Privacy of Synthetic Data: a Statistical Framework

March Boedihardjo, Thomas Strohmer, and Roman Vershynin

Synthetic data, differential privacy, Renyi divergene, marginals.

Abstract

Privacy-preserving data analysis is emerging as a challenging problem with far-reaching impact. In particular, synthetic data are a promising concept toward solving the aporetic conflict between data privacy and data sharing. Yet, it is known that accurately generating private, synthetic data of certain kinds is NP-hard. We develop a statistical framework for differentially private synthetic data, which enables us to circumvent the computational hardness of the problem. We consider the true data as a random sample drawn from a population Ω according to some unknown density. We then replace Ω by a much smaller random subset Ω^*, which we sample according to some known density. We generate synthetic data on the reduced space Ω^* by fitting the specified linear statistics obtained from the true data. To ensure privacy we use the common Laplacian mechanism. Employing the concept of Rényi condition number, which measures how well the sampling distribution is correlated with the population distribution, we derive explicit bounds on the privacy and accuracy provided by the proposed method.

I. Introduction

Data science and artificial intelligence play a key role in successfully tackling many of the grand challenges our society is facing over the coming years. Data sharing and data democratization will feature prominently in these endeavors. At the same time, data colonialism [1] and surveillance capitalism [2] emerge as increasingly concerning developments that threaten the potential benefits of data-driven advancements and that highlight the utmost importance of data rights and privacy. For instance, the WHO emphasized in its recent report the importance of data management methods that improve the utility and accuracy of health-care data, while not compromising privacy [3]. However, data democratization and responsible data sharing are not likely to be accommodated by more efficient deidentification or strict security/privacy processes alone.

Synthetic data is a promising ingredient toward solving the aporetic conflict between data privacy and data sharing. The goal of synthetic data is to create an as-realistic-as-possible data set, one that not only maintains the nuances of the original data, but does so without risk of exposing sensitive information. The problem of making private and accurate synthetic data is NP-hard in the worst case [4], [5].

In this paper we take a different route. We will show that the problem of making private and accurate synthetic data is tractable in the statistical framework, where the true data is seen as a random sample drawn from some probability space. Our method comes with guarantees of privacy, accuracy, and computational efficiency. We will discuss how our method has the potential to improve upon existing techniques in Section II-H. This paper focuses on the theoretical aspect of the proposed statistical framework. Specific details regarding numerics and an experimental validation are devoted to future work.

II. Problem setup and main results

A. The problem

We model the true data X as a sequence of n elements from some ground set Ω. E.g., for an electronic health record these elements might represent patients. For example, $\Omega = \{0,1\}^p$ allows each patient to have p binary parameters, while $\Omega = \mathbb{R}^p$ allows the parameters to be real. Multimodal data are possible, too: some parameters may be categorical, some real, some may consist of text strings, etc. We would like to manufacture a synthetic dataset Y, which is another sequence of k elements from Ω. We want the synthetic data to be private and accurate.

B. Defining accuracy

By "accuracy" we mean the accuracy of linear statistics of the data. Consider a finite class $\mathcal{F}$ of *test functions*, which are functions from Ω to $[-1,1]$. Linear statistics of the data $X = (x_1,\dots,x_n)$ are the sums of the form $\frac{1}{n}\sum_{i=1}^n f(x_i)$ for $f \in \mathcal{F}$. We would like the synthetic data Y to approximately preserve all these sums, up to a given additive error δ:

$$\max_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k} f(y_i) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right| \le \delta. \tag{1}$$

In this case we say that the synthetic dataset is δ-*accurate*.

As an important example, linear statistics are capable of encoding *marginals* of high-dimensional data. Indeed, let us consider Boolean data where $\Omega = \{0,1\}^p$. In the context of electronic health records, the data $X = (x_1,\dots,x_n)$ consists of records of

M.B. and R.V. are with the Department of Mathematics, University of California Irvine, T.S. is with the Department of Mathematics & Center of Data Science and Artificial Intelligence Research, University of California, Davis

n patients each having p binary parameters. The fraction of the number of patients whose first and second parameters equal 1 and third parameter equals 0 is a three-dimensional marginal. It can be expressed as the linear statistic $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$, where $f : \{0,1\}^p \to \{0,1\}$ is the indicator function $f(x) = \mathbf{1}_{\{x(1)=x(2)=1,\, x(3)=0\}}$. One-dimensional marginals capture the means of the parameters, jointly with two-dimensional marginals they determine the correlations, and higher dimensional marginals capture higher-order dependencies.

In many situations, $|\Omega|$ is too large for computations while $|\mathcal{F}|$ is reasonable. For example, if $\mathcal{F}$ encodes all d-dimensional marginals of p-dimensional Boolean data as in the previous example, $|\Omega| = 2^p$ is exponential in p, while

$$|\mathcal{F}| = \binom{p}{\leq d} = \binom{p}{0} + \binom{p}{1} + \cdots + \binom{p}{d} \leq \left(\frac{ep}{d}\right)^d$$

is polynomial in p for any fixed d.

C. A statistical framework

Ullman and Vadhan [4] showed (under standard cryptographic assumptions) that in general it is NP-hard to make private synthetic Boolean data which approximately preserve all two-dimensional marginals. While this result may seem discouraging, it is a worst-case result.

Yet the *worst* kind of data, for which the problem is hard, are rarely seen in practice. More common in applications is the statistical framework, where the true data is seen as a *random* sample drawn from some probability space (Ω, Σ, ν). The probability distribution ν specifies the population model of the true data. We assume that we neither know ν, nor can we sample according to ν thereby generating more true data.

Suppose, however, that we can sample from Ω according to some other, known, probability measure μ. For example, while we may not know the underlying population distribution ν of the patients in the Boolean cube $\Omega = \{0,1\}^p$, we can still sample from the cube according to the uniform measure μ by choosing all coordinates at random and independently. Similarly, while we may not know the population distribution ν of written notes in patient health records, there do exist generative models that generate texts, which can be leveraged as prior information when constructing μ. In order to uphold privacy, we assume that the true data X may not be used to build the generative model μ, but it can be built using some other public data. For example, X may represent the Census 2020 data with associated underlying population distribution ν. To generate μ we can use publicly available datasets, such as the published (and thus sanitized) Census 2010 data. The idea of using a publicly available dataset to model the underlying distribution of the original dataset in a private manner is also discussed in [27].

Having put our problem into a statistical framework, we can try to circumvent the computational hardness of our problem in the most obvious way: *subsample* Ω. Namely, we replace Ω by a much smaller random subset Ω^* that is sampled according to the distribution μ. Then we generate synthetic data in Ω^* by fitting the desired linear statistics (e.g. all marginals up to a specified degree) of the true data as close as possible[1].

This idea may only work if the sampling distribution μ has some "correlation" with the population distribution ν. We can quantify this correlation using the notion of *Rényi divergence* [6]. Namely, if ν is absolutely continuous with respect to μ, we can utilize the Radon-Nikodym derivative $d\nu/d\mu$ to define the *Rényi condition number*

$$\kappa(\nu\|\mu) = \int \left(\frac{d\nu}{d\mu}\right)^2 d\mu = \int \frac{d\nu}{d\mu}\, d\nu, \tag{2}$$

a quantity that equals the exponential of $D_2(\nu\|\mu)$, the *Rényi divergence* of order 2.

Recall that in information theory the Rényi divergence of order 2 is also referred to as χ^2-divergence [7], [8], a special case of the f-divergence, which in turn has found various applications in the context of privacy, see e.g. [9]–[11].

Conceptually, $\kappa(\nu\|\mu)$ is similar to the notion of the condition number in numerical linear algebra: the smaller, the better. The best value of the Rényi condition number is 1, achieved when $\nu = \mu$.

If Ω is finite, the Radon-Nikodym derivative $d\nu/d\mu$ equals the ratio of the densities $\phi(x) = \nu(\{x\})$ and $\psi(x) = \mu(\{x\})$. In particular, if the sampling distribution μ is uniform, $\psi(x) = 1/|\Omega|$ for all x, and we have

$$\kappa(\nu\|\mu) = \int \phi(x)^2 |\Omega|^2 \, d\mu(x) = \left(\frac{\|\phi\|_{L^2(\mu)}}{\|\phi\|_{L^1(\mu)}}\right)^2. \tag{3}$$

Thus, the Rényi condition number in this case measures the regularity of the population density ϕ: the more spread out it is, the smaller its Rényi condition number.

[1]We will denote the space of densities on Ω^* by $\mathcal{D}(\Omega^*)$.

D. Our approach

Our method, in a nutshell, is the following: obtain a reduced space Ω^* by subsampling Ω according to the known probability measure μ, and generate synthetic data Y on Ω^* by fitting the linear statistics obtained from X.

Our results come with guarantees of privacy, accuracy, and efficiency. To achieve all this, we assume (roughly speaking) that the size of the true data is at least nearly linear in the number of statistics we seek to preserve:

$$|X| \gtrsim |\mathcal{F}| \log |\mathcal{F}| .$$

For accuracy, we need the size of the synthetic data to be at least logarithmic in the number of statistics (a mild assumption):

$$|Y| \gtrsim \log |\mathcal{F}| .$$

And, finally, we can make all computations in the reduced space Ω^* as long as its size is at least linear in the number of statistics:

$$\left|\Omega^*\right| \gtrsim |\mathcal{F}| .$$

If these three conditions are met, we can generate synthetic data while preserving privacy, accuracy, and efficiency (for the latter, we solve a linear program in dimension $|\Omega^*|$).

E. Differential privacy

In order to provide rigorous privacy guarantees, we will employ the concept of *differential privacy* [12], which has emerged as a de-facto standard for private data sharing.

Definition II.1 (Differential Privacy [12])**.** *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 $S \subseteq \mathrm{range}(\mathcal{M})$,*

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

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

A basic technique to achieve differential privacy is the *Laplacian mechanism*, which consists of adding Laplacian noise to the data. A Laplacian random variable λ is Laplacian with parameter σ, abbreviated $\lambda \sim \mathrm{Lap}(\sigma)$, if λ is a symmetric random variable with exponential tails in both directions:

$$\mathbb{P}\left\{|\lambda| > t\right\} = \exp(-t/\sigma), \quad t \geq 0.$$

It is well known and not hard to see that Laplacian mechanism achieves differential privacy; see Lemma III.1 for details.

F. Algorithm

We present a high level algorithmic description of our proposed method in Algorithm 1 below. See Section II-G for the role of the parameters arising in the algorithm.

Algorithm 1 Private synthetic data algorithm

Input: (a) the true data: a sequence $X = (x_1, \ldots, x_n) \in \Omega$;
(b) a family $\mathcal{F}$ of test functions from Ω to $[-1, 1]$;
(c) the reduced space $\Omega^* = \{z_1, \ldots, z_m\}$, made of points z_i chosen from Ω;
(d) parameter $\sigma > 0$.

1. **Add noise:** For each test function $f \in \mathcal{F}$, generate an independent Laplacian random variable $\lambda(f) \sim \mathrm{Lap}(\sigma)$.
2. **Reweight:** Compute a density h^* on Ω^* whose linear statistics are uniformly as close as possible to the linear statistics of the true data perturbed by Laplacian noise:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{D}(\Omega^*)} \left\{ \max_{f \in \mathcal{F}} \left| \sum_{i=1}^{m} f(z_i) h(z_i) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \lambda(f) \right| \right\}.$$

3. **Bootstrap:** Create a sequence $Y = (y_1, \ldots, y_k)$ of k elements drawn from Ω^* independently with density h^*.

Output: synthetic data $Y = (y_1, \ldots, y_k)$.

To implement Algorithm 1 in practice, we note the following:

- The Rényi condition number $\kappa(\nu \| \mu)$ quantifies how the similarity of the (usually unknown) population distribution ν and the generated distribution μ affects the accuracy of the synthetic data. It may be difficult (or even impossible) to compute

$\kappa(\nu\|\mu)$. Luckily, this does not affect the practicality of Algorithm 1. Note that $\kappa(\nu\|\mu)$ is not needed for achieving privacy (see Theorem II.2 below), but only for accuracy purposes (see Theorem II.3). This is essential. While it is impossible to verify differential privacy empirically, it is not difficult to measure accuracy empirically. We thus can proceed as follows: Create Ω^*, apply Algorithm 1 with parameters $\varepsilon, \delta, \sigma$ chosen such that the desired ε-DP of the synthetic data, guaranteed by Theorem II.2, is fulfilled. Measure the resulting accuracy. If the accuracy is too low, create a new Ω^* with larger cardinality, and repeat, until the required accuracy is met.

- Computing h^* in Algorithm 1 amounts to solving a linear program with $|\Omega^*| \le m$ variables[2] and at most $|\mathcal{F}| + m + 1$ constraints. The complexity of solving general linear programs is polynomial in the number of variables, see e.g. [13] and thus Algorithm 1 is in principle feasible. Nevertheless the computational complexity may still be too high for certain problems. We plan to investigate efficient implementations based on Algorithm 1 in our future work.

G. Privacy and accuracy guarantees

Theorem II.2 (Privacy)**.** *Choose $\delta > 0, \gamma > 0$ and set $\sigma = \delta / \log(|\mathcal{F}|/\gamma)$. If*

$$n \ge 2(\varepsilon\delta)^{-1}|\mathcal{F}|\log(|\mathcal{F}|/\gamma),$$

then Algorithm 1 is ε-differentially private.

We emphasize that this privacy guarantee holds for *any* choice of the reduced space Ω^*.

Theorem II.3 (Accuracy)**.** *Let $\min(n,k) \ge \delta^{-2}\log(|\mathcal{F}|/\gamma)$ and $m \ge \delta^{-2}K|\mathcal{F}|/\gamma$, where $\delta \in (0, 1/2]$ and $\gamma \in (0, 1/4)$. Set $\sigma = \delta/\log(|\mathcal{F}|/\gamma)$. Suppose the true data $X = (x_1, \dots, x_n)$ is sampled from Ω independently and according to some probability measure ν, and the reduced space $\Omega^* = \{z_1, \dots, z_m\}$ is sampled from Ω independently and according to some probability measure μ. Assume that the Rényi condition number satisfies $\kappa(\nu\|\mu) \le K$. Also assume that the family $\mathcal{F}$ contains the function that is identically equal to 1. Then with probability at least $1 - 4\gamma$ the synthetic data $Y = (y_1, \dots, y_k)$ generated by Algorithm 1 is (8δ)-accurate.*

Let us specialize our results to Boolean data. Here the sample space is $\Omega = \{0,1\}^p$ and we seek accuracy with respect to all $|\mathcal{F}| = \binom{p}{\le d}$ marginals up to degree d. Choose μ to be the uniform density on the cube, recall (3), and combine the two theorems above to get:

Corollary II.4 (Boolean data)**.** *Let $n \gg \binom{p}{\le d}\log\binom{p}{\le d}$ and $k \gg \log\binom{p}{\le d}$. Suppose that the true data $X = (x_1, \dots, x_n)$ are sampled from $\{0,1\}^p$ independently and according to some (unknown) density ϕ. Then one can generate synthetic data $Y = (y_1, \dots, y_k)$ that is $o(1)$-accurate with respect to all marginals of dimension at most d with probability $1 - o(1)$, and is also $o(1)$-differentially private. The algorithm that generates Y from X runs in time polynomial in n, k, and κ for a fixed d.*

The proofs of the claims above will be given in Section III.

In light of the aforementioned "no-go" result of Ullman and Vadhan [4], a thorough analysis of the privacy-utility tradeoff must also include the computational complexity of the algorithm. In the notation of Theorems II.2 and II.3 it is thus not just a question of ε vs. δ, but ε *and* δ vs. computational cost. If we nevertheless focus only on the relationship between privacy (measured by ε) and accuracy (measured by δ) we conclude from an inspection of the assumptions and conclusions of Theorems II.2 and II.3 that

$$\delta \ge \max\left\{\frac{A}{\varepsilon}, B\right\},$$

where A and B depend on $n, m, |\mathcal{F}|, K$, and γ.

H. Related work

There exists a fairly large body of work on privately releasing answers in the interactive and non-interactive query setting, a detailed review of which is beyond the scope of this paper. 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 [14], 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. This computational inefficiency comes as no surprise, if we recall that making differentially private Boolean synthetic data which preserves all of the two-dimensional marginals with accuracy $o(1)$ is NP-hard [4].

The papers [15], [16] 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.

[2]We have inequality here because the set Ω^* is formed of points z_i that are sampled independently, which may result in repetitions.

In [17], Barak et al. derive a method for producing accurate and private synthetic Boolean data based on linear programming. The method in [17] is conceptually similar to ours even though it is concerned with marginals, while our approach holds for general linear statistics. The key difference is in the computational complexity. The method in [17] involves solving a linear program on the entire domain $\Omega = \{0,1\}^p$ and thus its running time is exponential in p. The authors of [17] emphasize that *"one of the main algorithmic questions left open from this work is that of efficiency"*, for which our paper provides a solution. Our method works in the reduced space Ω^*, which, according to Theorem II.3, has size m slightly larger than $\binom{p}{\leq d}$, and thus it is only polynomial in p, thereby providing a positive answer to the aforementioned algorithmic question.

The method developed by Hardt and Talwar in [18] privately releases answers to linear queries (including, in particular, marginals). It applies to general data that needs not be Boolean, just like in our work. However, unlike our method, the method in [18] does not construct synthetic data. Also, unlike our work, the theoretical accuracy bounds in [18] hold for most but not all linear queries. Nikolov, Talwar, and Zhang in [19], follow up on the work [18] and improve the (lower and upper) bounds derived by Hardt and Talwar. The lack of efficiency of the method in [19] is addressed in [20], where the authors demonstrate empirically the computational efficiency of their method.

The paper [21] by Dwork, Nikolov, and Talwar is concerned with a convex relaxation based approach for private marginal release, and thus, unlike our method, does not construct synthetic data for a ground set Ω. Also, [21] gives "only" (ε, δ)-differential privacy.

Privacy-preserving data analysis (beyond marginals) in a statistical framework is the focus of [22], [23]. While these papers are quite intriguing, they are not concerned with synthetic data, and thus not directly related to this work. There is also an increasing body of work deals with privately releasing data via methods from deep learning, see e.g. [24], [25]. But these methods do not come with any accuracy guarantees.

Another method of constructing private synthetic data was proposed recently in [26]. To compare the two, recall that the no-go result of Ullhman says (roughly) that, for the *worst true data*, it is impossible to efficiently construct private synthetic Boolean data that approximately preserves *all marginals* of dimension 2. The work [26] and the present paper overcome this impossibility result, each in its own way: this paper relaxes "worst data" to "typical data", while [26] relaxes "all marginals" to "most marginals".

Closest in spirit to our paper is the paper [27]. After submitting our paper we became aware of [27], which was published shortly before our submission. In [27], the authors propose to use publicly available datasets as a kind of statistical prior to improve the accuracy of synthetic datasets. Instead of standard ε-DP (as is used in this paper) they use concentrated DP, a weaker notion of privacy[3] (and thus easier to achieve) than ε-DP. Their method to construct synthetic data employs a variation of the *Multipliciative Weights Exponential Mechanism* [16] (a greedy-type algorithm that iteratively tries to modify the synthetic data to fit the query with the largest error), using in an adaptive manner the distribution of the publicly available dataset as prior. The authors derive theoretical bounds for the privacy and accuracy of their method. Since [27] uses the Gaussian mechanism to achieve privacy, a different metric to measure similarity between distributions, and a weaker notion of privacy, it makes a direct comparison of their results with our results elusive.

Generating private synthetic data with tools from machine learning has gained much attention in recent years, see e.g [24], [28]–[34]. However, many of the these methods are just empirical. While some of these methods do come with differential privacy guarantees, they provide no utility guarantees of any kind.

III. Proofs

For an integrable function $f : \Omega \to \mathbb{R}$ on a measure space (Ω, Σ, ν), we denote

$$\langle\!\langle f, \nu \rangle\!\rangle = \int f \, d\nu. \tag{4}$$

Given a sequence of points $x_1, \ldots, x_n \in \Omega$, possibly with repetitions, we consider the empirical measure

$$\nu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

By definition, we have

$$\langle\!\langle f, \nu_n \rangle\!\rangle = \frac{1}{n} \sum_{i=1}^{n} f(x_i). \tag{5}$$

With this notation, the optimization part of Algorithm 1 can be expressed as follows:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{D}(\Omega^*)} \left\{ \max_{f \in \mathcal{F}} \left| \langle\!\langle f, h \rangle\!\rangle - \langle\!\langle f, \nu_n \rangle\!\rangle - \lambda(f) \right| \right\}. \tag{6}$$

[3]While ε-DP implies concentrated DP (albeit with a different ε), concentrated DP only implies (ε, δ)-DP with $\delta > 0$.

A. Privacy

The following lemma is well known, see e.g. Theorem 2 in [17].

Lemma III.1 (Laplacian mechanism)**.** *Let $\mathcal{A}$ be a mapping that transforms data D to a point $\mathcal{A}(D) \in \mathbb{R}^N$. Let*

$$\Delta = \max_{D_1, D_2} \big\|\mathcal{A}(D_1) - \mathcal{A}(D_2)\big\|_1$$

where the maximum is over all pairs of input data D_1 and D_2 that differ in a single element. Then the addition of i.i.d. Laplacian noise $\lambda_i \sim \mathrm{Lap}(\sigma)$ to each coordinate of $\mathcal{A}(D)$ preserves (Δ/σ)-differential privacy.

Consider the linear map $\mathcal{L}$ that associates to a measure ν on Ω the set of its linear statistics, namely

$$\mathcal{L}(\nu) = \big(\langle\!\langle f, \nu \rangle\!\rangle\big)_{f \in \mathcal{F}} \in \mathbb{R}^{|\mathcal{F}|}.$$

Consider two input sets $(x_1, \dots, x_n)$ and $(x_1, \dots, x_n, x_{n+1})$ that differ by exactly one element x_{n+1}. Then one can easily check that the corresponding empirical measures satisfy the identity

$$\nu_{n+1} - \nu_n = \frac{1}{n+1}\left(\delta_{x_{n+1}} - \nu_n\right).$$

Then, using linearity of $\mathcal{L}$ and the triangle inequality, we obtain

$$\big\|\mathcal{L}(\nu_{n+1}) - \mathcal{L}(\nu_n)\big\|_1 = \big\|\mathcal{L}(\nu_{n+1} - \nu_n)\big\|_1 \tag{7}$$

$$\le \frac{1}{n+1}\big\|\mathcal{L}(\delta_{x_{n+1}})\big\|_1 + \frac{1}{n+1}\big\|\mathcal{L}(\nu_n)\big\|_1. \tag{8}$$

To bound this quantity further, note that for every i the definition of $\mathcal{L}$ yields

$$\big\|\mathcal{L}(\delta_{x_i})\big\|_1 = \sum_{f \in \mathcal{F}} \big|\langle\!\langle f, \delta_{x_i} \rangle\!\rangle\big| = \sum_{f \in \mathcal{F}} \big|f(x_i)\big| \le |\mathcal{F}|, \tag{9}$$

where in the last step we used that each function $f \in \mathcal{F}$ takes values in $[-1, 1]$. Therefore, by linearity of $\mathcal{L}$ and the triangle inequality,

$$\big\|\mathcal{L}(\nu_n)\big\|_1 = \left\|\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\delta_{x_i})\right\|_1 \le \frac{1}{n}\sum_{i=1}^n \big\|\mathcal{L}(\delta_{x_i})\big\|_1 \le |\mathcal{F}|,$$

where in the last step we used (9). Substituting the bound (9) for $i = n+1$ and the last inequality into (7), we conclude that

$$\Delta := \big\|\mathcal{L}(\nu_{n+1}) - \mathcal{L}(\nu_n)\big\|_1 \le \frac{2|\mathcal{F}|}{n}.$$

Applying Lemma III.1, we see that the addition of the independent Laplacian random variable $\lambda(f) \sim \mathrm{Lap}(\sigma)$ to each coordinate $\langle\!\langle f, \nu_n \rangle\!\rangle$ of $\mathcal{L}(\nu_n)$ preserves (Δ/σ)-differential privacy. Due to the bound on Δ above, the choice of σ in the algorithm, and the assumption on n in Theorem II.2, we have

$$\frac{\Delta}{\sigma} \le \frac{2|\mathcal{F}|\log(|\mathcal{F}|/\gamma)}{n\delta} \le \varepsilon.$$

Hence, the family of perturbed coefficients $\langle\!\langle f, \nu_n \rangle\!\rangle + \lambda(f)$ is ε-differentially private. Finally, the function h^* in (6) computed by the algorithm is a function of these private perturbed coefficients. Hence the algorithm is ε-differentially private. Theorem II.2 is proved.

B. Accuracy

Here, our input data $X_1, \dots, X_n$ are i.i.d. points sampled from Ω according to the probability measure ν, and the reduced space Ω^* is formed by the points $Z_1, \dots, Z_m$ sampled from Ω according to the probability measure μ. Consider the corresponding empirical probability measures

$$\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \quad \text{and} \quad \mu_m = \frac{1}{m}\sum_{i=1}^m \delta_{Z_i}.$$

Let us reweigh the reduced space, introducing the measure

$$\nu'_m = \frac{1}{m}\sum_{i=1}^m \left(\frac{d\nu}{d\mu}\right)(Z_i)\,\delta_{Z_i}. \tag{10}$$

The point is that both ν_n and ν'_m are unbiased estimators of the population measure ν:

$$\mathbb{E}\,\nu_n = \mathbb{E}\,\nu'_m = \nu.$$

These identities can be easily deduced from the definition of the Radon-Nikodym derivative. In our argument, however, they will not be used. Instead, we need uniform deviation inequalities that would guarantee that with high probability, all linear statistics of ν_n, ν'_m and ν approximately match. This is the content of the next two lemmas.

Lemma III.2 (Deviation of linear statistics for ν_n)**.** *Let (Ω, Σ, ν) be a probability space, and let ν_n be an empirical probability measure corresponding to ν. If $n \geq \delta^{-2} \log(|\mathcal{F}|/\gamma)$ then, with probability at least $1-\gamma$, we have*

$$\max_{f \in \mathcal{F}} \left| \langle\!\langle f, \nu_n \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle \right| \leq \delta.$$

Proof. For each function $f \in \mathcal{F}$, recalling (4) and (5) we get

$$\langle\!\langle f, \nu \rangle\!\rangle = \int f \, d\nu = \mathbb{E}\, f(X), \quad \langle\!\langle f, \nu_n \rangle\!\rangle = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

where $X, X_1, X_2, \ldots$ are drawn from Ω independently according to probability measure ν. Therefore

$$\langle\!\langle f, \nu_n \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle = \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - \mathbb{E}\, f(X_i) \big)$$

is a normalized and centered sum of i.i.d. random variables, which are bounded by 1 in absolute value (by assumption on $\mathcal{F}$). Applying Hoeffding's inequality (see e.g. [35, Theorem 2.2.6]) and after simplification we get that for any $\delta \in (0,1)$

$$\mathbb{P}\left\{ \left| \langle\!\langle f, \nu_n \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle \right| > \delta \right\} \leq \exp(-\delta^2 n) \leq \gamma / |\mathcal{F}|,$$

where in the last step we used the assumption on n. The lemma is proved. □

Lemma III.3 (Deviation of linear statistics for ν'_m)**.** *If $m \geq \delta^{-2} K |\mathcal{F}|/\gamma$ and $\kappa(\nu\|\mu) \leq K$ then, with probability at least $1-\gamma$, we have*

$$\max_{f \in \mathcal{F}} \left| \langle\!\langle f, \nu'_m \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle \right| \leq \delta.$$

Proof. For each test function $f \in \mathcal{F}$, by definition of the Radon-Nikodym derivative, we have

$$\langle\!\langle f, \nu \rangle\!\rangle = \int f \, d\nu = \int f(z) \left(\frac{d\nu}{d\mu} \right)(z) \, d\mu(z) = \mathbb{E}\left(\frac{d\nu}{d\mu} \right)(Z) f(Z),$$

where Z is drawn from Ω according to probability measure μ. Furthermore, by definition of reweighting (10) we have

$$\langle\!\langle f, \nu'_m \rangle\!\rangle = \int f \, d\nu'_m = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{d\nu}{d\mu} \right)(Z_i) \, f(Z_i),$$

where Z_i are i.i.d. copies of Z. Therefore

$$\langle\!\langle f, \nu'_m \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle = \frac{1}{m} \sum_{i=1}^{m} (R_i - \mathbb{E}\, R_i),$$

where

$$R_i = \left(\frac{d\nu}{d\mu} \right)(Z_i) \, f(Z_i).$$

In other words, we have a normalized and centered sum of i.i.d. random variables. The variance of each term of the sum is bounded by the Rényi condition number $\kappa(\nu\|\mu)$. Indeed,

$$\begin{aligned} \mathrm{Var}(R_i) \leq \mathbb{E}\, R_1^2 = \mathbb{E}\left(\frac{d\nu}{d\mu} \right)(Z)^2 \, f(Z)^2 &\leq \mathbb{E}\left(\frac{d\nu}{d\mu} \right)(Z)^2 \\ &= \int \left(\frac{d\nu}{d\mu} \right)^2 d\mu = \kappa(\nu\|\mu) \leq K. \end{aligned}$$

Here, in the third step we used the assumption that f takes values in $[-1, 1]$.

We showed that the variance of $\langle\!\langle f, \nu'_m \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle$ is bounded by K/m. Applying Chebyshev's inequality, we get for any $\delta \in (0,1)$ that

$$\mathbb{P}\left\{ \left| \langle\!\langle f, \nu'_m \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle \right| > \delta \right\} \leq \frac{K}{\delta^2 m} \leq \frac{\gamma}{|\mathcal{F}|},$$

where in the last step we used the assumption on m. The lemma is proved. □

Proof of Theorem II.3. Assume that the events in the conclusions of Lemma III.2 and Lemma III.3 hold; this happens with probability at least $1 - 2\gamma$.

The measure ν'_m introduced in (10) need not be a probability measure, since its total mass

$$r \coloneqq \langle\!\langle \mathbf{1}, \nu'_m \rangle\!\rangle$$

does not need to equal 1. But it is not far from 1. Indeed, since the constant function $\mathbf{1}$ lies in $\mathcal{F}$ by assumption, the conclusion of Lemma III.3 gives

$$\left|\langle\!\langle \mathbf{1}, \nu'_m \rangle\!\rangle - \langle\!\langle \mathbf{1}, \nu \rangle\!\rangle\right| \le \delta.$$

Since ν is a probability measure, it satisfies $\langle\!\langle \mathbf{1}, \nu \rangle\!\rangle = 1$, and we get

$$|r - 1| \le \delta. \tag{11}$$

Now, ν'_m/r is a probability measure. Let us check that it satisfies a deviation inequality. To this end, first note that the conclusion of Lemma III.3 and triangle inequality give

$$\left|\langle\!\langle f, \nu'_m \rangle\!\rangle\right| \le \left|\langle\!\langle f, \nu \rangle\!\rangle\right| + \delta = \left|\int f \, d\nu\right| + \delta \le 1 + \delta \tag{12}$$

where we used the assumption that all $f \in \mathcal{F}$ take values in $[-1, 1]$. Thus, subtracting and adding the term $\langle\!\langle f, \nu'_m \rangle\!\rangle$, we obtain

$$\left|\langle\!\langle f, \nu'_m/r \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle\right| \le \left|1/r - 1\right| \left|\langle\!\langle f, \nu'_m \rangle\!\rangle\right| + \left|\langle\!\langle f, \nu'_m \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle\right|.$$

Since $\delta \in (0, 1/2]$, (11) yields $\left|1/r - 1\right| \le 2\delta$. Furthermore, (12) yields $\left|\langle\!\langle f, \nu'_m \rangle\!\rangle\right| \le 3/2$. Finally, the conclusion of Lemma III.3 yields $\left|\langle\!\langle f, \nu'_m \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle\right| \le \delta$. Substituting these bounds into the inequality above, we obtain the desired deviation inequality:

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, \nu'_m/r \rangle\!\rangle - \langle\!\langle f, \nu \rangle\!\rangle\right| \le 4\delta.$$

Combining this with the conclusion of Lemma III.2 via the triangle inequality, we obtain

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, \nu'_m/r \rangle\!\rangle - \langle\!\langle f, \nu_n \rangle\!\rangle\right| \le 5\delta.$$

A simple union bound over $|\mathcal{F}|$ Laplacian random variables shows that with probability at least $1 - \gamma$,

$$\max_{f \in \mathcal{F}} \left|\lambda(f)\right| \le \sigma \log(|\mathcal{F}|/\gamma) = \delta \tag{13}$$

where the last identity is due to the choice of σ in the algorithm. Combining the two bounds, with probability at least $1 - 3\gamma$, we have

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, \nu'_m/r \rangle\!\rangle - \langle\!\langle f, \nu_n \rangle\!\rangle - \lambda(f)\right| \le 6\delta.$$

Recall that, by construction, ν'_m/r is a probability measure on the set $\Omega^* = \{Z_1, \dots, Z_m\}$. Therefore, minimality of h^* in algorithm (6) implies that

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, h^* \rangle\!\rangle - \langle\!\langle f, \nu_n \rangle\!\rangle - \lambda(f)\right| \le 6\delta.$$

Using (13) again, we conclude that

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, h^* \rangle\!\rangle - \langle\!\langle f, \nu_n \rangle\!\rangle\right| \le 7\delta.$$

To complete the proof, we note that bootstrapping preserves the accuracy of linear statistics. Indeed, apply Lemma III.2 for the probability density h^* on Ω^* and its empirical counterpart $h^*_k = \frac{1}{k}\sum_{i=1}^k \delta_{Y_i}$ where Y_i are sampled independently from Ω^* according to the probability density h^*. Since $k \ge \delta^{-2} \log(|\mathcal{F}|/\gamma)$ by assumption, with probability at least $1 - \gamma$ we have

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, h^*_k \rangle\!\rangle - \langle\!\langle f, h^* \rangle\!\rangle\right| \le \delta.$$

Combining this with the previous bound, we obtain that with probability at least $1 - 4\gamma$,

$$\max_{f \in \mathcal{F}} \left|\langle\!\langle f, h^*_k \rangle\!\rangle - \langle\!\langle f, \nu_n \rangle\!\rangle\right| \le 8\delta.$$

This is an equivalent form of (8δ)-accuracy (1). Theorem II.3 is proved. □

IV. Open problems

While the method proposed in this paper provides a simple and efficient roadmap to construct private synthetic data that preserve with high accuracy linear statistics of the original data, we may require our synthetic data to accurately model other features of the data that are not (fully) captured by linear statistics. This poses numerous questions. For example, how well do linear statistics inform other kinds of data analysis tasks (e.g., clustering, classification, regression, etc., see also [22], [36])?

Another interesting direction is to see whether we can replace the condition $\kappa(\nu\|\mu) \le K$ in Theorem II.3 by the more relaxed assumption $e^{\mathrm{KL}(\nu\|\mu)} \le K$, or whether there is any other differentially private algorithm that works under this relaxed assumption. Here $\mathrm{KL}(\nu\|\mu)$ is the Kullback-Liebler divergence and by Jensen's inequality, $\mathrm{KL}(\nu\|\mu) \le \log \kappa(\nu\|\mu)$.

Yet another challenge is that we do not know the population distribution ν, and thus we may not know how to choose a good sampling distribution μ. Using various generative models seem a natural choice for certain types of data, such as text and images. Using those, we may hope to build the sampling distribution μ that has enough "overlap" with the population distribution ν (as measured by the Rényi condition number). Since we just need to be able to sample from ν, building an MCMC model for it is enough.

It is important, however, that we may not use the true data X to make any decisions about μ, as this could violate privacy. The sampling distribution μ should be estimated in some other way. We can either use private density estimation for that purpose, or estimate μ from some publicly available data that does not need to be protected by privacy. As mentioned earlier, consider the example where X represents the Census 2020 data with associated underlying population distribution ν. To generate μ we could use the published Census 2010 data, see also [27].

Acknowledgement

M.B. acknowledges support from NSF DMS-2140592. T.S. acknowledges support from NSF-DMS-1737943, NSF DMS-2027248, NSF CCF-1934568 and a CeDAR Seed grant. 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] N. Couldry and U. A. Mejias, "Data colonialism: Rethinking big data's relation to the contemporary subject," *Television & New Media*, vol. 20, no. 4, pp. 336–349, 2019.
[2] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. PublicAffairs, 2019.
[3] W. H. Organization, "Ethics and governance of artificial intelligence for health: WHO guidance," https://apps.who.int/iris/rest/bitstreams/1352854/retrieve, 2021.
[4] J. Ullman and S. Vadhan, "PCPs and the hardness of generating private synthetic data," in *Theory of Cryptography Conference*. Springer, 2011, pp. 400–416.
[5] J. Ullman, "Answering $n^2 + o(1)$ counting queries with differential privacy is hard," *SIAM Journal on Computing*, vol. 45, no. 2, pp. 473–496, 2016.
[6] T. Liu, G. Vietri, T. Steinke, J. Ullman, and S. Wu, "Leveraging public data for practical private query release," in *International Conference on Machine Learning*. PMLR, July 2021, pp. 6968–6977.
[7] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 1961, pp. 547–561.
[8] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2013.
[9] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
[10] G. Barthe and F. Olmedo, "Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2013, pp. 49–60.
[11] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "Privacy under hard distortion constraints," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, 2018, pp. 1–5.
[12] K. Chaudhuri, J. Imola, and A. Machanavajjhala, "Capacity bounded differential privacy," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[13] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.
[14] N. Megiddo, *Progress in Mathematical Programming: Interior-Point and Related Methods*. Springer Science & Business Media, 2012.
[15] 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.
[16] 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.
[17] M. Hardt, K. Ligett, and F. 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.
[18] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. 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*, 2007, pp. 273–282.
[19] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proceedings of the 42nd ACM symposium on Theory of computing, STOC '10*, New York, NY, USA, 2010, pp. 705–714.
[20] A. Nikolov, K. Talwar, and L. Zhang, "The geometry of differential privacy: the sparse and approximate cases," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 2013, pp. 351–360.
[21] S. Aydore, W. Brown, M. Kearns, K. Kenthapadi, L. Melis, A. Roth, and A. Siva, "Differentially private query release through adaptive projection," 2021.
[22] C. Dwork, A. Nikolov, and K. Talwar, "Efficient algorithms for privately releasing marginals via convex relaxations," *Discrete & Computational Geometry*, vol. 53, no. 3, pp. 650–673, 2015.

[23] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 182–201, 2018.

[24] T. T. Cai, Y. Wang, and L. Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," *The Annals of Statistics*, 2020, to appear.

[25] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 510–526.

[26] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[27] M. Boedihardjo, T. Strohmer, and R. Vershyin, "Covariance's Loss is Privacy's Gain: Computationally Efficient, Private and Accurate Synthetic Data," *Foundations of Computational Mathematics*, to appear.

[28] P.-H. Lu and C.-M. Yu, "Poster: A unified framework of differentially private synthetic data release with generative adversarial network," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 2547–2549.

[29] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[30] F. Zhu, F. Ye, Y. Fu, Q. Liu, and B. Shen, "Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[31] A. M. Delaney, E. Brophy, and T. E. Ward, "Synthesis of realistic ECG using generative adversarial networks," *arXiv preprint arXiv:1909.09150*, 2019.

[32] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.

[33] J. Jordon, J. Yoon, and M. Van Der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations*, 2018.

[34] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, p. e005122, 2019.

[35] R. Vershynin, *High-dimensional probability. An introduction with applications in data science*. Cambridge University Press, 2018.

[36] D. Wang and J. Xu, "On sparse linear regression in the local differential privacy model," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 1182–1200, 2021.