dimensional tasks as well. By instantiating Theorem 1.3 with existing algorithms for robust estimation, we give ε-DP algorithms for the same tasks. At high level, Theorem 1.3 says that if there is a robust algorithm with accuracy $\alpha(\tau)$ as a function of the fraction of corruptions, then there is an ε-DP algorithm with accuracy $\alpha(D/n\varepsilon)$, where D is the dimension of the parameter we aim to estimate. Since usually $\alpha(\tau) = \tilde{O}(\tau)$, this implies that the error due to privacy is $\tilde{O}(D/n\varepsilon)$. In particular, we apply our theorem to give pure differentially private algorithms for Gaussian (sparse and non-sparse) linear regression and subgaussian PCA, which, to the best of our knowledge, are the first optimal algorithms for these tasks satisfying pure (rather than approximate) differential privacy.

1.2. Related Work

General transformations between robust and private algorithms. Dwork and Lei (2009) were the first to observe the intuitive connection between differential privacy and robust statistics. That work also introduced a generic framework for differentially private algorithms called *propose-test-release (PTR)* that can be used to transform any robust estimator into an approximately DP estimator. However, compared to our optimal transformation, the error of the resulting private algorithm will be larger by a factor of $\approx 1/\varepsilon$. An even earlier work of Nissim, Raskhodnikova, and Smith (Nissim et al., 2007) presented a framework called *smooth sensitivity* that can be used to obtain a similar transformation from robust to pure DP estimators, again losing a factor of $\approx 1/\varepsilon$ compared to our transformation.

In the other direction, Dwork and Lei (2009) also observed

that differentially private algorithms are robust with certain parameters. However, private estimators are mostly studied in a regime where $1/\varepsilon = o(n)$, so they do not give robustness to corrupting a constant fraction of inputs, which is the most commonly studied regime for robust estimation. More recently, Georgiev and Hopkins (2022) observed that private algorithms with sufficiently small failure probability and privacy parameter are robust to corrupting a constant fraction of inputs, and use this fact to give evidence of computational hardness of certain private estimation tasks.

Private estimators inspired by robust estimators. Although prior black-box transformations from robust to private estimators give suboptimal error rates, many optimal private algorithms are nonetheless inspired by methods from robust statistics, albeit with task-specific analyses (Kamath et al., 2019; Bun and Steinke, 2019; Avella-Medina and Brunel, 2019; 2020; Kamath et al., 2020; Liu et al., 2021; Brown et al., 2021; Ghazi et al., 2021; Liu et al., 2022; Tsfadia et al., 2022; Hopkins et al., 2022a; Ashtiani and Liaw, 2022; Kothari et al., 2022; Alabi et al., 2022; Georgiev and Hopkins, 2022). Their algorithms often leverage the structure of specific robust estimators such as medians, high-dimensional generalizations of the median, trimmed-means, or sum-of-squared-based certificates of robustness.

An elegant work of Liu et al. (2022) proposed a generalization of PTR that can be used to give near-optimal approximate differentially private estimators for many tasks. Although the framework is fairly general, their analysis relies on specific properties of the estimation tasks rather than merely the existence of a robust estimator.

Another line of work is that of algorithms that specifically aim to satisfy optimal robustness and privacy guarantees simultaneously for high-dimensional problems (Liu et al., 2021; Ghazi et al., 2021; Alabi et al., 2022).

Concurrent work by Hopkins et al. (2022b). In concurrent and independent work, Hopkins et al. (2022b) proposed the same black-box transformation from robust to DP algorithms based on the smooth-inverse-sensitivity mechanism. In contrast to our work, they also show that in some cases their method can be implemented in a computationally efficient way by instantiating the smooth-inverse-sensitivity mechanism with robust estimators based on the sum-of-squares paradigm. In particular they construct computationally efficient pure DP algorithms for estimating a Gaussian distribution with optimal error. In contrast to their work, we demonstrate that our transformation gives optimal error for low-dimensional problems and establishes a tight connection between privacy and robustness for these problems.

1.3. Organization

In Section 2, we provide background on DP and the inverse-sensitivity mechanism. In Section 3, we present and prove the guarantees of our transformation from robust to pure DP algorithms (and vice versa, for completeness). In Section 4, we show the optimality of our transformation for low-dimensional tasks and the equivalence of robustness and privacy for those tasks. In Section 5, we extend our transformation to convert robust to approximate DP algorithms. In Section 6, we show that our transformation gives us pure DP algorithms with near-optimal error for PCA for subgaussian data, deferring more applications to Appendix C.

2. Preliminaries and Background

Additional Notation For a finite set $\mathcal{T}$, we denote its cardinality by $\mathrm{card}(\mathcal{T})$. For any (continuous or discrete) set $\mathcal{T}$, we denote its diameter by $\mathsf{diam}(\mathcal{T}) = \sup_{s,t\in\mathcal{T}} \|s-t\|$, where the choice of norm will be clear from context. We denote by d_{H} the Hamming distance between two vectors or datasets. We denote by $\mathbb{B}^d(v, R)$ the d-dimensional ball with radius R and center v (with respect to some norm $\|\cdot\|$). We also let $\mathbb{B}^d(R) = \mathbb{B}^d(0^d, R)$ and omit d when it is clear from context.

2.1. Differential Privacy

Let $\mathcal{S}, \mathcal{S}' \in \mathcal{Z}^n$ be two datasets of size n. We say that $\mathcal{S}, \mathcal{S}'$ are *neighboring datasets* if $d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \leq 1$. Differentially private algorithms have indistinguishable output distributions on neighboring datasets.

Definition 2.1 (Differential Privacy (Dwork et al., 2006))**.** A (possibly randomized) algorithm $\mathcal{A}\colon \mathcal{Z}^n \to \mathcal{T}$ is (ε, δ)-*differentially private* (DP) if for all neighboring datasets $\mathcal{S}, \mathcal{S}'$ and any measurable output space $T \subseteq \mathcal{T}$ we have

$$\Pr[\mathcal{A}(\mathcal{S}) \in T] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{S}') \in T] + \delta.$$

We say algorithm $\mathcal{A}$ satisfies *pure* DP if it satisfies the definition for $\delta = 0$, which we denote by ε-DP. Otherwise, we say it satisfies *approximate* DP.

The exponential mechanism (McSherry and Talwar, 2007) is a ubiquitous building block for constructing DP algorithms. The inverse-sensitivity mechanism, on which our transformation is based, is an instantiation of this mechanism.

Definition 2.2 (Exponential Mechanism, (McSherry and Talwar, 2007))**.** Let input data set $\mathcal{S} \in \mathcal{Z}^n$, range $\mathcal{T}$, and score function $q : \mathcal{Z}^n \times \mathcal{T} \to \mathbb{R}$ with sensitivity $\Delta_q = \max_{t\in\mathcal{T}} \max_{\mathcal{S}':d_{\mathrm{H}}(\mathcal{S}',\mathcal{S})\leq 1} |q(\mathcal{S};t) - q(\mathcal{S}';t)|$. The *exponential mechanism* selects and outputs an element $t \in \mathcal{T}$ with probability $\pi_{\mathcal{S}}(t) \propto e^{(\varepsilon\cdot q(\mathcal{S};t)/2\Delta_q)}$. The exponential mechanism is ε-DP.

2.2. Inverse-sensitivity Mechanism

Let $f : \mathcal{Z}^n \to \mathcal{T}$ be a (deterministic) algorithm that we aim to compute on dataset $\mathcal{S}$. Define the path-length function

$$\mathsf{len}_f(\mathcal{S}; t) \coloneqq \inf_{S'} \{d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \mid f(\mathcal{S}') = t\},$$

which is the minimum number of points in $\mathcal{S}$ that need to be replaced so that the value of function f on the modified input is t. Given black-box access to function f, the *inverse-sensitivity mechanism* with input dataset $\mathcal{S}$, denoted by $M_{\mathrm{Inv}}(\mathcal{S}; f)$, is then defined as follows: the probability that $M_{\mathrm{Inv}}(\mathcal{S}; f)$ returns $t \in \mathcal{T}$ is

$$\Pr[M_{\mathrm{Inv}}(\mathcal{S}; f) = t] \coloneqq \frac{e^{-\mathsf{len}_f(\mathcal{S};t)\varepsilon/2}}{\sum_{s\in\mathcal{T}} e^{-\mathsf{len}_f(\mathcal{S};s)\varepsilon/2}}.$$

The error of this mechanism in some norm $\|\cdot\|$ depends on the *local modulus of continuity* of a function $f : \mathcal{Z}^n \to \mathcal{T}$ at $\mathcal{S} \in \mathcal{Z}^n$ with respect to $\|\cdot\|$, defined by

$$\omega_f(\mathcal{S}; K) = \sup_{\mathcal{S}'\in\mathcal{Z}^n} \{\|f(\mathcal{S}) - f(\mathcal{S}')\| : d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \le K\}.$$

For a finite set $\mathcal{T}$, the inverse-sensitivity mechanism has the following guarantees.

Theorem 2.3 (Discrete functions, Th.3 (Asi and Duchi, 2020a))**.** *Let $f : \mathcal{Z}^n \to \mathcal{T}$ and $D = \mathsf{diam}(\mathcal{T})$. Then for any $\mathcal{S} \in \mathcal{Z}^n$ and $\beta > 0$, with probability at least $1 - \beta$, the inverse-sensitivity mechanism has error*

$$\|M_{\mathrm{Inv}}(\mathcal{S}) - f(\mathcal{S})\| \le \omega_f\left(\mathcal{S}; \frac{2}{\varepsilon}\log\frac{2D\,\mathrm{card}(\mathcal{T})}{\beta\varepsilon}\right).$$

For continuous functions $f : \mathcal{Z}^n \to \mathbb{R}^d$, which is the main setting in this paper, we use a smooth version of the inverse-sensitivity mechanism. To this end, we define the ρ-smooth inverse-sensitivity of t with respect to norm $\|\cdot\|$:

$$\mathsf{len}^\rho(\mathcal{S}; t) = \inf_{s\in\mathbb{R}^d:\|s-t\|\le\rho} \mathsf{len}(\mathcal{S}; s). \tag{1}$$

The ρ-smooth inverse-sensitivity mechanism $M^\rho_{\mathrm{Inv}}(\cdot; f)$ then has the following density given input dataset $\mathcal{S}$:

$$\pi_{\mathcal{S}}(t) = \frac{e^{-\mathsf{len}^\rho(\mathcal{S};t)\varepsilon/2}}{\int_{s\in\mathbb{R}^d} e^{-\mathsf{len}^\rho(\mathcal{S};s)\varepsilon/2} ds} \tag{2}$$

For our setting of interest where $\mathcal{T} = \{v \in \mathbb{R}^d : \|v\| \le R\}$, we have the following upper bound on the error of M^ρ_{Inv}. Its proof follows similar ideas as in (Asi and Duchi, 2020a;b) and is in Appendix A .

Theorem 2.4 (Continuous functions)**.** *Let $f : \mathcal{Z}^n \to \mathcal{T}$ where $\mathcal{T} = \{v \in \mathbb{R}^d : \|v\| \le R\}$. Then for any $\mathcal{S} \in \mathcal{Z}^n$, and $\beta > 0$, with probability at least $1 - \beta$, the ρ-smooth inverse-sensitivity mechanism with norm $\|\cdot\|$ has error*

$$\|M^\rho_{\mathrm{Inv}}(\mathcal{S}; f) - f(\mathcal{S})\| \le \omega_f\left(\mathcal{S}; \frac{2\left(d\log\left(\frac{R}{\rho}+1\right)+\log\frac{1}{\beta}\right)}{\varepsilon}\right) + \rho.$$

3. Transformations between Robust and Private Algorithms

In this section, we provide transformations between differentially private and robust algorithms. We begin in Section 3.1 with our main result: a general transformation from robust algorithms to private algorithms with roughly the same error for a specified number of corruptions. In Section 3.2, we consider the other direction, and show for completeness that any private algorithm is inherently robust as well, which was already observed since (Dwork and Lei, 2009).

3.1. Robust to Private

Our first result shows how to transform a determinitstic robust algorithm into a private algorithm with roughly the same error. The main idea is to apply the ρ-smooth inverse-sensitivity mechanism (Asi and Duchi, 2020a) with the input function f being the robust algorithm itself.

Theorem 3.1 (Robust-to-private)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ such that $\mu(P) \in \mathbb{R}^d$. Let $\varepsilon, \beta \in (0, 1)$. Let $\mathcal{A}_{\mathrm{rob}} : (\mathbb{R}^d)^n \to \{t \in \mathbb{R}^d : \|t\| \le R\}$ be a deterministic (τ, β, α)-robust algorithm. Let $\alpha_0 \le \alpha$ and*

$$\tau^\star = \frac{2\left(d\log\left(\frac{R}{\alpha_0}+1\right)+\log\frac{1}{\beta}\right)}{n\varepsilon}. \tag{3}$$

If $\tau \ge \tau^\star$, then $M^\rho_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}})$ with $\rho = \alpha_0$ is ε-DP and, with probability at least $1 - 2\beta$, has error $\|M^\rho_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}) - \mu\| \le 4\alpha$. In particular, this implies that for $\tau \ge \frac{2\left(d\log\left(\frac{R}{\alpha_0}+1\right)+\log\frac{2}{\beta}\right)}{n\varepsilon}$, $\alpha_{\mathrm{priv}}(\varepsilon, \beta) \le 4\alpha_{\mathrm{rob}}(\tau, \beta/2)$.

Proof. First note that the privacy guarantee is immediate from the guarantees of the exponential mechanism (Definition 2.2) and the fact that the sensitivity of the ρ-smooth path-length function in Equation (1) is 1. Now we prove utility. Let $K = n\tau^\star$. The error of $M^\rho_{\mathrm{Inv}}(\mathcal{S}, \mathcal{A}_{\mathrm{rob}})$ is then bounded as follows:

$$\begin{aligned}
&\|M^\rho_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}) - \mu\| \\
&\le \|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) - \mu\| + \|M^\rho_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}) - \mathcal{A}_{\mathrm{rob}}(\mathcal{S})\| \\
&\le \|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) - \mu\| + \omega_{\mathcal{A}_{\mathrm{rob}}}(\mathcal{S}; K) + \alpha_0 \\
&\qquad\qquad \text{(w.p. } 1-\beta \text{ by Theorem 2.4)} \\
&\le 2\|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) - \mu\| + \|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}') - \mu\| + \alpha_0,
\end{aligned}$$

for $\mathcal{S}' = \operatorname{argmax}_{\mathcal{S}':d_{\mathrm{H}}(\mathcal{S}',\mathcal{S})\le K} \|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) - \mathcal{A}_{\mathrm{rob}}(\mathcal{S}')\|$. Recall that, by assumption, $\mathcal{A}_{\mathrm{rob}}$ is (τ,β,α)-robust for $\tau \ge K/n$. Thus, with probability $1-\beta$, $\|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}') - \mu\| \le \alpha$ for any τ-corrupted dataset $\mathcal{S}'$. By union bound, we have that with probability $1-2\beta$, $\|M^{\rho}_{\mathrm{Inv}}(\mathcal{S};\mathcal{A}_{\mathrm{rob}}) - \mu\| \le 3\alpha+\alpha_0 \le 4\alpha$. This completes the proof of the theorem. □

The parameter α_0 determines the smallest fraction of corruptions $\tau^\star$, which in turn determines the smallest α so that $\mathcal{A}_{\mathrm{rob}}$ is $(\tau^\star,\beta,\alpha)$-robust. A simple choice for α_0 is the minimax error for estimating the statistic $\mu(P)$, without adversarial corruptions or privacy constraints.

Remark 3.2. We can extend this transformation to hold for randomized robust algorithms, by first converting $\mathcal{A}_{\mathrm{rob}}$ into a deterministic algorithm, albeit doubling the error and failure probability, as shown in Theorem A.1, Appendix A.1.

3.2. Private to Robust

In this section, we state the folklore fact that any ε-differentially private algorithm is also $\tau \approx \frac{1}{n\varepsilon}$-robust with the same accuracy. This follows directly from the definition of differential privacy which states that changing $1/\varepsilon$ users does not change the output distribution by much (by *group privacy*) and was observed in (Dwork and Lei, 2009).

Theorem 3.3 (Private-to-robust)**.** *Let $\mathcal{S} = (S_1,\dots,S_n)$ where $S_i \overset{\text{iid}}{\sim} P$. Let $\varepsilon,\beta \in (0,1)$. Let $\mathcal{A}_{\mathrm{priv}}$ be an $(\varepsilon,\beta,\alpha)$-private algorithm for estimating the statistic μ. Let $\gamma \in (0,1)$ and $\tau = \frac{\log(1/\gamma)}{n\varepsilon}$. Then $\mathcal{A}_{\mathrm{priv}}$ is $(\tau,\beta/\gamma,\alpha)$-robust. In particular, $\alpha_{\mathrm{rob}}(\tau,\beta/\gamma) \le \alpha_{\mathrm{priv}}(\varepsilon,\beta)$, which is equivalent to $\alpha_{\mathrm{rob}}(\tau,\beta) \le \alpha_{\mathrm{priv}}(\log(1/\gamma)/n\tau, \gamma\beta)$.*

Proof. Let $W = \{t \in \mathbb{R}^d : \|t-\mu\| > \alpha\}$ be the set of bad outputs for the distribution P. The accuracy guarantee of $\mathcal{A}_{\mathrm{priv}}$ implies that $\Pr[\mathcal{A}_{\mathrm{priv}}(\mathcal{S}) \in W] \le \beta$. Now assume $\mathcal{S}'$ is a τn-corrupted version of $\mathcal{S}$ where $\tau = \log(1/\gamma)/(n\varepsilon)$, that is, $d_{\mathrm{H}}(\mathcal{S},\mathcal{S}') \le \log(1/\gamma)/\varepsilon$. The definition of differential privacy now implies that

$$\begin{aligned}\Pr[\|\mathcal{A}_{\mathrm{priv}}(\mathcal{S}') - \mu\| > \alpha] &= \Pr[\mathcal{A}_{\mathrm{priv}}(\mathcal{S}') \in W] \\ &\le e^{d_{\mathrm{H}}(\mathcal{S},\mathcal{S}')\varepsilon} \Pr[\mathcal{A}_{\mathrm{priv}}(\mathcal{S}) \in W] \\ &\le e^{\varepsilon\tau n}\beta \le \beta/\gamma,\end{aligned}$$

for $\tau \le \frac{\log(1/\gamma)}{n\varepsilon}$. Thus $\mathcal{A}_{\mathrm{priv}}$ is τ-robust for $\tau = \frac{\log(1/\gamma)}{n\varepsilon}$ with accuracy α and failure probability β/γ. □

We note that for $\gamma = 1/e$, $\alpha_{\mathrm{rob}}(\tau,\beta) \le \alpha_{\mathrm{priv}}(1/n\tau, \beta/e)$, that is, the minimax error of any τ-robust algorithm with failure probability β is bounded by the minimax error of a ε-DP algorithm, for $\varepsilon = 1/n\tau$, with the same failure probability up to constant factors.

As private algorithms are often randomized, this transformation would result in a randomized robust algorithm. As in the previous section, we can convert it into a deterministic one via Theorem A.1 in Appendix A.1.

4. Implications of our transformations

4.1. Equivalence between Private and Robust Estimation

In this section, we show that the (high-probability) minimax rates for ε-DP and τ-robustness are on the same order when the problem is low-dimensional and $\tau = \log(n)/n\varepsilon$. The following corollary states the result. For simplicity, we assume that the dimension $d = 1$ and the range $R = 1$.

Corollary 4.1 (Equivalence)**.** *Let $\mathcal{P}$ be a family of distributions and $P \in \mathcal{P}$. Let μ be a 1-dimensional statistic where $|\mu(P)| \le 1$ such that $\alpha_{\mathrm{rob}}(\tau,\beta)$ is a continuous function of β for all $\tau \le 1/2$. Let $n > 1$, $\varepsilon = \omega(\log(n)/n)$, and $\tau = \Theta(\log(n)/n\varepsilon)$. Suppose there exists a constant c such that the error $\alpha_{\mathrm{rob}}(0,\beta) \ge \frac{1}{n^c}$ for any $\beta \le 1/4$. Then there are constants $c_1 \ge c_2 > 0$ such that for $\beta_p = 1/n^{c_1}$ and $\beta_r = 1/n^{c_2}$,*

$$\alpha_{\mathrm{priv}}(\varepsilon,\beta_p) = \Theta\left(\alpha_{\mathrm{rob}}(\tau,\beta_r)\right).$$

Proof. First, we observe that $\alpha_0 = \alpha_{\mathrm{rob}}(0,\beta_p) \le \alpha_{\mathrm{rob}}(\tau,\beta_p)$ by the monotonicity of α_{rob} and $\alpha_0 \ge \frac{1}{n^c}$ since $\beta_p = \frac{1}{n^{c_1}} \le 1/2$. By Theorem 3.1[1], we have that, for $\tau_1 = \frac{2(c+c_1)\log(2n)}{n\varepsilon} \ge \frac{2\log(1/\alpha_0+1)+\log(2/\beta_p)}{n\varepsilon}$,

$$\alpha_{\mathrm{priv}}(\varepsilon,\beta_p) \le 4\alpha_{\mathrm{rob}}(\tau_1,\beta_p/2). \tag{4}$$

Setting $\gamma = 1/(2n)^{2(c+c_1)}$, we have that $\tau_1 = \frac{\log(1/\gamma)}{n\varepsilon}$. By Theorem 3.3,

$$\alpha_{\mathrm{rob}}(\tau_1,(2n)^{2(c+c_1)}\beta) \le \alpha_{\mathrm{priv}}(\varepsilon,\beta).$$

Note that if $\alpha_{\mathrm{priv}}(\varepsilon,\beta_p) \ge \alpha_{\mathrm{rob}}(\tau_1,\beta_p/2)$ then the claim follows from Equation (4), using $\beta_r = \beta_p/2$. Otherwise we have that

$$\alpha_{\mathrm{rob}}(\tau_1,(2n)^{2(c+c_1)}\beta_p) \le \alpha_{\mathrm{priv}}(\varepsilon,\beta_p) \le \alpha_{\mathrm{rob}}(\tau_1,\beta_p/2)$$

As $\alpha_{\mathrm{rob}}(\tau_1,\beta)$ is a continuous function of β, there is $\beta_r \in [\beta_p/2,(2n)^{2(c+c_1)}\beta_p] \subset [\frac{1}{n^{2c_1}},\frac{1}{n^{2c+3c_1}}]$ such that $\alpha_{\mathrm{rob}}(\tau,\beta_r) = \alpha_{\mathrm{priv}}(\varepsilon,\beta_p)$. □

Note that in most settings, $\alpha_{\mathrm{rob}}(\tau,\beta_r)$ has the same order when $\beta_r \in [\beta_p,\mathrm{poly}(n)\cdot\beta_p]$ as it depends on $\log(1/\beta)$ (see for example Section 6).

[1] If the robust algorithm achieving the minimax error α_{rob} is randomized, we can transform it into a deterministic one as Theorem 3.1 requires, via Theorem A.1, by losing only constant factors.

Algorithm 1 Robust-to-Private ((ε, δ)-DP)

Require: $\mathcal{S} = (S_1, \ldots, S_n)$, (τ, β, α)-robust algorithm $\mathcal{A}_{\text{rob}}$, local modulus bound B

Let $K = n\tau/2 - 1$
Let $S_{\mathsf{bad}} = \{\mathcal{S} \in \mathcal{Z}^n : \omega_f(\mathcal{S}; K+1) > B\}$
Calculate $d = \min_{\mathcal{S}' \in S_{\mathsf{bad}}} d_{\text{H}}(\mathcal{S}, \mathcal{S}')$
Set $\hat{d} = d + \zeta$ where $\zeta \sim \mathsf{Laplace}(2/\varepsilon)$
if $\hat{d} > 2\log(1/\min(\delta, \beta))/\varepsilon$ **then**
    Sample t from the truncated inverse-sensitivity mechanism with threshold K, privacy parameter $\varepsilon/2$, smoothness parameter $\rho = 2B$, and return t.
else
    Return $\perp$
end if

4.2. Optimality of Black-Box Transformation

An immediate corollary of the previous transformations is that—for some choice of robust algorithm $\mathcal{A}_{\text{rob}}$—our robust-to-private transformation achieves the minimax optimal rate among the family of private algorithms, for low-dimensional statistics. See Appendix A for its proof.

Corollary 4.2 (Optimality). *Let $\mathcal{P}$ be a family of distributions and $P \in \mathcal{P}$. Let μ be a 1-dimensional statistic where $|\mu(P)| \le 1$. Let $\alpha_{\text{priv}}(\varepsilon, \beta)$ be the minimax error of any ε-DP algorithm with failure probability $\beta \le 1/4$ that estimates statistic $\mu(P)$. Let $n > 1$. Suppose that there exists a constant c such that the non-private error $\alpha_{\text{priv}}(\infty, \beta) \ge \frac{1}{n^c}$ for any $\beta \le 1/2$. Then there are constants $c_1 \ge c_2 > 0$ such that $\beta_p = 1/n^{c_1}$ and $\beta'_p = 1/n^{c_2}$, robust algorithm $\mathcal{A}_{\text{rob}}$, and a choice of ρ, such that Algorithm $M^{\rho}_{\text{Inv}}(\cdot; \mathcal{A}_{\text{rob}})$ with privacy parameter ε achieves the optimal error $O(\alpha_{\text{priv}}(\varepsilon, \beta_p))$ with probability $1 - \beta'_p$.*

5. A Transformation for Approximate DP

In this section, we propose a different transformation for (ε, δ)-DP that avoids the necessary dependence on diameter for pure ε-DP , by using the following truncated version of the inverse-sensitivity:

$$\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S}; t) := \begin{cases} \mathsf{len}^{\rho}_f(\mathcal{S}; t) & \text{if } \mathsf{len}^{\rho}_f(\mathcal{S}; t) \le K \\ \infty & \text{otherwise.} \end{cases}$$

Algorithm 1 uses a private test to check whether $\mathcal{S}$ is far from the set $S_{\mathsf{bad}} = \{\mathcal{S} \in \mathcal{Z}^n : \omega_f(\mathcal{S}; K+1) > B\}$. If it is not, then it fails. Crucially, if $\mathcal{S} \overset{\text{iid}}{\sim} P^n$ then the robust algorithm guarantees that $\omega_f(\mathcal{S}'; K+1) \le B$ for $\mathcal{S}'$ in a neighborhood of $\mathcal{S}$, allowing the test to succeed in this case (Theorem 5.1, proven in Appendix B).

Theorem 5.1 (Robust-to-private, approximate DP). *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ such that $\mu(P) \in \mathbb{R}^d$. Let $\varepsilon, \delta, \beta \in (0, 1)$. Let $\mathcal{A}_{rob} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ be a deterministic (τ, β, α)-robust algorithm for the statistic μ. If $\tau \ge \frac{8(d + \log(1/\min\{\delta, \beta\}))}{n\varepsilon}$ then Algorithm 1 with $B = 2\alpha$ and $\rho = 2B$ is (ε, δ)-DP and, with probability at least $1 - 2\beta$ returns $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le 7\alpha$.*

6. Applications for Pure DP

6.1. Principal Component Analysis

In this section, we apply our transformation to obtain a pure DP algorithm for PCA under Gaussian data. We note that the result holds as-is for subgaussian distributions more generally, because Theorem 6.3 (Jambulapati et al., 2020) does. We assume w.l.o.g. that the distribution is zero-mean.

To the best of our knowledge, Corollary 6.1 gives the first (computationally inefficient) algorithm for pure DP with error $\tilde{O}(\frac{d}{n\varepsilon})$. PCA with a spectral gap has been studied under pure DP in (Chaudhuri et al., 2013), where the result can be translated to yield a suboptimal error of $\tilde{O}(\frac{d^2}{n\varepsilon})$. A long line of work studies PCA under approximate DP (Blum et al., 2005; Hardt and Roth, 2012; 2013; Chaudhuri et al., 2013; Kapralov and Talwar, 2013; Dwork et al., 2014) with the recent result by Liu et al. (2022) achieving the optimal error of $\tilde{O}(\frac{d}{n\varepsilon})$ for subgaussian distributions.

Corollary 6.1 (Gaussian PCA). *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \Sigma)$. Let $\varepsilon, \beta \in (0, 1)$. Suppose n is such that $\alpha \le 1$ in Equation (5). There exists a constant $C > 0$ and an ε-DP algorithm $\mathcal{M}$ such that, with probability at least $1 - \beta$, returns unit vector $\mathcal{M}(\mathcal{S}) = v$ such that $1 - \frac{v^\top \Sigma v}{\|\Sigma\|_2} \le \alpha$ for*

$$\alpha = C\left(\sqrt{\frac{d + \log(\frac{1}{\beta})}{n}} + \frac{d\log^2(\frac{n}{d}) + \log(\frac{1}{\beta})\log(\frac{n}{d})}{n\varepsilon}\right). \tag{5}$$

We will need a more general transformation, proven in Appendix A.2, and stated in Theorem 6.2below.

Theorem 6.2 (Robust-to-private, general loss). *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ such that $\mu(P) \in \mathbb{R}^d$. Let $\varepsilon, \beta \in (0, 1)$. Let $L : (\mathbb{R}^d)^2 \to \mathbb{R}$ be a loss function which satisfies the triangle inequality. Let $\mathcal{A}_{\text{rob}} : (\mathbb{R}^d)^n \to \{t \in \mathbb{R}^d : \|t\| \le R\}$ be a (deterministic) (τ, β, α)-robust algorithm with respect to L. Let $\alpha_0 \le \alpha$. Suppose n is such that the smallest value τ satisfying Equation (6) is at most 1. Suppose for all $u, v \in \mathbb{B}(R + \alpha_0)$, $L(u, v) \le c_L \|u - v\|$ for some constant c_L. If*

$$\tau \ge \frac{2\left(d\log\left(\frac{R}{\alpha_0} + 1\right) + \log\frac{1}{\beta}\right)}{n\varepsilon}, \tag{6}$$

then Algorithm $M^{\rho}_{\text{Inv}}(\mathcal{S}; \mathcal{A}_{\text{rob}})$ with $\rho = \alpha_0$ in norm $\|\cdot\|$ is

ε-DP and, with probability at least $1-2\beta$, has error

$$L\left(M^{\rho}_{\mathrm{Inv}}(\mathcal{S};\mathcal{A}_{\mathrm{rob}}),\mu\right) \le (3+c_L)\alpha = O(\alpha).$$

We will instantiate our transformation with the robust PCA algorithm from (Jambulapati et al., 2020). An alternative with the same guarantees is returning the unit vector that minimizes the surrogate cost function proposed by (Liu et al., 2022).

Theorem 6.3 (Theorem 1, (Jambulapati et al., 2020))**.** *Let $\mathcal{S} = (S_1,\dots,S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(0,\Sigma)$. Let $\beta \in (0,1), \tau \in (0,1/2)$. Let*

$$\alpha' = C'\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \tau\log\left(\frac{1}{\tau}\right)\right),$$

for a known constant $C'>0$. There exists algorithm $\mathcal{A}_{\mathrm{rob}}$ which is (τ,β,α')-robust.

Proof of Corollary 6.1. We define the following loss function $L(u,v) = \frac{u^\top\Sigma u}{\|\Sigma\|_2} - \frac{v^\top\Sigma v}{\|\Sigma\|_2}$. If v_1 is the top eigenvector of Σ, then our goal is to return vector v with small error $L(v_1,v)$. Let $\alpha_0 = C'\sqrt{\frac{d+\log(1/\beta)}{n}}$ and assume it is less than 1, to be confirmed later. Then L satisfies the triangle inequality and for all $u,v \in \mathbb{B}(1+\alpha_0) \subset \mathbb{B}(2)$,

$$\begin{aligned}
L(u,v) &= \frac{\left\|\Sigma^{1/2}u\right\|_2^2}{\|\Sigma\|_2} - \frac{\left\|\Sigma^{1/2}v\right\|_2^2}{\|\Sigma\|_2} \\
&= \frac{\left(\left\|\Sigma^{1/2}u\right\|_2 - \left\|\Sigma^{1/2}v\right\|_2\right)\cdot\left(\left\|\Sigma^{1/2}u\right\|_2 + \left\|\Sigma^{1/2}v\right\|_2\right)}{\|\Sigma\|_2} \\
&\le \|u-v\|_2\cdot(\|u\|_2+\|v\|_2) \le 4\|u-v\|_2.
\end{aligned}$$

Let $\mathcal{A}_{\mathrm{rob}} : (\mathbb{R}^d)^n \to \mathbb{S}^{d-1}$ be the algorithm established by Theorem 6.3, where $\mathbb{S}^{d-1}$ denotes the unit sphere. Then L satisfies the requirements of Theorem 6.2. For $\tau = \frac{2d\log(n/d)+2\log(1/\beta)}{n\varepsilon}$, $\mathcal{A}_{\mathrm{rob}}$ is (τ,β,α')-robust with

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

We then have that $M^{\rho}_{\mathrm{Inv}}(\cdot,\mathcal{A}_{\mathrm{rob}})$ with $\rho=\alpha_0$ is ε-DP and with probability $1-2\beta$, returns $v \in \mathbb{B}(1+\alpha_0)$, such that

$$L(v_1,v) = 1 - \frac{v^\top\Sigma v}{\|\Sigma\|_2} \le 7\alpha'.$$

Let $\hat{v} = \frac{v}{\|v\|_2}$ be the unit vector in the direction of v. We have that $L(v_1,\hat{v}) = 1 - \frac{v^\top\Sigma v}{\|\Sigma\|_2\|v\|_2^2}$. If $\|v\|_2 \le 1$ then $L(v_1,\hat{v}) \le L(v_1,v)$. Suppose $\|v\|_2 > 1$.

$$\begin{aligned}
L(v_1,\hat{v}) &= \frac{1}{\|v\|_2^2}\left(\|v\|_2^2 - \frac{v^\top\Sigma v}{\|\Sigma\|_2}\right) \\
&= \frac{1}{\|v\|_2^2}\left((\|v\|_2^2-1) + L(v_1,v)\right) \\
&\le \frac{\|v\|_2^2-1}{\|v\|_2^2} + L(v_1,v) \qquad (\text{since } \|v\|_2 > 1) \\
&\le \frac{\alpha_0(\alpha_0+2)}{(\alpha_0+1)^2} + L(v_1,v) \\
&\qquad\qquad (\text{since } (x-1)/x \nearrow) \\
&\le 2\alpha_0 + L(v_1,v) \le 9\alpha',
\end{aligned}$$

since $\alpha_0 \le \alpha'$. Therefore, we return a unit vector $\hat{v}$ with $1 - \frac{\hat{v}^\top\Sigma\hat{v}}{\|\Sigma\|_2} \le \alpha$ for $\alpha = 9\alpha'$. By assumption n is sufficiently large so that $\alpha \le 1$, and as such $\alpha' < 1$. The proof is complete by rescaling $\beta \leftarrow \beta/2$ and adjusting the constants. □

6.2. More Applications

We apply our transformation to Gaussian mean and covariance estimation, instantiated by the Tukey median and the robust algorithm by (Diakonikolas et al., 2017) respectively, retrieving the known near-optimal error. We also apply it to Gaussian linear regression (Corollary 6.4 below), instantiated by the robust algorithm by Gao (2020) to give the first algorithm with optimal error under pure DP.

Corollary 6.4 (Gaussian Linear Regression)**.** *Let $\mathcal{S} = (S_1,\dots,S_n)$ where for all $i \in [n]$, $S_i = (X_i,y_i) \in \mathbb{R}^d\times\mathbb{R}$ is generated by a linear model $y_i = X_i^\top\theta + \eta_i$ for some unknown $\theta \in \mathbb{B}^d(R)$, where $X_i \overset{\text{iid}}{\sim} \mathsf{N}(0,\Sigma)$, $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, and $\eta_i \overset{\text{iid}}{\sim} \mathsf{N}(0,\sigma^2)$, independent from X_i. Let $\varepsilon,\beta \in (0,1)$. There exists an ε-DP algorithm $\mathcal{M}$ such that, with probability $1-\beta$, $\left\|\Sigma^{-1/2}(\mathcal{M}(\mathcal{S})-\theta)\right\|_2 \le \alpha\sigma$ for*

$$\alpha = O\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \frac{d\log\left(\frac{(R/\sigma+\sqrt{\kappa})n}{d}\right)+\log(\frac{1}{\beta})}{n\varepsilon}\right).$$

We extend our main transformation to handle sparse estimation in Appendix A.3, which allows us to prove the equivalent result for the case of sparse linear regression where $\|\theta\|_0 \le k$. We show that in this case, the error is in the order of $\sqrt{k/n} + k/(n\varepsilon)$, as expected. All remaining statements and proofs are in Appendix C.

7. Conclusions and Future Work

We gave the first black-box transformation that converts an arbitrary robust algorithm into a differentially private one. We proved that this transformation gives an optimal

strategy for designing a differentially private algorithm for low-dimensional tasks, and that the minimax errors for robustness and privacy are equivalent for these tasks.

We also showed that this transformation often gives near-optimal error rates for several canonical high-dimensional tasks under (sub)Gaussian distributions (including under sparsity assumptions) and expect that it achieves similar results under other families of distributions, such as heavy-tailed. In particular, we note that the dependence on dimension d cannot be improved in the general setting. This follows from the optimal rates we obtain in our applications using this transformation, as any further improvements in the dependence on d in the general setting will result in a contradiction to existing lower bounds for the applications we consider. However, it would still be interesting to explore specialized settings where this dependence can be improved (for example, as we show for the case of sparse linear regression), as well as to determine the conditions under which this or another black-box transformation yields private algorithms with optimal error in high dimensions.

A drawback of our transformation is that it produces a computationally inefficient algorithm, even if the robust algorithm we instantiate it with is computationally efficient. However, there are some approaches which allow for efficient implementation of this transformation. One approach is to use accurate approximations for the inverse sensitivity mechanism that can be implemented efficiently in certain settings such as PCA and linear regression, as in (Asi and Duchi, 2020b). Moreover, as the inverse sensitivity is an application of the exponential mechanism with a specific score function, it is possible to use existing results (Hopkins et al., 2022a) which can be applied as long as the score function satisfies certain properties. Specifically, Hopkins et al. (2022b) use the sum-of-squares paradigm, to make this transformation computationally efficient specifically for the task of Gaussian estimation in TV distance, an approach that has been recently successful when applied to several problems (Hopkins et al., 2022a; Ashtiani and Liaw, 2022; Kothari et al., 2022; Alabi et al., 2022).

Acknowledgements

We thank Chao Gao for helpful discussions and clarifications regarding results in the robustness literature and the anonymous ICML reviewers for helpful comments that improved the presentation of our paper. JU and LZ were supported by NSF awards CCF-1750640, CNS-1816028, and CNS-2120603. LZ was additionally supported by a Facebook (Meta) Fellowship.

References

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

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, 2022. URL https://arxiv.org/abs/2203.16654.

Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT 2021)*, ALT '21. JMLR, Inc., March 2021. URL https://arxiv.org/abs/2010.09929.

Daniel Alabi, Pravesh K. Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a gaussian: Efficient, robust and optimal, 2022. URL https://arxiv.org/abs/2212.08018.

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

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1075–1076. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/ashtiani22a.html.

Hilal Asi and John C. Duchi. Near instance-optimality in differential privacy, May 2020a. URL https://arxiv.org/abs/2005.10630.

Hilal Asi and John C. Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14106–14117. Curran Associates, Inc., Dec 2020b. URL https://proceedings.neurips.cc/paper/2020/file/a267f936e54d7c10a2bb70dbe6ad7a89-Paper.pdf.

Marco Avella-Medina and Victor-Emmanuel Brunel. Differentially private sub-gaussian location estimators. *arXiv preprint arXiv:1906.11923*, 2019.

Marco Avella-Medina and Victor-Emmanuel and Brunel. Propose, test, release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774*, 2020.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Mach. Learn.*, 94(3):401–437, mar 2014. ISSN 0885-6125. doi: 10.1007/s10994-013-5404-1. URL https://doi.org/10.1007/s10994-013-5404-1.

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

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *Proceedings of the 24th Annual ACM Symposium on Principles of Database Systems*, PODS '05, pages 128–138, Baltimore, MD, USA, 2005. ACM.

Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7950–7964. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/42778ef0b5805a96f9511e20b5611fce-Paper.pdf.

Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Michael A Burr and Robert J Fabrizio. Uniform convergence rates for halfspace depth. *Statistics & Probability Letters*, 124:33–40, 2017.

Clément Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private identity testing for high dimensional distributions. In *Advances in Neural Information Processing Systems*, NeurIPS '20, 2020. https://arxiv.org/abs/1905.11947.

Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *J. Mach. Learn. Res.*, 14(1):2905–2943, jan 2013. ISSN 1532-4435.

Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 727–757. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/cheng19a.html.

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

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

David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827, 1992.

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

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

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

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

Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139 – 1170, 2020. doi: 10.3150/

19-BEJ1144. URL https://doi.org/10.3150/19-BEJ1144.

Kristian Georgiev and Samuel B. Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=g-OkeNXPy-X.

Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thao Nguyen. Robust and private learning of halfspaces. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1603–1611. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/ghazi21a.html.

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

Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. STOC '12, page 1255–1268, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312455. doi: 10.1145/2213977.2214088. URL https://doi.org/10.1145/2213977.2214088.

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 331–340, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488650. URL https://doi.org/10.1145/2488608.2488650.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC*, 2010.

Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. STOC 2022, page 1406–1417, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519947. URL https://doi.org/10.1145/3519935.3519947.

Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation, 2022b. URL https://arxiv.org/abs/2212.05015.

Peter J. Huber. A Robust Version of the Probability Ratio Test. *The Annals of Mathematical Statistics*, 36(6):1753 – 1758, 1965. doi: 10.1214/aoms/1177699803. URL https://doi.org/10.1214/aoms/1177699803.

Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15689–15701. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b58144d7e90b5a43edcce1ca9e642882-Paper.pdf.

Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 1079–1087, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2487687. URL https://doi.org/10.1145/2487575.2487687.

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

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. *https://arxiv.org/abs/2002.09464*, 2020.

Michael Kapralov and Kunal Talwar. *On differentially private low rank approximation*, pages 1395–1414. 2013. doi: 10.1137/1.9781611973105.101. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973105.101.

Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 723–777. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/kothari22a.html.

Jerry Li. Lecture notes in robustness in machine learning (cse 599-m), 2019. URL https://jerryzli.github.io/robust-ml-fall19.html.

Jerry Li and Guanghao Ye. Robust gaussian covariance estimation in nearly-matrix multiplication time. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12649–12659. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9529fbba677729d3206b3b9073d1e9ca-Paper.pdf.

Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3887–3901. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/1fc5309ccc651bf6b5d22470f67561ea-Paper.pdf.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1167–1246. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/liu22b.html.

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

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *ACM Symposium on Theory of computing*, pages 75–84, 2007.

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

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

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. FriendlyCore: Practical differentially private aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21828–21863. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tsfadia22a.html.

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

Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

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

A. Additional Proofs for Transformations

A.1. Randomized to Deterministic Robust Algorithm

Here, we present a transformation from a randomized algorithm $\mathcal{A}$ to a *deterministic* robust algorithm whose error and failure probability are larger by a factor of 2. Intuitively, Algorithm 2 finds a small ball where the randomized algorithm has the largest density, and returns its center. This transformation is computationally inefficient because it requires running the randomized algorithm with all possible choices of random coins.

Algorithm 2 Randomized-to-Deterministic Robust

Require: $\mathcal{S} = (S_1, \ldots, S_n)$, Algorithm $\mathcal{A}$, accuracy α.
1: Let $P_{\mathcal{S}}$ denote the probability distribution of $\mathcal{A}(\mathcal{S})$ over the randomness of the algorithm
2: Find the center $v^\star$ of a ball $\mathbb{B}(v) = \{u \in \mathbb{R}^d : \|u - v\| \leq \alpha\}$ of radius α that maximizes $P_{\mathcal{S}}(\mathbb{B}(v))$
3: Return $v^\star$

We have the following guarantees for the transformation in Algorithm 2.

Theorem A.1 (Randomized-to-deterministic-robust)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$. Let $\tau, \beta \in (0, 1)$. Let $\mathcal{A}$ be a (τ, β, α)-robust algorithm for estimating the statistic μ. Then Algorithm 2 is $(\tau, 2\beta, 2\alpha)$-robust.*

Proof. Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ and let $\mathcal{S}'$ be a τ-corrupted version of $\mathcal{S}$, that is, $d_{\text{H}}(\mathcal{S}, \mathcal{S}') \leq n\tau$. We show that running Algorithm 2 over $\mathcal{S}'$ returns an accurate estimate with high probability over $\mathcal{S}$. Let $\mathbb{B}(\mu) = \{u \in \mathbb{R}^d : \|u - \mu\| \leq \alpha\}$ be the ball of radius α around μ, and let $\mathbb{B}(v^\star)$ be the ball that maximizes $P_{\mathcal{S}'}(\mathbb{B}(v))$. Let $E = \{\mathcal{S} : \forall \mathcal{S}'\ d_{\text{H}}(\mathcal{S}', \mathcal{S}) \leq n\tau, P_{\mathcal{S}'}(\mathbb{B}(\mu)^c) \leq 1/2\}$ denote the set of good input datasets $\mathcal{S}$ such that for all τ-corrupted $\mathcal{S}'$, the robust algorithm $\mathcal{A}$ returns bad answers with probability less than $1/2$. We show that if $\mathcal{S} \in E$ then Algorithm 2 returns an accurate answer for any τ-corrupted $\mathcal{S}'$. Indeed, if $\mathcal{S} \in E$, then for any τ-corrupted $\mathcal{S}'$, $P_{\mathcal{S}'}(\mathbb{B}(\mu)) > 1/2$. Moreover, the definition of $\mathbb{B}(v^\star)$ implies

$$P_{\mathcal{S}'}(\mathbb{B}(v^\star)) \geq P_{\mathcal{S}'}(\mathbb{B}(\mu)) > 1/2.$$

As a result, we have that $\mathbb{B}(v^\star) \cap \mathbb{B}(\mu) \neq \emptyset$. Let $u \in \mathbb{B}(v^\star) \cap \mathbb{B}(\mu)$. We have that

$$\|v^\star - \mu\| \leq \|v^\star - u\| + \|u - \mu\| \leq 2\alpha.$$

Thus, if $\mathcal{S} \in E$ then Algorithm 2 is 2α accurate. It remains to show that $\Pr[\mathcal{S} \notin E] \leq 2\beta$. This follows from the fact that $\mathcal{A}$ has failure probability β:

$$\begin{aligned}
\beta &\geq \Pr_{\mathcal{S},\mathcal{A}}[\exists \mathcal{S}' : d_{\text{H}}(\mathcal{S}', \mathcal{S}) \text{ and } \mathcal{A}(\mathcal{S}') \in \mathbb{B}(\mu)^c] \\
&\geq \Pr_{\mathcal{S}}[\mathcal{S} \notin E] \cdot \Pr_{\mathcal{S},\mathcal{A}}[\exists \mathcal{S}' : d_{\text{H}}(\mathcal{S}', \mathcal{S}) \leq n\tau \text{ and } \mathcal{A}(\mathcal{S}') \in \mathbb{B}(\mu)^c \mid \mathcal{S} \notin E] \\
&= \Pr_{\mathcal{S}}[\mathcal{S} \notin E] \cdot \mathbb{E}_{\mathcal{S},\mathcal{A}}\left[\max_{\mathcal{S}' : d_{\text{H}}(\mathcal{S}', \mathcal{S}) \leq n\tau} \mathbb{1}\{\mathcal{A}(\mathcal{S}') \in \mathbb{B}(\mu)^c\} \mid \mathcal{S} \notin E\right] \\
&\geq \Pr_{\mathcal{S}}[\mathcal{S} \notin E] \cdot \mathbb{E}_{\mathcal{S}}\left[\max_{\mathcal{S}' : d_{\text{H}}(\mathcal{S}', \mathcal{S}) \leq n\tau} \mathbb{E}_{\mathcal{A}}[\mathbb{1}\{\mathcal{A}(\mathcal{S}') \in \mathbb{B}(\mu)^c\}] \mid \mathcal{S} \notin E\right] \qquad \text{(by Jensen's inequality)} \\
&= \Pr_{\mathcal{S}}[\mathcal{S} \notin E] \cdot \mathbb{E}_{\mathcal{S}}\left[\max_{\mathcal{S}' : d_{\text{H}}(\mathcal{S}', \mathcal{S}) \leq n\tau} P_{\mathcal{S}'}(\mathbb{B}(\mu)^c) \mid \mathcal{S} \notin E\right] \\
&> \Pr_{\mathcal{S}}[\mathcal{S} \notin E] \cdot \frac{1}{2}.
\end{aligned}$$

The claim follows. □

A.2. Robust-to-Private Transformation for General Loss

In the statement of Theorem 3.1, the error is measured in some norm $\|\cdot\|$, the range of the robust algorithm $\mathcal{A}_{\text{rob}}$ is bounded in the same norm, and the smoothness ρ allowed in the inverse sensitivity score function (Equation (1)) is again bounded in the same norm. We can prove a more general theorem, where the last two norms are the same, but the error is instead measured with respect to a general loss function which satisfies the triangle inequality.

Theorem A.2 (Robust-to-private, general loss, restatement of Theorem 6.2)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ such that $\mu(P) \in \mathbb{R}^d$. Let $\varepsilon, \beta \in (0,1)$. Let $L : (\mathbb{R}^d)^2 \to \mathbb{R}$ be a loss function which satisfies the triangle inequality. Let $\mathcal{A}_{rob} : (\mathbb{R}^d)^n \to \{t \in \mathbb{R}^d : \|t\| \le R\}$ be a (deterministic) (τ, β, α)-robust algorithm with respect to L. Let $\alpha_0 \le \alpha$. Suppose n is such that the smallest value τ satisfying Equation (7) is at most 1. Suppose $\forall u, v \in \mathbb{B}(R + \alpha_0)\ L(u, v) \le c_L \|u - v\|$ for some constant c_L. If*

$$\tau \ge \frac{2\left(d \log\left(\frac{R}{\alpha_0} + 1\right) + \log\frac{1}{\beta}\right)}{n\varepsilon}, \tag{7}$$

then Algorithm $M^{\rho}_{\text{Inv}}(\mathcal{S}; \mathcal{A}_{rob})$ with $\rho = \alpha_0$ in norm $\|\cdot\|$ is ε-DP and, with probability at least $1 - 2\beta$, has error

$$L\left(M^{\rho}_{\text{Inv}}(\mathcal{S}; \mathcal{A}_{rob}), \mu\right) \le (3 + c_L)\alpha = O(\alpha).$$

Before proving Theorem A.2, we need the equivalent of Theorem 2.4 for general loss functions.

Theorem A.3 (Continuous functions, general loss)**.** *Let $\varepsilon, \beta \in (0,1)$, $\rho > 0$. Let $f : \mathcal{Z}^n \to \mathcal{T}$ where $\mathcal{T} = \{t \in \mathbb{R}^d : \|t\| \le R\}$. Let $L : (\mathbb{R}^d)^2 \to \mathbb{R}$ be a loss function which satisfies the triangle inequality and $\forall u, v \in \mathbb{B}(R+\rho)\ L(u, v) \le c_L \|u - v\|$ for some constant c_L. Suppose n is larger than the smallest K satisfying Equation (8) below. Then for any $\mathcal{S} \in \mathcal{Z}^n$, with probability $1 - \beta$, the ρ-smooth-inverse-sensitivity mechanism with norm $\|\cdot\|$ has error*

$$L\left(M^{\rho}_{\text{Inv}}(\mathcal{S}; f), f(\mathcal{S})\right) \le \omega^L_f\left(\mathcal{S}; K\right) + c_L\rho,$$

where

$$K \ge \frac{2d \log(R/\rho + 1) + 2\log(1/\beta)}{\varepsilon} \tag{8}$$

and $\omega^L_f\left(\mathcal{S}; K\right) = \sup_{\mathcal{S}' : d_{\text{H}}(\mathcal{S}, \mathcal{S}') \le K} L\left(f(\mathcal{S}), f(\mathcal{S}')\right)$ denotes the local modulus of continuity of f at $\mathcal{S}$ with respect to loss function L.

Proof. We define the good set of outputs $A = \{t \in \mathbb{R}^d : \mathsf{len}^{\rho}(\mathcal{S}; t) \le K\}$, where $\mathsf{len}^{\rho}(\mathcal{S}; t)$ is defined with respect to norm $\|\cdot\|$ and $K \in \mathbb{N}$. By the definition of ρ-smooth-inverse sensitivity, for any $t \in A$, there exists $s \in \mathcal{T}$ with $\mathsf{len}(\mathcal{S}; s) = K$ and $\|s - t\| \le \rho$. We will show that $\Pr[M^{\rho}_{\text{Inv}}(\mathcal{S}) \notin A] \le \beta$ for sufficiently large K. This implies the desired upper bound as we have that for $t \in A$

$$\begin{aligned}
L(t, f(\mathcal{S})) &\le L(t, s) + L(s, f(\mathcal{S})) && \text{(by triangle inequality)} \\
&\le c_L \|t - s\| + \omega^L_f(\mathcal{S}; K) && \text{(since } s \in \mathbb{B}(R), \|t - s\| \le \rho, \text{ so } s, t \in \mathbb{B}(R + \rho)) \\
&\le c_L\rho + \omega^L_f(\mathcal{S}; K).
\end{aligned}$$

Now we upper bound $\Pr[M^{\rho}_{\text{Inv}}(\mathcal{S}; f) \notin A]$. First, note that $\mathsf{len}^{\rho}(\mathcal{S}; u) = 0$ for u such that $\|u - f(\mathcal{S})\| \le \rho$. This implies that for any t such that $\mathsf{len}^{\rho}(\mathcal{S}; t) \ge K$, the density is upper bounded by

$$\pi_{\mathcal{S}}(t) \le \frac{e^{-K\varepsilon/2}}{\int_{u : \|u - f(\mathcal{S})\| \le \rho} du}$$

Overall, this implies that

$$\begin{aligned}
\Pr[M^{\rho}_{\text{Inv}}(\mathcal{S}; f) \notin A] &\le e^{-K\varepsilon/2} \frac{\int_{u : \|u\| \le R + \rho} du}{\int_{u : \|u - f(\mathcal{S})\| \le \rho} du} \\
&\le e^{-K\varepsilon/2}(R/\rho + 1)^d.
\end{aligned}$$

Setting $K \ge \frac{2d \log(R/\rho + 1) + 2\log(1/\beta)}{\varepsilon}$, we get that $\Pr(M^{\rho}_{\text{Inv}}(\mathcal{S}; f) \notin A) \le \beta$. □

We are now ready to prove Theorem A.2.

Proof of Theorem A.2. First note that the claim about privacy is immediate from the guarantees of the smooth-inverse-sensitivity mechanism. Now we prove utility. Let $K = \frac{2\left(d\log\left(\frac{R}{\alpha_0}+1\right)+\log\frac{1}{\beta}\right)}{\varepsilon}$. The error of $M^{\rho}_{\mathrm{Inv}}(\mathcal{S}, \mathcal{A}_{\mathrm{rob}})$ is then bounded as follows:

$$
\begin{aligned}
&L\left(M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}), \mu\right) \\
&\le L\left(\mathcal{A}_{\mathrm{rob}}(\mathcal{S}), \mu\right) + L\left(M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}), \mathcal{A}_{\mathrm{rob}}(\mathcal{S})\right) && \text{(by triangle inequality)} \\
&\le L\left(\mathcal{A}_{\mathrm{rob}}(\mathcal{S}), \mu\right) + \sup_{\mathcal{S}': d_{\mathrm{H}}(\mathcal{S}', \mathcal{S}) \le K} L\left(\mathcal{A}_{\mathrm{rob}}(\mathcal{S}'), \mathcal{A}_{\mathrm{rob}}(\mathcal{S})\right) + c_L \alpha_0 && \text{(w.p. } 1-\beta \text{ by Theorem A.3)} \\
&\le 2L\left(\mathcal{A}_{\mathrm{rob}}(\mathcal{S}), \mu\right) + \sup_{\mathcal{S}': d_{\mathrm{H}}(\mathcal{S}', \mathcal{S}) \le K} L\left(\mu, \mathcal{A}_{\mathrm{rob}}(\mathcal{S}')\right) + c_L \alpha_0 && \text{(by triangle inequality)}
\end{aligned}
$$

Recall that, by assumption, $\mathcal{A}_{\mathrm{rob}}$ is (τ, β, α)-robust for $\tau \ge K/n$, and $\alpha_0 \le \alpha$. Thus, with probability $1-\beta$, $L\left(\mathcal{A}_{\mathrm{rob}}(\mathcal{S}'), \mu\right) \le \alpha$ for any τ-corrupted dataset $\mathcal{S}'$. By union bound, we have that with probability $1-2\beta$, $L\left(M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}), \mu\right) \le 3\alpha + c_L\alpha_0 \le (3+c_L)\alpha = O(\alpha)$. □

A.3. Robust-to-Private Transformation for Sparse Estimators

In this section, we extend our transformation to work for k-sparse statistical estimation problems with improved dependence on the dimension. To this end, we define a variant of the smooth inverse sensitivity which is non-zero only for k-sparse outputs,

$$
\mathsf{len}^{\mathsf{sp}}(\mathcal{S}; t) = \begin{cases} \inf_{s \in \mathbb{R}^d : \|s-t\| \le \rho} \mathsf{len}(\mathcal{S}; s) & \text{if } \|t\|_0 \le k \\ \infty & \text{if } \|t\|_0 > k \end{cases}
$$

Then, our sparse-variant of the inverse sensitivity mechanism M^{ρ}_{sp} applies the exponential mechanism with $\mathsf{len}^{\mathsf{sp}}$ as the score function,

$$
\pi_{\mathcal{S}}(t) = \frac{e^{-\mathsf{len}^{\mathsf{sp}}(\mathcal{S}; t)\varepsilon/2}}{\int_{s \in \mathbb{R}^d} e^{-\mathsf{len}^{\mathsf{sp}}(\mathcal{S}; s)\varepsilon/2} ds} \tag{9}
$$

We have the following upper bound for this mechanism.

Theorem A.4. *Let $f : \mathcal{Z}^n \to \mathcal{T}$ where $\mathcal{T} = \{v \in \mathbb{R}^d : \|v\| \le R\}$ such that $\|f(\mathcal{S})\|_0 \le k$ for all $\mathcal{S} \in \mathcal{Z}^n$. Then for any $\mathcal{S} \in \mathcal{Z}^n$, and $\beta > 0$, with probability at least $1-\beta$, the (sparse) inverse-sensitivity mechanism (9) with norm $\|\cdot\|$ has error*

$$
L\left(M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f), f(\mathcal{S})\right) \le \omega^L_f\left(\mathcal{S}; K\right) + c_L\rho,
$$

where

$$
K \ge \frac{2\left(k\left(\log(ed/k) + \log(R/\rho + 1)\right) + \log\frac{1}{\beta}\right)}{\varepsilon} \tag{10}
$$

and $\omega^L_f\left(\mathcal{S}; K\right) = \sup_{\mathcal{S}': d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \le K} L\left(f(\mathcal{S}), f(\mathcal{S}')\right)$ denotes the local modulus of continuity of f at $\mathcal{S}$ with respect to loss function L.

Proof. The proof follows similar steps to the proof of Theorem A.3. We define the good set of outputs $A = \{t \in \mathbb{R}^d : \mathsf{len}^{\mathsf{sp}}(\mathcal{S}; t) \le K\}$, where $\mathsf{len}^{\mathsf{sp}}(\mathcal{S}; t)$ is defined with respect to norm $\|\cdot\|$ and $K \in \mathbb{N}$. By the definition of sparse inverse sensitivity, for any $t \in A$, t is k-sparse and there exists $s \in \mathcal{T}$ with $\mathsf{len}(\mathcal{S}; s) = K$ and $\|s-t\| \le \rho$. We will show that $\Pr[M^{\rho}_{\mathrm{Inv}}(\mathcal{S}) \notin A] \le \beta$ for sufficiently large K. This implies the desired upper bound as we have that for $t \in A$

$$
\begin{aligned}
L(t, f(\mathcal{S})) &\le L(t, s) + L(s, f(\mathcal{S})) && \text{(by triangle inequality)} \\
&\le c_L \|t-s\| + \omega^L_f(\mathcal{S}; K) && \text{(since } s \in \mathbb{B}(R), \|t-s\| \le \rho, \text{ so } s, t \in \mathbb{B}(R+\rho)) \\
&\le c_L\rho + \omega^L_f(\mathcal{S}; K).
\end{aligned}
$$

Now we upper bound $\Pr[M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) \notin A]$. First, note that $\mathsf{len}^{\rho}(\mathcal{S}; u) = 0$ for u such that $\|u - f(\mathcal{S})\| \le \rho$. This implies that for any t such that $\mathsf{len}^{\rho}(\mathcal{S}; t) \ge K$, the density is upper bounded by 0 if $\|t\|_0 > k$, and otherwise,

$$
\pi_{\mathcal{S}}(t) \le \frac{e^{-K\varepsilon/2}}{\int_{u : \|u\|_0 \le k, \|u - f(\mathcal{S})\| \le \rho} du}
$$

Overall, this implies that

$$\begin{aligned}\Pr[M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) \notin A] &\leq e^{-K\varepsilon/2} \frac{\int_{u\in\mathbb{R}^d:\|u\|_0\leq k, \|u\|\leq R+\rho} du}{\int_{u\in\mathbb{R}^d:\|u\|_0\leq k, \|u-f(\mathcal{S})\|\leq\rho} du} \\ &\leq e^{-K\varepsilon/2}\binom{d}{k}\frac{\int_{u\in\mathbb{R}^k:\|u\|\leq R+\rho} du}{\int_{u\in\mathbb{R}^k:\|u\|\leq\rho} du} \\ &\leq e^{-K\varepsilon/2}\binom{d}{k}(R/\rho+1)^k \\ &\leq e^{-K\varepsilon/2}(ed/k)^k(R/\rho+1)^k.\end{aligned}$$

Setting $K \geq \frac{2k(\log(ed/k)+\log(R/\rho+1))+2\log(1/\beta)}{\varepsilon}$, we get that $\Pr(M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) \notin A) \leq \beta$. □

Using this mechanism, we now have the following transformation from robust-to-private for sparse estimators.

Theorem A.5 (Robust-to-private for sparse estimators)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ such that $\mu(P) \in \mathbb{R}^d$. Let $\varepsilon, \beta \in (0,1)$. Let $L : (\mathbb{R}^d)^2 \to \mathbb{R}$ be a loss function which satisfies the triangle inequality. Let $\mathcal{A}_{rob} : (\mathbb{R}^d)^n \to \{t \in \mathbb{R}^d : \|t\| \leq R\}$ be a (deterministic) (τ, β, α)-robust algorithm with respect to L such that $\|\mathcal{A}_{rob}(\mathcal{S})\|_0 \leq k$ for all $\mathcal{S}$. Let $\alpha_0 \leq \alpha$. Suppose n is such that the smallest value τ satisfying Equation (11) is at most 1. Suppose $\forall u, v \in \mathbb{B}(R+\alpha_0)\ L(u,v) \leq c_L \|u-v\|$ for some constant c_L. If*

$$\tau \geq \frac{2\left(k\left(\log(ed/k) + \log(R/\alpha_0+1)\right) + \log\frac{1}{\beta}\right)}{n\varepsilon} \tag{11}$$

then Algorithm $M^{\rho}_{\mathsf{sp}}(\mathcal{S}; \mathcal{A}_{rob})$ with $\rho = \alpha_0$ in norm $\|\cdot\|$ is ε-DP and, with probability at least $1-2\beta$, has error

$$L\left(M^{\rho}_{\mathsf{sp}}(\mathcal{S}; \mathcal{A}_{rob}), \mu\right) \leq (3+c_L)\alpha = O(\alpha).$$

We leave the proof as an exercise for the reader as it is identical to the proof of Theorem A.2 using the upper bounds for the sparse variant of the inverse sensitivity mechanism (Theorem A.4).

A.4. Omitted proofs of Section 2 and Section 4.2

Theorem A.6 (Continuous functions, restatement of Theorem 2.4)**.** *Let $f : \mathcal{Z}^n \to \mathcal{T}$ where $\mathcal{T} = \{v \in \mathbb{R}^d : \|v\| \leq R\}$. Then for any $\mathcal{S} \in \mathcal{Z}^n$, and $\beta > 0$, with probability at least $1-\beta$, the ρ-smooth inverse-sensitivity mechanism with norm $\|\cdot\|$ has error*

$$\|M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) - f(\mathcal{S})\| \leq \omega_f\left(\mathcal{S}; \frac{2\left(d\log\left(\frac{R}{\rho}+1\right)+\log\frac{1}{\beta}\right)}{\varepsilon}\right) + \rho.$$

Proof. We define the good set of outputs $A = \{t \in \mathbb{R}^d : \mathsf{len}^{\rho}(\mathcal{S}; t) \leq K\}$. By the definition of ρ-smooth inverse-sensitivity, for any $t \in A$, there exists $s \in \mathbb{R}^d$ with $\mathsf{len}(\mathcal{S}; s) \leq K$ and $\|s-t\| \leq \rho$. We will show that $\Pr[M^{\rho}_{\mathrm{Inv}}(\mathcal{S}) \notin A] \leq \beta$ for sufficiently large K. This implies the desired upper bound as we have that for $t \in A$

$$\begin{aligned}\|t - f(\mathcal{S})\| &= \|t - s + s - f(\mathcal{S})\| \\ &\leq \|t-s\| + \|s - f(\mathcal{S})\| \\ &\leq \rho + \omega_f(\mathcal{S}; K).\end{aligned}$$

Now we upper bound $\Pr[M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) \notin A]$. First, note that $\mathsf{len}^{\rho}(\mathcal{S}; s) = 0$ for s such that $\|s - f(\mathcal{S})\| \leq \rho$. This implies that for any t such that $\mathsf{len}^{\rho}(\mathcal{S}; t) \geq K$, the density is upper bounded by

$$\pi_{\mathcal{S}}(t) \leq \frac{e^{-K\varepsilon/2}}{\int_{s:\|s-f(\mathcal{S})\|\leq\rho} ds}.$$

Overall, this implies that

$$\Pr[M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) \notin A] \le e^{-K\varepsilon/2} \frac{\int_{s:\|s\| \le R+\rho} ds}{\int_{s:\|s-f(\mathcal{S})\| \le \rho} ds} \le e^{-K\varepsilon/2}(R/\rho + 1)^d.$$

Setting $K \ge \frac{2d\log(R/\rho+1)+2\log(1/\beta)}{\varepsilon}$, we get that $\Pr(M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; f) \notin A) \le \beta$. □

Corollary A.7 (Optimality, restatement of Corollary 4.2)**.** *Let $\mathcal{P}$ be a family of distributions and $P \in \mathcal{P}$. Let μ be a 1-dimensional statistic where $|\mu(P)| \le 1$. Let $\alpha_{\mathrm{priv}}(\varepsilon, \beta)$ be the minimax error of any ε-DP algorithm with failure probability $\beta \le 1/4$ that estimates statistic $\mu(P)$. Let $n > 1$. Suppose that there exists a constant c such that the non-private error $\alpha_{\mathrm{priv}}(\infty, \beta) \ge \frac{1}{n^c}$ for any $\beta \le 1/2$. Then there are constants $c_1 \ge c_2 > 0$ such that $\beta_p = 1/n^{c_1}$ and $\beta'_p = 1/n^{c_2}$, robust algorithm $\mathcal{A}_{\mathrm{rob}}$, and a choice of ρ, such that the ρ-smooth inverse-sensitivity mechanism $M^{\rho}_{\mathrm{Inv}}(\cdot; \mathcal{A}_{\mathrm{rob}})$ with privacy parameter ε achieves the minimax optimal error $O(\alpha_{\mathrm{priv}}(\varepsilon, \beta_p))$ with probability $1 - \beta'_p$.*

Proof. Let $\alpha_{\mathrm{priv}}(\varepsilon, \beta)$ be the minimax error for estimating μ under family $\mathcal{P}$ for any $\beta \le 1/4$. By Theorem 3.3 and Theorem A.1, there exists a deterministic τ-robust algorithm with $\tau = \log(1/\gamma)/n\varepsilon$, with accuracy $\alpha_1 = 2\alpha_{\mathrm{priv}}(\varepsilon, \beta)$ and failure probability $\beta_1 = 2\beta/\gamma$. Via the transformation of Theorem 3.1, choosing $\rho = \frac{1}{n^c} \le \alpha_{\mathrm{priv}}(\varepsilon, \beta)$, we can construct an ε-DP algorithm with failure probability $2\beta_1 = 4\beta/\gamma$ and accuracy $4\alpha_1 = 8\alpha_{\mathrm{priv}}(\varepsilon, \beta)$, if $\tau = \frac{\log(1/\gamma)}{n\varepsilon} \ge \frac{2\log(n^c+1)+2\log(\gamma/(4\beta))}{n\varepsilon}$. Setting $\gamma = (4\beta/(2n)^c)^{2/3}$ satisfies the requirement. Thus, the ε-DP algorithm constructed via the transformation in Theorem 3.1 has error at most $8\alpha_{\mathrm{priv}}(\varepsilon, \beta)$ with failure probability at most $\beta'_p = 4\beta/\gamma = (4\beta)^{1/3} \cdot (2n)^{2c/3}$. There exist constants $c_1 \ge c_2 > 0$ such that $\beta = \beta_p = \frac{1}{n^{c_1}}$ and $\beta'_p = \frac{1}{n^{c_2}}$. □

B. Improved Transformation for Approximate DP

In this section, we propose a different transformation for (ε, δ)-DP that avoids the necessary dependence on diameter for pure ε-DP. Our transformation uses a truncated version of the inverse-sensitivity mechanism which only outputs values with bounded inverse sensitivity. This mechanism is not differentially private for all inputs, therefore, in order to guarantee privacy, we use a private test to verify that the input is well-behaved before running the truncated inverse-sensitivity mechanism.

This approach can be viewed as a special case of the restricted exponential mechanism of Brown et al. (2021) (or the even more general HPTR framework (Liu et al., 2022)), which in turn has been inspired by the propose-test-release (PTR) framework (Dwork and Lei, 2009). However, we choose a simplified algorithm and presentation, which is tailored to our case, where we have the smooth inverse sensitivity as our cost function.

B.1. Truncated Inverse-Sensitivity Mechanism

We develop a truncated version of the inverse-sensitivity mechanism which is (ε, δ)-DP. This mechanism uses a truncated version of the inverse sensitivity as follows: given a function $f : \mathcal{Z}^n \to \mathbb{R}^d$ and threshold K,

$$\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S}; t) := \begin{cases} \mathsf{len}^{\rho}_f(\mathcal{S}; t) & \text{if } \mathsf{len}^{\rho}_f(\mathcal{S}; t) \le K \\ \infty & \text{otherwise.} \end{cases}$$

The truncated inverse-sensitivity mechanism $M^{\rho}_{\mathsf{trunc}}(\cdot; f)$ then applies the exponential mechanism using this score function, resulting in the following density given an input dataset $\mathcal{S}$:

$$\pi_{\mathcal{S}}(t) = \frac{e^{-\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S};t)\varepsilon/2}}{\int_{s\in\mathbb{R}^d} e^{-\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S};s)\varepsilon/2} ds} \tag{12}$$

Before proving the guarantees of the truncated inverse-sensitivity mechanism, we need to define (ε, δ)-indistinguishable distributions:

Definition B.1 ((ε, δ)-indistinguishability)**.** Two distributions P, Q over domain $\mathcal{T}$ are (ε, δ)-indistinguishable, denoted by $P \approx_{\varepsilon,\delta} Q$, if for any measurable subset $T \subseteq \mathcal{T}$,

$$\Pr_{t\sim P}[t \in T] \le e^{\varepsilon} \Pr_{t\sim Q}[t \in T] + \delta \quad \text{and} \quad \Pr_{t\sim Q}[t \in T] \le e^{\varepsilon} \Pr_{t\sim P}[t \in T] + \delta.$$

Note that if $\mathcal{A}(\mathcal{S}) \approx_{\varepsilon,\delta} \mathcal{A}(\mathcal{S}')$ for any neighboring datasets $\mathcal{S}, \mathcal{S}'$, then $\mathcal{A}$ is (ε, δ)-differentially private. We have the following guarantees for the truncated inverse-sensitivity mechanism.

Proposition B.2. *Let $n \geq 1$, $\varepsilon, \delta \in (0,1)$, $B > 0$, and $f : \mathcal{Z}^n \to \mathbb{R}^d$. Let $K \geq \frac{d+\log(1/\delta)}{\varepsilon}$ and $S_{\mathsf{bad}} = \{\mathcal{S} \in \mathcal{Z}^n : \omega_f(\mathcal{S}; K+1) > B\}$. For any $\mathcal{S} \notin S_{\mathsf{bad}}$, the truncated inverse-sensitivity mechanism (12) with $\rho = 2B$ has error*

$$\|M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) - f(\mathcal{S})\| \leq 3B.$$

Moreover, for any $\mathcal{S} \notin S_{\mathsf{bad}}$ and neighboring dataset $\mathcal{S}'$, $M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \approx_{\varepsilon,\delta} M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f)$.

Proof. The claim about utility follows directly from the definition of the truncated inverse-sensitivity as the probability of returning t such that $\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S}; t) \geq K$ is zero. Now we proceed to prove the privacy claim. Let $\mathcal{S} \in S^c_{\mathsf{bad}}$ and $\mathcal{S}' \in \mathcal{Z}^n$ be two neighboring datasets and $T \subset \mathcal{T}$. Since $\omega_f(\mathcal{S}; K+1) \leq B$, we have that $\omega_f(\mathcal{S}'; K) \leq 2B$. Thus, it suffices to show that for any two neighboring datasets $\mathcal{S}$ and $\mathcal{S}'$ such that $\omega_f(\mathcal{S}'; K) \leq 2B$ and $\omega_f(\mathcal{S}; K) \leq 2B$, we have $\Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \in T] \leq e^{\varepsilon} \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f) \in T] + \delta$. Let $T_0 = \{t \in \mathcal{T} : \mathsf{len}^{\rho}_f(\mathcal{S}; t) = K\}$. Now we have

$$\begin{aligned}
\Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \in T] &= \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \in T \setminus T_0] + \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \in T \cap T_0] \\
&\leq e^{\varepsilon} \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f) \in T \setminus T_0] + \frac{e^{-K\varepsilon}}{\mathsf{Vol}(\mathbb{B}^d(\rho))} \mathsf{Vol}(T \cap T_0) \\
&\leq e^{\varepsilon} \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f) \in T] + \frac{e^{-K\varepsilon}}{\mathsf{Vol}(\mathbb{B}^d(\rho))} \mathsf{Vol}(T \cap T_0),
\end{aligned}$$

where the first inequality follows since for $t \notin T_0$ we have that either $|\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S}; t) - \mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S}'; t)| \leq 1$ or $\mathsf{len}^{\mathsf{trunc}}_f(\mathcal{S}; t) = \infty$. Since $\omega_f(\mathcal{S}; K) \leq 2B$, we get $\mathsf{Vol}(T \cap T_0) \leq \mathsf{Vol}(T_0) \leq \mathsf{Vol}(\mathbb{B}^d(2B + \rho))$. For $\rho = 2B$, this implies that $\mathsf{Vol}(T \cap T_0)/\mathsf{Vol}(\mathbb{B}^d(\rho)) \leq \mathsf{Vol}(\mathbb{B}^d(4B))/\mathsf{Vol}(\mathbb{B}^d(2B)) = 2^d$. Therefore, for $K \geq \frac{d+\log(1/\delta)}{\varepsilon}$, we have that $\Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \in T] \leq e^{\varepsilon} \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f) \in T] + \delta$. By symmetry, we can use the same argument, to show that $\Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f) \in T] \leq e^{\varepsilon} \Pr[M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \in T] + \delta$. Thus, overall we show that $M^{\rho}_{\mathsf{trunc}}(\mathcal{S}; f) \approx_{\varepsilon,\delta} M^{\rho}_{\mathsf{trunc}}(\mathcal{S}'; f)$ for all $\mathcal{S} \notin S_{\mathsf{bad}}$ and neighboring dataset $\mathcal{S}' : d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \leq 1$. □

While it may seem that the truncated inverse sensitivity provides the desired transformation for (ε, δ)-DP, note that it requires the condition $\omega_f(\mathcal{S}; K+1) \leq B$ to hold for all inputs $\mathcal{S} \in \mathcal{Z}^n$. However, robust algorithms only guarantee boundedness of $\omega_f(\mathcal{S}; K+1)$ for $\mathcal{S} \overset{\text{iid}}{\sim} P^n$. To this end, in the next section we show how to use propose-test-release (PTR) in order to overcome this barrier.

B.2. PTR-based Transformation

Building on the truncated inverse-sensitivity mechanism, in this section we use propose-test-release (PTR) to design a transformation from robust algorithms into approximate (ε, δ)-DP algorithms where the error does not depend on the diameter.

An equivalent approach would be to use the restricted exponential mechanism from (Brown et al., 2021) with the smooth inverse-sensitivity $\mathsf{len}^{\rho}_f(\mathcal{S}; t)$ as its cost function. The main idea of this approach is to perform a private test to check if the input $\mathcal{S}$ is far from "unsafe", before running the exponential mechanism restricted to points t with $\mathsf{len}^{\rho}_f(\mathcal{S}; t) \leq K$. The set "unsafe" consists of datasets on which running the restricted exponential mechanism would not produce (ε, δ)-indistinguishable outputs. However, our specific score function allows us to simplify the "unsafe" set, and this is the algorithm we present in this section. The following theorem states its guarantees. We present our transformation in Algorithm 3.

Theorem B.3 (Robust-to-private, approximate DP, restatement of Theorem 5.1)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} P$ such that $\mu(P) \in \mathbb{R}^d$. Let $\varepsilon, \delta, \beta \in (0,1)$. Let $\mathcal{A}_{rob} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ be a deterministic (τ, β, α)-robust algorithm for the statistic μ. If $\tau \geq \frac{8(d+\log(1/\min\{\delta,\beta\}))}{n\varepsilon}$ then Algorithm 1 with $B = 2\alpha$ and $\rho = 2B$ is (ε, δ)-DP and, with probability at least $1 - 2\beta$ returns $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \leq 7\alpha$.*

Proof. We start by proving the privacy guarantees of Algorithm 1. Note that the Laplace mechanism (Dwork et al., 2006) implies that $\hat{d}$ is $\varepsilon/2$-DP as the function d has sensitivity 1. By assumption on τ, $K = n\tau/2 - 1 \geq \frac{2(d+\log(2/\delta))}{\varepsilon}$. Thus, by Proposition B.2, for input dataset $\mathcal{S}$, if $\omega_f(\mathcal{S}; K+1) \leq B$ then the truncated inverse-sensitivity is $(\varepsilon/2, \delta/2)$-DP. On the

Algorithm 3 Robust-to-Private ((ε, δ)-DP), restatement of Algorithm 1

Require: $\mathcal{S} = (S_1, \ldots, S_n)$, (τ, β, α)-robust algorithm $\mathcal{A}_{\text{rob}}$, local modulus bound B

Let $K = n\tau/2 - 1$
Let $S_{\mathsf{bad}} = \{\mathcal{S} \in \mathcal{Z}^n : \omega_f(\mathcal{S}; K+1) > B\}$
Calculate $d = \operatorname{dist}(\mathcal{S}, S_{\mathsf{bad}})$
Set $\hat{d} = d + \zeta$ where $\zeta \sim \mathsf{Laplace}(2/\varepsilon)$
if $\hat{d} > 2\log(1/\min(\delta, \beta))/\varepsilon$ **then**
    Sample t from the truncated inverse-sensitivity mechanism (12) with threshold K, privacy parameter $\varepsilon/2$, smoothness parameter $\rho = 2B$, and return t.
else
    Return $\perp$
end if

other hand, if $\omega_f(\mathcal{S}; K+1) > B$ then $d = 0$ and therefore $\hat{d} \le 2\log(1/\delta)/\varepsilon$ with probability $1 - \delta/2$ and the algorithm returns $\perp$. Overall, by composition, Algorithm 1 is (ε, δ)-DP.

We now prove the accuracy guarantee. Let $\mathcal{S} \overset{\text{iid}}{\sim} P^n$. By the guarantee of the robust algorithm, with probability $1 - \beta$, for all $\mathcal{S}'$ such that $d_{\text{H}}(\mathcal{S}, \mathcal{S}') \le \tau n$, we get that $\|\mathcal{A}_{\text{rob}}(\mathcal{S}') - \mu(P)\| \le \alpha$. Therefore, for any $\mathcal{S}'_1$ such that $d_{\text{H}}(\mathcal{S}, \mathcal{S}'_1) \le \tau n/2$, we have

$$\begin{aligned}
\omega_f(\mathcal{S}'_1; n\tau/2) &= \sup_{\mathcal{S}'_2 : d_{\text{H}}(\mathcal{S}'_1, \mathcal{S}'_2) \le n\tau/2} \|\mathcal{A}_{\text{rob}}(\mathcal{S}'_1) - \mathcal{A}_{\text{rob}}(\mathcal{S}'_2)\| \\
&\le \sup_{\mathcal{S}'_2 : d_{\text{H}}(\mathcal{S}'_1, \mathcal{S}'_2) \le n\tau/2} (\|\mathcal{A}_{\text{rob}}(\mathcal{S}'_1) - \mu(P)\| + \|\mu(P) - \mathcal{A}_{\text{rob}}(\mathcal{S}'_2)\|) \\
&\le \alpha + \sup_{\mathcal{S}'_2 : d_{\text{H}}(\mathcal{S}, \mathcal{S}'_2) \le n\tau} \|\mu(P) - \mathcal{A}_{\text{rob}}(\mathcal{S}'_2)\| \\
&\le 2\alpha.
\end{aligned}$$

Since $B = 2\alpha$, we have that with probability $1 - \beta$, $d => n\tau/2$ and in particular $\mathcal{S} \notin S_{\mathsf{bad}}$. By the concentration guarantees of the Laplace distribution, we have that $\hat{d} > n\tau/2 - 2\log(1/\beta)/\varepsilon$ with probability at least $1 - \beta$, and thus $\hat{d} > \frac{2\log(1/\min\{\beta,\delta\})}{\varepsilon}$, which implies that the algorithm will run the truncated inverse-sensitivity mechanism. Proposition B.2 now implies that the latter will return $\hat{\mu}$ such that $\|\hat{\mu} - \mathcal{A}_{\text{rob}}(\mathcal{S})\| \le 3B$. Moreover, $\|\mathcal{A}_{\text{rob}}(\mathcal{S}) - \mu\| \le \alpha$. Overall we get that with probability $1 - 2\beta$,

$$\|\hat{\mu} - \mu\| \le \|\hat{\mu} - \mathcal{A}_{\text{rob}}(\mathcal{S})\| + \|\mathcal{A}_{\text{rob}}(\mathcal{S}) - \mu\| \le 3B + \alpha = 7\alpha.$$

This completes the proof of the theorem. □

C. More Applications for Pure DP

In this section we apply our main transformation in Theorem 3.1 to fundamental tasks in private statistics to demonstrate that for all these tasks near-optimal error can be achieved by instantiating our black-box reduction with a robust estimator for the same task. In Appendix C.1 and Appendix C.2 we show that we can retrieve known optimal results for mean and covariance estimation of Gaussian distributions up to logarithmic factors. In Appendix C.3 and Section 6.1, we show that our transformation gives the first algorithms with optimal error for linear regression (including the sparse case) and PCA for Gaussian distributions. Our results for PCA hold for subgaussian distributions more generally.

For the majority of this section we will use the more general transformation, proven in Appendix A.2Theorem A.2.

The general strategy we follow in our applications is simple. We choose a known robust algorithm $\mathcal{A}$ for the statistic $\mu \in \mathbb{B}(R)$ we want to estimate. Informally, let us denote its accuracy by $\alpha(\tau)$, as it will be a function of the fraction of corruptions in the dataset τ (among other parameters). Applying our robust-to-private transformation from Theorem 3.1, we retrieve an ε-DP algorithm with accuracy roughly $\alpha(\tau^\star)$ for $\tau^\star \approx \frac{d\log(R'/\alpha_0) + \log(1/\beta)}{n\varepsilon}$. More precisely, we let $\mathcal{A}_{\text{rob}}$ be the algorithm that runs $\mathcal{A}$ and then projects its output on $\mathbb{B}(R')$, where R' is such that, with high probability, the projection

will have no effect and will maintain the accuracy guarantees of $\mathcal{A}$. Let α_0 be the error rate for learning statistic μ without privacy constraints or corruptions, which is always smaller than $\alpha(\tau)$. We run the ρ-smooth-inverse-sensitivity mechanism instantiating it with the projected robust algorithm $\mathcal{A}_{\text{rob}}$ and with smoothness parameter $\rho = \alpha_0$. In most applications, $\alpha(\tau) = \tilde{O}(\tau)$ so the error we incur on top of the non-private error α_0 is $\tilde{O}(d/n\varepsilon)$.

Theorem A.2extends Theorem 3.1 allowing us to measure the error of the algorithm with respect to a loss function L that may depend on unknown parameters and thus can not be computed directly. As long as this loss satisfies the triangle inequality, and any error we incur due to the smoothness ρ in norm $\|\cdot\|$ upper-bounds the error in L up to constants, the statement of our main theorem still holds.

For useful linear algebra facts and definitions, see Appendix D.

C.1. Mean Estimation

C.1.1. Known Covariance

We start with the task of estimating the mean μ of a d-dimensional Gaussian distribution with known covariance Σ. By applying $\Sigma^{-1/2}$ to all the points, this case can be reduced to spherical Gaussian mean estimation, where we can assume $\Sigma = \mathbb{I}$. We also assume that we know *a priori* a bound R such that $\|\mu\|_2 \le R$.[2] Corollary C.1 states that via our transformation, we can retrieve the optimal error for Gaussian mean estimation with known covariance under pure DP, matching optimal bounds (Bun et al., 2019; Liu et al., 2021).

Corollary C.1 (Spherical Gaussian mean)**.** *Let $\mathcal{S} = (S_1, \dots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \mathbb{I})$ such that $\mu \in \mathbb{B}(R)$. Let $\varepsilon, \beta \in (0,1)$ and $C \ge 1$ a known constant. Suppose n is such that $\alpha \le 1$ in Equation (13). There exists an ε-DP algorithm $\mathcal{M}$ such that, with probability at least $1-\beta$, has error $\|\mathcal{M}(\mathcal{S}) - \mu\|_2 \le \alpha$ for*

$$\alpha = C \cdot \left(\sqrt{\frac{d + \log(\frac{1}{\beta})}{n}} + \frac{d \log\left(\frac{Rn}{d}\right) + \log(\frac{1}{\beta})}{n\varepsilon} \right). \tag{13}$$

Since we are not concerned with computational efficiency, we will use the *Tukey median* as the robust Gaussian mean estimation algorithm for our transformation. The Tukey depth (Tukey, 1960) of a point t with respect to a distribution P is defined by

$$T_P(t) \coloneqq \inf_{v \in \mathbb{R}^d} \Pr_{S \sim P}[\langle S, v\rangle \ge \langle t, v\rangle].$$

We denote by $T_{\mathcal{S}}(t) \coloneqq \frac{1}{n} \min_v \sum_{i\in[n]} [\langle S_i, v\rangle \ge \langle t, v\rangle]$ the (normalized) Tukey depth of t with respect to dataset $\mathcal{S}$. The Tukey median with respect to any dataset $\mathcal{S}$ is then $t_m(\mathcal{S}) = \operatorname{argmax}_{t\in\mathbb{R}^d} T_{\mathcal{S}}(t)$. Let $\Pi_{\mathbb{C}}(t) = \operatorname{argmin}_{v\in\mathbb{C}} \|v - t\|_2$ be the euclidean projection of a point t to convex set $\mathbb{C}$. The next proposition states the robustness guarantees of (projected) Tukey median, which have been long-established (for a complete proof see e.g. (Li, 2019) or the more general Proposition D.4 in Appendix D).

Proposition C.2. *Let $\mathcal{S} = (S_1, \dots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \mathbb{I})$ such that $\mu \in \mathbb{B}(R)$. Let $\beta \in (0,1)$, $\tau \le 0.05$, and $\alpha_0 = C_0 \cdot \sqrt{(d + \log(1/\beta))/n}$ for a known constant $C_0 \ge 1$. Suppose n is such that $\alpha_0 \le 0.05$. Let $\alpha = 7(\alpha_0 + \tau) \le 1$. The projected Tukey median algorithm $\mathcal{A}_{\text{rob}}(\mathcal{S}) = \Pi_{\mathbb{B}(R+1)}(t_m(\mathcal{S}))$ is (τ, β, α)-robust. That is, with probability $1-\beta$, for any τ-corrupted $\mathcal{S}'$, such that $d_{\text{H}}(\mathcal{S}, \mathcal{S}') \le n\tau$, it holds that, $\|\mathcal{A}_{\text{rob}}(\mathcal{S}') - \mu\|_2 \le \alpha$.*

Using the above proposition, the proof of Corollary C.1 is a straightforward application of Theorem 3.1.

Proof of Corollary C.1. Consider the ρ-smooth-inverse-sensitivity mechanism $M^{\rho}_{\text{Inv}}(\cdot; \mathcal{A}_{\text{rob}})$ with norm $\|\cdot\| = \|\cdot\|_2$, $\mathcal{A}_{\text{rob}}(\mathcal{S}) = \Pi_{\mathbb{B}(R+1)}(t_m(\mathcal{S}))$ and $\rho = \alpha_0 = C_0\sqrt{(d + \log(1/\beta))/n}$, as in Proposition C.2 above. We apply Theorem 3.1 to obtain a bound on the error of $M^{\rho}_{\text{Inv}}(\mathcal{S}; \mathcal{A}_{\text{rob}})$. Let

$$\tau = \frac{2d \log\left(\frac{R+1}{\alpha_0} + 1\right) + 2\log(\frac{1}{\beta})}{n\varepsilon}.$$

[2] Knowledge of R is necessary for mean estimation under pure DP (Hardt and Talwar, 2010; Beimel et al., 2014; Bun et al., 2019).

Assume $\tau \leq 0.05$ and $\alpha_0 \leq 0.05$, which we will confirm later. By Proposition C.2, $\mathcal{A}_{\mathrm{rob}}$ is (τ, β, α)-robust for

$$\alpha = 7C_0 \cdot \sqrt{\frac{d + \log(\frac{1}{\beta})}{n}} + 7\left(\frac{2d\log\left(\frac{R+1}{\alpha_0} + 1\right) + 2\log(\frac{1}{\beta})}{n\varepsilon}\right)$$
$$\leq C' \cdot \left(\sqrt{\frac{d + \log(\frac{1}{\beta})}{n}} + \frac{d\log\left(\frac{Rn}{d}\right) + \log(\frac{1}{\beta})}{n\varepsilon}\right),$$

for constant $C' = 42C_0$. Notice that $\alpha_0 \leq \alpha$. Therefore, by Theorem 3.1, it holds that, with probability at least $1 - 2\beta$, $\|M^{\rho}_{\mathrm{Inv}}(\mathcal{S}; \mathcal{A}_{\mathrm{rob}}) - \mu\|_2 \leq 4\alpha \leq C \cdot \left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \frac{d\log(\frac{Rn}{d})+\log(\frac{1}{\beta})}{n\varepsilon}\right)$, for $C = 168C_0$. By assumption, n is sufficiently large so that the latter is less than 1 and as such, it also ensures that $\alpha_0 \leq 0.05$ and $\tau \leq 0.05$. The proof is complete by rescaling $\beta \leftarrow \beta/2$ and adjusting the constants. □

C.1.2. Unknown Covariance

We now move to the more general task of Gaussian mean estimation with unknown mean μ and covariance Σ, but with known *a priori* bounds R, κ such that $\mu \in \mathbb{B}^d(R)$ and $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. The error metric is the affine-invariant *Mahalanobis distance* with respect to Σ, defined by $\|\hat{\mu} - \mu\|_\Sigma := \sqrt{(\hat{\mu} - \mu)^\top \Sigma^{-1}(\hat{\mu} - \mu)}$. In Corollary C.3, we show that via our transformation, we retrieve known error bounds for Gaussian mean estimation with known parameters R, κ under pure DP (Bun et al., 2019; Liu et al., 2021).[3]

Corollary C.3 (Gaussian mean)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \Sigma)$ such that $\mu \in \mathbb{B}(R)$ and $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. Let $\varepsilon, \beta \in (0, 1)$ and $C \geq 1$ a known constant. Suppose n is such that $\alpha \leq 1$ in Equation (14). There exists an ε-DP algorithm $\mathcal{M}$ such that, with probability at least $1 - \beta$, has error $\|\mathcal{M}(\mathcal{S}) - \mu\|_\Sigma \leq \alpha$ for*

$$\alpha = C \cdot \left(\sqrt{\frac{d + \log(\frac{1}{\beta})}{n}} + \frac{d\log\left(\frac{(R+\sqrt{\kappa})n}{d}\right) + \log(\frac{1}{\beta})}{n\varepsilon}\right). \tag{14}$$

Again, we choose the projected Tukey median as our robust mechanism for this task. We state its guarantees for the Mahalanobis loss (proven in Appendix D).

Proposition C.4. *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \Sigma)$ such that $\mu \in \mathbb{B}(R)$ and $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. Let $\beta \in (0, 1)$, $\tau \leq 0.05$, and $\alpha_0 = C_0 \cdot \sqrt{(d + \log(1/\beta))/n}$ for known constant C_0. Suppose n is such that $\alpha_0 \leq 0.05$. Let $\alpha = 7(\alpha_0 + \tau) \leq 1$. The projected Tukey median algorithm $\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) = \Pi_{\mathbb{B}\left(R+\sqrt{\kappa}\right)}(t_m(\mathcal{S}))$ is (τ, β, α)-robust with respect to the Mahalanobis loss. That is, with probability $1 - \beta$, for any τ-corrupted $\mathcal{S}'$, such that $d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \leq n\tau$, it holds that $\|\mathcal{A}_{\mathrm{rob}}(\mathcal{S}') - \mu\|_\Sigma \leq \alpha$.*

Using the above proposition, the proof of Corollary C.3 is a straightforward application of Theorem A.2.

Proof of Corollary C.3. We let $L(u, v) = \|u - v\|_\Sigma$ be the loss function. As a norm, L satisfies the triangle inequality. Moreover, $\forall s, t \in \mathbb{R}^d$ $L(s, t) \leq c_L \|s - t\|_2$ for $c_L = 1$ since $\mathbb{I} \preceq \Sigma$ (by Proposition D.1 in Appendix D). Consider the ρ-smooth-inverse-sensitivity mechanism $M^{\rho}_{\mathrm{Inv}}(\cdot; \mathcal{A}_{\mathrm{rob}})$ with norm $\|\cdot\| = \|\cdot\|_2$, $\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) = \Pi_{\mathbb{B}\left(R+\sqrt{\kappa}\right)}(t_m(\mathcal{S}))$ and $\rho = \alpha_0 = C_0\sqrt{(d + \log(1/\beta))/n}$. We apply Theorem A.2 to obtain a bound on the mechanism's error with respect to L. Let

$$\tau = \frac{2d\log\left(\frac{R+\sqrt{\kappa}}{\alpha_0} + 1\right) + 2\log(\frac{1}{\beta})}{n\varepsilon}.$$

[3] These results are stated for $\Sigma = \mathbb{I}$, but can be extended to the case of unknown Σ such that $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, achieving the same error as in Corollary C.3 up to logarithmic factors.

Assume $\tau \leq 0.05$ and $\alpha_0 \leq 0.05$, which we will confirm later. By Proposition C.2, $\mathcal{A}_{\text{rob}}$ is (τ, β, α)-robust for

$$\alpha = 7C_0\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + 7\frac{2d\log\left(\frac{R+\sqrt{\kappa}}{\alpha_0}+1\right)+2\log(\frac{1}{\beta})}{n\varepsilon}$$
$$\leq C' \cdot \left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \frac{d\log\left(\frac{(R+\sqrt{\kappa})n}{d}\right)+\log(\frac{1}{\beta})}{n\varepsilon}\right),$$

for constant $C' = 28C_0$. Notice that $\alpha_0 \leq \alpha$. Therefore, by Theorem A.2, it holds that, with probability at least $1-2\beta$,

$$L(M^{\rho}_{\text{Inv}}(\mathcal{S};\mathcal{A}_{\text{rob}}),\mu) = \|M^{\rho}_{\text{Inv}}(\mathcal{S};\mathcal{A}_{\text{rob}})-\mu\|_{\Sigma} \leq 4\alpha \leq C\cdot\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \frac{d\log\left(\frac{(R+\sqrt{\kappa})n}{d}\right)+\log(\frac{1}{\beta})}{n\varepsilon}\right),$$

for $C = 112C_0$. By assumption, n is sufficiently large so that the latter is less than 1, and as such, it also ensures that $\alpha_0 \leq 0.05$ and $\tau \leq 0.05$. The proof is complete by rescaling $\beta \leftarrow \beta/2$ and adjusting the constants. □

C.2. Covariance Estimation

Given dataset $\mathcal{S} \overset{\text{iid}}{\sim} \mathsf{N}(0,\Sigma)^n$, where $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, our goal is to return an estimate $\hat{\Sigma} \in \mathbb{R}^{d\times d}$, with small error, measured by the *relative Frobenius norm*: $\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}-\mathbb{I}\right\|_{\text{F}}$. The task of covariance estimation for Gaussian distributions has been extensively studied both under robustness and differential privacy, and is particularly useful as a first step for learning a Gaussian distribution in total variation distance (see e.g. Corollary 2.14 in (Diakonikolas et al., 2016)). Note that the fact that the distribution is assumed to be zero-mean is w.l.o.g., as the general case can be reduced to the zero-mean case up to constant factors in the error, by letting the difference between a pair of nonzero-mean samples be a single zero-mean sample.

In Corollary C.5, we show that via our transformation, we retrieve the optimal known error bounds for Gaussian covariance estimation with known parameter κ under pure DP (Bun et al., 2019; Aden-Ali et al., 2021).[4]

Corollary C.5 (Gaussian covariance)**.** *Let $\mathcal{S} = (S_1,\ldots,S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(0,\Sigma)$ such that $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. Let $\varepsilon,\beta \in (0,1)$. Suppose n is such that $\alpha \leq 1$ in Equation (15). There exists an ε-DP algorithm $\mathcal{M}$ such that, with probability at least $1-\beta$, has error $\left\|\Sigma^{-1/2}\mathcal{M}(\mathcal{S})\Sigma^{-1/2}-\mathbb{I}\right\|_{\text{F}} \leq \alpha$ for*

$$\alpha = O\left(\left(\sqrt{\frac{d^2}{n}}+\frac{d^2}{n\varepsilon}\right)\cdot \text{polylog}(n\kappa/\beta)\right). \tag{15}$$

There are several algorithms in the robust statistics literature that achieve near-optimal bounds for robust covariance estimation of Gaussian distributions, which can serve as a good instantiation of our transformation. The next theorem states the robust accuracy guarantees of the algorithm proposed in (Diakonikolas et al., 2017).[5]

Theorem C.6 ((Diakonikolas et al., 2017))**.** *Let $\mathcal{S} = (S_1,\ldots,S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(0,\Sigma)$. Let $\beta \in (0,1), \tau \in (0,1)$. Suppose $n \geq \Omega\left(\frac{d^2\log^5(d/\tau\beta)}{\tau^2}\right)$. Let $\alpha' = O\left(\tau\log(1/\tau)\right)$. There exists algorithm $\mathcal{A}_{\text{rob}}$ which is (τ,β,α')-robust. That is, with probability $1-\beta$, for any τ-corrupted $\mathcal{S}'$, such that $d_{\text{H}}(\mathcal{S},\mathcal{S}') \leq n\tau$, it returns matrix $\mathcal{A}_{\text{rob}}(\mathcal{S}') = \hat{\Sigma}$ such that $\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}-\mathbb{I}\right\|_{\text{F}} \leq \alpha'$.*

Proof of Corollary C.5. We will run the ρ-smooth-inverse-sensitivity mechanism over vectors $\mathbb{R}^D$, $D = d^2$, with $\|\cdot\| = \|\cdot\|_2$. We denote by $\text{vec}(V) \in \mathbb{R}^{d^2}$ the *flattening* of a matrix $V \in \mathbb{R}^{d\times d}$, so that if $\text{vec}(V) = v$, then $V_{i,j} = v_{d(i-1)+j}$. Then $\|\text{vec}(U)-\text{vec}(V)\|_2 = \|U-V\|_{\text{F}}$. Let $\mathcal{A}$ be the robust algorithm established in Theorem C.6. We will instantiate our

[4] Knowledge of parameter κ is necessary for this task under pure DP (Bun et al., 2019; Alabi et al., 2022).

[5] This algorithm as well as other alternatives (Cheng et al., 2019; Li and Ye, 2020) are computationally efficient. It is possible that by using a computationally inefficient algorithm we would achieve smaller error up to logarithmic factors, but since we are not aiming to optimize for those factors, we chose the clearer statement from (Diakonikolas et al., 2017).

transformation with $\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) = \Pi_{\mathbb{B}^D(R')}(\mathrm{vec}(\mathcal{A}(\mathcal{S})))$ for $R' = 2\sqrt{d}\kappa$, that is, after flattening the output of $\mathcal{A}$, we take its euclidean projection on the $D = d^2$-dimensional ball of radius R' in ℓ_2 norm. Let $\hat{\Sigma} = \mathcal{A}(\mathcal{S})$. We have that

$$\begin{aligned}
\left\|\mathrm{vec}(\hat{\Sigma})\right\|_2 &= \left\|\hat{\Sigma}\right\|_{\mathrm{F}} \\
&\le \|\Sigma\|_{\mathrm{F}} + \left\|\hat{\Sigma} - \Sigma\right\|_{\mathrm{F}} && \text{(triangle inequality)} \\
&= \|\Sigma\|_{\mathrm{F}} + \left\|\Sigma\Sigma^{-1}(\hat{\Sigma} - \Sigma)\right\|_{\mathrm{F}} \\
&\le \|\Sigma\|_{\mathrm{F}} + \|\Sigma\|_{\mathrm{F}} \cdot \left\|\Sigma^{-1/2}(\hat{\Sigma} - \Sigma)\Sigma^{-1/2}\right\|_{\mathrm{F}} && (\|\cdot\|_{\mathrm{F}} \text{ sub-multiplicative, rotation-invariant}) \\
&\le \sqrt{d}\kappa\left(1 + \left\|\Sigma^{-1/2}(\hat{\Sigma} - \Sigma)\Sigma^{-1/2}\right\|_{\mathrm{F}}\right). && (\|\cdot\|_{\mathrm{F}} \le \sqrt{d}\,\|\cdot\|_2)
\end{aligned}$$

By Theorem C.6, the latter is at most $\sqrt{d}\kappa(1+\alpha')$ with probability $1-\beta$. Suppose $\alpha' \le 1$, which we will confirm last. Thus, with probability $1-\beta$, the projection on the euclidean ball with radius $R' = 2\sqrt{d}\kappa$ will not affect the output of the algorithm and $\mathcal{A}_{\mathrm{rob}}$ will have the same accuracy guarantees as stated in Theorem C.6.

We will let the loss function $L : (\mathbb{R}^{d\times d})^2 \to \mathbb{R}$ be $L(U,V) = \left\|\Sigma^{-1/2}(U-V)\Sigma^{-1/2}\right\|_{\mathrm{F}}$ over matrices U, V. Our goal is then to return a matrix U with small error $L(U,\Sigma)$.[6] Note that L satisfies the triangle inequality since the Frobenius norm does. For all $u, v \in \mathbb{R}^{d^2}$, let $V, U \in \mathbb{R}^{d\times d}$ denote their corresponding matrices. We have that $L(U,V) = \left\|\Sigma^{-1}(U-V)\right\|_{\mathrm{F}} \le \|(U-V)\|_{\mathrm{F}} = \|u-v\|_2$, since $\Sigma^{-1} \preceq \mathbb{I}$ and $\|\cdot\|_{\mathrm{F}}$ is monotone. It follows that L satisfies all the requirements of Theorem A.2.

Let $\alpha_0 = O(\sqrt{(d^2 + \log(1/\beta))/n}) < 1$, by assumption on n. We take τ which satisfies both $\tau = \Omega\left(\frac{2D\log(R'/\alpha_0+1)+2\log(1/\beta)}{n\varepsilon}\right) = \Omega\left(\frac{d^2\log(\kappa n/d)+\log(1/\beta)}{n\varepsilon}\right)$ (required by Theorem A.2) and $\tau = \Omega\left(\sqrt{\frac{d^2\log^5(d/\tau\beta)}{n}}\right)$ (required by Theorem C.6). Then $\mathcal{A}_{\mathrm{rob}}$ is (τ,β,α')-robust with

$$\alpha' = O\left(\left(\sqrt{\frac{d^2}{n}} + \frac{d^2}{n\varepsilon}\right)\cdot \mathrm{polylog}(n\kappa/\beta)\right).$$

We then have that $M^{\rho}_{\mathrm{Inv}}(\cdot, \mathcal{A}_{\mathrm{rob}})$ with $\rho = \alpha_0$ is ε-DP and with probability at least $1-2\beta$, returns matrix $\hat{V}$, which has error

$$L(\hat{V},\Sigma) = \left\|\Sigma^{-1/2}\hat{V}\Sigma^{1/2} - \mathbb{I}\right\|_{\mathrm{F}} \le 4\alpha' = \alpha.$$

By assumption, n is large enough so that $\alpha \le 1$ and as such $\alpha' < 1$ as well. The statement follows by rescaling $\beta \leftarrow \beta/2$ and adjusting the constants. □

C.3. Linear Regression

In this section, we apply our transformation to obtain an algorithm for linear regression for Gaussian data under pure DP. To the best of our knowledge, Corollary C.7 gives the first (computationally inefficient) algorithm for pure DP which achieves the optimal error rate up to logarithmic factors for Gaussian distributions. Liu et al. (2022) gave the analogous result under approximate DP.

Corollary C.7 (Gaussian Linear Regression). *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where for all $i \in [n]$, $S_i = (X_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ is generated by a linear model $y_i = X_i^\top\theta + \eta_i$ for some unknown $\theta \in \mathbb{B}^d(R)$, where $X_i \overset{\mathrm{iid}}{\sim} \mathsf{N}(0,\Sigma)$, $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, and $\eta_i \overset{\mathrm{iid}}{\sim} \mathsf{N}(0,\sigma^2)$, independent from X_i. Let $\varepsilon, \beta \in (0,1)$. Let*

$$\alpha = C\sigma\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \frac{d\log\left(\frac{(R/\sigma+\kappa)n}{d}\right)+\log(\frac{1}{\beta})}{n\varepsilon}\right), \tag{16}$$

for a known constant $C > 0$. Suppose n is such that $\alpha/\sigma \le c$ for a known constant $c \in (0,1)$. Then there exists an ε-DP algorithm $\mathcal{M}$ such that, with probability at least $1-\beta$, returns $\mathcal{M}(\mathcal{S}) = \hat{\theta}$ such that $\left\|\Sigma^{-1/2}(\hat{\theta}-\theta)\right\|_2 \le \alpha$.

[6] We can straightforwardly convert any vector $v \in \mathbb{R}^{d^2}$ to a unique matrix $V \in \mathbb{R}^{d\times d}$ such that $v = \mathrm{vec}(V)$.

Since the running time of the robust algorithm is not the bottleneck for the computational complexity of our proposed approach, we will instantiate our transformation with the (computationally inefficient) robust linear regression algorithm from (Gao, 2020). This algorithm achieves the information-theoretic optimal error for Gaussian distributions and is based on the notion of multivariate regression depth, similar to the Tukey depth we used for Gaussian mean estimation in Appendix C.1.[7]

Theorem C.8 (Theorem 3.2, (Gao, 2020))**.** *Consider the setting of Corollary C.7. Let $\beta \in (0,1), \tau \in (0,1)$. Suppose n and τ are such that $\tau + \sqrt{d/n} < c$ for a known constant $c \in (0,1)$. Then there exists constant $C' > 0$ and algorithm $\mathcal{A}_{\mathrm{rob}}$ which is (τ, β, α')-robust, for*

$$\alpha' = C'\sigma\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \tau\right). \tag{17}$$

That is, with probability $1-\beta$, for any τ-corrupted $\mathcal{S}'$, such that $d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \le n\tau$, it returns $\mathcal{A}_{\mathrm{rob}}(\mathcal{S}') = \hat{\theta} \in \mathbb{R}^d$ such that $\left\|\Sigma^{-1/2}(\hat{\theta} - \theta)\right\|_2 \le \alpha'$.

Proof of Corollary C.7. We will run the ρ-smooth-inverse-sensitivity mechanism in $\mathbb{R}^d$ with $\|\cdot\| = \|\cdot\|_2$. Let $\mathcal{A}$ be the robust algorithm established in Theorem C.8. We will instantiate our transformation with $\mathcal{A}_{\mathrm{rob}}(\mathcal{S}) = \Pi_{\mathbb{B}(R')}(\mathcal{A}(\mathcal{S}))$ for $R' = R + \sigma\sqrt{\kappa}$, that is, we take the euclidean projection of $\mathcal{A}(\mathcal{S})$ on the ball of radius R' in ℓ_2 norm. Let $\hat{\theta} = \mathcal{A}(\mathcal{S})$. We have that

$$\begin{aligned}
\left\|\hat{\theta}\right\|_2 &\le \|\theta\|_2 + \left\|\hat{\theta} - \theta\right\|_2 && \text{(triangle inequality)} \\
&\le R + \left\|\Sigma^{1/2}\right\|_2 \left\|\Sigma^{-1/2}(\hat{\theta} - \theta)\right\|_2 \\
&\le R + \sqrt{\kappa}\left\|\Sigma^{-1/2}(\hat{\theta} - \theta)\right\|_2.
\end{aligned}$$

Let $\alpha_0 = C'\sigma\sqrt{(d + \log(1/\beta))/n}$ for $C' > 0$ as in Theorem C.8 and

$$\tau = \frac{2d\log(R'/\alpha_0 + 1) + 2\log(1/\beta)}{n\varepsilon}.$$

Assume that n is such that $C'\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \tau\right) < c$ for $c \in (0,1)$ and for $C' > 0$ as in Theorem C.8, which we will confirm last. Then, the conditions of Theorem C.8 are satisfied, and with probability $1-\beta$, $R + \sqrt{\kappa}\left\|\Sigma^{-1/2}(\hat{\theta} - \theta)\right\|_2 \le R + \sqrt{\kappa}\alpha' \le R + \sigma\sqrt{\kappa} = R'$ and the projection will not affect the output of the algorithm $\mathcal{A}_{\mathrm{rob}}$.

We will let the loss function $L : (\mathbb{R}^d)^2 \to \mathbb{R}$ be $L(u,v) = \left\|\Sigma^{-1/2}(u-v)\right\|_2$. Our goal is then to return a vector u with small error $L(u,\theta)$. Note that L satisfies the triangle inequality. For all $u, v \in \mathbb{R}^d$, we have that $L(u,v) = \left\|\Sigma^{-1/2}(u-v)\right\|_2 \le \|u-v\|_2$, since $\Sigma^{-1/2} \preceq \mathbb{I}$. It follows that L satisfies all the requirements of Theorem A.2. Thus, $M^{\rho}_{\mathrm{Inv}}(\cdot, \mathcal{A}_{\mathrm{rob}})$ with $\rho = \alpha_0 < \alpha'$ is ε-DP and with probability at least $1-2\beta$, returns $\hat{u}$, which has error

$$L(\hat{u}, \theta) = \left\|\Sigma^{-1/2}(\hat{u} - \theta)\right\|_2 \le 4\alpha'.$$

That is, there exists $C > C'$, such that $\left\|\Sigma^{-1/2}(\hat{u} - \theta)\right\|_2 \le \alpha$, for

$$\alpha = C\sigma\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \frac{d\log((R+\sigma\sqrt{\kappa})n/(\sigma d)) + \log(1/\beta)}{n\varepsilon}\right).$$

By assumption n is sufficiently large so that the latter is smaller than σc, and as such, it ensures that $C'\left(\sqrt{\frac{d+\log(\frac{1}{\beta})}{n}} + \tau\right) < c$ as well. The proof is complete by rescaling $\beta \leftarrow \beta/2$ and adjusting the constants. □

[7] The result is stated for the weaker Huber's contamination model, but it holds under the strong contamination model as well.

C.3.1. SPARSE LINEAR REGRESSION

We now apply our transformation to obtain an algorithm for sparse linear regression for Gaussian data under pure DP, which, to the best of our knowledge, is the first (computationally inefficient) algorithm this case that achieves near-optimal error rate. When the solution is known to be k-sparse, our transformation allows us to improve the dependence on dimension from $d/n\varepsilon$ to $k \log d/(n\varepsilon)$ as we show in the next corollary.

Corollary C.9 (Sparse Linear Regression)**.** *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where for all $i \in [n]$, $S_i = (X_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ is generated by a linear model $y_i = X_i^\top \theta + \eta_i$ for some unknown $\theta \in \mathbb{B}^d(R)$, $\|\theta\|_0 \le k$, where $X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \Sigma)$, $\mathbb{I} \preceq \Sigma \preceq \kappa \mathbb{I}$, and $\eta_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$, independent from X_i. Let $\varepsilon, \beta \in (0,1)$. Let*

$$\alpha = C\sigma \left(\sqrt{\frac{k \log(\frac{ed}{k}) + \log(\frac{1}{\beta})}{n}} + \frac{k \log(\frac{ed}{k}) + k \log\left(\frac{(R/\sigma+\kappa)n}{d}\right) + \log(\frac{1}{\beta})}{n\varepsilon} \right), \tag{18}$$

for a known constant $C > 0$. Suppose n is such that $\alpha/\sigma \le c$ for a known constant $c \in (0,1)$. Then there exists an ε-DP algorithm $\mathcal{M}$ such that, with probability at least $1-\beta$, returns $\mathcal{M}(\mathcal{S}) = \hat{\theta}$ such that $\left\|\Sigma^{-1/2}(\hat{\theta} - \theta)\right\|_2 \le \alpha$.

We use the robust algorithm for sparse linear regression by (Gao, 2020).

Theorem C.10 (Theorem 3.2, (Gao, 2020))**.** *Consider the setting of Corollary C.9. Let $\beta \in (0,1), \tau \in (0,1)$. Suppose n and τ are such that $\tau + \sqrt{k \log(ed/k)/n} < c$ for a known constant $c \in (0,1)$. Then there exists constant $C' > 0$ and algorithm $\mathcal{A}_{\text{rob}}$ which is (τ, β, α')-robust, for*

$$\alpha' = C'\sigma \left(\sqrt{\frac{k \log(\frac{ed}{k}) + \log(\frac{1}{\beta})}{n}} + \tau \right). \tag{19}$$

That is, with probability $1-\beta$, for any τ-corrupted $\mathcal{S}'$, such that $d_{\text{H}}(\mathcal{S}, \mathcal{S}') \le n\tau$, it returns $\mathcal{A}_{\text{rob}}(\mathcal{S}') = \hat{\theta} \in \mathbb{R}^d$ such that $\left\|\Sigma^{-1/2}(\hat{\theta} - \theta)\right\|_2 \le \alpha'$.

The proof of Corollary C.9 follows exactly the same steps as Corollary C.7, but uses the slightly modified inverse-sensitivity mechanism for sparse estimation and its guarantees in Theorem A.5 instead of Theorem 6.2.

D. Useful Facts and Proofs for Applications

D.1. Linear Algebra Facts and Definitions

We denote by $\|v\|_M = \|M^{-1/2}v\| = \sqrt{v^\top M^{-1} v}$ the Mahalanobis norm of vector v with respect to M for any positive definite matrix M. Observe that $\|v\|_{\mathbb{I}} = \|v\|_2$.

Proposition D.1. *For positive definite matrices Σ_1, Σ_2, if $\Sigma_1 \preceq \Sigma_2$, then for any vector v, $\|v\|_{\Sigma_2} \le \|v\|_{\Sigma_1}$.*

Let $A \in \mathbb{R}^{d \times d}$. We denote the *spectral norm* of A by $\|A\|_2 = \sup\{\|Ax\|_2 : x \in \mathbb{R}^d \text{ s.t. } \|x\|_2 = 1\}$ and its *Frobenius norm* by $\|A\|_F = \sqrt{\sum_{j=1}^d \sum_{i=1}^d |A_{i,j}|^2}$. It holds that $\|A\|_2 \le \|A\|_{\text{F}} \le \sqrt{d}\,\|A\|_2$.

D.2. Robustness Guarantee of Tukey Median

We first state known properties of the Tukey depth for Gaussian datasets. The next proposition relates the Tukey depth of a point to its Mahalanobis distance from the mean (see e.g. Proposition D.2 in (Brown et al., 2021) for a proof). Here, Φ is the CDF of the univariate standard Gaussian.

Proposition D.2. *For any $\mu, y \in \mathbb{R}^d$ and positive definite Σ, $T_{\mathsf{N}(\mu,\Sigma)}(y) = \Phi(-\|y - \mu\|_\Sigma)$.*

The next proposition states the uniform convergence property of Tukey depth. It follows from standard uniform convergence of halfspaces (Vapnik and Chervonenkis, 1971), extended to the definition of Tukey depth (Donoho and Gasko, 1992; Burr and Fabrizio, 2017) (see e.g. (Liu et al., 2021) for a complete proof).

Proposition D.3 (Convergence of Tukey Depth). *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \Sigma)$. There exists constant C_0 such that, with probability $1-\beta$, for any $v \in \mathbb{R}^d$, $|T_{\mathsf{N}(\mu,\Sigma)}(v) - T_{\mathcal{S}}(v)| \leq C_0 \cdot \sqrt{\frac{d+\log(1/\beta)}{n}}$.*

Proposition D.4 (Robust Accuracy of Tukey Median, Restatement of Proposition C.4). *Let $\mathcal{S} = (S_1, \ldots, S_n)$ where $S_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \Sigma)$ such that $\mu \in \mathbb{B}(R)$ and $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. Let $\beta \in (0,1)$, $\tau \leq 0.05$, and $\alpha_0 = C_0 \cdot \sqrt{(d+\log(1/\beta))/n}$ as in Proposition D.3. Suppose n is such that $\alpha_0 \leq 0.05$. Let $\alpha = 7(\alpha_0 + \tau) \leq 1$. The projected Tukey median algorithm $\mathcal{A}_{rob}(\mathcal{S}) = \Pi_{\mathbb{B}\left(R+\sqrt{\kappa}\right)}(t_m(\mathcal{S}))$ is (τ, β, α)-robust with respect to the Mahalanobis loss. That is, with probability $1-\beta$, for any τ-corrupted $\mathcal{S}'$, such that $d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \leq n\tau$, it holds that $\|\mathcal{A}_{rob}(\mathcal{S}') - \mu\|_\Sigma \leq \alpha$.*

Proof. Let $\mathcal{S}'$ be any τ-corruption of $\mathcal{S}$, that is, $d_{\mathrm{H}}(\mathcal{S}, \mathcal{S}') \leq n\tau$. Observe that $|T_{\mathcal{S}}(v) - T_{\mathcal{S}'}(v)| \leq \tau$ for any $v \in \mathbb{R}^d$ by the definition of Tukey depth. Let $t'_m = \operatorname{argmax}_{v\in\mathbb{R}^d} T_{\mathcal{S}'}(v)$ be the Tukey median of the corrupted dataset. We condition on the event that the bound of Proposition D.3 holds, which occurs with probability $1-\beta$. We have that

$$
\begin{aligned}
& T_{\mathsf{N}(\mu,\Sigma)}(\mu) = \frac{1}{2} && \text{(by Proposition D.2 since } \Phi(0) = \tfrac{1}{2}\text{)} \\
& \Rightarrow T_{\mathcal{S}}(\mu) \geq \frac{1}{2} - \alpha_0 && \text{(by Proposition D.3)} \\
& \Rightarrow T_{\mathcal{S}'}(\mu) \geq \frac{1}{2} - \alpha_0 - \tau \\
& \Rightarrow T_{\mathcal{S}'}(t'_m) \geq \frac{1}{2} - \alpha_0 - \tau && \text{(by definition of } t'_m\text{)} \\
& \Rightarrow T_{\mathcal{S}}(t'_m) \geq \frac{1}{2} - \alpha_0 - 2\tau \\
& \Rightarrow T_{\mathsf{N}(\mu,\Sigma)}(t'_m) \geq \frac{1}{2} - 2\alpha_0 - 2\tau && \text{(by Proposition D.3)} \\
& \Rightarrow \Phi(-\|t'_m - \mu\|_\Sigma) \geq \frac{1}{2} - 2\alpha_0 - 2\tau && \text{(by Proposition D.2)} \\
& \Rightarrow \frac{1}{2}\mathrm{Erf}\left(\frac{\|t'_m - \mu\|_\Sigma}{\sqrt{2}}\right) \leq 2(\alpha_0 + \tau) && \text{(since } \Phi(-z) = \tfrac{1}{2} - \tfrac{1}{2}\mathrm{Erf}\left(\tfrac{z}{\sqrt{2}}\right)\text{)}
\end{aligned}
$$

It is easy to see that the following bound holds for the error function $0.84z \leq \mathrm{Erf}(z)$ for $z \in [0,1]$ (see e.g. Lemma 3.2 in (Canonne et al., 2020)). It follows that, $\|t'_m - \mu\|_\Sigma \leq \frac{4\sqrt{2}}{0.84}(\alpha_0 + \tau) \leq 7(\alpha_0 + \tau)$ for $\alpha_0 + \tau \leq 1/7$, which holds by assumption. Thus, with probability $1-\beta$, $\|t'_m - \mu\|_\Sigma \leq \alpha$, for $\alpha = 7(\alpha_0 + \tau) \leq 1$. Since we have assumed that $\|\mu\|_2 \leq R$, it follows that $\|t'_m\|_2 \leq \|\mu\|_2 + \|t'_m - \mu\|_2 \leq R + \sqrt{\kappa}\,\|t'_m - \mu\|_\Sigma \leq R + \sqrt{\kappa}\alpha \leq R + \sqrt{\kappa}$, where the second inequality holds due to Proposition D.1. Then $t'_m \in \mathbb{B}(R + \sqrt{\kappa})$ and the projection will not affect the output. □