Covariance's Loss is Privacy's Gain: Computationally Efficient, Private and Accurate Synthetic Data

March Boedihardjo[1], Thomas Strohmer[2], and Roman Vershynin[3]

[1]Department of Mathematics, ETH Zurich
[2]Center of Data Science and Artificial Intelligence Research and Department of Mathematics, University of California Davis
[3]Department of Mathematics, University of California Irvine

August 2, 2022

Abstract

The protection of private information is of vital importance in data-driven research, business, and government. The conflict between privacy and utility has triggered intensive research in the computer science and statistics communities, who have developed a variety of methods for privacy-preserving data release. Among the main concepts that have emerged are anonymity and differential privacy. Today, another solution is gaining traction, synthetic data. However, the road to privacy is paved with NP-hard problems. In this paper we focus on the NP-hard challenge to develop a synthetic data generation method that is computationally efficient, comes with provable privacy guarantees, and rigorously quantifies data utility. We solve a relaxed version of this problem by studying a fundamental, but a first glance completely unrelated, problem in probability concerning the concept of covariance loss. Namely, we find a nearly optimal and constructive answer to the question how much information is lost when we take conditional expectation. Surprisingly, this excursion into theoretical probability produces mathematical techniques that allow us to derive constructive, approximately optimal solutions to difficult applied problems concerning microaggregation, privacy, and synthetic data.

1 Introduction

"Sharing is caring", we are taught. But if we care about privacy, then we better think twice what we share. As governments and companies are increasingly collecting vast amounts of personal information (often without the consent or knowledge of the user [37]), it is crucial to ensure that fundamental rights to privacy of the subjects the data refer to are guaranteed[1]. We are facing the problem of how to release data that are useful to make accurate decisions and predictions without disclosing sensitive information on specific identifiable individuals.

The conflict between privacy and utility has triggered intensive research in the computer science and statistics communities, who have developed a variety of methods for privacy-preserving data release. Among the main concepts that have emerged are *anonymity* and *differential privacy* [6]. Today, another solution is gaining traction, *synthetic data* [4]. However, the road to privacy is paved

[1]The importance of individual privacy is underscored by the fact that Article 12 of the Universal Declaration of Human Rights is concerned with privacy.

with NP-hard problems. For example, finding the optimal partition into k-anonymous groups is NP-hard [24]. Optimal multivariate microaggregation is NP-hard [26, 33] (albeit, the error metric used in these papers is different from the one used in our paper). Moreover, assuming the existence of one-way functions, there is no polynomial time, differentially private algorithm for generating boolean synthetic data that preserves all two-dimensional marginals with accuracy $o(1)$ [35].

No matter which privacy preserving strategy one pursues, in order to implement that strategy the challenge is to navigate this NP-hard privacy jungle and develop a method that is computationally efficient, comes with provable privacy guarantees, and rigorously quantifies data utility. This is the main topic of our paper.

1.1 State of the art

Anonymity captures the understanding that it should not be possible to re-identify any individual in the published data [6]. One of the most popular ways in trying to ensure anonymity is via the concept of *k-anonymity* [32, 31]. A dataset has the k-anonymity property if the information for each person contained in the dataset cannot be distinguished from at least $k-1$ individuals whose information also appears in the dataset. Although the privacy guarantees offered by k-anonymity are limited, its simplicity has made it a popular part of the arsenal of privacy enhancing technologies, see e.g. [6, 10, 21, 16]. k-anonymity is often implemented via the concept of microaggregation [7, 19, 30, 17, 6]. The principle of microaggregation is to partition a data set into groups of at least k similar records and to replace the records in each group by a prototypical record (e.g. the centroid).

Finding the optimal partition into k-anonymous groups is an NP-hard problem [24]. Several practical algorithms exists that produce acceptable empirical results, albeit without any theoretical bounds on the information loss [7, 6, 25]. In light of the popularity of k-anonymity, it is thus quite surprising that it is an open problem to design a computationally efficient algorithm for k-anonymity that comes with theoretical utility guarantees.

As k-anonymity is prone to various attacks, *differential privacy* is generally considered a more robust type of privacy.

Differential privacy formalizes the intuition that the presence or absence of any single individual record in the database or data set should be unnoticeable when looking at the responses returned for the queries [9]. Differential privacy is a popular and robust method that comes with a rigorous mathematical framework and provable guarantees. It can protect aggregate information, but not sensitive information in general. Differential privacy is usually implemented via noise injection, where the noise level depends on the query sensitivity. However, the added noise will negatively affect utility of the released data.

As pointed out in [6], microaggregation is a useful primitive to find bridges between privacy models. It is a natural idea to combine microaggregation with differential privacy [30, 29] to address some of the privacy limitations of k-anonymity. As before, the fundamental question is whether there are computationally efficient methods to implement this scheme while also maintaining utility guarantees.

Synthetic data are generated (typically via some randomized algorithm) from existing data such that they maintain the statistical properties of the original data set, but do so without risk of exposing sensitive information. Combining synthetic data with differential privacy is a promising means to overcome key weaknesses of the latter [13, 4, 15, 20]. Clearly, we want the synthetic data to be faithful to the original data, so as to preserve utility. To quantify the faithfulness, we need some similarity metrics. A common and natural choice for tabular data is to try to (approximately) preserve low-dimensional marginals [3, 34, 8].

We model the *true data* $x_1, \dots, x_n$ as a sequence of n points from the Boolean cube $\{0,1\}^p$, which is a standard benchmark data model [3, 35, 13, 27]. For example, this can be n health records of patients, each containing p binary parameters (smoker/nonsmoker, etc.)[2] We are seeking to transform the true data into *synthetic data* $y_1, \dots, y_m \in \{0,1\}^p$ that is both differentially private and accurate.

As mentioned before, we measure accuracy by comparing the *marginals* of true and synthetic data. A d-dimensional marginal of the true data has the form

$$\frac{1}{n}\sum_{i=1}^{n} x_i(j_1)\cdots x_i(j_d)$$

for some given indices $j_1, \dots, j_d \in [p]$. In other words, a low-dimensional marginal is the fraction of the patients whose d given parameters all equal 1. The one-dimensional marginals encode the means of the parameters, and the two-dimensional marginals encode the covariances.

The *accuracy* of the synthetic data for a given marginal can be defined as

$$E(j_1, \dots, j_d) := \frac{1}{n}\sum_{i=1}^{n} x_i(j_1)\cdots x_i(j_d) - \frac{1}{m}\sum_{i=1}^{m} y_i(j_1)\cdots y_i(j_d). \tag{1.1}$$

Clearly, the accuracy is bounded by 1 in absolute value.

1.2 Our contributions

Our goal is to design a randomized algorithm that satisfies the following list of desiderata:

(i) **(synthetic data):** the algorithm outputs a list of vectors $y_1, \dots, y_m \in \{0,1\}^p$;

(ii) **(efficiency):** the algorithm requires only polynomial time in n and p;

(iii) **(privacy):** the algorithm is differentially private;

(iv) **(accuracy):** the low-dimensional marginals of $y_1, \dots, y_m$ are close to those of $x_1, \dots, x_n$.

There are known algorithms that satisfy any three of the above four requirements if we restrict the accuracy condition (iv) to two-dimensional marginals.

Indeed, if (i) is dropped, one can first compute the mean $\frac{1}{n}\sum_{k=1}^n x_k$ and the covariance matrix $\frac{1}{n}\sum_{k=1}^n x_k x_k^T - (\frac{1}{n}\sum_{k=1}^n x_k)(\frac{1}{n}\sum_{k=1}^n x_k)^T$, add some noise to achieve differential privacy, and output i.i.d. samples from the Gaussian distribution with the noisy mean and covariance.

Suppose (ii) is dropped. It suffices to construct a differentially private probability measure μ on $\{0,1\}^p$ so that $\int_{\{0,1\}^p} x\, d\mu(x) \approx \frac{1}{n}\sum_{k=1}^n x_k$ and $\int_{\{0,1\}^p} xx^T\, d\mu(x) \approx \frac{1}{n}\sum_{k=1}^n x_k x_k^T$. After μ is constructed, one can generate i.i.d. samples $y_1, \dots, y_m$ from μ. The measure μ can be constructed as follows: First add Laplace noises to $\frac{1}{n}\sum_{k=1}^n x_k$ and $\frac{1}{n}\sum_{k=1}^n x_k x_k^T$ (see Lemma 2.4 below) and then set μ to be a probability measure on $\{0,1\}^p$ that minimizes $\|\int_{\{0,1\}^p} x\, d\mu(x) - (\frac{1}{n}\sum_{k=1}^n x_k + \text{noise})\|_\infty + \|\int_{\{0,1\}^p} xx^T\, d\mu(x) - (\frac{1}{n}\sum_{k=1}^n x_k x_k^T + \text{noise})\|_\infty$, where $\|\ \|_\infty$ is the ℓ^∞ norm on $\mathbb{R}^p$ or $\mathbb{R}^{p^2}$. However, this requires exponential time in p, since the set of all probability measures on $\{0,1\}^p$ can be identified as a convex subset of $\mathbb{R}^{2^p}$. See [3].

If (iii) or (iv) is dropped, the problem is trivial: in the former case, we can output either the original true data; in the latter, all zeros.

[2] More generally, one can represent any categorical data (such as gender, occupation, etc.), genomic data, or numerical data (by splitting them into intervals) on the Boolean cube via binary or one-hot encoding.

While there are known algorithms that satisfy (i)–(iii) with proofs and empirically satisfy (iv) in simulations (see e.g., [22, 36, 18, 19]), the challenge is to develop an algorithm that provably satisfies all four conditions.

Ullman and Vadhan [35] showed that, assuming the existence of one-way functions, one cannot achieve (i)–(iv) even for $d = 2$, if we require in (iv) that *all* of the d-dimensional marginals be preserved accurately. More precisely, there is no polynomial time, differentially private algorithm for generating synthetic data in $\{0,1\}^p$ that preserves all of the two-dimensional marginals with accuracy $o(1)$ if one-way functions exist. This remarkable no-go result by Ullman and Vadhan already could put an end to our quest for finding an algorithm that rigorously can achieve conditions (i)–(iv).

Surprisingly, however, a slightly weaker interpretation of (iv) suffices to put our quest on a more successful path. Indeed, we will show in this paper that one can achieve (i)–(iv), if we require in (iv) that *most* of the d-dimensional marginals be preserved accurately. Remarkably, our result does not only hold for two-dimensional marginals, but for marginals of *any given degree*.

Note that even if the differential privacy condition in (iii) is replaced by the condition of anonymous microaggregation, it is still a challenging open problem to develop an algorithm that fulfills all these desiderata. In this paper we will solve this problem by deriving a computationally efficient anonymous microaggregation framework that comes with provable accuracy bounds.

Covariance loss. We approach the aforementioned goals by studying a fundamental, but a first glance completely unrelated, problem in probability. This problem is concerned with the most basic notion of probability: conditional expectation. We want to answer the fundamental question:

"How much information is lost when we take conditional expectation?"

The law of total variance shows that taking conditional expectation of a random variable underestimates the variance. A similar phenomenon holds in higher dimensions: taking conditional expectation of a random vector underestimates the covariance (in the positive-semidefinite order). We may ask: how much covariance is lost? And what sigma-algebra of given complexity minimizes the covariance loss?

Finding an answer to this fundamental probability question turns into a quest of finding among all sigma-algebras of given complexity that one which minimizes the covariance loss. We will derive a nearly optimal bound based on a careful explicit construction of a specific sigma-algebra. Amazingly, this excursion into theoretical probability produces mathematical techniques that are most suitable to solve the previously discussed challenging practical problems concerning microaggregation and privacy.

1.3 Private, synthetic data?

Now that we described the spirit of our main results, let us introduce them in more detail.

As mentioned before, it is known from Ullman and Vadhan [35] that it is generally impossible to efficiently make private synthetic data that accurately preserves all low-dimensional marginals. However, as we will prove, it is possible to efficiently construct private synthetic data that preserves *most* of the low-dimensional marginals.

To state our goal mathematically, we average the accuracy (in the L^2 sense) over all $\binom{p}{d}$ subsets of indices $\{i_1, \ldots, i_d\} \subset [p]$, then take the expectation over the randomness in the algorithm. In

other words, we would like to see

$$\mathbb{E}\binom{p}{d}^{-1} \sum_{1\le i_1<\cdots<i_d\le p} E(i_1,\dots,i_d)^2 \le \delta^2 \tag{1.2}$$

for some small δ, where $E(i_1,\dots,i_d)$ is defined in (1.1). If this happens, we say that the synthetic data is *δ-accurate for d-dimensional marginals on average*. Using Markov inequality, we can see that the synthetic data is $o(1)$-accurate for d-dimensional marginals on average if and only if with high probability, most of the d-dimensional marginals are asymptotically accurate; more precisely, with probability $1-o(1)$, a $1-o(1)$ fraction of the d-dimensional marginals of the synthetic data is within $o(1)$ of the corresponding marginals of the true data.

Let us state our result informally.

Theorem 1.1 (Private synthetic Boolean data)**.** *Let $\varepsilon, \kappa \in (0,1)$ and $n, m \in \mathbb{N}$. There exists an ε-differentially private algorithm that transforms input data $x_1,\dots,x_n \in \{0,1\}^p$ into output data $y_1,\dots,y_m \in \{0,1\}^p$. Moreover, if $d = O(1)$, $d \le p/2$, $m \gg 1$, $n \gg (p/\varepsilon)^{1+\kappa}$, then the synthetic data is $o(1)$-accurate for d-dimensional marginals on average. The algorithm runs in time polynomial in p, n and linear in m, and is independent of d.*

Theorem 5.15 gives a formal and non-asymptotic version of this result.

Our method is not specific to Boolean data. It can be used to generate synthetic data with *any predefined convex constraints* (Theorem 5.14). If we assume that the input data $x_1,\dots,x_n$ lies in some known convex set $K \subset \mathbb{R}^p$, one can make private and accurate synthetic data $y_1,\dots,y_m$ that lies in the same set K.

1.4 Covariance loss

Our method is based on a new problem in probability theory, a problem that is interesting on its own. It is about the most basic notion of probability: conditional expectation. And the question is: *how much information is lost when we take conditional expectation*?

The law of total expectation states that for a random variable X and a sigma-algebra $\mathcal{F}$, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ gives an unbiased estimate of the mean: $\mathbb{E}\,X = \mathbb{E}\,Y$. The *law of total variance*, which can be expressed as

$$\mathrm{Var}(X) - \mathrm{Var}(Y) = \mathbb{E}\,X^2 - \mathbb{E}\,Y^2 = \mathbb{E}(X-Y)^2,$$

shows that taking conditional expectation underestimates the variance.

Heuristically, the simpler the sigma-algebra $\mathcal{F}$ is, the more variance gets lost. What is the best sigma-algebra $\mathcal{F}$ with a given complexity? Among all sigma-algebras $\mathcal{F}$ that are generated by a partition of the sample space into k subsets, which one achieves the smallest loss of variance, and what is that loss?

If X is bounded, let us say $|X| \le 1$, we can decompose the interval $[-1,1]$ into k subintervals of length $2/k$ each, take F_i to be the preimage of each interval under X, and let $\mathcal{F} = \sigma(F_1,\dots,F_k)$ be the sigma-algebra generated by these events. Since X and Y takes values in the same subinterval a.s., we have $|X-Y| \le 2/k$ a.s. Thus, the law of total variance gives

$$\mathrm{Var}(X) - \mathrm{Var}(Y) \le \frac{4}{k^2}. \tag{1.3}$$

Let us try to generalize this question to higher dimensions. If X is a random vector taking values in $\mathbb{R}^p$, the law of total expectation holds unchanged. The law of total variance becomes the *law of total covariance*:

$$\Sigma_X - \Sigma_Y = \mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T} = \mathbb{E}(X-Y)(X-Y)^\mathsf{T}$$

where $\Sigma_X = \mathbb{E}(X - \mathbb{E}\, X)(X - \mathbb{E}\, X)^\mathsf{T}$ denotes the covariance matrix of X, and similarly for Σ_Y (see Lemma 3.1 below). Just like in the one-dimensional case, we see that taking conditional expectation underestimates the covariance (in the positive-semidefinite order).

However, if we naively attempt to bound the loss of covariance like we did to get (1.3), we would face a curse of dimensionality. The unit Euclidean ball in $\mathbb{R}^p$ cannot be partitioned into k subsets of diameter, let us say, $1/4$, unless k is exponentially large in p (see e.g. [2]). The following theorem[3] shows that a much better bound can be obtained that does not suffer the curse of dimensionality.

Theorem 1.2 (Covariance loss)**.** *Let X be a random vector in $\mathbb{R}^p$ such that $\|X\|_2 \le 1$ a.s. Then, for every $k \ge 3$, there exists a partition of the sample space into at most k sets such that for the sigma-algebra $\mathcal{F}$ generated by this partition, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies*

$$\|\Sigma_X - \Sigma_Y\|_2 \le C\sqrt{\frac{\log\log k}{\log k}}. \tag{1.4}$$

Here C is an absolute constant. Moreover, if the probability space has no atoms, then the partition can be made with exactly k sets, all of which have the same probability $1/k$.

Remark 1.3 (Optimality)**.** The rate in Theorem 1.2 is in general optimal up to a $\sqrt{\log\log k}$ factor; see Proposition 3.14.

Remark 1.4 (Higher moments)**.** Theorem 1.2 can be automatically extended to higher moments via the following *tensorization principle* (Theorem 3.10), which states that for all $d \ge 2$,

$$\big\|\mathbb{E}\, X^{\otimes d} - \mathbb{E}\, Y^{\otimes d}\big\|_2 \le 4^d \big\|\mathbb{E}\, X^{\otimes 2} - \mathbb{E}\, Y^{\otimes 2}\big\|_2 = 4^d \big\|\Sigma_X - \Sigma_Y\big\|_2. \tag{1.5}$$

Remark 1.5 (Hilbert spaces)**.** The bound (1.4) is dimension-free. Indeed, Theorem 1.2 can be extended to hold for infinite dimensional Hilbert spaces.

1.5 Anonymous microaggregation

Let us apply these abstract probability results to the problem of making synthetic data. As before, denote the true data by $x_1, \dots, x_n \in \mathbb{R}^p$. Let $X(i) = x_i$ be the random variable on the sample space $[n]$ equipped with uniform probability distribution. Obtain a partition $[n] = I_1 \cup \dots \cup I_m$ from the Covariance Loss Theorem 1.2, where $m \le k$, and let us assume for simplicity that $m = k$ and that all sets I_j have the same cardinality $|I_j| = n/k$ (this can be achieved whenever k divides n, a requirement that can easily be dropped as we will discuss later). The conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ on the sigma-algebra $\mathcal{F} = \sigma(I_1, \dots, I_k)$ generated by this partition takes values

$$y_j = \frac{k}{n}\sum_{i \in I_j} x_i, \quad j = 1, \dots, k. \tag{1.6}$$

with probability $1/k$ each. In other words, the synthetic data $y_1, \dots, y_k$ is obtained by taking local averages, or by *microaggregation* of the input data $x_1, \dots, x_n$. The crucial point is that the synthetic data is obviously generated via (n/k)-*anonymous microaggregation*. Here, we use the following formal definition of r-anonymous microaggregation.

[3] The ℓ_2 norm and the tensor notation used in this section are defined in Section 2.2.

Definition 1.6. Let $x_1, \dots, x_n \in \mathbb{R}^p$ be a dataset. Let $r \in \mathbb{N}$. A r-anonymous averaging of $x_1, \dots, x_n$ is a dataset consisting of the points $\sum_{i \in I_1} x_i, \dots, \sum_{i \in I_m} x_i$ for some partition $[n] = I_1 \cup \dots \cup I_m$ such that $|I_i| \geq r$ for each $1 \leq i \leq m$. A r-anonymous microaggregation algorithm $\mathcal{A}(\,)$ with input dataset $x_1, \dots, x_n \in \mathbb{R}^p$ is the composition of a r-anonymous averaging procedure followed by any algorithm.

For any notion of privacy, any post-processing of a private dataset should still be considered as private. While a post-processing of a r-anonymous averaging of a dataset is not necessarily a r-anonymous averaging of the original dataset (it might not even consist of vectors), the notion of r-anonymous microaggregation allows a post-processing step after r-anonymous averaging.

What about the accuracy? The law of total expectation $\mathbb{E}\, X = \mathbb{E}\, Y$ becomes $\frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{k}\sum_{j=1}^k y_j$. As for higher moments, assume that $\|x_i\|_2 \leq 1$ for all i. Then Covariance Loss Theorem 1.2 together with tensorization principle (1.5) yields

$$\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \frac{1}{k}\sum_{j=1}^k y_j^{\otimes d} \right\|_2 \lesssim 4^d \sqrt{\frac{\log\log k}{\log k}}.$$

Thus, if $k \gg 1$ and $d = O(1)$, the synthetic data is accurate in the sense of the mean square average of marginals.

This general principle can be specialized to Boolean data. Doing appropriate rescaling, bootstrapping (Section 4.2) and randomized rounding (Section 4.3), we can conclude the following:

Theorem 1.7 (Anonymous synthetic Boolean data)**.** *Suppose k divides n. There exists a randomized (n/k)-anonymous microaggregation algorithm that transforms input data $x_1, \dots, x_n \in \{0,1\}^p$ into output data $y_1, \dots, y_m \in \{0,1\}^p$. Moreover, if $d = O(1)$, $d \leq p/2$, $k \gg 1$, $m \gg 1$, the synthetic data is $o(1)$-accurate for d-dimensional marginals on average. The algorithm runs in time polynomial in p, n and linear in m, and is independent of d.*

Theorem 4.6 gives a formal and non-asymptotic version of this result.

1.6 Differential privacy

How can we pass from anonymity to differential privacy and establish Theorem 1.1? The microaggregation mechanism by itself is not differentially private. However, it *reduces sensitivity* of synthetic data. If a single input data point x_i is changed, microaggregation (1.6) suppresses the effect of such change on the synthetic data y_j by the factor k/n. Once the data has low sensitivity, the classical *Laplacian mechanism* can make it private: one has simply to add Laplacian noise.

This is the gist of the proof of Theorem 1.1. However, several issues arise. One is that we do not know how to make all blocks I_j of the same size while preserving their privacy, so we allow them to have arbitrary sizes in the application to privacy. However, small blocks I_j may cause instability of microaggregation, and diminish its beneficial effect on sensitivity. We resolve this issue by downplaying, or *damping*, the small blocks (Section 5.4). The second issue is that adding Laplacian noise to the vectors y_i may move them outside the set K where the synthetic data must lie (for Boolean data, K is the unit cube $[0,1]^p$.) We resolve this issue by metrically projecting the perturbed vectors back onto K (Section 5.5).

1.7 Related work

There exists a large body of work on privately releasing answers in the interactive and non-interactive query setting. But a major advantage of releasing a synthetic data set instead of just

the answers to specific queries is that synthetic data opens up a much richer toolbox (clustering, classification, regression, visualization, etc.), and thus much more flexibility, to analyze the data.

In [5], Blum, Ligett, and Roth gave an ε-differentially private synthetic data algorithm whose accuracy scales logarithmically with the number of queries, but the complexity scales exponentially with p. The papers [12, 13] propose methods for producing private synthetic data with an error bound of about $\tilde{\mathcal{O}}(\sqrt{n}p^{1/4})$ per query. However, the associated algorithms have running time that is at least exponential in p. In [3], Barak et al. derive a method for producing accurate and private synthetic Boolean data based on linear programming with a running time that is exponential in p. This should be contrasted with the fact that our algorithm runs in time polynomial in p and n, see Theorem 1.7.

We emphasize that the our method is designed to produce synthetic data. But, as suggested by the reviewers, we briefly discuss how well d-way marginals can be preserved by our method in the non-synthetic data regime. Here, we consider the dependence of n on p as well as the accuracy we achieve versus existing methods.

Dependence of n on p: In order to release 1-way marginals with any nontrivial accuracy on average and with ε-differential privacy, one already needs $n \gtrsim p$ [9, Theorem 8.7]. Our main result Theorem 1.1 on differential privacy only requires n to grow slightly faster than linearly in p.

If we just want to privately release the d-way marginals without creating synthetic data, and moreover relax ε-differential privacy to (ε, δ)-differential privacy, one might relax the dependence of n on p. Specifically, $n \gg p^{\lceil d/2 \rceil/2}\sqrt{\log(1/\delta)}/\varepsilon$ suffices [8]. In particular, when $d = 2$, this means that $n \gg \sqrt{p\log(1/\delta)}/\varepsilon$ suffices. On the other hand, our algorithm does not depend on d. Moreover, for $d \geq 5$, the dependence of n on p required in Theorem 1.1 is less restrictive than in [8].

Accuracy in n: As mentioned above, Theorem 5.15 gives a formal and non-asymptotic version of Theorem 1.1. The average error of d-way marginals we achieve has decay of order $\sqrt{\log\log n/\log n}$ in n. In Remark 5.16, we show that even for $d = 1, 2$, no polynomial time differentially private algorithm can have average error of the marginals decay faster than n^{-a} for any $a > 0$, assuming the existence of one-way functions.

However, if we only need to release d-way marginals with differential privacy but without creating synthetic data, then one can have the average error of the d-way marginals decay at the rate $1/\sqrt{n}$ [8].

1.8 Outline of the paper

The rest of the paper is organized as follows. In Section 2 we provide basic notation and other preliminaries. Section 3 is concerned with the concept of covariance loss. We give a constructive and nearly optimal answer to the problem of how much information is lost when we take conditional expectation. In Section 4 we use the tools developed for covariance loss to derive a computationally efficient microaggregation framework that comes with provable accuracy bounds regarding low-dimensional marginals. In Section 5 we obtain analogous versions of these results in the framework of differential privacy.

2 Preliminaries

2.1 Basic notation

The approximate inequality signs $\lesssim$ hide absolute constant factors; thus $a \lesssim b$ means that $a \leq Cb$ for a suitable absolute constant $C > 0$. A list of elements $\nu_1, \ldots, \nu_k$ of a metric space M is an

α-covering, where $\alpha > 0$, if every element of M has distance less than α from one of $\nu_1, \dots, \nu_k$. For $p \in \mathbb{N}$, define

$$B_2^p = \{x \in \mathbb{R}^p : \|x\|_2 \le 1\}.$$

2.2 Tensors

The marginals of a random vector can be conveniently represented in tensor notation. A tensor is a d-way array $X \in \mathbb{R}^{p \times \cdots \times p}$. In particular, 1-way tensors are vectors, and 2-way tensors are matrices. A simple example of a tensor is the rank-one tensor $x^{\otimes d}$, which is constructed from a vector $x \in \mathbb{R}^p$ by multiplying its entries:

$$x^{\otimes d}(i_1, \dots, i_d) = x(i_1) \cdots x(i_d), \quad \text{where } i_1, \dots, i_d \in [p].$$

In particular, the tensor $x^{\otimes 2}$ is the same as the matrix xx^{T}.

The ℓ^2 norm of a tensor X can be defined by regarding X as a vector in $\mathbb{R}^{p^d}$, thus

$$\|X\|_2^2 \coloneqq \sum_{i_1, \dots i_d \in [p]} \left|X(i_1, \dots, i_d)\right|^2.$$

Note that when $d = 2$, the tensor X can be identified as a matrix and $\|X\|_2$ is the Frobenius norm of X.

The errors of the marginals (1.1) can be thought of as the coefficients of the error tensor

$$E = \frac{1}{n} \sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m} \sum_{i=1}^m y_i^{\otimes d}, \tag{2.1}$$

A tensor $X \in \mathbb{R}^{p \times \cdots \times p}$ is *symmetric* if the values of its entries are independent of the permutation of the indices, i.e. if

$$X(i_1, \dots, i_d) = X(i_{\pi(1)}, \dots, i_{\pi(d)})$$

for any permutation π of $[p]$. It often makes sense to count each distinct entry of a symmetric tensor once instead of $d!$ times. To make this formal, we may consider the restriction operator P_{sym} that preserves the $\binom{p}{d}$ entries whose indices satisfy $1 \le i_1 < \cdots < i_d \le p$, and zeroes out all other entries. Thus

$$\left\|P_{\text{sym}} X\right\|_2^2 = \sum_{1 \le i_1 < \cdots < i_d \le p} X(i_1, \dots, i_d)^2.$$

Thus, the goal we stated in (1.2) can be restated as follows: for the error tensor (2.1), we would like to bound the quantity

$$\mathbb{E} \binom{p}{d}^{-1} \sum_{1 \le i_1 < \cdots < i_d \le p} E(i_1, \dots, i_d)^2 = \mathbb{E} \binom{p}{d}^{-1} \left\|P_{\text{sym}} E\right\|_2^2. \tag{2.2}$$

The operator P_{sym} is related to another restriction operator P_{off}, which retains the $\binom{p}{d} d!$ off-diagonal entires, i.e. those for which all indices $i_1, \dots i_d$ are distinct, and zeroes out all other entries. Thus,

$$\|P_{\text{off}} X\|_2^2 = \sum_{i_1, \dots i_d \in [p] \text{ distinct}} X(i_1, \dots, i_d)^2 = d! \left\|P_{\text{sym}} X\right\|_2^2, \tag{2.3}$$

for all symmetric tensor X.

Lemma 2.1. *If $p \geq 2d$, we have*

$$\binom{p}{d}^{-1} \|P_{\mathrm{sym}} X\|_2^2 \leq \left(\frac{2}{p}\right)^d \|P_{\mathrm{off}} X\|_2^2, \tag{2.4}$$

for all symmetric d-way tensor X.

Proof. According to (2.3), the left hand side of (2.4) equals $\left(\binom{p}{d} d!\right)^{-1} \|P_{\mathrm{off}} X\|_2^2$, and $\binom{p}{d} d! = p(p-1)\cdots(p-d+1) \geq (p/2)^d$ if $p \geq 2d$. This yields the desired bound. □

2.3 Differential privacy

We briefly review some basic facts about differential privacy. The interested reader may consult [9] for details.

Definition 2.2 (Differential Privacy [9])**.** A randomized function $\mathcal{M}$ gives ε-differential privacy if for all databases D_1 and D_2 differing on at most one element, and all measurable $S \subseteq \mathrm{range}(\mathcal{M})$,

$$\mathbb{P}\left\{\mathcal{M}(D_1) \in S\right\} \leq e^{\varepsilon} \cdot \mathbb{P}\left\{\mathcal{M}(D_2) \in S\right\},$$

where the probability is with respect to the randomness of $\mathcal{M}$.

Almost all existing mechanisms to implement differential privacy are based on adding noise to the data or the data queries, e.g. via the Laplacian mechanism [3]. Recall that a random variable has the (centered) Laplacian distribution $\mathrm{Lap}(\sigma)$ if its probability density function at x is $\frac{1}{2\sigma}\exp(-|x|/\sigma)$.

Definition 2.3. For $f : \mathcal{D} \to \mathbb{R}^d$, the L_1-sensitivity is

$$\Delta f := \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1,$$

for all D_1, D_2 differing in at most one element.

Lemma 2.4 (Laplace mechanism, Theorem 2 in [3])**.** *For any $f : \mathcal{D} \to \mathbb{R}^d$, the addition of $\mathrm{Lap}(\sigma)^d$ noise preserves $(\Delta f/\sigma)$-differential privacy.*

The proof of the following lemma, which is similar in spirit to the Composition Theorem 3.14 in [9], is left to the reader.

Lemma 2.5. *Suppose that an algorithm $\mathcal{A}_1 : \mathcal{D} \to \mathcal{Y}_1$ is ε_1-differentially private and an algorithm $\mathcal{A}_2 : \mathcal{D} \times \mathcal{Y}_1 \to \mathcal{Y}_2$ is ε_2-differentially private in the first component in $\mathcal{D}$. Assume that $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent. Then the composition algorithm $\mathcal{A} = \mathcal{A}_2(\,\cdot\,, \mathcal{A}_1(\cdot))$ is $(\varepsilon_1 + \varepsilon_2)$-differentially private.*

Remark 2.6. As outlined in [3], any function applied to private data, without accessing the raw data, is privacy-preserving.

The following observation is a special case of Lemma 2.5.

Lemma 2.7. *Suppose the data Y_1 and Y_2 are independent with respect to the randomness of the privacy-generating algorithm and that each is ε-differentially private, then (Y_1, Y_2) is 2ε-differentially private.*

3 Covariance loss

The goal of this section is to prove Theorem 1.2 and its higher-order version, Corollary 3.12. We will establish the main part of Theorem 1.2 in Sections 3.1–3.4, the "moreover" part (equipartition) in Sections 3.5–3.6, the tensorization principle (1.5) in Section 3.7, and then immediately yields Corollary 3.12. Finally, we show optimality in Section 3.8.

3.1 Law of total covariance

Throughout this section, X is an arbitrary random vector in $\mathbb{R}^p$, $\mathcal{F}$ is an arbitrary sigma-algebra and $Y = \mathbb{E}[X|\mathcal{F}]$ is the conditional expectation.

Lemma 3.1 (Law of total covariance)**.** *We have*

$$\Sigma_X - \Sigma_Y = \mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T} = \mathbb{E}(X-Y)(X-Y)^\mathsf{T}.$$

In particular, $\Sigma_X \succeq \Sigma_Y$*.*

Proof. The covariance matrix can be expressed as

$$\Sigma_X = \mathbb{E}(X - \mathbb{E}\, X)(X - \mathbb{E}\, X)^\mathsf{T} = \mathbb{E}\, XX^\mathsf{T} - (\mathbb{E}\, X)(\mathbb{E}\, X)^\mathsf{T}$$

and similarly for Y. Since $\mathbb{E}\, X = \mathbb{E}\, Y$ by the law of total expectation, we have $\Sigma_X - \Sigma_Y = \mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T}$, proving the first equality in the lemma. Next, one can check that

$$\mathbb{E}\left[XX^\mathsf{T}|\mathcal{F}\right] - YY^\mathsf{T} = \mathbb{E}\left[(X-Y)(X-Y)^\mathsf{T}|\mathcal{F}\right] \quad \text{almost surely}$$

by expanding the product in the right hand side and recalling that Y is $\mathcal{F}$-measurable. Finally, take expectation on both sides to complete the proof. □

Lemma 3.2 (Decomposing the covariance loss)**.** *For any orthogonal projection P in $\mathbb{R}^p$ we have*

$$\left\|\mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T}\right\|_2 \le \mathbb{E}\|PX - PY\|_2^2 + \left\|(I-P)(\mathbb{E}\, XX^\mathsf{T})(I-P)\right\|_2.$$

Proof. By the law of total covariance (Lemma 3.1), the matrix

$$A \coloneqq \mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T} = \mathbb{E}(X-Y)(X-Y)^\mathsf{T}$$

is positive semidefinite. Then we can use the following inequality, which holds for any positive-semidefinite matrix A (see e.g. in [1, p.157]):

$$\|A\|_2 \le \|PAP\|_2 + \left\|(I-P)A(I-P)\right\|_2. \tag{3.1}$$

Let us bound the two terms in the right hand side. Jensen's inequality gives

$$\|PAP\|_2 \le \mathbb{E}\left\|P(X-Y)(X-Y)^\mathsf{T}P\right\|_2 = \mathbb{E}\|PX - PY\|_2^2.$$

Next, since the matrix $\mathbb{E}\, YY^\mathsf{T}$ is positive-semidefinite, we have $0 \preceq A \preceq \mathbb{E}\, XX^\mathsf{T}$ in the semidefinite order, so $0 \preceq (I-P)A(I-P) \preceq (I-P)(\mathbb{E}\, XX^\mathsf{T})(I-P)$, which yields

$$\left\|(I-P)A(I-P)\right\|_2 \le \left\|(I-P)(\mathbb{E}\, XX^\mathsf{T})(I-P)\right\|_2.$$

Substitute the previous two bounds into (3.1) to complete the proof. □

3.2 Spectral projection

The two terms in Lemma 3.2 will be bounded separately. Let us start with the second term. It simplifies if P is a spectral projection:

Lemma 3.3 (Spectral projection)**.** *Assume that* $\|X\|_2 \le 1$ *a.s. Let* $t \in \mathbb{N} \cup \{0\}$*. Let* P *be the orthogonal projection in* $\mathbb{R}^p$ *onto the* t *leading eigenvectors of the second moment matrix* $S = \mathbb{E}\, XX^\mathsf{T}$*. Then*

$$\big\|(I-P)S(I-P)\big\|_2 \le \frac{1}{\sqrt{t}}.$$

Proof. We have

$$\big\|(I-P)S(I-P)\big\|_2^2 = \sum_{i>t} \lambda_i(S)^2 \tag{3.2}$$

where $\lambda_i(S)$ denote the eigenvalues of S arranged in a non-increasing order. Using linearity of expectation and trace, we get

$$\sum_{i=1}^{p} \lambda_i(S) = \mathbb{E}\,\mathrm{tr}\, XX^\mathsf{T} = \mathbb{E}\|X\|_2^2 \le 1.$$

It follows that at most t eigenvalues of S can be larger than $1/t$. By monotonicity, this yields $\lambda_i(S) \le 1/t$ for all $i > t$. Combining this with the bound above, we conclude that

$$\sum_{i>t} \lambda_i(S)^2 \le \sum_{i>t} \lambda_i(S) \cdot \frac{1}{t} \le \frac{1}{t}. \tag{3.3}$$

Substitute this bound into (3.2) to complete the proof. □

3.3 Nearest-point partition

Next, we bound the first term in Lemma 3.2. This is the only step that does not hold generally but for a specific sigma-algebra, which we generate by a nearest-point partition.

Definition 3.4 (Nearest-point partition)**.** Let X be a random vector taking values in $\mathbb{R}^p$, defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. A nearest-point partition $\{F_1, \dots, F_s\}$ of Ω with respect to a list of points $\nu_1, \dots, \nu_s \in \mathbb{R}^p$ is a partition $\{F_1, \dots, F_s\}$ of Ω such that

$$\|\nu_j - X(\omega)\|_2 = \min_{1 \le i \le s} \|\nu_i - X(\omega)\|_2,$$

for all $\omega \in F_j$ and $1 \le j \le s$. (Some of the F_j could be empty.)

Remark 3.5. A nearest-point partition can be constructed as follows: For each $\omega \in \Omega$, choose a point ν_j nearest to $X(\omega)$ in the ℓ^2 metric and put ω into F_j. Break any ties arbitrarily as long as the F_j are measurable.

Lemma 3.6 (Approximation)**.** *Let* X *be a random vector in* $\mathbb{R}^p$ *such that* $\|X\|_2 \le 1$ *a.s. Let* P *be an orthogonal projection on* $\mathbb{R}^p$*. Let* $\nu_1, \dots, \nu_s \in \mathbb{R}^p$ *be an* α*-covering of the unit Euclidean ball of* $\mathrm{ran}(P)$*. Let* $\Omega = F_1 \cup \cdots \cup F_s$ *be a nearest-point partition for* PX *with respect to* $\nu_1, \dots, \nu_s$*. Let* $\mathcal{F} = \sigma(F_1, \dots, F_s)$ *be the sigma-algebra generated by the partition. Then the conditional expectation* $Y = \mathbb{E}[X|\mathcal{F}]$ *satisfies*

$$\|PX - PY\|_2 \le 2\alpha \quad \textit{almost surely.}$$

Proof. If $\omega \in F_j$ then, by the definition of the nearest-point partition, $\|\nu_j - PX(\omega)\|_2 = \min_{1\le i\le s} \|\nu_i - PX(\omega)\|_2$. So by the definition of the α-covering, we have $\big\|PX(\omega) - \nu_j\big\|_2 \le \alpha$. Hence, by the triangle inequality we have

$$\big\|P(X(\omega) - X(\omega'))\big\|_2 \le 2\alpha \quad \text{whenever } \omega, \omega' \in F_j. \tag{3.4}$$

Furthermore, by definition of Y, we have

$$Y(\omega) = \frac{1}{\mathbb{P}(F_j)} \int_{F_j} X(\omega')\, d\mathbb{P}(\omega') \quad \text{whenever } \omega \in F_j.$$

Thus, for such ω we have

$$X(\omega) - Y(\omega) = \frac{1}{\mathbb{P}(F_j)} \int_{F_j} \big(X(\omega) - X(\omega')\big)\, d\mathbb{P}(\omega').$$

Applying the projection P and taking norm on both sides, then using Jensen's inequality, we conclude that

$$\big\|PX(\omega) - PY(\omega)\big\|_2 \le \frac{1}{\mathbb{P}(F_j)} \int_{F_j} \Big\|P\big(X(\omega) - X(\omega')\big)\Big\|_2\, d\mathbb{P}(\omega') \le 2\alpha,$$

where in the last step we used (3.4). Since the bound holds for each $\omega \in F_j$ and the events F_j form a partition of Ω, it holds for all $\omega \in \Omega$. The proof is complete. □

3.4 Proof of the main part of Theorem 1.2

The following simple (and possibly known) observation will come handy to bound the cardinality of an α-covering that we will need in the proof of Theorem 1.2.

Proposition 3.7 (Number of lattice points in a ball)**.** *For all $\alpha > 0$ and $t \in \mathbb{N}$,*

$$\left|B_2^t \cap \frac{\alpha}{\sqrt{t}}\mathbb{Z}^t\right| \le \left(\Big(\frac{1}{\alpha} + \frac{1}{2}\Big)\sqrt{2e\pi}\right)^t.$$

In particular, for any $\alpha \in (0,1)$, it follows that

$$\left|B_2^t \cap \frac{\alpha}{\sqrt{t}}\mathbb{Z}^t\right| \le \left(\frac{7}{\alpha}\right)^t.$$

Proof. The open cubes of side length $\alpha/\sqrt{t}$ that are centered at the points of the set $\mathcal{N} := B_2^t \cap \frac{\alpha}{\sqrt{t}}\mathbb{Z}^t$ are all disjoint. Thus the total volume of these cubes equals $|\mathcal{N}|\,(\alpha/\sqrt{t})^t$.

On the other hand, since each such cube is contained in a ball of radius $\alpha/2$ centered at some point of $\mathcal{N}$, the union of these cubes is contained in the ball $(1+\alpha/2)B_2^t$. So, comparing the volumes, we obtain

$$|\mathcal{N}|\,(\alpha/\sqrt{t})^t \le (1+\alpha/2)^t \operatorname{Vol}(B_2^t),$$

or

$$|\mathcal{N}| \le \left(\Big(\frac{1}{\alpha} + \frac{1}{2}\Big)\sqrt{t}\right)^t \operatorname{Vol}(B_2^t).$$

Now, it is well known that [23]

$$\mathrm{Vol}(B_2^t) = \frac{\pi^{t/2}}{\Gamma(t/2+1)}.$$

Using Stirling's formula we have

$$\Gamma(x+1) \geq \sqrt{\pi}(x/e)^x, \quad x \geq 0.$$

This gives

$$\mathrm{Vol}(B_2^t) \leq (2e\pi/t)^{t/2}.$$

Substitute this into the bound on $|\mathcal{N}|$ above to complete the proof. □

It follows now from Proposition 3.7 that for every $\alpha \in (0,1)$, there exists an α-covering in the unit Euclidean ball of dimension t of cardinality at most $(7/\alpha)^t$.

Fix an integer $k \geq 3$ and choose

$$t := \left\lfloor \frac{\log k}{\log(7/\alpha)} \right\rfloor, \quad \alpha := \left(\frac{\log\log k}{\log k}\right)^{1/4}. \tag{3.5}$$

The choice of t is made so that we can find an α-covering of the unit Euclidean ball of $\mathrm{ran}(P)$ of cardinality at most $(7/\alpha)^t \leq k$.

We decompose the covariance loss $\Sigma_X - \Sigma_Y$ in Lemma 3.1 into two terms as in Lemma 3.2 and bound the two terms as in Lemma 3.3 and Lemma 3.6. This way we obtain

$$\|\Sigma_X - \Sigma_Y\|_2 = \left\|\mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T}\right\|_2 \leq 4\alpha^2 + \frac{1}{\sqrt{t}} \lesssim \sqrt{\frac{\log\log k}{\log k}},$$

where the last bound follows from our choice of α and t. If $t = 0$ then $k \leq C$, for some universal constant $C > 0$, so $\|\Sigma_X - \Sigma_Y\|_2$ is at most $O(1)$ and $\sqrt{\log\log k/\log k} = O(1)$. The main part of Theorem 1.2 is proved. □

3.5 Monotonicity

Next, we are going to prove the "moreover" (equipartition) part of Theorem 1.2. This part is crucial in the application for anonymity, but it can be skipped if the reader is only interested in differential privacy.

Before we proceed, let us first note a simple monotonicity property:

Lemma 3.8 (Monotonicity)**.** *Conditioning on a larger sigma-algebra can only decrease the covariance loss. Specifically, if Z is a random variable and $\mathcal{B} \subset \mathcal{G}$ are sigma-algebras then*

$$\left\|\Sigma_Z - \Sigma_{\mathbb{E}[Z|\mathcal{G}]}\right\|_2 \leq \left\|\Sigma_Z - \Sigma_{\mathbb{E}[Z|\mathcal{B}]}\right\|_2.$$

Proof. Denoting $X = \mathbb{E}[Z|\mathcal{G}]$ and $Y = \mathbb{E}[Z|\mathcal{B}]$, we see from the law of total expectation that $Y = \mathbb{E}[X|\mathcal{B}]$. The law of total covariance (Lemma 3.1) then yields $\Sigma_Z \succeq \Sigma_X \succeq \Sigma_Y$, which we can rewrite as $0 \preceq \Sigma_Z - \Sigma_X \preceq \Sigma_Z - \Sigma_Y$. From this relation, it follows that $\|\Sigma_Z - \Sigma_X\|_2 \leq \|\Sigma_Z - \Sigma_Y\|_2$, as claimed. □

Passing to a smaller sigma-algebra may in general increase the covariance loss. The additional covariance loss can be bounded as follows:

Lemma 3.9 (Merger). *Let Z be a random vector in $\mathbb{R}^p$ such that $\|Z\|_2 \le 1$ a.s. If a sigma-algebra is generated by a partition, merging elements of the partition may increase the covariance loss by at most the total probability of the merged sets. Specifically, if $\mathcal{G} = \sigma(G_1, \dots, G_m)$ and $\mathcal{B} = \sigma(G_1 \cup \dots \cup G_r, G_{r+1}, \dots, G_m)$, then the random vectors $X = \mathbb{E}[Z|\mathcal{G}]$ and $Y = \mathbb{E}[Z|\mathcal{B}]$ satisfy*

$$0 \le \|\Sigma_Z - \Sigma_Y\|_2 - \|\Sigma_Z - \Sigma_X\|_2 \le \mathbb{P}(G_1 \cup \dots \cup G_r).$$

Proof. The lower bound follows from monotonicity (Lemma 3.8). To prove the upper bound, we have

$$\|\Sigma_Z - \Sigma_Y\|_2 - \|\Sigma_Z - \Sigma_X\|_2 \le \|\Sigma_X - \Sigma_Y\|_2 \le \mathbb{E}\|X - Y\|_2^2 \tag{3.6}$$

where the first bound follows by triangle inequality, and the second from Lemma 3.1 and Lemma 3.2 for $P = I$.

Denote by $\mathbb{E}_G$ the conditional expectation on the set $G = G_1 \cup \dots \cup G_r$, i.e. $\mathbb{E}_G[Z] = \mathbb{E}[Z|G] = \mathbb{P}(G)^{-1}\,\mathbb{E}[Z\mathbf{1}_G]$. Then

$$Y(\omega) = \begin{cases} \mathbb{E}_G X, & \omega \in G \\ X(\omega), & \omega \in G^c \end{cases}$$

Indeed, to check the first case, note that since $\mathcal{B} \subset \mathcal{G}$, the law of total expectation yields $Y = \mathbb{E}[X|\mathcal{B}]$; then the case follows since $G \in \mathcal{B}$. To check the second case, note that since the sets $G_{r+1}, \cdots, G_m$ belong to both sigma-algebras $\mathcal{G}$ and $\mathcal{B}$, so the conditional expectations X and Y must agree on each of these sets and thus on their union G^c. Hence

$$\mathbb{E}\|X - Y\|_2^2 = \mathbb{E}\|X - Y\|_2^2\,\mathbf{1}_G = \mathbb{P}(G) \cdot \mathbb{E}_G\|X - \mathbb{E}_G X\|_2^2 \le \mathbb{P}(G) \cdot \mathbb{E}_G\|X\|_2^2 \le \mathbb{P}(G).$$

Here we bounded the variance by the second moment, and used the assumption that $\|X\|_2 \le 1$ a.s. Substitute this bound into (3.6) to complete the proof. □

3.6 Proof of equipartition (the "moreover" part of Theorem 1.2)

Let $k' = \lfloor \sqrt{k} \rfloor$. Assume that $k' \ge 3$. (Otherwise $k < 9$ and the result is trivial by taking arbitrary partition into k of the same probability.) Applying the first part of Theorem 1.2 for k' instead of k, we obtain a sigma-algebra $\mathcal{F}'$ generated by a partition of a sample space into at most k' sets F_i, and such that

$$\left\|\Sigma_X - \Sigma_{\mathbb{E}[X|\mathcal{F}']}\right\|_2 \lesssim \sqrt{\frac{\log\log k'}{\log k'}} \lesssim \sqrt{\frac{\log\log k}{\log k}}.$$

Divide each set F_i into subsets with probability $1/k$ each using division with residual. Thus we partition each F_i into a certain number of subsets (if any) of probability $1/k$ each and one residual subset of probability less than $1/k$. By Monotonicity Lemma 3.8, any partitioning can only reduce the covariance loss.

This process results in the creation of a lot of good subsets – each having probability $1/k$ – and at most k' residual subsets that have probability less than $1/k$ each. Merge all residuals into one new "residual subset". While a merger may increase the covariance loss, Lemma 3.9 guarantees that the additional loss is bounded by the probability of the set being merged. Since we chose $k' = \lfloor \sqrt{k} \rfloor$, the probability of the residual subset is less than $k' \cdot (1/k) \le 1/\sqrt{k}$. So the additional covariance loss is bounded by $1/\sqrt{k}$.

Finally, divide the residual subset into further subsets of probability $1/k$ each. By monotonicity, any partitioning may not increase the covariance loss. At this point we partitioned the sample space into subsets of probability $1/k$ each and one smaller residual subset. Since k is an integer, the

residual must have probability zero, and thus can be added to any other subset without affecting the covariance loss.

Let us summarize. We partitioned the sample space into k subsets of equal probability such that the covariance loss is bounded by

$$\frac{1}{\sqrt{k}} + C\sqrt{\frac{\log\log k}{\log k}} \lesssim \sqrt{\frac{\log\log k}{\log k}}.$$

The proof is complete. □

3.7 Higher moments: tensorization

Recall that Theorem 1.2 provides a bound on the covariance loss[4]

$$\Sigma_X - \Sigma_Y = \mathbb{E}\, XX^\mathsf{T} - \mathbb{E}\, YY^\mathsf{T} = \mathbb{E}\, X^{\otimes 2} - \mathbb{E}\, Y^{\otimes 2}. \tag{3.7}$$

Perhaps counterintuitively, the bound on the covariance loss can automatically be lifted to higher moments, at the cost of multiplying the error by at most 4^d.

Theorem 3.10 (Tensorization)**.** *Let X be a random vector in $\mathbb{R}^p$ such that $\|X\|_2 \le 1$ a.s., let $\mathcal{F}$ be a sigma-algebra, and let $d \ge 2$ be an integer. Then the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies*

$$\left\|\mathbb{E}\, X^{\otimes d} - \mathbb{E}\, Y^{\otimes d}\right\|_2 \le 2^{d-2}(2^d - d - 1)\left\|\mathbb{E}\, X^{\otimes 2} - \mathbb{E}\, Y^{\otimes 2}\right\|_2.$$

For the proof, we need an elementary identity:

Lemma 3.11. *Let U and V be independent and identically distributed random vectors in $\mathbb{R}^p$. Then*

$$\mathbb{E}\langle U, V\rangle^2 = \left\|\mathbb{E}\, UU^\mathsf{T}\right\|_2^2.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\langle U, V\rangle^2 &= \mathbb{E}(V^\mathsf{T}U)(U^\mathsf{T}V) = \mathbb{E}\operatorname{tr} V^\mathsf{T}UU^\mathsf{T}V \\
&= \mathbb{E}\operatorname{tr} UU^\mathsf{T}VV^\mathsf{T} \quad \text{(cyclic property of trace)} \\
&= \operatorname{tr}\mathbb{E}\, UU^\mathsf{T}VV^\mathsf{T} \quad \text{(linearity)} \\
&= \operatorname{tr}\mathbb{E}[UU^\mathsf{T}]\ \mathbb{E}[VV^\mathsf{T}] \quad \text{(independence)} \\
&= \operatorname{tr}(\mathbb{E}\, UU^\mathsf{T})^2 \quad \text{(identical distribution)} \\
&= \left\|\mathbb{E}\, UU^\mathsf{T}\right\|_2^2 \quad \text{(the matrix } \mathbb{E}\, UU^\mathsf{T} \text{ is symmetric).}
\end{aligned}$$

□

Proof of Theorem 3.10. Step 1: binomial decomposition. Denoting

$$X_0 = Y, \quad X_1 = X - Y,$$

we can represent

$$X^{\otimes d} = (X_0 + X_1)^{\otimes d} = \sum_{i_1,\dots,i_d \in \{0,1\}} X_{i_1} \otimes \cdots \otimes X_{i_d}.$$

[4]Recall Lemma 3.1 for the first identity, and refer to Section 2.2 for the tensor notation.

Since $Y^{\otimes d} = X_0 \otimes \cdots \otimes X_0$, it follows that

$$X^{\otimes d} - Y^{\otimes d} = \sum_{\substack{i_1,\dots,i_d \in \{0,1\} \\ i_1+\cdots+i_d \geq 1}} X_{i_1} \otimes \cdots \otimes X_{i_d}.$$

Taking expectation on both sides and using triangle inequality, we obtain

$$\left\|\mathbb{E}\, X^{\otimes d} - \mathbb{E}\, Y^{\otimes d}\right\|_2 \leq \sum_{\substack{i_1,\dots,i_d \in \{0,1\} \\ i_1+\cdots+i_d \geq 1}} \left\|\mathbb{E}\, X_{i_1} \otimes \cdots \otimes X_{i_d}\right\|_2. \tag{3.8}$$

Let us look at each summand on the right hand side separately.

Step 2: Dropping trivial terms. First, let us check that all summands for which $i_1 + \cdots + i_d = 1$ vanish. Indeed, in this case exactly one term in the product $X_{i_1} \otimes \cdots \otimes X_{i_d}$ is X_1, while all other terms are X_0. Let $\mathbb{E}_{\mathcal{F}}$ denote conditional expectation with respect to $\mathcal{F}$. Since $\mathbb{E}_{\mathcal{F}} X_1 = \mathbb{E}_{\mathcal{F}}[X - \mathbb{E}_{\mathcal{F}} X] = 0$ and $X_0 = Y = \mathbb{E}_{\mathcal{F}} X$ is $\mathcal{F}$-measurable, it follows that $\mathbb{E}_{\mathcal{F}} X_{i_1} \otimes \cdots \otimes X_{i_d} = 0$. Thus, $\mathbb{E}\, X_{i_1} \otimes \cdots \otimes X_{i_d} = 0$ as we claimed.

Step 3: Bounding nontrivial terms. Next, we bound the terms for which $r = i_1 + \cdots + i_d \geq 2$. Let (X_0', X_1') be an independent copy of the pair of random variables (X_0, X_1). Then $\mathbb{E}\, X_{i_1} \otimes \cdots \otimes X_{i_d} = \mathbb{E}\, X_{i_1}' \otimes \cdots \otimes X_{i_d}'$, so

$$\begin{aligned}
\left\|\mathbb{E}\, X_{i_1} \otimes \cdots \otimes X_{i_d}\right\|_2^2 &= \langle \mathbb{E}\, X_{i_1} \otimes \cdots \otimes X_{i_d}, \mathbb{E}\, X_{i_1}' \otimes \cdots \otimes X_{i_d}' \rangle \\
&= \mathbb{E}\langle X_{i_1} \otimes \cdots \otimes X_{i_d}, X_{i_1}' \otimes \cdots \otimes X_{i_d}' \rangle \quad \text{(by independence)} \\
&= \mathbb{E}\langle X_{i_1}, X_{i_1}' \rangle \cdots \langle X_{i_d}, X_{i_d}' \rangle \\
&= \mathbb{E}\langle X_0, X_0' \rangle^{d-r} \langle X_1, X_1' \rangle^r.
\end{aligned}$$

By assumption, we have $\|X\|_2 \leq 1$ a.s., which implies by Jensen's inequality that $\|X_0\|_2 = \|\mathbb{E}_{\mathcal{F}} X\|_2 \leq \mathbb{E}_{\mathcal{F}} \|X\|_2 \leq 1$ a.s. These bounds imply by the triangle inequality that $\|X_1\|_2 = \|X - X_0\|_2 \leq 2$ a.s. By identical distribution, we also have $\|X_0'\|_2 \leq 1$ and $\|X_1'\|_2 \leq 2$ a.s. Hence,

$$\left|\langle X_0, X_0' \rangle\right| \leq 1, \quad \left|\langle X_1, X_1' \rangle\right| \leq 4 \quad \text{a.s.}$$

Returning to the term we need to bound, this yields

$$\begin{aligned}
\left\|\mathbb{E}\, X_{i_1} \otimes \cdots \otimes X_{i_d}\right\|_2^2 &\leq 4^{r-2}\, \mathbb{E}\langle X_1, X_1' \rangle^2 \\
&\leq 4^{d-2} \left\|\mathbb{E}\, X_1 X_1^{\mathsf{T}}\right\|_2^2 \quad \text{(by Lemma 3.11)} \\
&= 4^{d-2} \left\|\mathbb{E}(X - Y)(X - Y)^{\mathsf{T}}\right\|_2^2 \\
&= 4^{d-2} \left\|\mathbb{E}\, X^{\otimes 2} - \mathbb{E}\, Y^{\otimes 2}\right\|_2^2 \quad \text{(by Lemma 3.1)}.
\end{aligned}$$

Step 4: Conclusion. Let us summarize. The sum on the right side of (3.8) has $2^d - 1$ terms. The d terms corresponding to $i_1 + \cdots + i_d = 1$ vanish. The remaining $2^d - d - 1$ terms are bounded by $K := 2^{d-2} \left\|\mathbb{E}\, X^{\otimes 2} - \mathbb{E}\, Y^{\otimes 2}\right\|_2$ each. Hence the entire sum is bounded by $(2^d - d - 1)K$, as claimed. The theorem is proved. □

Combining the Covariance Loss Theorem 1.2 with Theorem 3.10 in view of (3.7), we conclude:

Corollary 3.12 (Tensorization). *Let X be a random vector in $\mathbb{R}^p$ such that $\|X\|_2 \le 1$ a.s. Then, for every $k \ge 3$, there exists a partition of the sample space into at most k sets such that for the sigma-algebra $\mathcal{F}$ generated by this partition, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies for all $d \in \mathbb{N}$,*

$$\left\|\mathbb{E}\, X^{\otimes d} - \mathbb{E}\, Y^{\otimes d}\right\|_2 \lesssim 4^d \sqrt{\frac{\log\log k}{\log k}}.$$

Moreover, if the probability space has no atoms, then the partition can be made with exactly k sets, all of which have the same probability $1/k$.

Remark 3.13. A similar bound can be deduced for the higher-order version of covariance matrix, $\Sigma_X^{(d)} := \mathbb{E}(X - \mathbb{E}\, X)^{\otimes d}$. Indeed, applying Theorem 1.2 and Theorem 3.10 for $X - \mathbb{E}\, X$ instead of X (and so for $Y - \mathbb{E}\, Y$ instead of Y), we conclude that

$$\left\|\Sigma_X^{(d)} - \Sigma_Y^{(d)}\right\|_2 \le 8^d \left\|\Sigma_X^{(2)} - \Sigma_Y^{(2)}\right\|_2 \lesssim 8^d \sqrt{\frac{\log\log k}{\log k}}.$$

(The extra 2^d factor appears because from $\|X\|_2 \le 1$ we can only conclude that $\|X - \mathbb{E}\, X\|_2 \le 2$, so the bound needs to be normalized accordingly.)

3.8 Optimality

The following result shows that the rate in Theorem 1.2 is in general optimal up to a $\sqrt{\log\log k}$ factor.

Proposition 3.14 (Optimality). *Let $p > 16\ln(2k)$. Then there exists a random vector X in $\mathbb{R}^p$ such that $\|X\|_2 \le 1$ a.s. and for any sigma-algebra $\mathcal{F}$ generated by a partition of a sample space into at most k sets, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies*

$$\|\Sigma_X - \Sigma_Y\|_2 \ge \frac{1}{80\sqrt{\ln(2k)}}.$$

We will make X uniformly distributed on a well-separated subset of the Boolean cube $p^{-1/2}\{0,1\}^p$ of cardinality $n = 2k$. The following well known lemma states that such a subset exists:

Lemma 3.15 (A separated subset). *Let $p > 16\ln n$. Then there exist points $x_1, \dots, x_n \in p^{-1/2}\{0,1\}^p$ such that*

$$\left\|x_i - x_j\right\|_2 > \frac{1}{2} \quad \text{for all distinct } i, j \in [n].$$

Proof. Let X and X' be independent random vectors uniformly distributed on $\{0,1\}^p$. Then $\left\|X - X'\right\|_2^2 = \sum_{r=1}^p (X(r) - X'(r))^2$ is a sum of i.i.d. Bernoulli random variables with parameter $1/2$. Then Hoeffding's inequality [11] yields

$$\mathbb{P}\left\{\left\|X - X'\right\|_2^2 \le p/4\right\} \le e^{-p/8}.$$

Let $X_1, \dots, X_n$ be independent random vectors uniformly distributed on $\{0,1\}^p$. Applying the above inequality for each pair of them and then taking the union bound, we conclude that

$$\mathbb{P}\left\{\exists i, j \in [n] \text{ distinct } : \left\|X_i - X_j\right\|_2^2 \le p/4\right\} \le n^2 e^{-p/8} < 1$$

due to the condition on n. Therefore, there exists a realization of these random vectors that satisfies

$$\big\|X_i - X_j\big\|_2 > \frac{\sqrt{p}}{2} \quad \text{for all distinct } i, j \in [n].$$

Divide both sides by $\sqrt{p}$ to complete the proof. □

We will also need a high-dimensional version of the identity $\operatorname{Var}(X) = \frac{1}{2}\,\mathbb{E}(X - X')^2$ where X and X' are independent and identically distributed random variables. The following generalization is straightforward:

Lemma 3.16. *Let X and X' be independent and identically distributed random vectors taking values in $\mathbb{R}^p$. Then*

$$\mathbb{E}\|X - \mathbb{E}\,X\|_2^2 = \frac{1}{2}\,\mathbb{E}\big\|X - X'\big\|_2^2.$$

Proof of Proposition 3.14. Let $n = 2k$. Consider the sample space $[n]$ equipped with uniform probability and the sigma-algebra that consists of all subsets of $[n]$. Define the random variable X by

$$X(i) = x_i, \quad i \in [n]$$

where $\{x_1, \dots, x_n\}$ is the $(1/2)$-separated subset of $p^{-1/2}\{0,1\}^p$ from Lemma 3.15. Hence, X is uniformly distributed on the set $\{x_1, \dots, x_n\}$.

Now, if $\mathcal{F}$ is the sigma-algebra generated by a partition $\{F_1, \dots, F_{k_0}\}$ of $[n]$ with $k_0 \le k$, then

$$\begin{aligned}
\sqrt{p}\|\Sigma_X - \Sigma_Y\|_2 &\ge \operatorname{tr}(\Sigma_X - \Sigma_Y) \\
&= \operatorname{tr}\mathbb{E}(X - Y)(X - Y)^{\mathsf{T}} \quad \text{(by Lemma 3.1 again)} \\
&= \mathbb{E}\operatorname{tr}(X - Y)(X - Y)^{\mathsf{T}} = \mathbb{E}\|X - Y\|_2^2 \\
&= \mathbb{E}\left[\mathbb{E}_{\mathcal{F}}\|X - \mathbb{E}_{\mathcal{F}}\,X\|_2^2\right] \quad \text{(where } \mathbb{E}_{\mathcal{F}} \text{ denotes conditional expectation)} \\
&= \sum_{j=1}^{k_0} \mathbb{P}(F_j)\ \mathbb{E}\big\|X_j - \mathbb{E}\,X_j\big\|_2^2
\end{aligned}$$

where the random variable X_j is uniformly distributed on the set $\{x_i\}_{i \in F_j}$.

$$= \frac{1}{2}\sum_{j=1}^{k_0} \mathbb{P}(F_j)\ \mathbb{E}\big\|X_j - X_j'\big\|_2^2$$

where X_j' is an independent copy of X_j, by Lemma 3.16. Since the X_j and X_j' are independent and uniformly distributed on the set of $\big|F_j\big|$ points, $\big\|X_j - X_j'\big\|_2$ can either be zero (if both random vectors hit the same point, which happens with probability $1/\big|F_j\big|$) or it is greater than $1/2$ by separation. Hence

$$\mathbb{E}\big\|X_j - X_j'\big\|_2^2 \ge \frac{1}{4}\Big(1 - \frac{1}{\big|F_j\big|}\Big).$$

Moreover, $\mathbb{P}(F_j) = \big|F_j\big|/n$, so substituting in the bound above yields

$$\sqrt{p}\|\Sigma_X - \Sigma_Y\|_2 \ge \frac{1}{2}\sum_{j=1}^{k_0} \frac{\big|F_j\big|}{n} \cdot \frac{1}{4}\Big(1 - \frac{1}{\big|F_j\big|}\Big) = \frac{1}{8n}(n - k_0) \ge \frac{1}{16},$$

where we used that the sets F_j form a partition of $[n]$ so their cardinalities sum to n, our choice of $n = 2k$ and the fact that $k_0 \le k$.

We proved that

$$\|\Sigma_X - \Sigma_Y\|_2 \ge \frac{1}{16\sqrt{p}}.$$

If $p \le 25 \ln n$, this quantity is further bounded below by $1/(80\sqrt{\ln n}) = 1/(80\sqrt{\ln(2k)})$, completing the proof in this range. For larger p, the result follows by appending enough zeros to X and thus embedding it into higher dimension. Such embedding obviously does not change $\|\Sigma_X - \Sigma_Y\|_2$. □

4 Anonymity

In this section, we use our results on the covariance loss to make anonymous and accurate synthetic data by microaggregation. To this end, we can interpret microaggregation probabilistically as conditional expectation (Section 4.1) and deduce a general result on anonymous microaggregation (Theorem 4.1). We then show how to make synthetic data with custom size by bootstrapping (Section 4.2) and Boolean synthetic data by randomized rounding (Section 4.3).

4.1 Microaggregation as conditional expectation

For discrete probability distributions, conditional expectation can be interpreted as microaggregation, or local averaging.

Consider a finite sequence of points $x_1, \dots, x_n \in \mathbb{R}^p$, which we can think of as true data. Define the random variable X on the sample space $[n]$ equipped with the uniform probability distribution by setting

$$X(i) = x_i, \quad i \in [n].$$

Now, if $\mathcal{F} = \sigma(I_1, \dots, I_k)$ is the sigma-algebra generated by some partition $[n] = I_1 \cup \dots \cup I_k$, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ must take a constant value on each set I_j, and that value is the average of X on that set. In other words, Y takes values y_j with probability w_j, where

$$w_j = \frac{|I_j|}{n}, \quad y_j = \frac{1}{|I_j|} \sum_{i \in I_j} x_i, \quad j = 1, \dots, k. \tag{4.1}$$

The law of total expectation $\mathbb{E}\, X = \mathbb{E}\, Y$ in our case states that

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{j=1}^{k} w_j y_j. \tag{4.2}$$

Higher moments are handled using Corollary 3.12. This way, we obtain an effective anonymous algorithm that creates synthetic data while accurately preserving most marginals:

Theorem 4.1 (Anonymous microaggregation)**.** *Suppose k divides n. There exists a (deterministic) algorithm that takes input data $x_1, \dots, x_n \in \mathbb{R}^p$ such that $\|x_i\|_2 \le 1$ for all i, and computes a partition $[n] = I_1 \cup \dots \cup I_k$ with $|I_j| = n/k$ for all j, such that the microaggregated vectors*

$$y_j = \frac{k}{n} \sum_{i \in I_j} x_i, \quad j = 1, \dots, k,$$

satisfy for all $d \in \mathbb{N}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} x_i^{\otimes d} - \frac{1}{k} \sum_{j=1}^{k} y_j^{\otimes d} \right\|_2 \lesssim 4^d \sqrt{\frac{\log \log k}{\log k}}.$$

The algorithm runs in time polynomial in p *and* n*, and is independent of* d*.*

Proof. Most of the statement follows straightforwardly from Corollary 3.12 in light of the discussion above. However, the "moreover" part of Corollary 3.12 requires the probability space to be atomless, while our probability space $[n]$ does have atoms. Nevertheless, if the sample space consists of n atoms of probability $1/n$ each, and k divides n, then it is obvious that the divide-and-merge argument explained in Section 3.6 works, and so the "moreover" part of Corollary 3.12 also holds in this case. Thus, we obtain the (n/k)-anonymity from the microaggregation procedure. It is also clear that the algorithm (which is independent of d) runs in time polynomial in p and n. See the Microaggregation part of Algorithm 1.

□

Remark 4.2. The requirement that k divides n appearing in Theorem 4.1 as well as in other theorems makes it possible to partition $[n]$ into k sets of *exactly* the same cardinality. While convenient for proof purposes, this assumption is not strictly necessary. One can drop this assumption and make one set slightly larger than others. The corresponding modifications are left to the reader.

The use of spectral projection in combination with microaggregation has also been proposed in [25], although without any theoretical analysis regarding privacy or utility.

4.2 Synthetic data with custom size: bootstrapping

A seeming drawback of Theorem 4.1 is that the anonymity strength n/k and the cardinality k of the output data $y_1, \dots, y_k$ are tied to each other. To produce synthetic data of arbitrary size, we can use the classical technique of *bootstrapping*, which consists of sampling new data $u_1, \dots, u_m$ from the data $y_1, \dots, y_k$ independently and with replacement. The following general lemma establishes the accuracy of resampling:

Lemma 4.3 (Bootstrapping)**.** *Let* Y *be a random vector in* $\mathbb{R}^p$ *such that* $\|Y\|_2 \le 1$ *a.s. Let* $Y_1, \dots, Y_m$ *be independent copies of* Y*. Then for all* $d \in \mathbb{N}$ *we have*

$$\mathbb{E} \left\| \frac{1}{m} \sum_{i=1}^{m} Y_i^{\otimes d} - \mathbb{E}\, Y^{\otimes d} \right\|_2^2 \le \frac{1}{m}.$$

Proof. We have

$$\begin{aligned}
\mathbb{E} \left\| \frac{1}{m} \sum_{i=1}^{m} Y_i^{\otimes d} - \mathbb{E}\, Y^{\otimes d} \right\|_2^2 &= \frac{1}{m^2} \sum_{i=1}^{m} \mathbb{E} \left\| Y_i^{\otimes d} - \mathbb{E}\, Y^{\otimes d} \right\|_2^2 \quad \text{(by independence and zero mean)} \\
&= \frac{1}{m}\, \mathbb{E} \left\| Y^{\otimes d} - \mathbb{E}\, Y^{\otimes d} \right\|_2^2 \quad \text{(by identical distribution)} \\
&= \frac{1}{m} \left(\mathbb{E} \left\| Y^{\otimes d} \right\|_2^2 - \left\| \mathbb{E}\, Y^{\otimes d} \right\|_2^2 \right) \le \frac{1}{m}\, \mathbb{E} \left\| Y^{\otimes d} \right\|_2^2 = \frac{1}{m}\, \mathbb{E} \|Y\|_2^{2d}.
\end{aligned}$$

Using the assumption $\|Y\|_2 \le 1$ a.s., we complete the proof. □

Going back to the data $y_1, \dots, y_k$ produced by Theorem 4.1, let us consider a random vector Y that takes values y_j with probability $1/k$ each. Then obviously $\mathbb{E}\, Y^{\otimes d} = \frac{1}{k}\sum_{j=1}^k y_j^{\otimes d}$. Moreover, the assumption that $\|x_i\|_2 \le 1$ for all i implies that $\|y_j\|_2 \le 1$ for all j, so we have $\|Y\|_2 \le 1$ as required in Bootstrapping Lemma 4.3. Applying this lemma, we get

$$\mathbb{E}\left\| \frac{1}{m}\sum_{i=1}^m Y_i^{\otimes d} - \frac{1}{k}\sum_{j=1}^k y_j^{\otimes d} \right\|_2^2 \le \frac{1}{m}.$$

Combining this with the bound in Theorem 4.1, we obtain:

Theorem 4.4 (Anonymous microaggregation: custom data size)**.** *Suppose k divides n. Let $m \in \mathbb{N}$. There exists a randomized (n/k)-anonymous microaggregation algorithm that transforms input data $x_1, \dots, x_n \in B_2^p$ to the output data $u_1, \dots, u_m \in B_2^p$ in such a way that for all $d \in \mathbb{N}$,*

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m}\sum_{i=1}^m u_i^{\otimes d} \right\|_2^2 \lesssim 16^d \frac{\log\log k}{\log k} + \frac{1}{m}.$$

The algorithm consists of anonymous averaging (described in Theorem 4.1) followed by bootstrapping (described above). It runs in time polynomial in p and n, and is independent of d.

Remark 4.5 (Convexity)**.** Microaggregation respects convexity. If the input data $x_1, \dots, x_n$ lies in some given convex set K, the output data $u_1, \dots, u_m$ will lie in K, too. This can be useful in applications where one often needs to preserve some natural constraints on the data, such as positivity.

4.3 Boolean data: randomized rounding

Let us now specialize to Boolean data. Suppose the input data $x_1, \dots, x_n$ is taken from $\{0,1\}^p$. We can use Theorem 4.4 (and obvious renormalization by the factor $\|x_i\|_2 = \sqrt{p}$) to make (n/k)-anonymous synthetic data $u_1, \dots, u_m$ that satisfies

$$\mathbb{E}\, p^{-d}\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m}\sum_{i=1}^k u_i^{\otimes d} \right\|_2^2 \lesssim 16^d \frac{\log\log k}{\log k} + \frac{1}{m}. \tag{4.3}$$

According to Remark 4.5, the output data $u_1, \dots, u_m$ lies in the cube $K = [0,1]^p$. In order to transform the vectors u_i into Boolean vectors, i.e. points in $\{0,1\}^p$, we can apply the known technique of *randomized rounding* [28]. We define the randomized rounding of a number $x \in [0,1]$ as a random variable $r(x) \sim \mathrm{Ber}(x)$. Thus, to compute $r(x)$, we flip a coin that comes up heads with probability x and output 1 for a head and 0 for a tail. It is convenient to think of $r : [0,1] \to \{0,1\}$ as a random function. The randomized rounding $r(x)$ of a vector $x \in [0,1]^p$ is obtained by randomized rounding on each of the p coordinates of x independently.

Theorem 4.6 (Anonymous synthetic Boolean data)**.** *Suppose k divides n. There exists a randomized (n/k)-anonymous microaggregation algorithm that transforms input data $x_1, \dots, x_n \in \{0,1\}^p$ into output data $z_1, \dots, z_m \in \{0,1\}^p$ in such a way that the error $E = \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m}\sum_{i=1}^m z_i^{\otimes d}$ satisfies*

$$\mathbb{E}\binom{p}{d}^{-1} \sum_{1 \le i_1 < \dots < i_d \le p} E(i_1, \dots, i_d)^2 \lesssim 32^d \Big(\frac{\log\log k}{\log k} + \frac{1}{m} \Big)$$

for all $d \le p/2$. The algorithm consists of anonymous averaging and bootstrapping (as in Theorem 4.4) followed by independent randomized rounding of all coordinates of all points. It runs in time polynomial in p, n and linear in m, and is independent of d.

For convenience of the reader, Algorithm 1 below gives a pseudocode description of the algorithm described in Theorem 4.6.

Algorithm 1 Boolean n/k-anonymous synthetic data via microaggregation

Input: a sequence of points $x_1, \dots, x_n$ in the cube $\{0,1\}^p$ (true data); $k \ge 9$, where k divides n; $m \in \mathbb{N}$ (number of points in the synthetic data).

Microaggregation

1. Compute the second-moment matrix $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T}$.
2. Let $k' = \lfloor \sqrt{k} \rfloor$. Let $t \coloneqq \left\lfloor \frac{\log k'}{\log(7/\alpha)} \right\rfloor$ and $\alpha \coloneqq \left(\frac{\log\log k'}{\log k'}\right)^{1/4}$.
3. Let $P : \mathbb{R}^d \to \mathbb{R}^d$ be the orthogonal projection onto the span of the eigenvectors associated with the t largest eigenvalues of S.
4. Choose an α-covering $\nu_1, \dots, \nu_s \in \mathbb{R}^p$ of the unit Euclidean ball of the subspace $\mathrm{ran}(P)$. This is done by enumerating $B_2^t \cap (\alpha/\sqrt{t})\mathbb{Z}^t$ and mapping it into $\mathrm{ran}(P)$ using any linear isometry.
5. Construct a nearest-point partition $[n] = F_1 \cup \dots \cup F_s$ for $Px_1, \dots, Px_n$ with respect to $\nu_1, \dots, \nu_s$ as follows. For each $\ell \in [n]$, choose a point ν_j nearest to x_ℓ in the ℓ^2 metric and put ℓ into F_j. Break any ties arbitrarily.
6. Transform the partition $[n] = F_1 \cup \dots \cup F_s$ into the equipartition $[n] = I_1 \cup \dots \cup I_k$ with $|I_j| = \frac{n}{k}\ \forall j$ following the steps in Section 3.6: Divide each non-empty set F_i into subsets with probability $1/k$ each using division with residual, then merge all residuals into one new residual subset and divide the residual subset into further subsets of probability $1/k$ each.
7. Perform microaggregation: compute $y_j = \frac{k}{n}\sum_{i \in I_j} x_i, \quad j = 1, \dots, k$.

Bootstrapping creates new data $u_1, \dots, u_m$ by sampling (independently and with replacement) m points from the data $y_1, \dots, y_k$.

Randomized rounding maps the data $\{u_\ell\}_{\ell=1}^m \in [0,1]^p$ to data $\{z_j\}_{j=1}^m \in \{0,1\}^p$.

Output: a sequence of points $z_1, \dots, z_m$ in the cube $\{0,1\}^p$ (synthetic data) that satisfy the properties outlined in Theorem 4.6.

To prove Theorem 4.6, first note:[5]

Lemma 4.7 (Randomized rounding is unbiased)**.** *For any $x \in [0,1]^p$ and $d \in \mathbb{N}$, all off-diagonal entries of the tensors $\mathbb{E}\, r(x)^{\otimes d}$ and $x^{\otimes d}$ match:*

$$P_{\mathrm{off}}\big(\mathbb{E}\, r(x)^{\otimes d} - x^{\otimes d}\big) = 0,$$

where P_{off} is the orthogonal projection onto the subspace of tensors supported on the off-diagonal entries.

Proof. For any tuple of distinct indices $i_1, \dots, i_d \in [p]$, the definition of randomized rounding implies that $r(x)_{i_1}, \dots, r(x)_{i_d}$ are independent $\mathrm{Ber}(x_{i_1}), \dots, \mathrm{Ber}(x_{i_d})$ random variables. Thus

$$\mathbb{E}\, r(x)_{i_1} \cdots r(x)_{i_d} = x_{i_1} \cdots x_{i_d},$$

[5] This may be a good time for the reader to refer to Section 2.2 for definitions of restriction operators on tensors.

completing the proof. □

Proof of Theorem 4.6. Condition on the data $u_1, \dots, u_m$ obtained in Theorem 4.4. The output data of our algorithm can be written as $z_i = r_i(u_i)$, where the index i in r_i indicates that we perform randomized rounding on each point u_i independently. Let us bound the error introduced by randomized rounding, which is

$$a \coloneqq \mathbb{E}\, p^{-d} \Big\| P_{\text{off}} \Big(\frac{1}{m} \sum_{i=1}^m z_i^{\otimes d} - \frac{1}{m} \sum_{i=1}^m u_i^{\otimes d} \Big) \Big\|_2^2 = \frac{p^{-d}}{m^2}\, \mathbb{E} \Big\| \sum_{i=1}^m Z_i \Big\|_2^2$$

where $Z_i \coloneqq P_{\text{off}} \big(r_i(u_i)^{\otimes d} - u_i^{\otimes d} \big)$ are independent mean zero random variables due to Lemma 4.7. Therefore,

$$a = \frac{p^{-d}}{m^2} \sum_{i=1}^m \mathbb{E} \| Z_i \|_2^2 \,.$$

Since the variance is bounded by the second moment, we have

$$\mathbb{E} \| Z_i \|_2^2 \le \mathbb{E} \Big\| P_{\text{off}} \big(r_i(u_i)^{\otimes d} \big) \Big\|_2^2 \le \mathbb{E} \Big\| r_i(u_i)^{\otimes d} \Big\|_2^2 = \mathbb{E} \big\| r_i(u_i) \big\|_2^{2d} \le p^d$$

since $r_i(u_i) \in \{0,1\}^p$. Hence

$$a \le \frac{1}{m}.$$

Lifting the conditional expectation (i.e. taking expectation with respect to $u_1, \dots, u_m$) and combining this with (4.3) via triangle inequality, we obtain

$$\mathbb{E}\, p^{-d} \Big\| P_{\text{off}} \Big(\frac{1}{n} \sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m} \sum_{i=1}^m z_i^{\otimes d} \Big) \Big\|_2^2 \lesssim 16^d\, \frac{\log\log k}{\log k} + \frac{2}{m}.$$

Finally, we can replace the off-diagonal norm by the symmetric norm using Lemma 2.1. If $p \ge 2d$, it yields

$$\mathbb{E} \binom{p}{d}^{-1} \Big\| P_{\text{sym}} \Big(\frac{1}{n} \sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m} \sum_{i=1}^m z_i^{\otimes d} \Big) \Big\|_2^2 \lesssim 2^d \Big(16^d\, \frac{\log\log k}{\log k} + \frac{2}{m} \Big).$$

In view of (2.2), the proof is complete. □

5 Differential Privacy

Here we pass from anonymity to differential privacy by noisy microaggregation. In Section 5.1, we construct a "private" version of the PCA projection using repeated applications of the "exponential mechanism" [14]. This "private" projection is needed to make the PCA step in Algorithm 1 in Section 4.3 differentially private. In Sections 5.2–5.6, we show that the microaggregation is sufficiently stable with respect to additive noise, as long as we damp small blocks F_j (Section 5.4) and project the weights w_j and the vectors y_j back to the unit simplex and the convex set K, respectively (Section 5.5). We then establish differential privacy in Section 5.7 and accuracy in Section 5.6, with Theorem 5.13 being the most general result on private synthetic data. Just like we did for anonymity, we then show how to make synthetic data with custom size by bootstrapping (Section 5.9) and Boolean synthetic data by randomized rounding (Section 5.10).

5.1 Differentially private projection

If A is a self-adjoint linear transformation on a real inner product space, then the ith largest eigenvalue of A is denoted by $\lambda_i(A)$; the spectral norm of A is denoted by $\|A\|$; and the Frobenius norm of A is denoted by $\|A\|_2$. If $v_1, \ldots, v_t \in \mathbb{R}^p$ then $P_{v_1,\ldots,v_t}$ denotes the orthogonal projection from $\mathbb{R}^p$ onto $\mathrm{span}\{v_1, \ldots, v_t\}$. In particular, if $v \in \mathbb{R}^p$ then P_v denotes the orthogonal projection from $\mathbb{R}^p$ onto $\mathrm{span}\{v\}$.

In this section, we construct for any given $p \times p$ positive semidefinite A and $1 \leq t \leq p$, a random projection P that behaves like the projection onto the t leading eigenvectors of A and, at the same time, is "differentially private" in the sense that if A is perturbed a little, the distribution of P changes a little. Something like this is done [14]. However, in [14], a PCA approximation of A is produced in the output rather than the projection. The error in the operator norm for this approximation is estimated in [14], whereas in this paper, we need to estimate the error in the Frobenius norm.

Thus, we will do a modification of the construction in [14]. But the general idea is the same: first construct a vector that behaves like the principal eigenvector (i.e., 1-dimensional PCA) and, at the same time, is "differentially private." Repeatedly doing this procedure gives a "differentially private" version of the t-dimensional PCA projection.

The following algorithm is referred to as the "exponential mechanism" in [14]. As shown in Lemma 5.1 below, this algorithm outputs a random vector that behaves like the principal eigenvector (see part 1) and is "differentially private" in the sense of part 3.

Algorithm 2 PVEC(A)

Input: positive semidefinite linear transformation $A : V \to V$, where V is a finite dimensional real inner product space.

Output: x sampled from the unit sphere of V according to the density proportional to $e^{\langle Ax,x\rangle}$.

Lemma 5.1 ([14])**.** *Suppose that A is a positive semidefinite linear transformation on a finite dimensional vector space V.*

(1) If v is an output of PVEC(A)*, then*

$$\mathbb{E}\langle Av, v\rangle \geq (1-\gamma)\lambda_1(A)$$

for all $\gamma > 0$ such that $\lambda_1(A) \geq C \dim(V)\frac{1}{\gamma}\log(\frac{1}{\gamma})$, where $C > 0$ is an absolute constant.

(2) PVEC(A) *can be implemented in time* $\mathrm{poly}(\dim V, \lambda_1(A))$.

(3) Let $B : V \to V$ be a positive semidefinite linear transformation. If $\|A - B\| \leq \beta$, then

$$\mathbb{P}\left\{\mathrm{PVEC}(A) \in \mathcal{S}\right\} \leq e^{\beta} \cdot \mathbb{P}\left\{\mathrm{PVEC}(B) \in \mathcal{S}\right\}$$

for any measurable subset $\mathcal{S}$ of V.

Let us restate part 1 of Lemma 5.1 more conveniently:

Lemma 5.2. *Suppose that A is a positive semidefinite linear transformations on a finite dimensional vector space V. If v is an output of* PVEC(A)*, then*

$$\lambda_1(A)^2 - \mathbb{E}\langle Av, v\rangle^2 \leq 2\gamma\lambda_1(A)^2 + C\dim(V)^2\frac{1}{\gamma^2}\log^2\left(\frac{1}{\gamma}\right)$$

for all $\gamma > 0$, where $C > 0$ is an absolute constant.

Proof. Fix $\gamma > 0$, and let us consider two cases.

If $\lambda_1(A) \geq C \dim(V) \frac{1}{\gamma} \log(\frac{1}{\gamma})$, then by part 1 of Lemma 5.1, we have $\lambda_1(A) - \mathbb{E}\langle Av, v\rangle \leq \gamma \lambda_1(A)$. Therefore, keeping in mind that the inequality $\langle Av, v\rangle \leq \lambda_1(A)$ always hold, we obtain $\lambda_1(A)^2 - \mathbb{E}\langle Av, v\rangle^2 = \mathbb{E}\left[(\lambda_1(A) + \langle Av, v\rangle)(\lambda_1(A) - \langle Av, v\rangle)\right] \leq 2\gamma\lambda_1(A)^2$.

If $\lambda_1(A) < C \dim(V) \frac{1}{\gamma} \log(\frac{1}{\gamma})$, then $\lambda_1(A)^2 - \mathbb{E}\langle Av, v\rangle^2 \leq \lambda_1(A)^2 \leq C^2 (\dim V)^2 \frac{1}{\gamma^2} \log^2(\frac{1}{\gamma})$. The proof is complete. □

We now construct a "differentially private" version of the t-dimensional PCA projection. This is done by repeated applications of PVEC in Algorithm 2.

Algorithm 3 PROJ(A, t)

Input: $p \times p$ positive semidefinite real matrix A; and $1 \leq t \leq p$
Apply PVEC(A) to obtain $v_1 \in \mathbb{R}^p$ with $\|v_1\|_2 = 1$
for $i = 1, \ldots, t-1$,
consider the linear transformation $A_i = (I - P_{v_1,\ldots,v_i}) A (I - P_{v_1,\ldots,v_i})$ on the space $V_i = \text{ran}(I - P_{v_1,\ldots,v_i})$;
Apply PVEC(A_i) to obtain $v_{i+1} \in V_i$ with $\|v_{i+1}\|_2 = 1$.
end for
Output: Orthogonal projection $P_{v_1,\ldots,v_t}$ on $\mathbb{R}^p$.

The following lemma shows that the algorithm PROJ is "differentially private" in the sense of part 3 of Lemma 5.1, except that e^{β} is replaced by $e^{t\beta}$.

Lemma 5.3. *Suppose that A and B are $p \times p$ positive semidefinite matrices and $1 \leq t \leq p$. If $\|A - B\| \leq \beta$ then*

$$\mathbb{P}\left\{\text{PROJ}(A, t) \in \mathcal{S}\right\} \leq e^{t\beta} \cdot \mathbb{P}\left\{\text{PROJ}(B, t) \in \mathcal{S}\right\}$$

for any measurable subset $\mathcal{S}$ of $\mathbb{R}^{p \times p}$.

Proof. Fix β. We first define a notion of privacy similar to the one in [14]. A randomized algorithm $\mathcal{M}$ with input being a $p \times p$ positive semidefinite real matrix A is θ-DP if whenever $\|A - B\| \leq \beta$, we have $\mathbb{P}\{\mathcal{M}(A) \in \mathcal{S}\} \leq e^{\theta} \mathbb{P}\{\mathcal{M}(B) \in \mathcal{S}\}$ for all measurable subset $\mathcal{S}$ of $\mathbb{R}^{p \times p}$.

In the algorithm PROJ, the computation of v_1 as an algorithm is β-DP by Lemma 5.1(3).

Similarly, if we fix v_1, the computation of v_2 as an algorithm is also β-DP. So by some version of Lemma 2.5, the computation of (v_1, v_2) as an algorithm (without fixing v_1) is 2β-DP.

And so on. By induction, we have that the computation of $(v_1, \ldots, v_t)$ as an algorithm is $t\beta$-DP. Thus, PROJ$(\cdot, t)$ is $t\beta$-DP. The result follows. □

Next, we show that the output of the algorithm PROJ behaves like the t-dimensional PCA projection in the sense of Lemma 5.5 below. Observe that if P is the projection onto the t leading eigenvectors of a $p \times p$ positive semidefinite matrix A, then $\|(I - P)A(I - P)\|_2^2 = \sum_{i=t+1}^{p} \lambda_i(A)^2$. To prove Lemma 5.5, we first prove the following lemma and then we apply this lemma repeatedly to obtain Lemma 5.5.

Lemma 5.4. *Let A be a $p \times p$ positive semidefinite matrix. Let $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$. Then*

$$\sum_{i=j}^{p} \lambda_i((I - P_v)A(I - P_v))^2 \leq \sum_{i=j+1}^{p} \lambda_i(A)^2 + \lambda_1(A)^2 - \langle Av, v\rangle^2,$$

for every $1 \leq j \leq p$.

Proof. For every $p \times p$ real symmetric matrix B and every $1 \le i \le p$, we have

$$\lambda_i(B) = \inf_{\dim W = p-i+1} \sup_{x \in W,\, \|x\|_2 = 1} \langle Bx, x \rangle,$$

where the infimum is over all subspaces W of $\mathbb{R}^p$ with dimension $p - i + 1$. Thus, since P_v is a rank-one orthogonal projection,

$$\begin{aligned} \lambda_i((I - P_v)A(I - P_v)) &= \inf_{\dim W = p-i+1} \sup_{x \in W,\, \|x\|_2 = 1} \langle A(I - P_v)x, (I - P_v)x \rangle \\ &\ge \inf_{\dim W = p-i+1} \sup_{x \in W \cap \mathrm{ran}(I - P_v),\, \|x\|_2 = 1} \langle Ax, x \rangle \ge \lambda_{i+1}(A), \end{aligned}$$

for every $1 \le i \le p - 1$. Thus,

$$\sum_{i=1}^{j-1} \lambda_i((I - P_v)A(I - P_v))^2 \ge \sum_{i=2}^{j} \lambda_i(A)^2,$$

so

$$\sum_{i=j}^{p} \lambda_i((I - P_v)A(I - P_v))^2 \le \sum_{i=j+1}^{p} \lambda_i(A)^2 + \lambda_1(A)^2 + \|(I - P_v)A(I - P_v)\|_2^2 - \|A\|_2^2.$$

Since $\|A\|_2^2 - \|(I - P_v)A(I - P_v)\|_2^2 \ge \|P_v A P_v\|_2^2 = \langle Av, v \rangle^2$, the result follows. □

Lemma 5.5. *Suppose that A is a $p \times p$ positive semidefinite matrix and $1 \le t \le p$. If P is an output of* PROJ(A, t)*, then*

$$\mathbb{E}\|(I - P)A(I - P)\|_2^2 \le \sum_{i=t+1}^{p} \lambda_i(A)^2 + 2t\gamma\|A\|^2 + Ct\frac{p^2}{\gamma^2} \log^2\left(\frac{1}{\gamma}\right),$$

for all $\gamma > 0$, where $C > 0$ is an absolute constant.

Proof. Let $v_1, \ldots, v_t$ be those vectors defined in the algorithm PROJ(A, t). Let $A_0 = A$. For $1 \le k \le t$, let $A_k = (I - P_{v_1,\ldots,v_k})A(I - P_{v_1,\ldots,v_k})$. Since v_{k+1} is an output of PVEC(A_k), by Lemma 5.2, we have

$$\lambda_1(A_k)^2 - \mathbb{E}_{v_{k+1}}(\langle A_k v_{k+1}, v_{k+1} \rangle^2) \le 2\gamma\lambda_1(A_k)^2 + C\frac{p^2}{\gamma^2} \log^2\left(\frac{1}{\gamma}\right),$$

for all $1 \le j \le p$ and $0 \le k \le t - 1$, where the expectation $\mathbb{E}_{v_{k+1}}$ is over v_{k+1} conditioning on $v_1, \ldots, v_k$. By Lemma 5.4, we have

$$\sum_{i=j}^{p} \lambda_i((I - P_{v_{k+1}})A_k(I - P_{v_{k+1}}))^2 \le \sum_{i=j+1}^{p} \lambda_i(A_k)^2 + \lambda_1(A_k)^2 - \langle A_k v_{k+1}, v_{k+1} \rangle^2,$$

for all $1 \le j \le p$ and $0 \le k \le t - 1$. Therefore,

$$\mathbb{E}_{v_{k+1}} \sum_{i=j}^{p} \lambda_i((I - P_{v_{k+1}})A_k(I - P_{v_{k+1}}))^2 \le \sum_{i=j+1}^{p} \lambda_i(A_k)^2 + 2\gamma\lambda_1(A_k)^2 + C\frac{p^2}{\gamma^2} \log^2\left(\frac{1}{\gamma}\right),$$

for all $1 \le j \le p$ and $0 \le k \le t-1$.

In the algorithm PROJ(A,t), each v_{k+1} is chosen from the unit sphere of $\mathrm{ran}(I - P_{v_1,\dots,v_k})$. Hence, the vectors $v_1,\dots,v_t$ are orthonormal, so $I - P_{v_1,\dots,v_{k+1}} = (I - P_{v_{k+1}})(I - P_{v_1,\dots,v_k})$ for all $1 \le k \le t-1$. Thus, $(I - P_{v_{k+1}})A_k(I - P_{v_{k+1}}) = A_{k+1}$. So we have

$$\mathbb{E}_{v_{k+1}} \sum_{i=j}^{p} \lambda_i(A_{k+1})^2 \le \sum_{i=j+1}^{p} \lambda_i(A_k)^2 + 2\gamma\lambda_1(A_k)^2 + C\frac{p^2}{\gamma^2}\log^2\left(\frac{1}{\gamma}\right),$$

for all $1 \le j \le p$ and $0 \le k \le t-1$. Taking the full expectation $\mathbb{E}$ on both sides, we get

$$\mathbb{E}\sum_{i=j}^{p} \lambda_i(A_{k+1})^2 \le \mathbb{E}\sum_{i=j+1}^{p} \lambda_i(A_k)^2 + 2\gamma\|A\|^2 + C\frac{p^2}{\gamma^2}\log^2\left(\frac{1}{\gamma}\right),$$

for all $1 \le j \le p$ and $0 \le k \le t-1$, where we used the fact that $\lambda_1(A_k) = \|A_k\| \le \|A\|$. Repeated applications of this inequality yields

$$\mathbb{E}\sum_{i=1}^{p} \lambda_i(A_t)^2 \le \sum_{i=t+1}^{p} \lambda_i(A_0)^2 + 2t\gamma\|A\|^2 + Ct\frac{p^2}{\gamma^2}\log^2\left(\frac{1}{\gamma}\right).$$

Note that $A_t = (I - P_{v_1,\dots,v_t})A(I - P_{v_1,\dots,v_t})$ and $P = P_{v_1,\dots,v_t}$ is the output of PROJ(A,t). Thus, the left hand side is equal to $\mathbb{E}\|A_t\|_2^2 = \mathbb{E}\|(I-P)A(I-P)\|_2^2$. The result follows. □

5.2 Microaggregation with more control

We will protect privacy by adding noise to the microaggregation mechanism. To make this happen, we will need a version of Theorem 4.1 with more control.

We adapt the microaggregation mechanism from (4.1) to the current setting. Given a partition $[n] = F_1 \cup \cdots \cup F_s$ (where some F_j could be empty), we define for $1 \le j \le s$ with F_j being non-empty,

$$w_j = \frac{|F_j|}{n}, \quad y_j = \frac{1}{|F_j|}\sum_{i \in F_j} x_i; \tag{5.1}$$

and when F_j is empty, set $w_j = 0$ and y_j to be an arbitrary point.

Theorem 5.6 (Microaggregation with more control)**.** *Let $x_1,\dots,x_n \in \mathbb{R}^p$ be such that $\|x_i\|_2 \le 1$ for all i. Let $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T}$. Let P be an orthogonal projection on $\mathbb{R}^p$. Let $\nu_1,\dots,\nu_s \in \mathbb{R}^p$ be an α-covering of the unit Euclidean ball of* $\mathrm{ran}(P)$*. Let $[n] = F_1 \cup \cdots \cup F_s$ be a nearest-point partition of (Px_i) with respect to $\nu_1,\dots,\nu_s$. Then the weights w_j and vectors y_j defined in (5.1) satisfy for all $d \in \mathbb{N}$:*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} x_i^{\otimes d} - \sum_{j=1}^{s} w_j y_j^{\otimes d} \right\|_2 \le 4^d\left(4\alpha^2 + \|(I-P)S(I-P)\|_2\right). \tag{5.2}$$

Proof of Theorem 5.6. We explained in Section 4.1 how to realize microaggregation probabilistically as conditional expectation. To reiterate, we consider the sample space $[n]$ equipped with the uniform probability distribution and define a random variable X on $[n]$ by setting $X(i) = x_i$ for $i = 1,\dots,n$. If $\mathcal{F} = \sigma(F_1,\dots,F_s)$ is the sigma-algebra generated by some partition $[n] = F_1 \cup \cdots \cup F_s$, the

conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ is a random vector that takes values y_j with probability w_j as defined in (5.1). Then the left hand side of (5.2) equals

$$\begin{aligned}\left\|\mathbb{E}\, X^{\otimes d} - \mathbb{E}\, Y^{\otimes d}\right\|_2 &\le 4^d \left\|\mathbb{E}\, XX^{\mathsf{T}} - \mathbb{E}\, YY^{\mathsf{T}}\right\|_2 \quad \text{(by the Higher Moment Theorem 3.10)} \\ &\le 4^d \Big(\mathbb{E}\|PX - PY\|_2^2 + \left\|(I-P)S(I-P)\right\|_2 \Big) \quad \text{(by Lemma 3.2)} \\ &\le 4^d \Big(4\alpha^2 + \left\|(I-P)S(I-P)\right\|_2 \Big) \quad \text{(by Lemma 3.6)}.\end{aligned}$$

□

5.3 Perturbing the weights and vectors

Theorem 5.6 makes the first step towards noisy microaggregation. Next, we will add noise to the weights (w_j) and vectors (y_j) obtained by microaggregation. To control the effect of such noise on the accuracy, the following two simple bounds will be useful.

Lemma 5.7. *Let $u, v \in \mathbb{R}^n$ be such that $\|u\|_2 \le 1$ and $\|v\|_2 \le 1$. Then, for every $d \in \mathbb{N}$,*

$$\left\|u^{\otimes d} - v^{\otimes d}\right\|_2 \le d\|u - v\|_2 \,.$$

Proof. For $d = 1$ the result is trivial. For $d \ge 2$, we can represent the difference as a telescopic sum

$$u^{\otimes d} - v^{\otimes d} = \sum_{k=0}^{d-1} \Big(u^{\otimes(d-k)} \otimes v^{\otimes k} - u^{\otimes(d-k-1)} \otimes v^{\otimes(k+1)} \Big) = \sum_{k=0}^{d-1} u^{\otimes(d-k-1)} \otimes (u - v) \otimes v^{\otimes k}.$$

Then, by triangle inequality,

$$\left\|u^{\otimes d} - v^{\otimes d}\right\|_2 \le \sum_{k=0}^{d-1} \|u\|_2^{d-k-1} \|u - v\|_2 \|v\|_2^k \le d\|u - v\|_2 \,,$$

where we used the assumption on the norms of u and v in the last step. The lemma is proved. □

Lemma 5.8. *Consider numbers $\lambda_j, \mu_j \in \mathbb{R}$ and vectors $u_j, v_j \in \mathbb{R}^p$ such that $\left\|u_j\right\|_2 \le 1$ and $\left\|v_j\right\|_2 \le 1$ for all $j = 1, \dots, m$. Then, for every $d \in \mathbb{N}$,*

$$\left\| \sum_j \left(\lambda_j u_j^{\otimes d} - \mu_j v_j^{\otimes d} \right) \right\|_2 \le d \sum_j |\lambda_j| \left\|u_j - v_j\right\|_2 + \sum_j |\lambda_j - \mu_j| \,. \tag{5.3}$$

Proof. Adding and subtracting the cross term $\sum_j \lambda_j v_j^{\otimes d}$ and using triangle inequality, we can bound the left side of (5.3) by

$$\left\| \sum_j \lambda_j \left(u_j^{\otimes d} - v_j^{\otimes d} \right) \right\|_2 + \left\| \sum_j (\lambda_j - \mu_j) v_j^{\otimes d} \right\|_2 \le \sum_j |\lambda_j| \left\|u_j^{\otimes d} - v_j^{\otimes d}\right\|_2 + \sum_j |\lambda_j - \mu_j| \left\|v_j^{\otimes d}\right\|_2 .$$

It remains to use Lemma 5.7 and note that $\left\|v_j^{\otimes d}\right\|_2 = \left\|v_j\right\|_2^d$. □

5.4 Damping

Although the microaggregation mechanism (5.1) is stable with respect to additive noise in the weights w_j or the vectors y_j as shown in Section 5.3, there are still two issues that need to be resolved.

The first issue is the potential instability of the microaggregation mechanism (5.1) for small blocks F_j. For example, if $\big|F_j\big| = 1$, the microaggregation does not do anything for that block and returns the original input vector $y_j = x_i$. To protect the privacy of such vector, a lot of noise is needed, which might be harmful to the accuracy.

One may wonder why can we not make all blocks F_j of the same size like we did in Theorem 4.1. Indeed, in Section 3.6 we showed how to transform a potentially imbalanced partition $[n] = F_1 \cup \cdots \cup F_s$ into an *equipartition* (where all F_j have the same cardinality) using a divide-an-merge procedure; could we not apply it here? Unfortunately, an equipartition might be too sensitive[6] to changes even in a single data point x_i. The original partition $F_1, \dots, F_s$, on the other hand, is sufficiently stable.

We resolve this issue by suppressing, or *damping*, the blocks F_j that are too small. Whenever the cardinality of F_j drops below a predefined level b, we divide by b rather than $\big|F_j\big|$ in (5.1). In other words, instead of vanilla microaggregation (5.1), we consider the following damped microaggregation:

$$w_j = \frac{\big|F_j\big|}{n}, \quad \tilde{y}_j = \frac{1}{\max\big(\big|F_j\big|, b\big)} \sum_{i \in F_j} x_i, \quad j = 1, \dots, s. \tag{5.4}$$

5.5 Metric projection

And here is the second issue. Recall that the numbers w_j returned by microaggregation (5.4) are probability weights: the weight vector $w = (w_j)_{j=1}^s$ belongs to the unit simplex

$$\Delta \coloneqq \Big\{ a = (a_1, \dots, a_s) : \sum_{i=1}^s a_i = 1;\ a_i \ge 0\ \forall i \Big\}.$$

This feature may be lost if we add noise to w_j. Similarly, if the input vectors x_j are taken from a given convex set K (for Boolean data, this is $K = [0,1]^p$), we would like the synthetic data to belong to K, too. Microaggregation mechanism (5.1) respects this feature: by convexity, the vectors y_j do belong to K. However, this property may be lost if we add noise to y_j.

We resolve this issue by projecting the perturbed weights and vectors back onto the unit simplex Δ and the convex set K, respectively. For this purpose, we utilize *metric projections* mappings that return a proximal point in a given set. Formally, we let

$$\pi_{\Delta,1}(w) \coloneqq \operatorname*{argmin}_{\bar{w} \in \Delta} \|\bar{w} - w\|_1\,; \quad \pi_{K,2}(y) \coloneqq \operatorname*{argmin}_{\bar{y} \in K} \|\bar{y} - y\|_2\,. \tag{5.5}$$

(If the minimum is not unique, break the tie arbitrarily. One valid choice of $\pi_{\Delta,1}(w)$ can be defined by setting all the negative entries of w to be 0 and then normalize it so that it is in Δ. In the case when all entries of w are negative, set $\pi_{\Delta,1}(w)$ to be any point in Δ.)

[6]The divide-and-merge procedure described in Section 3.6/Proof of Theorem 4.1 for producing the I blocks from the F blocks is sensitive to even a change in a single data point x_i. For example, suppose that one block $F_1 = \{x_1, \dots, x_n\}$ contains all the points and it is divided into $I_1 = \{x_1, \dots, x_{n/k}\}, \cdots, I_k = \{x_{n-n/k+1}, \dots, x_n\}$. If x_1 is changed to another point so that it becomes a new point in another block F_2, then the new I blocks could become $I_1 = \{x_2, \dots, x_{n/k+1}\}, \cdots, I_{k-1} = \{x_{n-2n/k+2}, \dots, x_{n-n/k+1}\}$, $I_k = \{x_{n-n/k+2}, \dots, x_n, x_1\}$ and so every I block is changed by two points.

Thus, here is our plan: given input data $(x_i)_{i=1}^s$, we apply damped microaggregation (5.4) to compute weights and vectors $(w_j, \tilde{y}_j)_{j=1}^s$, add noise, and project the noisy vectors back to the unit simplex Δ and the convex set K respectively. In other words, we compute

$$\bar{w} = \pi_{\Delta,1}\left(w + \rho\right), \quad \bar{y}_j = \pi_{K,2}\left(\tilde{y}_j + r_j\right), \tag{5.6}$$

where $\rho \in \mathbb{R}^s$ and $r_j \in \mathbb{R}^p$ are noise vectors (which we will set to be random Laplacian noise in the future).

5.6 The accuracy guarantee

Here is the accuracy guarantee of our procedure. This is a version of Theorem 5.6 with noise, damping, and metric projection:

Theorem 5.9 (Accuracy of damped, noisy microaggregation)**.** *Let K be a convex set in $\mathbb{R}^p$ that lies in the unit Euclidean ball B_2^p. Let $x_1, \dots, x_n \in K$. Let $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T}$. Let P be an orthogonal projection on $\mathbb{R}^p$. Let $\nu_1, \dots, \nu_s \in \mathbb{R}^p$ be an α-covering of the unit Euclidean ball of* $\operatorname{ran}(P)$*. Let $[n] = F_1 \cup \dots \cup F_s$ be a nearest-point partition of (Px_i) with respect to $\nu_1, \dots, \nu_s$. Then the weights $\bar{w}_j$ and vectors $\bar{y}_j$ defined in (5.6) satisfy for all $d \in \mathbb{N}$:*

$$\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \sum_{j=1}^s \bar{w}_j \bar{y}_j^{\otimes d} \right\|_2 \le 4^d \left(4\alpha^2 + \|(I-P)S(I-P)\|_2\right) + \frac{2dsb}{n} + 2\|\rho\|_1 + 2d\sum_{j=1}^s w_j \|r_j\|_2 . \tag{5.7}$$

Proof. Adding and subtracting the cross term $\sum_j w_j y_j^{\otimes d}$ and using triangle inequality, we can bound the left hand side of (5.7) by

$$\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \sum_{j=1}^s w_j y_j^{\otimes d} \right\|_2 + \left\| \sum_{j=1}^s \left(w_j y_j^{\otimes d} - \bar{w}_j \bar{y}_j^{\otimes d} \right) \right\|_2 \tag{5.8}$$

The first term can be bounded by Theorem 5.6. For the second term we can use Lemma 5.8 and note that

$$\|y_j\|_2 \le 1, \quad \|\bar{y}_j\|_2 \le 1 \quad \text{for all } j \in [s]. \tag{5.9}$$

Indeed, definition (5.1) of y_j and the assumption that x_i lie in the convex set K imply that $y_j \in K$. Also, definition (5.6) implies that $\bar{y}_j \in K$ as well. Now the bounds in (5.9) follow from the assumption that $K \subset B_2^p$. So, applying Theorem 5.6 and Lemma 5.8, we see that the quantity in (5.8) is bounded by

$$4^d \left(4\alpha^2 + \|(I-P)S(I-P)\|_2\right) + d\sum_{j=1}^s w_j \|y_j - \bar{y}_j\|_2 + \sum_{j=1}^s |w_j - \bar{w}_j| . \tag{5.10}$$

We bound the two sums in this expression separately.

Let us start with the sum involving y_j and $\bar{y}_j$. We will handle large and small blocks differently. For a large block, one for which $|F_j| \ge b$, by (5.4) we have $\tilde{y}_j = |F_j|^{-1} \sum_{i \in F_j} x_i = y_j \in K$. By definition (5.6), $\bar{y}_j$ is the closest point in K to $y_j + r_j$. Since $y_j \in K$, we have

$$\begin{aligned} \|y_j - \bar{y}_j\|_2 &\le \|y_j + r_j - \bar{y}_j\|_2 + \|r_j\|_2 \quad \text{(by triangle inequality)} \\ &\le \|y_j + r_j - y_j\|_2 + \|r_j\|_2 \quad \text{(by minimality property of } \bar{y}_j) \\ &= 2\|r_j\|_2 . \end{aligned}$$

Hence

$$\sum_{j\in[s]:\,|F_j|\geq b} w_j\|y_j-\bar{y}_j\|_2 \leq 2\sum_{j=1}^{s} w_j\|r_j\|_2\,.$$

Now let us handle small blocks. By (5.9) we have $\|y_j-\bar{y}_j\|_2\leq 2$, so

$$\sum_{j\in[s]:\,|F_j|<b} w_j\|y_j-\bar{y}_j\|_2 \leq \sum_{j\in[s]:\,|F_j|<b} \frac{|F_j|}{n}\cdot 2 \leq \frac{2sb}{n}.$$

Combining our bounds for large and small blocks, we conclude that

$$\sum_{j=1}^{s} w_j\|y_j-\bar{y}_j\|_2 \leq 2\sum_{j=1}^{s} w_j\|r_j\|_2 + \frac{2sb}{n}. \tag{5.11}$$

Finally, let us bound the last sum in (5.10). By definition (5.6), $\bar{w}$ is a closest point in the unit simplex Δ to $w+\rho$ in the ℓ^1 metric. Since $w\in\Delta$, we have

$$\begin{aligned}\sum_{j=1}^{s}|w_j-\bar{w}_j| &= \|w-\bar{w}\|_1\\ &\leq \|w+\rho-\bar{w}\|_1+\|\rho\|_1 \quad \text{(by triangle inequality)}\\ &\leq \|w+\rho-w\|_1+\|\rho\|_1 \quad \text{(by minimality property of } \bar{w})\\ &= 2\|\rho\|_1\,. \end{aligned}\tag{5.12}$$

Substitute (5.11) and (5.12) into (5.10) to complete the proof. □

5.7 Privacy

Now that we analyzed the accuracy of the synthetic data, we prove differential privacy. To that end, we will use Laplacian mechanism, so we need to bound the sensitivity of the microaggregation.

Lemma 5.10 (Sensitivity of damped microaggregation)**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^p$. Consider vectors $x_1,\dots,x_n\in\mathbb{R}^p$. Let I and I' be subsets of $[n]$ that differ in exactly one element. Then, for any $b>0$, we have*

$$\left\|\frac{1}{\max(|I|,b)}\sum_{i\in I}x_i-\frac{1}{\max(|I'|,b)}\sum_{i\in I'}x_i\right\|\leq\frac{2}{b}\max_{i\in[n]}\|x_i\|\,. \tag{5.13}$$

Proof. Without loss of generality, we can assume that $I'=I\setminus\{n_0\}$ for some $n_0\in I$.

Case 1: $|I|\geq b+1$

In this case, $|I'|=|I|-1\geq b$. Denoting by ξ the difference vector whose norm we are estimating in (5.13), we have

$$\xi=\frac{1}{|I|}\sum_{i\in I}x_i-\frac{1}{|I|-1}\sum_{i\in I\setminus\{n_0\}}x_i=\frac{1}{|I|\,(|I|-1)}\sum_{i\in I\setminus\{n_0\}}(x_{n_0}-x_i).$$

The sum in the right hand side consists of $|I|-1$ terms, each satisfying $\|x_{n_0}-x_i\|\leq 2\max_i\|x_i\|$. This yields $\|\xi\|\leq(2/|I|)\max_i\|x_i\|$. Since $|I|\geq b+1$ by assumption, we get even a better bound than we need in this case.

Case 2: $|I| \le b$

In this case, $\left|I'\right| = |I| - 1 < b$. Hence the difference vector of interest equals

$$\xi = \frac{1}{b}\sum_{i\in I} x_i - \frac{1}{b}\sum_{i\in I\setminus\{n_0\}} x_i = \frac{x_{n_0}}{b}.$$

Therefore, $\|\xi\| \le (1/b)\max_i\|x_i\|$. The lemma is proved. □

Lemma 5.11 (Sensitivity of damped microaggregation II)**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^p$. Let I be a subset of $[n]$ and let $n_0 \in I$. Consider vectors $x_1, \dots, x_n \in \mathbb{R}^p$ and $x'_1, \dots, x'_n \in \mathbb{R}^p$ such that $x_i = x'_i$ for all $i \ne n_0$. Then, for any $b > 0$, we have*

$$\left\| \frac{1}{\max(\left|I\right|, b)}\sum_{i\in I} x_i - \frac{1}{\max(\left|I\right|, b)}\sum_{i\in I} x'_i \right\| \le \frac{1}{b}\left\|x_{n_0} - x'_{n_0}\right\|.$$

Proof.

$$\left\| \frac{1}{\max(\left|I\right|, b)}\sum_{i\in I} x_i - \frac{1}{\max(\left|I\right|, b)}\sum_{i\in I} x'_i \right\| = \left\| \frac{1}{\max(\left|I\right|, b)}(x_{n_0} - x'_{n_0}) \right\| \le \frac{1}{b}\left\|x_{n_0} - x'_{n_0}\right\|.$$

□

Theorem 5.12 (Privacy)**.** *In the situation of Theorem 5.9, suppose that all coordinates of the vectors ρ and r_j are independent Laplacian random variables, namely*

$$\rho_i \sim \mathrm{Lap}\left(\frac{6}{n\varepsilon}\right) \text{ for } i \in [s]; \quad r_{ji} \sim \mathrm{Lap}\left(\frac{12\sqrt{p}}{b\varepsilon}\right) \text{ for } i \in [p],\ j \in [s],$$

and P is an output of $\mathrm{PROJ}(\frac{n\varepsilon}{6t}S, t)$ where $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$. Then the output data $(\bar{w}_j, \bar{y}_j)_{j=1}^s$ is ε-differentially private in the input data $(x_i)_{i=1}^n$.

Proof. First we check that the projection P is private. To do this, let us bound the sensitivity of the second moment matrix $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ in the spectral norm. Consider two input data $(x_i)_{i=1}^n$ and $(x'_i)_{i=1}^n$ that differ in exactly one element, i.e. $x_i = x'_i$ for all i except some $i = n_0$. Then the difference in the spectral norm of the corresponding matrices S and S' satisfy

$$\begin{aligned}
\left\|S - S'\right\| &= \frac{1}{n}\left\|x_{n_0}x_{n_0}^T - x'_{n_0}(x'_{n_0})^T\right\| \\
&\le \frac{1}{n}\left(\left\|x_{n_0}\right\|_2^2 + \left\|x'_{n_0}\right\|_2^2\right) \quad \text{(by triangle inequality)} \\
&\le \frac{2}{n} \quad \text{(since all } x_i \in K \subset B_2^p\text{)}.
\end{aligned}$$

Thus, $\|\frac{n\varepsilon}{6t}S - \frac{n\varepsilon}{6t}S'\| \le \frac{\varepsilon}{3t}$. So by Lemma 5.3, the projection P is $(\varepsilon/3)$-differentially private.

Due to Lemma 2.5, it suffices to prove that for any fixed projection P, the output data $(\bar{w}_j, \bar{y}_j)_{j=1}^s$ is $(2\varepsilon/3)$-differentially private in the input data $(x_i)_{i=1}^n$. Fixing P fixes also the covering $(\nu_j)_{j=1}^s$.

Consider what happens if we change exactly one vector in the input data $(x_i)_{i=1}^n$. The effect of that change on the nearest-point partition $[n] = F_1 \cup \dots \cup F_s$ is minimal: at most one of the indices can move from one block F_j to another block (thereby changing the cardinalities of those two blocks by 1 each) or to another point in the same block, and the rest of the blocks stay the

same. Thus, the weight vector $w = (w_j)_{j=1}^s$, $w_j = \left|F_j\right|/n$, can change by at most $2/n$ in the ℓ^1 norm. Due to the choice of ρ, it follows by Lemma 2.4 that $w + \rho$ is $(\varepsilon/3)$-differentially private.

For the same reason, all vectors $\tilde{y}_j$ defined in (5.4), except for at most two, stay the same. Moreover, by Lemma 5.10 and Lemma 5.11, the change of each of these (at most) two vectors in the ℓ^1 norm is bounded by

$$\frac{2}{b}\max_{i\in[n]}\|x_i\|_1 \le \frac{2\sqrt{p}}{b}\max_{i\in[n]}\|x_i\|_2 \le \frac{2\sqrt{p}}{b}$$

since all $x_i \in K \subset B_2^p$. Hence, the change of the tuple $(\tilde{y}_1,\dots,\tilde{y}_s) \in \mathbb{R}^{ps}$ in the ℓ^1 norm is bounded by $4\sqrt{p}/b$. Due to the choice of r_j, it follows by Lemma 2.4 that $(\tilde{y}_j + r_j)_{j=1}^s$ is $(\varepsilon/3)$-differentially private.

Since ρ and r_j are all independent vectors, it follows by Lemma 2.7 that the pair $(w+\rho, (\tilde{y}_j + r_j)_{j=1}^s)$ is $(2\varepsilon/3)$-differentially private. The output data $(\bar{w}_j, \bar{y}_j)_{j=1}^s$ is a function of that pair, so it follows by Remark 2.6 that for any fixed projection P, that the output data must be $(2\varepsilon/3)$-differentially private. Applying Lemma 2.5, the result follows. □

5.8 Accuracy

We are ready to combine privacy and accuracy guarantees provided by Theorem 5.12 and Theorem 5.9.

Choose the noises $\rho \in \mathbb{R}^s$ and $r_j \in \mathbb{R}^p$ as in the Privacy Theorem 5.12; then

$$\left(\mathbb{E}\|\rho\|_1^2\right)^{1/2} \lesssim \frac{s}{n\varepsilon}; \quad \left(\mathbb{E}\left\|r_j\right\|_2^2\right)^{1/2} \lesssim \frac{p}{b\varepsilon} \text{ for all } j \in [s]. \tag{5.14}$$

To check the first bound, use triangle inequality as follows:

$$\left(\mathbb{E}\|\rho\|_1^2\right)^{1/2} = \left(\mathbb{E}(|\rho_1| + \dots + |\rho_s|)^2\right)^{1/2} \le (\mathbb{E}\,|\rho_1|^2)^{\frac{1}{2}} + \dots + (\mathbb{E}\,|\rho_s|^2)^{\frac{1}{2}},$$

which is the sum of the standard deviations of the Laplacian distribution.

The second bound follows from summing the variances of the Laplace distribution over all entries.

Choose P to be an output of $\mathrm{PROJ}(\frac{n\varepsilon}{6t}S, t)$ as in Theorem 5.12, where $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$. Take $A = \frac{n\varepsilon}{6t}S$ in Lemma 5.5. We obtain

$$\mathbb{E}\|(I-P)A(I-P)\|_2^2 \le \sum_{i=t+1}^p \lambda_i(A)^2 + 2t\gamma\|A\|^2 + Ct\frac{p^2}{\gamma^2}\log^2\left(\frac{1}{\gamma}\right),$$

and so

$$\mathbb{E}\|(I-P)S(I-P)\|_2^2 \le \sum_{i=t+1}^p \lambda_i(S)^2 + 2t\gamma\|S\|^2 + C\frac{t^3p^2}{n^2\varepsilon^2\gamma^2}\log^2\left(\frac{1}{\gamma}\right),$$

for all $\gamma > 0$, where $C > 0$ is an absolute constant. Since $x_1,\dots,x_n \in K \subset B_2^p$, by (3.3), we have $\sum_{i=t+1}^p \lambda_i(S)^2 \le \frac{1}{t}$. We also have $\|S\| \le 1$. Take $\gamma = \frac{1}{t^2}$. We have

$$\mathbb{E}\|(I-P)S(I-P)\|_2^2 \le \frac{3}{t} + C\frac{t^7p^2}{n^2\varepsilon^2}\log^2 t, \tag{5.15}$$

where $C > 0$ is an absolute constant.

Choose the accuracy α of the covering and its dimension t as follows:

$$t := \left\lfloor \frac{\kappa \log n}{\log(7/\alpha)} \right\rfloor; \quad \alpha = \frac{1}{(\log n)^{1/4}},$$

where $\kappa \in (0,1)$ is a fixed constant that will be introduced later. (See Theorem 5.13.)

By Proposition 3.7, there exists an α-covering in a unit ball of dimension t of cardinality s, where

$$s \le \left(\frac{7}{\alpha}\right)^t \le n^\kappa.$$

Since $t \sim \frac{\kappa \log n}{\log\log n}$, from (5.15), we have

$$\mathbb{E}\|(I-P)S(I-P)\|_2^2 \lesssim \frac{\log\log n}{\kappa \log n} + \frac{p^2(\kappa \log n)^7}{n^2\varepsilon^2}, \tag{5.16}$$

assuming that $t \ge 1$. In the case, when $t = 0$, we have $n \le C$, for some universal constant $C > 0$, so the left hand side is at most $\|S\|_2 \le 1$ and the right hand side is at least $O(1)$.

Apply the Accuracy Theorem 5.9 for this choice of parameters, square both sides and take expectation. Use (5.14) and (5.16). Since the weights $w_j = |F_j|/n$ satisfy $\sum_{j=1}^s w_j = 1$, we have $(\mathbb{E}(\sum_{j=1}^s w_j \|r_j\|_2)^2)^{\frac{1}{2}} \le (\mathbb{E}\sum_{j=1}^s w_j \|r_j\|_2^2)^{\frac{1}{2}} \lesssim \frac{p}{b\varepsilon}$. The error bound in the theorem becomes

$$\begin{aligned}
E :=& \left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \sum_{j=1}^s \bar{w}_j \bar{y}_j^{\otimes d}\right\|_2^2\right)^{1/2} \\
\lesssim& 4^d\left(4\alpha^2 + \|(I-P)S(I-P)\|_2\right) + \frac{2dsb}{n} + 2(\mathbb{E}\|\rho\|_1^2)^{\frac{1}{2}} + 2d\big(\mathbb{E}\,\big(\sum_{j=1}^s w_j \|r_j\|_2\big)^2\big)^{\frac{1}{2}}. \\
\lesssim& 4^d\left(\frac{1}{\sqrt{\log n}} + \sqrt{\frac{\log\log n}{\kappa\log n}} + \frac{p(\kappa\log n)^{\frac{7}{2}}}{n\varepsilon}\right) + \frac{d\cdot b\cdot n^\kappa}{n} + \frac{n^\kappa}{n\varepsilon} + \frac{dp}{b\varepsilon}.
\end{aligned}$$

Optimizing b leads to the following choice:

$$b = \sqrt{\frac{pn^{1-\kappa}}{\varepsilon}}, \tag{5.17}$$

and with this choice we can simplify the error bound as follows:

$$E \lesssim 4^d\left(\sqrt{\frac{\log\log n}{\kappa\log n}} + \frac{p(\kappa\log n)^{\frac{7}{2}}}{n\varepsilon}\right) + d\sqrt{\frac{p}{n^{1-\kappa}\varepsilon}} + \frac{1}{n^{1-\kappa}\varepsilon}.$$

Note that $\kappa \log n = \frac{7}{2}\log(n^{2\kappa/7}) \le \frac{7}{2}n^{2\kappa/7}$. So $\frac{p(\kappa\log n)^{\frac{7}{2}}}{n\varepsilon} \lesssim \frac{pn^\kappa}{n\varepsilon} = \frac{p}{n^{1-\kappa}\varepsilon}$. Thus, in the range where $n \ge (p/\varepsilon)^{1/(1-\kappa)}$ for $\varepsilon \in (0,1)$, where $0 < \kappa \le 1$, we have $\frac{p}{n^{1-\kappa}\varepsilon} \le 1$, so the error can finally be simplified to

$$E \lesssim 4^d\left(\sqrt{\frac{\log\log n}{\kappa\log n}} + \sqrt{\frac{p}{n^{1-\kappa}\varepsilon}}\right).$$

Note that in the complement range where $n < (p/\varepsilon)^{1/(1-\kappa)}$, the second term is greater than one, so such error bound is trivial to achieve by outputting $\bar{y}_j$ to be an arbitrary point in K for all j. Thus we proved:

Theorem 5.13 (Privacy and accuracy)**.** *Let K be a convex set in $\mathbb{R}^p$ that lies in the unit ball B_2^p, and $\varepsilon \in (0,1)$. Fix $\kappa \in (0,1)$. There exists an ε-differentially private algorithm that transforms input data $(x_i)_{i=1}^n$ where all $x_i \in K$ into the output data $(\bar{w}_j, \bar{y}_j)_{j=1}^s$ where $s \le n$, all $\bar{w}_j \ge 0$, $\sum_j \bar{w}_j = 1$, and all $\bar{y}_j \in K$, in such a way that for all $d \in \mathbb{N}$:*

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \sum_{j=1}^s \bar{w}_j \bar{y}_j^{\otimes d} \right\|_2^2 \lesssim 16^d \left(\frac{\log\log n}{\kappa \log n} + \frac{p}{n^{1-\kappa}\varepsilon} \right).$$

The algorithm runs in time polynomial in p, n and linear in the time to compute the metric projection onto K, and it is independent of d.

5.9 Bootstrapping

To get rid of the weights $\bar{w}_j$ and make the synthetic data to have custom size, we can use bootstrapping introduced in Section 4.2, i.e., we can sample new data $u_1, \dots, u_m$ independently and with replacement by choosing $\bar{y}_j$ with probability $\bar{w}_j$ at every step.

Thus, we consider the random vector Y that takes value $\bar{y}_j$ with probability $\bar{w}_j$. Let $Y_1, \dots, Y_m$ be independent copies of Y. Then obviously $\mathbb{E}\, Y^{\otimes d} = \sum_{j=1}^s \bar{w}_j \bar{y}_j^{\otimes d}$, so Bootstrapping Lemma 4.3 yields

$$\mathbb{E}\left\| \frac{1}{m}\sum_{i=1}^m Y_i^{\otimes d} - \sum_{j=1}^s \bar{w}_j \bar{y}_j^{\otimes d} \right\|_2^2 \le \frac{1}{m}.$$

Combining this with the bound in Theorem 5.13, we obtain:

Theorem 5.14 (Privacy and accuracy: custom data size)**.** *Let K be a convex set in $\mathbb{R}^p$ that lies in the unit ball B_2^p, and $\varepsilon \in (0,1)$. Fix $\kappa \in (0,1)$ There exists an ε-differentially private algorithm that transforms input data $x_1, \dots, x_n \in K$ into the output data $u_1, \dots, u_m \in K$, in such a way that for all $d \in \mathbb{N}$:*

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m}\sum_{i=1}^m u_i^{\otimes d} \right\|_2^2 \lesssim 16^d \left(\frac{\log\log n}{\kappa \log n} + \frac{p}{n^{1-\kappa}\varepsilon} \right) + \frac{1}{m}.$$

The algorithm runs in time polynomial in p, n and linear in m and the time to compute the metric projection onto K, and it is independent of d.

5.10 Boolean data: randomized rounding

Now we specialize to Boolean data, i.e., data from $\{0,1\}^p$. If the input data $x_1, \dots, x_n$ is Boolean, the output data $u_1, \dots, u_m$ is in $[0,1]^p$ (for technical reasons we may need to rescale the data by $\sqrt{p}$ because Theorem 5.14 requires K to be in B_2^p). To transform it to Boolean data, we can use randomized rounding as described in Section 4.3. Thus, each coefficient of each vector u_i is independently and randomly rounded to 1 with probability equal to that coefficient, (and to 0 with the complementary probability). Exactly the same analysis as we did in Section 4.3 applies here, and we conclude:

Theorem 5.15 (Boolean private synthetic data)**.** *Let $\varepsilon, \kappa \in (0,1)$. There exists an ε-differentially private algorithm that transforms input data $x_1, \dots, x_n \in \{0,1\}^p$ into the output data $z_1, \dots, z_m \in \{0,1\}^p$ in such a way that the error $E = \frac{1}{n}\sum_{i=1}^n x_i^{\otimes d} - \frac{1}{m}\sum_{i=1}^m z_i^{\otimes d}$ satisfies*

$$\mathbb{E}\binom{p}{d}^{-1} \sum_{1 \le i_1 < \dots < i_d \le p} E(i_1, \dots, i_d)^2 \lesssim 32^d \Big(\frac{\log\log n}{\kappa \log n} + \frac{p}{n^{1-\kappa}\varepsilon} + \frac{1}{m} \Big)$$

for all $d \le p/2$. The algorithm runs in time polynomial in p, n and linear in m, and is independent of d.

A pseudocode description is given in Algorithm 4 below.

Remark 5.16. When $n \ge (p/\varepsilon)^2$, we can take $\kappa = 1/3$ and $m = n$ in Theorem 5.15 and we have the following accuracy

$$\mathbb{E}\binom{p}{d}^{-1} \sum_{1 \le i_1 < \cdots < i_d \le p} E(i_1, \ldots, i_d)^2 \lesssim 32^d \frac{\log\log n}{\log n}, \tag{5.18}$$

for all $d \le p/2$. Can this decay in n on the right hand side be improved if we use other polynomial time differentially private algorithms? Even for $d = 1, 2$, one cannot replace the right hand side by n^{-a}, for any $a > 0$, with any polynomial time differentially private algorithm, assuming the existence of one-way functions. This is because if we could achieve this, then

$$\frac{1}{p^2}\mathbb{E} \max_{1 \le i_1, i_2 \le p} E(i_1, i_2) \le \frac{1}{p^2}\,\mathbb{E} \sum_{i_1=1}^{p} \sum_{i_2=1}^{p} E(i_1, i_2)^2 \lesssim \frac{1}{n^a},$$

so $\mathbb{E}\max_{1 \le i_1, i_2 \le p} E(i_1, i_2) \lesssim p^2/n^a$. But when $n \gg p^{2/a}$, this is impossible to achieve using any polynomial time differentially private algorithm if we assume the existence of one-way functions [35]. It remains an open question: among all polynomial time differentially private algorithms, what is the optimal decay in n on the right hand side of (5.18)?

Acknowledgement

The authors would like to thank the referees and Hao Xing for their detailed and constructive feedback, which has led to definite improvements of some aspects of this paper. M.B. acknowledges support from NSF DMS-2140592. T.S. acknowledges support from NSF-DMS-1737943 and NSF DMS-2027248. R.V. acknowledges support from NSF DMS-1954233, NSF DMS-2027299, U.S. Army 76649-CS, and NSF+Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning.

References

[1] Koenraad MR Audenaert. A norm compression inequality for block partitioned positive semidefinite matrices. *Linear algebra and its applications*, 413(1):155–176, 2006.

[2] Afonso Bandeira, Amit Singer, and Thomas Strohmer. Mathematics of Data Science. https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf, 2020.

[3] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. 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*, pages 273–282, 2007.

[4] Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019.

Algorithm 4 Differentially private Boolean synthetic data via microaggregation

Input: a sequence of points $x_1, \dots, x_n$ in the cube $\{0,1\}^p$ (true data); $\varepsilon, \kappa \in (0,1)$ (privacy); $m \in \mathbb{N}$ (number of points in the synthetic data).

Damped microaggregation

1. Redefine $x_i = \frac{1}{\sqrt{p}} x_i$ for $i = 1, \dots, n$.
2. Compute the second-moment matrix $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$.
3. Let $t \coloneqq \left\lfloor \frac{\kappa \log n}{\log(7/\alpha)} \right\rfloor$ and $\alpha \coloneqq \frac{1}{(\log n)^{1/4}}$.
4. Compute the second-moment matrix $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$.
5. Generate an orthogonal projection P on $\mathbb{R}^p$ from $\mathrm{PROJ}(\frac{n\varepsilon}{6t} S, t)$.
6. Choose an α-covering $\nu_1, \dots, \nu_s \in \mathbb{R}^p$ of the unit Euclidean ball of the subspace $\mathrm{ran}(P)$. This is done by enumerating $B_2^t \cap (\alpha/\sqrt{t})\mathbb{Z}^t$ and mapping it into $\mathrm{ran}(P)$ using any linear isometry.
7. Construct the nearest-point partition $[n] = F_1 \cup \dots \cup F_s$ for $Px_1, \dots, Px_n$ with respect to $\nu_1, \dots, \nu_s$ as follows. For each $\ell \in [n]$, choose a point ν_j nearest to x_ℓ in the ℓ^2 metric and put ℓ into F_j. Break any ties arbitrarily.
8. Let $b = \sqrt{\frac{pn^{1-\kappa}}{\varepsilon}}$.
9. Perform damped microaggregation: compute $w_j = \frac{|F_j|}{n}$ and $\tilde{y}_j = \frac{1}{\max(|F_j|, b)} \sum_{i \in F_j} x_i \in \mathbb{R}^p, \quad j = 1, \dots, s$.
10. Generate an independent noise vector $\rho \in \mathbb{R}^k$ with $\rho_i \sim \mathrm{Lap}\left(\frac{6}{n\varepsilon}\right)$ for $i = 1, \dots, s$.
11. For each $j = 1, \dots, s$, generate noise vectors $r_j \in \mathbb{R}^p$ with $r_{ji} \sim \mathrm{Lap}\left(\frac{12\sqrt{p}}{b\varepsilon}\right)$ for $i = 1, \dots, p$.
12. Consider the simplex $\Delta = \left\{a = (a_1, \dots, a_s) : \ \sum_{i=1}^s a_i = 1; \ a_i \geq 0 \ \forall i\right\}$ and cube $K = \frac{1}{\sqrt{p}}[0,1]^p$.
13. Compute the metric projections $\overline{w} = \pi_{\Delta,1}(w + \rho)$ and $\overline{y}_j = \pi_{K,2}(\tilde{y}_j + r_j), \quad j = 1, \dots, s$, by solving the convex minimizations in (5.5)

Bootstrapping creates new data $u_1, \dots, u_m$ by sampling from the points $\overline{y}_1, \dots, \overline{y}_s$ with weights $\overline{w}_1, \dots, \overline{w}_s$, respectively.

Randomized rounding maps the data $\{\sqrt{p} u_\ell\}_{\ell=1}^m \in [0,1]^p$ to data $\{z_j\}_{j=1}^m \in \{0,1\}^p$.

Output: a sequence of points $z_1, \dots, z_m$ in the cube $\{0,1\}^p$ (synthetic data) that satisfy the properties outlined in Theorem 5.15.

[5] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy," *Journal of the ACM (JACM)*, vol. 60, no. 2, pp. 1–25, 2013.

[6] Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(1):1–136, 2016.

[7] Josep Domingo-Ferrer and Vicenç Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[8] Cynthia Dwork, Aleksandar Nikolov, and Kunal Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *Discrete & Computational Geometry*, 53.3 (2015): 650-673.

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

[10] Fan Fei, Shu Li, Haipeng Dai, Chunhua Hu, Wanchun Dou, and Qiang Ni. A k-anonymity based schema for location privacy preservation. *IEEE Transactions on Sustainable Computing*, 4(2):156–167, 2017.

[11] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford University Press, 2020.

[12] M. Hardt and G. N. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 61–70.

[13] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, 2012.

[14] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics*, 2013.

[15] Michael Kearns and Aaron Roth. How much still needs to be done to make algorithms more ethical. *URL: https://www.shine.cn/opinion/2008214615/*, 2020.

[16] Razaullah Khan, Xiaofeng Tao, Adeel Anjum, Tehsin Kanwal, Abid Khan, Carsten Maple, et al. θ-sensitive k-anonymity: An anonymization model for IoT based electronic health records. *Electronics*, 9(5):716, 2020.

[17] Michael Laszlo and Sumitra Mukherjee. Iterated local search for microaggregation. *Journal of Systems and Software*, 100:15–26, 2015.

[18] Haoran Li, Li Xiong, and Xiaoqian Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Advances in database technology: proceedings. International conference on extending database technology*, volume 2014, page 475. NIH Public Access, 2014.

[19] Ninghui Li, Wahbeh H Qardaji, and Dong Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. *CoRR, abs/1101.2604*, 49:55, 2011.

[20] Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, and Zhiwei Steven Wu. Leveraging public data for practical private query release. *Preprint, arXiv:2102.08598*, 2021.

[21] Yining Liu and Quanyu Zhao. E-voting scheme using secret sharing and k-anonymity. *World Wide Web*, 22(4):1657–1667, 2019.

[22] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pages 4435–4444. PMLR, 2019.

[23] Duncan McLaren-Young-Sommerville. *An Introduction to the Geometry of N Dimensions*. Dover Publications, 2020.

[24] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, 2004.

[25] David Rebollo Monedero, Ahmad Mohamad Mezher, Xavier Casanova Colomé, Jordi Forné, and Miguel Soriano. Efficient k-anonymous microaggregation of multivariate numerical data via principal component analysis. *Information Sciences*, 503:417–443, 2019.

[26] Anna Oganian and Josep Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commission for Europe*, 18(4):345–353, 2001.

[27] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–5, 2017.

[28] Prabhakar Raghavan and Clark D Tompson. Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374, 1987.

[29] David Sánchez, Josep Domingo-Ferrer, Sergio Martínez, and Jordi Soria-Comas. Utility-preserving differentially private data releases via individual ranking microaggregation. *Information Fusion*, 30:1–14, 2016.

[30] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and Sergio Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5):771–794, 2014.

[31] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[32] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[33] Florian Thaeter and Rüdiger Reischuk. Hardness of k-anonymous microaggregation. *Discrete Applied Mathematics*, 2020.

[34] Justin Thaler, Jonathan Ullman, and Salil Vadhan. Faster algorithms for privately releasing marginals. In *International Colloquium on Automata, Languages, and Programming*, pages 810–821. Springer, 2012.

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

[36] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

[37] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. PublicAffairs, 2019.