Perturb-and-Project: Differentially Private Similarities and Marginals

Vincent Cohen-Addad [*1] **Tommaso d'Orsi** [*1 2] **Alessandro Epasto** [*1] **Vahab Mirrokni** [*1] **Peilin Zhong** [*1]

Abstract

We revisit the input perturbations framework for differential privacy where noise is added to the input $A \in \mathcal{S}$ and the result is then projected back to the space of admissible datasets $\mathcal{S}$. Through this framework, we first design novel efficient algorithms to privately release pair-wise cosine similarities. Second, we derive a novel algorithm to compute k-way marginal queries over n features. Prior work could achieve comparable guarantees only for k even. Furthermore, we extend our results to t-sparse datasets, where our efficient algorithms yields novel, stronger guarantees whenever $t \leqslant n^{5/6}/\log n$. Finally, we provide a theoretical perspective on why *fast* input perturbation algorithms works well in practice. The key technical ingredients behind our results are tight sum-of-squares certificates upper bounding the Gaussian complexity of sets of solutions.

1. Introduction

Differential privacy (DP) (Dwork et al., 2014) has become the golden standard to define how much information an algorithm leaks about users' data. Thus, designing differentially private algorithms is a central problem in modern machine learning. One of the key considerations in designing DP algorithms is to determine the minimum amount of noise needed to ensure privacy to maximize the accuracy of the output: Insufficient noise may result in a non-private algorithm while excessive noise can significantly degrade the algorithm's output quality. It is thus key to quantify how much a worst-case user impacts (Dwork et al., 2006) the final output of the algorithm to add the appropriate noise.

An interesting example is the case of empirical risk minimization (ERM) problems. Iterative optimization algorithms such as stochastic gradient descent (SGD) are central algorithms in modern machine learning. However, making these algorithms differentially private is particularly challenging (Chaudhuri et al., 2011); repeated noise addition is necessary, leading to an accumulation of noise that can significantly degrade the final solution's quality. Furthermore, precisely quantifying the required noise at each iteration is often difficult, potentially resulting in overly conservative noise addition.

Another widely-used approach to ERM problems is objective perturbation (Chaudhuri & Monteleoni, 2008; Chaudhuri et al., 2011; Iyengar et al., 2019; Kifer et al., 2012), which involves perturbing the objective function so that optimizing the perturbed objective ensures that the output is private. One very generic way of achieving objective perturbation is through *input perturbation* which consists in adding noise to the input dataset to obtain a private perturbed input. This permits the use of any non-DP algorithms on the perturbed input and hence simplifies practical implementation. One benefit over perturbing the objective is that the properties (e.g., convexity) of the objective are unchanged and so the state-of-the-art non-DP optimizers can be used. Experimentation with various non-DP algorithms becomes possible, and privacy guarantees are immediate. A crucial research direction is thus to investigate input perturbation methods that preserve the intrinsic properties of the data that are beneficial for the downstream task.

Our Contribution In this work, we study input perturbation techniques from a theoretical perspective. We focus on the problem of designing differentially private projection functions (such as e.g., dimensionality reduction techniques). We observe that, quite surprisingly, input perturbations yield the best known (and conjecturally optimal) guarantees for a large class of projection functions. This challenges the expectation that a general approach might incur sub-optimal utility loss. Our analysis reveals that the algorithm's utility guarantees are contingent upon the "richness" (metric entropy) of the target set (the set into which the input is projected). Inspired by (d'Orsi et al., 2023), we use sum-of-squares to obtain tight bounds on the Gaussian complexity of set of solutions. To further improve, we introduce new sum-of-square certificates (e.g. for sparse injective tensor norms). Finally, while our sum-of-squares

[*]Equal contribution [1]Google Research [2]BIDSA, Bocconi. Correspondence to: Tommaso d'Orsi <tommaso.dorsi@unibocconi.it>, Alessandro Epasto <aepasto@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

solution might appear complex, we extract an interesting explanation for the practical success of algorithms like linear projection. At a high-level, we exploit the fact that the sum-of-squares projections onto convex sets can be broken down into projections onto simpler convex sets (this is a consequence of results on alternating projections, see (Bauschke & Borwein, 1993; Sakai, 1995)). The result is that for polynomially bounded inputs, this can thus be achieved by using a small (often only logarithmic) number of basic projections, a popular approach in practice.

2. Results

Privately releasing pair-wise distances Our first result concerns the task of privately releasing pair-wise cosine similiarities of a set of n vectors (Yang et al., 2017; Fernandes et al., 2021). This is a fundamental subroutine of many algorithms (e.g. nearest neighbors search). Without any constraint on the vectors structure, one should not expect to output a reasonable estimate without sacrificing privacy. This is because even a single change in a single entry of a vector can drastically alter all its inner products (for example if the vector is sparse). We consider the following natural notion of adjacency over matrices $V, V' \in \mathbb{R}^{n\times m}$, related notions appeared in (Blocki et al., 2012; Imola et al., 2023),

$$\begin{aligned}\Delta &\geqslant \sum_{ij} \left(\langle V_i, V_j\rangle - \langle V'_i, V'_j\rangle\right)^2 \\ &= \left\|VV^\intercal - V'V'^\intercal\right\|_{\mathrm{F}} .\end{aligned}$$

That is, the sum of the squared differences between the inner products of the rows of V and V' is bounded by some chosen parameter Δ. Notice that, for example, for set of vectors in $\{\pm 1/\sqrt{n}\}^n$, this definition captures $O(\sqrt{n})$ entry-wise changes.

Theorem 2.1. *Let $V =: \{v_1, \ldots, v_n\} \subseteq \mathbb{R}^m$ be a set of unit vectors. There exists an (ε, δ)-differentially private algorithm that, on input V, returns a matrix $\hat{\mathbf{X}} \in \mathbb{R}^{n\times n}$ satisfying*[1]

$$\mathbb{E}\left\|VV^\intercal - \hat{\mathbf{X}}\right\|_{\mathrm{F}}^2 \leqslant O\left(\frac{\Delta\cdot\sqrt{\log(2/\delta)}}{\varepsilon}\right)\cdot n^{3/2} .$$

Moreover, the algorithm runs in polynomial time.

Prior works focused either on vector level differential privacy for restricted classes of vectors (Kenthapadi et al., 2012; Blocki et al., 2012), or considered local extended differential privacy (Fernandes et al., 2021), making a fair comparison somewhat hard. In the latter case, when extended to standard differential privacy, the resulting guarantees are worse than Theorem 2.1.In the former, we observe that (Kenthapadi et al., 2012) improves over the naive Gaussian mechanism only for sparse vectors with $o(n)$ non-zero entries. In contrast, compared to Theorem 2.1 the Gaussian mechanism would yield guarantees worse by a $\sqrt{n}$ factor, for any set of vectors Moreover, as we discuss in Remark 4.1, evidence suggests the bound of Theorem 2.1 may be information theoretically optimal.

K-way marginals We also apply our framework to privately compute k-way marginal queries (Kasiviswanathan et al., 2010). Roughly, speaking, in these settings, the input is a dataset of vectors in $\{0,1\}^n$ and the goal is to answer queries of the form: *"How many vectors are non-zero in the entries in $V \subseteq [n], |V| \leqslant k$?* The definition of adjacency here is the natural notion in which one vector in the dataset is replaced. For k-way marginal queries we obtain the following two results (formally stated in Appendix A).

Theorem 2.2 (K-way marginals, informal)**.** *Let D be a dataset of m elements, each a binary vector in $\{0,1\}^n$. There exists an (ε, δ)-differentially private algorithm that, on input D, k, answers all k-way marginal queries with expected average query-wise squared error*

$$O\left(m\cdot n^{k/4}\cdot\frac{\sqrt{\log(2/\delta)}}{\varepsilon}\right)$$

when k is even,

$$O\left(m\cdot (n^k\cdot k\cdot\log n)^{1/4}\cdot\frac{\sqrt{\log(2/\delta)}}{\varepsilon}\right)$$

when k is odd. Moreover, the algorithm runs in time $O(m)\cdot n^{O(k)}$.

Notice the algorithm runs in time polynomial in the number of queries and the dataset size.

It is known that simply applying the Gaussian mechanism to the input achieves error of the order $O\left(\frac{\log(2/\delta)}{\varepsilon^2}\right)\cdot n^k$, which is information theoretically optimal when $m \geqslant n^{3k/4}$, up to the dependency on ε, δ. In the more realistic settings of $m \leqslant n^{3k/4}$, Theorem 2.2 non-trivially improves over this naive algorithm. In comparison, existing polynomial time algorithms (Talwar et al., 2014) could obtain guarantees comparable to Theorem 2.2 only for *even* k. In the odd case, these algorithms yielded an average query wise squared error larger by an $n^{1/2}/(k\log n)$ multiplicative factor.[2] Furthermore, the algorithm of (Talwar et al., 2014) required constrained convex optimization methods in which random noise is injected at each iteration.

[1] We use boldface to denote random variables.

[2] This gap appears in (Talwar et al., 2014) due to the fact that even order k tensors can be flattened into $k/2 \times k/2$ matrices and odd order tensors cannot.

Disregarding computational complexity, exponential time algorithms (Gupta et al., 2011; Hardt et al., 2012; Hardt & Rothblum, 2010) are known to achieve entry-wise error $m \cdot n^{1/4}$. This bound is known to be tight (Kasiviswanathan et al., 2010) and it is matched, up to constant factors, by Theorem 2.2 for $k = 2$. Given that the k-way marginal queries are the entries of the k-th order tensor of the dataset (see Section 4), this information-computation gap appears to be related to the well-known conjectural hardness of computing injective tensor norms of random tensors (Bhattiprolu et al., 2016; Hopkins et al., 2017; Brennan & Bresler, 2020).

Despite these strong negative results, it turns out that, for t-sparse datasets,[3] one can achieve significantly higher utility than Theorem 2.2 whenever $t \leqslant n^{5/6}/\log n$.

Theorem 2.3 (K-way marginals for sparse-datasets, informal)**.** *Let D be a dataset of m elements, each a a binary vector in $\{0,1\}^n$ with at most t non-zero entries. There exists an (ε, δ)-differentially private algorithm that, on input D , k , t, answers all k-way marginal queries with expected average query-wise squared error*

$$O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon}\right) \cdot \left(m \cdot (t^{3/2}/n)^k \cdot \sqrt{k \log n / t^2}\right)$$

when k is even,

$$O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon}\right) \cdot \left(m \cdot (t^{3/2}/n)^k \cdot \sqrt{k \log n}\right)$$

when k is odd. Moreover, the algorithm runs in time $O(m) \cdot n^{O(k)}$.

The error bound of Theorem 2.3 may seem counter-intuitive as it becomes vanishing small as $t^{3/2}/n$ decreases. However, this is a consequence of the fact that in such settings most queries must have value 0. In particular, the following trivial algorithm achieves vanishing small error for sufficiently small t^2/n ratio: add Gaussian noise and remove all but the largest $m \cdot t^k$ entries.

Similarly to the dense settings, the error of Theorem 2.3 is related to the existing (conditional) computational lower bounds for certifying sparse injective tensor norms of random tensors (Choo & d'Orsi, 2021) (see Section 4).

3. Related work

Differential privacy (Dwork et al., 2006) has emerged in the past decades as the *de facto* privacy standard in the design of private algorithms. There is an ever-growing literature on DP for which we refer to the survey of Dwork et al. (2014). Our work spans many areas of DP algorithms including similarity and distance approximation and k-way marginal estimation. We now review the most relevant work in each of these areas.

Similarity and distance approximation A related area is that of approximately (and privately) preserving distance and similarity measures among metric data. A highly celebrated result in the non-private literature is that the Johnson-Lindenstrauss (JL) linear projection preserves approximately the distances of Euclidean points (Dasgupta & Gupta, 2003). More recently Kenthapadi et al. (2012) showed that JL projection with additional noise can be used to preserve distances while achieving vector level DP. Blocki et al. (2012) showed instead that, under certain assumptions on the data, the JL projection itself is sufficient to achieve DP (without any additional noise). Subsequent work extended such analysis to private sparse JL projections (Stausholm, 2021).

Related projection-based methods have found wide application in privacy settings for instance in solving PCA problems (Chaudhuri et al., 2013), answering statistical queries (Nikolov, 2023), releasing synthetic data (Xu et al., 2017), answering distance based queries (Yang et al., 2018), as well as addressing computer vision problems (Zhu et al., 2020).

For the related problem of preserving similarity (as opposed to distance) measures, Aumüller et al. (2020) provided Local DP (LDP) algorithms for Jaccard Similarity estimation in item sets. Yang et al. (2017) instead applies JL-based methods for approximating the cosine similarity. On a related problem, Fernandes et al. (2021) designed approximate cosine similarity computations in an LDP relaxation known as local extended DP (or d_χ privacy). Our work provides improved bounds for cosine similarity computation over such work for the model of central DP (as opposed to LDP).

On a different direction, Alaggan et al. (2011); Choi et al. (2023) used cryptographic schemes for cosine similarity based on public-private key encryption and fully-homomorphic encryption. These works require cryptographic assumptions for privacy and focus on the 2-party computations (i.e., two users interested in computing the similarity of their data). Schoppmann et al. (2018) studied instead secure multi-party computation for performing similarity search in documents.

From an application point of view, Battaglia et al. (2021) studied the problem of inferring a distance measure over categorical values in central DP setting using the DILCA method; Xue & Yuqing (2016); Zhang et al. (2020) studied recommender systems with differential privacy, while Ding et al. (2019); Chen et al. (2019); Wang et al. (2020) presented privacy-aware methods for finding similar items.

[3] We say a dataset is t-sparse if all vectors in the dataset have at most t non-zero entries.

k-way marginals Marginal tables and contingency tables are a high level synopsis of a multi-dimensional dataset. The problem has been studied in the context of DP for a long time. Gupta et al. (2011); Hardt et al. (2012); Hardt & Rothblum (2010); Kasiviswanathan et al. (2010) showed lower-bounds and algorithms for answering such queries. Chandrasekaran et al. (2014); Dwork et al. (2015) focused on efficient algorithms for answering high dimensional marginals with DP. Nikolov (2023) used JL projections to answer 2-way marginals while Cormode et al. (2018) focused on a local DP version of the problem. From an application point of view marginal tables can be used to produce synthetic data (Li et al., 2017).

Finally a related problem to 2-way marginal is covariance matrix estimation. Several authors (Dong et al., 2022; Blocki et al., 2012; Mangoubi & Vishnoi, 2023) studied the problem in a private setting.

Sum-of-squares-based algorithms Our work is tightly linked to recent advances in robust statistics (Hopkins et al., 2015; Hopkins & Li, 2018; d'Orsi et al., 2020; Kothari et al., 2018; Diakonikolas & Kane, 2019; Ding et al., 2022; 2023; Bakshi et al., 2020; 2022; Klivans et al., 2018; Liu & Moitra, 2022)). Very recently, the sum-of-squares framework over which we build upon, has found surprising applications in the context of differential privacy (Hopkins et al., 2023; Georgiev & Hopkins, 2022; Hopkins et al., 2022) including clustering (Chen et al., 2023) and moment estimation (Kothari et al., 2022). Relevant to our techniques are also the results on consistent estimation of (Tsakonas et al., 2014; d'Orsi et al., 2021; d'Orsi et al., 2021; 2023), see Section 4.

Organization

The rest of the paper is organized as follows. We provide a technical overview in Section 4. In Section 5 we prove general guarantees for the perturb-and-project framework. Here we also show how to design practical implementations of the algorithm and what the resulting guarantees are. We then use these results to obtain Theorem 2.1 in Section 6 and to obtain Theorem 2.2, Theorem 2.3 in Appendix A. Background notions and definitions used throughout the paper can be found in Appendix B.

Notation

We describe here the notation used throughout the paper and some relevant preliminary notions. We hide absolute constant multiplicative factors using the standard notation $O(\cdot)\,,\Omega(\cdot)\,,\Theta(\cdot)$. We denote random variables in **boldface**. We write $o(1),\omega(1)$ for functions tending to zero (resp. infinity) as n grows. We say that an event happens with high probability if this probability is at least $1-o(1)$. Throughout the paper, when we say "an algorithm runs in time $O(q)$" we mean that the number of basic arithmetic operations involved is $O(q)$.

Vectors, matrices, tensors We use Id_n to denote the n-by-n dimensional identity matrix, $\bar{0}_n \in \mathbb{R}^n$ to denote the zero vector. When the context is clear we drop the subscript. For matrices $A, B \in \mathbb{R}^{n\times n}$ we write $A \succeq B$ if $A - B$ is positive semidefinite. We denote by $\|M\|$ the spectral norm of M, by $\|M\|_{\text{nuc}}$ its nuclear norm, by $\|M\|_{\text{F}}$ its frobenius norm and by $\|X\|_{\text{max}}$ the largest absolute value of any of its entries. We use M_i and $M_{-,i}$ to respectively denote the i-th row and column of M . We denote by $(\mathbb{R}^n)^{\otimes k}$ the set of real-valued order-k tensors. For a $n \times n$ matrix M, we denote by $M^{\otimes k}$ the *t-fold Kronecker product* $\underbrace{M \otimes M \otimes \cdots \otimes M}_{k \text{ times}}$. We define the *flattening*, or *vectorization*, of M to be the n^k-dimensional vector, whose entries are the entries of M appearing in lexicographic order. With a slight abuse of notation we refer to this flattening with M, ambiguities will be clarified form context. We denote by $N\left(0,\sigma^2\right)^{n^{\otimes k}}$ the distribution over Gaussian tensors with n^k independent entries with standard deviation σ.

Sparse vectors and norm We denote number of non-zero entries of a vector $v \in \mathbb{R}^n$ by $\|v\|_0$. Hence a t-sparse vector v is a vector satisfying $\|v\|_0 \leqslant t$. For an order-k tensor M, we define its symmetric injective t-sparse norm as $\max_{v\in\mathbb{R}^n\,,\|v\|=1\,,\|v\|_0\leqslant t}\langle M, v^{\otimes k}\rangle$.

Sets and projections We use calligraphic letters to denote sets. Given $A \in \mathbb{R}^n$ and a compact convex set $\mathcal{S} \in \mathbb{R}^n$, we denote by $\Pi_{\mathcal{S}}(A)$ the projection of A onto $\mathcal{S}$, $\Pi_{\mathcal{S}}(A) := \arg\min_{X\in\mathcal{S}} \|A - X\|_2^2$. For two sets $\mathcal{S}, \mathcal{S}'$ we denote their symmetric difference by $S \triangle S'$. We write $\mathcal{G}(\mathcal{S})$ for the Gaussian complexity of $\mathcal{S}$.

Notions about Gaussian complexity, differential privacy and sum-of-squares are deferred to Appendix B.

4. Techniques

We present here the main ideas behind our results.

The perturb-and-project framework The general problem we consider can be described as follows: given a vector $A \in \mathbb{R}^n$, and a fixed space $\mathcal{S} \subseteq \mathbb{R}^n$, we would like to output the projection of A onto $\mathcal{S}$ while preserving differential privacy with respect to A (for some problem dependent notion of adjacency). Our algorithmic approach can be then described as follows.

Notice that the set $\mathcal{S}$ is a parameter that is *independent* of the input vector A. It is not hard to show that, for compact

Algorithm 1 Perturb-and-project

Input: $\varepsilon, \delta, \Delta > 0$, $A \in \mathbb{R}^n$, compact convex set $\mathcal{S} \subseteq \mathbb{R}^n$.
Output: $\hat{\mathbf{\Pi}} \in \mathcal{S}$.
Let $\mathbf{W} \sim N\left(0, \mathrm{Id}_n \cdot \frac{2\log(2/\delta)}{\varepsilon^2} \cdot \Delta^2\right)$.
Return: $\hat{\mathbf{\Pi}} := \arg\min_{X \in \mathcal{S}} \|X - A + \mathbf{W}\|_2^2$.

convex sets, the strong convexity of the projection guarantees

$$\mathbb{E}\,\|\Pi_{\mathcal{S}}(A) - \Pi_{\mathcal{S}}(A + \mathbf{W})\|_2^2 \leqslant O(\tfrac{2\Delta\sqrt{\log(2/\delta)}}{\varepsilon^2}) \cdot \mathcal{G}(\mathcal{S}) \tag{1}$$

where $\mathcal{G}(\mathcal{S}) := \mathbb{E}_{\mathbf{W}^* \sim N(0,\mathrm{Id}_n)}[\sup_{X \in \mathcal{S}} \langle X, \mathbf{W}^* \rangle]$ is the Gaussian complexity of $\mathcal{S}$ and $\Pi_{\mathcal{S}}(\cdot)$ denotes the orthogonal projection onto $\mathcal{S}$. (Related results exist in the literature, see Section 5 for a proof and a proper comparison). The significance of Equation (1) is in the implication that what governs the error is the structure of the space $\mathcal{S}$. Indeed, by Sudakov's minoration (Wainwright, 2019) the minimal size of any α-net over $\mathcal{S}$ is bounded by $\exp(\mathcal{G}(\mathcal{S})^2/\varepsilon^2)$. If $\mathcal{S}$ in this set are heavily structured, then the resulting error will be significantly better than naively using the Gaussian mechanism.

A simple application: pair-wise cosine similarities As a concrete example, let $\{v_1, \ldots, v_n\}$ be a set of vectors in $\{\pm 1/\sqrt{m}\}^m$ (this assumption is only for simplicity of the exposition) and let $V \in \mathbb{R}^{n \times m}$ be the matrix with rows $V_i = v_i$. The problem of privately releasing the cosine similarities between these vectors is equivalent to the problem of privately realising VV^T. Notice that this matrix falls in the set

$$\mathcal{S} := \left\{ X \in \mathbb{R}^{n \times n} \;\middle|\; X \succeq 0\,, X_{ii} \leqslant 1\,, \forall i \right\}. \tag{2}$$

In other words, this task adheres perfectly to the perturb-and-project framework with $A = VV^\mathsf{T}$. Hence we can apply the Gaussian mechanism and then project onto the set in Equation (2). Since by Holder's inequality for the nuclear and the spectral norm,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{W} \sim N(0,\mathrm{Id}_n)^{n \times n}} \left[\sup_{X \in \mathcal{S}} \langle X, \mathbf{W} \rangle\right] &\leqslant \mathbb{E}\left[\|X\|_{\text{nuc}}\, \|\mathbf{W}\|\right] \\ &\leqslant n \cdot 2\sqrt{n} \cdot \sigma\,,\end{aligned}$$

we immediately recover the guarantees of Theorem 2.1 for the appropriate choice of σ.

Tight relaxations and sum-of-squares In the above examples, the projection onto the desired set $\mathcal{S}$ could be computed efficiently. However, this is not the case in general. For example, if A is (the flattening of) a rank-m k-th order symmetric tensor, it stands to reason one may want to project the perturbed input onto the set of rank-m k-order symmetric tensors[4]

$$\left\{ X \in (\mathbb{R}^n)^{\otimes k} \;\middle|\; \sum_{i \in [m]} v_i^{\otimes k}\,, \|X\|_{\max} \leqslant 1 \right\}. \tag{3}$$

Unfortunately, carrying out such a projection is (conjecturally) hard even on average (Bhattiprolu et al., 2016; Hopkins et al., 2017; Brennan & Bresler, 2020). To overcome this issue, one may replace $\mathcal{S}$ with a superset $\mathcal{S}'$ over which we can efficiently project. The hope is that, if $\mathcal{S}'$ is not too "large", then $\mathcal{G}(\mathcal{S}')$ is close to $\mathcal{G}(\mathcal{S})$ and therefore the error is not significantly larger. Following (d'Orsi et al., 2023) our idea to find a tight set $\mathcal{S}'$, is to use the natural degree-$O(1)$ sum-of-squares relaxation of $\mathcal{S}$.

An advanced application: k-way marginal queries The problem of releasing k-way marginal queries can indeed be recast as the above example (see Appendix A). Given a set $\{e_1, \ldots, e_n\}$ in $\{0,1\}^m$, privately output the tensor $\frac{1}{m}\sum_i e_i^{\otimes k}$. Here, adjacent datasets are usually defined as datasets differing in one vector. The even k case is easy as one can always flatten the tensor into a $n^{k/2}$-by-$n^{k/2}$ matrix. Indeed this is the approach of (Dwork et al., 2015). Hence we discuss the odd case. Ideally, one would like now to project the perturbed input exactly onto the set of rank-m tensors in Equation (3). However, to run the algorithm efficiently, we replace this set with its sum of squares relaxation:

$$\begin{aligned}\{\; X \in (\mathbb{R}^n)^{\otimes k} \mid \exists \text{ deg-}O(k)\ \mu \text{ over } \mathcal{Q} \\ \text{s.t. } \tilde{\mathbb{E}}_\mu[\sum_{i \in [m]} v_i^{\otimes k}] = X \;\}\,,\end{aligned}$$

where $\mathcal{Q}$ is the set of polynomial inequalities over vector variables $v_1, \ldots, v_n$

$$\mathcal{Q} := \left\{ \left\| \sum_i v_i^{\otimes k} \right\|_{\max} \leqslant 1 \right\}.$$

The Gaussian complexity of $\mathcal{S}$ is bounded by $O(\sqrt{k \log n})$, in contrast for this set $\mathcal{S}'$

$$\begin{aligned}\mathcal{G}(\mathcal{S}') &= \mathop{\mathbb{E}}_{\mathbf{W} \sim N(0,\mathrm{Id}_n)} \left[\sup_{X \in \mathcal{S}'} \langle X, \mathbf{W} \rangle\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{W} \sim N(0,\mathrm{Id}_n)} \left[\sup_{\text{deg-O(1) } \mu \text{ over } \mathcal{Q}} \langle \tilde{\mathbb{E}}_\mu \left[\sum_i v_i^{\otimes k}\right], \mathbf{W} \rangle\right].\end{aligned}$$

This last term is the minimum value of any degree-$O(1)$ sum-of-squares certificate of the injective tensor norm of

[4] The careful reader may argue that this set is not convex. However, notice we may obtain comparable guarantees by projecting onto the convex set of k-th order tensors that are expectations of distributions over rank-m tensors.

Gaussian tensors, which is known to be $O(n^{k/2})$ up to multiplicative logarithmic factors (Hopkins et al., 2015; Bhattiprolu et al., 2016). This approach yields the guarantees of Theorem 2.2

Sparse k-way marginal queries The aforementioned ideas can be pushed even further. For instance, real world datasets over which k-way marginal queries are relevant are often sparse. For instance, consider the cause of a store interested in estimating the number of users that have purchased any specific basket of item. It is natural to assume that no user has purchased a large fraction of the whole inventory. Therefore, it is natural to ask whether one could improve over the above guarantees when each vector e_i in the datasets has at most t non-zero entries: $e_i \in \{v \in \mathbb{R}^n \mid \|v\|_0 \leqslant t\}$. Here one would like to project onto the intersection between this set and the one in Equation (3). The Gaussian complexity of this intersection can be verified to be $O(\sqrt{t \log n})$ (Choo & d'Orsi, 2021) but again, projecting onto this set is computationally hard. Our improvement in these settings come from studying the sum-of-squares relaxations of this set (see Appendix A.1). Specifically, we introduce *novel* sum-of-squares certificates for sparse injective tensor norms of Gaussian tensors. Our bounds match existing lower bounds (against the class of algorithms captured by low degree polynomials) up to logarithms (Choo & d'Orsi, 2021).

Practical perturb-and-project via alternating projections The algorithms discussed so far show how effective the perturb-and-project framework can be. However, even though these projections can be computed in polynomial time, at first sight comparable results appear unattainable in practice given that the computational budget required to naively implement these projections is too large for real-world applications. This warrant the question of whether comparable guarantees could be achieved in practice. It turns out that this is the case to a large extent.

A long line of work—see (Bauschke & Borwein, 1993; Sakai, 1995) and references therein—initiated by (Neumann, 1949) showed how the orthogonal projection onto a set that is the intersection of multiple compact convex sets $\mathcal{S} = \bigcap_{\ell \in [q]} \mathcal{S}_\ell$ can be computed by repeatedly projecting onto each of these sets and averaging the resulting projections. Remarkably, the convergence of this iterative algorithm is linear in the number of iterations, implying that whenever $\|A\|_2^2 \leqslant n^{O(1)}$ a logarithmic number of steps suffices. In our settings, this amounts to replacing step 2 in Algorithm 1 as described (Algorithm 2).

At each iteration i , Algorithm 2 projects the current vector $\mathbf{X}_{i-1}$ to *each* of the sets $\mathcal{S}_1, \ldots, \mathcal{S}_q$ and then computes the average. A single iteration of Algorithm 2 can be significantly faster than Algorithm 1 as well as easier to implement.

Algorithm 2 Perturb-and-alternately-project

Input: $\varepsilon, \delta, \Delta, t > 0$, $A \in \mathbb{R}^n$ compact convex sets $\mathcal{S}, \mathcal{S}_1, \ldots, \mathcal{S}_q \subseteq \mathbb{R}^n$ with $\bigcap_{\ell \in [q]} \mathcal{S}_\ell = \mathcal{S}$.

Output: $\hat{\mathbf{\Pi}}$.

Let $\mathbf{X}_0 = A + \mathbf{W}$ for $\mathbf{W} \sim N\left(0, \mathrm{Id}_n \cdot \frac{2\log(2/\delta)}{\varepsilon^2} \cdot \Delta^2\right)$.

for $i = 1$ **to** t **do**

  Compute $\mathbf{X}_i = \mathbb{E}_{\ell \sim [q]} \arg\min_{X \in \mathcal{S}_\ell} \|X - \mathbf{X}_{i-1}\|_2^2$

end for

Return $\hat{\mathbf{\Pi}} := \mathbf{X}_t$.

This is because each of the projections may be fast to compute by itself. Furthermore, since by linear convergence few iterations suffice, the whole algorithm turns out to be faster than Algorithm 1.

As a concrete example, observe that the set of Equation (2), is the intersection of[5] $\{X \in \mathbb{R}^{n \times n} \mid, X \succeq 0\}$, and $\left\{X \in \mathbb{R}^{n \times n} \mid \|X\|_{\max} \leqslant 1\right\}$. Hence we can compute the output of our algorithm repeatedly projecting onto these two sets and averaging. The projection onto the first set can be obtained zeroing out negative eigenvalues. The latter projection amounts to clipping entries that are larger than 1 in absolute values, a typical subroutine of practical differentially private algorithms!

That is, not only the perturb-and-alternately-project framework can yield practical algorithms, it also gives an alternative perspective on typical subroutines used in the context of privacy.

Remark 4.1 (On the error guarantees of Theorem 2.1). The task of computing 2-way marginal queries under differential privacy is related to the task of privately computing pairwise cosine similarities. One can see the vectors over which we compute the cosine similarities as the Gram vectors of the 2-way marginal queries matrix. Indeed, for certain family of vectors, the set over which we perform the projection in the perturb-and-project framework is the same. By optimality of the algorithm for k-way marginal queries (Kasiviswanathan et al., 2010), this suggests that the guarantees of Theorem 2.1 may be optimal among algorithms captured by the perturb-and-project framework.

5. Guarantees of the perturb-and-project paradigm for differential privacy

We present here a general statement about the privacy and utility performance of the *perturb-and-project* framework defined in Algorithm 1.

[5]Technically speaking we would like all sets considered to be explicitly bounded. This can be easily achieved by replacing the positive semidefinite cone with its intersection with a large enough Euclidean ball.

Both the privacy and utility guarantees of Algorithm 1 can be conveniently expressed in terms of ε , δ and the convex set $\mathcal{S}$. In what follows we say $A, A' \in \mathbb{R}^n$ are *adjacent* if $\|A - A'\|_2 \leqslant \Delta$, notice this applies to higher order tensors via flattening. For a convex set $\mathcal{S}$ and $A \in \mathbb{R}^n$ we define $\Pi_{\mathcal{S}}(A) := \arg\min_{X \in \mathcal{S}} \|A - X\|_2^2$.

Theorem 5.1 (Guarantees of perturb-and-project)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a compact convex set and let $\varepsilon, \delta > 0$. Then, on input $A \in \mathbb{R}^n, \varepsilon, \delta$, Algorithm 1 returns $\hat{\mathbf{\Pi}} \in \mathcal{S}$ satisfying*

$$\mathbb{E}\left\|\hat{\mathbf{\Pi}} - \Pi_{\mathcal{S}}(A)\right\|_2^2 \leqslant \frac{8\sqrt{\log(2/\delta)}}{3\varepsilon} \cdot \Delta \cdot \mathcal{G}(\mathcal{S}) .$$

Moreover, the algorithm is (ε, δ)-differentially private.

Theorem 5.1 states that for any given set of privacy parameters (ε, δ) , the error guarantees of Algorithm 1 are governed by the metric entropy of the set considered.[6] Hence more structured sets yields stronger guarantees. This phenomenon introduces a trade-off between the computational complexity required by the projection and the accuracy of the estimate.

It is important to remark that similar results appeared in the risk minimization literature (Chaudhuri & Monteleoni, 2008; Chaudhuri et al., 2011; Jain et al., 2012; Thakurta & Smith, 2013; Duchi et al., 2013; Jain & Thakurta, 2014; Bassily et al., 2014; Ullman, 2015; Chourasia et al., 2021; Nikolov et al., 2013). Specifically, the perturb-and-project framework was studied in (Kifer et al., 2012; Chaudhuri et al., 2011; Dwork et al., 2009), resulting in weaker guarantees than Theorem 5.1. In (Dwork et al., 2015), the special case of Theorem 5.1 for the set of positive semidefinite matrices with bounded nuclear norm was shown. Perhaps, most relevant to our settings is (Talwar et al., 2014). Here the authors studied objective perturbations and bounded in terms of Gaussian complexity of the solution space the change in the objective value for strongly convex functions. Furthermore, (Talwar et al., 2014) also obtained comparable guarantees for the mirror descent algorithm, in which noise is added at each iteration. This difference leads to an important algorithmic consequence: since in Algorithm 1 both the privacy and error guarantees do not depend on the particular steps involved in the computation of $\hat{\mathbf{\Pi}}$, different techniques may be used to effectively carry out the computation of Algorithm 1 (see Corollary 5.3 and the related discussion).

The privacy guarantees follows from the definition of adjacent inputs and the use of the Gaussian mechanism (Lemma B.6). The utility guarantees follows from an argument about the stability of projections onto compact convex sets, as shown below.

[6] Recall, by Sudakov's minoration, we can bound the metric entropy of a set with an exponential function of its squared Gaussian complexity.

Lemma 5.2 (Stability of projections)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a compact convex set and for any $A \in \mathbb{R}$ let*

$$\Pi_{\mathcal{S}}(A) := \arg\min_{X \in \mathcal{S}} \|X - A\|_2^2 .$$

Then

$$\mathop{\mathbb{E}}_{\mathbf{W} \sim N(0, \mathrm{Id}_n)} \|\Pi_{\mathcal{S}}(A + \mathbf{W}) - \Pi_{\mathcal{S}}(A)\|_2^2 \leqslant \frac{4}{3}\mathcal{G}(\mathcal{S}) .$$

Proof. The statement follows by strong convexity of the underlying function. Indeed we have (we drop the subscript $\mathcal{S}$ for simplicity)

$$\begin{aligned}
&\|A - \Pi(A)\|_2^2 \\
&\geqslant \|A - \Pi(A + \mathbf{W})\|_2^2 \\
&+ \langle \nabla \|A - \Pi(A + \mathbf{W})\|_2^2 , \Pi(A) - \Pi(A + \mathbf{W})\rangle \\
&+ \frac{1}{2} \|\Pi(A + \mathbf{W}) - \Pi(A)\|_2^2
\end{aligned}$$

and

$$\begin{aligned}
&\|A + \mathbf{W} - \Pi(A + \mathbf{W})\|_2^2 \\
&\geqslant \|A + \mathbf{W} - \Pi(A)\|_2^2 \\
&+ \langle \nabla \|A + \mathbf{W} - \Pi(A)\|_2^2 , \Pi(A + \mathbf{W}) - \Pi(A)\rangle \\
&+ \frac{1}{2} \|\Pi(A + \mathbf{W}) - \Pi(A)\|_2^2 .
\end{aligned}$$

Putting the two inequalities together and using the definition of $\Pi(A + \mathbf{W}) , \Pi(A)$

$$\begin{aligned}
&\|\Pi(A + \mathbf{W}) - \Pi(A)\|_2^2 \\
&\leqslant \langle \nabla \|A - \Pi(A + \mathbf{W})\|_2^2 , \Pi(A) - \Pi(A + \mathbf{W})\rangle \\
&+ \langle \nabla \|A + \mathbf{W} - \Pi(A)\|_2^2 , \Pi(A + \mathbf{W}) - \Pi(A)\rangle \\
&= -2 \|\Pi(A) - \Pi(A + \mathbf{W})\|_2^2 \\
&+ 2\langle \mathbf{W}, \Pi(A + \mathbf{W}) - \Pi(A)\rangle .
\end{aligned}$$

Rearranging and taking expectations

$$\begin{aligned}
&\|\Pi(A + \mathbf{W}) - \Pi(A)\|_2^2 \\
&\leqslant \frac{2}{3} \mathbb{E}\langle \mathbf{W}, \Pi(A + \mathbf{W}) - \Pi(A)\rangle \\
&\leqslant \frac{4}{3} \mathbb{E}\, \mathcal{G}(\mathcal{S}) .
\end{aligned}$$

□

5.1. Practical perturb-and-project algorithms

As mentioned in the previous sections, a crucial consequence of Theorem 5.1 is that the projection steps of Algorithm 1 can be broken down into a sequence of simpler projections as in Algorithm 2. We study here the guarantees of this algorithm.

Corollary 5.3 (Guarantees of the perturb-and-alternately-project)**.** *Let $\varepsilon, \delta \geqslant 0$. Let*

$$\mathcal{B}_C := \left\{ X \in \mathbb{R}^n \;\middle|\; \|X\|_2^2 \leqslant n^C \right\} \subseteq \mathbb{R}^n$$

and let $\mathcal{S}_1, \ldots, \mathcal{S}_n \subseteq \mathbb{R}^n$ be compact convex sets with intersection $\bigcap_{\ell \in [q]} \mathcal{S}_\ell = \mathcal{S} \subseteq \mathcal{B}_C$ for some $C \geqslant 0$. Then, on input $A \in \mathbb{R}^n, \varepsilon, \delta, t \geqslant O(C \log n), \Delta \leqslant n^{O(1)}$, Algorithm 2 returns $\hat{\mathbf{\Pi}}_t$ satisfying

$$\mathbb{E} \left\| \hat{\mathbf{\Pi}}_t - \Pi_{\mathcal{S}}(A) \right\|_2^2 \leqslant \frac{16\sqrt{\log(2/\delta)}}{3\varepsilon} \cdot \Delta \cdot \left(\mathcal{G}(\mathcal{S}) + 2^{-t} \|A\|_2^2 + n^{-\Omega(1)} \right) .$$

Moreover, the algorithm is (ε, δ)-differentially private.

The proof of Corollary 5.3 rely on a result in the theory of alternating projections (Bauschke & Borwein, 1993) (see (Sakai, 1995) and references therein for variations of Algorithm 2).

Theorem 5.4 (Linear convergence of alternating projections, (Bauschke & Borwein, 1993))**.** *Let $\mathcal{S}_1, \ldots, \mathcal{S}_q \subseteq \mathbb{R}^n$ be compact convex sets with intersection $\mathcal{S} = \bigcap_{i \in [q]} \mathcal{S}_i$. For $x \in \mathbb{R}^n$, let $\{\Pi_i(x)\}_{t \geqslant 1}$ be the sequence with $\Pi_0(x) = x$ and $\Pi_t(x) = \mathbb{E}_{\mathbf{i} \overset{u.a.r}{\sim} [q]} \Pi_{\mathcal{S}_\mathbf{i}}(\Pi_{t-1}(x))$. Then, there exists a constant $c \in (0,1)$, depending only on $\mathcal{S}_1, \ldots, \mathcal{S}_q$, such that*

$$\|\Pi_{\mathcal{S}}(x) - \Pi_t(x)\|_2 \leqslant c^t \|\Pi_{\mathcal{S}}(x) - x\|_2 .$$

Theorem 5.4 states that the output of Algorithm 2 converges to the output of Algorithm 1 linearly in the number of iterations. We are ready to prove Corollary 5.3.

Proof of Corollary 5.3. For a matrix $M \in \mathbb{R}^{n \times n}$, let $\Pi_t(M)$ be the output of step 2 of Algorithm 2 with t iterations and $\Pi(M)$ the true projection of M onto $\mathcal{S}$. On input A, we can bound the expected squared error of Algorithm 2 as follows:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{W}} \|\Pi_t(A + \mathbf{W}) - \Pi(A)\|_2^2 \\ &\quad \leqslant 2 \mathop{\mathbb{E}}_{\mathbf{W}} \|\Pi_t(A + \mathbf{W}) - \Pi(A + \mathbf{W})\|_2^2 \\ &\quad + 2 \mathop{\mathbb{E}}_{\mathbf{W}} \|\Pi(A + \mathbf{W}) - \Pi(A)\|_2^2 \end{aligned}$$

where we used Cauchy-Schwarz. By Theorem 5.4

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{W}} \|\Pi_t(A + \mathbf{W}) - \Pi(A + \mathbf{W})\|_2^2 \\ &\quad \leqslant \mathop{\mathbb{E}}_{\mathbf{W}} c^t \|A + \mathbf{W} - \Pi(A + \mathbf{W})\|_2^2 \\ &\quad \leqslant 2c^t \left(\|A\|_2^2 + \Delta^2 \cdot n^2 + n^C \right) \\ &\quad \leqslant 2c^t \|A\|_2^2 + 1/n^{-10} , \end{aligned}$$

where in the last step we used the fact $t \geqslant (12 + 12C) \log_c n$. Combining this with Theorem 5.1 the result follows as desired. □

6. Cosine similarity release

We introduce here our differentially private algorithm to compute the pair-wise similarity between a set of vectors and prove Theorem 2.1.

Algorithm 3 Private pair-wise cosine similarities

Input: $\varepsilon, \delta, \Delta > 0, V \in \mathbb{R}^{n \times m}$.
Output: $\hat{\mathbf{X}} \in \mathbb{R}^{n \times n}$.
Run Algorithm 1 with input VV^T, parameters $\varepsilon, \delta, \Delta$ and $\mathcal{S} = \{X \in \mathbb{R}^{n \times n} \mid X \succeq 0, X_{ii} \leqslant 1\}$.

Equipped with Algorithm 3 we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Notice that $VV^\mathsf{T} \in \mathbb{R}^{n \times n}$ is a positive semidefinite matrix satisfying $\left\|VV^\mathsf{T}\right\|_{\max} \leqslant 1$. Moreover, by definition of adjacency, for any $V' \in \mathbb{R}^{n \times m}$ adjacent to V we have

$$\Delta^2 \geqslant \left\|VV^\mathsf{T} - V'V'^\mathsf{T}\right\|_\mathrm{F}^2 .$$

Privacy of Algorithm 3 then follows by Theorem 5.1. Concerning utility, notice that by Holder's inequality,

$$\begin{aligned} \mathcal{G}(\mathcal{S}) &= \mathop{\mathbb{E}}_{\mathbf{W}^* \sim N(0,1)^{n \times n}} \sup_{X \in \mathcal{S}} \langle X, \mathbf{W}^* \rangle \\ &\leqslant \|X\|_{\text{nuc}} \mathbb{E} \|\mathbf{W}^*\| \leqslant n \cdot 2\sqrt{n} . \end{aligned}$$

The desired error bound follows again by Theorem 5.1. □

Remark 6.1 (On using Algorithm 2 for pair-wise cosine similarity). Notice that in practice it is hard to implement the projection used in the last step of Algorithm 1. By Corollary 5.3 we may replace this step with an application of Algorithm 2 with $\mathcal{S}_1 := \{X \in \mathbb{R}^{n \times n} \mid \|X\|_\mathrm{F} \leqslant n\}$ and $\mathcal{S}_2 := \left\{X \in \mathbb{R}^{n \times n} \;\middle|\; \|X\|_{\max} \leqslant 1\right\}$. Indeed it always holds that $VV^\mathsf{T} \in \mathcal{S}_1 \cap \mathcal{S}_2$. Now the projection of a matrix $X \in \mathbb{R}^{n \times n}$ onto $\mathcal{S}_1$ can be obtained removing negative eigenvalues and rescaling the positive eigenvalues $\lambda_1, \ldots, \lambda_n \geqslant 0$ so that $\sum_i \lambda_i^2 \leqslant n$. The projection onto $\mathcal{S}_2$ can be obtained by clipping entries to 1 in absolute value. Moreover, by Corollary 5.3 a logarithmic number of iterations suffices. The algorithm then is the following.

Acknowledgments

Tommaso d'Orsi is partially supported by the project MUR FARE2020 PAReCoDi.

Algorithm 4 Practical pair-wise cosine similarities

Input: $\varepsilon, \delta, \Delta, t > 0\,, V \in \mathbb{R}^{n\times m}$.
Output: $\hat{\mathbf{X}} \in \mathbb{R}^{n\times n}$.
Let $X^{(0)} = VV^{\mathsf{T}} + \mathbf{W}$, where $\mathbf{W} \sim N(0, \frac{\delta^2 \cdot \log(2/\delta)}{\varepsilon^2})^{n\times n}$.
for $i = 1$ to t **do**
  Compute $\mathbf{X}^{(i)} = \frac{1}{2}\Pi_{\mathcal{S}_1}(\mathbf{X}^{(i-1)}) + \frac{1}{2}\Pi_{\mathcal{S}_2}(\mathbf{X}^{(i-1)})$.
end for
Return: $\mathbf{X}^{(t)}$.

Impact statement

This paper presents work whose goal is to advance the field of differentially private machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alaggan, M., Gambs, S., and Kermarrec, A.-M. Private similarity computation in distributed systems: from cryptography to differential privacy. In *Principles of Distributed Systems: 15th International Conference, OPODIS 2011, Toulouse, France, December 13-16, 2011. Proceedings 15*, pp. 357–377. Springer, 2011.

Aumüller, M., Bourgeat, A., and Schmurr, J. Differentially private sketches for jaccard similarity estimation. 8 2020. URL http://arxiv.org/abs/2008.08134. Provides local DP algorithms for the problem of estimating the Jaccard Similarity of item-sets. Their algorithms are based on the local computation of an appropriately randomized version (using Generalized Randomized Response) of the popular MinHash LSH. They work in an adjancency notion allowing l items changes for a user.

Bakshi, A., Diakonikolas, I., Hopkins, S. B., Kane, D., Karmalkar, S., and Kothari, P. K. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 149–159. IEEE, 2020.

Bakshi, A., Diakonikolas, I., Jia, H., Kane, D. M., Kothari, P. K., and Vempala, S. S. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1234–1247, 2022.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 273–282, 2007.

Barak, B., Brandao, F. G., Harrow, A. W., Kelner, J., Steurer, D., and Zhou, Y. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 307–326, 2012.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization, revisited. *rem*, 3:19, 2014.

Battaglia, E., Celano, S., and Pensa, R. G. Differentially private distance learning in categorical data. *Data Mining and Knowledge Discovery*, 35:2050–2088, 2021.

Bauschke, H. H. and Borwein, J. M. On the convergence of von neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1:185–212, 1993.

Bhattiprolu, V., Guruswami, V., and Lee, E. Sum-of-squares certificates for maxima of random tensors on the sphere. *arXiv preprint arXiv:1605.00903*, 2016.

Blocki, J., Blum, A., Datta, A., and Sheffet, O. The johnson-lindenstrauss transform itself preserves differential privacy. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 410–419. IEEE, 2012.

Brennan, M. and Bresler, G. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pp. 648–847. PMLR, 2020.

Chandrasekaran, K., Thaler, J., Ullman, J., and Wan, A. Faster private release of marginals on small databases. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 387–402, 2014.

Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Chaudhuri, K., Sarwate, A. D., and Sinha, K. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 14, 2013.

Chen, H., Cohen-Addad, V., d'Orsi, T., Epasto, A., Imola, J., Steurer, D., and Tiegel, S. Private estimation algorithms for stochastic block models and mixture models. *arXiv preprint arXiv:2301.04822*, 2023.

Chen, X., Liu, H., and Yang, D. Improved lsh for privacy-aware and robust recommender system with sparse data in edge environment. *EURASIP Journal on Wireless Communications and Networking*, 2019:1–11, 2019.

Choi, S. G., Dachman-Soled, D., Liang, M., Liu, L., and Yerukhimovich, A. On the privacy of sublinear-communication jaccard index estimation via min-hash sketching. *Cryptology ePrint Archive*, 2023.

Choo, D. and d'Orsi, T. The complexity of sparse tensor pca. *Advances in Neural Information Processing Systems*, 34:7993–8005, 2021.

Chourasia, R., Ye, J., and Shokri, R. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34: 14771–14781, 2021.

Cormode, G., Kulkarni, T., and Srivastava, D. Marginal release under local differential privacy. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 131–146, 2018.

Dasgupta, S. and Gupta, A. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Diakonikolas, I. and Kane, D. M. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

Ding, J., d'Orsi, T., Nasser, R., and Steurer, D. Robust recovery for stochastic block models. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 387–394. IEEE, 2022.

Ding, J., d'Orsi, T., and Hua, Y. Node robust recovery for stochastic block models. 2023.

Ding, X., Yang, W., Choo, K.-K. R., Wang, X., and Jin, H. Privacy preserving similarity joins using mapreduce. *Information Sciences*, 493:20–33, 2019.

Dong, W., Liang, Y., and Yi, K. Differentially private covariance revisited. *Advances in Neural Information Processing Systems*, 35:850–861, 2022.

d'Orsi, T., Kothari, P. K., Novikov, G., and Steurer, D. Sparse pca: algorithms, adversarial perturbations and certificates. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 553–564. IEEE, 2020.

d'Orsi, T., Liu, C.-H., Nasser, R., Novikov, G., Steurer, D., and Tiegel, S. Consistent estimation for pca and sparse regression with oblivious outliers. *Advances in Neural Information Processing Systems*, 34:25427–25438, 2021.

d'Orsi, T., Nasser, R., Novikov, G., and Steurer, D. Higher degree sum-of-squares relaxations robust against oblivious outliers. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 3513–3550. SIAM, 2023.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Dwork, C., Gopalan, P., Lin, H., Pitassi, T., Rothblum, G., Smith, A., and Yekhanin, S. An analysis of the chaudhuri and monteleoni algorithm. *Innovations in Computer Science (poster)*, 2009.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Dwork, C., Nikolov, A., and Talwar, K. Efficient algorithms for privately releasing marginals via convex relaxations. *Discrete & Computational Geometry*, 53:650–673, 2015.

d'Orsi, T., Novikov, G., and Steurer, D. Consistent regression when oblivious outliers overwhelm. In *International Conference on Machine Learning*, pp. 2297–2306. PMLR, 2021.

Fernandes, N., Kawamoto, Y., and Murakami, T. Locality sensitive hashing with extended differential privacy. In *European Symposium on Research in Computer Security*, pp. 563–583. Springer, 2021.

Georgiev, K. and Hopkins, S. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in Neural Information Processing Systems*, 35:6829–6842, 2022.

Grötschel, M., Lovász, L., and Schrijver, A. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1:169–197, 1981.

Gupta, A., Hardt, M., Roth, A., and Ullman, J. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 803–812, 2011.

Hardt, M. and Rothblum, G. N. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st annual symposium on foundations of computer science*, pp. 61–70. IEEE, 2010.

Hardt, M., Ligett, K., and McSherry, F. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25, 2012.

Hopkins, S. B. and Li, J. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1021–1034, 2018.

Hopkins, S. B., Shi, J., and Steurer, D. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pp. 956–1006. PMLR, 2015.

Hopkins, S. B., Kothari, P. K., Potechin, A., Raghavendra, P., Schramm, T., and Steurer, D. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 720–731. IEEE, 2017.

Hopkins, S. B., Kamath, G., and Majid, M. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1406–1417, 2022.

Hopkins, S. B., Kamath, G., Majid, M., and Narayanan, S. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 497–506, 2023.

Imola, J., Epasto, A., Mahdian, M., Cohen-Addad, V., and Mirrokni, V. Differentially private hierarchical clustering with provable approximation guarantees. In *International Conference on Machine Learning*, pp. 14353–14375. PMLR, 2023.

Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., and Wang, L. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 299–316. IEEE, 2019.

Jain, P. and Thakurta, A. G. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pp. 476–484. PMLR, 2014.

Jain, P., Kothari, P., and Thakurta, A. Differentially private online learning. In *Conference on Learning Theory*, pp. 24–1. JMLR Workshop and Conference Proceedings, 2012.

Kasiviswanathan, S. P., Rudelson, M., Smith, A., and Ullman, J. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 775–784, 2010.

Kenthapadi, K., Korolova, A., Mironov, I., and Mishra, N. Privacy via the johnson-lindenstrauss transform. *arXiv preprint arXiv:1204.2606*, 2012.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Klivans, A., Kothari, P. K., and Meka, R. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pp. 1420–1430. PMLR, 2018.

Kothari, P., Manurangsi, P., and Velingker, A. Private robust estimation by stabilizing convex relaxations. In *Conference on Learning Theory*, pp. 723–777. PMLR, 2022.

Kothari, P. K., Steinhardt, J., and Steurer, D. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1035–1046, 2018.

Lasserre, J. B. New positive semidefinite relaxations for nonconvex quadratic programs. *Advances in Convex Analysis and Global Optimization: Honoring the Memory of C. Caratheodory (1873–1950)*, pp. 319–331, 2001.

Li, N., Lyu, M., Su, D., and Yang, W. Publishing marginals. In *Differential Privacy: From Theory to Practice*, pp. 79–91. Springer, 2017.

Liu, A. and Moitra, A. Minimax rates for robust community detection. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 823–831. IEEE, 2022.

Mangoubi, O. and Vishnoi, N. K. Private covariance approximation and eigenvalue-gap bounds for complex gaussian perturbations. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1522–1587. PMLR, 2023.

Nesterov, Y. Squared functional systems and optimization problems. *High performance optimization*, pp. 405–440, 2000.

Neumann, J. V. On rings of operators. reduction theory. *Annals of Mathematics*, 50(2):401–485, 1949.

Nikolov, A. Private query release via the johnson-lindenstrauss transform. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4982–5002. SIAM, 2023.

Nikolov, A., Talwar, K., and Zhang, L. The geometry of differential privacy: the sparse and approximate cases. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 351–360, 2013.

Parrilo, P. A. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. California Institute of Technology, 2000.

Sakai, M. Strong convergence of infinite products of orthogonal projections in hilbert space. *Applicable Analysis*, 59 (1-4):109–120, 1995.

Schoppmann, P., Vogelsang, L., Gascón, A., and Balle, B. Secure and scalable document similarity on distributed databases: Differential privacy to the rescue. *Cryptology ePrint Archive*, 2018.

Shor, N. Z. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25:1–11, 1987.

Stausholm, N. M. Improved differentially private euclidean distance approximation. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 42–56, 2021.

Talwar, K., Thakurta, A., and Zhang, L. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

Thakurta, A. G. and Smith, A. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pp. 819–850. PMLR, 2013.

Tsakonas, E., Jaldén, J., Sidiropoulos, N. D., and Ottersten, B. Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214, 2014.

Ullman, J. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 303–312, 2015.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, Y., Lu, S., and Zhang, L. Searching privately by imperceptible lying: A novel private hashing method with differential privacy. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2700–2709, 2020.

Xu, C., Ren, J., Zhang, Y., Qin, Z., and Ren, K. Dppro: Differentially private high-dimensional data release via random projection. *IEEE Transactions on Information Forensics and Security*, 12(12):3081–3093, 2017.

Xue, Z. and Yuqing, S. Differential privacy for collaborative filtering recommender algorithm [c]. In *Proceedings of the 2016 ACM on International Workshop on Security And Privacy Analytics*, 2016.

Yang, M., Zhu, T., Ma, L., Xiang, Y., and Zhou, W. Privacy preserving collaborative filtering via the johnson-lindenstrauss transform. In *2017 IEEE Trustcom/BigDataSE/ICESS*, pp. 417–424. IEEE, 2017.

Yang, M., Zhu, T., Liu, B., Xiang, Y., and Zhou, W. Differential private poi queries via johnson-lindenstrauss transform. *IEEE Access*, 6:29685–29699, 2018.

Zhang, Y., Gao, N., Chen, J., Tu, C., and Wang, J. Privrec: user-centric differentially private collaborative filtering using lsh and kd. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pp. 113–121. Springer, 2020.

Zhu, Y., Yu, X., Chandraker, M., and Wang, Y.-X. Private-knn: Practical differential privacy for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11854–11862, 2020.

A. K-way-marginal queries and sparse tensor certificates

In this section we combine Theorem 5.1 with novel sum-of-squares certificates for injective tensor norms to obtain tight results for k-way marginal queries. Before formally stating the theorems we introduce some definitions.

Let $[n]$ denote a set of n features. A user can be represented by a binary vector $e \in \{0,1\}^n$. There are at most 2^n distinct potential users with features in $[n]$. We denote the set of all possible 2^n distinct potential users by $\mathcal{E}$: and the multiplicity of $e \in \mathcal{E}$ in a dataset D by $m_D(e)$. Hence, a dataset D can also be described by the 2^n dimensional vector $m(D)$ with entries $m(D)_e = m_D(e)$ for each $e \in \mathcal{E}$. When the context is clear we drop the subscript. We denote the set of all possible datasets over m users as $\mathcal{D}_{m,n}$, and by $\mathcal{M}_m$ the set of all possible vectors $m(D)$ with $D \in \mathcal{D}_{m,n}$. Notice that $|\mathcal{D}_{m,n}| = |\mathcal{M}_n|$. We are interested in differential privacy with respect to the following natural notion of adjacent datasets.

Definition A.1 (Adjacent datasets)**.** Two datasets $D, D' \in \mathcal{D}_{m,n}$ are said to be *adjacent* if

$$\|m(D) - m(D')\|_1 \leqslant 2\,. \tag{4}$$

In other words, D' can be obtained from D replacing one user in the dataset.

We will formally study so-called *k-wise parity* queries. Indeed, as shown in (Barak et al., 2007; Dwork et al., 2015), k-way parities form an orthonormal base of all k-way marginals. In other words, it suffices to reconstruct k-wise parity queries to answer all k-way marginal queries.

Definition A.2 (k-wise parity query)**.** For $D \in \mathcal{D}_{m,n}$, a k-wise parity query is specified by some $\alpha \subseteq [n]$ and given by

$$\sum_{e \in \mathcal{E}} m_D(e) \cdot \prod_{i \in \alpha} e_i\,.$$

In particular, Definition A.2 implies that the tensor

$$\sum_{e \in \mathcal{E}} m_D(e) e^{\otimes k}$$

contains *all* k'-wise parity queries for any $k' \leqslant k$. Hence, reconstructing this tensor suffices to answer all k-wise parity queries. For sparse datasets, we can capture the sparsity through the following definition.

Definition A.3 (t-sparse dataset)**.** Let $\mathcal{E}_t := \{e \in \mathcal{E} \mid \|e\|_0 \leqslant t\} \subseteq \mathcal{E}$. That is, the set of t-sparse users. A dataset $D \in \mathcal{D}_{m,n}$ is said to be t-sparse if $m_D(e) = 0$ for all $e \in \mathcal{E} \setminus \mathcal{E}_t$. We write $\mathcal{D}_{m,n,t} \subseteq \mathcal{D}_{m,n}$ for the subset of t-sparse datasets.

We are ready to state the main theorem of the section, which implies Theorem 2.2 and Theorem 2.3.

Theorem A.4. *Let $D \in \mathcal{D}_{m,n,t}$. There exists an (ε, δ)-differentially private algorithm that, on input D , k , t , returns a tensor $\mathbf{T} \in (\mathbb{R}^n)^{\otimes k}$ satisfying*

$$\mathbb{E}\left\|\mathbf{T} - \sum_{e \in \mathcal{E}} m_D(e) e^{\otimes k}\right\|_2^2 \leqslant O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon} \cdot |D|\right) \cdot \tau\,,$$

where if k is even $\tau := \min\left\{n^{5k/4}, t^{3k/2}\sqrt{k \log n / t^2}\right\}$, if k is odd $\tau := \min\left\{\left(n^{5k} k \log n\right)^{\frac{1}{4}}, \left(t^{3k} k \log n\right)^{\frac{1}{2}}\right\}$.

Moreover, the algorithm runs in time $m \cdot n^{O(k)}$.

Theorem A.4 immediately implies Theorem 2.2, Theorem 2.3 by taking the entry-wise average. The proof of Theorem A.4 requires novel sum-of-squares certificates for sparse tensor norms. We first focus on those.

A.1. Sparse tensor norms certificates via sum-of-squares

We make use of the following system of polynomial inequalities in n dimensional vector variables x, s, y:

$$\mathcal{P}_t := \left\{ \begin{array}{ll} \text{unit norm:} & \|x\|^2 \leqslant 1 \\ \text{indicators:} & s_i^2 = s_i \\ & x_i \cdot s_i = x_i \\ \text{sparsity:} & \sum_i s_i \leqslant t \\ \text{absolute value:} & x_i \leqslant y_i \\ & -x_i \leqslant y_i \\ & x_i^2 = y_i^2 \end{array} \right\} \tag{$\mathcal{P}_t$}$$

$\mathcal{P}_t$ is the natural sum-of-squares relaxation for the set of t-sparse vectors over the unit ball, with the addition of the vector variable y meant to capture the absolute values of entries in x . Indeed, it is easy to see that for any p-sparse unit vector z in $\mathbb{R}^n$ there exists a corresponding solution (x, s, y) to $\mathcal{P}_t$ with $x = z$. We are ready to state the main theorem of the section.

Theorem A.5. *Let $\mathbf{W} \in (\mathbb{R}^n)^{\otimes k}$ be a tensor with i.i.d entries sampled from $N(0,1)$. Let $\Omega_{k,n}$ be the set of degree-k pseudo-distributions μ satisfying $\mathcal{P}_t$, let*

$$\mathcal{S}_{4k-4,n} := \left\{ X \in (\mathbb{R}^n)^{\otimes k} \;\middle|\; \exists \mu \in \Omega_{4k-4,n} \text{ with } \tilde{\mathbb{E}}_\mu \left[x^{\otimes k} \right] = X \right\} .$$

Then, it holds

- *if k is even $\mathcal{G}(\mathcal{S}_{4k-4,n}) \leqslant O\left(t^{k/2} \cdot \sqrt{k \log n / t^2}\right)$,*
- *if k is odd $\mathcal{G}(\mathcal{S}_{4k-4,n}) \leqslant O\left(t^{k/2} \cdot \sqrt{k \log n}\right)$.*

As already discussed, the even case follows directly from known sum-of-squares bounds for sparse vector norms. The odd case relies crucially on the next lemma.

Lemma A.6. *Let $A \in (\mathbb{R}^n)^{\otimes k}$. Then for any degree $4k-4$ pseudo-distribution satisfying $\mathcal{P}_t$*

$$\tilde{\mathbb{E}} \langle A, x^{\otimes k} \rangle \leqslant \|A\|_\infty \cdot t^{\frac{k}{2}} .$$

Proof. For a n-dimensional vector v and a multi-set $\alpha \in [n]$ we write $v^\alpha = \prod_{i \in \alpha} v^\alpha$. We have

$$\begin{aligned} \tilde{\mathbb{E}} \langle A, x^{\otimes k} \rangle &= \tilde{\mathbb{E}} \sum_{\alpha \in [n]^k} A_\alpha x^\alpha \cdot s^\alpha \\ &\leqslant \tilde{\mathbb{E}} \sum_{\alpha \in [n]^k} \|A\|_\infty y^\alpha \cdot s^\alpha \\ &\leqslant \|A\|_\infty \cdot \left(\tilde{\mathbb{E}} \sum_{\alpha \in [n]^k} y^{2\alpha} \right)^{1/2} \cdot \left(\tilde{\mathbb{E}} \sum_{\alpha \in [n]^k} s^{2\alpha} \right)^{1/2} \\ &= \|A\|_\infty \cdot \left(\tilde{\mathbb{E}} \sum_{\alpha \in [n]^k} x^{2\alpha} \right)^{1/2} \cdot \left(\tilde{\mathbb{E}} \sum_{\alpha \in [n]^k} s^{2\alpha} \right)^{1/2} , \end{aligned}$$

using $\mathcal{P}_t$ and Cauchy-Schwarz. Then, since $\mathcal{P}_t \restriction_{\frac{x}{2q}} \{\|x\|_2^{2q} \leqslant 1\}$ and $\mathcal{P}_t \restriction_{\frac{s}{2q}} \{\|s\|_2^{2q} \leqslant t^q\}$ we may conclude

$$\tilde{\mathbb{E}}\langle A, x^{\otimes k}\rangle = \|A\|_\infty \cdot \left(\tilde{\mathbb{E}}\left[\|x\|_2^{2k}\right] \cdot \tilde{\mathbb{E}}\left[\|s\|_2^{2k}\right]\right)^{1/2} \leqslant \|A\|_\infty \cdot t^{k/2}$$

as desired. □

We can now obtain Theorem A.5.

Proof of Theorem A.5. First notice that by standard concentration arguments, $\mathbb{E}\|\mathbf{W}\|_\infty \leqslant O\left(\sqrt{k\log n}\right)$. Now, consider the settings with even k . Using the natural matrix flattening of $\tilde{\mathbb{E}}x^{\otimes k}$ and $\mathbf{W}$, the statement follows by Fact B.15. For odd k the argument is an immediate corollary of Lemma A.6. □

A.2. Putting things together

We are ready to prove Theorem A.4.

Proof of Theorem A.4. By closure under post-processing of differential privacy, we need to add noise once and may then perform the two projections without any additional privacy loss. So consider first the dense case and let $T^* := \frac{1}{|D|}\sum_{e\in\mathcal{E}} m_D(e)(e/\sqrt{n})^{\otimes k}$. This normalization is chosen so that the ℓ_2^2-sensitivity is $4/|D|^2$ and $\|T^*\|_2^2 = 1$. Let $\Omega_{k,n}$ be the set of degree-k pseudo-distribution over $\left\{\|x\|^2 \leqslant 1\right\}$. We use Algorithm 1 with Gaussian noise having standard deviation $\sigma = \frac{\sqrt{4\log(2/\delta)}}{\varepsilon\cdot|D|}$ and projection set

$$\mathcal{S}'_{2k,n} := \left\{X \in (\mathbb{R}^n)^{\otimes k} \;\middle|\; \exists \mu \in \Omega_{2k,n} \text{ s.t. } \tilde{\mathbb{E}}_\mu\left[x^{\otimes k}\right] = X\right\}$$

with T^* as input. Let $\mathbf{T}$ be the algorithm's output. By Fact B.16, $\mathcal{G}(\mathcal{S}'_{2k,n})$ is bounded by $n^{k/4}$ if k is even and by $\left(n^k \cdot k \cdot \log n\right)^{1/4}$ if k is odd. Hence Theorem 5.1 yields for k even

$$\mathbb{E}\|\mathbf{T} - T^*\|_2^2 \leqslant O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon\cdot|D|}\cdot n^{k/4}\right),$$

for k odd

$$\mathbb{E}\|\mathbf{T} - T^*\|_2^2 \leqslant O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon\cdot|D|}\cdot \left(n^k \cdot k \cdot \log n\right)^{1/4}\right).$$

Rescaling, the dense part of the Theorem follows.

For the sparse settings, let $T^* := \frac{1}{|D|}\sum_{e\in\mathcal{E}} m_D(e)(e/\sqrt{t})^{\otimes k}$. Again this ensures the ℓ_2^2 sensitivity is $4/|D|^2$ and $\|T^*\|_2^2 = 1$ Let $\mathcal{S}_{4k-4,n}$ be the set defined in Theorem A.5 (i.e. not related to the set defined above in the proof) . Notice T^* yields a valid distribution over solutions to Equation ($\mathcal{P}_t$) and hence $T^* \in \mathcal{S}_{4k-4,n}$. We use Algorithm 1 with projection set $\mathcal{S}_{4k-4,n}$, standard deviation $\frac{\sqrt{4\log(2/\delta)}}{\varepsilon\cdot|D|}$ and the rescaled tensor T^* as input. Combining Theorem 5.1 and Theorem A.5 we get for k even

$$\|\mathbf{T} - T^*\|_2^2 \leqslant O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon\cdot|D|}\cdot t^{k/2}\cdot\sqrt{k\log\frac{n}{t^2}}\right),$$

for k odd

$$\|\mathbf{T} - T^*\|_2^2 \leqslant O\left(\frac{\sqrt{\log(2/\delta)}}{\varepsilon\cdot|D|}\cdot t^{k/2}\cdot\sqrt{k\log n}\right).$$

Rescaling the result follows. □

B. Background

We introduce here background notions required in other sections of the paper.

The Gaussian complexity of a set is defined as follows:

Definition B.1 (Gaussian complexity)**.** Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a set, the Gaussian complexity of $\mathcal{S}$ is defined as

$$\mathcal{G}(\mathcal{S}) := \mathop{\mathbb{E}}_{\mathbf{W}\sim N(0,\mathrm{Id}_n)} \max_{v\in\mathcal{S}} \langle \mathbf{W}, v\rangle \,.$$

B.1. Differential privacy

In this section we introduce standard notions of differential privacy (Dwork et al., 2006). Unless specified otherwise, we use the notion of adjacency defined below.

Definition B.2 (Adjacent vectors)**.** Two vectors $u, v \in \mathbb{R}^n$ are said to be *adjacent* if

$$\|v - u\|_2 \leqslant 1 \,.$$

Definition B.3 (Differential privacy)**.** An algorithm $\mathcal{M} : \mathcal{V} \to \mathcal{O}$ is said to be (ε, δ)-differentially private for $\varepsilon, \delta > 0$ if and only if, for every $S \subseteq \mathcal{O}$ and every neighboring $u, v \in \mathcal{V}$ we have

$$\mathbb{P}\left[\mathcal{M}(u) \in S\right] \leqslant e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(v) \in S\right] + \delta \,.$$

Differential privacy is closed under post-processing and composition.

Lemma B.4 (Post-processing)**.** *If $\mathcal{M} : \mathcal{V} \to \mathcal{O}$ is an (ε, δ)-differentially private algorithm and $\mathcal{M}' : \mathcal{O} \to \mathcal{O}'$ is any randomized function. Then the algorithm $\mathcal{M}'(\mathcal{M}(v))$ is (ε, δ)-differentially private.*

B.1.1. Basic differential privacy mechanisms

The Gaussian mechanism is among the most widely used mechanisms in differential privacy. It works by adding a noise drawn from the Gaussian distribution to the output of the function one wants to privatize. The magnitude of the noise depends on the sensitivity of the function.

Definition B.5 (Sensitivity of function)**.** Let $f : \mathcal{Y} \to \mathbb{R}^n$ be a function, its ℓ_2-sensitivity is

$$\Delta_{f,2} := \max_{\substack{Y, Y' \in \mathcal{Y} \\ Y, Y' \text{ are adjacent}}} \|f(Y) - f(Y')\|_2 \,.$$

The Gaussian mechanism provides privacy proportional to the ℓ_2-sensitivity of the function.

Lemma B.6 (Gaussian mechanism)**.** *Let $f : \mathcal{Y} \to \mathbb{R}^n$ be any function with ℓ_2-sensitivity at most $\Delta_{f,2}$. Let $0 < \varepsilon, \delta \leqslant 1$. Then the algorithm that adds $N\left(0, \frac{\Delta_{f,2}^2 \cdot 2\log(2/\delta)}{\varepsilon^2} \cdot \mathrm{Id}_n\right)$ to f is (ε, δ)-DP.*

B.2. Sum-of-squares and pseudo-distributions

We present here necessary background notion about the sum-of-squares framework. We borrow the description and all statements in this section from (d'Orsi et al., 2020; 2023).

Let $w = (w_1, w_2, \ldots, w_n)$ be a tuple of n indeterminates and let $\mathbb{R}[w]$ be the set of polynomials with real coefficients and indeterminates $w, \ldots, w_n$. We say that a polynomial $p \in \mathbb{R}[w]$ is a *sum-of-squares (sos)* if there are polynomials $q_1, \ldots, q_r$ such that $p = q_1^2 + \cdots + q_r^2$.

B.2.1. Pseudo-distributions

Pseudo-distributions are generalizations of probability distributions. We can represent a discrete (i.e., finitely supported) probability distribution over $\mathbb{R}^n$ by its probability mass function $\mu\colon \mathbb{R}^n \to \mathbb{R}$ such that $\mu \geqslant 0$ and $\sum_{w\in\mathrm{supp}(\mu)} \mu(w) = 1$. Similarly, we can describe a pseudo-distribution by its mass function. Here, we relax the constraint $\mu \geqslant 0$ and only require that μ passes certain low-degree non-negativity tests.

Concretely, a *level-ℓ pseudo-distribution* is a finitely-supported function $\mu : \mathbb{R}^n \to \mathbb{R}$ such that $\sum_w \mu(w) = 1$ and $\sum_w \mu(w) f(w)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$. (Here, the summations are over the support of μ.) A straightforward polynomial-interpolation argument shows that every level-∞-pseudo distribution satisfies $\mu \geqslant 0$ and is thus an actual probability distribution. We define the *pseudo-expectation* of a function f on $\mathbb{R}^n$ with respect to a pseudo-distribution D, denoted $\tilde{\mathbb{E}}_{\mu(w)} f(w)$, as

$$\tilde{\mathbb{E}}_{\mu(w)} f(w) = \sum_w \mu(w) f(w) \ . \tag{5}$$

The degree-ℓ moment tensor of a pseudo-distribution D is the tensor $\mathbb{E}_{\mu(w)}(1, w_1, w_2, \ldots, w_n)^{\otimes \ell}$. In particular, the moment tensor has an entry corresponding to the pseudo-expectation of all monomials of degree at most ℓ in w. The set of all degree-ℓ moment tensors of probability distribution is a convex set. Similarly, the set of all degree-ℓ moment tensors of degree d pseudo-distributions is also convex. Key to the algorithmic utility of pseudo-distributions is the fact that while there can be no efficient separation oracle for the convex set of all degree-ℓ moment tensors of an actual probability distribution, there's a separation oracle running in time $n^{O(\ell)}$ for the convex set of the degree-ℓ moment tensors of all level-ℓ pseudodistributions.

Fact B.7 ((Shor, 1987; Parrilo, 2000; Nesterov, 2000; Lasserre, 2001)). For any $n, \ell \in \mathbb{N}$, the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of (Grötschel et al., 1981)):

$$\left\{ \tilde{\mathbb{E}}_{\mu(w)}(1, w_1, w_2, \ldots, w_n)^{\otimes \ell} \mid \text{ degree-}\ell \text{ pseudo-distribution } \mu \text{ over } \mathbb{R}^n \right\} \ . \tag{6}$$

This fact, together with the equivalence of weak separation and optimization (Grötschel et al., 1981) allows us to efficiently optimize over pseudo-distributions (approximately)—this algorithm is referred to as the sum-of-squares algorithm.

The *level-ℓ sum-of-squares algorithm* optimizes over the space of all level-ℓ pseudo-distributions that satisfy a given set of polynomial constraints—we formally define this next.

Definition B.8 (Constrained pseudo-distributions)**.** Let μ be a level-ℓ pseudo-distribution over $\mathbb{R}^n$. Let $\mathcal{A} = \{f_1 \geqslant 0, f_2 \geqslant 0, \ldots, f_m \geqslant 0\}$ be a system of m polynomial inequality constraints. We say that μ *satisfies the system of constraints $\mathcal{A}$ at degree r*, denoted $\mu \models_r \mathcal{A}$, if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\} \leqslant \ell$,

$$\tilde{\mathbb{E}}_\mu h \cdot \prod_{i \in S} f_i \geqslant 0 \, .$$

We write $\mu \models \mathcal{A}$ (without specifying the degree) if $\mu \models_0 \mathcal{A}$ holds. Furthermore, we say that $\mu \models_r \mathcal{A}$ holds *approximately* if the above inequalities are satisfied up to an error of $2^{-n^\ell} \cdot \|h\| \cdot \prod_{i \in S} \|f_i\|$, where $\|\cdot\|$ denotes the Euclidean norm[7] of the coefficients of a polynomial in the monomial basis.

We remark that if μ is an actual (discrete) probability distribution, then we have $\mu \models \mathcal{A}$ if and only if μ is supported on solutions to the constraints $\mathcal{A}$.

We say that a system $\mathcal{A}$ of polynomial constraints is *explicitly bounded* if it contains a constraint of the form $\{\|w\|^2 \leqslant M\}$. The following fact is a consequence of Fact B.7 and (Grötschel et al., 1981),

Fact B.9 (Efficient Optimization over Pseudo-distributions). There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[8] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.

B.2.2. Sum-of-squares proof

Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in w. A *sum-of-squares proof* that the constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ imply the constraint $\{g \geqslant 0\}$ consists of sum-of-squares polynomials $(p_S)_{S \subseteq [m]}$ such that

$$g = \sum_{S \subseteq [m]} p_S \cdot \Pi_{i \in S} f_i \, . \tag{7}$$

[7] The choice of norm is not important here because the factor 2^{-n^ℓ} swamps the effects of choosing another norm.

[8] Here, we assume that the bit complexity of the constraints in $\mathcal{A}$ is $(n+m)^{O(1)}$.

We say that this proof has *degree* ℓ if for every set $S \subseteq [m]$, the polynomial $p_S \Pi_{i \in S} f_i$ has degree at most ℓ. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \vdash_{\ell} \{g \geqslant 0\}. \tag{8}$$

Sum-of-squares proofs satisfy the following inference rules. For all polynomials $f, g \colon \mathbb{R}^n \to \mathbb{R}$ and for all functions $F \colon \mathbb{R}^n \to \mathbb{R}^m$, $G \colon \mathbb{R}^n \to \mathbb{R}^k$, $H \colon \mathbb{R}^p \to \mathbb{R}^n$ such that each of the coordinates of the outputs are polynomials of the inputs, we have:

$$\frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \vdash_{\ell} \{f + g \geqslant 0\}}, \frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0\}, \mathcal{A} \vdash_{\ell'} \{g \geqslant 0\}}{\mathcal{A} \vdash_{\ell+\ell'} \{f \cdot g \geqslant 0\}} \tag{addition and multiplication}$$

$$\frac{\mathcal{A} \vdash_{\ell} \mathcal{B}, \mathcal{B} \vdash_{\ell'} C}{\mathcal{A} \vdash_{\ell \cdot \ell'} C} \tag{transitivity}$$

$$\frac{\{F \geqslant 0\} \vdash_{\ell} \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \vdash_{\ell \cdot \deg(H)} \{G(H) \geqslant 0\}}. \tag{substitution}$$

Low-degree sum-of-squares proofs are sound and complete if we take low-level pseudo-distributions as models.

Concretely, sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact B.10 (Soundness). If $\mu \models_{r} \mathcal{A}$ for a level-ℓ pseudo-distribution μ and there exists a sum-of-squares proof $\mathcal{A} \vdash_{r'} \mathcal{B}$, then $\mu \models_{r \cdot r' + r'} \mathcal{B}$.

If the pseudo-distribution μ satisfies $\mathcal{A}$ only approximately, soundness continues to hold if we require an upper bound on the bit-complexity of the sum-of-squares $\mathcal{A} \vdash_{r'} B$ (number of bits required to write down the proof).

In our applications, the bit complexity of all sum of squares proofs will be $n^{O(\ell)}$ (assuming that all numbers in the input have bit complexity $n^{O(1)}$). This bound suffices in order to argue about pseudo-distributions that satisfy polynomial constraints approximately.

The following fact shows that every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact B.11 (Completeness). Suppose $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^{n} w_i^2 \leqslant B\}$ for some finite B.

Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $\mu \models_{r} \mathcal{A}$ also satisfies $\mu \models_{r'} \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \vdash_{d} \{g \geqslant -\varepsilon\}$.

B.2.3. Sum-of-squares toolkit

Here we present some well-known sum-of-squares bound that we will use throughout the paper.

Fact B.12 (Cauchy-Schwarz for pseudo-distributions (Barak et al., 2012)). Let f, g be vector polynomials of degree at most ℓ in indeterminate $x \in \mathbb{R}^n$. Then, for any degree 2ℓ pseudo-distribution μ,

$$\tilde{\mathbb{E}}_{\mu}\left[\langle f, g \rangle\right] \leqslant \sqrt{\tilde{\mathbb{E}}_{\mu}[\|f\|^2]} \cdot \sqrt{\tilde{\mathbb{E}}_{\mu}[\|g\|^2]}.$$

We will also repeatedly use the following SoS version of Cauchy-Schwarz inequality and its generalization, Hölder's inequality:

Fact B.13 (Sum-of-Squares Cauchy-Schwarz). Let $x, y \in \mathbb{R}^d$ be indeterminites. Then,

$$\vdash_{4}^{x,y} \left\{ \left(\sum_i x_i y_i \right)^2 \leqslant \left(\sum_i x_i^2 \right) \left(\sum_i y_i^2 \right) \right\}$$

We will use the following fact that shows how spectral certificates are captured within the SoS proof system.

Fact B.14 (Spectral Certificates). For any $m \times m$ matrix A,

$$\vdash^{u}_{2} \left\{ \langle u, Au \rangle \leqslant \|A\| \, \|u\|_2^2 \right\} .$$

For sparse vectors, the statement below is often tighter.

Fact B.15 (Spare vector certificates (d'Orsi et al., 2020)). Let $\boldsymbol{W} \sim N(0,1)^{\times n}$ and let

$$\mathcal{A} := \left\{ X \in \mathbb{R}^{n \times n} \;\middle|\; X \succeq 0 \,, \operatorname{Tr} X = 1 \,, \|X\|_1 \leqslant t \right\} .$$

Then with high probability over $\mathbf{W}$

$$\max_{X \in \mathcal{X}} \langle X, \mathbf{W} \rangle \leqslant O\left(t \sqrt{\log n / t^2} \right) .$$

The next statement captures known certificates for the injective norms of random tensors.

Fact B.16 ((Hopkins et al., 2015; d'Orsi et al., 2023)). Let $p \geqslant 3$ be an odd number, and let $\boldsymbol{W} \in (\mathbb{R}^n)^{\otimes p}$ be a tensor with i.i.d. entries from $N(0,1)$. Then with probability $1 - \delta$ (over $\boldsymbol{W}$) every pseudo-distribution μ of degree at least $2p - 2$ on indeterminates $x = (x_1, \ldots, x_n)$ satisfies

$$\tilde{\mathbb{E}} \langle x^{\otimes p}, \boldsymbol{W} \rangle \leqslant C \cdot \left((n^p \cdot p \cdot \ln n)^{1/4} + n^{p/4} \left(\ln(1/\delta)\right)^{1/4} + n^{1/4} \left(\ln(1/\delta)\right)^{3/4} \right) \cdot \left(\tilde{\mathbb{E}} \|x\|^{2p-2} \right)^{\frac{p}{2p-2}}$$

for some absolute constant C.